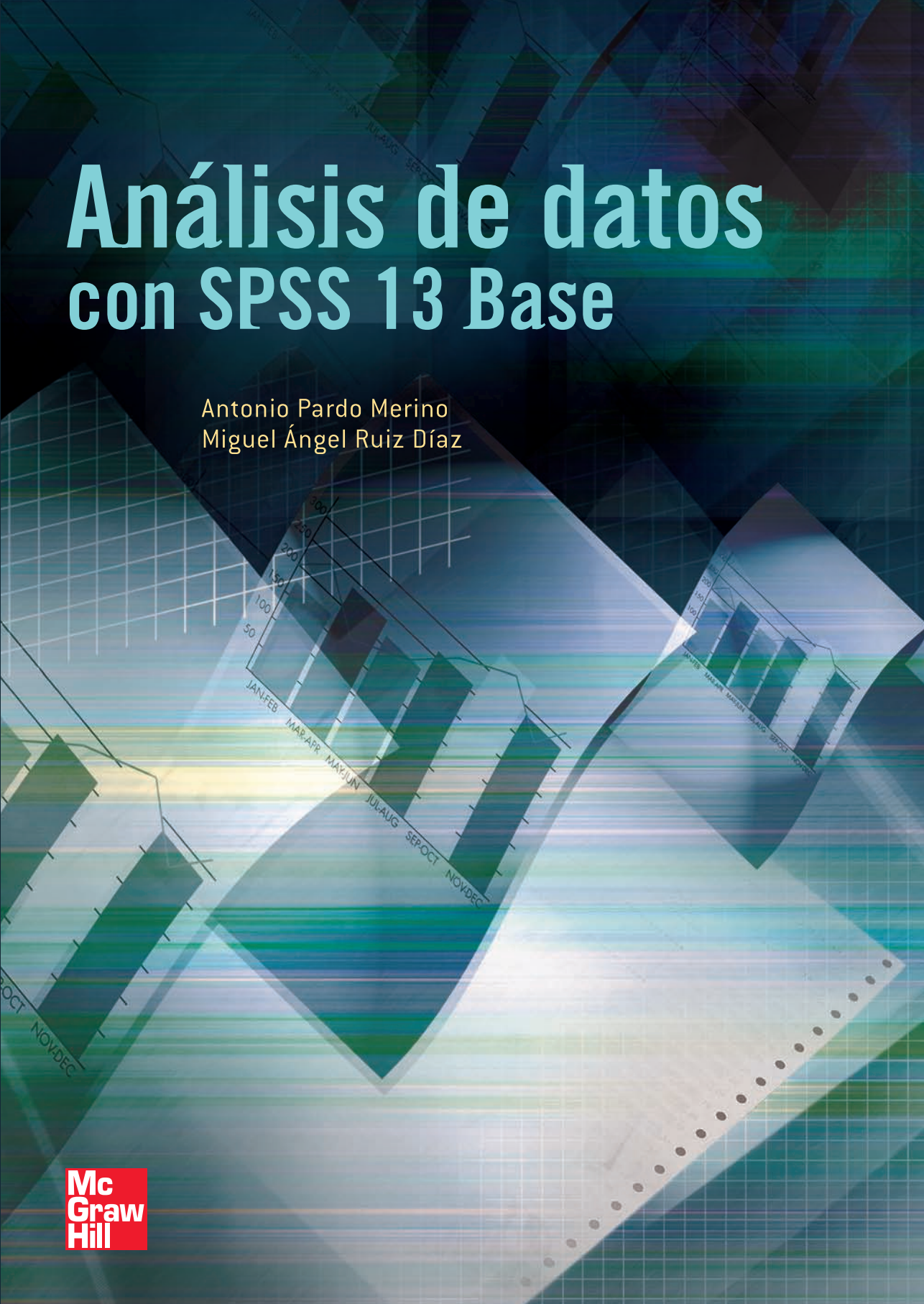


# Análisis de datos con SPSS 13 Base

Antonio Pardo Merino  
Miguel Ángel Ruiz Díaz



# **Análisis de datos con SPSS 13 Base**



# **Análisis de datos con SPSS 13 Base**

**Antonio Pardo**  
**Miguel Ángel Ruiz**

Profesores titulares de Metodología de las Ciencias del Comportamiento  
Universidad Autónoma de Madrid



**MADRID • BOGOTÁ • BUENOS AIRES • CARACAS • GUATEMALA • LISBOA  
MÉXICO • NUEVA YORK • PANAMÁ • SAN JUAN • SANTIAGO • SÃO PAULO  
AUKLAND • HAMBURGO • LONDRES • MILÁN • MONTREAL • NUEVA DELHI • PARÍS  
SAN FRANCISCO • SIDNEY • SINGAPUR • ST. LOUIS • TOKIO • TORONTO**

La información contenida en este libro procede de una obra original entregada por los autores. No obstante, McGraw-Hill no garantiza la exactitud o perfección de la información publicada. Tampoco asume ningún tipo de garantía sobre los contenidos y las opiniones vertidas en dichos textos.

Este trabajo se publica con el reconocimiento expreso de que se está proporcionando una información, pero no tratando de prestar ningún tipo de servicio profesional o técnico. Los procedimientos y la información que se presentan en este libro tienen sólo la intención de servir como guía general.

McGraw-Hill ha solicitado los permisos oportunos para la realización y el desarrollo de esta obra.

## ANÁLISIS DE DATOS CON SPSS 13 BASE

No está permitida la reproducción total o parcial de este libro, ni su tratamiento informático, ni la transmisión de ninguna forma o por cualquier medio, ya sea electrónico, mecánico, por fotocopia, por registro u otros métodos, sin el permiso previo y por escrito de los titulares del Copyright.



**McGraw-Hill / Interamericana  
de España S. A. U.**

DERECHOS RESERVADOS © 2005, respecto a la primera edición en español  
por MCGRAW-HILL/INTERAMERICANA DE ESPAÑA, S. A. U.

Edificio Valrealty, 1ª planta  
Basauri, 17  
28023 Aravaca (Madrid)

<http://www.mcgraw-hill.es>  
[universidad@mcgraw-hill.com](mailto:universidad@mcgraw-hill.com)

La marca SPSS® es propiedad de SPSS, Inc. Los cuadros de diálogo, las tablas de resultados y los gráficos se han capturado de la versión 13.0 con permiso de SPSS, Inc.  
<http://www.spss.com> y <http://www.spss.com/iberica>

ISBN: 84-481-4536-4  
Depósito legal:

Editor: Carmelo Sánchez González  
Diseño de cubierta: Luis Sanz Cantero  
Compuesto por: Antonio Pardo  
Impreso por:

IMPRESO EN ESPAÑA - PRINTED IN SPAIN

*A María y Alberto*  
*A Leticia, Rocío y Macarena*

## Otros títulos de interés relacionados



1. 84-481-3670-5 – LEÓN y MONTERO – Métodos de investigación en psicología y educación (3ª Edición)

Tercera edición de la obra anteriormente titulada "Diseño de investigaciones". Incluye varios capítulos nuevos que tratan las bases de la investigación, las metodologías cualitativas y los diseños con los mismos sujetos. Todos los recuadros y ejemplos que aparecen en la obra han sido revisados con respecto a la anterior edición. Incluye referencias a direcciones de Internet para conseguir información sobre métodos de investigación en psicología y educación. Es un excelente complemento de los manuales sobre análisis de datos.



2. 84-481-9825-5 – BERGANZA y RUIZ – Investigar en comunicación

Se trata de un manual práctico para las asignaturas obligatorias de Métodos de Investigación en Comunicación, Investigación de Audiencias e Investigación de Medios de las Facultades de Periodismo, Comunicación Audiovisual y Publicidad y para la materia Análisis del Entorno Social (troncal en las licenciaturas de estas mismas Facultades). También puede emplearse como manual para los alumnos de Doctorado de las Facultades de Comunicación. Su carácter eminentemente práctico (explicaciones con abundantes ejemplos y casos prácticos) lo hacen particularmente útil para estudiantes de Comunicación, profesionales de la investigación y consultores en Comunicación.



3. 84-481-379-14 – CORBETTA – Metodología y técnicas de investigación social

A partir de su larga experiencia docente y su continua investigación en el ámbito de las Ciencias Sociales, el autor ha elaborado un manual con una gran claridad expositiva y una gran riqueza de referencias en el campo de la investigación social. El texto se caracteriza por una amplia oferta de herramientas, sin dejar de considerar la dimensión metodológica en su sentido estricto: el discurso crítico sobre los principios lógicos, las condiciones y las normas fundamentales de la investigación científica. Se hace hincapié en las técnicas, esto es, en todos aquellos procedimientos elaborados, codificados y practicados por los investigadores, que son susceptibles de ser transmitidos didácticamente.

# Índice de contenidos

Presentación .....	XIX
--------------------	-----

## Primera parte Introducción al SPSS

### 1. ESTRUCTURA DEL SPSS

Tipos de ventanas SPSS .....	3
El <i>Editor de datos</i> .....	3
El <i>Visor de resultados</i> .....	4
El <i>Editor de sintaxis</i> .....	5
El <i>Editor de procesos</i> .....	5
Ventana <i>designada</i> y ventana <i>activa</i> .....	5
Cuadros de diálogo .....	6
Subcuadros de diálogo .....	9
Barra de menús .....	9
Menús .....	10
El <i>Editor de menús</i> .....	11
Ejemplo: Utilidades > <i>Editor de menús</i> .....	13
Barra de herramientas .....	15
Cómo personalizar una barra de herramientas .....	16
Barra de estado .....	21
Una sesión con el SPSS .....	22
Abrir un archivo de datos .....	22
Utilizar un procedimiento estadístico .....	23
Examinar los resultados .....	24

### 2. CÓMO UTILIZAR LA AYUDA

La ayuda por temas .....	27
Contenido .....	28
Índice .....	29
Búsqueda .....	30
El <i>Tutorial</i> .....	31
El <i>Asesor estadístico</i> .....	31
La <i>Guía de sintaxis</i> .....	32
La ayuda contextual .....	32
El <i>Asesor de resultados</i> .....	33
Estudios de casos .....	33
Los botones de ayuda .....	34



**3. ARCHIVOS DE DATOS**

Archivos nuevos .....	35
Abrir archivos de datos .....	35
Abrir base de datos .....	38
Leer datos de texto .....	45
Guardar archivos de datos .....	50
Guardar como .....	53
Marcar archivos como de sólo lectura .....	53
Mostrar información de datos .....	53
Hacer una <i>caché</i> de datos .....	54
Detener procesador SPSS .....	55
Presentación preliminar .....	55
Imprimir archivos de datos .....	55
Datos/archivos usados recientemente .....	58
Salir del SPSS .....	58

**4. EL EDITOR DE DATOS**

Definir variables .....	60
Asignar nombre a una variable .....	61
Definir el tipo de variable .....	61
Asignar etiquetas .....	64
Definir valores perdidos .....	65
Definir el formato de columna .....	66
Alinear texto .....	66
Asignar un nivel de medida .....	67
Definir variables de forma automática .....	67
Copiar propiedades de datos .....	71
Definir fechas .....	72
Entrar datos .....	73
Editar datos .....	74
Deshacer/rehacer .....	74
Seleccionar datos .....	75
Mover y copiar datos .....	75
Borrar datos .....	76
Buscar datos .....	76
Buscar casos .....	77
Buscar variables .....	77
Insertar variables nuevas .....	78
Insertar casos nuevos .....	78
Modificar el aspecto del <i>Editor de datos</i> .....	79
Adosar comentarios al archivo de datos .....	80
Trabajar con conjuntos de variables .....	81
Definir conjuntos de variables .....	81
Usar conjuntos de variables .....	82

## 5. TRANSFORMAR DATOS

Calcular .....	85
Variable de destino .....	86
Tipo de variable y etiqueta .....	87
Expresión numérica .....	87
Calculadora .....	87
Funciones .....	88
Expresiones condicionales .....	90
Ejemplo: Calcular > Si .....	91
Recodificar .....	92
Recodificar en las mismas variables .....	92
Recodificar en distintas variables .....	94
Categorizar variables .....	96
Categorizador visual .....	97
Definir categorías manualmente .....	99
Definir categorías automáticamente .....	100
Contar apariciones .....	102
Asignar rangos .....	105
Tipos de rangos .....	106
Rangos empatados .....	108
Recodificación automática .....	109
Operaciones con fechas y horas .....	111
Crear serie temporal .....	115
Funciones .....	116
Reemplazar valores perdidos .....	118
Métodos de estimación .....	119
Generadores de números aleatorios .....	119
Ejecutar transformaciones pendientes .....	121

## 6. MODIFICAR ARCHIVOS DE DATOS

Ordenar casos .....	123
Transponer archivos .....	124
Reestructurar el archivo de datos .....	125
Convertir variables en casos .....	126
Convertir casos en variables .....	132
Fundir archivos .....	135
Añadir casos .....	135
Añadir variables .....	138
Agregar datos .....	140
Identificar casos duplicados .....	143
Diseño ortogonal .....	145
Generar un diseño ortogonal .....	147
Mostrar un diseño ortogonal .....	150
Segmentar archivo .....	152
Seleccionar casos .....	153
Ponderar casos .....	156

## 7. ARCHIVOS DE RESULTADOS: EL VISOR

El <i>Visor</i> de resultados . . . . .	161
El menú <i>Archivo</i> . . . . .	162
Editar resultados . . . . .	163
Seleccionar resultados . . . . .	163
Mover, copiar y borrar resultados . . . . .	164
Cambiar de nivel un titular . . . . .	164
Mostrar y ocultar resultados . . . . .	165
Tamaño y fuente de los titulares . . . . .	166
Saltos de página . . . . .	166
Insertar texto y gráficos . . . . .	167
Alinear resultados . . . . .	169
Editar tablas . . . . .	169
La barra de herramientas del <i>Editor de tablas</i> . . . . .	170
Seleccionar . . . . .	170
Agrupar y desagrupar casillas . . . . .	171
Mostrar y ocultar casillas . . . . .	171
Modificar y añadir texto . . . . .	172
Pivotar tablas . . . . .	173
Paneles de pivotado . . . . .	174
Señalizadores . . . . .	176
Modificar las propiedades de una tabla . . . . .	177
General . . . . .	178
Notas al pie . . . . .	179
Formatos de casilla . . . . .	180
Bordes . . . . .	181
Impresión . . . . .	181
Modificar las propiedades de una casilla . . . . .	183
Valor . . . . .	183
Alineación . . . . .	184
Márgenes . . . . .	185
Sombreado . . . . .	185
Seleccionar el aspecto de una tabla . . . . .	186
Características del texto . . . . .	187
Anchura de las casillas . . . . .	188
Imprimir resultados . . . . .	189
Preparar página . . . . .	189
Controlar la ruptura de las tablas grandes . . . . .	191
Presentación preliminar . . . . .	191
Imprimir . . . . .	192
Copiar resultados en otras aplicaciones . . . . .	193
Copiar texto y gráficos . . . . .	193
Copiar tablas . . . . .	193
Copiar más de un objeto . . . . .	194
Incrustar tablas . . . . .	194
Exportar resultados . . . . .	194

**8. ARCHIVOS DE SINTAXIS**

Abrir y guardar archivos de sintaxis .....	200
Generar sintaxis SPSS .....	201
El botón <i>Pegar</i> de los cuadros de diálogo .....	201
Las anotaciones de los archivos de resultados .....	202
El archivo <i>spss.jnl</i> .....	203
Ejecutar sintaxis .....	204
Algunas reglas sintácticas básicas .....	204

## Segunda parte

# Análisis estadístico con el SPSS

**9. INTRODUCCIÓN AL ANÁLISIS ESTADÍSTICO**

Qué es el análisis estadístico .....	207
Para qué sirve el análisis estadístico .....	208
Tipos de variables .....	209
Conceptos básicos .....	212
Población .....	212
Muestra .....	213
Parámetro .....	213
Estadístico .....	214
Muestreo .....	215
Distribución muestral .....	217
La inferencia estadística .....	221
El contraste de hipótesis .....	221
La estimación de parámetros .....	231

**10. ANÁLISIS DESCRIPTIVO****Los procedimientos *Frecuencias*, *Descriptivos*, *Razón* y *Cubos OLAP***

Frecuencias .....	233
Ejemplo: Frecuencias .....	234
Estadísticos .....	235
Ejemplo: Frecuencias > Estadísticos .....	238
Gráficos .....	239
Ejemplo: Frecuencias > Gráficos .....	240
Formato .....	242
Descriptivos .....	243
Opciones .....	244
Ejemplo: Descriptivos > Opciones .....	246
Puntuaciones típicas y curva normal .....	247
Descriptivos para el cociente entre dos variables ( <i>Razón</i> ) .....	249
Estadísticos .....	251
Ejemplo: <i>Razón</i> .....	253

Cubos OLAP .....	255
Estadísticos .....	256
Diferencias .....	256
Encabezados y pies de tabla .....	258
Ejemplo: Cubos OLAP .....	258

## 11. ANÁLISIS EXPLORATORIO

### El procedimiento *Explorar*

El procedimiento <i>Explorar</i> .....	261
Estadísticos .....	262
Ejemplo: Explorar > Estadísticos .....	264
Gráficos .....	266
Diagramas de caja .....	267
Diagramas descriptivos .....	270
Cómo contrastar supuestos .....	272
Opciones .....	278

## 12. ANÁLISIS DE VARIABLES CATEGÓRICAS

### El procedimiento *Tablas de contingencias*

Tablas de contingencias .....	279
Ejemplo: Tablas de contingencias .....	281
Tablas segmentadas .....	282
Ejemplo: Tablas segmentadas .....	282
Estadísticos .....	283
<i>Chi</i> -cuadrado .....	284
Ejemplo: Tablas de contingencias > Estadísticos > <i>Chi</i> -cuadrado .....	285
Correlaciones .....	286
Datos nominales .....	287
Ejemplo: Tablas de contingencias > Estadísticos > Datos nominales .....	290
Datos ordinales .....	291
Ejemplo: Tablas de contingencias > Estadísticos > Datos ordinales .....	292
Nominal por intervalo .....	294
Índice de acuerdo (kappa) .....	294
Ejemplo: Tablas de contingencias > Estadísticos > Kappa .....	295
Índices de riesgo .....	296
Diseños prospectivos o de cohortes (hacia adelante) .....	297
Diseños retrospectivos o de caso-control (hacia atrás) .....	298
Ejemplo: Tablas de contingencias > Estadísticos > Riesgo .....	299
Proporciones relacionadas (prueba de McNemar) .....	301
Ejemplo: Tablas de contingencias > Estadísticos > McNemar .....	303
Combinación de tablas 2×2 (Cochran y Mantel-Haenszel) .....	306
Ejemplo: Tablas de cont. > Estadísticos > Cochran y Mantel-Haenszel .....	307
Contenido de las casillas .....	309
Ejemplo: Tablas de cont. > Casillas > Frecuencias, porcentajes y residuos .....	311
Formato de las casillas .....	313

**13. CONTRASTES SOBRE MEDIAS****Los procedimientos *Medias y Prueba T***

Medias .....	315
Opciones .....	316
Ejemplo: Comparar medias > Medias .....	317
Prueba <i>T</i> para una muestra .....	319
Opciones .....	320
Ejemplo: Comparar medias > Prueba <i>T</i> para una muestra .....	321
Prueba <i>T</i> para muestras independientes .....	322
Opciones .....	325
Ejemplo: Comparar medias > Prueba <i>T</i> para muestras independientes .....	326
Prueba <i>T</i> para muestras relacionadas .....	328
Opciones .....	330
Ejemplo: Comparar medias > Prueba <i>T</i> para muestras relacionadas .....	330

**14. ANÁLISIS DE VARIANZA (I)****El procedimiento *ANOVA de un factor***

El modelo lineal general .....	333
Introducción al análisis de varianza .....	335
Modelos de ANOVA .....	335
Número de factores .....	336
Tipo de aleatorización .....	337
Muestreo de niveles .....	338
Lógica del NOVA .....	339
ANOVA de un factor .....	342
Ejemplo: ANOVA de un factor .....	344
Opciones .....	346
Ejemplo: ANOVA de un factor > Opciones .....	347
Comparaciones <i>post hoc</i> o <i>a posteriori</i> .....	349
Ejemplo: ANOVA de un factor > Comparaciones <i>post hoc</i> .....	352
Comparaciones planeadas o <i>a priori</i> .....	354
Ejemplo: ANOVA de un factor > Contrastes polinómicos .....	356
Ejemplo: ANOVA de un factor > Contrastes personalizados .....	358

**15. ANÁLISIS DE VARIANZA (II)****El procedimiento *Modelo lineal general: Univariante***

Análisis de varianza factorial .....	361
Ejemplo: MLG > Univariante .....	363
Comparaciones <i>post hoc</i> o <i>a posteriori</i> .....	366
Ejemplo: MLG > Univariante > Comparaciones <i>post hoc</i> .....	367
Gráficos de perfil para la interacción .....	368
Ejemplo: MLG > Univariante > Gráficos de perfil .....	369
Análisis de covarianza .....	370
Ejemplo: MLG > Univariante > Covariables .....	372
Opciones .....	373

Contrastes personalizados .....	379
Contrastes predefinidos .....	379
La sentencia LMATRIX .....	380
Modelos personalizados .....	386
Modelos con bloques aleatorios .....	387
Modelos jerárquicos o anidados .....	388
Análisis de regresión lineal .....	389
Homogeneidad de las pendientes de regresión .....	390
Guardar pronósticos y residuos .....	391

## 16. ANÁLISIS DE VARIANZA (III)

### El procedimiento *Modelo lineal general: Medidas repetidas*

Medidas repetidas .....	395
Modelo de un factor .....	396
Datos .....	396
Análisis básico .....	397
Ejemplo: MLG > ANOVA de un factor con medidas repetidas .....	399
Aspectos complementarios .....	401
Más de una variable dependiente .....	408
Modelo de dos factores, ambos con medidas repetidas .....	408
Datos .....	409
Análisis básico .....	410
Ejemplo: MLG > ANOVA de dos factores, ambos con medidas repetidas ...	412
Aspectos complementarios del análisis .....	414
Modelo de dos factores con medidas repetidas en un factor .....	418
Datos .....	418
Análisis básico .....	419
Ejemplo: MLG > ANOVA de dos factores con medidas repetidas en un factor .....	420
Aspectos complementarios del análisis .....	422

## 17. RELACIÓN ENTRE VARIABLES

### El procedimiento *Correlaciones*

Correlaciones bivariadas .....	429
Opciones .....	434
Ejemplo: Correlaciones > Bivariadas .....	435
Correlaciones parciales .....	437
Opciones .....	439
Ejemplo: Correlaciones > Parciales .....	440
Distancias .....	442
Medidas de similaridad .....	444
Medidas de disimilaridad .....	449
Transformación de los valores .....	452
Transformación de las medidas .....	453

**18. ANÁLISIS DE REGRESIÓN LINEAL****El procedimiento *Regresión lineal***

La recta de regresión .....	455
La mejor recta de regresión .....	457
Bondad de ajuste .....	458
Análisis de regresión lineal simple .....	459
Bondad de ajuste .....	460
Ecuación de regresión .....	462
Coeficientes de regresión tipificados .....	462
Pruebas de significación .....	463
Análisis de regresión lineal múltiple .....	463
Bondad de ajuste .....	465
Ecuación de regresión .....	466
Coeficientes de regresión tipificados .....	466
Pruebas de significación .....	467
Información complementaria .....	467
Supuestos del modelo de regresión lineal .....	471
Análisis de los residuos .....	471
Independencia .....	473
Homocedasticidad .....	474
Normalidad .....	477
Linealidad .....	478
Colinealidad .....	479
Puntos de influencia .....	482
Análisis de regresión por pasos (regresión <i>stepwise</i> ) .....	485
Criterios de selección de variables .....	485
Métodos de selección de variables .....	487
Regresión por pasos .....	488
Qué variables debe incluir la ecuación de regresión .....	492
Cómo efectuar pronósticos .....	493
Validez del modelo de regresión .....	496

**19. ANÁLISIS DE REGRESIÓN CURVILÍNEA****El procedimiento *Estimación curvilínea***

Estimación curvilínea .....	498
Pronósticos y residuos .....	500
Ejemplo: Estimación curvilínea .....	501

**20. FIABILIDAD DE LAS ESCALAS****El procedimiento *Análisis de fiabilidad***

Concepto de fiabilidad .....	508
Análisis de fiabilidad .....	509
Ejemplo: Análisis de fiabilidad .....	511



Modelos de fiabilidad .....	512
Modelo <i>alfa</i> .....	512
Ejemplo: Análisis de fiabilidad > Modelo <i>alfa</i> .....	513
Modelo de dos mitades .....	514
Ejemplo: Análisis de fiabilidad > Modelo de dos mitades .....	515
Modelo de Guttman .....	516
Ejemplo: Análisis de fiabilidad > Modelo de Guttman .....	517
Modelo de medidas paralelas .....	517
Ejemplo: Análisis de fiabilidad > Modelo de medidas paralelas .....	518
Estadísticos .....	519
Descriptivos .....	520
Ejemplo: Análisis de fiabilidad > Estadísticos > Descriptivos .....	520
Resúmenes .....	522
Ejemplo: Análisis de fiabilidad > Estadísticos > Resúmenes .....	522
Entre elementos .....	523
Ejemplo: Análisis de fiabilidad > Estadísticos > Entre elementos .....	523
Tabla de ANOVA .....	524
Ejemplo: Análisis de fiabilidad > Estadísticos > Tabla de ANOVA .....	526
Prueba $T^2$ de Hotelling .....	526
Ejemplo: Análisis de fiabilidad > Estadísticos > $T^2$ de Hotelling .....	527
Prueba de aditividad de Tukey .....	528
Ejemplo: Análisis de fiabilidad > Estadísticos > Prueba de aditividad de Tukey. ....	528
Coeficiente de correlación intraclase .....	529
Ejemplo: Análisis de fiabilidad > Estadísticos > Coeficiente de correlación intraclase .....	532

## 21. ANÁLISIS NO PARAMÉTRICO

### El procedimiento *Pruebas no paramétricas*

Pruebas para una muestra .....	534
Prueba <i>chi</i> -cuadrado para una muestra .....	534
Ejemplo: Pruebas no paramétricas > <i>Chi</i> -cuadrado .....	537
Prueba binomial .....	538
Ejemplo: Pruebas no paramétricas > Binomial .....	540
Prueba de las rachas .....	541
Ejemplo: Pruebas no paramétricas > Rachas .....	543
Prueba de Kolmogorov-Smirnov para una muestra .....	544
Ejemplo: Pruebas no paramétricas > Kolmogorov-Smirnov .....	545
Pruebas para dos muestras independientes .....	546
Prueba <i>U</i> de Mann-Whitney .....	547
Prueba de reacciones extremas de Moses .....	549
Prueba de Kolmogorov-Smirnov para dos muestras .....	550
Prueba de las rachas de Wald-Wolfowitz .....	551
Ejemplo: Pruebas no paramétricas > Dos muestras independientes .....	552

Pruebas para varias muestras independientes .....	554
Prueba $H$ de Kruskal-Wallis .....	555
Prueba de la mediana .....	556
Ejemplo: Pruebas no paramétricas > Varias muestras independientes .....	557
Pruebas para dos muestras relacionadas .....	558
Prueba de Wilcoxon .....	559
Prueba de los signos .....	560
Ejemplo: Pruebas no paramétricas > Dos muestras relacionadas .....	561
Pruebas para varias muestras relacionadas .....	563
Prueba de Friedman .....	564
Coeficiente de concordancia $W$ de Kendall .....	565
Prueba de Cochran .....	567
Ejemplo: Pruebas no paramétricas > Varias muestras relacionadas .....	567
 <b>22. ANÁLISIS DE VARIABLES DE RESPUESTA MÚLTIPLE</b>	
<b>El procedimiento <i>Respuestas múltiples</i></b>	
Variables de respuesta múltiple .....	573
Definir conjuntos de respuestas múltiples .....	576
Ejemplo: Respuestas múltiples > Definir conjuntos de resp. múltiples .....	577
Tablas de frecuencias .....	578
Ejemplo: Respuestas múltiples > Frecuencias .....	579
Tablas de contingencias .....	580
Ejemplo: Respuestas múltiples > Tablas de contingencias .....	583
 <b>REFERENCIAS BIBLIOGRÁFICAS .....</b>	<b>585</b>
 <b>ÍNDICE DE MATERIAS .....</b>	<b>591</b>



## Presentación

En 2002 publicamos una *Guía para el análisis de datos* basada en la versión 11 del SPSS. Las rápidas actualizaciones a las que se ve sometido el programa, por un lado, y la insistencia de muchos de nuestros lectores en solicitar la incorporación de nuevos procedimientos, por otro, nos ha animado a ofrecer esta nueva guía basada en la versión 13.

Al igual que antes, no se trata de un material diseñado exclusivamente para prestar ayuda al usuario más básico en el manejo de aplicaciones informáticas y en el conocimiento de herramientas estadísticas, sino para servir de ayuda también al usuario más avanzado. Todo ello, sin embargo, prestando más atención a los aspectos prácticos o aplicados que a los teóricos o formales, aunque sin descuidar estos últimos. Por tanto, el propósito de este material es doble: pretende servir de apoyo al usuario más novato en el manejo del paquete estadístico SPSS en su versión para Windows y, al mismo tiempo, ayudar al usuario más experimentado a comprender e interpretar los detalles asociados a cada procedimiento SPSS.

Esta nueva guía se presenta en dos volúmenes. El contenido de ambos viene determinado por los procedimientos que el SPSS incluye en sus módulos *Base* y *Avanzado*, que son los de mayor difusión tanto en el ámbito académico como en el profesional. Al primer volumen lo hemos llamado *Análisis de datos con SPSS 13 Base*; al segundo, *Análisis de datos con SPSS 13 Avanzado*. No obstante, el contenido de ambos volúmenes no se ajusta exactamente a la división establecida en el SPSS entre sus módulos *Base* y *Avanzado*, sino que se basa en la necesidad de hacer de ellos dos volúmenes relativamente independientes por los motivos que se explican a continuación.

El primer volumen (*Base*) contiene dos partes bien diferenciadas. La primera incluye, fundamentalmente, los procedimientos que permiten utilizar el SPSS como *gestor de datos*: describe cómo construir o importar un archivo de datos y cómo preparar los datos para el análisis, y contiene una descripción pormenorizada de las tres ventanas principales del SPSS: el *Editor de datos*, el *Visor de resultados* y el *Editor de sintaxis*. La segunda parte incluye varios procedimientos que permiten utilizar el SPSS como *programa de análisis estadístico*. En esta segunda parte se incluyen las técnicas de análisis de datos más comúnmente ofertadas en los planes de estudio de las licenciaturas en las que se utiliza la estadística como herramienta de apoyo: estadística descriptiva, análisis exploratorio, contrastes sobre medias, análisis de varianza, análisis de correlación y regresión, estadística no paramétrica y fiabilidad de las escalas (exceptuando los modelos de análisis de varianza de medidas repetidas, todos los procedimientos incluidos en este volumen se encuentran en el módulo *Base* del SPSS).

El segundo volumen (*Avanzado*) recoge los procedimientos estadísticos que, en parte por ser algo más refinados o menos básicos, en parte por estar diseñados para tratar un mayor número de variables, se incluyen en asignaturas optativas o cursos de posgrado: análisis de conglomerados, análisis factorial, escalamiento multidimensional, análisis discriminante (módulo *Base*), modelos lineales mixtos, análisis de regresión ordinal, análisis de regresión logística, análisis de supervivencia y modelos loglineales (módulo *Avanzado*).

Los procedimientos SPSS se explican utilizando archivos de datos que el lector puede consultar y descargar, si lo desea, en la dirección [http://www.mhe.es/pardo\\_spss](http://www.mhe.es/pardo_spss). Siempre que se hace referencia a esta dirección a lo largo del texto se utiliza la expresión *página web del manual*.

Todas las técnicas de análisis están descritas intentando ajustar la exposición a la estructura de los cuadros de diálogo del SPSS; por tanto, al igual que ocurría en la *Guía para el análisis de datos* de la versión 11, este nuevo manual no puede considerarse un libro convencional de análisis estadístico, sino un libro de análisis estadístico con el SPSS.

Conviene tener muy presente que, aunque las herramientas informáticas pueden realizar cálculos con suma facilidad, todavía no están capacitadas para tomar algunas decisiones. Un programa estadístico no aclara si el diseño aplicado es correcto, o si las medidas utilizadas son apropiadas; tampoco decide qué prueba estadística conviene aplicar en cada caso, ni interpreta los resultados del análisis. Los programas estadísticos todavía no permiten prescindir del analista de datos. Es él, el analista, quien debe mantener el control de todo el proceso. El éxito de un análisis depende de él y no de la máquina o del programa informático. El hecho de que sea posible ejecutar las técnicas de análisis más complejas con la simple acción de pulsar un botón sólo significa que es necesario haber atado bien todos los cabos del proceso (diseño, medida, análisis, etc.) antes de pulsar el botón.

Por último, queremos aprovechar la oportunidad que nos brinda esta presentación para agradecer a nuestro compañero Ludgerio Espinosa las aportaciones hechas para mejorar la exposición; y a muchos de nuestros alumnos y a no pocos lectores de la versión anterior, la ayuda prestada en la caza de erratas. Los errores y deficiencias que todavía permanezcan son, sin embargo, atribuibles sólo a nosotros.

Antonio Pardo  
Miguel Ángel Ruiz

Madrid, febrero de 2005

**Primera parte**

# **Introducción al SPSS**



# Estructura del SPSS

El SPSS (*Statistical Product and Service Solutions*) es una potente herramienta de tratamiento de datos y análisis estadístico. Al igual que el resto de aplicaciones que utilizan como soporte el sistema operativo Windows, el SPSS funciona mediante menús desplegables y cuadros de diálogo que permiten hacer la mayor parte del trabajo utilizando el puntero del ratón.

Al iniciar una sesión con el SPSS aparece una ventana de aspecto similar al de una hoja de cálculo: el *Editor de datos* (ver Figura 1.1). El *Editor de datos* es la ventana principal del SPSS, pero no la única. En los próximos capítulos se explican con detalle las diferentes ventanas SPSS, pero antes, en este capítulo, se ofrece una descripción general de todas ellas. Conocer las distintas ventanas del SPSS es, probablemente, la mejor manera de aproximarse por primera vez al programa y obtener una idea global sobre su estructura. También se estudiarán en este capítulo introductorio las *barras de menús*, las *barras de herramientas* y las *barras de estado*.

## Tipos de ventanas SPSS

Existen ocho tipos de ventanas SPSS, aunque no todas ellas poseen la misma importancia desde el punto de vista de su utilidad para el usuario.

Las dos ventanas principales (imprescindibles para trabajar con el SPSS) son el *Editor de datos*, que se estudia con detalle en los Capítulos 4, 5 y 6, y el *Visor de resultados*, que se estudia con detalle en el Capítulo 7).

Tres de las ocho ventanas (los editores de texto, tablas y gráficos) se encuentran *dentro* del propio *Visor de resultados*. Y otra de las ventanas no es más que una versión *borrador* del *Visor*.

Las dos ventanas restantes son el *Editor de sintaxis*, que se explica en el Capítulo 8 y cuya importancia se irá haciendo más evidente a medida que se vaya aprendiendo a manejar el SPSS, y el *Editor de procesos*, al que se le prestará poca atención aquí.

### El *Editor de datos*

Contiene el archivo de datos sobre el que se basa la mayor parte de las acciones que es posible llevar a cabo con el SPSS. El *Editor de datos* se abre automáticamente (vacío, sin datos; ver Figura 1.1) cuando se entra en el SPSS.



La ventana del *Editor de datos* puede mostrar dos contenidos diferentes: los **datos** propiamente dichos y las **variables** del archivo acompañadas del conjunto de características que las definen. Al igual que el resto de ventanas SPSS, el *Editor de datos* contiene una **barra de menús** (un conjunto de menús desplegables), una **barra de herramientas** (una serie de botones-íconos que facilitan el acceso rápido a muchas de las funciones SPSS) y una **barra de estado** (con información puntual sobre diferentes aspectos relacionados con el estado del programa). Es posible abrir más de un *Editor de datos* y, por tanto, trabajar con varios archivos de datos simultáneamente.

Figura 1.1. Ventana del *Editor de datos*



## El *Visor de resultados*

Recoge la información (tablas, gráficos, etc.) que el SPSS genera como consecuencia de las acciones que lleva a cabo. El *Visor* no sólo muestra los resultados del SPSS, sino que permite editarlos y guardarlos para su uso posterior.

Los resultados del *Visor* adoptan tres formatos distintos: *tablas*, *gráficos* y *texto*. El SPSS dispone de un editor (por tanto, de una ventana distinta) para cada uno de estos tres formatos básicos (todos ellos se describen con detalle en el Capítulo 7):

- El **Editor de tablas**. Ofrece múltiples posibilidades de edición de los resultados presentados en formato de *tabla pivotante* (un tipo de formato propio de SPSS).
- El **Editor de gráficos**. Permite modificar los colores, los tipos de letra, las etiquetas, la posición de los ejes y muchos otros detalles de los gráficos del *Visor*.
- El **Editor de texto**. Permite modificar los diferentes atributos (tipo, tamaño, color, etc., de las fuentes) de los resultados tipo *texto*: títulos, subtítulos y notas.

Una cuarta ventana corresponde al *Visor* en formato *borrador*: el **Borrador del Visor de resultados**. Ofrece la misma información que el *Visor*, pero en formato texto, es decir, con un aspecto menos depurado y sin las posibilidades de edición propias del *Visor* (no es posible, por ejemplo, pivotar tablas o editar gráficos).

Es posible tener abiertas simultáneamente varias ventanas del *Visor de resultados*. Es decir, es posible asociar a un mismo *Editor de datos* varias ventanas (por tanto, varios archivos) del *Visor de resultados*.

## El *Editor de sintaxis*

Permite utilizar las posibilidades de programación del SPSS. Las acciones que el SPSS lleva a cabo como resultado de las selecciones hechas en los menús y cuadros de diálogo se basan en un conjunto de sentencias construidas con una sintaxis propia del SPSS.

Estas sentencias en sintaxis SPSS (se puede abreviar diciendo simplemente *sintaxis SPSS*) pueden *pegarse* en una ventana de sintaxis desde cualquier cuadro de diálogo. El botón **Pegar** disponible en la mayor parte de los cuadros de diálogo siempre tiene el mismo efecto: convierte en sintaxis SPSS las selecciones hechas. La sintaxis SPSS *pegada* puede editarse para, por ejemplo, ejecutar algunas acciones no disponibles desde los cuadros de diálogo, o para salvarla en un archivo y volver a utilizarla en una sesión diferente. Es posible tener abiertas simultáneamente varias ventanas de sintaxis. Aunque el *Editor de sintaxis* no es imprescindible para trabajar con el SPSS, su capacidad para, entre otras cosas, automatizar trabajos repetitivos, lo convierte en una ventana de especial utilidad (ver Capítulo 8).

## El *Editor de procesos*


Un proceso (*script*) es una secuencia de sentencias que sirve para personalizar y automatizar algunas tareas SPSS, especialmente en lo relacionado con el contenido y el aspecto de las tablas de resultados. Esta ventana no será objeto de atención de este manual.

## Ventana designada y ventana activa

Según se ha señalado ya, el *Editor de datos* (ventana principal del SPSS) puede tener asociados más de un *Visor de resultados* y más de un *Editor de sintaxis*. Esto no es algo muy habitual, pero en ocasiones puede interesar. Cuando ocurre, es decir, cuando se tienen abiertas más de una ventana del *Visor de resultados* o del *Editor de sintaxis*, el SPSS debe saber a qué ventana del *Visor* debe enviar los resultados que produce y en qué ventana del *Editor de sintaxis* debe pegar la sintaxis cuando se le solicita que lo haga. Pues bien, los resultados y la sintaxis generados por el SPSS son automáticamente transferidos a las ventanas **designadas**. Estas ventanas se distinguen de las no designadas por la presencia de un signo de admiración rojo (!) en la *barra de estado*.

No debe confundirse la ventana designada con la ventana **activa**. La ventana activa es la que Windows selecciona como tal: es la ventana en la que se está trabajando (la que contiene el cursor). Si se está trabajando con el *Editor de datos*, la ventana activa es el *Editor de datos*; si se está editando un gráfico, la ventana activa es el *Editor de gráficos*; si se está editando una tabla, la ventana activa es el *Editor de tablas*; etc. Una ventana activa no tiene por qué ser la ventana designada. Si existen varias ventanas del *Visor* o del *Editor de sintaxis* abiertas, la ventana activa es aquella que aparece en primer plano (la que contiene el cursor), pero la ventana designada es o la última que se haya abierto (pues al abrir una ventana del *Visor* o

del *Editor de sintaxis* se convierte en ventana designada), o la que explícitamente se haya designado como tal. Si se desea que una ventana no designada se convierta en designada, deben darse instrucciones explícitas. Para convertir una ventana en *ventana designada*:

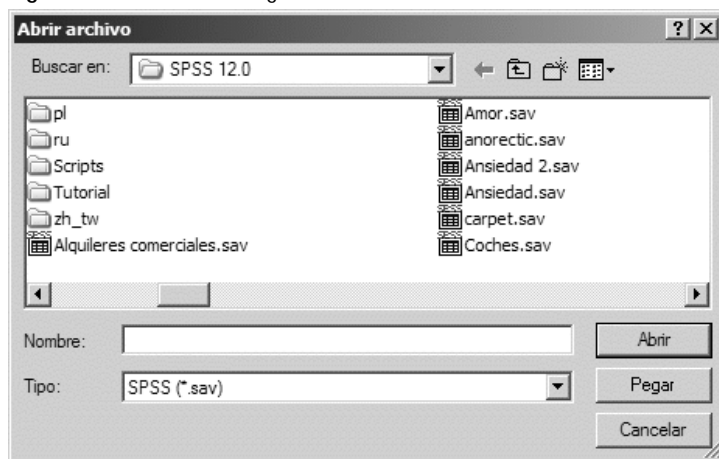
- Seleccionar la opción **Designar ventana** del menú **Utilidades**. Se obtiene el mismo resultado pulsando el botón *Designar ventana*  de la barra de herramientas.

## Cuadros de diálogo

Además de las ventanas SPSS ya descritas, existe también otro tipo de ventanas (propias de toda aplicación que funciona en el entorno Windows) llamadas **cuadros de diálogo**. Estas ventanas son las que el SPSS, al igual que otros programas informáticos, utiliza para que el usuario pueda llevar a cabo cualquier acción: son las ventanas que permiten al usuario *dialogar* con el SPSS para darle instrucciones sobre lo que desea hacer; permiten utilizar la mayoría de las funciones del SPSS simplemente utilizando el puntero del ratón

Algunos de estos cuadros de diálogo (particularmente los relacionados con abrir, guardar, imprimir y exportar archivos; o los relacionados con el establecimiento de las opciones que funcionan por defecto; etc.) mantienen un formato muy similar al del resto de aplicaciones que funcionan en el entorno Windows. No son, por tanto, cuadros de diálogo típicos del SPSS. La Figura 1.2 muestra, como ejemplo, un cuadro de diálogo de este tipo. Se accede a él cuando se intenta abrir un archivo (ver Capítulo 3).

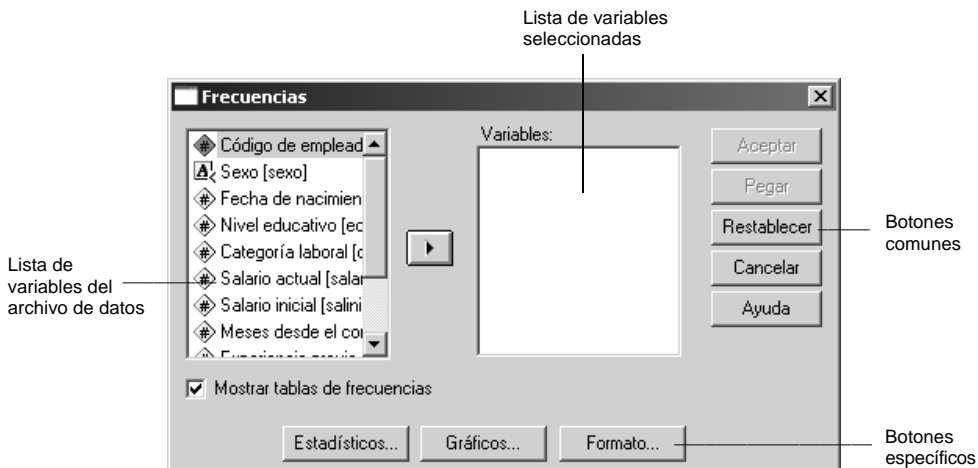
Figura 1.2. Cuadro de diálogo *Abrir archivo*



Pero la mayoría de los cuadros de diálogo, o si se prefiere, los cuadros de diálogo *típicos* del SPSS, son los que permiten realizar las tareas propias del SPSS: efectuar modificaciones en el archivo de datos, ejecutar procedimientos estadísticos, obtener gráficos, etc. Y todos estos cuadros de diálogo comparten una estructura peculiar. Al intentar, por ejemplo, ejecutar el procedimiento Frecuencias (dentro del menú **Analizar > Descriptivos**), el SPSS abre el cuadro

de diálogo *Frecuencias*, cuyos detalles muestra la Figura 1.3. Los cuadros de diálogo típicos del SPSS poseen esta misma estructura.


Figura 1.3. Contenido del cuadro de diálogo *Frecuencias*



- **Lista de variables del archivo de datos.** El primer recuadro del cuadro de diálogo ofrece un listado de todas las variables del archivo de datos: las variables numéricas van precedidas del símbolo «#»; las variables de cadena corta, del símbolo «A<sub>c</sub>»; y las de cadena larga, del símbolo «A<sub>l</sub>» (para entender la diferencia entre variables *numéricas* y de *cadena*, ver más adelante, en el Capítulo 4, el apartado *Definir variables*).

Este listado de variables puede mostrar el *nombre* de las variables o su *etiqueta*. Y las variables pueden aparecer en orden alfabético o en el orden en el que se encuentran en el *Editor de datos*. Ambos detalles pueden controlarse desde el menú Edición > Opciones..., en la pestaña General, dentro del recuadro Listas de variables.

Situando el puntero del ratón sobre el nombre o la etiqueta de cualquiera de las variables del listado y pulsando el botón secundario, puede obtenerse información adicional sobre esa variable: nivel de medida y etiquetas de los valores, si existen (para más detalles sobre esta ayuda en línea, consultar el Capítulo 2 sobre *Cómo utilizar la ayuda*).

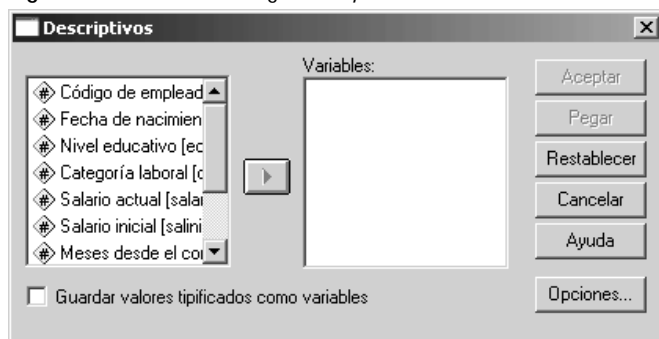
- **Lista de variables seleccionadas.** Lista (a veces más de una) a la que deben trasladarse las variables con las que se desea trabajar.
  - Para trasladar variables desde el listado de *variables del archivo* hasta el listado de *variables seleccionadas*, seleccionar con el puntero de ratón la variable que se desea trasladar y pulsar el botón flecha  situado entre ambos listados.
  - Para devolver al listado de *variables del archivo* una variable previamente seleccionada, seleccionar esa variable en el listado de *variables seleccionadas* y pulsar el botón flecha, el cual apunta ahora en la dirección contraria.

Cuando existe un único listado de *variables seleccionadas* (como ocurre, por ejemplo, en el cuadro de diálogo *Frecuencias* de la Figura 1.3), es posible desplazar va-

riables de un listado a otro simplemente pulsando dos veces el botón principal del ratón tras situar el puntero sobre la variable deseada. Y en algunos cuadros de diálogo (como en los correspondientes a los gráficos interactivos o a las propiedades de las variables y los datos, etc.) las variables se desplazan simplemente arrastrándolas tras seleccionarlas con el puntero del ratón.

- **Botones comunes.** Son botones que se encuentran en la mayoría de los cuadros de diálogo y siempre con el mismo significado:
  - **Aceptar.** Cierra el cuadro de diálogo y ejecuta el procedimiento seleccionado teniendo en cuenta las opciones y variables seleccionadas.
  - **Pegar.** Genera la sintaxis SPSS correspondiente a las selecciones hechas en un cuadro de diálogo y las pega en la ventana de sintaxis designada. Si no existe ninguna ventana de sintaxis abierta, el botón **Pegar** abre una y le asigna el nombre *Sintaxis#*. En el Capítulo 8 se tratan algunos aspectos relacionados con la sintaxis SPSS. Este botón cierra el cuadro de diálogo pero no ejecuta el procedimiento.
  - **Restablecer.** Limpia el listado de variables seleccionadas y cualquier otra opción marcada, es decir, restaura todas las opciones del cuadro de diálogo a sus valores originales (los valores por defecto). No cierra el cuadro de diálogo.
  - **Cancelar.** Cancela todos los cambios introducidos en el cuadro de diálogo desde la última vez que fue abierto y lo cierra. Es decir, cierra el cuadro de diálogo dejándolo como estaba antes de entrar en él.
  - **Ayuda.** Ofrece ayuda específica sobre los contenidos del cuadro de diálogo (ver *Cómo obtener ayuda* en el Capítulo 2).
- **Botones específicos.** Además de los botones comunes, existen botones específicos que van cambiando de un cuadro de diálogo a otro. Así, por ejemplo, en el cuadro de diálogo *Frecuencias* de la Figura 1.3, los botones específicos son **Estadísticos...**, **Gráficos...** y **Formato...** Pero si se abre otro cuadro de diálogo como, por ejemplo, **Descriptivos** (ver Figura 1.4), puede observarse que los botones específicos se limitan a uno: **Opciones...** Los botones comunes son ahora exactamente los mismos que antes, pero los botones específicos han cambiado. El número de botones específicos de un cuadro de diálogo depende básicamente de la complejidad del procedimiento.

Figura 1.4. Cuadro de diálogo *Descriptivos*



## Subcuadros de diálogo

Los botones específicos recién mencionados poseen la peculiaridad de ir acompañados de puntos suspensivos: Estadísticos..., Opciones... Esto sirve para recordar que se trata de botones que conducen a *subcuadros de diálogo* que están colgando del cuadro de diálogo principal. La Figura 1.5 muestra uno de estos subcuadros de diálogo.

Los subcuadros de diálogo permiten seguir seleccionando opciones no contenidas en el cuadro de diálogo principal hasta conseguir personalizar al máximo la ejecución de un determinado procedimiento. Estas opciones pueden ir precedidas de un cuadrado ( ), en cuyo caso se trata de opciones entre las que es posible seleccionar tantas como se quiera, incluso ninguna, o de un círculo ( ), en cuyo caso se trata de opciones exclusivas de las cuales sólo es posible seleccionar una.

Por lo general, los subcuadros de diálogo contienen tres botones: Continuar, Cancelar y Ayuda. Los dos últimos son idénticos a los de los cuadros de diálogo ya descritos en el apartado anterior. El botón Continuar permite volver al cuadro de diálogo principal tras marcar las opciones deseadas.

Figura 1.5. Subcuadro de diálogo *Descriptivos: Opciones*



## Barra de menús

Las barras de menús contienen una serie de menús desplegable que permiten controlar la mayoría de las acciones que el SPSS puede llevar a cabo. Se encuentran situadas en la parte superior de cada ventana, inmediatamente debajo del nombre de la ventana (ver, por ejemplo, la Figura 1.1). Cada tipo de ventana SPSS tiene su propia barra de menús, con opciones particulares para las funciones relacionadas con ella. Algunos de estos menús, como *Estadísticos*, *Gráficos* o *Ventana*, se repiten en todas las ventanas. Otros menús son específicos de un tipo particular de ventana. A continuación se describen todos ellos, indicando a qué ventanas pertenecen y en cuáles no están disponibles.

## Menús

**Archivo.** Desde este menú pueden crearse nuevos archivos de datos, resultados, sintaxis, etc.; abrir y guardar todo tipo de archivos; importar/exportar archivos desde/a otros programas (hojas de cálculo, bases de datos, procesadores de texto, etc.); imprimir archivos o partes de un archivo; obtener una vista previa del resultado de la impresión; recuperar archivos utilizados recientemente; controlar el servidor con el que se está trabajando; salir del programa; etc.

**Edición.** Permite editar (cortar, copiar, pegar, buscar, seleccionar, reemplazar, etc.) el contenido de un archivo; y deshacer y rehacer acciones de edición; este menú también ofrece la posibilidad de modificar algunas de las especificaciones iniciales (denominadas *opciones*) con las que arranca el programa.

**Ver.** Controla el aspecto de las distintas ventanas SPSS mediante una serie de opciones que permiten mostrar/ocultar la barra de estado, personalizar la barra de herramientas, seleccionar el tipo y tamaño de las fuentes utilizadas, etc. En el *Editor de datos*, este menú, además, permite controlar el aspecto de las celdas (con líneas o sin líneas) y mostrar/ocultar las etiquetas de los valores. En el *Visor de resultados*, sirve para mostrar/ocultar resultados concretos y para contraer/expandir bloques de resultados.

**Datos.** Contiene funciones propias del *Editor de datos*: fusionar archivos de datos, trasponer las filas y las columnas, seleccionar sólo una parte del archivo, dividir el archivo en subgrupos, insertar filas o columnas nuevas, etc.

**Transformar.** Las opciones de este menú permiten crear variables nuevas y cambiar los valores de las variables ya existentes poniendo a disposición del usuario una gran cantidad de funciones.

**Insertar.** Disponible en el *Visor de resultados*, en el *Editor de tablas* y en el *Editor de texto*. Contiene opciones para insertar texto, gráficos, títulos, encabezamientos, notas, saltos de página, objetos de otras aplicaciones, etc.

**Pivotar.** Sólo disponible en el *Editor de tablas*. Permite modificar la ubicación de las entradas (filas, columnas, capas) de las tablas de resultados.

**Galería.** Sólo disponible en el *Editor de gráficos*. Permite seleccionar diferentes tipos de gráficos para unos mismos datos.

**Diseño.** Sólo disponible en el *Editor de gráficos*. Contiene múltiples opciones para controlar las características de un gráfico: la escala y los rótulos de los ejes, los títulos y las leyendas, los tipos de letra, los colores, etc.

**Series.** Sólo disponible en el *Editor de gráficos*. Desde este menú es posible seleccionar las categorías que se desea mostrar/ocultar en el eje de abscisas. También es posible optar entre gráficos de barras, de líneas y de áreas para una misma serie de datos.

**Proceso.** Sólo disponible en el *Editor de procesos*. Sirve para crear nuevas funciones y rutinas de procesamiento, acceder a un editor de cuadros de diálogo, ejecutar procesos previamente definidos, controlar el color y el tipo de letra de las palabras clave y de los comentarios de los archivos de proceso, etc.

**Depurar.** Sólo disponible en el *Editor de procesos*. Permite depurar procesos básicos y acceder a un editor de objetos.

**Formato.** Disponible en el *Visor de resultados* y en sus tres editores (texto, gráficos y tablas). En el *Visor de resultados* permite cambiar el alineamiento de los objetos. En el *Editor de tablas* contiene opciones para controlar el formato y las propiedades de las tablas y de sus celdas, ajustar automáticamente las dimensiones de la tabla, rotar las cabeceras de las filas y de las columnas, etc. En el *Editor de gráficos*, permite controlar el color y la trama del relleno, el estilo de las líneas y de las barras, y el tipo de letra; también permite intercambiar los ejes de un gráfico, efectuar rotaciones 3D en los diagramas de dispersión, desgajar uno o más sectores de un diagrama de sectores, etc. En el *Editor de texto*, sirve para modificar el tipo de letra (fuente, tamaño, aspecto) y el alineamiento de los objetos de texto.

**Analizar.** Contiene todos los procedimientos estadísticos. El contenido de este menú depende de la cantidad de módulos SPSS que se tengan instalados (el programa SPSS se distribuye por módulos: *Base*, *Avanzado*, *Tablas*, *Modelos de regresión*, *Tendencias*, *Categorías*, etc.). Y otros productos SPSS (como *Answer Tree*, *Amos*, etc.) también pueden aparecer listados en este menú.

**Gráficos.** Desde este menú es posible generar todo tipo de gráficos: de barras, de líneas, de sectores, diagramas de dispersión, histogramas, gráficos de control de calidad, etc.

**Utilidades.** No disponible en el *Editor de gráficos*. Permite obtener información sobre las variables o sobre el archivo de datos, controlar las variables que aparecen en las listas de variables de los cuadros de diálogo, ejecutar procesos (o crear y editar autoprocesos desde el *Visor de resultados*) y editar las barras de menús. En el *Visor de resultados* y en el *Editor de sintaxis*, ofrece la posibilidad de cambiar de ventana designada.

**Ejecutar.** Sólo disponible en el *Editor de sintaxis*. Contiene varias opciones para ejecutar total o parcialmente las sentencias de un archivo de sintaxis.

**Ventana.** No disponible en el *Editor de gráficos*. Permite cambiar de una ventana a otra dentro del SPSS y minimizar todas las ventanas abiertas.

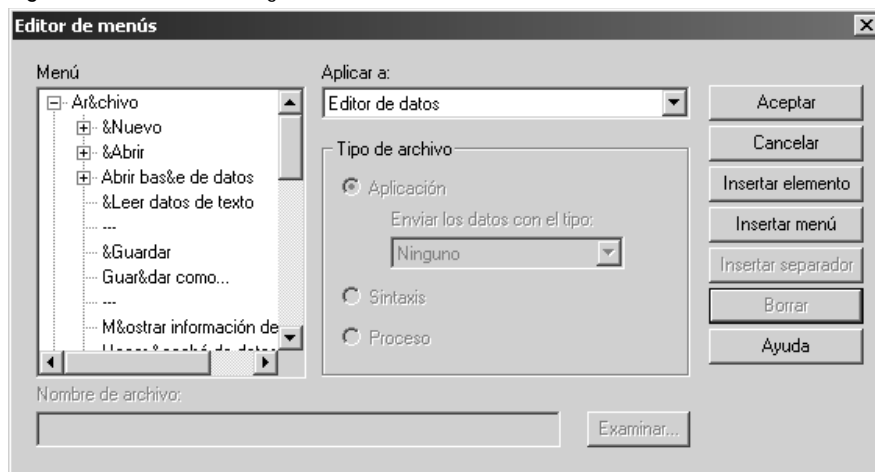
**Ayuda.** Ofrece ayuda general sobre las funciones más importantes del SPSS y proporciona acceso a la página principal de SPSS en Internet. También contiene el *Tutorial* y el *Asesor estadístico* (ver Capítulo 2) y detalles sobre la versión y propiedad del programa.

## El Editor de menús

El *Editor de menús* del SPSS sirve para personalizar las barras de menús de las principales ventanas SPSS. Mediante el *Editor de menús* es posible añadir (o eliminar) uno o varios menús nuevos a las barras de menús disponibles y añadir (o eliminar) uno o varios elementos de menú a los menús ya existentes. Aunque no es necesario conocer el *Editor de menús* para utilizar correctamente el SPSS, su uso podría facilitar, en ocasiones, la ejecución de tareas repetitivas no incluidas en los menús del sistema. Para utilizar el *Editor de menús*:

- Seleccionar la opción **Editor de menús** del menú **Utilidades** para acceder al cuadro de diálogo *Editor de menús* que muestra la Figura 1.6.



Figura 1.6. Cuadro de diálogo *Editor de menús*

**Menú.** Este listado contiene un esquema con los menús de la barra correspondiente a la ventana SPSS desde la cual se ha entrado en el *Editor de menús*. Es posible cambiar la barra de menús que aparece en este esquema seleccionando la ventana correspondiente en el menú desplegable **Aplicar a**.

- Al pulsar el signo más (+) situado delante de cada menú, el esquema se expande y muestra los elementos de ese menú (se obtiene el mismo resultado situando el puntero del ratón sobre el nombre del menú y pulsando dos veces el botón principal). Una vez expandido, el signo más se transforma en menos (–); al pulsarlo, el esquema se contrae ocultando los elementos del menú (se obtiene el mismo resultado situando el puntero del ratón sobre el nombre del menú y pulsando dos veces el botón principal).
- Para añadir un *menú nuevo* basta con pulsar el botón **Insertar menú** tras situar el cursor sobre el menú delante del cual se desea crear el nuevo menú. El nuevo menú recibe, por defecto, el nombre *Nuevo menú*, pero es posible asignarle cualquier otro nombre utilizando el teclado.
- Para insertar un *elemento de menú nuevo* se debe expandir el menú dentro del cual se desea incluir ese nuevo elemento, seleccionar el elemento delante del cual (por encima del cual) se quiere que aparezca el nuevo elemento, y pulsar el botón **Insertar elemento**. El elemento recién insertado recibe, por defecto, el nombre *Nuevo elemento de menú*, pero es posible asignarle cualquier otro nombre utilizando el teclado.
- El Botón **Borrar** elimina el menú o los elementos de menú seleccionados.
- El botón **Separador** crea un elemento de menú vacío que, en el menú desplegable de la barra de menús, aparece como una línea separadora.

**Aplicar a.** Permite seleccionar la ventana concreta cuya barra de menús se desea editar. El menú desplegable al que se accede con el botón flecha ▼ permite seleccionar una de estas ventanas: *Editor de datos*, *Visor de resultados*, *Editor de procesos* y *Editor de sintaxis*.

**Nombre de archivo.** Un elemento de menú nuevo no se convierte en operativo hasta que se le asocia un archivo ejecutable. Este archivo ejecutable debe ser nombrado, junto con su ruta completa, en esta casilla. Para asociar un archivo ejecutable al nuevo elemento de menú:

- Pulsar el botón **Examinar...** para acceder al subcuadro de diálogo *Abrir* y utilizar las opciones del cuadro de diálogo para buscar el archivo ejecutable deseado y asignarlo al nuevo elemento de menú.

**Tipo de archivo.** El archivo ejecutable seleccionado puede ser de tres tipos:

**Aplicación.** Una aplicación externa (archivos con extensión *.exe*).

**Sintaxis.** Un *archivo de sintaxis* de SPSS (archivos con extensión *.sps*).

**Proceso.** Un *archivo de proceso* de SPSS (archivos con extensión *.sbs*).

Si el archivo ejecutable seleccionado es el de una aplicación externa, el nuevo elemento de menú no sólo permite iniciar esa aplicación externa, sino que, al iniciarla, transfiere automáticamente a ella el archivo de datos (osea, el contenido del *Editor de datos*). Ahora bien, para que esta transferencia sea correcta, es necesario indicar el formato en el que debe ser transferido el archivo. Para seleccionar un formato:

- Pulsar el botón **Enviar los datos con el tipo** y seleccionar del menú desplegable el formato en el que el SPSS debe transferir los datos: SPSS, Excel 4.0, Lotus 1-2-3 versión 3, SYLK, ASCII delimitado por tabulaciones y dBaseIV.

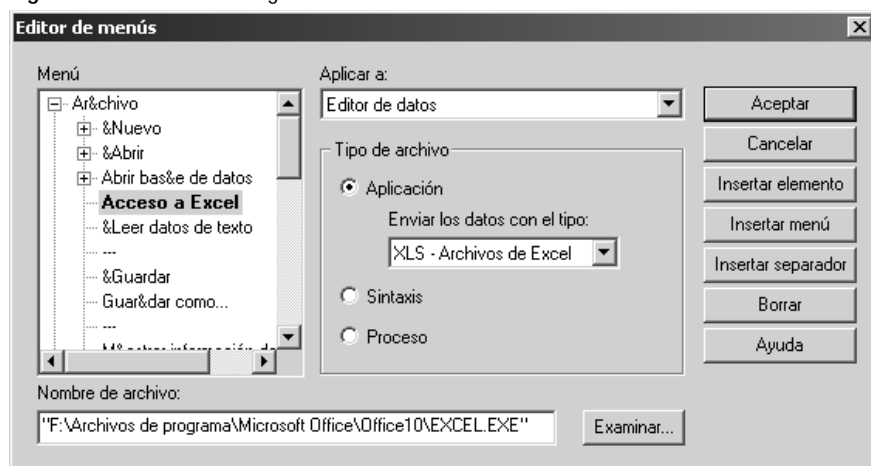
### ***Ejemplo: Utilidades > Editor de menús***

Este ejemplo muestra cómo crear un nuevo elemento de menú llamado *Acceso a Excel* dentro del menú *Archivo* del *Editor de datos*. Este nuevo elemento, al que se le asociará el archivo ejecutable *Excel.exe*, será ubicado entre los elementos *Abrir bases de datos* y *Leer datos de texto*:

- En la ventana del *Editor de datos*, seleccionar **Utilidades > Editor de menús** para acceder al cuadro de diálogo del *Editor de menús* (Figura 1.6).
- En el listado **Menú**, pulsar sobre el signo más (+) situado delante del menú *Archivo* (para expandirlo) y seleccionar el elemento *Leer datos de texto* (el nuevo elemento de menú se desea insertar en la posición inmediatamente anterior a *Leer datos de texto*).
- Para crear el nuevo elemento de menú, pulsar el botón **Insertar elemento** y, utilizando el teclado, sustituir el nombre *Nuevo elemento de menú* (que el sistema asigna por defecto) por el nombre *Acceso a Excel*.
- Pulsar el botón **Examinar...** para buscar el archivo ejecutable de la aplicación Excel, que es la aplicación que se desea asociar al nuevo elemento de menú *Acceso a Excel*.
- En el recuadro **Tipo de archivo**, seleccionar la opción *XLS - Archivos de Excel* para indicar al SPSS el formato en el que debe transferir el archivo de datos a la aplicación externa asociada al nuevo elemento de menú creado.

Al llegar a este punto, el cuadro de diálogo del *Editor de menús* ha quedado configurado tal como muestra la Figura 1.7. El nuevo elemento *Acceso a Excel* está ubicado en la posición seleccionada.

Figura 1.7. Cuadro de diálogo *Editor de menús* con el nuevo elemento de menú *Acceso a Excel*



Pulsando el botón *Aceptar*, el nuevo elemento de menú *Acceso a Excel* pasa a formar parte del menú *Archivo* (en la barra de menús del *Editor de datos*). En el *Editor de datos* puede comprobarse que el nuevo elemento se encuentra efectivamente en la posición en la que ha sido creado, es decir, en el menú *Archivo* detrás del elemento *Abrir bases de datos* (ver Figura 1.8). A partir de este momento, al seleccionar *Acceso a Excel* se iniciará la aplicación Excel y el archivo de datos será automáticamente transferido a Excel.

Figura 1.8. Elementos del menú desplegable *Archivo* de la barra de menús del *Editor de datos*

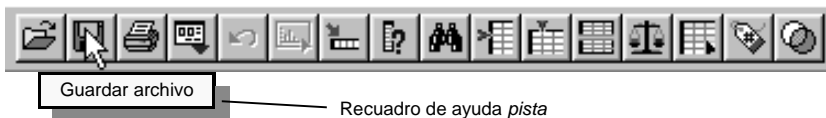


## Barra de herramientas

Una *barra de herramientas* es un conjunto de botones-iconos que permiten ejecutar algunas de las funciones del SPSS de forma rápida y sencilla (por supuesto, funciones que también es posible ejecutar con las opciones de la *barra de menús*).

Cada ventana tiene su propia barra de herramientas, con botones específicos adaptados a sus funciones básicas. Al pasar con el puntero del ratón (sin pulsar ningún botón) por encima de una herramienta, se abre automáticamente un pequeño recuadro de ayuda *pista* ofreciendo una breve descripción de la función asociada a esa herramienta. La Figura 1.9 muestra un ejemplo de estos recuadros de ayuda *pista*.

**Figura 1.9.** Barra de herramientas del *Editor de datos* con un ejemplo del recuadro de ayuda *pista* que aparece al situar el puntero del ratón sobre el botón-icono de la herramienta *Guardar archivo*



A continuación se describe, como ejemplo de barra de herramientas, la barra del *Editor de datos* (Figura 1.9). El resto de las barras de herramientas son similares a ésta (algunas de las herramientas son, incluso, las mismas) y se aprenderán a manejar en los próximos capítulos a medida que se vayan explicando cada una de las ventanas del SPSS.



**Abrir archivo.** Muestra el cuadro de diálogo *Abrir archivo*, el cual permite abrir un archivo SPSS de cualquier tipo. Con esta herramienta, los archivos listados por defecto son archivos de datos en formato SPSS con extensión *.sav*; sin embargo, el cuadro de diálogo contiene opciones que permiten seleccionar el tipo de archivos que se desea ver listados.



**Guardar archivo.** Guarda el archivo de datos. Si el archivo no tiene nombre, abre el cuadro de diálogo *Guardar archivo*, el cual permite asignar nombre y ruta al archivo de la ventana activa.



**Imprimir.** Abre el cuadro de diálogo *Imprimir documento*, el cual permite imprimir el archivo completo o sólo una parte seleccionada.



**Recuperar cuadros de diálogo.** Muestra una lista con los últimos cuadros de diálogo abiertos y permite elegir cualquiera de ellos.



**Deshacer/rehacer.** deshace o rehace las últimas acciones de edición llevadas a cabo: borrar o introducir un dato, o una variable, o un caso, cambiar el nombre de una variable, etc.



**Ir a gráfico.** Convierte en ventana activa la ventana del *Editor de gráficos* (si es que existe una ventana de este tipo abierta).



**Ir a caso.** Abre el cuadro de diálogo *Ir a caso*, el cual permite desplazar el cursor a un caso concreto del archivo de datos.



**Variables.** Abre el cuadro de diálogo *Variables*, que contiene información sobre el formato, las etiquetas y los valores perdidos de las variables del archivo de datos.



**Buscar.** Abre el cuadro de diálogo *Buscar datos*, el cual permite buscar valores concretos en la variable en la que se encuentra el cursor.



**Insertar caso.** Inserta una fila nueva delante (inmediatamente más arriba) de la fila en la que se encuentra situado el cursor.



**Insertar variable.** Inserta una columna nueva delante (inmediatamente a la izquierda) de la columna en la que se encuentra situado el cursor.



**Segmentar archivo.** Abre el cuadro de diálogo *Segmentar archivo*, el cual permite dividir el archivo de datos en grupos (segmentos) utilizando una o más variables.



**Ponderar casos.** Abre el cuadro de diálogo *Ponderar casos*, el cual permite utilizar una variable para ponderar los casos del archivo de datos (es decir, para dar un peso específico a cada caso).



**Seleccionar casos.** Abre el cuadro de diálogo *Seleccionar casos*, el cual permite seleccionar una parte del archivo de datos utilizando diferentes criterios (pueden establecerse filtros, extraer una muestra aleatoria, etc.).



**Mostrar etiquetas de valor.** Hace que las celdas del *Editor de datos* muestren las etiquetas de los valores. Al presionarlo de nuevo, las celdas muestran los valores.



**Usar conjuntos.** Abre el cuadro de diálogo *Usar conjuntos*, el cual permite seleccionar conjuntos de variables previamente definidos.

## Cómo personalizar una barra de herramientas

Las barras de herramientas están situadas, por defecto, en la parte superior de las ventanas, justo debajo de las barras de menús, pero pueden reubicarse en cualquier otro lugar de la pantalla. Para cambiar de lugar una barra de herramientas:

- Situar el puntero del ratón dentro de la barra en un lugar no ocupado por botones, pulsar el botón principal del ratón y arrastrarla (manteniendo pulsado el botón) hasta el lugar deseado.

Si una barra es arrastrada al borde izquierdo o derecho de su ventana, adopta una forma *vertical*. Si es arrastrada al borde superior o inferior de su ventana, adopta una forma *horizontal*. Si es arrastrada a cualquier otra parte dentro o fuera de su ventana, adopta una forma variable (generalmente *rectangular*) que puede ser modificada pinchando con el puntero del ratón sobre sus lados y arrastrándolos hasta darle la forma deseada.

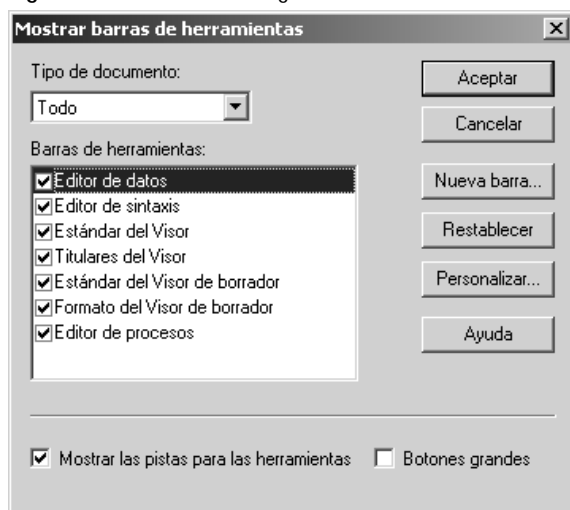
Las barras de herramientas no sólo pueden cambiarse de lugar. El cuadro de diálogo *Mostrar barras de herramientas* también permite ocultarlas cuando están visibles, mostrarlas cuando están ocultas, controlar el tamaño de los botones-iconos asociados a cada herramienta,


modificar las barras de herramientas existentes a gusto del usuario (añadiendo o quitando herramientas) y crear nuevas barras de herramientas (con herramientas disponibles o con otras nuevas creadas por el usuario).

Para editar una barra de herramientas:

- Seleccionar la opción **Barras de herramientas...** del menú **Ver** para acceder al cuadro de diálogo *Mostrar barras de herramientas* que muestra la Figura 1.10. Se consigue el mismo efecto situando en puntero del ratón sobre la barra de herramientas, pulsando el botón secundario del ratón para obtener un menú desplegable y seleccionando en ese menú la opción **Barras de herramientas...**

Figura 1.10. Cuadro de diálogo *Mostrar barras de herramientas*



**Tipo de documento.** Pulsando el botón de menú desplegable  se puede seleccionar el tipo de ventana cuya barra se desea ocultar/mostrar o personalizar. En el cuadro de diálogo de la Figura 1.10 se han seleccionado todas las ventanas.

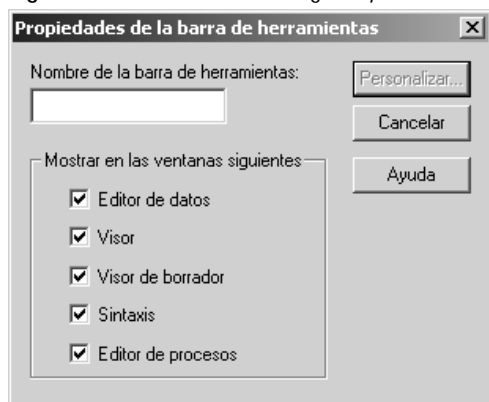
**Barra de herramientas.** Las barras visibles aparecen junto a casillas marcadas.

- Marcando o desmarcando una casilla de selección con el puntero del ratón, la correspondiente barra de herramientas queda activada o desactivada. De esta forma es posible decidir, para cada ventana, si estará o no visible una determinada barra de herramientas.
- “ **Mostrar las pistas para las herramientas.** Al colocar el puntero del ratón (sin pulsar ningún botón) sobre los botones de una barra de herramientas, aparece un pequeño recuadro de ayuda *pista* (ver Capítulo 2) que ofrece información sobre la función que realiza cada herramienta (ver Figura 1.9). Esta opción se encuentra activa por defecto. Para desactivarla basta con pinchar con el puntero del ratón en la opción **Mostrar las pistas para las herramientas**.

- “ **Botones grandes.** Esta opción permite controlar el tamaño de los botones de las barras de herramientas. Desactivada, muestra los botones en su tamaño estándar. Activada, los aumenta de tamaño.

**Nueva barra...** Este botón conduce al subcuadro de diálogo *Propiedades de la barra de herramientas* que muestra la Figura 1.11.

Figura 1.11. Subcuadro de diálogo *Propiedades de la barra de herramientas*



**Nombre de la barra de herramientas.** Permite asignar un nombre a la nueva barra de herramientas.

**Mostrar en las ventanas siguientes.** Contiene opciones para decidir en qué tipo de ventanas estará visible la nueva barra de herramientas: activando y desactivando las casillas correspondientes a cada ventana es posible decidir en cuál de ellas estará visible la nueva barra de herramientas.

**Personalizar...** Tras asignar un nombre a la nueva barra y decidir en qué ventanas estará visible, las opciones asociadas al botón **Personalizar...** permiten decidir qué herramientas concretas formarán parte de la nueva barra. Para seleccionar las herramientas de la nueva barra:

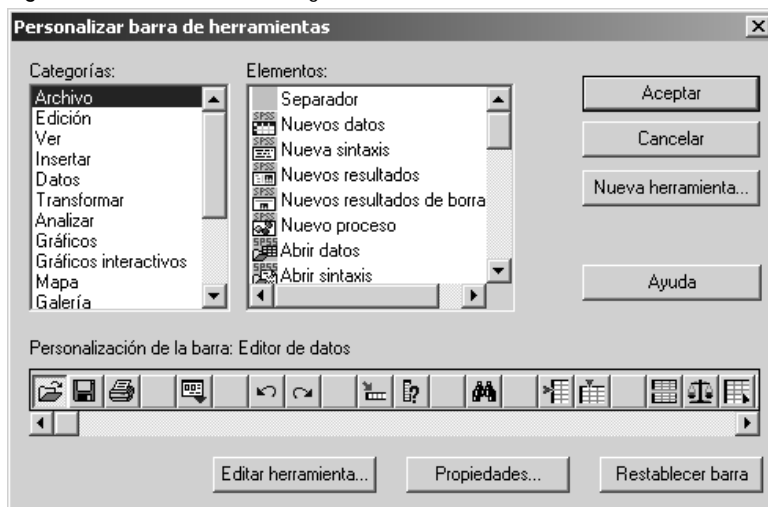
- Pulsar el botón **Personalizar...** para acceder al subcuadro de diálogo *Personalizar barra de herramientas* que muestra la Figura 1.12.

**Personalizar...** Tanto con el botón **Personalizar...** del subcuadro de diálogo *Propiedades de la barra de herramientas* (ver Figura 1.11), como con el botón **Personalizar...** del cuadro de diálogo *Mostrar barras de herramientas* (ver Figura 1.10), se accede al subcuadro de diálogo *Personalizar barra de herramientas* que muestra la Figura 1.12. Este subcuadro de diálogo permite:

- Personalizar barras de herramientas ya existentes (añadiendo o quitando herramientas disponibles).
- Crear nuevas barras de herramientas, con nuevos nombres, tomando como base alguna de las barras de herramientas disponibles.

- Crear nuevas herramientas (botón incluido) utilizando el *Editor de mapa de bits*.
- Decidir en qué ventanas aparecerá cada barra de herramientas.

Figura 1.12. Subcuadro de diálogo *Personalizar barra de herramientas*



**Categorías.** Las herramientas disponibles se encuentran agrupadas en categorías. Este listado ofrece todas esas categorías (*Archivo*, *Edición*, *Ver*, *Insertar*, *Datos*, *Transformar*, etc.). Al seleccionar una de estas categorías, las herramientas agrupadas en esa categoría aparecen en la lista **Elementos**.

**Elementos.** Contiene varios listados con todas las herramientas disponibles en el SPSS. Los elementos incluidos en cada listado dependen de la categoría seleccionada en la lista **Categorías**.

**Personalización de la barra...** Contiene los botones-íconos de la barra de herramientas que se está intentando personalizar. Si se ha llegado hasta aquí desde el cuadro de diálogo *Mostrar barras de herramientas* (ver Figura 1.10), aparecerán los botones-íconos de la barra de herramientas seleccionada. Si se ha llegado aquí desde el subcuadro de diálogo *Propiedades de la barra de herramientas* (ver Figura 1.11), la barra de herramientas estará vacía.

- Para eliminar un botón (una herramienta), arrastrarlo fuera de la barra de herramientas.
- Para añadir un botón, arrastrarlo desde la lista **Elementos** hasta la posición deseada de la barra.

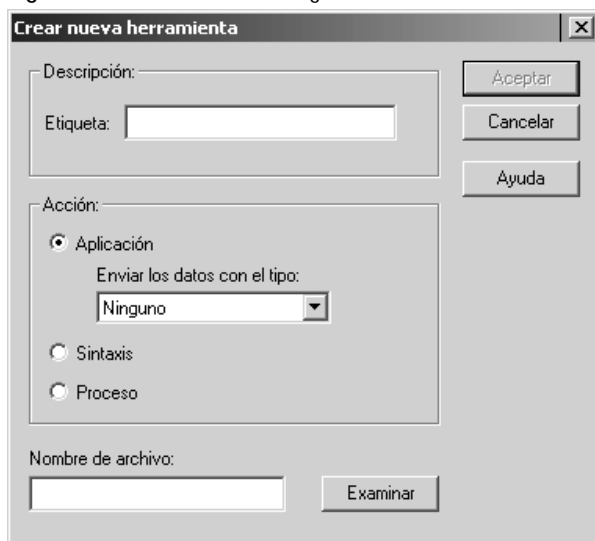
**Nueva herramienta.** Este botón permite asignar una etiqueta descriptiva a la nueva herramienta y asociarle un archivo ejecutable. Para ello:

- Pulsar el botón **Nueva herramienta...** para acceder al subcuadro de diálogo *Crear nueva herramienta* (ver Figura 1.13).



**Descripción.** El cuadro de texto **Etiqueta** permite asignar a la nueva herramienta una etiqueta descriptiva que será la que más tarde mostrará el recuadro de ayuda *pista* (ver Figura 1.9) al señalar la herramienta con el puntero del ratón.

Figura 1.13. Subcuadro de diálogo *Crear nueva herramienta*




**Acción.** Una nueva herramienta no queda definida hasta que se le asigna una de las siguientes tres acciones:

**Aplicación.** Ejecutar una aplicación externa.

**Sintaxis.** Ejecutar un archivo de sintaxis SPSS.

**Proceso.** Ejecutar un archivo de proceso SPSS.

En el caso de asignar una aplicación externa, la nueva herramienta, además de iniciar la aplicación seleccionada, puede hacer que el SPSS transfiera a esa aplicación el contenido del *Editor de datos*; para ello es necesario seleccionar el formato apropiado:

- Pulsar el botón de menú desplegable  del recuadro **Enviar los datos con el tipo** y seleccionar del listado uno de estos formatos: ninguno (para no transferir datos) SPSS, Excel 4.0, Lotus 1-2-3 versión 3, SYLK, ASCII delimitado por tabulaciones y dBaseIV.

**Nombre de archivo.** Una herramienta nueva no es operativa hasta que se le asocia un archivo ejecutable. Ese archivo ejecutable debe ser nombrado, junto con su ruta completa, en esta casilla. Para asociar un archivo ejecutable a la nueva herramienta:

- Pulsar el botón **Examinar...** para acceder al subcuadro de diálogo *Abrir*, buscar el archivo ejecutable que se desea incluir en el menú y asignarlo al nuevo elemento de menú.

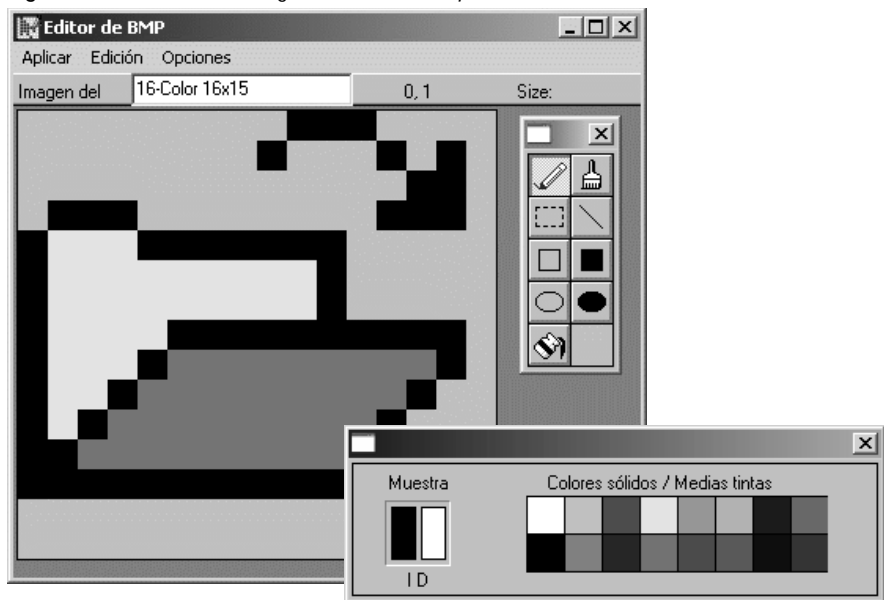
**Restablecer barra.** Restablece la barra de herramientas, es decir, la devuelve a su forma original.

**Propiedades...** Conduce al subcuadro de diálogo *Propiedades de la barra de herramientas* (ver Figura 1.11). Desde este subcuadro puede cambiarse el nombre a la herramienta y decidir en qué tipo de ventanas estará visible la barra.

**Editar herramienta...** El SPSS dispone de un *Editor del mapa de bits* que permite editar los botones-iconos de las herramientas. Este editor resulta especialmente útil para crear botones-iconos para las nuevas herramientas. Para crear o cambiar un botón:

- Pulsar el botón **Editar herramienta...** para acceder al cuadro de diálogo *Editor de BMP* que muestra la Figura 1.14.

Figura 1.14. Cuadro de diálogo del *Editor del mapa de bits*



La paleta de colores y la caja de herramientas del *Editor del mapa de bits* permiten trabajar como en una pantalla de dibujo (trazando y coloreando) hasta dar al nuevo botón el aspecto deseado.

## Barra de estado

Todas las barras de estado están situadas en la parte inferior de las ventanas del SPSS (ver Figura 1.1). Pueden ocultarse/mostrarse seleccionando la opción **Barra de estado** del menú **Ver**. Ofrecen información sobre diferentes aspectos del SPSS:

- **Estado del procesador.** Cuando el SPSS está ocupado, la barra de estado muestra el nombre del procedimiento que se está ejecutando y un contador indicando el número de casos

procesados. Si el procedimiento que se está ejecutando utiliza un método iterativo, el contador muestra el número de iteraciones. Cuando el procesador está inactivo, la barra de estado muestra el mensaje *Procesador de SPSS para Windows preparado*.


- **Estado del filtrado de casos.** Si se ha seleccionado una muestra aleatoria de casos o un subconjunto de casos que cumplen cierta condición, la barra de estado muestra el mensaje *Filtrado*, el cual indica que el filtro (la selección de casos) está activo.
- **Estado de la ponderación de casos.** Si se utiliza alguna variable para ponderar el número de veces que se repite cada caso, la barra de estado muestra el mensaje *Ponderado*, el cual indica que la ponderación de casos está activa.
- **Estado de la segmentación del archivo.** Si se utiliza alguna variable para segmentar (dividir en subgrupos) el archivo de datos, la barra de estado muestra el mensaje *Segmentado*, el cual indica que la segmentación del archivo está activa.
- **Indicador de ventana designada.** La barra de estado del *Visor de resultados* y del *Editor de sintaxis* utiliza un signo de admiración rojo para indicar cuál de las diferentes ventanas abiertas es la designada.

## Una sesión con el SPSS

Como una primera aproximación al funcionamiento del programa, este apartado propone realizar algunas tareas básicas para que el lector vaya familiarizándose con la forma de trabajar en el SPSS. En una sesión estándar, estas tareas básicas suelen ser tres: abrir un archivo de datos, ejecutar un procedimiento estadístico y examinar los resultados.

### Abrir un archivo de datos

Una vez abierto el *Editor de datos* (ver Figura 1.1), la primera acción suele consistir en introducir datos desde el teclado o en abrir un archivo de datos existente. Puesto que no es el momento de ponerse a introducir datos, se va a abrir un archivo llamado *Datos de empleados*, que se encuentra en la misma carpeta en la que se ha instalado el SPSS. Para ello:

- Seleccionar la opción **Abrir > Datos...** del menú **Archivo** para acceder al cuadro de diálogo *Abrir archivo* (ver Figura 1.2). Este cuadro de diálogo muestra, por defecto, un listado de los archivos con extensión *.sav* (los archivos con extensión *.sav* son archivos de datos en formato SPSS; ver Capítulo 3). Si el archivo buscado no se encuentra en la carpeta que el SPSS abre por defecto, puede utilizarse el botón *carpeta-arriba*  hasta llegar a la carpeta raíz (*Mi PC*) y, a partir de ella, continuar buscando la carpeta apropiada: el archivo *Datos de empleados* se encuentra en la misma carpeta en la que se ha instalado el SPSS.
- Buscar en el cuadro de diálogo *Abrir archivo* el archivo *Datos de empleados* y, para abrirlo, seleccionarlo y pulsar dos veces el botón secundario del ratón (o, alternativamente, seleccionarlo y pulsar el botón **Abrir**).

Al abrir un archivo de datos, el *Editor de datos*, hasta ahora vacío, toma el aspecto que muestra la Figura 1.15.

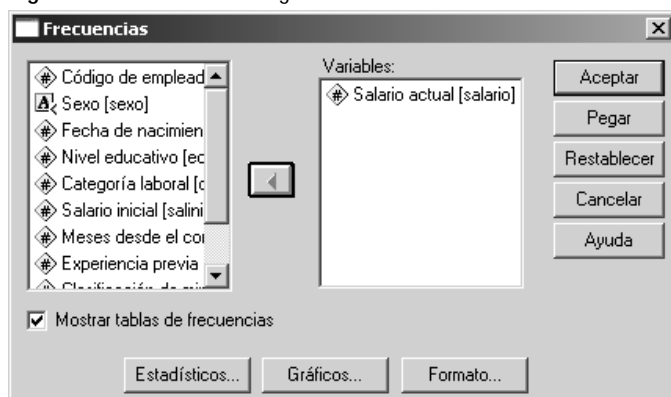
Figura 1.15. Editor de datos (archivo: Datos de empleados)

	id	sexo	fechnac	educ	catlab	salario	salini	tiempemp	expprev
1	1	h	03.02.1952	15	3	\$57,000	\$27,000	98	144
2	2	h	23.05.1958	16	1	\$40,200	\$18,750	98	36
3	3	m	26.07.1929	12	1	\$21,450	\$12,000	98	381
4	4	m	15.04.1947	8	1	\$21,900	\$13,200	98	190
5	5	h	09.02.1955	15	1	\$45,000	\$21,000	98	138
6	6	h	22.08.1958	15	1	\$32,100	\$13,500	98	67
7	7	h	26.04.1956	15	1	\$36,000	\$18,750	98	114
8	8	m	06.05.1966	12	1	\$21,900	\$9,750	98	0
9	9	m	23.01.1946	15	1	\$27,900	\$12,750	98	115

## Utilizar un procedimiento estadístico

Una vez abierto un archivo de datos, la siguiente acción que suele llevarse a cabo en una sesión con el SPSS consiste en seleccionar algún procedimiento para modificar el contenido de la base de datos, ejecutar un procedimiento estadístico, obtener un gráfico, etc. En este ejemplo se va a seleccionar el procedimiento Frecuencias para obtener un histograma de la variable *salario* (salario actual). El procedimiento Frecuencias no se describe con detalle en este capítulo; de momento, sólo se va a utilizar para ilustrar cómo funciona el SPSS. Para abrir el cuadro de diálogo *Frecuencias*:

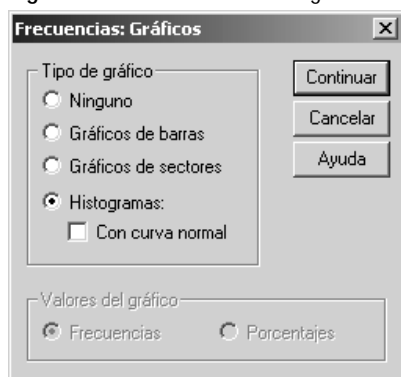
- Seleccionar la opción Estadísticos descriptivos... > Frecuencias del menú Analizar para acceder al cuadro de diálogo *Frecuencias* que muestra la Figura 1.16.

Figura 1.16. Cuadro de diálogo *Frecuencias*

Trasladando ahora la variable *salario* desde el listado de *variables del archivo* al listado de *variables seleccionadas* (mediante el botón *flecha*) y pulsando el botón **Aceptar** se obtiene el resultado que el procedimiento **Frecuencias** ofrece por defecto: una distribución o tabla de frecuencias de la variable *salario*. Sin embargo, puesto que *salario* es una variable cuantitativa, no interesa obtener una distribución de frecuencias (demasiado larga y, por tanto, poco informativa). Lo que interesa obtener, según se ha señalado ya, es un histograma de la variable *salario* (salario actual). Para ello:

- Seleccionar la variable *salario* (salario actual) en la lista de variables del archivo de datos y trasladarla a la lista **Variables**.
- Desactivar la opción **Mostrar tablas de frecuencias** (no se tiene intención de obtener la distribución de frecuencias de la variable *salario*, sino un histograma).
- Pulsar el botón **Gráficos...** para acceder al cuadro de diálogo *Frecuencias: Gráficos* que muestra la Figura 1.17.

Figura 1.17. Subcuadro de diálogo *Frecuencias: Gráficos*



- Marcar las opciones **Histogramas** y **Con curva normal** del apartado **Tipo de gráfico** y pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Una vez marcadas las opciones deseadas, el botón **Aceptar** (ver cuadro de diálogo *Frecuencias*; Figura 1.16), hace que el procesador del SPSS ejecute el procedimiento **Frecuencias** y mande el resultado de esa ejecución al *Visor de resultados*.

## Examinar los resultados

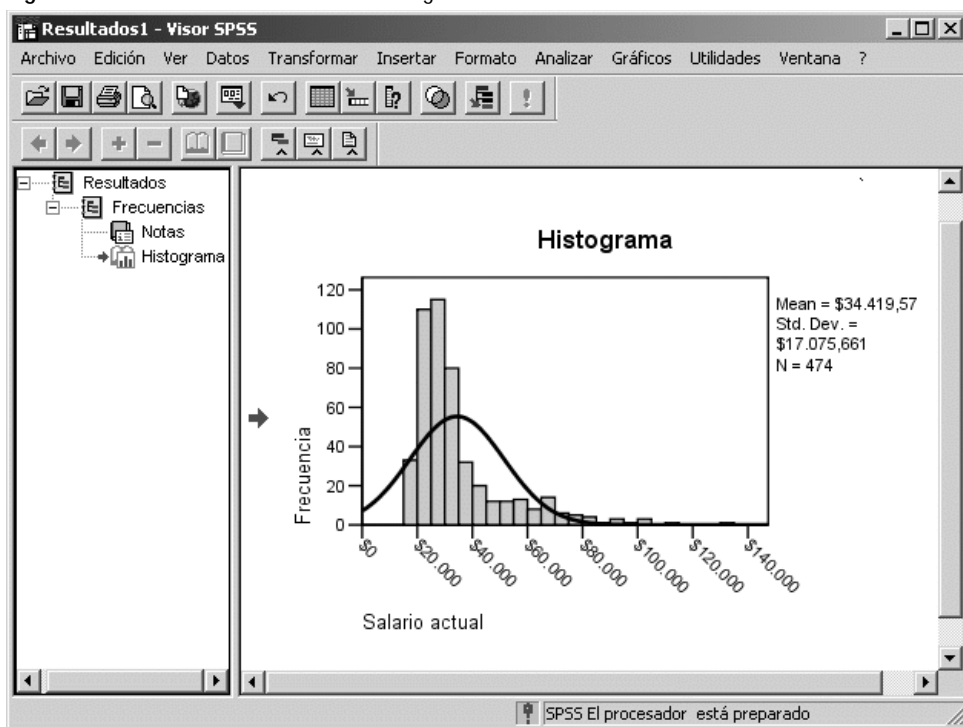
El *Visor de resultados* se encarga de recoger la información que el SPSS genera como consecuencia de los procedimientos que ejecuta. Puesto que en el procedimiento **Frecuencias** únicamente se ha solicitado un histograma de la variable *salario*, el *Visor de resultados* sólo contiene el histograma que muestra la Figura 1.18.

Los detalles del *Visor de resultados* se explican en el Capítulo 7. No obstante, conviene empezar ya a familiarizarse con él. La ventana del *Visor de resultados* se encuentra dividida verticalmente en dos paneles:

- El **Esquema** del *Visor* (panel izquierdo), que contiene un índice o esquema de los resultados generados por el SPSS.
- El **Contenido** del *Visor* (panel derecho), que contiene los resultados (texto, tablas y gráficos) generados por el SPSS.

El tamaño de cada panel puede cambiarse simplemente pinchando con el puntero del ratón sobre la línea vertical que separa ambos paneles y arrastrándola hacia la izquierda o la derecha).

Figura 1.18. *Visor de resultados* con un histograma de la variable *salario*



En el ejemplo de la Figura 1.18, el *Esquema* del *Visor* (panel izquierdo) contiene un *encabezado* que identifica el procedimiento utilizado (Frecuencias) y, colgando de ese encabezado, una serie de *titulares* que informan sobre los resultados asociados a ese procedimiento e incluidos en el *Contenido* del *Visor* (panel derecho). El panel derecho, es decir, el *Contenido* del *Visor*, muestra el resultado solicitado al procedimiento Frecuencias: un histograma de la variable *salario* con una curva normal superpuesta.



## Cómo utilizar la ayuda

El SPSS incluye un completo sistema de ayuda al que puede accederse desde cualquier ventana o cuadro de diálogo. Este sistema de ayuda adopta diferentes formatos dependiendo de la ubicación concreta desde la que se solicita la ayuda:

- El **menú ayuda (?)** de la barra de menús incluye:
  - Un sistema de ayuda por **temas** que permite acceder a todos los contenidos de la ayuda utilizando tres estrategias diferentes: *contenido*, *índice* y *búsqueda*.
  - Un **tutorial** que explica, paso a paso, cómo llevar a cabo muchas de las tareas propias del SPSS.
  - Un **asesor estadístico** que ayuda al usuario a decidir qué procedimiento estadístico debe utilizar para analizar sus datos.
- La **ayuda contextual** de los cuadros de diálogo, de las tablas pivotantes y de las barras de herramientas permite obtener ayuda puntual sobre diferentes aspectos de un cuadro de diálogo, de una tabla del *Visor de resultados* o de una herramienta concreta.
- El **botón ayuda** de los cuadros de diálogo permite obtener ayuda específica sobre el procedimiento SPSS concreto al que se refiere un cuadro de diálogo.
- El **asesor de resultados** ofrece ayuda similar al tutorial, pero referida a las diferentes partes de las tablas pivotantes del *Visor de resultados*.
- La **guía de sintaxis** contiene ayuda sobre la sintaxis concreta de cada procedimiento SPSS.

### La ayuda por temas

Para utilizar la ayuda por temas:

- Seleccionar la opción **Temas** del menú **Ayuda** para acceder al sistema de ayuda en formato HTML que muestra la Figura 2.1.

Este cuadro de diálogo permite acceder a la ayuda del SPSS utilizando tres estrategias diferentes (que se corresponden con las tres pestañas del cuadro de diálogo): **Contenido**, **Índice** y **Búsqueda**.



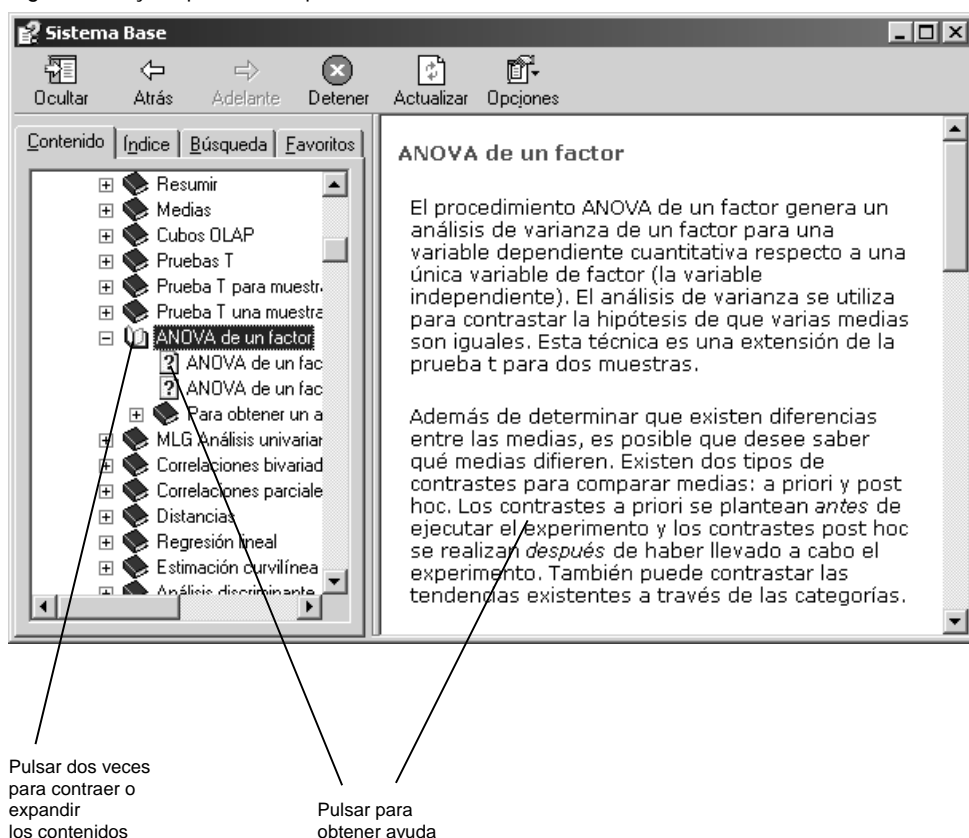
## Contenido

Esta opción proporciona ayuda a través de un listado ordenado de todos los contenidos de la ayuda, al estilo del índice de contenidos de un libro.

Las entradas principales de este índice se refieren a los distintos módulos del SPSS: *Base*, *Modelos avanzados*, *Modelos de regresión*, *Categorías*, *Tendencias*, etc. La Figura 2.1 muestra cómo obtener ayuda sobre el procedimiento ANOVA de un factor, que se encuentra en el módulo *Base*:

- Seleccionar la opción **Temas** del menú **Ayuda** para acceder al sistema de ayuda del SPSS.
- Pulsar la pestaña **Contenido** en el caso de que la pestaña activa sea otra.
- Seleccionar **Sistema base** y pulsar dos veces con el puntero del ratón sobre su icono (libro cerrado) para expandir los contenidos que incluye (o pulsar una vez sobre el signo + de su cabecera).
- Seleccionar cualquiera de las opciones acompañadas de un signo de interrogación para obtener la ayuda buscada.

Figura 2.1. Ayuda por temas. Opción *Contenido*



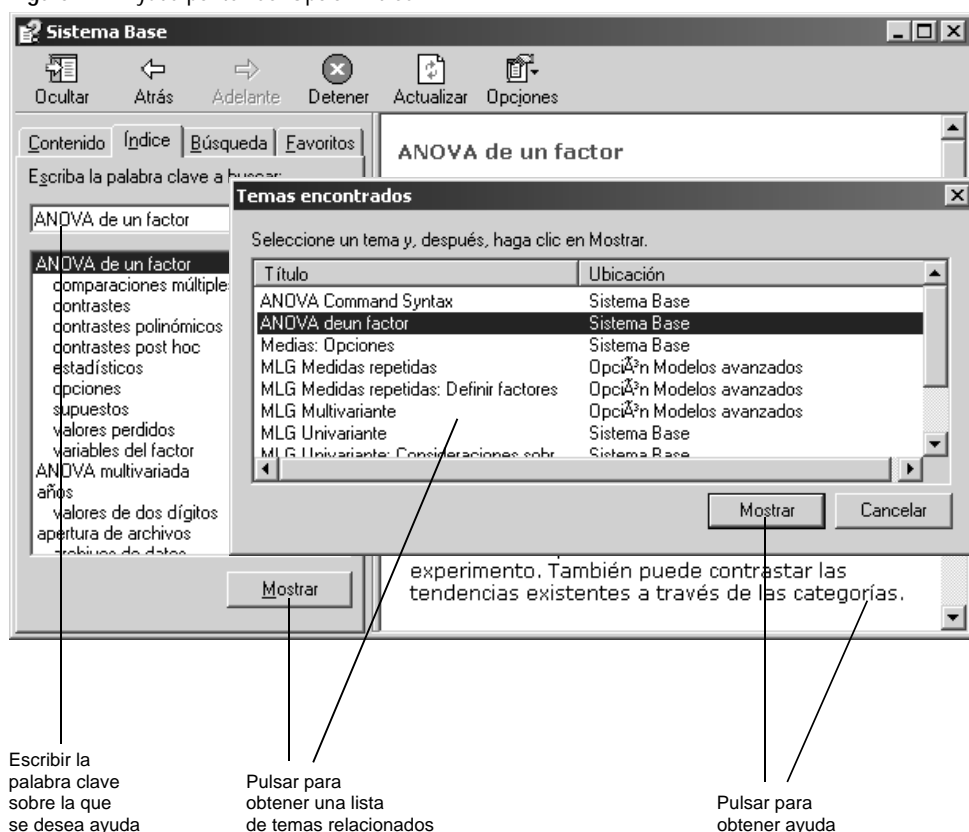
## Índice

La opción **Índice** ofrece un listado alfabéticamente ordenado de todos los temas y conceptos incluidos en el sistema de ayuda del SPSS.

Para obtener ayuda, por ejemplo, sobre el tema *análisis de varianza de un factor* (ver Figura 2.2):

- Seleccionar la opción **Temas** del menú **Ayuda** para acceder al sistema de ayuda del SPSS.
- Pulsar la pestaña **Índice** en el caso de que la pestaña activa sea otra.
- Escribir la palabra sobre la que se desea obtener ayuda. El sistema de ayuda buscará esa palabra en las entradas principales del índice y posicionará el cursor en ella, si existe.
- Pulsar el botón **Mostrar** (o pulsar dos veces sobre esa entrada principal) para obtener un listado de temas relacionados o procedimientos en los que se encuentra la entrada solicitada.
- Seleccionar cualquiera de las entradas que cuelgan de la entrada principal y pulsar el botón **Mostrar** (o pulsar dos veces sobre esa entrada) para obtener ayuda.

Figura 2.2. Ayuda por temas. Opción *Índice*



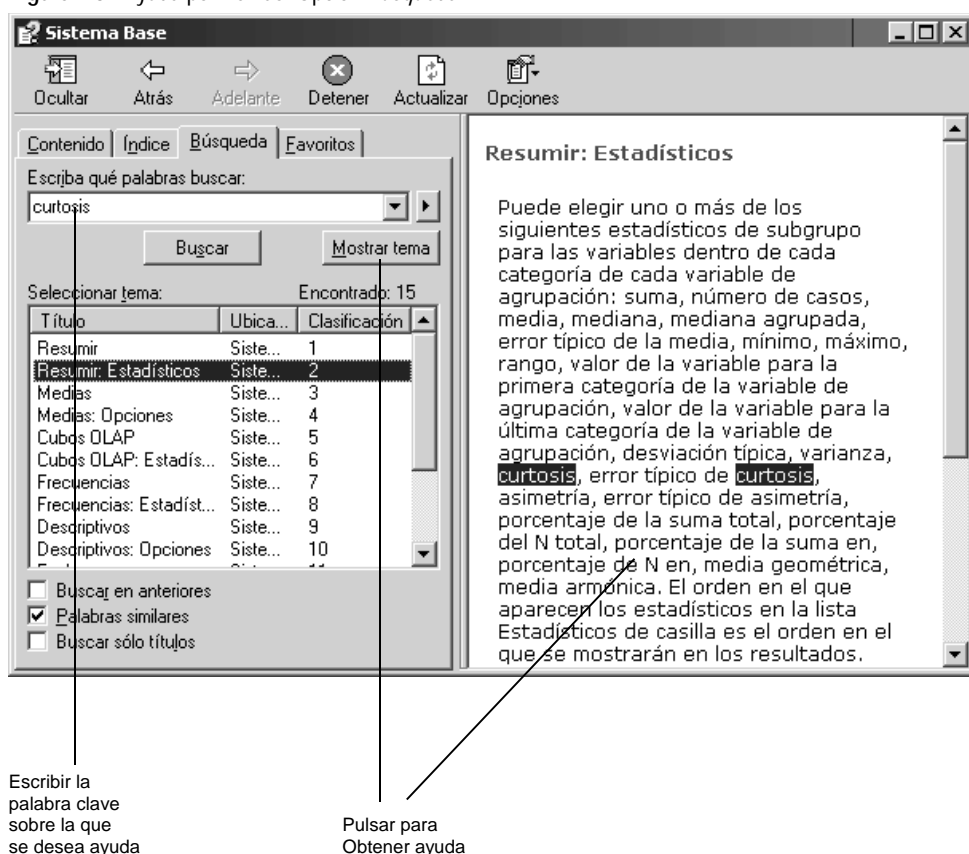
## Búsqueda

Esta opción permite buscar una o más palabras a través de todo el sistema de ayuda del SPSS. Puede resultar útil cuando se está seguro de que debe existir ayuda sobre el tema buscado y no se ha conseguido encontrar utilizando las estrategias **Índice** o **Contenido**. Mientras la opción **Índice** permite buscar palabras dentro de las entradas del índice, la opción **Búsqueda** busca palabras en todo el sistema de ayuda.

Para buscar ayuda, por ejemplo, sobre el término *curtosis* (ver Figura 2.3):

- Seleccionar la opción **Temas** del menú **Ayuda** para acceder al sistema de ayuda del SPSS.
- Pulsar la pestaña **Índice** en el caso de que la pestaña activa sea otra.
- Escribir la palabra sobre la que se desea obtener ayuda. El sistema de ayuda buscará esa palabra en el sistema de ayuda del SPSS y posicionará el cursor en la primera entrada que contenga la palabra buscada.
- Pulsar el botón **Mostrar tema** (o pulsar dos veces con el puntero del ratón sobre esa entrada) para obtener ayuda.

Figura 2.3. Ayuda por Temas. Opción *Búsqueda*



## El Tutorial

El *Tutorial* del SPSS es un sistema de ayuda que explica paso a paso, a modo de demostración, cómo llevar a cabo algunas de las muchas tareas que es posible realizar con el SPSS. Las pantallas del tutorial se visualizan en el explorador de Internet que se tenga instalado, o en cualquier otra aplicación que pueda trabajar con archivos *html*. Para utilizar el *Tutorial*:

- Seleccionar la opción **Tutorial** del menú **Ayuda** para acceder a los cuadros de diálogo del *Tutorial*.

A la información que contiene el *Tutorial* se accede mediante los cuatro botones que aparecen en la parte inferior derecha de la pantalla. La Figura 2.4 muestra estos botones.

Figura 2.4. Barra de botones del *Tutorial*



- Pulsar sobre los botones **Índice** o **Contenido** para localizar en el sistema de ayuda del *Tutorial* la lección sobre la que se desea ayuda. Estos botones funcionan de forma similar a como lo hacen las pestañas **Contenido** e **Índice** de la *ayuda por temas*.
- Una vez localizada la tarea sobre la que se desea ayuda, utilizar los botones **Página siguiente** y **Página anterior** para moverse por la lección.

## El Asesor estadístico

El SPSS incluye un *Asesor estadístico* que ayuda al usuario a decidir cuál es el procedimiento estadístico más apropiado para analizar sus datos.

Por supuesto, este *Asesor estadístico* no es un sustituto del analista de datos, de modo que es necesario tener algunas ideas claras para poder sacarle partido. Entre otras cosas, es necesario tener claro si se desea, por ejemplo, describir datos, comparar grupos o estudiar relaciones (entre otras cosas); o si las variables que se van a analizar son categóricas o cuantitativas; o si los datos proceden de muestras relacionadas o independientes; o si se cumplen o no determinados supuestos, etc. Y, aun conociendo todos estos detalles, las recomendaciones del *Asesor estadístico* sólo serán útiles si los conocimientos del usuario le permiten sacar partido de ellas. Para utilizar el *Asesor estadístico*:

- Seleccionar la opción **Asesor estadístico** del menú **Ayuda**.
- Seleccionar las opciones pertinentes en cada cuadro de diálogo del *Asesor estadístico* y pulsar el botón **Siguiente...** hasta que aparezca un mensaje indicando cuál es el procedimiento que el SPSS considera apropiado para analizar los datos propuestos.

## La Guía de sintaxis

La versión SPSS en CD-ROM incluye una guía con todo lo necesario para utilizar la sintaxis SPSS. El usuario avanzado encontrará de gran utilidad esta guía que, en realidad, no es otra cosa que el manual *SPSS Syntax Reference Guide* volcado en archivos de ayuda (sin traducir) en formato PDF. Para acceder a la *Guía de sintaxis* es necesario disponer del CD-ROM del programa o haber hecho una instalación personalizada. Para utilizar la *Guía de sintaxis*:

- Seleccionar la opción **Command Syntax Reference** del menú **Ayuda**.

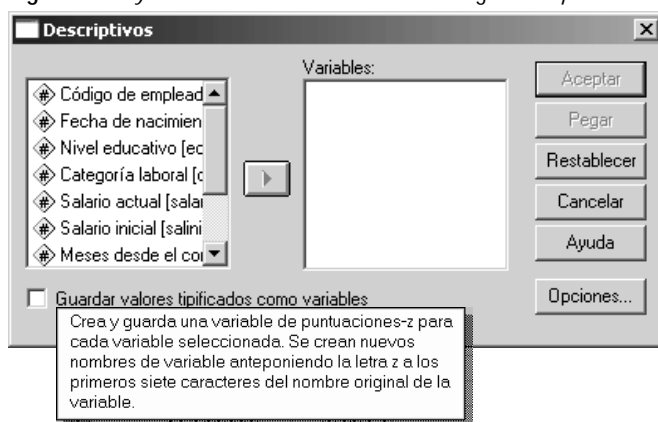
Esta opción conduce a un macroarchivo en el que se explican los detalles sintácticos de todos los procedimientos SPSS. Para poder visualizar el archivo es necesario tener instalado el programa *Acrobat Reader*.

## La ayuda contextual

Consiste en un sistema de ayuda rápida que proporciona información puntual sobre la parte concreta del programa desde la que se solicita la ayuda. Generalmente, esta ayuda contextual se obtiene pulsando el *botón secundario del ratón*, pero a veces es necesario hacer algo más y en algunos casos no es necesario pulsar ningún botón:

- En los **cuadros y subcuadros de diálogo**, al pulsar el botón secundario del ratón cuando éste se encuentra situado sobre una opción o un botón, el sistema abre un cuadro de ayuda con información sobre esa opción o ese botón. En el ejemplo de la Figura 2.5 se ha situado el ratón sobre la opción **Guardar valores tipificados como variables** (cuadro de diálogo *Descriptivos*) y, al pulsar el botón secundario del ratón, se ha obtenido el recuadro de ayuda contextual que muestra la figura.

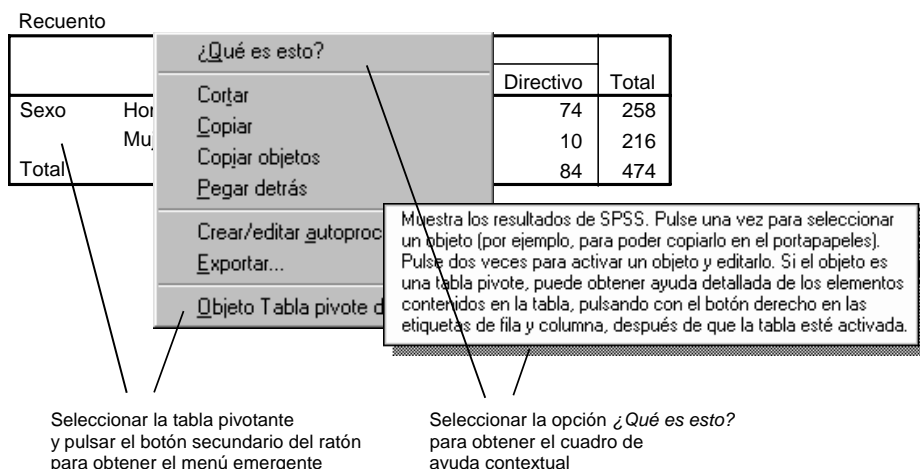
Figura 2.5. *Ayuda contextual* en el cuadro de diálogo *Descriptivos*



- En los botones-iconos de las **barras de herramientas**, basta con situar el puntero del ratón (sin pulsar ningún botón) sobre una herramienta concreta para obtener una descripción rápida de la misma en forma de cuadro de ayuda *pista* (ver Figura 1.9).

- En los **objetos** (texto, tablas, gráficos) del *Visor de resultados* y en las **listas de variables** de los cuadros de diálogo, al pulsar el botón secundario del ratón aparece un **menú emergente** con varias opciones, la primera de las cuales es *¿Qué es esto?* Seleccionando esta opción aparece un cuadro de ayuda contextual (ver Figura 2.6).

Figura 2.6. *Ayuda contextual* en una tabla pivotante del *Visor de resultados*



Los cuadros de ayuda contextual se desactivan pulsando el botón principal del ratón tras situar el puntero del ratón en cualquier lugar de la pantalla (incluido el propio cuadro de ayuda).

## El *Asesor de resultados*

El *Asesor de resultados* proporciona ayuda sobre el significado de las distintas partes de que consta una tabla de resultados. Su objetivo es el de facilitar la interpretación de los resultados. Para acceder al *Asesor de resultados*:

- Seleccionar la tabla de resultados sobre la que se desea obtener ayuda y entrar en modo de edición (situando el puntero del ratón sobre la tabla y pulsando dos veces el botón principal).
- Situar el cursor en la casilla sobre la que se desea información y seleccionar la opción *Asesor de resultados* del menú *Ayuda*.

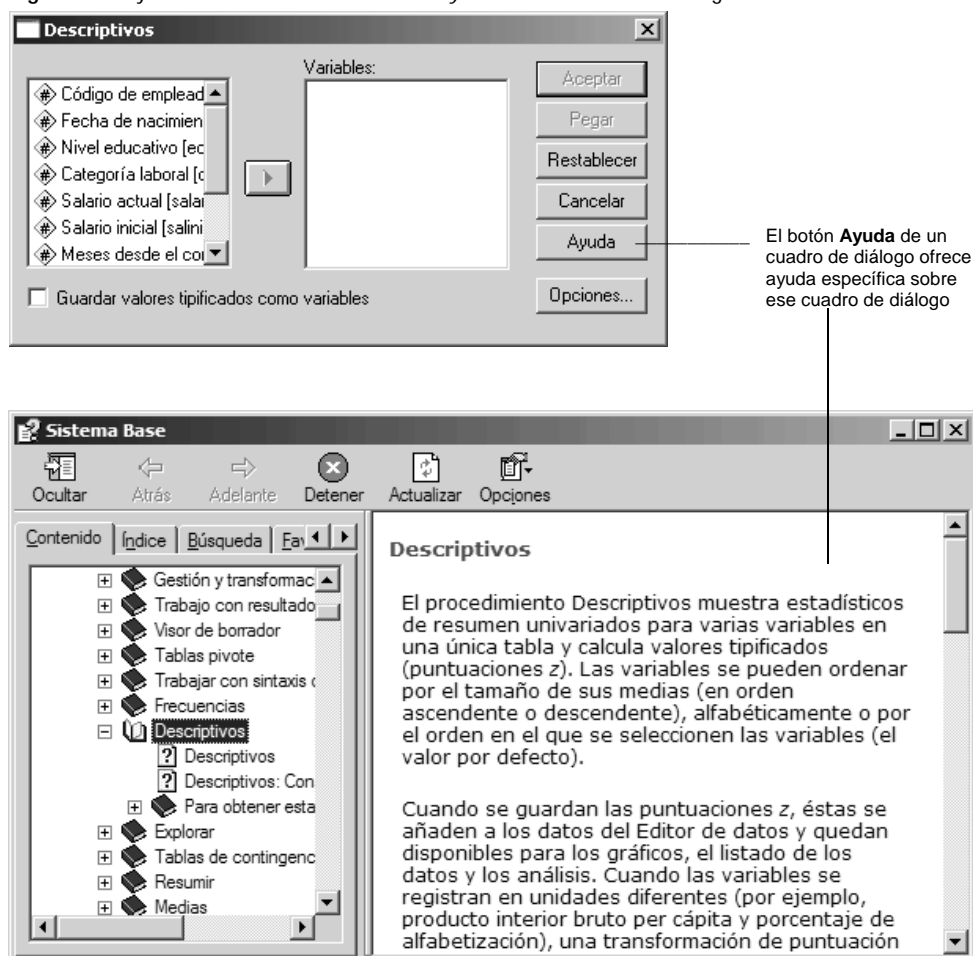
## Estudios de casos

Esta opción del menú *Ayuda* incluye varios ejemplos acerca de cómo aplicar diferentes procedimientos estadísticos (están incluidos prácticamente todos los procedimientos SPSS) y cómo interpretar correctamente los resultados.

## Los botones de ayuda

Todos los cuadros y subcuadros de diálogo poseen un botón **Ayuda** mediante el cual es posible obtener información específica sobre el procedimiento SPSS al que se refiere el cuadro de diálogo. Este botón **Ayuda** conduce directamente al sistema de ayuda del SPSS, que no es otro que el descrito en las Figuras 2.1 a la 2.3. La Figura 2.7 muestra un ejemplo de la ayuda obtenida al pulsar el botón **Ayuda** en el cuadro de diálogo *Descriptivos*.

Figura 2.7. Ayuda obtenida mediante el *botón ayuda* de los cuadros de diálogo



## Archivos de datos

Las opciones del menú **Archivo** permiten abrir, guardar, imprimir, exportar, etc., el contenido de las diferentes ventanas del SPSS. Este capítulo describe las opciones disponibles en el menú **Archivo** del *Editor de datos* (opciones éstas esencialmente referidas a los archivos de datos). Este menú también incluye la opción **Salir**.

### Archivos nuevos

La opción **Nuevo** del menú **Archivo** crea un archivo (ventana) nuevo. Para que esta acción tenga efecto es necesario seleccionar el tipo de ventana que se desea crear:

- **Datos.** Vacía el contenido del *Editor de datos* y lo deja preparado para introducir nuevos datos o para abrir un archivo de datos existente (ver Capítulos 3 y 4).
- **Sintaxis.** Abre una ventana del *Editor de sintaxis* (ver Capítulo 8).
- **Resultados.** Abre el *Visor de resultados*. Las ventanas o archivos de resultados recogen la información que genera el SPSS: tablas, gráficos, etc. (ver Capítulo 7).
- **Resultados de borrador.** Abre el *Visor de resultados* en formato *borrador*.
- **Proceso.** Abre el *Editor de procesos* del SPSS, el cual permite crear archivos capaces de personalizar algunos aspectos del funcionamiento del programa.

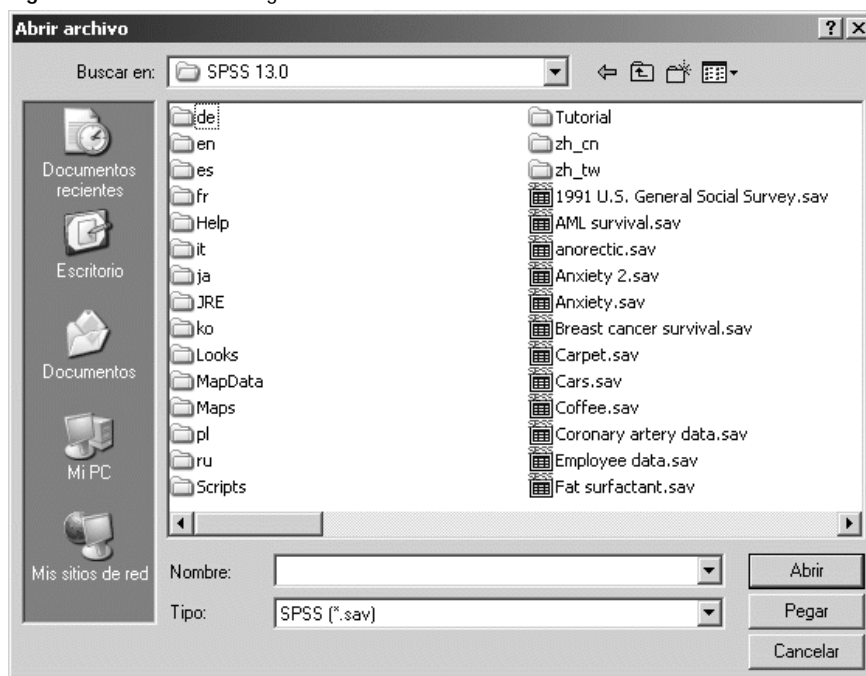
En el *Editor de datos* sólo es posible tener abierto un archivo de datos, pero es posible abrir más de un *Editor de datos* simultáneamente. También pueden abrirse simultáneamente varias ventanas del *Visor* y del *Editor de sintaxis*; ahora bien, en el caso de tener abiertas varias ventanas del mismo tipo, sólo una de ellas actúa como ventana *designada* (ver Capítulo 1).

### Abrir archivos de datos

Abre un archivo guardado en disco en cualquiera de los formatos SPSS o en otros formatos como Excel, Lotus, dBase, etc. Para abrir un archivo:

- Seleccionar la opción **Abrir** del menú **Archivo** y marcar una de las opciones del menú emergente (**Datos...**, **Sintaxis...**, **Resultados...**, **Proceso...**, **Otro...**) para acceder al cuadro de diálogo *Abrir archivo* que muestra la Figura 3.1.



Figura 3.1. Cuadro de diálogo *Abrir archivo*

El cuadro de diálogo ofrece un listado de los archivos cuya extensión se corresponde con el tipo de archivo seleccionado en la opción **Abrir**. Si se ha decidido abrir un archivo de datos, los archivos listados serán archivos de datos (archivos con extensión *.sav*). Si se ha optado por abrir un archivo de resultados, los archivos listados tendrán extensión *.spo*. Si se ha optado por abrir un archivo de sintaxis, los archivos listados tendrán extensión *.sps*. Etc.

**Buscar en.** La carpeta abierta por defecto es la carpeta en la que se encuentra instalado el SPSS. Si el archivo buscado se encuentra en otro lugar, el menú desplegable **Buscar en** permite indicar la unidad y la carpeta en la que se encuentra el archivo que se desea abrir.

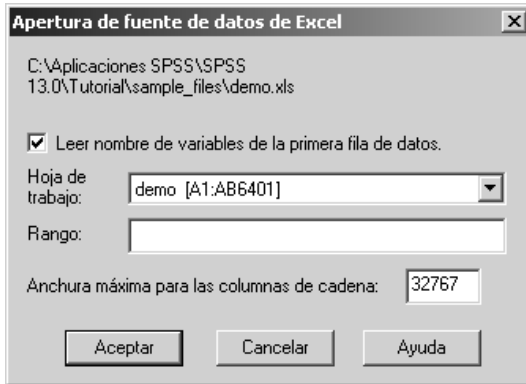
**Nombre.** Para abrir un archivo hay que seleccionarlo de la lista (o teclear su nombre en el cuadro de texto **Nombre**) y pulsar el botón **Abrir**.

**Tipo.** Independientemente del tipo de archivo por el que se haya optado al entrar en el cuadro de diálogo *Abrir archivo*, el menú desplegable **Tipo** permite seleccionar archivos de datos de diferentes formatos:

- **SPSS (\*.sav).** Archivos creados con el SPSS para Windows, Macintosh o UNIX.
- **SPSS/PC+ (\*.sys).** Archivos creados con el SPSS para MS-DOS.
- **SYSTAT.** Archivos de datos del programa estadístico SYSTAT.
- **SPSS portable (\*.por).** Archivos salvados desde SPSS en formato transportable.
- **Lotus (\*.w\*).** Archivos de la hoja de cálculo Lotus 1-2-3. Este tipo de archivos se abre siguiendo las mismas reglas que se utilizan para abrir archivos Excel (ver a continuación).

- SYLK (\*.slk). Archivos de hojas de cálculo (Excel, Multiplan, etc.) guardados en formato SYLK. Los archivos SYLK se abren siguiendo las mismas reglas que se describen a continuación para los archivos Excel.
- Excel (\*.xls). Archivos de la hoja de cálculo Microsoft Excel. Al intentar abrir un archivo con formato Excel, el SPSS muestra el subcuadro de diálogo *Apertura de fuente de datos de Excel* que recoge la Figura 3.2.

Figura 3.2. Subcuadro de diálogo *Apertura de fuente de datos de Excel*



Este cuadro contiene varias opciones que permiten identificar la *hoja de trabajo* que se desea abrir y el rango de casillas:

“ **Leer nombres de variable de la primera fila de datos.** Si se utiliza esta opción, el SPSS toma, como nombres para las variables del nuevo archivo, el contenido de los campos de la primera fila de la hoja de cálculo. En ese caso, si algún campo de la primera fila está vacío, el SPSS le asigna un nombre por defecto (V#). Si no se utiliza esta opción, el SPSS asigna a cada columna los siguientes nombres de variables: V1, V2, V3, etc.

**Hoja de trabajo.** Puesto que a partir de la versión 5 de Excel es posible definir más de una *hoja* dentro del mismo *libro*, la opción **Hoja de trabajo** permite seleccionar la hoja que se desea abrir. Para poder trabajar en el SPSS con varias hojas de Excel (en el caso de que esto tuviera sentido), sería necesario crear un archivo de datos para cada hoja y, tras esto, fundir en uno los archivos correspondientes a las distintas hojas (ver, en el Capítulo 6, el apartado *Fundir archivos*).

**Rango.** Esta opción permite leer sólo un rango determinado de celdas de la hoja de cálculo. Si no se indica nada, se leen todas las celdas del archivo. El rango de celdas se indica con la letra de la primera columna, el número de la primera fila, dos puntos, la letra de la última columna y el número de la última fila; por ejemplo: A1:D16.

Al abrir un archivo con formato Excel las filas se convierten en casos y las columnas en variables. El tipo de formato de las nuevas variables viene definido por el tipo de formato de cada columna en Excel. Si existen columnas mixtas (con casillas numéricas y alfanuméricas), el SPSS asigna a las nuevas variables formato de *cadena* (ver siguiente capítulo). Las celdas en blanco (vacías) de las columnas con formato numérico se convierten

en valores perdidos. Pero si las celdas vacías tienen formato de cadena, los espacios en blanco se consideran valores válidos y son tratados como tales.

Por supuesto, también puede trasladarse información parcial o total desde Excel o desde otras aplicaciones Windows hasta el SPSS *cortando y copiando* las celdas que contienen la información de interés (con las opciones de edición *cortar* y *copiar* propias de todas las aplicaciones que funcionan en entorno Windows). El problema de esta forma de proceder es que se pierden los nombres de las variables.

- **SAS.** Archivos de distintas versiones del programa estadístico SAS: archivos de las versiones 7 y 8 para Windows con extensión larga (SAS Long File Name); archivos de las versiones 7 y 8 para Windows con extensión corta (SAS Short File Name); archivos de la versión 6 para Windows y OS2 (SAS v6 for Windows); archivos de la versión 6 para Unix (SAS v6 for UNIX); y archivos en formato portátil o comprimido (SAS Transport).
- **dBase (\*.dbf).** Archivos de la base de datos dBase II, III, III+ y IV. Al abrir un archivo dBase, los registros se convierten en casos y los campos en variables. Cuando el nombre de un campo excede de 8 caracteres, el SPSS crea el nombre de la nueva variable con los 8 primeros; si estos 8 primeros caracteres no generan un nombre de variable único, el SPSS deja fuera ese campo. Los dos puntos utilizados en los nombres de los campos del dBase se transforman en caracteres de subrayado en el SPSS. Por otro lado, los registros marcados en dBase para ser eliminados pero todavía no eliminados definitivamente se leen como casos válidos. El SPSS crea una nueva variable de cadena (ver siguiente capítulo) llamada *D\_R* y asigna asteriscos a los casos correspondientes a registros marcados.
- **Texto.** Archivos de texto con formato ASCII (ver más adelante, en este mismo capítulo, el apartado *Leer datos de texto*).

Por supuesto, además de archivos de datos en todos estos formatos recién mencionados, la opción **Abrir** del menú **Archivo** también permite abrir los distintos tipos de archivos SPSS: *sintaxis*, *documentos del Visor*, *documentos del Visor borrador* y *procesos de SPSS* (el Capítulo 1 contiene una descripción de estos archivos).

## Abrir bases de datos

Esta opción permite abrir bases de datos y hojas de cálculo de varios formatos siempre que se disponga del controlador ODBC apropiado. Los controladores ODBC pueden instalarse desde el CD-ROM del SPSS: los controladores del **Data Access Pack** de SPSS se instalan automáticamente al instalar el SPSS; el usuario puede decidir instalar, además, los controladores para productos Microsoft incluidos en el **Data Access Pack** de Microsoft. Para abrir una base de datos:

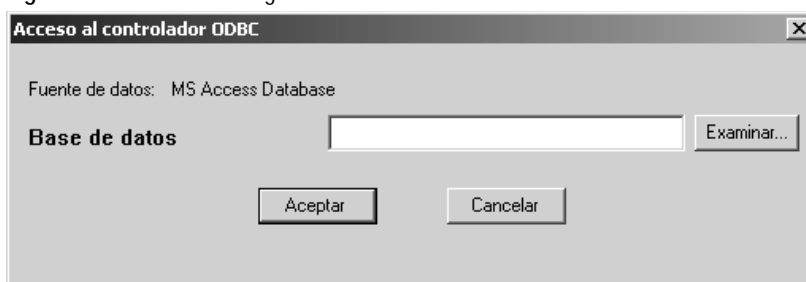
1. Seleccionar la opción **Abrir bases de datos > Nueva consulta...** del menú **Archivo** para acceder al cuadro de diálogo *Asistente para bases de datos* que muestra la Figura 3.3.

Este primer cuadro de diálogo muestra un listado con los distintos tipos de formatos que es posible importar mediante el *Asistente para bases de datos*. El contenido de este listado depende de los controladores que cada usuario tenga instalados en el ordenador. En el ejemplo de la Figura 3.3 aparecen listados tres tipos de formato: *dBase*, *Excel* y *MS Access*.

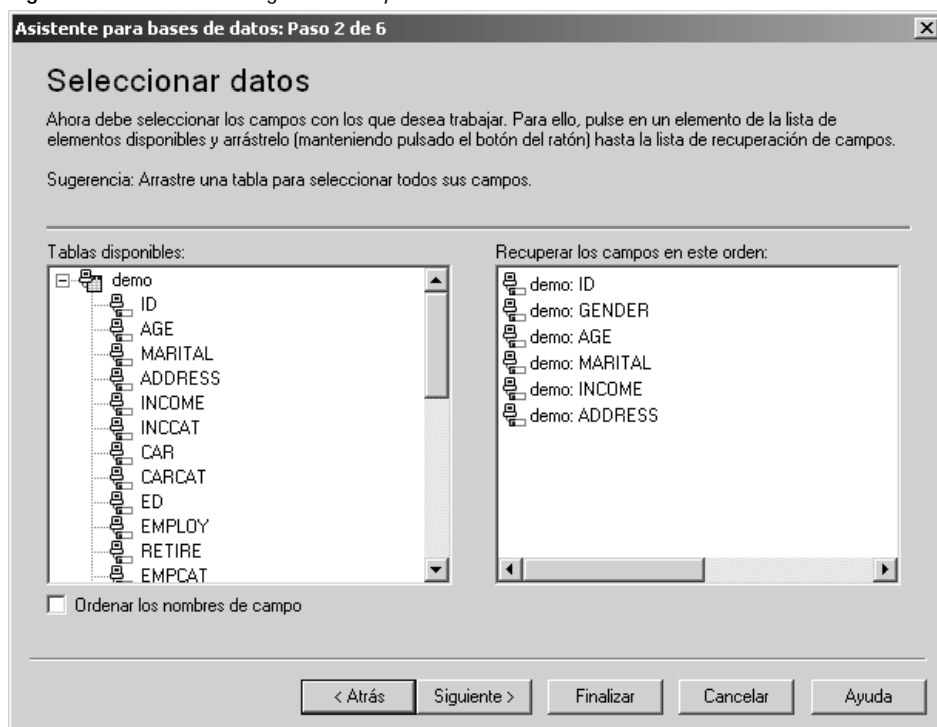
Figura 3.3. Cuadro de diálogo *Asistente para bases de datos*

Para continuar con el paso siguiente:

- Seleccionar el tipo de archivo que se desea abrir y pulsar el botón **Siguiente>** para acceder al cuadro de diálogo *Acceso al controlador ODBC* que muestra la Figura 3.4.

Figura 3.4. Cuadro de diálogo *Acceso al controlador ODBC*

Este cuadro de diálogo permite seleccionar el archivo concreto en el que se encuentran los datos. Para facilitar la tarea de nombrar ese archivo, el botón **Examinar...** permite buscar la ubicación exacta en la que se encuentra la base de datos. Una vez nombrado el archivo, el botón **Aceptar** permite acceder al cuadro de diálogo de selección de datos (paso 2 de 6) que muestra la Figura 3.5.

Figura 3.5. Cuadro de diálogo *Asistente para bases de datos: Seleccionar datos*

Este cuadro de diálogo permite decidir qué campos (variables) se desea incluir en el nuevo archivo de datos y el orden en el que deben aparecer.

**Tablas disponibles.** Esta lista recoge las tablas de datos de que consta la base de datos seleccionada. En el ejemplo, se ha elegido una base de datos llamada *demo.mdb* que se incluye entre los ejemplos que el SPSS graba durante la instalación; se encuentra en la carpeta *Tutorial > Sample\_files*).

Al entrar en el cuadro de diálogo *Seleccionar datos*, la tabla *demo* aparece contraída (no muestran su contenido). La primera acción que debe llevarse a cabo es la de expandir o desplegar el contenido de la tabla cuyos datos se desea importar. Esto se hace pulsando sobre el signo más (+) que precede al nombre de la tabla. Una vez desplegados los campos, el signo más se convierte en signo menos (–); pulsando ahora este signo, el contenido de la tabla vuelve a contraerse (ocultarse).

**Recuperar campos en este orden.** Seleccionada la tabla que se desea abrir, debe concretarse qué campos pasarán a ser las variables del nuevo archivo de datos. Para seleccionar un campo hay que trasladarlo desde el cuadro *Tablas disponibles* al cuadro *Recuperar campos en este orden*. Esto se consigue situando el puntero del ratón en el campo deseado y pulsando dos veces el botón principal del ratón; o seleccionando un campo y arrastrándolo con el puntero del ratón. Para seleccionar todos los campos de una tabla hay que pulsar dos veces en el nombre de la tabla o arrastrar el nombre de la tabla. Una vez seleccionado un campo, puede deshacerse la

selección repitiendo la misma operación pero en sentido inverso; es decir, pulsando dos veces sobre el nombre de ese campo en el cuadro **Recuperar los campos en este orden**, o arrastrando el campo fuera del recuadro.

“ **Ordenar los nombres de campo.** Activando esta opción, el SPSS ordena alfabéticamente los campos seleccionados.

Tras seleccionar los campos que se van a recuperar, el botón **Finalizar** permite leer la base de datos y construir el correspondiente archivo de datos SPSS. Pero de esta forma se importa toda la base de datos (todos los casos del archivo). Si se desea importar sólo parte de los casos, debe pulsarse el botón **Siguiente>** para acceder al cuadro de diálogo de selección de casos (paso 4 de 6) que muestra la Figura 3.6.

Figura 3.6. Cuadro de diálogo *Asistente para bases de datos: Limitar la recuperación de casos*

**Asistente para bases de datos: Paso 4 de 6**

### Limitar la recuperación de casos

Es posible limitar el número de casos recuperados especificando uno o varios criterios.

Sugerencia: Para añadir campos y funciones a una expresión, puede arrastrarlos y colocarlos en la casilla de la expresión.

**Campos:**

- demo
  - ID
  - AGE
  - MARITAL
  - ADDRESS
  - INCOME

**Funciones:**

- ASCII(expr\_cadena)
- CHAR(código ASCII)
- CONCAT(expr\_cadena)
- LEFT(expr\_cadena,rec)
- LTRIM(expr\_cadena)
- LENGTH(expr\_cadena)

Pedir el valor al usuario...

**Criterios:**

	Expresión 1	Relación	Expresión 2
1	demo: GENDER	=	h
2	demo: AGE	>	18
3			
4			
5			
6			

☒ Utilizar muestreo aleatorio

☐ Muestreo aleatorio nativo

☒ Muestreo aleatorio de SPSS

☒ Aproximadamente  % de todos los casos

☐ Exactamente  casos desde el primero  casos

< Atrás Siguiente > Finalizar Cancelar Ayuda

Este cuadro de diálogo permite establecer las condiciones (filtro) bajo las cuales un *registro* de la base de datos pasará a ser un *caso* en el nuevo archivo de datos SPSS.

**Campos.** Ofrece un listado de todos los campos (variables) de la base de datos, independientemente de que hayan sido o no elegidos para ser importados.

**Criterios.** Este recuadro está diseñado para facilitar al usuario la definición de los criterios de selección. Cada criterio de selección consta de dos *expresiones* y de una *relación* entre ellas. Las expresiones pueden incluir nombres de campo (incluidos los no seleccionados para ser

importados), constantes, operadores aritméticos, variables lógicas y funciones de todo tipo (aritméticas, lógicas, de cadena, de fecha y de hora; ver, en el Capítulo 5, el apartado *Calcular: Expresiones condicionales*).

La relación entre las expresiones se establece, la mayor parte de las veces, mediante uno de los seis operadores relacionales:  $<$ ,  $>$ ,  $<=$ ,  $>=$ ,  $=$  y  $<>$ . Para crear un criterio es necesario introducir al menos dos expresiones y una relación entre ellas.

Si se desea utilizar más de un criterio es necesario conectarlos mediante los operadores lógicos *and* y *or*.

Para introducir el valor de cada casilla puede utilizarse el teclado o el menú desplegable asociado a cada casilla. En el ejemplo de la Figura 3.6 se han introducido algunos criterios de selección: que la variable *género* sea igual a *h* y que la variable *edad* sea mayor que 18.

“ **Utilizar muestreo aleatorio.** Esta opción permite limitar la importación de casos seleccionando únicamente una muestra aleatoria.

**Muestreo aleatorio nativo.** La selección de la muestra aleatoria se realiza en la base de datos original (no todas las aplicaciones pueden hacer esto).

**Muestreo aleatorio de SPSS.** La selección de casos se basa en el generador de números aleatorios del SPSS (ver Capítulo 6, apartado *Seleccionar casos*).

**Aproximadamente \_\_\_ % de todos los casos.** Selecciona, aproximadamente, el porcentaje de casos indicado en el cuadro de texto.

**Exactamente \_\_\_ casos de los primeros \_\_\_ casos.** Selecciona exactamente el número de casos indicado en el primer cuadro de texto. Selecciona los casos de entre los  $n$  primeros, siendo  $n$  el número definido por el usuario en el segundo cuadro de texto. Este número debe ser menor o igual que el número de casos del archivo de datos; si es mayor, el tamaño de la muestra aleatoria seleccionada será menor que el solicitado.

**Pedir el valor al usuario.** Esta opción permite crear una consulta con parámetros. Cuando se ejecuta una consulta con parámetros, el proceso de importación se detiene en un momento dado para solicitar al usuario que indique el valor que debe adoptar una determinada expresión.

Si se crea una consulta con parámetros, la misma consulta puede utilizarse, por ejemplo, para importar sólo los casos que cumplen la condición *A* (por ejemplo, *hombres*) o sólo los que cumplen la condición *B* (por ejemplo, *mujeres*). El proceso de ejecución de una consulta de este tipo se detiene en el momento preciso para solicitar al usuario que decida si desea importar los casos que cumplen la condición *A* o los que cumplen la condición *B*. Por supuesto, en una misma consulta pueden definirse más de dos condiciones. Para crear una consulta con parámetros:

- Introducir en la casilla **Expresión1** la variable que se desea utilizar como filtro (por ejemplo, «género»).
- Introducir en la casilla **Relación** el operador que debe relacionar **Expresión1** con **Expresión2** (por ejemplo, el signo « $\Rightarrow$ »).
- Situar el cursor en la casilla **Expresión2** y pulsar el botón **Pedir el valor al usuario...** para acceder al subcuadro de diálogo *Pedir el valor al usuario* que muestra la Figura 3.7.

Figura 3.7. Subcuadro de diálogo *Pedir el valor al usuario*

El texto del cuadro **Cadena de petición** es justamente el mensaje que más tarde se ofrece solicitando información al usuario. El mensaje por defecto es *Introduzca el valor*, pero puede introducirse cualquier otro mensaje. Cuando se ejecuta una consulta con parámetros, el proceso se detiene para solicitar al usuario que introduzca un valor mediante el cuadro de diálogo *Seleccionar el valor para la consulta* que muestra la Figura 3.8.

Figura 3.8. Cuadro de diálogo *Seleccionar el valor para la consulta*

El procedimiento exige introducir un valor por defecto (por ejemplo, «h») en el cuadro de texto **Valor predeterminado** (ver Figura 3.7), independientemente de que más tarde se pueda elegir entre distintos valores. Y el cuadro de texto **Permitir al usuario seleccionar el valor de la lista** (ver Figura 3.7) permite introducir los valores entre los que más tarde se podrá optar al recibir el mensaje *Introduzca el valor*; los valores guardados en esta lista son los que más tarde se presentan en el menú desplegable de la Figura 3.8.

El botón **Finalizar** (ver Figura 3.6) permite leer la base de datos y crear el nuevo archivo de datos en el *Editor de datos* del SPSS. Los nombres de las variables del nuevo archivo se generan automáticamente a partir de los nombres de cada campo; si estos nombres no son válidos (ver, en el siguiente capítulo, el apartado *Definir variables: Asignar nombre a una variable*), se asigna un nombre formado por el prefijo *var* y un número secuencial.

Antes de comenzar la captura de datos, existe la posibilidad de modificar el nombre de las variables que formarán parte del nuevo archivo. Si en lugar de pulsar el botón **Finalizar** se pulsa el botón **Siguiente>** se accede a un cuadro de diálogo (*Definir las variables*) que permite



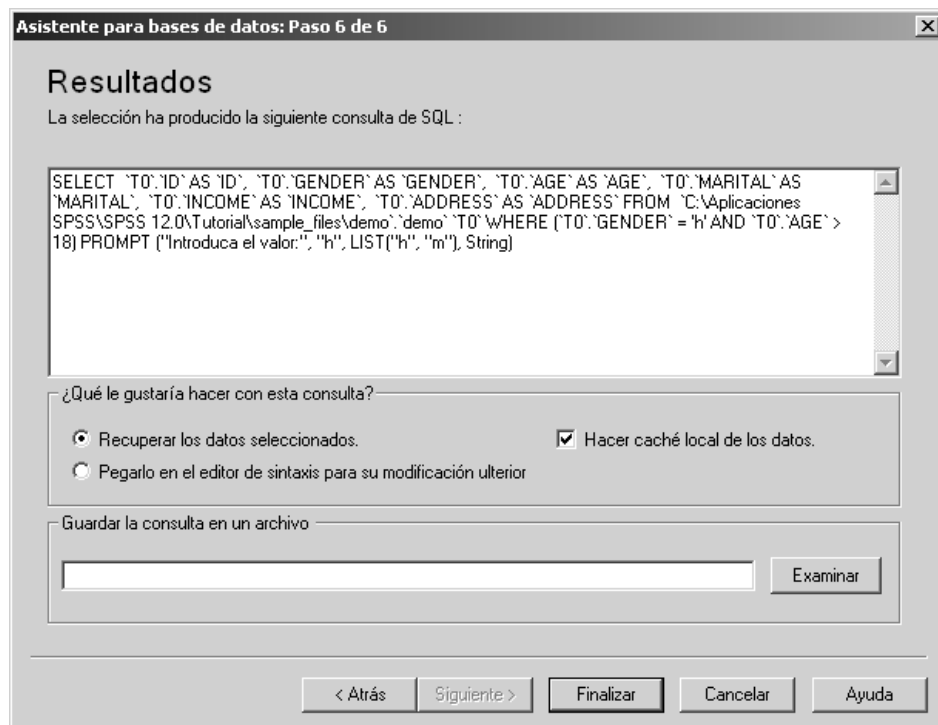
cambiar los nombres de las variables y, en el caso de que exista alguna variable no numérica (o sea, de cadena; ver, en el siguiente capítulo, el apartado *Definir variables: Definir el tipo de variable*), cambiarla a numérica conservando los valores originales como etiquetas de valor (ver, en el siguiente capítulo, el apartado *Definir variables: Asignar etiquetas*).

Por último, el usuario algo avanzado puede estar interesado en consultar la sintaxis SPSS correspondiente a las elecciones hechas. En ese caso, el botón **Siguiente >** permite acceder al cuadro de diálogo *Resultados* (paso 6 de 6) que muestra la Figura 3.9. El cuadro **La selección ha generado la siguiente consulta de SQL** muestra la sintaxis en la que el SPSS se basará para llevar a cabo la captura de la base de datos. Esta sintaxis admite cualquier tipo de modificación desde el teclado y puede pegarse en un archivo de sintaxis (ver Capítulo 8).

Si se elige la opción **Recuperar los datos seleccionados** del recuadro *¿Qué le gustaría hacer con esta consulta?*, el botón **Finalizar** inicia la captura de la base de datos (sin guardar la sintaxis ni la consulta). Si se elige la opción **Pegarla en el editor de sintaxis para su modificación ulterior**, el botón **Finalizar** abre el *Editor de sintaxis* y pega en él la sintaxis correspondiente a las elecciones hechas. Con esta segunda opción activa no se inicia la captura de la base de datos: para iniciar la captura es necesario ejecutar la sintaxis pegada (ver, en el Capítulo 8, el apartado *Ejecutar sintaxis*).

La opción **Hacer caché local de datos** permite hacer una copia temporal del archivo en el disco duro local, lo cual puede mejorar el rendimiento cuando se está trabajando con un servidor. Y la opción **Guardar la consulta en un archivo** permite guardar la consulta para ejecutarla o editarla más tarde.

Figura 3.9. Cuadro de diálogo *Asistente para bases de datos: Resultados*

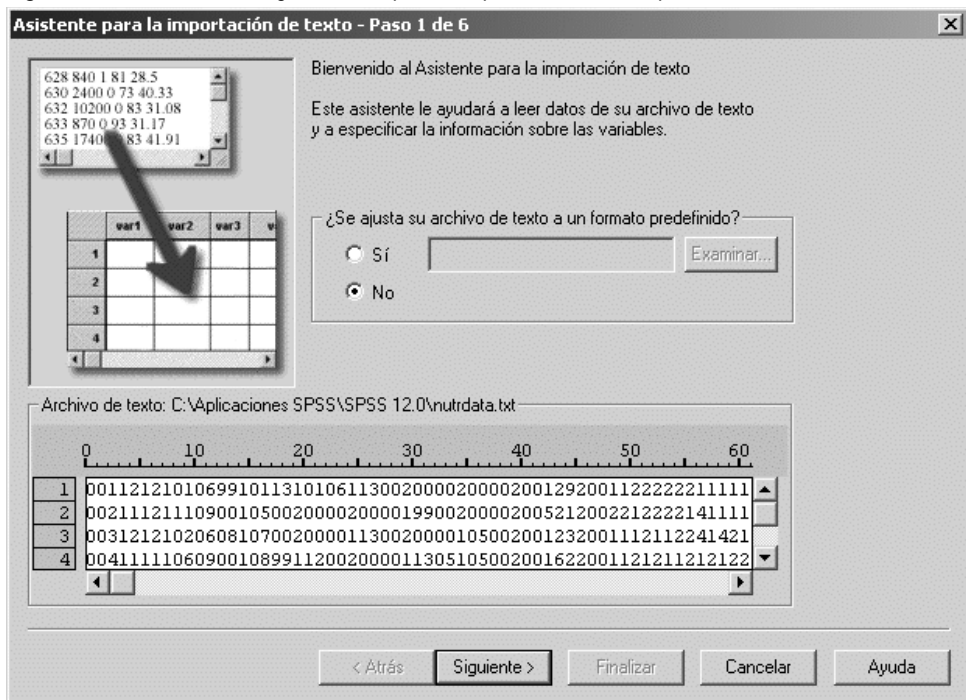


## Leer datos de texto

La opción **Leer datos de texto** del menú **Archivo** permite leer archivos de datos con formato de texto ASCII. Esto significa que, puesto que la mayor parte de las aplicaciones informáticas (procesadores de texto, bases de datos, hojas de cálculo, etc.) son capaces de grabar datos en formato ASCII, siempre será posible importar un archivo de datos independientemente de la aplicación con la que haya sido creado.

Al seleccionar la opción **Leer datos de texto**, el SPSS muestra el cuadro de diálogo **Abrir archivo** (ver Figura 3.1) para permitir seleccionar el archivo de texto que se desea leer. Una vez seleccionado el archivo, el botón **Abrir** conduce al cuadro de diálogo *Asistente para la importación de texto* (paso 1 de 6) que muestra la Figura 3.10.

Figura 3.10. Cuadro de diálogo *Asistente para la importación de texto* (paso 1 de 6)



### Paso 1

El *Asistente para la importación de texto* (Figura 3.10) ofrece, en **Archivo de texto**, una vista previa de los datos del archivo, junto con dos reglas a modo de guía, una vertical y otra horizontal, que permiten identificar la posición exacta (fila y columna) de cada dígito.

¿Se ajusta su archivo de texto a un formato predefinido? Si el archivo de texto se ajusta a un formato predefinido (previamente guardado desde el *Asistente para la importación de texto*), se

puede marcar la opción **Sí** de este recuadro y, a continuación, utilizar el botón **Examinar...** para buscar el archivo en el que se guardó ese formato. Si no existe tal formato predefinido, se debe marcar la opción **No** (viene marcada por defecto).

Tras seleccionar la opción **Sí** (o **No**), el botón **Siguiente...** conduce al segundo paso del cuadro de diálogo *Asistente para la importación de texto*. El aspecto del *Asistente* en este segundo paso sigue siendo idéntico al del primero, pero añade algunas opciones nuevas.

## Paso 2

¿Cómo están organizadas sus variables? Para leer los datos correctamente, es necesario indicar el lugar en el que empiezan y terminan los valores de cada variable. La forma de hacer esto depende de cómo se encuentren las variables en el archivo de texto:

- **Delimitadas:** separadas por espacios, comas, tabulaciones u otros caracteres. En los archivos delimitados (también llamados archivos con *formato libre*), las variables se encuentran en el mismo orden para todos los casos, pero no necesariamente en la misma posición vertical.
- **Ancho fijo:** cada variable ocupa las mismas columnas (la misma posición) del mismo registro (fila) en todos los casos del archivo de datos. No se requiere la presencia de un carácter delimitador entre valores. La posición exacta de cada valor es la que determina qué valor se está leyendo.

¿Están incluidos los nombres de las variables en la parte superior del archivo? Si la primera fila del archivo de texto contiene etiquetas descriptivas de las variables, puede utilizarse la opción **Sí** para que el SPSS lea correctamente esas etiquetas y las utilice como nombres de variables. Si la etiqueta descriptiva de una variable tiene más de 64 caracteres, sólo se toman los 64 primeros. Y si esos 64 primeros caracteres no crean un nombre de variable único, el SPSS crea un nombre nuevo para esa variable.

## Paso 3

Tras decidir cómo están organizadas las variables y si la primera fila contiene los nombres de las variables, el botón **Siguiente...** conduce al paso 3 del *Asistente para la importación de texto*. Las opciones del paso 3 son distintas dependiendo de que en el paso 2 se hayan definido variables *delimitadas* o variables de *ancho fijo*. Si se han definido variables de *ancho fijo*, aparecen las opciones que muestra la Figura 3.11.

¿En qué número de línea comienza el primer caso de los datos? Este cuadro de texto permite indicar en qué número de línea del archivo de texto comienzan los datos correspondientes al primer caso.

¿Cuántas líneas representan un caso? Puesto que un caso puede ocupar más de una línea del archivo de texto, esta opción permite hacer explícito el número exacto de líneas que ocupa cada caso. El número de líneas determina el final de un caso y el comienzo del siguiente.

¿Cuántos casos desea importar? Estas opciones permiten decidir si se desea leer todo el archivo de texto o sólo una parte: los primeros *n* casos o un determinado porcentaje.

Figura 3.11. Cuadro de diálogo *Asistente para la importación de texto: Ancho fijo* (paso 3 de 6)

Asistente para la importación de texto: Ancho fijo - Paso 3 de 6

¿En qué número de línea comienza el primer caso de los datos? 1

¿Cuántas líneas representan un caso? 1

¿Cuántos casos desea importar?

☒ Todos los casos

☐ Los primeros 1000 casos.

☐ Un porcentaje de los casos: 10 %

Vista previa de datos:

	0	10	20	30	40	50	60
1	001121210106991011310106113002000020000200129200112222211111						
2	0021112111090010500200002000019900200002005212002212222141111						
3	0031212102060810700200001130020000105002001232001112112241421						
4	00411111060900108991120020000113051050020016220011212112122						

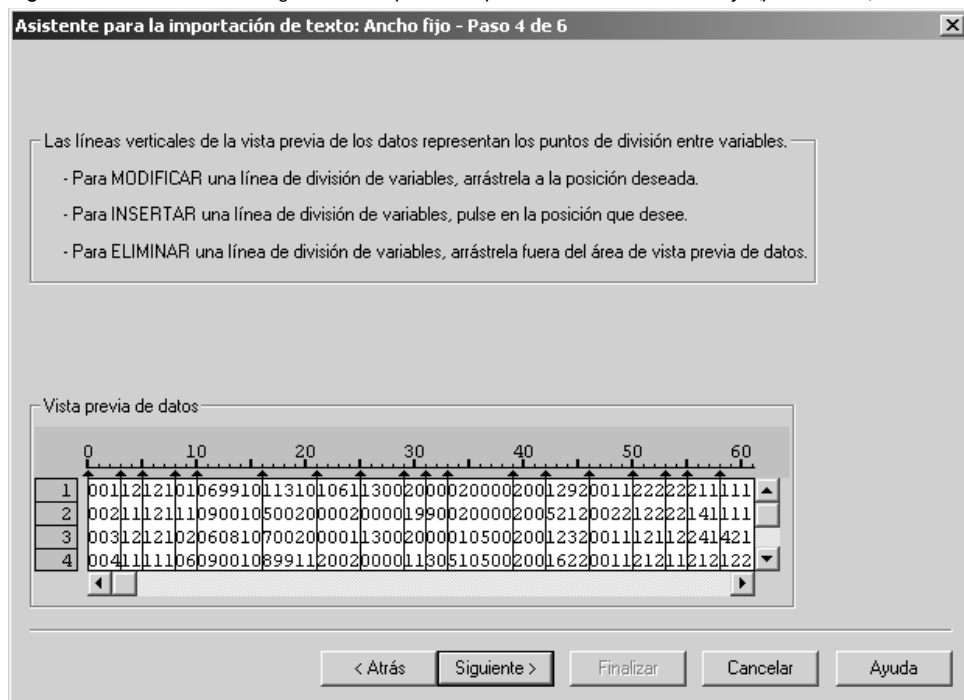
< Atrás    **Siguiente >**    Finalizar    Cancelar    Ayuda

Si en el paso 2, en lugar de variables de *ancho fijo*, se han definido variables *delimitadas*, en lugar de la opción *¿Cuántas líneas representan un caso?* aparece la opción *¿Cómo se encuentran representados los casos?*, la cual permite indicar si cada caso ocupa una línea distinta o viene definido por un número concreto de variables. En este último caso, cuando el SPSS comienza la lectura, asigna el primer valor que encuentra a la primera variable definida; el segundo valor, a la segunda variable; el tercer valor, a la tercera variable, etc. Cuando ha leído tantos valores como variables definidas, el siguiente valor pasa a ser el segundo valor (el valor del segundo caso) de la primera variable; el siguiente valor, el segundo de la segunda variable; el siguiente valor, el segundo de la tercera variable; etc. Por tanto, el SPSS determina el final de un caso y el comienzo del siguiente a partir del número de variables definidas. Por tanto, no debe quedar sin definir ninguna de las variables del archivo; si se omite alguna, los datos se leerán incorrectamente (cuando ocurre esto, aparece un mensaje de aviso).

## Paso 4

Si se está trabajando con variables de *ancho fijo*, las opciones del paso 4 (ver Figura 3.12) permiten definir el ancho concreto de cada variable del archivo. Para ello, basta con pulsar con el puntero del ratón en la columna apropiada dentro del recuadro *Vista previa de datos*.

Si se está trabajando con variables *delimitadas* (datos ASCII con formato *libre*), las opciones del *Asistente para la importación de texto* permiten, al llegar al paso 4, seleccionar el carácter utilizado como separador de variables.

Figura 3.12. Cuadro de diálogo *Asistente para la importación de texto: Ancho fijo* (paso 4 de 6)

## Paso 5

El siguiente paso permite, en el caso de que así se desee, cambiar el nombre o el formato a las variables recién definidas (ver Figura 3.13). Si el archivo de texto que se desea importar contiene en la primera fila los nombres de las variables y en el paso 2 se ha marcado la correspondiente opción, los nombres que el SPSS muestra en este cuadro de diálogo son los que toma del propio archivo de texto (truncados a 40 caracteres en caso necesario y utilizando los nombres *V1*, *V2*, etc., si al truncar no resulta un nombre válido y único). Si el archivo de texto no contiene los nombres de las variables, el SPSS asigna, como nombres de variables, *V1* a la primera, *V2* a la segunda, *V3* a la tercera, etc.

Las opciones del paso 5 permiten cambiar esos nombres (en el cuadro de texto **Nombre**) y asignar un tipo de formato a cada variable (en el menú desplegable **Formato de datos**). No obstante, ésta no es la mejor forma de llevar a cabo esta tarea: en el apartado *Definir variables* del Capítulo 4 se ofrece una descripción detallada de cómo asignar un nombre y un tipo de formato a las variables en el *Editor de datos* del SPSS. Para asignar un mismo formato a más de una variable, pueden seleccionarse varias (pinchando en la cabecera de la primera variable cuyo nombre se desea cambiar y arrastrando el puntero del ratón; o con las teclas de movimiento acompañadas de la tecla de mayúsculas) antes de aplicar el formato deseado.

Entre las opciones del menú **Formato de datos** también se encuentra la opción **No importar**, la cual puede aplicarse a la variable o conjunto de variables que no se desee incluir en el nuevo archivo de datos.

Figura 3.13. Cuadro de diálogo *Asistente para la importación de texto (paso 5 de 6)*

**Asistente para la importación de texto - Paso 5 de 6**

Especificaciones para la(s) variable(s) seleccionada(s) en la vista previa de datos:

Nombre de variable:

Formato de datos:

Vista previa de datos:

V1	V2	V3	V4	V5	V6	V7
001	12	121	01	069910	11310	1061
002	11	121	11	090010	50020	0002
003	12	121	02	060810	70020	0001
004	11	111	06	090010	89911	2002

< Atrás   **Siguiente >**   Finalizar   Cancelar   Ayuda

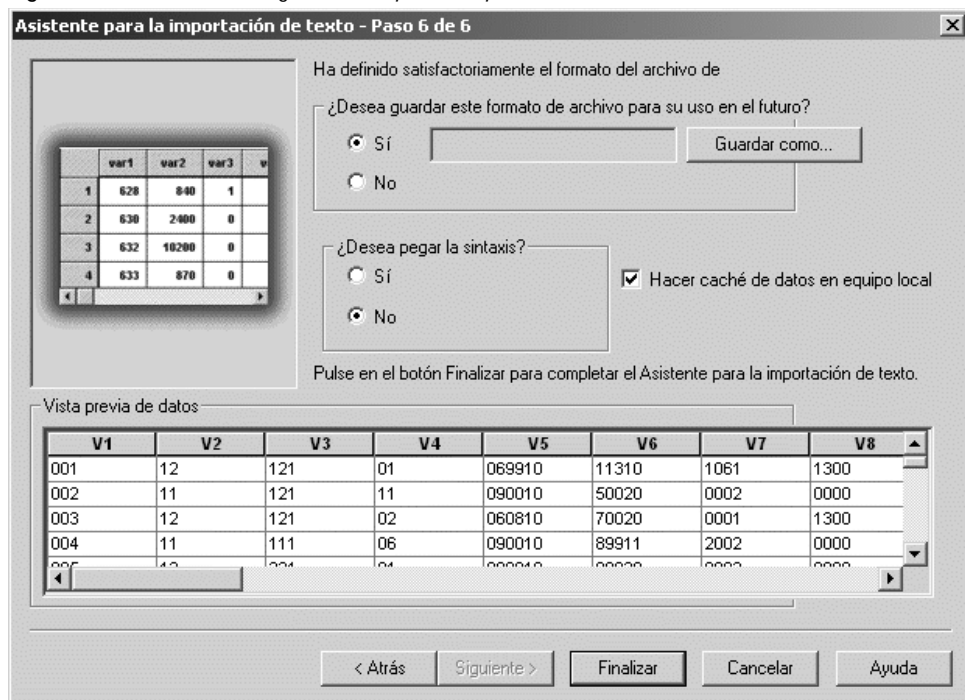
## Paso 6

Una vez definido el nombre y el formato de cada variable, ya se ha completado todo lo relacionado con la definición del archivo de texto. Sin embargo, el *Asistente para la importación de texto* permite controlar un par de detalles más (ver Figura 3.14).

¿Desea guardar este formato de archivo para su uso en el futuro? Si se desea, es posible optar por guardar en un archivo todas las especificaciones correspondientes a las selecciones hechas en los pasos previos. Para ello, basta con seleccionar la opción **Sí** y asignar un nombre al archivo. Este archivo puede utilizarse más tarde simplemente indicando su nombre en el cuadro de texto ¿Se ajusta su archivo de texto a un formato predefinido? del cuadro de diálogo correspondiente al paso 1 del *Asistente para la importación de texto* (ver Figura 3.10).

¿Desea pegar la sintaxis? Activando la opción **No** y pulsando el botón **Finalizar**, el SPSS comienza la lectura del archivo de texto y crea, a partir de él, un archivo de datos en el *Editor de datos* (sustituyendo al archivo activo, si es que existe alguno). Activando la opción **Sí**, el botón **Finalizar** abre el *Editor de sintaxis* y pega en él la sintaxis SPSS correspondiente a las elecciones hechas en los pasos previos.

Conviene señalar que, en el caso de que se opte por pegar la sintaxis, el botón **Finalizar** no inicia la lectura del archivo de texto; para iniciar la lectura es necesario ejecutar la sintaxis desde el *Editor de sintaxis* (puede consultarse, en el Capítulo 8, el apartado *Ejecutar sintaxis* para una explicación de cómo se trabaja con archivos de sintaxis).

Figura 3.14. Cuadro de diálogo *Asistente para la importación de texto - Paso 6 de 6*

## Guardar archivos de datos

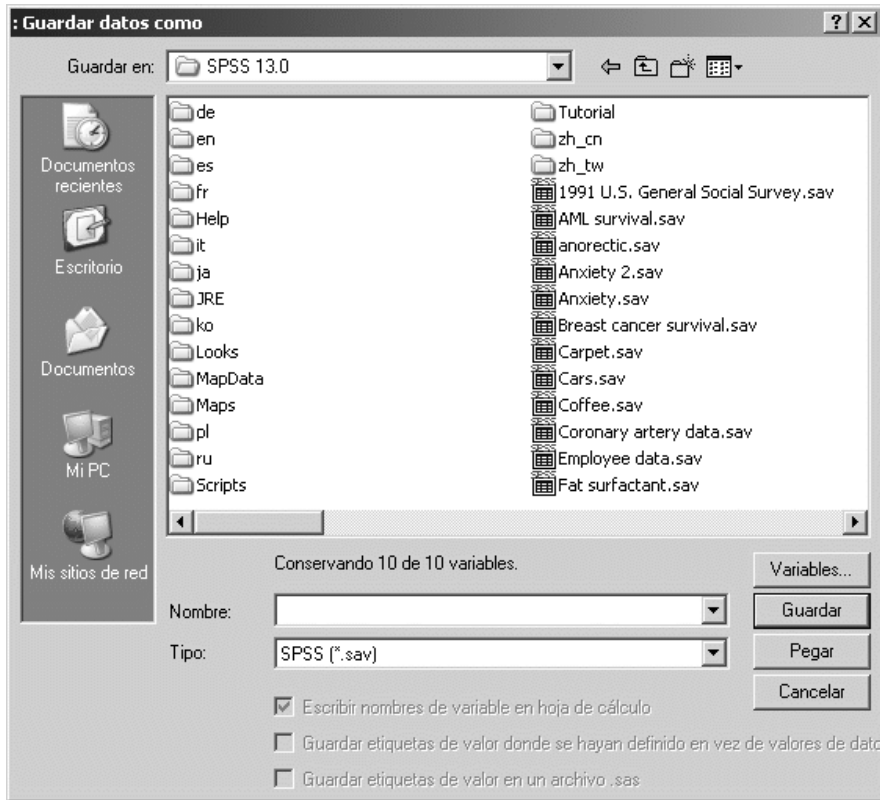
Si el archivo del *Editor de datos* ya tiene nombre pero se desea guardar algunas modificaciones, la opción **Guardar** del menú **Archivo** hace que el archivo del *Editor de datos* sea automáticamente grabado en disco con el mismo nombre. Si el archivo del *Editor de datos* no tiene nombre (o lo que es lo mismo, tiene el nombre que el sistema asigna por defecto: *Sin título*), la opción **Guardar** del menú **Archivo** conduce al cuadro de diálogo *Guardar datos como* que muestra la Figura 3.15.

**Guardar en.** Incluye un menú desplegable que ayuda a buscar la unidad y la carpeta en la que se desea guardar el archivo de datos.

**Nombre.** Cuadro de texto que permite asignar un nombre al archivo que se desea guardar. El SPSS asigna extensiones por defecto a cada tipo de archivo. A los archivos de datos les asigna la extensión *.sav*.

**Tipo.** Menú desplegable que permite seleccionar el formato en el que se desea guardar el archivo de datos. Por defecto, el SPSS guarda los archivos de datos en formato SPSS (extensión *\*.sav*), pero es posible elegir cualquiera de los siguientes formatos:

- **SPSS (\*.sav).** Formato SPSS. Los archivos guardados con este formato no se pueden leer con versiones del SPSS anteriores a la 7.5.

Figura 3.15. Cuadro de diálogo *Guardar datos como*

- **SPSS 7.0 (\*.sav).** Formato SPSS 7.0 para Windows. Los archivos de datos guardados en este formato se pueden leer con versiones anteriores del SPSS para Windows.
- **SPSS/PC+ (\*.sys).** Formato SPSS/PC+. Si el archivo de datos contiene más de 500 variables, sólo se guardan las 500 primeras.
- **SPSS portátil (\*.por).** Los archivos de datos con este formato pueden leerse con versiones del SPSS para otros sistemas operativos (Macintosh, UNIX, etc.).
- **Delimitado por tabulaciones (\*.dat).** Archivos de texto ASCII con los valores separados por tabulaciones.
- **ASCII en formato fijo (\*.dat).** Archivos de texto ASCII con formato fijo: sin tabulaciones ni espacios entre los valores.
- **Excel (\*.xls).** Archivos con formato Excel. Puede elegirse entre uno de los tres siguientes formatos Excel: 2.1, 5.0/95 y 97/2000/XP. Los archivos Excel están limitados a 256 columnas: si el archivo SPSS contiene más de 256 variables, sólo se exportan las 256 primeras. Los archivos Excel 4.0 y Excel 5.0/95 están limitados a 16.384 registros o filas y los archivos Excel 97/2000/XP a 65.536: si el archivo SPSS posee más casos, sólo se exportan los permitidos por Excel.



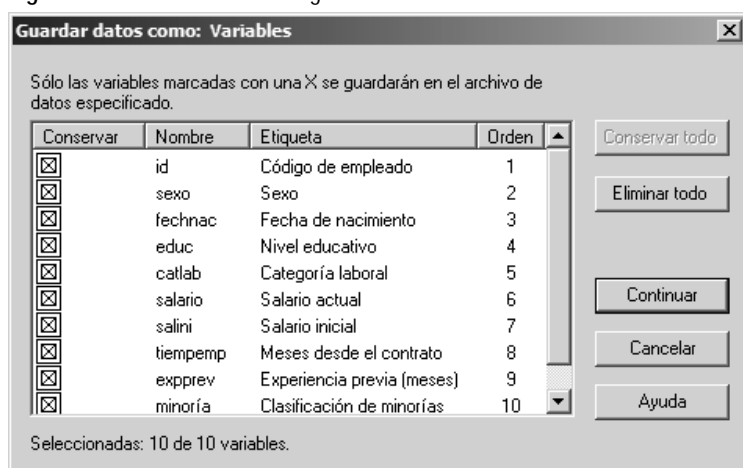
- 1-2-3 versión 3.0 (\*.wk?). Archivos de hoja de cálculo de Lotus 1-2-3. Lotus sólo admite un máximo de 256 variables.
- SYLK (\*.slk). Archivos con formato de vínculo simbólico (utilizado en hojas de cálculo como Excel y Multiplan). Este formato sólo permite guardar 256 variables.
- dBASE II, III y IV (\*.dbf). Archivos con formato dBASE.
- SAS. Archivos con formato SAS v6 para Windows/OS2 (\*.sd2), SAS v6 para UNIX (\*.ssd01), SAS v6 para Alpha/OSF (\*.ssd04), SAS v7-8 para Windows extensión corta (\*.sd7), SAS v7-8 para Windows extensión larga (\*.sas7bdat), SAS v7-8 para UNIX (\*.ssd01), y SAS portátil o transportable (\*.xpt).

Al guardar un archivo con formato SAS se aplica un tratamiento especial a determinadas características de los datos. Algunos caracteres que se permiten en los nombres de variables de SPSS (por ejemplo @, # y \$) no son válidos en SAS; al exportar los datos, estos caracteres no válidos se reemplazan por un carácter de subrayado.

- “ Escribir nombres de variables en hoja de cálculo. Si se ha elegido guardar el archivo de datos en el formato de alguna hoja de cálculo, esta opción hace que los nombres de las variables del archivo de datos pasen a ocupar el primer registro de la hoja de cálculo.
- “ Guardar etiquetas de valor donde se hayan definido en vez de los valores de los datos. Si se ha elegido guardar los datos en el formato de alguna hoja de cálculo, esta opción permite guardar como valores las etiquetas de valor, si existen.
- “ Guardar etiquetas de valor en un archivo .sas. Si se guardan los datos en formato SAS, esta opción permite guardar los valores y sus etiquetas en un archivo de sintaxis SAS.

**Variables.** Si el formato elegido es SPSS, el botón **Variables...** permite acceder al subcuadro de diálogo *Guardar datos como: Variables* que muestra la Figura 3.16, el cual permite elegir las variables que se desea guardar. Marcando y desmarcando la correspondiente casilla de verificación en la columna **Conservar** puede decidirse qué variables guardar. Por defecto, se guardan todas las variables.

Figura 3.16. Subcuadro de diálogo *Guardar datos como: Variables*



Una vez especificados el nombre, la ruta y el formato del archivo que se desea guardar, el botón **Guardar** (ver Figura 3.15) permite guardar el archivo de datos.

## Guardar como

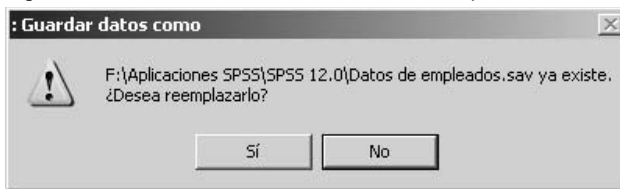
Para guardar el archivo de datos con un nombre nuevo y una ruta nueva (es decir, para cambiar el nombre a un archivo):

- Seleccionar la opción **Guardar como...** del menú **Archivo** para acceder al cuadro de diálogo *Guardar como* (idéntico al cuadro de diálogo de la Figura 3.15).

Este cuadro de diálogo muestra, por defecto, los archivos cuya extensión se corresponde con el formato del archivo que se está intentando guardar. Por tanto, al seleccionar la opción **Guardar como...** desde el *Editor de datos*, los archivos listados son archivos de datos con extensión *.sav*.

Si se le asigna al archivo un nombre ya existente, aparece un mensaje (ver Figura 3.17) indicando tal circunstancia.

Figura 3.17. Advertencia de nombre de archivo duplicado



## Marcar archivo como de sólo lectura

Existe la posibilidad de proteger un archivo de datos contra modificaciones accidentales. La opción **Marcar archivo como de sólo lectura** del menú **Archivo** impide que los cambios hechos en el *Editor de datos* puedan grabarse en el archivo original.

Con esta opción activa, al intentar guardar un archivo en el que se han hecho modificaciones aparece un mensaje indicando que el archivo es de sólo lectura y que, por tanto, no puede escribirse sobre él.

Si se desea guardar las modificaciones hechas en un archivo definido como *de sólo lectura*, puede optarse entre: (1) guardar el archivo con un nombre distinto y (2) restablecer el estatus del archivo seleccionando la opción **Marcar archivo como de lectura y escritura** del menú **Archivo**.

## Mostrar información de datos

El SPSS permite obtener información rápida sobre las diferentes características de un archivo de datos. Para obtener información sobre el archivo activo (es decir, sobre el archivo que se encuentra abierto en el *editor de datos*):

- Seleccionar la opción **Mostrar información de datos > Archivo de trabajo...** del menú **Archivo**.

Para obtener información sobre un archivo de datos almacenado en disco (el archivo sobre el que se solicita información con esta acción debe ser un archivo de datos en formato SPSS, es decir, un archivo previamente creado con las opciones **Guardar** o **Guardar como...** y, por tanto, un archivo que puede abrirse directamente con la opción **Abrir > Datos...** del menú **Archivo**).

- Seleccionar la opción **Mostrar información de datos > Archivo externo...** del menú **Archivo** para acceder al cuadro de diálogo *Mostrar información de datos*.

Este cuadro de diálogo es similar al cuadro de diálogo *Abrir archivo* de la Figura 3.1. Para ver la información relativa a un archivo de datos basta con indicar el nombre del archivo, y la unidad y la carpeta en que se encuentra. Al hacer esto, el *Visor de resultados* muestra, entre otras cosas, la siguiente información: el nombre del archivo (con su ruta completa); el formato del archivo; la fecha y la hora en que fue creado; la etiqueta del archivo, si existe; el número de casos y de variables de que consta; algunos detalles relacionados con la ponderación de casos, la presencia de conjuntos de variables y de respuestas múltiples, etc.; los nombres de las variables y sus etiquetas, si existen; el formato de las variables (numérico, cadena, etc.), incluyendo su longitud; las etiquetas de los valores, si existen; etc.

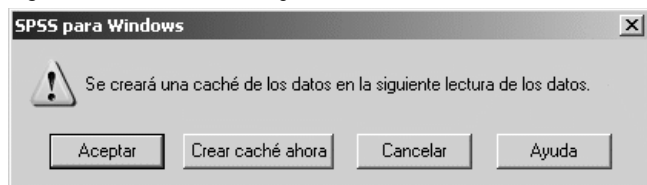
## Hacer una caché de datos

El hecho de que el SPSS trabaje sin copia temporal del archivo de datos activo hace que los datos tengan que ser releídos cada vez que se ejecuta un nuevo procedimiento. Si el archivo de datos ha sido importado de una base de datos, la consulta SQL creada para capturar la información de la base de datos se activa con cada procedimiento que necesita releer el archivo de datos. Puesto que la mayor parte de los procedimientos SPSS necesitan releer el archivo de datos, trabajar con archivos procedentes de bases de datos tiene como consecuencia un notable incremento del tiempo de procesamiento (especialmente si se utilizan grandes archivos y se ejecutan muchos procedimientos).

Este problema puede evitarse creando una *caché* de datos, que no es más que una copia temporal del archivo de datos. Para ello, basta con que el disco duro en el que se está trabajando tenga suficiente espacio libre. Para crear una caché de datos:

- Seleccionar la opción **Hacer una caché de datos...** del menú **Archivo** para acceder al cuadro de diálogo que muestra la Figura 3.18.

Figura 3.18. Cuadro de diálogo *Hacer una caché de datos*



A partir del momento en que se pulsa el botón **Aceptar**, el SPSS crea un archivo temporal del archivo de datos la primera vez que se ejecuta un procedimiento SPSS que requiere leer los

datos del archivo. El botón **Crear caché ahora** crea un archivo temporal de forma instantánea, sin esperar a la primera lectura.

## Detener procesador SPSS

Detener el procesador SPSS significa interrumpir cualquier acción que el SPSS esté llevando a cabo. Esta opción tiene ese efecto. Las transformaciones pendientes quedan canceladas. Detener el procesador SPSS no significa salir del SPSS. Se sigue dentro del SPSS y, por tanto, se puede seguir trabajando con él. Esta opción puede resultar útil cuando se ha solicitado al SPSS que ejecute algún procedimiento en el que el procesador está invirtiendo demasiado tiempo y no se desea continuar, o cuando, como consecuencia de alguna instrucción sintáctica incorrecta, el procesador ha quedado atrapado en un bucle. Etc.

## Presentación preliminar

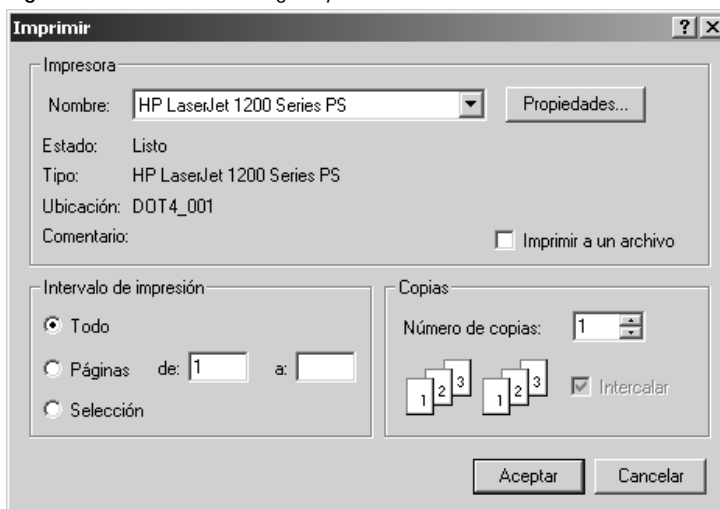
Para visualizar el aspecto exacto que adoptará el archivo una vez impreso, la opción **Presentación preliminar** del menú **Archivo** permite acceder a la ventana de *Presentación preliminar*, la cual muestra, página a página, todo el contenido del archivo de datos con el mismo aspecto que adoptará una vez impreso.

## Imprimir archivos de datos

Para imprimir el archivo de datos y controlar diferentes aspectos del proceso de impresión:

- Seleccionar la opción **Imprimir...** del menú **Archivo** para acceder al cuadro de diálogo *Imprimir* que muestra la Figura 3.19.

Figura 3.19. Cuadro de diálogo *Imprimir*



**Intervalo de impresión.** Las opciones de este recuadro permiten decidir qué parte del archivo de datos se desea imprimir:

**Todo.** Imprime todo el contenido del archivo de datos.

**Páginas de \_\_ a \_\_.** Permite especificar un rango de páginas.

**Selección.** Imprime sólo la parte del archivo de datos que se encuentra seleccionada (en el caso de que exista alguna parte seleccionada).

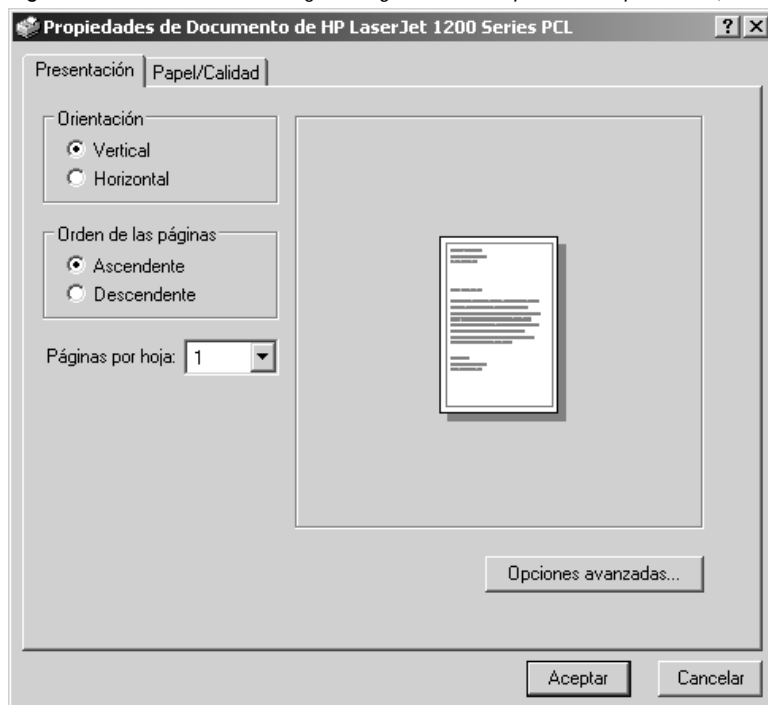
**Copias.** Permite seleccionar el número de copias que se desea imprimir. En el caso de seleccionar más de una copia, la opción **Intercalar** permite elegir cómo obtener las copias. Con esta opción desactivada, se hacen primero todas las copias de la primera página, después todas las copias de la segunda página, etc; con esta opción activada, se hace primero una copia de todas las páginas, después una segunda copia de todas las páginas, etc.

**Impresora.** El menú desplegable **Nombre** permite seleccionar la impresora con la que se desea imprimir. En ese menú desplegable están identificadas todas las impresoras previamente definidas en Windows.

" **Imprimir en un archivo.** Esta opción guarda el archivo de datos en un archivo en disco y le asigna la extensión *prn*.

**Propiedades...** Este botón conduce a un subcuadro de diálogo (ver Figura 3.20) cuyos nombre y aspecto dependen de la impresora seleccionada.

Figura 3.20. Subcuadro de diálogo *Configuración de impresión: Propiedades* (HP láser 1200)



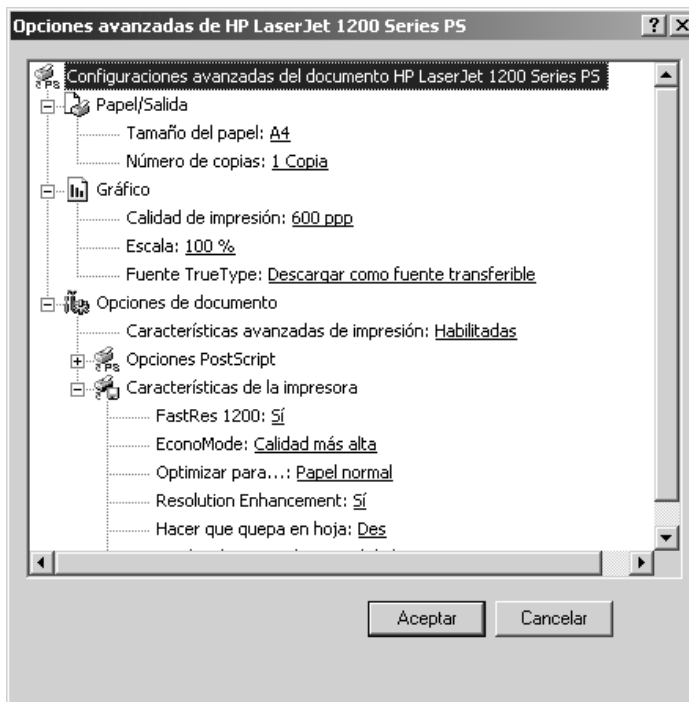
Desde este subcuadro de diálogo (utilizando las diferentes pestañas que incluye) es posible controlar múltiples aspectos de la impresión. Debe tenerse en cuenta que las posibilidades de control que ofrece este cuadro de diálogo y los que cuelgan de él dependen del tipo de impresora que se tenga instalada. En realidad, estos cuadros de diálogo no son cuadros SPSS, sino Windows. Por tanto, el aspecto y contenido de los mismos depende del sistema operativo instalado y de las impresoras definidas en él.

**Presentación.** Las opciones de esta pestaña permiten controlar la orientación del papel (*vertical* u *horizontal*), el orden en el que se desea imprimir las páginas (*ascendente*: desde la primera a la última; o *descendente*: desde la última a la primera), y el número de páginas que se desea incluir en cada hoja (una por defecto).

**Papel/Calidad.** Desde aquí es posible seleccionar el origen de la bandeja de entrada, es decir, es posible indicar dónde se encuentra ubicado el papel (cuando la impresora admite varias ubicaciones diferentes). Al igual que ocurre con muchas otras opciones de impresión, las de este recuadro dependen del tipo de impresora que se haya seleccionado.

El botón **Opciones avanzadas...** conduce al subcuadro de diálogo de *Opciones avanzadas de...* que muestra la Figura 3.21. Estas opciones permiten seleccionar el tamaño del papel (A4, ejecutivo, carta, folio, sobre, etc.) y el número de copias que se desea imprimir; controlar algunos detalles relacionados con la impresión de gráficos (calidad, escala, etc.); habilitar o deshabilitar las características avanzadas de impresión; seleccionar el modo de impresión económico; etc.

Figura 3.21. Subcuadro de diálogo *Opciones avanzadas de impresión*



## Datos/archivos usados recientemente

Esta opción del menú **Archivo** recoge, si así se desea, un listado de los últimos archivos utilizados. La opción **Datos usados recientemente** muestra únicamente *archivos de datos SPSS*. La opción **Archivos usados recientemente** muestra un listado del resto de archivos utilizados (syntax, resultados, etc.). El hecho de que esta opción muestre o no un listado de archivos y, en caso de hacerlo, la longitud del listado, depende de las especificaciones establecidas en el cuadro de diálogo **Opciones**, en el recuadro **Lista de archivos recientes**. A este cuadro de diálogo se accede mediante la opción **Opciones...** del menú **Edición**.

## Salir del SPSS

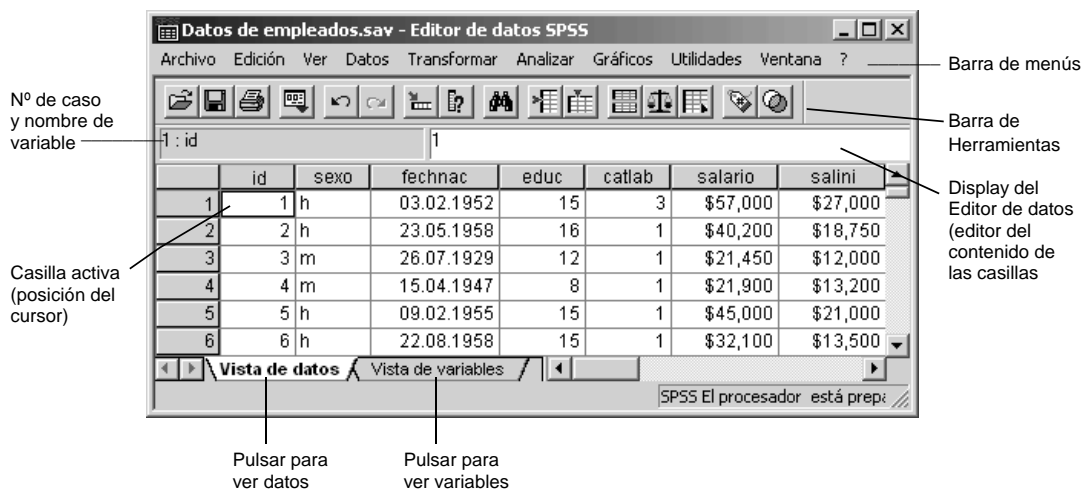
La opción **Salir** cierra el programa. Esta acción permite abandonar el SPSS y volver al escritorio de Windows o a otra aplicación abierta. Si el procesador está ocupado, puede que sea necesario detenerlo para poder salir del sistema.

Antes de cerrarse, el SPSS pregunta, mediante un cuadro de diálogo, si se desea guardar alguno de los archivos que se encuentran abiertos, pero sólo en el caso de que alguno de estos archivos haya sufrido alguna modificación. Si se decide no guardar un archivo que haya sufrido modificaciones, éstas se perderán definitivamente.

## El Editor de datos

Para analizar datos es necesario, en primer lugar, disponer de los datos sobre los que poder efectuar el análisis. El *Editor de datos* es la ventana que contiene el archivo de datos en que se basan todos los análisis. Se trata de una ventana con aspecto de hoja de cálculo diseñada para crear y editar archivos de datos SPSS (ver Figura 4.1). El *Editor de datos* se abre automáticamente al iniciar una sesión. Entrar en el *Editor de datos* equivale a entrar en el SPSS. Cerrar el *Editor de datos* equivale a salir del SPSS.

Figura 4.1. Ventana del *Editor de datos* (vista de datos)



El *Editor de datos* permite visualizar dos ventanas distintas mediante dos pestañas o solapas situadas en la parte inferior izquierda del propio *Editor*. La solapa *Vista de datos* muestra el *Editor de datos*, es decir, el contenido del archivo de datos (ver Figura 4.1). La solapa *Vista de variables* muestra el *Editor de variables*, es decir, los nombres de las variables acompañados del conjunto de características que las definen (ver Figura 4.2).

Un archivo de datos puede crearse de dos formas distintas: utilizando el teclado para introducir datos directamente en el *Editor de datos*, o importando la información ya existente en alguna fuente externa tal como un archivo de texto, una hoja de cálculo o una base de datos. En el Capítulo 3 ya se ha descrito cómo importar datos de una fuente externa. La segunda parte de este capítulo explica cómo introducir y editar datos utilizando el *Editor de datos*.



Unas pocas ideas generales ayudarán a comprender cuál es la estructura de un archivo de datos en formato SPSS o, lo que es lo mismo, la estructura del *Editor de datos* del SPSS:

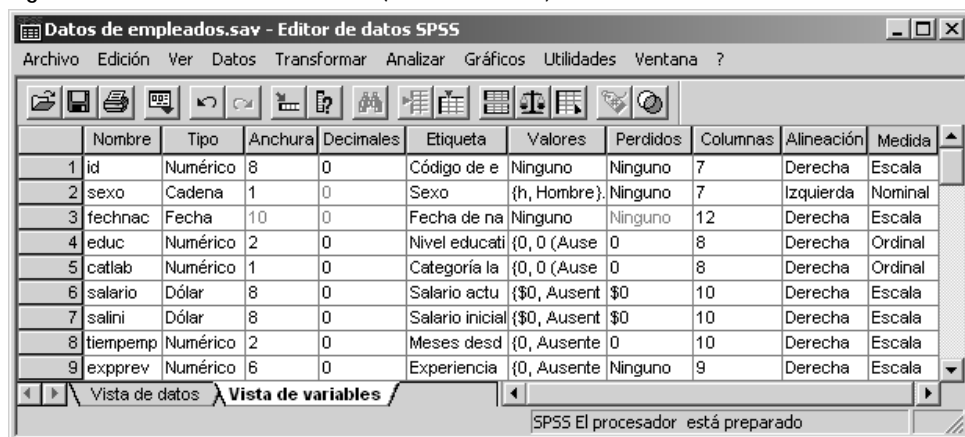
- Las *filas* representan *casos*. Cada fila es un caso (generalmente, un sujeto). Por ejemplo, cada sujeto que responde a un cuestionario es un caso.
- Las *columnas* representan *variables*. Cada columna es una variable. Por ejemplo, cada pregunta del cuestionario es una variable.
- Cada *casilla* contiene un *valor*. Una casilla concreta contiene el valor individual que corresponde a un determinado caso en una determinada variable. O sea, cada casilla es la intersección de un caso con una variable. A diferencia de lo que ocurre en las hojas de cálculo, una casilla no puede contener fórmulas, sino sólo valores individuales.
- El archivo del *Editor de datos* es siempre *rectangular*. Sus dimensiones vienen determinadas por el número de casos y de variables. Si se introduce algún valor en una casilla situada fuera de los límites del rectángulo definido por el número de casos y de variables, el SPSS extiende los límites del archivo de datos (los límites del rectángulo) para incluir cualquier casilla comprendida entre el nuevo valor y los límites anteriores (incrementando así el número de casos y/o de variables). Esto significa que no existen casillas vacías dentro de los límites del rectángulo: las casillas vacías se consideran valores *perdidos* (en inglés, *missing*) si corresponden a una variable numérica y valores *válidos* si corresponden a una variable de cadena (enseguida se tratarán los tipos de variables).

## Definir variables

Para definir una variable:

- Pulsar la solapa *Vista de variables* (ver Figura 4.1) para que el *Editor de datos* muestre la ventana de definición de variables (*Editor de variables*) que aparece en la Figura 4.2. También se puede acceder al *Editor de variables* situando el puntero del ratón en la cabecera de una variable y pulsando dos veces el botón principal del ratón.

Figura 4.2. Ventana del *Editor de datos* (vista de variables)



La ventana *Vista de variables* del *Editor de datos* (es decir, el *Editor de variables*) permite llevar a cabo todas las tareas relacionadas con la definición de una variable: asignarle nombre y etiqueta; definir el tipo de variable (numérica, fecha, cadena, etc.); asignar, en caso necesario, etiquetas a los valores; identificar si existen o no valores perdidos y de qué tipo; establecer el formato de columna del *Editor de datos*; asignar un nivel de medida. Las diez columnas de la ventana *Vista de variables* del *Editor de datos* contienen todos los detalles que el SPSS utiliza para definir una variable.

## Asignar nombre a una variable

Existen varias formas diferentes de crear una variable nueva. Se crea una variable nueva al introducir un valor en alguna de las casillas de una columna vacía del *Editor de datos* (ver Figura 4.1), o al introducir un valor en alguna casilla en blanco del *Editor de datos-variables* (ver Figura 4.2).

Cualquiera que sea el modo elegido para crear una nueva variable, al crearla, el SPSS le asigna un nombre por defecto compuesto por el prefijo *var* y una secuencia de cinco dígitos: *var00001*, *var00002*, etc. No obstante, es posible asignar a una variable cualquier otro nombre simplemente escribiéndolo con el teclado. Para ello:

- Situar el cursor, dentro de la columna **Nombre**, en la casilla correspondiente a la variable cuyo nombre se desea crear o editar (para esto puede utilizarse el teclado o el ratón) y escribir el nuevo nombre.

Al asignar nombre a las variables, es necesario tener en cuenta unas pocas reglas:

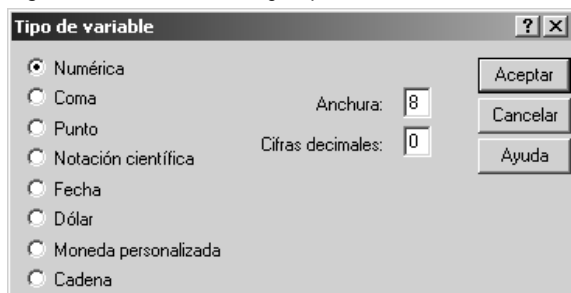
- Los nombres de las variables en el SPSS constan de un máximo de 64 caracteres.
- Son caracteres *válidos*: todas las letras, todos los números, el punto y los caracteres @, #, \$, %, \_.
- Son caracteres *no válidos*: el espacio en blanco, los signos de admiración e interrogación (!, ?), el apóstrofo (') y el asterisco (\*).
- Los nombres de las variables deben comenzar siempre con una letra, pero pueden terminar con cualquier carácter válido exceptuando el punto.
- Dos variables no pueden tener el mismo nombre. Respecto a esto, debe tenerse en cuenta que el SPSS no hace distinción entre mayúsculas y minúsculas; por tanto, los nombres *VARIABLE*, *Variable*, *VariaBle* y *variable* se consideran el mismo nombre (si bien el *Editor de datos* preserva la apariencia original).
- Existen unas cuantas palabras *reservadas* que no pueden utilizarse como nombres de variables: ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO y WITH. Si se utiliza alguna de estas palabras, aparece un mensaje indicando tal circunstancia.

## Definir el tipo de variable

Si no se dan instrucciones en otro sentido, el SPSS asume que todas las variables son numéricas. Pero es posible asignar a una variable cualquiera de los formatos disponibles. Para cambiar el *tipo* de una variable:

Situar el cursor en la columna **Tipo** (ver Figura 4.2) sobre la casilla correspondiente a la variable cuyo tipo se desea cambiar y pulsar el botón *puntos suspensivos* [...] de esa casilla para acceder al cuadro de diálogo *Tipo de variable* que muestra la Figura 4.3.

Figura 4.3. Cuadro de diálogo *Tipo de variable*



**Numérica.** Este formato de variable acepta como caracteres válidos cualquier número, el signo más (+), el signo menos (–) y el separador decimal (el punto o la coma, dependiendo de las especificaciones internacionales establecidas en Windows). El cuadro de texto **Anchura** permite establecer el total de dígitos de la variable, incluyendo una posición para el separador decimal y otra más para cada coma de millar. El cuadro de texto **Cifras decimales** permite establecer el número de decimales que se desea visualizar en el *Editor de datos*. La anchura máxima permitida para las variables numéricas es 40; el número máximo de decimales es 16.

**Coma.** Son caracteres válidos: cualquier número, el signo más (+), el signo menos (–), el punto como separador decimal y múltiples comas insertadas como separadores de los millares. Las comas separadoras de los millares se insertan automáticamente. El cuadro de texto **Anchura** permite establecer el total de dígitos de la variable, incluyendo una posición para el separador decimal y otra más para cada coma de millar. El cuadro de texto **Cifras decimales** permite establecer el número de decimales que se desea visualizar en el *Editor de datos*.

**Punto.** Son caracteres válidos: cualquier número, el signo más (+), el signo menos (–), la coma como separador decimal y múltiples puntos insertados como separadores de los millares. Los puntos separadores de los millares son automáticamente insertados. El cuadro de texto **Anchura** permite establecer el total de dígitos de la variable, incluyendo una posición para el separador decimal y otra más para cada coma de millar. El cuadro de texto **Cifras decimales** permite establecer el número de decimales que se desea visualizar en el *Editor de datos*.

**Notación científica.** Son caracteres válidos cualquier número acompañado de una expresión en notación científica: la letra D, la letra E, el signo más o el signo menos. Por ejemplo, 1.234E2 equivale a 123.4; 1.234+2 equivale a 123.4; 1.234E+2 equivale a 123.4; etc.

**Fecha.** Este formato admite como valores válidos fechas y horas. Al marcar esta opción aparece una lista con los formatos de fecha disponibles. Estas variables son procesadas, en su mayor parte, como el número de segundos transcurridos desde el 14 de octubre de 1582.

**Dólar.** Permite introducir como caracteres válidos cualquier número, el símbolo \$, el punto como separador decimal y comas como separadores de los millares. Pueden fijarse la *Anchura* y el número de *Cifras decimales*, o puede seleccionarse un formato concreto de la lista desplegable que se obtiene al marcar esta opción. El símbolo \$ y las comas separadoras de los millares se insertan de forma automática.

**Moneda personalizada.** Esta opción permite modificar el aspecto de las definiciones hechas (si es que se ha hecho alguna) en el menú **Edición > Opciones > Moneda**. El cuadro de texto **Anchura** permite establecer el total de dígitos de la variable, incluyendo el carácter específico de la moneda. El cuadro de texto **Cifras decimales** permite señalar el número de decimales que se desea visualizar en el *Editor de datos*. Al introducir datos no es necesario (tampoco posible) incluir el carácter específico de la moneda: el SPSS lo asigna automáticamente.

**Cadena.** En este tipo de variables es válido cualquier carácter: se admiten todos los caracteres que puedan introducirse desde el teclado. En **Anchura** debe especificarse el número máximo de caracteres que se le asigna a la variable. Si la anchura definida es de 8 caracteres o menos, la variable se considera de *cadena corta*. Si es de más de 8 caracteres, la variable se considera de *cadena larga*. Las variables de *cadena corta* pueden utilizarse en muchos más procedimientos SPSS que las de *cadena larga*. Como norma general de actuación, es preferible evitar en lo posible las variables de cadena.

La *anchura* y el número de *cifras decimales* de una variable puede cambiarse sin necesidad de entrar en el cuadro de diálogo *Tipo de variable*. Basta con introducir los valores deseados en las casillas de las columnas encabezadas **Anchura** y **Decimales** del *Editor de variables* (ver Figura 4.2).

No debe confundirse el *formato* asignado a una variable en el cuadro de diálogo *Tipo de variable* (ver Figura 4.3) con el *aspecto* que la variable adopta en el *Editor de datos*. Formato y aspecto no tienen por qué coincidir. No obstante, el aspecto que la variable adopta en el *Editor de datos* viene condicionado por el formato asignado a la variable. Un par de reglas generales pueden ayudar a comprender esta relación:

- Una variable con formato *numérico*, *coma o punto*, acepta cualquier número de decimales (hasta un máximo de 16), independientemente de la restricción establecida en la columna **Decimales** (ver Figura 4.2) o, lo que es lo mismo, en el cuadro de texto **Cifras decimales** del cuadro de diálogo *Tipo de variable* (ver Figura 4.3). El SPSS procesa y utiliza en los cálculos todos los decimales introducidos (o los resultantes de una transformación) hasta un máximo de 16. Sin embargo, las casillas del *Editor de datos* muestran los valores con el número de decimales establecido en la columna **Decimales** (redondeando, en caso necesario, el último decimal de los mostrados). El *Display* del *Editor de datos* (ver Figura 4.1), por el contrario, muestra todos los decimales de la casilla en la que se encuentra el cursor.
- Las variables de *cadena* se almacenan y procesan respetando siempre la anchura establecida en la columna **Anchura** (ver Figura 4.2). Por tanto, un valor no aceptará caracteres más allá de la anchura establecida. Y si un valor es más corto que la anchura establecida, se le asignarán espacios en blanco a la derecha hasta igualar esa anchura. Así, por ejemplo, si se define una variable con ancho 6, el valor *NO* será «*NO* ». Lo cual es diferente de, por ejemplo, «*NO* ».

## Asignar etiquetas

El nombre de una variable es, en muchos casos, insuficiente para recordar de qué variable se trata y, por supuesto, para reconocer el significado de los valores que toma. Esto, sin embargo, no constituye un problema importante, pues el SPSS permite asignar etiquetas descriptivas tanto a los nombres de las variables como a sus valores.

Para asignar *etiqueta a una variable*:

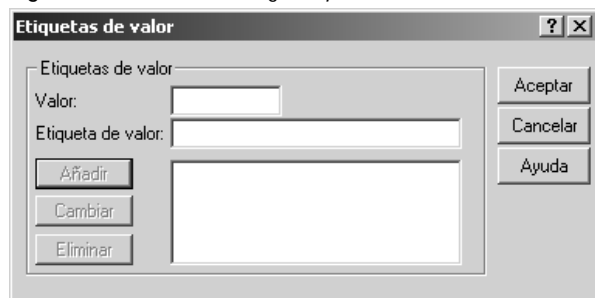
- Situar el cursor en la correspondiente casilla de la columna **Etiquetas** (ver Figura 4.2) y escribir una etiqueta descriptiva para la variable.

La etiqueta descriptiva de una variable puede contener hasta 256 caracteres (aunque muchos procedimientos SPSS muestran menos de 256 caracteres en las tablas de resultados). Puede utilizarse cualquier carácter del teclado, incluso espacios en blanco: la etiqueta aparecerá en las tablas de resultados tal como sea introducida desde el teclado (mayúsculas, tildes, espacios, eñes, etc.).

Para asignar *etiquetas a los valores* de una variable:

- Situar el cursor en la columna **Valores** (ver Figura 4.2) sobre la casilla correspondiente a la variable cuyos valores se desea etiquetar y pulsar el botón *puntos suspensivos* [...] que contiene esa casilla para acceder al cuadro de diálogo *Etiquetas de valor* que muestra la Figura 4.4.

Figura 4.4. Cuadro de diálogo *Etiquetas de valores*



Este cuadro de diálogo permite asignar etiquetas descriptivas a los valores de una variable (las variables de *cadena larga* no admiten etiquetas de valor). Las etiquetas de valores resultan especialmente útiles cuando se utilizan códigos numéricos para representar categorías no numéricas (como, por ejemplo, cuando se utilizan los códigos 1 y 2 para representar los valores *masculino* y *femenino*). Las etiquetas de valor admiten hasta 60 caracteres (aunque muchos procedimientos SPSS muestran menos de 60 caracteres en las tablas de resultados) y puede utilizarse cualquier carácter, incluso espacios en blanco: la etiqueta aparecerá tal como sea introducida desde el teclado. Para asignar etiquetas a los valores de una variable:

- Escribir el valor de la variable en el cuadro de texto **Valor** (por ejemplo, *h*).
- Escribir la etiqueta que se le quiere asignar a ese valor en el cuadro de texto **Etiqueta de valor** (por ejemplo, *Hombre*).

- Pulsar el botón **Añadir** para trasladar el valor y su etiqueta a la lista de valores y etiquetas (recuadro inferior).
- Repetir la operación para cada valor. Al final, todos los valores y las etiquetas asignadas quedan listados en el recuadro inferior.
- Los botones **Cambiar** y **Borrar** permiten modificar y eliminar, respectivamente, etiquetas previamente definidas.

El SPSS ofrece la posibilidad de automatizar el proceso de asignación de etiquetas mediante un *asistente* que efectúa un barrido de todos los casos del archivo y ofrece un listado con todos los valores de la variable. Para una descripción de cómo utilizar este *asistente*, ver más adelante, en este mismo capítulo, el apartado *Definir variables de forma automática*.

## Definir valores perdidos

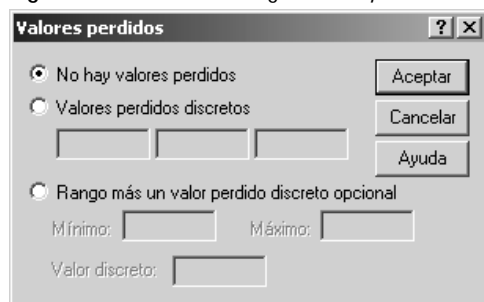
Existen dos tipos de valores perdidos en el SPSS:

- Valores perdidos **Definidos por el sistema**. Las casillas vacías del *Editor de datos* (las cuales aparecen con un punto) son automáticamente interpretadas por el SPSS como valores perdidos.
- Valores perdidos **Definidos por el usuario**. A veces resulta útil distinguir entre diferentes tipos de valores perdidos. En las respuestas a una pregunta puede interesar distinguir, por ejemplo, entre los sujetos que no conocen la respuesta, los que simplemente no responden y los que no desean responder; si se dispone de esta información, no hay por qué tratarla como si fuera un único valor: puede optarse por definir varios tipos de valores perdidos.

Para definir valores perdidos:

- Situar el cursor en la columna **Perdidos** (ver Figura 4.2) sobre la casilla correspondiente a la variable en la que se desea definir valores perdidos y pulsar el botón *puntos suspensivos* ... que contiene esa casilla para acceder al cuadro de diálogo *Valores perdidos* que muestra la Figura 4.5.

Figura 4.5. Cuadro de diálogo *Valores perdidos*



Todos los *tipos* de variable admiten valores perdidos definidos por el usuario excepto las variables de cadena *larga*:

**No hay valores perdidos.** Esta opción indica que no existen valores perdidos definidos por el usuario. Todos los valores se consideran válidos. Sólo las casillas que contienen un punto (casillas vacías) se consideran valores perdidos. Es la opción que se encuentra activa por defecto.

**Valores perdidos discretos.** Permite definir como valores perdidos hasta tres valores concretos. Todos los valores que coincidan con los establecidos en esta opción serán considerados valores perdidos. Esta opción es útil para variables categóricas y cuantitativas discretas. Sólo es posible utilizarla con variables numéricas y de cadena *corta*.

**Rango más un valor perdido discreto opcional.** Permite definir como valores perdidos un determinado rango de valores (comprendido entre el **Mínimo** y el **Máximo** fijados) y, opcionalmente, un valor concreto no perteneciente al rango. Todos los valores comprendidos entre los límites del rango establecido, incluidos los límites, serán considerados valores perdidos. Esta opción es útil para variables cuantitativas continuas. No es posible utilizarla con variables de cadena.

## Definir el formato de columna

Para cambiar la anchura de las columnas del *Editor de datos*:

- Situar el cursor en la columna **Columnas** (ver Figura 4.2) sobre la casilla correspondiente a la variable cuya anchura se desea modificar y utilizar las flechas que contiene esa casilla para establecer la anchura deseada.

La anchura de una columna viene determinada, por defecto, por la anchura asignada a la variable (ver, más arriba, el apartado *Definir el tipo de variable*), pero puede cambiarse introduciendo el valor deseado. También puede cambiarse la anchura de una columna desde el *Editor de datos*, situando el puntero del ratón en el borde derecho de la cabecera de la columna y arrastrando el puntero hasta obtener la anchura deseada.

Es importante tener en cuenta que la anchura de una columna afecta únicamente al aspecto del *Editor de datos*. Cambiar la anchura de una columna no cambia la anchura de la variable. Si a una variable se le ha asignado una anchura mayor que la anchura definida para la columna, el *Editor de datos* mostrará los valores truncados.

La nueva anchura de columna no sólo se mantiene activa mientras el archivo de datos permanece abierto. Si se guarda el archivo tras alterar la anchura de una columna, al abrir de nuevo el archivo se mantiene la anchura guardada.

## Alinear texto

La opción **Alinear texto** permite controlar la posición (izquierda, centro, derecha) que adoptan los valores dentro de sus casillas. Para alinear el contenido de una columna:

- Situar el cursor en la columna **Alineación** (ver Figura 4.2) sobre la casilla correspondiente a la variable cuyo texto se desea alinear y pulsar el botón de menú desplegable que contiene esa casilla para elegir una de las tres opciones disponibles: izquierda, derecha o centrado.

La alineación que actúa por defecto para las variables numéricas es *derecha* y, para las variables de cadena, *izquierda*. Pero todas las variables admiten cualquiera de las tres alineaciones disponibles.

## Asignar un nivel de medida

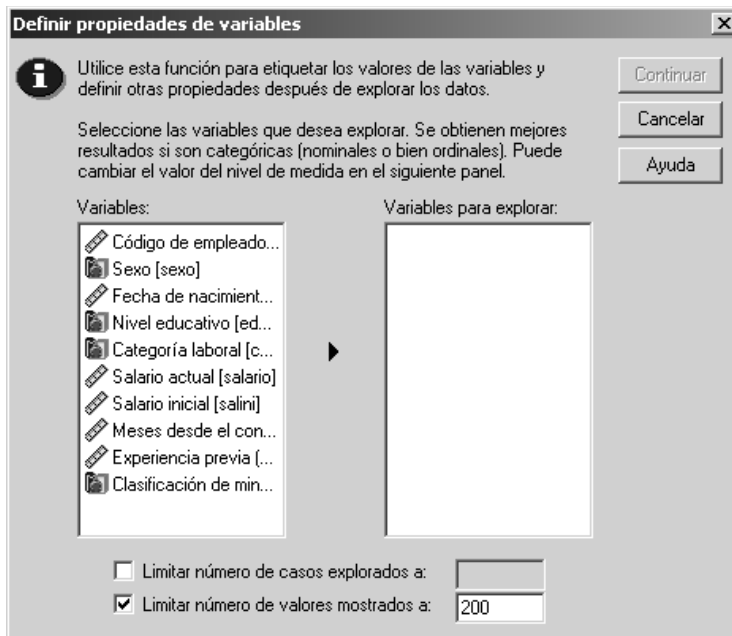
Para terminar de definir una variable numérica es necesario asignarle uno de los siguientes niveles de medida: *escala* (para variables cuantitativas continuas obtenidas con una escala de intervalo o razón: edad, salario, altura, temperatura, etc.); *ordinal* (para variables cuantitativas obtenidas con una escala ordinal: nivel educativo, clase social, etc.); y *nominal* (para variables categóricas medidas con una escala nominal: sexo, clasificación étnica, lugar de procedencia, tipo de tratamiento, etc.). La asignación de un nivel de medida no afecta a los procedimientos estadísticos, pues la elección de variables depende del usuario; pero sí afecta a los gráficos interactivos, pues en determinadas posiciones sólo pueden ir ciertos tipos de variables.

## Definir variables de forma automática

Esta opción permite el acceso a un *asistente* diseñado, básicamente, para ayudar al usuario a automatizar el proceso de asignación de *etiquetas de valor*. Para acceder a este asistente:

- Seleccionar la opción **Definir propiedades de variable...** del menú **Datos** para acceder al cuadro de diálogo *Definir propiedades de variable* (paso 1) que muestra la Figura 4.6.

Figura 4.6. Cuadro de diálogo *Definir propiedades de variable* (paso 1)





La lista de variables del archivo de datos ofrece un listado con todas las variables numéricas y de cadena *corta* del archivo de datos (las variables de cadena *larga* no admiten etiquetas de valor). Para asignar etiquetas de valor:

- Seleccionar la variable a la que se desea asignar etiquetas y trasladarla a la lista **Variables para explorar**. En el ejemplo de la Figura 4.6 se han seleccionado las variables *sexo* y *catlab*.
- “ Limitar el número de casos explorados a. Para que el SPSS pueda ayudar a automatizar el proceso de asignación de etiquetas necesita explorar el archivo de datos. Si el archivo es demasiado largo y no se desea explorar todos los casos, esta opción permite indicar al SPSS cuántos casos debe explorar. Aunque esta opción puede tener alguna utilidad, debe tenerse en cuenta que limitar el número de casos que son explorados podría conducir a que el SPSS no reconociera todos los valores distintos de una variable (particularmente si el archivo se encuentra ordenado por esa variable).
- “ Limitar el número de valores mostrados a. En general, las etiquetas de valor únicamente tienen sentido con variables categóricas (es decir, variables nominales o variables ordinales con pocos niveles). Por tanto, lo habitual será que no haya que preocuparse por limitar el número de valores listados. No obstante, si una variable tiene demasiados valores y todavía se tiene interés en asignarle etiquetas de valor, esta opción permite limitar el número de valores distintos que ofrecerá el asistente después de explorar los datos.

Una vez seleccionada la variable (o variables) a la que se desea asignar etiquetas de valor, el botón **Continuar** conduce al segundo paso del *Asistente* para la definición de propiedades de variable (ver Figura 4.7).

Figura 4.7. Cuadro de diálogo *Definir propiedades de variable* (paso 2)

**Definir propiedades de las variables**

Lista de variables exploradas: S... M... Variable

☐ sexo

☒ fechnac

Variable actual: sexo Etiqueta: Sexo

Nivel de medida: Nominal Sugerir Tipo: Cadena

Valores sin etiqueta: 0 Ancho: 1 Decimales: 0

Rejilla etiq. valores: **i** Añada etiquetas a la rejilla o editelas. Puede añadir valores abajo.

	Cambiado	Perdido	Recuento	Valor	Etiqueta
1	<input type="checkbox"/>	<input type="checkbox"/>	258	h	Hombre
2	<input type="checkbox"/>	<input type="checkbox"/>	216	m	Mujer
3	<input type="checkbox"/>	<input type="checkbox"/>			

Copiar propiedades: De otra variable... A otras variables... Etiquetas automáticas

Valores sin etiquetas: Etiquetas automáticas

Casos explorados: 474 Límite lista valores: 200

Aceptar Pegar Restablecer Cancelar Ayuda

**Lista de variables exploradas.** El cuadro de diálogo ofrece, en primer lugar, un listado de las variables previamente seleccionadas y que acaban de ser exploradas. Pinchando sobre la cabecera de la columna **Variable**, las variables listadas pueden ordenarse alfabéticamente (por orden ascendente o descendente, pinchando sucesivamente).

En la primera columna de este listado **S...** (ampliando la columna puede verse el encabezado **Sin etiqueta**) aparece una marca en forma de cruz en aquellas variables que contienen valores sin etiquetas de valor. En el ejemplo de la Figura 4.7 la variable *sexo* conserva sus etiquetas, de modo que en esta primera columna no aparece la marca (en la variable *catlab* aparece la marca porque en esta variable se ha eliminado la etiqueta correspondiente al código de valor perdido. Pinchando sobre la cabecera de esta columna pueden colocarse al principio de la lista las variables sin marca (es decir, las variables que poseen valores sin etiqueta).

La segunda columna de este listado (encabezada **Medida**) indica, mediante el correspondiente símbolo, el nivel de medida de la variable. Pinchando sobre la cabecera de esta columna pueden ordenarse las variables por su nivel de medida.

**Propiedades de variable.** El tercio superior del cuadro de diálogo muestra algunas de las propiedades de la variable seleccionada en la **Lista de variables exploradas**. En concreto: el nombre de la variable (**Variable actual**), su **Etiqueta** (si la tiene), su **Nivel de medida**, el **Tipo** de variable que le ha sido asignado y el número de **Valores sin etiqueta**. Desde este cuadro de diálogo no es posible cambiar el nombre de una variable ni su formato básico (numérico o cadena). Si el formato de una variable es *numérico*, es posible asignarle cualquiera de los formatos numéricos disponibles (coma, punto, fecha, dólar, etc.) y controlar su anchura y número de decimales; pero si una variable tiene formato de *cadena*, únicamente es posible cambiar su etiqueta.

Si no se tiene claro qué nivel de medida asignar a una variable (nominal, ordinal, escala), el botón **Sugerir** ayuda a tomar la decisión (a este respecto, conviene recordar que el SPSS asigna por defecto a todas las variables con formato numérico un nivel de medida de *escala*, es decir, un nivel de medida de intervalo o razón). El botón **Sugerir** se encuentra inactivo si en el cuadro de diálogo inicial se ha indicado que no se explore el archivo (**Limitar el número de casos explorados a** = 0).

**Etiquetas de valor.** La rejilla central contiene las opciones necesarias para asignar o modificar las etiquetas de los valores:

**Valor.** Contiene un listado de todos los valores de la variable seleccionada. Si la variable posee algún valor no listado (por ejemplo, por haber puesto límite al número de casos explorados; ver Figura 4.6) al que se le desea asignar etiqueta, esta columna permite introducir valores nuevos a partir del último valor listado. Si el número de casos explorados se ha limitado a 0, esta columna únicamente mostrará los valores que ya posean etiqueta y los que estén definidos como valores perdidos. De modo similar a como ocurre en el *Editor de datos*, esta columna muestra un asterisco cuando un valor tiene una anchura mayor que la anchura definida para la variable.

**Etiqueta.** Muestra las etiquetas de valor previamente definidas. En esta columna pueden introducirse o modificarse las etiquetas utilizando el teclado.

**Recuento.** Muestra el número de veces que se repite cada valor explorado.

**Perdido.** Indica, mediante una marca, qué valores de la variable están codificados como valor perdido (lógicamente se trata de valores perdidos *definidos por el usuario*; los valo-

res perdidos *definidos por el sistema* son casillas vacías y, por tanto, no están listados aquí como valores de la variable). Si en una variable ya se ha definido un *rango* de valores perdidos, esta columna queda bloqueada, de modo que no es posible definir otros valores de la variable como valores perdidos ni borrar los existentes.

**Cambiado.** Las marcas de esta columna identifican valores cuyas etiquetas han sido modificadas (creadas, borradas o cambiadas).

**Etiquetas automáticas.** Este botón permite obtener, de forma automática, etiquetas para los valores que no poseen etiqueta. Esta etiqueta automática no es otra cosa que el propio valor de la variable.

**Copiar propiedades.** Es posible asignar propiedades a la variable seleccionada utilizando otra variable del archivo de datos; también es posible asignar a otras variables del archivo las propiedades de la variable seleccionada. Para ello es necesario que las variables que se van a intercambiar propiedades hayan sido seleccionadas en el primer cuadro de diálogo del *asistente* (ver Figura 4.6). Para copiar las propiedades de una variable en la variable seleccionada:

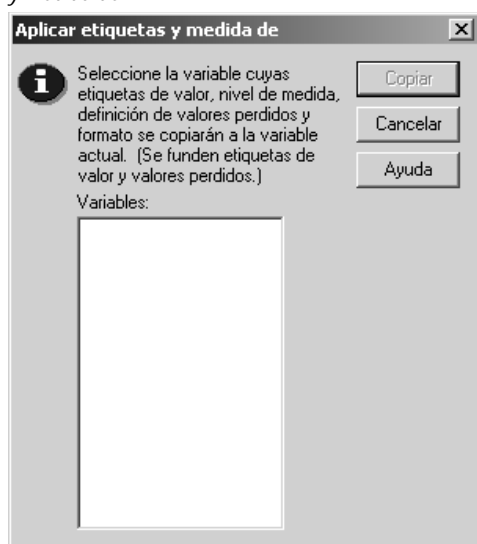
- Pulsar el botón **De otra variable...** para acceder al subcuadro de diálogo *Aplicar etiquetas y medida de...* que muestra la Figura 4.8.a.

Para copiar las propiedades de la variable seleccionada en otra u otras variables:

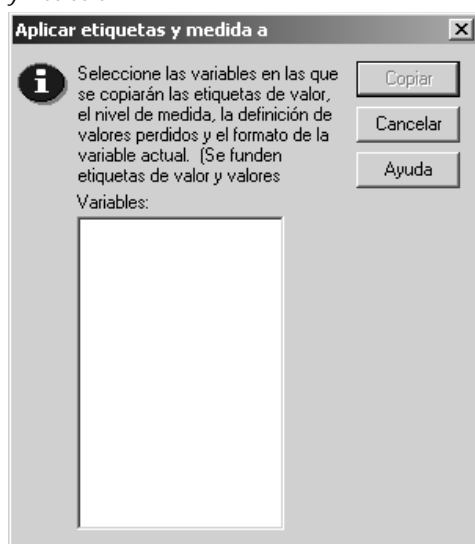
- Pulsar el botón **A otras variables...** para acceder al subcuadro de diálogo *Aplicar etiquetas y medida a...* que muestra la Figura 4.8.b.

Estos cuadros de diálogo ofrecen un listado con las variables que, habiendo sido seleccionadas previamente (ver Figura 4.6), poseen el mismo tipo de formato básico (numérico o cadena) que la variable actualmente seleccionada.

**Figura 4.8.a.** Subcuadro de diálogo *Aplicar etiquetas y medida de...*



**Figura 4.8.b.** Subcuadro de diálogo *Aplicar etiquetas y medida a...*



## Copiar propiedades de datos

El *diccionario* de un archivo de datos contiene información general sobre el archivo e información específica sobre las etiquetas, los valores perdidos, el formato de las variables, etc. Es decir, información sobre las características del archivo y sobre las propiedades de las variables que contiene. La información del diccionario de un archivo puede utilizarse para dar formato rápido a otro archivo de datos. Esta información se transfiere automáticamente desde un archivo origen o *fuentes* hasta un archivo receptor o destino que siempre es el archivo de *trabajo* (el archivo que se encuentra abierto en el *Editor de datos*). Para transferir la información del diccionario de datos:

- Abrir el archivo al que se desea dar formato y seleccionar la opción **Copiar propiedades de datos** del menú **Archivo** para acceder al *Asistente para copiar propiedades de los datos*.
- En el **primer paso**, el *asistente* solicita información sobre el archivo *fuentes*. Este archivo puede ser el propio archivo de *trabajo* o un archivo distinto almacenado en disco.
- En el **segundo paso** debe indicarse qué criterio se desea utilizar para transferir la información del diccionario. Existen tres opciones. Con la primera opción, la información del diccionario se transfiere utilizando como criterio el *nombre* y el *tipo básico* (numérico o de cadena) de las variables; esta opción ofrece la posibilidad de crear en el archivo de *trabajo*, si así se desea, nuevas variables basadas en las variables del archivo *fuentes* que no se encuentran en el archivo de *trabajo*. La segunda opción permite transferir la información del diccionario desde una variable del archivo *fuentes* hasta una o más variables del archivo de *trabajo*; el único requisito es que la variable *fuentes* y las variables *destino* tengan el mismo *tipo básico* de formato (numérico o de cadena). La tercera opción únicamente transfiere características generales del archivo; con esta opción no se realiza transferencia de información entre variables.
- En el **tercer paso** debe decidirse qué aspectos concretos del diccionario se desea transferir (etiquetas, valores perdidos, nivel de medida, etc.). Esto incluye decidir si las etiquetas de valor de las variables *fuentes* reemplazarán o se añadirán a las de las variables *destino*.
- En el **cuarto paso** debe decidirse si se desea transferir la etiqueta del archivo y, en el caso de haber elegido la tercera opción del segundo paso, qué características del archivo se desea transferir: conjuntos de respuestas múltiples, conjuntos de variables, documento, ponderación de casos.

La aplicación del diccionario de datos se ajusta a las siguientes claves:

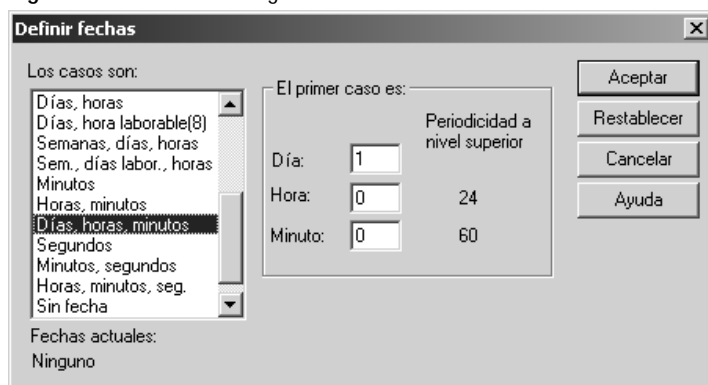
- Si una variable del archivo *fuentes* carece de algunas propiedades (etiquetas, valores perdidos), la variable del archivo de *trabajo* conservará sus propiedades.
- Si una variable del archivo *fuentes* no contiene especificaciones para los valores perdidos, la variable del archivo de *trabajo* perderá las especificaciones que posea.
- El *tipo básico* de formato no cambia: una variable numérica no puede cambiar a variable de cadena, ni al revés.
- Si el archivo *fuentes* está ponderado por una variable existente en el archivo de *trabajo*, la ponderación también se activa en el archivo de *trabajo*. Si tal ponderación no existe en el archivo *fuentes*, el estado de ponderación del archivo de *trabajo* no se altera.

## Definir fechas

Esta opción permite generar variables *fecha* cuyos valores progresan a lo largo de los casos con un incremento constante. Puede utilizarse, entre otras cosas, para establecer la periodicidad de una serie temporal o para etiquetar los resultados de un análisis de series temporales. Para definir una variable fecha:

- Seleccionar la opción Definir fechas... del menú Datos para acceder al cuadro de diálogo *Definir fechas* que muestra la Figura 4.9.

Figura 4.9. Cuadro de diálogo *Definir fechas*



**Los casos son.** Este listado contiene varias opciones para definir el intervalo de tiempo con el que será generada la serie: horas-minutos-segundos, semanas-días horas, etc. Entre estas opciones existen dos que merecen consideración especial. La opción *Sin fecha* elimina las variables *fecha* previamente definidas, es decir, elimina cualquier variable que tenga uno de estos nombres: *year\_*, *quarter\_*, *month\_*, *week\_*, *day\_*, *hour\_*, *minute\_*, *second\_* y *date\_*. La opción *Personalizado* permite definir variables *fecha* personalizadas creadas con el *Editor de sintaxis* del SPSS (por ejemplo, una semana de cuatro días de trabajo, etc.).

**El primer caso es.** Define el valor inicial de la variable *fecha* (el contenido de este recuadro depende de la opción marcada en **Los casos son**). Este valor inicial se asigna al primer caso. Los casos siguientes reciben valores que van incrementándose de forma secuencial tomando como base el intervalo de tiempo seleccionado.

**Periodicidad a nivel superior.** Informa sobre la variación cíclica asociada a los componentes de la opción seleccionada (el número de meses de un año, el número de días de una semana, el número de minutos de una hora, etc.). El valor informado indica el valor máximo que puede introducirse en cada casilla.

Conviene conocer algunas de las reglas que rigen la creación de variables *fecha*:

- El SPSS crea una variable numérica nueva para cada uno de los componentes que forman parte de la opción seleccionada en la lista **Los casos son**. Los nombres de estas nuevas variables terminan con un carácter de subrayado. Además de las variables

correspondientes a cada componente, el SPSS crea también una variable de cadena descriptiva a la que asigna el nombre *date\_*. Al seleccionar, por ejemplo, la opción *Semanas, días, horas*, se crean cuatro nuevas variables: *week\_*, *day\_*, *hour\_* y *date\_*.

- Si el archivo de datos ya contiene variables *fecha*, éstas son reemplazadas al definir nuevas variables *fecha* con los mismos nombres que las ya existentes.
- No deben confundirse las *variables fecha* creadas con la opción **Definir fechas** con las *variables tipo fecha* (variables con formato de fecha; ver, en este mismo capítulo, el apartado *Definir el tipo de variable*). Las *variables tipo fecha* contienen fechas codificadas en distintos formatos (por ejemplo, la fecha de nacimiento de los sujetos) y son procesadas como el número de segundos transcurridos desde el 14 de octubre de 1582. Las *variables fecha*, por el contrario, constituyen una serie temporal: son enteros que representan el número de horas, días, semanas, etc., transcurridos a partir de un valor inicial establecido por el usuario.

## Entrar datos

El *Editor de datos* permite introducir datos en cualquier orden: por casos, por variables, por áreas determinadas o sólo en casillas individuales. Para introducir un dato en una casilla pueden seguirse dos estrategias distintas: (1) introducir el dato directamente en la casilla deseada; (2) introducir el dato en el *Display del Editor de datos* (ver Figura 4.1).

Para introducir un dato ***directamente en una casilla***:

- Colocar el cursor en la casilla. Para situar el cursor en una casilla pueden utilizarse las flechas del teclado o el puntero del ratón. Se sabe dónde se encuentra el cursor porque la casilla correspondiente tiene los bordes resaltados. También se sabe porque en la parte izquierda del *Display del Editor de datos* aparece el número de caso y el nombre de variable correspondientes a esa casilla (ver Figura 4.1).
- Introducir el dato. Los valores que se van escribiendo van apareciendo tanto en la casilla seleccionada como en el *Display del Editor de datos* (ver Figura 4.1). Pueden utilizarse las teclas de borrado para corregir errores.

Para introducir un dato ***a través del Display del Editor de datos***:

- Situar el cursor en la casilla en la que se desea introducir el dato.
- Pinchar con el puntero del ratón sobre el *Display del Editor de datos*.
- Introducir el dato en el *Display* (la casilla no muestra el dato introducido).
- Pulsar la tecla de retorno de carro (o cualquiera de las flechas, o la tecla del tabulador) para que los valores introducidos en el *Display* pasen a la casilla activa.

Al introducir datos hay que tener en cuenta que el *tipo de variable* condiciona el tipo de valores que admite una casilla:

- Una variable no admite caracteres que no sean compatibles con su formato. Y una variable de cadena no admite caracteres que excedan del ancho especificado.

- Una variable con formato numérico admite valores con una anchura superior a la establecida, pero en ese caso la casilla muestra el valor en notación científica (o muestra un asterisco), indicando esto que la anchura del valor supera la anchura establecida para la variable. Puede verse el valor completo ensanchando la columna, pero ensanchar una columna no altera la anchura definida para la variable; para cambiar la anchura de una variable es necesario utilizar las columnas **Tipo** o **Anchura** del *Editor de variables* (ver Figura 4.2).
- Al introducir un valor en una columna vacía, el SPSS crea una nueva variable y le asigna un nombre por defecto. Si el valor introducido es un número, el SPSS asigna formato numérico a la nueva variable. Si el valor introducido no es un número, le asigna formato de cadena.

## Editar datos

El *Editor de datos* ofrece la posibilidad de modificar el archivo de datos de múltiples maneras. Para modificar, por ejemplo, el valor de una casilla cualquiera:

- Colocar el cursor en la casilla en la que se encuentra el valor que se desea modificar (el valor de la casilla seleccionada aparece en el *Display del Editor de datos*) y escribir el valor deseado.
- Alternativamente, pinchar con el puntero del ratón en el *Display del Editor de datos* y editar el valor original, es decir, el valor correspondiente a la casilla activa (pueden utilizarse las teclas de borrado para corregir errores). Tras ello, presionar la tecla de retorno de carro (o cualquiera de las flechas, o la tecla de tabulador) para que el nuevo valor ocupe la casilla seleccionada.

Además de modificar valores, el *Editor de datos* permite cortar, copiar y pegar valores individuales o áreas rectangulares, borrar casos y variables, buscar datos, etc. Todas estas funciones, similares a las de otras aplicaciones Windows, se encuentran disponibles en el menú **Edición**. Otras funciones de edición, como insertar casos o variables nuevas, o localizar un caso de forma rápida, se encuentran en el menú **Datos**.

Por último, el menú **Ver** incluye algunas funciones relacionadas con el aspecto del *Editor de datos*.

## Deshacer/rehacer

Para anular el efecto de las últimas acciones de edición:

- Seleccionar la opción **Deshacer** (o **Rehacer**) del menú **Edición**. Se consigue el mismo efecto pulsando los botones *Deshacer* y *Rehacer* de la barra de herramientas.

Tras eliminar un valor o un caso, tras reemplazar el contenido de una casilla, tras crear o eliminar una variable, etc., la opción **Deshacer** deja las cosas exactamente como estaban justo antes de la(s) última(s) acción(es). Y después de deshacer una acción, la opción **Rehacer** la restaura.

## Seleccionar datos

La opción **Seleccionar** del menú **Edición** no está disponible cuando la ventana activa es el *Editor de datos* (sí lo está en el resto de ventanas). Pero puede seleccionarse un valor, un caso, una variable, o un conjunto de casillas, tanto con el ratón como con el teclado.

Para **seleccionar un valor**:

- Situar el cursor en la casilla que lo contiene.

Para **seleccionar un caso**:

- Pinchar con el puntero del ratón sobre la cabecera de la fila que contiene ese caso. Se consigue el mismo efecto situando el cursor en cualquier casilla correspondiente a ese caso y pulsando simultáneamente la tecla *mayúsculas* y la *barra espaciadora*.

Para **seleccionar una variable**:

- Pinchar con el puntero del ratón sobre la cabecera de la variable. Se consigue el mismo efecto situando el cursor en cualquier casilla correspondiente a esa variable y pulsando simultáneamente la tecla *control* y la *barra espaciadora*.

Para **seleccionar un área rectangular** (un conjunto de casillas):

- Situar el cursor en un extremo del rectángulo y arrastrar el puntero del ratón hasta el extremo opuesto. Se consigue idéntico efecto yendo hasta el extremo opuesto con las flechas de movimiento mientras se mantiene pulsada la tecla de *mayúsculas*.

## Mover y copiar datos

Para mover y copiar datos puede procederse de forma similar a como se hace en otras aplicaciones que funcionan en entorno Windows:

- La opción **Edición > Cortar** (teclado: *control + x*) elimina el texto seleccionado (ya sea una casilla, un caso, una variable o un conjunto de casillas) y lo lleva al portapapeles de Windows.
- La opción **Edición > Copiar** (teclado: *control + c*) hace una copia del texto seleccionado (ya sea una casilla, un caso, una variable o un conjunto de casillas) y la lleva al portapapeles de Windows.
- La opción **Edición > Pegar** (teclado: *control + v*) inserta el contenido del portapapeles en la ventana activa a partir del punto en el que se encuentra el cursor. La ventana activa puede ser tanto el *Editor de datos* como una ventana de resultados o de sintaxis. También puede pegarse el contenido del portapapeles en una aplicación externa que funcione en entorno Windows.

Al mover o copiar datos, el formato original es sustituido por el formato de las nuevas columnas que pasan a ocupar. Si la conversión de formato de un dato no es posible, el dato se convierte en un valor perdido definido por el sistema.



## Borrar datos

Para eliminar el texto seleccionado (una casilla, un caso, una variable, o un conjunto de casillas, de casos o de variables):

- Seleccionar la opción **Eliminar** del menú **Edición**. La tecla *suprimir* produce el mismo efecto.

## Buscar datos

Esta opción permite buscar un valor concreto en los casos de la variable seleccionada (es decir, en los casos de la columna en la que se encuentra el cursor). Para buscar un dato:

- Seleccionar la opción **Buscar...** del menú **Edición** (o pulsar el botón *Buscar* de la barra de herramientas) para acceder al cuadro de diálogo *Buscar datos en variable...* que muestra la Figura 4.10.

Figura 4.10. Cuadro de diálogo *Buscar datos*



Para buscar un *dato* concreto:

- Introducir el valor buscado en el cuadro de texto **Buscar qué**.
- Pulsar el botón **Buscar siguiente**. La búsqueda se realiza desde la posición del cursor *hacia adelante*; cuando la búsqueda llega al final del archivo continúa con el primer caso.

Para buscar una *etiqueta de valor* en lugar de un valor:

- Activar la opción **Etiquetas de valor** del menú **Ver** antes de entrar en el cuadro de diálogo *Buscar datos* (ver más adelante, en este mismo capítulo, el apartado *Modificar el aspecto del Editor de datos*).
- **Coincidir mayúsculas y minúsculas**. Al activar esta casilla, la búsqueda distingue entre mayúsculas y minúsculas.

El botón **Detener** permite detener la búsqueda sin abandonar el cuadro de diálogo. El botón **Cancelar** detiene la búsqueda y cierra el cuadro de diálogo.

## Buscar casos

Cuando se trabaja con archivos de datos muy grandes, puede ocurrir que sea necesario invertir demasiado tiempo en buscar un caso concreto. Para evitar este problema, el SPSS incluye una función de búsqueda que permite posicionar el cursor de forma rápida en el lugar deseado. Para buscar un caso:

- Seleccionar la opción **Ir a caso...** del menú **Datos** (o pulsar el botón *Ir a caso* de la barra de herramientas), para acceder al cuadro de diálogo *Ir a caso* que muestra la Figura 4.11.

Figura 4.11. Cuadro de diálogo *Ir a caso*



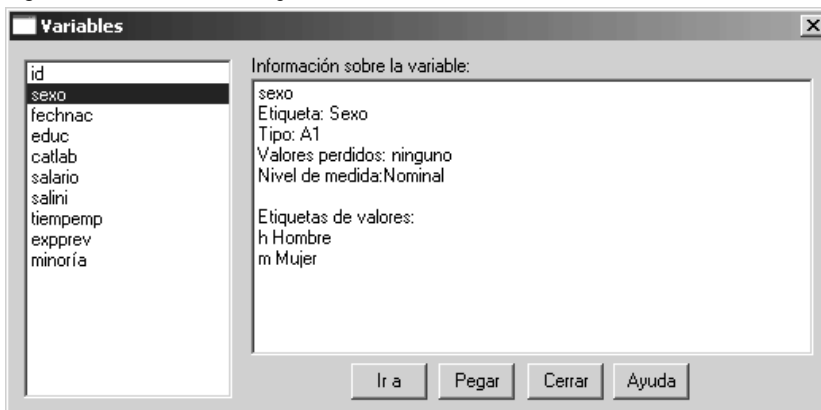
Para situar el cursor en un caso concreto del archivo de datos basta con introducir su número de fila en el cuadro de texto **Número de caso** y pulsar el botón **Aceptar**.

## Buscar variables

Si el archivo de datos contiene muchas variables y éstas no se encuentran en orden alfabético, puede que, para encontrar una variable, sea necesario invertir demasiado tiempo recorriendo el archivo de datos. Afortunadamente, la tarea de encontrar una variable concreta puede convertirse en algo rápido y sencillo utilizando la opción de búsqueda de variables del *Editor de datos*. Para buscar una variable:

- Seleccionar la opción **Variables...** del menú **Utilidades** (o pulsar el botón *Variables* de la barra de herramientas) para acceder al cuadro de diálogo *Variables* que muestra la Figura 4.12.

Figura 4.12. Cuadro de diálogo *Variables*



Este cuadro de diálogo ofrece, además de información detallada sobre cada variable, la posibilidad de posicionar el cursor de forma instantánea en cualquier variable del archivo de datos. Para ello:

- Seleccionar, en la lista de variables del archivo de datos, la variable en la que se desea colocar el cursor.
- Pulsar el botón **Ir a** para cerrar el cuadro de diálogo, volver al *Editor de datos* y situar el cursor en la variable seleccionada.

El recuadro **Información sobre la variable** muestra la siguiente información sobre la variable seleccionada: el nombre de la variable, su etiqueta (si la tiene), su formato, incluyendo la anchura, el número de valores perdidos definidos por el usuario, el nivel de medida de la variable y las etiquetas de los valores.

El botón **Pegar** cierra el cuadro de diálogo y pega, en la ventana designada del *Editor de sintaxis*, el nombre de las variables seleccionadas. Si no existe ninguna ventana del *Editor de sintaxis* abierta, el botón **Pegar** abre una ventana nueva y pega en ella los nombres de las variables.

## Insertar variables nuevas

Para insertar una variable nueva (una columna nueva) entre dos variables existentes:

- Situar el cursor en la columna donde se desea insertar la nueva variable.
- Seleccionar la opción **Insertar variable** del menú **Datos**, o pulsar el botón *Insertar variable* de la barra de herramientas.

La variable recién insertada pasa a ocupar la columna inmediatamente anterior (a la izquierda) a la de la variable en la que se encuentra el cursor. Todas las variables situadas a la derecha de la posición del cursor (incluida la variable en la que se encuentra el cursor) son desplazadas una columna hacia la derecha.

La primera variable insertada durante una sesión recibe un nombre por defecto: *var00001*. Conforme se van insertando o creando nuevas variables, el prefijo *var* del nombre por defecto va siendo acompañado de números enteros consecutivos: *var00002*, *var00003*, etc.

## Insertar casos nuevos

Para insertar un caso nuevo (una fila nueva) entre dos casos existentes:

- Situar el cursor en la fila en la que se desea ubicar el nuevo caso.
- Seleccionar la opción **Insertar caso** del menú **Datos**, o pulsar el botón-ícono *Insertar caso* de la barra de herramientas.

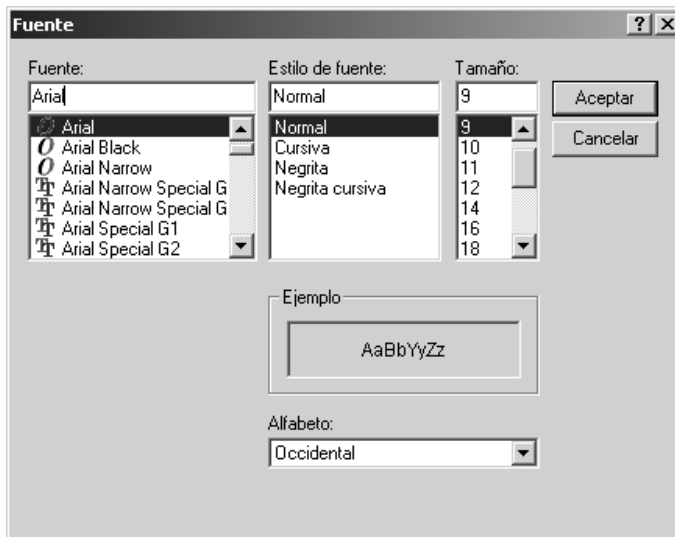
El caso recién insertado pasa a ocupar la fila inmediatamente anterior (por encima) a la del caso en el que se encuentra el cursor. Todos los casos situados por debajo de la posición del cursor (incluido el caso en el que se encuentra el cursor) se desplazan una fila hacia abajo.

## Modificar el aspecto del Editor de datos

Los menús **Ver** y **Ventana** contienen una serie de opciones que permiten cambiar el aspecto del *Editor de datos*. Los cambios de aspecto afectan tanto a la forma en que los datos son presentados en la pantalla, como al resultado de *imprimir* el contenido del *Editor de datos*. Para modificar el aspecto del *Editor de datos*:

- Seleccionar la opción **Barra de estado** del menú **Ver** para ocultar/mostrar la barra de estado (ver, en el Capítulo 1, el apartado *Las barras de estado*).
- Seleccionar la opción **Barra de herramientas...** del menú **Ver** para, entre otras cosas, ocultar/ mostrar la barra de herramientas (ver, en el Capítulo 1, el apartado *Las barras de herramientas*).
- Seleccionar la opción **Fuentes...** del menú **Ver** para acceder al cuadro de diálogo *Fuentes* que muestra la Figura 4.13. Este cuadro de diálogo permite controlar la fuente (el tipo, el tamaño y el estilo de la letra) de los diferentes componentes del *Editor de datos*. La fuente seleccionada afecta a las cabeceras de los casos y de las variables, a los valores y a sus etiquetas, y al contenido del *Display del Editor de datos*.

Figura 4.13. Cuadro de diálogo *Fuentes*



- Seleccionar la opción **Cuadrícula** del menú **Ver** para activar y desactivar la presencia del reticulado del *Editor de datos* (el reticulado se refiere a las líneas o bordes que delimitan las casillas).
- Seleccionar la opción **Etiquetas** del menú **Ver** (o pulsar el botón *Mostrar etiquetas de valor* de la barra de herramientas) para controlar el contenido visualizado en las casillas del *Editor de datos* (no disponible en el *Editor de variables*). Con las variables a cuyos valores se les han asignado etiquetas, es posible optar entre visualizar los valores o las etiquetas.

- Seleccionar la opción **Segmentar** del menú **Ver** para dividir en sub-ventanas la ventana del *Editor de datos* (no disponible en el *editor de variables*). Esta opción permite dividir en 4 partes la ventana del *Editor de datos*. De este modo, es posible desplazarse por el archivo de datos viendo simultáneamente partes diferentes del mismo.

Si el cursor se encuentra en la primera casilla del archivo, la división se realiza dividiendo la pantalla horizontal y verticalmente por la mitad. Si el cursor se encuentra fuera de la primera casilla, la división se realiza insertando un corte horizontal por encima de la casilla seleccionada y otro vertical a la izquierda de la casilla seleccionada.

Para eliminar la segmentación de la ventana del *Editor de datos*, seleccionar la opción **Eliminar segmentación** del menú **Ventana**.

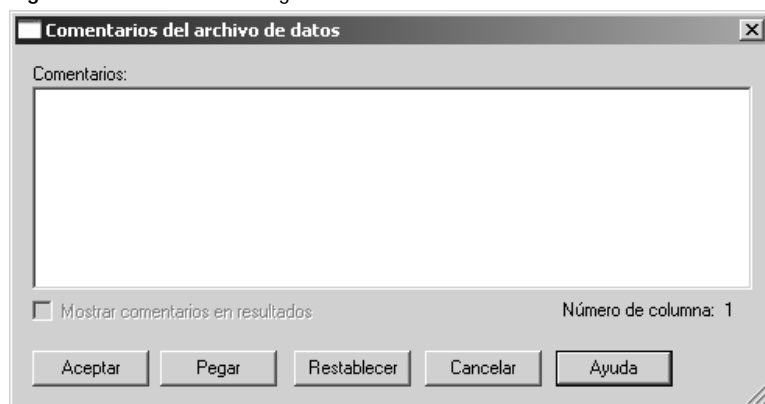
## Adosar comentarios al archivo de datos

El SPSS ofrece la posibilidad de añadir comentarios descriptivos a un archivo de datos. Estos comentarios pueden incluir cualquier información que se desee y pueden consultarse tanto con el *Editor de datos* como con el *Visor de resultados*.

Para añadir un comentario a un archivo de datos o para consultar comentarios previamente añadidos:

- Seleccionar la opción **Comentarios del archivo de datos...** del menú **Utilidades** para acceder al cuadro de diálogo *Comentarios del archivo de datos* que muestra la Figura 4.14.

Figura 4.14. Cuadro de diálogo *Comentarios del archivo de datos*



El cuadro de texto **Comentarios** permite introducir comentarios de cualquier longitud (no existe limitación en el tamaño de los comentarios). Una vez introducidos los comentarios, el botón **Aceptar** hace que el comentario pase a formar parte del archivo de datos. Al guardar el archivo, los comentarios se guardan con él.

“ **Mostrar comentarios en resultados**. Si se pulsa el botón **Aceptar** con esta opción activada, el *Visor de resultados* muestra los comentarios en formato de tabla (con la fuente utilizada por defecto para generar tablas).

Cada vez que se crea o modifica un comentario, el SPSS inserta la fecha actual al final del comentario. Por tanto, si se modifica un comentario existente o el comentario nuevo se inserta entre otros comentarios ya existentes, la fecha original es alterada.

## Trabajar con conjuntos de variables

Es habitual que los archivos de datos incluyan una gran cantidad de variables. Sin embargo, también es habitual que sólo interese efectuar análisis de datos sobre un conjunto reducido de variables. Trabajar con sólo unas pocas variables cuando el archivo de datos contiene muchas resulta algo engorroso, pues es necesario ir buscando las variables que interesan dentro de la lista de variables que ofrecen los distintos cuadros de diálogo. Este problema puede resolverse trabajando con *conjuntos de variables*.

### Definir conjuntos de variables

Para definir conjuntos de variables:

- Seleccionar la opción **Definir conjuntos...** del menú **Utilidades** para acceder al cuadro de diálogo *Definir conjuntos de variables* que muestra la Figura 4.14.

Figura 4.14. Cuadro de diálogo *Definir conjuntos de variables*



En el cuadro de texto **Nombre del conjunto** de la Figura 4.14 se ha definido el conjunto SALARIO (el conjunto DEMOGR viene predefinido con el archivo *Datos de empleados*). Para definir un conjunto:

- Escribir el nombre del conjunto en el cuadro de texto **Nombre del conjunto**. En el ejemplo de la Figura 4.14 se ha introducido el nombre SALARIO.

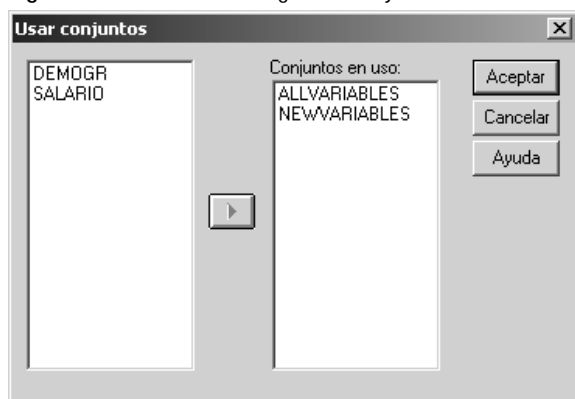
- Seleccionar las variables que se desea que formen parte del conjunto y trasladarlas a la lista **Variables del conjunto** (esto puede hacerse con el botón flecha o pulsando dos veces sobre cada variable con el botón principal del ratón). En el ejemplo de la Figura 4.14 se han seleccionado las variables *salario* y *salini*.
- Pulsar el botón **Añadir conjunto** para que el conjunto recién definido (SALARIO en el ejemplo) pase a formar parte de la lista de conjuntos. Los botones **Cambiar conjunto** y **Borrar conjunto** permiten modificar y eliminar, respectivamente, conjuntos previamente definidos.

## Usar conjuntos de variables

Una vez definido un conjunto es necesario activarlo para conseguir que las listas de variables de los cuadros de diálogo sólo muestren las variables de ese conjunto. Para ello:

- Seleccionar la opción **Usar conjuntos...** del menú **Utilidades** para acceder al cuadro de diálogo *Definir conjuntos* que muestra la Figura 4.15.

Figura 4.15. Cuadro de diálogo *Usar conjuntos*



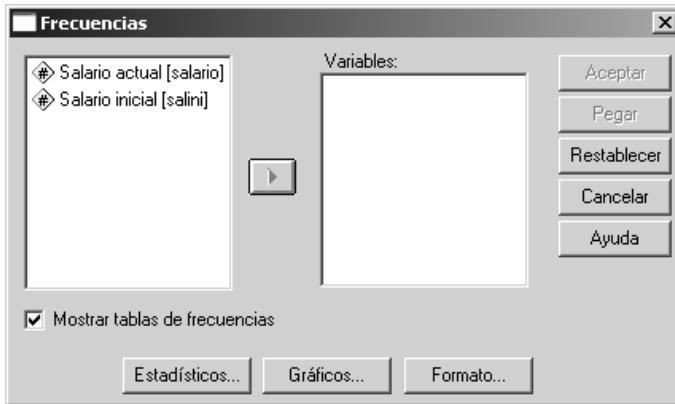
El listado **Conjuntos en uso**, contiene un listado de los conjuntos que están siendo utilizados. El SPSS tiene predefinidos dos conjuntos de variables y ambos se encuentran activos por defecto: el conjunto **ALL VARIABLES**, que incluye todas las variables del archivo de datos, y el conjunto **NEW VARIABLES**, que incluye todas las nuevas variables que se van creando durante una sesión.

Para activar un conjunto distinto de los que están actuando por defecto, basta con seleccionarlo en el listado de la izquierda y trasladarlo (mediante el botón flecha o pulsando dos veces el botón principal del ratón) al listado **Conjuntos en uso**.

Con el conjunto **ALL VARIABLES** (=todas las variables) en uso, al utilizar un cuadro de diálogo cualquiera, por ejemplo, el cuadro de diálogo *Frecuencias*, la lista de variables del archivo de datos muestra *todas las variables* del archivo de datos. Si se desea que la lista de variables sólo muestre las variables nuevas que se van creando a lo largo de una sesión (y que todavía no han sido guardadas), puede dejarse en uso únicamente el conjunto **NEW VARIA-**

BLES (= variables nuevas). Sin embargo, si se activa el conjunto SALARIO y se desactivan los dos conjuntos predefinidos (retirándolos de la lista **Conjuntos en uso**), el listado de variables que aparece en los distintos cuadros de diálogo SPSS sólo mostrará las variables de ese conjunto (la Figura 4.16 refleja esta circunstancia).

Figura 4.16. Cuadro de diálogo *Frecuencias*



Por supuesto, el beneficio de utilizar conjuntos de variables es más evidente cuando el archivo de datos contiene gran cantidad de variables y sólo se tiene intención de trabajar con algunas de ellas.

**Nota:** Otros aspectos relacionados con el *Editor de datos*, tales como el cálculo de variables nuevas o la recodificación de variables ya existentes, la categorización de variables, la selección de casos mediante filtros, la fusión de archivos de datos, etc., se tratan en los Capítulos 5 y 6.





## Transformar datos

Puede ocurrir que, al introducir los datos de un estudio en el *Editor de datos* o al importarlos desde una fuente externa, el resultado obtenido sea de tal índole que resulte posible y tenga sentido aplicar directamente el análisis estadístico deseado. Pero esto sólo ocurrirá en una situación más bien ideal; y las situaciones ideales raramente se presentan.

Más bien al contrario, lo habitual será encontrarse con archivos de datos que necesitarán ser cuidadosamente preparados antes de poder aplicar un análisis estadístico con las mínimas garantías.

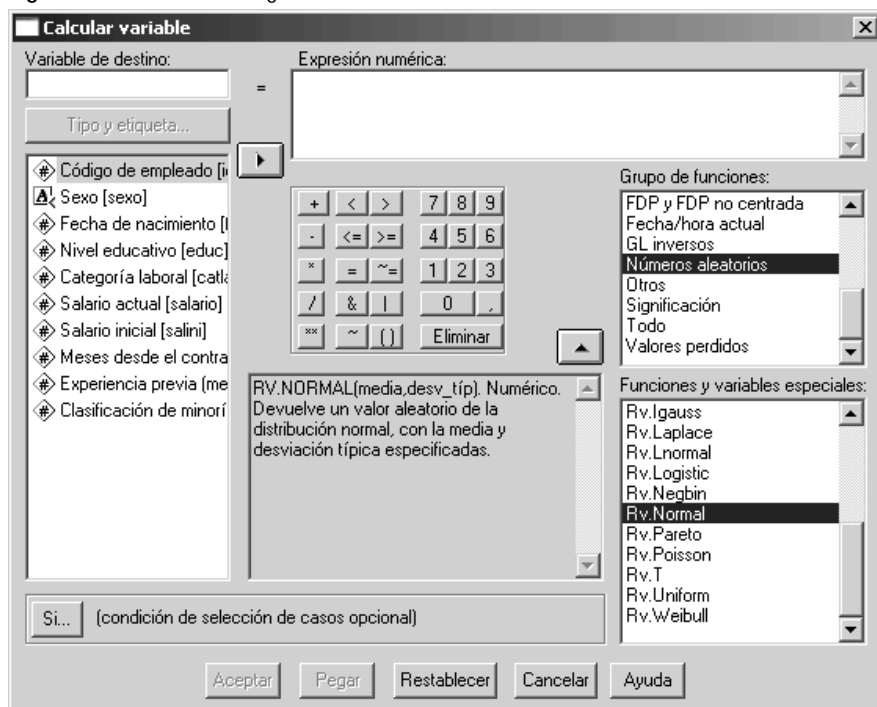
La *preparación* del archivo de datos incluye desde la simple detección y corrección de los posibles errores cometidos al introducir datos, hasta sofisticadas transformaciones (necesarias a veces para llegar a obtener las variables que realmente interesa analizar), pasando por la recodificación de los códigos utilizados para los valores de alguna variable, o la creación de nuevas variables a partir de otras ya existentes.

El menú **Transformar** de la barra de menús principal incluye una serie de opciones que permiten efectuar diferentes tipos de transformaciones. Este capítulo ofrece una descripción de todas ellas: **Calcular** (para crear variables nuevas, bien a partir de otras variables existentes, bien a partir de algún tipo de función), **Recodificar** (para cambiar los códigos de las variables), **Categorizador visual** (para convertir variables cuantitativas en categóricas), **Contar apariciones** (para contar el número de veces que aparece un valor o conjunto de valores en una o más variables), **Asignar rangos** (para asignar enteros consecutivos a los valores de una variable), **Recodificación automática** (para asignar enteros consecutivos de forma secuencial, es decir, tratando los empates como un valor único), **Fecha/Hora** (para hacer cálculos y transformaciones con fechas y unidades de tiempo), **Crear serie temporal** (para suavizar los valores de una serie temporal), **Reemplazar valores perdidos** (para sustituir valores perdidos utilizando diferentes funciones), **Generadores de números aleatorios** (para elegir el generador de números aleatorios y controlar el valor inicial) y **Ejecutar transformaciones pendientes** (para ejecutar transformaciones pendientes previamente definidas mediante sintaxis).

### Calcular

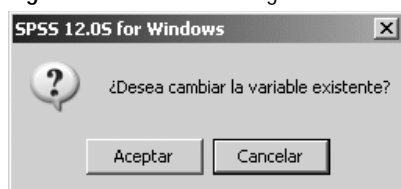
El SPSS contiene una potente opción que permite crear variables nuevas a partir de otra u otras existentes o a partir de alguna de las casi 200 funciones que incluye. Para crear una variable nueva:

- Seleccionar la opción **Calcular...** del menú **Transformar** para acceder al cuadro de diálogo *Calcular variable* (ver Figura 5.1).

Figura 5.1. Cuadro de diálogo *Calcular variable*

## Variable de destino

El cuadro de texto **Variable de destino** (ver Figura 5.1) permite introducir el nombre de la variable que recibirá los valores calculados. El nombre de esta variable puede ser nuevo o puede ser el de una variable ya existente. Si se propone un nombre nuevo, éste debe respetar las reglas de los nombres de variable (ver, en el Capítulo 4, el apartado *Definir variables*). Si el nombre propuesto para la variable de destino coincide con el de una variable ya existente, al pulsar el botón **Aceptar** aparece un mensaje de aviso (ver Figura 5.2) solicitando confirmar o cancelar la acción.

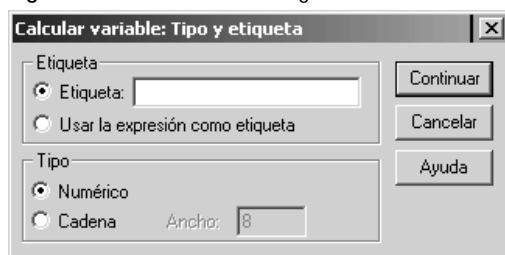
Figura 5.2. Cuadro de diálogo de *advertencia de variable duplicada*

## Tipo de variable y etiqueta

A la variable de destino se le asigna, por defecto, formato numérico. No obstante, es posible controlar el *tipo* de las nuevas variables. Para cambiar el *tipo* de la nueva variable y/o asignarle una etiqueta:

- Pulsar el botón **Tipo y etiqueta...** (ver Figura 5.1) para acceder al subcuadro de diálogo *Calcular variable: Tipo y etiqueta* que muestra la Figura 5.3.

Figura 5.3. Subcuadro de diálogo *Calcular variable: Tipo y etiqueta*



Las opciones del recuadro **Etiqueta** permiten asignar una etiqueta descriptiva a la nueva variable: la opción **Etiqueta** permite introducir una etiqueta descriptiva de 256 caracteres como máximo; la opción **Usar la expresión como etiqueta** permite utilizar como etiqueta los primeros 110 caracteres de la expresión numérica.

Las opciones del recuadro **Tipo** sirven para cambiar el tipo de formato de la nueva variable. Las nuevas variables reciben, por defecto, formato **Numérico**; pero también es posible seleccionar un formato de **Cadena** (si se opta por este formato es necesario especificar el ancho de la cadena en el cuadro de texto **Ancho**).

## Expresión numérica

En este cuadro de texto (ver Figura 5.1) debe escribirse la expresión numérica encargada de generar los valores de la variable de destino.

Una expresión numérica es una expresión matemática similar a la que puede construirse con una calculadora de bolsillo, con la diferencia de que, además de constantes y operadores aritméticos, admite nombres de variables ya existentes, operadores relacionales y lógicos, y una gran variedad de funciones matemáticas. Todos estos detalles se describen a continuación.

## Calculadora

Para facilitar la tarea de construir la expresión numérica, el cuadro de diálogo *Calcular variable* incluye un teclado de calculadora con números y operadores aritméticos, relacionales y lógicos (ver Tabla 5.1). Este teclado funciona exactamente igual que el de una calculadora convencional (pero pinchando con el puntero del ratón).

En relación con los operadores *aritméticos* hay que tener presente, sobre todo, el *orden* en el que operan: en primer lugar se evalúan las funciones (ver siguiente apartado), después las potencias, después la multiplicación y la división y, por último, la suma y la resta. Cuando hay varios operadores *lógicos* en una expresión, primero actúa NOT, después OR y por último AND. Para alterar el orden natural de los operadores es necesario utilizar paréntesis. Los operadores *relacionales* y *lógicos* son útiles, sobre todo, para efectuar transformaciones condicionales (enseguida serán tratadas). Algunos ejemplos de expresiones numéricas:

- 1.  $2 * (var1 - var2) ** 2$
- 2.  $(var1 + var2 + var3) / 3$
- 3.  $var1 * 0.05$
- 4.  $var1 * 1.25 + 32$


La primera expresión calcula el doble de la diferencia al cuadrado entre las variables *var1* y *var2*: cada caso del archivo de datos pasa a tener, en la nueva variable, el doble del valor resultante de elevar al cuadrado la resta de sus puntuaciones en *var1* y *var2*. La segunda expresión calcula la media aritmética de las variables *var1*, *var2* y *var3*: cada caso del archivo de datos pasa a tener, en la nueva variable, la media aritmética de sus puntuaciones en esas tres variables. La tercera expresión calcula el 5 por ciento de la variable *var1*: el valor de cada caso en la nueva variable será el 5 por ciento de su puntuación en *var1*. La expresión 4 incrementa el valor de la nueva variable en un 25 por ciento y al resultado le suma la constante 32.

Tabla 5.1. Operadores del cuadro del diálogo *Calcular variable*

Operadores aritméticos	Operadores relacionales	Operadores lógicos
+ : suma	= o EQ : igual que	& o AND : y
- : resta	< o LT : menor que	u OR : o
* : multiplicación	> o GT : mayor que	~ o NOT : no
/ : división	<= o LE : menor o igual que	
** : potencia	>= o GE : mayor o igual que	
	~= o NE : distinto de	

Funciones

El cuadro de diálogo *Calcular variable* incluye cerca de 200 funciones aritméticas, estadísticas, de cadena, de fecha, etc. Para trasladar una función al cuadro de texto **Expresión numérica**:

- ' Seleccionar el **Grupo de funciones** en el que se encuentra la función que se desea utilizar y elegir la función en la lista **Funciones y variables especiales**.
- ' Pulsar el botón *flecha*  (o pulsar dos veces sobre la función) para trasladarla al cuadro de texto **Expresión numérica**.
- **Funciones aritméticas**: valor absoluto, seno, coseno, arcoseno, arcotangente, logaritmo natural, logaritmo en base 10, exponente, raíz cuadrada, etc. Por ejemplo, la expresión

ARSIN(SQRT(*var1*)) calcula la función arcoseno de la raíz cuadrada de la variable *var1* (transformación ésta muy utilizada cuando las puntuaciones de la variable dependiente de un análisis de varianza son proporciones). A cada caso del archivo de datos se le asigna el valor resultante de la expresión.

- **Funciones de búsqueda:** para encontrar valores dentro de una cadena, señalar la posición de determinados valores dentro de una cadena, encontrar el valor más pequeño o más grande de una variable numérica, etc. Por ejemplo la expresión INDEX(*var1*, 'abc') permite obtener un número que indica cuál es la posición que ocupa la primera aparición de los caracteres abc dentro de la cadena *var1*.
- **Funciones de cadena:** para concatenar argumentos, convertir mayúsculas en minúsculas o al revés, buscar elementos, extraer elementos, añadir y truncar elementos, etc. Por ejemplo, la expresión LTRIM(*var1*, '0') elimina de los valores de la variable *var1* (que debe ser una variable con formato de *cadena*) los ceros situados a la izquierda. Y la expresión CONCAT(*var1*, *var2*) fusiona en una sola cadena las cadenas *var1* y *var2*.
- **Funciones de conversión:** para convertir variables numéricas en cadenas y cadenas en variables numéricas. Por ejemplo, si *var1* es una cadena cuyos códigos son números de cuatro dígitos, para transformar *var1* en una variable numérica podría utilizarse la expresión NUMBER(*var1*,f4).
- **Funciones de fecha:** para leer fechas en distintos formatos, extraer parte de la fecha, convertir fechas a unidades de tiempo, convertir unidades de tiempo a fechas, etc. Por ejemplo, en la expresión CTIME.DAYS(DATE.DMY(20,02,1990)–fechnac)/365.25, si *fechnac* es la fecha de nacimiento (variable con formato *fecha*), la función DATE.DMY calcula la edad a día 20-02-1990 (en segundos), la función CTIME.DAYS convierte los segundos en días, y al dividir por 365.25 se obtiene la edad en años.
- **Funciones estadísticas:** suma, media aritmética, desviación típica, varianza, etc. Por ejemplo, la expresión MEAN(*var1*, *var2*, *var3*) calcula la media aritmética de las variables *var1*, *var2* y *var3*. A cada caso del archivo de datos se le asigna la media de sus valores en las tres variables. La función MEAN, al igual que la función SUM, es especialmente útil para combinar variables, tal como se hace, por ejemplo, cuando se suman o promedian las puntuaciones de varias preguntas de un cuestionario para obtener una puntuación total; si un caso tiene valor perdido en alguna de las variables que forman parte del argumento (*var1*, *var2* y *var3* en el ejemplo), la media o la suma se calcula (no arroja valor perdido) con las puntuaciones del resto de las variables del argumento.
- **Funciones de variables aleatorias y de distribuciones de probabilidad:** función de distribución, función de distribución inversa, funciones de probabilidad y de densidad de probabilidad, funciones de distribución no centradas, generación de variables aleatorias, etc. La mayoría de estas funciones están disponibles para múltiples modelos teóricos de probabilidad: Bernoulli, Binomial, Binomial negativa, Poisson, Geométrica, Hipergeométrica, Normal, *t*, *F*, Exponencial, Chi-cuadrado, Gamma, Uniforme, Cauchy, Logística, Log-normal, Pareto, etc. Por ejemplo, la expresión RV.NORMAL (0,1) genera una variable aleatoria (*RV* = *random variable*) distribuida normalmente con media 0 y desviación típica 1; cada caso del archivo de datos recibe un valor aleatorio; este conjunto de valores se distribuye de forma aproximadamente normal; y la media y la desviación típica se aproximan a 0 y a 1, respectivamente, tanto más cuanto mayor es el número de valores generados. La expresión CDFNOR(1.96) calcula la probabilidad de encontrar valores

iguales o menores que 1,96 en la distribución normal tipificada (media 0 y desviación típica 1). La expresión `IDF.CHI(0.95, 30)` calcula el valor que deja por debajo de sí una probabilidad de 0,95 en la distribución *chi*-cuadrado con 30 grados de libertad. La expresión `PDF.BINOM(3,10,0.5)` calcula la probabilidad de obtener 3 éxitos en 10 ensayos cuando la probabilidad de éxito en cada ensayo es de 0,5.

- **Funciones de valores perdidos:** para identificar valores perdidos, contar el número de valores perdidos o el número de casos válidos en un conjunto de variables, etc. Por ejemplo, la expresión `NMIS(var1 to var2)` cuenta, para cada caso, el número de valores perdidos que existen en el conjunto de variables incluidas en el argumento.

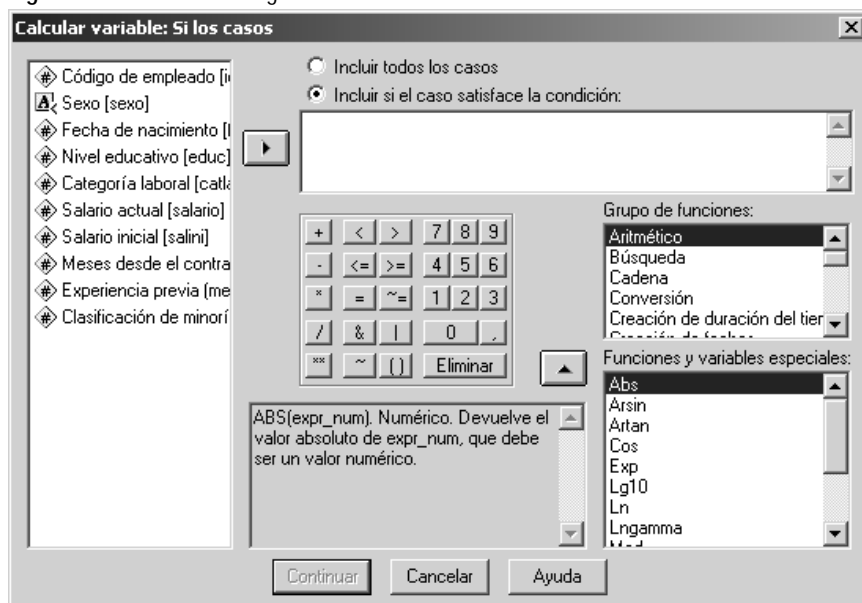
Es importante recordar que una expresión numérica no está completa hasta que se ha incluido entre paréntesis el *argumento* correspondiente a la función seleccionada. La ayuda específica del cuadro de diálogo contiene información puntual sobre cada función.

## Expresiones condicionales

Definida la expresión numérica, el botón **Aceptar** crea, en el *Editor de datos*, una nueva variable en la que *todos los casos válidos* del archivo de datos adoptan el valor resultante de la expresión numérica propuesta. Ahora bien, la expresión numérica no tiene por qué afectar a todos los casos del archivo. De hecho, existe la posibilidad de establecer una condición cualquiera y hacer que las transformaciones propuestas afecten sólo a los casos que cumplan esa condición. Para ello:

- Pulsar el botón **Si...** del cuadro de diálogo *Calcular variable* (ver Figura 5.1) para acceder al subcuadro de diálogo *Calcular variable: Si los casos* que muestra la Figura 5.4.

Figura 5.4. Cuadro de diálogo *Calcular variable: Si los casos*



Este subcuadro de diálogo permite establecer una gran variedad de condiciones para seleccionar sólo los casos que interesen:

**Incluir todos los casos.** Las transformaciones propuestas afectan a todos los casos. Es la opción por defecto.

**Incluir si el caso satisface la condición.** Hace que las transformaciones propuestas afecten únicamente a los casos que cumplan la condición establecida. La condición puede incluir nombres de variables, constantes, operadores aritméticos, relacionales y lógicos, y funciones matemáticas. Para construir la condición se dispone de un teclado de calculadora y de una lista de funciones idénticos a los del cuadro de diálogo *Calcular variable* (ver Figura 5.1).

Si la variable de destino es una variable nueva, los casos seleccionados (aquellos que cumplen la condición establecida) adoptan en ella el valor resultante de la expresión numérica; y los casos no seleccionados (los que no cumplen la condición establecida) se consideran valores perdidos (aparecen puntos en las casillas del *Editor de datos*). Si la variable de destino es una variable ya existente, los casos seleccionados adoptan el valor resultante de la expresión numérica y los casos no seleccionados quedan como estaban.

La expresión numérica puede construirse utilizando la lista de variables, los botones de la calculadora y la lista de funciones. Pero también puede construirse utilizando el teclado, en cuyo caso hay que tener en cuenta unas pocas reglas para no cometer errores:

- Los valores de las variables de cadena deben escribirse entre apóstrofes o entre comillas. Hay que tener especial cuidado en no dejar espacios en blanco donde no deba haberlos.
- Los argumentos de una función deben ir entre paréntesis y, cuando haya más de uno, separados por comas. Puede insertarse un espacio en blanco entre un paréntesis y un argumento, o entre un argumento y otro, pero no es necesario.
- Dentro de una expresión compleja, todas las expresiones simples deben estar completas. Por ejemplo,  $\text{edad} > 18 \& < 30$  es una expresión incorrecta, por incompleta. La expresión correcta es:  $\text{edad} > 18 \& \text{edad} < 30$ .
- El punto es el único separador decimal válido, independientemente de las especificaciones internacionales seleccionadas en Windows.

### **Ejemplo: Calcular > Si...**

Un sencillo ejemplo servirá para formarse una idea clara acerca de cómo se crean variables utilizando expresiones condicionales. Se va a crear la variable *salario2* aplicando a la variable *salario* una subida del 10% para las *mujeres administrativas* y del 5% para el resto de los casos. Para ello:

- Seleccionar la opción **Calcular...** del menú **Transformar** para acceder al cuadro de diálogo *Calcular variable* (ver Figura 5.1).
- Introducir el nombre *salario2* en el cuadro de texto **Variable de destino**.
- Introducir en el cuadro de texto **Expresión numérica** una expresión que permita incrementar la variable *salario* en un 10 por ciento; por ejemplo:  $\text{salario} * 1.10$ .



- Pulsar el botón Si... para acceder al subcuadro de diálogo *Calcular variable: Si los casos* (ver Figura 5.4) y establecer la condición necesaria para aplicar la expresión numérica sólo a las mujeres administrativas: `sexo = 'm'` and `catlab = 1` (los valores de una variable de cadena deben ir entre comillas o apóstrofes). Pulsar el botón Continuar (ver Figura 5.4) para volver al cuadro de diálogo principal y el botón Aceptar (ver Figura 5.1) para hacer efectiva esta primera transformación.
- Volver a entrar en el cuadro de diálogo *Calcular variable* (ver Figura 5.1) e introducir en el cuadro de texto **Expresión numérica** una expresión que permita incrementar la variable *salario* en un 5 por ciento; por ejemplo: `salario * 1.05`.
- Pulsar el botón Si... para acceder al subcuadro de diálogo *Calcular variable: Si los casos* (Figura 5.4) y establecer la condición necesaria para aplicar esta segunda expresión numérica a todos los casos excepto a las mujeres administrativas: `sexo = 'h'` and `catlab = 1` or `catlab > 1`. Pulsar el botón Continuar (ver Figura 5.4) para volver al cuadro de diálogo principal y el botón Aceptar (ver Figura 5.1) para hacer efectiva esta segunda transformación.

## Recodificar

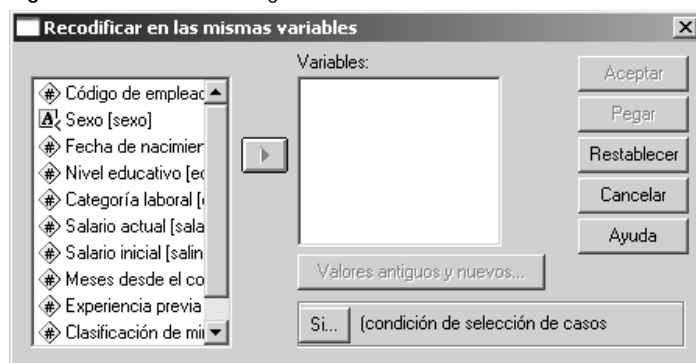
La opción **Recodificar** del menú **Transformar** permite cambiar los códigos asignados a los valores de una variable. La recodificación resulta especialmente útil para agrupar en un único valor diferentes valores de una variable y para transformar variables con formato de cadena en variables con formato numérico. La recodificación puede efectuarse sobre *las mismas variables* (cambiando los códigos de una variable existente sin cambiar su nombre) o sobre *variables distintas* (creando una variable nueva a partir de otra ya existente).

### Recodificar en las mismas variables

Para cambiar los códigos de una variable sin cambiar su nombre:

- Seleccionar la opción **Recodificar > En las mismas variables...** del menú **Transformar** para acceder al cuadro de diálogo *Recodificar en las mismas variables* que muestra la Figura 5.5.

Figura 5.5. Cuadro de diálogo *Recodificar en las mismas variables*



**Variables.** Las variables cuyos códigos se desea cambiar deben trasladarse a esta lista. Para trasladar variables:

- Seleccionar la variable en la lista de variables del archivo de datos y desplazarla a la lista **Variables** mediante el botón flecha. Puede seleccionarse más de una variable si es que interesa efectuar *la misma recodificación* a más de una variable.

En la recodificación pueden utilizarse tanto variables numéricas como variables de cadena, pero no al mismo tiempo. Es decir, todas las variables llevadas a la lista **Variables** deben tener el mismo formato. Al trasladar la primera variable, si es numérica, la lista **Variables** cambia su nombre a **Variables numéricas**; si esa primera variable es de cadena, el nombre de la lista cambia a **Variables de cadena**.

**Si...** La recodificación puede efectuarse de forma condicional, es decir, de forma que sólo afecte a los casos que cumplan determinada condición. Para establecer una condición:

- Pulsar el botón **Si...** para acceder al subcuadro de diálogo *Recodificar en las mismas variables: Si los casos* (idéntico al de la Figura 5.4). Este subcuadro contiene todas las opciones necesarias para efectuar transformaciones condicionales.

**Valores antiguos y nuevos...** Una vez seleccionadas las variables que se van a recodificar, es necesario indicar la recodificación concreta que se desea llevar a cabo. Para ello, el botón **Valores antiguos y nuevos...** conduce al subcuadro de diálogo *Recodificar en las mismas variables: Valores antiguos y nuevos*: *Valores antiguos y nuevos* que muestra la Figura 5.6.

Figura 5.6. Subcuadro de diálogo *Recodificar en las mismas variables: Valores antiguos y nuevos*

**Valor antiguo.** Las opciones de este recuadro permiten especificar el valor o valores de la variable original que se desea recodificar. Ofrece varias alternativas para facilitar la identificación del valor o valores originales (antiguos):

**Valor.** Un valor individual.

**Perdido por el sistema.** Valores perdidos definidos por el sistema (aparecen como SYSMIS en la lista de valores **Antiguo ! Nuevo**). Esta opción no está disponible con variables de cadena.

**Perdido por el sistema o usuario.** Valores perdidos de cualquier tipo: definidos por el sistema o definidos por el usuario (aparecen como MISSING en la lista de valores **Antiguo ! Nuevo**).

**Rango (\_\_\_ hasta\_\_\_).** Rango de valores comprendidos entre los dos valores indicados. Esta opción no está disponible con variables de cadena.

**Rango (Del menor hasta\_\_\_).** Rango comprendido entre el valor más pequeño de la variable y el valor indicado. Opción no disponible con variables de cadena.

**Rango (\_\_\_ hasta el mayor).** Rango comprendido entre el valor indicado y el valor más grande. Opción no disponible con variables de cadena.

**Todos los demás valores.** Todos los valores de la variable original no definidos previamente (aparece como ELSE en la lista de valores).

**Valor nuevo.** En este recuadro debe especificarse el nuevo valor que se desea asignar al valor o valores antiguos. Ofrece las siguientes alternativas:

**Valor.** El valor introducido en este cuadro de texto sustituye al valor o valores antiguos recién definidos. Si se trata de una variable de cadena, el valor introducido en este cuadro de texto aparece entre apóstrofes o comillas al trasladarlo a la lista de valores: por tanto, no hay que poner apóstrofes o comillas a los valores de las variables de cadena, como ocurre, por ejemplo, en los cuadros de diálogo que permiten definir una expresión numérica.

**Perdido por el sistema.** Asigna un valor perdido definido por el sistema (aparece como SYSMIS en la lista de valores). Estos valores perdidos aparecen en el *Editor de datos* como un punto.

Para cada valor o rango de valores que se desee recodificar:

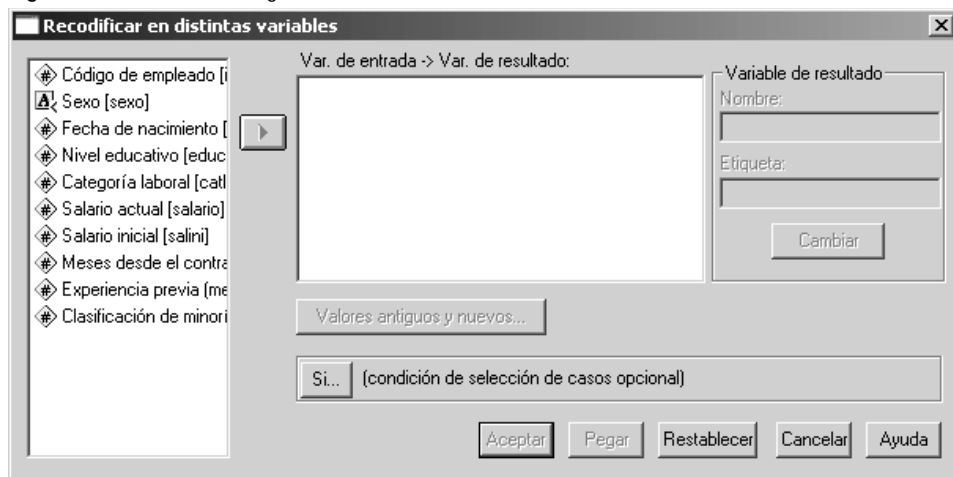
- Indicar el valor o rango de valores antiguos que se desea recodificar utilizando alguna de las opciones del recuadro **Valor antiguo**.
- Especificar el nuevo código en el cuadro de texto **Valor nuevo**.
- Pulsar el botón **Añadir** para trasladar la recodificación recién definida a la lista **Antiguo ! Nuevo**.
- Utilizar los botones **Cambiar** y **Borrar** para modificar o eliminar, respectivamente, recodificaciones previamente añadidas.

Debe tenerse presente que, aunque es posible asignar *el mismo valor nuevo* a *más de un valor antiguo* (lo cual puede resultar especialmente útil, por ejemplo, para categorizar variables), no es posible asignar *más de un valor nuevo* a *un solo valor antiguo*.

## Recodificar en distintas variables

Para cambiar los códigos de una variable y, al mismo tiempo, crear una variable nueva con los nuevos códigos (dejando intacta la variable original):

- Seleccionar la opción **Recodificar > En distintas variables...** del menú **Transformar** para acceder al cuadro de diálogo *Recodificar en distintas variables* que muestra la Figura 5.7.

Figura 5.7. Cuadro de diálogo *Recodificar en distintas variables*

Este cuadro de diálogo permite crear variables nuevas a partir de los valores de las variables ya existentes en el *Editor de datos*. El cuadro es muy similar al de la Figura 5.5; la diferencia entre ambos está, únicamente, en que aquí es necesario asignar nombre (y, opcionalmente, etiqueta) a las nuevas variables que se desea crear. Para iniciar la recodificación en variables diferentes:

- Seleccionar, en la lista de variables, la variable que se desea recodificar y trasladarla, mediante el botón flecha, a la lista **Var. de entrada ! Var. de resultado**.
- En el recuadro **Variable de resultado**, introducir el nombre elegido para la nueva variable en el cuadro de texto **Nombre** y, si se desea, una etiqueta descriptiva en el cuadro de texto **Etiqueta** (256 caracteres como máximo).
- Pulsar el botón **Cambiar** para activar el nuevo nombre y situarlo en la lista **Var. de entrada ! Var. de resultado** junto al nombre de la variable original.

**Si...** La recodificación puede efectuarse de forma condicional, es decir, de forma que sólo afecte a los casos que cumplan determinada condición. Para ello:

- Pulsar el botón **Si...** para acceder al subcuadro de diálogo *Recodificar en distintas variables: Si los casos* (este cuadro de diálogo es idéntico al de la Figura 5.4, y contiene todas las opciones necesarias para efectuar transformaciones condicionales).

**Valores antiguos y nuevos...** Una vez seleccionadas las variables que se van a recodificar, es necesario definir la recodificación concreta que se desea llevar a cabo. Para ello:

- Pulsar el botón **Valores antiguos y nuevos...** para acceder al subcuadro de diálogo *Recodificar en distintas variables: Valores antiguos y nuevos* (ver Figura 5.8).

Este subcuadro de diálogo es idéntico al de la Figura 5.6, pero contiene un elemento adicional en el recuadro **Valor nuevo**:

**Copiar valores antiguos.** Al marcar esta opción, el valor o valores antiguos seleccionados se mantienen sin cambios en la nueva variable.

Figura 5.8. Subcuadro de diálogo *Recodificar en distintas variables: Valores antiguos y nuevos*

Este subcuadro de diálogo también contiene dos elementos adicionales referidos a las variables de cadena:

- " **Las variables de resultado son cadenas.** Si la variable receptora de los nuevos códigos es una variable de cadena, es necesario marcar esta opción y especificar el ancho de la cadena en el cuadro de texto **Ancho**.
- " **Convertir cadenas numéricas en números.** Esta opción permite transformar una variable de cadena en una variable numérica. Para que la transformación sea posible, los valores de la cadena deben ser únicamente números (opcionalmente acompañados de un signo +, de un signo –, o de un separador decimal). Si se activa esta opción y la cadena contiene algún carácter no numérico, la recodificación arroja un valor perdido definido por el sistema.

Los valores antiguos no seleccionados pasan a ser, en la nueva variable, valores perdidos definidos por el sistema. La opción **Todos los demás valores** del recuadro **Valores antiguos** combinada con la opción **Copiar valores antiguos** del recuadro **Valores nuevos** resulta especialmente útil para conseguir que los valores antiguos no seleccionados pasen a formar parte de la nueva variable sin ser convertidos en valores perdidos.

## Categorizar variables

Categorizar una variable consiste en crear una variable categórica (es decir, una variable con un número limitado de niveles) agrupando los valores contiguos de una variable cuantitativa. Por ejemplo, la *edad* de los sujetos medida en años y meses, que es una variable cuantitativa, puede transformarse en una variable categórica creando, por ejemplo, cinco grupos o tramos de edad. Este tipo de transformación lleva asociada una inevitable pérdida de información, pero puede resultar útil si se tiene interés en construir tablas de contingencias cruzando la variable *edad* con otras variables categóricas como el *sexo* o la *categoría laboral*; o si se desea utilizar la *edad* como variable independiente en, por ejemplo, un análisis de varianza.

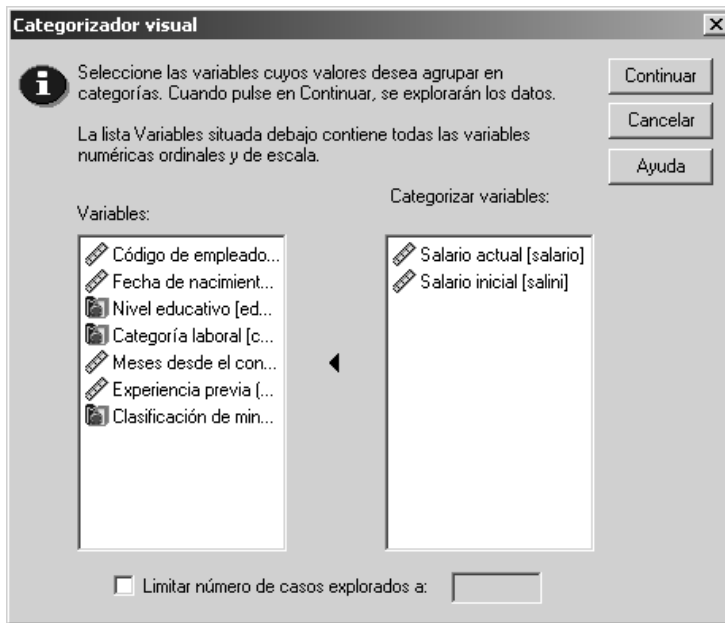
También puede categorizarse reduciendo el número de categorías de una variable ordinal. Por ejemplo, una variable medida con una escala ordinal de 1 a 9 podría categorizarse agrupando las puntuaciones de tres en tres para crear una variable con tres niveles: «bajo», «medio» «alto».

## Categorizador visual

Para categorizar variables:

- Seleccionar la opción **Categorizador visual...** del menú **Transformar** para acceder al cuadro de diálogo *Categorizador visual* (paso 1) que muestra la Figura 5.9.

Figura 5.9 Cuadro de diálogo del *Categorizador visual* (paso 1)



La lista de variables del archivo de datos muestra un listado con todas las variables numéricas que tienen asignada una medida *ordinal* o de *escala*. La lista no incluye las variables numéricas *nominales* ni las variables de *cadena* porque el *Categorizador visual* no permite categorizar este tipo de variables. Sólo es posible categorizar variables numéricas ordinales y de escala: el proceso de categorización asume que los valores de la variable que se desea categorizar están ordenados de alguna forma lógica que garantiza el agrupamiento de valores contiguos con sentido. Para categorizar variables:

- Seleccionar la(s) variable(s) que se desea categorizar de la lista **Variables** y trasladarla(s), mediante el botón flecha o arrastrándola(s) con el puntero del ratón, a la lista **Categorizar variables** (en el ejemplo de la Figura 5.9 se han seleccionado las variables *salario actual* y *salario inicial*) y pulsar el botón **Continuar** para acceder al cuadro de diálogo *Categorizador visual* (paso 2) que muestra la Figura 5.10.

- “ Limitar el número de casos explorados a. Para que el SPSS pueda ayudar a automatizar el proceso de categorización de variables necesita explorar el archivo de datos. Si el archivo es demasiado largo y no se desea explorar todos los casos, esta opción permite indicar al SPSS cuántos casos debe explorar (introduciendo el valor  $n$  en el correspondiente cuadro de texto se exploran los primeros  $n$  casos). Aunque esta opción puede tener alguna utilidad (especialmente con archivos muy grandes), debe tenerse en cuenta que limitar el número de casos que son explorados podría conducir a que el SPSS no reconociera todos los valores distintos de una variable (particularmente si el archivo se encuentra ordenado por esa variable).

Figura 5.10. Cuadro de diálogo del *Categorizador visual* (paso 2)

**Categorizador visual**

Lista de variables exploradas

Variable actual:  Etiqueta:

Variable categorizada:

Mínimo:  Valores no perdidos Máximo:

Introduzca puntos de corte de los intervalos o pulse en Crear puntos de corte para generar los intervalos automáticamente. Por ejemplo, un valor de 10 define un intervalo que comienza encima del intervalo previo y finaliza en 10.

Rejilla:

	Valor	Etiqueta
1	SUPERIOR	
2		

Límites superiores

☒ Incluidos (<=)

☐ Excluidos (<)

☐ Invertir escala

Copiar categorías

**Lista de variables exploradas.** El cuadro de diálogo ofrece, en primer lugar, un listado de las variables previamente seleccionadas (ver Figura 5.9) y que acaban de ser exploradas. Pinchando con el puntero del ratón sobre la cabecera de la columna **Variable**, las variables listadas pueden ordenarse alfabéticamente (por orden ascendente o descendente, pinchando sucesivamente). La primera columna de este listado (encabezada **M...**; ensanchando la columna puede verse el encabezado **Medida**) indica, mediante el correspondiente símbolo, el nivel de medida de la variable (ordinal o escala). Las variables también pueden ordenarse por su nivel de medida pinchando con el puntero del ratón sobre la cabecera de esta columna.

**Casos explorados.** Indica el número de casos explorados. Tanto los cálculos (percentiles, desviaciones típicas) como el histograma se basan en el número de casos válidos. Por tanto, en

el procedimiento intervienen todos los casos explorados exceptuando los que tienen valor perdido de cualquier tipo (definidos por el sistema o definidos por el usuario).

**Valores perdidos.** Indica el número detectado de valores perdidos de cualquier tipo: definidos por el sistema y definidos por el usuario. Aunque los valores perdidos no se incluyen en ninguna de las categorías de la nueva variable, los valores perdidos definidos por el usuario pasan a ser también valores perdidos definidos por el usuario en la nueva variable (con las mismas etiquetas, si existen). Si el código de alguno de estos valores perdidos entra en conflicto con alguno de los códigos asignados a la nueva variable, el código del valor perdido es recodificado añadiendo 100 al código más alto de la nueva variable (a no ser que los valores perdidos de la variable original hayan sido definidos como un rango de valores, en cuyo caso el código de valor perdido en la nueva variable será un número negativo).

**Variable actual.** Muestra el nombre y la etiqueta (si existe) de la variable seleccionada en la Lista de variables exploradas.

**Variable categorizada.** Estos cuadros de texto ofrecen la posibilidad de introducir un nombre (obligatorio) y una etiqueta (opcional) para la nueva variable, es decir, para la variable categorizada. Lógicamente, el nombre de la nueva variable debe ajustarse a las reglas de los nombres de variable del SPSS (ver, en este mismo capítulo, el apartado *Definir variables*). Si no se introduce ninguna etiqueta, el procedimiento utiliza la de la variable original (o el nombre de la variable original, si no existe etiqueta) seguido de la palabra «Categorizada» entre paréntesis.

**Mínimo y Máximo.** Informan del valor más pequeño y del más grande encontrados en la variable seleccionada en la Lista de variables exploradas. Sólo se tienen en cuenta los casos válidos. Entre ambos valores se encuentra un título (**Valores no perdidos**) que indica que, en el histograma de la variable seleccionada en la Lista de variables exploradas, únicamente intervienen los casos válidos.

## Definir categorías manualmente

Las casillas de la **Rejilla** central permiten introducir manualmente (en la columna **Valor**) los puntos de corte que se desea utilizar para categorizar la variable, así como las etiquetas que se desea asignar a cada categoría (en la columna **Etiqueta**). Para introducir estos valores o etiquetas manualmente basta con situar el cursor en una casilla en blanco y utilizar el teclado. Para borrar valores o etiquetas previamente introducidos, basta con situar el cursor en la correspondiente casilla y pulsar la tecla suprimir (o las opciones del menú emergente que se obtiene pulsando el botón secundario del ratón).

Los puntos de corte son los límites superiores que definen cada categoría de la nueva variable. Las dos opciones situadas a la derecha de la rejilla permiten decidir si se desea que el punto de corte quede incluido en la categoría inferior (**Incluido**,  $\leq$ ) o en la superior (**Excluido**,  $<$ ). A las categorías definidas por estos puntos de corte ( $k$  = número de puntos de corte) se les asigna, en la nueva variable, valores enteros consecutivos de 1 a  $k+1$  (a no ser que se marque la opción **Escala inversa** situada en la parte inferior derecha del cuadro de diálogo, en cuyo caso los códigos asignados serán enteros consecutivos de  $k+1$  a 1).



Al introducir un número en la columna **Valor**, el dibujo del histograma muestra una línea vertical justamente por el punto de corte introducido. Estos puntos de corte pueden modificarse arrastrando esas líneas verticales hacia derecha e izquierda. Y un punto de corte ya introducido puede eliminarse arrastrando la correspondiente línea vertical fuera del histograma.

La primera casilla de la columna **Valor** de la rejilla central incluye, por defecto, el valor **SUPERIOR**; este valor es necesario para definir la última categoría. De hecho, sirve para definir la categoría que contendrá todos los valores más altos que el mayor de los puntos de corte definidos. Si se elimina este valor, los casos cuyo valor sea mayor que el mayor punto de corte se convierten en valores perdidos definidos por el sistema (casillas vacías en el *Editor de datos*).

La columna **Etiquetas** permite introducir una etiqueta (opcional) para describir el contenido de cada nueva categoría. El botón **Crear etiquetas** sirve para asignar automáticamente como etiquetas los puntos de corte utilizados.

## Definir categorías automáticamente

Los puntos de corte pueden ser definidos de forma automática. Para ello, el botón **Crear puntos de corte** conduce al subcuadro de diálogo *Crear puntos de corte* que muestra la Figura 5.11. Las opciones de este subcuadro de diálogo permiten definir categorías de forma automática a partir de puntos de corte basados en tres criterios diferentes: categorías o intervalos de la misma amplitud, categorías o intervalos con el mismo número de casos y categorías o intervalos basados en la media y la desviación típica.

Figura 5.11. Subcuadro de diálogo *Crear puntos de corte*

El subcuadro de diálogo **Crear puntos de corte** presenta tres métodos para definir automáticamente los puntos de corte:

- Intervalos de igual amplitud** (seleccionado): Requiere rellenar al menos dos campos:
  - Posición del primer punto de corte: [campo de texto]
  - Número de puntos de corte: [campo de texto]
  - Amplitud: [campo de texto]
  - Posición del último punto de corte: [campo de texto]
- Percentiles iguales basados en los casos explorados**: Requiere rellenar cualquiera de los dos campos:
  - Número de puntos de corte: [campo de texto]
  - % de casos: [campo de texto]
- Puntos de corte en media y desviaciones típicas seleccionadas, basadas en casos explorados**: Permite seleccionar una o más opciones:
  - ☐ +/- 1 Desv. típica
  - ☐ +/- 2 Desv. típicas
  - ☐ +/- 3 Desv. típicas

En la parte inferior, un icono de información indica: "Aplicar reemplazará las definiciones de los puntos de corte actuales con esta especificación. Un intervalo final incluirá todos los valores restantes: N puntos de corte generan N+1 intervalos."

Botones de acción: **Aplicar**, **Cancelar**, **Ayuda**.

**Intervalos de igual amplitud.** Genera categorías o intervalos de la misma amplitud (por ejemplo: 1–3, 4–6, 7–9, etc.) a partir de tres valores (de los cuales es necesario introducir al menos dos):

- " **Posición del primer punto de corte.** Límite superior de la primera categoría o intervalo. Todos los casos cuyo valor sea menor que este primer punto de corte se incluyen en la primera categoría.
- " **Número de puntos de corte.** El número de categorías o intervalos de la nueva variable depende del número de puntos de corte elegido:  $k$  puntos de corte definen  $k+1$  categorías o intervalos.
- " **Amplitud.** Anchura de cada categoría o intervalo, excepto de la primera (cuya amplitud viene definida por el primer punto de corte) y de la última (cuya amplitud viene definida por el primer punto de corte y por el número de puntos de corte).

**Percentiles iguales basados en los casos explorados.** Genera categorías o intervalos con aproximadamente el mismo número de casos. Estas categorías o intervalos se basan en los percentiles calculados con el método AEMPIRICAL (ver, en el Capítulo 11 sobre *Análisis exploratorio*, el apartado *Estadísticos*).

Cuando la variable que se desea categorizar mediante esta estrategia tiene poca definición métrica (es decir, tiene un pequeño número de valores distintos o un gran número de casos con el mismo valor), el procedimiento genera menos categorías o intervalos de los solicitados (informando de ello en un cuadro de advertencia):

- " **Número de puntos de corte.** El número de categorías o intervalos de la nueva variable depende del número de puntos de corte elegido:  $k$  puntos de corte definen  $k+1$  categorías o intervalos. El número de puntos de corte determina qué percentiles se utilizarán para efectuar la categorización. Una especificación de, por ejemplo, 1 punto de corte crea la nueva variable categórica asignando un 1 a los casos con puntuaciones iguales o menores que la mediana (percentil 50) y un 2 a los casos con puntuaciones mayores que la mediana. Y una especificación de, por ejemplo, 3 puntos de corte crea la nueva variable categórica asignando un 1 a los casos que se encuentran por debajo del percentil 25, el valor 2 a los casos comprendidos entre el percentil 25 y el 50, el valor 3 a los casos comprendidos entre el percentil 50 y el 75, y el valor 4 a los casos situados por encima del percentil 75. De este modo, cada categoría de la nueva variable pasa a tener aproximadamente un 25 % de los casos. Esta forma de categorizar variables basada en los valores de los percentiles se encuentra también disponible en la opción *Ntiles* del procedimiento *Asignar rangos* que se describe más adelante en este mismo capítulo.
- " **% de casos.** Anchura de cada categoría o intervalo expresada como un porcentaje del número total de casos. Si se elige, por ejemplo, un porcentaje de 25, la nueva variable categórica tendrá 4 categorías basadas en los cuartiles.

**Puntos de corte basados en la media y en la desviación típica de los casos explorados.** Genera categorías o intervalos basados en la media y en la desviación típica de la variable que se desea categorizar (estos valores se calculan teniendo en cuenta únicamente los valores explorados). Las opciones de este recuadro permiten elegir a qué distancia de la media se desea colocar los puntos de corte: a una, a dos o a tres desviaciones típicas.

**Copiar categorías.** Las dos opciones de este recuadro (situado en la parte inferior izquierda de la Figura 5.10) permiten intercambiar entre variables puntos de corte previamente definidos. Para poder utilizar estas opciones es necesario que la **Lista de variables exploradas** contenga más de una variable. Para asignar las categorías de otra variable a la variable seleccionada:

- Seleccionar la variable a la cual se desea asignar las categorías.
- Pulsar el botón **De otra variable** del recuadro **Copiar categorías** para acceder al subcuadro de diálogo *Copiar categorías a la variable actual* (ver Figura 5.12.a). Este subcuadro permite asignar a la variable previamente seleccionada (variable *actual*) los puntos de corte de cualquier variable de la lista.

Para asignar las categorías de la variable actual a otra(s) variable(s):

- Seleccionar la variable que contiene las categorías que se desea copiar.
- Pulsar el botón **A otra variable** del recuadro **Copiar categorías** para acceder al subcuadro de diálogo *Copiar categorías de la variable actual* (ver Figura 5.12.b). Este subcuadro permite asignar las categorías de la variable previamente seleccionada (variable *actual*) a cualquiera de las variables de la lista.

Figura 5.12.a. Subcuadro de diálogo *Copiar categorías a la variable actual*

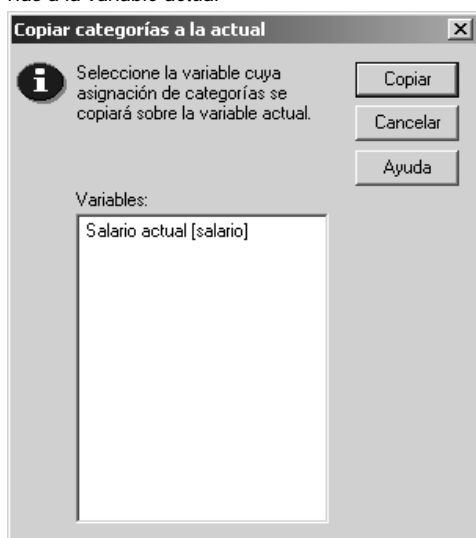
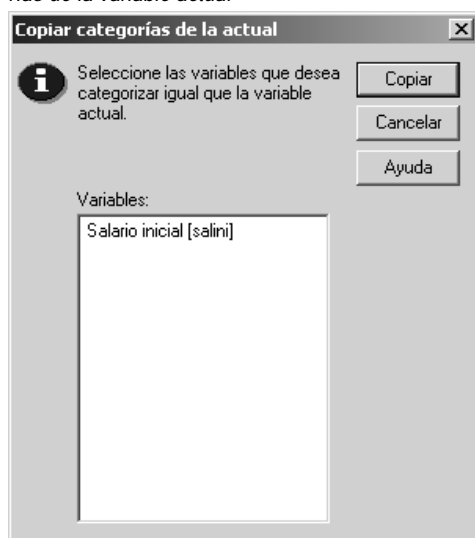


Figura 5.12.b. Subcuadro de diálogo *Copiar categorías de la variable actual*



## Contar apariciones

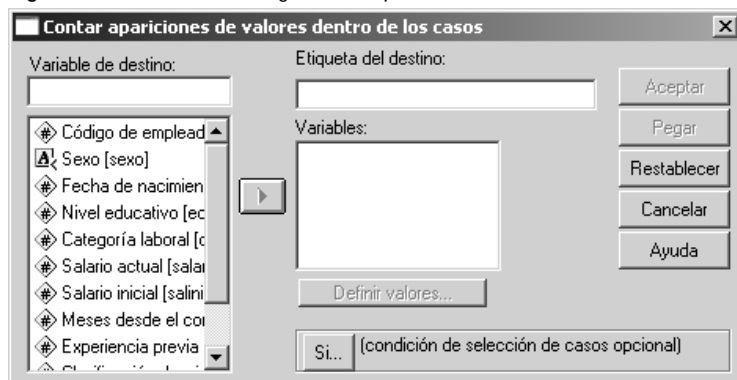
Esta opción sirve para crear variables nuevas a partir del número de veces que se repite uno o más valores determinados en un conjunto de variables. Una variable creada con esta opción contiene, para cada caso, el valor resultante de contar el número de veces que el valor o valores seleccionados aparecen en el conjunto de variables seleccionadas.

La opción **Contar apariciones** es útil, por ejemplo, para contar el número de valores perdidos que acumula cada caso del archivo en un determinado número de variables, o para calcular el número de aciertos en un conjunto de preguntas de opción múltiple (preguntas en las que los sujetos responden seleccionando una alternativa entre varias posibles, y más tarde es necesario averiguar cuántas respuestas de cada tipo ha dado cada sujeto para poder calcular el número de aciertos), o para contar el número de veces que un conjunto de jueces puntúan a cada caso por encima de un determinado valor, etc.

Para contar valores:

- Seleccionar la opción **Contar apariciones...** del menú **Transformar** para acceder al cuadro de diálogo *Contar apariciones de valores dentro de los casos* que muestra la Figura 5.13.

Figura 5.13. Cuadro de diálogo *Contar apariciones de valores dentro de los casos*



**Variable de destino.** Este cuadro de texto permite asignar un nombre a la nueva variable, es decir, a la variable que recogerá el resultado del recuento.

El nombre de la variable de destino puede ser nuevo o puede ser el de una variable ya existente. Si el nombre elegido es nuevo, éste debe respetar las reglas de los nombres de variable del SPSS (ver, en el Capítulo 4, el apartado *Definir variables*). Si el nombre elegido para la variable de destino es el de una variable ya existente, al pulsar el botón **Aceptar** aparece un mensaje de aviso (ver Figura 5.2) solicitando confirmar o cancelar la acción. Lógicamente, puesto que la variable de destino tiene que recoger el resultado de un recuento, debe tener formato numérico.

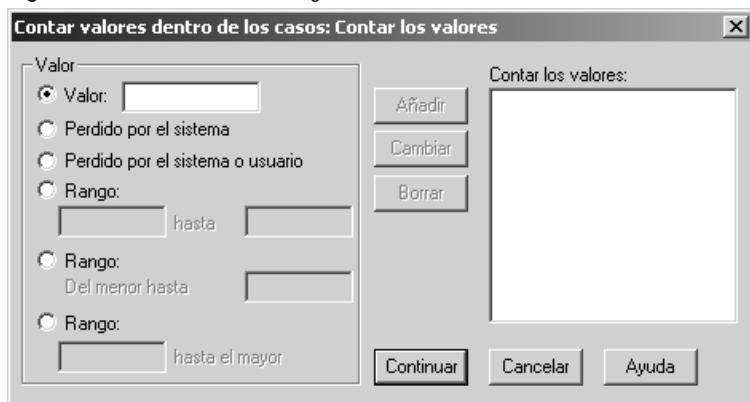
**Etiqueta del destino.** Permite asignar una etiqueta descriptiva a la variable de destino (256 caracteres como máximo). Si la variable de destino ya existe y tiene etiqueta, la etiqueta original aparece en este cuadro de texto.

**Variables.** Las variables seleccionadas, es decir, las variables sobre las que se desea efectuar el recuento, deben trasladarse a esta lista. Estas variables pueden ser numéricas o de cadena, pero no es posible mezclar ambos tipos.

**Definir valores.** Antes de iniciar el recuento de casos es necesario indicar qué valores se desea contar. Para ello:

- Pulsar el botón **Definir valores...** para acceder al subcuadro de diálogo *Contar valores dentro de los casos: Contar los valores* que muestra la Figura 5.14.

Figura 5.14. Subcuadro de diálogo *Contar valores dentro de los casos: Contar los valores*



El recuadro **Valor** contiene varias opciones para especificar el valor o valores que se desea contar. Permite seleccionar valores individuales, rangos de valores o una combinación de ambas cosas:

**Valor.** Para contar el número de veces que aparece un valor concreto.

**Perdido por el sistema.** Para contar el número de valores perdidos definidos por el sistema. En la lista **Contar valores** aparece **SYSMIS**.

**Perdido por el sistema o usuario.** Para contar el número de valores perdidos de cualquier tipo (definidos por el sistema o definidos por el usuario). En la lista **Contar valores** aparece **MISSING**.

**Rango (\_\_\_ hasta \_\_\_).** Para contar el número de veces que aparecen valores comprendidos entre los límites del rango definido. Opción no disponible para variables de cadena.

**Rango (Del menor hasta \_\_\_).** Para contar el número de veces que aparecen valores comprendidos entre el valor más pequeño y el valor especificado. Opción no disponible para variables de cadena.

**Rango (\_\_\_ hasta el mayor).** Para contar el número de veces que aparecen valores comprendidos entre el valor especificado y el mayor. Opción no disponible para variables de cadena.

La función **Contar** efectúa, para cada caso del archivo de datos, un recuento del número de veces que se repite, en el conjunto de variables seleccionadas en la lista **Variables** (Figura 5.13), cualquier valor de los añadidos a la lista **Contar valores** (Figura 5.14). Para construir la lista de valores que se desea contar:

- Seleccionar la opción deseada en el recuadro **Valor** y, en caso necesario, introducir el valor o valores en los correspondientes cuadros de texto.

- Llevar a la lista **Contar los valores**, mediante el botón **Añadir**, el valor o valores definidos.
- Utilizar los botones **Cambiar** y **Borrar** para modificar o eliminar, respectivamente, valores previamente añadidos.

La función **Contar** puede afectar a todos los casos del archivo o a sólo un conjunto de casos que cumplan determinada condición. Si se desea establecer algún filtro:

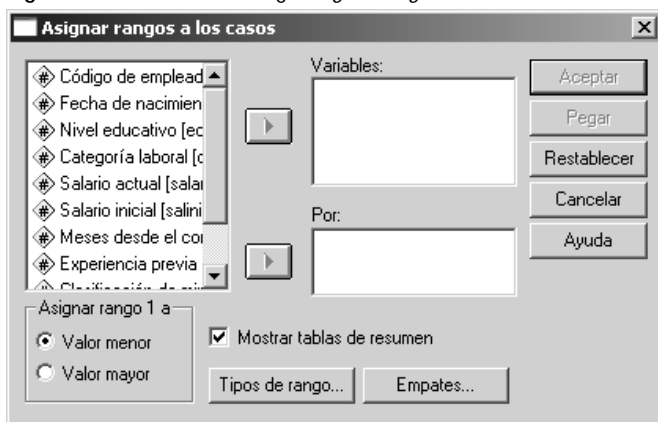
- Pulsar el botón **Si...** (ver Figura 5.13) para acceder al subcuadro de diálogo *Contar apariciones: Si los casos* (idéntico al de la Figura 5.4). En este subcuadro de diálogo es posible establecer las condiciones que deben cumplir los casos que se desea seleccionar.

## Asignar rangos

Asignar rangos consiste en sustituir los valores originales de una variable numérica por enteros consecutivos de 1 a  $n$ . El SPSS incluye diferentes métodos para llevar a cabo esta sustitución. El SPSS también permite asignar puntuaciones normales y de Savage, y agrupar los casos según el percentil que les corresponde. Para asignar rangos:

- Seleccionar la opción **Asignar rangos a casos...** del menú **Transformar** para acceder al cuadro de diálogo *Asignar rangos a los casos* que muestra la Figura 5.15.

Figura 5.15. Cuadro de diálogo *Asignar rangos a los casos*



**Variables.** Para asignar rangos, debe comenzarse seleccionando la variable a la cual se desea asignar rangos. Para ello:

- Marcar la variable en la lista de variables del archivo de datos y trasladarla a la lista **Variables** mediante el botón flecha.

Al pulsar el botón **Aceptar**, el SPSS crea una nueva variable cuyos valores son enteros consecutivos de 1 a  $n$ . La nueva variable recibe de forma automática un nuevo nombre (el nombre de la variable original precedido por una *r*) y una etiqueta. La variable origi-

nal queda intacta. Si no se marca ninguna otra opción, la asignación de rangos se hace en orden *ascendente* (al valor más pequeño se le asigna un 1) y los empates se resuelven asignando a cada valor empatado la media de los rangos que les corresponden (ver más adelante el apartado *Rangos empatados*).

**Por.** La asignación de rangos puede organizarse por subgrupos (es decir, separadamente para cada subgrupo). Para ello, es necesario trasladar a la lista **Por** la variable o variables que definen los subgrupos de interés.

**Asignar rango 1 a.** Las opciones de este recuadro permiten cambiar el orden en el que son asignados los rangos:

**Valor menor.** Asigna los rangos en orden ascendente: al valor más pequeño se le asigna un 1, al valor más pequeño de los restantes se le asigna un 2, etc. Es la opción por defecto.

**Valor mayor.** Asigna los rangos en orden descendente: al valor más grande se le asigna un 1, al valor más grande de los restantes se le asigna un 2, etc.

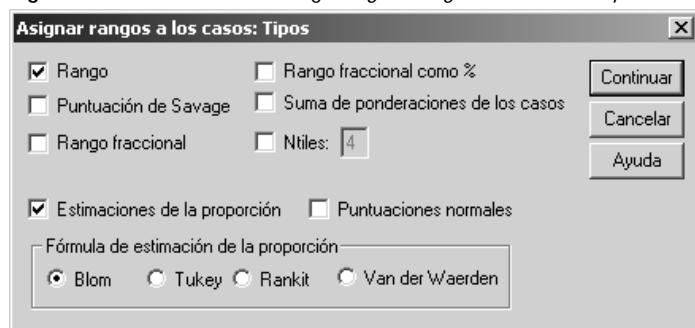
“ **Mostrar tablas de resumen.** Con esta opción activa (se encuentra activa por defecto), al asignar rangos a una variable el SPSS muestra en el *Visor de resultados* una tabla con el nombre de la variable original, el nombre de la nueva variable y una etiqueta descriptiva incluyendo el tipo de rangos utilizados. El *Visor* no muestra toda esta información si se desactiva esta casilla.

## Tipos de rangos

El procedimiento **Asignar rangos** permite elegir entre varios tipos de rangos. Para decidir qué tipo de rangos se desea asignar:

- Pulsar el botón **Tipos de rangos...** (ver Figura 5.15) para acceder al subcuadro de diálogo *Asignar rangos a los casos: Tipos* que muestra la Figura 5.16.

Figura 5.16. Subcuadro de diálogo *Asignar rangos a los casos: Tipos*



Este subcuadro de diálogo permite seleccionar diferentes métodos de asignación de rangos. Si se elige más de un método, el SPSS crea una variable diferente con los rangos correspon-

dientes a cada método seleccionado. La etiqueta que el SPSS asigna de forma automática a cada nueva variable informa sobre el método de asignación de rangos utilizado. El SPSS ofrece seis métodos distintos de asignación de rangos:

- " **Rango.** Asigna enteros consecutivos de 1 a  $n$ . Es la opción por defecto.
- " **Puntuación de Savage.** Asigna puntuaciones basadas en una distribución exponencial.
- " **Rango fraccional.** Asigna el resultado de dividir cada rango por el número de casos válidos
- " **Rango fraccional como %.** Asigna el rango fraccional multiplicado por 100.
- " **Suma de ponderaciones de los casos.** Asigna a cada caso, como único rango, el número de casos válidos. Si se utiliza una variable de agrupación (ver Figura 5.15, lista Por), a los casos de cada grupo se les asigna como puntuación el número de sujetos del grupo al que pertenecen (por tanto, la puntuación asignada es constante para los casos de un mismo grupo).
- " **Ntiles.** Esta opción divide la distribución en  $k$  grupos (grupos de aproximadamente el mismo tamaño basados en el cálculo de percentiles) y asigna a cada caso, como puntuación, un rango de 1 a  $k$  dependiendo del grupo al que es asignado. Una especificación de, por ejemplo, 4 grupos asigna un 1 a los casos que se encuentran por debajo del percentil 25, el valor 2 a los comprendidos entre el percentil 25 y el 50, el valor 3 a los comprendidos entre el percentil 50 y el 75, y el valor 4 a los situados por encima del centil 75. De este modo, cada grupo pasa a tener aproximadamente el 25 % de los casos. Se puede obtener el mismo resultado con el **Categorizador visual** del menú Transformar (ver, en este mismo capítulo, el apartado *Categorizador visual: Definir categorías automáticamente*).

Además de estas opciones para asignar rangos a los casos, el cuadro de diálogo *Asignar rangos* también contiene opciones para: (1) obtener estimaciones de la proporción de casos acumulada hasta cada rango y (2) calcular las puntuaciones típicas normales que corresponden a esas proporciones:

- " **Estimaciones de la proporción.** Estima la proporción acumulada (el área acumulada de la distribución) que corresponde a cada rango concreto.
- " **Puntuaciones normales.** Asigna las puntuaciones  $Z$  que corresponden a las proporciones acumuladas de cada rango en la curva normal tipificada. Por ejemplo, si la proporción acumulada es de 0,50, la puntuación  $Z$  asignada será de 0; si la proporción acumulada es de 0,75, la puntuación  $Z$  estimada será de 0,67; etc.

**Fórmula de estimación de la proporción.** La estimación de la proporción acumulada que corresponde a cada rango puede efectuarse utilizando diferentes procedimientos (en todos los casos,  $R_i$  se refiere al rango asignado y  $n$  al número de casos válidos):

$$\text{Blom: } (R_i - 3/8) / (n + 1/4).$$

$$\text{Tukey: } (R_i - 1/3) / (n + 1/3).$$

$$\text{Rankit: } (R_i - 1/2) / n.$$

$$\text{Van der Waerden: } R_i / (n + 1).$$

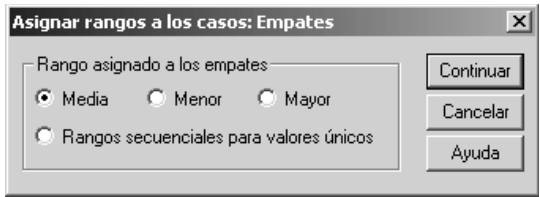


# Rangos empatados

Cuando existen casos con los mismos valores, es decir, casos *empatados*, a cada caso se le asigna, por defecto, el promedio de los rangos que corresponden a esos casos. Para tratar los empates de distinta manera:

- Pulsar el botón **Empates...** (ver Figura 5.15) para acceder al subcuadro de diálogo *Asignar rangos a los casos: Empates* que muestra la Figura 5.17.

**Figura 5.17.** Cuadro de diálogo *Asignar rangos a los casos: Empates*



**Rango asignado a los empates.** El SPSS ofrece cuatro formas distintas de asignar rangos a los empates:

**Media.** Asigna, a los valores de un mismo bloque empates, la media de los rangos que corresponden a los casos de ese bloque. Es la opción por defecto.

**Menor.** Asigna, a los valores de un mismo bloque empates, el menor de los rangos que corresponden a los casos de ese bloque.

**Mayor.** Asigna, a los valores de un mismo bloque empates, el mayor de los rangos que corresponden a los casos de ese bloque.

**Rangos secuenciales para valores únicos.** Asigna rangos de 1 a  $m$ , siendo  $m$  el número de valores distintos. Los casos empatados reciben el mismo rango y cuentan como un único caso para el cómputo del rango siguiente.

La Tabla 5.2 ofrece un ejemplo con 8 valores entre los que se dan cuatro empates. Los rangos asignados a estos 8 valores ilustran cómo afecta a la asignación de rangos cada una de estas formas de tratar los empates.

**Tabla 5.2.** Resultado obtenido con las distintas formas de tratar los empates en la asignación de rangos

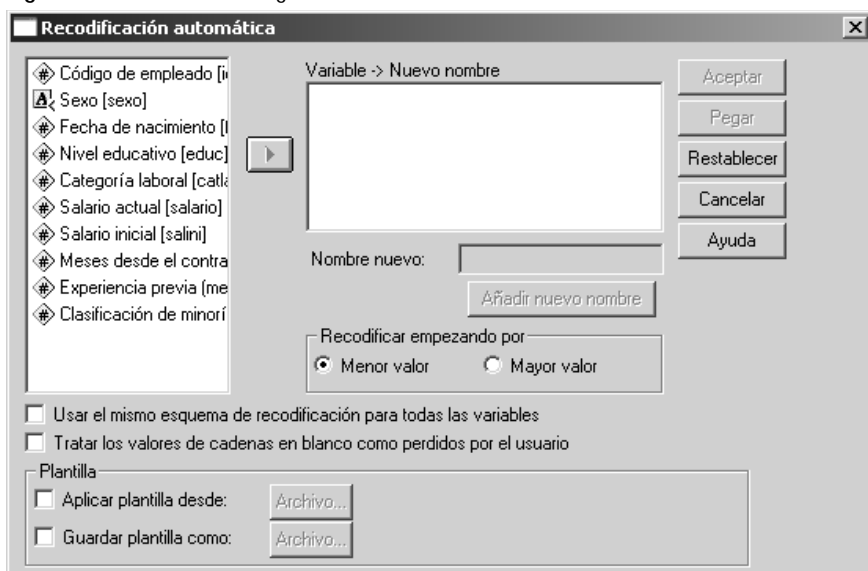
<i>Valor</i>	<i>Media</i>	<i>Menor</i>	<i>Mayor</i>	<i>Secuencial</i>
7	1	1	1	1
9	2	2	2	2
11	4,5	3	6	3
11	4,5	3	6	3
11	4,5	3	6	3
11	4,5	3	6	3
14	7	7	7	4
17	8	8	8	5

## Recodificación automática

La recodificación automática asigna enteros consecutivos de 1 a  $m$  a los  $m$  valores diferentes de una variable (sea ésta numérica o de cadena). Se trata, por tanto, de una asignación de rangos equivalente al *método secuencial* de tratamiento de los empates visto en el apartado anterior. Resulta particularmente útil para crear variables numéricas a partir de variables de cadena (el análisis de varianza, por ejemplo, no admite variables de cadena). También resulta particularmente útil para modificar, por ejemplo, los códigos de las respuestas dadas en una escala de actitudes cuando las puntuaciones altas en unas preguntas indican actitud favorable y en otras preguntas actitud desfavorable y lo que interesa es obtener la puntuación total de la escala sumando las diferentes preguntas; la recodificación automática permite resolver este problema de forma rápida y sencilla. Para asignar  $m$  enteros consecutivos a los  $m$  valores distintos de una variable:

- Seleccionar la opción **Recodificación automática...** del menú **Transformar** para acceder al cuadro de diálogo *Recodificación automática* que muestra la Figura 5.18.

Figura 5.18. Cuadro de diálogo *Recodificación automática*



**Variable!** **Nuevo nombre.** Las variables cuyos códigos se van a recodificar en enteros consecutivos deben trasladarse a esta lista. Para ello:

- Seleccionar, en la lista de variables del archivo de datos, la variable o variables que se desea recodificar y pulsar el botón flecha para trasladarla(s) a la lista **Variable!** **Nuevo nombre.**

**Nombre nuevo.** El botón **Aceptar** no está disponible hasta que se da un nombre a la variable receptora de los nuevos códigos. Por tanto, es necesario asociar un nombre nuevo al nombre de la variable original. Para ello:

- ' Seleccionar, en la lista **Variable!** **Nuevo nombre**, la variable a la que se quiere asignar nombre e introducir el nombre deseado en el cuadro de texto **Nombre nuevo**.
- ' Pulsar el botón **Añadir nuevo nombre** para validar el nombre elegido (es decir, para asociarlo al nombre original).

El SPSS crea una variable con los nuevos códigos y el nuevo nombre asignado. La variable original queda intacta. Si la variable original posee etiquetas, a la nueva variable y a los nuevos códigos se les asignan las mismas etiquetas. Si los valores de la variable original no poseen etiquetas, los nuevos códigos adoptan como etiquetas los valores de la variable original.

**Recodificar empezando por.** Las opciones de este recuadro permiten cambiar el orden en el que son asignados los nuevos códigos:

**Menor valor.** Asigna los nuevos códigos en orden ascendente: al valor más pequeño de la variable original se le asigna un 1, al valor más pequeño de los restantes se le asigna un 2, etc. Es la opción que se encuentra activa por defecto.

**Mayor valor.** Asigna los nuevos códigos en orden descendente: al valor más grande de la variable original se le asigna un 1, al valor más grande de los restantes se le asigna un 2, etc.

Si la variable es de cadena, los nuevos códigos se asignan de acuerdo con el orden alfabético de las categorías (las mayúsculas preceden a las minúsculas, y los números a las letras). Los valores perdidos se codifican después de los válidos.

- " **Usar el mismo esquema de recodificación para todas las variables.** Esta opción permite asignar enteros consecutivos de 1 a  $m$  tomando todas las variables seleccionadas como si fueran una única variable. Por ejemplo, si se han seleccionado tres variables y orden ascendente, la autorecodificación asigna un 1 al valor más pequeño de las tres variables, un 2 al valor más pequeño de los restantes en las tres variables, etc. Esta opción sólo es válida si todas las variables seleccionadas son del mismo tipo (numérico, cadena, etc.).
- " **Tratar los valores de cadenas en blanco como perdidos por el usuario.** Esta opción permite decidir qué tratamiento se desea dar a los valores vacíos (o espacios en blanco) de las variables de cadena (no tiene efecto sobre las variables numéricas). Si no se marca esta opción, los valores vacíos se consideran valores *válidos* (que es el valor que adoptan por defecto en una variable de cadena) y se recodifican como un valor más. Si se marca esta opción, los valores vacíos se consideran valores perdidos y se les asigna un código de valor perdido definido por el usuario.

**Plantilla.** Las opciones de este recuadro permiten trabajar con plantillas que incluyen esquemas de recodificación previamente definidos:

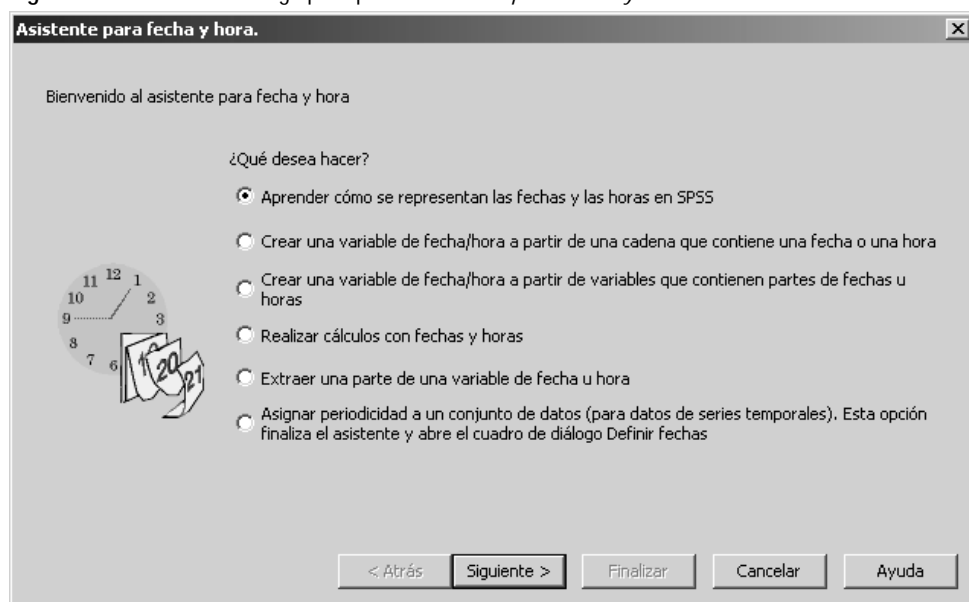
- " **Aplicar plantilla desde.** Permite aplicar a las variables seleccionadas un esquema de recodificación previamente guardado.
- " **Guardar plantilla como.** Guarda en un archivo externo el esquema de recodificación utilizado en el cuadro de diálogo actual. Estas plantillas sólo registran la información referida a los valores no perdidos.

## Operaciones con fechas y horas

El asistente para fechas y horas permite simplificar muchas de las operaciones que es posible llevar a cabo con variables en formato *fecha* y con variables que representan unidades de tiempo (segundos, minutos, horas, días, etc.). Estas operaciones incluyen la suma y resta de fechas o unidades de tiempo, la creación de variables con formato *fecha* a partir de variables con formato *cadena*, la obtención de una fecha a partir de variables que contienen partes de la fecha, etc. Para utilizar este asistente:

- Seleccionar la opción **Fecha/Hora...** del menú **Transformar** para acceder al cuadro de diálogo principal del *Asistente para fechas y horas* que muestra la Figura 5.19.

Figura 5.19. Cuadro de diálogo principal del *Asistente para fechas y horas*



Este primer cuadro de diálogo permite elegir entre varias acciones. La elección apropiada depende del tipo de variables con el que se desea trabajar y de la operación concreta que se desea llevar a cabo. Debe tenerse en cuenta que algunas de las opciones de este cuadro de diálogo o de los siguientes podrían no estar disponibles: esto ocurre cuando en el archivo de datos no existen variables del tipo (numérico, fecha, cadena, etc.) necesario para utilizar esas opciones. Por ejemplo, la opción que permite extraer una fecha de una cadena no estará disponible si el archivo de datos no contiene variables de cadena.

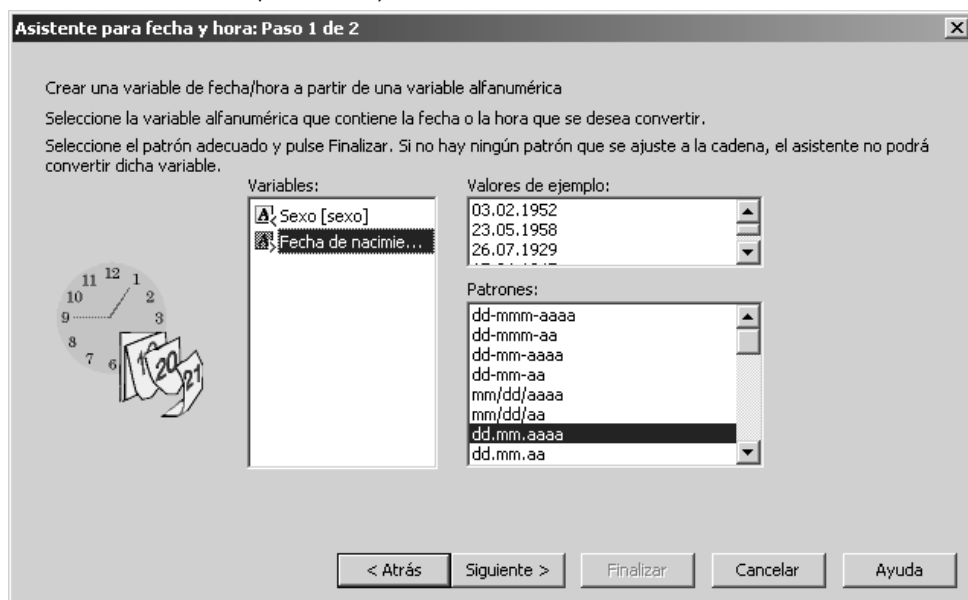
**Aprender cómo se representan las fechas y las horas en SPSS.** Esta primera opción (que aparece marcada por defecto) conduce a un sencillo cuadro de diálogo en el que se ofrece una breve explicación del significado y características de las variables con formato *fecha*. A este respecto, quizá convenga recordar que una variable con formato fecha es, en realidad, una variable numérica, aunque con un aspecto particular; en concreto, el SPSS inter-

preta una fecha como el número de segundos transcurridos desde la medianoche del día 14 de octubre de 1582 (día de implantación del calendario gregoriano) hasta el momento indicado por la fecha.

Crear una variable fecha/hora a partir de una variable de cadena. Esta opción permite transformar una variable con formato *cadena* en una variable numérica con formato *fecha*. Para ello es necesario que la cadena contenga códigos que representen fechas o unidades de tiempo. Para transformar una cadena en una variable numérica con formato fecha:

- Seleccionar la opción **Crear una variable fecha/hora a partir de una variable de cadena** (ver Figura 5.19) para acceder al cuadro de diálogo que muestra la Figura 5.20.

**Figura 5.20.** *Asistente para fechas y horas. Cuadro de diálogo Crear una variable de fecha/hora a partir de una variable de cadena (alfanumérica): Paso 1 de 2*



La lista **Variables** muestra un listado de todas las variables del archivo de datos que tienen formato *cadena*. Al seleccionar una de estas variables, la lista **Valores de ejemplo** ofrece los valores de la variable seleccionada (en el ejemplo propuesto en la Figura 5.20, al archivo *Datos de empleados* se le ha añadido una variable con formato *cadena* con las fechas de nacimiento). Para convertir una cadena de estas características en una variable numérica con formato fecha:

- Seleccionar, en la lista **Variables**, la variable con formato *cadena* que contiene las fechas o unidades de tiempo que se desea transformar.
- Seleccionar de la lista **Patrones** el formato correspondiente a los valores de la cadena, es decir, a los valores presentados en la lista **Valores de ejemplo**. Los valores de la cadena que no se ajusten al patrón seleccionado serán convertidos en valores perdidos.

- Pulsar el botón **Siguiente>** para acceder al siguiente cuadro de diálogo.

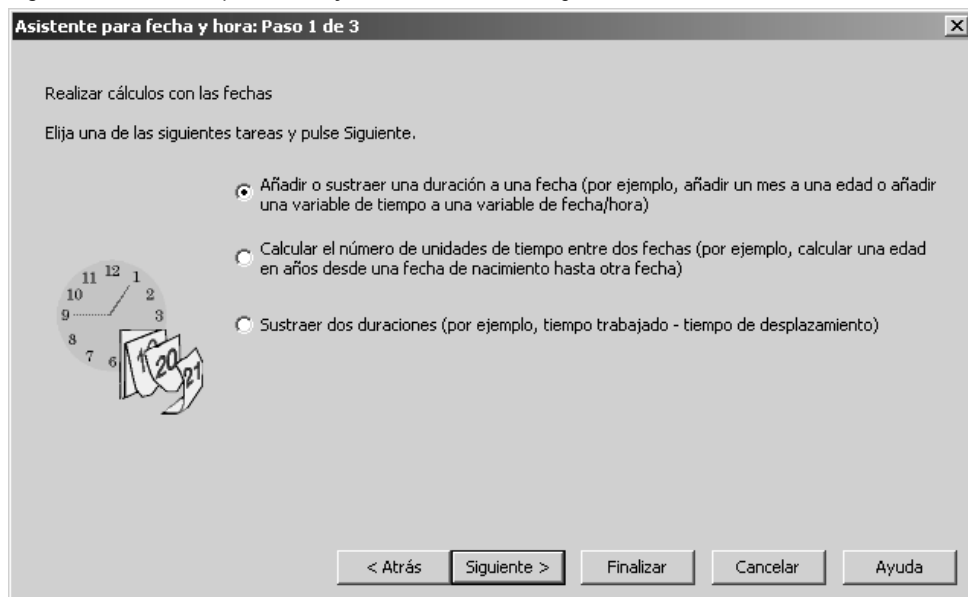
Este segundo y último cuadro de diálogo (no se incluye aquí) contiene las opciones necesarias para dar *nombre* y *etiqueta* a la nueva variable. Si así se desea, se puede cambiar el formato de la fecha original para darle otro aspecto en la nueva variable. Las opciones del cuadro de diálogo también permiten elegir entre crear la variable inmediatamente o pegar la sintaxis correspondiente a las selecciones hechas.

**Crear una variable fecha/hora a partir de variables que contienen partes de la fecha o la hora.** Esta opción conduce a un cuadro de diálogo (no se incluye aquí) que permite crear una variable con formato fecha uniendo varias variables *numéricas* con información parcial sobre una fecha. Por ejemplo, si la variable *año* contiene el año de nacimiento, la variable *mes* contiene el mes de nacimiento y la variable *día* contiene el día de nacimiento (todas ellas variables con formato *numérico*), es posible combinar toda esa información en una única variable con formato *fecha* que incluya el día, mes y año de nacimiento.

**Realizar cálculos con fechas y horas.** Las opciones de este apartado están diseñadas para facilitar las operaciones aritméticas que es posible llevar a cabo con las variables que tienen formato fecha. Para operar (sumar, restar) con este tipo de variables:

- Seleccionar la opción **Realizar cálculos con fechas y horas** (ver Figura 5.19) para acceder al cuadro de diálogo que muestra la Figura 5.21.

**Figura 5.21.** *Asistente para fechas y horas. Cuadro de diálogo Realizar cálculos con las fechas*



**Añadir o sustraer una duración a una fecha.** Permite sumar o restar una cantidad en unidades de tiempo (segundos, minutos, horas, días, etc.) a una variable con formato fecha. Por ejemplo, si se suman 18 años (una constante) a una variable con formato

fecha que contenga las fechas de nacimiento, se obtendrá una nueva variable cuyos valores representarán la fecha en la que cada sujeto ha alcanzado la mayoría de edad.

El botón **Siguiente>** conduce a un cuadro de diálogo (no se incluye aquí) que permite seleccionar: (1) la variable con formato fecha a cuyos valores se desea sumar o restar una cantidad; (2) la variable numérica (o la constante) cuyos valores representan las cantidades que se desea sumar o restar a la variable con formato fecha; y (3) las unidades de tiempo (segundos, minutos, horas, días, etc.) que representan los valores de la variable numérica o el valor de la constante. Un tercer cuadro de diálogo (no se incluye aquí) contiene las opciones necesarias para dar *nombre y etiqueta* a la nueva variable y elegir entre crear la variable inmediatamente o pegar la sintaxis correspondiente a las selecciones hechas.

**Calcular el número de unidades de tiempo entre dos fechas.** Permite restar dos variables con formato fecha. Por ejemplo, si a la fecha actual se le resta la fecha de nacimiento se obtiene la edad.

El botón **Siguiente>** conduce a un cuadro de diálogo (no se incluye aquí) que permite seleccionar: (1) la variable con formato fecha del minuendo; (2) la variable con formato fecha del sustraendo; y (3) las unidades de tiempo (segundos, minutos, horas, días, etc.) en las que se desea expresar el resultado. Un tercer cuadro de diálogo (no se incluye aquí) contiene las opciones necesarias para dar *nombre y etiqueta* a la nueva variable y elegir entre crear la variable inmediatamente o pegar la sintaxis correspondiente a las selecciones hechas.

**Sustraer dos duraciones.** Permite restar dos variables con formato fecha cuyos valores representan duraciones (hh:mm, hh:mm:ss, etc.). Esta opción no estará disponible si el archivo de datos, aun conteniendo variables con formato fecha, ninguna de ellas se ajusta a un patrón de duración.

El botón **Siguiente>** conduce a un cuadro de diálogo (no se incluye aquí) que permite seleccionar: (1) la variable que contiene las duraciones del minuendo; y (2) la variable que contiene las duraciones del sustraendo. Un tercer cuadro de diálogo (no se incluye aquí) contiene las opciones necesarias para dar *nombre y etiqueta* a la nueva variable y para elegir el formato concreto que se desea dar a la duración resultante. También permite elegir entre crear la variable inmediatamente o pegar la sintaxis correspondiente a las selecciones hechas.

**Extraer una parte de una variable de fecha u hora.** Mediante esta opción es posible extraer parte del contenido de una variable con formato fecha. Por ejemplo, de la fecha de nacimiento (dd:mm:aaa) es posible extraer el día de nacimiento, o el mes y el año de nacimiento, o el número de semana correspondiente a la fecha de nacimiento, o el día de la semana, etc.

El botón **Siguiente>** conduce a un cuadro de diálogo (no se incluye aquí) que permite seleccionar: (1) la variable que contiene las fechas sobre las que se desea realizar la extracción; y (2) la parte que se desea extraer de esa variable. Un tercer cuadro de diálogo (no se incluye aquí) contiene las opciones necesarias para dar *nombre y etiqueta* a la nueva variable y elegir entre crear la variable inmediatamente o pegar la sintaxis correspondiente a las selecciones hechas. Este último cuadro de diálogo también permite indicar qué parte de la fecha elegir en el caso de que en el cuadro de diálogo anterior se haya se-

leccionado extraer una porción de fecha u hora: cuando se decide extraer el día, o el mes, o el trimestre, o el año, o el día de la semana, etc., la extracción se realiza automáticamente; pero cuando se decide extraer una porción compuesta como el día y el mes, o el mes y el año, etc., el último cuadro de diálogo permite indicar la porción concreta que se desea extraer.

**Asignar periodicidad a un conjunto de datos.** Esta opción permite generar variables *fecha* cuyos valores progresan a lo largo de los casos con un incremento constante. Puede utilizarse, entre otras cosas, para establecer la periodicidad de una serie temporal o para etiquetar los resultados de un análisis de series temporales.

Al pulsar el botón **Siguiente>** se accede al cuadro de diálogo *Definir fechas* (ver apartado *Definir fechas* del Capítulo 4). Por tanto, esta opción tiene exactamente el mismo efecto que la opción *Definir fechas...* del menú **Datos**.

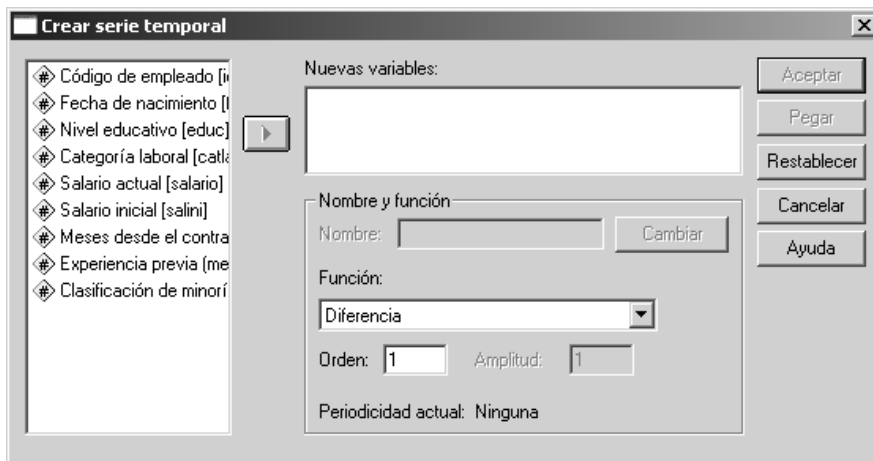
## Crear serie temporal

Esta opción permite crear series temporales nuevas a partir de variables existentes. Una serie temporal es una variable numérica cuyos valores progresan caso a caso a intervalos regulares de tiempo. En una serie temporal, cada caso (cada fila) representa un momento en el tiempo.

Estas variables pueden ser utilizadas más tarde en muchos de los procedimientos de análisis de series temporales (disponibles en el módulo *Tendencias* del SPSS). Para crear una serie temporal:

- Seleccionar la opción **Crear serie temporal...** del menú **Transformar** para acceder al cuadro de diálogo *Crear serie temporal* que muestra la Figura 5.22.

Figura 5.22. Cuadro de diálogo *Crear serie temporal*



La variable o variables del archivo de datos a partir de las cuales se crearán las nuevas series temporales deben trasladarse a la lista **Nuevas variables**. Para ello:



- Seleccionar, en la lista de variables del archivo de datos, la variable o variables a partir de las cuales se van a crear las series temporales (sólo se admiten variables numéricas), y pulsar el botón flecha.

La(s) variable(s) seleccionada(s) pasa(n) a la lista **Nuevas variables** con los seis primeros caracteres del nombre original seguidos del guión de subrayado y un número secuencial (en el ejemplo de la Figura 5.22, la variable *salini* ha pasado con el nombre *salini\_1*).


El nombre de la nueva variable aparece acompañado (con un signo «=») de la función que será utilizada para crear la serie temporal. Tanto el nuevo nombre como la función (DIFF en el ejemplo) los asigna el SPSS por defecto. Las nuevas variables conservan las etiquetas de las variables originales.

## Funciones

Tanto el nombre de la nueva variable como la función asignada por defecto pueden cambiarse utilizando las opciones del recuadro **Nombre y función**. Para cambiar el nombre de la nueva variable:

- Introducir el nombre de la nueva variable en la casilla **Nombre** y pulsar el botón **Cambiar**.

Para cambiar la función:

- Pulsar el botón flecha  del menú desplegable **Función** y seleccionar una función de la lista.

Es posible seleccionar una de las siguientes funciones:

- **Diferencia.** Calcula, para cada caso, la diferencia no estacional entre el valor de ese caso y el valor del caso situado un número determinado de posiciones anteriores. El cuadro de texto **Orden** permite establecer el número de posiciones que se utilizarán para calcular la diferencia. Dado que al comienzo de la serie se pierden tantos valores como el número de posiciones establecidas en el cuadro de texto **Orden**, a los primeros casos se les asignan valores perdidos definidos por el sistema. Por ejemplo, si en el cuadro de texto **Orden** se establece para la diferencia un valor de 3, los primeros 3 casos del archivo tendrán valor perdido en la nueva variable.
- **Diferencia estacional.** Calcula, para cada caso, la diferencia estacional entre el valor de ese caso y el valor del caso situado un número determinado de posiciones anteriores. Ese número de posiciones se basa en la amplitud de un periodo estacional previamente establecido: para calcular diferencias estacionales deben definirse previamente variables *fecha* (ver Capítulo 4, apartado *Definir fechas*) que incluyan un componente periódico (como, por ejemplo, los meses del año, las horas del día, etc.). El cuadro de texto **Orden** recoge el número de periodos estacionales utilizados para calcular la diferencia. El número de casos a los que se asigna un valor perdido al comienzo de la serie es igual a la amplitud del periodo estacional multiplicada por el *orden* establecido. Si, por ejemplo, la periodicidad actual vale 12 (meses del año) y el orden vale 2, se asignará valor perdido a los primeros 24 casos.

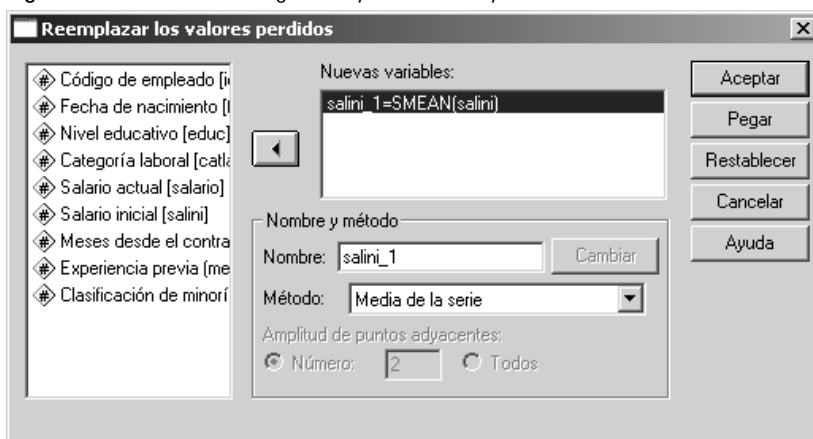
- **Media móvil centrada.** Calcula, para cada caso, la media de los valores de los casos que rodean a ese caso (incluido el propio caso). La opción **Amplitud** permite establecer el número de casos que serán utilizados para calcular la media. Si la *amplitud* es par, la media móvil se calcula como el promedio de las dos medias no centradas que corresponden a cada caso. El número de casos a los que se asigna valor perdido al comienzo y al final de la serie es igual a  $k/2$  ( $k = \text{amplitud}$ ) si la amplitud es par, e igual a  $(k-1)/2$  si la amplitud es impar. Si, por ejemplo, la amplitud vale 5, el número de casos con valor perdido al comienzo y al final de la serie es  $(5-1)/2 = 2$ .
- **Media móvil anterior.** Calcula, para cada caso, la media de los valores de los casos que preceden a ese caso. El cuadro de texto **Amplitud** permite establecer el número de valores que serán utilizados para calcular la media. El número de casos a los que se asigna valor perdido al comienzo de la serie es igual al valor de la *amplitud*.
- **Medianas móviles.** Calcula, para cada caso, la mediana de los valores de los casos que rodean a ese caso (incluido el propio caso). El cuadro de texto **Amplitud** permite establecer el número de casos que serán utilizados para calcular la mediana. Si la *amplitud* es par, la mediana se calcula como el promedio de las dos medianas no centradas que corresponden a cada caso. El número de casos a los que se asigna valor perdido al comienzo y al final de la serie es igual a  $k/2$  ( $k = \text{amplitud}$ ) si la amplitud es par, e igual a  $(k-1)/2$  si la amplitud es impar. Si, por ejemplo, la amplitud vale 5, el número de casos con valor perdido al comienzo y al final de la serie es  $(5-1)/2 = 2$ .
- **Suma acumulada.** Calcula, para cada caso, la suma del valor de ese caso y de todos los que le preceden en la serie.
- **Retardo.** Asigna a cada caso el valor del caso situado un determinado número de posiciones por delante de él. El cuadro de texto **Orden** permite establecer el número de posiciones en que se basará la asignación. El número de casos a los que se asigna valor perdido al comienzo de la serie es igual al *orden* establecido.
- **Adelanto.** Asigna a cada caso el valor del caso situado un determinado número de posiciones por detrás de él. El cuadro de texto **Orden** permite establecer el número de posiciones en que se basará la asignación. El número de casos a los que se asigna valor perdido al comienzo de la serie es igual al *orden* establecido.
- **Suavizado.** Calcula, para cada caso, un nuevo valor basado en un largo proceso de suavizado. Este proceso de suavizado comienza con una mediana móvil de amplitud 4, que se centra por una mediana móvil de amplitud 2. Los valores resultantes se vuelven a suavizar aplicando una mediana móvil de amplitud 5 y una mediana móvil de amplitud 3, y obteniendo el promedio ponderado de ambas medianas móviles. A continuación se calculan las diferencias (residuos) entre los valores de la serie suavizada y los de la serie original. Después se repite todo el proceso sobre los residuos obtenidos para suavizarlos. Por último, se obtienen las diferencias entre los residuos suavizados y los valores suavizados obtenidos en el primer paso del proceso. A este procedimiento de suavizado se le suele denominar T4253H.

## Reemplazar valores perdidos

Los valores perdidos casi siempre constituyen una fuente de problemas en la mayor parte de los procedimientos estadísticos (y, muy especialmente, en algunos como el análisis de series temporales). Lo ideal para el análisis y para el analista es que no existan valores perdidos. No obstante, si existen, los problemas que se derivan de su tratamiento pueden resolverse, en parte, sustituyéndolos por alguna estimación del valor que podrían haber adoptado (aunque este tipo de sustituciones siempre deben realizarse con la máxima cautela). Para reemplazar valores perdidos:

- Seleccionar la opción **Reemplazar valores perdidos...** del menú **Transformar** para acceder al cuadro de diálogo *Reemplazar valores perdidos* que muestra la Figura 5.23.

Figura 5.23. Cuadro de diálogo *Reemplazar valores perdidos*



El cuadro **Nuevas variables** muestra el nombre de la nueva variable junto con el método de estimación que se utilizará para reemplazar los valores perdidos. Para trasladar una variable a esta lista:

- Seleccionar la variable en la lista de variables del archivo de datos y pulsar el botón flecha (o pulsar dos veces el botón principal del ratón). En el ejemplo de la Figura 5.23 se ha seleccionado y trasladado a la lista **Nuevas variables** la variable *salini* (salario inicial).

Al seleccionar una variable y trasladarla a la lista **Nuevas variables**, el SPSS le asigna automáticamente un nombre nuevo formado por los primeros 6 caracteres de la variable original, el guión de subrayado y un número secuencial. En el ejemplo de la Figura 5.23, a la variable *salini* se le ha asignado el nombre *salini\_1*. Las nuevas variables conservan las etiquetas de las variables originales.

El nombre de la nueva variable aparece acompañado (con un signo «=») del método que será utilizado para reemplazar los valores perdidos. En el ejemplo, la variable *salini\_1* aparece acompañada del método MEAN (media de puntos adyacentes). Tanto el nuevo nombre como el método de estimación los asigna el SPSS por defecto.

## Métodos de estimación

Las opciones de este recuadro permiten cambiar el nombre asignado automáticamente a la nueva variable y el método de estimación utilizado por defecto. Para cambiar el nombre:

- Introducir el nombre deseado en el cuadro de texto **Nombre** y pulsar el botón **Cambiar**.

Para cambiar el método de estimación:

- Pulsar el botón de menú desplegable de la opción **Método** y seleccionar cualquiera de las opciones listadas:
  - **Media de la serie.** Sustituye los valores perdidos por la media aritmética de la serie completa (es decir, por la media aritmética de la variable).
  - **Media de puntos adyacentes.** Sustituye cada valor perdido por la media aritmética de sus valores adyacentes válidos.
  - **Mediana de puntos adyacentes.** Sustituye cada valor perdido por la mediana de sus valores adyacentes válidos.
  - **Interpolación lineal.** Sustituye los valores perdidos utilizando una interpolación lineal basada en el último valor válido anterior al valor perdido y en el primer valor válido posterior al valor perdido. El primer y el último valor de la serie no son reemplazados (en caso de que sean valores perdidos).
  - **Tendencia lineal en el punto.** Sustituye cada valor perdido por los pronósticos obtenidos al llevar a cabo un análisis de regresión lineal utilizando la propia serie como variable dependiente y una variable índice escalada de 1 a  $n$  como variable independiente.

**Amplitud de puntos adyacentes.** En los métodos *Media de puntos adyacentes* y *Mediana de puntos adyacentes* es posible establecer el número de puntos adyacentes con los que se desea calcular la media o la mediana:

**Número.** Número de valores válidos, por encima y por debajo del valor perdido, que serán utilizados para el cálculo de la media o la mediana. El número de valores por defecto es 2 (es decir, 2 valores por encima y 2 por debajo). Si la amplitud solicitada es mayor que el número de valores válidos que hay por debajo o por encima de un valor perdido concreto, ese valor perdido no es reemplazado.

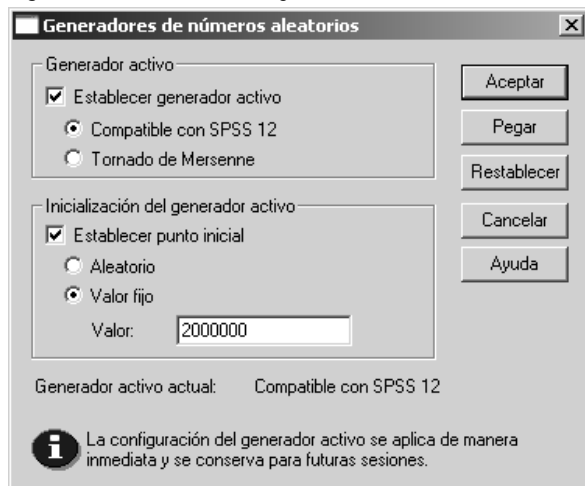
**Todos.** Sustituye los valores perdidos por la media o la mediana de la serie completa. Equivale a seleccionar la opción **Media de la serie**.

## Generadores de números aleatorios

Las funciones del SPSS que utilizan números aleatorios (como, por ejemplo, la función de probabilidad *RV.NORMAL* o la selección aleatoria de una muestra del archivo de datos) se basan en un *generador de números aleatorios*. El SPSS incorpora dos de estos generadores. La opción **Generadores de números aleatorios** permite elegir entre ambos generadores y establecer la semilla de aleatorización. Para ello:

Seleccionar la opción **Generadores de números aleatorios...** del menú **Transformar** para acceder al cuadro de diálogo *Generadores de números aleatorios* que muestra la Figura 5.24.

Figura 5.24. Cuadro de diálogo *Generadores de números aleatorios*



**Generador activo.** El SPSS incluye dos generadores de números aleatorios. Las opciones de este recuadro permiten elegir con cuál de ellos se desea trabajar:

**Compatible con SPSS 12.** Generador original (utilizado en las versiones 12 y anteriores). Es el generador que actúa por defecto. Para reproducir series aleatorias generadas con versiones del SPSS anteriores a la 13, es necesario utilizar este generador.

**Tornado de Mersenne.** Se trata de un nuevo generador que está optimizado para llevar a cabo trabajos de simulación.

**Inicialización del generador activo.** Un generador de números aleatorios siempre comienza a generar una serie aleatoria a partir de un valor inicial llamado *semilla*. Esta semilla, por defecto, cambia aleatoriamente cada vez que se solicita generar una serie aleatoria durante la misma sesión. Esto significa que las distintas series aleatorias generadas durante la misma sesión no serán siempre las mismas (justamente por ser aleatorias). No obstante, existe la posibilidad de replicar una serie aleatoria si se fuerza al generador de números aleatorios a comenzar con la misma semilla.

“ **Establecer punto inicial.** Con esta opción desactivada, la semilla o punto inicial se fija aleatoriamente cada vez que se genera una serie:

**Aleatorio.** Marcando esta opción, el *Visor de resultados* muestra la semilla aleatoria que será utilizada. De este modo es posible conocer la semilla que se está utilizando. Cada vez que se marca esta opción y se pulsa el botón **Aceptar**, cambia la semilla.

**Valor fijo.** Esta opción permite seleccionar como semilla un entero positivo entre 1 y 2.000.000.000. Siempre que se utiliza la misma semilla, se obtiene la misma serie de números aleatorios.

## Ejecutar transformaciones pendientes

Cuando se intenta efectuar algún tipo de transformación mediante los cuadros de diálogo del SPSS (*Calcular*, *Recodificar*, *Contar apariciones*, etc.), el botón **Aceptar** ejecuta de forma inmediata la transformación solicitada. Sin embargo, cuando se lleva a cabo alguna transformación utilizando el *Editor de sintaxis*, es necesario que la sentencia EXECUTE sea la última del conjunto de sentencias ejecutadas. Si no se incluye esta sentencia, las transformaciones incluidas en las sentencias ejecutadas quedan pendientes. La opción **Ejecutar transformaciones pendientes** del menú **Transformar** hace que las transformaciones pendientes sean efectivamente ejecutadas.



## Modificar archivos de datos

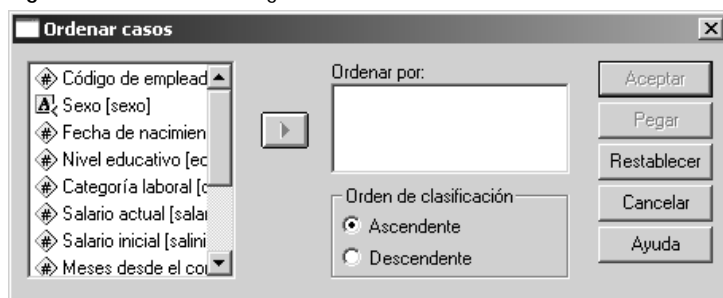
Los archivos de datos no siempre están organizados de forma idónea. En ocasiones puede interesar cambiar el orden de los casos, o transponer las filas y las columnas, o mezclar en uno varios archivos diferentes, o seleccionar sólo unos pocos casos para el análisis, etc. En este capítulo se describen varios procedimientos relacionados con el archivo de datos, todos los cuales se encuentran en el menú **Datos**.

### Ordenar casos

La opción **Ordenar casos** permite cambiar el orden de los casos (es decir, el orden de las filas del *Editor de datos*) utilizando como criterio una o más variables. Es una opción necesaria para preparar la fusión de archivos (ver más adelante, en este mismo capítulo, el apartado *Fun-dir archivos*) y puede resultar útil, por ejemplo, si se desea imprimir el archivo de datos con los casos ordenados siguiendo algún criterio de interés. Para ordenar casos:

- Seleccionar la opción **Ordenar casos...** del menú **Datos** para acceder al cuadro de diálogo *Ordenar casos* que muestra la Figura 6.1.

Figura 6.1. Cuadro de diálogo *Ordenar casos*



**Ordenar por.** Los casos del archivo se ordenan utilizando como criterio la(s) variable(s) trasladada(s) a esta lista desde la lista de variables. Si se utiliza más de una variable de ordenación, el orden resultante viene determinado por la secuencia de las variables seleccionadas: los casos se ordenan por los valores de la primera variable; a continuación, los casos empatados en esa primera variable, si existen, se ordenan por los valores de la segunda; si todavía quedan casos empatados, se ordenan por los valores de la tercera variable; etc.



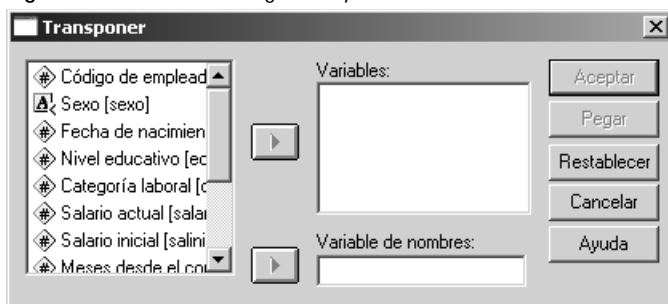
**Orden de clasificación.** Si se utiliza como criterio de ordenación una variable numérica, la opción **Ascendente** ordena los casos desde el valor menor al mayor. Si el criterio de ordenación es una variable de cadena, los casos se ordenan alfabéticamente (las mayúsculas preceden a las minúsculas; los números preceden a las letras; el resto de caracteres –asterisco, admiración, interrogación, etc.– preceden a los números). La opción **Descendente** realiza la ordenación de forma inversa.

## Transponer archivos

En los archivos de datos SPSS se asume que las filas representan *casos* y las columnas *variables*. Esto es así también en la mayor parte de las bases de datos y hojas de cálculo. Sin embargo, en ocasiones puede ocurrir, por ejemplo, que surja la necesidad de importar datos de alguna fuente externa en la que las filas representen variables y las columnas casos; también puede ocurrir que sea necesario reordenar el archivo para efectuar algunos cálculos o para aplicar algunos procedimientos. En estos casos, la opción **Transponer** permite crear un nuevo archivo de datos con las filas convertidas en columnas y las columnas en filas. Para transponer el archivo de datos:

- Seleccionar la opción **Transponer...** del menú **Datos** para acceder al cuadro de diálogo *Transponer* que muestra la Figura 6.2.

Figura 6.2. Cuadro de diálogo *Transponer*



**Variables.** En el archivo transpuesto, cada caso del archivo original (todos) pasa a ser una variable. Pero sólo las variables trasladadas a esta lista pasan a formar parte del nuevo archivo transpuesto. En concreto, cada variable seleccionada en esta lista pasa a ser un caso del archivo transpuesto. El procedimiento crea una variable adicional (con formato de cadena y nombre *case\_lbl*) cuyos valores son los nombres de las variables del archivo original.

**Variable de nombres.** Por defecto, las nuevas variables (una por cada caso) son nombradas *var001*, *var002*, etc. Pero, opcionalmente, los nuevos nombres pueden ser los valores de alguna de las variables del archivo original. Para ello, la variable cuyos valores serán los nombres de las nuevas variables debe desplazarse a este cuadro. Si la variable propuesta es numérica, los nuevos nombres comienzan con la letra *V*. Si los valores de la variable propuesta no son *únicos*, a cada valor repetido se le asigna un número secuencial al final. El *Visor de resultados* ofrece un listado con los nombres asignados a las nuevas variables.

# Reestructurar el archivo de datos

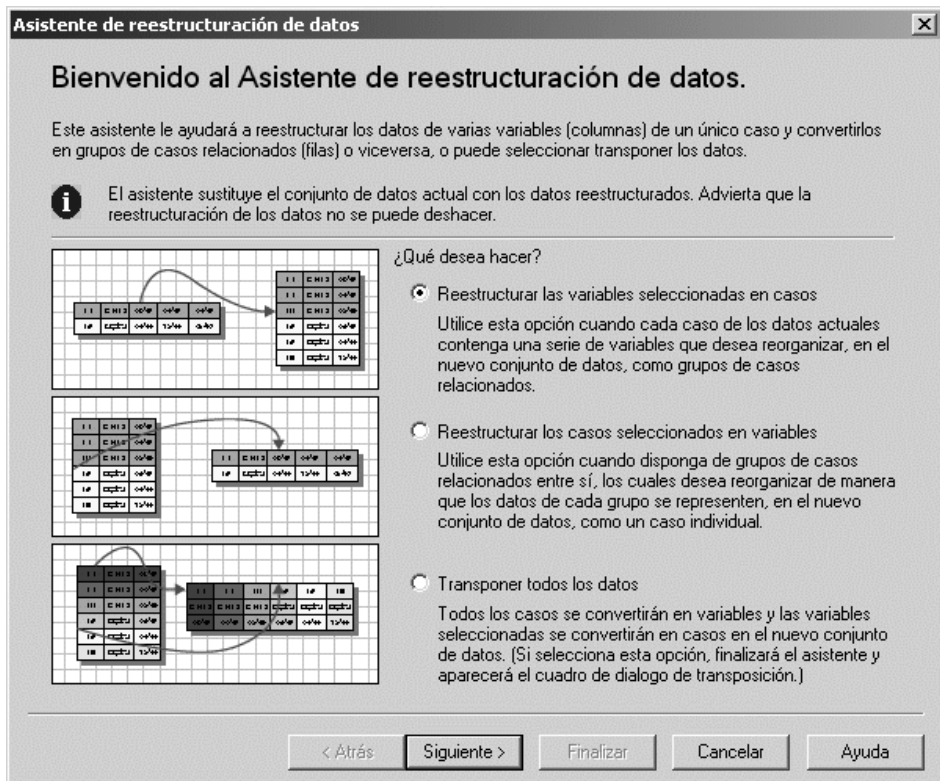
La opción Reestructurar del menú Datos permite reordenar total o parcialmente el archivo de datos a través del *Asistente de reestructuración de datos*. Este *Asistente* ofrece la posibilidad de convertir variables en casos y casos en variables.

En una estructura de datos *simple*, cada fila representa un caso y cada columna una variable. Pero no todos los archivos se ajustan a una estructura de datos *simple*. En ocasiones, una variable puede estar codificada en más de una columna y un mismo caso en más de una fila. Esta estructura de datos *compleja* se da, por ejemplo, cuando se mide la misma variable a lo largo del tiempo a los mismos sujetos; estas medidas en el tiempo pueden codificarse como varias variables (*grupo de variables*) o como varias filas (*grupo de casos*).

Algunos procedimientos SPSS (como la opción Medidas repetidas del procedimiento Modelo lineal general) requieren que los datos estén codificados como *grupos de variables*. Otros procedimientos (como Modelos mixtos lineales) requieren que los datos estén codificados como *grupos de casos*. Cuando los datos no están dispuestos tal como requiere el procedimiento que se desea utilizar, es necesario reestructurarlos. Para ello:

- Seleccionar la opción Reestructurar... del menú Datos para acceder al cuadro de diálogo *Asistente para la reestructuración de datos (paso 1)* que muestra la Figura 6.3.

Figura 6.3. Cuadro de diálogo del *Asistente para la reestructuración de datos (paso 1)*



**Reestructurar las variables seleccionadas en casos.** Reestructura el archivo de datos convirtiendo variables en casos. Con este tipo de reestructuración, cada caso del archivo de datos original pasa a ocupar más de una fila en el archivo reestructurado.

**Reestructurar los casos seleccionados en variables.** Reestructura el archivo de datos convirtiendo casos en variables. Con este tipo de reestructuración, el número de casos del archivo reestructurado es menor que el del archivo original.

**Transponer todos los datos.** Si se desea transponer el archivo de datos completo, el *Asistente para la reestructuración de datos* abre el cuadro de diálogo *Transponer* descrito en el apartado anterior (ver Figura 6.2).

## Convertir variables en casos

La opción **Reestructurar las variables seleccionadas en casos** del paso 1 (ver Figura 6.3) permite convertir variables en casos. El botón **Siguiente** de ese cuadro de diálogo conduce al **Paso 2** del *Asistente para la reestructuración de datos*, el cual permite indicar cuántos *grupos de variables* se desean convertir en casos (no se incluye aquí este cuadro de diálogo). El procedimiento permite elegir entre **Uno** o **Más de uno**. Los datos de las Figuras 6.4 y 6.5 pueden ayudar a entender el significado de esta elección.

La Figura 6.4 ofrece un ejemplo en el que las variables  $x_1$  y  $x_2$  constituyen el *grupo de variables* que se desea convertir en casos. Al convertir estas *dos* variables en casos, cada caso del archivo original pasa a ocupar *dos* filas del archivo reestructurado. Si el grupo de variables elegido estuviera formado por tres, cuatro, etc., variables, cada caso del archivo original pasaría a ocupar tres, cuatro, etc., filas en el archivo reestructurado. En el ejemplo de la Figura 6.4, la variable  $x$  recoge los valores de las dos variables reestructuradas; y la variable  $id$  permite identificar cada caso antes y después de la reestructuración.

**Figura 6.4.** Reestructurar convirtiendo variables en casos: un solo grupo de variables

	id	x1	x2
1	1	1	2
2	2	3	4
3	3	5	6

→

	id	x
1	1	1
2	1	2
3	2	3
4	2	4
5	3	5
6	3	6

La Figura 6.5 muestra un ejemplo en el que se han reestructurado *dos grupos de variables*: las variables  $x_1$  y  $x_2$  constituyen el primer grupo de variables; las variables  $y_1$  y  $y_2$ , el segundo grupo (lógicamente, todos los grupos de variables que se desee reestructurar deben tener el mismo número de variables, pues cada grupo de variables quedará referido al mismo grupo de casos). Al reestructurar dos grupos de variables, cada caso del archivo original pasa a ocupar, en el archivo reestructurado, tantas filas como número de variables tiene cada grupo. En el ejemplo de la Figura 6.5, a las nuevas variables (una por cada grupo de variables del archivo original), se les ha asignado los nombres  $x$  e  $y$ . La variable  $id$  permite identificar cada caso antes y después de la reestructuración.

Figura 6.5. Reestructurar convirtiendo variables en casos: dos grupos de variables

	id	x1	x2	y1	y2
1	1	1	2	10	11
2	2	3	4	12	13
3	3	5	6	14	15

→

	id	x	y
1	1	1	10
2	1	2	11
3	2	3	12
4	2	4	13
5	3	5	14
6	3	6	15

Pulsando el botón **Siguiente** del cuadro de diálogo correspondiente al paso 2 se accede al **Paso 3** del *Asistente para la reestructuración de datos*. La Figura 6.6 muestra el aspecto que adopta el *Asistente* en el paso 3.

 Figura 6.6. Cuadro de diálogo del *Asistente para la reestructuración de datos: variables a casos* (paso 3)

**Asistente de reestructuración de datos: Paso 3 de 7**

### Variables a casos: Seleccionar variables

Por cada grupo de variables contenidas en los datos actuales, el archivo reestructurado contendrá una variable de destino.

En este paso, seleccione el modo de identificar grupos de casos en los datos reestructurados y seleccione las variables que pertenecen a cada variable de destino.

De forma opcional, también puede seleccionar variables que desee copiar en el nuevo archivo como variables fijas.

Variables del archivo actual:

- # id
- # x1
- # x2
- # y1
- # y2

Identificación de grupos de casos:

Utilizar número del caso

Nombre:

Variables que se van a transponer:

Variable de destino:

▲

▼

←

- # x1
- # x2

Variables fijas:

Las opciones de este cuadro de diálogo permiten seleccionar las variables que se desea reestructurar y qué variable se desea utilizar para identificar los casos del nuevo archivo reestructurado:

**Variables del archivo actual.** Ofrece un listado con todas las variables incluidas en el archivo de datos.

**Identificación de grupos de casos.** Al convertir variables en casos, el nuevo archivo reestructurado pasa a tener más casos que el archivo original. Esto es debido a que cada caso del archivo original está duplicado, triplicado, etc., en el archivo reestructurado. La opción **Identificación de grupos de casos** permite decidir qué variable se desea utilizar para identificar en el archivo reestructurado cada caso del archivo original. El menú desplegable de este cuadro permite elegir entre:

- **Utilizar número de caso.** Crea una variable cuyo valor para cada caso del archivo reestructurado es el número de fila (número de caso) que ocupa en el archivo original. El procedimiento asigna a esta variable el nombre *id* (si esta variable ya existe en el archivo original, le asigna el nombre *id1*), pero puede elegirse cualquier nombre válido utilizando el teclado. El botón **Etiqueta** abre un pequeño cuadro de diálogo que permite introducir una etiqueta descriptiva asociada al nombre de la variable.
- **Utilizar variable seleccionada.** Permite seleccionar como variable de identificación cualquiera de las variables del archivo de datos. Esta opción es útil cuando el archivo incluye una variable con códigos de identificación especiales para cada sujeto (DNI, teléfono, número de la seguridad social, número de cliente, etc.). Si la variable de identificación de casos del archivo original coincide con el número de caso (número de fila), tal como ocurre en los ejemplos de las Figuras 6.4 y 6.5, la opción **Utilizar variable seleccionada** tiene el mismo efecto que la opción **Utilizar número de caso**.
- **Ninguno.** No se crea ninguna variable de identificación de grupos de casos.

**Variables que se van a transponer.** Las opciones de este recuadro permiten llevar a cabo dos acciones: (1) seleccionar las variables que se desea transponer y (2) asignar nombre a las nuevas variables del archivo reestructurado. Las **Variables de destino** son las variables del archivo reestructurado que reciben los valores de las variables del archivo original. Es necesario definir tantas variables de destino como *grupos de variables* se haya elegido reestructurar en el paso anterior. El botón de menú desplegable permite mostrar y elegir cada una de las variables de destino. A la primera variable de destino se le asigna el nombre *trans1*; esta primera variable de destino es la que recibirá los valores del primer grupo de variables (si se ha optado por reestructurar un solo grupo de variables, únicamente habrá que definir una variable de destino). A la segunda variable de destino se le asigna el nombre *trans2*; esta segunda variable de destino es la que recibirá los valores del segundo grupo de variables. Etc. Los nombres que el procedimiento asigna por defecto a las variables de destino pueden cambiarse introduciendo cualquier otro nombre válido. Para definir variables de destino:

- En el menú desplegable **Variable de destino**, seleccionar la variable de destino que se quiere definir y, si se desea, cambiarle el nombre.
- Seleccionar de la lista **Variables del archivo actual** el primer *grupo de variables* que se desea transponer y trasladarlas (arrastrándolas o utilizando el botón flecha) a la lista **Variables que se van a transponer**. Estas variables pueden ser de cualquier tipo, pero todas deben ser del mismo (numéricas o de cadena). Una misma variable puede estar repetida en el mismo grupo de variables, pero una variable de un grupo no puede estar repetida en un grupo distinto.
- Repetir las dos acciones anteriores para cada variable de destino que se desee definir, es decir, para cada grupo de variables que se desee transponer. Debe tenerse en cuen-

ta que todos los grupos de variables deben tener el mismo número de variables y que es necesario definir tantas variables de destino como grupos de variables se hayan seleccionado en el primer paso.

**Variables fijas.** Con las variables que no se van transponer (las variables no seleccionadas en el paso 3 del *Asistente*) es posible hacer dos cosas: eliminarlas o conservarlas. Si se desea eliminar o conservar *todas* las variables no seleccionadas, en este cuadro de diálogo no hay que hacer nada (más adelante, en el paso 6 del *Asistente*, se ofrecen opciones para decidir qué hacer con *todas* las variables no seleccionadas). Si se desea conservar sólo *algunas* de las variables no seleccionadas, esas variables deben ser trasladadas a la lista **Variables fijas**. En el archivo reestructurado, las variables fijas duplican, triplican, etc., sus valores para los casos del nuevo archivo que se corresponde con el mismo caso del archivo original.

Los siguientes dos pasos del *Asistente para la reestructuración de datos*, **Paso 4** y **Paso 5**, permiten crear, si así se desea, una o más **variables índice** (no se incluyen aquí estos dos cuadros de diálogo). Una variable *índice* es una variable con valores secuenciales que permiten identificar las filas del archivo reestructurado a partir de las filas y columnas del archivo original (son especialmente útiles para identificar los distintos niveles de las variables *factor*, es decir, de las variables que definen grupos de puntuaciones).

La Tabla 6.1 y las Figuras 6.7 y 6.8 ofrecen un ejemplo que puede ayudar a entender en qué consiste una variable *índice*. La Tabla 6.1 ofrece las puntuaciones obtenidas por una muestra aleatoria de 6 sujetos a los que se les ha hecho memorizar dos listas distintas: una de *letras* y otra de *números*; más tarde, al cabo de una *hora*, de un *día*, de una *semana* y de un *mes*, se les ha pedido que intenten repetir ambas listas. Las puntuaciones reflejan las valoraciones obtenidas por cada sujeto tras ser evaluados por un grupo de expertos.

**Tabla 6.1.** Datos de un diseño con dos factores: *tiempo* y *contenido*

<i>Sujetos</i>	<i>Hora</i>		<i>Día</i>		<i>Semana</i>		<i>Mes</i>	
	<i>Números</i>	<i>Letras</i>	<i>Números</i>	<i>Letras</i>	<i>Números</i>	<i>Letras</i>	<i>Números</i>	<i>Letras</i>
1	6	8	6	6	3	4	2	3
2	7	10	5	8	5	5	5	2
3	4	7	2	7	1	2	3	2
4	7	11	5	9	3	3	4	6
5	6	10	4	6	4	4	5	3
6	5	9	2	4	1	3	1	5

Los datos de la Tabla 6.1 pueden introducirse en el *Editor de datos* con la misma disposición que tienen en la tabla: 6 casos y 8 variables. La Figura 6.7. muestra esta forma convencional de introducir los datos. Esta es la disposición típica para utilizar la mayoría de los procedimientos SPSS (por ejemplo, la opción **Medidas repetidas** del procedimiento **Modelo lineal general**), pero no es apropiada para todos ellos (no es apropiada, por ejemplo, para el procedimiento **Modelos mixtos lineales**). Para poder utilizar estos otros procedimientos no es suficiente con reestructurar el archivo de datos convirtiendo las variables en casos; es necesario, además, crear variables *índice* que permitan identificar correctamente cada nueva fila del archivo reestructurado.

**Figura 6.7.** Datos de la Tabla 6.1 con la disposición convencional que suelen adoptar en el *Editor de datos*

	id	hora_n	hora_l	día_n	día_l	semana_n	semana_l	mes_n	mes_l
1	1	6	8	6	6	3	4	2	3
2	2	7	10	5	8	5	5	5	2
3	3	4	7	2	7	1	2	3	2
4	4	7	11	5	9	3	3	4	6
5	5	6	10	4	6	4	4	5	3
6	6	5	9	2	4	1	3	1	5

Para crear estas variables *índice* hay que tener en cuenta que las 8 variables de la Tabla 6.1 no son más que los 8 *niveles* resultantes de combinar los 4 niveles de la variable *tiempo* (hora, día, semana y mes) con los 2 niveles de la variable *contenido* (números y letras). Esto significa que si las 8 variables se convierten en casos sin utilizar una variable *índice* se perderá la información referida a las variables *tiempo* y *contenido*.

La Figura 6.8 (parte izquierda) muestra las variables de la Figura 6.7 convertidas en casos *sin utilizar variables índice*. La Figura 6.8 (parte derecha) muestra las variables de la Figura 6.7 convertidas en casos utilizando *dos variables índice: tiempo y contenido*. Por supuesto, el archivo reestructurado contiene 48 casos (los 6 casos del archivo original multiplicados por las 8 variables). La Figura 6.8. únicamente reproduce los 3 primeros casos, los cuales ocupan las primeras 24 filas. Las puntuaciones de los sujetos se encuentran en la columna nombrada *destino*.

**Figura 6.8.** Datos de la Figura 6.7 reestructurados *sin variables índice* (izquierda) y *con dos variables índice* (derecha)

	id	destino
1	1	6
2	1	8
3	1	6
4	1	6
5	1	3
6	1	4
7	1	2
8	1	3
9	2	7
10	2	10
11	2	5
12	2	8
13	2	5
14	2	5
15	2	5
16	2	2
17	3	4
18	3	7
19	3	2
20	3	7
21	3	1
22	3	2
23	3	3
24	3	2

	id	tiempo	contenido	destino
1	1	1	1	6
2	1	1	2	8
3	1	2	1	6
4	1	2	2	6
5	1	3	1	3
6	1	3	2	4
7	1	4	1	2
8	1	4	2	3
9	2	1	1	7
10	2	1	2	10
11	2	2	1	5
12	2	2	2	8
13	2	3	1	5
14	2	3	2	5
15	2	4	1	5
16	2	4	2	2
17	3	1	1	4
18	3	1	2	7
19	3	2	1	2
20	3	2	2	7
21	3	3	1	1
22	3	3	2	2
23	3	4	1	3
24	3	4	2	2

Para crear estas dos variables *índice* (*tiempo* y *contenido*), en el paso 4 del *Asistente para la reestructuración de datos* debe indicarse que se desea crear 2 variables *índice* (el procedimiento permite, mediante botones de selección, elegir entre *Una*, *Varias* –hay que decir cuántas– o *Ninguna*).

También hay que elegir un nombre (obligatorio) y una etiqueta descriptiva (opcional) para las variables *índice*. El procedimiento asigna, por defecto, los nombres *índice1* e *índice2*, pero estos nombres pueden cambiarse introduciendo cualquier nombre válido (en el ejemplo se han utilizado los nombres *tiempo* y *contenido*). En el caso de que se vaya a crear una sola variable *índice*, en el paso 5 también debe decidirse si se desea asignar, a la nueva variable *índice*, números secuenciales (es lo habitual) o los valores de alguna variable existente en el archivo de datos.

En el paso 5 del *Asistente* también es necesario especificar cuántos niveles tiene cada variable *índice*. En el ejemplo, a la variable *tiempo* (primera variable *índice*) se le han asignado 4 niveles y a la variable *contenido* (segunda variable *índice*) se le han asignado 2 niveles. El resultado puede apreciarse en la parte derecha de la Figura 6.8: los 2 niveles de la segunda variable *índice* se combinan secuencialmente con los 4 niveles de la primera variable *índice*. Por supuesto, el producto de niveles asignados debe ser igual al del número de variables dentro del grupo (8 en el ejemplo); si esto no es así, el *Asistente* ofrece una señal de advertencia indicando tal circunstancia y no deja continuar.

El cuadro de diálogo correspondiente al **Paso 6** del *Asistente para la reestructuración de datos* (no se incluye aquí este cuadro de diálogo) ofrece algunas opciones para decidir, entre otras cosas, qué se desea hacer con las variables no seleccionadas y con los valores perdidos.

Respecto al **Tratamiento de las variables no seleccionadas**, puede optarse entre: (1) *desecharlas*, o (2) *mantenerlas tratándolas como variables fijas*. Si se desea desechar algunas variables y mantener otras, esta decisión debe tomarse en el paso 3 del *Asistente* (ver Figura 6.6, recuadro **Variables fijas**).

Respecto a si uno o más casos tienen **Valores perdidos definidos por el sistema en todas las variables transpuestas**, puede optarse por **Crear un caso nuevo** con esos valores perdidos o por **Desechar los datos**.

Por último, entre las opciones del paso 6 está la de crear una variable que refleje el número de nuevos casos que corresponden a cada caso del archivo original. Esta variable tomará el mismo valor en todos los casos del archivo reestructurado: el número de variables de cada *grupo de variables* previamente definido. El cuadro de diálogo permite asignar un nombre (obligatorio) y una etiqueta descriptiva (optativa) a esta nueva variable.

En el cuadro de diálogo correspondiente al **Paso 7** del *Asistente para la reestructuración de datos* (no se incluye aquí este cuadro de diálogo) es necesario decidir entre **Reestructurar los datos ahora** o **Pegar la sintaxis generada por el Asistente en una ventana de sintaxis**. Si se elige la primera opción, el botón **Finalizar** inicia la reestructuración del archivo y el archivo original es sustituido en el *Editor de datos* por el archivo reestructurado. Si se elige la segunda opción, el botón **Finalizar** abre una ventana de sintaxis (si no existe ninguna abierta) y pega en ella la sintaxis SPSS correspondiente a todas las elecciones hechas. Conviene señalar que, en el caso de que se opte por pegar la sintaxis, el botón **Finalizar** no inicia la reestructuración del archivo original; para iniciar la reestructuración es necesario ejecutar la sintaxis desde el *Editor de sintaxis* (puede consultarse el Capítulo 8 para una aclaración de cómo se trabaja con archivos de sintaxis).



## Convertir casos en variables

La opción **Reestructurar los casos seleccionados en variables** del paso 1 del *Asistente para la reestructuración de datos* (ver Figura 6.3) permite convertir casos en variables. El botón **Siguiente** de ese cuadro de diálogo conduce al **Paso 2 del Asistente**, el cual adopta el aspecto que muestra la Figura 6.9.

**Figura 6.9.** Cuadro de diálogo del *Asistente para la reestructuración de datos: casos en variables (paso 2)*



Las opciones de este cuadro de diálogo permiten seleccionar las variables que se desea reestructurar. Convertir casos en variables tiene el efecto de transformar, por ejemplo, el archivo representado en la Figura 6.8 (parte derecha) en el archivo representado en la Figura 6.7.

**Variables del archivo actual.** Ofrece un listado con todas las variables del archivo de datos.

**Variables de identificación.** Para que sea posible la conversión de casos a variables, el archivo original debe contener al menos una variable de identificación (como *id* en los datos de la Figura 6.8). La(s) variable(s) de identificación debe(n) trasladarse a esta lista. Conviene que el archivo original esté ordenado por la(s) variable(s) de identificación; si no lo está, puede ordenarse en el siguiente paso del *Asistente*.

**Variables índice.** Si existen variables *índice* (como *tiempo* y *contenido* en la Figura 6.8) y se desea incluirlas en el nuevo archivo de datos, deben trasladarse a esta lista. Si existe más de una variable *índice*, el orden en el que se trasladen a esta lista determinará el orden que seguirán las variables dentro de cada grupo de variables.

El resto de variables del archivo de datos (es decir, las variables no seleccionadas como **Variables de identificación** ni como **Variables índice**) son incluidas en el archivo reestructurado. El propio *Asistente* decide de forma automática qué hacer con las variables no seleccionadas: si los valores de una variable no seleccionada varían dentro de cada grupo de casos (es lo habitual), los casos se reestructuran como un grupo de variables; si cada grupo de casos tiene el mismo valor en esa variable, se crea una sola variable con el valor correspondiente a cada caso.

El cuadro de diálogo correspondiente al **Paso 3** del *Asistente para la reestructuración de datos* (no se incluye aquí) permite decidir cómo se ordenarán los datos en el archivo reestructurado. Si se elige la opción **Sí** (es la opción que se encuentra activa por defecto), los datos se ordenan por las variables de identificación y de *índice*; esto significa, básicamente, que las filas que pertenezcan al mismo caso aparecerán juntas y que las variables del archivo reestructurado se ordenarán de acuerdo con las variables *índice* seleccionadas; lo cual implica que, si todos los casos puntúan el mismo número de veces en todas las variables (es lo habitual), cada fila del archivo reestructurado corresponderá a un caso distinto. Si se elige la opción **No**, los datos conservarán el orden del archivo original; esto significa, básicamente, que las filas que perteneciendo al mismo caso estén separadas en el archivo original, seguirán separadas en el archivo reestructurado. Por tanto, si el archivo original no está ordenado por las variables de identificación, es necesario seleccionar la opción **Sí**.

En el cuadro de diálogo correspondiente al **Paso 4** del *Asistente* (ver Figura 6.10) hay que tomar tres decisiones. La primera de ellas se refiere al orden que adoptarán los grupos de variables en el archivo reestructurado. El archivo reestructurado va a contener tantos grupos de variables nuevas como variables se transpongan del archivo original. Y cada grupo de variables incluirá tantas variables como veces se repita cada caso en el archivo original. Las opciones del cuadro de diálogo correspondiente al paso 4 del *Asistente* permiten decidir cómo se ordenarán las nuevas variables: según las variables del archivo original o según las variables *índice* seleccionadas.

Si se transpone una sola variable, el archivo reestructurado contendrá un solo grupo de variables (como ocurre al convertir los datos de la Figura 6.8 en los de la Figura 6.7); el orden de esas variables puede controlarse tanto desde aquí como desde el cuadro de diálogo correspondiente al paso 3. Si se transponen dos variables, el archivo reestructurado contendrá dos grupos de variables: la opción **Agrupar por variable original** coloca juntas las variables creadas a partir de la misma variable original (coloca primero las variables del primer grupo y a continuación las variables del segundo grupo); la opción **Agrupar por variable índice** ordena las variables según los valores de las variables *índice* seleccionadas (coloca primero la primera variable del primer grupo y la primera del segundo; a continuación, la segunda variable del primer grupo y la segunda del segundo grupo; a continuación, la tercera variable del primer grupo y la tercera del segundo; etc.).

La segunda decisión que debe tomarse en el paso 4 del *Asistente* se refiere a si se desea o no crear una **Variable de recuento de casos**. Se trata de una variable que refleja el número de

filas del archivo original que se han utilizado para crear cada caso nuevo del archivo reestructurado. El procedimiento permite elegir un nombre y una etiqueta para la nueva variable.

Y la tercera decisión que debe tomarse en el paso 4 del *Asistente* se refiere a la creación de **Variables indicador**. Una variable *indicador* es una variable dicotómica cuyos valores son unos y ceros; los unos indican la presencia de la característica valorada; los ceros, la ausencia de la característica. En el contexto de la reestructuración de datos, las variables *indicador* se crean a partir de las variables *índice* (una variable *indicador* por cada valor único de la(s) variable(s) *índice*) y ofrecen información acerca de si un determinado caso tiene o no puntuación en cada una de las combinaciones de las variables *indicador*.

Figura 6.10. Cuadro de diálogo del *Asistente para la reestructuración de datos: casos a variables* (paso 3)

En el cuadro de diálogo correspondiente al **Paso 5** del *Asistente para la reestructuración de datos* (no se incluye aquí este cuadro de diálogo) es necesario decidir entre **Reestructurar los datos ahora** o **Pegar la sintaxis generada por el *Asistente* en una ventana de sintaxis**. Si se elige la primera opción, el botón **Finalizar** inicia la reestructuración del archivo y el archivo original es sustituido en el *Editor de datos* por el archivo reestructurado. Si se elige la segunda opción, el botón **Finalizar** abre una ventana de sintaxis (si no existe ninguna abierta) y pega en ella la sintaxis correspondiente a todas las elecciones hechas. Conviene señalar que, en el caso de que se opte por pegar la sintaxis, el botón **Finalizar** no inicia la reestructuración del archivo original; para iniciar la reestructuración es necesario ejecutar la sintaxis desde el *Editor de sin-*

*taxis* (puede consultarse el Capítulo 8 para una aclaración de cómo se trabaja con archivos de sintaxis).

## Fundir archivos

El SPSS permite combinar en un solo archivo los datos de archivos diferentes. Esta acción es necesaria cuando se desea incluir en el mismo análisis datos almacenados en archivos distintos. Existen dos posibilidades de fusión de archivos: (1) *añadir casos*, que consiste en combinar archivos que contienen las mismas variables, pero casos diferentes; y (2) *añadir variables*, que consiste en combinar archivos que contienen los mismos casos, pero distintas variables.

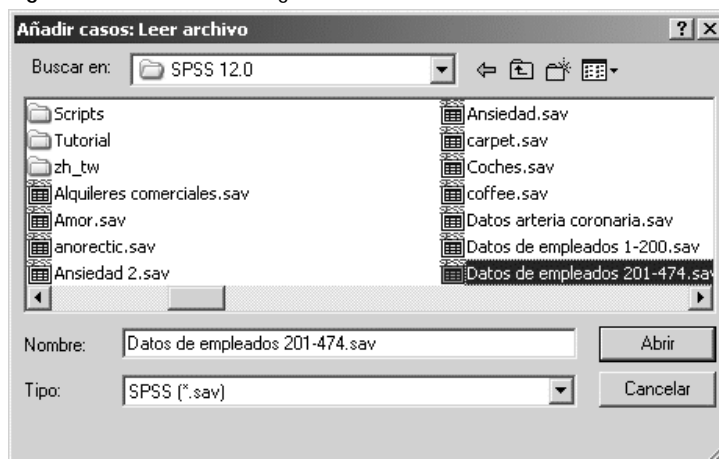
### Añadir casos

*Nota:* Para ilustrar el procedimiento **Fundir archivos > Añadir casos...** se ha partido en dos el archivo *Datos de empleados*. Los primeros 200 casos se han guardado en el archivo *Datos de empleados 1-200*; y los casos restantes se han guardado en el archivo *Datos de empleados 201-474*. Además, la variable *educ* mantiene el nombre en el archivo *Datos de empleados 1-200* pero ha cambiado a *estudios* en el archivo *Datos de empleados 201-474*. Así pues, aunque ambos archivos contienen las mismas variables (si bien una de ellas con nombre diferente), contienen casos distintos.

Para añadir casos se debe comenzar abriendo en el *Editor de datos* uno de los archivos que se desea combinar. A este primer archivo se le llama **archivo de trabajo**. Este ejemplo utiliza como archivo de trabajo *Datos de empleados 1-200*. Tras abrir este archivo, debe seleccionarse el archivo que se va a fundir con el de trabajo. Para ello:

- Seleccionar la opción **Fundir archivos > Añadir casos...** del menú **Datos** para acceder al cuadro de diálogo *Añadir casos: Leer archivo* que muestra la Figura 6.11.

Figura 6.11. Cuadro de diálogo *Añadir casos: Leer archivo*

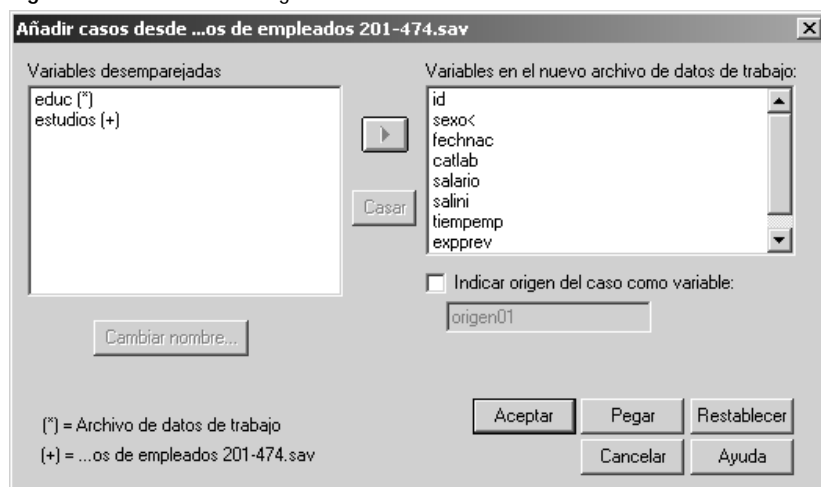


Este cuadro de diálogo contiene un listado de los archivos de datos (archivos con extensión *.sav*) de la carpeta a la que el SPSS accede por defecto. Puede verse que en ese listado aparecen, entre otros, los dos archivos que se han creado a partir del archivo *Datos de empleados*, es decir, el archivo *Datos de empleados 1-200* (que se está utilizando como archivo de trabajo) y el archivo *Datos de empleados 201-474*. Desde este cuadro de diálogo es posible seleccionar el archivo de datos cuyos casos se desea añadir al archivo de trabajo. A este segundo archivo se le llama **archivo externo**. Al archivo resultante de la fusión se le llama **archivo combinado**. Para realizar la fusión:

- Seleccionar el archivo *Datos de empleados 201-474* y pulsar el botón **Abrir** para acceder al cuadro de diálogo *Añadir casos desde...* que muestra la Figura 6.12.

En el ejemplo de la Figura 6.12, el archivo de datos externo seleccionado es *Datos de empleados 201-474*, lo cual queda reflejado en el título del cuadro de diálogo y en la parte inferior izquierda, precedido del signo +.

Figura 6.12. Cuadro de diálogo *Añadir casos desde...*



**Variables desemparejadas.** Ofrece un listado de las variables que no serán incluidas en el archivo combinado. Esta lista contiene:

- Variables que se encuentran sólo en uno de los dos archivos (caso del ejemplo).
- Variables definidas con formato *numérico* en uno de los archivos y con formato de *cadena* en el otro; las variables numéricas no pueden combinarse con las de cadena.
- Variables con formato de cadena que poseen distinto ancho. El ancho definido para una variable de cadena debe ser el mismo en ambos archivos.

Las variables del archivo de trabajo aparecen acompañadas de un asterisco (\*). Las variables del archivo externo aparecen acompañadas de un signo más (+). Para recordar esto, en la parte inferior izquierda del cuadro de diálogo aparecen los archivos de trabajo y externo precedidos de los símbolos que los identifican. En el ejemplo de la Figura 6.12, la

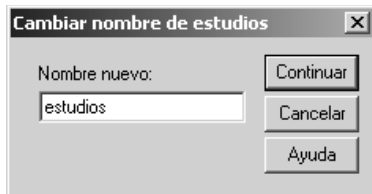
lista **Variables desemparejadas** contiene dos variables: la variable *educ*, que pertenece al archivo de trabajo (\*) y la variable *estudios*, que pertenece al archivo externo (+).

**Casar.** Si, como ocurre en el ejemplo, dos variables distintas contienen la misma información, el SPSS ofrece la posibilidad de emparejar esas dos variables para formar una sola. Para ello, basta con marcar las dos variables que se desea emparejar y pulsar el botón **Casar** situado debajo del botón flecha. Las dos variables son entonces trasladadas a la lista **Variables en el nuevo archivo de datos de trabajo** y situadas en la misma fila (en el archivo combinado prevalece el nombre de la variable perteneciente al archivo de trabajo). También es posible incluir en el archivo combinado cualquier variable individual de la lista **Variables desemparejadas** (aunque ya se sabe que estas variables se encuentran en sólo uno de los dos archivos). Para ello, basta con marcar la variable que se desea incluir y pulsar el botón flecha. Ahora bien, puesto que la variable *desemparejada* sólo se encuentra en uno de los dos archivos, los casos del archivo que no contiene esa variable serán, en el archivo combinado, casos con valor perdido en esa variable.

**Cambiar nombre.** Previamente a la acción de emparejar dos variables, puede resultar conveniente cambiar el nombre a cualquiera de ellas para que ambas posean el mismo nombre. Para cambiar el nombre de una variable:

- En la lista **Variables desemparejadas**, seleccionar la variable cuyo nombre se desea cambiar y pulsar el botón **Cambiar nombre...** para acceder al subcuadro de diálogo *Cambiar nombre de...* que muestra la Figura 6.13.

Figura 6.13. Subcuadro de diálogo *Cambiar nombre de...*



Este subcuadro de diálogo permite cambiar el nombre a la variable seleccionada (*estudios* en el ejemplo). Por supuesto, una variable puede ser renombrada independientemente de que se tenga o no intención de emparejarla con otra.

**Variables en el nuevo archivo de datos de trabajo.** Contiene un listado de las variables que pasarán a formar parte del nuevo archivo combinado. Por defecto, este listado incluye las variables que tienen el mismo nombre y formato en ambos archivos (de trabajo y externo). Las variables de este listado que no se desee incluir en el nuevo archivo combinado pueden eliminarse seleccionándolas y desplazándolas a la lista **Variables desemparejadas**.

- Indicar origen del caso como variable.** Esta opción permite crear una variable *indicador* dicotómica para identificar a qué archivo (de trabajo o externo) pertenecía originalmente cada caso del nuevo archivo combinado. Esta variable *indicador* toma el valor 0 para los casos del archivo de trabajo y el valor 1 para los casos del archivo externo. El nombre por defecto para esta variable es *origen01*, pero es posible introducir cualquier nombre válido en el cuadro de texto destinado a tal efecto.

## Añadir variables

*Nota:* Para ilustrar el procedimiento **Fundir archivos > Añadir variables...** se ha partido en dos el archivo *Datos de empleados*. Las variables *id*, *sexo*, *fechnac*, *educ*, *catlab*, *salario* y *salini* se han guardado en el archivo *Datos de empleados con salario*; y las variables *id*, *tiempemp*, *expprev* y *minoría* se han guardado en el archivo *Datos de empleados sin salario*. Así pues, aunque ambos archivos contienen los mismos casos, contienen distintas variables (excepto por lo que se refiere a la variable *id*, que se encuentra en ambos archivos).

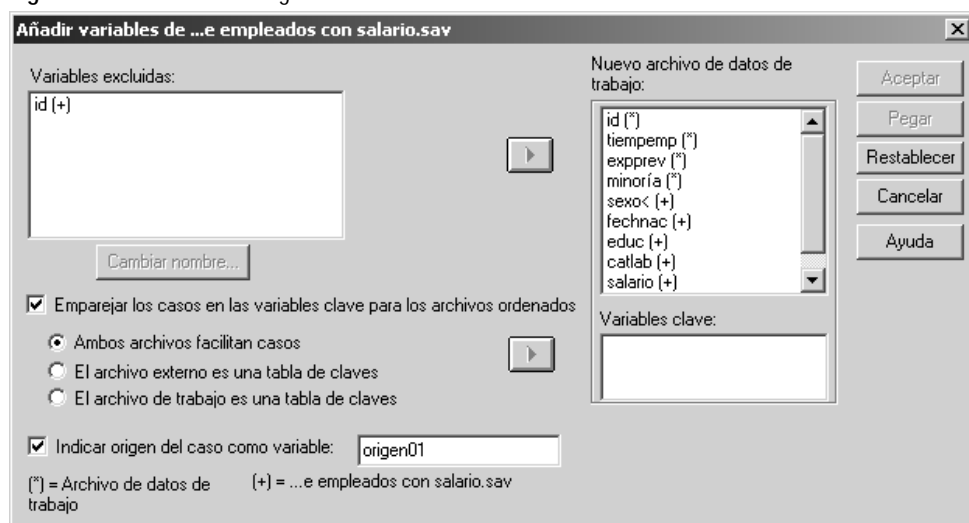
Para combinar dos archivos de datos SPSS con los mismos casos pero con distintas variables es necesario que los casos de ambos archivos estén ordenados con el mismo criterio.

Para fundir dos archivos añadiendo variables debe comenzarse abriendo en el *Editor de datos* uno de los archivos que se desea combinar; a este primer archivo se le llama **archivo de trabajo**. Este ejemplo utiliza como archivo de trabajo el archivo *Datos de empleados con salario*. Tras abrir este archivo:

- Seleccionar la opción **Fundir archivos > Añadir variables...** del menú **Datos** para acceder al cuadro de diálogo *Añadir variables: Leer archivo* (idéntico al de la Figura 6.11).

Este cuadro de diálogo, que contiene un listado de los archivos de datos de la carpeta que se abre por defecto, permite seleccionar el **archivo externo** que se desea combinar con el de trabajo. Al archivo resultante de la fusión se le llama **archivo combinado**. Seleccionando el archivo *Datos de empleados sin salario* y pulsando el botón **Abrir** se accede al cuadro de diálogo *Añadir variables de...* que muestra la Figura 6.14.

Figura 6.14. Cuadro de diálogo *Añadir variables de...*



En el ejemplo de la Figura 6.14, el archivo de datos externo seleccionado es *Datos de empleados sin salario*, lo cual queda reflejado en el título del cuadro de diálogo y en la parte inferior izquierda del cuadro, donde aparece el nombre precedido del signo +.

**Variables excluidas.** Este cuadro ofrece un listado de las variables que, en principio, no formarán parte del archivo combinado. Incluye, por defecto, las variables del archivo externo cuyo nombre coincide con el de alguna variable del archivo de trabajo. Si se desea incluir en el archivo combinado el contenido de una variable con nombre duplicado, es necesario renombrar esa variable (utilizando el botón **Cambiar nombre...**) y añadirla a la lista **Nuevo archivo de datos de trabajo** mediante el botón flecha. Al igual que antes, las variables del archivo de trabajo aparecen acompañadas de un asterisco (\*), mientras que las del archivo externo lo están de un signo más (+).

**Nuevo archivo de datos de trabajo.** Este cuadro ofrece un listado de las variables que pasarán a formar parte del archivo combinado. Por defecto, el listado recoge: (1) todas las variables del archivo de trabajo y (2) las variables del archivo externo cuyo nombre no está duplicado en el de trabajo.

“ **Emparejar los casos en las variables clave para los archivos ordenados.** Si los dos archivos que se van a fundir no contienen exactamente los mismos casos, es posible utilizar una o más variables clave para emparejar correctamente los casos de ambos archivos. Esta variable clave debe tener el mismo nombre en los dos archivos y los casos deben estar, en ambos archivos, ordenados de forma ascendente según esa variable clave. Si se utiliza más de una variable clave, los casos deben estar ordenados de forma ascendente en el conjunto de variables clave utilizadas.

Puesto que las variables clave deben ser variables duplicadas, aparecerán en la lista **Variables excluidas**. Para utilizar una variable clave, es necesario marcar la opción **Emparejar los casos en las variables clave para los archivos ordenados**, seleccionar las variables que se van a utilizar como claves, y trasladarlas a la lista **Variables clave** mediante el correspondiente botón flecha. Hecho esto, el procedimiento permite elegir entre tres métodos diferentes de emparejamiento de casos:

**Ambos archivos facilitan casos.** Esta opción supone que existe una correspondencia uno a uno entre los casos de ambos archivos. Es decir, se supone que cada caso posee un valor único en la(s) variable(s) clave(s). Si existen dos o más casos con el mismo valor en la(s) variable(s) clave(s), el SPSS los empareja en orden secuencial y ofrece un mensaje advirtiendo de tal circunstancia.

**El archivo externo es una tabla de claves.** El archivo externo actúa como una tabla de claves. Cada caso del archivo externo puede emparejarse con más de un caso del archivo de trabajo.

**El archivo de trabajo es una tabla de claves.** El archivo de trabajo actúa como una tabla de claves. Cada caso del archivo de trabajo puede emparejarse con más de un caso del archivo externo.

“ **Indicar origen del caso como variable.** Esta opción permite crear una *variable indicador* para identificar a qué archivo (de trabajo o externo) pertenecía originalmente cada caso del nuevo archivo combinado. Esta variable *indicador* toma el valor 0 para los casos del archivo de trabajo ausentes del archivo externo y el valor 1 para los casos del archivo externo (pertenecan o no al archivo de trabajo). El nombre por defecto para esta variable *indicador* es *origen01*, pero es posible asignar cualquier otro nombre válido utilizando el cuadro de texto destinado a tal efecto.



## Agregar datos

Agregar datos consiste en agrupar varios casos en uno. Un archivo agregado tiene, por tanto, menos casos que el archivo original.

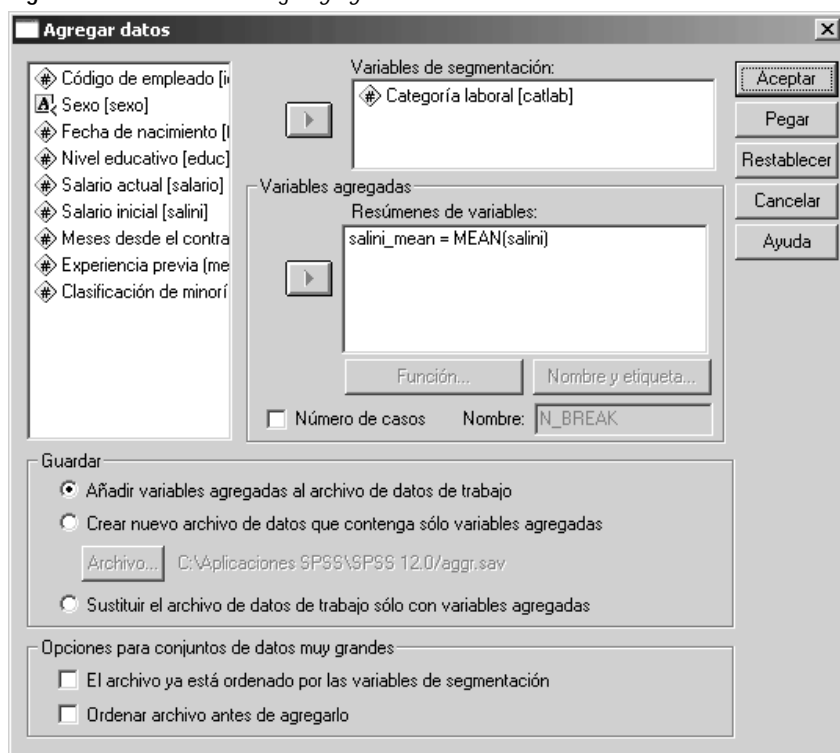
También es posible agregar datos haciendo que el archivo agregado tenga tantos casos como el archivo original; en ese caso, todos los casos del archivo original agrupados en el mismo caso agregado tendrán el mismo valor en las nuevas variables agregadas.

Más adelante, en este mismo capítulo, se explica la opción **Ponderar casos**, la cual, permite hacer justamente lo contrario de lo que hace la opción **Agregar**.

Para agregar casos:

- Seleccionar la opción **Agregar...** del menú **Datos** para acceder al cuadro de diálogo *Agregar datos* que muestra la Figura 6.15.

Figura 6.15. Cuadro de diálogo *Agregar datos*



**Variables de segmentación.** Los casos del archivo original se agrupan tomando como referencia los niveles de una o más variables de segmentación. En el ejemplo de la Figura 6.15 se ha utilizado la variable *catlab* (categoría laboral) como variable de segmentación, lo que significa que todos los sujetos que tengan la misma categoría laboral (el mismo valor en *catlab*) pasarán a formar un único caso en el nuevo archivo de datos agregados. Las variables de segmentación pueden ser numéricas o de cadena.

**Variables agregadas.** Las variables del archivo agregado se obtienen a partir de las variables del archivo original. Los valores de las nuevas variables son el resultado de aplicar alguna **Función** (ver más abajo) a las variables del archivo original. En el ejemplo de la Figura 6.15 se ha decidido que el nuevo archivo contenga una variable: *salini\_1* (el nombre es automáticamente asignado por el SPSS; ver, más abajo, **Nombre y etiquetas...**). Los valores de la variable *salini\_1* serán el resultado de obtener, para cada nuevo caso agregado, la media aritmética (MEAN) de la variable *salini* en todos los casos de cada segmento definido por *catlab*.

“ **Número de casos** Al marcar esta opción, el nuevo archivo incluye una variable que contiene el número de casos de cada segmento. Esta variable recibe, por defecto, el nombre *n\_break*, pero puede utilizarse el cuadro de texto **Nombre** para asignarle cualquier nombre válido.

**Guardar.** En lo referente a la ubicación del nuevo archivo de datos agregados, el SPSS ofrece tres alternativas:

**Añadir variables agregadas al nuevo archivo de trabajo.** Las nuevas variables son añadidas al final del archivo de trabajo. Todos los casos pertenecientes al mismo segmento adoptan el mismo valor en las variables agregadas.

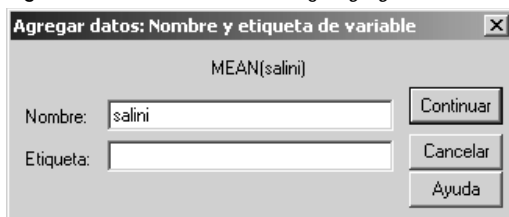
**Crear nuevo archivo de datos que contenga sólo variables agregadas.** Crea un nuevo archivo con los datos agregados y le asigna el nombre *aggr.sav*. El botón **Archivo...** permite cambiar el nombre y la ruta del nuevo archivo agregado.

**Sustituir el archivo de datos de trabajo sólo variables agregadas.** Alternativamente, en lugar de crear un nuevo archivo y almacenarlo en disco, puede decidirse que el nuevo archivo con los datos agregados pase a ser el archivo de trabajo (sustituyendo al archivo de trabajo actual).

**Nombre y etiquetas.** Las nuevas variables son nombradas, por defecto, añadiendo «\_1» a los seis primeros caracteres de la variable original. En la Figura 6.15, la variable *salini* ha sido renombrada como *salini\_1*. Si se desea cambiar el nombre asignado por defecto:

- Pulsar el botón **Nombre y etiquetas...** (ver Figura 6.15) para acceder al cuadro de diálogo *Agregar datos: Nombre y etiqueta de variable* que muestra la Figura 6.16.

Figura 6.16. Subcuadro de diálogo *Agregar datos: Nombre y etiqueta de variable*



**Nombre.** Este cuadro de texto permite cambiar el nombre asignado por defecto. El nuevo nombre debe cumplir con las especificaciones de los nombres de variable (ver, en el Capítulo 4, el apartado *Definir variables: asignar nombre a una variable*).

**Etiqueta.** Permite asignar una etiqueta descriptiva de hasta 120 caracteres.

**Función.** Si no se indica otra cosa, el SPSS asume que la función estadística que se desea utilizar es la media aritmética. No obstante, es posible elegir entre una gran variedad de funciones diferentes. Para seleccionar una función distinta de la media aritmética:

- Pulsar el botón **Función...** (ver Figura 6.15) para acceder al subcuadro de diálogo *Agregar datos: Función de agregación* que muestra la Figura 6.17.

Este subcuadro de diálogo permite seleccionar varias funciones, todas las cuales están apropiadamente descritas en la ayuda contextual que ofrece el SPSS al pinchar con el botón secundario del ratón en cualquiera de ellas. Las funciones se calculan para cada segmento definido por la(s) variable(s) de agrupación, es decir para cada nuevo caso del archivo agregado.

Figura 6.17. Subcuadro de diálogo *Agregar datos: Función de agregación*

**Agregar datos: Función de agregación**

Estadísticos de resumen	Valores específicos	Número de casos
<input checked="" type="radio"/> Media	<input type="radio"/> Primero	<input type="radio"/> Ponderado
<input type="radio"/> Mediana	<input type="radio"/> Último	<input type="radio"/> Perdido ponderado
<input type="radio"/> Suma	<input type="radio"/> Mínimo	<input type="radio"/> Sin ponderar
<input type="radio"/> Desviación típica	<input type="radio"/> Máximo	<input type="radio"/> Perdido sin ponderar

**Porcentajes**

☐ Por encima Valor:

☐ Por debajo

☐ Dentro Menor:  Mayor:

☐ Fuera

**Fracciones**

☐ Por encima Valor:

☐ Por debajo

☐ Dentro Mínimo:  Superior:

☐ Fuera

Continuar Cancelar Ayuda

Las funciones disponibles están agrupadas en cinco bloques. **Estadísticos de resumen:** algunos estadísticos descriptivos: media aritmética, mediana, suma y desviación típica. **Valores específicos:** valor correspondiente al primer caso de cada segmento, valor correspondiente al último caso de cada segmento, valor más pequeño de cada segmento y el valor más grande de cada segmento. **Número de casos:** número de casos teniendo en cuenta la ponderación de casos (si existiese), número de casos con valor perdido teniendo en cuenta la ponderación de casos (si existiese), número de casos y número de casos con valor perdido. **Porcentajes:** porcentaje de casos que quedan por encima o por debajo de un determinado valor, y dentro y fuera de un determinado rango. **Fracciones:** porcentajes divididos entre 100.

La función o funciones seleccionadas en este subcuadro de diálogo (sólo es posible seleccionar una por bloque) aparecen en la lista **Agregar variables** del cuadro de diálogo *Agregar datos* (ver Figura 6.15), junto al nombre de la nueva variable.

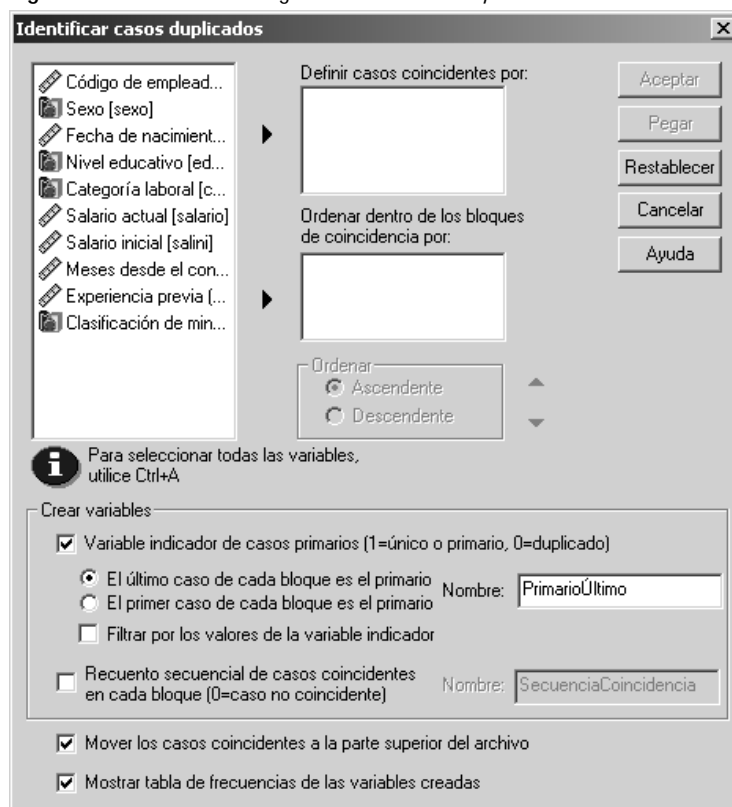
## Identificar casos duplicados

En un archivo de datos pueden aparecer casos duplicados por diferentes razones. En ocasiones puede tratarse simplemente de un error cometido en la variable de identificación al introducir los datos. Otras veces puede tratarse del mismo caso que está presente más de una vez en el archivo porque se ha introducido dos veces por error. También puede ocurrir que haya varios casos que compartan un código de identificación primario pero que posean distinto código de identificación secundario, como los miembros de una familia que viven en el mismo hogar, o los empleados de un mismo departamento. Un ejemplo más de duplicidad se da cuando un mismo caso se encuentra repetido en el archivo (mismo código de identificación) pero posee valores distintos en las variables que no son de identificación, como cuando un mismo cliente realiza varias compras en distintos momentos, o un paciente realiza varias consultas en un mismo mes.

El SPSS incluye un procedimiento diseñado para ayudar a encontrar casos duplicados y realizar varias acciones con ellos. Para identificar casos duplicados:

- Seleccionar la opción **Identificar casos duplicados...** del menú **Datos** para acceder al cuadro de diálogo *Identificar casos duplicados* que muestra la Figura 6.18.

Figura 6.18. Cuadro de diálogo *Identificar casos duplicados*



**Definir casos coincidentes por.** Las variables en las que se desea basar la búsqueda de duplicados deben trasladarse a esta lista. El procedimiento considera que dos casos son duplicados cuando sus valores son iguales en todas las variables incluidas en esta lista. Los casos con valor perdido en una variable numérica o con valor *vacío* en una variable de cadena se considera que tienen el mismo valor en esa variable.

**Ordenar dentro de bloques de coincidencias por.** Opcionalmente, puede seleccionarse una o más variables (cuyos niveles definan subgrupos) para que cada bloque de casos duplicados quede secuencialmente ordenado según los niveles de esa o esas variables. Si la variable utilizada para definir casos duplicados es *id* (código de empleado), los casos que tengan el mismo valor en *id*, si existen, aparecerán juntos en el archivo de datos. Si además se selecciona una variable de ordenación, por ejemplo *catlab* (categoría laboral), los casos con el mismo valor en *id* (es decir, los casos de cada bloque de casos duplicados), si existen, aparecerán ordenados por *catlab*.

Si se elige orden **Ascendente**, los casos se ordenan de menor a mayor (variables numéricas) o alfabéticamente (variables de cadena). Si se elige orden **Descendente**, los casos se ordenan de forma inversa. A este respecto, debe tenerse en cuenta que en el orden alfabético, las mayúsculas preceden a las minúsculas, los números preceden a las letras y el resto de caracteres (asterisco, interrogación, admiración, etc.) preceden a los números.

Si se elige más de una variable de ordenación, los casos se ordenan atendiendo al orden en el que se han seleccionado las variables. Las flechas situadas a la derecha de las opciones **Ascendente** y **Descendente** permiten cambiar el orden inicial de las variables. Si no se selecciona ninguna variable de ordenación, los casos duplicados de cada bloque se ordenan de acuerdo con el orden original del archivo.

**Crear variables.** Las opciones de este recuadro permiten crear nuevas variables con información sobre la presencia de casos duplicados:

- " **Variable indicador de casos primarios (1=único o primario; 0=duplicado).** Esta opción, que se encuentra activa por defecto, crea una variable *indicador* con unos y ceros. A los casos únicos o no duplicados y al caso primario de cada bloque de casos duplicados se les asigna un uno; a los casos duplicados, un cero. El procedimiento permite decidir a qué caso de cada bloque de casos duplicados se le desea asignar un uno (caso primario):

El último caso de cada bloque es el **primario**. Asigna un 1 al último caso de cada bloque de casos duplicados. Si se elige esta opción, el procedimiento asigna a la variable indicador el nombre *PrimarioÚltimo*. Este nombre puede cambiarse introduciendo cualquier nombre que se ajuste a las reglas de los nombres de variable del SPSS.

El primer caso de cada bloque es el **primario**. Asigna un uno al primer caso de cada bloque de casos duplicados. Si se elige esta opción, el procedimiento asigna a la variable indicador el nombre *PrimarioPrimero*. Este nombre puede cambiarse introduciendo cualquier nombre que se ajuste a las reglas de los nombres de variable del SPSS.

- " **Filtrar por los valores de la variable indicador.** Si se opta por filtrar el archivo de datos, los casos identificados como duplicados son automáticamente fil-

trados y, consecuentemente, excluidos de cualquier procedimiento estadístico o gráfico.

- " **Recuento secuencial de casos coincidentes en cada bloque (0=caso no coincidente).** Crea una variable con valores secuenciales de 1 a  $k$  que reflejan el número ( $k$ ) de casos iguales dentro de cada bloque de casos duplicados. En esta variable, a los casos únicos o no duplicados se les asigna un cero. A los casos de un bloque de, por ejemplo, 3 casos duplicados, se les asigna los valores 1, 2 y 3. Si en un bloque de casos duplicados existen, por ejemplo, 7 casos, al primero se le asigna un 1, al segundo un 2, ..., al séptimo un 7.

El valor que recibe cada caso depende del tipo de ordenación seleccionado. Si no se ha seleccionado ninguna variable de ordenación, el orden de los casos dentro de cada bloque de casos duplicados es el orden original del archivo de datos. Si se ha seleccionado una o más variables de ordenación, el orden de los casos dentro de cada bloque de casos duplicados viene impuesto por estas variables.

A esta variable de valores secuenciales de 1 a  $k$  se le asigna, por defecto, el nombre *SecuenciaCoincidencia*, pero este nombre puede cambiarse introduciendo cualquier nombre que se ajuste a las reglas de los nombres de variable del SPSS (ver, en el Capítulo 4, el apartado *Definir variables: asignar nombre a una variable*).

- " **Mover lo casos coincidentes a la parte superior del archivo.** Marcando esta opción los casos identificados como duplicados son colocados al principio del archivo de datos.
- " **Mostrar la tabla de frecuencias de las variables creadas.** Ofrece una tabla de frecuencias con información sobre las variables que se han creado. La tabla de frecuencias de las variables *PrimarioÚltimo* o *PrimarioPrimero* muestra el número de unos (número de casos únicos o primarios) y ceros (número casos duplicados). La tabla de frecuencias de la variable *SecuenciaCoincidencia* muestra el número de ceros (número de casos únicos), el número de unos (número de bloques con al menos 1 caso duplicado), el número de doses (número de bloques con al menos 2 casos duplicados), el número de treses (número de bloques con al menos 3 casos duplicados), etc.

## Diseño ortogonal

La opción **Diseño ortogonal** permite crear un archivo de datos con información relativa a los efectos principales de un diseño ortogonal. Está especialmente diseñada para generar y preparar los datos antes de utilizar el procedimiento *análisis conjunto* (*conjoint analysis*). Este procedimiento puede adquirirse como un módulo adicional del SPSS. El *análisis conjunto* es una técnica diseñada para valorar las preferencias de los consumidores de un producto o de los usuarios de un servicio a partir del estudio de las características o atributos de ese producto o servicio. Aunque no es propósito de este apartado explicar en qué consiste el *análisis conjunto*, es conveniente conocer qué es lo que pretende el *análisis conjunto* para entender mejor lo que significa un diseño ortogonal.

Supongamos que un cliente desea comprar un coche y que puede elegir entre un coche de 20.000 euros o un coche de 40.000 euros; si ambos coches fueran idénticos en todo excepto en el precio, la elección tendría pocas complicaciones. Supongamos ahora que la elección

debe hacerse entre un coche que gasta poca gasolina y otro que gasta mucha; si el consumo fuera la única diferencia entre ambos coches, de nuevo la elección sería igualmente clara. Supongamos, finalmente, que la elección debe hacerse entre un coche poco potente y otro muy potente; una vez más, si la potencia fuera la única diferencia, la mayor parte de las personas tendrían clara su elección. El inconveniente de los tres supuestos anteriores es que, en el mundo real, las diferentes alternativas de elección no se dan por separado, sino que aparecen combinadas en un único producto.

Pues bien, para conocer las preferencias de las personas, pueden agruparse las diferentes alternativas de elección en conjuntos de tres atributos tal como muestra la Tabla 6.2. Combinando los diferentes niveles de los tres atributos (2 *precios*, 2 *consumos* y 2 *potencias*) se obtienen 8 conjuntos de atributos.

Probablemente, la combinación 2 será la más preferida y la número 7 la menos preferida. Pero, para saber lo que realmente prefieren los clientes, lo apropiado es presentar estos 8 conjuntos de atributos a un grupo de sujetos solicitando de ellos que manifiesten su opinión ordenando las 8 combinaciones desde la más preferida a la menos preferida, o desde la que es más probable que compren a la que es menos probable que compren, etc. Una vez obtenidas las preferencias de los sujetos, el *análisis conjunto* permite establecer cuál es la importancia relativa de cada atributo y cuál es el nivel más preferido de cada atributo.

**Tabla 6.2.** Conjuntos de atributos: combinaciones de *precio*, *consumo* y *potencia*

<i>Combinaciones</i>	<i>Precio</i>	<i>Consumo</i>	<i>Potencia</i>
1	20.000	Bajo	Baja
2	20.000	Bajo	Alta
3	20.000	Alto	Baja
4	20.000	Alto	Alta
5	40.000	Bajo	Baja
6	40.000	Bajo	Alta
7	40.000	Alto	Baja
8	40.000	Alto	Alta

El problema de esta aproximación es que el número de conjuntos de atributos se incrementa rápidamente al añadir algún atributo más y algún nivel más por atributo. Por ejemplo, si en lugar de 3 atributos y 2 niveles por atributo (como en el ejemplo de la Tabla 6.2) se utilizan 5 atributos y 3 niveles por atributo (situación ésta bastante realista), el número de combinaciones que resultan sube hasta  $3^5 = 243$ . Esto significa que, con sólo 5 atributos y 3 niveles por atributo, la solución se hace prácticamente intratable: si se decidiera presentar todas estas combinaciones a un sujeto, además del tiempo y coste que esto supondría, generaría en él un estado de fatiga que podría invalidar el resultado; por otro lado, no parece que exista forma razonable de ordenar con significado 243 combinaciones.

Este problema puede resolverse creando un *diseño ortogonal*. Un diseño ortogonal es un conjunto reducido de combinaciones de atributos que conservan la propiedad de permitir evaluar los efectos que interesa evaluar sin necesidad de presentar a los sujetos todas las combinaciones posibles. Es decir, en un diseño ortogonal sólo se seleccionan unas pocas combinaciones de todas las posibles; la selección se realiza de tal forma que las combinaciones elegi-

das permitan estudiar los efectos principales (los efectos de cada atributo o *factor*), desestimando las interacciones entre atributos. Esto se consigue haciendo que cada nivel de un atributo o factor se combine el mismo número de veces (o un número proporcional de veces) con el resto de niveles de los demás atributos o factores. El lector poco familiarizado con esta terminología puede consultar el Capítulo 15 sobre *ANOVA factorial*; en él se describen los conceptos de *factor*, *efecto principal* e *interacción*.

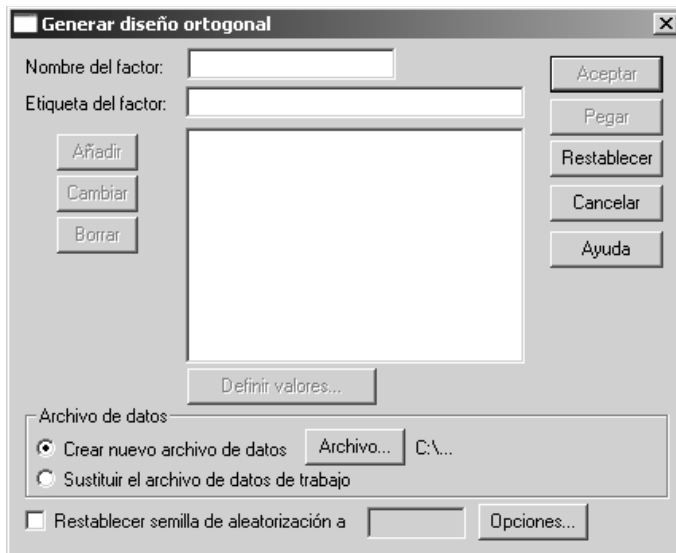
Un diseño ortogonal representa la forma más parsimoniosa de evaluar los efectos principales (pues utiliza el número mínimo de combinaciones necesarias para estimarlos). Pero, puesto que no permite obtener información sobre la interacción entre atributos, debe tenerse en cuenta que la valoración de un nivel concreto se hace independientemente de la presencia de los niveles del resto de atributos.

## Generar un diseño ortogonal

Para crear un diseño ortogonal:

- Seleccionar la opción **Diseño ortogonal > Generar...** del menú **Analizar** para acceder al cuadro de diálogo *Generar diseño ortogonal* que muestra la Figura 6.19.

Figura 6.19. Cuadro de diálogo *Generar diseño ortogonal*



A diferencia de lo que ocurre en la mayor parte de los procedimientos SPSS, la opción **Generar** del procedimiento **Diseño ortogonal** no requiere que el *Editor de datos* contenga un archivo de datos. Para definir los *factores* que formarán parte del diseño ortogonal:

- Asignar un nombre al primer factor en el cuadro de texto **Nombre del factor**. En el contexto de los diseños ortogonales, los factores se refieren a los atributos del producto



o servicio que se desea evaluar (en el ejemplo de la Tabla 6.2, los factores del producto *coche* son: *precio*, *consumo* y *potencia*). Es necesario definir al menos un factor. Y el nombre utilizado debe atenerse a las reglas de los nombres de variable del SPSS (ver, en el Capítulo 4, el apartado *Definir variables*). Los nombres *status\_* y *card\_* están reservados (los utiliza el SPSS para identificar las variables que crea al generar el diseño ortogonal).

- Asignar, si así se desea, una etiqueta descriptiva en el cuadro de texto **Etiqueta del factor**.
- Pulsar el botón **Añadir** para validar el nombre y la etiqueta del primer factor. Los botones **Cambiar** y **Borrar** permiten modificar y eliminar factores previamente añadidos.
- Repetir los pasos anteriores para cada uno de los factores que se desee definir.

Tras definir un factor (es decir, tras asignarle nombre y etiqueta y pulsar el botón **Añadir**), es necesario indicar cuáles son sus niveles. Para ello.

- Seleccionar el factor cuyos niveles se desea establecer y pulsar el botón **Definir valores...** para acceder al subcuadro de diálogo *Generar diseño ortogonal: Definir valores* que muestra la Figura 6.20.

Figura 6.20. Subcuadro de diálogo *Generar diseño ortogonal: Definir valores*

Valores y etiquetas para factor1	
Valor	Etiqueta
1: <input type="text"/>	<input type="text"/>
2: <input type="text"/>	<input type="text"/>
3: <input type="text"/>	<input type="text"/>
4: <input type="text"/>	<input type="text"/>
5: <input type="text"/>	<input type="text"/>
6: <input type="text"/>	<input type="text"/>
7: <input type="text"/>	<input type="text"/>
8: <input type="text"/>	<input type="text"/>
9: <input type="text"/>	<input type="text"/>

Continuar Cancelar Ayuda

Auto-relleno  
Del 1 al  Rellenar

- Introducir el código del primer nivel del factor en la primera fila de la columna **Valor**. Los códigos que se asignan a los niveles de un factor suelen ser enteros consecutivos de 1 a  $k$  ( $k$  = número de niveles). El botón **Rellenar** del recuadro **Auto-relleno** permite asignar automáticamente estos enteros consecutivos de 1 a  $k$ ; para ello, basta con introducir el valor  $k$  en el cuadro de texto **Del 1 al  $k$**  y pulsar el botón **Rellenar**.
- Introducir, en la columna **Etiqueta**, una etiqueta descriptiva asociada a cada nivel del factor. Si no se asignan etiquetas, el procedimiento asigna automáticamente enteros consecutivos de 1 a  $k$ .

- Una vez introducidos todos los valores y etiquetas, pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

**Archivo de datos.** Con los casos y variables del diseño ortogonal generado puede hacerse una de estas dos cosas:

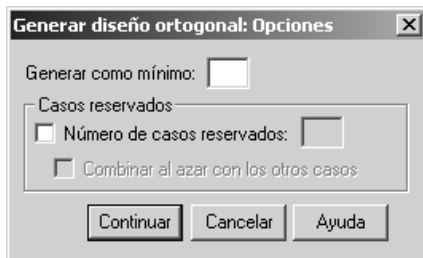
**Crear nuevo archivo de datos.** Crea un nuevo archivo en disco, dejando el *Editor de datos* tal como se encuentra. El procedimiento asigna, por defecto, a este nuevo archivo el nombre *orto.sav* y lo guarda en la carpeta activa (la última carpeta utilizada). El botón **Archivo...** abre un cuadro de diálogo que permite asignar al nuevo archivo el nombre y la ruta deseados.

**Sustituir el archivo de datos de trabajo.** Coloca el diseño ortogonal generado en el *Editor de datos* (reemplazando, si existe, el archivo del *Editor de datos*).

- “ **Restablecer semilla de aleatorización a \_\_.** Esta opción permite establecer un valor entre 0 y 2.000.000.000 para la semilla aleatoria (ver, en el capítulo anterior, el apartado *Generador de números aleatorios*). Durante la misma sesión, la semilla aleatoria cambia aleatoriamente cada vez que el SPSS genera una serie aleatoria. Esto significa que las distintas series aleatorias solicitadas al SPSS no serán siempre las mismas (justamente por ser aleatorias). No obstante, existe la posibilidad de replicar una serie aleatoria si se fuerza al generador de números aleatorios a comenzar con la misma semilla.

El botón **Opciones...** del cuadro de diálogo principal (ver Figura 6.19) conduce al subcuadro de diálogo *Generar diseño ortogonal: Opciones* que muestra la Figura 6.21. Este subcuadro de diálogo permite controlar algunos detalles del procedimiento.

**Figura 6.21.** Subcuadro de diálogo *Generar diseño ortogonal: Opciones*



**Generar como mínimo \_\_.** El procedimiento genera un diseño ortogonal basado en el mínimo número de casos necesario para estimar los efectos principales de cada factor. A este respecto, debe tenerse en cuenta que cada caso representa una combinación entre los niveles de los factores.

Si se desea que este número mínimo de casos (combinaciones) sea mayor que el que el procedimiento determina automáticamente, puede especificarse aquí ese número mínimo de casos introduciendo un valor entero igual o menor que el número total de posibles combinaciones entre los niveles de los factores. Si el número elegido representa un mínimo insuficiente para obtener las estimaciones de interés, el procedimiento generará un número mayor de casos (que seguirá siendo el mínimo necesario).

**Casos reservados.** Los casos *reservados* son casos diferentes (añadidos) de los casos del diseño ortogonal. Se presentan a los sujetos, pero no se incluyen en el análisis conjunto. Se utilizan para validar las estimaciones que el análisis conjunto ofrece a partir de los casos del diseño ortogonal.

Estos casos (tantos como se soliciten en la opción **Número de casos reservados**) se colocan, por defecto, al final del archivo (detrás de los casos correspondientes al diseño ortogonal), pero, opcionalmente, pueden intercalarse entre los casos del diseño ortogonal marcando la opción **Combinar al azar con los otros casos**.

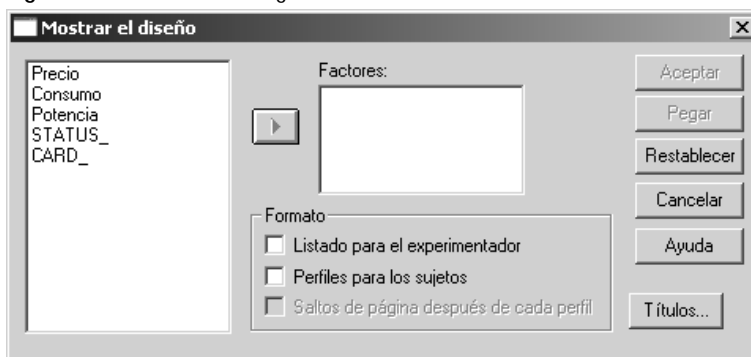
Definidos los factores y sus niveles, y decidido si se desea o no crear casos reservados, el botón **Aceptar** del cuadro de diálogo principal genera un diseño ortogonal con el número mínimo de casos necesario para la estimación de los efectos principales (o el número mínimo de casos establecido por el usuario; el mayor de ambos mínimos) y con tantas variables como factores, más dos variables nuevas: *status\_* y *card\_*. La variable *status\_* indica si un caso pertenece al diseño ortogonal (0=Diseño) o es un caso reservado (1=Exclusión). La variable *car\_* simplemente asigna un número secuencial a cada caso.

## Mostrar un diseño ortogonal

Una vez que se ha generado un diseño ortogonal, es necesario colocar cada combinación de niveles (perfil) en una tarjeta para poder presentarla a los sujetos. Cada caso del diseño ortogonal puede obtenerse como un *perfil* que incluye una combinación particular de niveles. Para obtener estos perfiles:

- Abrir en el *Editor de datos* el archivo que contiene el diseño ortogonal.
- Seleccionar la opción **Diseño ortogonal > Mostrar...** del menú **Analizar** para acceder al cuadro de diálogo *Mostrar el diseño* que muestra la Figura 6.22.

Figura 6.22. Cuadro de diálogo *Mostrar el diseño*



La lista de variables del archivo de datos muestra un listado de todas las variables excepto *status\_* y *card\_*. Para mostrar los perfiles de un diseño ortogonal es necesario seleccionar los factores que se desea incluir y trasladarlos a la lista **Factores**.

**Formato.** Las opciones de este recuadro permiten controlar el aspecto que debe adoptar el listado de los perfiles en el *Visor de resultados*:

- " **Listado para el experimentador.** Muestra los perfiles en un formato borrador que permite realizar una inspección rápida del resultado. En este formato aparecen identificados los perfiles que pertenecen a casos del diseño ortogonal y los que pertenecen a casos reservados.

La parte izquierda de la Tabla 6.3 muestra un ejemplo de este tipo de formato. En este ejemplo se ha creado un diseño ortogonal con los mismos tres factores de la Tabla 6.2, pero con tres niveles por factor (*precio* = «20.000», «30.000», «40.000»; *consumo* = «bajo», «medio», «alto»; *potencia* = «baja», «media», «alta») y solicitando dos casos reservados.

De los  $3^3=27$  perfiles posibles, el procedimiento ha generado un diseño ortogonal con sólo 9. Los perfiles 10 y 11 pertenecen a casos reservados (*Holdout*).

- " **Listado para los sujetos.** Muestra los perfiles en un formato que puede ser presentado los sujetos. En este formato no aparecen diferenciados los casos del diseño ortogonal y los casos reservados (ver Tabla 6.3, parte derecha).

**Tabla 6.3.** Listado de los perfiles para el experimentador (izquierda) y para los sujetos (derecha)

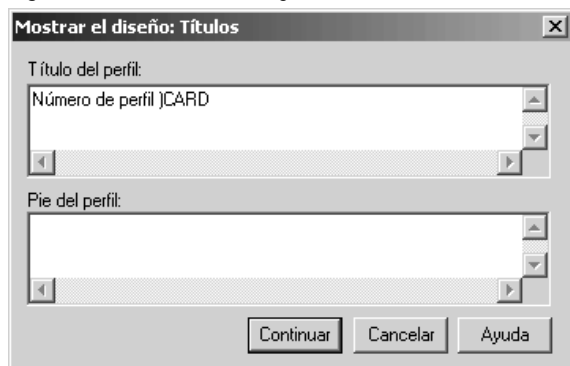
Title: COCHES	COCHES
Card 1	Precio 20.000
Precio 20.000	Consumo Medio
Consumo Medio	Potencia Media
Potencia Media	
Card 2	COCHES
Precio 40.000	Precio 40.000
Consumo Medio	Consumo Medio
Potencia Alta	Potencia Alta
Card 3	...
Precio 60.000	COCHES
Consumo Medio	Precio 40.000
Potencia Baja	Consumo Bajo
...	Potencia Alta
Card 10 (Holdout)	COCHES
Precio 40.000	Precio 40.000
Consumo Bajo	Consumo Bajo
Potencia Alta	Potencia Baja
Card 11 (Holdout)	
Precio 40.000	
Consumo Bajo	
Potencia Baja	

- " **Salto de página después de cada perfil.** Esta opción permite insertar saltos de página en el listado de los perfiles. De este modo, cada perfil puede imprimirse en una página diferente.

El botón **Títulos...** del cuadro de diálogo principal (ver Figura 6.22) conduce al subcuadro de diálogo *Mostrar el diseño: Títulos* que recoge la Figura 6.23. El texto que se introduzca en el cuadro **Título del perfil** (hasta un máximo 80 caracteres) aparecerá al comienzo del listado para el experimentador y al comienzo de cada perfil del listado para los sujetos. En el título que se ofrece por defecto, la variable “)CARD” inserta el número de perfil en el listado para

los sujetos. El texto introducido en el cuadro **Pie del perfil** (hasta un máximo 80 caracteres) aparece al final del listado para el experimentador y al final de cada perfil del listado para los sujetos.

Figura 6.23. Cuadro de diálogo *Mostrar el diseño: Títulos*

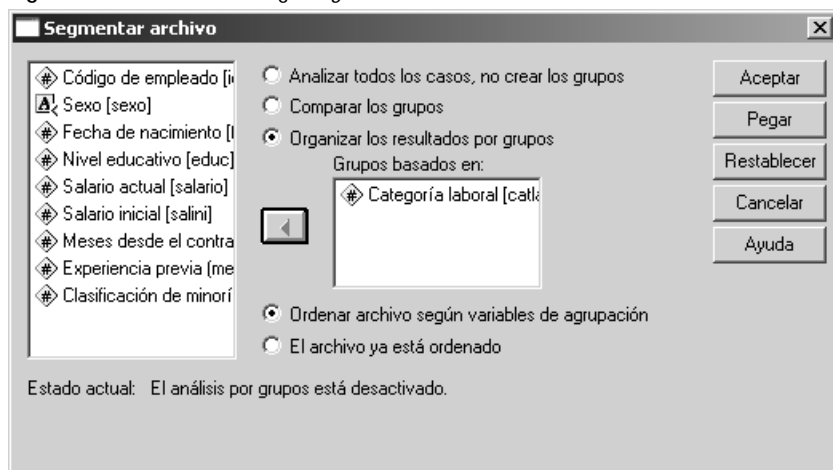


## Segmentar archivo

Segmentar un archivo consiste en dividirlo en subgrupos (esta división es virtual: en el archivo de datos no se aprecia tal división). Los análisis estadísticos que se llevan a cabo mientras un archivo se encuentra segmentado se repiten para cada subgrupo resultante de la segmentación. Para segmentar un archivo:

- Seleccionar la opción **Segmentar archivo...** del menú **Datos** para acceder al cuadro de diálogo *Segmentar archivo* que muestra la Figura 6.24.

Figura 6.24. Cuadro de diálogo *Segmentar archivo*



**Analizar todos los casos, no crear los grupos.** Esta opción está activa mientras no se solicita segmentar el archivo. Una vez segmentado, el archivo permanece segmentado hasta que se activa esta opción, hasta que se realiza una segmentación diferente, o hasta que se inicia una nueva sesión.

**Comparar los grupos.** Activa la segmentación. Si se solicitan varios análisis, el *Visor de resultados* los ofrece organizados de la siguiente manera: el primer análisis para todos los grupos, el segundo análisis para todos los grupos, etc.

**Organizar los resultados por grupos.** Activa la segmentación. Si se solicitan varios análisis, el *Visor de resultados* los ofrece organizados de la siguiente manera: todos los análisis para el primer grupo, todos los análisis para el segundo grupo, etc.

**Grupos basados en.** Para segmentar el archivo de datos es necesario seleccionar una o más variables de segmentación. Para ello:

- Seleccionar la(s) variable(s) en la lista de variables del archivo de datos y trasladarla(s) a la lista **Grupos basados en**.

Las variables de segmentación pueden ser numéricas o de cadena. No es posible seleccionar más de 8 variables (a efectos de este límite, cada 8 caracteres de una variable de cadena larga cuentan como una variable).

**Ordenar archivo según variables de segmentación.** Ordena por segmentos los casos del archivo de datos. Es la opción por defecto. Siempre es conveniente ordenar el archivo de datos por las variables utilizadas en la segmentación. En el caso de que se utilice más de una variable de segmentación, no ordenar el archivo puede conducir a segmentos inconsistentes.

**El archivo ya está ordenado.** Impide que los casos sean reordenados.

Cuando el archivo se encuentra segmentado, la barra de estado de la ventana principal muestra el mensaje *Segmentado*. Si se está trabajando con una pantalla de dimensiones reducidas (menos de 17 pulgadas) y con baja resolución (por debajo de 800×600) es posible que este mensaje se *salga* de la pantalla; es decir, es posible que la parte derecha de la barra de estado del SPSS quede fuera de los límites de la pantalla.

## Seleccionar casos

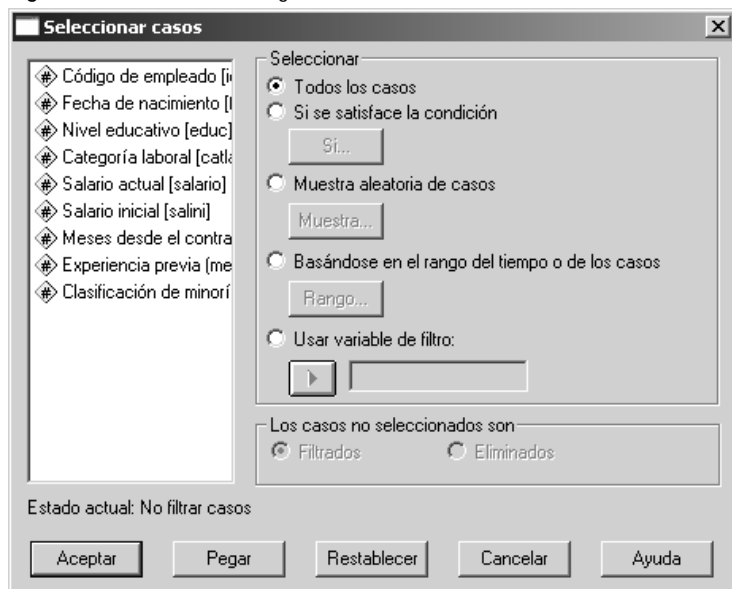
En ocasiones puede interesar centrar el análisis en sólo un grupo de casos que cumplan determinada condición. Otras veces habrá que aplicar determinado tipo de transformaciones a unos sujetos y no a otros. También puede ocurrir que sólo interese analizar una muestra aleatoria del total de casos del archivo de datos.

El SPSS permite seleccionar un conjunto de casos utilizando diferentes criterios: valores o rangos de valores de una variable, números de registro, expresiones aritméticas y lógicas, funciones matemáticas, etc.

La selección de casos es una opción a la que todo usuario del SPSS termina encontrando gran utilidad. Para seleccionar casos:

- Seleccionar la opción **Seleccionar casos...** del menú **Datos** para acceder al cuadro de diálogo *Seleccionar casos* que muestra la Figura 6.25.

Figura 6.25. Cuadro de diálogo *Seleccionar casos*



El recuadro **Seleccionar** contiene todas las opciones de selección de casos disponibles en el SPSS. Puede optarse por cualquiera de las siguientes cinco alternativas:

**Todos los casos.** Selecciona todos los casos del archivo de datos. Es la opción que se encuentra activa por defecto. Cuando la selección de casos se encuentra activa, esta opción la anula.

**Si se satisface la condición.** Permite seleccionar los casos que cumplen una condición, es decir, permite definir *filtros* y aplicarlos al archivo de datos para trabajar únicamente con los casos que cumplen una determinada condición. Para definir una condición:

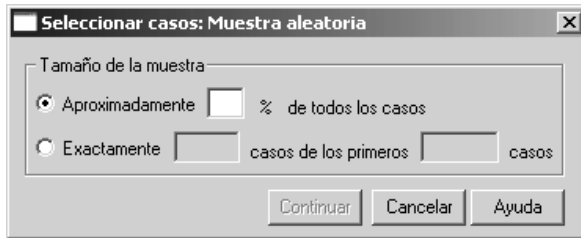
- Pulsar el botón **Si...** (ver Figura 6.25) para acceder al subcuadro de diálogo *Seleccionar casos si*, similar al ya estudiado en el Capítulo 5 (ver Figura 5.4).

Este cuadro de diálogo permite utilizar expresiones condicionales basadas en funciones aritméticas, lógicas, estadísticas, etc. (ver, en el Capítulo 5, los apartados *Calcular: Funciones* y *Calcular: Expresiones condicionales*).

**Muestra aleatoria de casos.** Esta opción selecciona aleatoriamente un porcentaje o un número de casos. Para obtener una muestra aleatoria:

- Pulsar el botón **Muestra...** (ver Figura 6.25) para acceder al subcuadro de diálogo *Seleccionar casos: Muestra aleatoria* que recoge la Figura 6.26.

Figura 6.26. Subcuadro de diálogo *Seleccionar casos: Muestra aleatoria*



**Tamaño de la muestra.** Existen dos formas de obtener una muestra aleatoria:

**Aproximadamente** \_\_\_ % de todos los casos. Selecciona, aproximadamente, el porcentaje de casos indicado.

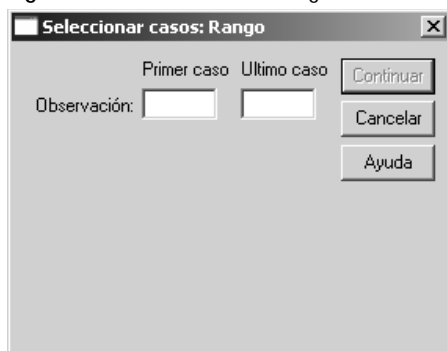
**Exactamente** \_\_\_ casos de los primeros \_\_\_ casos. Selecciona exactamente el número de casos indicado. Selecciona los casos de entre los  $n$  primeros, siendo  $n$  un número definido por el usuario. Este número debe ser menor o igual que el número de casos del archivo de datos; si es mayor, el tamaño de la muestra aleatoria seleccionada será menor que el indicado.

Cada vez que el SPSS genera una muestra aleatoria utiliza una *Semilla de aleatorización* diferente, lo que da lugar a muestras aleatorias diferentes. Si se desea replicar la misma muestra es necesario establecer la misma semilla de aleatorización (ver, en el Capítulo 5, el apartado *Semilla de aleatorización*).

**Basándose en el rango del tiempo o de los casos.** Permite seleccionar un rango de casos a partir del número de registro (fila) que ocupan en el *Editor de datos*. Cuando los casos constituyen una serie temporal y existen variables *fecha*, es posible establecer un rango de fechas u horas para seleccionar casos. Para ello:

- Pulsar el botón **Rango...** para acceder al subcuadro de diálogo *Seleccionar casos: Rango* que muestra la Figura 6.27.

Figura 6.27. Subcuadro de diálogo *Seleccionar casos: Rango*



- Fijar el rango de valores que se desea seleccionar utilizando los cuadros de texto **Primer caso** y **Último caso**.



**Usar variable de filtro.** Esta opción permite utilizar como variable de filtro para efectuar la selección de casos (o para eliminar casos de forma selectiva) cualquier variable numérica del archivo de datos. Al utilizar esta opción, quedan seleccionados todos los casos cuyo valor en la variable de *filtro* es distinto de cero; los casos cuyo valor es cero quedan excluidos.

**Los casos no seleccionados son.** Las opciones de este recuadro permiten decidir qué se desea hacer con los casos no seleccionados:

**Filtrados.** Los casos no seleccionados no son incluidos ni en los procedimientos estadísticos ni en los gráficos, pero permanecen en el archivo de datos. Por tanto, pueden volverse a utilizar durante la misma sesión si se desactiva la selección de casos.

Al seleccionar una muestra aleatoria o al seleccionar casos utilizando una expresión condicional, la opción **Filtrados** crea una variable nueva en el *Editor de datos* llamada *filter\_\$* con «unos» para los casos seleccionados y «ceros» para los no seleccionados. Al efectuar una nueva selección, el contenido de la variable *filter\_\$* previamente creada cambia para reflejar el resultado de la nueva selección. Si se desea guardar la variable *filter\_\$* y utilizarla como variable de filtro en otro momento (por ejemplo, para no tener que volver a definir de nuevo una expresión condicional compleja que ha costado bastante tiempo y trabajo construir), se puede asignar un nombre distinto a la variable *filter\_\$* y guardar el archivo de datos con la nueva variable; de este modo, cuando se cree un nuevo filtro, el filtro previamente creado no será sustituido. Los casos no seleccionados pueden identificarse en el archivo de datos por una línea diagonal (a modo de tachadura) sobre el número de caso, es decir, sobre la cabecera de la fila correspondiente al caso no seleccionado.

**Eliminados.** Los casos no seleccionados son eliminados del archivo de datos. Esto permite reducir el archivo de datos y ganar tiempo de procesamiento, lo cual, con archivos muy grandes, posee cierta utilidad. Los casos eliminados pueden recuperarse abriendo de nuevo el archivo de datos sin haberlo salvado tras la selección (es decir, la eliminación de casos sólo es permanente si, una vez eliminados los casos no seleccionados, se guarda el archivo de datos sin cambiarle el nombre).

Cuando la selección de casos se encuentra activa, la barra de estado de la ventana principal muestra el mensaje *Filtrado*. Si se está trabajando con una pantalla de dimensiones reducidas (menos de 17 pulgadas) y con baja resolución (por debajo de 800x600) es posible que este mensaje se *salga* de la pantalla; es decir, es posible que la parte derecha de la barra de estado del SPSS quede fuera de los límites de la pantalla.

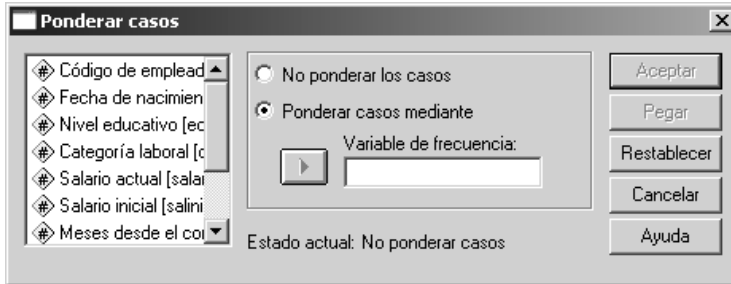
## Ponderar casos

En un archivo de datos estándar, cada registro corresponde a un caso. Ponderar casos consiste en hacer que un registro (caso) represente a más (o a menos) de un caso. El resultado de la ponderación es justamente el inverso de la agregación. Este apartado contiene ejemplos que ayudan a comprender el significado de la ponderación.

Ponderar casos exige utilizar una *variable de ponderación* que es justamente la que contiene los pesos que serán asignados a cada caso. Para ponderar casos:

- Seleccionar la opción **Ponderar casos...** del menú **Datos** para acceder al cuadro de diálogo *Ponderar casos* que muestra la Figura 6.28.

Figura 6.28. Cuadro de diálogo *Ponderar casos*



La lista de variables del archivo de datos muestra únicamente las variables que poseen formato numérico (no es posible ponderar casos utilizando variables de cadena). Para seleccionar una variable de ponderación:

- Activar la opción **Ponderar casos mediante**.
- Seleccionar en la lista de variables del archivo de datos la variable que se desea utilizar como variable de ponderación y trasladarla al cuadro **Variable de frecuencia**.

Una vez activada la ponderación, ésta permanece activa hasta que se cambia de variable de ponderación o hasta que se marca la opción **No ponderar casos**.

La parte inferior del cuadro de diálogo, **Estado actual**, informa sobre si la ponderación está activa o no. Además, cuando la ponderación se encuentra activa, la barra de estado de la ventana principal muestra el mensaje *Ponderado*. Si se está trabajando con una pantalla de dimensiones reducidas (menos de 17 pulgadas) y con baja resolución (por debajo de 800x600 píxeles) es probable que este mensaje se *salga* de la pantalla; es decir, es probable que la parte derecha de la barra de estado del SPSS quede fuera de los límites de la pantalla.

La ponderación de casos posee una utilidad especial relacionada con la reproducción de tablas de contingencias. Una tabla de contingencias es una forma concreta de organizar las frecuencias conjuntas resultantes de cruzar dos variables categóricas (para una aclaración del concepto de *tabla de contingencias* y del tipo de análisis que es posible efectuar en ellas, puede consultarse el Capítulo 12). La Tabla 6.4 muestra un ejemplo de tabla de contingencias en la que se han cruzado las variables *sexo* (filas) y *tabaco* (columnas).

Tabla 6.4. Tabla de contingencias de *sexo* por *tabaco*

<b>Sexo</b>	<b>Tabaco</b>			<b>Totales</b>
	<i>Fumadores</i> (1)	<i>No fumadores</i> (2)	<i>Exfumadores</i> (3)	
<i>Varones</i> (1)	360	600	240	1200
<i>Mujeres</i> (2)	340	300	160	800
<i>Totales</i>	700	900	400	2000

Esta tabla se ha obtenido clasificando una muestra de 2.000 casos en las variables *sexo* (varones, mujeres) y *tabaco* (fumadores, no fumadores, exfumadores). Los valores de cada casilla reflejan el número de casos que cumplen las condiciones de las correspondientes filas y columnas. Por ejemplo, el valor 360 de la primera casilla (1, 1) significa que, en la muestra de 2.000 personas, 360 son *varones fumadores*; el valor de la segunda casilla (1, 2) significa que 300 personas son *varones no fumadores*; etc.

Si se desea reproducir los datos de esta tabla en el *Editor de datos* para efectuar los análisis pertinentes, no es necesario crear un archivo de datos con 2.000 casos. Gracias a la posibilidad de ponderar casos sólo es necesario crear un número de casos igual al número de casillas que contiene la tabla. La Figura 6.29 muestra cómo hacerlo.

**Figura 6.29.** Datos de la Tabla 6.4 reproducidos en el *Editor de datos* (izquierda: valores; derecha: etiquetas)

	sexo	tabaco	ncasos		sexo	tabaco	ncasos
1	1	1	360	1	Varones	Fumadores	360
2	1	2	600	2	Varones	No fumadores	600
3	1	3	240	3	Varones	Exfumadores	240
4	2	1	340	4	Mujeres	Fumadores	340
5	2	2	300	5	Mujeres	No fumadores	300
6	2	3	160	6	Mujeres	Exfumadores	160

Puesto que la Tabla 6.4 tiene 6 casillas, el archivo de datos incluye 6 casos. La parte izquierda de la Figura 6.29 muestra los valores; la parte derecha muestra las etiquetas de los valores. Para las categorías de la variable *sexo* se han utilizado los valores 1 y 2 con las siguientes etiquetas: 1 = «varones» y 2 = «mujeres»; y para las categorías de la variable *tabaco* se han utilizado los valores 1, 2 y 3, con las siguientes etiquetas: 1 = «fumadores», 2 = «no fumadores» y 3 = «exfumadores».

Para que los 6 casos del archivo de la Figura 6.29 puedan convertirse en los 2.000 de la Tabla 6.4, además de las variables *sexo* y *tabaco* es necesario crear una tercera variable con las frecuencias de cada casilla (en el ejemplo, esta nueva variable se ha nombrado *ncasos*).

Hecho esto, para reproducir la Tabla 6.4 en el *Visor de resultados* a partir de los datos que muestra el *Editor de datos* de la Figura 6.29:

- Seleccionar la opción **Ponderar casos...** del menú **Datos** para acceder al cuadro de diálogo *Ponderar casos* (ver Figura 6.28).
- Marcar la opción **Ponderar casos mediante**, trasladar la variable *ncasos* al cuadro **Variable de frecuencia** y pulsar el botón **Aceptar**.
- Seleccionar la opción **Estadísticos descriptivos > Tablas de contingencias...** del menú **Analizar** para acceder al cuadro de diálogo *Tablas de contingencias*.
- Seleccionar la variable *sexo* y trasladarla al recuadro **Fila**; seleccionar la variable *tabaco* y trasladarla al recuadro **Columna**.

Aceptando estas elecciones, el *Visor de resultados* ofrece la tabla de contingencias que muestra la Tabla 6.5. Puede comprobarse que esta tabla es idéntica a la Tabla 6.4. A pesar de que el archivo de datos sólo contiene 6 casos, al ponderar el archivo mediante la variable *ncasos*, los 6 casos del archivo reproducido en la Figura 6.29 se han convertido en los 2.000 de las Tablas 6.4 y 6.5.

**Tabla 6.5.** Tabla de contingencias de *sexo* por *tabaco*

Recuento

		Tabaco			Total
		Fumadores	No fumadores	Exfumadores	
Sexo	Varones	360	600	240	1200
	Mujeres	340	300	160	800
Total		700	900	400	2000



## Archivos de resultados: el *Visor*

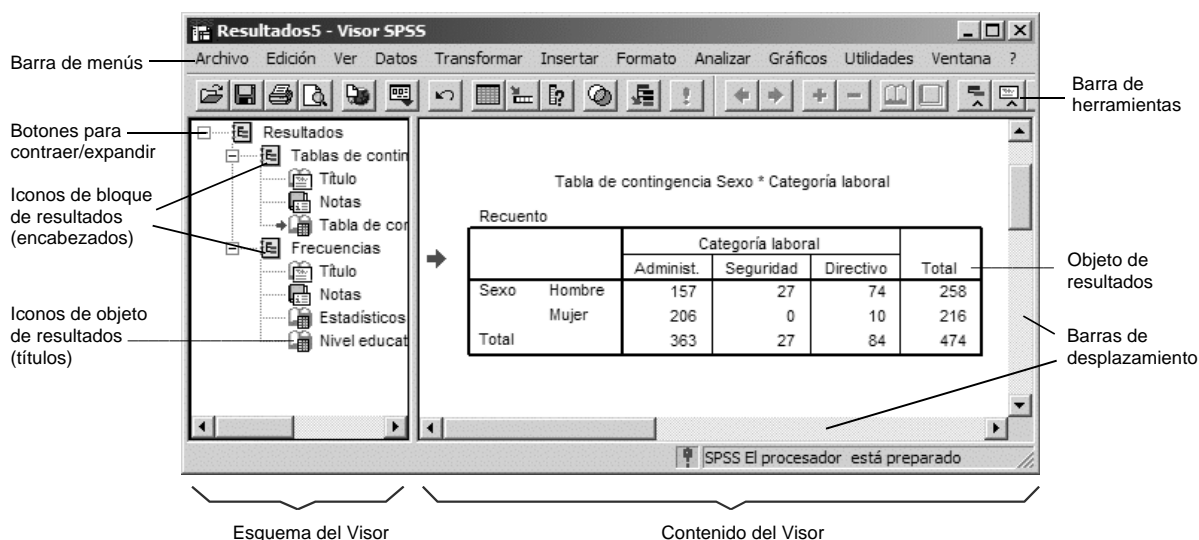
Los archivos de resultados se encargan de recoger toda la información que el SPSS genera (tablas, gráficos, texto, etc.) como consecuencia de las acciones que lleva a cabo. Estos archivos de resultados se muestran en una ventana llamada *Visor* que permite aplicar una amplia variedad de acciones de edición sobre los resultados, moverse fácilmente por ellos, guardarlos para su uso posterior, imprimirlos, exportarlos a una aplicación externa, etc.

### El Visor de resultados

El SPSS abre automáticamente una ventana del *Visor* de resultados la primera vez que se ejecuta un procedimiento que requiere de la presentación de resultados. Esta ventana recibe, por defecto, el nombre *Resultado1*.

La ventana del *Visor* (ver Figura 7.1) se encuentra dividida verticalmente en dos paneles: el *Esquema* del *Visor* y el *Contenido* del *Visor*.

Figura 7.1. Ventana del *Visor de resultados*.



El panel izquierdo, al que puede llamarse *Esquema del Visor*, muestra un índice con los *titulares* correspondientes a los resultados generados por el SPSS. Estos *titulares* son de dos clases: *encabezados* y *títulos*. Los *encabezados* van precedidos de un icono amarillo y sirven para indicar el comienzo de un *bloque de resultados* (un bloque de resultados es el conjunto de resultados generados por un único procedimiento SPSS). Los encabezados no están asociados a un resultado concreto, sino a todo un bloque. Los *títulos* cuelgan de los *encabezados* y van precedidos de iconos en forma de libro (abierto o cerrado, según se verá más adelante). Cada título se corresponde con un resultado concreto. Puede cambiarse el tamaño de este panel pinchando con el puntero del ratón sobre su borde derecho y arrastrándolo.

El panel derecho muestra el *Contenido del Visor*, es decir, los resultados generados por el SPSS. Cada *bloque de resultados* consta de un conjunto de *objetos* (títulos, notas aclaratorias, tablas, gráficos, advertencias, etc.). A cada uno de esos *objetos* le corresponde un *título* (una línea de texto) en el *Esquema del Visor*. Los distintos *objetos de resultados* del *Visor* adoptan tres formatos básicos: texto, tablas y gráficos; el SPSS incluye, para cada uno de estos tres formatos, un editor que permite efectuar todo tipo de modificaciones.

La versión *borrador* del *Visor* tiene aspecto de procesador de texto (la ventana no está partida en dos paneles) y, según se ha señalado ya, muestra los resultados en formato texto. Para trabajar con el *Visor* en formato borrador, seleccionar Edición > Opciones... y, en el recuadro Tipo de resultados al arrancar, marcar la opción Visor de borrador.

## El menú Archivo

La mayor parte de las opciones que contiene el menú Archivo del *Visor* son idénticas a las del menú Archivo del *Editor de datos* (ver Capítulo 3). Pero el menú Archivo del *Visor* contiene algunas opciones propias no incluidas en el menú Archivo del *Editor de datos*: Cerrar, Guardar con contraseña, Enviar mensaje, Exportar y Preparar página. Todas ellas se tratan en este capítulo. En este mismo apartado se explican las tres primeras. La opción Exportar se trata al final del capítulo. Y la opción Preparar página se incluye dentro del apartado *Imprimir resultados*.

### Cerrar

Esta opción permite cerrar el *Visor* de resultados sin salir del SPSS (es decir, sin cerrar el *Editor de datos*). Al igual que ocurre en otras aplicaciones Windows, si el archivo del *Visor* tiene nombre y no ha sufrido modificaciones, la opción Cerrar del menú Archivo cierra el *Visor*. Si el archivo del *Visor* tiene ya nombre pero ha sufrido alguna modificación, o no tiene nombre (es decir, tiene el nombre provisional que el sistema asigna por defecto: *Resultado#*), la opción Cerrar abre un cuadro de diálogo de *advertencia* preguntando si se desea o no guardar las modificaciones hechas (a los archivos de resultados se les asigna, por defecto, la extensión *.spo*).

### Enviar mensaje

La opción Enviar mensaje abre automáticamente el programa de correo electrónico que el sistema reconoce por defecto y adjunta el archivo de resultados como un *attach*.

## Guardar con contraseña

La opción **Guardar con contraseña...** del menú **Archivo** conduce a un sencillo cuadro de diálogo que permite asignar una contraseña al guardar un archivo del *Visor*. El archivo guardado con contraseña sólo podrá abrirse si se introduce la contraseña correcta.

## Editar resultados

El *Visor* ofrece múltiples posibilidades de edición. Es posible seleccionar todos los resultados o parte de ellos, moverlos, copiarlos, borrarlos, cambiar su aspecto, mostrarlos, ocultarlos, añadir objetos, insertar saltos de página, etc.

## Seleccionar resultados

El *Visor* ofrece varios métodos para llevar a cabo selecciones totales (de todo el contenido del *Visor*) o parciales (de una o varias tablas, uno o varios gráficos, una tabla y un gráfico, etc.) de los resultados.

Utilizando el **puntero del ratón**:

- Para seleccionar *un objeto* del *Visor* (un cuadro de texto, una tabla o un gráfico), pulsar sobre él con el puntero del ratón.
- Para seleccionar *varios objetos contiguos*, pulsar sobre el primero de ellos con el puntero del ratón y, manteniendo pulsada la tecla *Mayúsculas*, pulsar sobre el último.
- Para seleccionar *varios objetos no contiguos*, pulsar sobre ellos con el puntero del ratón manteniendo pulsada la tecla *Control*.

Utilizando la **barra de menús**:

- La opción **Seleccionar todo** del menú **Edición** (teclado: *Control + S*) permite seleccionar todos los contenidos del *Visor*.
- La opción **Seleccionar** del menú **Edición** ofrece varias posibilidades de selección parcial de resultados (básicamente agrupados por tipo de resultados: títulos, notas, tablas, gráficos, advertencias, etc.).

También es posible seleccionar resultados (y, al mismo tiempo, desplazar el cursor) utilizando el **Esquema del Visor**:

- Pulsando con el puntero del ratón sobre el icono de un *objeto* de resultados o sobre su título, el cursor se desplaza a ese objeto y lo selecciona.
- Pulsando con el puntero del ratón sobre el icono de un *bloque* de resultados o sobre su encabezado, queda seleccionado todo el bloque.
- Pulsando con el puntero del ratón sobre el primer encabezado del *Esquema del Visor* (*Resultados*), quedan seleccionados todos los contenidos del *Visor*.



## Mover, copiar y borrar resultados

Los contenidos de un archivo de resultados y el orden en el que aparecen en el *Visor* pueden ser fácilmente alterados utilizando el puntero del ratón o algunas de las opciones del menú Edición.

Para **mover** un objeto pueden seguirse dos procedimientos:

- Seleccionar el objeto o conjunto de objetos que se desea mover.
- Arrastrarlo con el puntero del ratón hasta el lugar deseado (justo sobre el objeto detrás del cual se desea colocar).

O bien:

- Seleccionar el objeto o conjunto de objetos que se desea mover.
- Seleccionar la opción **Cortar** del menú **Edición** (teclado: *Control* + *X*).
- Situar el cursor en el punto detrás del cual se desea colocar el objeto.
- Seleccionar la opción **Pegar después** del menú **Edición** (teclado: *Control* + *V*).

Para **copiar** un objeto pueden seguirse dos procedimientos:

- Seleccionar el objeto o conjunto de objetos que se desea copiar.
- Arrastrarlo con el puntero del ratón hasta el lugar deseado (justo sobre el objeto detrás del cual se desea colocarlo) mientras se mantiene pulsada la tecla *Control*.

O bien:

- Seleccionar el objeto o conjunto de objetos que se desea copiar.
- Seleccionar la opción **Copiar** del menú **Edición** (teclado: *Control* + *C*).
- Situar el cursor en el punto detrás del cual se desea copiar el objeto.
- Seleccionar la opción **Pegar después** del menú **Edición** (teclado: *Control* + *V*).

Para **borrar** un objeto:

- Seleccionar el objeto o conjunto de objetos que se desea borrar.
- Seleccionar la opción **Eliminar** del menú **Edición** (o pulsar la tecla *Supr*).

## Cambiar de nivel un titular

Al mover un objeto o un bloque de resultados de un lugar a otro dentro del *Visor* puede ocurrir que el nivel de su titular dentro del *Esquema* del *Visor* se vea alterado. Este problema puede corregirse fácilmente pues tanto la barra de menús como la barra de herramientas contienen opciones que permiten cambiar el nivel de un titular. No obstante, estas opciones no afectan a todos los titulares del *Esquema*. Sólo es posible cambiar de nivel algunos títulos y encabezados.

Para **ascender** un titular:

- Seleccionar el titular en el *Esquema* del *Visor*.
- Seleccionar la opción **Ascender** del menú **Edición**, o pulsar el botón *Ascender* (flecha apuntando hacia la izquierda) de la barra de herramientas.

Para **degradar** un titular:

- Seleccionar el titular en el *Esquema* del *Visor*.
- Seleccionar la opción **Degradar** del menú **Edición**, o pulsar el botón *Degradar* (flecha apuntando hacia la derecha) de la barra de herramientas.

## Mostrar y ocultar resultados

El *Visor de resultados* permite mostrar y ocultar selectivamente objetos y bloques de resultados. Ocultar parte de los resultados resulta útil para reducir (sin eliminar) la cantidad de contenidos del *Visor* o para imprimir sólo determinadas partes del *Visor*.

Para **mostrar/ocultar un objeto** (una tabla, un gráfico, etc.) del *Contenido* del *Visor*:

- Seleccionar el objeto y utilizar la opción **Ocultar** o **Mostrar** del menú **Ver**.
- Se consigue el mismo efecto si, tras seleccionar un objeto, se pulsa el botón *Mostrar* (libro abierto) o el botón *Ocultar* (libro cerrado) de la barra de herramientas.
- También se consigue el mismo efecto pulsando dos veces con el puntero del ratón sobre el icono del objeto en el *Esquema* del *Visor* (el icono de un objeto tiene forma de libro y aparece abierto o cerrado para indicar si el contenido asociado a él se encuentra en modo **Mostrar** o en modo **Ocultar**).

Para **mostrar/ocultar un bloque de resultados** del *Contenido* del *Visor*:

- Seleccionar el bloque y utilizar la opción **Ocultar** o **Mostrar** del menú **Ver**.
- Se consigue el mismo efecto si, tras seleccionar un bloque, se pulsa el botón *Mostrar* (libro abierto) o el botón *Ocultar* (libro cerrado) de la barra de herramientas.

Para **mostrar/ocultar un bloque de resultados** del *Contenido* del *Visor* y, al mismo tiempo, **expandir/contraer los títulos** que cuelgan del encabezado de ese bloque en el *Esquema* del *Visor*:

- Seleccionar el bloque y utilizar la opción **Contraer** (o **Expandir**) del menú **Ver**.
- Se consigue el mismo efecto pulsando dos veces con el puntero del ratón sobre el icono del bloque (en el *Esquema* del *Visor*).
- Se consigue el mismo efecto pulsando el signo «+» (para expandir) y el signo «-» (para contraer) situados delante del icono del bloque (en el *Esquema* del *Visor*).
- Se consigue el mismo efecto pulsando el signo «+» (para expandir) y el signo «-» (para contraer) situados en la barra de herramientas.

El estado inicial de la acción mostrar/ocultar puede controlarse desde la opción **Edición > Opciones**. Entre las opciones relacionadas con el *Visor* existe una referida a qué objetos, por defecto, serán mostrados/ocultados cuando el SPSS genere resultados.

## Tamaño y fuente de los titulares

El menú **Ver** contiene dos opciones para controlar el tipo y el tamaño de la letra con la que aparecen los titulares del *Esquema* del *Visor*.

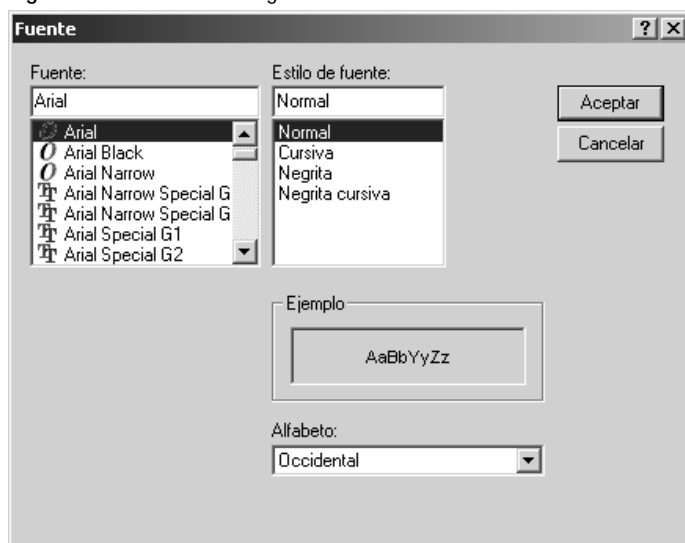
Para **cambiar el tamaño de la letra** de los titulares del *Esquema* del *Visor*:

- Seleccionar la opción **Tamaño de los titulares** del menú **Ver**.
- Seleccionar **Pequeño** (es el valor por defecto), **Mediano** o **Grande** en el menú desplegable.

Para **cambiar el tipo de letra** del *Esquema* del *Visor*:

- Seleccionar la opción **Fuente de los titulares** del menú **Ver** para acceder al cuadro de diálogo *Fuentes* que muestra la Figura 7.2. Desde este cuadro de diálogo es posible seleccionar el tipo de letra deseado y su aspecto.

Figura 7.2. Cuadro de diálogo *Fuentes*



## Saltos de página

Al imprimir el contenido del *Visor* el SPSS controla automáticamente el contenido de cada página. En ocasiones puede resultar útil insertar saltos de página manuales para decidir dónde debe empezar una nueva página.

Para **insertar** un salto de página:

- ' Seleccionar el objeto sobre el que se desea insertar el salto de página.
- ' Seleccionar la opción **Salto de página** del menú **Insertar**.

Para **eliminar** un salto de página:

- ' Seleccionar el objeto sobre el que se encuentra el salto de página.
- ' Seleccionar la opción **Eliminar salto de página** del menú **Insertar**.

## Insertar texto y gráficos

El menú **Insertar** contiene opciones para insertar texto y gráficos en cualquier punto del panel de *Contenidos del Visor*. También es posible insertar líneas (encabezados) en el *Esquema del Visor*.

Para **insertar texto**:

- ' Situar el cursor en el objeto detrás del cual se desea insertar texto.
- ' Seleccionar la opción **Nuevo título**, **Nuevo título de página** o **Nuevo texto** del menú **Insertar**.

Las opciones **Nuevo título** y **Nuevo texto** tienen idéntico efecto: insertan un título en el *Esquema del Visor* (justo a continuación del título o encabezado seleccionado) y abren un cuadro de texto en el *Contenido del Visor*. En ese momento el cursor se encuentra situado en un cuadro de texto en el que es posible escribir con el teclado. La única diferencia entre ambas opciones está en las características de la fuente utilizada por defecto.

La opción **Nuevo título de página** también inserta un título en el *Esquema* y un cuadro de texto en los *Contenidos*, pero inserta un salto de página y coloca el cuadro de texto como cabecera de página.

Para **insertar texto en un encabezado nuevo**:

- ' Situar el cursor en el objeto debajo del cual se desea insertar el nuevo encabezado.
- ' Seleccionar la opción **Nuevo encabezado** del menú **Insertar**.
- ' Seleccionar la opción **Nuevo título**, o **Nuevo título de página**, o **Nuevo texto** del menú **Insertar** e introducir el texto deseado.

La opción **Nuevo encabezado** inserta un encabezado en el *Esquema del Visor*, pero no tiene ningún efecto sobre el *Contenido del Visor*. Ahora bien, el nuevo título o el nuevo texto estarán colgando del nuevo encabezado.

Para **insertar un gráfico**:

- ' Situar el cursor en el objeto del *Visor* debajo del cual se va a insertar el nuevo gráfico.
- ' Seleccionar la opción **Gráfico 2D interactivo** o **Gráfico 3D interactivo** para insertar la plantilla de un gráfico interactivo.

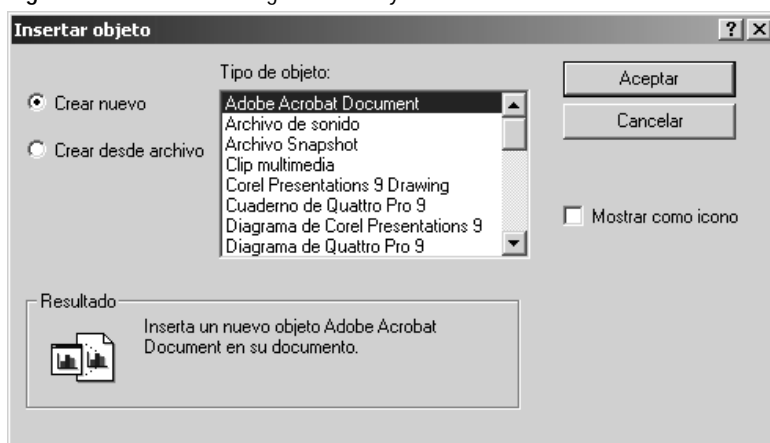
Para **insertar un archivo de texto** creado con el editor de sintaxis del SPSS o con cualquier otro procesador de textos, pero guardado en formato *texto*:

- Situar el cursor en el objeto del *Visor* a partir del cual se desea insertar el nuevo archivo de texto.
- Seleccionar la opción **Archivo de texto...** del menú **Insertar**. Esta opción conduce a un cuadro de diálogo idéntico al cuadro *Abrir* de la Figura 3.1 (ver Capítulo 3), con la diferencia de que, ahora, los archivos listados en este cuadro de diálogo tienen extensión *.lst* (que es la extensión que las versiones anteriores del SPSS asignan a los archivos de resultados; los archivos de resultados de las versiones del SPSS anteriores a la 7.5 se guardan en formato de texto ASCII). Cambiando la extensión se puede obtener un listado de los archivos con la extensión buscada.
- Seleccionar el nombre del archivo que contiene el texto que se desea insertar y pulsar el botón **Abrir**. Una vez insertado el texto, es posible editarlo con las funciones de edición del *Visor*.

Para **insertar un objeto** (un archivo de sonido, una imagen, una ecuación, un gráfico de otra aplicación, una secuencia MIDI, etc.):

- Situar el cursor en el objeto del *Visor* a partir del cual se desea insertar el nuevo objeto.
- Seleccionar la opción **Objeto...** del menú **Insertar** para acceder al cuadro de diálogo *Insertar objeto* que muestra la Figura 7.3.

Figura 7.3. Cuadro de diálogo *Insertar objeto*





Las opciones de este cuadro de diálogo permiten crear un objeto nuevo o insertar el contenido de un archivo ya existente. Si se marca la opción **Mostrar como icono**, el *Visor* muestra un icono que permite acceder al objeto insertado. Por ejemplo, si se inserta un documento de WordPerfect, el *Visor* abre un cuadro de texto para el documento. Pero si el documento se inserta como un icono, el *Visor* únicamente muestra un icono de acceso al documento de WordPerfect.

## Alinear resultados

Todos los objetos del *Visor* aparecen alineados, por defecto, a la izquierda. Las opciones del menú **Formato** permiten cambiar el alineamiento de todos o parte de los objetos. Para cambiar el alineamiento de un objeto o de un bloque de objetos:

- Seleccionar el objeto o bloque de objetos cuyo alineamiento se desea cambiar.
- En el menú **Formato**, seleccionar una de las tres opciones disponibles: **Alinear a la izquierda**, **Centrar**, o **Alinear a la derecha**.

Estos cambios sólo afectan a los resultados cuando se imprimen. No obstante, aunque en el *Visor* todos los objetos se muestran alineados a la izquierda, cuando se altera el alineamiento de un objeto, aparece un símbolo delante de ese objeto: el símbolo «» indica que el objeto está centrado; y el símbolo «» indica que el objeto está alineado a la derecha.

## Editar tablas

Los objetos del *Visor de resultados* adoptan, según se ha explicado ya, tres tipos de formato: texto, tablas y gráficos. Pero la mayor parte de los objetos adoptan formato de *tabla*. El SPSS ofrece, a través del *Editor de tablas*, múltiples posibilidades de edición. Quizá la más importante de las funciones de edición del *Editor de tablas* sea su capacidad para reordenar de forma rápida y sencilla las filas, columnas y capas de las tablas. De ahí que las tablas del SPSS reciban el nombre de **tablas pivotantes**. Las opciones de edición para las tablas pivotantes se encuentran en el *Editor de tablas*.

Para **editar una tabla** con el *Editor de tablas*:

- Situar el cursor sobre la tabla y pulsar el botón secundario del ratón.
- En el menú emergente que aparece, seleccionar la opción **Objeto tabla pivote de SPSS > Editar**.
- Se obtiene el mismo resultado colocando el puntero del ratón sobre la tabla y pulsando dos veces el botón principal.

Para **editar más de una tabla** al mismo tiempo:

- Situar el cursor sobre la tabla y pulsar el botón secundario del ratón.
- En el menú emergente que aparece, seleccionar la opción **Objeto tabla pivote de SPSS > Abrir**.
- Repetir la acción para cada una de las tablas que se desee editar.

Las tablas pueden editarse en la propia ventana del *Visor* o en una ventana distinta para cada tabla. El recuadro **Modo de edición por defecto** del menú **Edición > Opciones... > Tablas pivote** permite seleccionar el tipo de edición preferido. Una vez dentro del *Editor de tablas* se tiene acceso a un conjunto de funciones de edición no disponibles en otras ventanas. Algunos menús del *Editor de tablas* no cambian respecto de los ya estudiados (como, por ejemplo, los menús **Archivo**, **Analizar** y **Gráficos**); pero otros adoptan importantes peculiaridades cuya descripción constituye el objetivo de los apartados que siguen.

## La barra de herramientas del Editor de tablas

La barra de herramientas del *Editor de tablas* (ver Figura 7.4) contiene unos cuantos botones-íconos con algunas funciones de edición especialmente útiles para la edición de tablas: deshacer la última acción de edición, activar/desactivar los paneles de pivotado, cambiar el tipo y el tamaño de la letra, los atributos de letra (negrita, cursiva, subrayado, color) y el tipo de justificación del texto dentro de las casillas.

Figura 7.4. Barra de herramientas del *Editor de tablas*



Al abrir la barra de herramientas del *Editor de tablas*, el SPSS le asigna el nombre *Barra de herramientas de formato#*, donde # es un número que indica el número de orden que ocupa la tabla que se está editando en el conjunto de las tablas editadas en una sesión.

Para **activar/desactivar** la barra de herramientas:

- Seleccionar la opción **Barra de herramientas** del menú **Ver**.

## Seleccionar

- Para seleccionar una casilla, situar el cursor en ella.
- Para seleccionar más de una casilla, mover el cursor manteniendo pulsada la tecla *Mayúsculas*, o pinchar y arrastrar con el puntero del ratón.

Además de estas formas básicas de selección, la opción **Seleccionar** del menú **Edición** ofrece otras posibilidades. Para **seleccionar toda la tabla**, incluyendo el **texto** y las **líneas**:

- Situar el cursor en cualquier casilla de la tabla.
- Seleccionar la opción **Seleccionar > Tabla** del menú **Edición**. Se obtiene el mismo efecto pulsando simultáneamente las teclas *Control + T*.

Para seleccionar **toda la tabla**, pero **sólo el texto** (sin las líneas):

- Situar el cursor en cualquier casilla de la tabla.
- Seleccionar la opción **Seleccionar > Cuerpo de tabla** del menú **Edición**.

Para seleccionar **los datos de una o varias filas** (o de una o varias columnas):

- Situar el cursor en la cabecera de la fila o filas deseadas (o en la cabecera de la columna o columnas deseadas). Para seleccionar más de una fila o columna, utilizar el cursor junto con la tecla *Mayúsculas*.
- Seleccionar la opción **Seleccionar > Casillas de datos** del menú **Edición**. Se obtiene el mismo efecto pinchando con el puntero del ratón en la cabecera de la fila o de la columna mientras se mantiene pulsada la tecla *Alt*.

Para seleccionar *los datos y las etiquetas de una o varias filas* (o de una o varias columnas):

- Situar el cursor en la cabecera de la fila o filas que se desea seleccionar (o en la cabecera de la columna o columnas que se desea seleccionar). Para seleccionar más de una fila o columna, utilizar el cursor junto con la tecla *Mayúscula*.
- Seleccionar la opción **Seleccionar > Casillas de datos y etiquetas** del menú **Edición**. Se obtiene el mismo efecto pinchando con el puntero del ratón en la cabecera de la fila o de la columna mientras se mantienen pulsadas las teclas *Control* y *Alt*.

## Agrupar y desagrupar casillas

La opción **Agrupar** del menú **Edición** permite agrupar varias filas o columnas en una sola categoría (fila o columna). A la nueva categoría fruto de la agrupación se le puede asignar la etiqueta deseada. Las filas (o columnas) previamente agrupadas en una categoría pueden desagruparse (pueden devolverse a su estado original) mediante la opción **Desagrupar** del menú **Edición**.

Para **agrupar** varias filas o columnas:

- Seleccionar las cabeceras de las filas (o de las columnas) que se desea agrupar.
- Seleccionar la opción **Agrupar** del menú **Edición**.

Para **desagrupar** filas o columnas previamente agrupadas:

- Seleccionar la cabecera de las filas (o de las columnas) agrupadas.
- Seleccionar la opción **Desagrupar** del menú **Edición**.

## Mostrar y ocultar casillas

La mayor parte de las opciones del menú **Ver** permiten mostrar y ocultar diferentes partes (casillas, filas, columnas, etiquetas, etc.) del contenido de una tabla pivotante.

Para **ocultar filas o columnas**:

- Seleccionar la fila o columna que se desea ocultar (ver, en este mismo capítulo, el apartado *Seleccionar resultados*).
- Utilizar la opción **Ocultar** del menú **Ver**. Se consigue el mismo efecto utilizando la opción **Ocultar categoría** del menú emergente que se obtiene al pulsar el botón secundario del ratón.

Para **mostrar filas o columnas** previamente ocultas:

- Situar el cursor en una fila o columna de la misma dimensión que la fila o columna que se desea mostrar.
- En el menú **Ver**, seleccionar cualquiera de estas dos opciones: **Mostrar todo** o **Mostrar todas las categorías de *nombre-de-variable***.



Para **mostrar/ocultar la etiqueta de una dimensión**:

- Situar el cursor en la etiqueta de la dimensión que se desea mostrar/ocultar, o en cualquiera de las etiquetas de las categorías de esa dimensión.
- Seleccionar la opción **Mostrar (Ocultar) etiqueta de dimensión** del menú **Ver**.

Para **mostrar/ocultar las etiquetas de las categorías de una dimensión**:

- Activar los paneles de pivotado seleccionando **Paneles de pivotado** en el menú **Pivotar** (ver, más adelante, el apartado *Paneles de pivotado* y la Figura 7.5).
- Situar el puntero del ratón sobre el icono correspondiente a la dimensión cuyas categorías se desea mostrar/ocultar y pulsar el botón secundario del ratón.
- Seleccionar, en el menú emergente que se obtiene, la opción **Ocultar todas las etiquetas de las categorías**.

Para **ocultar notas a pie** de tabla:

- Situar el cursor sobre la nota que se desea ocultar.
- Utilizar la opción **Ocultar** del menú **Ver**.

Para **mostrar notas a pie** de tabla previamente ocultas:

- Utilizar la opción **Mostrar todas las notas a pie** del menú **Ver**.

Para **mostrar/ocultar líneas divisorias** entre todas las casillas de la tabla:

- Utilizar la opción **Cuadrícula** del menú **Ver**.

## Modificar y añadir texto

El *Editor de tablas* permite modificar total o parcialmente el texto existente (cabeceras, etiquetas, datos, etc.) y añadir nuevo texto en las casillas que se encuentran vacías o en los encabezados y pies de las tablas.

Para **modificar total o parcialmente el texto de una casilla**, o para **añadir texto en una casilla vacía**:

- Situar el puntero del ratón en la casilla cuyo texto se desea modificar, pulsar dos veces el botón principal del ratón (o situar el cursor en esa casilla y pulsar la tecla **F2**) e introducir desde el teclado las modificaciones deseadas.
- Pulsar la tecla **Intro** para aceptar las modificaciones.
- Pulsar la tecla **Escape** para cancelar las modificaciones hechas y abandonar la casilla dejándola como estaba.

Si se desea introducir texto *fuera de las casillas* (es decir, en espacios nuevos), el *Editor de tablas* ofrece tres posibilidades de inserción de texto: títulos, notas a pie de tabla y texto a pie de tabla.

Para *añadir un título*:

- Situar el cursor en cualquier parte dentro de la tabla.
- Utilizar la opción **Título** del menú **Ver** (esta opción sólo está disponible si la tabla no tiene título).

Para *añadir texto a pie de tabla*:

- Situar el cursor en cualquier parte de la tabla.
- Utilizar la opción **Texto al pie** del menú **Ver** (esta opción no está disponible si ya existe un cuadro de texto a pie de tabla).

Para *añadir una nota a pie de tabla*:

- Situar el cursor en cualquier parte de la tabla, excepto en una nota a pie de tabla.
- Utilizar la opción **Nota al pie** del menú **Ver**.

## Pivotar tablas

El *Editor de tablas* posee la capacidad de reordenar de varias maneras y de forma sencilla las dimensiones (filas, columnas y capas) de las tablas pivotantes. A esta capacidad de reorganización de las dimensiones se le llama *pivotado*.

Este apartado describe las posibilidades que ofrece el pivotado de tablas; pero antes es necesario aclarar qué es una tabla pivotante. Una tabla pivotante es un conjunto de información (encabezados, etiquetas, datos, etc.) organizada en *filas*, *columnas* y, opcionalmente, *capas*. La Tabla 7.1 muestra un ejemplo de tabla pivotante con tres criterios de clasificación o dimensiones: la variable *categoría laboral* (con tres categorías o niveles: administrativo, seguridad y directivo), la variable *sexo* (con dos categorías o niveles: hombre y mujer) y la dimensión *estadísticos* (con dos categorías o niveles: media y *N*).

**Tabla 7.1.** Tabla pivotante con tres *dimensiones*: dos dimensiones definidas por las variables *categoría laboral* (filas) y *sexo* (columnas) y una tercera dimensión de *estadísticos* (*Media* y *N*)

Salario actual		Sexo		
Categoría laboral		Hombre	Mujer	Total
Administrativo	Media	31.558,15	25.003,69	27.838,54
	N	157,00	206,00	363,00
Seguridad	Media	30.938,89		30.938,89
	N	27,00		27,00
Directivo	Media	66.243,24	47.213,50	63.977,80
	N	74,00	10,00	84,00
Total	Media	41.441,78	26.031,92	34.419,57
	N	258,00	216,00	474,00

Aunque en la Tabla 7.1 las tres dimensiones (*categoría laboral*, *sexo* y *estadísticos*) están visibles simultáneamente, esto no tiene por qué ser así siempre. De hecho, es posible utilizar

las categorías de cualquiera de esas tres dimensiones para dividir la tabla en *capas* y ver sólo una de ellas: todas las capas, excepto una, están situadas *detrás* (y por tanto no se ven) de la capa visible.

Las Tablas 7.2.a, 7.2.b y 7.2.c muestran los mismos datos que la Tabla 7.1, pero en un formato diferente: la tabla original (7.1) ha quedado descompuesta en tres *capas* utilizando los niveles o categorías de la variable *sexo*: la primera capa (Tabla 7.2.a) recoge los datos referidos a toda la muestra; la segunda capa (Tabla 7.2.b), los datos referidos a los hombres; y la tercera capa (Tabla 7.2.c), los datos referidos a las mujeres (en esta última capa no aparece una categoría laboral –seguridad– porque en la muestra no existen mujeres que pertenezcan a esa categoría laboral).

Tabla 7.2.a. Primera capa: *Total*

Salario actual  
Sexo: Total

Categoría laboral	Media	N
Administrativo	27.838,54	363
Seguridad	30.938,89	27
Directivo	63.977,80	84
Total	34.419,57	474

Tabla 7.2.b. Segunda capa: *Hombre*

Salario actual  
Sexo: Hombre

Categoría laboral	Media	N
Administrativo	31.558,15	157
Seguridad	30.938,89	27
Directivo	66.243,24	74
Total	41.441,78	258

Tabla 7.2.c. Tercera capa: *Mujer*


Salario actual  
Sexo: Mujer

Categoría laboral	Media	N
Administrativo	25.003,69	206
Directivo	47.213,50	10
Total	26.031,92	216

Cuando una tabla pivotante contiene varias capas, sólo está visible una de ellas. Las capas restantes están ocultas detrás de la capa visible. Para ver el contenido de las capas ocultas es necesario recorrer una a una las distintas capas o, si se prefiere, transformar la dimensión que define las capas en una dimensión fila o en una dimensión columna. Estas y otras acciones de pivotado pueden ejecutarse de forma rápida y sencilla recurriendo a los *paneles de pivotado*.

## Paneles de pivotado

Para activar los paneles de pivotado:

- Seleccionar la opción **Paneles de pivotado** del menú **Pivotar** del *Editor de tablas* o del menú emergente que se obtiene al pulsar el botón secundario del ratón cuando el puntero del ratón se encuentra sobre una tabla pivotante. Se consigue el mismo efecto pinchando sobre el icono  de la barra de herramientas del *Editor de tablas*.

Esta acción conduce al cuadro de diálogo *Paneles de pivotado#* que muestra la Figura 7.5. El símbolo # se refiere al número de orden que ocupa cada panel de pivotado abierto durante una misma sesión.

Figura 7.5. Detalles de un cuadro de *Paneles de pivotado*



Cada icono representa una dimensión de la tabla (por ejemplo, una variable, un conjunto de estadísticos, etc.). Todas las dimensiones (iconos) se encuentran ubicadas en uno de los tres paneles disponibles: *fila*, *columna* o *capa*. Cualquier cambio de lugar de un icono supone el mismo cambio de lugar de la correspondiente dimensión en la tabla pivotante.

Para **identificar qué dimensión** de la tabla pivotante corresponde a cada icono de los paneles de pivotado:

- Situar el puntero del ratón sobre un icono y mantener pulsado el botón principal. La dimensión correspondiente a ese icono aparece sombreada en la tabla pivotante.

Para conocer el **significado de las etiquetas** utilizadas por el SPSS en las dimensiones que no corresponden a variables (por ejemplo, la categoría *Media* de la dimensión *Estadísticos*):

- Situar el puntero del ratón sobre la etiqueta, pulsar el botón secundario del ratón y, en el menú emergente que se obtiene, seleccionar la opción ¿Qué es esto?

Para **mover una dimensión** de un panel a otro:

- Pinchar sobre el icono de esa dimensión y arrastrarlo al panel deseado. Las dimensiones del panel *capa* pueden llevarse al panel *fila* o al panel *columna*, además de arrastrando los correspondientes iconos, utilizando las opciones **Mover capas a filas** y **Mover capas a columnas** del menú **Pivotar**.

Para **cambiar el orden** en el que aparecen las dimensiones de un mismo panel:

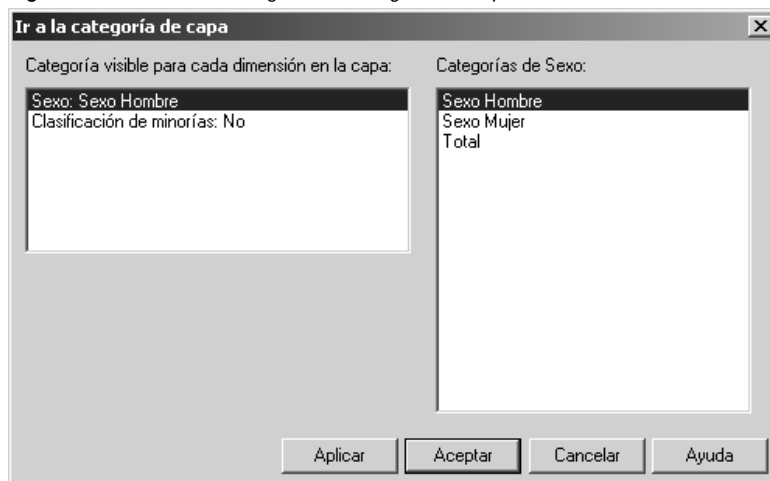
- Pinchar sobre los iconos de ese panel y arrastrarlos dentro del propio panel hasta obtener el orden deseado.

Para **moverse por las distintas capas** haciendo visible su contenido:

- Pinchar con el puntero del ratón sobre las flechas adosadas al (los) icono(s) de capa.
- También puede visualizarse una capa concreta seleccionando la opción **Ir a capa...** del menú **Pivotar** (esta opción es particularmente útil cuando se utiliza un gran número de dimensiones como capas o cuando alguna dimensión posee un gran número de

capas). La opción **Ir a capa...** conduce al cuadro de diálogo *Ir a la categoría de capa* que muestra la Figura 7.6.

Figura 7.6. Cuadro de diálogo *Ir a la categoría de capa*



La lista **Categoría visible para cada dimensión en la capa** ofrece un listado de las dimensiones que están siendo utilizadas como capas. Al seleccionar una dimensión de este listado, la lista **Categorías de Categoría laboral** muestra las categorías de la dimensión seleccionada. Al seleccionar una de estas categorías, el botón **Aplicar** coloca en el primer plano de la tabla pivotante la categoría seleccionada. El botón **Aceptar** hace lo mismo, pero además, cierra el cuadro de diálogo.

Para **transponer la posición de las filas y de las columnas**:

- Utilizar la opción **Transponer filas y columnas** del menú **Pivotar**. Se consigue el mismo efecto con los paneles de pivotado, arrastrando todos los iconos del panel fila al panel columna y todos los iconos del panel columna al panel fila (ver Figura 7.5).

Después de efectuar una o más acciones de pivotado, es posible dejar todo exactamente igual que estaba originalmente. Para devolver todos los iconos a su posición original:

- Utilizar la opción **Restablecer pivotes por defecto** del menú **Pivotar**.

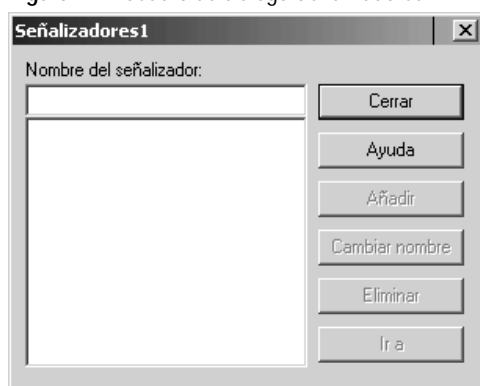
## Señalizadores

Los señalizadores sirven para guardar diferentes configuraciones de una misma tabla. El *Editor de tablas* permite crear varios señalizadores memorizando: (1) las posiciones de los elementos en las filas, columnas y capas; (2) el orden de presentación de los elementos en cada dimensión; y (3) la capa visualizada. Así, es posible, por ejemplo, guardar un señalizador con la configuración de tabla que el SPSS genera por defecto; otro con la disposición de las filas y las columnas cambiada; otro más cambiando el aspecto de las casillas o el color de las cabeceras; etc.

Para crear un señalizador:

- Seleccionar, dentro del *Editor de tablas*, la opción **Señalizadores** del menú **Pivotar** para acceder al cuadro de diálogo *Señalizadores#* que muestra la Figura 7.7 (el título del cuadro de diálogo aparece acompañado de un número que representa el número de orden que ocupa la tabla que se está editando en el conjunto de tablas editadas en una sesión).

Figura 7.7. Cuadro de diálogo *Señalizadores*



**Nombre del señalizador.** Permite asignar un nombre al señalizador en el que quedará memorizada la configuración actual de la tabla. Una vez nombrado un señalizador:

- Pulsar el botón **Añadir** para registrar el nombre en la lista de señalizadores.
- Pulsar el botón **Cambiar nombre** para modificar el nombre de un señalizador previamente añadido.
- Pulsar el botón **Eliminar** para eliminar un señalizador previamente añadido.
- Elegir un señalizador y pulsar el botón **Ir a** para que la tabla adopte la configuración memorizada en ese señalizador.

Cada tabla pivotante tiene su propio conjunto de señalizadores (tantos como se desee). Dentro de un mismo conjunto, el nombre de cada señalizador debe ser único; pero es posible repetir nombres en conjuntos de señalizadores diferentes.

## Modificar las propiedades de una tabla

El *Editor de tablas* contiene varias opciones diseñadas para poder controlar diferentes aspectos de una tabla. Todas estas opciones se encuentran en el cuadro de diálogo *Propiedades de tabla*. Este cuadro de diálogo contiene opciones agrupadas en cuatro bloques:

- Propiedades *Generales*.
- Propiedades relacionadas con las *notas a pie de tabla*.
- Propiedades relacionadas con el *formato de las casillas*.
- Propiedades relacionadas con los *bordes* de las tablas.
- Propiedades relacionadas con la *impresión* de tablas.

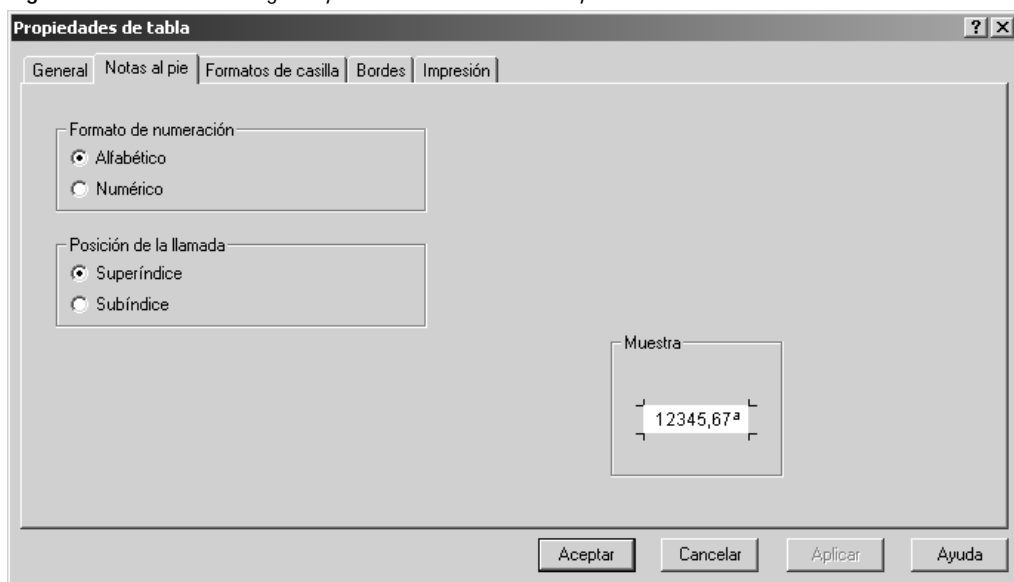


## Notas al pie

Para cambiar las propiedades de las notas a pie de tabla:

- Pulsar sobre la solapa **Notas al pie** del cuadro de diálogo *Propiedades de tabla* (ver Figura 7.8) para cambiar la vista del cuadro a *Notas al pie* (ver Figura 7.9).

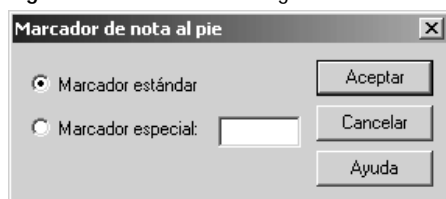
Figura 7.9. Cuadro de diálogo *Propiedades de tabla: Notas al pie*



**Formato de numeración.** Controla el estilo de los marcadores de las notas a pie de tabla: *alfabético* (a, b, c, ...), o *numérico* (1, 2, 3, ...).

**Posición de la llamada.** Las llamadas de las notas a pie de tabla pueden aparecer como *superíndices* o como *subíndices*. Si se elimina o añade una nota a pie de tabla, es posible que el orden de numeración de las notas quede alterado. En ese caso, la opción **Renumerar notas al pie** del menú **Formato** permite asignar a las notas caracteres correlativos (a, b, c, ...; 1, 2, 3, ...). También existe la posibilidad de cambiar el aspecto de un marcador concreto; para ello, entrar en el *Editor de tablas*, seleccionar la nota cuyo marcador se desea cambiar y elegir la opción **Marcador de notas al pie** del menú **Formato** para acceder al cuadro de diálogo *Marcador de nota al pie* que muestra la Figura 7.10.

Figura 7.10. Cuadro de diálogo *Marcador de nota al pie*

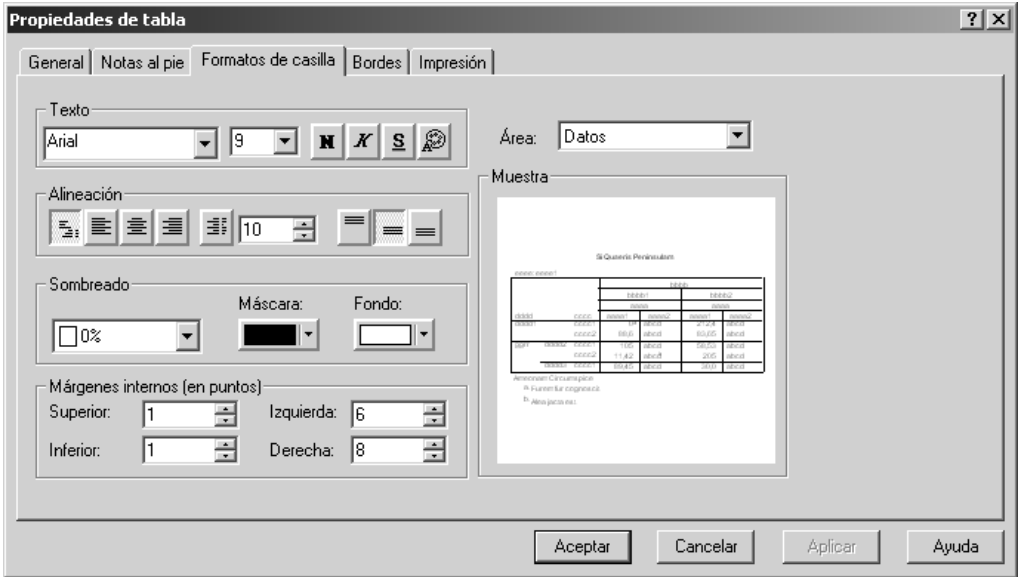




## Formatos de casilla

- Pulsar la solapa **Formatos de casilla** del cuadro de diálogo *Propiedades de tabla* (ver Figura 7.8) para cambiar la vista del cuadro a *Formatos de casilla* (ver Figura 7.11).

Figura 7.11. Cuadro de diálogo *Propiedades de tabla: Formatos de casilla*



**Área.** Una tabla se divide en varias áreas: título, capas, etiquetas de esquina, etiquetas de fila, etiquetas de columna, datos, texto a pie de tabla y notas a pie de tabla (ver Figura 7.12). Cada área puede recibir un formato de casilla distinto. Este formato afecta a todas las casillas de la misma área (desde aquí no puede asignarse formato a una casilla individual).

Figura 7.12. Áreas de una tabla pivotante

Título	
Capas	
Etiquetas de esquina	Etiquetas de columna
Etiquetas de fila	Datos

Texto a pie de tabla  
Notas a pie de tabla

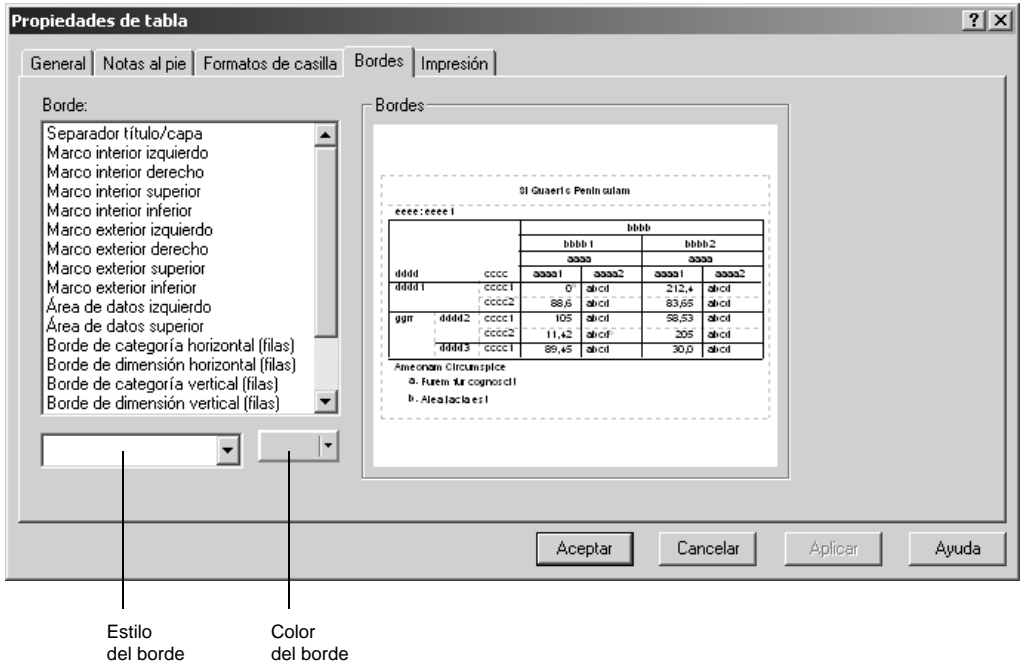
Los formatos de casilla que es posible controlar son cuatro: las características del **Texto** (el tipo de fuente, el tamaño, el color, el estilo), la **Alineación** horizontal y vertical del texto dentro de las casillas, el **Sombreado** de las casillas (colores de la máscara y del fondo) y los **Márgenes internos** de las casillas (en centímetros; pero pueden seleccionarse puntos o pulgadas en el menú Edición > Opciones... > General, recuadro Sistema de medida).

## Bordes

Para cambiar el estilo y el color de los bordes de una tabla:

- Pulsar la solapa **Bordes** del cuadro de diálogo *Propiedades de tabla* (ver Figura 7.8) para cambiar la vista del cuadro a *Bordes* (ver Figura 7.13).

Figura 7.13. Cuadro de diálogo *Propiedades de tabla: Bordes*



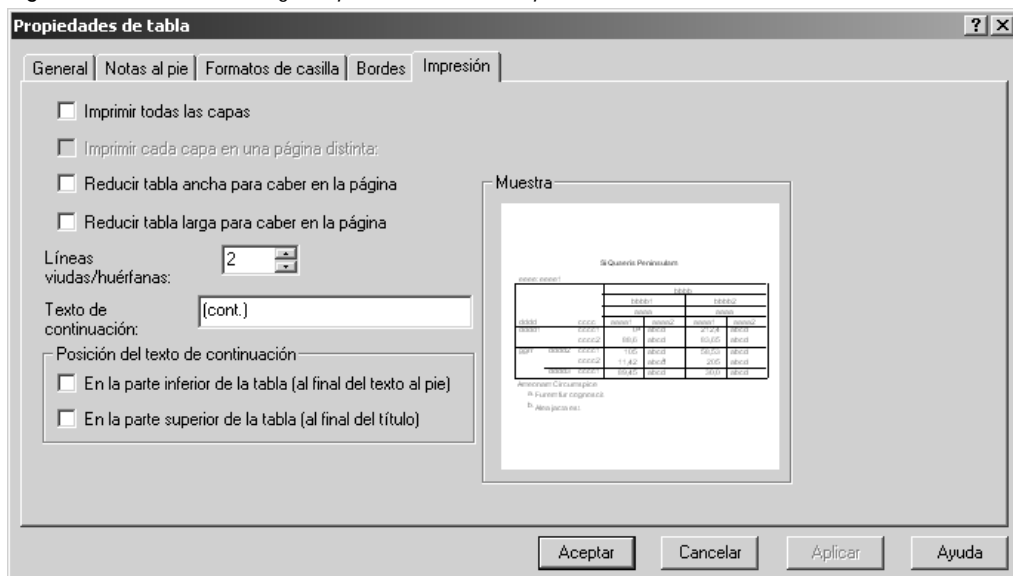
La lista **Borde** contiene un listado con todos los bordes posibles de una tabla. Para modificar un borde:

- Seleccionar, en la lista **Borde**, la opción que identifica el borde que se desea modificar y, tras esto, seleccionar un estilo y un color para el borde utilizando los correspondientes botones de menú desplegable.

## Impresión

Las propiedades de impresión permiten controlar, básicamente, algunos aspectos relacionados con la impresión de las distintas capas de una tabla, así como la forma de imprimir tablas demasiado largas o demasiado anchas para entrar en una página. Para modificar las propiedades de impresión:

- Pulsar la solapa **Impresión** del cuadro de diálogo *Propiedades de tabla* (ver Figura 7.8) para cambiar la vista del cuadro a *Impresión* (ver Figura 7.14).

Figura 7.14. Cuadro de diálogo *Propiedades de tabla: Impresión*

- “ **Imprimir todas las capas.** Con esta opción activa se imprimen todas las capas (esto afecta a la impresión, pero no al aspecto de la tabla en el *Visor*). Desactivando esta opción sólo se imprime la capa visible, es decir, la situada en primer plano.
- “ **Imprimir cada capa en una página distinta.** Activando esta opción, cada capa se imprime en una página diferente (sólo disponible seleccionando *Imprimir todas las capas*).
- “ **Reducir tabla ancha para caber en la página.** Permite reducir la tabla horizontalmente para conseguir que se ajuste al ancho de la página.
- “ **Reducir tabla larga para caber en la página.** Permite reducir la tabla verticalmente para conseguir que se ajuste a la longitud de la página.

**Líneas viudas/huérfanas.** Esta opción permite especificar el número mínimo de filas y columnas que podrá incluir cualquier sección impresa de una tabla si ésta es demasiado ancha y/o larga para ajustarse al tamaño de página establecido. Si una tabla no cabe en una página porque hay otros resultados por encima de la tabla en esa página, pero cabe dentro de la longitud de página establecida, se imprimirá automáticamente en una nueva página, independientemente del valor especificado en *Líneas viudas/huérfanas*.

**Texto de continuación.** Cuando una tabla se imprime en más de una página, esta opción permite personalizar el texto que aparecerá indicando tal circunstancia. Este texto no aparece en las tablas del *Visor*, sino sólo en las páginas impresas, y su ubicación puede controlarse mediante las opciones del recuadro *Posición del texto de continuación*:

- “ **En la parte inferior de la tabla.** Al final de la tabla, como texto a pie de tabla.
- “ **En la parte superior de la tabla.** Al comienzo de la tabla, como título. Si la tabla ya tiene título, el texto de continuación se imprime a continuación del título.

## Modificar las propiedades de una casilla

A diferencia de lo que ocurre con las propiedades de las tablas, las propiedades de las casillas pueden aplicarse a casillas individuales. La opción **Propiedades de casilla** del menú **Formato** permite controlar varios aspectos relacionados con el formato de las casillas. En concreto, permite controlar:

- El formato del *valor* de la casilla.
- La *alineación* vertical y horizontal del texto.
- Los *márgenes* entre el texto y los bordes de la casilla.
- El *sombreado* de la máscara y el fondo.

Las propiedades de casilla tienen preferencia sobre las propiedades de tabla. Una vez asignada una propiedad de casilla, ésta prevalecerá sobre cualquier cambio efectuado en las propiedades de tabla.

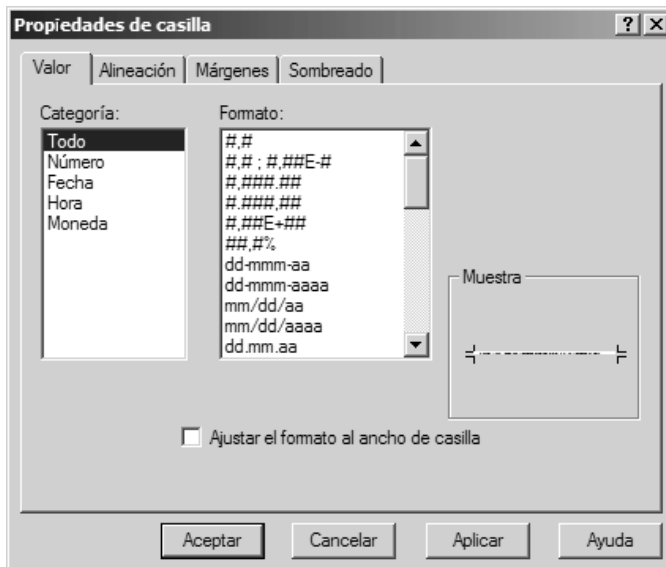
Para cambiar las propiedades de una casilla hay que situar el cursor en esa casilla antes de seleccionar la opción **Propiedades de casilla**. Para cambiar las propiedades de una fila o una columna, es necesario seleccionar previamente esa fila o esa columna.

### Valor

Para cambiar el formato del *valor* de una casilla:

- Seleccionar, dentro del *Editor de tablas*, la opción **Propiedades de casilla...** del menú **Formato** para acceder al cuadro de diálogo *Propiedades de casilla: Valor* que muestra la Figura 7.15.

Figura 7.15. Cuadro de diálogo *Propiedades de casilla: Valor*



Hay cuatro formatos básicos disponibles: *número*, *fecha*, *hora* y *moneda*. Seleccionando uno de estos formatos básicos en la lista **Categoría**, la lista **Formato** muestra las variantes disponibles para ese formato. Seleccionando la categoría *Todo*, la lista **Formato** muestra un listado con todos los formatos disponibles.

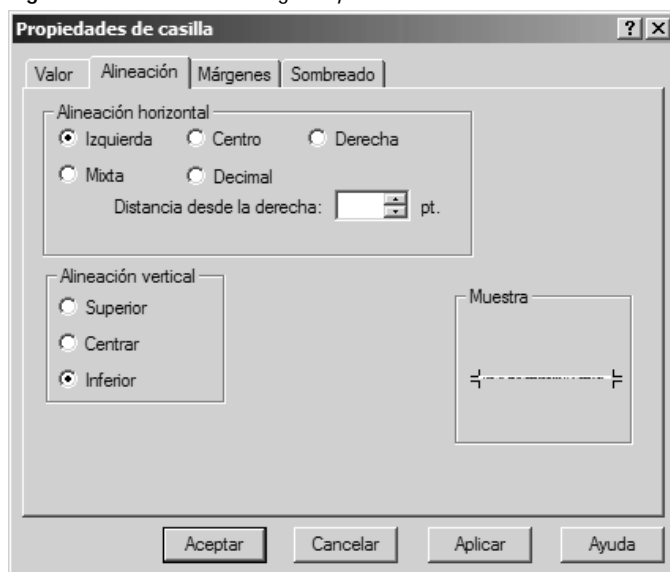
Este cuadro de diálogo también permite establecer el número de decimales que se desea visualizar y, dentro de una categoría de formato, conseguir que el formato concreto visualizado sea aquel que se ajuste al ancho de la casilla.

## Alineación

Para controlar la posición horizontal y vertical del texto dentro de una casilla:

- Pulsar sobre la solapa **Alineación** del cuadro de diálogo *Propiedades de casilla* (ver Figura 7.15) para cambiar la vista del cuadro a *Alineación* (ver Figura 7.16).

Figura 7.16. Cuadro de diálogo *Propiedades de casilla: Alineación*



**Alineación horizontal.** Controla la posición horizontal del texto dentro de las casillas: a la *izquierda*, en el *centro*, a la *derecha*, *mixta* (el texto se alinea dependiendo del formato de la casilla: numérico, fecha, moneda, etc.) y *decimal* (el texto se alinea tomando como referencia el separador decimal).

La opción **Distancia desde la derecha** permite situar el texto en una posición exacta tomando como referencia el borde derecho de la casilla (la distancia al borde derecho se calcula en puntos).

**Alineación vertical.** Controla la posición vertical del texto: en la parte *superior*, en el *centro* y en la parte *inferior* de la casilla.

## Márgenes

Para especificar la distancia exacta que debe existir entre el contenido (el texto) de una casilla y cada uno de sus bordes (*superior*, *inferior*, *izquierda* y *derecha*):

- Pulsar sobre la solapa **Márgenes** del cuadro de diálogo *Propiedades de casilla* (ver Figura 7.15) para cambiar la vista del cuadro a *Márgenes* (ver Figura 7.17).

Figura 7.17. Cuadro de diálogo *Propiedades de casilla: Márgenes*



Los márgenes internos de una casilla o del conjunto de casillas seleccionadas pueden modificarse simplemente introduciendo el valor deseado en los cuadros de texto destinados a tal efecto (la distancia del texto a los bordes se calcula en puntos).

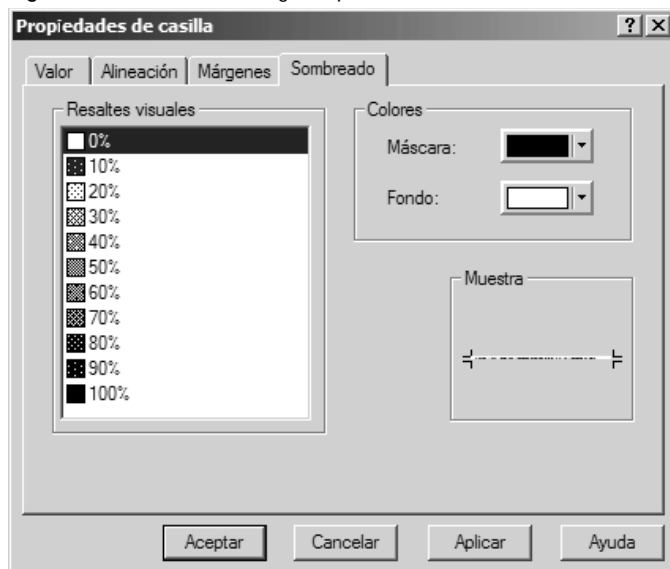
## Sombreado

Para cambiar el sombreado de una casilla:

- Pulsar sobre la solapa **Sombreado** del cuadro de diálogo *Propiedades de casilla* (ver Figura 7.15) para cambiar la vista del cuadro a *Sombreado* (ver Figura 7.18).

**Resaltos visuales.** Controla el porcentaje de *sombreado* de una casilla o del conjunto de casillas seleccionadas.

**Colores.** Permite seleccionar el color de la *máscara* (contorno) y del *fondo* de una casilla o del conjunto de casillas seleccionadas. El color de la máscara y del fondo no altera el color del texto.

Figura 7.18. Cuadro de diálogo *Propiedades de casilla: Sombreado*

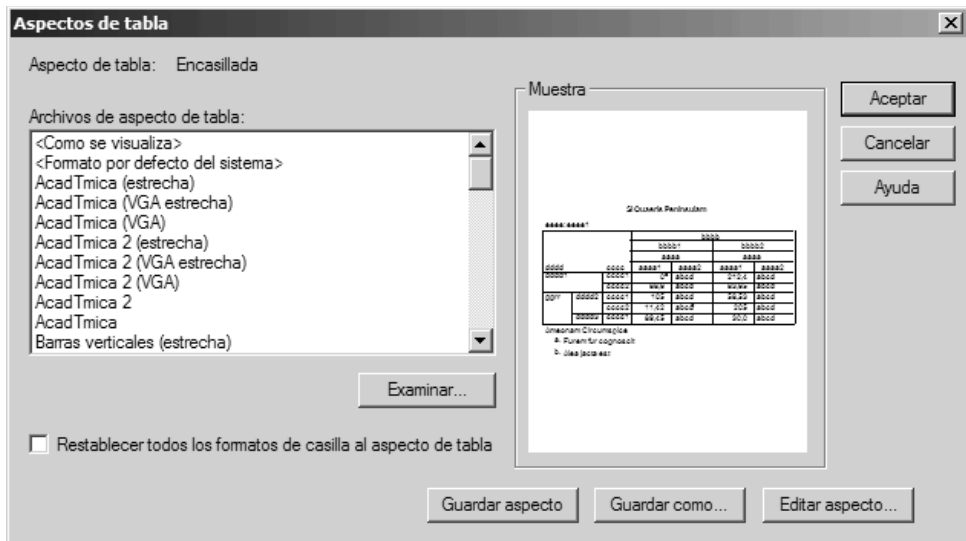
## Seleccionar el aspecto de una tabla

La apariencia de una tabla puede modificarse de dos maneras: editando sus propiedades (ver los dos apartados anteriores) o seleccionando y aplicando un *aspecto de tabla* existente. Un *aspecto de tabla* es un conjunto de propiedades de tabla, es decir, un conjunto de formatos generales, de casilla, de notas a pie y de bordes. Estos *aspectos de tabla* pueden seleccionarse entre los proporcionados por el programa o pueden ser creados y guardados por el propio usuario. Para crear o aplicar un aspecto de tabla:

- Seleccionar la opción **Aspectos de tabla...** del menú **Formato** para acceder al cuadro de diálogo *Aspectos de tabla* que muestra la Figura 7.19.

**Archivos de aspecto de tabla.** Cada *aspecto de tabla* está guardado en un archivo con extensión *.tlo*. Al seleccionar uno de estos archivos queda activado su correspondiente *aspecto de tabla*. Si existen *aspectos de tabla* guardados en archivos ubicados en una carpeta distinta de la que el SPSS utiliza por defecto, el botón **Examinar...** ayuda a buscar esos archivos (la carpeta utilizada por defecto es la seleccionada como tal en la solapa **Tablas pivote** del cuadro de diálogo *Opciones de SPSS*, al cual se accede desde el menú **Edición > Opciones...**).

**Restablecer todos los formatos de casilla al aspecto de tabla.** Según se ha señalado ya, las *propiedades de casilla* prevalecen sobre las *propiedades de tabla* (estas propiedades están descritas en los dos apartados anteriores). Puesto que los *aspectos de tabla* se basan en las *propiedades de tabla*, las casillas editadas con la opción *propiedades de casilla* no se verán alteradas al aplicar un *aspecto de tabla*. Ahora bien, marcando esta opción, el *aspecto de tabla* afecta a todas las casillas, independientemente de que hayan sido o no editadas mediante *propiedades de casilla*.

Figura 7.19. Cuadro de diálogo *Aspectos de tabla*

**Editar aspecto...** Esta opción conduce a los cuadros de diálogo ya estudiados en el apartado *Propiedades de tabla*. Estos cuadros de diálogo contienen opciones para establecer las propiedades (generales, de las notas a pie de tabla, del formato de las casillas y de los bordes de la tabla; ver Figuras 7.10 a 7.13) que formarán parte del *aspecto de tabla*. Una vez seleccionado y/o editado un *aspecto de tabla*, éste puede guardarse en un archivo para su uso posterior.

*Nota.* La solapa **Tablas pivote** del cuadro de diálogo *Opciones de SPSS*, al cual se accede mediante **Edición > Opciones...**, permite seleccionar el *aspecto de tabla* que será utilizado como *aspecto por defecto* para las tablas pivotantes del *Visor*. Este *aspecto por defecto* no se aplica a las tablas del *Visor* ya creadas, sino sólo a las que se generan a partir de la selección de ese aspecto de tabla.

## Características del texto

Una de las opciones que incluye el menú **Propiedades de tabla** (ver, en este mismo capítulo, el apartado *Modificar las propiedades de una tabla*) se refiere a la posibilidad de seleccionar las características del texto de una tabla (tipo, estilo y tamaño de la *fuerza* del texto). Pero la selección de estas características desde el menú **Propiedades de tabla** (y, por tanto, desde el menú **Aspectos de tabla**) afecta, no a una casilla individual, sino a una o varias de las *áreas* de la tabla. Por el contrario, la opción **Fuente** del menú **Formato** permite controlar el tipo, el estilo y el tamaño de la fuente del texto de una casilla individual o del conjunto de casillas seleccionadas. Para modificar las características del texto de una casilla:

- Seleccionar la opción **Fuente...** del menú **Formato** del *Editor de tablas* para acceder al cuadro de diálogo *Fuente* (este cuadro de diálogo es idéntico al ya visto en este mismo capítulo en el apartado *Tamaño y fuente de los titulares*; ver Figura 7.2).





Además del tipo, estilo y tamaño de la fuente, el cuadro de diálogo *Fuente* permite seleccionar un *color* para la fuente, *ocultar* el texto de la casilla o casillas seleccionadas y *subrayar* el texto. También permite seleccionar un estilo de escritura (en *Alfabeto*). Las características seleccionadas para una casilla concreta afectan a todas las *capas* de la tabla (si existen).

## Anchura de las casillas

El *Visor de resultados* controla la anchura de las casillas basándose en las especificaciones establecidas en la solapa **Tablas** pivote del cuadro de diálogo *Opciones de SPSS* (en el menú **Edición > Opciones...**).

Por defecto, la anchura de las casillas se ajusta de forma automática a la anchura de las etiquetas de las columnas. La consecuencia de este tipo de ajuste es que los valores demasiado anchos no pueden visualizarse (aparecen asteriscos). No obstante, el cuadro de diálogo *Opciones de SPSS* permite cambiar ese criterio para hacer que la anchura de las casillas se ajuste tanto a la anchura de las *etiquetas* como a la anchura de los *valores* (la mayor de ambas).

Al margen de las especificaciones iniciales sobre el ajuste automático de la anchura de las casillas (especificaciones que afectan a todas las tablas generadas por el *Visor de resultados*), la anchura de las casillas puede ser alterada de diferentes formas. Para cambiar la anchura de las casillas de **una sola columna**:

- Situar el puntero del ratón sobre el borde derecho de la columna cuya anchura se desea cambiar. El puntero abandona su forma habitual () para convertirse en un *puntero de selección horizontal* (). Para conseguir que el cursor cambie de forma no es necesario que el borde de la columna sea visible.
- Pulsar el botón principal del ratón y arrastrar el borde hasta dar a la columna la anchura deseada.

Para dar la misma anchura a **todas las columnas que contienen casillas de datos**:

- Seleccionar la opción **Ancho de casillas de datos...** del menú **Formato**. Se accede así al cuadro de diálogo *Establecer ancho de las casillas de datos*. Este cuadro de diálogo permite establecer una anchura fija para todas las columnas que contienen casillas con datos (esta anchura no afecta a las columnas que sólo contienen etiquetas).

Después de haber modificado la anchura de algunas casillas, es posible hacer que la tabla vuelva a **adoptar su aspecto original**. Para ello:

- Seleccionar la opción **Autoajuste** del menú **Formato**. Esta opción ajusta automáticamente la anchura de las casillas al contenido de las mismas.

Cuando las etiquetas de las columnas de datos son más anchas que los propios datos, es posible **reducir la anchura de las columnas** colocando las etiquetas de forma *vertical*. Para rotar las etiquetas:

- Seleccionar la opción **Rotar etiquetas de columna interior** del menú **Formato**.

También es posible **reducir la anchura de la primera columna** colocando verticalmente sus etiquetas. Para ello:

- Seleccionar la opción **Rotar etiquetas de fila exterior** del menú **Formato** (esta opción sólo está disponible si las filas de la tabla están definidas por más de una dimensión).

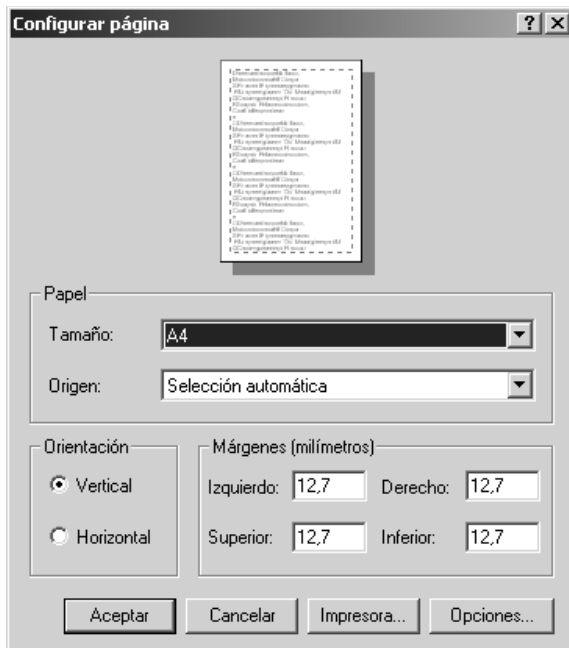
## Imprimir resultados

### Preparar página

Antes de imprimir el archivo de resultados puede interesar controlar algunos detalles relacionados con el proceso de impresión. Para ello:

- Dentro del *Visor de resultados*, seleccionar la opción **Preparar página...** del menú **Archivo** para acceder al cuadro de diálogo *Configurar página* que muestra la Figura 7.20.

Figura 7.20. Cuadro de diálogo *Configurar página*



Este cuadro de diálogo permite seleccionar el tipo de papel (A4, carta, ejecutivo, etc.), la orientación del mismo (vertical, horizontal) y el tamaño de los márgenes entre el texto y los cuatro bordes del papel.

El botón **Impresora...** conduce a un subcuadro de diálogo (también llamado *Configurar página*) que permite: (1) seleccionar la impresora con la que se desea imprimir (la elección puede hacerse entre las impresoras definidas en Windows) y (2) establecer sus propiedades. El botón **Propiedades** conduce al subcuadro de diálogo *Propiedades de nombre-de-impresora* que ya se ha descrito en el apartado *Imprimir archivos de datos* del Capítulo 3 (ver Figura 3.17).

Y el botón **Opciones...** conduce al subcuadro de diálogo *Configurar página: Opciones (Cabecera/Pie)* que muestra la Figura 7.21.

Figura 7.21. Subcuadro de diálogo *Preparar página: Opciones (Cabecera/Pie)*



Este cuadro de diálogo permite editar la **Cabecera** y el **Pie** de página que aparecerán en todas las páginas impresas. La barra de herramientas central contiene opciones para controlar el *tipo de letra* y la *justificación* del texto, insertar la *fecha*, la *hora* y el *número de página*, y establecer varios *niveles* para los encabezados. Para controlar otros detalles de la página impresa:

- Pulsar la solapa **Opciones** del cuadro de diálogo *Preparar página: Opciones* (ver Figura 7.21) para acceder al subcuadro de diálogo *Preparar página: Opciones (Opciones)*.

Dentro de este cuadro de diálogo (no se incluye aquí), el recuadro **Tamaño del gráfico impreso** contiene opciones que permiten seleccionar uno de cuatro tamaños diferentes para los gráficos que se imprimen. Estos cuatro tamaños de gráfico pueden tomar como referencia el propio gráfico (es decir, su tamaño original) o el tamaño de la página. Las proporciones del gráfico se mantienen independientemente del tamaño seleccionado en este cuadro de diálogo (es decir, el tamaño seleccionado para los gráficos en este cuadro de diálogo no altera la relación existente entre su anchura y su altura).

Este subcuadro de diálogo también incluye dos opciones que permiten, por un lado, establecer el espacio que se desea que exista entre cada objeto del *Visor* (es decir, la separación vertical entre objetos) y, por otro, el número de página con el que se desea empezar a numerar la primera página.

## Controlar la ruptura de las tablas grandes

Al imprimir archivos de resultados existe la posibilidad de encontrarse con que el SPSS divide automáticamente una tabla y utiliza más de una página para imprimirla. Esto ocurre cuando la tabla que se está imprimiendo es más ancha o más larga que el tamaño de página definido.

Si no se desea que esta división se realice de forma automática, el *Editor de tablas* del *Visor de resultados* contiene una serie de opciones para que el usuario pueda decidir dónde romper la tabla o qué partes de la misma conservar juntas.

Para ***mantener juntas*** dos o más filas, o dos o más columnas:

- Dentro del *Editor de tablas*, seleccionar las *etiquetas* de las filas o columnas que se quiere mantener juntas.
- Seleccionar la opción **Mantener juntas** del menú **Formato**.

Para indicar ***por dónde debe romperse la tabla*** en el caso de que sea demasiado ancha o demasiado larga:

- Dentro del *Editor de tablas*, situar el cursor en la etiqueta de la fila por encima de la cual se desea romper la tabla o en la etiqueta de la columna a la izquierda de la cual se desea romper la tabla.
- Seleccionar la opción **Ruptura aquí** del menú **Formato**.

Para eliminar los códigos ***Mantener juntas*** y ***Ruptura aquí***:

- Seleccionar las opciones **Eliminar Mantener juntas** o **Eliminar Ruptura aquí** del menú **Formato** (dentro del *Editor de tablas*).

En lugar de introducir códigos de ruptura, es posible que una recomposición de la tabla consiga hacer que ésta quepa en una sola página. Para ***recomponer la tabla*** intentando ajustarla al tamaño de la página:

- Seleccionar la opción **Reducir tabla ancha para caber en la página** (o **Reducir tabla larga para caber en la página**) del menú **Formato > Propiedades de tabla > Impresión** (esta opción sólo está disponible en el *Editor de tablas*; por tanto, para poder utilizarla, no basta con seleccionar la tabla, sino que es necesario estar en modo de edición; ver próximo capítulo).

## Presentación preliminar

Después de personalizar las opciones del cuadro de diálogo *Configurar página* y de establecer los puntos de ruptura de las tablas (en el caso de que fuera necesario), suele resultar conveniente visualizar, antes de imprimir, el aspecto exacto que adoptará el archivo de resultados una vez impreso. Para ello:

- Seleccionar la opción **Presentación preliminar** del menú **Archivo**. Se obtiene el mismo resultado pulsando con el puntero del ratón sobre el botón *Presentación preliminar* de la barra de herramientas (un folio con una lupa).

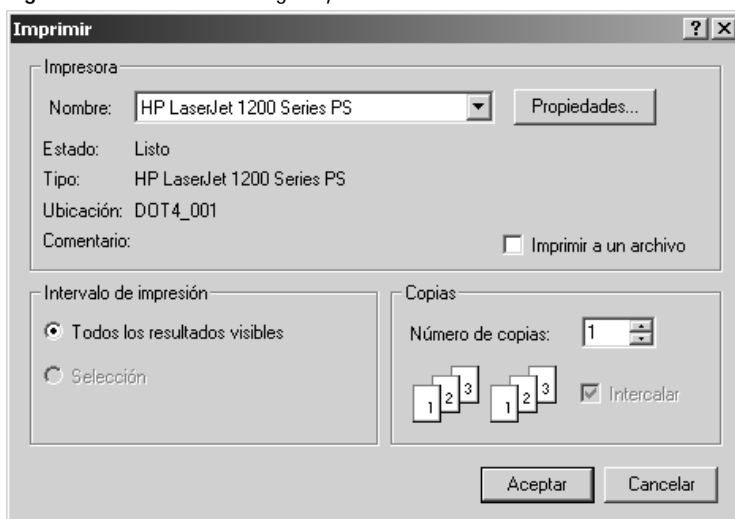
La ventana de *presentación preliminar* sólo muestra los objetos seleccionados. Si se desea visualizar todo el archivo de resultados (página a página), hay que asegurarse de que no se encuentre seleccionado ningún objeto del *Visor*.

## Imprimir

Para imprimir el archivo de resultados:

- Seleccionar la opción **Imprimir...** del menú **Archivo** para acceder al cuadro de diálogo *Imprimir* que muestra la Figura 7.22.

Figura 7.22. Cuadro de diálogo *Imprimir*



**Impresora.** Las opciones de este recuadro permiten seleccionar una de las impresoras disponibles en la lista desplegable **Nombre** y establecer sus **Propiedades** (para esto último, ver la Figura 3.17, en el Capítulo 3).

- **Imprimir en un archivo.** Activando esta opción, el archivo de resultados se guarda en un archivo acompañado de todas las propiedades de impresión establecidas.

**Intervalo de impresión.** Las opciones de este recuadro permiten decidir si se desea imprimir *todos los resultados visibles* (no se imprimen los objetos ocultos del *Visor*) o sólo los resultados que en ese momento se encuentran *seleccionados*.

En las tablas pivotantes con varias capas sólo es visible la primera de ellas, de modo que, si no se indica lo contrario, sólo se imprimirá la capa visible. Para *imprimir todas las capas* de una tabla:

- Seleccionar la opción **Imprimir todas las capas** del cuadro de diálogo *Propiedades de tabla: General* (ver Figura 7.8). Opción sólo disponible desde el *Editor de tablas*.

**Copias.** Este recuadro contiene opciones para controlar el número de copias que se desea imprimir y el orden en el que se imprimirán, pero estas opciones sólo están disponibles cuando la impresora seleccionada incluye esas funciones.

- " **Intercalar.** Si se opta por imprimir varias copias, éstas pueden obtenerse ordenadas de dos formas distintas: (1) todas las copias de la primera página, todas las copias de la segunda página, etc.; (2) la primera copia de todas las páginas, la segunda copia de todas las páginas, etc. Para esto último, marcar la opción **Intercalar** (es la opción por defecto).

## Copiar resultados en otras aplicaciones

Ya se ha señalado repetidamente que los objetos del *Visor de resultados* adoptan tres formatos básicos: texto, gráficos y tablas. El texto y los gráficos pueden copiarse en otras aplicaciones externas (procesadores de texto, hojas de cálculo, etc.) siguiendo la estrategia habitualmente utilizada en las aplicaciones que funcionan en entorno Windows: copiando y pegando. Pero las tablas pivotantes y los gráficos interactivos se copian siguiendo una estrategia algo diferente.

### Copiar texto y gráficos

Para copiar un cuadro de texto o un gráfico:

- ' Seleccionar, en el *Visor*, el texto y/o el gráfico que se desea copiar.
- ' Seleccionar la opción **Copiar** del menú **Edición** del *Visor*.
- ' Seleccionar la opción **Pegar** del menú **Edición** de la aplicación externa (es decir, del menú **Edición** de la aplicación en la cual se desea copiar el texto o el gráfico). En la mayoría de los procesadores de texto, esta opción pega los gráficos como un mapa de bits.

### Copiar tablas

Para copiar una tabla de resultados (una tabla pivotante):

- ' Seleccionar, en el *Visor*, la tabla que se desea copiar.
- ' Seleccionar la opción **Copiar** del menú **Edición** del *Visor*.
- ' Si se quiere copiar una tabla de tal forma que ésta adopte el formato de tabla propio de la aplicación externa, se debe utilizar la opción **Pegar** del menú **Edición** de la aplicación externa.
- ' Si se desea copiar una tabla de tal forma que ésta mantenga el formato original del *Visor* (filas, columnas, líneas, bordes, tamaño y tipo de letra, etc.), debe utilizarse la opción **Pegado especial...** del menú **Edición** de la aplicación externa. El cuadro de diálogo *Pegado especial* de la aplicación externa ofrecerá un listado de formatos dispo-

nibles entre los que podrá elegirse el apropiado (generalmente, se consigue un buen resultado pegando el objeto como un gráfico o como una imagen, es decir, como un *meta-archivo*).

- Si sólo se desea copiar el contenido de la tabla, sin formato, se debe utilizar la opción **Pegado especial...** del menú **Edición** de la aplicación externa y, en la lista de formatos disponibles, seleccionar **Texto sin formato**.

## Copiar más de un objeto

Para copiar más de un objeto de resultados:

- Seleccionar, en el *Visor*, los objetos que se desea copiar.
- Seleccionar la opción **Copiar objetos** del menú **Edición**. Utilizar esta opción sólo para copiar varios objetos en otras aplicaciones; para copiar varios objetos en otra parte del *Visor* o en otra ventana del *Visor*, utilizar la opción **Copiar**.
- Proceder de la forma ya descrita para pegar una tabla o un gráfico.

## Incrustar tablas

Las tablas del *Visor de resultados* pueden copiarse en otras aplicaciones de tal forma que, al pulsar dos veces sobre ellas con el puntero del ratón, puedan editarse como si se encontraran en el *Visor de resultados* del SPSS. A esta forma particular de copiar tablas se le llama *incrustar* tablas. Para incrustar tablas es necesario que la aplicación externa soporte el trabajo con objetos Active-X.

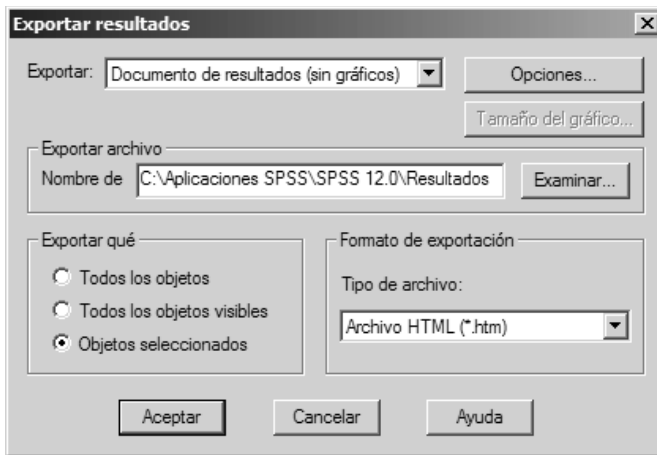
Para incrustar una tabla de resultados en una aplicación externa:

- Ejecutar el archivo *objs-on.bat* (que se encuentra en la misma carpeta en la que se ha instalado el SPSS). La aplicación Active-X queda activada al ejecutar este archivo. Para desactivar Active-X, ejecutar el archivo *objs-off.bat*.
- Seleccionar la tabla que se desea copiar y copiarla utilizando la opción **Copiar** del menú **Edición** del *Visor*.
- En la aplicación externa en la que se desea incrustar la tabla, utilizar la opción **Pegado especial...** del menú **Edición** y, de la lista de formatos, seleccionar *Objeto tabla pivote de SPSS*.

## Exportar resultados

La opción **Exportar** permite guardar los objetos del *Visor* en una amplia variedad de formatos. Para exportar total o parcialmente el archivo de resultados:

- Seleccionar la opción **Exportar...** del menú **Archivo** para acceder al cuadro de diálogo *Exportar resultados* que muestra la Figura 7.23.

Figura 7.23. Cuadro de diálogo *Exportar resultados*

**Exportar.** Este menú contiene tres opciones (pulsando el botón de menú desplegable) para decidir qué parte del archivo de resultados se desea exportar: *documento completo*, *documento sin gráficos* y *sólo los gráficos*.

- Al seleccionar la opción *Sólo gráficos*, y dependiendo del formato de exportación elegido, se activa el botón **Tamaño del gráfico...** Pulsando este botón se accede al cuadro de diálogo *Tamaño de exportación del gráfico* que muestra la Figura 7.24, el cual permite personalizar el tamaño del gráfico.

Figura 7.24. Cuadro de diálogo *Tamaño de exportación del gráfico*

**Exportar archivo.** El SPSS asigna un nombre por defecto al archivo al que van a ser exportados los resultados. Pero ese nombre puede cambiarse introduciendo el nuevo nombre en el cuadro de texto **Nombre de archivo**.

- El botón **Examinar...** ayuda a buscar el nombre y la carpeta del archivo al que se desea exportar los resultados.

**Exportar qué.** Independientemente de la opción elegida en el recuadro **Exportar**, las opciones de este recuadro permiten seleccionar:

**Todos los objetos.** Exporta todos los objetos del archivo de resultados, incluidos los ocultos.

**Todos los objetos visibles.** Excluye de la exportación los objetos ocultos.

**Objetos seleccionados.** Sólo exporta los objetos seleccionados.



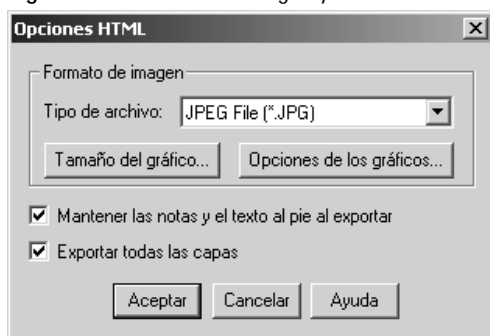
**Formato de exportación.** Las opciones contenidas en el botón de menú desplegable de este recuadro dependen de la opción elegida en el recuadro **Exportar**.

- Si se ha elegido **Documento de resultados**, el texto y las tablas se exportan en uno de los dos formatos de exportación disponibles: *HTML* (versión 3.0 o posterior) o *texto* (separado por tabuladores o separado por espacios); y los gráficos se exportan en el formato de exportación de gráficos seleccionado en el cuadro de diálogo *Opciones* (ver Figura 7.25).
- Si se ha elegido **Documento de resultados sin gráficos**, el texto y las tablas se exportan en uno de los dos formatos de exportación disponibles: *HTML* (versión 3.0 o posterior) o *texto* (separado por tabuladores o separado por espacios).
- Si se ha elegido **Sólo gráficos**, los gráficos se exportan en uno de los formatos de exportación disponibles para gráficos: Enhanced Metafile (\*.emf), JPEG (\*.jpe), PNG File (\*.png), Macintosh PICT (\*.pct), PostScript (\*.eps), Tagged Image File (\*.tif), Windows Bitmap (\*.bmp) y Windows Metafile (\*.wmf).

**Opciones.** El cuadro de diálogo *Exportar resultados* de la Figura 7.23 contiene un botón de opciones que permite seguir personalizando el formato de exportación del texto, de las tablas y de los gráficos. Este botón de opciones conduce a cuadros de diálogo que van variando en su aspecto y contenido dependiendo de la selección establecida en los recuadros **Exportar** y **Formato de exportación**.

- Si en **Exportar** se ha seleccionado *Documento de resultados* y en **Formato de exportación** se ha seleccionado *HTML File (\*.htm)* como **Tipo de archivo**:
- El botón **Opciones...** conduce al cuadro de diálogo *Opciones HTML* que muestra la Figura 7.25.

Figura 7.25. Cuadro de diálogo *Opciones HTML*



**Formato de imagen.** Permite seleccionar, para los gráficos, uno de los formatos de exportación ya mencionados (el texto y las tablas se exportan en formato HTML).

**Tamaño del gráfico...** Conduce al cuadro de diálogo *Tamaño de exportación del gráfico* (ver Figura 7.24), el cual permite personalizar el tamaño del gráfico.

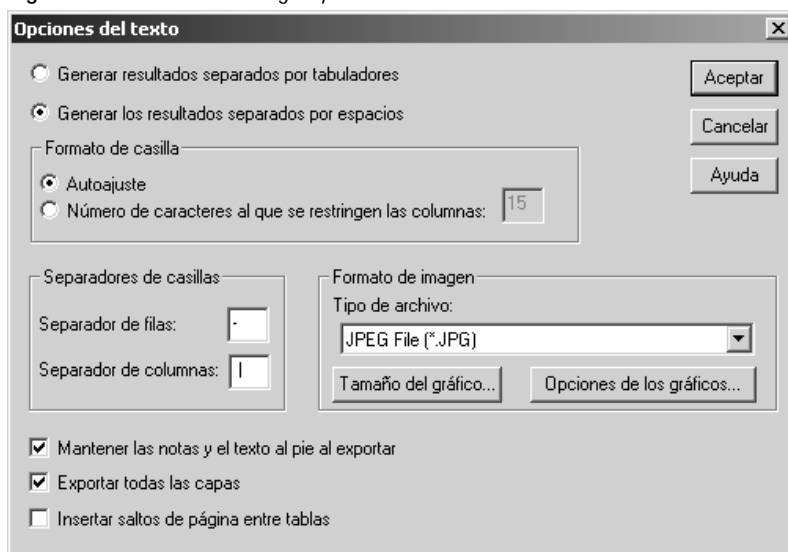
**Opciones de los gráficos.** Conduce a un subcuadro de diálogo (distinto para cada *formato de imagen* seleccionado) que contiene opciones relacionadas, básicamente, con el color y la resolución de los gráficos que se desea exportar.

" **Mantener las notas y el texto al pie al exportar.** Al activar esta opción, las tablas pivotantes exportadas incluyen las notas y el texto a pie de tabla.

" **Exportar todas las capas.** Activar esta opción para exportar todas las capas de las tablas pivotantes (si existen). Cada capa se exporta como una tabla distinta.

- Si en **Exportar** se ha seleccionado *Documento de resultados* y en **Formato de exportación** se ha seleccionado *Text File (\*.txt)*:
  - El botón **Opciones...** conduce al cuadro de diálogo *Opciones del texto* que muestra la Figura 7.26.

Figura 7.26. Cuadro de diálogo *Opciones del texto*



Las dos primeras opciones de este cuadro de diálogo permiten decidir si el texto y las tablas se exportarán como *texto delimitado por tabuladores* o *texto delimitado por espacios*. Las opciones del recuadro **Formato de casilla** permiten controlar la anchura de las casillas (automática o fija) y los caracteres que se deben utilizar como separadores de las filas y de las columnas. El resto de opciones coinciden con las descritas en el cuadro de diálogo *Opciones HTML* de la Figura 7.25, con excepción de la opción **Insertar salto de página entre las tablas**, cuyo efecto es el de exportar cada tabla en una página separada.

- Si en **Exportar** se ha seleccionado *Documento de resultados (sin gráficos)* y en **Formato de exportación** se ha seleccionado *HTML File (\*.htm)*:

- ' El botón **Opciones...** conduce al cuadro de diálogo *Opciones HTML* ya recogido en la Figura 7.25 (aunque, lógicamente, con las opciones del recuadro **Formato de imagen** no disponibles).
- Si en **Exportar** se ha seleccionado *Documento de resultados (sin gráficos)* y en **Formato de exportación** se ha seleccionado *Text File (\*.txt)*:
  - ' El botón **Opciones...** conduce al cuadro de diálogo *Opciones del texto* ya recogido en la Figura 7.26 (aunque, lógicamente, con las opciones del recuadro **Formato de imagen** no disponibles).
- Si en **Exportar** se ha seleccionado *Sólo gráficos*:
  - ' El botón **Opciones...** conduce a un cuadro de diálogo distinto para cada **Formato de exportación** seleccionado.

Recuérdese que los formatos de exportación para los gráficos son: Enhanced Metafile (\*.emf), JPEG (\*.jpe), Macintosh PICT (\*.pct), PNG File (\*.png), PostScript (\*.eps), Tagged Image File (\*.tif), Windows Bitmap (\*.bmp) y Windows Metafile (\*.wmf). Para cada uno de estos formatos de exportación (excepto para Windows Metafile) existe un cuadro de diálogo *Opciones* que ofrece diferentes posibilidades relacionadas, básicamente, con el color y la resolución de los gráficos que se desea exportar.

## Archivos de sintaxis

El SPSS permite generar y editar archivos de texto con sintaxis SPSS, es decir, archivos de texto con sentencias de programación en un lenguaje propio del SPSS. Esta sintaxis es la que de hecho ejecuta el SPSS cuando se le solicita desde los cuadros de diálogo que lleve a cabo alguna acción. Generalmente, estos archivos de *sintaxis* no son *necesarios* para trabajar con el SPSS. Sin embargo, aprender a manejar la sintaxis SPSS aporta al analista de datos dos beneficios muy destacables:

- En primer lugar, aunque la mayor parte de los procedimientos SPSS pueden ejecutarse utilizando los cuadros de diálogo a los que se accede desde la barra de menús del programa, algunas de las posibilidades del SPSS sólo están accesibles a través del lenguaje de sintaxis SPSS.
- En segundo lugar, y esto es lo verdaderamente útil, los archivos de sintaxis pueden guardarse y volverse a utilizar en sesiones diferentes. Cuando un investigador comienza a trabajar con un archivo de datos, la primera tarea que debe abordar es la de preparar las variables para el análisis. Esta preparación exige, con frecuencia, invertir gran cantidad de tiempo en etiquetar variables, recodificar algunas de las variables originales, calcular variables nuevas a partir de las existentes, etc.

Después de preparar el archivo de datos comienza el análisis. Primero suelen utilizarse procedimientos exploratorios para detectar posibles errores o pautas extrañas en los datos y para obtener una primera idea sobre las características de las variables. Tras esto, se comienza a comparar medias, a estudiar relaciones, etc. Muchos de estos análisis se hacen tras segmentar el archivo, o se aplican a sólo los casos que cumplen ciertas condiciones. Hecho esto, no es infrecuente descubrir que se ha cometido un error en alguna parte del proceso (se ha recodificado mal tal variable, tal otra variable debería crearse de otra manera, en tal análisis se quedó fuera una variable, etc.); tampoco es infrecuente verse en la necesidad de añadir nuevos casos al archivo de datos. En cualquiera de estas (u otras) circunstancias, el investigador descubre que es necesario comenzar desde el principio con todo el proceso (efectuar las mismas recodificaciones, crear las mismas nuevas variables, llevar a cabo los mismos análisis, etc.).

Pues bien, todo este trabajo extra puede evitarse si se aprende a trabajar con la *sintaxis* del SPSS.

La utilidad de los archivos de sintaxis se ve reforzada por el hecho de que, para generar la sintaxis SPSS capaz de ejecutar los procedimientos SPSS, no es necesario aprender ni una sola

regla sintáctica. El SPSS genera esa sintaxis de forma automática y permite modificarla fácilmente utilizando el *Editor de sintaxis*.

Para poder trabajar con los archivos de sintaxis hay que aprender, en primer lugar, a abrirlos y a guardarlos. En segundo lugar, hay que conocer las diferentes formas que existen de generar sintaxis automáticamente. Por último, conviene familiarizarse con algunas reglas sintácticas básicas para poder moverse cómodamente por los archivos de sintaxis.

## Abrir y guardar archivos de sintaxis

Si no existe ningún archivo (ventana) de sintaxis abierto (el hecho de que haya o no un archivo de sintaxis abierto al iniciar una sesión depende de la opción seleccionada en la pestaña **Visor** del menú **Edición > Opciones**), el SPSS abre uno automáticamente la primera vez que se pulsa el botón **Pegar** (se verá esto enseguida). Pero también es posible abrir archivos de forma manual.


Para **abrir un archivo de sintaxis nuevo**:

- Seleccionar la opción **Nuevo > Sintaxis** del menú **Archivo**. Esta opción abre un archivo nuevo (vacío) en una ventana del *Editor de sintaxis*. El primer archivo nuevo abierto durante una sesión recibe el nombre *Syntax1*. El segundo archivo abierto durante una sesión recibe el nombre *Syntax2* (independientemente de que se haya cerrado o no el anterior). Y así sucesivamente.


Para **abrir un archivo de sintaxis previamente guardado**:

- Seleccionar la opción **Abrir...** del menú **Archivo** para acceder al cuadro de diálogo *Abrir archivo* (ver Figura 3.1).
- En caso necesario, ir hasta la unidad o la carpeta donde se encuentra el archivo de sintaxis que se desea abrir.
- En el menú desplegable **Archivos de tipo**, seleccionar la opción *Sintaxis* para obtener un listado de los archivos de sintaxis disponibles. Los archivos de sintaxis tienen, por defecto, extensión *.sps*.
- Pulsar el botón **Abrir**.

Para **guardar un archivo de sintaxis que todavía no tiene nombre** (es decir, que tiene el nombre que el sistema asigna por defecto: *Syntax#*), o para guardar un archivo de sintaxis con un nombre diferente:

- Seleccionar la opción **Guardar** o **Guardar como...** del menú **Archivo** para acceder al cuadro de diálogo *Guardar como*. Este cuadro de diálogo permite asignar un nombre al archivo y seleccionar la unidad y/o la carpeta donde se desea guardar. Se obtiene el mismo resultado pulsando el botón *Guardar*  de la barra de herramientas.

Para **guardar un archivo de sintaxis que ya tiene nombre**:

- Seleccionar la opción **Guardar** del menú **Archivo**. Se obtiene el mismo resultado pulsando el botón *Guardar*  de la barra de herramientas.

## Generar sintaxis

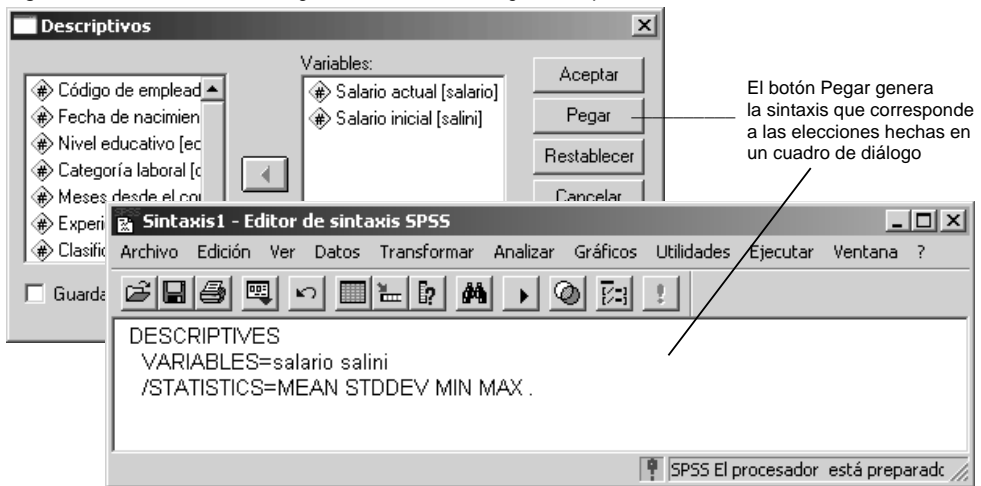
Aunque es posible abrir un archivo de sintaxis y escribir directamente en él, el SPSS ofrece tres procedimientos alternativos para obtener de forma automática la sintaxis correspondiente a un procedimiento o conjunto de procedimientos:

- El botón **Pegar** de los cuadros de diálogo del SPSS.
- Las *anotaciones* incluidas en los archivos de resultados del *Visor*.
- El archivo *spss.jnl* que el SPSS genera recogiendo todo el historial de una sesión.

### El botón *Pegar* de los cuadros de diálogo

La forma más sencilla y rápida de generar la sintaxis correspondiente a un procedimiento SPSS consiste en entrar en un cuadro de diálogo, hacer las elecciones deseadas y pulsar el botón **Pegar**. Al pulsar el botón **Pegar**, el SPSS abre una ventana de sintaxis (si es que todavía no hay ninguna abierta) y pega en ella la sintaxis correspondiente a las elecciones hechas en el cuadro de diálogo (ver Figura 8.1).

Figura 8.1. Efecto del botón *Pegar* del cuadro de diálogo *Descriptivos*



El botón **Pegar** no ejecuta ningún procedimiento SPSS (ver, más adelante, el apartado *Ejecutar sintaxis*). Únicamente pega sintaxis en la ventana designada del *Editor de sintaxis*. Al igual que ocurre con los archivos de resultados, también es posible abrir simultáneamente varios archivos o ventanas de sintaxis. En ese caso, sólo uno de ellos, el abierto en último lugar, es el archivo *designado*. No obstante, el usuario puede cambiar de archivo designado utilizando el menú *Utilidades* o el botón *Designar ventana* (icono de una ventana con una flecha) de la barra de herramientas (ver el apartado *Ventana designada versus ventana activa* del Capítulo 1).

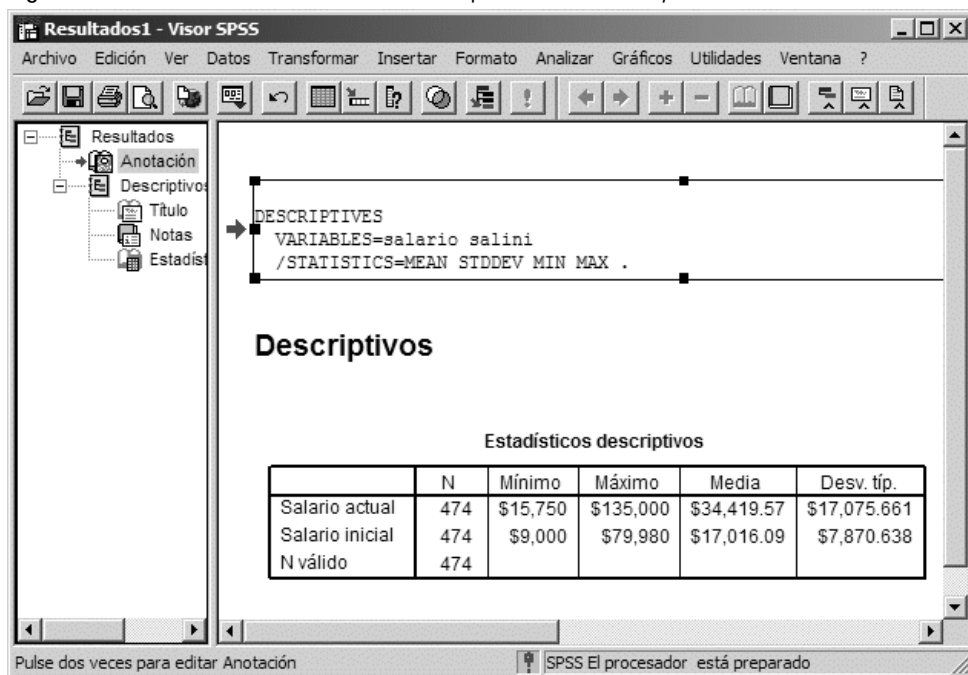
En el cuadro de diálogo *Variables* (*Utilidades > Variables*), el botón **Pegar** permite copiar en la ventana de sintaxis los nombres de las variables seleccionadas.

## Las anotaciones de los archivos de resultados

Cada vez que se ejecuta un procedimiento desde un cuadro de diálogo con el botón **Aceptar** es posible obtener, junto con cada bloque de resultados, la sintaxis que ha generado ese bloque de resultados.

La Figura 8.2 muestra, justo antes de la tabla de estadísticos descriptivos, una anotación (cuadro de texto) con la sintaxis SPSS que ha generado esa tabla de estadísticos.

Figura 8.2. Visor de resultados con la sintaxis del procedimiento *Descriptives* en una anotación



De acuerdo con las especificaciones iniciales del SPSS, los archivos de resultados no incluyen la sintaxis correspondiente a los procedimientos que se utilizan. Para que un archivo de resultados muestre la sintaxis asociada a un determinado procedimiento, es necesario dar instrucciones explícitas. Para ello:

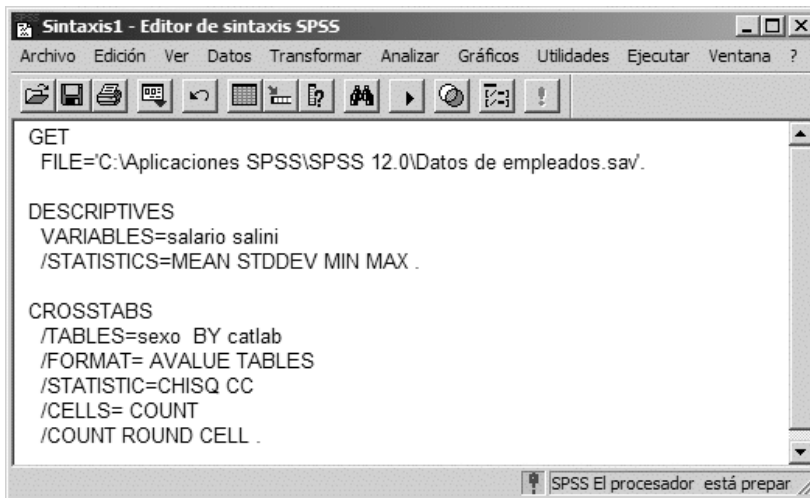
- Seleccionar **Opciones...** dentro del menú **Edición** (en la barra principal de menús) para acceder al cuadro de diálogo *Opciones*.
- Seleccionar la pestaña **Visor**.
- En la parte inferior del recuadro **Estado inicial de los resultados**, marcar la opción **Mostrar comandos en anotaciones**.

Una vez obtenida la anotación sintáctica, ésta puede copiarse y pegarse en una ventana del *Editor de sintaxis*. En ese momento, la anotación pasa a ser un archivo de sintaxis y, por tanto, el contenido puede modificarse, ejecutarse, guardarse, etc.

## El archivo *spss.jnl*

El SPSS guarda en un archivo temporal llamado *spss.jnl* toda la sintaxis correspondiente a los procedimientos utilizados durante una sesión. Este archivo temporal es una especie de *diario* (*jnl*  $\Rightarrow$  *journal*) en el que queda registrado todo el trabajo desarrollado durante una sesión con el SPSS. La Figura 8.3 muestra una parte del archivo *spss.jnl* en la que aparecen sentencias relacionadas con abrir un archivo (GET), obtener una tabla de contingencias (CROSSTAB) y calcular unos estadísticos descriptivos (DESCRIPTIVES).

Figura 8.3. Archivo *spss.jnl* con parte de la sintaxis-historial (*diario*) de una sesión



El archivo *spss.jnl* puede abrirse, editarse y guardarse exactamente igual que cualquier otro archivo de sintaxis (también tiene formato de texto estándar). Por tanto, puede utilizarse para repetir transformaciones o análisis previos de forma rápida y sencilla. Y posee la importante utilidad adicional de sacar del apuro cuando, por un descuido o por un fallo eléctrico, se ha salido del SPSS sin grabar el archivo de datos o el de resultados.

Generalmente, el archivo *spss.jnl* se encuentra en el carpeta *Windows > Temp*, pero tanto el nombre del archivo como su ubicación pueden modificarse a gusto del usuario asignando un nombre y una ubicación nuevos. Para ello:

- En el menú **Edición**, seleccionar **Opciones...** para acceder al cuadro de diálogo *Opciones*.
- Utilizar el botón **Examinar...** del recuadro **Diario de la sesión** de la pestaña **General** para asignar una ruta y/o un nombre nuevo.

El recuadro **Diario de la sesión** de la pestaña **General** también contiene opciones para decidir si se desea o no guardar en el *diario* la sintaxis de una sesión y la forma de hacerlo:

**Añadir.** El archivo *spss.jnl* va recogiendo una sesión tras otra.


**Sobrescribir.** El archivo *spss.jnl* se vacía (se pone a cero) cada vez que se inicia una nueva sesión.



## Ejecutar sintaxis

Un archivo de sintaxis puede ejecutarse de forma completa o por partes. El menú **Ejecutar** de la barra de menús del *Editor de sintaxis* ofrece varias posibilidades:

- **Todo:** ejecuta el archivo completo, independientemente de la posición del cursor.
- **Selección:** ejecuta sólo las sentencias seleccionadas. A estos efectos, una sentencia se considera seleccionada cuando lo está cualquier parte de la misma, aunque sea un solo carácter (las distintas sentencias van separadas por puntos).
- **Actual:** ejecuta sólo la sentencia en la que se encuentra el cursor.
- **Hasta el final:** ejecuta las sentencias comprendidas entre la posición del cursor y el final del documento.

El botón *Ejecutar comando actual*  de la barra de herramientas del *Editor de sintaxis* permite ejecutar, de forma rápida, las sentencias seleccionadas o, en el caso de no existir selección, la sentencia en la que se encuentra el cursor.

## Algunas reglas sintácticas básicas

Aunque, según se ha señalado, la sintaxis SPSS puede generarse de forma automática, si se desea hacer algunas modificaciones en un archivo de sintaxis conviene conocer unas pocas reglas básicas:

- Cada sentencia debe comenzar en una línea nueva y terminar con un punto.
- La mayor parte de las sub-sentencias que forman parte de una sentencia deben ir precedidas de una barra (/), aunque la primera sub-sentencia no suele necesitarla.
- Las variables se identifican por el nombre completo (no sirven las etiquetas).
- Independientemente de las especificaciones establecidas en la configuración regional de Windows, siempre debe utilizarse el punto como separador decimal.
- Pueden utilizarse tantas líneas como se desee para escribir una sola sentencia.
- Pueden utilizarse tantos espacios en blanco como se desee, e incluso cambios de línea, allí donde pueda figurar un solo espacio en blanco. Por ejemplo, entre nombres de variables, antes y después de un paréntesis o de un operador aritmético, etc.

**Segunda parte**

**Análisis estadístico  
con el SPSS**



**Segunda parte**

**Análisis estadístico  
con el SPSS**



## Introducción al análisis estadístico

### Qué es el análisis estadístico

El *análisis estadístico* o *análisis de datos* engloba un conjunto de procedimientos diseñados para (1) *seleccionar datos*, (2) *describirlos* y (3) *extraer conclusiones de ellos*. Este conjunto de procedimientos, aun siendo una *herramienta* de la que todas las ciencias empíricas (medicina, biología, psicología, sociología, economía, antropología, etc.) hacen uso, no pertenece a ninguna de ellas, sino a una rama de las matemáticas conocida con el nombre de *estadística*. Esta moderna ciencia, la estadística, es el resultado de la confluencia de dos disciplinas independientes: el *cálculo de probabilidades*, que nace como aproximación matemática a los juegos de azar, y la *estadística*, o ciencia del Estado, dedicada a llevar registros ordenados (contar, tabular, clasificar, censar, etc.) de los datos del Estado. La unión de ambas en el siglo XIX dio lugar a una nueva ciencia interesada, fundamentalmente, en estudiar cómo obtener conclusiones de la investigación empírica mediante el uso de modelos matemáticos (ver Hays, 1995) y que puede definirse como una *ciencia que recoge, ordena y analiza los datos de una muestra extraída de una determinada población, para hacer inferencias acerca de esa población valiéndose del cálculo de probabilidades* (Amón, 1979, pág. 37).

Es común encontrar la estadística dividida en dos partes diferentes: la estadística *descriptiva* y la estadística *inferencial* o *inductiva*. La *estadística descriptiva* consta de una serie de procedimientos diseñados para organizar y resumir la información contenida en un conjunto (muestra) de datos empíricos; es lo que se corresponde con lo que se ha llamado, en el primer párrafo de este apartado, *descripción* de los datos.

La *estadística inferencial* o *inductiva*, por su parte, engloba una serie de estrategias que permiten generalizar (inferir, inducir) las propiedades de ese conjunto de datos empíricos (muestra) al conjunto total de datos (población) a los que representan; se corresponde con lo que anteriormente se ha llamado *extracción de conclusiones*. Por supuesto, para poder efectuar esta generalización (inferencia) de lo concreto a lo general es imprescindible que el conjunto de datos utilizados para obtener información (muestra) sea *representativo* del conjunto total de datos (población) sobre el que se desea realizar la inferencia (es decir, es necesario efectuar una correcta *selección* de los datos). Esto se consigue mediante las técnicas de *muestreo*, las cuales, como se verá, también pertenecen al ámbito de la estadística.

En ocasiones se habla del *cálculo de probabilidades* como de una parte de la estadística; no obstante, lo habitual es considerarlo como una parte de la estadística inferencial: podría decirse que el cálculo de probabilidades es, según se tendrá ocasión de comprobar, el aparato matemático utilizado por la estadística inferencial para dar el salto (hacer inferencia) de lo concreto a lo general, de lo *observado* a lo *desconocido*.

## Para qué sirve el análisis estadístico

Las ciencias pueden ser clasificadas en *formales* y *empíricas*. En las ciencias formales (las matemáticas, por ejemplo) no hay necesidad de entrar en contacto con el mundo real; basta con establecer un conjunto de postulados sobre entidades abstractas y proceder a partir de ellos por deducción lógica.

En las ciencias empíricas, por el contrario, el objetivo fundamental es el de encontrar relaciones de tipo general (leyes) capaces de explicar el comportamiento de uno o varios eventos reales cuando se dan las circunstancias apropiadas. Y, a diferencia de lo que ocurre en las ciencias formales, esas leyes sólo pueden ser descubiertas y verificadas observando el mundo real. Sin embargo, no existe ningún científico o grupo de científicos capaces de observar todos los eventos posibles relacionados con una determinada ley. Las conclusiones sobre lo que ocurrirá con la totalidad de una clase particular de eventos se extraen a partir de la observación de sólo unos pocos eventos concretos de esa clase. Esto es lo que se conoce como *inducción* o generalización inductiva.

Mientras las leyes de la deducción lógica (propias de las ciencias formales) permiten llegar a conclusiones verdaderas a partir de premisas verdaderas, la generalización inductiva (propia de las ciencias empíricas) intenta ir desde lo que se considera que es verdad para un *conjunto* reducido de observaciones hasta la afirmación de que eso mismo es verdad también para el *total* de observaciones posibles de la misma clase.

Este *salto* de lo concreto a lo general posee un *riesgo* nada despreciable. Multitud de factores influyen sobre los eventos observables alterando las similitudes y diferencias entre ellos. Podría decirse que cada observación es, en algún sentido, diferente de la siguiente. En ciencias como la física (en algunas de sus parcelas, al menos), esta diferencia entre observaciones consecutivas es, generalmente, bastante reducida, de modo que unas pocas observaciones de un mismo evento suelen producir resultados muy parecidos, si no idénticos. Bajo estas circunstancias, la generalidad de las conclusiones obtenidas inductivamente no constituye un problema importante. Pero ese no es el caso en las demás ciencias empíricas (medicina, biología, psicología, sociología, economía, etc.). En estas ciencias, la variación existente entre las distintas observaciones de un mismo evento no puede ser sometida, habitualmente, a un control riguroso. Las fuentes de variación existentes son muy numerosas y resultan extremadamente difíciles de identificar, medir y controlar. Bajo estas circunstancias, las conclusiones a las que es posible llegar inductivamente requieren la utilización de una metodología en cierto sentido especial. Y es precisamente la estadística, mediante el conjunto de procedimientos o herramientas englobadas bajo la denominación de *análisis estadístico*, quien proporciona a las ciencias empíricas esa metodología.

La más importante aplicación del análisis estadístico está, por tanto, relacionada con el concepto de *incertidumbre*, entendida ésta como la tendencia de un resultado a variar cuando se efectúan repetidas observaciones del mismo bajo condiciones idénticas. En situaciones *deterministas*, en las que una misma causa produce siempre un mismo resultado (un cuerpo desplazado a una velocidad constante  $v$  durante un tiempo  $t$  recorre un espacio  $e$ ), el álgebra o el análisis matemático bastan para alcanzar el nivel de comprensión buscado. Por el contrario, en situaciones *aleatorias*, en las que una misma causa puede producir cualquiera de un conjunto de resultados posibles (lanzar una moneda al aire, observar la respuesta de un paciente a un tratamiento, etc.), es necesario recurrir al análisis estadístico (a las herramientas que proporciona la estadística) para poder extraer conclusiones fiables.

## Tipos de variables

El análisis estadístico o análisis de datos se basa, obviamente, en *datos*. Pero un dato no es otra cosa que un *número*. Lo cual significa que, para poder analizar datos, es necesario asignar números a las características de las personas u objetos que se desea estudiar. Sin embargo, ese proceso consistente en asignar números a las características objeto de estudio, proceso denominado *medida* o *medición*, es ajeno a la estadística. De ese proceso se encarga la *teoría de la medida*, la cual tiene por objeto el estudio de los diferentes modelos que permiten establecer las reglas que es necesario seguir para una correcta asignación de números.

Si la característica o propiedad (es decir, la *variable*) que se desea medir existe en una cierta cantidad, la medición consiste simplemente en asignar a esa variable, de acuerdo con alguna regla, un número que exprese esa cantidad con la mayor precisión posible. Así es como se hace con variables tales como la longitud o el tiempo; disponiendo de un instrumento de medida apropiado, esto no constituye un problema importante. El problema surge cuando se desea medir variables que no están tan claro que puedan ser cuantificadas. No es éste, por supuesto, el lugar adecuado para entrar en el debate histórico que ha suscitado este problema, pero sí es conveniente señalar que, gracias al persistente esfuerzo de muchos científicos, a partir del congreso sobre Medición para el Avance de la Ciencia y la Tecnología, celebrado en Moscú en 1979, la medición de variables aparentemente poco cuantificables dejó de ser prohibitiva y empezó a adquirir el reconocimiento por el que tanto tiempo estuvo luchando.

Ahora, la medición no se concibe exactamente como la asignación de un numeral que exprese la magnitud de cierta propiedad. Medir consiste en hacer corresponder dos sistemas de relaciones: uno *empírico* (el de las propiedades que se desea medir) y otro *formal* (el de los números que se asignan en la medición). Es necesario que las relaciones presentes en el sistema formal reflejen las presentes en el sistema empírico para que la correspondencia efectuada se considere una medición.

Consideremos, como ejemplo, la variable *sexo*. Para analizar datos referidos a esa variable, puede atribuirse el número 1 a la modalidad *varón* y el número 2 a la modalidad *mujer*. Consideremos ahora dos individuos y la variable sexo. O los dos individuos son varones, o los dos son mujeres, o uno es varón y el otro mujer. Desde el punto de vista del análisis de datos, tras la medición, se tendrá dos unos, dos doses, o un uno y un dos. La relación que se establezca entre estos números sólo podrá ser de igualdad o desigualdad. No se podrá, por ejemplo, establecer una relación de orden (es decir, de *mayor* o *menor*), pues el valor 2 no indica mayor *cantidad* de variable (ser mujer no indica, como es evidente, mayor posesión de la característica sexo que ser hombre, a pesar de que  $1 < 2$ ).

En este caso, los números sólo sirven para identificar o distinguir las dos modalidades de la variable sexo. Sin embargo, en otros casos, con otras variables, los números permiten establecer otro tipo de relaciones. Los números que se asignan a la variable *altura*, por ejemplo, reflejan relaciones diferentes de las que reflejan los asignados a la variable sexo. Un individuo que mide 1,80 m. posee más cantidad de la variable altura que otro sujeto que mide 1,60 m. Es decir, las variables no se miden todas de la misma forma (los números que se asignan no significan siempre lo mismo) porque entre sus valores no se da siempre el mismo tipo de relación. La medición será en unos casos *mejor* que en otros, en el sentido de que en unos casos permitirá establecer mayor número de relaciones que en otros. Por tanto, dependiendo de la riqueza de las relaciones que se puedan establecer entre los diferentes valores de una variable, existirán diferentes *niveles* o *escalas de medida*.



Tradicionalmente se han distinguido cuatro escalas o niveles de medida: nominal, ordinal, de intervalo y de razón. La medida **nominal** consiste en clasificar en categorías a los sujetos u objetos que se desea medir, de modo que todos los sujetos u objetos clasificados dentro de la misma categoría sean equivalentes respecto a la variable o propiedad que se está midiendo. Tras esto, se asignan números a las categorías establecidas. Las categorías utilizadas (que serán tantas como niveles o categorías tenga la variable que se desea medir) deben reunir dos propiedades: *exhaustividad* (todos los sujetos u objetos pueden ser clasificados en alguna de las categorías establecidas), y *exclusividad* (cada sujeto u objeto puede ser clasificado en sólo una de las categorías establecidas; las categorías no se solapan). Esta escala de medida es la más *débil* de todas: la única relación que es posible establecer entre los sujetos u objetos medidos es la de *igualdad-desigualdad*. Los números asignados actúan simplemente como nombres o rótulos para identificar cada una de las categorías establecidas: en lugar de números podría utilizarse nombres o símbolos y nada cambiaría. Hay muchos ejemplos de variables en las que sólo puede conseguirse un nivel de medida nominal: el sexo (masculino, femenino), el estado civil (soltero, casado, divorciado, etc.), el lugar de procedencia (Madrid, Galicia, Andalucía, etc.), la nacionalidad, la raza, el tipo de enfermedad, el tipo de tratamiento aplicado, el resultado de una tarea (éxito, fracaso), la actitud mantenida hacia un objeto (a favor, en contra), etc.

La medida **ordinal** consiste en asignar a los sujetos u objetos medidos un número que permita *ordenarlos* según la cantidad de variable que poseen. En la escala ordinal, además de estar presente la relación de igualdad-desigualdad propia de la escala nominal, los números asignados permiten afirmar si la cantidad de variable que posee un sujeto u objeto es *mayor que o menor que* la cantidad de variable que posee otro sujeto u objeto cualquiera. En las ciencias sociales y de la salud es frecuente encontrarse con variables en las que resulta apropiado utilizar una escala de medida ordinal. Es posible ordenar, por ejemplo, a un conjunto de sujetos según su grado de satisfacción con un determinado servicio: asignando un 1 al sujeto más satisfecho, un 2 al sujeto más satisfecho de los restantes, un 3 al siguiente, etc. Al final se tendrá  $n$  sujetos ordenados según su grado de satisfacción. Al hacer esto, ya no sólo es posible afirmar que dos sujetos a los que se les ha asignado un número diferente poseen un grado de satisfacción diferente (como se hacía en el nivel de medida nominal), sino, además, que el grado de satisfacción de tal sujeto es mayor o menor que el de tal otro. Sin embargo, no es posible afirmar nada acerca de la magnitud de la diferencia existente entre dos sujetos medidos. En la escala ordinal se desconoce si la diferencia existente entre los sujetos a los que se les ha asignado un 1 y un 2 es igual (o distinta) que la diferencia existente entre los sujetos a los que se les ha asignado un 3 y un 4. De modo que la diferencia en grado de satisfacción entre los sujetos a los que se les ha asignado un 1 y un 2 puede no ser (y normalmente, en este nivel de medida, no lo será) la misma que entre los sujetos a los que se les ha asignado un 2 y un 3.

En la medida de **intervalo**, además de poder afirmar que un objeto posee más o menos cantidad de variable que otro (relación alcanzada ya en la escala ordinal), también es posible determinar la magnitud de la diferencia existente entre dos objetos medidos, es decir, la cantidad de variable en la que difieren dos objetos. En la escala de intervalo se define una unidad de medida y, tras ello, se asigna a cada objeto medido un número indicativo de la cantidad de variable que posee en términos de las unidades de medida definidas. Así, un objeto al que se le asigna la puntuación 12 en una escala de intervalo tiene, en cantidad de variable, 2 unidades de medida más que un objeto al que se le asigna la puntuación 10; del mismo modo, un objeto al que se le asigna la puntuación 6 tiene 2 unidades de medida más que un objeto al

que se le asigna la puntuación 4. Entre 10 y 12 existe la misma diferencia, en cantidad de variable, que entre 4 y 6. Sin embargo, en la escala de intervalo no se puede afirmar que 12 es el doble de 6. En la escala de intervalo no existe el cero absoluto, es decir, no existe un valor numérico que indique ausencia total de cantidad de variable. El valor numérico 0 es un punto más de la escala, un punto arbitrario. La temperatura, por ejemplo, es una variable que se mide utilizando una escala de intervalo. Cuando se dice, en escala Celsius, que ayer hubo 20 grados de temperatura máxima y hoy 25, se está diciendo no sólo que hoy hubo más temperatura que ayer (afirmación propia de la escala ordinal), sino que hoy hubo 5 grados más de temperatura que ayer. Del mismo modo, 20 grados son 5 más que 15. La diferencia entre 15 y 20 grados es la misma que entre 20 y 25, y esto va más allá de lo que puede afirmarse con una escala ordinal. Sin embargo, no es posible afirmar que 20 grados representen el doble de temperatura que 10. Esto es debido a que, en la escala Celsius, el punto cero es un punto arbitrario de la escala y, por tanto, no indica ausencia de cantidad de variable.

La medida de *razón* añade a la de intervalo la presencia del cero absoluto: en la escala de razón el cero indica ausencia total de cantidad de variable. Aquí, el cero no es un punto arbitrario de la escala (como ocurría en la escala de intervalo: la temperatura medida en escala Celsius), sino un punto fijo: el punto que indica que no existe cantidad alguna de variable. Al igual que en la escala de intervalo, también aquí las diferencias entre los objetos medidos son constantes (existe una unidad de medida), pero, además, la presencia del cero absoluto permite afirmar si un objeto posee el doble, el triple, etc., de cantidad de variable que otro. El tiempo, la extensión, el peso, por ejemplo, son variables medidas en escala de razón. No sólo es posible afirmar que la diferencia existente entre 300 y 600 metros es la misma que entre 600 y 900 (afirmación válida también en la escala de intervalo), sino, además, que 600 metros son el doble de 300 metros.

La importancia de distinguir apropiadamente las diferentes escalas de medida radica en que la utilización de las técnicas de análisis de datos está, en buena medida, mediatizada por el tipo de mediciones de que se dispone. No obstante, a pesar de la necesidad de distinguir apropiadamente las diferentes escalas de medida, existen multitud de variables de diferente índole en las que no resulta nada fácil determinar el nivel de medida alcanzado. El hecho de que las cuatro escalas de medida descritas sean exhaustivas (cualquier variable puede ser medida con alguna de ellas) y mutuamente exclusivas (no se solapan), constituye un verdadero problema a la hora de trabajar con algunas variables.

Supongamos que se mide la *percepción subjetiva de dolor* de 3 sujetos en una escala de 0 a 100, y que se obtiene una puntuación de 10 para el primero de ellos, de 20 para el segundo y de 90 para el tercero. Si se interpretan las escalas de medida en sentido estricto, no se podrá considerar que la distancia existente entre un *dolor* de 10 y otro de 20 (10 puntos) sea equivalente a la distancia que existe entre *dolor* de 50 y otro de 60 (también 10 puntos). Y no se podrán considerar equivalentes esas distancias porque en una escala de percepción subjetiva del dolor no existe una unidad de medida que garantice tal equivalencia. Según esto, se debería considerar que la medida de *dolor* obtenida se encuentra a nivel ordinal, lo que permitiría concluir, tan sólo, que al tercer sujeto le *duele* más que al segundo y a éste más que al primero. Sin embargo, si se solicitara opinión a algunos expertos, seguramente contestarían que el tercer sujeto (90) manifiesta más *dolor* que los otros dos (10 y 20) y que las respuestas de estos dos sujetos se parecen entre sí más de lo que se parecen a la respuesta del tercero (lo cual excede el alcance de las propiedades de la escala ordinal). Es razonable pensar, según esto, que una escala de percepción subjetiva del dolor (al igual que otras muchas) no puede identificarse con la escala ordinal común. Desde el punto de vista práctico, entre la escala ordinal y la

de intervalo existe una línea de separación bastante difusa que hace que muchas variables de naturaleza ordinal puedan ser analizadas como variables de intervalo.

Para terminar este apartado conviene insistir en una idea importante. En principio, cualquier conjunto de números es susceptible de ser manipulado por cualquiera de las técnicas de análisis de datos que se describen en este libro: es decir, no existe ninguna técnica de análisis de datos cuya mecánica no pueda seguirse por el motivo de que los números asignados al efectuar la medición sean o no los apropiados. Pero una técnica de análisis de datos no quita ni pone significado a los números que manipula. El hecho de que los números asignados en la medición posean o no algún significado no es un problema que pueda resolverse con la utilización de una u otra técnica de análisis, sino desde la teoría de la medida y desde el conocimiento por parte del investigador de las propiedades de las variables estudiadas. Por esta razón, es muy importante conocer la problemática relacionada con las escalas de medida: el conocimiento de esta problemática puede servir, al menos, para saber si, con los números disponibles, tiene o no sentido efectuar determinado tipo de operaciones.

## Conceptos básicos

El objetivo fundamental del análisis estadístico es el de extraer conclusiones de tipo general a partir de unos pocos datos particulares. A este salto de lo concreto a lo general es a lo que se llama, según se verá enseguida, *inferencia estadística*. Este salto exige la utilización de, por un lado, procedimientos que ayuden a efectuarlo correctamente y, por otro, procedimientos que garanticen que el salto se apoya en una buena base. Tan importante como disponer de una buena técnica de análisis de datos (para realizar la inferencia) es *seleccionar* apropiadamente los datos que se van a analizar (para proporcionar una buena base de apoyo a la inferencia). Las *técnicas de muestreo* se encargan de garantizar que la inferencia se apoya en una buena base. Y las herramientas estadísticas englobadas bajo la denominación general de *análisis estadístico* se encargan de garantizar que la inferencia se desarrolla correctamente. De todo ello se hablará en este capítulo, pero antes de seguir adelante es necesario repasar algunos conceptos fundamentales de especial utilidad para entender lo demás.

## Población

Una *población* (o *universo*) es un conjunto de elementos (sujetos, objetos, entidades abstractas, etc.) que poseen una o más características específicas en común.

En general, el término población hace referencia al conjunto total de elementos que se desea estudiar, de manera que una población queda definida cuando se hace explícita la característica (o características) que esos elementos comparten. Ejemplos de poblaciones son: las personas empadronadas en una comunidad autónoma, todos los varones mayores de 30 años, los pacientes que sufren hipertensión, las posibles respuestas que un sujeto podría emitir en una escala de satisfacción, los diferentes tipos de terapia disponibles para afrontar el tratamiento de un determinado trastorno, el censo de votantes en unas elecciones, los números múltiplos de 3; etc.

Dependiendo del número de elementos de que constan, unas poblaciones son *finitas* (contienen un número finito de elementos) y otras *infinitas* (contienen un número infinito de ele-

mentos). Normalmente, las poblaciones con las que tiene sentido trabajar son finitas, pero tan grandes que a todos los efectos podrán ser consideradas infinitas. Es precisamente el hecho de que las poblaciones, por lo general, sean infinitas o estén formadas por un gran número de elementos, lo que hace que la descripción exacta de sus propiedades sea un objetivo prácticamente inaccesible. Por esta razón, lo habitual es trabajar con muestras.

## Muestra

Una *muestra* es un subconjunto de elementos de una población. A diferencia de las poblaciones, que suelen ser conjuntos de elementos de gran tamaño, las muestras suelen ser conjuntos de elementos de tamaño reducido. Por supuesto, para poder describir con exactitud las propiedades de una población cualquiera, sería necesario examinar cada uno de los elementos que componen esa población. Pero, dado que las poblaciones que habitualmente interesa estudiar son tan grandes que, normalmente, resulta muy difícil (si no imposible) tener acceso a todos sus elementos, son las muestras quienes proporcionan toda la información necesaria para poder describir las propiedades de las poblaciones objeto de estudio.

El conocimiento que se va generando en la vida cotidiana acerca del mundo está, muy frecuentemente, basado en muestras: con sólo comer una vez en un restaurante nos formamos una opinión acerca de la calidad de su cocina y de su servicio; con sólo conocer a un par de personas de un determinado colectivo nos formamos una idea sobre el tipo de personas que forman ese colectivo; etc. Con el análisis de datos se hace algo parecido: se extraen conclusiones referidas a todos los elementos (población) a partir de la observación de sólo unos pocos elementos (muestra). Ahora bien, para que esto sea posible, es necesario, que la muestra utilizada sea *representativa* de la población; esto se consigue mediante las técnicas de *muestreo* (más adelante, en este mismo capítulo, se incluye un apartado sobre muestreo).

## Parámetro

Un *parámetro* es un valor numérico que describe una característica poblacional. Ya se ha definido una población como un conjunto de elementos que poseen una o más características en común. Pero los elementos de una población poseen, además, otras muchas características que no comparten o en las que no coinciden. La población de varones españoles mayores de 30 años está formada por elementos que tienen en común *ser varones, españoles y de edad superior a 30 años*, pero en esa población es posible considerar otras muchas características en las que no todos los elementos poblacionales coinciden: el estado civil, el nivel educativo, el peso, la altura, la presión arterial, la actitud hacia la eutanasia, etc. Al medir, por ejemplo, el *nivel de colesterol en sangre*, se obtendrán tantos valores numéricos como elementos formen parte de la población (suponiendo que se tenga acceso a todos los elementos). Si ahora se calcula el promedio (un solo número) de todos esos valores numéricos se habrá definido un parámetro, pues se habrá descrito numéricamente una característica de la población: *el nivel de colesterol medio* de los varones españoles mayores de 30 años.

En la población de clientes de un servicio, todos los elementos de esa población comparten una característica específica: *son clientes de ese servicio*. Pero existen, obviamente, otras características que no comparten. Por ejemplo, puede ocurrir que unos clientes sean varones y otros mujeres. Si se tuviera acceso a todos los elementos de esa población, se podría contar

el número de clientes que son varones (o mujeres) y eso permitiría definir un parámetro; es decir, permitiría describir numéricamente una característica de la población: *la proporción de varones (o mujeres)* en la población de clientes de un servicio.

Así pues, existen valores numéricos como la media o la proporción (además de otros muchos), que cuando se refieren a alguna característica poblacional reciben el nombre de parámetros.

Hay algunos aspectos de los parámetros que interesa resaltar. En primer lugar, los parámetros son, en general, valores poblacionales *desconocidos*: puesto que las poblaciones con las que se suele trabajar son tan grandes que sus elementos raramente resultan accesibles en su totalidad, no es posible calcular un valor numérico basado en el total de los elementos. En segundo lugar, los parámetros son valores numéricos *constantes* (es decir, no son variables): definida una población cualquiera y un parámetro en ella, ese parámetro sólo puede tomar un valor numérico concreto (la *proporción de varones* en la población de clientes de un servicio viene determinada por el número de varones que son clientes de ese servicio).

## Estadístico

Un *estadístico* es un valor numérico que describe una característica muestral. Se acaba de ver que en una población cualquiera, además de las características que la definen y que son comunes a todos los elementos, es posible definir otras muchas características en las que no todos los elementos coinciden. De una muestra, lógicamente, cabe decir lo mismo. Y una vez definida una o más de esas características en las que no todos los elementos coinciden, es posible obtener un valor numérico que las describa: a ese valor numérico se le llama *estadístico*.

De la población de varones españoles mayores de 30 años puede extraerse una muestra de  $n$  sujetos. En esa muestra de  $n$  sujetos se puede definir y medir, por ejemplo, el nivel de colesterol en sangre. Hecho esto, es posible realizar diferentes transformaciones con las puntuaciones obtenidas. Cada una de estas transformaciones es un valor numérico que describe un aspecto diferente de la característica medida (el nivel de colesterol en sangre). Es decir, cada una de estas transformaciones es un estadístico. Pero no todos los estadísticos poseen la misma utilidad. Hay algunos, como la *media*, la *mediana*, la *varianza*, la *proporción*, la *correlación*, etc., cuya utilidad ya ha sido contrastada.

Definido un estadístico, cualquiera que éste sea, su valor concreto depende de los valores concretos de la muestra en la que es calculado. Pero es evidente que de una población cualquiera es posible extraer más de una muestra diferente del mismo tamaño. Es decir, el valor de un estadístico *varía* de una muestra a otra. Esto quiere decir que un estadístico no es un valor numérico constante (como lo es un parámetro), sino que es una *variable*: su valor concreto depende de la muestra en la que es calculado.

Resumiendo, mientras un parámetro es un valor poblacional, un estadístico es un valor muestral; mientras un parámetro es, por lo general, desconocido, un estadístico es calculable a partir de unos datos muestrales y, por tanto, conocido; mientras un parámetro es un valor numérico constante, un estadístico es una variable.

Estas diferencias se hacen patentes en la notación habitualmente utilizada para representar a unos y a otros. Mientras que los parámetros se suelen representar con letras griegas minúsculas ( $\mu$ ,  $\sigma$ ,  $\pi$ ,  $\beta$ , etc.), los estadísticos se suelen representar con letras latinas mayúsculas ( $\bar{X}$ ,  $S$ ,  $P$ ,  $B$ , etc.).

## Muestreo

Para extraer conclusiones sobre las propiedades de una población a partir de la información contenida en una muestra extraída de esa población es necesario, en primer lugar, utilizar muestras *representativas del total de la población*, es decir, muestras en las que exista alguna garantía de que *cualquier* elemento de la población *ha podido* (ha tenido la oportunidad de) estar representado en ellas. El no trabajar con muestras apropiadas llevará inevitablemente a que nuestras predicciones estén, ya desde el principio, condenadas al fracaso.

El término *muestreo* se refiere al *proceso seguido para extraer una muestra de una población*. El muestreo puede ser de dos tipos: probabilístico y no-probabilístico. En el muestreo *probabilístico* se conoce (o puede calcularse) la probabilidad asociada a cada una de las muestras que es posible extraer de una determinada población; cada elemento poblacional posee una probabilidad conocida (o calculable) de pertenecer a la muestra. En el muestreo *no-probabilístico* se desconoce o no se tiene en cuenta la probabilidad asociada a cada una de las muestras posibles; el investigador selecciona aquella muestra que, en su opinión, más representativa le parece o, simplemente, aquella que considera que puede extraer con mayor comodidad o menor coste (voluntarios que responden a un anuncio, alumnos matriculados en un curso, clientes que compran un producto, pacientes que acuden a un servicio, etc.). Lógicamente, sólo el muestreo probabilístico, por estar basado en la teoría de la probabilidad, permite obtener una idea sobre el grado de representatividad de una muestra. Por tanto, sólo él proporciona una base adecuada para inducir las propiedades de una población a partir de una muestra.

Esto no significa que el muestreo no probabilístico no pueda generar muestras representativas; lo que ocurre es que al utilizar un muestreo de tipo no probabilístico no se tiene ninguna información sobre el grado de representatividad de la muestra obtenida. En consecuencia, ya desde ahora, se dejará a un lado el muestreo no-probabilístico y se considerará en todo momento que los datos de que se dispone constituyen una muestra aleatoriamente seleccionada de su respectiva población, es decir, una muestra *aleatoria*.

Ahora bien, aunque el muestreo aleatorio permite obtener una muestra apropiada (representativa de la población) en la mayor parte de los contextos, en ocasiones es posible que surja la necesidad de trabajar con poblaciones cuyas características estén aconsejando una estrategia diferente. No es este el lugar para describir con detalle cada uno de los tipos de muestreo aleatorio que existen, pero sí parece recomendable ofrecer una breve descripción de los más utilizados.

### Muestreo aleatorio sistemático

En el muestreo aleatorio sistemático se comienza elaborando una lista con los  $N$  elementos poblacionales numerados de 1 a  $N$ . A continuación se determina el tamaño de la muestra que se desea obtener ( $n$ ) y se efectúa una extracción al azar de entre los  $k = N/n$  primeros elementos (si  $k$  no es un número entero se redondea al entero más próximo). Llamemos  $i$  al lugar ocupado por ese primer elemento extraído. Hecho esto, el resto de los  $n-1$  elementos que configurarán la muestra se obtienen a partir de  $k$ : la muestra estará formada por los elementos poblacionales que ocupen las posiciones  $i, i+k, i+2k, i+3k, \dots, i+(n-1)k$ .

Así, si se desea extraer una muestra aleatoria de tamaño 100 de una población formada por 2.000 elementos, se comenzará elaborando una lista asignando a cada elemento un número de 1 a 2.000. La constante que se deberá utilizar será  $k = N/n = 2.000/100 = 20$ . Después,

se seleccionará al azar un elemento entre los 20 primeros. Si, por ejemplo, el elemento seleccionado es el que ocupa la posición  $i = 9$ , el resto de los elementos de la muestra serán los que ocupen en la lista las posiciones 29, 49, 69, 89, ..., 1949, 1969, 1989. Obviamente, la utilización de este tipo de muestreo cobra especial sentido cuando se dispone de un listado de toda la población y se desea obtener una muestra aleatoria homogéneamente distribuida a lo largo de toda la lista.

## Muestreo aleatorio estratificado

Una población puede estar formada por diferentes subpoblaciones o *estratos*. En la población de varones españoles mayores de 30 años, por ejemplo, se pueden definir diferentes estratos: según el nivel socioeconómico, según el tipo de profesión, según el nivel de estudios, según el estado civil, según el sexo, etc. Es posible que, en ocasiones, interese utilizar una muestra en la que todos los estratos de la población tengan una adecuada representación. Con el muestreo aleatorio simple existe la posibilidad de que, al extraer una muestra aleatoria, alguno de los estratos no esté suficientemente representado en la muestra (particularmente si existen estratos de muy distinto tamaño). En estos casos resulta útil hacer uso del muestreo aleatorio estratificado.

Se comienza definiendo los estratos e identificando los elementos que pertenecen a cada estrato. Se tendrán, de esta forma,  $k$  estratos con tamaños  $N_1, N_2, \dots, N_k$  (obviamente,  $N_1 + N_2 + \dots + N_k = N$ ). A continuación se elaboran  $k$  listas (una por estrato) con los elementos de cada estrato debidamente numerados y se procede a extraer aleatoriamente una muestra de cada estrato mediante muestreo aleatorio simple o mediante muestreo aleatorio sistemático. La muestra total estará formada por las  $k$  submuestras extraídas.

El tamaño de las submuestras puede o no ser proporcional al tamaño de los estratos. Si la variabilidad de la característica estudiada es similar en todos los estratos, el tamaño de las submuestras se fija de forma proporcional al tamaño de los estratos: *afijación proporcional*. Si esa variabilidad cambia ostensiblemente de estrato a estrato conviene extraer submuestras más grandes de los estratos con mayor varianza: *afijación no proporcional*.

Si se quiere, por ejemplo, extraer una muestra aleatoria de tamaño 100 de una población de 20.000 personas formada por un 40 % de varones y un 60 % de mujeres y se desea que esas proporciones poblacionales se mantengan en la muestra (afijación proporcional), se deben formar dos estratos (es decir, dos grupos: uno con los varones y otro con las mujeres) y seleccionar aleatoriamente 40 sujetos del primer estrato y 60 del segundo. Si la varianza de los varones en la variable estudiada fuera muy diferente de la de las mujeres (lo que sólo se puede saber si se conocen o se estiman tales varianzas poblacionales), se debería seleccionar más sujetos del estrato con mayor varianza.

## Muestreo aleatorio por conglomerados

En este tipo de muestreo las unidades muestrales no son elementos individuales, sino grupos de elementos a los que se les llama *conglomerados*. En lugar de considerar que la población está formada por  $N$  elementos se considera que está formada por  $k$  conjuntos o conglomerados de elementos. La forma de proceder consiste en seleccionar aleatoriamente uno o varios de esos conglomerados y aceptar como muestra el conjunto de *todos* los elementos que forman parte de ese o esos conglomerados seleccionados.

En un estudio sobre desarrollo cognitivo se toma como población de referencia la de todos los alumnos de educación primaria de la comunidad de Madrid. En lugar de seleccionar una muestra aleatoria de un listado de todos los alumnos de educación primaria, se podría seleccionar uno o varios colegios y utilizar como muestra todos los alumnos de los colegios seleccionados. Las ventajas de este muestreo son evidentes cuando se trabaja con poblaciones muy grandes: no se necesita un listado de todos los elementos de la población, sino sólo de aquellos que forman parte de los conglomerados seleccionados.

En el muestreo aleatorio por conglomerados puede procederse por etapas; se habla entonces de muestreo *polietápico*. En la primera etapa se divide la población en  $k$  conglomerados y se selecciona uno o varios de ellos (unidades muestrales primarias). En la segunda etapa, los conglomerados seleccionados se dividen en conglomerados más pequeños y se vuelve a seleccionar uno o varios de ellos (unidades muestrales secundarias). Se continúa así hasta que se considera necesario. La muestra definitiva la componen todos los elementos de los conglomerados definitivamente seleccionados en la última etapa. Obviamente, al proceder por etapas sólo es necesario un listado de los elementos que forman parte de los conglomerados seleccionados en la última etapa.

Si en el estudio anterior sobre desarrollo cognitivo la población de referencia fuese la de todos los alumnos españoles de enseñanza primaria, se podría comenzar seleccionando unas pocas comunidades autónomas; después, una provincia de cada comunidad autónoma seleccionada; después, un pueblo o ciudad de esa provincia; por último, un colegio de cada pueblo o ciudad seleccionados.

Proceder por etapas posee la ventaja de que, en cada etapa, dependiendo de las características de los conglomerados que se van a muestrear, es posible utilizar cualquiera de los métodos de muestreo aleatorios estudiados: simple, sistemático o estratificado.

## Distribución muestral

La inferencia estadística es un tipo de razonamiento que procede de lo concreto a lo general: intenta extraer conclusiones sobre los *parámetros* de una población a partir de la información contenida en los *estadísticos* de una muestra procedente de esa población. Ese razonamiento está basado en el conocimiento de la *variabilidad de un estadístico de una muestra a otra*, es decir, en el conocimiento de cómo un estadístico se comporta en las diferentes muestras que es posible extraer de una determinada población. El concepto de distribución muestral se refiere precisamente al comportamiento de un estadístico.

Los estadísticos son variables aleatorias. Como tales, tienen, al igual que cualquier variable aleatoria, su propia función de probabilidad. Pues bien, el término *distribución muestral* se refiere a la *función de probabilidad (o de densidad de probabilidad) de un estadístico*. Por tanto, una distribución muestral puede definirse como *una distribución teórica que asigna una probabilidad concreta a cada uno de los valores que puede tomar un estadístico en todas las muestras del mismo tamaño que es posible extraer de una determinada población*.

### Un caso concreto

Uno de los estadísticos más utilizados es la media aritmética:  $\bar{X}$ . En cuanto estadístico que es, su valor depende de la muestra concreta en la que es calculado. De una población cual-



quiera es posible extraer más de una muestra de tamaño  $n$  (en una población infinita es posible extraer un número infinito de muestras de cualquier tamaño). Si en cada una de las muestras se calcula  $\bar{X}$ , podrá comprobarse que no siempre toma el mismo valor, sino que varía de una muestra a otra.

El ejemplo que se propone a continuación\* para explicar la distribución muestral del estadístico  $\bar{X}$  está basado en una población formada por  $N = 5$  puntuaciones:  $X_i = \{1, 2, 3, 4, 5\}$ . Si de esa población se seleccionan aleatoriamente y con reposición todas las muestras posibles de tamaño  $n = 2$ , se obtendrán  $N^n = 5^2 = 25$  muestras (variaciones con repetición de  $N$  elementos tomados de  $n$  en  $n$ ), todas las cuales tendrán la misma probabilidad de ser extraídas:  $1/25$ . Si ahora se calcula en cada una de esas 25 muestras el estadístico  $\bar{X}$ , se llegará al resultado presentado en la Tabla 9.1. En la tabla aparecen las 25 posibles muestras y el valor que toma el estadístico  $\bar{X}$  en cada una de ellas.

**Tabla 9.1.** Muestras de tamaño  $n = 2$  que es posible extraer con reposición de una población de  $N = 5$  elementos, y valor del estadístico  $\bar{X}$  en cada una de ellas

Muestras	Valores muestrales	$\bar{X}$
1	1, 1	1,0
2	1, 2	1,5
3	1, 3	2,0
4	1, 4	2,5
5	1, 5	3,0
6	2, 1	1,5
7	2, 2	2,0
8	2, 3	2,5
9	2, 4	3,0
10	2, 5	3,5
11	3, 1	2,0
12	3, 2	2,5
13	3, 3	3,0
14	3, 4	3,5
15	3, 5	4,0
16	4, 1	2,5
17	4, 2	3,0
18	4, 3	3,5
19	4, 4	4,0
20	4, 5	4,5
21	5, 1	3,0
22	5, 2	3,5
23	5, 3	4,0
24	5, 4	4,5
25	5, 5	5,0

\* El ejemplo utilizado en este apartado es a todas luces un ejemplo irreal sin ningún tipo de relación con la investigación empírica. Sin embargo, su simplicidad le confiere la virtud de permitir explicar con claridad el importantísimo concepto de distribución muestral.

En la Tabla 9.1 es posible observar diferentes cosas. Por ejemplo, aunque en sólo una de las 25 muestras se obtiene  $\bar{X} = 1$ , en cuatro de ellas se obtiene  $\bar{X} = 2,5$ . Lo cual significa que el estadístico  $\bar{X}$  puede tomar el mismo valor en más de una muestra diferente. Por tanto, aunque las 25 muestras son equiprobables (pues todas tienen la misma probabilidad de ocurrir), los posibles valores del estadístico  $\bar{X}$  no lo son: hay unos valores de  $\bar{X}$  que son más probables que otros porque unos pueden obtenerse en mayor número de muestras que otros; puede comprobarse en los valores de la tabla que, efectivamente, existen más muestras en las que se obtiene, por ejemplo,  $\bar{X} = 2,5$  que  $\bar{X} = 1,5$ .

Estas consideraciones sugieren que los datos de la Tabla 9.1 pueden resumirse tal como se muestra en la Tabla 9.2. En ella aparecen los distintos valores que puede tomar el estadístico  $\bar{X}$  y la probabilidad asociada a cada uno de esos valores; es decir, los diferentes valores de la variable aleatoria  $\bar{X}$  y su función de probabilidad: *la distribución muestral de la media*.

**Tabla 9.2.** Distribución muestral de la media formada a partir de los valores de la Tabla 9.1

<i>Nº de muestras</i>	$\bar{X}$	$f(\bar{x})$
1	1,0	1/25
2	1,5	2/25
3	2,0	3/25
4	2,5	4/25
5	3,0	5/25
4	3,5	4/25
3	4,0	3/25
2	4,5	2/25
1	5,0	1/25
1		

La distribución muestral de la media puede obtenerse por procedimientos puramente matemáticos, sin necesidad de tener que extraer todas las posibles muestras de tamaño  $n$  de una determinada población (lo cual, por otra parte, resultaría imposible si se estuviera trabajando con una población infinita). Sin embargo, la obtención de la distribución muestral de un estadístico a partir de la extracción de todas las posibles muestras de tamaño  $n$  tiene la ventaja de ayudar a reparar en ciertos detalles que de otro modo podrían pasar desapercibidos. En la Tabla 9.2 puede comprobarse, por ejemplo, que, si de una población formada por los elementos 1, 2, 3, 4 y 5, se selecciona aleatoriamente una muestra de tamaño 2, lo más probable es que el estadístico  $\bar{X}$  tome el valor 3, pues  $P(\bar{X} = 3) = 5/25 = 0,20$  es la probabilidad más alta de entre todas las asociadas a los diferentes valores de  $\bar{X}$  (curiosamente, la media de la población vale  $\mu = (1+2+3+4+5)/5 = 3$ ). Se sabe que, si se decide utilizar el estadístico  $\bar{X}$  para estimar el parámetro  $\mu$ , existe una probabilidad de 0,20 de que el valor de  $\bar{X}$  sea exactamente el mismo que el valor de  $\mu$ ; es decir, existe una probabilidad de 0,20 de efectuar una estimación correcta. Pero también se sabe, por ejemplo, que la probabilidad de que una estimación concreta se separe del verdadero valor de  $\mu$  en medio punto como máximo ( $3 \pm 0,5$ ) vale  $4/25 + 5/25 + 4/25 = 13/25 = 0,52$  (es decir, la suma de las probabilidades asociadas a los valores 2,5, 3 y 3,5 de  $\bar{X}$ ). El razonamiento puede seguirse argumentando que la probabilidad de que el valor  $\bar{X}$  obtenido en una muestra concreta, con  $n=2$ , no se separe del verdadero valor de  $\mu$  en más de

1 punto ( $3 \pm 1$ ) vale  $3/25 + 4/25 + 5/25 + 4/25 + 3/25 = 19/25 = 0,76$ . Etc. Todo esto sirve para recordar una vez más que la distribución muestral de un estadístico, en cuanto función de probabilidad que es, ofrece la probabilidad asociada a cada uno de los valores que ese estadístico puede tomar en las diferentes muestras de tamaño  $n$  que es posible extraer de una determinada población.

## El caso general

Según lo visto en el apartado anterior, parece claro que con poblaciones y muestras pequeñas resulta relativamente sencillo conocer la distribución muestral de un estadístico y, a partir de ella, el comportamiento de ese estadístico en las diferentes muestras de tamaño  $n$  en las que puede ser calculado. Pero ocurre que las poblaciones con las que habitualmente es necesario trabajar no son, ni mucho menos, tan pequeñas como las del ejemplo del apartado anterior. Lo que generalmente ocurre es, más bien, todo lo contrario: las poblaciones que habitualmente tiene sentido estudiar suelen ser muy grandes e incluso, en ocasiones, infinitas. Lo cual significa que, para obtener la distribución muestral de un estadístico cualquiera, por simple que éste sea, no resulta posible proceder de la forma que se ha hecho hasta ahora.

Sin embargo, el concepto de distribución muestral sigue siendo el mismo cualquiera que sea el tamaño de la población y de la muestra con las que se trabaje. En una población infinita, la distribución muestral de, por ejemplo, el estadístico  $\bar{X}$  sigue siendo la distribución resultante de extraer infinitas muestras de tamaño  $n$  y calcular en todas ellas  $\bar{X}$ . Por supuesto, no es posible extraer las infinitas muestras de tamaño  $n$  de una población para conocer la distribución muestral de un estadístico, pero eso no significa que se tenga que renunciar a conocer la distribución muestral de un estadístico. Cuando se está trabajando con una población infinita (o muy grande) es posible utilizar procedimientos matemáticos que informan con exactitud sobre las características de las distribuciones muestrales de diferentes estadísticos.

Una distribución muestral es la función de probabilidad de un estadístico. Como tal, una distribución muestral puede quedar caracterizada, al igual que cualquier otra distribución de probabilidad, haciendo explícitas su forma, su valor esperado y su varianza. Y dado que un estadístico es una variable aleatoria, su valor esperado y su varianza pueden ser definidos de la forma en que se acostumbra a definir el valor esperado y la varianza de cualquier variable aleatoria:

$$E(X) = \sum Xf(x) \quad (\text{variables discretas})$$

$$E(X) = \int Xf(x) d(x) \quad (\text{variables continuas})$$

$$\sigma_x^2 = E(X^2) - [E(X)]^2$$

Conviene señalar por último que, cuando se está trabajando con distribuciones muestrales, es habitual utilizar, en lugar de la varianza ( $\sigma_x^2$ ), la desviación típica ( $\sigma_x$ ); y es habitual, también, para referirse a esa desviación típica, utilizar el término *error típico*. Así pues, a la desviación típica de la distribución muestral de la media se le llama *error típico de la media*:  $\sigma_{\bar{x}}$ . Y lo mismo cabe decir de cualquier otra distribución muestral: para nombrar, por ejemplo, la desviación típica de la distribución muestral del estadístico *proporción* se utiliza la expresión *error típico de la proporción*:  $\sigma_p$ . Etc.

## La inferencia estadística

Ya se ha señalado repetidamente que el objetivo final del análisis de datos es el de extraer conclusiones de tipo general a partir de unos pocos datos particulares. Pues bien, ya se han repasado suficientes conceptos como para poder concretar que al hablar de *conclusiones de tipo general* se está queriendo decir alguna *población* y/o alguno de sus *parámetros*, y al hablar de *datos particulares* se está queriendo decir alguna *muestra* de esa población y/o alguno de sus *estadísticos*.

También se ha señalado que lo que garantiza que la muestra obtenida es representativa de la población es la utilización de un método de muestreo apropiado. Esto significa que la información contenida en una muestra es, en mayor o menor medida, un reflejo de la información que caracteriza a la población. Es posible, por tanto, utilizar la información muestral para formarse una idea sobre las propiedades de la población. Es decir, es posible utilizar muestras para hacer *inferencias* sobre poblaciones.

Estas inferencias pueden realizarse utilizando dos estrategias distintas: la *estimación de parámetros* y el *contraste de hipótesis*. Ambas formas de inferencia son equivalentes en el sentido de que ambas permiten llegar a la misma conclusión. Podría pensarse en ellas como en las dos caras de una misma moneda.

El contraste de hipótesis ha constituido, tradicionalmente, la esencia de lo que se conoce como análisis estadístico. Sin embargo, en la década de los noventa ha habido una fuerte corriente crítica que ha insistido en la importancia de acompañar cualquier contraste con su correspondiente estimación (ver Chow, 1996; Hagen, 1997; Harlow, Mulaik y Steiger, 1997). A pesar de que ambas estrategias son equivalentes, la información que ofrecen es algo diferente: mientras el contraste de hipótesis pone el énfasis en intentar detectar la presencia de un efecto significativo (grupos que difieren, variables que correlacionan, etc.), la estimación de parámetros pone el énfasis más bien en intentar cuantificar el tamaño del efecto detectado (cuánto difieren dos grupos, cómo de intensa es la relación entre dos variables, etc.). En los apartados que siguen se explica la lógica general en la que se basan ambas estrategias.

### El contraste de hipótesis

El *contraste de hipótesis* puede ser entendido como un método de *toma de decisiones*: un contraste de hipótesis, también llamado *prueba de significación* o *prueba estadística*, es un procedimiento que permite decidir si una proposición acerca de una población puede ser mantenida o debe ser rechazada.

En la investigación empírica es frecuente encontrarse con *problemas de conocimiento* surgidos a partir de conocimientos ya existentes o a partir de la observación de nuevas situaciones: ¿Es la técnica terapéutica *a* más apropiada que la *b* para aliviar los síntomas de los pacientes con problemas vasculares? ¿Son los sujetos que se sienten inseguros más agresivos que los que se sienten seguros? ¿Difieren los varones y las mujeres en intención de voto? Etc. Estos interrogantes son sólo un pequeño ejemplo de la multitud de *problemas* que se generan en la investigación empírica. Tales interrogantes surgen, en general, en el seno de una teoría que intenta dar cuenta de alguna parcela de la realidad y se plantean con la intención de cubrir alguna laguna concreta de conocimiento que esa teoría no cubre o para corroborar una parte o el total de esa teoría.

Surgido el problema, el paso siguiente consiste en aventurar algún tipo de solución al mismo. Esta solución provisional suele tomar forma de afirmación directamente verificable (es decir, empíricamente contrastable; de no ser así, nos moveríamos en el terreno de la especulación y no en el de la ciencia) en la que se establece de forma operativa el comportamiento de la variable o las variables involucradas en el problema. Esa afirmación verificable recibe el nombre de *hipótesis científica*. Así, ante la pregunta (problema de conocimiento) «¿difieren los varones y las mujeres en el nivel medio de colesterol en sangre?», podría aventurarse la hipótesis de que «los varones no difieren de las mujeres». Por supuesto, se debería definir con precisión (operativamente) qué se entiende por «nivel de colesterol en sangre» y cómo medirlo. Sólo entonces la afirmación sería una hipótesis científica.

Hecho esto, ya se estaría en condiciones de iniciar el proceso de verificación de esa hipótesis. Y el proceso de verificación habitualmente utilizado en las ciencias empíricas sigue los pasos que en este apartado se describen como contraste de hipótesis.

El primer paso del proceso de verificación de una hipótesis consiste en *formular estadísticamente la hipótesis científica que se desea contrastar*; es decir, en transformar la hipótesis científica en *hipótesis estadística*. Esto supone que una hipótesis científica puede ser formulada en términos de la forma de una o varias distribuciones poblacionales, o en términos del valor de uno o más parámetros de esa o esas distribuciones. Así, por ejemplo, la hipótesis científica «el nivel medio de colesterol en sangre de los varones no difiere del de las mujeres» podría formularse, en términos estadísticos, de la siguiente manera:  $\mu_v = \mu_m$ ; es decir: el promedio  $\mu$  de la distribución de la variable «nivel de colesterol en sangre» en la población de varones es igual al promedio  $\mu$  de esa misma distribución en la población de mujeres.

Formulada la hipótesis estadística, el segundo paso del proceso de verificación consiste en *buscar evidencia empírica relevante capaz de informar sobre si la hipótesis establecida es o no sostenible*. Esto, en general, no resulta demasiado complicado de conseguir: parece razonable pensar que, si una hipótesis concreta referida a una distribución poblacional es *correcta*, al extraer una muestra de esa población debe encontrarse un resultado muestral similar al que esa hipótesis propone para la distribución poblacional. O lo que es lo mismo: una hipótesis será compatible con los datos empíricos cuando a partir de ella sea posible deducir o predecir un resultado muestral (un estadístico) con cierta precisión.

Si una hipótesis afirma que «los varones y las mujeres no difieren en el nivel medio de colesterol en sangre» (formulada en términos estadísticos:  $\mu_v = \mu_m$ ) y se asume que esa hipótesis es correcta, debe esperarse que, al extraer una muestra aleatoria de la población de varones y otra de la población de mujeres, el nivel medio de colesterol en sangre observado en ambas muestras,  $\bar{X}_v$  y  $\bar{X}_m$ , sea similar. Una discrepancia importante entre la afirmación propuesta en la hipótesis y el resultado muestral encontrado puede estar indicando dos cosas diferentes: bien la hipótesis es correcta y la discrepancia observada sólo es producto de las fluctuaciones propias del azar muestral; bien la hipótesis es incorrecta y, por tanto, incapaz de ofrecer predicciones acertadas. La cuestión clave que se plantea en este momento es la de determinar cuándo la discrepancia encontrada es lo bastante grande como para poder considerar que el resultado muestral observado es incompatible con la hipótesis formulada y, en consecuencia, para hacer pensar que esa discrepancia no es explicable por fluctuaciones debidas al azar sino por el hecho de que la hipótesis planteada es incorrecta.

Se necesita, y este es el tercer paso del proceso, una *regla de decisión*. Y esa regla debe establecerse en términos de *probabilidad*. Si en el ejemplo planteado sobre el nivel de colesterol en sangre de los varones y de las mujeres se pudiera trabajar con las poblaciones comple-

tas de varones y mujeres (es decir, si se pudiera medir el colesterol en sangre de *todos* los varones y *todas* las mujeres), no habría que recurrir a la teoría de la probabilidad porque tampoco sería necesario efectuar ningún tipo de contraste de hipótesis: se conocerían los valores de  $\mu_v$  y  $\mu_m$ , y se sabría si son iguales o no. Pero el tener que trabajar con *muestras* en lugar de poblaciones obliga a establecer una regla de decisión en términos de probabilidad.

Ahora bien, el número de reglas de decisión que se pueden establecer en una situación particular es casi ilimitado. Por supuesto, unas reglas serán mejores o más útiles que otras y, probablemente, ninguna de ellas será lo bastante buena como para resultar de utilidad en todo tipo de situaciones. Afortunadamente, la *teoría de la decisión* se ha encargado de elaborar unos cuantos principios elementales que pueden trasladarse al contexto del contraste de hipótesis. En general, la regla de decisión que se utiliza en los contrastes de hipótesis es una afirmación de este tipo: si el resultado muestral observado es, suponiendo correcta la hipótesis, muy poco probable, se considerará que la hipótesis es incompatible con los datos; por el contrario, si el resultado muestral observado es, suponiendo correcta la hipótesis, probable, se considerará que la hipótesis es compatible con los datos. Por tanto, se trata de una regla de decisión que se basa en el grado de compatibilidad existente entre la *hipótesis* y los *datos*, expresada ésta en términos de probabilidad.

Imaginemos que se desea averiguar si un grafólogo posee o no la capacidad de detectar, por medio de la escritura, la presencia de trastornos depresivos. Puede formularse la hipótesis de que «el grafólogo no posee tal capacidad». Si esta hipótesis es correcta, al presentar al grafólogo un par de muestras de escritura (una perteneciente a un paciente con trastorno y otra a un paciente sin trastorno) para que elija la que pertenece al paciente con trastorno, cabe esperar que responda al azar (se está asumiendo que la hipótesis es correcta), por lo que la probabilidad de que acierte será de 0,5. Por el contrario, si la hipótesis es incorrecta (y, por tanto, el grafólogo sí posee la mencionada capacidad), al presentarle el mismo par de muestras de escritura, la probabilidad de que acierte será mayor de 0,5, es decir, mayor que la probabilidad de acertar por azar.

En una situación como ésta, la hipótesis de que «el grafólogo no posee la capacidad de diagnosticar trastornos depresivos a través de la escritura» se puede plantear la siguiente manera:  $\pi_{\text{acierto}} \leq 0,5$ . Para contrastar esta hipótesis se pueden presentar, en lugar de un par de muestras de escritura, 10 pares. Si la hipótesis es correcta, cabe esperar encontrar no más de 5 aciertos (es decir, no más del número de aciertos esperable por azar). Por el contrario, si la hipótesis es incorrecta, cabe esperar encontrar un número de aciertos superior a 5 (es decir, más del número de aciertos esperable por azar). Ahora bien, si el grafólogo obtiene 6 aciertos, ¿podrá decirse que ese resultado es mayor que el esperable por azar? ¿y si obtiene 7? La clave consiste en utilizar la teoría de la probabilidad para establecer una regla que permita decidir cuándo un resultado muestral es compatible con la hipótesis y cuándo no. En consecuencia, un número de aciertos esperable por azar nos llevará a pensar que la hipótesis planteada es compatible con los datos y a concluir que el grafólogo no posee la capacidad de diagnosticar a partir de la escritura; mientras que un número de diagnósticos correctos superior al esperable por azar nos llevará a pensar que la hipótesis planteada es incompatible con los datos y a concluir que el grafólogo sí posee esa capacidad (pues si  $\pi_{\text{acierto}} \leq 0,5$  es una afirmación incorrecta, entonces la afirmación correcta debe ser  $\pi_{\text{acierto}} > 0,5$ ).

Así pues, resumiendo: un **contraste de hipótesis** es un *proceso de decisión en el que una hipótesis formulada en términos estadísticos es puesta en relación con los datos empíricos para determinar si es o no compatible con ellos*.

## Las hipótesis estadísticas

Una hipótesis estadística es una afirmación sobre una o más distribuciones de probabilidad; más concretamente, sobre la *forma* de una o más distribuciones de probabilidad, o sobre el valor de uno o más *parámetros* de esas distribuciones. Las hipótesis estadísticas se suelen representar por la letra  $H$  seguida de una afirmación que da contenido a la hipótesis:

$H$ : la variable  $X$  se distribuye normalmente con  $\mu = 100$  y  $\sigma = 15$ .

$H$ :  $\pi = 0,5$ .

$H$ :  $\mu \leq 30$ .

$H$ :  $Mdn_1 \neq Mdn_2$ .

$H$ :  $\mu_1 = \mu_2 = \mu_3 = \mu_4$ .

En general, una hipótesis estadística surge a partir de una hipótesis científica. Pero entre una hipótesis científica y una hipótesis estadística no existe una correspondencia exacta. La primera proporciona la base para la formulación de la segunda, pero no son la misma cosa. Mientras una hipótesis científica se refiere a algún aspecto de la realidad, una hipótesis estadística se refiere a algún aspecto o detalle de una distribución de probabilidad. Esto significa, por ejemplo, que la expresión  $\mu_v = \mu_m$  presentada anteriormente no es la única formulación estadística posible de la hipótesis científica «los varones y las mujeres no difieren en el nivel medio de colesterol en sangre». En lugar del promedio  $\mu$  podría utilizarse el promedio  $Mdn$  y establecer esta otra formulación estadística:  $Mdn_v = Mdn_m$ . Y todavía podría transformarse esa hipótesis científica en hipótesis estadística utilizando otras estrategias; por ejemplo:  $F_v(x) = F_m(x)$ , es decir, la función de distribución de la variable  $X$  = «nivel de colesterol en sangre» es la misma en la población de varones y en la población de mujeres.

Existen, por tanto, varias formas diferentes de expresar estadísticamente una hipótesis científica concreta. A lo largo de este capítulo y de los que siguen se irá viendo qué hipótesis estadísticas es posible plantear y cómo deben plantearse. De momento, basta con saber que el primer paso en el proceso de verificación de una hipótesis consiste en formular en términos estadísticos la afirmación contenida en la hipótesis científica que se desea verificar.

Dicho esto, es necesario advertir que, aunque hasta ahora se han venido proponiendo ejemplos en los que se ha formulado una sola hipótesis, lo cierto es que todo contraste de hipótesis se basa en la formulación de dos hipótesis:

1. La *hipótesis nula*, representada por  $H_0$ .
2. La *hipótesis alternativa*, representada por  $H_1$ .

La **hipótesis nula**  $H_0$  es la hipótesis que se somete a contraste. Consiste generalmente en una afirmación concreta sobre la forma de una distribución de probabilidad o sobre el valor de alguno de los parámetros de esa distribución:

$H_0$ : La variable  $X$  se distribuye normalmente con  $\mu = 100$  y  $\sigma = 15$ .

$H_0$ :  $\pi_1 = \pi_2$ .

$H_0$ :  $\mu_1 = \mu_2$ .

$H_0$ :  $\rho = 0$ .

$H_0$ :  $\pi = 0,5$ .

La **hipótesis alternativa**  $H_1$  es la negación de la nula.  $H_1$  incluye todo lo que  $H_0$  excluye. Mientras  $H_0$  es una hipótesis *exacta* (tal cosa es *igual* a tal otra),  $H_1$  es *inexacta* (tal cosa es *distinta*, *mayor* o *menor* que tal otra):

$H_1$ : La variable  $X$  no se distribuye normalmente con  $\mu = 100$  y  $\sigma = 15$ .

$H_1$ :  $\pi_1 > \pi_2$ .

$H_1$ :  $\mu_1 < \mu_2$ .

$H_1$ :  $\rho \neq 0$ .

$H_1$ :  $\mu < 0,5$ .

Cuando en  $H_1$  aparece el signo *distinto* ( $\neq$ ), se dice que el contraste es *bilateral* o bidireccional. Cuando en  $H_1$  aparece el signo *menor que* ( $<$ ) o *mayor que* ( $>$ ) se dice que el contraste es *unilateral* o unidireccional. Más adelante se volverá sobre esta cuestión.

La hipótesis nula y la hipótesis alternativa suelen plantearse como hipótesis rivales. Son hipótesis exhaustivas y mutuamente exclusivas, lo cual implica que si una es verdadera, la otra es necesariamente falsa. Según esto, en los ejemplos propuestos anteriormente pueden plantearse las siguientes hipótesis:

a)  $H_0$ :  $\mu_v = \mu_m$ .

$H_1$ :  $\mu_v \neq \mu_m$ .

b)  $H_0$ :  $\pi_{\text{acierto}} \leq 0,5$ .

$H_1$ :  $\pi_{\text{acierto}} > 0,5$ .

Las hipótesis del párrafo *a* se refieren al ejemplo sobre diferencias en el nivel de colesterol en sangre entre varones y mujeres. La hipótesis nula afirma que los varones y las mujeres no difieren en el nivel medio de colesterol en sangre; la hipótesis alternativa afirma que sí difieren. Son hipótesis exhaustivas y mutuamente exclusivas. Las hipótesis del párrafo *b* se refieren al ejemplo del grafólogo capaz de diagnosticar a través de la escritura. La hipótesis nula afirma que el grafólogo no posee tal capacidad; la hipótesis alternativa afirma que sí la posee. También estas dos hipótesis son exhaustivas y mutuamente exclusivas.

Conviene no pasar por alto un detalle de especial importancia: el signo *igual* ( $=$ ), tanto si va solo ( $\mu_v = \mu_m$ ) como si va acompañado ( $\pi \leq 0,5$ ), *siempre* va en la hipótesis nula. Según se ha dicho,  $H_0$  es la hipótesis que se somete a contraste. Esto significa que es a partir de la afirmación *concreta* establecida en  $H_0$  (y la única afirmación concreta establecida es la que corresponde al signo  $=$ ) desde donde se inicia todo el contraste. Es decir, tanto si  $H_0$  es *exacta* ( $\mu_v = \mu_m$ ) como si es *inexacta* ( $\pi < 0,5$ ), todo el proceso de decisión va a estar basado en un modelo probabilístico construido a partir de la afirmación concreta correspondiente al signo « $=$ » presente en  $H_0$ . Ese modelo probabilístico, que enseguida será tratado, es el que proporciona la información necesaria para tomar una decisión sobre  $H_0$ .

## Los supuestos

Para que una hipótesis estadística pueda predecir un resultado muestral con cierta exactitud es necesario, en primer lugar, que la distribución poblacional con la que se va a trabajar esté completamente especificada. Por ejemplo, hipótesis del tipo:



$H$ : La variable  $X$  se distribuye normalmente con  $\mu = 100$  y  $\sigma = 15$ ,

$H$ :  $\mu = 0,5$ ,

son hipótesis que especifican por completo las distribuciones poblacionales a las que hacen referencia. La primera hipótesis define una distribución normal con parámetros conocidos. La segunda hipótesis permitiría especificar por completo una distribución binomial una vez establecido el tamaño de la muestra. A este tipo de hipótesis se les llama *simples*.

Las hipótesis en las que la distribución poblacional no queda completamente especificada reciben el nombre de *compuestas*. Hipótesis del tipo:

$H$ : La variable  $X$  se distribuye normalmente con  $\mu = 100$ ,

$H$ :  $\pi < 0,50$ ,

son hipótesis compuestas pues en ninguna de ellas quedan completamente especificadas las distribuciones poblacionales a las que hacen referencia. La primera hipótesis define una distribución normal con media conocida pero con varianza desconocida. La segunda hipótesis, referida a una distribución binomial, no define una única distribución sino muchas diferentes.

Lo ideal, por supuesto, sería poder plantear, siempre, hipótesis nulas *simples*, pues eso permitiría definir con precisión la distribución poblacional a partir de la cual se efectuarán las predicciones muestrales. Pero ocurre que ni los intereses del investigador se corresponden siempre con el contenido de una hipótesis simple, ni en todas las situaciones resulta posible formular hipótesis de ese tipo. Esto significa que, con frecuencia, la hipótesis nula planteada no será simple, sino compuesta. Lo cual obligará a establecer un conjunto de *supuestos* que, junto con la hipótesis, permitan especificar por completo la distribución poblacional de referencia. Sólo entonces será posible predecir con cierta precisión qué es lo que cabe esperar encontrar al extraer una muestra aleatoria de esa población.

En el ejemplo del grafólogo supuestamente capaz de detectar trastornos depresivos a través de la escritura, para verificar si el grafólogo posee o no esa capacidad, se han planteado las hipótesis estadísticas:  $H_0$ :  $\pi_{\text{acierto}} \leq 0,5$ ;  $H_1$ :  $\pi_{\text{acierto}} > 0,5$ . Y para contrastar esas hipótesis se presentaban al grafólogo 10 pares de muestras de escritura. Pues bien, si los 10 pares de muestras de escritura se presentan de forma independiente y en cada presentación sólo hay dos resultados posibles (acierto-error) con  $\pi_{\text{acierto}} = 0,5$  en cada presentación, la variable *número de aciertos* tendrá una distribución de probabilidad completamente especificada (la binomial, con parámetros  $n = 10$  y  $\pi = 0,5$ ) y eso permitirá poder tomar una decisión respecto a  $H_0$  en términos de probabilidad.

Por tanto, los supuestos de un contraste de hipótesis hacen referencia al conjunto de condiciones que deben cumplirse para poder tomar una decisión sobre la hipótesis nula  $H_0$  basada en una distribución de probabilidad conocida. Pero ese conjunto de condiciones que ha sido necesario establecer no se refieren únicamente a la distribución poblacional de partida. También hacen referencia a ciertas características de los datos muestrales: *si la muestra es aleatoria...*, *si las presentaciones son independientes...* Esto significa que, para apoyar la decisión en una distribución de probabilidad conocida, es necesario, por un lado, especificar por completo la distribución poblacional a partir de la cual se establecen las predicciones formuladas en  $H_0$  y, por otro, definir las características de los datos con los que se contrastarán esas predicciones (muestra aleatoria, nivel de medida, etc.).

Resumiendo: los supuestos de un contraste de hipótesis son un conjunto de afirmaciones que hay que establecer (sobre la población de partida y sobre la muestra utilizada) para conseguir determinar la distribución de probabilidad en la que se basará la decisión sobre  $H_0$ .

## El estadístico de contraste

Un *estadístico de contraste* es un resultado muestral que cumple la doble condición de (1) proporcionar información empírica relevante sobre la afirmación propuesta en la hipótesis nula y (2) poseer una distribución muestral conocida.

Si la hipótesis que se desea contrastar es  $H_0: \mu = 30$ , debe recurrirse a un estadístico capaz de detectar cualquier desviación empírica de la afirmación establecida en  $H_0$ . Obviamente, ni  $S_n$ , ni  $P$ , ni  $r_{xy}$ , por citar algunos estadísticos conocidos, proporcionarán información relevante sobre el parámetro  $\mu$ . Para contrastar la hipótesis  $H_0: \mu = 30$ , lo razonable será utilizar la información muestral ofrecida por el estadístico  $\bar{X}$ . Del mismo modo, si la hipótesis que se desea contrastar es  $H_0: \pi \leq 0,5$ , lo razonable será recurrir a un estadístico que pueda proporcionar información relevante sobre  $\pi$ , por ejemplo,  $P =$  «proporción de aciertos». Etc.

La segunda condición que debe cumplir un resultado muestral para poder ser utilizado como estadístico de contraste es la de poseer una *distribución muestral conocida*. Un estadístico es una variable aleatoria y, como tal, tiene su propia función de probabilidad denominada distribución muestral. Es precisamente la distribución muestral del estadístico de contraste la que contiene las probabilidades en que se basa la decisión sobre  $H_0$ .

Por tanto, una vez planteadas las hipótesis, es necesario seleccionar el estadístico de contraste capaz de proporcionar información relevante sobre ellas y establecer los supuestos necesarios para conseguir determinar la distribución muestral de ese estadístico. En el ejemplo sobre el grafólogo supuestamente capaz de diagnosticar trastornos depresivos a través de la escritura se habían planteado las hipótesis:  $H_0: \pi_{\text{acierto}} \leq 0,5$ ;  $H_1: \pi_{\text{acierto}} > 0,5$ . Existen dos estadísticos (en realidad los dos son el mismo, pues uno es transformación lineal del otro) capaces de proporcionar información relevante sobre las hipótesis planteadas (se utiliza la letra  $T$  para nombrar, de forma genérica, a un estadístico de contraste cualquiera):

$T_1 = X$  («número de aciertos o de diagnósticos correctos»).

$T_2 = P$  («proporción de aciertos o de diagnósticos correctos»).

Suponiendo, según se ha señalado antes, que las presentaciones de los 10 pares de muestras de escritura son independientes y que la probabilidad de cada uno de los dos resultados posibles (acierto-error) es la misma en cada presentación, la distribución muestral de las variables o estadísticos de contraste  $X$  y  $P$  será la binomial con parámetros  $n = 10$  y  $\pi = 0,5$ . Según esto, la probabilidad asociada a cada uno de los valores de  $X$  y  $P$  (Tabla 9.3) vendrá dada por la función:

$$f(x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} = \binom{10}{x} 0,5^x (0,5)^{10-x}$$

La distribución muestral de  $X$  o de  $P$  (Tabla 9.3) proporciona la probabilidad asociada a cada uno de los posibles valores de  $X$  o  $P$  bajo la condición  $H_0: \pi = 0,5$ . En la tabla puede verse, por ejemplo, que la probabilidad de encontrar, suponiendo  $\pi = 0,5$ , 10 aciertos (es decir, la probabilidad de  $X = 10$ , o  $P = 1$ ) vale 0,001. Y también puede verse, por ejemplo, que la probabilidad de encontrar 9 aciertos o más (es decir, la probabilidad de  $X \geq 9$ , o  $P \geq 0,9$ ), siempre suponiendo  $\pi = 0,5$ , vale  $0,010 + 0,001 = 0,011$ . Justamente en estas probabilidades se basará la decisión sobre  $H_0$ .

Tabla 9.3. Distribución muestral de  $X$  y  $P$ , con  $\pi = 0,5$  y  $n = 10$ 

$X$	$P$	$f(x) = f(p)$
0	0,0	0,001
1	0,1	0,010
2	0,2	0,044
3	0,3	0,117
4	0,4	0,205
5	0,5	0,246
6	0,6	0,205
7	0,7	0,117
8	0,8	0,044
9	0,9	0,010
10	1,0	0,001

Así pues, los estadísticos  $X$  y  $P$  sirven como estadísticos de contraste para poner a prueba la hipótesis  $H_0: \pi_{\text{acierto}} \leq 0,5$  porque ambos cumplen las condiciones exigidas a un estadístico de contraste: (1) proporcionan información relevante sobre  $H_0$  y (2) poseen distribución muestral conocida.

### La regla de decisión

La regla de decisión es el criterio que se utiliza para decidir si la hipótesis nula planteada debe o no ser rechazada. Esta regla se basa en la partición de la distribución muestral del estadístico de contraste en dos zonas exclusivas y exhaustivas: la *zona de rechazo* y la *zona de aceptación*.

La *zona de rechazo*, también llamada *zona crítica*, es el área de la distribución muestral (distribución del estadístico) que corresponde a los valores del estadístico de contraste que se encuentran tan alejados de la afirmación establecida en  $H_0$  que es muy poco probable que ocurran si  $H_0$ , como se supone, es verdadera. Su probabilidad se denomina *nivel de significación* o *nivel de riesgo* y se representa generalmente con la letra griega  $\alpha$ .

La *zona de aceptación* es el área de la distribución muestral que corresponde a los valores del estadístico de contraste próximos a la afirmación establecida en  $H_0$ . Es, por tanto, el área correspondiente a los valores del estadístico de contraste que es probable que ocurran si  $H_0$ , como se supone, es verdadera. Su probabilidad se denomina *nivel de confianza* y se representa mediante  $1 - \alpha$ .

Definidas las zonas de rechazo y aceptación, la **regla de decisión** consiste en *rechazar*  $H_0$  si el estadístico de contraste toma un valor perteneciente a la zona de rechazo o crítica; *mantener*  $H_0$  si el estadístico de contraste toma un valor perteneciente a la zona de aceptación. Por tanto, se rechaza una hipótesis nula sometida a contraste *cuando* el valor del estadístico de contraste cae en la zona crítica; y se rechaza *porque* eso significa que el valor del estadístico de contraste se aleja demasiado de la predicción establecida en esa hipótesis, es decir, *porque*, si la hipótesis planteada fuera verdadera, el estadístico de contraste no debería haber tomado ese valor (sería muy poco probable que lo tomara); como de hecho lo ha tomado, la conclusión más razonable será que la hipótesis planteada no es verdadera.

El tamaño de las zonas de rechazo y aceptación se determina fijando el valor de  $\alpha$ , es decir, fijando el nivel de significación con el que se desea trabajar. Por supuesto, si se tiene en cuenta que  $\alpha$  es la probabilidad que se va a considerar como lo bastante pequeña para que valores con esa probabilidad o menor no ocurran bajo  $H_0$ , se comprenderá que  $\alpha$  será, necesariamente, un valor pequeño. Cómo de pequeño es algo que debe establecerse de forma arbitraria, si bien los niveles de significación habitualmente propuestos para  $\alpha$  en la literatura estadística y aceptados por acuerdo de la comunidad científica son 0,01 y 0,05 (también referidos como 1 % y 5 %, respectivamente).

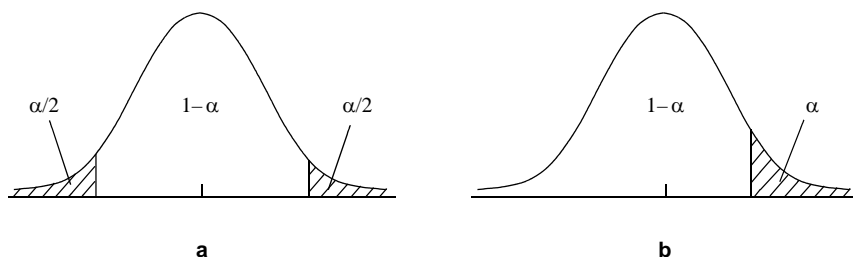
Dependiendo de cómo se formule  $H_1$ , los contrastes de hipótesis pueden ser *bilaterales* o *unilaterales*:

1. Contraste bilateral:  $H_0: \mu_v = \mu_m$ ;  $H_1: \mu_v \neq \mu_m$ .
2. Contraste unilateral:  $H_0: \pi_{\text{acierto}} \leq 0,5$ ;  $H_1: \pi_{\text{acierto}} > 0,5$ .

La forma de dividir la distribución muestral en zona de rechazo o crítica y zona de aceptación depende de que el contraste sea bilateral o unilateral. La zona crítica debe situarse donde puedan aparecer los valores muestrales incompatibles con  $H_0$ , es decir, donde puedan aparecer los valores muestrales que apunten en la dirección propuesta en  $H_1$ . Así, en el contraste 1, dada la afirmación establecida en  $H_1$ , la zona crítica debe recoger los valores muestrales que vayan tanto en la dirección  $\bar{X}_v - \bar{X}_m > 0$  como en la dirección  $\bar{X}_v - \bar{X}_m < 0$ . Dicho de otro modo, si  $H_0: \mu_v = \mu_m$  es falsa, lo será tanto si  $\mu_v$  es mayor que  $\mu_m$  como si  $\mu_v$  es menor que  $\mu_m$ , por lo que la zona crítica deberá recoger ambas posibilidades. Por esta razón, *en los contrastes bilaterales, la zona crítica se encuentra, generalmente\*, repartida a partes iguales entre las dos colas de la distribución muestral* (Figura 9.1.a).

En el contraste 2, por el contrario, los únicos valores muestrales incompatibles con  $H_0$  serán los que vayan en la dirección  $P > 0,5$ , que es la dirección apuntada en  $H_1$ . Los valores muestrales que estén por debajo de  $P = 0,5$  no serán incompatibles con  $H_0$  y la zona crítica deberá reflejar esta circunstancia quedando ubicada en la cola derecha de la distribución muestral. Por tanto, *en los contrastes unilaterales, la zona crítica se encuentra en una de las dos colas de la distribución muestral* (Figura 9.1.b).

**Figura 9.1.** Ejemplo de zonas críticas en un contraste bilateral (Figura a) y unilateral derecho (Figura b) con una distribución muestral de forma normal



\* «Generalmente» quiere decir que existen excepciones a esta regla. Dependiendo del tipo de hipótesis nula planteada, del estadístico de contraste elegido y de la distribución muestral utilizada, puede ocurrir que la zona crítica de un contraste bilateral esté, toda ella, situada en la cola derecha de la distribución.

Según esto, las reglas de decisión para los dos contrastes utilizados de ejemplo (el referido a las diferencias en colesterol entre varones y mujeres, y el referido al grafólogo capaz de diagnosticar trastornos depresivos a través de la escritura) pueden concretarse de la siguiente manera:

- (1) Rechazar  $H_0$ :  $\mu_v = \mu_m$  si el estadístico de contraste cae en la zona crítica, es decir, si toma un valor mayor que el percentil  $100(1-\alpha/2)$  o menor que el percentil  $100(\alpha/2)$  de su distribución muestral.

O bien: rechazar  $H_0$ :  $\mu_v = \mu_m$  si el estadístico de contraste toma un valor tan grande o tan pequeño que la probabilidad de obtener un valor tan extremo o más que el encontrado es menor que  $\alpha/2$ .

- (2) Rechazar  $H_0$ :  $\pi_{\text{acierto}} \leq 0,5$  si el estadístico de contraste cae en la zona crítica, es decir, si toma un valor mayor que el percentil  $100(1-\alpha)$  de su distribución muestral.

O bien: rechazar  $H_0$ :  $\pi_{\text{acierto}} \leq 0,5$  si el estadístico de contraste toma un valor tan grande que la probabilidad de obtener un valor como ese o mayor es menor que  $\alpha$ .

Por tanto, en un contraste de hipótesis se toman decisiones a partir de la probabilidad asociada al resultado muestral obtenido, es decir a la probabilidad asociada al valor del estadístico de contraste. A esta probabilidad se le llama *nivel crítico* y se representa por la letra  $p$ . La decisión siempre consiste en rechazar  $H_0$  cuando esta probabilidad es *pequeña* ( $p < \alpha$ ) y mantener  $H_0$  cuando esta probabilidad es *grande* ( $p > \alpha$ ).

## La decisión

Planteada la hipótesis, formulados los supuestos, obtenido el estadístico de contraste y su distribución muestral, y establecida la regla de decisión, el paso siguiente de un contraste consiste en tomar una decisión. Tal decisión se toma, *siempre*, respecto a  $H_0$ , y consiste en rechazarla o mantenerla de acuerdo con las condiciones establecidas en la regla de decisión: si el estadístico de contraste cae en la zona crítica ( $p < \alpha$ ), se rechaza  $H_0$ ; si el estadístico de contraste cae en la zona de aceptación ( $p > \alpha$ ), se mantiene  $H_0$ .

La decisión, así planteada, parece no revestir ningún tipo de problema. Pero eso no es del todo cierto. Conviene resaltar un aspecto importante de este proceso de decisión que no siempre es adecuadamente tenido en cuenta en la investigación empírica. Una decisión, en el contexto del contraste de hipótesis, siempre consiste en *rechazar* o *mantener* una  $H_0$  particular. Si se rechaza, se está afirmando que esa hipótesis es *falsa*; es decir, se está afirmando que ha quedado probado que esa hipótesis es falsa. Por el contrario, si se mantiene, no se está afirmando que ha quedado probado que esa hipótesis es verdadera; simplemente se está afirmando que no se dispone de evidencia empírica suficiente para rechazarla y que, por tanto, puede considerarse compatible con los datos. Así pues:

- Cuando se decide ***mantener*** una hipótesis nula, se quiere significar con ello que se considera que esa hipótesis es compatible con los datos.
- Cuando se decide ***rechazar*** una hipótesis nula, se quiere significar con ello que se considera probado que esa hipótesis es falsa.

La razón de que esto sea así es doble. Por un lado, dada la naturaleza inespecífica de  $H_1$ , raramente es posible afirmar que  $H_1$  no es verdadera; las desviaciones pequeñas de  $H_0$  forman parte de  $H_1$ , por lo que al mantener una  $H_0$  particular, también se están manteniendo, muy probablemente, algunos valores de  $H_1$ ; debe concluirse, por tanto, que se mantiene o no rechaza  $H_0$ , pero nunca que se acepta como verdadera. Por otro lado, en el razonamiento estadístico que lleva a la toma de una decisión respecto a  $H_0$ , puede reconocerse el argumento deductivo *modus tollens*, aunque de tipo probabilístico: *si  $H_0$  es verdadera, entonces, muy probablemente, el estadístico de contraste  $T$  tomará valores comprendidos entre  $a$  y  $b$ ;  $T$  no toma un valor comprendido entre  $a$  y  $b$ ; luego, muy probablemente,  $H_0$  no es verdadera*. Este argumento es impecable, nada hay en él que lo invalide desde el punto de vista lógico. Sin embargo, si una vez establecida la primera premisa se continúa de esta otra manera:  *$T$  toma un valor comprendido entre  $a$  y  $b$ ; luego  $H_0$ , muy probablemente, es verdadera*, se comete un error lógico llamado *falacia de la afirmación del consecuente*: obviamente,  $T$  puede haber tomado un valor comprendido entre  $a$  y  $b$  por razones diferentes de las contenidas en  $H_0$ .

## Resumen

Probablemente ahora se entenderá mejor la definición propuesta para el contraste de hipótesis: *proceso de toma de decisiones en el que una afirmación sobre alguna característica poblacional (hipótesis nula) es puesta en relación con lo datos empíricos (estadístico de contraste) para determinar si es o no compatible con ellos (compatibilidad que se establece en términos de probabilidad:  $p$ )*.

Todos los contrastes de hipótesis siguen la lógica expuesta. Es decir, cualquier técnica de análisis de datos se ajusta al proceso descrito: hipótesis, supuestos, estadístico de contraste y distribución muestral, y decisión basada en la teoría de la probabilidad. Ahora bien, puesto que las situaciones concretas que interesa analizar poseen características particulares, el proceso general recién descrito necesita ser adaptado a las particularidades de cada una de ellas. Esto es lo que hacen las técnicas de análisis que se describen en los capítulos siguientes: cada técnica de análisis (cada prueba estadística o prueba de significación) es una adaptación de este proceso general a una situación concreta.

Dependiendo del número de variables, del tipo de variables, de la forma de recoger los datos, etc., habrá que utilizar una prueba concreta u otra. Todas estas pruebas se describen en los capítulos que siguen.

## La estimación de parámetros

La **estimación de parámetros** consiste en utilizar la información muestral para inferir alguna propiedad de la población; es decir, en utilizar un estadístico (la media muestral, la proporción muestral, la correlación muestral, etc.), que recibe el nombre de *estimador*, para inferir el valor de algún parámetro (la media poblacional, la proporción poblacional, la correlación poblacional, etc.). A esta estimación directa (utilizar la media muestral para estimar la media poblacional) se le llama *estimación puntual*.

Si al valor muestral o estimador puntual (por ejemplo, la media) se le suma y resta una cantidad para estimar no un valor concreto, sino un rango o intervalo de valores, se habla de *estimación por intervalos*.

Esta cantidad que se suma y se resta a un estimador para obtener un intervalo de estimación recibe el nombre de *error máximo* ( $E_{\max}$ ) y depende de la distribución muestral del estadístico utilizado como estimador. El error máximo se calcula intentando que el intervalo construido incluya, con una probabilidad alta y conocida, el valor del parámetro que se desea estimar. Esta probabilidad recibe el nombre de *nivel de confianza* y, generalmente suele establecerse en 0,95.

Al intervalo de valores se le llama *intervalo de confianza* y viene definido por los dos valores resultantes de sumar y restar el error máximo al estimador puntual. Estos dos valores que definen el intervalo de confianza reciben el nombre de *límite inferior* y *límite superior*.

Para entender algo mejor cómo se construye un intervalo de confianza, recuérdese el ejemplo de la distribución muestral de la media propuesto en la Tabla 9.2. La tabla recoge las probabilidades asociadas a todos los posibles valores del estadístico *media* al extraer muestras de tamaño 2 de una población de 5 elementos. Por los datos de la tabla se sabe, por ejemplo, que la probabilidad de obtener una media igual a 2 vale  $3/25 = 0,12$ . Pero también se sabe, por ejemplo, que la probabilidad de obtener una media comprendida entre 1,5 y 4,5 (es decir, una media comprendida entre  $\mu \pm 1,5$ ) vale  $1 - 2/25 = 0,92$ .

Supongamos que se desea estimar la media poblacional  $\mu$  a partir de la media obtenida en una muestra concreta. Por supuesto, la media de esa muestra concreta podría tomar cualquier valor de los posibles, por ejemplo,  $\bar{x}$ . Si al valor  $\bar{x}$  obtenido se le suma y se le resta 1,5 puntos para obtener un intervalo de valores, el verdadero valor de la media poblacional (que sabemos que es  $\mu = 3$ ) estará incluido entre los límites del intervalo construido cualquiera que sea el valor de  $\bar{x}$ , excepto si  $\bar{x}$  vale 1 o 5. Ahora bien, la probabilidad de obtener en una muestra cualquiera una media de 1 o 5 vale  $2/25 = 0,08$ . O lo que es lo mismo, la probabilidad de obtener un valor muestral que no sea 1 o 5 vale 0,92. En consecuencia, en una situación como la descrita, de cada 100 muestras que se utilicen para construir intervalos sumando y restando 1,5 puntos a la media muestral, cabe esperar que con 92 de ellas se construyan intervalos entre cuyos límites se encuentre el verdadero valor del parámetro  $\mu$ .

Lógicamente, en una situación real no se conoce el valor de la media poblacional, razón por la cual se desea obtener una estimación. Sin embargo, sí se conoce el comportamiento (distribución muestral) del estadístico *media*. De modo que puede utilizarse un intervalo que permita incluir un determinado porcentaje del total de posibles valores del estadístico *media* a partir justamente de las probabilidades de su distribución muestral.

Todo lo dicho para la *media* vale para cualquier otro estadístico que se desee utilizar como estimador, siempre que posea distribución muestral conocida.

## Análisis descriptivo

# Los procedimientos *Frecuencias*, *Descriptivos*, *Razón* y *Cubos OLAP*

Generalmente, lo primero que conviene hacer con una variable, sea ésta categórica o cuantitativa, es formarse una idea lo más exacta posible acerca de sus características. Si la variable que se desea describir es *categórica*, una *distribución de frecuencias* y un *diagrama de barras* o *sectores* ofrecen información suficiente para formarse una idea lo bastante precisa sobre las características de esa variable. Si la variable que se desea describir es *cuantitativa* es necesario prestar atención a tres aspectos básicos: *tendencia central*, *dispersión* y *forma de la distribución*.

Así pues, las medidas de tendencia central y de dispersión, y los índices y gráficos sobre la forma de la distribución resultan más o menos útiles dependiendo de la naturaleza de las variables. Con variables categóricas, las medidas de tendencia central y de dispersión carecen de utilidad comparadas con una distribución de frecuencias o un gráfico sobre la forma de la distribución. Por el contrario, con variables continuas, una distribución de frecuencias pierde importancia comparada con la capacidad informativa de las medidas de tendencia central y de dispersión. Por otro lado, los diagramas que informan sobre la forma de una distribución son diferentes dependiendo de que la variable estudiada sea categórica o continua.

En este capítulo se describen dos procedimientos SPSS que permiten obtener la información necesaria para caracterizar apropiadamente tanto variables categóricas como cuantitativas: *Frecuencias* y *Descriptivos*. También se describen los procedimientos *Estadísticos de la razón*, que ofrece diversos estadísticos para el cociente entre dos variables, y *Cubos OLAP*, que permite obtener estadísticos por subgrupos.

## Frecuencias

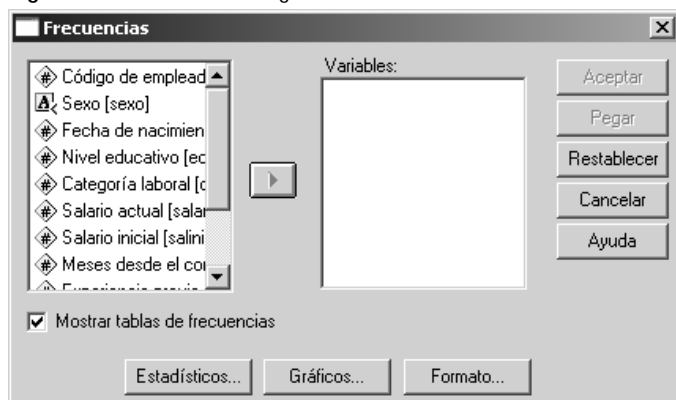
Una distribución de frecuencias informa sobre los valores concretos que adopta una variable y sobre el número (y porcentaje) de veces que se repite cada uno de esos valores. El procedimiento *Frecuencias* permite obtener distribuciones de frecuencias, pero además contiene opciones para: (1) calcular algunos de los estadísticos descriptivos más utilizados (tendencia central, posición, dispersión, asimetría y curtosis); (2) obtener algunos gráficos básicos (de barras, de sectores e histogramas); (3) controlar el formato de presentación de las distribuciones de frecuencias. La utilización de todas estas opciones depende, en gran medida, del



hecho de que la variable estudiada sea categórica o continua. Para obtener una distribución de frecuencias:

- Seleccionar la opción **Estadísticos descriptivos > Frecuencias...** del menú **Analizar** para acceder al cuadro de diálogo *Frecuencias* que muestra la Figura 10.1.

Figura 10.1. Cuadro de diálogo *Frecuencias*



Este cuadro de diálogo permite obtener distribuciones de frecuencias absolutas y porcentuales, varios estadísticos descriptivos y algunos gráficos básicos. Para ello:

- Seleccionar la(s) variable(s) cuya distribución de frecuencias se desea obtener y trasladarla(s) a la lista **Variables**. La especificación mínima requerida es una variable numérica o de cadena corta; las variables de cadena larga no están disponibles en la lista de variables del archivo de datos.
- “ **Mostrar tablas de frecuencias**. Esta opción (activa por defecto) permite decidir si se desea o no obtener la distribución de frecuencias. Puede desactivarse si, por ejemplo, sólo interesa ver algún gráfico o algún estadístico descriptivo. Si se desactiva esta opción y no se efectúa ninguna otra selección, los resultados sólo muestran el número total de casos y de valores perdidos del archivo de datos.

### ***Ejemplo: Estadísticos descriptivos > Frecuencias***

Este ejemplo muestra cómo obtener una distribución de frecuencias con las especificaciones que el procedimiento *Frecuencias* tiene establecidas por defecto. Se basa en el archivo *Datos de empleados*, el cual se encuentra en la misma carpeta en la que está instalado el SPSS:

- Seleccionar la opción **Estadísticos descriptivos > Frecuencias...** del menú **Analizar** para acceder al cuadro de diálogo *Frecuencias* (ver Figura 10.1).
- Seleccionar la variable *catlab* (categoría laboral) y trasladarla a la lista **Variables** mediante el botón *flecha* o pulsando dos veces sobre ella.

Pulsando el botón **Aceptar**, el *Visor* ofrece los resultados que muestra la Tabla 10.1. La tabla contiene información sobre: los valores de la variable *catlab* (sus etiquetas), la frecuencia ab-

soluto de cada valor (*Frecuencia*), la frecuencia porcentual calculada sobre el número total de casos del archivo (*Porcentaje*), la frecuencia porcentual calculada sobre el número de casos válidos, es decir, sin tener en cuenta los casos con valor perdido (*Porcentaje válido*) y la frecuencia porcentual acumulada (*Porcentaje acumulado*). La última línea ofrece el número total de casos.

Tabla 10.1. Tabla de frecuencias de la variable *catlab* (categoría laboral)

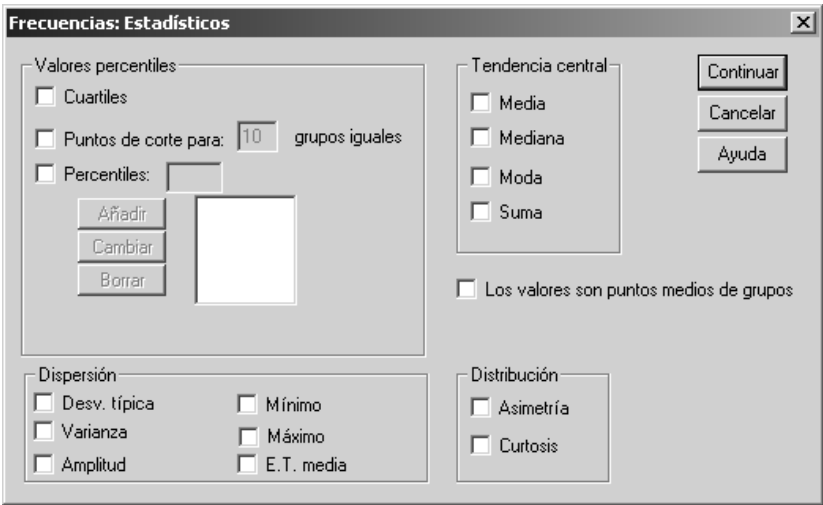
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Administrativo	363	76,6	76,6	76,6
	Seguridad	27	5,7	5,7	82,3
	Directivo	84	17,7	17,7	100,0
	Total	474	100,0	100,0	

Estadísticos

Puede obtenerse información adicional utilizando los botones específicos del cuadro de diálogo *Frecuencias* (ver Figura 10.1). Para obtener los estadísticos descriptivos más habituales:

- Pulsar el botón *Estadísticos...* para acceder al subcuadro de diálogo *Frecuencias: Estadísticos* que muestra la Figura 10.2.

Figura 10.2. Subcuadro de diálogo *Frecuencias: Estadísticos*



**Valores percentiles.** Este recuadro contiene varias opciones para solicitar cuantiles: cuantiles, deciles, percentiles, etc.

- **Cuartiles.** Calcula los percentiles 25, 50 y 75, es decir, los valores por debajo de los cuales se encuentra el 25 %, el 50 % y el 75 % de los casos, respectivamente. Para sa-

ber cómo calcula el SPSS estos cuantiles, puede consultarse el apartado *Estadísticos* del siguiente capítulo sobre análisis exploratorio.

- " **Puntos de corte para  $k$  grupos iguales.** Calcula los  $k-1$  valores que dividen la muestra en  $k$  grupos del mismo tamaño. El valor por defecto de  $k$  es 10 (= *deciles*), pero puede escribirse cualquier otro valor entre 2 y 100.
- " **Percentiles.** Permite solicitar percentiles concretos (valores que acumulan un determinado porcentaje de casos). Para obtener un percentil concreto:
  - ' Escribir el valor deseado en el cuadro de texto **Percentiles**.
  - ' Pulsar el botón **Añadir** para trasladar ese valor a la lista de percentiles.
  - ' Utilizar los botones **Cambiar** y **Borrar** para modificar o eliminar, respectivamente, valores previamente añadidos.

**Tendencia central.** Puede seleccionarse uno o más de los siguientes estadísticos:

- " **Media.** Media aritmética: suma de todas las puntuaciones dividida por el número de puntuaciones.
- " **Mediana.** Valor por debajo del cual se encuentra el 50% de los casos (equivale al percentil 50). Si el número de casos es par, la mediana se calcula como el promedio de los dos casos centrales cuando éstos se encuentran ordenados. Si el número de casos es impar, la mediana es el valor del caso central de la distribución.
- " **Moda.** Valor que más se repite. Si existen dos o más valores empatados en el número de repeticiones, sólo se muestra el más pequeño de ellos.
- " **Suma.** Suma de todos los valores.

**Dispersión.** Puede seleccionarse uno o más de los siguientes estadísticos:

- " **Desv. típica.** Desviación típica: raíz cuadrada de la varianza. Mide el grado en que las puntuaciones de la variable se alejan de su media.
- " **Varianza.** Medida de dispersión que se obtiene dividiendo por  $n-1$  ( $n$  = número de casos) la suma de los cuadrados de las diferencias entre cada puntuación y la media.
- " **Amplitud.** Diferencia entre el valor más grande (máximo) y el más pequeño (mínimo). También recibe el nombre de *rango*.
- " **Mínimo.** Valor más pequeño.
- " **Máximo.** Valor más grande.
- " **E.T. media.** Error típico de la media: desviación típica de la distribución muestral de la media. Se obtiene dividiendo la desviación típica por la raíz cuadrada del número de casos.

**Distribución.** Puede seleccionarse uno o más de los siguientes estadísticos:

- " **Asimetría.** Índice que expresa el grado de asimetría de la distribución. La asimetría positiva indica que los valores más extremos se encuentran por encima de la media. La asimetría negativa indica que los valores más extremos se encuentran por debajo de la media. Los índices de asimetría próximos a cero indican simetría.

El *Visor* también ofrece el *error típico* del índice de asimetría (es decir, la desviación típica de la distribución muestral del índice de asimetría), el cual permite tipificar el valor del índice de asimetría e interpretarlo como una puntuación *Z* con distribución aproximadamente normal (ver más adelante, en este mismo capítulo, el apartado *Puntuaciones típicas y curva normal*). Índices tipificados mayores que 1,96 en valor absoluto permiten afirmar que la distribución es asimétrica (positiva o negativa, dependiendo del signo del índice).

- " **Curtosis.** Índice que expresa el grado en que una distribución acumula casos en sus colas en comparación con los casos acumulados en las colas de una distribución normal con la misma varianza. La curtosis positiva indica que en las colas de la distribución hay acumulados más casos que en una distribución normal (lo cual suele coincidir con distribuciones más *puntiagudas* que una distribución normal). Los índices de curtosis próximos a cero indican semejanza con la curva normal.

El *Visor* también ofrece el *error típico* del índice de curtosis, el cual puede utilizarse para tipificar el valor del índice de curtosis y poder interpretarlo como una puntuación *Z* con distribución aproximadamente normal tipificada. Índices mayores que 1,96 en valor absoluto permiten afirmar que la distribución analizada se aleja de la distribución normal.

- " **Los valores son puntos medios de grupos.** En el caso de que la variable que se desea estudiar se encuentre agrupada en intervalos, esta opción permite calcular la mediana y los percentiles interpolando valores, es decir, considerando que los valores de la variable son los puntos medios de intervalos uniformemente distribuidos.

Puesto que esta opción afecta a todas las variables de la lista **Variables** (ver Figura 10.1), no debería marcarse si una o más variables de las listadas no se encuentran agrupadas en intervalos.

## Cuándo utilizar cada estadístico

Por lo que se refiere a los **percentiles**, sólo tiene sentido calcularlos con variables al menos ordinales. Carecen de significado con variables nominales.

Entre las medidas de **tendencia central**, la media requiere variables cuantitativas (de intervalo o razón, aunque también suele calcularse con datos ordinales). La mediana es un estadístico típicamente ordinal (requiere variables al menos ordinales). Al contrario de lo que ocurre con la media, la mediana es insensible a la presencia de valores extremos y, por tanto, es preferible a la media cuando la distribución es asimétrica. La moda sirve para todo tipo de variables, pero es más apropiada para caracterizar datos categóricos porque, por un lado, es un estadístico que sólo aprovecha información nominal y, por otro, con variables continuas es esperable que todos los valores tengan una frecuencia muy pequeña.

En cuanto a las medidas de **dispersión**, la desviación típica, la varianza y el error típico de la media únicamente poseen significado con variables cuantitativas (de intervalo o razón, aunque también suelen calcularse con datos ordinales). La amplitud o rango es apropiada para todo tipo de variables, excepto para las nominales, en las que no tiene sentido hablar de dispersión en el sentido de amplitud.

En lo relativo a los índices de **asimetría** y **curtosis**, de nuevo sólo tiene sentido calcularlos con variables cuantitativas pues en ambos interviene la media.

**Ejemplo: Estadísticos descriptivos > Frecuencias > Estadísticos**

Este ejemplo muestra cómo obtener algunos estadísticos descriptivos utilizando el procedimiento Frecuencias (se sigue utilizando el archivo *Datos de empleados*):

- En el cuadro de diálogo principal (ver Figura 10.1), seleccionar la variable *salario* (salario actual) y trasladarla a la lista **Variables**.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Frecuencias: Estadísticos* (ver Figura 10.2) y marcar todas las opciones del subcuadro excepto **Percentiles** y **Los valores son puntos medios de grupos**.
- Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones el *Visor de resultados* muestra los estadísticos que recoge la Tabla 10.2. Observando la tabla se aprecia, por ejemplo, que el salario medio es de 34.419,57 dólares (*Media*), que la mitad de los sujetos tienen salarios por debajo de 28.875,00 dólares (*Mediana*, *Percentil 50*), que entre el sujeto que gana más y el que gana menos existe una diferencia de 119.250,00 dólares (*Rango* o *amplitud*), que el 25 % de los sujetos tiene un salario de 24.000 dólares o menos (*Percentil 25*), que el 50 % de los sujetos tiene salarios comprendidos entre 24.000,00 y 37.162,50 dólares (*Percentiles 25 y 75*), etc.

**Tabla 10.2.** Estadísticos descriptivos

Salario actual		
N	Válidos	474
	Perdidos	0
Media		34.419,568
Error típ. de la media		784,311
Mediana		28.875,000
Moda		30.750,000
Desv. típ.		17.075,661
Varianza		291578214,453
Asimetría		2,125
Error típ. de asimetría		,112
Curtosis		5,378
Error típ. de curtosis		,224
Rango		119.250,000
Mínimo		15.750,000
Máximo		135.000,000
Suma		16.314.875,000
Percentiles	10	21.000,000
	20	22.950,000
	25	24.000,000
	30	24.825,000
	40	26.700,000
	50	28.875,000
	60	30.750,000
	70	34.500,000
	75	37.162,500
	80	41.100,000
	90	59.700,000

El cociente entre el índice de asimetría (también el de curtosis) y su error típico puede interpretarse como una puntuación típica (puntuación *Z*) distribuida normalmente (ver, más adelante, en este mismo capítulo, el apartado *Puntuaciones típicas y curva normal*). Para considerar que una distribución es simétrica, el índice de asimetría dividido por su error típico debe tomar un valor comprendido entre  $-1,96$  y  $1,96$  (valores correspondientes a los cuantiles 2,5 y 97,5 en una distribución normal tipificada). Los valores que se salen de esos límites delatan la presencia de asimetría (positiva si el valor es positivo, o negativa si el valor es negativo). En el ejemplo, el grado de asimetría es acusadamente positivo, pues el cociente entre el índice de asimetría y su error típico vale  $2,125/0,112 = 18,97$ , y este valor es demasiado grande para pensar que pertenece a una distribución con valor esperado cero.

Y lo mismo vale decir del índice de curtosis. La acumulación de casos en las colas es mayor que la que corresponde a una distribución normal, pues tipificando el índice de curtosis se obtiene  $5,378/0,224 = 24,01$ , y este valor es demasiado grande para pensar que pertenece a una distribución con valor esperado cero (que es el valor que indica una curtosis equivalente a la de una curva *normal*).

## Gráficos

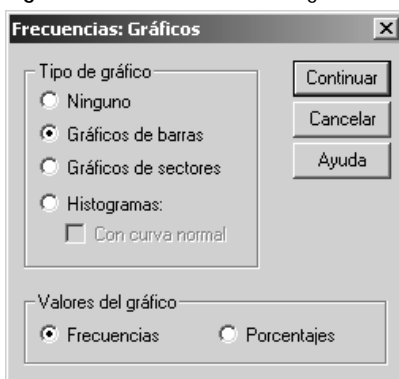
Además de las tablas o distribuciones de frecuencias y de los estadísticos descriptivos estudiados, el procedimiento Frecuencias también ofrece la posibilidad de obtener algunos gráficos básicos. En concreto, permite obtener gráficos de *barras*, gráficos de *sectores* e *histogramas* (el menú Gráficos de la barra de menús principal también permite obtener estos mismos gráficos, además de otros muchos).

Una vez que el *Visor de resultados* ha generado un gráfico, es posible introducir modificaciones en él (cambiar los colores, los ejes, las etiquetas, etc.) a través del *Editor de gráficos*. Para entrar en el *Editor de gráficos* basta con seleccionar, en el *Visor*, el gráfico que se desea editar y pulsar dos veces sobre él con el botón principal del ratón.

Para obtener un gráfico de *barras*, un gráfico de *sectores*, o un *histograma* con el procedimiento Frecuencias:

- Pulsar el botón Gráficos... del cuadro de diálogo principal (ver Figura 10.1) para acceder al subcuadro de diálogo *Frecuencias: Gráficos* que muestra la Figura 10.3.

Figura 10.3. Subcuadro de diálogo *Frecuencias: Gráficos*



**Tipo de gráfico.** Puede elegirse entre:

**Ninguno.** No se genera ningún gráfico. Es la opción por defecto.

**Gráficos de barras.** En un gráfico de barras a cada valor de la variable se le asigna una barra cuya altura es proporcional a su frecuencia absoluta o porcentual. La escala de la altura de las barras se ajusta automáticamente teniendo en cuenta la frecuencia más alta de las representadas. Son gráficos apropiados para representar variables categóricas.

**Gráficos de sectores.** Gráficos circulares en los que a cada valor de la variable se le asigna un sector de tamaño equivalente a su frecuencia absoluta o porcentual. Son gráficos apropiados para representar variables categóricas.

**Histogramas.** Similares a los gráficos de barras, pero con las barras juntas, dando así la impresión de continuidad. Sólo pueden obtenerse con variables con formato numérico. Para construir el histograma, el SPSS agrupa la variable en 21 intervalos (o menos, si el rango o amplitud de la variable es menor que 21). Son gráficos apropiados para representar variables cuantitativas.

“ **Con curva normal.** Ofrece una curva normal superpuesta sobre el histograma (generada a partir de la media y la desviación típica de la variable representada).

**Valores del gráfico.** En los gráficos de barras y de sectores es posible decidir qué tipo de frecuencia se desea representar:

**Frecuencias.** La escala y la etiqueta del eje correspondiente a la altura de las barras (o al tamaño de los sectores) están expresadas en frecuencias absolutas. Es la opción por defecto.

**Porcentajes.** La escala y la etiqueta del eje correspondiente a la altura de las barras (o al tamaño de los sectores) están expresadas en frecuencias porcentuales.

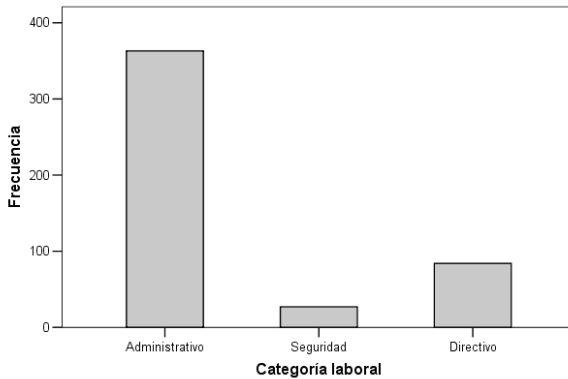
### ***Ejemplo: Estadísticos descriptivos > Frecuencias > Gráficos***

Este ejemplo muestra cómo obtener gráficos para variables categóricas y cuantitativas utilizando la opción **Gráficos...** del procedimiento **Frecuencias**. En concreto, muestra cómo obtener un gráfico de barras de la variable *catlab* (categoría laboral) y un histograma de la variable *salario* (salario actual). Se sigue utilizando el archivo *Datos de empleados*. Para obtener un gráfico de barras de la variable *catlab* (categoría laboral):

- En el cuadro de diálogo principal (ver Figura 10.1), seleccionar la variable *catlab* y trasladarla a la lista **Variables**.
- Pulsar el botón **Gráficos...** para acceder al subcuadro de diálogo *Frecuencias: Gráficos* (ver Figura 10.3) y seleccionar la opción **Gráficos de barras**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones se obtiene el gráfico de barras que muestra la Figura 10.4. En él puede apreciarse con claridad el tamaño relativo de las frecuencias anteriormente obtenidas en la Tabla 10.1.

Figura 10.4. Gráfico de barras de la variable *catlab* (categoría laboral)



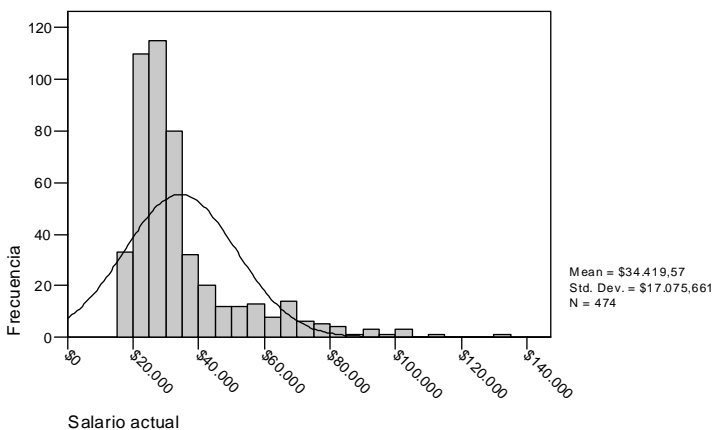
Pulsando dos veces sobre el gráfico se accede al *Editor de gráficos*, el cual permite introducir múltiples modificaciones relacionadas con el tamaño del gráfico, sus colores, los elementos que se desea visualizar, el tamaño y tipo de las fuentes, etc.

Para obtener un histograma de la variable *salario* (salario actual):

- En el cuadro de diálogo principal (ver Figura 10.1), seleccionar la variable *salario* (salario actual) y trasladarla a la lista **Variables**.
- Pulsar el botón **Gráficos...** para acceder al subcuadro de diálogo *Frecuencias: Gráficos* (ver Figura 10.3), seleccionar la opción **Histogramas** y marcar la opción **Con curva normal**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones se obtiene el histograma que muestra la Figura 10.5. Tal como habían anticipado ya los índices de asimetría y curtosis del ejemplo anterior, el histograma muestra asimetría positiva (casos extremos por la cola derecha), con una evidente desviación de la normalidad.

Figura 10.5. Histograma de la variable *salario actual*



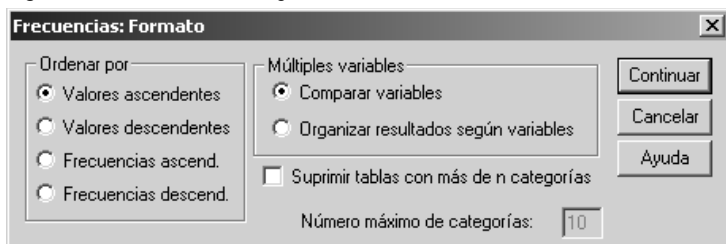


## Formato

Las opciones de formato permiten controlar algunos aspectos relacionados con la forma en que aparecerán en el *Visor* las tablas de frecuencias y los estadísticos solicitados. Para controlar el formato de presentación de las tablas de frecuencias:

- Pulsar el botón **Formato...** del cuadro de diálogo *Frecuencias* (ver Figura 10.1) para acceder al subcuadro de diálogo *Frecuencias: Formato* que muestra la Figura 10.6.

Figura 10.6. Cuadro de diálogo *Frecuencias: Formato*



**Ordenar por.** Las opciones de este recuadro sirven para establecer el orden en el que aparecerán los valores o categorías de la variable en la distribución de frecuencias:

**Valores ascendentes.** Los valores o categorías de la variable se ordenan desde el más pequeño al más grande. Es la opción por defecto.

**Valores descendentes.** Los valores o categorías de la variable se ordenan desde el más grande al más pequeño.

**Frecuencias ascendentes.** Los valores o categorías de la variable se ordenan de forma ascendente tomando como criterio el tamaño de cada frecuencia.

**Frecuencias descendentes.** Los valores o categorías de la variable se ordenan de forma descendente tomando como criterio el tamaño de cada frecuencia.

Si se solicita algún percentil o algún histograma, los valores se ordenan de forma ascendente (la opción por defecto) independientemente del orden seleccionado.

**Múltiples variables.** Al solicitar gráficos o estadísticos para más de una variable:

**Comparar variables.** Muestra una sola tabla de resultados para todas las variables.

**Organizar resultados según variables.** Muestra una tabla para cada variable.

- **Suprimir tablas con más de  $k$  categorías.** Esta opción elimina de la salida las distribuciones con más de  $k$  valores o categorías. El valor por defecto para  $k$  es 10, pero puede introducirse cualquier valor igual o mayor que 1.

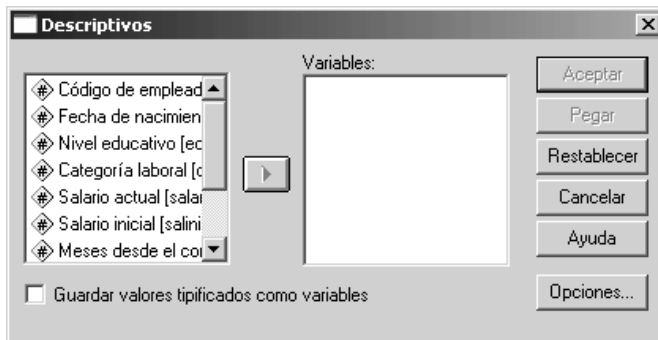
Esta opción es particularmente útil si en la lista **Variables** (Figura 10.1) se han seleccionado variables continuas y categóricas; permite eliminar selectivamente las distribuciones de frecuencias de las variables continuas al tiempo que permite obtener las de las variables categóricas. Generalmente, con variables cuantitativas continuas interesa obtener estadísticos descriptivos o histogramas, pero no distribuciones de frecuencias, pues éstas son demasiado largas y, por tanto, poco informativas.

## Descriptivos

A diferencia de lo que ocurre con el procedimiento **Frecuencias**, que contiene opciones para describir tanto variables categóricas como cuantitativas continuas, el procedimiento **Descriptivos** está diseñado únicamente para variables cuantitativas continuas. Contiene algunos de los estadísticos descriptivos (tendencia central, dispersión y forma de la distribución) que también incluye el procedimiento **Frecuencias**, pero añade una opción especialmente importante: la posibilidad de obtener *puntuaciones típicas* (comúnmente conocidas como *puntuaciones Z*). Para obtener estadísticos descriptivos y puntuaciones típicas:

- Seleccionar la opción **Estadísticos descriptivos > Descriptivos...** del menú **Analizar** para acceder al cuadro de diálogo *Descriptivos* que muestra la Figura 10.7.

Figura 10.7. Cuadro de diálogo *Descriptivos*



La lista de variables del archivo de datos ofrece un listado de todas las variables con formato *numérico*. Las variables con formato de *cadena* no están disponibles en la lista.

Para obtener los estadísticos básicos (media aritmética, desviación típica, valor mínimo y valor máximo) que el procedimiento **Descriptivos** ofrece por defecto:

- Trasladar una o más variables a la lista **Variables**. La especificación mínima requerida es una variable numérica.
- “ **Guardar valores tipificados como variables**. Los valores tipificados, también llamados puntuaciones típicas o puntuaciones *Z* (*Z scores*), expresan el número de desviaciones típicas que cada valor se aleja de su media (ver, más adelante en este mismo capítulo, el apartado *Puntuaciones típicas y curva normal*).

Marcando esta opción, el SPSS crea en el *Editor de datos* una nueva variable con las puntuaciones típicas correspondientes a cada caso del archivo. Esta nueva variable recibe, por defecto, el nombre de la variable original con el prefijo *Z*. Si el nombre de la variable original posee 64 caracteres (que es la longitud máxima de los nombres de variable), el nuevo nombre utiliza los 63 primeros. Por ejemplo, al solicitar las puntuaciones típicas de la variable *salario*, el SPSS crea esas puntuaciones típicas en una nueva variable a la que asigna el nombre *Zsalario*. Si la variable *Zsalario* ya existe, entonces el nombre que se asigna a la nueva variable es *Zsco01* (y así sucesivamente hasta *Zsco99*). El prefijo *Zsco* proviene de *Z score*.

La etiqueta de la nueva variable se obtiene a partir de la etiqueta de la variable original. Si la variable original tiene etiqueta, la etiqueta de la nueva variable se forma con los caracteres de la etiqueta original acompañados del prefijo *Puntua*; si el nombre de la nueva variable no contiene parte del nombre original (lo que ocurre cuando el nombre asignado es, por ejemplo, *Zscore01*), la etiqueta de la nueva variable se forma con *Puntua* (*nombre-de-la-variable-original*) y los caracteres de la etiqueta original. Si la variable original no tiene etiqueta, la etiqueta de la nueva variable se forma con el prefijo *Puntua* acompañado del nombre de la variable original entre paréntesis.

## Opciones

Para seleccionar los estadísticos descriptivos que se desea obtener y para controlar algunos detalles de la presentación:

- Pulsar el botón **Opciones...** del cuadro de diálogo principal (ver Figura 10.7) para acceder al subcuadro de diálogo *Descriptivos: Opciones* que muestra la Figura 10.8.

Figura 10.8. Subcuadro de diálogo *Descriptivos: Opciones*



- **Media.** Media aritmética: suma de todas las puntuaciones dividida por el número de puntuaciones.
- **Suma.** Suma de todos los valores.

**Dispersión.** Puede seleccionarse uno o más de los siguientes estadísticos de dispersión:

- **Desv. típica.** Desviación típica: raíz cuadrada de la varianza. Mide el grado de dispersión de las puntuaciones respecto de su media.
- **Varianza.** Medida de dispersión que se obtiene dividiendo por  $n-1$  ( $n$  = número de casos) la suma de los cuadrados de las diferencias entre cada puntuación y la media.

- " **Amplitud.** Diferencia entre el valor más grande (máximo) y el más pequeño (mínimo). También recibe el nombre de *rango*.
- " **Mínimo.** Valor más pequeño.
- " **Máximo.** Valor más grande.
- " **E.T. media.** Error típico de la media: desviación típica de la distribución muestral de la media. Se obtiene dividiendo la desviación típica de la variable por la raíz cuadrada del número de casos.

**Distribución.** Puede seleccionarse uno o más de los siguientes estadísticos:

- " **Curtosis (o *apuntamiento*).** Índice que expresa el grado en que una distribución acumula casos en sus colas en comparación con los casos acumulados en las colas de una distribución normal con la misma varianza.  
Un índice de curtosis positivo indica que la distribución acumula en una o en las dos colas más casos que la curva normal (lo cual suele coincidir con un mayor *apuntamiento*). Un índice de curtosis negativo indica que la distribución acumula en las colas menos casos que la curva normal. Un índice de curtosis próximo a cero indica *apuntamiento* similar al de la curva normal.  
El *Visor* también ofrece el *error típico* del índice de curtosis, el cual puede utilizarse para tipificar el valor del índice (dividiéndolo por su error típico) e interpretarlo como una puntuación  $Z$  distribuida de forma aproximadamente normal  $N(0, 1)$ . Índices con un valor tipificado mayor que 1,96 en valor absoluto permiten afirmar que la distribución observada se aleja de la distribución normal.
- " **Asimetría.** Índice que expresa el grado de asimetría de la distribución. La asimetría positiva indica que los valores más extremos tienden a situarse por encima de la media. La asimetría negativa indica que los valores más extremos tienden a situarse por debajo de la media. Los valores en torno a cero indican simetría.  
El *Visor* también ofrece el *error típico* del índice de asimetría, el cual puede utilizarse para tipificar el valor del índice de asimetría (dividiéndolo por su error típico) e interpretarlo como una puntuación  $Z$  distribuida de forma aproximadamente normal  $N(0, 1)$ . Índices con un valor tipificado mayor que 1,96 en valor absoluto permiten afirmar que existe asimetría (positiva o negativa, dependiendo del signo del índice).

**Orden de visualización.** Esta opción permite establecer el orden en el que serán listadas las variables en la tabla de descriptivos que ofrece el *Visor*:

**Lista de variables.** Las variables aparecen listadas en el mismo orden en el que han sido seleccionadas en el cuadro de diálogo *Descriptivos*, es decir, en el mismo orden que aparecen en el listado *Variables* de la Figura 10.7. Es la opción que se encuentra activa por defecto.

**Alfabético.** Las variables aparecen listadas en orden alfabético.

**Medias ascendentes.** Las variables se ordenan por el tamaño de sus medias, desde la más pequeña hasta la más grande.

**Medias descendentes.** Las variables se ordenan por el tamaño de sus medias, desde la más grande hasta la más pequeña.

**Ejemplo: Estadísticos descriptivos > Descriptivos > Opciones**

Este ejemplo muestra cómo obtener (guardar) puntuaciones típicas y algunos estadísticos descriptivos utilizando el procedimiento **Descriptivos** (se sigue utilizando el archivo *Datos de empleados*):

- En el cuadro de diálogo principal (ver Figura 10.7), seleccionar las variables *salini* (salario inicial), *salario* (salario actual) y *tiempemp* (meses desde el contrato) y trasladarlas a la lista **Variables**.
- Marcar la opción **Guardar valores tipificados como variables**.
- Pulsar el botón **Opciones...** para acceder al subcuadro de diálogo *Descriptivos: Opciones* (ver Figura 10.8) y marcar las opciones **Media** y **Suma**, y todas las opciones de los recuadros **Dispersión** y **Distribución** (dejar la opción marcada por defecto en el recuadro **Orden de visualización**).

Aceptando estas elecciones, el *Visor de resultados* ofrece la información que muestra la Tabla 10.3. Se ha pivotado la tabla (llevando el icono de *estadísticos* desde la dimensión *columna* a la dimensión *fila*) para poder ajustarla mejor al tamaño de la página.

**Tabla 10.3.** Estadísticos descriptivos

		Estadístico	Error típico
Salario actual	N	474	
	Rango	119.250	
	Mínimo	15.750	
	Máximo	135.000	
	Suma	16.314.875	
	Media	34.419,57	784,311
	Desv. típ.	17.075,661	
	Varianza	291578214,453	
	Asimetría	2,125	,112
	Curtosis	5,378	,224
Meses desde el contrato	N	474	
	Rango	35	
	Mínimo	63	
	Máximo	98	
	Suma	38446	
	Media	81,11	,462
	Desv. típ.	10,061	
	Varianza	101,223	
	Asimetría	-,053	,112
	Curtosis	-1,153	,224
N válido (según lista)	N	474	

Comparando estos resultados con los obtenidos en el ejemplo del procedimiento **Frecuencias** (ver Tabla 10.2), se puede observar que la diferencia radica en que, además de los estadísticos que ofrece el procedimiento **Descriptivos** (todos ellos útiles y apropiados para describir variables cuantitativas continuas), el procedimiento **Frecuencias** ofrece otros estadísticos susceptibles de ser utilizados para describir variables nominales y ordinales (moda, mediana, percentiles, etc.).

Puede comprobarse en el *Editor de datos* que, además de los estadísticos descriptivos de la Tabla 10.3, el procedimiento **Descriptivos** ha generado dos nuevas variables (a las que ha asignado automáticamente los nombres *Zsalario* y *Ztiempemp*) que ha colocado al final del archivo del *Editor de datos*; estas nuevas variables contienen las puntuaciones típicas de las dos variables seleccionadas en el ejemplo: *salario* y *tiempemp* (recuérdese que el nombre de las nuevas variables se forma añadiendo el prefijo Z al nombre original).

Como ejercicio, puede utilizarse el procedimiento **Descriptivos** o el procedimiento **Frecuencias** para comprobar si, tal como se explica en el siguiente apartado, las variables recién tipificadas tienen una media de cero y una desviación típica de uno. (*Nota:* en el caso de que el valor de las medias de las variables tipificadas no sea exactamente 0 y el de las desviaciones típicas no sea exactamente 1, debe tenerse en cuenta que el SPSS trabaja con 16 decimales y que, consecuentemente, el valor del decimosexto decimal está redondeado.)

## Puntuaciones típicas y curva normal

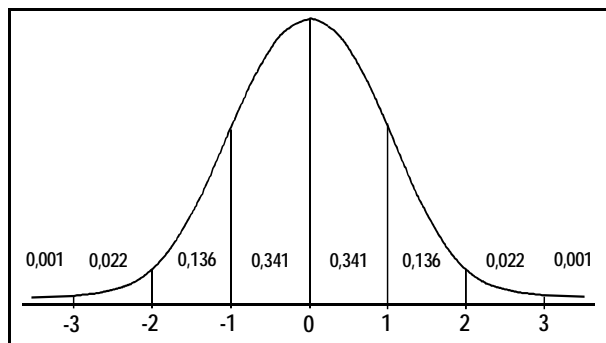
En muchas de las variables que es posible medir, la mayoría de los valores se encuentran próximos al centro de la distribución y van siendo menos frecuentes a medida que va aumentando la distancia al centro. Este tipo de distribuciones tienen forma de campana, y el ejemplo más típico es una distribución teórica llamada *curva normal*. Muchos de los fenómenos que se estudian en la sociedad y en la naturaleza tienen distribuciones muy similares a la distribución teórica normal.

La distribución normal es, probablemente, la distribución teórica más importante en estadística y sirve como punto de referencia para describir cómo se distribuyen muchos de los datos muestrales que se recogen. La explicación de por qué esto es así se encuentra en el *teorema del límite central*, que, formulado en palabras, afirma lo siguiente: si los datos que se recogen son debidos a la suma de cierto número de causas independientes entre sí, cada una con un efecto parcial, siempre que la desviación típica de estos efectos sea finita, la distribución de los datos recogidos se asemejará tanto más a la curva normal cuantos más datos se recojan (cualquiera que sea la distribución original de esos efectos parciales).

La importancia de la distribución normal como referente del comportamiento de los datos que se recogen obliga a describir, aunque sólo sea brevemente, algunas de sus características (ver Figura 10.9):

- Tiene forma de campana (de ahí que sea conocida también como *campana de Gauss*). Esto implica que los valores centrales de la distribución son más probables que los valores que se van alejando del centro de la distribución.
- Es simétrica respecto a su valor central. Al ser simétrica, las medidas de tendencia central (media, mediana, moda, etc.) coinciden.
- Es asintótica respecto al eje de abscisas (por mucho que se extienda, nunca llega a tocarlo), por lo que los valores mínimo y máximo del eje de abscisas son  $-\infty$  y  $+\infty$ .
- Los puntos de inflexión de la curva se encuentran a una desviación típica por encima y por debajo de la media.
- Cualquier combinación lineal de variables normalmente distribuidas también se distribuye según el modelo de probabilidad normal.

Figura 10.9. Curva normal (con probabilidades para algunas puntuaciones típicas)



La mayor parte del trabajo relacionado con variables aleatorias distribuidas normalmente consiste en hallar las probabilidades asociadas a sus valores. Estas probabilidades se obtienen integrando la función de densidad normal. Pero para evitar este tipo de cálculos se han construido tablas con las probabilidades ya calculadas. En cualquier libro de estadística puede encontrarse una tabla con las probabilidades de la curva normal. Ahora bien, esas tablas recogen las probabilidades de una curva normal muy especial: la que tiene media 0 y desviación típica 1; lo cual se expresa así:  $N(0, 1)$ .

Por supuesto, el hecho de que la distribución normal tabulada tenga media 0 y desviación típica 1 no es un problema sino, de hecho, una ventaja, pues cualquier variable puede ser transformada en otra variable equivalente con media 0 y desviación típica 1 sin que se alteren sus propiedades. A esta transformación se le llama *tipificación* o *estandarización* y se realiza de la siguiente manera:

$$Z_i = \frac{X_i - E(X_i)}{S_x}$$

Donde:  $X_i$  se refiere a las puntuaciones originales de la variable

$E(X_i)$  es el valor esperado de  $X_i$  (es decir, su media:  $\bar{X}$ )

$S_x$  es la desviación típica de  $X_i$ :  $S_x = \left[ \sum (X_i - \bar{X})^2 / (n-1) \right]^{1/2}$

Ya se ha señalado que el procedimiento **Descriptivos** permite obtener puntuaciones típicas. Es posible transformar en puntuaciones  $Z_i$  las puntuaciones originales de cualquier variable, siempre que en esa variable tenga sentido el cálculo de la media y de la desviación típica; y una vez obtenidas las puntuaciones  $Z_i$ , es posible, por un lado, describir con precisión la posición relativa de un sujeto dentro de su distribución, pues informan sobre el número de desviaciones típicas que una determinada puntuación se aleja de su media (por arriba o por abajo, dependiendo de que la puntuación típica sea positiva o negativa, respectivamente); y, por otro, conocer la probabilidad asociada a cualquier valor  $X_i$  de una variable normalmente distribuida a partir de la probabilidad asociada a su correspondiente puntuación  $Z_i$ .

La Figura 10.9 muestra las probabilidades asociadas a algunos valores  $Z_i$  (una, dos y tres desviaciones típicas por encima y por debajo del valor central). Recuerdese que los valores

$Z_i$  tienen media 0 y desviación típica 1. Puede observarse en la figura que, aunque el eje de abscisas de una curva normal admite valores desde  $-\infty$  hasta  $+\infty$ , el 99,8% de los casos está comprendido entre las puntuaciones  $\pm 3$ . Además, se sabe que en una curva normal  $N(0, 1)$ , entre  $\pm 1,645$  puntuaciones típicas se encuentra el 90% de los casos; entre  $\pm 1,96$ , el 95% de los casos; entre  $\pm 2,575$ , el 99% de los casos; etc.

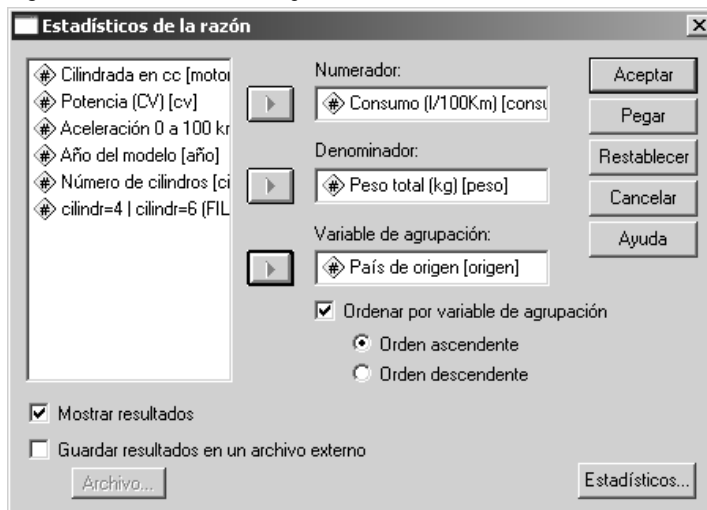
## Descriptivos para el cociente entre dos variables

El estudio del cociente entre dos variables puede resultar útil en muy diversos contextos. Por ejemplo: en un estudio sobre seguridad vial podría interesar analizar el cociente entre el número de accidentes y el número de kilómetros que realiza cada vehículo; en un estudio clínico podría interesar describir el cociente entre la altura y el peso de los pacientes obesos; en un estudio socio-económico podría interesar estudiar, por marcas, el cociente entre el valor de los coches de segunda mano al cabo de dos años y el valor de esos mismos coches cuando son nuevos; en un estudio sobre percepción de un estímulo podría interesar describir el cociente entre el tiempo de reacción de los sujetos y el tiempo de exposición del estímulo; etc.

El procedimiento Estadísticos de la razón ofrece la posibilidad de obtener varios estadísticos descriptivos del cociente (razón) entre dos variables cuantitativas. También permite obtener estos descriptivos para los distintos subgrupos definidos por una variable categórica. Para obtener esta información:

- Seleccionar la opción Estadísticos descriptivos > Razón... del menú Analizar para acceder al cuadro de diálogo *Estadísticos de la razón* que muestra la Figura 10.10.

Figura 10.10. Cuadro de diálogo *Estadísticos de la razón*



La lista de variables del archivo de datos muestra un listado de todas las variables del archivo (cualquiera que sea su formato). Para obtener los descriptivos del procedimiento Estadísticos de la razón:



- ' Seleccionar una variable con formato numérico y trasladarla al cuadro **Numerador**.
  - ' Seleccionar una variable con formato numérico y trasladarla al cuadro **Denominador**.
  - ' Opcionalmente, seleccionar una variable categórica (numérica o de cadena; no importa la longitud de la cadena) y trasladarla al cuadro **Variable de agrupación** para obtener información separada para los distintos subgrupos definidos por los niveles de esa variable categórica.
- " **Ordenar por variable de agrupación.** Si se selecciona una **Variable de agrupación**, esta opción, que se encuentra activa por defecto, hace que los casos del archivo de datos se ordenen según los códigos de esa variable (si se desmarca esta opción, el orden de los casos en el archivo permanece inalterado):
- Orden ascendente.** Los códigos de la **Variable de agrupación** se ordenan de menor a mayor (si la variable de agrupación es numérica) o por orden alfabético de la *a* a la *z* (si la variable es de cadena).
- Descendente.** Los códigos de la **Variable de agrupación** se ordenan de mayor a menor (si la variable de agrupación es numérica) o por orden alfabético de la *z* a la *a* (si la variable es de cadena).
- " **Mostrar resultados.** Esta opción, que se encuentra activa por defecto, hace que el *Visor* muestre los resultados del análisis. Si se desactiva esta opción, es necesario activar la opción **Guardar resultados en un archivo externo**.
- " **Guardar resultados en un archivo externo.** Guarda los resultados del análisis en un archivo de datos SPSS. Tras marcar esta opción es necesario utilizar el botón **Archivo...** para asignar nombre y ubicación al nuevo archivo.
- Si no se ha seleccionado ninguna variable de agrupación, el nuevo archivo de datos contiene un solo caso y tantas variables como estadísticos se hayan solicitado más una variable adicional con el número total de casos. Si se ha seleccionado una variable de agrupación, el nuevo archivo de datos contiene tantos casos como subgrupos defina la variable de agrupación más uno (el referido a la muestra total) y tantas variables como estadísticos se hayan solicitado más una variable para identificar a cada grupo y otra más para indicar el número de casos de cada grupo.

Antes de enumerar los descriptivos disponibles en el procedimiento **Estadísticos de la razón**, conviene tener en cuenta algunos detalles relacionados con los *valores perdidos* y con las *variables de ponderación*.

Si un caso tiene valor perdido *definido por el sistema* en la variable del *numerador* y/o en la variable del *denominador*, el caso es excluido del análisis. Los valores perdidos *definidos por el usuario* pueden incluirse en el análisis modificando la sintaxis (añadiendo la sentencia **MISSING = INCLUDE**). Los valores perdidos *definidos por el sistema* no hay forma de incluirlos en el análisis.

Cuando se utiliza una variable para ponderar el archivo de datos, el SPSS utiliza los valores de la variable de ponderación como frecuencias. Si un caso tiene un peso nulo (cero) o negativo en la variable de ponderación, el caso es excluido del análisis. Puesto que los pesos se consideran frecuencias, los pesos con valor decimal se redondean al entero más próximo. Por ejemplo, un peso de 0,5 se redondea a 1; y un peso de 3,4 se redondea a 3.

## Estadísticos

Para seleccionar los estadísticos descriptivos disponibles en el procedimiento (recuérdese que los estadísticos que se exponen a continuación se aplican a la variable resultante de dividir la variable seleccionada en el *numerador* y la variable seleccionada en el *denominador*):

- En el cuadro de diálogo principal (ver Figura 10.10), pulsar el botón Estadísticos... para acceder al subcuadro de diálogo *Estadísticos de la razón: Estadísticos* que muestra la Figura 10.11.

Figura 10.11. Subcuadro de diálogo *Estadísticos de la razón: Estadísticos*

### Tendencia central. Estadísticos de tendencia central:

- " **Mediana.** Valor por debajo del cual se encuentra el 50 % de los casos. Si el número de casos es par, la mediana se calcula como el promedio de los dos casos centrales cuando éstos se encuentran ordenados. Si el número de casos es impar, la mediana es el valor del caso central de la distribución.
- " **Media.** Media aritmética: suma de todas las puntuaciones dividida por el número de puntuaciones.
- " **Media ponderada.** Cociente entre la media aritmética de la variable del *numerador* y la media aritmética de la variable del *denominador*.
- " **Intervalo de confianza.** Intervalos de confianza para la mediana, para la media y para la media ponderada (siempre que se soliciten estos estadísticos). Los límites del intervalo se calculan, por defecto, con una confianza del 95 %, pero es posible modificar el nivel de confianza introduciendo un valor comprendido entre 0 y 99,99.

**Dispersión.** Las opciones de este recuadro ofrecen algunos estadísticos de dispersión que no se encuentran disponibles en otros procedimientos SPSS:

- " **DPA. Desviación promedio absoluta.** Media de los valores absolutos de las desviaciones de todas las puntuaciones respecto de la media.
- " **CDD. Coeficiente de dispersión.** Desviación promedio absoluta expresada como un porcentaje respecto de la mediana.
- " **DRP. Diferencial relativo al precio** (también llamado *índice de regresibilidad*). Cociente entre la media y la media ponderada.
- " **CDV centrado en la mediana. Coeficiente de variación centrado en la mediana.** Desviaciones centradas en la mediana (es decir, raíz cuadrada de las desviaciones al cuadrado respecto de la mediana divididas por el número de casos menos 1) divididas entre la mediana y multiplicadas por 100.
- " **CDV centrado en la media. Coeficiente de variación centrado en la media.** Desviación típica dividida entre la media y multiplicada por 100.
- " **Desviación típica.** Raíz cuadrada de la varianza. La varianza se obtiene dividiendo por el número de casos menos 1 la suma de los cuadrados de las diferencias entre cada puntuación (cociente) y la media.
- " **Rango.** Diferencia entre el valor más grande (máximo) y el más pequeño (mínimo). También recibe el nombre de *amplitud*.
- " **Mínimo.** Valor más pequeño.
- " **Máximo.** Valor más grande.

**Índice de concentración.** Este índice refleja el porcentaje de casos que caen dentro de un determinado intervalo de valores. Este intervalo puede definirse fijando los límites (mediante la opción **Razones entre...**) o utilizando la mediana como punto de referencia (mediante la opción **Razones dentro del...**):

**Razones entre...** Esta opción permite definir los límites concretos del intervalo cuyo porcentaje de casos se desea conocer. Para definir estos límites es necesario introducir el valor del límite inferior en la casilla **Proporción inferior** y el valor del límite superior en la casilla **Proporción superior** y pulsar el botón **Añadir**.

A pesar de que a los límites del intervalo se les está llamando *proporciones*, estos límites se refieren a los valores de la variable que se está analizando, es decir, a los valores que toma el cociente entre la variable del *numerador* y la del *denominador*. Pueden definirse varios intervalos. Los botones **Borrar** y **Cambiar** permiten eliminar y modificar, respectivamente, intervalos previamente añadidos.

**Razones dentro del...** Esta opción permite definir un intervalo centrado sobre la mediana. Los límites de este intervalo se calculan como un porcentaje del valor de la mediana. Si en la casilla **\_\_ % de la mediana** se introduce, por ejemplo, el valor 20, el límite inferior se calcula como  $[(1-0,01*20)*\text{mediana}]$  y el límite superior como  $[(1+0,01*20)*\text{mediana}]$ . Por tanto, si la mediana vale, por ejemplo, 50, y el porcentaje seleccionado es 20, el límite inferior del intervalo valdrá  $(1-0,001*20)*50=40$  (es decir, el valor de la mediana menos un 20% de su valor) y el límite superior valdrá  $(1+0,001*20)*50=60$  (es decir, el va-

lor de la mediana más un 20 % de su valor). El cuadro de diálogo permite definir varios intervalos. Y los botones **Borrar** y **Cambiar** permiten eliminar y modificar, respectivamente, intervalos previamente añadidos.

### Ejemplo: Razón

Este ejemplo muestra cómo obtener e interpretar los resultados del procedimiento **Estadísticos de la razón**. Se basa en el archivo *Coches*, que se encuentra en la misma carpeta en la que está instalado el SPSS. El interés del análisis se centra en describir el cociente entre el *consumo* de los vehículos y su *peso*, distinguiendo entre los vehículos fabricados en EE.UU., Europa y Japón.

Ahora bien, puesto que la variable *consumo* está expresada en «litros/100 km» y la variable *peso* está expresada en «kg», para no tratar con la variable «litros/kg», que ofrecería cantidades muy pequeñas, se ha creado la variable *peso100* dividiendo la variable *peso* por 100 (mediante la opción **Calcular** del menú **Transformar**). De este modo, el cociente entre *consumo* y *peso100* (que son las dos variables que se van a utilizar en el análisis) expresa el «consumo en litros por cada 100 kg de peso». Para llevar a cabo este análisis:

- En el cuadro de diálogo principal (ver Figura 10.10), seleccionar la variable *consumo* y trasladarla al cuadro **Numerador**.
- Seleccionar la variable *peso100* y trasladarla al cuadro **Denominador**.
- Seleccionar la variable *origen* y trasladarla al cuadro **Variable de agrupación**.
- Pulsar el botón **Estadísticos...** para acceder al cuadro de diálogo *Estadísticos de la razón: Estadísticos* (ver Figura 10.11) y marcar todas las opciones de los recuadros **Tendencia central** y **Dispersión**.
- En el recuadro **Índice de concentración**, introducir los valores 0,5 y 1 en las casillas **Proporción inferior** y **Proporción superior**, respectivamente, y pulsar el correspondiente botón **Añadir**. Introducir el valor 25 en la casilla **% sobre la mediana** y pulsar el correspondiente botón **Añadir**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestran las Tablas 10.4 y 10.5. La Tabla 10.4 informa sobre el número de casos procesados. Indica el tamaño de cada grupo en frecuencia absoluta (*Recuento*) y en frecuencia relativa (*Porcentaje*). También informa sobre el número de casos excluidos del análisis: de los 406 vehículos de que consta el archivo, se han desechado 9 y se han incluido en el análisis los restantes 397.

**Tabla 10.4.** Resumen de los casos procesados

		Recuento	Porcentaje
País de origen	EE.UU.	248	62,5%
	Europa	70	17,6%
	Japón	79	19,9%
Global		397	100,0%
Excluido		9	
Total		406	

La Tabla 10.5 contiene todos los estadísticos disponibles en el procedimiento **Estadísticos de la razón** (esta tabla se ha pivotado para poder ajustarla al tamaño de la página). Puesto que se ha utilizado la variable *origen* como *variable de agrupación*, la tabla muestra los estadísticos tanto para la muestra total como para cada uno de los grupos definidos por la variable *origen* (EE.UU, Europa y Japón).

Tabla 10.5. Estadísticos para la razón *consumo/peso*

		Grupo			
		EE.UU.	Europa	Japón	Global
Media		1,142	1,112	1,091	1,127
IC para la media al 95%	Límite inferior	1,121	1,061	1,045	1,108
	Límite superior	1,164	1,163	1,137	1,145
Mediana		1,120	1,147	1,069	1,121
IC para la mediana al 95%	Límite inferior	1,097	1,100	1,028	1,101
	Límite superior	1,144	1,210	1,171	1,142
	Cobertura real	95,1%	95,9%	95,8%	95,5%
Media ponderada		1,148	1,101	1,088	1,132
IC para la media ponderada al 95%	Límite inferior	1,127	1,048	1,041	1,114
	Límite superior	1,169	1,153	1,134	1,151
Mínimo		,597	,610	,711	,597
Máximo		1,720	1,473	1,836	1,836
Desviación típica		,170	,213	,204	,186
Rango		1,123	,863	1,125	1,239
Desviación absoluta media		,131	,166	,161	,144
Diferencial relacionado con el precio		,995	1,010	1,003	,995
Coefficiente de dispersión		,117	,145	,150	,128
Coefficiente de variación	Media centrada	14,9%	19,1%	18,7%	16,5%
	Mediana centrada	15,4%	18,8%	19,2%	16,6%
Coefficiente de concentración	Porcentaje entre ,5 y 1	20,2%	25,7%	34,2%	23,9%
	Dentro del 25% de la mediana	91,9%	78,6%	86,1%	88,7%

El intervalo de confianza para la mediana se crea sin ningún supuesto acerca de la distribución. El nivel de cobertura real puede ser mayor que el nivel especificado. Otros intervalos de confianza se crean con el supuesto de una distribución normal para las razones.

Tanto las medias como las medianas indican que los vehículos analizados consumen, en promedio, en torno a 1,1 litros por cada 100 kg de peso. Y esto parece ser así en los tres grupos considerados, pues las medias oscilan entre 1,09 y 1,14, y las medianas entre 1,07 y 1,15. En principio, y a falta de otra información, no parece que la razón *consumo/peso* sea muy distinta en los grupos considerados. El valor de las medias ponderadas también es similar en los tres grupos. Los intervalos de confianza son bastante estrechos, lo cual indica que las estimaciones son bastante precisas.

Los vehículos estadounidenses muestran una dispersión *consumo/peso* ligeramente menor: tanto la desviación típica como el coeficiente de dispersión y los dos coeficientes de variación (no así el rango) indican que la variabilidad de las puntuaciones en el grupo estadounidense es menor que en el resto de los grupos.

El vehículo que menos consume en relación a su peso (valor *mínimo*) consume 0,597 litros por cada 100 kg y pertenece al grupo estadounidense. El vehículo que más consume en relación a su peso (valor *máximo*) consume 1,836 litros por cada 100 kg de peso y pertenece al grupo japonés.

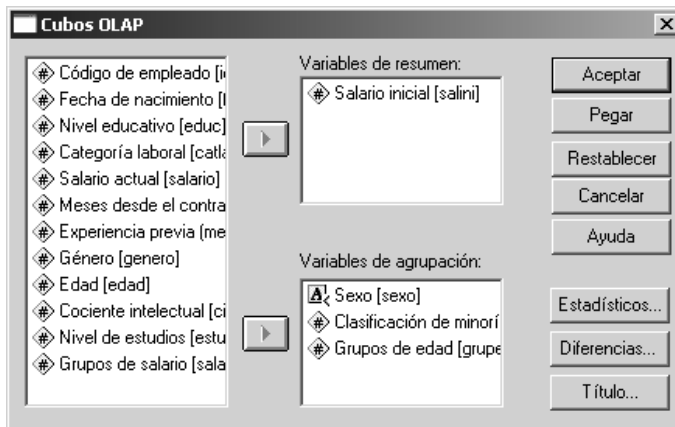
Por último, el coeficiente de concentración indica, en primer lugar, que el 23,93 % de los vehículos consume entre 0,5 y 1 litro por cada 100 kg de peso. Este porcentaje se incrementa sensiblemente en el grupo de vehículos japoneses (sube hasta el 34,18 %). Y en el intervalo definido por el límite inferior equivalente a restar a la mediana el 25 % de su valor y el límite superior equivalente a sumar a la mediana el 25 % de su valor, se encuentra el 88,66 % de los vehículos analizados.

## Cubos OLAP

En ocasiones es necesario obtener estadísticos descriptivos basados en subgrupos de casos. Aunque los procedimientos **Frecuencias** y **Descriptivos** ya estudiados permiten obtener todo tipo de estadísticos descriptivos, ninguno de ellos está diseñado para trabajar con subgrupos. Por el contrario, el procedimiento **Cubos OLAP** (*OnLine Analytical Processing: procesamiento analítico interactivo*) permite obtener fácilmente una amplia variedad de estadísticos descriptivos para los subgrupos resultantes de combinar múltiples variables categóricas. Para obtener estadísticos por subgrupos:

- Seleccionar la opción **Informes > Cubos OLAP** del menú **Analizar** para acceder al cuadro de diálogo *Cubos OLAP* que muestra la Figura 10.12.

Figura 10.12. Cuadro de diálogo *Cubos OLAP*



La lista de variables del archivo de datos ofrece un listado de todas las variables del archivo (numéricas, de cadena corta y de cadena larga). Las variables numéricas pueden utilizarse como variables de *resumen* y como variables de *agrupación*; las variables de cadena únicamente pueden utilizarse como variables de *agrupación*.

**Variables de resumen.** A esta lista deben desplazarse las variables cuantitativas (de intervalo o razón) que se desea describir. Puesto que el procedimiento también incluye estadísticos robustos como la mediana y el rango o amplitud, pueden utilizarse variables ordinales si sólo se desea obtener este tipo de estadísticos.

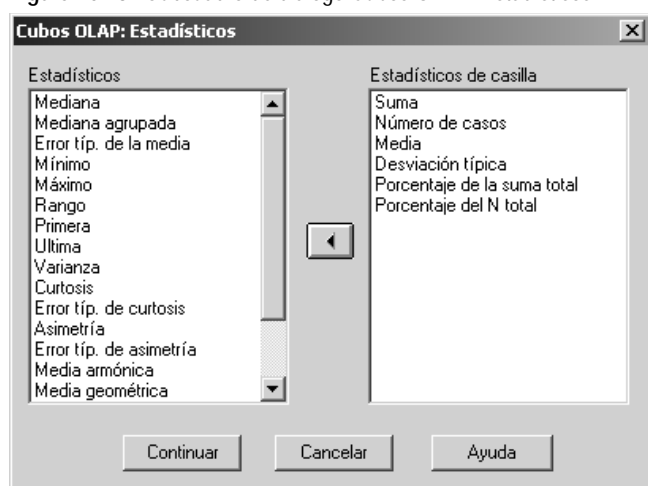
**Variables de agrupación.** Una variable de agrupación es una variable categórica que define grupos. El procedimiento definirá tantos subgrupos como combinaciones resulten de combinar todos los niveles de las variables categóricas seleccionadas como variables de agrupación.

## Estadísticos

Para decidir qué estadísticos se desea obtener:

- Pulsar el botón **Estadísticos...** del cuadro de diálogo principal (ver Figura 10.12) para acceder al subcuadro de diálogo *Cubos OLAP: Estadísticos* que muestra la Figura 10.13.

Figura 10.13. Subcuadro de diálogo *Cubos OLAP: Estadísticos*



**Estadísticos.** Contiene un listado de todos los estadísticos disponibles en el procedimiento. Pulsando sobre cada estadístico con el botón secundario del ratón se obtiene una breve descripción del mismo.

**Estadísticos de casillas.** Estadísticos seleccionados. Los estadísticos incluidos en esta lista son los que más tarde aparecerán en las tablas de resultados. Aunque el procedimiento ya tiene algunos estadísticos seleccionados por defecto, el botón flecha permite seleccionar los estadísticos que se desea utilizar.

## Diferencias

El procedimiento incluye algunas opciones útiles para describir la diferencia entre dos variables. Para utilizar estas opciones:

- Pulsar el botón **Diferencias...** del cuadro de diálogo principal (ver Figura 10.12) para acceder al subcuadro de diálogo *Cubos OLAP: Diferencias* que muestra la Figura 10.14.

Figura 10.14. Subcuadro de diálogo *Cubos OLAP: Diferencias*

**Cubos OLAP: Diferencias**

**Diferencias para estadísticos de resumen**

☐ Ninguna  
☒ Diferencias entre variables  
☐ Diferencias entre grupos

**Tipo de diferencia**

☒ Diferencia de porcentaje  
☒ Diferencia aritmética

Continuar  
Cancelar  
Ayuda

**Diferencias entre variables**

Variable:

Menos variable:

Pares:

Etiqueta de porcentaje:

Etiqueta aritmética:

Eliminar par

**Diferencias entre grupos de casos**

Variable de agrupación:

Categoría:

Menos categoría:

Etiqueta de porcentaje:

Etiqueta aritmética:

Eliminar par

**Diferencias para estadísticos de resumen.** Las opciones de este recuadro permiten decidir qué tipo de diferencias se desea estudiar. Activando la opción *Ninguna* (se encuentra activa por defecto) no es posible obtener estadísticos para la diferencia entre variables. La opción *Diferencias entre variables* permite activar las opciones del recuadro *Diferencias entre variables*. Y la opción *Diferencias entre grupos* permite activar las opciones del recuadro *Diferencias entre grupos de casos*.

**Tipo de diferencia.** El procedimiento *Cubos OLAP* permite trabajar con dos tipos de diferencias: la opción *Diferencia de porcentaje* calcula el porcentaje que representa la diferencia entre dos valores respecto del segundo valor; la opción *Diferencia aritmética* calcula la diferencia entre dos valores.

**Diferencias entre variables.** Para definir una diferencia entre dos variables cuantitativas es necesario: (1) seleccionar en el menú desplegable *Variable* el primer término de la diferencia (la primera variable), (2) seleccionar en el menú desplegable *Menos variable* el segundo término de la diferencia (la segunda variable) y (3) trasladar la selección hecha a la lista *Pares*. El botón *Eliminar par* permite eliminar pares previamente definidos. Los cuadros de texto *Etiqueta de porcentaje* y *Etiqueta aritmética* permiten introducir (opcionalmente) la etiqueta de variable con la que aparecerán en las tablas de resultados la diferencia porcentual y la diferencia aritmética, respectivamente.

**Diferencias entre grupos de casos.** Para definir una diferencia entre dos categorías de una variable categórica es necesario: (1) seleccionar en el menú desplegable *Variable de agrupación* la



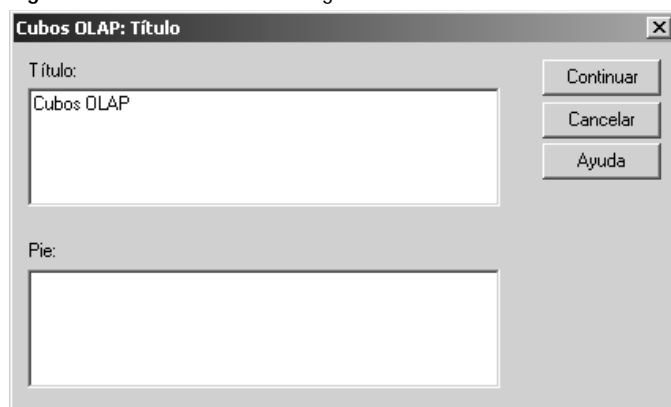
variable categórica cuyos niveles se desea restar, (2) introducir en el cuadro de texto **Categoría** el código que identifica a la categoría minuendo, (3) introducir en el cuadro de texto **Menos categoría** el código que identifica a la categoría sustraendo y (4) trasladar la selección hecha a la lista **Pares**. El botón **Eliminar par** permite eliminar pares previamente definidos. Los cuadros de texto **Etiqueta de porcentaje** y **Etiqueta aritmética** permiten introducir (opcionalmente) la etiqueta de variable con la que aparecerán en las tablas de resultados la diferencia porcentual y la diferencia aritmética, respectivamente.

## Encabezados y pies de tabla

Para modificar el encabezado y el pie de una tabla de resultados:

- En el cuadro de diálogo principal (ver Figura 10.12), pulsar el botón **Título...** para acceder al subcuadro de diálogo *Cubos OLAP: Título* que muestra la Figura 10.15.

Figura 10.15. Subcuadro de diálogo *Cubos OLAP: Título*



El cuadro de texto **Título** permite introducir el título o encabezado de la tabla de resultados; el título por defecto es *Cubos OLAP*. El cuadro de texto **Pie** permite introducir el texto que se desea que aparezca como pie de tabla.

## Ejemplo: *Cubos OLAP*

Este ejemplo muestra cómo obtener algunos estadísticos descriptivos de la variable *salini* (salario inicial) en los subgrupos resultantes de combinar los niveles de tres variables categóricas: *sexo*, *minoría* (clasificación de minorías) y *grupedad* (grupos de edad). Todas estas variables se encuentran en el archivo *Datos de empleados ampliado*, el cual puede obtenerse en la página *web* del manual.

- En el cuadro de diálogo principal (ver Figura 10.12), seleccionar la variable *salini* y trasladarla a la lista **Variables de resumen**.
- Seleccionar las variables *sexo*, *minoría* y *grupedad* y trasladarlas a la lista **Variables de agrupación**.

Pulsar el botón Estadísticos... para acceder al subcuadro de diálogo *Cubos OLAP: Estadísticos* y, en la lista *Estadísticos de casilla*, dejar únicamente los estadísticos **Número de casos**, **Media** y **Desviación típica**. Pulsar el botón *Continuar* para volver al cuadro de diálogo principal.

Aceptando estas selecciones el *Visor* ofrece los resultados que muestran las Tablas 10.6 y 10.7. La Tabla 10.6 contiene un resumen de los casos procesados: número de casos incluidos en el análisis, número de casos excluidos y número total de casos (todo ello en valor absoluto y porcentual).

La Tabla 10.7 contiene los estadísticos solicitados (*n*, *media* y *desviación típica*), calculados para cada subgrupo resultante de combinar los niveles de las tres variables seleccionadas.

**Tabla 10.6.** Resumen de los casos procesados

				Casos		
				Incluidos	Excluidos	Total
Salario inicial * Sexo * Grupos de edad * Clasificación étnica	N		Porcentaje	473	1	474
				99,8%	,2%	100,0%

**Tabla 10.7.** Cubos OLAP

Salario inicial			Grupos de edad					
Clasificación de minorías	Sexo		Menos de 25 años	Entre 25 y 30 años	Entre 30 y 35 años	Entre 35 y 40 años	Más de 40 años	Total
No	Hombre	N	27	92	23	17	34	193
		Media	17,141.85	19,836.63	25,308.26	31,671.18	22,372.06	21,600.78
		Desv. típ.	5,388.79	6,221.57	9,101.71	12,187.40	14,264.14	9,702.44
	Mujer	N	84	26		4	62	176
		Media	12,610.06	16,528.85		13,507.50	13,012.74	13,351.22
		Desv. típ.	2,255.00	3,378.70		4,411.83	3,035.77	3,064.90
	Total	N	111	118	23	21	96	369
		Media	13,712.39	19,107.80	25,308.26	28,211.43	16,327.50	17,666.03
		Desv. típ.	3,809.91	5,868.65	9,101.71	13,234.85	9,840.85	8,402.74
Sí	Hombre	N	8	20	9	9	18	64
		Media	13,856.25	15,679.50	20,703.33	17,193.33	15,725.00	16,383.75
		Desv. típ.	2,048.77	3,624.45	10,925.58	2,988.86	4,891.57	5,570.36
	Mujer	N	10	3	4	3	20	40
		Media	11,355.00	11,850.00	13,200.00	14,250.00	11,670.00	11,951.25
		Desv. típ.	674.72	259.81	714.14	1,299.04	2,391.40	1,928.55
	Total	N	18	23	13	12	38	104
		Media	12,466.67	15,180.00	18,394.62	16,457.50	13,590.79	14,678.94
		Desv. típ.	1,898.68	3,618.06	9,628.01	2,928.44	4,259.17	5,008.24
Total	Hombre	N	35	112	32	26	52	257
		Media	16,390.86	19,094.29	24,013.13	26,659.62	20,071.15	20,301.60
		Desv. típ.	5,002.99	6,044.83	9,696.39	12,134.99	12,240.35	9,129.56
	Mujer	N	94	29	4	7	82	216
		Media	12,476.54	16,044.83	13,200.00	13,825.71	12,685.24	13,091.97
		Desv. típ.	2,175.70	3,507.17	714.14	3,232.98	2,935.71	2,935.60
	Total	N	129	141	36	33	134	473
		Media	13,538.57	18,467.09	22,811.67	23,937.27	15,551.42	17,009.25
		Desv. típ.	3,625.02	5,741.17	9,756.87	12,057.88	8,703.60	7,877.56

Al construir cubos OLAP es importante tener en cuenta que el número de casos de cada subgrupo disminuye rápidamente con cada nueva variable de agrupación que se va añadiendo. En el ejemplo de la Tabla 10.7, del total de 20 subgrupos resultantes, 7 de ellos (un 35%) tienen menos de 10 casos; y un subgrupo está vacío (*minoría*=«no», *sexo*=«mujer» y *grupedad*=«entre 30 y 35 años»).

## Análisis exploratorio

### El procedimiento *Explorar*

Independientemente de la complejidad de los datos disponibles y del procedimiento estadístico que se desee utilizar, una exploración minuciosa de los datos previa al inicio de cualquier análisis posee importantísimas ventajas que un analista de datos no puede pasar por alto (ver Behrens, 1997). Una exploración de los datos permite identificar, entre otras cosas: posibles errores (datos mal introducidos, respuestas mal codificadas, etc.), valores atípicos (valores que se alejan demasiado del resto), pautas extrañas en los datos (valores que se repiten demasiado o que no aparecen nunca, etc.), variabilidad no esperada (demasiados casos en una de las dos colas de la distribución, demasiada concentración en torno a determinado valor), etc. El procedimiento **Explorar** permite detectar este tipo de problemas.

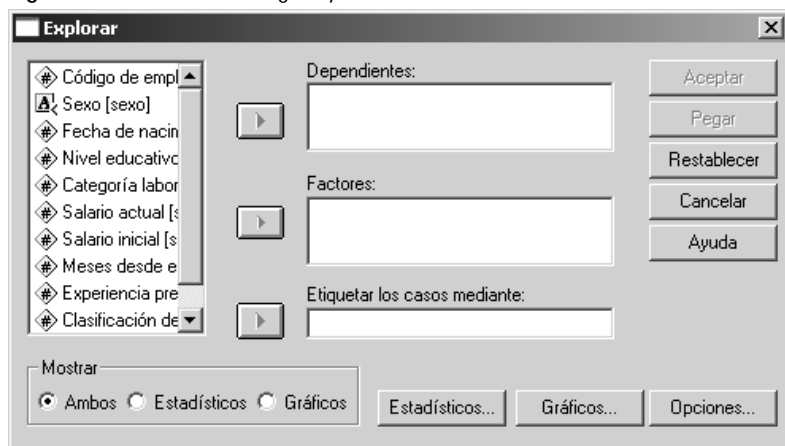
### El procedimiento *Explorar*

Además de incluir gran parte de los estadísticos descriptivos ya estudiados en los procedimientos **Frecuencias** y **Descriptivos**, el procedimiento **Explorar** permite obtener nuevos estadísticos descriptivos, identificar casos atípicos y estudiar con mayor precisión la forma y otras características de una distribución. También permite contrastar dos de los supuestos en que se basan muchas de las técnicas de análisis que se estudiarán más adelante: normalidad y homogeneidad de varianzas. Para obtener todos estos estadísticos:

- Seleccionar la opción **Estadísticos descriptivos > Explorar** del menú **Analizar** para acceder al cuadro de diálogo *Explorar* que muestra la Figura 11.1.

**Dependientes.** Trasladando a esta lista una o más variables y pulsando el botón **Aceptar** se obtienen los estadísticos y gráficos que el procedimiento ofrece por defecto: varios estadísticos descriptivos, un diagrama de tallo y hojas, y un diagrama de caja.

**Factores.** Si en lugar de analizar juntos todos los casos del archivo se desea analizar por separado diferentes grupos de casos (por ejemplo, hombres y mujeres, o cada una de las categorías laborales, etc.), debe introducirse en la lista **Factores** la variable que define esos grupos. Si se introduce más de una variable *factor* se obtendrá, para cada variable *dependiente*, un análisis completo referido a cada uno de los grupos definidos por cada variable *factor*. Mediante sintaxis es posible obtener información referida a subgrupos definidos por la combinación de dos o más *factores*.

Figura 11.1. Cuadro de diálogo *Explorar*

**Etiquetar los casos mediante.** En algunos gráficos (como en los diagramas de caja) y en las tablas de resultados que incluyen listados de casos, los casos individuales se identifican por el número de registro (fila) que ocupan en el *Editor de datos*. Si se desea identificar los casos mediante los valores de alguna variable del archivo de datos, hay que trasladar esa variable a este cuadro.

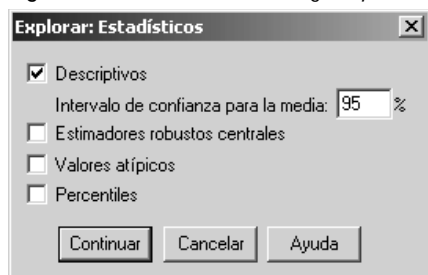
**Mostrar.** Las opciones de este recuadro permiten seleccionar qué parte de los resultados se desea que muestre el *Visor*: estadísticos y gráficos, sólo estadísticos o sólo gráficos.

## Estadísticos

La opción **Estadísticos** (ver Figura 11.1) permite obtener algunos estadísticos adicionales a los que ofrece el procedimiento por defecto. Para seleccionar estos estadísticos:

- Pulsar el botón **Estadísticos...** del cuadro de diálogo principal (ver Figura 11.1) para acceder al subcuadro de diálogo *Explorar: Estadísticos* que muestra la Figura 11.2.

*Nota:* el botón **Estadísticos...** sólo está activo si en el recuadro **Mostrar** está marcada la opción **Ambos** o la opción **Estadísticos**.

Figura 11.2. Subcuadro de diálogo *Explorar: Estadísticos*

- " **Descriptivos.** Esta opción (activada por defecto), ofrece la media aritmética, la mediana, la media truncada o recortada al 5 % (media aritmética calculada eliminando el 5 % de los casos con valores más pequeños y el 5 % de los casos con valores más grandes con el objetivo de obtener una media menos sensible a la presencia de valores extremos), el intervalo de confianza para la media, el error típico de la media, la varianza, la desviación típica, el valor mínimo, el valor máximo, la amplitud, la amplitud intercuartílica, los índices de asimetría y curtosis, y los errores típicos de los índices de asimetría y curtosis.

**Intervalo de confianza para la media:  $k$  %.** Permite fijar el nivel de confianza con el que se desea obtener el intervalo de confianza para la media. El valor de  $k$  por defecto es 95, pero es posible introducir cualquier otro valor entre 1 y 99,99.

- " **Estimadores robustos centrales.** Son estimadores de tendencia central basados en el método de máxima verosimilitud (de ahí que también sean conocidos como estimadores  $M$ ). En realidad, un estimador robusto central o estimador  $M$  no es más que una media ponderada en la que los pesos asignados a los casos dependen de la distancia de cada caso al centro de la distribución: los casos centrales reciben un peso de 1 y los demás valores reciben un peso tanto menor cuanto más alejados se encuentran del centro.

Al igual que ocurre con la media truncada, los estimadores  $M$  son menos sensibles que la media aritmética a la presencia de valores extremos. Por tanto, cuando las distribuciones son muy asimétricas, es preferible utilizar como índices de tendencia central, en lugar de la media aritmética, estos estimadores robustos.

Existen varios estimadores  $M$  que difieren entre sí por la forma concreta de asignar pesos a los casos. El procedimiento **Explorar** incluye cuatro de estos estimadores: Huber, Andrew, Hampel y Tukey (puede encontrarse una descripción detallada de estos estimadores en Norusis y SPSS Inc., 1993, págs. 192-194; y en Palmer, 1999, págs. 124-162).

- " **Valores atípicos.** Muestra los 5 casos con valores más pequeños y los 5 casos con valores más grandes. Si existen empates en los valores ocupados por el quinto caso más pequeño o el quinto más grande, el *Visor* muestra un mensaje indicando tal circunstancia (el número de casos atípicos listados puede controlarse mediante sintaxis).

- " **Percentiles.** Muestra los percentiles 5, 10, 25, 50, 75, 90 y 95. El SPSS incluye diferentes métodos para calcular percentiles. Marcando esta opción se obtienen percentiles calculados con el método **HAVERAGE**, que consiste en asignar al percentil buscado el valor que ocupa la posición  $i = p(n + 1)$  cuando los casos están ordenados de forma ascendente;  $p$  se refiere a la proporción de casos que acumula el percentil buscado (por ejemplo, el percentil 30 acumula una proporción de casos de 0,30), y  $n$  se refiere al tamaño de la muestra. Si el valor de  $i$  no es un número entero, el valor del percentil se obtiene por interpolación:  $X_i(1 - d) + X_{i+1}(d)$ , donde  $X_i$  se refiere al valor que ocupa la posición correspondiente a la parte entera de  $i$ , y  $d$  se refiere a la parte decimal de  $i$ .

El SPSS incluye (disponibles mediante sintaxis) otros cuatro métodos de cálculo de percentiles:

- El método **WAVEREAGE** es idéntico en todo al método **HAVERAGE** excepto en un detalle:  $i = np$ .
- El método **ROUND** asigna al percentil buscado el valor que ocupa la posición correspondiente a la parte entera de  $i = np + 0,5$  (o, lo que es lo mismo, la posición entera más próxima a  $i = np$ ).

- El método EMPIRICAL asigna el valor que ocupa la posición  $i = np$  cuando  $i$  es un número entero, y el valor que ocupa la posición siguiente a la parte entera de  $i$  cuando  $i = np$  es un número decimal.
- El método AEMPIRICAL asigna la media de  $X_i$  y  $X_{i+1}$  cuando  $i = np$  un número entero, y asigna el valor que ocupa la posición siguiente a la parte entera de  $i$  cuando  $i = np$  es un número decimal.

El *Visor* de resultados también ofrece las *bisagras* de Tukey: una versión distinta de los cuartiles clásicos. La primera bisagra (similar al percentil 25) es el valor que ocupa la posición intermedia entre la mediana y el valor más pequeño de la distribución; la segunda bisagra es la mediana; la tercera bisagra (similar al percentil 75) es el valor que ocupa la posición intermedia entre la mediana y el valor más grande de la distribución.

### Ejemplo: Explorar > Estadísticos

Este ejemplo muestra cómo obtener e interpretar los estadísticos del procedimiento **Explorar** utilizando las variables *salario* y *sexo* (del archivo *Datos de empleados*).

- En el cuadro de diálogo principal (ver Figura 11.1), trasladar la variable *salario* (salario actual) a la lista **Dependientes** y la variable *sexo* a la lista **Factores**.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Explorar: Estadísticos* (ver Figura 11.2) y marcar todas las opciones: **Descriptivos**, **Estimadores robustos**, **Valores atípicos** y **Percentiles**.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 11.1 a la 11.4. La primera de ellas (Tabla 11.1) recoge los estadísticos asociados a la opción **Descriptivos**. Se trata de los estadísticos descriptivos clásicos: media, mediana, varianza, desviación típica, rango, índices de asimetría y curtosis, etc.

**Tabla 11.1.** Estadísticos descriptivos

Salario actual		Sexo	
		Hombre	Mujer
Estadístico	Media	\$41,441.78	\$26,031.92
	Intervalo de confianza para la media al 95%	\$39,051.19	\$25,018.29
		\$43,832.37	\$27,045.55
	Media recortada al 5%	\$39,445.87	\$25,248.30
	Mediana	\$32,850.00	\$24,300.00
	Varianza	380219336	57123688,3
	Desv. típ.	\$19,499.214	\$7,558.021
	Mínimo	\$19,650	\$15,750
	Máximo	\$135,000	\$58,125
	Rango	\$115,350	\$42,375
	Amplitud intercuartil	\$22,675	\$7,013
	Asimetría	1,639	1,863
	Curtosis	2,780	4,641
Error típ.	Media	\$1,213.968	\$514.258
	Asimetría	,152	,166
	Curtosis	,302	,330

Como novedad, además de los descriptivos clásicos, aparecen dos estadísticos no recogidos en los procedimientos **Frecuencias y Descriptivos**: la *media recortada* al 5 por ciento (se calcula desechando el 5 por ciento de los valores de cada cola), que adopta un valor más cercano a la mediana que a la media aritmética porque es menos sensible que ésta a la presencia de valores extremos; y la *amplitud intercuartílica*, que refleja la distancia existente entre los cuartiles 1 y 3. La tabla también ofrece los límites del intervalo de confianza para la media al 95 por ciento: se puede estimar, con una confianza del 95 por ciento, que el salario medio de la población de los *hombres* se encuentra entre 39.051,19 y 43.832,37 dólares.

La Tabla 11.2 recoge los estimadores robustos centrales o *estimadores-M*. Todos ellos oscilan en torno a 33.000 \$ en el grupo de varones y en torno a 24.000 \$ en el grupo de mujeres, lo cual representa una estimación de la tendencia central más próxima a los valores de la mediana y de la media recortada que a los de la media aritmética (ésta es más sensible que la mediana a la presencia de valores extremos por una de las dos colas). Las notas a pie de tabla ofrecen el valor de las constantes utilizadas en las ecuaciones de cada estimador.

**Tabla 11.2.** Estimadores-M

Salario actual				
Sexo	Estimador-M de Huber <sup>a</sup>	Biponderado de Tukey <sup>b</sup>	Estimador-M de Hampel <sup>c</sup>	Onda de Andrews <sup>d</sup>
Hombre	\$34,820.15	\$31,779.76	\$34,020.57	\$31,732.27
Mujer	\$24,607.10	\$24,014.73	\$24,421.16	\$24,004.51

a. La constante de ponderación es 1,339.

b. La constante de ponderación es 4,685.

c. Las constantes de ponderación son 1,700, 3,400 y 8,500.

d. La constante de ponderación es  $1,340 \cdot \pi$ .

La Tabla 11.3 muestra los percentiles 5, 10, 25, 50, 75, 90 y 95 calculados con el método W-AVERAGE. Estudiando detenidamente el valor de los percentiles, puede observarse, por ejemplo, que la distribución del *salario actual* es distribución asimétrica positiva: la distancia entre el percentil 10 y el 50 es de 2.200 \$, mientras que la distancia entre el percentil 50 y el 90 es de 6.216 \$ (más del triple). Junto con los percentiles aparecen las tres *bisagras* de Tukey: en el grupo de *hombres*, la tercera bisagra difiere ligeramente del tercer cuartil (percentil 75) calculado con el método W-AVERAGE; en el grupo de mujeres, la primera bisagra difiere ligeramente del primer cuartil (percentil 25).

**Tabla 11.3.** Percentiles

Salario actual				
Percentiles	Sexo			
	Hombre		Mujer	
	Promedio ponderado (definición 1)	Bisagras de Tukey	Promedio ponderado (definición 1)	Bisagras de Tukey
5	\$23,212.50		\$16,950.00	
10	\$25,500.00		\$18,660.00	
25	\$28,050.00	\$28,050.00	\$21,487.50	\$21,525.00
50	\$32,850.00	\$32,850.00	\$24,300.00	\$24,300.00
75	\$50,725.00	\$50,550.00	\$28,500.00	\$28,500.00
90	\$69,325.00		\$34,890.00	
95	\$81,312.50		\$40,912.50	



La Tabla 11.4 ofrece un listado de los cinco casos con los valores más pequeños y los cinco casos con los valores más grandes. Por la parte alta de la distribución de los hombres hay cinco casos con salarios de 100.000 \$ o más, mientras que en la distribución de las mujeres los casos con mayor salario no llegan a 60.000 \$. Estos valores *extremos* no deben confundirse con los valores *atípicos* y *extremos* que se estudian en el siguiente apartado al describir los *diagramas de caja*. Aunque a los valores listados en la Tabla 11.4 también se les llama *extremos*, se trata simplemente de los valores más *grandes* y más *pequeños* de la distribución.

Tabla 11.4. Valores extremos

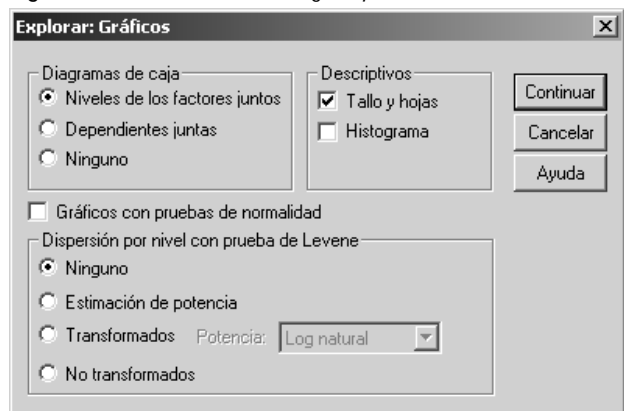
Salario actual		Sexo							
		Hombre				Mujer			
		Mayores		Menores		Mayores		Menores	
		Nº de caso	Valor	Nº de caso	Valor	Nº de caso	Valor	Nº de caso	Valor
1	29	\$135,000	192	\$19,650	371	\$58,125	378	\$15,750	
2	32	\$110,625	372	\$21,300	348	\$56,750	338	\$15,900	
3	18	\$103,750	258	\$21,300	468	\$55,750	411	\$16,200	
4	343	\$103,500	22	\$21,750	240	\$54,375	224	\$16,200	
5	446	\$100,000	65	\$21,900	72	\$54,000	90	\$16,200	

# Gráficos

La opción Gráficos (ver Figura 11.1) ofrece la posibilidad de obtener varios tipos de gráficos (diagramas de caja, diagramas de tallo y hojas, histogramas, gráficos de normalidad y de dispersión) y algunos estadísticos relacionados con los supuestos de normalidad y homogeneidad de varianzas. Para obtener estos gráficos y estadísticos:

- Pulsar el botón Gráficos... del cuadro de diálogo principal (ver Figura 11.1) para acceder al subcuadro de diálogo *Explorar: Gráficos* que muestra la Figura 11.3 (el botón Gráficos... sólo está disponible si en el recuadro **Mostrar** está marcada la opción **Ambos** o la opción **Gráficos**).

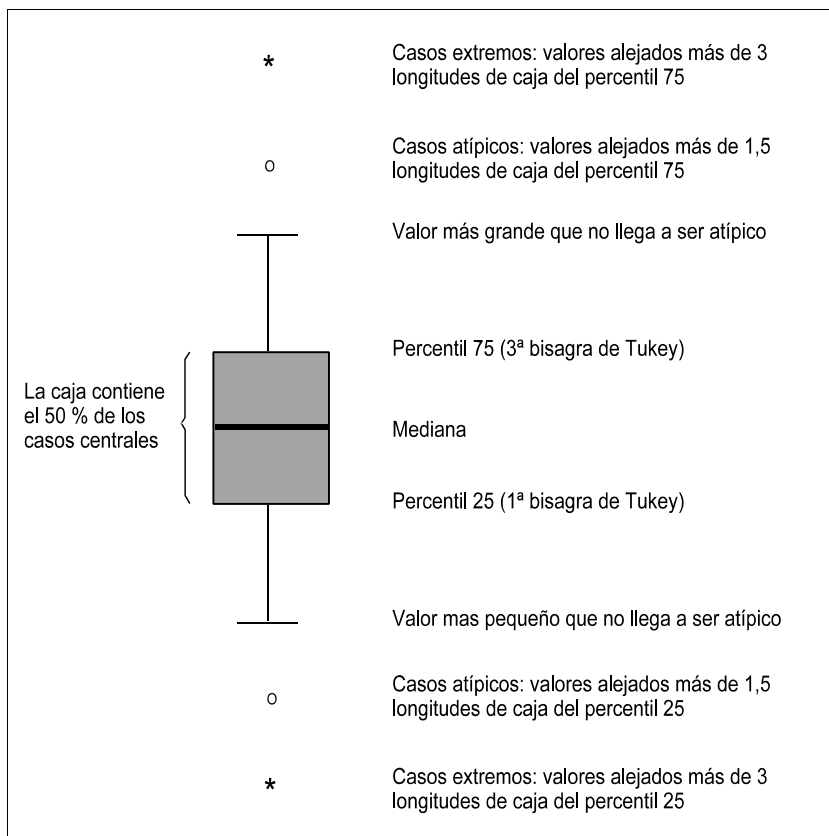
Figura 11.3. Subcuadro de diálogo *Explorar: Gráficos*



## Diagramas de caja

La Figura 11.4 describe los detalles de un *diagrama de caja*. El diagrama incluye la mediana, los percentiles 25 y 75 (en realidad son las bisagras de Tukey), y una serie de valores (atípicos, extremos) que junto con la mediana y la propia caja proporcionan información bastante completa sobre, entre otras cosas, el *grado de dispersión* de los datos y el *grado de asimetría* de la distribución (ver Tukey, 1977).

Figura 11.4. Detalles de un diagrama de caja



Las opciones del recuadro **Diagramas de caja** (ver Figura 11.3) permiten decidir si se desea o no obtener diagramas de caja y, en caso afirmativo, optar entre dos formas diferentes de organizar los diagramas solicitados:

**Niveles de los factores juntos.** Muestra un gráfico diferente para cada variable dependiente. En cada uno de esos gráficos aparecen juntos los diagramas de caja correspondientes a los grupos definidos por una variable *factor*. Si no se ha seleccionado ninguna variable *factor*, cada gráfico muestra un solo diagrama de caja: el correspondiente a toda la muestra. Esta opción resulta útil para comparar distintos grupos en la misma variable. Es la opción por efecto. Para obtener estos diagramas de caja:

- En el cuadro de diálogo principal (ver Figura 11.1), trasladar las variables *salini* (salario inicial) y *salario* (salario actual) a la lista **Dependientes** y la variable *sexo* a la lista **Factores**.
- Pulsar el botón **Gráficos...** para acceder al subcuadro de diálogo *Explorar: Gráficos* (ver Figura 11.3) y marcar la opción **Niveles de los factores juntos** del recuadro **Diagramas de caja**.

Aceptando estas elecciones, se obtienen los diagramas de caja que muestran las Figuras 11.5.a y 11.5.b.

Figura 11.5.a. Diagramas de caja de la variable *salini* (salario inicial) en *hombres* y *mujeres*

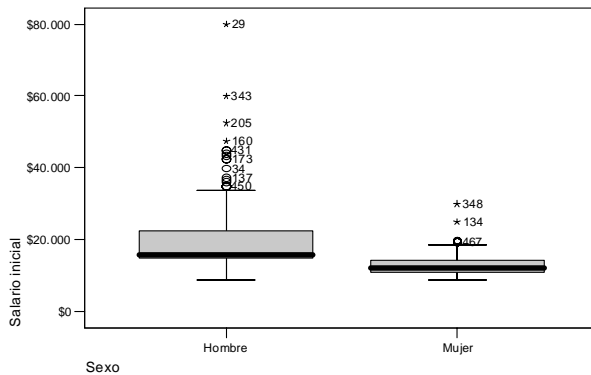
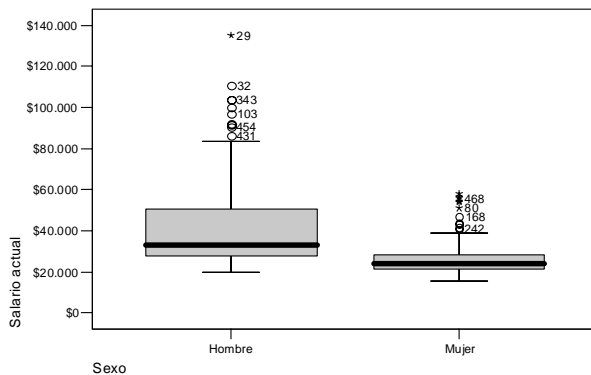


Figura 11.5.b. Diagramas de caja de la variable *salario* (salario actual) en *hombres* y *mujeres*



Puede verse, en primer lugar, que la opción **Niveles de los factores juntos** genera dos gráficos: uno para cada variable dependiente seleccionada. En cada gráfico aparecen juntos los diagramas de caja correspondientes a los dos grupos definidos por la variable *factor* (hombres y mujeres). Las medianas informan del salario medio de cada grupo: la mediana de los *hombres* es mayor que la de las *mujeres* en ambas variables dependientes. Una mediana desplazada del centro de la caja delata la presencia de asimetría: los diagramas obtenidos indican que tanto en el grupo de hombres como en el de mujeres (aunque de forma

más acusada en el de hombres), la mediana está desplazada hacia abajo, lo que está delatando la presencia de asimetría positiva.

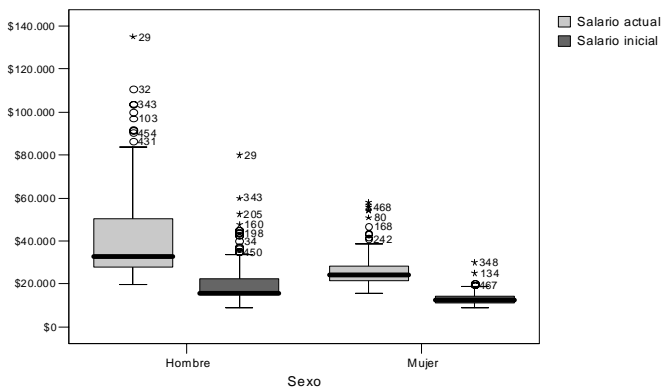
Las cajas (cuya altura representa la amplitud intercuartílica; es decir, la distancia existente entre el primer y el tercer cuartil) muestran el grado de dispersión del 50 por ciento de los casos centrales: en ambas variables, las cajas correspondientes al grupo de *hombres* reflejan una amplitud mayor que las cajas del grupo de mujeres.

Los bigotes y los casos atípicos y extremos indican hacia dónde se desplazan los valores más alejados del centro. En todos los casos se observan valores extremos (asteriscos) y atípicos (círculos) por la parte alta de las distribuciones, lo cual indica, de nuevo, asimetría positiva (de forma más acusada en el grupo de *hombres*).

**Dependientes juntas.** Muestra un gráfico diferente para cada uno de los grupos definidos por la variable *factor*. En cada uno de esos gráficos aparecen juntos los diagramas de caja correspondientes a cada variable *dependiente* seleccionada. Si el número de variables dependientes y de grupos es pequeño, el SPSS intenta mostrar en un solo gráfico todos los diagramas solicitados. Si no se ha seleccionado ninguna variable *factor*, aparece un solo gráfico con tantos diagramas de caja como variables *dependientes* se hayan seleccionado. Esta opción resulta útil para comparar diferentes variables dentro del mismo grupo. Para obtener estos diagramas de caja:

- Seleccionar las mismas variables dependientes (*salini* y *salario*) y la misma variable factor (*sexo*), pero marcar en el recuadro **Diagramas de caja** la opción **Dependientes juntas** para obtener los diagramas de caja que muestra la Figura 11.6.

**Figura 11.6.** Diagramas de caja de las variables *salini* (salario inicial) y *salario* (salario actual) en cada nivel de la variable *sexo*



Ahora aparecen juntos los diagramas de caja correspondientes a cada variable dependiente. Como el número de variables dependientes y de factores es pequeño, en lugar de generar dos gráficos, uno para cada grupo, el SPSS ha construido un solo gráfico con los dos grupos. Con estos diagramas pueden extraerse las mismas conclusiones que con los diagramas de las Figuras 11.5.a y 11.5.b. Pero ahora, además, puede constatar que los promedios de la variable *salario actual* son más altos que los de la variable *salario inicial*.

**Ninguno.** Suprime de los resultados los diagramas de caja.

<sup>u</sup> **Tallo y hojas.** Esta opción, que se encuentra activa por defecto, permite obtener gráficos similares a los histogramas, pero con información más precisa que éstos (ver Tukey, 1977). La Figura 11.7 muestra el *diagrama de tallo y hojas* de la variable *edad* obtenido con una muestra de 148 sujetos.

**Figura 11.7.** Diagrama de tallo y hojas de la variable *edad*

```

Frequency      Stem & Leaf
 12.00         2 s  666677777777
 28.00         2 .  8888888888888888888899999999
 30.00         3 *  000000000000000111111111111111
 13.00         3 t  2222222222333
   7.00         3 f  4445555
   4.00         3 s  6666
 12.00         3 .  8888999999999
 15.00         4 *  0000111111111111
   4.00         4 t  3333
   4.00         4 f  4444
   7.00         4 s  6666666
   4.00         4 .  8889
   3.00         5 *  011
   3.00         5 t  222
   2.00 Extremes (59) (64)

Stem width:      10.00
Each leaf:       1 case(s)

```

Al igual que en un histograma, la longitud de las líneas refleja el número de casos que pertenecen a cada intervalo. Cada caso (o grupo de casos) está representado por un número que coincide con el valor de ese caso en la variable. En un diagrama de tallo y hojas cada valor se descompone en dos partes: el primer o primeros dígitos forma(n) el *tallo* (*Stem*) y el dígito que sigue a los utilizados en el tallo forma las hojas (*Leaf*). Por ejemplo, el valor 23 puede descomponerse en un tallo de 2 y una hoja de 3; el valor 12.300 puede descomponerse en un tallo de 12 y una hoja de 3; etc.

Cada tallo puede ocupar una sola fila o más de una. Si ocupa una sola fila, sus hojas contienen dígitos del 0 al 9; si ocupa dos filas, las hojas de la primera fila contienen dígitos del 0 al 4 y las de la segunda fila dígitos del 5 al 9. Etc. En el diagrama de la Figura 11.7, los tallos (excepto el primero y el último) ocupan cinco filas: la primera fila contiene los dígitos 0 y 1 (encabezadas con un asterisco); la segunda, los dígitos 2 y 3 (con el encabezado «*t*» = *two, three*); la tercera, los dígitos 4 y 5 (con el encabezado «*f*» = *four, five*); la cuarta, los dígitos 6 y 7 (con el encabezado «*s*» = *six, seven*); y la quinta, los dígitos 8 y 9 (encabezadas con un punto).

La anchura del tallo viene indicada en la parte inferior del diagrama (*Stem width*) y es un dato imprescindible para interpretar correctamente el diagrama. En el ejemplo de la Figura 11.7 el tallo tiene una anchura de 10, lo que significa que los valores del tallo hay que multiplicarlos por 10. Así, un tallo de 1 vale 10, un tallo de 2 vale 20, un tallo de 5 vale 50, etc.

Las hojas completan la información del tallo. Así, un tallo de 4 con una hoja de 3 representa una edad de 43 años; un tallo de 5 con una hoja de 0 representa una edad de 50 años; etc. El número de casos que representa cada hoja (cada hoja puede representar a más de un caso) viene indicado en la parte inferior del diagrama, en *Each leaf*. Así, en la Figura 11.7: *Each leaf* = 1 caso; en la Figura 11.8: *Each leaf* = 3 casos.

Cuando la anchura del tallo vale 10 (como en el diagrama de la Figura 11.7), los dígitos de las hojas son unidades; cuando la anchura del tallo vale 100, los dígitos de las hojas son decenas; cuando la anchura del tallo vale 1.000 (como en el diagrama de la Figura 11.8), los dígitos de las hojas son centenas. Etc.

La Figura 11.8 muestra el diagrama de tallo y hojas de la variable *salario* (salario actual). Ahora, la anchura del tallo vale 10.000, lo que significa que un tallo de 1 equivale a 10.000 \$, un tallo de 2 equivale a 20.000 \$, etc. Así, por ejemplo, en el tallo 1 hay 1(3) = 3 casos cuyo salario es de 15.000 \$; 3(3) = 9 casos cuyo salario es 16.000 \$; 2(3) = 6 casos cuyo salario es 17.000 \$; 1(3) = 3 casos cuyo salario es de 18.000 \$; y 4(3) = 12 casos cuyo salario es de 19.000 \$; todos estos casos suman 33, que es justamente la frecuencia informada en la primera columna del diagrama. No obstante, conviene señalar que estos valores no siempre son exactos, sino aproximados: en el tallo 5, por ejemplo, se informa de una frecuencia de 7 y sin embargo aparecen 3 hojas (que equivalen a 9 casos).

**Figura 11.8.** Diagrama de tallo y hojas de la variable *salario actual*

```
Salario actual Stem-and-Leaf Plot

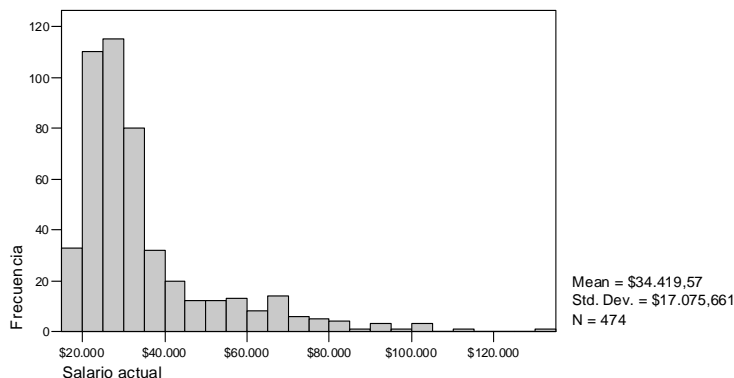
Frequency      Stem & Leaf

  33.00         1 .  56667789999
 110.00         2 .  0000111111122222222233334444444444
 115.00         2 .  555555566666666667777777778888889999999
   80.00         3 .  000000000001111112233333444
   32.00         3 .  55556677889
   20.00         4 .  0001233&
   12.00         4 .  5678&
   12.00         5 .  0124&
    7.00         5 .  556
   53.00 Extremes      (>=56750)

Stem width:      10000
Each leaf:       3 case(s)
```

La última fila del diagrama muestra el número de casos con valores extremos y los valores concretos que toman esos casos (entre paréntesis). Así, por ejemplo, en el diagrama de la Figura 11.7 aparecen 2 casos extremos, con edades de 59 y 64 años; y en el diagrama de la Figura 11.8, 53 casos extremos con un salario de al menos 56.750 \$.

- “ **Histograma.** Un histograma se construye agrupando los datos en intervalos de la misma amplitud y levantando barras de altura proporcional al número de casos de cada intervalo. Aunque esta opción permite obtener histogramas con amplitud calculada de forma automática, tanto la amplitud de los intervalos como otros aspectos del histograma pueden controlarse utilizando el *Editor de gráficos*. La Figura 11.9 muestra un histograma de la variable *salario* (salario actual). Se trata de la misma variable representada en el diagrama de tallo y hojas de la Figura 11.8, por lo que es posible comparar ambos diagramas y observar las coincidencias y diferencias existentes entre ellos.

Figura 11.9. Histograma de la variable *salario actual*

## Cómo contrastar supuestos

Muchos de los procedimientos estadísticos que se estudiarán en los próximos capítulos se apoyan en dos supuestos básicos: (1) *normalidad*: las muestras con las que se trabaja proceden de poblaciones distribuidas normalmente; y (2) *homocedasticidad* u *homogeneidad de varianzas*: esas poblaciones normales poseen la misma varianza. El subcuadro de diálogo *Explorar: Gráficos* (ver Figura 11.3) incluye varios estadísticos y gráficos para contrastar estos supuestos.

## Normalidad

“ **Gráficos con pruebas de normalidad.** Esta opción permite obtener dos gráficos de normalidad (*Q-Q normal* y *Q-Q normal sin tendencia*) junto con dos pruebas de significación: *Kolmogorov-Smirnov* (Kolmogorov, 1933; Smirnov, 1948; Lilliefors, 1967) y *Shapiro-Wilk* (Shapiro y Wilk, 1965).

Las pruebas de significación permiten contrastar la hipótesis nula de que las muestras utilizadas han sido extraídas de poblaciones normales. Para contrastar esta hipótesis, el SPSS ofrece, por defecto, el estadístico de Kolmogorov-Smirnov con las probabilidades de Lilliefors para el caso en el que la media y la varianza poblacionales son desconocidas y necesitan ser estimadas. Y en el caso de que el tamaño muestral sea igual o menor que 50 ofrece, además, el estadístico de Shapiro-Wilk. Para obtener estos gráficos y estadísticos:

- En el cuadro de diálogo principal (ver Figura 11.1), trasladar la variable *salario* (salario actual) a la lista **Dependientes** y la variable *estudios* (nivel de estudios) a la lista **Factores** (estas variables están disponibles en el archivo *Datos de empleados ampliado*, el cual puede obtenerse en la página web del manual).
- Pulsar el botón **Gráficos...** (ver Figura 11.1) para acceder al subcuadro de diálogo *Explorar: Gráficos* (ver Figura 11.3), y marcar la opción **Gráficos con pruebas de normalidad** para obtener el resultado que muestra la Tabla 11.5.

La Tabla 11.5 ofrece los estadísticos de Kolmogorov-Smirnov y de Shapiro-Wilk acompañados de sus correspondientes niveles críticos (*Sig.*). Ambos permiten contrastar la hipótesis nula de que los datos muestrales proceden de poblaciones normales: se rechaza la hipótesis de normalidad cuando el nivel crítico (*Sig.*) es menor que el nivel de significación establecido (generalmente 0,05).

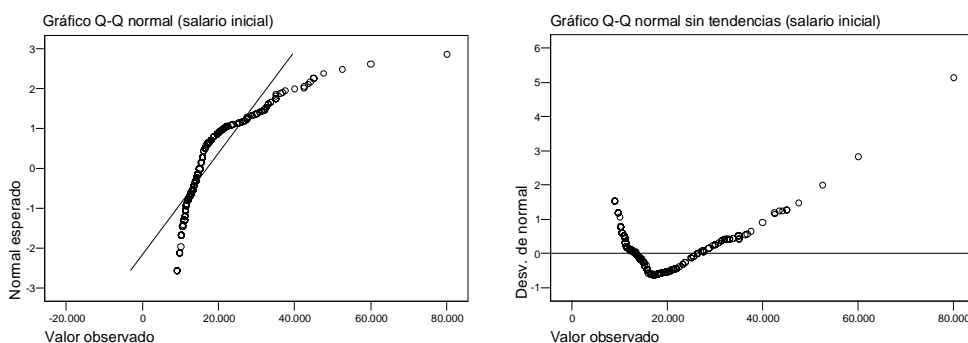
En el ejemplo, los estadísticos del grupo *secundarios* y del grupo *medios* tienen asociados niveles críticos menores que 0,05, lo que debe llevar a concluir que el salario de los grupos *secundarios* y *medios* no procede de poblaciones normales. El estadístico de Shapiro-Wilk sólo aparece con el nivel de estudios *superiores* porque es el único grupo con un tamaño igual o menor que 50.

Tabla 11.5. Contrastes de normalidad

Salario actual		Nivel de estudios			
		Primarios	Secundarios	Medios	Superiores
Kolmogorov-Smirnov	Estadístico	,119	,079	,154	,113
	gl	53	190	181	50
	Sig.	,057	,006	,000	,148
Shapiro-Wilk	Estadístico	,946	,915	,807	,945
	gl	53	190	181	50
	Sig.	,018	,000	,000	,021

El problema de estos y otros estadísticos de normalidad es que, con muestras muy grandes, son demasiado sensibles a pequeñas desviaciones de la normalidad. Por esta razón, es conveniente acompañar estos estadísticos con algún gráfico de normalidad. Según se ha señalado ya, el procedimiento **Explorar** ofrece dos gráficos de normalidad: el *Q-Q normal* y el *Q-Q normal sin tendencias*. Ambos gráficos se ofrecen en la Figura 11.10.

Figura 11.10. Gráficos de normalidad



En un gráfico *Q-Q normal*, cada valor observado ( $Y_i$ ) es comparado con la puntuación típica  $NZ_i$  que teóricamente le correspondería a ese valor en una distribución normal estandarizada (para comprender cómo se calculan esas puntuaciones típicas normales puede consultarse, en el Capítulo 5, el apartado *Asignar rangos: Tipos de rangos*). En el eje de abscisas están representados los valores observados ordenados desde el más pequeño al más grande ( $Y_i$ ); en el de



ordenadas están representadas las puntuaciones típicas normales ( $NZ_i$ ). Cuando una muestra procede de una población normal, los puntos correspondientes a cada par se encuentran agrupados en torno a la diagonal representada en el diagrama. Las desviaciones de la diagonal indican desviaciones de la normalidad.

Un gráfico *Q-Q normal sin tendencias* muestra las *diferencias* existentes entre la puntuación típica observada de cada valor ( $Z_i$ ) y su correspondiente puntuación típica normal ( $NZ_i$ ). Es decir, muestra las distancias verticales existentes entre cada punto del gráfico *Q-Q normal* y la recta diagonal que atraviesa ese gráfico. En el eje de abscisas están representados los valores observados ( $Y_i$ ) y en el de ordenadas el tamaño de las diferencias entre las puntuaciones típicas observadas y las esperadas ( $Z_i - NZ_i$ ). Si las puntuaciones proceden de una población normal, esas diferencias deben oscilar de forma aleatoria en torno al valor cero (línea recta horizontal). La presencia de pautas de variación no aleatorias indica desviaciones de la normalidad.

Los diagramas de las Figuras 11.11.a, 11.11.b y 11.11.c ofrecen varios ejemplos que pueden ayudar a comprender el significado de los gráficos de normalidad. Estos diagramas muestran el comportamiento de tres muestras de puntuaciones aleatoriamente extraídas de tres distribuciones de probabilidad diferentes: una distribución *normal*, una distribución *uniforme* y una distribución *ji-cuadrado* (para obtener estas muestras de puntuaciones se han utilizado las funciones RV.NORMAL, RV.UNIFORM y RV.CHISQ del procedimiento Calcular del menú Transformar).

Figura 11.11.a. Gráficos de normalidad: muestra extraída de una distribución *normal* (media=10, des. típ.=3)

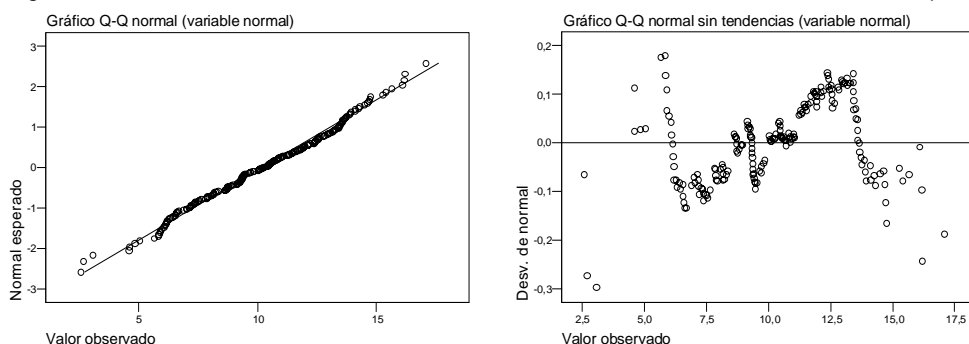


Figura 11.11.b. Gráficos de normalidad: muestra extraída de una distribución *uniforme* (rango 0, 1)

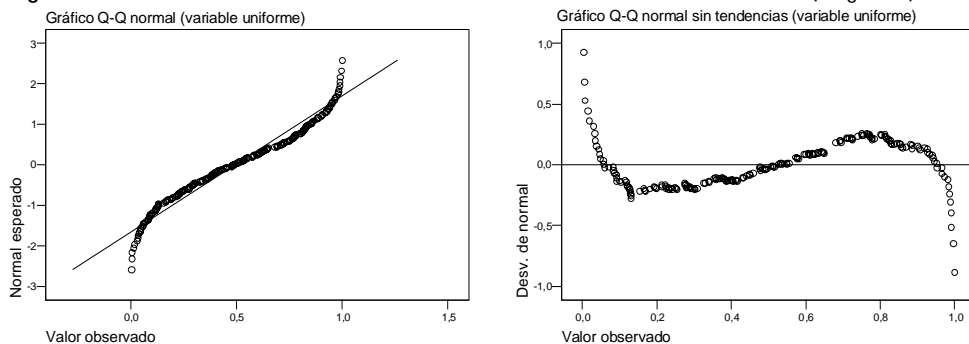
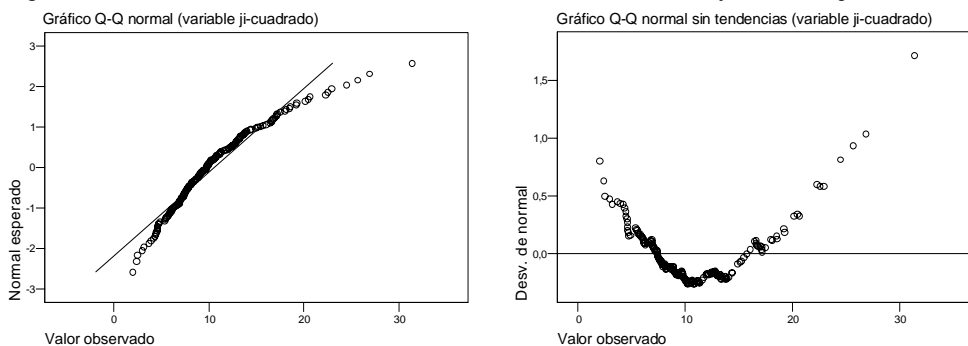


Figura 11.11.c. Gráficos de normalidad: muestra extraída de una distribución *ji-cuadrado* ( $gl = 10$ )



Puede observarse que, cuando una muestra de puntuaciones se distribuye normalmente (Figura 11.11.a), los puntos del diagrama *Q-Q normal* se ajustan a la diagonal y los puntos del diagrama *Q-Q normal sin tendencia* se distribuyen aleatoriamente sin mostrar una pauta clara. Por el contrario, cuando una muestra de puntuaciones procede de una distribución uniforme (Figura 11.11.b) o de una distribución *ji-cuadrado* (Figura 11.11.c), los puntos del diagrama *Q-Q normal* no se ajustan a la diagonal y los puntos del diagrama *Q-Q normal sin tendencia* muestran una pauta de variación claramente no aleatoria.

## Homogeneidad de varianzas

Además del supuesto de normalidad, el procedimiento **Explorar** también permite contrastar el supuesto de homogeneidad de varianzas (lo que requiere, obviamente, haber seleccionado al menos una variable en la lista **Factor** del cuadro de diálogo *Explorar* –ver Figura 11.1).

El recuadro **Dispersión por nivel con prueba de Levene** del subcuadro de diálogo *Explorar: Gráficos* (ver Figura 11.3) proporciona: (1) la prueba de *Levene* (1960) para contrastar la hipótesis de que los grupos definidos por la variable *factor* proceden de poblaciones con la misma varianza y (2) un gráfico de dispersión de la variable dependiente en cada nivel definido por la variable *factor* (gráfico de *dispersión por nivel*).

La **prueba de Levene** consiste en llevar a cabo un análisis de varianza de un factor (ver Capítulo 14) utilizando como variable dependiente la diferencia en valor absoluto entre cada puntuación individual y la media (o la mediana, o la media recortada) de su grupo. Para obtener la prueba de Levene:

- En el cuadro de diálogo principal (ver Figura 11.1), trasladar la variable *salario* (salario actual) a la lista **Dependientes** y la variable *estudios* (nivel de estudios) a la lista **Factores** (estas variables están disponibles en el archivo *Datos de empleados ampliado*, el cual puede obtenerse en la página *web* del manual).
- Pulsar el botón **Gráficos** para acceder al cuadro de diálogo *Explorar: Gráficos* (ver Figura 11.3) y marcar la opción **No transformados**.

*Nota:* para poder obtener el estadístico de Levene es necesario que en el recuadro **Mostrar** del cuadro de diálogo *Explorar* (ver Figura 11.1) esté marcada la opción **Ambos**.

Aceptando estas elecciones, se obtiene el estadístico de Levene que recoge la Tabla 11.6. El nivel crítico (*Sig.*) asociado al estadístico permite contrastar la hipótesis de homogeneidad de varianzas: si el valor del nivel crítico es menor que 0,05, debe rechazarse la hipótesis de homogeneidad. En el ejemplo, el nivel crítico (cualquiera que sea el estimador de tendencia central a partir del cual se obtengan las diferencias) es menor que 0,0005, por lo que puede afirmarse que la varianza de la variable *salario actual* no es la misma en las cuatro poblaciones definidas por la variable *estudios*.

Tabla 11.6. Pruebas de homogeneidad de varianzas

Salario actual				
	Estadístico de Levene	gl1	gl2	Sig.
Basándose en la media	28,085	3	470	,000
Basándose en la mediana	21,799	3	470	,000
Basándose en la mediana y con gl corregidos	21,799	3	266,893	,000
Basándose en la media recortada	24,767	3	470	,000

Además de la prueba de Levene, el recuadro **Dispersión por nivel con prueba de Levene** (ver Figura 11.3) contiene algunas opciones relacionadas con el **gráfico de dispersión por nivel**:

**Ninguno.** Con esta opción activa, el *Visor* no ofrece ni la prueba de Levene sobre homogeneidad de varianzas ni el gráfico de *dispersión por nivel*. Es la opción que se encuentra activa por defecto.

**Estimación de potencia.** Cuando se incumple el supuesto de homogeneidad de varianzas (supuesto necesario para poder utilizar con garantía algunos procedimientos estadísticos como el *análisis de varianza*), es práctica frecuente aplicar algún tipo de transformación a los datos originales para conseguir homogeneizar las varianzas.

Una transformación basada en *potencias* consiste en elevar las puntuaciones originales a una potencia específica. Para determinar la potencia apropiada, el SPSS genera un gráfico de dispersión comparando, para cada grupo, el logaritmo neperiano de la mediana (en el eje de abscisas) con el logaritmo neperiano de la amplitud intercuartílica (en el eje de ordenadas). Cuando las varianzas son iguales, los puntos del gráfico se encuentran a la misma altura, es decir, alineados horizontalmente.

La Figura 11.12 muestra un gráfico de *dispersión por nivel* referido a las variables *salario actual* (dispersión) y *nivel de estudios* (nivel). El gráfico contiene 4 puntos, uno por cada nivel de la variable *estudios*. El hecho de que los puntos no se encuentren horizontalmente alineados indica que las varianzas no son homogéneas (lo cual coincide con la información obtenida con el estadístico de Levene).

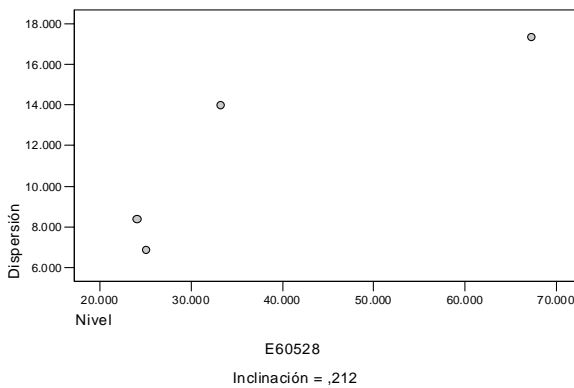
El gráfico también muestra el valor de la pendiente de la recta de regresión (ver Capítulo 18) obtenida por el método de mínimos cuadrados (*Inclinación* = 0,212). Basándose en el valor de la pendiente, el SPSS ofrece una estimación de la potencia a la que habría que elevar las puntuaciones de la variable *dependiente* (*salario actual*) para homogeneizar las varianzas de esa variable en cada nivel de la variable *factor* (*nivel de estudios*); o, mejor, para intentar homogeneizarlas, porque lo cierto es que no siempre se consigue.

La estimación del valor de esa potencia se obtiene restando a uno el valor de la pendiente de la recta de regresión; en el ejemplo:  $1 - 0,212 = 0,788$ . La potencia así estimada

puede tomar cualquier valor. Sin embargo, lo habitual es utilizar potencias redondeadas a múltiplos de 0,5 (incluyendo el cero). Algunas de las potencias más utilizadas para transformar datos son las siguientes:  $-1 = \text{recíproco}$ ;  $-1/2 = \text{recíproco de la raíz cuadrada}$ ;  $0 = \text{logaritmo natural}$ ;  $1/2 = \text{raíz cuadrada}$ ;  $1 = \text{sin transformación}$ ;  $2 = \text{cuadrado}$ ;  $3 = \text{cubo}$ . Todas estas transformaciones, que son las habitualmente recomendadas en la literatura estadística, están recogidas en la opción **Transformados**.

Con el valor de potencia obtenido en el ejemplo (0,788), se utilizaría una potencia redondeada de 1, que equivale a no efectuar ningún tipo de transformación; lo cual significa que el procedimiento no ha encontrado un valor de potencia que permita homogeneizar las varianzas.

**Figura 11.12.** Gráfico de dispersión (*salario actual*) por nivel (*nivel de estudios*)



**Transformados.** Una vez estimada la potencia apropiada para la homogeneización de las varianzas, puede utilizarse la opción **Transformados** para aplicar la transformación sugerida por el SPSS. Esta opción incluye, dentro de la lista desplegable **Potencia**, las siguientes transformaciones: logaritmo natural, recíproco de la raíz cuadrada, recíproco, raíz cuadrada, cuadrado y cubo. Todas estas transformaciones intentan homogeneizar las varianzas alterando (aumentando en unos casos y disminuyendo en otros) las varianzas de las distribuciones y corrigiendo el grado de asimetría.

La transformación *logarítmica* es apropiada para corregir distribuciones positivamente asimétricas. Y lo mismo vale decir de la transformación *raíz cuadrada* (si los valores de la variable que se desea transformar son pequeños, es conveniente sumar 0,5 a cada puntuación antes de obtener la raíz cuadrada). La transformación de los valores en sus *recíprocos* es adecuada cuando existen valores muy extremos por el lado positivo (cosa que ocurre, por ejemplo, con tiempos de reacción, donde los tiempos muy largos indican, probablemente, falta de atención más que otra cosa). Las transformaciones *cuadrado* y *cubo* permiten corregir, cada una en distinto grado, la asimetría negativa.

Al solicitar un gráfico de *dispersión por nivel* seleccionando algún tipo de transformación, tanto la prueba de Levene como el gráfico de dispersión se obtienen a partir de los datos transformados. Pero, excepto en el caso de la transformación logarítmica, al solicitar una transformación basada en alguna de las potencias disponibles, el gráfico de *dispersión por nivel* se obtiene a partir de la mediana y de la amplitud intercuartílica, no a

partir de sus logaritmos (estos logaritmos son los que se utilizan en las opciones **Estimación de potencia** y **No transformados**).

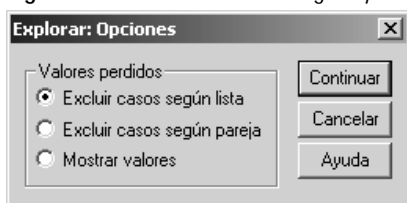
**No transformados.** Esta opción permite obtener la prueba de Levene y el gráfico de *dispersión por nivel* a partir de los datos originales, sin ningún tipo de transformación (lo cual es equivalente a utilizar una potencia de 1). El gráfico de *dispersión por nivel* se obtiene representando, para cada grupo, el logaritmo neperiano de la mediana en el eje horizontal y el logaritmo neperiano de la amplitud intercuartílica en el vertical (ver Figura 11.12).

## Opciones

Las opciones del procedimiento **Explorar** permiten decidir qué tipo de tratamiento se desea dar a los valores perdidos. Para ello:

- Pulsar el botón **Opciones...** del cuadro de diálogo principal (Figura 11.1) para acceder al subcuadro de diálogo *Explorar: Opciones* que muestra la Figura 11.13.

Figura 11.13. Subcuadro de diálogo *Explorar: Opciones*



**Valores perdidos.** Las opciones de este recuadro permiten elegir una de las siguientes tres formas de tratar los valores perdidos:

**Excluir casos según lista.** Se excluyen de todos los análisis solicitados los casos con algún valor perdido en cualquiera de las variables introducidas en la lista *dependientes* o en la lista *factores*. Es la opción por defecto.

**Excluir casos según pareja.** Se excluyen de cada análisis concreto los casos con algún valor perdido en las variables que intervienen en ese análisis (no se excluyen de un análisis concreto los casos que, aun teniendo algún valor perdido en alguna variable de las listadas, no lo tienen en las variables objeto de ese análisis).

**Mostrar valores.** Los casos con valores perdidos en la(s) variable(s) *factor* son tratados como una categoría más de esa(s) variable(s). Las tablas de frecuencias muestran los valores perdidos como una categoría más aunque esta opción no esté marcada.

## Análisis de variables categóricas

### El procedimiento *Tablas de contingencias*

En las ciencias sociales, del comportamiento y de la salud es muy frecuente encontrarse con variables categóricas. El sexo, la clase social, el lugar de procedencia, la categoría laboral, participar o no en un programa de intervención, el tipo de tratamiento aplicado, padecer o no una enfermedad o un determinado síntoma, los distintos departamentos de una empresa, etc., son algunos ejemplos de este tipo de variables. Son variables sobre las que únicamente es posible obtener una medida de tipo nominal (u ordinal, pero con pocos valores).

En este capítulo se expone el procedimiento *Tablas de contingencias*, el cual permite describir este tipo de variables y estudiar diferentes pautas de asociación entre ellas.

### Tablas de contingencias

Cuando se trabaja con variables categóricas, los datos suelen organizarse en tablas de doble (triple,...) entrada en las que cada entrada representa un criterio de clasificación (una variable categórica). Como resultado de esta clasificación, las frecuencias (el número o porcentaje de casos) aparecen organizadas en casillas que contienen información sobre la relación existente entre ambos criterios. A estas tablas de frecuencias se les llama *tablas de contingencias*.

La Tabla 12.1 muestra un ejemplo de tabla de contingencias con 474 sujetos clasificados utilizando dos criterios: *sexo* y *salario* (tabla *bidimensional*). Los valores de la tabla no son puntuaciones, sino frecuencias absolutas (número de casos): 19 hombres tienen salarios de menos de 25.000 \$; 86 mujeres tienen salarios comprendidos entre 25.000 y 50.000 \$; etc.

Tabla 12.1. Tabla de contingencias de *sexo* por *grupos de salario*

Recuento		Grupos de salario				Total
		Hasta 25.000 \$	Entre 25.001 y 50.000 \$	Entre 50.001 y 75.000 \$	Más de 75.000 \$	
Sexo	Hombre	19	174	48	17	258
	Mujer	124	86	6	0	216
Total		143	260	54	17	474

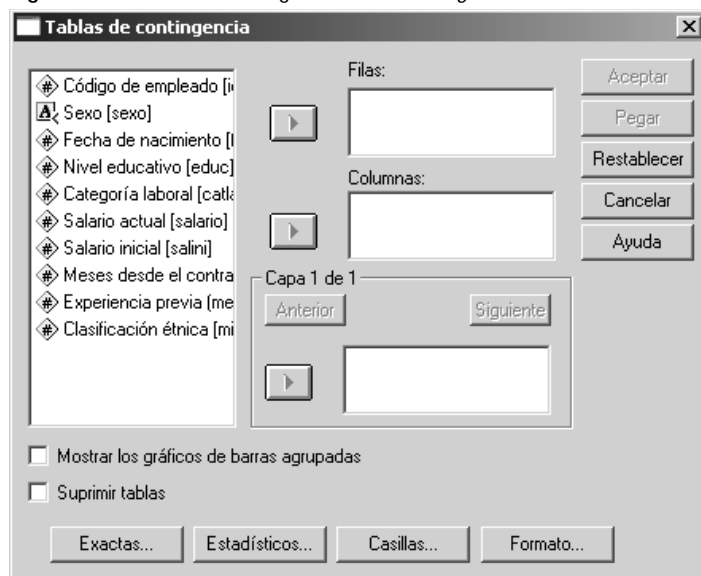
Por supuesto, en lugar de utilizar sólo dos criterios de clasificación para generar una tabla de contingencias *bidimensional*, también se podría utilizar tres o más criterios, lo que llevaría

a obtener tablas *tridimensionales*, *cuatridimensionales*, etc. El procedimiento **Tablas de contingencias** del SPSS permite generar tablas con cualquier número de dimensiones. No obstante, todos los estadísticos que incluye (con excepción de los de Mantel-Haenszel y Cochran) sólo sirven para analizar tablas *bidimensionales*. El análisis de tablas de contingencias con más de dos criterios de clasificación se aborda en otros procedimientos SPSS (por ejemplo, en el procedimiento **Modelos loglineales**).

Así pues, el procedimiento **Tablas de contingencias** permite obtener tablas de contingencias *bidimensionales*. Pero, además, incluye la posibilidad de añadir terceras variables (variables de segmentación) para definir subgrupos o capas y obtener así tablas de más de dos dimensiones. También incluye varios estadísticos que proporcionan la información necesaria para estudiar las posibles pautas de asociación existentes entre las variables que conforman una tabla de contingencias bidimensional. Para utilizar el procedimiento **Tablas de contingencias**:

- Seleccionar la opción **Estadísticos descriptivos > Tablas de contingencias...** del menú **Analizar** para acceder al cuadro de diálogo *Tablas de contingencias* que muestra la Figura 12.1.

Figura 12.1. Cuadro de diálogo *Tablas de contingencias*



La lista de variables del archivo de datos muestra todas las variables numéricas y de cadena corta del archivo de datos. Para obtener una tabla de contingencias:

- Seleccionar una variable categórica y trasladarla a la lista **Filas**; seleccionar otra variable categórica y trasladarla a la lista **Columnas**; pulsar el botón **Aceptar**.
- **Mostrar los gráficos de barras agrupadas.** Activando esta opción, el *Visor de resultados* muestra un gráfico de barras agrupadas con las categorías de la variable *fila* en el eje horizontal y las categorías de la variable *columna* anidadas dentro de las categorías de la variable *fila*. Cada barra del diagrama, por tanto, representa una casilla; y la altura de cada barra viene dada por la frecuencia de la correspondiente casilla.

- “ **Suprimir tablas.** Esta opción puede activarse si no se desea obtener tablas de contingencias. Esto tendría sentido si únicamente interesara obtener un gráfico de barras o alguno de los estadísticos o medidas de asociación disponibles en el procedimiento.

### Ejemplo: Tablas de contingencias

Este ejemplo muestra cómo obtener una tabla de contingencias y un diagrama de barras agrupadas mediante el procedimiento **Tablas de contingencias**, utilizando la variable *sexo* como variable *fila* y la variable *catlab* (categoría laboral) como variable *columna* (del archivo *Datos de empleados*, que se encuentra en la misma carpeta en la que está instalado el SPSS):

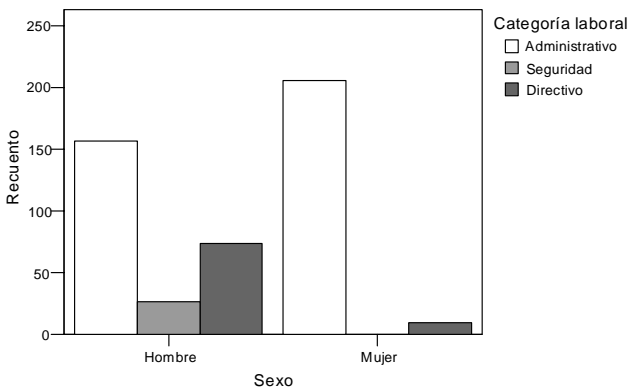
- En el cuadro de diálogo principal (ver Figura 12.1), seleccionar la variable *sexo* y trasladarla a la lista **Filas**; seleccionar la variable *catlab* y trasladarla a la lista **Columnas**.
- Marcar la opción **Mostrar los gráficos de barras agrupadas**.

Aceptando estas elecciones el *Visor* genera la Tabla 12.2 y la Figura 12.2. La Tabla 12.2 ofrece las frecuencias (número de casos) resultantes de cruzar cada categoría de la variable *sexo* con cada categoría de la variable *catlab*. También ofrece las frecuencias (*Total*) correspondientes a cada variable individualmente considerada (frecuencias *marginales*). La Figura 12.2 muestra el gráfico de barras agrupadas correspondiente a los datos de la Tabla 12.2. Cada barra corresponde a una casilla de la tabla; y la altura de las barras refleja el tamaño de las frecuencias de las casillas.

**Tabla 12.2.** Tabla de contingencias de *sexo* por *categoría laboral*

Recuento		Categoría laboral			Total
		Administrativo	Seguridad	Directivo	
Sexo	Hombre	157	27	74	258
	Mujer	206	0	10	216
Total		363	27	84	474

**Figura 12.2.** Gráfico de barras agrupadas de las variables *sexo* y *categoría laboral*





Puede trasladarse más de una variable tanto a la lista **Filas** como a la lista **Columnas**. En ese caso, cada variable *fila* se cruza con cada variable *columna* para formar una tabla distinta. Seleccionando, por ejemplo, 2 variables *fila* y 3 variables *columna*, se obtienen  $2 \times 3 = 6$  tablas de contingencias diferentes.

# Tablas segmentadas

El procedimiento **Tablas de contingencias** también permite cruzar variables categóricas teniendo en cuenta los niveles o categorías de una o más variables adicionales. Al cruzar, por ejemplo, las variables *sexo* y *categoría laboral*, existe la posibilidad de solicitar tablas separadas para cada nivel de, por ejemplo, *nivel de estudios*; en ese caso, la variable *nivel de estudios* estaría actuando como variable de *segmentación*. Para obtener una tabla de contingencias segmentada:

- Una vez seleccionadas las variables *fila* y *columna*, trasladar la variable de segmentación a la lista situada en el recuadro **Capa 1 de 1** (ver Figura 12.1).

Al seleccionar una variable de segmentación, el SPSS genera una tabla con tres dimensiones: la variable *fila*, la variable *columna* y la variable de *segmentación*.

Es posible trasladar más de una variable de segmentación a la lista del recuadro **Capa 1 de 1**. Si se traslada más de una variable, el SPSS genera una tabla de contingencias tridimensional separada para cada variable de segmentación seleccionada. Y si se seleccionan variables en distintas capas, la tabla de contingencias pasa a tener una nueva dimensión por cada capa adicional (ver siguiente ejemplo).

## Ejemplo: Tablas segmentadas

Este ejemplo muestra cómo incluir una variable de segmentación para obtener varias capas de una misma tabla de contingencias. Se sigue utilizando el archivo *Datos de empleados*. Manteniendo las variables *sexo* y *categoría laboral* como variables *fila* y *columna*, respectivamente:

- En el cuadro de diálogo principal (ver Figura 12.1), seleccionar la variable *minoría* (clasificación de minorías) y trasladarla a la lista del recuadro **Capa 1 de 1**.

Al incluir *minoría* como variable de segmentación el *Visor* ofrece los resultados que muestra la Tabla 12.3.

**Tabla 12.3.** Tabla de contingencias de *sexo* por *categoría laboral*, segmentada por *clasificación de minorías*

Recuento			Categoría laboral			Total
Clasificación de minorías			Administrativo	Seguridad	Directivo	
No	Sexo	Hombre	110	14	70	194
		Mujer	166	0	10	176
	Total		276	14	80	370
Sí	Sexo	Hombre	47	13	4	64
		Mujer	40	0	0	40
	Total		87	13	4	104

Utilizando los botones **Siguiente** y **Anterior** del recuadro **Capa # de #** (ver Figura 12.1) pueden obtenerse tablas de contingencias para los distintos niveles resultantes de combinar dos o más variables de segmentación. Para cruzar, por ejemplo, las variables *sexo* y *categoría laboral* y obtener una tabla separada para cada uno de los niveles resultantes de combinar las variables *minoría* (clasificación de minorías) y *estudios* (nivel de estudios):

- Seleccionar la variable *minoría* como variable de segmentación en la *primera capa* y pulsar el botón **Siguiente** para acceder a la *segunda capa*.
- Seleccionar la variable *estudios* como variable de segmentación en la *segunda capa*.
- Utilizar el botón **Anterior** para ver o cambiar la variable seleccionada en la capa previa.

Al hacer esto se obtiene una tabla de contingencias con cuatro dimensiones: *sexo*, *catlab*, *minoría* y *estudios*. Conforme se van creando capas, los valores # del recuadro **Capa # de #** van indicando el número de la capa actual y el número total de capas definidas.

## Estadísticos

El grado de relación existente entre dos variables categóricas no puede ser establecido simplemente observando las frecuencias de una tabla de contingencias. Incluso aunque la tabla recoja las frecuencias porcentuales en lugar de las absolutas, la simple observación de las frecuencias no puede conducir a una conclusión definitiva. Para determinar si dos variables se encuentran relacionadas debe utilizarse alguna medida de asociación, preferiblemente acompañada de su correspondiente prueba de significación. Para obtener medidas de asociación:

- Pulsar el botón **Estadísticos...** del cuadro de diálogo principal (ver Figura 12.1) para acceder al subcuadro de diálogo *Tablas de contingencias: Estadísticos* que muestra la Figura 12.3.

Figura 12.3. Subcuadro de diálogo *Tablas de contingencias: Estadísticos*

**Tablas de contingencia: Estadísticos**

☐ Chi-cuadrado

☐ Correlaciones

☐ Continuar

☐ Cancelar

☐ Ayuda

**Nominal**

☐ Coeficiente de contingencia

☐ Phi y V de Cramer

☐ Lambda

☐ Coeficiente de incertidumbre

**Ordinal**

☐ Gamma

☐ d de Somers

☐ Tau-b de Kendall

☐ Tau-c de Kendall

**Nominal por intervalo**

☐ Eta

☐ Kappa

☐ Riesgo

☐ McNemar

☐ Estadísticos de Cochran y de Mantel-Haenszel

Contrastar la razón de ventajas común igual a:

Este cuadro de diálogo contiene una amplia variedad de procedimientos estadísticos (medidas de asociación para variables nominales y ordinales, índices de acuerdo y de riesgo, etc.) diseñados para evaluar el grado de asociación existente entre dos variables categóricas en diferentes tipos de situaciones. Todos estos estadísticos se describen en los apartados que siguen.

## Chi-cuadrado

La opción **Chi-cuadrado** proporciona un estadístico (también conocido como  $X^2$  y *ji-cuadrado*) propuesto por Pearson (1911) que permite contrastar la hipótesis de independencia entre los dos criterios de clasificación utilizados (las dos variables categóricas). Para ello, compara las frecuencias *observadas* (las frecuencias de hecho obtenidas:  $n_{ij}$ ) con las frecuencias *esperadas* (las frecuencias que teóricamente debería haber en cada casilla si los dos criterios de clasificación fueran independientes:  $m_{ij}$ ). Cuando dos criterios de clasificación son independientes, las frecuencias esperadas se estiman de la siguiente manera:

$$\hat{m}_{ij} = \frac{(\text{total de la fila } i) \times (\text{total de la columna } j)}{\text{n}^\circ \text{ total de casos}} = \frac{n_{i+} n_{+j}}{n}$$

( $i+$  se refiere a una fila cualquiera;  $+j$  a una columna cualquiera;  $ij$  a una casilla cualquiera;  $i = 1, 2, \dots, I$ ;  $j = 1, 2, \dots, J$ ). Es decir, bajo la condición de independencia, la frecuencia esperada de una casilla concreta se obtiene dividiendo el producto de las frecuencias marginales correspondientes a esa casilla (su total de fila y su total de columna) por el número total de casos (este razonamiento se deriva de la teoría de la probabilidad; en concreto, del concepto de independencia entre sucesos: si dos sucesos son independientes, su probabilidad conjunta es igual al producto de sus probabilidades individuales).

Obtenidas las frecuencias esperadas para cada casilla, el estadístico  $X^2$  o *chi-cuadrado* de Pearson se obtiene de la siguiente manera:

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

( $n_{ij}$  se refiere a las frecuencias observadas y  $m_{ij}$  a las esperadas). De la ecuación se desprende que el estadístico  $X^2$  valdrá cero cuando las variables sean completamente independientes (pues las frecuencias observadas y las esperadas serán iguales), y que el valor del estadístico  $X^2$  será tanto mayor cuanto mayor sea la discrepancia existente entre las frecuencias observadas y las esperadas (discrepancia que será tanto mayor cuanto mayor sea la relación entre las variables).

El estadístico  $X^2$  se distribuye según el modelo de probabilidad  $\chi^2$  con los grados de libertad resultantes de multiplicar el número de filas menos uno por el número de columnas menos uno ( $gl = [I-1][J-1]$ ). Por tanto, puede utilizarse la distribución  $\chi^2$  para establecer el grado de compatibilidad existente entre el valor del estadístico  $X^2$  y la hipótesis de independencia. Si los datos son compatibles con la hipótesis de independencia, la probabilidad asociada al estadístico  $X^2$  será alta (mayor de 0,05). Si esa probabilidad es muy pequeña (menor que 0,05), se considerará que los datos son incompatibles con la hipótesis de independencia y se podrá concluir que las variables estudiadas están relacionadas.

Para que las probabilidades de la distribución  $\chi^2$  constituyan una buena aproximación a la distribución del estadístico  $X^2$  conviene que se cumplan algunas condiciones; entre ellas, que las frecuencias esperadas no sean demasiado pequeñas. Suele asumirse (siguiendo a Cochran, 1952) que, si existen frecuencias esperadas menores que 5, éstas no deben superar el 20 % del total de frecuencias de la tabla. El SPSS muestra en una nota a pie de tabla un mensaje indicando el valor de la frecuencia esperada más pequeña; si existe alguna casilla con frecuencia esperada menor que 5, la nota a pie de tabla también informa acerca del porcentaje que éstas representan sobre el total de casillas de la tabla. En el caso de que ese porcentaje supere el 20 %, el estadístico de Pearson debe interpretarse con cautela.

### Ejemplo: Tablas de contingencias > Estadísticos > Chi-cuadrado

Este ejemplo muestra cómo obtener e interpretar el estadístico *chi*-cuadrado de Pearson en una tabla de contingencias bidimensional (las variables se han tomado del archivo *Datos de empleados*, que se encuentra en la misma carpeta en la que está instalado el SPSS):

- En el cuadro de diálogo principal (ver Figura 12.1), seleccionar las variables *sexo* y *categoría laboral* como variables *fila* y *columna*, respectivamente.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Tablas de contingencias: Estadísticos* (ver Figura 12.3) y marcar la opción **Chi-cuadrado**.

Aceptando estas elecciones, el *Visor de resultados* ofrece los estadísticos que muestra la Tabla 12.4. El estadístico *chi*-cuadrado de Pearson toma un valor de 79,277, el cual, en la distribución  $\chi^2$  con 2 grados de libertad (*gl*), tiene asociada una probabilidad (*Sig. asintótica*) menor que 0,0005. Puesto que esta probabilidad (denominada *nivel crítico*, o *nivel de significación observado*, o *valor p*) es muy pequeña, se puede rechazar la hipótesis de independencia y concluir que las variables *sexo* y *catlab* están relacionadas.

Tabla 12.4. Estadísticos

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	79,277 <sup>a</sup>	2	,000
Razón de verosimilitud	95,463	2	,000
N de casos válidos	474		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5.  
La frecuencia mínima esperada es 12,30.

Además del estadístico *chi*-cuadrado, la tabla muestra otro estadístico denominado *razón de verosimilitudes*  $G^2$  (Fisher, 1924; Neyman y Pearson, 1928), que se obtiene mediante:

$$G^2 = 2 \sum_i \sum_j n_{ij} \log \left( \frac{n_{ij}}{m_{ij}} \right)$$

Se trata de un estadístico asintóticamente equivalente a  $X^2$  (se distribuye e interpreta igual que  $X^2$ ) y es muy utilizado para estudiar la relación entre variables categóricas, particularmente en el contexto de los modelos loglineales.

Cuando la tabla de contingencias se construye con dos variables dicotómicas (tablas 2×2), los resultados incluyen información adicional. Utilizando, por ejemplo, la variable *sexo* como variable *fila* y la variable *minoría* (clasificación de minorías) como variable *columna* se obtiene la tabla 2×2 y los estadísticos que muestran las Tablas 12.5 y 12.6. Puede verse en ellas que siguen estando presentes tanto el estadístico de Pearson como la razón de verosimilitudes. Pero ahora hay dos líneas nuevas: la *corrección por continuidad* y el *estadístico exacto de Fisher*. También aparece, en una nota a pie de tabla, el valor de la frecuencia esperada más pequeña, que en este ejemplo es 47,39.

La *corrección por continuidad* de Yates (1934) consiste en restar 0,5 puntos al valor absoluto de las diferencias  $n_{ij} - m_{ij}$  del numerador del estadístico  $X^2$  (antes de elevarlas al cuadrado). Aunque algunos autores sugieren que, con muestras pequeñas, esta corrección hace que la función de probabilidad de  $X^2$  se parezca más a las probabilidades de la distribución *chi-cuadrado*, lo cierto es que no existe un consenso generalizado sobre su utilización. Más bien parece que su uso debería limitarse al caso en el que los totales marginales de ambas variables son fijos (ver Haviland, 1990).

El *estadístico exacto de Fisher* (1935) ofrece, basándose en la distribución hipergeométrica y en la hipótesis nula de independencia, la probabilidad exacta de obtener las frecuencias de hecho obtenidas o cualquier otra combinación de frecuencias más alejada de la hipótesis de independencia. Es una solución apropiada para tablas 2×2 cuando los totales de ambas variables son fijos.

**Tabla 12.5.** Tabla de contingencias de *sexo* por *clasificación de minorías*

Recuento		Clasificación de minorías		Total
		No	Sí	
Sexo	Hombre	194	64	258
	Mujer	176	40	216
Total		370	104	474

**Tabla 12.6.** Estadísticos

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	2,714 <sup>b</sup>	1	,099		
Corrección por continuidad <sup>a</sup>	2,359	1	,125		
Razón de verosimilitud	2,738	1	,098		
Estadístico exacto de Fisher				,119	,062
N de casos válidos	474				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 47,39.

## Correlaciones

La opción **Correlaciones** (ver Figura 12.3) permite obtener dos coeficientes de correlación: el de *Pearson* y el de *Spearman*. El coeficiente de correlación de Pearson es una medida de asociación *lineal* especialmente apropiada para estudiar la relación entre variables de intervalo

o razón; por tanto posee escasa utilidad con tablas de contingencias (aunque aplicado a una variable de intervalo o razón y a una variable dicotómica permite obtener el coeficiente de correlación *biserial-puntual*). El coeficiente de correlación de Spearman es también una medida de asociación *lineal*, pero para variables ordinales. Ambos coeficientes se explican con detalle en el Capítulo 17 sobre *Análisis de correlación lineal*.

## Datos nominales

Aunque, según se ha señalado, el estadístico *chi*-cuadrado de Pearson permite contrastar la hipótesis de independencia en una tabla de contingencias, nada dice sobre la *fuerza de la asociación* existente entre las variables estudiadas. Esto es debido a que su valor depende no sólo del grado en que los datos se ajustan al modelo de independencia, sino también del número de casos de que consta la muestra. Con tamaños muestrales muy grandes, diferencias relativamente pequeñas entre las frecuencias observadas y las esperadas pueden dar lugar a valores *chi*-cuadrado demasiado altos. Por esta razón, para estudiar el grado de relación existente entre dos variables se utilizan medidas de asociación que intentan cuantificar ese grado de relación eliminando el efecto del tamaño muestral.

Existen diversas medidas de asociación que no sólo difieren en la forma de definir lo que es asociación perfecta e intermedia, sino en la forma en que cada una se ve afectada por factores tales como las distribuciones marginales. De hecho, una medida puede arrojar un valor bajo en una situación concreta, no porque las variables estudiadas no estén relacionadas, sino porque esa medida no sea sensible al tipo de relación presente en los datos. Para seleccionar una medida concreta, además de las características particulares de cada medida, hay que tener en cuenta cosas tales como el tipo de variables estudiadas y la hipótesis que interesa contrastar. En ningún caso está justificado obtener todas las medidas disponibles para seleccionar aquella cuyo valor se ajusta mejor a los intereses del usuario.

Conviene señalar que las medidas *nominales* sólo aprovechan información nominal. Por tanto, únicamente informan del grado de asociación existente, no de la dirección o naturaleza de tal asociación. Con variables nominales no tiene sentido hablar de relación positiva o negativa.

**Medidas basadas en *chi*-cuadrado.** Son medidas que intentan corregir el valor  $X^2$  haciéndole tomar un valor entre 0 y 1; de este modo se elimina el efecto del tamaño de la muestra sobre la cuantificación del grado de asociación (Pearson, 1913; Cramer, 1946).

- " Coeficiente de contingencia:  $C = \sqrt{X^2/(X^2 + n)}$ . Toma valores entre 0 y 1, pero difícilmente llega a 1. Su valor máximo depende del número de filas y de columnas. Si el número de filas y de columnas es el mismo ( $k$ ), entonces el valor máximo de  $C$  se obtiene de la siguiente manera:  $C_{\max} = \sqrt{(k-1)/k}$ . Un coeficiente de 0 indica independencia, mientras que un coeficiente que alcanza su valor máximo indica asociación perfecta.
- " Phi y V de Cramer. El coeficiente *phi* se obtiene de la siguiente manera:  $\phi = \sqrt{X^2/n}$ . En una tabla de contingencias  $2 \times 2$ , *phi* adopta valores entre 0 y 1, y su valor es idéntico al del coeficiente de correlación de Pearson (ver Capítulo 17). En tablas en las que una de las variables tiene más de dos niveles, *phi* puede tomar valores mayores que 1 (pues el valor de  $X^2$  puede ser mayor que el tamaño muestral).

La  $V$  de Cramer incluye una ligera modificación de  $\phi$ :  $V_{\text{Cramer}} = \sqrt{X^2/[n(k-1)]}$ , donde  $k$  se refiere al valor menor del número de filas y de columnas. A diferencia de  $\phi$ , la  $V_{\text{Cramer}}$  nunca excede de 1. En una tabla de contingencias  $2 \times 2$  los valores  $V_{\text{Cramer}}$  y  $\phi$  son idénticos.

**Medidas basadas en la reducción proporcional del error (RPE).** Son medidas de asociación que expresan la proporción en que se consigue reducir la probabilidad de cometer un error de predicción cuando, al intentar clasificar un caso o grupo de casos como pertenecientes a una u otra categoría de una variable, en lugar de utilizar únicamente las probabilidades asociadas a cada categoría de esa variable, se efectúa la clasificación teniendo en cuenta las probabilidades de las categorías de esa variable en cada categoría de una segunda variable.

“ **Lambda.** Esta opción permite obtener dos medidas de asociación desarrolladas por Goodman y Kruskal: *lambda* y *tau* (ver Goodman y Kruskal, 1979).

La medida de asociación **lambda** se basa en la siguiente lógica: si al intentar predecir a qué categoría de una determinada variable ( $X$ ) pertenece un caso concreto se decide que pertenece a la categoría más probable de todas, se estará cometiendo un error de predicción igual a la probabilidad de pertenecer a una cualquiera de las restantes categorías; si, en lugar de esto, se clasifica a ese caso en una u otra categoría de la variable  $X$  dependiendo de cuál sea la categoría de una segunda variable ( $Y$ ) a la que pertenece, se puede estar consiguiendo una reducción en el error de predicción (lo cual sólo ocurrirá si las dos variables están relacionadas). El coeficiente *lambda* expresa el grado en que se consigue reducir la proporción de error de clasificación al utilizar la segunda estrategia (información proporcionada por una segunda variable) en lugar de la primera (categoría más probable sin tener en cuenta otra información).

La Tabla 12.7 (ver siguiente ejemplo) recoge las frecuencias resultantes de cruzar las variables *sexo* y *grupos de salario*. Si se conoce la distribución de la variable *grupos de salario*, al estimar a qué grupo de salario pertenece un sujeto cualquiera, se decidirá que pertenece al grupo de «entre 25.000 y 50.000 \$» porque la probabilidad de pertenecer a ese grupo, que vale  $260/474 = 0,5485$ , es más alta que la probabilidad de pertenecer a cualquiera de los grupos restantes. Procediendo de esta manera, se estará cometiendo un error de clasificación de  $1 - 0,5485 = 0,4515$ .

Si ahora se tiene en cuenta la variable *sexo* para efectuar esa estimación y se clasifica a los varones en el grupo de «entre 25.000 y 50.000 \$» porque ése es el grupo de salario más probable entre los varones (con un error de  $(19 + 48 + 17)/474 = 0,1772$ ), y a las mujeres en el grupo de «menos de 25.000 \$» porque ése es el grupo salarial más probable entre las mujeres (con un error de  $(86 + 6 + 0)/474 = 0,1941$ ), se estará cometiendo un error de clasificación de  $0,1772 + 0,1941 = 0,3713$ . Utilizando esta segunda estrategia se ha conseguido reducir el error de clasificación en  $0,0802$  (de  $0,4515$  a  $0,3713$ ), lo cual representa una proporción de reducción de  $0,0802/0,4515 = 0,1776$ , que es justamente el valor que toma *lambda* en los estadísticos de la Tabla 12.9 cuando se considera la variable *grupos de salario* como variable dependiente.

*Lambda* tiene tres versiones: dos *asimétricas* (para cuando una de las dos variables se considera independiente y la otra dependiente) y una *simétrica* (para cuando no existe razón para distinguir entre variable independiente y dependiente). La salida del SPSS incluye las tres versiones.

*Lambda* toma valores entre 0 y 1. Un valor de 0 indica que la variable independiente (la variable utilizada para efectuar pronósticos) no contribuye en absoluto a reducir el error de clasificación. Un valor de 1 indica que el error de clasificación se ha conseguido reducir por completo, es decir, que la variable independiente permite predecir con toda precisión a qué categoría de la variable dependiente pertenecen los casos clasificados.

Cuando dos variables son estadísticamente independientes, *lambda* vale 0. Pero un valor de 0 no implica independencia estadística, pues *lambda* únicamente es sensible a un tipo particular de asociación: a la derivada de la reducción en el error de clasificación que se consigue cuando para predecir a qué categoría de una variable pertenece un determinado caso, se utiliza la información de una segunda variable. Recuérdese que no existe ningún índice de asociación sensible a todos los posibles tipos de asociación.

La medida de asociación *tau* se parece a *lambda*, pero su lógica es algo diferente. Al pronosticar a qué categoría de la variable *grupos de salario* pertenece un grupo de sujetos, se puede optar por asignar aleatoriamente el 100(143/474)=30,17 % a la categoría «menos de 25.000 \$», el 100(260/474)=54,85 % a la categoría «entre 25.000 y 50.000 \$», etc., tomando como referencia la probabilidad de pertenecer a cada categoría, en lugar de considerar sólo la categoría más probable, como se hace con *lambda*. Con esta estrategia se estará clasificando correctamente al 30,17 % de los 143 sujetos del grupo «menos de 25.000 \$», al 54,85 % de los 260 sujetos del grupo «entre 25.000 y 50.000 \$», etc. Lo cual representa una proporción de clasificación correcta global de 0,4061 y, por tanto, una proporción de clasificación errónea de  $1-0,4061=0,5939$ .

En lugar de esto, se puede optar por aprovechar la información de la variable *sexo* y, entre los varones, asignar aleatoriamente el 100(19/258)=7,36 % de los casos a la categoría «menos de 25.000 \$», el 100(174/258)=67,44 % a la categoría «entre 25.000 y 50.000 \$», etc.; y entre las mujeres, asignar aleatoriamente el 100(124/216)=57,41 % de los casos a la categoría «menos de 25.000 \$», el 100(86/216)=39,81 % a la categoría «entre 25.000 y 50.000 \$»; etc. Con esta estrategia se estará clasificando de forma correcta al 49,45 % de los sujetos y, por tanto, se estarán efectuando pronósticos de clasificación erróneos con una probabilidad de  $1-0,4945=0,5055$ .

Utilizando esta segunda estrategia se reduce la probabilidad de efectuar pronósticos erróneos en  $0,5939-0,5055=0,0884$ . Por lo que se habrá conseguido reducir la probabilidad de clasificación errónea en una proporción de  $0,0884/0,5939=0,149$ , que es justamente el valor que toma la *tau* de Goodman y Kruskal en los estadísticos de la Tabla 12.9 cuando se considera la variable *grupos de salario* como dependiente.

Al igual que *lambda*, *tau* también toma valores entre 0 y 1, significando el 0 ausencia de reducción del error de clasificación y el 1 reducción completa.

“ Coeficiente de incertidumbre (Theil, 1970). Al igual que *lambda* y *tau*, el *coeficiente de incertidumbre* es una medida de asociación basada en la reducción proporcional del error. Por tanto, es una medida que expresa el grado de incertidumbre que se consigue reducir cuando se utiliza una variable para efectuar pronósticos sobre otra.

Posee dos versiones *asimétricas* (dependiendo de cuál de las dos variables se considere dependiente) y una *simétrica* (para cuando no se hace distinción entre variable independiente y dependiente). Se obtiene de la siguiente manera:

$$I_{Y|X} = [I(X) + I(Y) - I(XY)]/I(Y)$$



donde:  $I(X) = -\sum_i [(n_{i+}/n) \ln(n_{i+}/n)]$  ( $n_{i+}$  = frec. marginales de las filas)

$I(Y) = -\sum_j [(n_{+j}/n) \ln(n_{+j}/n)]$  ( $n_{+j}$  = frec. marginales de las columnas)

$I(XY) = -\sum_i \sum_j [(n_{ij}/n) \ln(n_{ij}/n)]$  ( $n_{ij}$  = frec. de las casillas [con  $n_{ij} > 0$ ])

Para obtener  $I_{X/Y}$  basta con intercambiar los papeles de  $I(X)$  e  $I(Y)$ . Y la versión *simétrica* se obtiene multiplicando  $I_{Y/X}$  por 2 después de añadirle  $I(X)$  al denominador.

### Ejemplo: Tablas de contingencias > Estadísticos > Datos nominales

Este ejemplo muestra cómo obtener e interpretar los estadísticos para datos nominales del procedimiento *Tablas de contingencias* (las variables se han tomado del archivo *Datos de empleados ampliado*, el cual puede obtenerse en la página *web* del manual).

- En el cuadro de diálogo principal (ver Figura 12.1), seleccionar las variables *sexo* y *salargr* (grupos de salario) como variables *fila* y *columna*, respectivamente.
- Pulsar el botón *Estadísticos...* para acceder al subcuadro de diálogo *Tablas de contingencias: Estadísticos* (ver Figura 12.3) y marcar las cuatro opciones del recuadro *Nominal*.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 12.7 a la 12.9. La primera tabla (12.7) recoge las frecuencias resultantes de cruzar las variables *sexo* y *grupos de salario*.

Tabla 12.7. Tabla de contingencias de *sexo* por *grupos de salario*

Recuento		Grupos de salario				Total
		Hasta 25.000 \$	Entre 25.001 y 50.000 \$	Entre 50.001 y 75.000 \$	Más de 75.000 \$	
Sexo	Hombre	19	174	48	17	258
	Mujer	124	86	6	0	216
Total		143	260	54	17	474

Las Tablas 12.8 y 12.9 muestran las medidas de asociación para datos nominales recién estudiadas. Cada medida aparece acompañada de su correspondiente nivel crítico (*Sig. aproximada*), el cual permite decidir sobre la hipótesis de independencia: puesto que el nivel crítico de todas las medidas listadas es muy pequeño (menor que 0,05 en todos los casos), se puede rechazar la hipótesis nula de independencia y concluir que las variables *sexo* y *grupos de salario* están relacionadas.

En la Tabla 12.9, junto con el valor de cada medida de asociación aparece una tipificación o estandarización del mismo (*T aproximada*) que se obtiene dividiendo el valor de la medida entre su error típico (calculado éste suponiendo independencia entre las variables). También aparece el error típico de cada medida calculado sin suponer independencia (*Error típico asintótico*).

Además de las medidas de asociación, las tablas recogen *notas* aclaratorias acerca de aspectos tales como bajo qué condiciones se hacen algunos cálculos, cómo se obtienen algunos de los niveles críticos que se ofrecen, cuál es el motivo de que no se puedan efectuar algunos cálculos, etc.

Tabla 12.8. Medidas de asociación *simétricas*

	Valor	Sig. aproximada
Nominal por nominal Phi	,570	,000
V de Cramer	,570	,000
Coefficiente de contingencia	,495	,000
N de casos válidos	474	

Tabla 12.9. Medidas de asociación *direccionales*

		Valor	Error típ. asint. <sup>a</sup>	T aproximada <sup>b</sup>	Sig. aproximada
Lambda	Simétrica	,333	,048	6,054	,000
	Sexo dependiente	,486	,040	9,596	,000
	Grupos de salario dependiente	,178	,061	2,641	,008
Tau de Goodman y Kruskal	Sexo dependiente	,325	,036		,000 <sup>c</sup>
	Grupos de salario dependiente	,149	,024		,000 <sup>c</sup>
Coeficiente de incertidumbre	Simétrica	,210	,026	7,948	,000 <sup>d</sup>
	Sexo dependiente	,266	,033	7,948	,000 <sup>d</sup>
	Grupos de salario dependiente	,173	,021	7,948	,000 <sup>d</sup>

a. Asumiendo la hipótesis alternativa.

b. Empleando el error típico asintótico basado en la hipótesis nula.

c. Basado en la aproximación chi-cuadrado.

d. Probabilidad del chi-cuadrado de la razón de verosimilitud.

## Datos ordinales

El recuadro **Datos ordinales** (ver Tabla 12.3) recoge una serie de medidas de asociación que permiten aprovechar la información ordinal que las medidas diseñadas para datos nominales pasan por alto. Con datos ordinales ya tiene sentido hablar de relación *lineal*: una relación *positiva* indica que los valores altos de una variable tienden a asociarse con valores altos de la otra, y los valores bajos, con valores bajos; una relación *negativa* indica que los valores altos de una variable tienden a asociarse con valores bajos de la otra, y los valores bajos con valores altos.

Muchas de las medidas de asociación diseñadas para estudiar la relación entre variables ordinales se basan en el concepto de *inversión* y *no inversión*. Si los dos valores de un caso en ambas variables son mayores (o menores) que los dos valores de otro caso, se dice que entre esos casos se da una *no inversión* (*P*). Si el valor de un caso en una de las variables es mayor que el de otro caso, y en la otra variable el valor del segundo caso es mayor que el del primero, se dice que se da una *inversión* (*Q*). Si dos casos tienen valores idénticos en una o en las dos variables, se dice que se da un *empate* (*E*). Cuando predominan las *no inversiones*, la

relación es positiva: conforme aumentan (o disminuyen) los valores de una de las variables, aumentan (o disminuyen) los de la otra. Cuando predominan las *inversiones*, la relación es negativa: conforme aumentan (o disminuyen) los valores de una de las variables, disminuyen (o aumentan) los de la otra. Todas las medidas de asociación recogidas en este apartado utilizan en el numerador la diferencia entre el número de *inversiones* y *no inversiones* resultantes de comparar cada caso con cada otro, pero se diferencian en el tratamiento dado a los *empates* (ver Somers, 1962; Kendall, 1963; Goodman y Kruskal, 1979).

- " **Gamma:**  $\gamma = (n_p - n_Q) / (n_p + n_Q)$ . Si la relación lineal entre dos variables es perfecta y positiva, todos los pares (todas las comparaciones dos a dos entre casos) serán *no inversiones* ( $n_p$ ) y, consiguientemente,  $n_Q$  valdrá cero, en cuyo caso,  $\gamma = 1$ . Si la relación lineal entre las variables es perfecta, pero negativa, todos los pares serán *inversiones* ( $n_Q$ ) y, en consecuencia,  $n_p$  valdrá cero, de donde  $\gamma = -1$ . Si las variables son linealmente independientes, habrá tantas *inversiones* como *no inversiones* ( $n_p = n_Q$ ); de modo que  $n_p - n_Q = 0$  y  $\gamma = 0$ . Así pues,  $\gamma$  oscila entre  $-1$  y  $1$ . Si dos variables son estadísticamente independientes,  $\gamma$  vale cero; pero una  $\gamma$  de cero no implica independencia (excepto en tablas de contingencias  $2 \times 2$ ).
- " **d de Somers:** cuando una de las variables se considera independiente ( $X$ ) y la otra dependiente ( $Y$ ), Somers ha propuesto una modificación del coeficiente  $\gamma$  que consiste en añadir en el denominador de *gamma* el número de pares empatados en la variable dependiente:  $d = (n_p - n_Q) / (n_p + n_Q + n_{E(Y)})$ . El SPSS ofrece tres versiones: dos asimétricas y una simétrica. La versión *simétrica* se obtiene utilizando en el denominador del coeficiente  $d$  el promedio de los denominadores correspondientes a las dos versiones asimétricas.
- " **Tau-b de Kendall:** tanto el coeficiente *tau-b* como el *tau-c* tienen en cuenta el número de empates, pero de distinta manera:  $\tau_b = (n_p - n_Q) / \sqrt{(n_p + n_Q + n_{E(X)})(n_p + n_Q + n_{E(Y)})}$ . El coeficiente *tau-b* toma valores entre  $-1$  y  $+1$  sólo en tablas de contingencias cuadradas y si ninguna frecuencia marginal vale cero.
- " **Tau-c de Kendall:**  $\tau_c = 2m(n_p - n_Q) / [n^2(m - 1)]$ , donde  $m$  se refiere al valor menor del número de filas y del número de columnas. *Tau-c* toma valores entre aproximadamente  $-1$  y  $+1$  sea cual sea el número de filas y de columnas de la tabla.

### **Ejemplo: Tablas de contingencias > Estadísticos > Datos ordinales**

Este ejemplo muestra cómo obtener e interpretar los estadísticos para datos ordinales del procedimiento *Tablas de contingencias* (las variables se han tomado del archivo *Datos de empleados ampliado*, el cual puede obtenerse en la página *web* del manual).

- ' En el cuadro de diálogo principal (ver Figura 12.1), seleccionar las variables *salargr* (grupos de salario) y *estudios* (nivel de estudios) como variables *fila* y *columna*, respectivamente.
- ' Pulsar el botón *Estadísticos...* para acceder al subcuadro de diálogo *Tablas de contingencias: Estadísticos* (ver Figura 12.3) y marcar las cuatro opciones del recuadro *Ordinal: gamma, d de Somers, tau-b y tau-c*.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 12.10 a la 12.12. La Tabla 12.10 recoge las frecuencias absolutas resultantes de cruzar *salargr* y *estudios*. La Tabla 12.11 ofrece el coeficiente *d* de *Somers* en sus tres versiones: (1) sin hacer distinción entre variable independiente y dependiente; (2) tomando la variable *grupos de salario* como variable dependiente; y (3) tomando la variable *nivel de estudios* como variable dependiente. La Tabla 12.12 contiene las medidas de asociación simétricas, es decir, los coeficientes *tau-b*, *tau-c* y *gamma*.

Cada coeficiente de correlación aparece con su correspondiente nivel crítico (*Sig. aproximada*), el cual permite decidir sobre la hipótesis nula de independencia. Puesto que todos estos niveles críticos son menores que 0,05, se puede rechazar la hipótesis de independencia y afirmar que las variables *salargr* y *estudios* están relacionadas. Y como el valor de las medidas es positivo (relación positiva), puede concluirse que a mayor nivel de estudios corresponde mayor salario.

Al igual que ocurría con las medidas de asociación para datos nominales, junto con el valor de cada coeficiente de correlación aparece también su valor estandarizado (*T aproximada*), es decir, el valor del coeficiente dividido por su error típico. La tabla también ofrece una estimación del error típico de cada coeficiente obtenida sin suponer independencia (*Error típico asintótico*).

**Tabla 12.10.** Tabla de contingencias de *salargr* (grupos de salario) por *estudios* (nivel de estudios)

Recuento		Nivel de estudios				Total
		Primarios	Secundarios	Medios	Superiores	
Grupos de salario	Hasta 25.000 \$	29	95	19	0	143
	Entre 25.001 y 50.000 \$	24	94	136	6	260
	Entre 50.001 y 75.000 \$	0	1	21	32	54
	Más de 75.000 \$	0	0	5	12	17
Total		53	190	181	50	474

**Tabla 12.11.** Medidas de asociación *direccionales* (*d* de *Somers*)

Ordinal por ordinal		Valor	Error típ. <sup>a</sup> asint.	T aproximada <sup>b</sup>	Sig. aproximada
d de Somer	Simétrica	,557	,029	16,075	,000
	Grupos de salario dependiente	,525	,030	16,075	,000
	Nivel de estudios dependiente	,593	,030	16,075	,000

a. Asumiendo la hipótesis alternativa.

b. Empleando el error típico asintótico basado en la hipótesis nula.

**Tabla 12.12.** Medidas de asociación *simétricas* (*tau-b*, *tau-c* y *gamma*)

Ordinal por ordinal	Valor	Error típ. <sup>a</sup> asint.	T aproximada <sup>b</sup>	Sig. aproximada
Tau-b de Kendall	,558	,029	16,075	,000
Tau-c de Kendall	,469	,029	16,075	,000
Gamma	,798	,031	16,075	,000
N de casos válidos	474			

a. Asumiendo la hipótesis alternativa.

b. Empleando el error típico asintótico basado en la hipótesis nula.

# Nominal por intervalo

El coeficiente de correlación *eta* sirve para cuantificar el grado de asociación existente entre una variable cuantitativa (medida en escala de intervalo o razón) y una variable categórica (medida en escala nominal u ordinal). Su mayor utilidad no está precisamente asociada a las tablas de contingencias, pues éstas se construyen, según se ha señalado ya, utilizando variables categóricas (nominales u ordinales). Pero, puesto que este coeficiente se encuentra disponible en el procedimiento SPSS *Tablas de contingencias*, puede marcarse la opción *Eta* (ver Figura 12.2) para obtener el valor de la relación entre dos variables cuando una de ellas es cuantitativa y la otra categórica. Se trata de un coeficiente de correlación que no supone *linealidad* y cuyo cuadrado puede interpretarse, si el diseño lo permite, como la proporción de varianza de la variable cuantitativa que está explicada por (que depende de) la variable categórica.

# Índice de acuerdo (kappa)

La opción **Kappa** (ver Figura 12.2) ofrece una medida del grado de acuerdo existente entre dos observadores o jueces al evaluar una serie de sujetos u objetos (Cohen, 1960). La Tabla 12.13 muestra el resultado obtenido por dos jueces al clasificar una muestra de 200 pacientes neuróticos según el tipo de neurosis padecida.

**Tabla 12.13.** Resultado obtenido por dos *jueces* al diagnosticar una muestra de 200 pacientes neuróticos

Recuento		segundo diagnóstico				Total
		fóbica	histérica	obsesiva	depresiva	
primer diagnóstico	fóbica	20	8	6	1	35
	histérica	7	36	14	4	61
	obsesiva	1	8	43	7	59
	depresiva	2	6	4	33	45
Total		30	58	67	45	200

Una forma intuitiva de medir el grado de acuerdo entre los dos jueces consiste en hacer un recuento del número de coincidencias existentes entre ambos (es decir, del número de casos que ambos jueces han clasificado de la misma manera). Sumando las frecuencias que indican acuerdo, es decir, las que se encuentran en la diagonal que va desde la parte superior izquierda de la tabla a la parte inferior derecha, se obtienen 132 coincidencias, lo que representa un porcentaje de acuerdo del  $100(132/200) = 66\%$ .

El problema de utilizar este porcentaje como índice de acuerdo es que no tiene en cuenta la probabilidad de obtener acuerdos por azar. Si se supone que ambos jueces son independientes, los casos que cabría esperar por azar en las casillas de la diagonal pueden obtenerse multiplicando las correspondientes frecuencias marginales y dividiendo por el total de casos. Así, en la primera casilla de la diagonal cabría esperar, por azar,  $35(30)/200 = 5,25$  casos; en la segunda casilla,  $61(58)/200 = 17,69$  casos; etc. Repitiendo la operación para todas las casillas de la diagonal se obtiene un total de 52,83 casos, lo que representa un 26,42% de acuerdo esperado por azar. La diferencia entre la proporción de *acuerdo observado* (0,66) y la proporción de *acuerdo esperado por azar* (0,2642) es 0,3958.

La *kappa* de Cohen se obtiene dividiendo esa diferencia por la proporción de acuerdo máximo que los dos jueces podrían alcanzar. Esta proporción máxima se obtiene restando a 1 la proporción de acuerdo esperado por azar:  $1 - 0,2642 = 0,7358$ . Dividiendo el acuerdo observado (0,3958) entre el acuerdo máximo posible (0,7358), se obtiene una proporción de acuerdo de 0,5379 que es justamente el valor de *kappa* si se aplica la ecuación:

$$\kappa = \frac{n \sum_i n_{ii} - \sum_i n_{i+} n_{+i}}{n^2 - \sum_i n_{i+} n_{+i}}$$

( $n_{ii}$  se refiere a las frecuencias de la diagonal principal:  $i = j$ ). El valor de *kappa* debe interpretarse teniendo en cuenta que toma valores entre 0 (acuerdo nulo) y 1 (acuerdo máximo). Si el acuerdo alcanzado es menor que el esperado por azar, *kappa* toma un valor negativo.

Fleiss, Cohen y Everitt (1969) han demostrado que el error típico del coeficiente *kappa* puede estimarse mediante:

$$\sigma_{\kappa}^2 = \frac{1}{n(n^2 - \sum_i n_{i+} n_{+i})^2} \left[ n^2 \sum_i n_{i+} n_{+i} + (\sum_i n_{i+} n_{+i})^2 - n \sum_i n_{i+} n_{+i} (n_{i+} + n_{+i}) \right]$$

La hipótesis de que los dos jueces son independientes (o, lo que es lo mismo, que el coeficiente *kappa* vale cero) puede contrastarse tipificando el valor de *kappa*. Dividiendo *kappa* por su error típico se obtiene un valor tipificado  $Z_{\kappa}$  que se distribuye de forma aproximadamente normal, con media 0 y desviación típica 1 (esta tipificación aparece en los resultados del *Visor* del SPSS con el nombre de *T aproximada*):

$$Z_{\kappa} = \kappa / \sigma_{\kappa} \rightarrow N(0, 1)$$

Al margen de la significación estadística del coeficiente *kappa*, Landis y Koch (1977) han argumentado que, en la mayor parte de los contextos, valores por encima de 0,75 suelen reflejar un acuerdo excelente; valores entre 0,40 y 0,75, un buen acuerdo; y valores por debajo de 0,40, un acuerdo más bien pobre.

## Ejemplo: Tablas de contingencias > Estadísticos > Kappa

Este ejemplo muestra cómo obtener e interpretar el índice de acuerdo *kappa* del procedimiento Tablas de contingencias.

- Reproducir en el *Editor de datos* los datos de la Tabla 12.13 tal como muestra la Figura 12.4 y ponderar el archivo con la variable *ncasos* (consultar, en el Capítulo 6, el apartado *Ponderar casos*). Alternativamente, abrir el archivo *Acuerdo kappa* que se encuentra en la página *web* del manual.
- En el cuadro de diálogo principal (ver Figura 12.1), seleccionar las variables *juez\_1* y *juez\_2* como variables *fila* y *columna*, respectivamente.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Tablas de contingencias: Estadísticos* (ver Figura 12.3) y marcar la opción **Kappa**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Figura 12.4. Datos de la Tabla 12.13 reproducidos en el *Editor de datos*

	juez1	juez2	ncasos
1	1	1	20
2	1	2	8
3	1	3	6
4	1	4	1
5	2	1	7
6	2	2	36
7	2	3	14
8	2	4	4
9	3	1	1
10	3	2	8
11	3	3	43
12	3	4	7
13	4	1	2
14	4	2	6
15	4	3	4
16	4	4	33

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestra la Tabla 12.14. La tabla recoge el valor del estadístico *kappa* (0,583) y su nivel crítico (*Sig. aproximada* < 0,0005), el cual permite decidir sobre la hipótesis nula de ausencia de acuerdo: puesto que el nivel crítico es muy pequeño, se puede rechazar la hipótesis de acuerdo nulo y concluir que existe un acuerdo mayor que el esperado por azar.

Al igual que ocurría con el resto de medidas de asociación estudiadas, el índice *kappa* aparece acompañado de su valor estandarizado (*T aproximada*), que se obtiene dividiendo el valor de *kappa* por su error típico, calculado éste bajo el supuesto de acuerdo nulo; la tabla también recoge el error típico calculado sin suponer acuerdo nulo (*Error típico asintótico*).

Tabla 12.14. Índice de acuerdo *kappa*

	Valor	Error típ. asint. <sup>a</sup>	T aproximada <sup>b</sup>	Sig. aproximada
Medida de acuerdo Kappa	.538	.046	12.921	.000
N de casos válidos	200			

a. Asumiendo la hipótesis alternativa.

b. Empleando el error típico asintótico basado en la hipótesis nula.

## Índices de riesgo

Las frecuencias de una tabla de contingencias pueden obtenerse utilizando dos estrategias básicas de recogida de datos. En la estrategia habitual, que es la que se ha asumido al aplicar todas las medidas de asociación estudiadas hasta aquí, los datos representan un corte temporal **transversal**: se recogen en el mismo o aproximadamente el mismo punto temporal y, consecuentemente, el tiempo no interviene como una variable relevante en el análisis. Si en lugar de esto se mide una o más variables en una muestra de sujetos y se hace seguimiento a esos sujetos para volver a tomar una medida de las mismas o de otras variables, se estará trabajando en una situación **longitudinal**: las medidas se toman en diferentes puntos temporales. Los

*índices de riesgo* que se estudian en este apartado resultan especialmente útiles para diseños longitudinales en los que se miden dos variables *dicotómicas* (ver Pardo y San Martín, 1998, págs. 511-514).

El seguimiento en los estudios longitudinales puede hacerse de dos formas: *hacia adelante* o *hacia atrás*. En los diseños longitudinales *hacia adelante*, llamados diseños *prospectivos* o de *cohortes*, los sujetos son clasificados en dos grupos dependiendo de la presencia o ausencia de algún factor desencadenante (por ejemplo, el hábito de fumar: fumadores y no fumadores) y se les hace seguimiento durante un periodo de tiempo hasta determinar la proporción de sujetos de cada grupo en los que se da un determinado *desenlace* objeto de estudio (por ejemplo, problemas vasculares). En los diseños longitudinales *hacia atrás*, también llamados *retrospectivos* o *caso-control*, se forman dos grupos de sujetos a partir de la presencia o ausencia de una determinada condición objeto de estudio (por ejemplo, sujetos sanos y pacientes con problemas vasculares) y se hace seguimiento hacia atrás intentando encontrar información sobre la proporción en la que se encuentra presente en cada muestra un determinado factor desencadenante (por ejemplo, el hábito de fumar). Lógicamente, cada uno de estos diseños de recogida de datos permite dar respuesta a diferentes preguntas y requiere la utilización de estadísticos particulares.

Es importante señalar aquí que los índices de riesgo no son válidos (como no lo es ningún otro procedimiento estadístico) para establecer relaciones de tipo causal; es decir, no son válidos para poder concluir que el factor desencadenante es la causa del desenlace estudiado. La razón de esto es que no existe control sobre la variable que se considera desencadenante. Para poder establecer relaciones de causalidad entre variables es necesario utilizar diseños experimentales (con asignación aleatoria que es imposible llevar a cabo en los diseños de cohortes o de caso control), o basar las conclusiones en teorías sólidas. Conviene tener esto presente justamente porque la palabra *riesgo*, tan utilizada en este tipo de diseños, podría llevar fácilmente a una conclusión equivocada.

### Diseños prospectivos o de cohortes (hacia adelante)

En los diseños *prospectivos* o de *cohortes* (en los cuales se establecen dos grupos de sujetos a partir de la presencia o ausencia de una condición que se considera desencadenante y se hace seguimiento *hacia adelante* para determinar en qué proporción de sujetos de cada grupo se produce un determinado desenlace), la medida de interés suele ser el *riesgo relativo* ( $R_r$ ): el grado en que la proporción de desenlaces es más alta en un grupo que en el otro:

$$R_r = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}$$

El valor del índice de riesgo relativo se interpreta de la siguiente manera: la proporción de desenlaces entre los sujetos expuestos al factor desencadenante es  $R_r$  veces más alta que entre los sujetos no expuestos. De otra manera, por cada desenlace observado entre los sujetos no expuestos, cabe esperar que aparezcan  $R_r$  desenlaces entre los sujetos expuestos. Un riesgo relativo de 1 indica que la probabilidad de encontrar el desenlace es la misma en el grupo de sujetos expuestos y en el grupo de sujetos no expuestos.

La Tabla 12.15 recoge los datos obtenidos en un estudio prospectivo (hacia adelante o de *cohortes*) sobre la relación entre el hábito de fumar, *tabaquismo* (fumadores y no fumado-



res), y la presencia de *problemas vasculares* (con problemas y sin problemas) en una muestra de 240 sujetos.

Tabla 12.15. Tabla de contingencias de *tabaquismo* por *problemas vasculares*

Recuento		Problemas vasculares		Total
		Con problemas	Sin problemas	
Tabaquismo	Fuman	23	81	104
	No fuman	9	127	136
Total		32	208	240

Entre los fumadores, la proporción de sujetos con problemas vasculares vale  $n_{11}/n_{1+} = 23/104 = 0,221$ . Entre los no fumadores,  $n_{21}/n_{2+} = 9/136 = 0,066$ . El riesgo relativo se obtiene dividiendo ambas proporciones:  $R_r = 0,221/0,066 = 3,34$ . Este valor indica que la proporción de problemas de tipo vascular entre los fumadores es 3,34 veces más alta que entre los no fumadores. O, de otra manera: por cada no fumador con problemas vasculares, cabe esperar encontrar 3,34 fumadores con problemas vasculares.

Para valorar si el índice de riesgo obtenido es significativamente distinto de 1, puede calcularse el intervalo de confianza para  $R_r$  mediante:

$$L_i = R_r e^{(z_{\alpha/2} \sqrt{n_{12}/n_{11}n_{1+} + n_{22}/n_{21}n_{2+}})} \quad \text{y} \quad L_s = R_r e^{(z_{1-\alpha/2} \sqrt{n_{12}/n_{11}n_{1+} + n_{22}/n_{21}n_{2+}})}$$

Si el intervalo de confianza no contiene el valor 1, puede concluirse que el riesgo de experimentar el desenlace estudiado no es el mismo en los dos grupos comparados.

### Diseños retrospectivos o de caso-control (hacia atrás)

En los diseños *retrospectivos* o de *caso-control*, tras formar dos grupos de sujetos a partir de alguna condición de interés, se va hacia atrás buscando la presencia de algún factor desencadenante. El mismo estudio sobre *tabaquismo* y *problemas vasculares* del apartado anterior podría diseñarse seleccionando dos grupos de sujetos (*con* y *sin* problemas vasculares) y buscando en la historia clínica de cada sujeto la presencia o ausencia del hábito de fumar. Puesto que el tamaño de los grupos se fija a partir de la presencia o ausencia de un determinado desenlace, no tiene sentido calcular un índice de riesgo basado en las proporciones de desenlaces (incidencias), pues el número de fumadores y no fumadores no ha sido previamente establecido sino que es producto del muestreo. Pero puede calcularse la proporción o *ventaja* (*odds*) de tener problemas vasculares respecto de no tenerlos, tanto en el grupo de fumadores como en el de no fumadores, y utilizar el cociente entre esas ventajas como una estimación del riesgo relativo:

$$O_r = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11} n_{22}}{n_{21} n_{12}}$$

Este cociente se conoce como *odds ratio* (*razón de las ventajas* o *razón de productos cruzados*) y suele utilizarse como una estimación del riesgo relativo en los diseños de *caso-control*

(justamente por la imposibilidad de estimar las incidencias). El valor de la *odds ratio* ( $O_r$ ) es tanto mejor estimador del *riesgo relativo* ( $R_r$ ) cuanto más pequeñas son las proporciones de desenlaces en cada grupo (cuanto más pequeñas son esas proporciones, más pequeña es la diferencia entre  $R_r$  y  $O_r$ ).

En los datos de la Tabla 12.15, la *odds* fumadores/no-fumadores en el grupo de sujetos *con problemas* vasculares vale:  $23/9=2,5556$ ; y en el grupo de sujetos *sin problemas* vasculares:  $81/127=0,6378$ . El riesgo en un diseño de *caso-control* se estima dividiendo ambas *odds*:  $O_r=0,284/0,071=4,007$ . Este valor suele interpretarse del mismo modo que el índice de riesgo relativo  $R_r$ , pues se utiliza como una estimación del mismo: la proporción de sujetos con problemas vasculares es 4 veces más alta entre los fumadores que entre los no fumadores (aunque lo que realmente se está calculando debería llevar a interpretar que la proporción de fumadores es 4,007 mayor entre los sujetos con problemas vasculares que entre los sujetos sin problemas vasculares). Un índice de riesgo de 1 indica que la probabilidad de encontrarse con el factor desencadenante es la misma en los dos grupos estudiados.

Para determinar si este índice de riesgo es significativamente distinto de 1, puede calcularse un intervalo de confianza mediante:

$$L_i = O_r e^{(z_{\alpha/2} \sqrt{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}})} \quad \text{y} \quad L_s = O_r e^{(z_{1-\alpha/2} \sqrt{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}})}$$

Por supuesto, siempre es posible utilizar más de dos grupos (factores desencadenantes con más de dos niveles). Sin embargo, las comparaciones deberán efectuarse entre un grupo y el resto, para lo cual habrá que efectuar las recodificaciones pertinentes en cada caso. Si se tienen tres grupos (*fumadores*, *no fumadores* y *exfumadores*), se pueden reunir en un único grupo a *no fumadores* y *exfumadores* y calcular el riesgo del grupo de *fumadores* respecto del resto de grupos tomados juntos. Y si lo que interesa es calcular el riesgo del grupo de *fumadores* respecto de cada grupo por separado, se tendrá que utilizar el filtrado de casos o, alternativamente, crear dos variables nuevas: una primera con códigos válidos (1 y 2, por ejemplo) para los grupos de *fumadores* y de *no fumadores* y códigos de valor perdido para el grupo de *exfumadores*; y una segunda variable con códigos válidos para el grupo de *fumadores* y el grupo de *exfumadores* y códigos de valor perdido para el grupo de *no fumadores*.

### Ejemplo: Tablas de contingencias > Estadísticos > Riesgo

Este ejemplo explica cómo obtener e interpretar los índices de riesgo del procedimiento *Tablas de contingencias*:

- Reproducir en el *Editor de datos* los datos de la Tabla 12.15 tal como muestra la Figura 12.5 y ponderar el archivo con la variable *ncasos* (o abrir el archivo *Riesgo tabaco-vascular* que se encuentra en la página web del manual).

Figura 12.5. Datos de la Tabla 12.15 reproducidos en el *Editor de datos*

	tabaco	vascular	ncasos
1	1	1	23
2	1	2	81
3	2	1	9
4	2	2	127

- En el cuadro de diálogo principal (ver Figura 12.1), seleccionar las variables *tabaco* (tabaquismo) y *vascular* (problemas vasculares) como variables *fila* y *columna*, respectivamente.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Tablas de contingencias: Estadísticos* (ver Figura 12.3) y marcar la opción **Riesgo**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestra la Tabla 12.16. Puesto que el SPSS *ignora* si los datos de la tabla han sido recogidos con un diseño de *cohortes* o con un diseño de *caso-control*, ofrece tanto el índice de riesgo como la razón de las ventajas.

La primera fila de la tabla indica que el riesgo estimado se refiere al de fumadores sobre el de no fumadores (*Fuman/No fuman*) en un diseño de *caso-control* (*Razón de las ventajas*). Su valor (4,007) significa que, entre los sujetos con problemas vasculares, la probabilidad (el riesgo) de encontrar fumadores es 4 veces mayor que la de encontrar no fumadores.

La *razón de las ventajas* suele interpretarse como una estimación del riesgo relativo (aunque no debe olvidarse que esto sólo es correcto si la proporción de desenlaces es muy pequeña). En ese sentido, puede concluirse que la proporción de sujetos con problemas vasculares es 4 veces mayor entre los fumadores que entre los no fumadores.

Los límites del intervalo de confianza calculado al 95 por ciento indican que el riesgo obtenido es mayor que 1: se concluye que el riesgo es significativamente mayor que 1 cuando, como en el ejemplo, el valor 1 no se encuentre entre los límites obtenidos.

Las dos filas siguientes ofrecen dos índices de riesgo para un diseño de *cohortes* (dos índices porque el desenlace que interesa evaluar puede encontrarse en cualquiera de las dos categorías de la variable). Si el desenlace que interesa estudiar es la *presencia* de problemas vasculares, la probabilidad o riesgo de encontrar tal desenlace entre los fumadores es 3,34 veces mayor que la de encontrarlo entre los no fumadores; es decir, por cada sujeto *con* problemas vasculares entre los no fumadores, cabe esperar encontrar 3,34 sujetos *con* problemas vasculares entre los fumadores. Si el desenlace que interesa estudiar es la *ausencia* de problema vascular, la probabilidad o riesgo de encontrar tal desenlace entre los fumadores es menor que entre los no fumadores: por cada sujeto *sin* problemas vasculares entre los no fumadores, cabe esperar encontrar 0,83 sujetos *sin* problemas vasculares entre los fumadores. De otra forma: entre los no fumadores cabe esperar encontrar un 20% ( $1/0,83 = 1,20$ ) más de sujetos *sin* problemas vasculares que entre los fumadores.

En ninguno de los dos casos el valor 1 está incluido en el intervalo de confianza, por lo que puede concluirse que el riesgo encontrado es distinto de 1.

**Tabla 12.16.** Índices de riesgo (*cohortes* y *caso-control*)

	Valor	Intervalo de confianza al 95%	
		Inferior	Superior
Razón de las ventajas para Tabaquismo (Fuman / No fuman)	4.007	1.766	9.093
Para la cohorte Problemas vasculares = Con problemas	3.342	1.615	6.915
Para la cohorte Problemas vasculares = Sin problemas	.834	.746	.933
N de casos válidos	240		

Es importante tener presente que los índices de riesgo siempre se calculan dividiendo la información de la primera fila de la tabla entre la información de la segunda fila (en el ejemplo, la fila *fuman* entre la fila *no fuman*; ver Tabla 12.15). Como el orden en el que el SPSS coloca

en la tabla las categorías de las filas (también de las columnas) viene determinado por los códigos que tienen asignados (se ordenan de menor a mayor), es importante vigilar que la categoría que se desea tomar como *desencadenante*, es decir, aquella cuyo riesgo se desea establecer, reciba un código menor que la otra categoría (en el ejemplo se ha utilizado el código 1 para los que fuman y el código 2 para los que no fuman). Este detalle es importante debido a que este tipo de variables suelen codificarse como variables *indicador* («fuman = 1», «no fuman = 0»); una codificación tipo *indicador* haría que el SPSS construyera la tabla con los no fumadores en la primera fila, con las consiguientes consecuencias para el análisis.

## Proporciones relacionadas (prueba de McNemar)

Una variante de los diseños longitudinales recién estudiados consiste en medir una misma variable dicotómica («acierto-error», «a favor-en contra», etc.) en dos momentos diferentes. Esta situación es propia de diseños *antes-después* y resulta especialmente útil para evaluar el cambio entre dos momentos. La forma de proceder es la siguiente: se toma una medida de una variable dicotómica, se aplica un tratamiento (o simplemente se deja pasar el tiempo) y se vuelve a tomar una medida de la *misma variable* a los *mismos sujetos*. La situación sería similar si en lugar de tomar dos medidas a los mismos sujetos se tomara una medida a *pares* de sujetos igualados en algún criterio de interés. Tal es el caso, por ejemplo, cuando se utilizan casos y controles en un estudio clínico, o los dos miembros de una pareja en un estudio social, etc.

Este diseño permite contrastar la hipótesis nula de *igualdad de proporciones antes-después*, es decir, la hipótesis de que la proporción de *éxitos* es la misma en la medida *antes* y en la medida *después* (la categoría *éxito* se refiere a una cualquiera de las dos categorías de la variable dicotómica estudiada). Esta hipótesis es equivalente a la hipótesis de *simetría completa*, es decir a la hipótesis de que la proporción de cambios en una dirección es la misma que la proporción de cambios en la otra dirección.

La Tabla 12.17 muestra una forma típica de presentar los datos provenientes de un diseño *antes-después* con una variable dicotómica. Se trata de una muestra de 240 sujetos a los que se les ha preguntado sobre su intención de voto antes y después de un debate televisado con candidatos del partido A y del partido B.

**Tabla 12.17.** Tabla de contingencias de *intención de voto antes* por *intención de voto después*

Recuento		Intención de voto (después)		Total
		Partido A	Partido B	
Intención de voto (antes)	Partido A	51	45	96
	Partido B	80	64	144
Total		131	109	240

En este ejemplo, la hipótesis sobre *igualdad de proporciones antes-después* puede formularse de la siguiente manera: la proporción de sujetos que tienen intención de votar al partido A en la medida *antes* es la misma que la proporción de sujetos que tienen intención de votar al partido A en la medida *después* (la hipótesis podría estar referida al partido B en lugar de estar referida al partido A; es del todo irrelevante cuál de las dos categorías de la variable dicotómica se toma como punto de referencia). La hipótesis de igualdad de proporciones también puede plantearse en términos del número o proporción de cambios: el número o proporción

de cambios que se producen en una dirección (sujetos que pasan de votar al partido *A* a votar al partido *B*) es igual al número o proporción de cambios que se producen en la otra dirección (sujetos que pasan de votar al partido *B* a votar al partido *A*).

Así pues, de acuerdo con la hipótesis nula de *igualdad de proporciones antes-después*, los cambios en una dirección deben ser los mismos que los cambios en la otra dirección. Para contrastar esta hipótesis, el estadístico de McNemar (1947) compara los cambios que se producen entre el *antes* y el *después* en ambas direcciones y determina la probabilidad asociada a ese número de cambios (o a cualquier otro número de cambios más alejado de la hipótesis de igualdad) asumiendo que las proporciones *antes-después* son iguales. Se rechaza la hipótesis nula de igualdad de proporciones cuando los cambios en una dirección son significativamente más numerosos que en la otra. El estadístico de McNemar adopta la forma:

$$X^2_{\text{McNemar}} = \frac{(n_1 - n_2 - 1)^2}{n_c}$$

donde  $n_1$  y  $n_2$  se refieren al número de cambios en una y otra dirección y  $n_c = n_1 + n_2$ . El estadístico de McNemar se distribuye aproximadamente según el modelo de probabilidad *chi-cuadrado* con 1 grado de libertad, de modo que puede utilizarse el valor de ese estadístico y la distribución *chi-cuadrado* para conocer la probabilidad aproximada de encontrar un número de cambios como el encontrado (o más alejado del valor esperado bajo hipótesis) y contrastar así la hipótesis de igualdad de proporciones.

En el caso de que la variable analizada tenga más de dos categorías (por ejemplo, añadiendo un partido o más a los datos de la Tabla 12.17), Bowker (1948) ha propuesto una modificación del estadístico de McNemar para evaluar la hipótesis de *igualdad de proporciones antes-después*:

$$X^2_{\text{McNemar-Bowker}} = \sum_{i < j} \sum \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$$

Este estadístico se distribuye según el modelo de probabilidad *chi-cuadrado* con grados de libertad igual al número de categorías de la variable menos uno.

Cuando la complejidad computacional no constituye una barrera, el SPSS utiliza la distribución binomial (con parámetros  $n_c$  y  $\pi = 0,5$ ) para obtener la probabilidad exacta asociada al número de cambios observado (la ecuación binomial está multiplicada por 2; esto significa que el procedimiento ofrece el nivel crítico bilateral):

$$p_{\text{binomial}} = 2 \sum_{i=0}^r \binom{n_1 + n_2}{i} 0,5^{n_1 + n_2}$$

donde:  $n_1$  = nº de casos en los que la medida *antes* es menor que la medida *después*.

$n_2$  = nº de casos en los que la medida *antes* es mayor que la medida *después*.

$r$  = el menor de  $n_1$  y  $n_2$ .

$i$  = 0, 1, 2, ...,  $r$ .

En los datos de la Tabla 12.17 se tiene:  $n_1 = 45$ ,  $n_2 = 80$ ,  $r = 45$ ,  $i$  = «todos los valores menores o iguales que 45». En los datos de la Tabla 12.19 se tiene:  $n_1 = 18 + 16 + 31 = 65$ ,  $n_2 = 12 + 14 + 9 = 37$ ,  $r = 37$ ,  $i$  = «todos los valores menores o iguales que 37».

### Ejemplo: Tablas de contingencias > Estadísticos > McNemar

Este ejemplo utiliza los datos de la Tabla 12.17 para ilustrar cómo obtener e interpretar el contraste sobre dos proporciones relacionadas (conocido como prueba de McNemar) del procedimiento Tablas de contingencias.

- Reproducir en el *Editor de datos* los datos de la Tabla 12.17 tal como muestra la Figura 12.6 y ponderar el archivo con la variable *ncasos* (o abrir el archivo *McNemar voto* (2) que se encuentra en la página *web* del manual).

Figura 12.6. Datos de la Tabla 12.17 reproducidos en el *Editor de datos*

	antes	despues	ncasos
1	1	1	51
2	1	2	45
3	2	1	80
4	2	2	64

- En el cuadro de diálogo principal (ver Figura 12.1), seleccionar las variables *antes* (intención de voto antes) y *después* (intención de voto después) como variables *fila* y *columna*, respectivamente.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Tablas de contingencias: Estadísticos* (ver Figura 12.3) y marcar la opción **McNemar**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestra la Tabla 12.18. La tabla informa sobre el nivel crítico asociado al número de cambios observados (*Sig. exacta bilateral*) y el número de casos válidos. El hecho de que la tabla no informe sobre el valor del estadístico de McNemar significa que el nivel crítico se ha calculado utilizando la distribución *binomial* (la cual permite obtener la probabilidad exacta en lugar de la aproximada).

Tabla 12.18. Prueba de *McNemar*

	Valor	Sig. exacta (bilateral)
Prueba de McNemar		.002 <sup>a</sup>
N de casos válidos	240	

a. Utilizada la distribución binomial

Cualquiera que sea la forma de obtener el nivel crítico, su significado siempre es el mismo: indica el grado de compatibilidad existente entre los datos muestrales y la hipótesis nula. En el ejemplo, puesto que el nivel crítico es menor que 0,05 (*Sig. exacta bilateral* = 0,002), se puede rechazar la hipótesis nula de igualdad de proporciones y concluir que la proporción de sujetos que piensan votar al partido A antes del debate televisado ( $96/240=0,40$ ) ha cambiado significativamente –ha aumentado– tras el debate ( $131/240=0,55$ ).

La opción **McNemar** también permite contrastar la hipótesis de igualdad de proporciones cuando la variable analizada tiene más de dos categorías. La Tabla 12.19 muestra un ejemplo similar al estudiado (ver Tabla 12.17) pero con una variable con tres categorías.

**Tabla 12.19.** Tabla de contingencias de *intención de voto antes* por *intención de voto después*

Recuento		Intención de voto después			Total
		Partido A	Partido B	Partido C	
Intención de voto antes	Partido A	54	18	16	88
	Partido B	12	42	31	85
	Partido C	14	9	63	86
Total		80	69	110	259

Para contrastar la hipótesis de *simetría* o de *igualdad de proporciones antes-después* con los datos de la Tabla 12.19:

- Reproducir en el *Editor de datos* los datos de la Tabla 12.19 tal como muestra la Figura 12.7 y ponderar el archivo con la variable *ncasos* (o abrir el archivo *McNemar voto* (3) que se encuentra la página *web* del manual).

**Figura 12.7.** Datos de la Tabla 12.19 reproducidos en el *editor de datos*

	antes	después	ncasos
1	1	1	54
2	1	2	18
3	1	3	16
4	2	1	12
5	2	2	42
6	2	3	31
7	3	1	14
8	3	2	9
9	3	3	63

- En el cuadro de diálogo principal (ver Figura 12.1), seleccionar las variables *antes* y *después* como variables *fila* y *columna*, respectivamente.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Tablas de contingencias: Estadísticos* (ver Figura 12.3) y marcar la opción **McNemar**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestra la Tabla 12.20. La tabla ofrece el valor del estadístico de *McNemar-Bowker* (*Valor* = 13,433) junto con sus grados de libertad (*gl* = 3) y su nivel crítico (*Sig. asintótica bilateral* = 0,004). Puesto que el nivel crítico es menor que 0,05, se puede rechazar la hipótesis nula de igualdad de proporciones y concluir que las proporciones de sujetos que piensan votar a los partidos A, B y C no son las mismas antes y después del debate televisado.

**Tabla 12.20.** Prueba de *McNemar-Bowker*

	Valor	gl	Sig. asintótica (bilateral)
Prueba de McNemar-Bowker	13.433	3	.004
N de casos válidos	259		

El problema de este contraste es que, dado que la variable categórica analizada tiene más de dos categorías, no es posible determinar qué proporción marginal difiere de qué otra. Por tanto, una vez rechazada la hipótesis de igualdad de proporciones, es necesario volver a aplicar la prueba de McNemar a cada par de categorías: primero al partido A y al B, a continuación al A y al C, y por último al B y al C. Para hacer estas comparaciones es necesario aplicar un filtro al archivo de datos (con la opción **Seleccionar casos...** del menú **Transformar**) de tal manera que en cada contraste únicamente intervengan las dos categorías que se desea comparar. La Figura 12.8 muestra los tres filtros creados para efectuar los tres contrastes. La variable *filtroab* permite aislar los partidos A y B; la variable *filtroac* permite aislar los partidos A y C; la variable *filtrobc* permite aislar los partidos B y C.

Aplicando la prueba de McNemar tras activar consecutivamente cada uno de los tres filtros definidos se obtienen los resultados que muestran las Tablas 12.21 a la 12.23. Los niveles críticos obtenidos (*Sig. exacta bilateral*) indican que las proporciones antes-después únicamente difieren entre los partidos B y C (*Sig.* = 0,003).

**Figura 12.8.** Datos de la Figura 12.7 con tres variables *filtro* añadidas

	antes	después	ncasos	filtroab	filtroac	filtrobc
1	1	1	54	1	1	0
2	1	2	18	1	0	0
3	1	3	16	0	1	0
4	2	1	12	1	0	0
5	2	2	42	1	0	1
6	2	3	31	0	0	1
7	3	1	14	0	1	0
8	3	2	9	0	0	1
9	3	3	63	0	1	1

**Tabla 12.21.** Tabla de contingencias de *intención de voto antes* por *intención de voto después* (izquierda) y prueba de *McNemar* (derecha). Partidos A y B

Recuento		Intención de voto después		Total
		Partido A	Partido B	
Intención de voto antes	Partido A	54	18	72
	Partido B	12	42	54
Total		66	60	126

	Valor	Sig. exacta (bilateral)
Prueba de McNemar		.362 <sup>a</sup>
N de casos válidos	126	

a. Utilizada la distribución binomial

**Tabla 12.22.** Tabla de contingencias de *intención de voto antes* por *intención de voto después* (izquierda) y prueba de *McNemar* (derecha). Partidos A y C

Recuento		Intención de voto después		Total
		Partido A	Partido C	
Intención de voto antes	Partido A	54	16	70
	Partido C	14	63	77
Total		68	79	147

	Valor	Sig. exacta (bilateral)
Prueba de McNemar		.856 <sup>a</sup>
N de casos válidos	147	

a. Utilizada la distribución binomial



**Tabla 12.23.** Tabla de contingencias de *intención de voto antes* por *intención de voto después* (izquierda) y prueba de *McNemar* (derecha). Partidos B y C

Recuento				
		Intención de voto después		Total
		Partido B	Partido C	
Intención de voto antes	Partido B	42	28	70
	Partido C	9	63	72
Total		51	91	142

	Valor	Sig. exacta (bilateral)
Prueba de McNemar		.003 <sup>a</sup>
N de casos válidos	142	

a. Utilizada la distribución binomial

## Combinación de tablas 2×2 (Cochran y Mantel-Haenszel)

En ocasiones, puede interesar analizar los diseños de *cohortes* y de *caso-control* (descritos ya en el apartado sobre los *índices de riesgo*) controlando el efecto de terceras variables. Estas situaciones se producen, por ejemplo, cuando se desea evaluar el efecto de un tratamiento sobre una determinada respuesta utilizando distintos grupos de pacientes.

En general, se trata de estudiar si existe o no asociación entre una variable *factor* y una variable *respuesta*, ambas dicotómicas, cuando se dispone de información referida a varios *estratos* (distintos grupos de edad o de sexo, pacientes con distinta sintomatología, distintas dosis de fármaco, distintos grupos étnicos, etc.). La Tabla 12.24 recoge datos referidos a las variables *tabaquismo* y *problemas vasculares* en dos estratos: *varones* y *mujeres*.

**Tabla 12.24.** Tabla de contingencias de *tabaquismo* por *problemas vasculares* en *varones* y *mujeres*

Recuento					
Sexo			Problemas vasculares		Total
			Con problemas	Sin problemas	
Varones	Tabaquismo	Fuman	22	103	125
		No fuman	17	151	168
	Total		39	254	293
Mujeres	Tabaquismo	Fuman	23	81	104
		No fuman	9	127	136
	Total		32	208	240

En situaciones de este tipo, utilizar el estadístico *chi-cuadrado* de Pearson sobre el conjunto de datos agrupados puede arrojar resultados equívocos. Y analizar separadamente cada estrato no proporciona una idea global del efecto de la variable *factor*. Se obtiene información más ajustada utilizando los estadísticos de Cochran y Mantel-Haenszel para contrastar la hipótesis de independencia condicional, es decir, la hipótesis de independencia entre las variables *factor* y *respuesta* una vez que se ha controlado el efecto de los *estratos*. El estadístico de Cochran (1954) adopta la siguiente forma:

$$X^2_{\text{Cochran}} = (\sum_k n_k - \sum_k m_k)^2 / \sum_k \sigma_{n_k}^2$$

donde  $k$  se refiere a cada uno de los estratos;  $n_k$  a la frecuencia observada en una cualquiera de las casillas del estrato  $k$  (sólo una y siempre la misma en todos los estratos);  $m_k$  a las frecuencias esperadas correspondientes a  $n_k$ ; y  $\sigma_{n_k}^2 = n_{1+k} n_{2+k} n_{+1k} n_{+2k} / n^3$  (siendo  $n_{1+k}$ ,  $n_{2+k}$ ,  $n_{+1k}$  y  $n_{+2k}$  las cuatro frecuencias marginales de las tablas  $2 \times 2$  de cada estrato).

El estadístico de Mantel-Haenszel (1959) es idéntico al de Cochran, excepto en lo que se refiere a dos detalles: (1) utiliza la corrección por continuidad (restando medio punto al numerador de la ecuación antes de elevar el paréntesis al cuadrado), y (2) en el denominador de la varianza utiliza  $n^2(n-1)$  en lugar de  $n^3$ .

Ambos estadísticos (el de Cochran y el de Mantel-Haenszel) se distribuyen según el modelo de probabilidad  $\chi^2$  con 1 grado de libertad. Si el nivel crítico asociado a ellos es menor que 0,05, se deberá rechazar la hipótesis nula de independencia condicional y concluir que, una vez controlado el efecto de la variable *estratos*, las variables *factor* y *respuesta* están asociadas.

### Ejemplo: Tablas de contingencias > Estadísticos > Cochran-Mantel-Haenszel

Este ejemplo muestra cómo obtener los estadísticos de Cochran y Mantel-Haenszel utilizando los datos de la tabla 12.24:

- Reproducir en el *Editor de datos* los datos de la Tabla 12.24 tal como muestra la Figura 12.9 y ponderar el archivo con la variable *ncasos* (o abrir el archivo *Riesgo sexo-tabaco-vascular* que se encuentra en la página web del manual).

Figura 12.9. Datos de la Tabla 12.24 reproducidos en el *Editor de datos*

	sexo	tabaco	vascular	ncasos
1	1	1	1	22
2	1	1	2	103
3	1	2	1	17
4	1	2	2	151
5	2	1	1	23
6	2	1	2	81
7	2	2	1	9
8	2	2	2	127

- En el cuadro de diálogo principal (ver Figura 12.1), seleccionar las variables *tabaco* (tabaquismo) y *vascular* (problemas vasculares) como variables *fila* y *columna*, respectivamente, y la variable *sexo* como variable *capa*.
- Pulsar el botón *Estadísticos...* para acceder al subcuadro de diálogo *Tablas de contingencias: Estadísticos* (ver Figura 12.3) y marcar la opción *Estadísticos de Cochran y de Mantel-Haenszel*.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestra la Tabla 12.25. El estadístico de Cochran vale 13,933 y tiene un nivel crítico asociado (*Sig. asintótica bilateral*) menor que 0,0005; puesto que el nivel crítico es muy pequeño, se puede rechazar la hipótesis nula de independencia condicional y concluir que, una vez controlado el efecto de la variable *sexo*, las variables *tabaquismo* y *problemas vasculares* están relacionadas. A idéntica conclu-

sión se llega con el estadístico de Mantel-Haenszel, cuyo estadístico toma un valor de 12,939 con un nivel crítico asociado menor que 0,0005.

**Tabla 12.25.** Pruebas de homogeneidad de la razón de las ventajas

Estadísticos		Chi-cuadrado	gl	Sig. asintótica (bilateral)
Independencia condicional	Cochran	13.933	1	.000
	Mantel-Haenszel	12.939	1	.000
Homogeneidad	Breslow-Day	1.911	1	.167
	De Tarone	1.910	1	.167

Si se rechaza la hipótesis de independencia condicional, el interés del investigador debe orientarse hacia la cuantificación del grado de relación existente entre las variables *factor* y *respuesta*. Para ello, el SPSS ofrece una estimación del riesgo (*odds-ratio* = *razón de las ventajas*) común para todos los estratos. Pero esta estimación *común* sólo tiene sentido si no existe interacción triple, es decir, si la relación detectada entre las variables *factor* y *respuesta* es homogénea en todos los *estratos*. Esta hipótesis nula de homogeneidad de las *odds-ratio* inter-estratos puede contrastarse utilizando los estadísticos de Breslow-Day (1980, 1987) y Tarone (1985; Tarone, Gart y Hauck, 1983; ver también Breslow, 1996).

La Tabla 12.25 indica que el nivel crítico asociado a ambos estadísticos vale 0,167, por lo que puede mantenerse la hipótesis de homogeneidad. Y, puesto que puede asumirse que el riesgo es homogéneo en todos los estratos, tiene sentido obtener una estimación común o global del riesgo.

La Tabla 12.26 ofrece estimación común del riesgo basada en un estadístico debido a Mantel y Haenszel (1959) que adopta la siguiente forma:

$$RV_{\text{común}} = [\sum_k (n_{11k}n_{22k}/n_{++k})] / [\sum_k (n_{12k}n_{21k}/n_{++k})]$$

En el ejemplo de la Tabla 12.26, el valor de la estimación del riesgo común (*Estimación*) es 2,068, con un intervalo de confianza al 95 % definido por los límites 1,555 y 4,373. Puesto que el intervalo de confianza no contiene el valor 1, se puede concluir que el riesgo común (el de todos los estratos tomados juntos) es significativamente mayor que 1.

La tabla ofrece esta misma información en escala logarítmica; en este caso, el valor de referencia para la interpretación ya no es el 1, sino el 0.

**Tabla 12.26.** Estimación de la razón de las ventajas común de Mantel-Haenszel

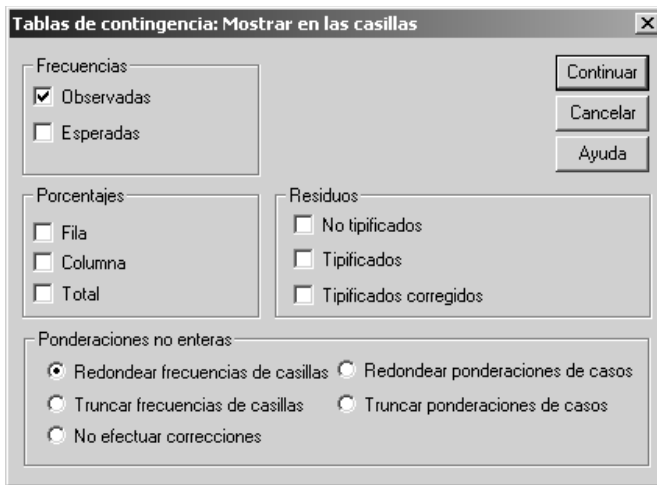
Estimación			2.608
ln(estimación)			.959
Error típ. de ln(estimación)			.264
Sig. asintótica (bilateral)			.000
Intervalo de confianza asintótico al 95%	Razón de ventajas común	Límite inferior	1.555
		Límite superior	4.373
	ln(Razón de ventajas común)	Límite inferior	.442
		Límite superior	1.475

## Contenido de las casillas

Todas las tablas de contingencias estudiadas hasta este momento se han construido utilizando únicamente las frecuencias absolutas, es decir, el número de casos resultantes de la clasificación. Pero las casillas o celdas de una tabla de contingencias pueden contener información muy variada. Parte de esta información es especialmente útil para poder interpretar apropiadamente las pautas de asociación presentes en una tabla después de que algún estadístico de los ya estudiados conduce al rechazo de la hipótesis de independencia. Para controlar el contenido de las casillas:

- Pulsar el botón **Casillas...** del cuadro de diálogo *Tablas de contingencias* (ver Figura 12.1) para acceder al subcuadro de diálogo *Tablas de contingencias: Mostrar en las casillas* que recoge la Figura 12.10.

Figura 12.10. Subcuadro de diálogo *Tablas de contingencias: Mostrar en las casillas*



**Frecuencias.** Es posible seleccionar uno o los dos siguientes tipos de frecuencias absolutas:

- **Observadas.** Número de casos resultantes de la clasificación.
- **Esperadas.** Número de casos que debería haber en cada casilla si las variables *fila* y *columna* fueran independientes.

**Porcentajes.** Las opciones de este recuadro permiten seleccionar una o más de las siguientes frecuencias porcentuales:

- **Fila.** Porcentaje que la frecuencia observada de una casilla representa respecto al total marginal de su fila.
- **Columna.** Porcentaje que la frecuencia observada de una casilla representa respecto al total marginal de su columna.
- **Total.** Porcentaje que la frecuencia observada de una casilla representa respecto al número total de casos de la tabla.

**Residuos.** Los residuos son las diferencias existentes entre las frecuencias observadas y esperadas de cada casilla. Son especialmente útiles para averiguar en qué grado se desvían de la hipótesis de independencia las frecuencias de cada casilla; consecuentemente, son útiles para interpretar la pautas de asociación presentes en una tabla. Es posible seleccionar una o más de las siguientes opciones:

- " **No tipificados.** Diferencias entre las frecuencias observadas y las esperadas (residuos en *bruto*):  $r_{ij} = n_{ij} - \hat{m}_{ij}$ .
- " **Tipificados.** Residuo no tipificado dividido por la raíz cuadrada de su correspondiente frecuencia esperada:

$$r_{ij(\text{tipificado})} = r_{ij} / \sqrt{\hat{m}_{ij}}$$

Su valor esperado vale 0, pero su desviación típica es menor que 1, lo cual hace que no puedan interpretarse como puntuaciones *Z*. Sin embargo, sirven como indicadores del grado en que cada casilla contribuye al valor del estadístico *chi*-cuadrado. De hecho, sumando los cuadrados de los residuos tipificados se obtiene el valor del estadístico *chi*-cuadrado. Razón por la cual estos residuos también reciben el nombre de residuos de *Pearson*.

- " **Tipificados corregidos.** Residuos tipificados corregidos de Haberman (1973). Estos residuos se distribuyen normalmente con media 0 y desviación típica 1. Se calculan dividiendo el residuo de cada casilla por su *error típico*, que en tablas bidimensionales se obtiene como la raíz cuadrada de:

$$r_{ij(\text{corregido})} = r_{ij} / \sqrt{\hat{m}_{ij}(1 - n_{i+}/n)(1 - n_{+j}/n)}$$

La gran utilidad de los residuos tipificados corregidos radica precisamente en que su distribución es normal con media cero y desviación típica uno:  $N(0, 1)$ . Una variable distribuida de esta forma es fácilmente interpretable: utilizando un nivel de confianza de, por ejemplo, 0,95, puede afirmarse que los residuos mayores que 1,96 (puntuación típica correspondiente al cuantil 97,5 en una distribución normal) delatan casillas con demasiados casos, es decir, con más casos de los que cabría esperar por azar si las dos variables estudiadas fueran realmente independientes; mientras que los residuos menores que -1,96 (puntuación típica correspondiente al cuantil 2,5 en una distribución normal) delatan casillas con pocos casos, es decir, con menos casos de los que cabría esperar si las dos variables estudiadas fueran realmente independientes (los cuantiles 2,5 y 97,5 son los que se utilizan cuando se trabaja con un nivel de confianza del 95 %).

En tablas de contingencias con variables nominales, una vez que se ha establecido que entre dos variables existe asociación significativa (mediante el estadístico *chi*-cuadrado) y que se ha cuantificado esa asociación con alguna medida de asociación (coeficiente de contingencia, phi, etc.), los residuos tipificados corregidos constituyen una de las mejores herramientas disponibles para poder interpretar con precisión el significado de la asociación detectada, pues permiten valorar hacia dónde y desde dónde se producen desplazamientos significativos de casos (ver siguiente ejemplo).

**Ponderaciones no enteras.** Puesto que el contenido de las casillas (frecuencias observadas) de una tabla de contingencias representa el número de casos resultantes de la clasificación, las frecuencias observadas suelen ser valores enteros (es decir, valores sin decimales). No obstante, si se está trabajando con un archivo de datos *ponderado* y los valores de la variable utilizada para efectuar la ponderación poseen decimales, las frecuencias observadas resultantes de la ponderación también podrán tener decimales. En estos casos, el SPSS ofrece la posibilidad de redondear o truncar las frecuencias observadas o los valores de la variable de ponderación mediante las siguientes opciones:

**Redondear las frecuencias de las casillas.** Redondea las frecuencias observadas a su valor entero más próximo.

**Redondear los pesos de los casos.** Redondea los valores de la variable de ponderación a su valor entero más próximo. Lógicamente, si los valores de la variable de ponderación son enteros, las frecuencias resultantes de la ponderación también serán valores enteros.

**Truncar las frecuencias de las casillas.** Elimina de las frecuencias observadas la parte decimal.

**Truncar los pesos de los casos.** Elimina de los valores de la variable de ponderación la parte decimal. Lógicamente, si los valores de la variable de ponderación son enteros, las frecuencias resultantes de la ponderación también serán valores enteros.

**No efectuar correcciones.** No se efectúan cambios ni en las frecuencias observadas ni en los valores de la variable de ponderación.

Si se utiliza alguna de estas opciones para redondear o truncar las frecuencias observadas o los pesos de la variable de ponderación, la tabla de contingencias muestra las frecuencias redondeadas o truncadas y los estadísticos solicitados se calculan sobre esas frecuencias.

### ***Ejemplo: Tablas de contingencias > Casillas > Frecuencias, Porcentajes y Residuos***

Este ejemplo muestra cómo obtener e interpretar los diferentes tipos de frecuencias y residuos que ofrece el procedimiento **Tablas de contingencias**. Se basa en el archivo *Datos de empleados*, que se encuentra en la misma carpeta en la que está instalado el SPSS.

- En el cuadro de diálogo principal (ver Figura 12.1), seleccionar las variables *sexo* y *catlab* como variables *fila* y *columna*, respectivamente.
- Pulsar el botón **Casillas...** para acceder al subcuadro de diálogo *Tablas de contingencias: Casillas* (ver Figura 12.10) y marcar todas las opciones de los recuadros **Frecuencias**, **Porcentajes** y **Residuos**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor de resultados* ofrece la información que muestra la Tabla 12.27. Cada casilla de la tabla contiene ocho valores que se corresponden exactamente con las ocho opciones con selector cuadrado del subcuadro de diálogo *Tablas de contingencias: Mostrar en las casillas* (ver Figura 12.10).

Los distintos porcentajes pueden ayudar a intuir posibles pautas de asociación. Por ejemplo, los porcentajes de fila (*% de sexo*) indican que en las *mujeres* se da una concentración

en la categoría *administrativos* (95,4 %) que no se da entre los *hombres* (60,0 %); y los porcentajes de columna (*% de categoría laboral*) indican que el 88,1 % de los *directivos* son *hombres*. Pero son los residuos tipificados corregidos los que permiten interpretar de forma precisa el significado de la relación existente entre las variables *sexo* y *catlab*. Trabajando con una confianza del 95 %, basta con fijarse en los residuos mayores que +1,96 o menores que -1,96: en el grupo de *Administrativos*, existe una desproporción significativa a favor de las mujeres (residuo tipificado corregido de 8,8 frente a -8,8), mientras que en los grupos de *Seguridad* y *Directivos* existe una desproporción significativa a favor de los varones (residuos tipificados corregidos de 4,9 y 6,8 frente a -4,9 y -6,8).

Lo que se está afirmando no es que haya más mujeres administrativas que hombres administrativos (aunque esto es cierto), sino que la proporción de mujeres administrativas es significativamente mayor y la proporción de hombres administrativos significativamente menor de lo que pronostica la hipótesis nula de independencia. Del mismo modo, tampoco se está afirmando que entre los agentes de seguridad y entre los directivos haya más hombres que mujeres (aunque esto también es cierto), sino que, en esas casillas, la proporción de hombres es significativamente mayor y la proporción de mujeres significativamente menor de lo que pronostica la hipótesis nula de independencia.

Tabla 12.27. Tabla de contingencias de *sexo* por *categoría laboral*

			Categoría laboral			Total
			Administrativo	Seguridad	Directivo	
Sexo	Hombre	Frecuencia observada	157	27	74	258
		Frecuencia esperada	197.6	14.7	45.7	258.0
		% de Sexo	60.9%	10.5%	28.7%	100.0%
		% de Categoría laboral	43.3%	100.0%	88.1%	54.4%
		% del total	33.1%	5.7%	15.6%	54.4%
		Residuo	-40.6	12.3	28.3	
		Residuos tipificados	-2.9	3.2	4.2	
		Residuos corregidos	-8.8	4.9	6.8	
	Mujer	Frecuencia observada	206	0	10	216
		Frecuencia esperada	165.4	12.3	38.3	216.0
		% de Sexo	95.4%	.0%	4.6%	100.0%
		% de Categoría laboral	56.7%	.0%	11.9%	45.6%
		% del total	43.5%	.0%	2.1%	45.6%
		Residuo	40.6	-12.3	-28.3	
		Residuos tipificados	3.2	-3.5	-4.6	
		Residuos corregidos	8.8	-4.9	-6.8	
Total	Frecuencia observada		363	27	84	474
	Frecuencia esperada		363.0	27.0	84.0	474.0
	% de Sexo		76.6%	5.7%	17.7%	100.0%
	% de Categoría laboral		100.0%	100.0%	100.0%	100.0%
	% del total		76.6%	5.7%	17.7%	100.0%

La Tabla 12.28 puede ayudar a precisar el significado de los residuos tipificados corregidos. La tabla se ha obtenido cruzando las variables *sexo* (hombre, mujer) y *postura sobre la pena de muerte por asesinato* (a favor, en contra) y solicitando los porcentajes de *fila* y los residuos tipificados corregidos. Los datos corresponden al archivo *GSS93 reducido* que se encuentra en la misma carpeta en la que está instalado el SPSS. Los porcentajes de *fila* referidos a toda la muestra indican que el porcentaje de entrevistados que están a favor y en contra de la pena

de muerte es del 77,4 % y del 22,6 %, respectivamente. Pero los residuos tipificados corregidos indican que entre los hombres y las mujeres esos porcentajes no se mantienen (de hecho, el estadístico *chi*-cuadrado de Pearson aplicado a esta tabla tiene asociado un nivel crítico menor que 0,0005). En el grupo de hombres, el porcentaje de entrevistados que se muestran a favor de la pena de muerte sube hasta el 82,7 %; y, puesto que el residuo tipificado corregido de esa casilla vale 4,2, puede afirmarse que ese incremento es significativo. En el grupo de mujeres, por el contrario, el porcentaje de entrevistados que se muestran a favor de la pena de muerte baja hasta el 73,2 %; y, puesto que el residuo tipificado corregido de esa casilla vale -4,2, puede afirmarse que esa disminución es significativa.

Lo que se está queriendo decir con esto es que el porcentaje de mujeres que se manifiesta en contra de la pena de muerte es significativamente más alto que ese mismo porcentaje referido al total de entrevistados (pasa del 22,6 % al 26,8 %, con un residuo corregido de 4,2); pero esto no significa que haya más mujeres en contra de la pena de muerte que a favor de ella: a pesar de que el porcentaje de mujeres que se manifiesta en contra de la pena de muerte ha aumentado, sigue habiendo más mujeres a favor (572) que en contra (209).

**Tabla 12.28.** Tabla de contingencias de *sexo por postura sobre la pena de muerte por asesinato*

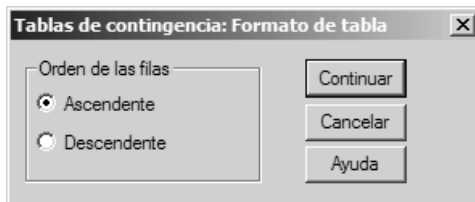
			Postura sobre la pena de muerte por asesinato		Total
			A favor	En contra	
Sexo	Hombre	Recuento	502	105	607
		% de Sexo	82,7%	17,3%	100,0%
		Residuos corregidos	4,2	-4,2	
	Mujer	Recuento	572	209	781
		% de Sexo	73,2%	26,8%	100,0%
		Residuos corregidos	-4,2	4,2	
Total	Recuento	1074	314	1388	
	% de Sexo	77,4%	22,6%	100,0%	

## Formato de las casillas

Las opciones de formato permiten controlar el orden en el que aparecerán las categorías de la variable que define las filas de la tabla de contingencias. Para cambiar este orden:

- Pulsar el botón **Formato...** del cuadro de diálogo principal (ver Figura 12.1) para acceder al subcuadro de diálogo *Tablas de contingencias: Formato de tabla* que muestra la Figura 12.5.

**Figura 12.11.** Subcuadro de diálogo *Tablas de contingencias: Formato de tabla*





**Orden de filas.** Las opciones de este recuadro permiten controlar el orden en el que aparecen las categorías de la variable *fila*:

**Ascendente.** Si la variable que define las filas es numérica, las categorías de esa variable se ordenan de menor a mayor. Si la variable que define las filas es de cadena, las categorías de esa variable se ordenan de la *a* a la *z* (los números preceden a las letras). Es la opción que se encuentra activa por defecto.

**Descendente.** Si la variable que define las filas es numérica, las categorías de esa variable se ordenan de mayor a menor. Si la variable que define las filas es de cadena, las categorías de esa variable se ordenan de la *z* a la *a* (los números preceden a las letras).

## Contrastes sobre medias

### Los procedimientos *Medias* y *Prueba T*

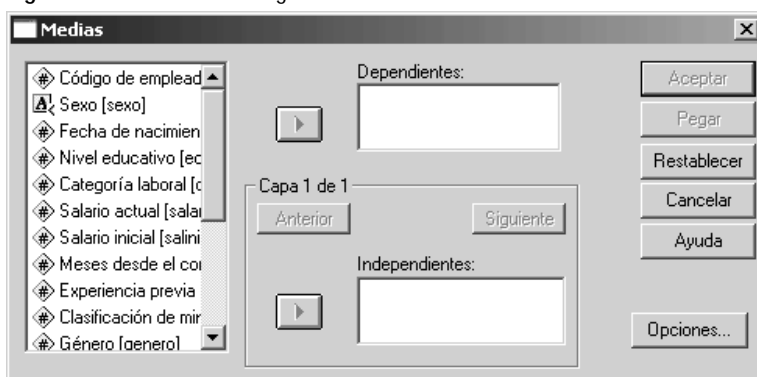
La opción **Comparar medias** del menú **Analizar** contiene cinco procedimientos estadísticos diseñados para efectuar contrastes de hipótesis sobre medias; entre ellos, la prueba *T* de *Student* y el análisis de varianza de un factor. En este capítulo se estudian cuatro de estos procedimientos: **Medias**, **Prueba T para una muestra**, **Prueba T para dos muestras independientes** y **Prueba T para dos muestras relacionadas**. El próximo capítulo está dedicado por completo al **Análisis de varianza de un factor**.

## Medias

El procedimiento **Medias** ofrece, como utilidad fundamental, estadísticos descriptivos que pueden calcularse para los distintos grupos y subgrupos definidos por una o más variables independientes. Opcionalmente, pueden solicitarse ANOVAs de un factor, obtener algunos estadísticos sobre proporción de varianza explicada y contrastar la hipótesis de linealidad (si bien estas opciones se estudiarán detalladamente en el próximo capítulo). Para utilizar el procedimiento **Medias**:

- Seleccionar la opción **Comparar medias > Medias...** del menú **Analizar** para acceder al cuadro de diálogo **Medias** que muestra la Figura 13.1.

Figura 13.1. Cuadro de diálogo *Medias*



Para obtener los estadísticos descriptivos que el procedimiento ofrece por defecto (*media aritmética, desviación típica y número de casos* de cada variable *dependiente* en cada uno de los grupos definidos por cada variable *independiente*):

- Seleccionar la variable o variables que interesa describir o aquellas en las que se van a comparar los grupos y trasladarlas a la lista **Dependientes**.
- Seleccionar la variable o variables que definen los grupos que interesa describir o comparar y trasladarlas a la lista **Independientes**.
- Pulsar el botón **Aceptar**.

Las variables *dependientes* son variables *cuantitativas*, mientras que las variables *independientes* son variables *categorías*.

Al seleccionar más de una variable independiente, el SPSS ofrece información descriptiva para cada uno de los grupos definidos por los distintos niveles de cada variable independiente seleccionada (es decir, no combina entre sí los niveles de las distintas variables independientes).

No obstante, también es posible combinar más de una variable independiente para, dentro de los grupos definidos por una primera variable, formar subgrupos definidos por una segunda variable (o una tercera, o una cuarta, etc.). De este modo, cada estadístico solicitado se calcula en cada uno de los subgrupos resultantes de la combinación. Esto se consigue definiendo *capas* con los botones **Siguiente** y **Anterior** del recuadro **Capa # de #**. Para obtener, por ejemplo, estadísticos descriptivos de la variable *salario* (salario actual) en cada uno de los subgrupos resultantes de combinar las variables *minoría* (clasificación de minorías) y *catlab* (categoría laboral):

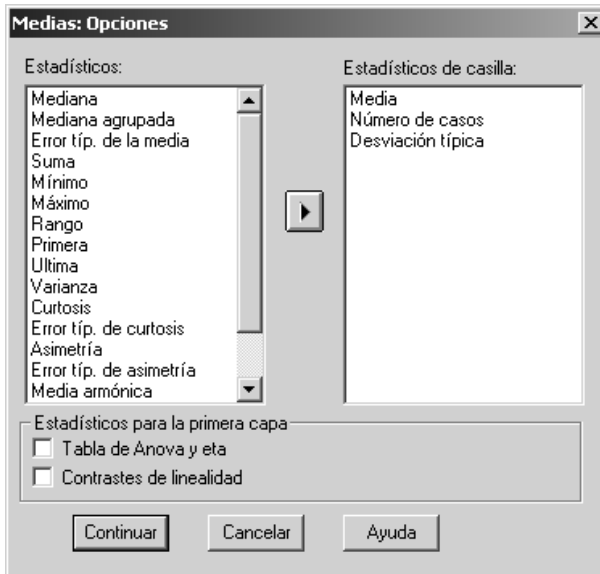
- Seleccionar la variable *salario* y trasladarla a la lista **Dependientes**.
- Seleccionar la variable *minoría* como variable **Independientes** en la *primera capa*.
- Pulsar el botón **Siguiente** para acceder a la *segunda capa*.
- Seleccionar la variable *catlab* como variable **Independiente** en la *segunda capa*.
- Utilizar el botón **Anterior** para moverse por capas previamente definidas.

Con estas especificaciones, el *Visor de resultados* ofrece algunos estadísticos descriptivos de la variable *salario* para cada uno de los 6 subgrupos resultantes de combinar los 2 niveles de la variable *minoría* con los 3 niveles de la variable *catlab*. Conforme se van creando capas, los valores # del recuadro **Capa # de #** van indicando el número de la capa en el que se está y el número total de capas definidas.

## Opciones

Las opciones del procedimiento **Medias** permiten seleccionar: (1) los estadísticos descriptivos concretos que interesa obtener, y (2) algunos contrastes (sobre medias, sobre linealidad) que el procedimiento no ofrece por defecto. Para solicitar estos estadísticos y contrastes:

- Pulsar el botón **Opciones...** del cuadro de diálogo principal (ver Figura 13.1) para acceder al subcuadro de diálogo *Medias: Opciones* que muestra la Figura 13.2.

Figura 13.2. Subcuadro de diálogo *Medias: Opciones*

**Estadísticos de casilla.** Contiene los estadísticos que el procedimiento **Medias** calcula por defecto. Cualquier estadístico adicional que se desee obtener debe trasladarse a esta lista desde la lista **Estadísticos**.

**Estadísticos para la primera capa.** Para los grupos definidos por las variables independientes seleccionadas en la primera capa (sólo en la primera capa), es posible marcar una o más de las siguientes opciones (ambas se describen con detalle en el próximo capítulo):

- **Tabla de ANOVA y eta.** Ofrece la tabla resumen del análisis de varianza de un factor y algunos estadísticos sobre la proporción de varianza explicada: el coeficiente de correlación de *Pearson* y su cuadrado (sólo si el formato de la variable independiente no es de *cadena*), y el coeficiente de correlación *eta* y su cuadrado.
- **Contrastes de linealidad.** Permite averiguar si una variable independiente *categorica* se relaciona linealmente con una variable dependiente *cuantitativa*. Estos contrastes no están disponibles para variables independientes con formato de *cadena*.

### **Ejemplo: Comparar medias > Medias**

Este ejemplo muestra cómo utilizar el procedimiento **Medias** para obtener algunos estadísticos de la variable *salario* en los subgrupos definidos por las variables *sexo* y *catlab* (del archivo *Datos de empleados*, que está en la misma carpeta en la que se ha instalado el SPSS):

- En el cuadro de diálogo principal (ver Figura 13.1), seleccionar la variable *salario* (salario actual) como **Dependiente** y la variable *sexo* como **Independiente**.
- Pulsar el botón **Siguiente** del recuadro **Capa 1 de 1** y seleccionar la variable *catlab* (categoría laboral) como variable **Independiente** en la *segunda capa*.

Pulsar el botón **Opciones...** para acceder al subcuadro de diálogo *Medias: Opciones* (ver Figura 13.2), seleccionar todos los estadísticos de la lista **Estadísticos** y trasladarlos a la lista **Estadísticos de casilla** (no marcar las opciones **Tabla de ANOVA** y **eta y Contrastes de linealidad**; estas dos opciones se describen con detalle en el próximo capítulo sobre *ANOVA de un factor*).

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestra la Tabla 13.1 (se ha pivotado la tabla para adaptarla al tamaño de la página).

**Tabla 13.1.** Estadísticos descriptivos del procedimiento *Medias*

Salario actual		Categoría laboral			
Sexo		Administrativo	Seguridad	Directivo	Total
Hombre	Media	\$31,558.15	\$30,938.89	\$66,243.24	\$41,441.78
	N	157	27	74	258
	Desv. típ.	\$7,997.978	\$2,114.616	\$18,051.570	\$19,499.214
	Mediana	\$29,850.00	\$30,750.00	\$63,750.00	\$32,850.00
	Mediana agrupada	\$29,737.50	\$30,725.00	\$63,125.00	\$32,850.00
	Error típ. de la media	\$638.308	\$406.958	\$2,098.452	\$1,213.968
	Suma	\$4,954,630	\$835,350	\$4,902,000	\$10,691,980
	Mínimo	\$19,650	\$24,300	\$38,700	\$19,650
	Máximo	\$80,000	\$35,250	\$135,000	\$135,000
	Rango	\$60,350	\$10,950	\$96,300	\$115,350
	Varianza	63967646,887	4471602,564	325859166,050	380219336,303
	Curtosis	9,850	3,652	2,116	2,780
	Error típ. de la curtosis	,385	,872	,552	,302
	Asimetría	2,346	-,368	1,193	1,639
	Error típ. de la asimetría	,194	,448	,279	,152
	Media armónica	\$30,070.04	\$30,793.00	\$62,096.13	\$35,392.56
	Media geométrica	\$30,750.38	\$30,867.29	\$64,071.30	\$37,972.18
	% de la suma total	30,4%	5,1%	30,0%	65,5%
	% del total de N	33,1%	5,7%	15,6%	54,4%
Mujer	Media	\$25,003.69		\$47,213.50	\$26,031.92
	N	206		10	216
	Desv. típ.	\$5,812.838		\$8,501.253	\$7,558.021
	Mediana	\$24,000.00		\$45,187.50	\$24,300.00
	Mediana agrupada	\$24,066.67		\$45,187.50	\$24,275.00
	Error típ. de la media	\$405.000		\$2,688.332	\$514.258
	Suma	\$5,150,760		\$472,135	\$5,622,895
	Mínimo	\$15,750		\$34,410	\$15,750
	Máximo	\$54,000		\$58,125	\$58,125
	Rango	\$38,250		\$23,715	\$42,375
	Varianza	33789086,810		72271294,722	57123688,268
	Curtosis	4,029		-1,554	4,641
	Error típ. de la curtosis	,337		1,334	,330
	Asimetría	1,421		-,019	1,863
	Error típ. de la asimetría	,169		,687	,166
	Media armónica	\$23,861.66		\$45,805.40	\$24,402.89
	Media geométrica	\$24,406.44		\$46,511.67	\$25,146.07
	% de la suma total	31,6%		2,9%	34,5%
	% del total de N	43,5%		2,1%	45,6%

## Prueba $T$ para una muestra

La prueba  $T$  para una muestra permite contrastar hipótesis referidas a una media poblacional. Al seleccionar una muestra aleatoria de tamaño  $n$  de una población en la que la variable  $Y_i$  se distribuye normalmente con media  $\mu$  y desviación típica  $\sigma$ , la media  $\bar{Y}$  de esa muestra puede tipificarse restándole su valor esperado (que es justamente la media de la población:  $E(\bar{Y}) = \mu$ ) y dividiendo esa diferencia por su error típico ( $\sigma_{\bar{Y}}$ ), es decir:

$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$$

Se obtiene así una puntuación  $Z$  normalmente distribuida con media 0 y desviación típica 1, que puede interpretarse utilizando la distribución normal estandarizada  $N(0, 1)$ . Recuérdese que en una distribución  $N(0, 1)$ , entre  $\pm 1,96$  puntuaciones típicas se encuentra el 95 % de los casos; entre  $\pm 2,58$  puntuaciones típicas se encuentra el 99 % de los casos, etc.

El error típico de la media ( $\sigma_{\bar{Y}}$ ) es la desviación típica de la distribución muestral de  $\bar{Y}$ , es decir, la desviación típica de las medias calculadas en todas las muestras de tamaño  $n$  que es posible extraer de una determinada población (ver, en el Capítulo 9, el apartado sobre distribuciones muestrales). Se obtiene mediante:  $\sigma_{\bar{Y}} = \sigma/\sqrt{n}$ , siendo  $\sigma$  la desviación típica de la población.

El problema que surge al intentar obtener  $\sigma_{\bar{Y}}$  es que el valor de la desviación típica poblacional  $\sigma$  es, generalmente, desconocido. Es necesario estimarlo utilizando la desviación típica muestral,  $S_{n-1}$ , en cuyo caso, el error típico de la media se obtiene mediante:  $\hat{\sigma}_{\bar{Y}} = S_{n-1}/\sqrt{n}$ . El hecho de tener que estimar la desviación típica poblacional hace que la tipificación del estadístico  $\bar{Y}$  ya no sea una puntuación  $Z$ , sino una puntuación  $T$ :

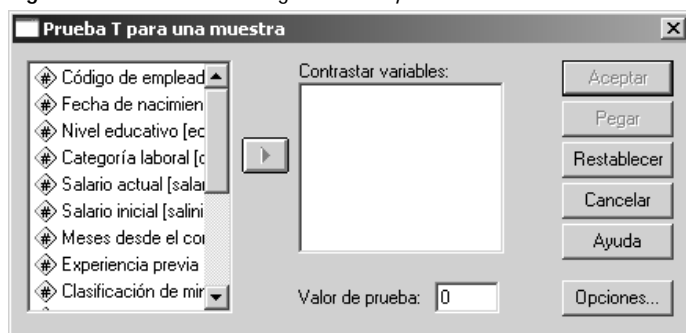
$$T = \frac{\bar{Y} - \mu}{\hat{\sigma}_{\bar{Y}}} = \frac{\bar{Y} - \mu}{S_{n-1}/\sqrt{n}}$$

distribuida según el modelo de probabilidad  $t$  de *Student* con  $n-1$  grados de libertad. Esta tipificación del estadístico  $\bar{Y}$  es lo que se conoce como *prueba  $T$  para una muestra*. Su principal utilidad radica en que permite conocer la probabilidad asociada a cada uno de los diferentes valores  $\bar{Y}$  que es posible obtener en muestras de tamaño  $n$  cuando: (1) se asume que el verdadero valor de la media poblacional es  $\mu$ ; y (2) se utiliza la media muestral  $S_{n-1}$  para estimar la desviación típica poblacional  $\sigma$ .

Para que el estadístico  $T$  se ajuste correctamente al modelo de distribución de probabilidad  $t$  de *Student*, es necesario que la población muestreada sea *normal*. No obstante, con tamaños muestrales grandes (a partir de 20 o 30 casos), el ajuste del estadístico  $T$  a la distribución  $t$  de *Student* es lo suficientemente bueno incluso con poblaciones originales sensiblemente alejadas de la normalidad.

Para contrastar hipótesis sobre una media:

- Seleccionar la opción **Comparar medias > Prueba  $T$  para una muestra...** del menú **Analizar** para acceder al cuadro de diálogo *Prueba  $T$  para una muestra* que recoge la Figura 13.3.

Figura 13.3. Cuadro de diálogo *Prueba T para una muestra*

La lista de variables contiene un listado con todas las variables *numéricas* del archivo de datos (no es posible, por tanto, utilizar variables con formato de *cadena*). Para llevar a cabo un contraste con las especificaciones que el procedimiento tiene establecidas por defecto:

- Seleccionar la variable cuya media poblacional se desea contrastar y trasladarla a la lista **Contrastar variables**.
- Introducir en **Valor de prueba** el valor poblacional concreto que se desea contrastar. Este valor se utiliza para todas las variables seleccionadas en la lista **Contrastar variables**.
- Pulsar el botón **Aceptar**.

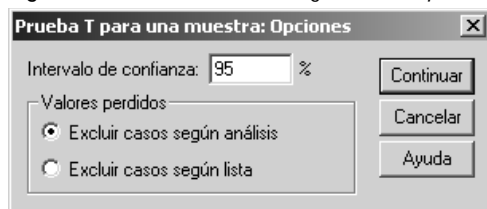
Cada variable trasladada a la lista **Contrastar variables** genera una *prueba T* acompañada de su correspondiente *nivel crítico bilateral* (el unilateral se obtiene dividiendo el bilateral por 2). El nivel crítico indica la probabilidad de obtener una media  $\bar{Y}$  tan alejada de  $\mu$  como la de hecho obtenida ( $\mu$  es el valor propuesto en **Valor de prueba**). Si esa probabilidad es pequeña (generalmente, menor que 0,05), se rechaza la hipótesis nula de que la media poblacional es el valor propuesto.

Entre los resultados que ofrece el *Visor* también se incluye el intervalo de confianza (calculado al 95 %) para la diferencia entre la media muestral  $\bar{Y}$  y el **Valor de prueba**. Los límites de confianza de este intervalo se calculan sumando y restando a la diferencia entre  $\bar{Y}$  y el **Valor de prueba** una cantidad que se obtiene multiplicando el error típico de la media ( $\sigma_{\bar{Y}}$ ) por el cuantil 97,5 de la distribución *t* de *Student* con  $n-1$  grados de libertad (es decir, multiplicando el error típico de la media por el cuantil  $100(1-\alpha/2) = 100(1-0,05/2) = 97,5$ ;  $\alpha$  se refiere al nivel de significación habitualmente utilizado).

## Opciones

Las opciones del procedimiento **Prueba T para una muestra** permiten controlar algunos aspectos del análisis. Para modificar las opciones por defecto:

- Pulsar el botón **Opciones...** del cuadro de diálogo principal (ver Figura 13.3) para acceder al subcuadro de diálogo *Prueba T para una muestra: Opciones* que recoge la Figura 13.4.

Figura 13.4. Subcuadro de diálogo *Prueba T para una muestra: Opciones*

**Intervalo de confianza:  $k$  %.** Esta opción permite establecer, en escala porcentual, el *nivel de confianza* ( $1-\alpha$ ) con el que se desea obtener el intervalo de confianza para la diferencia entre la media muestral y la media propuesta para la población en **Valor de prueba**. Un intervalo de confianza sirve para tomar una decisión sobre la misma hipótesis nula que permite contrastar el estadístico  $T$ : cuando el nivel crítico bilateral asociado al estadístico  $T$  es menor que 0,05, el intervalo calculado con una confianza del 95 % no incluye el valor cero entre sus límites. El valor de  $k$  es, por defecto, 95, pero es posible introducir con el teclado cualquier otro valor comprendido entre 0,01 y 99,99.

**Valores perdidos.** Es posible optar entre dos formas diferentes de tratar los casos con valores perdidos:

**Excluir casos según análisis.** Esta opción excluye de cada análisis (de cada prueba  $T$ ) los casos con valor perdido en la variable concreta que se está contrastando.

**Excluir casos según lista.** Esta opción excluye de todos los análisis los casos con algún valor perdido en una cualquiera de las variables seleccionadas en la lista **Contrastar variables**.

### ***Ejemplo: Comparar medias > Prueba T para una muestra***

Este ejemplo muestra cómo contrastar la hipótesis de que la media poblacional de la variable *ci* (cociente intelectual) vale 100. La variable *ci* se ha generado con la función *RV.NORMAL* del procedimiento **Calcular** (menú **Transformar**), utilizando una media de 100 y una desviación típica de 15 tras fijar la semilla aleatoria en 100 (esta variable forma parte del archivo *Datos de empleados ampliado*, el cual puede obtenerse en la página web del manual):

- En el cuadro de diálogo principal (ver Figura 13.3), seleccionar la variable *ci* y trasladarla a la lista **Contrastar variables**.
- Introducir el valor 100 en el cuadro de texto **Valor de prueba**.

Aceptando estos valores, el *Visor de resultados* ofrece la información que muestran las Tablas 13.2 y 13.3.

La Tabla 13.2 incluye el número de casos válidos sobre el que se basan los cálculos ( $N = 474$ ), la media de la variable *ci* (100,065), la desviación típica insesgada (15,128) y el error típico de la media (0,695; este error típico es el denominador de la *prueba T* y se obtiene dividiendo la desviación típica insesgada entre la raíz cuadrada del número de casos).



**Tabla 13.2.** Estadísticos descriptivos del procedimiento *Prueba T para una muestra*

	N	Media	Desviación típica	Error típ. de la media
Cociente intelectual	474	100,0653	15,12759	,69483

La Tabla 13.3 ofrece un resumen de la prueba *T* encabezado con un título que recuerda cuál es el valor propuesto para la media poblacional (*Valor de prueba* = 100) en la hipótesis nula. Las primeras columnas contienen el valor del estadístico ( $t=0,094$ ), sus grados de libertad ( $gl=473$ ) y el nivel crítico (*Significación bilateral* = 0,925). El nivel crítico indica el grado de compatibilidad existente entre el valor poblacional propuesto para la media y la información muestral disponible: si el nivel crítico es pequeño (menor que 0,05), puede concluirse que los datos son incompatibles con la hipótesis de que el verdadero valor de la media poblacional es el propuesto. En el ejemplo, el nivel crítico vale 0,925; puesto que es mayor que 0,05, no se puede rechazar la hipótesis nula. Puede concluirse, por tanto, que los datos muestrales pueden haber sido extraídos de una población con media 100.

Las siguiente columna de la Tabla 13.3 contiene la diferencia entre la media muestral y el valor de prueba (*Diferencia de medias* = 0,065). Esta diferencia es el numerador de la *prueba T*. Y a continuación aparecen los límites inferior (−1,300) y superior (1,431) del intervalo de confianza (calculado al 95 por ciento) para la diferencia entre la media muestral y el valor de prueba. Estos límites también permiten decidir sobre el valor propuesto para la media poblacional: si los límites incluyen el valor cero (como ocurre en el ejemplo), puede concluirse que los datos muestrales son compatibles con el valor poblacional propuesto y, en consecuencia, puede mantenerse  $H_0$ ; si los límites no incluyen el valor cero, debe concluirse que los datos son incompatibles con el valor propuesto y, consecuentemente, debe rechazarse  $H_0$ .

**Tabla 13.3.** Resumen del procedimiento *Prueba T para una muestra*

	Valor de prueba = 100					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% Intervalo de confianza para la diferencia	
					Límite inferior	Límite superior
Cociente intelectual	,094	473	,925	,06530	-1,3000	1,4306

Nada se ha dicho sobre el supuesto de normalidad en que se basa la prueba *T*, pero ya se ha señalado que el cumplimiento de este supuesto sólo es exigible con muestras pequeñas. Con una muestra de 474 casos, como en el ejemplo, el supuesto de normalidad carece de relevancia. No obstante, el SPSS permite contrastar la hipótesis de normalidad mediante el procedimiento *Explorar* ya estudiado en el Capítulo 11.

## Prueba *T* para muestras independientes

La prueba *T* para dos muestras independientes permite contrastar hipótesis referidas a la diferencia entre dos medias independientes. La situación típica que permite resolver esta prueba es la relativa a la comparación de dos grupos distintos de sujetos. Por ejemplo, un grupo de sujetos tratados (grupo experimental) y otro de no tratados (grupo control); o un grupo de sujetos sometido al tratamiento *A* y otro sometido al tratamiento *B*; etc.

Ahora se tienen dos poblaciones normales, con medias  $\mu_1$  y  $\mu_2$ , de las que se seleccionan sendas muestras aleatorias (de tamaños  $n_1$  y  $n_2$ ). Tras esto, se utilizan las medias muestrales  $\bar{Y}_1$  e  $\bar{Y}_2$  para contrastar la hipótesis nula de que las medias poblacionales  $\mu_1$  y  $\mu_2$  son iguales.

La prueba  $T$  que permite contrastar esta hipótesis de igualdad de medias no es otra cosa que una *tipificación de la diferencia entre las dos medias muestrales*. Esta tipificación se obtiene restando a esa diferencia su valor esperado (el propuesto en la hipótesis nula) y dividiendo el resultado entre el error típico de la diferencia:

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2}}$$

El estadístico o prueba  $T$  tiene dos versiones que difieren en la forma de estimar el error típico de la diferencia:  $\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2}$ . Si puede asumirse que las dos varianzas poblacionales son *iguales* (es decir, si puede asumirse que  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ), esa única varianza poblacional  $\sigma^2$  puede estimarse utilizando el promedio ponderado  $\hat{\sigma}^2$  de las varianzas (insesgadas) muestrales  $S_{n_1-1}^2$  y  $S_{n_2-1}^2$ . Con esta estimación *promedio* de la varianza poblacional, el error típico de la diferencia puede obtenerse como:  $\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2} = \hat{\sigma} \sqrt{1/n_1 + 1/n_2}$ . Al proceder de esta manera, el estadístico  $T$  resultante se distribuye según el modelo de probabilidad  $t$  de *Student* con  $n_1 + n_2 - 2$  grados de libertad.

Si no puede asumirse que las varianzas poblacionales son *iguales*, entonces  $\sigma_1^2$  debe estimarse mediante  $S_{n_1-1}^2$  y  $\sigma_2^2$  mediante  $S_{n_2-1}^2$ , en cuyo caso, el error típico de la diferencia puede estimarse mediante:

$$\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{S_{n_1-1}^2/n_1 + S_{n_2-1}^2/n_2}$$

Al utilizar este error típico el estadístico  $T$  sigue siendo una variable distribuida según el modelo de probabilidad  $t$  de *Student*, pero los grados de libertad de la distribución cambian. Para estimar los nuevos grados de libertad suele utilizarse una ecuación propuesta por Welch (1938):

$$gl = \left( \frac{S_{n_1-1}^2}{n_1} + \frac{S_{n_2-1}^2}{n_2} \right)^2 \left/ \left[ \left( \frac{S_{n_1-1}^2}{n_1} \right)^2 \frac{1}{(n_1-1)} + \left( \frac{S_{n_2-1}^2}{n_2} \right)^2 \frac{1}{(n_2-1)} \right] \right.$$

Para decidir si se puede o no asumir que las varianzas poblacionales son iguales, el procedimiento **Prueba  $T$  para muestras independientes**, además de incluir las dos versiones del estadístico  $T$ , ofrece la prueba de *Levene* sobre igualdad de varianzas (esta prueba ya ha sido descrito en el Capítulo 11, en el apartado *Cómo contrastar supuestos: homogeneidad de varianzas*). Se asumirán o no varianzas iguales (y, consecuentemente, se utilizará una u otra versión del estadístico  $T$ ) dependiendo de la conclusión a la que lleve la prueba de *Levene*.

Para comparar dos medias independientes:

- Seleccionar la opción **Comparar medias > Prueba  $T$  para dos muestras independientes...** del menú **Analizar** para acceder al cuadro de diálogo *Prueba  $T$  para muestras independientes* que recoge la Figura 13.5.

Figura 13.5. Cuadro de diálogo *Prueba T para muestras independientes*

La lista de variables contiene un listado con las variables *numéricas* y de *cadena corta* del archivo de datos. Para llevar a cabo un contraste con las especificaciones que el procedimiento tiene establecidas por defecto:

- Seleccionar la variable (o variables) en la que se desea comparar los grupos y trasladarla a la lista **Contrastar variables**. En el ejemplo de la Figura 13.5 se ha seleccionado la variable *salini* (salario inicial).  
Todas las variables trasladadas a esta lista deben tener formato numérico. Cada variable seleccionada genera una prueba *T* acompañada de su nivel crítico y del intervalo de confianza para la diferencia entre las dos medias comparadas.
- Seleccionar la variable que define los grupos que se desea comparar y trasladarla al cuadro **Variable de agrupación**. Esta variable puede tener formato numérico o de cadena corta. En el ejemplo de la Figura 13.5 se ha seleccionado la variable *sexo*, la cual define dos grupos: *h* = «hombres» y *m* = «mujeres».

**Definir grupos.** Tras seleccionar una variable de agrupación aparecen junto a ella dos interrogantes entre paréntesis que avisan de la necesidad de informar al procedimiento sobre los códigos (valores) que identifican a los dos grupos que se desea comparar (es necesario indicar estos códigos incluso aunque la variable de agrupación seleccionada sólo tenga dos valores). Para definir estos códigos:

- Pulsar el botón **Definir grupos...** para acceder al subcuadro de diálogo *Definir grupos* que muestra la Figura 13.6.

Figura 13.6. Subcuadro de diálogo *Definir grupos*

**Usar valores especificados.** Si la variable de agrupación es categórica, los códigos que definen los dos grupos que se desea comparar deben introducirse en los cuadros de texto **Grupo 1** y **Grupo 2**. Los casos que posean otros códigos serán excluidos del análisis. Si la **Variable de agrupación** es una variable con formato de cadena, es muy importante tener en cuenta que los códigos deben introducirse de forma exacta (vigilando la presencia de espacios, las mayúsculas y minúsculas, etc.).

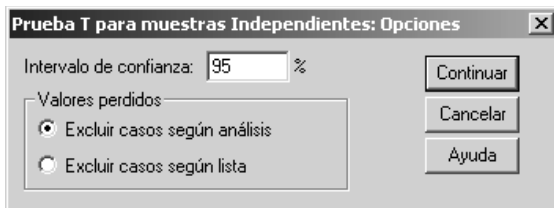
**Punto de corte.** Si se desea utilizar como variable de agrupación una variable cuantitativa *continua*, esta opción permite introducir un valor como punto de corte para definir dos grupos: los casos con puntuación igual o mayor que el punto de corte forman un grupo; el resto de los casos forman el otro grupo. Esta opción no está disponible si, como variable de agrupación, se elige una variable con formato de cadena.

## Opciones

Las opciones del cuadro de diálogo *Prueba T para muestras independientes* permiten controlar dos aspectos del análisis: el nivel de confianza con el que se desea construir los intervalos de confianza y el tratamiento que se desea dar a los valores perdidos. Para modificar las opciones por defecto:

- Pulsar el botón **Opciones...** del cuadro de diálogo principal (ver Figura 13.5) para acceder al subcuadro de diálogo *Prueba T para muestras independientes: Opciones* que muestra la Figura 13.7.

Figura 13.7. Subcuadro de diálogo *Prueba T para muestras independientes: Opciones*



**Intervalo de confianza:**  $k\%$ . Esta opción permite establecer, en escala porcentual, el *nivel de confianza*  $(1-\alpha)$  con el que se desea obtener el intervalo de confianza para la diferencia de medias. El valor de  $k$  es, por defecto, 95, pero es posible introducir cualquier otro valor comprendido entre 0,01 y 99,99.

**Valores perdidos.** Es posible optar entre dos formas diferentes de tratar los casos con valores perdidos:

**Excluir casos según análisis.** Esta opción excluye de cada análisis (de cada prueba  $T$ ) los casos con valor perdido en la variable de agrupación o en la variable que se está contrastando en ese análisis.

**Excluir casos según lista.** Esta opción excluye de todos los análisis (de todas las pruebas  $T$  solicitadas) los casos con algún valor perdido en la variable de agrupación o en una cualquiera de las variables incluidas en la lista **Contrastar variables**.

### Ejemplo: Comparar medias > Prueba T para muestras independientes

Este ejemplo muestra cómo contrastar hipótesis sobre dos medias poblacionales independientes utilizando el procedimiento **Prueba T para muestras independientes**. En concreto, muestra cómo comparar las medias de la variable *salini* (salario inicial) en los dos grupos definidos por la variable *sexo* (*h*=«hombres» y *m*=«mujeres»). Los datos pertenecen al archivo *Datos de empleados*, el cual se encuentra en la misma carpeta en la que está instalado el SPSS:

- En el cuadro de diálogo principal (ver Figura 13.5), trasladar la variable *salini* a la lista **Contrastar variables** y la variable *sexo* al cuadro de selección **Variable de agrupación**.
- Pulsar el botón **Definir grupos...** para acceder al subcuadro de diálogo *Definir grupos* (ver Figura 13.6) e introducir los códigos *h* y *m* en los cuadros de texto **Grupo 1** y **Grupo 2** (dado que la variable *sexo* es una variable de cadena, debe ponerse especial cuidado en introducir los códigos *h* y *m* con minúscula y sin espacios en blanco delante o detrás del código). Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestran las Tablas 13.4 y 13.5 (las cabeceras de las columnas de la Tabla 13.5 se han rotado para ajustar el tamaño de la tabla a las dimensiones de la página).

La Tabla 13.4 muestra la variable que está siendo contrastada (*salario inicial*) y los dos grupos que se están comparando (*hombre*, *mujer*). Para cada grupo, la tabla informa sobre el número de casos válidos, la media y la desviación típica insesgada del salario inicial, y el error típico de la media del salario inicial.

**Tabla 13.4.** Estadísticos descriptivos del procedimiento *Prueba T para muestras independientes*

Salario inicial				
Sexo	N	Media	Desviación típica	Error típ. de la media
Hombre	258	\$20,301.40	\$9,111.78	\$567.27
Mujer	216	\$13,091.97	\$2,935.60	\$199.74

La Tabla 13.5 ofrece, en primer lugar, el contraste de Levene (*F*) sobre homogeneidad o igualdad de varianzas. El resultado de este contraste es el que permite decidir si se puede o no asumir que las varianzas poblacionales son iguales: si la probabilidad asociada al estadístico de Levene es mayor que 0,05, puede asumirse que las varianzas poblacionales son iguales; si esa probabilidad es menor que 0,05, debe rechazarse la hipótesis de igualdad de varianzas y asumirse que son distintas.

Las columnas siguientes contienen el estadístico *t*, sus grados de libertad (*gl*), el nivel crítico bilateral (*Significación bilateral*; el nivel crítico unilateral se obtiene dividiendo el bilateral entre 2)), la diferencia observada entre los promedios salariales de ambos grupos, el error típico de esa diferencia, y los límites inferior y superior del intervalo de confianza (calculado al 95%) de la diferencia entre ambas medias. Toda esta información está calculada tanto para el caso de varianzas poblacionales iguales (línea encabezada *Asumiendo varianzas iguales*) como para el caso de varianzas poblacionales distintas (línea encabezada *No asumiendo varianzas iguales*).

En el ejemplo, dado que la probabilidad asociada al estadístico de Levene es muy pequeña ( $\text{Sig.} < 0,0005$ ), debe rechazarse la hipótesis de igualdad de varianzas y, consecuentemente, debe utilizarse la información de la fila encabezada *No asumiendo varianzas iguales*: el estadístico  $t$  toma el valor 11,987 y tiene asociado un nivel crítico bilateral menor que 0,0005. Este valor es justamente el que informa sobre el grado de compatibilidad existente entre la diferencia observada entre las medias muestrales de los grupos comprados y la hipótesis nula de que las medias poblacionales son iguales. Puesto que el nivel crítico es menor que 0,05, puede afirmarse que los datos muestrales son incompatibles con la hipótesis nula de igualdad de medias. Por tanto, se puede rechazar la hipótesis nula y concluir que el salario poblacional medio de los hombres y de las mujeres no es el mismo.

Los límites del intervalo de confianza permiten estimar que la verdadera diferencia entre el salario medio de la población de hombres y el salario medio de la población de mujeres se encuentra entre 6.026,19 y 8.392,67 dólares. El hecho de que el intervalo obtenido no incluya el valor cero también permite rechazar la hipótesis de igualdad de medias.

**Tabla 13.5.** Resumen del procedimiento *Prueba T para muestras independientes*

Salario inicial	Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error tip. de la diferencia	95% Intervalo de confianza para la diferencia	
								Inferior	Superior
Asumiendo varianzas iguales	105.969	.000	11.15	472	.000	\$7,209.43	\$646.45	\$5,939.16	\$8,479.70
No asumiendo varianzas iguales			11.99	319	.000	\$7,209.43	\$601.41	\$6,026.19	\$8,392.67

Los diagramas de caja de la Figura 13.8 ofrecen información descriptiva sobre el comportamiento de la variable *salini* en los dos grupos comparados. Los diagramas de caja muestran con claridad que el salario inicial de los hombres tiene mayor promedio (mayor mediana) y mayor dispersión (caja más alta y bigotes más largos) que el salario inicial de las mujeres. También muestran con claridad que el salario de los hombres es más disperso que el de las mujeres. Este gráfico se ha obtenido con la opción **Diagramas de caja > Simple** del menú **Gráficos**.

La Figura 13.9 muestra la posición del salario inicial medio de cada grupo junto con los límites del intervalo de confianza (95%) que corresponden a esos promedios. Aunque el intervalo de confianza para la diferencia entre dos medias (intervalo que ofrece la Tabla 13.5) no ofrece exactamente la misma información que los intervalos individuales para las dos medias individualmente consideradas, lo cierto es que los límites individuales de la Figura 13.9 permiten apreciar la enorme distancia existente entre los dos promedios comparados. Este gráfico se ha obtenido con la opción **Barras de error > Simple** del menú **Gráficos**.

Figura 13.8. Diagramas de caja de *salario inicial* en *hombres* y *mujeres*

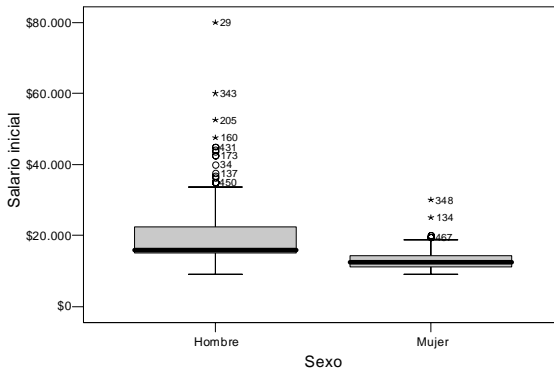
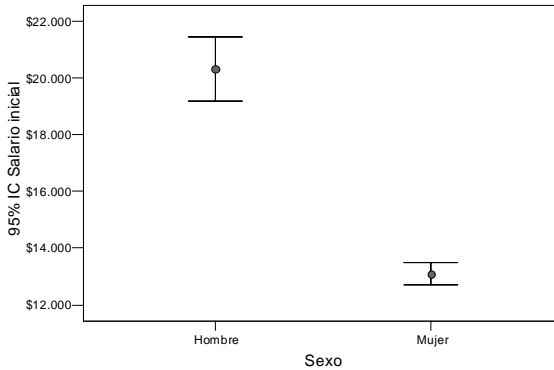


Figura 13.9. Barras de error de *salario inicial* en *hombres* y *mujeres*



## Prueba $T$ para muestras relacionadas

La prueba  $T$  para dos muestras relacionadas permite contrastar hipótesis referidas a la diferencia entre dos medias relacionadas. Ahora se dispone de una población de diferencias con media  $\mu_D$ , obtenida al restar las puntuaciones del mismo grupo de casos en dos variables diferentes o en la misma variable medida en dos momentos diferentes (de ahí que se hable de muestras relacionadas). De esa población de diferencias se extrae una muestra aleatoria de tamaño  $n$  y se utiliza la media  $\bar{Y}_D$  de esas  $n$  diferencias para contrastar la hipótesis nula de que la media  $\mu_D$  de la población de diferencias vale cero.

Una situación típica que permite resolver este contraste es la que se da en los diseños *antes-después* o *pre-post*: a una muestra de  $n$  sujetos se le toma una medida *pre*-tratamiento (línea base), se les aplica el tratamiento y se les vuelve a tomar una medida *post*-tratamiento para evaluar si se ha producido algún cambio.

Desde el punto de vista estadístico, este contraste es idéntico al presentado en el apartado *Prueba  $T$  para una muestra*. La única diferencia existente entre ambos contrastes es que allí

se tenía una muestra de puntuaciones obtenida al medir una sola variable y ahora se tienen dos muestras relacionadas (o una muestra de *pares* de puntuaciones) que se convierte en una sola muestra de *diferencias* restando las puntuaciones de cada par.

El estadístico o prueba  $T$  sigue siendo una tipificación de la media muestral de las diferencias  $\bar{Y}_D$ :

$$T = \frac{\bar{Y}_D - \mu_D}{\hat{\sigma}_{\bar{Y}_D}} = \frac{\bar{Y}_D - \mu_D}{S_D / \sqrt{n}}$$

( $S_D$  se refiere a la desviación típica insesgada de las  $n$  diferencias). Este estadístico  $T$  se distribuye según el modelo de probabilidad  $t$  de *Student*, con  $n-1$  grados de libertad; por tanto, permite conocer la probabilidad asociada a los diferentes valores  $\bar{Y}_D$  que es posible obtener en muestras aleatorias de tamaño  $n$ .

Al igual que antes, para que el valor  $T$  se ajuste apropiadamente al modelo de distribución de probabilidad  $t$  de *Student*, es necesario que la población de diferencias sea *normal*. No obstante, con tamaños muestrales grandes el ajuste del estadístico  $T$  a la distribución  $t$  de *Student* es lo suficientemente bueno incluso con poblaciones originales sensiblemente alejadas de la normalidad.

Para contrastar hipótesis sobre dos medias relacionadas:

- Seleccionar la opción **Comparar medias > Prueba  $T$  para muestras relacionadas...** del menú **Analizar** para acceder al cuadro de diálogo *Prueba  $T$  para muestras relacionadas* que recoge la Figura 13.10.

Figura 13.10. Cuadro de diálogo *Prueba  $T$  para muestras relacionadas*



Lógicamente, la lista de variables sólo contiene variables con formato numérico. Para llevar a cabo un contraste sobre dos medias relacionadas con las especificaciones que el programa tiene establecidas por defecto:

- Seleccionar en la lista de variables del archivo de datos las dos variables cuyas medias se desea comparar y trasladarlas a la lista **Variables relacionadas**.



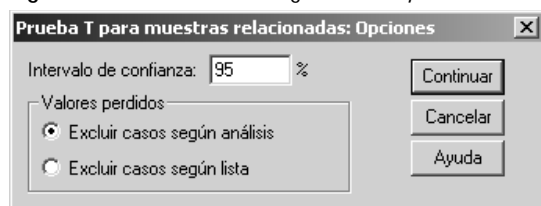
Las variables deben trasladarse a esta lista *por pares*. Es decir, es necesario marcar dos variables de la lista de variables para que el botón flecha esté activo. Pueden trasladarse tantos pares de variables como se desee. Una vez seleccionado uno o más pares de variables, el botón **Aceptar** genera tantas pruebas *T* como pares seleccionados (incluyendo los correspondientes niveles críticos bilaterales y los intervalos de confianza para las diferencias).

## Opciones

Las opciones del cuadro de diálogo *Prueba T para muestras relacionadas* permiten controlar el nivel de confianza con el que se desea trabajar y el tratamiento que se desea dar a los casos con valores perdidos. Para modificar estas opciones:

- Pulsar el botón **Opciones...** del cuadro de diálogo principal (ver Figura 13.10) para acceder al subcuadro de diálogo *Prueba T para muestras relacionadas: Opciones* que ofrece la Figura 13.11.

Figura 13.11. Cuadro de diálogo *Prueba T para muestras relacionadas: Opciones*



**Intervalo de confianza:**  $k\%$ . Esta opción permite establecer, en escala porcentual, el *nivel de confianza*  $(1 - \alpha)$  con el que se desea obtener el intervalo de confianza para la diferencia de medias. El valor de  $k$  es, por defecto, 95, pero puede introducirse cualquier otro valor comprendido entre 0,01 y 99,99.

**Valores perdidos.** Hay dos formas diferentes de tratar los casos con valores perdidos:

**Excluir casos según análisis.** Esta opción excluye de cada contraste los casos con valor perdido en cualquiera de las dos variables que están siendo contrastadas.

**Excluir casos según lista.** Esta opción excluye de todos los contrastes (de todas las pruebas *T* solicitadas) los casos con algún valor perdido en cualquiera de las variables seleccionadas en la lista **Variables relacionadas**.

## Ejemplo: Comparar medias > Prueba T para muestras relacionadas

Este ejemplo muestra cómo contrastar hipótesis sobre dos medias poblacionales utilizando el procedimiento *Prueba T para muestras relacionadas*. En concreto, muestra cómo comparar las medias de las variables *tiempemp* (meses desde el contrato) y *expprev* (experiencia previa) para averiguar si el promedio de meses que los empleados llevan contratados difiere o no del promedio de experiencia previa con la que fueron contratados. Los datos pertenecen al archi-

vo *Datos de empleados*, que se encuentra en la misma carpeta en la que está instalado el SPSS. Para efectuar el contraste:

- En el cuadro de diálogo principal (ver Figura 13.10), seleccionar las variables *tiempemp* y *expprev* y trasladarlas a la lista **Variables relacionadas**.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 13.6 a la 13.8. La Tabla 13.6 muestra, para cada variable, la media, el número de casos válidos, la desviación típica insesgada y el error típico de la media (la desviación típica dividida por la raíz cuadrada del número de casos).

**Tabla 13.6.** Estadísticos descriptivos del procedimiento *Prueba T para muestras relacionadas*

		Media	N	Desviación típica	Error típ. de la media
Par 1	Meses desde el contrato	81.11	474	10.06	.46
	Experiencia previa	95.86	474	104.59	4.80

La Tabla 13.7 contiene el coeficiente de correlación de Pearson entre ambas variables junto con el nivel crítico bilateral que le corresponde bajo la hipótesis de independencia (el coeficiente de correlación de Pearson se estudia en el Capítulo 17).

**Tabla 13.7.** Coeficiente de correlación de *Pearson*

	N	Correlación	Sig.
Par 1 Meses desde el contrato y Experiencia previa	474	.003	.948

La Tabla 13.8 incluye, en la primera mitad, tres estadísticos referidos a las *diferencias* entre cada par de puntuaciones: la media, la desviación típica y el error típico de la media. La siguiente columna contiene el intervalo de confianza para la diferencia entre las medias: puede estimarse, con una confianza del 95 por ciento, que la verdadera diferencia entre las medias de *meses de contrato* y *experiencia previa* se encuentra entre 5,27 y 24,2 meses (a favor de *experiencia previa*).

La segunda mitad de la tabla informa sobre el valor del estadístico *t*, sus grados de libertad (*gl*) y el nivel crítico bilateral (*Sig. bilateral*; el unilateral se obtiene dividiendo el bilateral entre 2). Puesto que el valor del nivel crítico es muy pequeño (0,002), se puede rechazar la hipótesis de igualdad de medias y concluir que el promedio de *meses de contrato* es significativamente menor que el promedio de meses de *experiencia previa*.

**Tabla 13.8.** Resumen del procedimiento *Prueba T para muestras relacionadas*

		Diferencias relacionadas				t	gl	Sig. (bilateral)
		Media	Desv. típica	Error típ. de la media	95% Intervalo de confianza para la diferencia			
					Inferior Superior			
Par 1	Meses desde el contrato - Experiencia previa	-14.75	105.04	4.82	-24.2 -5.27	-3.057	473	.002

Los gráficos de las Figuras 13.12 y 13.13 pueden ayudar a precisar los resultados obtenidos. De acuerdo con los diagramas de caja (Figura 13.12), aunque las medianas de ambas variables no son muy distintas, la dispersión de la variable *experiencia previa* es mucho mayor que la de *meses de contrato* y su distribución muestra una acusada asimetría positiva que no se da en *meses de contrato*. Esto último implica que la media de *experiencia previa* está muy inflada (razón por la cual las medias son mucho más distintas que las medianas).

Las barras de error de la Figura 13.13 reflejan con claridad esta circunstancia: las medias son muy distintas y los intervalos no se solapan. Cuando se dan estas circunstancias debería considerarse la posibilidad de comparar las medianas en lugar de las medias. En el Capítulo 21 se ofrecen procedimientos *no paramétricos* que constituyen una excelente alternativa cuando los datos no reúnen las condiciones apropiadas para la aplicación de la prueba *T*. Los diagramas de caja se han obtenido con la opción **Diagramas de caja...** > **Simple (Resúmenes para distintas variables)** del menú **Gráficos**. Las barras de error se han obtenido con la opción **Barras de error...** > **Simple (Resúmenes para distintas variables)** del menú **Gráficos**.

Figura 13.12. Diagramas de caja de *meses de contrato* y *experiencia previa*

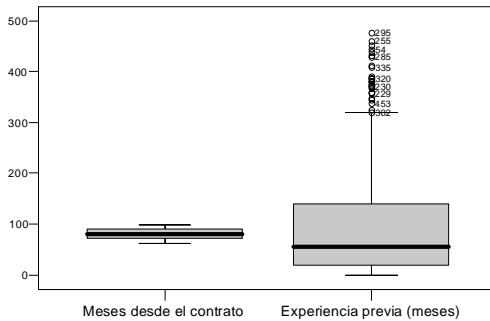
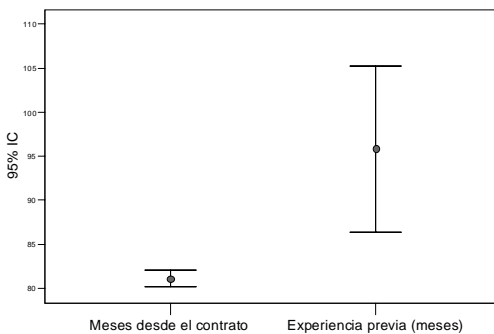


Figura 13.13. Barras de error de *meses de contrato* y *experiencia previa*



## Análisis de varianza (I)

### El procedimiento *ANOVA de un factor*

El análisis de varianza o ANOVA (del inglés **AN**alysis **Of** **V**ariance) no es realmente una única técnica de análisis, sino una familia de técnicas que comparten el objetivo de ayudar al investigador a formular modelos para interpretar los datos obtenidos en un estudio empírico. Estos modelos permiten explicar el comportamiento de una variable dependiente o respuesta (cuantitativa) a partir de una o más variables independientes o predictoras (categóricas). También permiten controlar el efecto de variables extrañas (variables ajenas al experimento) introduciéndolas como covariables.

Todas estas técnicas de análisis agrupadas bajo la denominación de ANOVA se basan en una estructura matemática relativamente simple conocida como *modelo lineal general*. Este capítulo incluye una breve descripción del modelo lineal general, así como una clasificación de los diferentes modelos de ANOVA. También incluye una explicación del modelo de un factor. En los siguientes capítulos se abordan los modelos factoriales (Capítulo 15) y los modelos de medidas repetidas (Capítulo 16).

### El modelo lineal general

En el contexto del ANOVA, un *modelo* es una afirmación algebraica (una ecuación matemática) acerca de cómo se relacionan dos o más variables. Por supuesto, existen muchas clases diferentes de formulaciones algebraicas o modelos capaces de representar la relación entre dos o más variables, pero el más simple y flexible de todos ellos se conoce con el nombre de *modelo lineal general*.

En esencia, un modelo lineal intenta describir una variable dependiente como el resultado de la suma ponderada de varios efectos. Ahora bien, las variables sometidas a estudio dependen de multitud de factores diferentes. Cuando un sujeto obtiene una puntuación en una variable cualquiera, es realista pensar que los factores (causas) que han determinado esa puntuación son numerosos y variados. Y también es realista pensar que en una investigación concreta sólo será posible manipular y medir un número reducido de las múltiples causas atribuibles a una variable cualquiera.

Estas ideas dan pie para formular la estructura de los modelos lineales, en su nivel más elemental, según muestra la Figura 14.1.

Figura 14.1. Estructura básica de un modelo lineal

$$\boxed{\text{valor observado en la variable dependiente}} = \boxed{\text{efecto debido a factores tenidos en cuenta}} + \boxed{\text{efecto debido a factores no tenidos en cuenta}}$$

De acuerdo con esta expresión, un modelo lineal intenta describir el valor observado en una variable dependiente recurriendo a: (1) un conjunto de efectos atribuibles a factores *tenidos en cuenta* (es decir, a factores explícitamente incluidos en el modelo) y (2) un conjunto de efectos atribuibles a factores *no tenidos en cuenta*. Los factores tenidos en cuenta se refieren a las variables que el investigador incluye en el experimento para estudiar su efecto sobre la variable dependiente. Los factores no tenidos en cuenta se refieren a variables cuyo efecto, aun pudiendo ser importante para describir la variable dependiente, no interesa estudiar de forma inmediata o no resulta posible hacerlo.

Sobre estos factores no tenidos en cuenta el investigador puede decidir ejercer o no algún tipo de control. Puede ejercerse control sobre una variable manteniéndola *constante*: seleccionando sujetos de la misma edad se puede controlar el efecto de la edad; utilizando las mismas condiciones ambientales se puede controlar el efecto del contexto; etc. Sobre otras variables no se ejerce control, bien por que no se desea (en un estudio sobre rendimiento, la inteligencia es una variable importante, pero el investigador puede no estar interesado en controlar su efecto, es decir, puede decidir utilizar sujetos con diferentes niveles de inteligencia, simplemente porque desea que sus resultados posean más generalidad), bien porque no resulta posible hacerlo (la historia individual de cada sujeto, por ejemplo, es algo en lo que los sujetos claramente difieren pero sobre lo que un investigador no tiene ningún tipo de control). Todas las variables no controladas son las responsables de la parte de la variable dependiente que no es capaz de describir el conjunto de variables controladas; constituyen, por tanto, *aquello que escapa al investigador*, razón por la cual se suele utilizar el término *error* para caracterizar al conjunto de efectos debidos a las variables no sujetas a control.

Estas consideraciones permiten retocar la primera formulación del modelo lineal expuesta en la Figura 14.1. La Figura 14.2 muestra estos retoques.

Figura 14.2. Estructura básica de un modelo lineal: efectos debidos a factores tenidos en cuenta desglosados

$$\boxed{\text{valor observado en la variable dependiente}} = \boxed{\text{efecto debido a factores constantes}} + \boxed{\text{efecto debido a factores tenidos en cuenta}} + \boxed{\text{efecto debido a factores no controlados (error)}}$$

Un ejemplo concreto puede ayudar a entender mejor la estructura de un modelo lineal. Entre los muchos factores de los que parece depender el rendimiento académico, en un estudio concreto puede interesar evaluar el efecto de dos variables: el nivel cultural de los padres y el cociente intelectual de los estudiantes. Si se formula este interés en términos de un modelo lineal (es decir, según la estructura de la Figura 14.2) se obtiene como resultado el modelo propuesto en la Figura 14.3.

Figura 14.3. Estructura básica de un modelo lineal: dos variables independientes

$$\boxed{\text{puntuación obtenida en la variable rendimiento}} = \boxed{\text{puntuación media en rendimiento (común a todos los sujetos)}} + \boxed{\text{efecto del nivel cultural de los padres + efecto del CI}} + \boxed{\text{efecto debido a factores no controlados (error)}}$$

Ahora puede darse un paso más e intentar formular matemáticamente el modelo propuesto en la Figura 14.3:

$$Y_i = \beta_0 X_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

$Y_i$  representa la puntuación obtenida en la variable dependiente por el sujeto  $i$  (el subíndice  $i$  se refiere, por tanto, a cada uno de los sujetos;  $i = 1, 2, \dots, n$ );  $X_{i0}$ ,  $X_{i1}$  y  $X_{i2}$  son los diferentes factores tenidos en cuenta en el modelo a la hora de intentar explicar el comportamiento de la variable dependiente; y  $\beta_0$ ,  $\beta_1$  y  $\beta_2$  son valores desconocidos (llamados *parámetros*) que es necesario estimar y que informan sobre la importancia de cada uno de los factores presentes en la ecuación. El primer término de la ecuación ( $\beta_0 X_{i0}$ ) recoge el conjunto de efectos debidos a los factores mantenidos *constantes*, es decir, aquellos factores que son comunes a todos los sujetos:  $X_{i0}$  suele tomar el valor 1 para todos los sujetos (lo que significa que todos los sujetos puntúan igual en los factores que se mantienen constantes) y  $\beta_0$  es, generalmente, la media poblacional (que es justamente la parte de la variable dependiente que es común a todos los sujetos). El término final ( $\epsilon_i$ ) representa el efecto debido al conjunto de factores no tenidos en cuenta y que se supone que varían aleatoriamente: refleja la diferencia existente entre la realidad y las predicciones que se derivan del modelo.

Si en lugar de dos factores tenidos en cuenta ( $X_{i1}$  y  $X_{i2}$ ) el modelo incluye cualquier número de factores (por ejemplo,  $p$ ), se llega a la formulación del modelo lineal general en su forma estándar:

$$Y_i = \beta_0 X_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \epsilon_i$$

Este modelo posee una gran utilidad; a pesar de su simplicidad, es lo bastante general para dar fundamento a gran parte de las técnicas de análisis de datos utilizadas en la investigación empírica. Y, desde luego, sirve para dar fundamento a los diferentes modelos de análisis de varianza que se estudian en este y en los próximos capítulos. Ahora bien, su formulación no siempre adopta el mismo formato: según se verá, cada situación analítica concreta requiere una reformulación particular del modelo general.

## Introducción al análisis de varianza

En esencia, los modelos de ANOVA sirven para analizar los datos provenientes de diseños con una o más variables *independientes* o *factores* (variables categóricas nominales u ordinales) y una variable *dependiente* o *respuesta* (variable cuantitativa medida con una escala de intervalo o razón). A diferencia de lo que ocurre con otras concreciones del modelo lineal general, los modelos de ANOVA permiten, básicamente, comparar medias.

### Modelos de ANOVA

Aunque existen muchos y muy diferentes modelos de ANOVA, puede obtenerse una clasificación bastante simple de los mismos atendiendo a unos pocos criterios. Tres criterios bastan para clasificar los modelos de ANOVA: (1) el número de factores, (2) el tipo de aleatorización utilizada y (3) el tipo de muestreo efectuado sobre los niveles de los factores.

## Número de factores

El término *factor* en el contexto del ANOVA es sinónimo de *variable independiente*. Así, al modelo de ANOVA diseñado para analizar los datos obtenidos utilizando un diseño con *una variable independiente* se le llama ANOVA de *un factor* (en inglés, *one-way ANOVA*). Si el diseño consta de *dos variables independientes*, al modelo de ANOVA que permite analizar los datos se le llama ANOVA de *dos factores* (*two-way ANOVA*). Etc. A los modelos de más de un factor se les llama modelos *factoriales*.

El **modelo de un factor** incluye, entre los efectos debidos a factores tenidos en cuenta, el término constante ( $\mu$ ) y el término referido a la variable independiente o factor ( $\alpha_j$ ). Y, por supuesto, también incluye los efectos debidos a factores no tenidos en cuenta ( $\epsilon_{ij}$ ):

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

( $i = 1, 2, \dots, n$ , se refiere a los sujetos;  $j = 1, 2, \dots, J$ , se refiere a los niveles del factor). Este modelo establece que la puntuación obtenida por el sujeto  $i$  bajo el nivel  $j$  puede ser explicada recurriendo a tres componentes que se combinan de forma aditiva: la *media total* de la variable dependiente en la población (que es la parte común a todas las puntuaciones  $Y_{ij}$ ), el efecto atribuible a la variable independiente (es decir, el efecto atribuible al nivel del factor bajo el cual se obtiene esa puntuación  $Y_{ij}$ ) y el efecto atribuible al conjunto de variables no contempladas en el modelo (variables extrañas cuyo efecto es desconocido; variables cuyo efecto, aun siendo conocido, no se desea estudiar; errores de medida; etc.).

El conjunto de efectos debidos a factores tenidos en cuenta constituye la parte del modelo que se utiliza para efectuar los pronósticos:

$$\mu_j = \mu + \alpha_j$$

El modelo ofrece, por tanto, un único pronóstico para cada nivel del factor (todos los casos agrupados bajo el mismo nivel del factor reciben el mismo pronóstico). En consecuencia, el efecto  $\alpha_j$  asociado a cada nivel del factor se interpreta como la diferencia entre la media de ese nivel y la media total (pues lo que cada nivel del factor tiene de específico es justamente aquello en lo que se diferencia del promedio total  $\mu$ ):

$$\alpha_j = \mu_j - \mu$$

Y los errores o residuos (el término que recoge los efectos debidos a los factores no tenidos en cuenta) se definen como las diferencias existentes entre los valores observados y los pronósticos del modelo:

$$\epsilon_{ij} = Y_{ij} - \mu_j$$

El **modelo de dos factores** incluye, entre los efectos debidos a factores tenidos en cuenta, el término constante ( $\mu$ ), el término referido a la primera variable independiente o primer factor ( $\alpha_j$ ), el término referido a la segunda variable independiente o segundo factor ( $\beta_k$ ) y el término referido a la interacción entre ambas variables independientes o factores ( $\alpha\beta_{jk}$ ). Y, al igual que cualquier otro modelo, también incluye los efectos debidos a factores no tenidos en cuenta ( $\epsilon_{ijk}$ ):

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \epsilon_{ijk}$$

( $k = 1, 2, \dots, K$ , se refiere a los niveles del segundo factor). Siguiendo la misma lógica que en el modelo de un factor, el efecto  $\beta_k$  asociado a cada nivel del segundo factor se interpreta como la diferencia entre la media de ese nivel y la media total:  $\beta_k = \mu_k - \mu$ . Y el efecto asociado a cada combinación entre los niveles de los factores (interacción) se define como la desviación de la media de esa combinación respecto de sus medias marginales:

$$\alpha\beta_{jk} = \mu_{jk} - \mu_j - \mu_k + \mu$$

En los próximos capítulos se tendrá ocasión de estudiar con algún detalle la formulación concreta que adopta el modelo lineal general en los diferentes modelos de ANOVA.

### Tipo de aleatorización

*Aleatorización* es el término utilizado para denominar el proceso consistente en asignar aleatoriamente (es decir, al azar) las unidades experimentales (generalmente sujetos) a cada uno de los niveles del factor. Con la aleatorización se intenta garantizar que todos los sujetos tengan la misma probabilidad de pertenecer a cada uno de los niveles del factor. Se pretende con ello que el conjunto de posibles variables extrañas asociadas a las características personales de los sujetos queden distribuidas de forma similar en todos los niveles del factor.

La aleatorización se puede llevar a cabo de diferentes formas. Si se realiza sobre cada uno de los sujetos, se tiene un ANOVA *completamente aleatorizado*: cada sujeto, uno a uno, es asignado al azar a cada uno de los niveles del factor. Supongamos que interesa establecer la cantidad de fármaco idónea para reducir el insomnio de determinado tipo de pacientes. Se tiene una variable independiente o factor (*cantidad de fármaco*) en la que se han definido cuatro niveles: 0 mg, 100 mg, 250 mg, 500 mg. Y una variable dependiente (*insomnio*) de la que se puede obtener una medida cuantitativa. Para determinar el efecto del fármaco sobre el insomnio se puede comenzar seleccionando una muestra aleatoria de, por ejemplo,  $N=40$  pacientes. Después se pueden formar 4 grupos de sujetos, de tamaños  $n_1, n_2, n_3$  y  $n_4$ , asignando al azar cada uno de los 40 sujetos a uno de los 4 grupos. Por último, se puede asignar, aleatoriamente también, cada grupo a uno de los cuatro niveles del factor. Procediendo de esta manera se habrá utilizado un diseño completamente aleatorizado.

Si se sospecha que existe alguna variable extraña que puede alterar de forma apreciable las conclusiones del experimento, se puede ejercer sobre ella un control directo modificando el tipo de aleatorización. Supongamos que el fármaco cuyo efecto sobre el insomnio se desea establecer posee la peculiaridad de afectar de forma diferenciada a los pacientes dependiendo del grado de insomnio que padecen. Se puede controlar ese efecto formando *bloques*: si se clasifica a los 40 sujetos de nuestra muestra como pacientes con insomnio *severo*, *moderado* o *leve* (tres bloques)\* y, tras esto, los sujetos de cada bloque se asignan a cada uno de los niveles del factor, se habrá conseguido que dentro de cada nivel haya tanto pacientes con insomnio severo, como pacientes con insomnio moderado y pacientes con insomnio leve: el efecto de la variable extraña habrá quedado controlado al estar todos los grupos *iguales* en grado de insomnio. Procediendo de esta manera se tiene un diseño de *bloques aleatorios*. Y el mo-

---

\* Aunque en este ejemplo concreto se han establecido 3 bloques, el número de bloques que pueden formarse es arbitrario. Oscila entre un mínimo de 2 (o formamos al menos 2 bloques o no formamos ninguno) y un máximo de  $N/k$ , siendo  $N$  el tamaño de la muestra y  $k$  el número de niveles del factor.



delo de ANOVA que permite analizar los datos así obtenidos recibe el nombre de ANOVA de un factor *aleatorizado en bloques* o con *bloques aleatorios*.

Un caso extremo de bloqueo es aquel en el que cada bloque está formado por un único sujeto: a todos y cada uno de los sujetos se le aplican todos y cada uno de los niveles de la variable independiente o factor. La homogeneidad dentro de cada bloque es máxima (y por tanto mínima la presencia de variables extrañas atribuibles a diferencias entre los sujetos) porque todas las puntuaciones dentro de un mismo bloque pertenecen a un mismo sujeto. En este caso ya no se habla de diseño de *bloques*, sino de diseño *intra-sujetos*; y al modelo de ANOVA que permite analizar estos datos se le llama ANOVA de *medidas repetidas*.

Estas distinciones basadas en el concepto de aleatorización son equivalentes a las que se establecen al hablar de muestras independientes y muestras relacionadas: hablar de diseños completamente aleatorizados es equivalente a hablar de muestras independientes (a cada nivel del factor se asigna un grupo distinto de sujetos); y hablar de diseños de bloques aleatorios o de diseños intra-sujetos es equivalente a hablar de muestras relacionadas (bien porque los sujetos de un mismo bloque han sido igualados –*emparejados*– atendiendo a algún criterio, bien porque todos los grupos están formados por los mismos sujetos).

## Muestreo de niveles

En los diseños experimentales un factor es, en general, una variable *controlada* por el propio experimentador. El número de valores (niveles) que toma dependen, normalmente, de los intereses del investigador. En el ejemplo del apartado anterior se han definido 4 niveles de fármaco, pero igualmente se podrían haber definido 3, o 5, o cualquier otro número. Estos niveles pueden establecerse de dos formas diferentes: (1) *fijando* sólo aquellos niveles del factor que realmente interesa estudiar, o (2) *seleccionando* aleatoriamente un conjunto de niveles entre todos los posibles niveles del factor.

Si se establecen, por ejemplo, 4 niveles de fármaco (0 mg, 100 mg, 250 mg y 500 mg) porque esos niveles de fármaco son justamente los que interesa estudiar, entonces el modelo de ANOVA es de *efectos fijos* (también llamado *modelo I*). Los niveles que interesa estudiar son justamente esos 4. De modo que, si se replicara el experimento, aunque los sujetos serían diferentes, los niveles del factor serían exactamente los mismos. Cuando se utiliza un factor de efectos fijos, el propósito del análisis consiste en determinar si los niveles concretos que se están utilizando difieren entre sí. Las inferencias se limitan a esos niveles.

Si en lugar de *fijar* los niveles que se desea estudiar se procede seleccionando al azar unos pocos niveles entre todos los posibles porque las inferencias que interesa realizar se refieren, no a unos niveles concretos, sino a cualquiera de los posibles, entonces el modelo de ANOVA es de *efectos aleatorios*\* (también llamado *modelo II*). Cuando se utiliza un factor

---

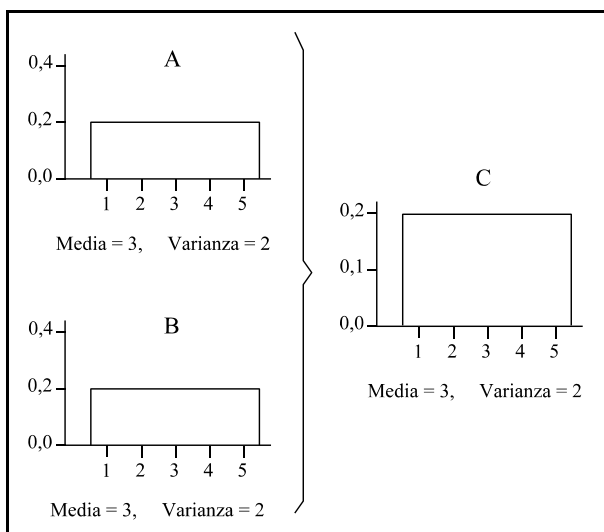
\* Los modelos utilizados con mayor frecuencia en la investigación social son los de efectos fijos. Pero existen situaciones concretas donde resulta apropiado recurrir a un modelo de efectos aleatorios. Por ejemplo, es posible que la eficacia de una determinada terapia venga condicionada por las características personales del terapeuta que la aplica; pero no porque haya algunas características personales conocidas que determinen tal efecto, sino, simplemente, porque distintos terapeutas obtienen diferentes resultados. Podrían seleccionarse aleatoriamente unos pocos terapeutas (no sería necesario seleccionar a todos los terapeutas posibles) y asignar una muestra aleatoria de pacientes a cada uno de ellos. Los resultados del experimento podrían informar, no sobre si tal terapeuta concreto difiere de tal otro, sino sobre si la variable *terapeuta* se relaciona con los resultados de la terapia. Si se eligieran otros terapeutas diferentes, el resultado al que se llegaría sería el mismo (cosa que no puede afirmarse cuando el factor es de efectos fijos).

de efectos aleatorios ya no interesa comparar unos niveles concretos del factor, sino estudiar cualquiera de sus posibles niveles. Si se llevara a cabo una réplica del mismo experimento, los sujetos serían diferentes y también serían diferentes (muy probablemente) los niveles seleccionados. Utilizando un modelo de efectos aleatorios se podría determinar si la utilización de diferentes niveles de fármaco produce efectos diferenciados sobre la reducción del insomnio.

## Lógica del ANOVA

Al combinar en una dos distribuciones con la *misma media* y la *misma varianza*, la distribución resultante de la combinación tiene la misma media y la misma varianza que las dos distribuciones originales. Así, si se combinan, por ejemplo, las distribuciones  $A = \{1, 2, 3, 4, 5\}$  y  $B = \{1, 2, 3, 4, 5\}$ , ambas con media 3 y varianza 2, la distribución resultante  $C = \{1, 2, 3, 4, 5, 1, 2, 3, 4, 5\}$  sigue teniendo media 3 y varianza 2 (ver Figura 14.4).

Figura 14.4. Fusión de dos distribuciones con la misma media y la misma varianza



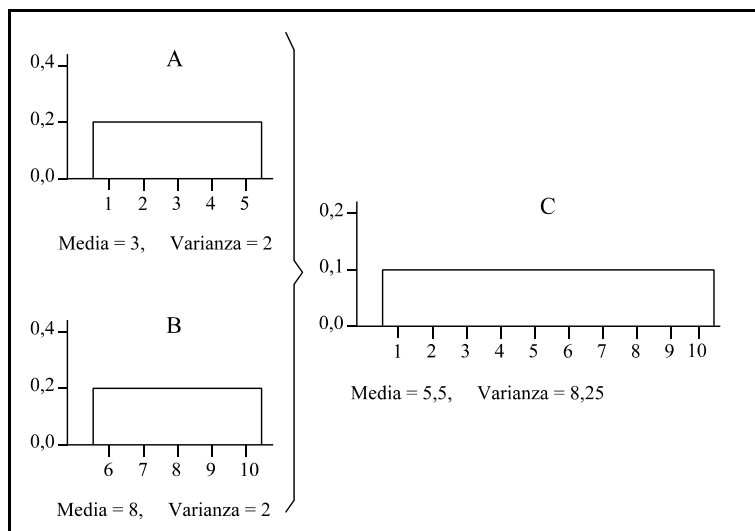
En una situación como ésta es razonable pensar que, si se estima la varianza poblacional a partir de una muestra de las distribuciones  $A$  o  $B$  se obtendrá un resultado similar al que se obtendría si la estimación se efectuara a partir de una muestra de la distribución  $C$ . Con más de dos distribuciones ocurriría exactamente lo mismo.

Consideremos ahora dos distribuciones con *distinta media* pero con la *misma varianza*. Al combinarlas, no sólo cambia la media de la nueva distribución, sino que también lo hace la varianza. Por ejemplo, si se combina la distribución  $A = \{1, 2, 3, 4, 5\}$ , con media 3 y varianza 2, con la distribución  $B = \{6, 7, 8, 9, 10\}$ , con media 8 y varianza 2, la distribución resultante  $C = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$  tiene una media de 5,5 y una varianza de 8,25. La Figura 14.5 refleja esta situación.

Ahora es razonable pensar que una estimación de la varianza poblacional a partir de una muestra extraída de las poblaciones *A* o *B* será sustancialmente diferente de una estimación efectuada a partir de una muestra extraída de la población *C*. Con más de dos poblaciones ocurre exactamente lo mismo.

Esta simple observación es el punto de partida del análisis de varianza, el cual va a permitir comparar las medias de varias distribuciones a partir del estudio de sus varianzas. Para ello, según se desprende de los párrafos anteriores, es necesario comenzar asumiendo que las poblaciones con las que se va a trabajar poseen la misma varianza.

Figura 14.5. Fusión de dos distribuciones con distinta media y con la misma varianza



Supongamos que de  $J$  poblaciones, todas ellas normales y con idéntica varianza (es decir,  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_J^2 = \sigma_e^2$ ), se extraen  $J$  muestras aleatorias de tamaño  $n$  en las cuales se mide una variable  $Y_{ij}$  ( $i$  se refiere a los diferentes elementos de la misma muestra:  $i = 1, 2, \dots, n$ ; y  $j$  se refiere a las diferentes muestras:  $j = 1, 2, \dots, J$ ; así,  $Y_{52}$ , por ejemplo, representa la puntuación obtenida por el 5º sujeto de la 2ª muestra).

En una situación como ésta, cada varianza muestral  $S_j^2$  puede tomarse como una estimación de la varianza de su correspondiente población. Pero como se está asumiendo que todas las poblaciones tienen la misma varianza, esa estimación puede mejorarse utilizando el promedio de las  $J$  varianzas muestrales para obtener una única estimación de esa varianza poblacional\*:

\* Si los tamaños muestrales  $n_j$  son distintos, los  $J$  estimadores  $S_j^2$  pueden promediarse ponderando cada uno con sus correspondientes grados de libertad:

$$\sigma_e^2 = \left[ \sum_{j=1}^J (n_j - 1) S_j^2 \right] / \left[ \sum_{j=1}^J (n_j - 1) \right]$$

$$\hat{\sigma}_e^2 = \frac{\sum_{j=1}^J S_j^2}{J}$$

Este promedio es un estimador de la varianza poblacional que suele recibir el nombre de *media cuadrática intra-grupos o error (MCE)*. Conviene ya desde ahora empezar a familiarizarse con este término.

Supongamos ahora por un momento que las  $J$  poblaciones normales que se están muestreando, además de la misma varianza, también tienen la misma media. En ese caso, las  $J$  muestras aleatorias obtenidas pueden ser consideradas muestras de la *misma* población (pues han sido extraídas de  $J$  poblaciones idénticas) y, en consecuencia, las medias de esas muestras podrán ser utilizadas para obtener un nuevo estimador de la varianza poblacional.

Sabemos\* que la varianza poblacional ( $\sigma_e^2$ ) se relaciona con la varianza de la distribución muestral de la media\*\* ( $\sigma_{\bar{X}}^2$ ) de la siguiente manera:

$$\sigma_{\bar{X}}^2 = \frac{\sigma_e^2}{n}$$

En consecuencia, la varianza de la distribución muestral de la media puede utilizarse para obtener un segundo estimador de la varianza poblacional:

$$\hat{\sigma}_e^2 = n \hat{\sigma}_{\bar{X}}^2$$

A este estimador se le suele llamar *media cuadrática inter-grupos (MCI)*. También con este término conviene empezar a familiarizarse.

Así pues, se tienen dos estimadores de la varianza poblacional  $\sigma_e^2$ . Uno de ellos, *MCE*, es independiente del valor de las medias poblacionales; se obtiene a partir de las varianzas de las puntuaciones individuales de cada muestra\*\*\*. El otro, *MCI*, depende del valor de las medias poblacionales: sólo es un estimador de  $\sigma_e^2$  cuando las  $J$  muestras son extraídas de la misma población (con media  $\mu$  y varianza  $\sigma_e^2$ ) o de  $J$  poblaciones idénticas (y, por tanto, todas ellas con la misma media, además de con la misma varianza).

\* Las medias obtenidas a partir de muestras aleatorias de tamaño  $n$  extraídas de una población normal se distribuyen normalmente con media  $\mu$  y varianza  $\sigma^2/n$ .

\*\* La distribución muestral de la media es la distribución de las medias calculadas en todas las muestras de tamaño  $n$  que es posible extraer de una determinada población. Puede repasarse el concepto de distribución muestral de la media en el Capítulo 9.

\*\*\* Conviene recordar que la varianza de un conjunto de puntuaciones no se ve alterada si a esas puntuaciones se le añade una constante. Por tanto, aunque las medias poblacionales difieran entre sí (es decir, aunque las puntuaciones de las distintas poblaciones difieran en un valor constante), como las varianzas poblacionales siguen siendo iguales, el estimador

$$\hat{\sigma}_e^2 = \frac{\sum_{j=1}^J S_j^2}{J}$$

no se verá afectado por el valor de las medias.

Por tanto, si  $MCI$  y  $MCE$  se calculan a partir de muestras aleatorias extraídas de poblaciones con la misma media, sus valores serán muy parecidos. Por el contrario, si  $MCI$  y  $MCE$  se calculan en muestras extraídas de poblaciones que no tienen la misma media, el valor de  $MCI$  será mayor que el valor de  $MCE$  (recuérdese el argumento expuesto más arriba a propósito de las Figuras 14.4 y 14.5).

Ahora bien, aunque las poblaciones muestreadas tengan la misma media, como las estimaciones obtenidas con  $MCI$  y  $MCE$  son valores muestrales, raramente tomarán valores idénticos. Cabe esperar que, aun siendo iguales las medias poblacionales, entre  $MCI$  y  $MCE$  existan ligeras diferencias atribuibles a las fluctuaciones propias del azar muestral. Y si las medias poblacionales son distintas, los valores de  $MCI$  y  $MCE$  también serán distintos. La clave está precisamente en encontrar un método que permita determinar cuándo la diferencia entre  $MCI$  y  $MCE$  es lo bastante grande como para pensar que no puede ser atribuida al azar muestral, sino a la diferencia entre las medias poblacionales. Justamente ese método es el que se conoce como análisis de varianza.

## ANOVA de un factor

El análisis de varianza de un factor sirve para comparar varios grupos en una variable cuantitativa. Se trata, por tanto, de una generalización de la *prueba T para dos muestras independientes* al caso de diseños con más de dos muestras.

A la variable categórica (nominal u ordinal) que define los grupos que se desea comparar se le llama *independiente* o *factor* y se representa por VI. A la variable cuantitativa (de intervalo o razón) en la que se desea comparar los grupos se le llama *dependiente* y se representa por VD.

Si se quiere, por ejemplo, averiguar cuál de tres programas distintos de incentivos aumenta de forma más eficaz el rendimiento de un determinado colectivo, puede seleccionarse tres muestras aleatorias de ese colectivo y aplicar a cada una de ellas uno de los tres programas. Después, se puede medir el rendimiento de cada grupo y averiguar si existen o no diferencias entre ellos. Se tendrá una VI categórica (el tipo de programa de incentivos) cuyos niveles se quieren comparar entre sí, y una VD cuantitativa (la medida del rendimiento), en la cual se desea comparar los tres programas. El ANOVA de un factor permite obtener información sobre el resultado de esa comparación. Es decir, permite concluir si los sujetos sometidos a distintos programas difieren en la medida de rendimiento utilizada.

La hipótesis que se pone a prueba en el ANOVA de un factor es que las medias poblacionales (las medias de la VD en cada nivel de la VI) son iguales. Si las medias poblacionales son iguales, eso significa que los grupos no difieren en la VD y que, en consecuencia, la VI o factor es independiente de la VD.

La estrategia para poner a prueba la hipótesis de igualdad de medias consiste en obtener un estadístico, llamado  $F$ , que refleja el grado de parecido existente entre las medias que se están comparando.

El numerador del estadístico  $F$  es una estimación de la varianza poblacional basada en la variabilidad existente entre las medias de cada grupo:  $MCI = n\hat{\sigma}_Y^2$ . El denominador del estadístico  $F$  es también una estimación de la varianza poblacional, pero basada en la variabilidad existente dentro de cada grupo:  $MCE = \bar{S}_j^2$  ( $j$  se refiere a los distintos grupos o niveles del factor):

$$F = \frac{MCI}{MCE} = \frac{n \hat{\sigma}_y^2}{\bar{S}_j^2}$$

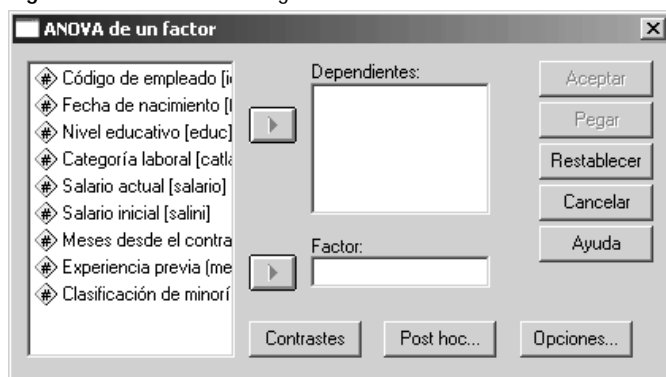
Si las medias poblacionales son iguales, las medias muestrales de los diferentes grupos serán parecidas, existiendo entre ellas tan sólo diferencias atribuibles al azar. En ese caso, la estimación  $\hat{\sigma}_1^2$  (basada en las diferencias entre las medias muestrales) reflejará el mismo grado de variación que la estimación  $\hat{\sigma}_2^2$  (basada en las diferencias entre las puntuaciones individuales dentro de cada grupo) y el cociente  $F$  tomará un valor próximo a 1. Por el contrario, si las medias muestrales son distintas, la estimación  $\hat{\sigma}_1^2$  reflejará mayor grado de variación que la estimación  $\hat{\sigma}_2^2$ , en cuyo caso el cociente  $F$  tomará un valor mayor que 1. Cuanto más diferentes sean las medias muestrales, mayor será el valor de  $F$ .

Si las poblaciones muestreadas son normales y sus varianzas iguales, el estadístico  $F$  se distribuye según el modelo de probabilidad  $F$  de *Fisher-Snedecor* (los grados de libertad del numerador son el número de grupos menos 1; los del denominador, el número total de observaciones menos el número de grupos). Suponiendo cierta la hipótesis de igualdad de medias, es posible conocer la probabilidad de obtener un valor  $F$  como el obtenido o mayor (ver Pardo y San Martín, 1998, págs. 248-250).

El estadístico  $F$  se interpreta de forma similar a como se ha hecho en el capítulo anterior con el estadístico  $T$ . Si el nivel crítico asociado al estadístico  $F$  (es decir, si la probabilidad de obtener valores como el obtenido o mayores) es menor que 0,05, se deberá rechazar la hipótesis de igualdad de medias y se podrá concluir que no todas las medias poblacionales comparadas son iguales. En caso contrario, no se podrá rechazar la hipótesis nula ni, consecuentemente, afirmar que los grupos comparados difieran en sus promedios poblacionales. Para llevar a cabo un ANOVA de un factor:

- Seleccionar la opción **Comparar medias > ANOVA de un factor...** del menú **Analizar** para acceder al cuadro de diálogo *ANOVA de un factor* que muestra la Figura 14.6.

Figura 14.6. Cuadro de diálogo *ANOVA de un factor*



La lista de variables contiene un listado de todas las variables numéricas del archivo de datos (no aparecen listadas las variables con formato de cadena). Para obtener un ANOVA de un factor con las especificaciones que el programa tiene establecidas por defecto:

- Seleccionar una variable *cuantitativa* y trasladarla a la lista **Dependientes**.
- Seleccionar una variable *categorica* y trasladarla al cuadro **Factor**.

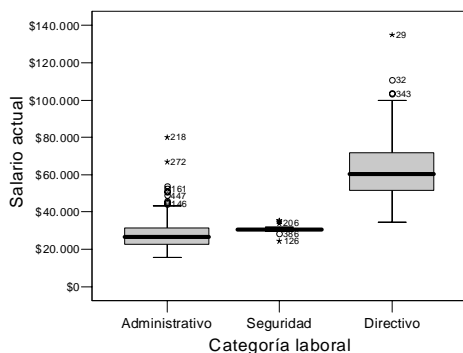
La variable *dependiente* es la variable cuantitativa en la cual se desea comparar los grupos. La variable *factor* es la variable categórica que define los grupos que se desea comparar. Puede seleccionarse más de una variable dependiente: el SPSS genera un análisis de varianza completo por cada variable dependiente seleccionada.

### Ejemplo: ANOVA de un factor

Este ejemplo muestra cómo llevar a cabo un análisis de varianza de un factor con las especificaciones que el programa tiene establecidas por defecto. Siguiendo con el archivo *Datos de empleados*, se desea averiguar si los diferentes grupos definidos por la variable *catlab* (categoría laboral) difieren en la variable *salario* (salario actual).

En general, antes de llevar a cabo un ANOVA, siempre es buena idea obtener alguna información descriptiva de las variables que se desea estudiar. Un diagrama de cajas, por ejemplo, puede ayudar a formarse una idea rápida y bastante completa de lo que está ocurriendo. La Figura 14.7 muestra las cajas correspondientes a *salario actual* en cada *categoría laboral*. El diagrama permite apreciar que la dispersión del *salario* es muy diferente en los tres grupos: el de directivos muestra gran dispersión, algo menos el de administrativos y muy poca el de agentes de seguridad. Por otro lado, mientras que las distribuciones de los administrativos y de los directivos son sensiblemente asimétricas, en la de agentes de seguridad no se aprecia una asimetría evidente. Por último, los promedios salariales de los tres grupos (que es lo que realmente interesa estudiar), no parece que sean iguales: el grupo de directivos tiene un salario medio (mediana) mayor que los otros dos grupos.

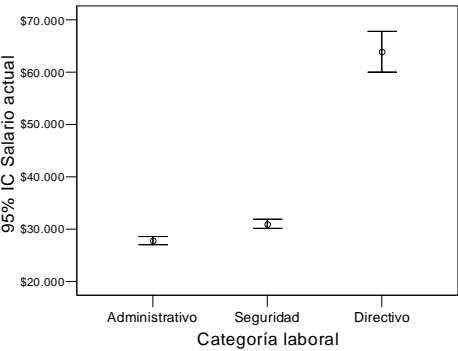
Figura 14.7. Diagramas de caja de *salario actual* por *categoría laboral*



En este sentido, un gráfico de barras de error (con las medias de cada grupo y sus correspondientes intervalos de confianza) sirve para formarse una idea sobre el grado de parecido existente entre las medias salariales que se desea comparar. La Figura 14.8 ofrece estas barras de error. En ellas se aprecia de nuevo que la media salarial del grupo de directivos es sensiblemente mayor que la de los otros dos grupos. El hecho de que los intervalos de confianza no se solapan indica que los tres promedios podrían ser significativamente distintos; pero estos

intervalos de confianza están calculados individualmente y no reflejan diferencias reales. Para poder determinar si existen diferencias y, en caso afirmativo, entre qué grupos existen, es necesario recurrir al análisis de varianza.

Figura 14.8. Barras de error del *salario actual* en cada *categoría laboral*



Para comparar los promedios salariales de las diferentes categorías laborales mediante un análisis de varianza:

- En el cuadro de diálogo principal (ver Figura 14.6), seleccionar la variable *salario* (salario actual) y trasladarla a la lista **Dependientes**.
- Seleccionar la variable *catlab* (categoría laboral) y trasladarla al cuadro **Factor**.

Aceptando estas selecciones, el *Visor* ofrece la información que muestra la Tabla 14.1. Según se ha explicado ya, el estadístico *F* es el cociente entre dos estimadores diferentes de la varianza poblacional: uno basado en la variación existente entre los grupos (variación *Inter-grupos*) y otro basado en la variación existente dentro de cada grupo (variación *Intra-grupos* o *error*). La Tabla 14.1 ofrece: una cuantificación de ambas fuentes de variación (*Sumas de cuadrados*), los grados de libertad asociados a cada suma de cuadrados (*gl*) y el valor concreto que adopta cada estimador de la varianza poblacional (*Medias cuadráticas*, que se obtienen dividiendo las sumas de cuadrados entre sus correspondientes grados de libertad).

El cociente entre estas dos medias cuadráticas (la *inter-grupos* y la *intra-grupos*) proporciona el valor del estadístico *F*, el cual aparece acompañado de su correspondiente nivel crítico o nivel de significación observado (*Sig.*), es decir, de la probabilidad de obtener valores *F* como el obtenido o mayores bajo la hipótesis nula de igualdad de medias.

Puesto que el valor del nivel crítico (*Sig.* < 0,0005) es muy pequeño, se puede rechazar la hipótesis de igualdad de medias y concluir que las poblaciones definidas por la variable *catlab* no poseen el mismo salario medio: hay al menos una población cuyo salario medio difiere del de al menos otra.

Tabla 14.1. Resumen del ANOVA de un factor

Salario actual					
	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	89438483925,943	2	44719241962,971	434,481	,000
Intra-grupos	48478011510,397	471	102925714,459		
Total	137916495436,340	473			

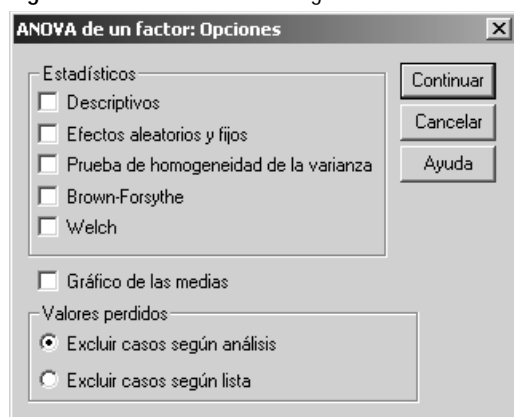


## Opciones

Las opciones del procedimiento permiten, entre otras cosas, seleccionar algunos estadísticos descriptivos básicos, obtener el contraste de *Levene* sobre igualdad de varianzas y decidir qué tratamiento se desea dar a los casos con valores perdidos. Para seleccionar estas opciones:

- Pulsar el botón **Opciones...** del cuadro de diálogo principal (ver figura 14.8) para acceder al subcuadro de diálogo *ANOVA de un factor: Opciones* que muestra la Figura 14.9.

Figura 14.9. Subcuadro de diálogo *ANOVA de un factor: Opciones*



**Estadísticos.** Este recuadro incluye algunos estadísticos descriptivos, estimaciones de la varianza de los componentes, el contraste de *Levene* sobre igualdad u homogeneidad de las varianzas poblacionales y dos alternativas robustas al estadístico *F* del ANOVA:

- **Descriptivos.** Ofrece estadísticos descriptivos referidos tanto a cada grupo como al total muestral: número de observaciones, media, desviación típica, error típico de la media, intervalo de confianza para la media y valores mínimo y máximo.
- **Efectos aleatorios y fijos.** Los niveles de una variable independiente o factor pueden establecerse de dos maneras distintas: *fijándolos* (se utilizan solamente los niveles que se desea estudiar o los niveles que posee la variable; es la forma habitual de proceder) o *seleccionándolos aleatoriamente* entre la población de posibles niveles del factor. En el primer caso se habla de factor de *efectos fijos*; en el segundo, de factor de *efectos aleatorios*.

Para los factores de *efectos fijos*, el *Visor* ofrece una estimación de la desviación típica poblacional (raíz cuadrada de la media cuadrática intra-grupos) junto con el error típico (desviación típica dividida por la raíz cuadrada del número de casos) y el intervalo de confianza para la media total basado en ese error típico. Para los factores de *efectos aleatorios*, el *Visor* ofrece una estimación de la varianza del factor (una de las dos varianzas en las que se descompone la varianza total), una estimación del error típico y un intervalo de confianza para la media total basado en ese error típico.

- " **Prueba de homogeneidad de varianzas.** El estadístico  $F$  del ANOVA de un factor se basa en el cumplimiento de dos supuestos fundamentales: *normalidad* y *homocedasticidad*. *Normalidad* significa que la variable dependiente se distribuye normalmente en las  $J$  poblaciones muestreadas (tantas como grupos definidos por la variable independiente o factor); si los tamaños de los grupos son grandes, el estadístico  $F$  se comporta razonablemente bien incluso con distribuciones poblacionales sensiblemente alejadas de la normalidad. *Homocedasticidad* o igualdad de varianzas significa que las  $J$  poblaciones muestreadas poseen la misma varianza; el incumplimiento de este supuesto debe ser cuidadosamente vigilado, particularmente cuando los grupos tienen distinto tamaño. La opción **Prueba de homogeneidad de varianzas** permite evaluar este supuesto mediante el contraste de *Levene* (ver Capítulo 11, apartado *Homogeneidad de varianzas*, para una descripción de este contraste).
- " **Brown-Forsythe.** El estadístico de Brown-Forsythe (1974) representa una alternativa robusta al estadístico  $F$  del ANOVA cuando no se puede asumir que las varianzas poblacionales son iguales. El numerador del estadístico de Brown-Forsythe es el mismo que el del estadístico  $F$ ; el denominador se obtiene estimando cada varianza poblacional por separado.
- " **Welch.** El estadístico de Welch (1951) también representa una alternativa robusta al estadístico  $F$  del ANOVA cuando no se puede asumir igualdad de varianzas. Tanto el estadístico de Welch como el de Brown-Forsythe se distribuyen según el modelo de probabilidad  $F$ , pero con los grados de libertad corregidos.
- " **Gráfico de las medias.** Esta opción permite obtener un gráfico de líneas con el factor en el eje de abscisas y la variable dependiente en el de ordenadas.

**Valores perdidos.** Los casos con valores perdidos pueden excluirse del análisis utilizando dos criterios diferentes:

**Excluir casos según análisis.** Esta opción excluye de cada ANOVA los casos que tienen algún valor perdido en la variable factor o en la variable dependiente que está siendo analizada. Es la opción por defecto.

**Excluir casos según lista.** Esta opción excluye de todos los ANOVAs solicitados los casos que tienen algún valor perdido en la variable factor o en cualquiera de las variables trasladadas a la lista Dependientes.

### **Ejemplo: ANOVA de un factor > Opciones**

Este ejemplo muestra cómo obtener e interpretar los estadísticos del subcuadro de diálogo *ANOVA de un factor: Opciones* (se sigue utilizando el archivo *Datos de empleados*):

- ' En el cuadro de diálogo principal (ver Figura 14.6), trasladar la variable *salario* a la lista Dependientes y la variable *catlab* al cuadro Factor.
- ' Pulsar el botón **Opciones...** para acceder al subcuadro de diálogo *ANOVA de un factor: Opciones* (ver Figura 14.9) y marcar todas las opciones. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor de resultados* ofrece la información que recogen las Tablas 14.2 a la 14.4 y el gráfico de líneas de la Figura 14.10.

La Tabla 14.2 (se ha pivotado para ajustarla al tamaño de la página) muestra, para cada grupo y para el total muestral, el número de casos, la media, la desviación típica insesgada, el error típico de la media, los límites del intervalo de confianza para la media (al 95 %) y los valores mínimo y máximo. Todo ello referido a la variable *salario actual*.

Además, para el modelo de efectos fijos, la tabla muestra la desviación típica (raíz cuadrada de la media cuadrática intra-grupos; ver Tabla 14.1), el error típico (la desviación típica dividida por la raíz cuadrada del número total de casos) y el intervalo de confianza para la media total basado en ese error típico. Y para el modelo de efectos aleatorios, la tabla ofrece una estimación de la varianza del factor (*Componentes de la varianza*), el error típico basado en esa estimación de la varianza y el intervalo de confianza para la media total basado en ese error típico. La estimación de la varianza del factor o varianza inter-grupos (que es uno de los *componentes* de la varianza total, junto con la varianza residual o intra-grupos) se obtiene a partir de la variabilidad existente entre las medias del factor; en concreto, restando las medias cuadráticas inter-grupos e intra-grupos (ver Tabla 14.1) y dividiendo esa diferencia entre el tamaño de los grupos (con tamaños distintos se utiliza una especie de promedio de todos los tamaños).

**Tabla 14.2.** Estadísticos descriptivos

Salario actual		Modelo				
		Administ.	Seguridad	Directivo	Total	
						Efectos fijos
						Efectos aleatorios
N		363	27	84	474	
Media		27.838,54	30.938,89	63.977,80	34.419,57	
Desviación típica		7.567,99	2.114,62	18.244,78	17.075,66	10.145,23
Error típico		397,22	406,96	1.990,67	784,31	465,99
Intervalo de confianza para la media al 95%	Límite inferior	27.057,40	30.102,37	60.018,44	32.878,40	33.503,90
	Límite superior	28.619,68	31.775,40	67.937,16	35.960,73	35.335,24
Mínimo		15.750	24.300	34.410	15.750	
Máximo		80.000	35.250	135.000	135.000	
Componentes de la varianza						496.889.967,38

La Tabla 14.3 contiene el contraste de Levene sobre igualdad de varianzas. Junto con el valor del estadístico (59,733) aparecen los grados de libertad de su distribución ( $gl1 = 2$ ,  $gl2 = 471$ ) y el nivel crítico ( $Sig. < 0,0005$ ). Puesto que el nivel crítico es muy pequeño, se debe rechazar la hipótesis de igualdad de varianzas y concluir que, en las poblaciones definidas por las tres categorías laborales, las varianzas de la variable *salario* no son iguales.

**Tabla 14.3.** Contraste de *Levene* sobre igualdad de varianzas

Salario actual			
Estadístico de Levene	gl1	gl2	Sig.
59.733	2	471	.000

La Tabla 14.4 ofrece los estadísticos de Brown-Forsythe y de Welch. Según se ha señalado ya, ambos constituyen una buena alternativa al estadístico *F* cuando no es posible asumir que

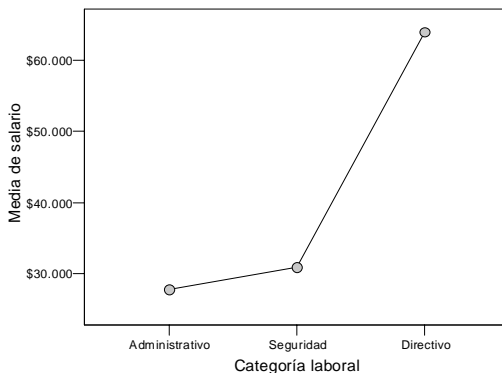
las varianzas poblacionales son iguales (como ocurre en el ejemplo). Puesto que el nivel crítico (*Sig.*) asociado a ambos estadísticos es menor que 0,05, se puede rechazar la hipótesis de igualdad de medias y concluir que los promedios salariales de las dos poblaciones comparadas no son iguales.

**Tabla 14.4.** Estadísticos robustos para el contraste de la hipótesis de igualdad de medias

Salario actual				
	Estadístico	gl1	gl2	Sig.
Welch	162,200	2	117,312	,000
Brown-Forsythe	306,810	2	93,906	,000

Por último, la Figura 14.10 muestra una representación de las medias del *salario actual* en cada *categoría laboral*. Si se desea, puede editarse el gráfico (pinchando dos veces sobre él) y transformarlo en un gráfico de barras, con la altura de las barras indicando el tamaño de las medias.

**Figura 14.10.** Gráfico de líneas: medias de *salario actual* en cada *categoría laboral*



## Comparaciones *post hoc* o *a posteriori*

El estadístico *F* del ANOVA únicamente permite contrastar la hipótesis general de que los *J* promedios comparados son iguales. Rechazar esa hipótesis significa que las medias poblacionales comparadas no son iguales, pero no permite precisar dónde en concreto se encuentran las diferencias detectadas: ¿difieren entre sí todas las medias?, ¿hay una sola media que difiere de las demás?, etc.

Para saber qué media difiere de qué otra se debe utilizar un tipo particular de contrastes denominados comparaciones múltiples *post hoc* o comparaciones *a posteriori*. Estas comparaciones permiten controlar la *tasa de error* al efectuar varios contrastes utilizando las mismas medias, es decir, permiten controlar la probabilidad de cometer errores tipo I al tomar varias decisiones (los errores tipo I se cometen cuando se decide rechazar una hipótesis nula que en realidad no debería rechazarse). Para efectuar comparaciones *post hoc*:

Pulsar el botón **Post Hoc...** del cuadro de diálogo principal (ver Figura 14.6) para acceder al subcuadro de diálogo *ANOVA de un factor: Comparaciones múltiples post hoc* que muestra la Figura 14.11. Todas las opciones de este cuadro de diálogo ofrecen información similar: permiten, una vez rechazada la hipótesis global de igualdad de medias, averiguar qué medias en concreto difieren de qué otras (para una revisión de estos procedimientos, ver Toothaker, 1991; o Pardo y San Martín, 1998, Capítulo 6).

Figura 14.11. Subcuadro de diálogo *ANOVA de un factor: Comparaciones múltiples post hoc*

**Asumiendo varianzas iguales.** Puede seleccionarse uno o más de los siguientes métodos de comparaciones *post hoc*:

- " **DMS.** *Diferencia Mínima Significativa* basada en la distribución *t* de Student. Este método, inicialmente propuesto por Fisher (1935), no ejerce ningún control sobre la *tasa de error*. Es decir, cada comparación se lleva a cabo utilizando el nivel de significación establecido (generalmente 0,05), por lo que la *tasa de error* para el conjunto de comparaciones puede llegar a  $1 - (1 - \alpha)^k$ , siendo  $\alpha$  el nivel de significación y  $k$  el número de comparaciones llevadas a cabo. Este método suele encontrarse en la literatura estadística con su acrónimo inglés: LSD = «Least Significant Difference».
- " **Bonferroni.** Método basado en la distribución *t* de Student y en la desigualdad de Bonferroni (también conocido como método de Dunn –su promotor en 1961– o de Dunn-Bonferroni). Controla la tasa de error dividiendo el nivel de significación ( $\alpha$ ) entre el número de comparaciones ( $k$ ) llevadas a cabo. Cada comparación se evalúa utilizando un nivel de significación  $\alpha_c = \alpha/k$ .
- " **Sidak** (1967). También se basa en la distribución *t* de Student, pero controla la tasa de error evaluando cada comparación con un nivel de significación  $\alpha_c = 1 - (1 - \alpha)^{1/k}$ . Esta solución es algo menos conservadora que la de Bonferroni, es decir, rechaza la hipótesis de igualdad de medias en más ocasiones que el método de Bonferroni.
- " **Scheffé** (1953, 1959). Este método, basado en la distribución *F*, permite controlar la tasa de error para el conjunto total de comparaciones que es posible diseñar con *J* medias (una con otra, una con todas las demás, dos con dos, etc.). Utilizado para efectuar comparacio-

nes por pares es un método muy conservador: tiende a considerar significativas menos diferencias de las que debería.

- " R-E-G-W *F*. Método de Ryan (1960), Einot-Gabriel (1975) y Welsch (1977) basado en la distribución *F*. Se trata de un método por pasos. Tras ordenar de forma ascendente las *J* medias por su tamaño, se efectúan todas las comparaciones posibles entre pares de medias teniendo en cuenta el número de escalones (*r*) que las separan: con *J* medias, la media más pequeña y la más grande están separadas  $r = J$  escalones; la media más pequeña y la segunda más grande están separadas  $r = J - 1$  escalones; la media más pequeña y la tercera más grande están separadas  $r = J - 2$  escalones; etc. Dos medias adyacentes tras la ordenación están separadas 2 escalones. El número de escalones existente entre las medias comparadas condiciona el nivel de significación de cada comparación, siendo éste mayor cuanto más alejadas se encuentran las medias después de ser ordenadas. En el método R-E-G-W *F*, cada comparación se evalúa utilizando un estadístico *F* y un nivel de significación  $\alpha_c = 1 - (1 - \alpha)^{r/J}$ . Es un método por pasos más potente que el de Duncan y el de Student-Newman-Keuls (ver más abajo), pero no es apropiado cuando los grupos tienen tamaños distintos.
- " R-E-G-W *Q*. Método de Ryan (1960), Einot-Gabriel (1975) y Welsch (1977) basado en la distribución del rango estudentizado. Se trata de un método por pasos que utiliza el mismo estadístico que, por ejemplo, el método de Student-Newman-Keuls o el método de Tukey, pero que controla el nivel de significación de cada comparación del mismo modo que el método R-E-G-W *F*. Es un método por pasos más potente que el de Duncan y el de Student-Newman-Keuls (ver más abajo), pero no apropiado cuando los grupos tienen tamaños distintos.
- " S-N-K. Student-Newman-Keuls (Newman, 1939; Keuls, 1952). Método basado en la distribución del rango estudentizado. Al igual que los métodos R-E-G-W *F* y *Q*, éste también se basa en una ordenación de las medias por su tamaño. Pero a diferencia de ellos, aquí el nivel de significación para cada conjunto de medias separadas *r* pasos es siempre  $\alpha$ . Cuantos más pasos existen entre dos medias, mayor es la *diferencia mínima* necesaria para considerar que esas medias difieren significativamente.
- " Tukey (1953). *Diferencia honestamente significativa* de Tukey. Equivale a utilizar el método de Student-Newman-Keuls con  $r = J = n^\circ \text{ de medias}$ . Por tanto, todas las comparaciones son referidas a una misma *diferencia mínima*. Es uno de los métodos de mayor aceptación.
- " Tukey-b (1953). Este método consiste en considerar como *diferencia mínima* el valor medio entre la *diferencia honestamente significativa* de Tukey y la *diferencia mínima* obtenida con el método de Student-Newman-Keuls para el caso de  $r = 2$ .
- " Duncan (1955). Prueba del rango múltiple de Duncan. Método de comparación por pasos basado en la distribución del rango estudentizado. Controla la tasa de error utilizando, para el conjunto de medias separadas *r* pasos, un nivel de significación  $\alpha_c = 1 - (1 - \alpha)^{r-1}$ . Cuantos más pasos existen entre dos medias, mayor es la *diferencia mínima* necesaria para considerar que esas medias difieren significativamente.
- " GT2 de Hochberg (1974). Es un método muy similar a la *Diferencia honestamente significativa* de Tukey, pero se basa en la distribución del módulo máximo estudentizado. El método de Tukey suele ser más potente.

- " Gabriel (1969). También se basa en la distribución del módulo máximo estudentizado. Con grupos del mismo tamaño, este método es más potente que el de Hochberg, pero con tamaños muy desiguales ocurre lo contrario.
- " Waller-Duncan (1969). Utiliza la distribución  $t$  de Student y una aproximación bayesiana. Si los tamaños muestrales son distintos, utiliza la media armónica.
- " Dunnett (1955). Sirve para comparar cada grupo con un grupo control. Por tanto, controla la tasa de error para  $k-1$  comparaciones. Por defecto, se considera que la última categoría del factor es la que define el grupo control, pero puede seleccionarse la primera categoría. Permite efectuar tanto contrastes bilaterales como unilaterales.

**No asumiendo varianzas iguales.** En el caso de que no pueda asumirse que las varianzas poblacionales son iguales, existe la posibilidad de elegir alguno de estos cuatro métodos:

- " T2 de Tamhane (1977, 1979). Método basado en la distribución del módulo máximo estudentizado.
- " T3 de Dunnett (1980). Modificación propuesta por Dunnett al estadístico T2 de Tamhane. Se basa también en la distribución del módulo máximo estudentizado.
- " Games-Howell (1976). Método similar al de Tukey. Se basa en la distribución del rango estudentizado y en un estadístico  $T$  en el que, tras estimar las varianzas poblacionales suponiendo que son distintas, se corrigen los grados de libertad mediante la ecuación de Welch (ver, en el Capítulo 13, el apartado *Prueba T para muestras independientes*). En términos generales, de los cuatro métodos de este recuadro, el de Games-Howell es el que mejor permite controlar la tasa de error en diferentes situaciones.
- " C de Dunnett (1980). Método idéntico al de Games-Howell excepto en la forma de corregir los grados de libertad de la distribución del rango estudentizado. Esta solución es más conservadora que la de Games-Howell.

**Nivel de significación.** Esta opción permite establecer el nivel de significación con el que se desea llevar a cabo las comparaciones múltiples.

### ***Ejemplo: ANOVA de un factor > Comparaciones post hoc***

Este ejemplo muestra cómo obtener e interpretar las comparaciones *pot hoc* del procedimiento ANOVA de un factor. Puesto que todas las comparaciones *post hoc* se obtienen e interpretan de la misma forma, bastará con marcar cualquiera de las disponibles y estudiar los resultados que genera (se sigue utilizando el archivo *Datos de empleados*).

- ' En el cuadro de diálogo principal (ver Figura 14.6), trasladar la variable *salario* a la lista Dependientes y la variable *catlab* al cuadro Factor.
- ' Pulsar el botón Post hoc... para acceder al subcuadro de diálogo ANOVA de un factor: Comparaciones múltiples post hoc (ver Figura 14.10).
- ' Marcar la opción Tukey del recuadro Asumiendo varianzas iguales y la opción Games-Howell del recuadro No asumiendo varianzas iguales.

Aceptando estas elecciones, el *Visor de resultados* ofrece la información que recogen las Tablas 14.5 y 14.6.

La primera columna de la Tabla 14.5 indica que se han seleccionado dos métodos *post hoc*: la *diferencia honestamente significativa* (HSD) de Tukey y el método de Games-Howell. A continuación aparecen todas las posibles combinaciones dos a dos entre los niveles o categorías de la variable factor (*categoría laboral*), las diferencias entre los salarios medios de cada dos grupos, el error típico de esas diferencias y el nivel crítico asociado a cada diferencia (*Significación*). Los grupos cuyas medias difieren significativamente al nivel de significación establecido (0,05 por defecto) están marcados con un asterisco.

Puede comprobarse que el resultado obtenido (es decir, el número de diferencias significativas detectadas) no es el mismo con los dos métodos utilizados. Pero, puesto que no puede asumirse que las varianzas poblacionales sean iguales (ver contraste de Levene en la Tabla 14.3), debe prestarse atención a la solución propuesta por el método de Games-Howell. Por tanto, puede concluirse que todos los promedios comparados difieren significativamente: los directivos poseen un salario medio mayor que el de los agentes de seguridad y éstos mayor que el de los administrativos.

Los límites del intervalo de confianza de las dos últimas columnas permiten estimar entre qué límites se encuentra la verdadera diferencia entre las medias de los grupos. Estos intervalos también permiten tomar decisiones sobre si dos promedios difieren o no significativamente (dependiendo de que el intervalo incluya o no el valor cero). Pero al utilizar estos intervalos para decidir sobre la hipótesis de igualdad de medias hay que tener en cuenta que el intervalo se obtiene individualmente para cada diferencia, sin establecer control sobre la tasa de error, por lo que las decisiones que puedan tomarse tomando como base estos intervalos serán demasiado arriesgadas.

**Tabla 14.5.** Comparaciones múltiples *post hoc*. Pruebas de *Tukey* y *Games-Howell*

Variable dependiente: Salario actual

	(I) Categoría laboral	(J) Categoría laboral	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
						Límite inferior	Límite superior
HSD de Tukey	Administrativo	Seguridad	-3.100,35	2.023,76	,277	-7.858,50	1.657,80
		Directivo	-36.139,26*	1.228,35	,000	-39.027,29	-33.251,22
	Seguridad	Administrativo	3.100,35	2.023,76	,277	-1.657,80	7.858,50
		Directivo	-33.038,91*	2.244,41	,000	-38.315,84	-27.761,98
	Directivo	Administrativo	36.139,26*	1.228,35	,000	33.251,22	39.027,29
		Seguridad	33.038,91*	2.244,41	,000	27.761,98	38.315,84
Games-Howell	Administrativo	Seguridad	-3.100,35*	568,68	,000	-4.454,82	-1.745,88
		Directivo	-36.139,26*	2.029,91	,000	-40.977,01	-31.301,51
	Seguridad	Administrativo	3.100,35*	568,68	,000	1.745,88	4.454,82
		Directivo	-33.038,91*	2.031,84	,000	-37.881,37	-28.196,45
	Directivo	Administrativo	36.139,26*	2.029,91	,000	31.301,51	40.977,01
		Seguridad	33.038,91*	2.031,84	,000	28.196,45	37.881,37

\*. La diferencia entre las medias es significativa al nivel .05.

La Tabla 14.6 ofrece una clasificación de los grupos basada en el grado de parecido existente entre sus medias. Así, en el *subconjunto* 1, están incluidos dos grupos (*Administrativos* y *Agentes de seguridad*) cuyas medias no difieren significativamente (*Significación* = 0,226), y en el *subconjunto* 2 está incluido un solo grupo (*Directivos*) que difiere de los dos anteriores



y que, obviamente, no difiere de sí mismo (*Significación* = 1,00). Esta clasificación por subconjuntos no está disponible con todos los métodos *post-hoc*, sino sólo con algunos: *S-N-K*, Tukey, Tukey-b, Duncan, Scheffé, Gabriel, *R-E-G-WF* y *Q*, GT2 de Hochberg y Waller-Duncan; esta es la razón por la cual, a pesar de que no puede asumirse que las varianzas poblacionales sean iguales, la clasificación en subconjuntos homogéneos de la Tabla 14.6 se ha realizado utilizando el método de Tukey en lugar del de Games-Howell.

Tabla 14.6. Subconjuntos homogéneos

		N	Subconjunto para alfa = .05	
			1	2
HSD de Tukey <sup>a,b</sup>	Administrativo	363	27.838,54	
	Seguridad	27	30.938,89	
	Directivo	84		63.977,80
	Sig.		,227	1,000

Se muestran las medias para los grupos en los subconjuntos homogéneos.

- a. Usa el tamaño muestral de la media armónica = 58,031.
- b. Los tamaños de los grupos no son iguales. Se utilizará la media armónica de los tamaños de los grupos. Los niveles de error de tipo I no están garantizados.

## Comparaciones planeadas o *a priori*

Las comparaciones entre pares de grupos (comparaciones *post hoc*) no son las únicas comparaciones múltiples que es posible llevar a cabo con el procedimiento **Anova de un factor**. La opción **Contrastes** permite efectuar comparaciones de tendencia y definir cualquier otro tipo de comparación entre medias que se desee plantear (contrastes personalizados). Para obtener comparaciones de tendencia y contrastes personalizados:

- Pulsar el botón **Contrastes...** del cuadro de diálogo principal (ver Figura 14.6), para acceder al subcuadro de diálogo **ANOVA de un factor: Contrastes** que muestra la Figura 14.12.

Figura 14.12. Subcuadro de diálogo **ANOVA de un factor: Contrastes**



- “ **Polinómico.** Esta opción permite obtener *comparaciones de tendencia*. Si el estadístico *F* lleva al rechazo de la hipótesis de igualdad de medias, eso significa que no todas las medias son iguales y, por tanto, que la variable independiente (VI) y la dependiente (VD) están relacionadas. En ese caso, si la VI es *cuantitativa* (la VD siempre lo es), la opción **Polinómico** permite determinar cuál es el *tipo de relación* (lineal, cuadrática, cúbica, etc.) existente entre la VI y la VD.

Cada polinomio o tendencia es un componente ortogonal (independiente) de la suma de cuadrados inter-grupos. El número máximo de polinomios que se puede obtener es el número de grados de libertad de la suma de cuadrados inter-grupos. Si todos los grupos tienen el mismo tamaño, el SPSS ofrece una solución *no ponderada* en la que cada polinomio o tendencia es, efectivamente, un componente independiente de la suma de cuadrados inter-grupos. Si los grupos no tienen el mismo tamaño, el SPSS ofrece, además de la no ponderada, una solución *ponderada* en la que, para conseguir componentes independientes, se tiene en cuenta el distinto tamaño de los grupos. Tanto en la solución ponderada como en la no ponderada se tiene en cuenta la distancia existente entre los niveles de la variable independiente o factor (ver Pardo y San Martín, 1998, págs. 298-303).

Las opciones del menú desplegable **Orden** permiten fijar cuál es el polinomio de mayor orden que se desea estudiar. Si se elige **Lineal** sólo se obtiene el componente lineal; si se elige **Cuadrático** se obtiene el componente lineal y el cuadrático; etc. El límite se encuentra en el número de grados de libertad de la suma de cuadrados inter-grupos.

**Coefficientes.** Este cuadro de texto permite definir *contrastes personalizados* asignando coeficientes concretos a cada uno de los grupos que se desea comparar. Así, en un diseño con, por ejemplo, 4 grupos de los que interesa comparar los dos primeros con el último, la comparación quedaría definida asignando estos coeficientes: 1, 1, 0, -2; o bien, de forma equivalente: 0.5, 0.5, 0, -1. Para comparar, por ejemplo, el primer grupo con todos los demás tomados juntos, habría que asignar estos otros coeficientes: 3, -1, -1, -1; o bien, de forma equivalente: 1, -1/3, -1/3, -1/3. En un contraste de este tipo siempre se están comparando *dos términos*: una media con otra, una media con varias, o varias medias con varias. El tamaño de los coeficientes utilizados es irrelevante, pero es necesario vigilar que los coeficientes asignados a los grupos de uno de los términos comparados sean positivos y que los coeficientes asignados a los grupos del otro término sean negativos; y que la suma de todos los coeficientes valga cero.

El orden en el que se asignan los coeficientes se corresponde con el orden ascendente de los códigos de los niveles de la variable independiente (el primer coeficiente corresponde al grupo con el código más pequeño). Hay que asignar tantos coeficientes como grupos; por tanto, a los grupos que no intervengan en un contraste concreto se les debe asignar un cero.

Para definir un contraste de tipo *lineal*, los coeficientes asignados deben sumar cero, pero es posible definir contrastes cuyos coeficientes no sumen cero (si es éste el caso, el SPSS muestra un mensaje de aviso). Para definir un contraste personalizado asignando coeficientes (ver Figura 14.12):

- Introducir el primer coeficiente en el cuadro de texto **Coefficientes** y pulsar el botón **Añadir** para trasladarlo a la lista de la parte inferior.
- Repetir la acción para cada uno de los coeficientes hasta añadir tantos como niveles o categorías tenga la variable factor.
- Utilizar los botones **Cambiar** y **Borrar** para modificar y eliminar, respectivamente, coeficientes previamente añadidos.

La línea **Total** para los coeficientes va mostrando la suma de los coeficientes añadidos. Tras asignar los coeficientes de un contraste, debe vigilarse que este total sume cero.

Por supuesto, es posible definir más de un contraste. De hecho, es posible definir hasta 10 contrastes diferentes con un máximo de 50 coeficientes por contraste. Para definir el segundo contraste:

- Pulsar el botón **Siguiente** del recuadro **Contraste 1 de 1**.
- Comenzar a introducir los coeficientes del segundo contraste del mismo modo que se ha hecho con el primero.

**Contraste # de ##** indica el contraste en el que se está (#) y el número total de contrastes definidos (##). El botón **Anterior** permite moverse por contrastes previamente definidos.

El *Visor de resultados* muestra, para cada uno de los contrastes definidos, los coeficientes asignados, el valor del contraste (que se obtiene sumando los productos de cada coeficiente por la media de su correspondiente grupo), el error típico, el estadístico *T*, los grados de libertad y el nivel crítico asociado al estadístico *T* bajo la hipótesis nula de que el valor poblacional del contraste es cero.

### **Ejemplo: ANOVA de un factor > Contrastes polinómicos**

Este ejemplo muestra cómo efectuar *comparaciones de tendencia* mediante la opción **Contrastes... > Polinómico** del procedimiento **ANOVA de un factor**.

- En el cuadro de diálogo principal (ver Figura 14.6), seleccionar la variable *salario* (salario actual) como variable **Dependiente** y la variable *grupedad* (grupos de edad) como variable **Factor** (estos datos están disponibles en el archivo *Datos de empleados ampliado*, el cual puede obtenerse en la página *web* del manual).
- Pulsar el botón **Contrastes...** para acceder al subcuadro de diálogo *ANOVA de un factor: Contrastes* (ver Figura 14.12), marcar la opción **Polinómico** y, en la lista desplegable **Orden**, seleccionar *cúbico* (aunque la variable independiente *grupedad* tiene 5 niveles y es posible, por tanto, evaluar hasta 4 tendencias, más allá de la tendencia cúbica no resulta fácil interpretar la relación).

Aceptando estas elecciones, el *Visor de resultados* ofrece la información que muestra la Tabla 14.7. La información referida a las comparaciones de tendencia aparece integrada en la tabla resumen del ANOVA como parte de la variación *inter-grupos*. Puesto que los grupos no tienen el mismo tamaño, la tabla ofrece tanto la solución *no ponderada* como la *ponderada*. Cada tendencia aparece acompañada de su correspondiente suma de cuadrados, grados de libertad, media cuadrática, estadístico *F* y nivel crítico asociado al estadístico *F*.

Puesto que los grupos no tienen el mismo tamaño, debe optarse por la solución *ponderada* (aunque en este ejemplo ambas soluciones llevan a la misma conclusión). La hipótesis nula que se contrasta con cada tendencia es que la relación representada por esa tendencia concreta es nula. La tendencia o término lineal tiene un nivel crítico asociado de 0,120; puesto que ese valor es mayor que 0,05, se mantiene la hipótesis de que la tendencia lineal es nula y se concluye que no es posible afirmar que entre la VI (grupos de edad) y la VD (salario actual) exista relación lineal significativa.

Tabla 14.7. Resumen del ANOVA de un factor incluyendo comparaciones de tendencia

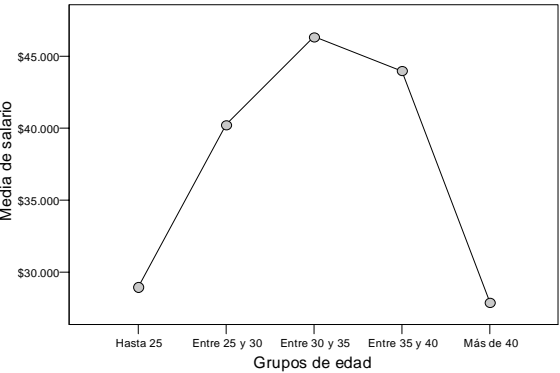
Salario actual			Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	(Combinados)		22441996648,30	4	5610499162,07	22,74	,000
	Término lineal	No ponderado	26306688,29	1	26306688,29	,11	,744
		Ponderado	597016688,56	1	597016688,56	2,42	,120
		Desviación	21844979959,74	3	7281659986,58	29,51	,000
	Término cuadrát.	No ponderado	19056784339,48	1	19056784339,48	77,23	,000
		Ponderado	21359533118,90	1	21359533118,90	86,57	,000
		Desviación	485446840,83	2	242723420,42	,98	,375
	Término cúbico	No ponderado	444357891,76	1	444357891,76	1,80	,180
		Ponderado	482894860,31	1	482894860,31	1,96	,162
		Desviación	2551980,52	1	2551980,52	,01	,919
Intra-grupos		115474216834,58	468	246739779,56			
Total		137916213482,88	472				

A continuación aparece información referida al resto de tendencias todavía no contrastadas (*Desviación*). Puesto que la única tendencia contrastada es la lineal, las todavía no contrastadas son la cuadrática, la cúbica y la de cuarto orden (pues con 5 grupos pueden definirse hasta 4 tendencias). El nivel crítico de estas tendencias ( $Sig. < 0,005$ ) es menor que 0,05, lo que está indicando que, entre las tendencias de orden mayor que el lineal, existe alguna que es significativa. Observando la información referida a la tendencia o término cuadrático se ve que el nivel crítico es menor que 0,0005, lo cual debe llevar al rechazo la hipótesis nula referida a esa tendencia y a concluir que entre la VI y la VD existe relación cuadrática significativa.

En el resto de tendencias todavía no contrastadas (*Desviación*) se incluyen las tendencias cúbica y de cuarto orden. El nivel crítico de estas dos tendencias tomadas juntas vale 0,375; puesto que este nivel crítico es mayor que 0,05, puede afirmarse que entre las tendencias de orden mayor que el cuadrático no existe ninguna significativa.

Puede concluirse, por tanto, que la relación entre las variables *grupos de edad* y *salario actual* es cuadrática. Un gráfico de líneas como el que muestra la Figura 14.13 puede ayudar a entender lo que está ocurriendo (se obtiene con la opción *Gráfico de las medias* del subcuadro de diálogo *ANOVA de un factor: Opciones*; ver Figura 14.9).

Figura 14.13. Gráfico de líneas: relación entre *grupos de edad* y *salario actual*



### Ejemplo: ANOVA de un factor > Contrastes personalizados

Este ejemplo muestra cómo efectuar contrastes personalizados utilizando la opción **Contrastes** del procedimiento **ANOVA de un factor**. Se sigue utilizando *grupedad* (grupos de edad) como variable independiente o factor y *salario* (salario actual) como variable dependiente (estas variables están disponibles en el archivo *Datos de empleados ampliado*, el cual puede obtenerse en la página *web* del manual).

Como ejemplo de contrastes personalizados, se comparan, en primer lugar, los dos primeros grupos de edad (grupos 1 y 2) con los dos últimos (grupos 4 y 5) y, en segundo lugar, los dos grupos extremos con los tres intermedios. Es decir, se van a llevar a cabo dos contrastes. Para ello es necesario introducir los coeficientes apropiados en el cuadro de texto **Coeficientes** y llevarlos a la lista inferior con el botón **Añadir** (ver Figura 14.12).

- Para definir el primer contraste, asignar los coeficientes: 1, 1, 0, -1 y -1.
- Pulsar el botón **Siguiente** y, para definir el segundo contraste, asignar los coeficientes: 3, -2, -2, -2 y 3.
- Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas especificaciones, el *Visor* ofrece los resultados que muestran las Tablas 14.8 y 14.9.

La Tabla 14.8 muestra los coeficientes que se han asignado al definir los dos contrastes. Esta información sirve para comprobar si los contrastes que se están llevando a cabo están correctamente definidos.

**Tabla 14.8.** Coeficientes para los contrastes personalizados

Contraste	Grupos de edad				
	Menos de 25 años	Entre 25 y 30 años	Entre 30 y 35 años	Entre 35 y 40 años	Más de 40 años
1	1	1	0	-1	-1
2	3	-2	-2	-2	3

La Tabla 14.9 ofrece la información sobre los contrastes planteados agrupada en dos bloques que deben utilizarse de forma alternativa: en el primer bloque, los contrastes propuestos están evaluados asumiendo que las varianzas poblacionales son iguales; en el segundo, sin asumir igualdad de varianzas. Aunque es frecuente que ambas estrategias lleven a la misma conclusión, debe utilizarse aquella que se ajuste a las características de los datos; para lo cual debe tenerse en cuenta la decisión tomada al evaluar la hipótesis de igualdad de varianzas mediante el contraste de Levene (en el ejemplo *ANOVA de un factor > Opciones* se ha explicado ya la forma de obtener e interpretar este contraste). Puesto que el contraste de Levene aplicado a los datos del ejemplo no permite asumir varianzas iguales (ver Tabla 14.3), la decisión sobre la hipótesis de que los promedios comparados son iguales debe basarse en la parte inferior de la Tabla 14.9 (*No asumiendo igualdad de varianzas*).

La tabla muestra, para cada uno de los dos contrastes definidos, el valor del contraste, su error típico, el estadístico de contraste *t*, sus grados de libertad y el nivel crítico (*Significación bilateral*). La hipótesis nula que se pone a prueba con cada contraste es que los promedios comparados son iguales. Teniendo en cuenta los niveles críticos asociados a cada contraste

debe decidirse: (1) mantener la hipótesis nula referida al primer contraste (pues  $0,545 > 0,05$ ) y (2) rechazar la hipótesis nula referida al segundo contraste (pues  $0,0005 < 0,05$ ).

En consecuencia, puede concluirse, en primer lugar, que el salario medio de los dos primeros grupos de edad no difiere del salario medio de los dos últimos grupos; y, en segundo lugar, que el salario medio de los grupos de menor y mayor edad difiere significativamente del salario medio de los tres grupos intermedios.

**Tabla 14.9.** Contrastes personalizados

Salario actual		Valor del contraste	Error típico	t	gl	Sig. (bilateral)
Contraste						
Asumiendo igualdad de varianzas	1	-2,676.10	3,602.91	-,743	468	,458
	2	-90,525.64	9,905.03	-9,139	468	,000
No asumiendo igualdad de varianzas	1	-2,676.10	4,388.46	-,610	52,745	,545
	2	-90,525.64	12,453.21	-7,269	101,853	,000



## Análisis de varianza (II)

### El procedimiento *Modelo lineal general: Univariante*

Los modelos *factoriales* de análisis de varianza (*factorial = más de un factor*) sirven para evaluar el efecto individual y conjunto de dos o más factores (variables independientes categóricas) sobre una variable dependiente cuantitativa.

Un ANOVA factorial permite estudiar, por ejemplo, si el salario (variable dependiente) es diferente entre los varones y las mujeres (efecto del primer factor) y, al mismo tiempo, si varios grupos de edad tienen distinto salario (efecto del segundo factor). Pero, además, también permite estudiar si las diferencias entre varones y mujeres se repiten o no en cada grupo de edad, es decir, permite determinar si la interacción entre los factores *sexo* y *grupos de edad* afecta a la variable dependiente *salario*. Incluir más de un factor en un mismo diseño posee la importante ventaja de poder estudiar el efecto conjunto (interacción) de los factores.

En un modelo de dos factores, los efectos de interés son tres: los dos efectos *principales* (uno por cada factor) y el efecto de la *interacción* entre ambos factores. En un modelo de tres factores, los efectos de interés son siete: los tres efectos principales, los tres efectos de las interacciones dobles (uno por cada interacción entre cada dos factores) y el efecto de la interacción triple (entre los tres factores). Etc.

El procedimiento **Univariante** permite ajustar este tipo de modelos. Además, ofrece la posibilidad de trabajar con factores de efectos fijos y con factores de efectos aleatorios. También permite llevar a cabo *análisis de covarianza y de regresión*, y utilizar *modelos aleatorizados en bloques y modelos jerárquicos* o con factores anidados.

### Análisis de varianza factorial

En un análisis de varianza factorial existe una hipótesis nula por cada factor y por cada posible combinación de factores. La hipótesis nula referida a un factor individual afirma que las medias de las poblaciones definidas por los niveles del factor son iguales. La hipótesis nula referida al efecto de una interacción entre factores afirma que tal efecto es nulo. Para contrastar estas hipótesis, el ANOVA factorial se sirve de estadísticos *F* basados en la lógica ya expuesta en el capítulo anterior al estudiar el modelo de un factor.

Así pues, para cada efecto existe una hipótesis y para cada hipótesis un estadístico *F* que permite contrastarla. Y, al igual que en el ANOVA de un factor, el nivel crítico asociado a cada estadístico *F* es el que permite decidir si se puede mantener o se debe rechazar una hipótesis.

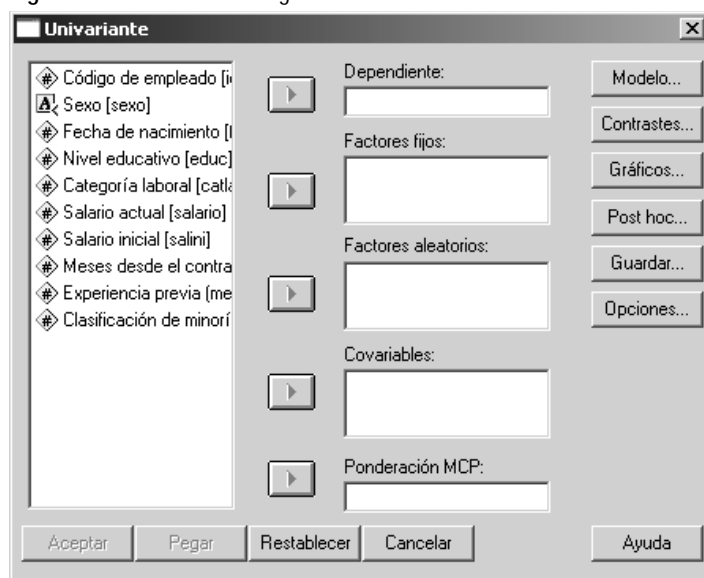


En un ANOVA factorial se trabaja con tantas poblaciones como casillas resultan de la combinación de todos los niveles de los factores involucrados. Por ejemplo, en un ANOVA de dos factores, con 2 niveles en un factor y 3 en otro, se trabaja con las  $2 \times 3 = 6$  poblaciones. El modelo asume que esas 10 poblaciones son normales y homocedásticas. También asume que las observaciones han sido aleatoriamente seleccionadas (una muestra de cada población) y que, por tanto, son independientes entre sí.

Para llevar a cabo un análisis de varianza con más de un factor:

- Seleccionar la opción **Modelo lineal general > Univariante...** del menú **Analizar** para acceder al cuadro de diálogo *Univariante* que muestra la Figura 15.1.

Figura 15.1. Cuadro de diálogo *Univariante*



La lista de variables contiene un listado con todas las variables del archivo de datos, incluidas las que poseen formato de cadena. Para obtener un ANOVA factorial con las especificaciones que el procedimiento **Univariate** tiene establecidas por defecto:

- Seleccionar una variable *cuantitativa* (de intervalo o razón, y, por tanto, con formato numérico) y trasladarla al cuadro **Dependiente**.
- Seleccionar dos o más variables *categorías* (nominales u ordinales; con formato numérico o de cadena, indistintamente) y trasladarlas a las listas **Factores fijos** o **Factores aleatorios**.

**Factores fijos.** Un factor de *efectos fijos* es aquel cuyos niveles los establece (fija) el investigador (por ejemplo, *cantidad de fármaco*; con niveles establecidos en: 0 mg, 100 mg, 250 mg) o vienen dados por la propia naturaleza del factor (por ejemplo, *sexo*; con niveles: varones y mujeres). Los niveles concretos de un factor de efectos fijos constituyen la población de niveles sobre los que se hace inferencia.

**Factores aleatorios.** Un factor de *efectos aleatorios* es aquel cuyos niveles son seleccionados de forma aleatoria entre todos los posibles niveles del factor (por ejemplo, *cantidad de fármaco*, con niveles 17 mg, 172 mg y 223 mg, obtenidos seleccionándolos aleatoriamente entre todos los posibles niveles de 0 a 250 mg). Los niveles concretos que toma un factor de efectos aleatorios constituyen sólo una muestra de la población de niveles sobre los que interesa hacer inferencia.

**Covariables.** En el caso de que se desee llevar a cabo un *análisis de covarianza*, la(s) covariable(s) debe(n) trasladarse a esta lista (ver, más adelante, el apartado *Análisis de covarianza*).

**Ponderación MCP.** El modelo lineal general asume que la varianza de la variable dependiente es la misma en todas las poblaciones objeto de estudio (en un diseño factorial hay tantas poblaciones como casillas resultan de combinar los niveles de los factores). Cuando las varianzas poblacionales no son iguales (por ejemplo, cuando las casillas con puntuaciones mayores muestran más variabilidad que las casillas con puntuaciones menores), el método de mínimos cuadrados no consigue ofrecer estimaciones óptimas. En estos casos, si las diferentes variabilidades de las casillas se pueden estimar o pronosticar a partir de alguna variable, el método de mínimos cuadrados ponderados (MCP) permite tener en cuenta esa variable de ponderación al estimar los parámetros de un modelo lineal, dando más importancia a las observaciones más precisas (es decir, a aquéllas con menos variabilidad). La variable de ponderación debe trasladarse al cuadro **Ponderación MCP**.

### **Ejemplo: MLG > Univariante**

Este ejemplo muestra cómo llevar a cabo un ANOVA factorial con las especificaciones que el procedimiento **Univariante** tiene establecidas por defecto. Siguiendo con el archivo *Datos de empleados* (que se encuentra en la misma carpeta en la que está instalado el SPSS) se desea estudiar si los grupos definidos por la variable *catlab* (categoría laboral) y los definidos por la variable *minoría* (clasificación de minorías) difieren en la variable *salario* (salario actual). Para ello:

- En el cuadro de diálogo principal (ver Figura 15.1) seleccionar la variable *salario* y trasladarla al cuadro **Dependiente**.
- Seleccionar las variables *catlab* y *minoría* y trasladarlas a la lista **Factores Fijos**.

Aceptando estas selecciones, el *Visor de resultados* ofrece la información que recogen las Tablas 15.1 y 15.2.

La Tabla 15.1 muestra el nombre de las variables independientes (factores), sus niveles, incluidas las etiquetas de los valores, y el tamaño de cada grupo (*N*).

**Tabla 15.1.** Factores inter-sujetos

		Etiqueta del valor	N
Categoría laboral	1	Administrativo	363
	2	Seguridad	27
	3	Directivo	84
Clasificación de minorías	0	No	370
	1	Sí	104

La tabla resumen del ANOVA (Tabla 15.2) contiene la misma información que la tabla resumen del modelo de un factor: las fuentes de variación, las sumas de cuadrados, los grados de libertad (*gl*), las medias cuadráticas, los estadísticos *F* y los niveles críticos (*Sig.*) asociados a cada estadístico *F*. Pero, ahora, toda esa información está referida no sólo a un factor, sino a los tres efectos presentes en un modelo de dos factores.

**Tabla 15.2.** Resumen del ANOVA. Contrastes de los efectos inter-sujetos

Variable dependiente: Salario actual

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Modelo corregido	90.341.824.425,07 <sup>a</sup>	5	18.068.364.885,01	177,74	,000
Intersección	153.679.697.553,40	1	153.679.697.553,40	1.511,77	,000
catlab	25.962.509.537,76	2	12.981.254.768,88	127,70	,000
minoría	237.964.814,40	1	237.964.814,40	2,34	,127
catlab * minoría	788.578.413,07	2	394.289.206,54	3,88	,021
Error	47.574.671.011,27	468	101.655.279,94		
Total	699.467.436.925,00	474			
Total corregida	137.916.495.436,34	473			

a. R cuadrado = ,655 (R cuadrado corregida = ,651)

La fila *Modelo corregido* se refiere a todos los efectos del modelo tomados juntos (el efecto de los dos factores, el de la interacción y el de la constante o intersección). El nivel crítico asociado al estadístico *F* (*Sig.* < 0,0005) está indicando que el modelo explica una parte significativa de la variabilidad de la variable dependiente (*salario*). El valor  $R^2=0,655$  que se ofrece en una nota a pie de tabla (éste valor se obtiene dividiendo la suma de cuadrados del *Modelo corregido* entre la suma de cuadrados *Total corregida*) indica que los tres efectos incluidos en el modelo (*catlab*, *minoría* y *catlab\*minoría*) están explicando el 65,5 % de la varianza de la variable dependiente *salario* (o, de otra forma, que los tres efectos incluidos en el modelo son capaces de predecir el 65,5 % del *salario*).

La fila *Intersección* se refiere a la constante del modelo. Esta constante forma parte del modelo y es necesaria para obtener las estimaciones de las medias de las casillas, pero no contiene información útil sobre los efectos incluidos en el modelo.

Las dos filas siguientes recogen los efectos principales, es decir, los efectos individuales de los dos factores incluidos en el modelo: *catlab* y *minoría*. Los niveles críticos (*Sig.*) indican que, mientras los grupos definidos por la variable *catlab* poseen salarios medios significativamente diferentes (*Sig.* < 0,0005), los salarios medios de los grupos definidos por la variable *minoría* no difieren (*Sig.* = 0,127).

La siguiente fila contiene información sobre el efecto de la interacción *catlab\*minoría*. El estadístico *F* correspondiente a este efecto tiene asociado un nivel crítico de 0,021, lo que indica que el efecto de la interacción es significativo. Sólo con este dato, ya se puede anticipar que las diferencias salariales que se dan entre las distintas categorías laborales (diferencias de las que informa el efecto individual del factor *catlab*) no son las mismas en los dos grupos étnicos considerados. Más adelante se precisará el significado del efecto de la interacción a través de una serie de estadísticos y de un gráfico de perfil.

La fila *Error* ofrece información referida a la fuente de variación *error* o *residual*. La media cuadrática *error* (que vale 101.655.279,939 y que es el divisor en cada cociente *F*), es un estimador insesgado de la varianza de las 6 poblaciones estudiadas, la cual se asume que es la misma en todas ellas.

La penúltima fila (*Total*) muestra la suma de los cuadrados de la variable dependiente (información ésta que, de momento, carece de utilidad); sus grados de libertad son el número total de casos utilizados en el análisis. La última fila (*Total corregida*) recoge la variación total: la suma de las variaciones de los dos factores, de la interacción y del término error.

Los resultados de un ANOVA siempre se entienden mejor con una representación gráfica de las medias. Las Figuras 15.2 y 15.3 ofrecen dos formas alternativas de representar las medias del *salario* en cada combinación de *catlab* por *minoría*. Las cajas de la Figura 15.2 muestran que la mediana del *salario* es mayor en el grupo de directivos que en el resto de categorías laborales; pero también muestran que el grado de dispersión y asimetría de las distribuciones del *salario* es más evidente en el grupo de administrativos y directivos que en el de agentes de seguridad; bajo estas circunstancias, podría considerarse la posibilidad de transformar las puntuaciones originales para obtener distribuciones menos asimétricas y con varianzas menos desiguales (los diagramas de caja de la Figura 15.2 se han obtenido mediante la opción **Diagramas de caja > Agrupado** del menú **Gráficos**).

Las barras de error de la Figura 15.3 permiten hacerse una idea más precisa del grado de solapamiento existente entre las medias salariales de cada subgrupo incluido en el análisis (se han obtenido mediante la opción **Barras de error > Agrupado** del menú **Gráficos**).

Figura 15.2. Diagramas de caja de *salario* en cada nivel de *catlab* por *minoría*

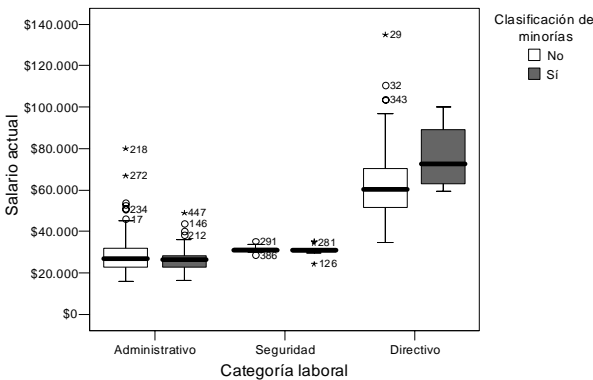
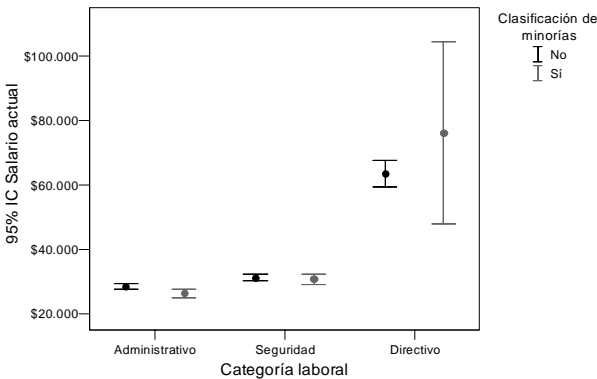


Figura 15.3. Barras de error de *salario* en cada nivel de *catlab* por *minoría*



Más adelante se ofrece una representación alternativa de las medias: los gráficos de líneas o gráficos de perfil (ver apartado *Gráficos de perfil para la interacción*). De hecho, el efecto de la interacción entre factores se entiende con mayor claridad cuando las medias se representan mediante un gráfico de líneas.

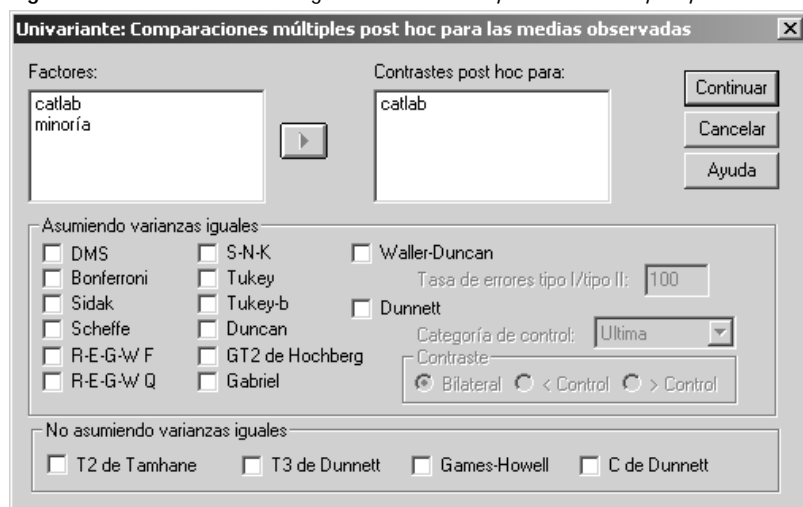
## Comparaciones *post hoc* o *a posteriori*

Si alguno de los estadísticos  $F$  asociados a los efectos principales resulta significativo, puede interesar efectuar comparaciones *post hoc*. Ya se ha explicado que los estadísticos  $F$  del ANOVA sólo permiten contrastar la hipótesis general de que los promedios comparados son iguales. Al rechazar esa hipótesis se sabe que existen diferencias, pero no se sabe dónde están.

Para averiguar qué media en concreto difiere de qué otra es necesario utilizar un tipo particular de contrastes denominados comparaciones múltiples *post hoc* o comparaciones *a posteriori*. Estas comparaciones permiten controlar la *tasa de error* al efectuar varias comparaciones utilizando las mismas medias, es decir, permiten controlar la probabilidad de cometer errores tipo I al tomar varias decisiones sobre los mismos datos (recuérdese que los errores de tipo I se cometen cuando se decide rechazar una hipótesis nula que en realidad es verdadera y que, por tanto, no debería rechazarse). Para llevar acabo comparaciones *post hoc*:

- Pulsar el botón Post Hoc... del cuadro de diálogo principal (ver Figura 15.1) para acceder al subcuadro de diálogo *Univariante: Comparaciones múltiples post hoc* que muestra la Figura 15.4.

Figura 15.4. Subcuadro de diálogo *Univariante: Comparaciones múltiples post hoc*



Este subcuadro de diálogo permite seleccionar las variables independientes cuyos niveles interesa comparar y elegir entre una amplia variedad de métodos *post hoc*. Estos procedimientos son los mismos que los ya descritos en el capítulo anterior sobre ANOVA de un factor, en el apartado *Comparaciones post hoc*.

### Ejemplo: MLG > Univariante > Comparaciones post hoc

Este ejemplo muestra cómo obtener e interpretar las comparaciones múltiples *pot hoc* del procedimiento **Univariante**. Aunque no todos los métodos disponibles se basan en la misma lógica, todos ellos se obtienen e interpretan de la misma forma; de modo que basta con marcar uno o dos de estos métodos y estudiar los resultados que genera. Se sigue trabajando con el archivo *Datos de empleados* (ubicado en la misma carpeta en la que está instalado el SPSS).

- En el cuadro de diálogo principal (ver Figura 15.1), trasladar la variable *salario* al cuadro **Dependiente** y las variables *catlab* y *minoría* a la lista **Factores fijos**.
- Pulsar el botón **Post hoc...** para acceder al subcuadro de diálogo *Univariante: Comparaciones múltiples post hoc* (ver Figura 15.4), seleccionar la variable *catlab* en la lista **Factores** y trasladarla a la lista **Contrastes post hoc para**.
- Marcar la opción **Tukey** del recuadro **Asumiendo varianzas iguales** y la opción **Games-Howell** del recuadro **No asumiendo varianzas iguales**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor de resultados* ofrece la información que recogen las Tablas 15.3 y 15.4. La Tabla 15.3 muestra el resultado obtenido con los dos métodos solicitados: Tukey y Games-Howell. Las conclusiones a las que se llega con ambos métodos no son las mismas. Según el método de Tukey, el grupo de administrativos no difiere del grupo de agentes de seguridad ( $Sig. = 0,272$ ), pero estos dos grupos sí difieren significativamente del grupo de directivos ( $Sig. < 0,0005$  en ambos casos). Según el método de Games-Howell, todos los grupos difieren significativamente entre sí ( $Sig. < 0,0005$  en las tres comparaciones). Puesto que el método de Tukey asume varianzas poblacionales iguales y el de Games-Howell no, debe elegirse el método que se ajuste a las características de los datos (ver más adelante, en este mismo capítulo, el párrafo sobre *Pruebas de homogeneidad* del apartado *Opciones*).

**Tabla 15.3.** Comparaciones múltiples *post hoc*. Pruebas de *Tukey* y *Games-Howell*

Variable dependiente: Salario actual

	(I) Categoría laboral	(J) Categoría laboral	Diferencia entre medias (I-J)	Error típ.	Sig.	Intervalo de confianza al 95%	
						Límite inferior	Límite superior
DHS de Tukey	Administrativo	Seguridad	-3.100,35	2.011,23	,272	-7.829,14	1.628,44
		Directivo	-36.139,26*	1.220,75	,000	-39.009,47	-33.269,05
	Seguridad	Administrativo	3.100,35	2.011,23	,272	-1.628,44	7.829,14
		Directivo	-33.038,91*	2.230,51	,000	-38.283,28	-27.794,54
	Directivo	Administrativo	36.139,26*	1.220,75	,000	33.269,05	39.009,47
		Seguridad	33.038,91*	2.230,51	,000	27.794,54	38.283,28
Games-Howell	Administrativo	Seguridad	-3.100,35*	568,68	,000	-4.454,82	-1.745,88
		Directivo	-36.139,26*	2.029,91	,000	-40.977,01	-31.301,51
	Seguridad	Administrativo	3.100,35*	568,68	,000	1.745,88	4.454,82
		Directivo	-33.038,91*	2.031,84	,000	-37.881,37	-28.196,45
	Directivo	Administrativo	36.139,26*	2.029,91	,000	31.301,51	40.977,01
		Seguridad	33.038,91*	2.031,84	,000	28.196,45	37.881,37

Basado en las medias observadas.

\*. La diferencia de medias es significativa al nivel ,05.

La Tabla 15.4 ofrece un resumen (basado en el método de Tukey) del resultado obtenido con las comparaciones múltiples. En este resumen, los grupos cuyas medias no difieren entre sí están agrupados en el mismo subconjunto y los grupos cuyas medias difieren forman parte de subconjuntos diferentes. En el ejemplo, existe un primer subconjunto de grupos homogéneos formado por el grupo de *administrativos* y por el de *agentes de seguridad* (los dos grupos cuyas medias no difieren entre sí; *Sig.* = 0,222), y un segundo subconjunto formado por el grupo de *directivos* (que difiere de los dos grupos anteriores y que, obviamente, no difiere de sí mismo; *Sig.* = 1,00).

Según se ha señalado ya en el capítulo anterior, esta clasificación en subgrupos homogéneos no está disponible con todos las pruebas *post-hoc*, sino sólo con algunas: *S-N-K*, Tukey, Tukey-b, Duncan, Scheffé, Gabriel, *R-E-G-W F* y *Q*, GT2 de Hochberg y Waller-Duncan. Esta es la razón por la cual, a pesar de que no es posible asumir varianzas poblacionales iguales (ver Tabla 15.10) la clasificación en subgrupos homogéneos de la Tabla 15.4 se ha realizado utilizando el método de Tukey en lugar del de Games-Howell.

Tabla 15.4. Subgrupos homogéneos

	Categoría laboral	N	Subconjunto	
			1	2
DHS de Tukey <sup>a,b,c</sup>	Administrativo	363	27.838,54	63.977,80 1,000
	Seguridad	27	30.938,89	
	Directivo	84		
	Significación		,223	

Basado en la suma de cuadrados tipo III

El término error es la Media cuadrática (Error) = 101655279,939.

a. Usa el tamaño muestral de la media armónica = 58,031

b. Los tamaños de los grupos son distintos. Se empleará la media armónica de los tamaños de los grupos. No se garantizan los niveles de error tipo I.

c. Alfa = ,05.

## Gráficos de perfil para la interacción

Las comparaciones múltiples *post hoc* suelen proporcionar toda la información necesaria para poder interpretar correctamente un efecto principal significativo. Pero no ocurre lo mismo con los efectos de las interacciones. La interpretación correcta de una interacción suele requerir (además del análisis de los efectos *simples*, que se explicará más adelante) la ayuda de un gráfico de líneas, también llamado gráfico de perfil.

En un gráfico de perfil sobre la interacción entre dos factores, en el eje vertical está representada la escala de las medias de la variable dependiente; en el eje horizontal están representados los niveles del primer factor; y las líneas del gráfico representan los niveles del segundo factor. Para representar una interacción triple es necesario hacer un gráfico de perfil para cada interacción doble en cada nivel del tercer factor. Ir más allá de las interacciones triples no suele tener mucho sentido, entre otras cosas, porque no resulta nada fácil la interpretación.

Para obtener gráficos de perfil representando el efecto de las interacciones:

- Pulsar el botón Gráficos... del cuadro de diálogo principal (ver Figura 15.1) para acceder al subcuadro de diálogo *Univariante: Gráficos de perfil* que muestra la Figura 15.5.

Figura 15.5. Subcuadro de diálogo *Univariante: Gráficos de perfil*

Este cuadro de diálogo permite obtener gráficos de perfil para combinaciones de dos y tres factores. Para obtener un gráfico de perfil referido a una **interacción doble**:

- Trasladar a los cuadros **Eje horizontal** y **Líneas distintas** los factores cuya interacción se desea representar.
- Pulsar el botón **Añadir** para hacer efectiva la selección.
- Utilizar los botones **Cambiar** y **Borrar** para modificar o eliminar combinaciones previamente añadidas.

Para obtener un gráfico de perfil referido a una **interacción triple**:

- Trasladar el tercer factor al cuadro **Gráficos distintos** y pulsar el botón **Añadir**.

### **Ejemplo: MLG Univariante > Gráficos de perfil**

Continuando con el ejemplo anterior, este ejemplo muestra cómo obtener un gráfico de perfil para el efecto de la interacción entre dos factores: *catlab* y *minoría*.

- En el cuadro de diálogo principal (ver Figura 15.1), trasladar la variable *salario* al cuadro **Dependiente** y las variables *catlab* y *minoría* a la lista **Factores fijos**.
- Pulsar el botón **Gráficos...** para acceder al subcuadro de diálogo *Univariante: Gráficos de perfil* (ver Figura 15.5).
- Trasladar la variable *catlab* al cuadro **Eje horizontal** y la variable *minoría* al cuadro **Líneas distintas**.
- Pulsar el botón **Añadir** para hacer efectiva la selección de variables.

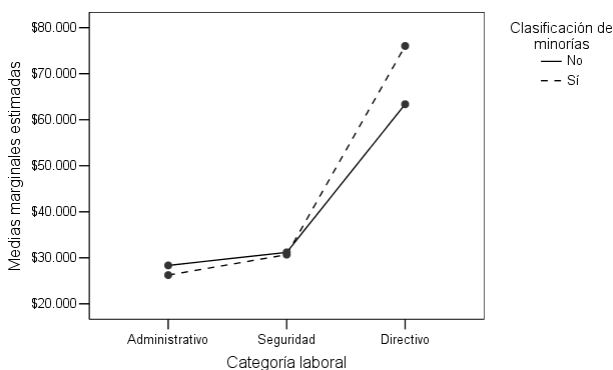
Aceptando estas elecciones, el *Visor de resultados* construye un gráfico de líneas como el que muestra la Figura 15.6.



En el gráfico aparecen representadas las *medias del salario actual* calculadas en cada subgrupo resultante de combinar cada nivel de la variable *categoría laboral* con cada nivel de la variable *clasificación de minorías*.

Una rápida inspección de las líneas aclara el significado de esta interacción: en principio, tanto en el grupo de blancos (*minoría* = «no») como en el de no blancos (*minoría* = «sí»), el salario medio parece mayor en los agentes de seguridad que en los administrativos, y mayor todavía en los directivos que en los agentes de seguridad; sin embargo, mientras en los dos primeros grupos (administrativos y agentes de seguridad) no parecen existir diferencias entre los dos grupos de *clasificación de minorías*, en el grupo de directivos las diferencias entre los dos grupos de *clasificación de minorías* parecen evidentes. Por tanto, las diferencias en salario entre los dos grupos de *clasificación de minorías* parece no ser la misma a lo largo de todas las *categorías laborales*.

Figura 15.6. Gráfico de perfil de *catlab* por *minoría*



Ese es el significado del efecto de la interacción. Pero una interpretación correcta del mismo no puede basarse en la apariencia del gráfico, sino que requiere utilizar *comparaciones múltiples* para determinar qué medias en concreto difieren de qué otras. Desafortunadamente, los cuadros de diálogo del procedimiento *Univariante* no contemplan la posibilidad de llevar a cabo estas comparaciones múltiples, por lo que es necesario utilizar la sintaxis *SPSS* (ver más adelante, en este mismo capítulo, el párrafo *Comparar los efectos principales* del apartado *Opciones*).

## Análisis de covarianza

El *análisis de covarianza* (ANCOVA) es una técnica de control estadístico que permite eliminar de la variable dependiente del ANOVA el efecto atribuible a variables no incluidas en el diseño como factores y, por tanto, no sometidas a control experimental.

La forma de controlar estadísticamente el efecto de estas variables *extrañas* consiste en efectuar un análisis de varianza utilizando como variable dependiente, no las puntuaciones originales de la variable dependiente, sino los errores en los pronósticos resultantes de llevar a cabo un *análisis de regresión lineal* (ver Capítulo 18) con las covariables como variables *independientes* y la propia variable dependiente del ANOVA como variable *dependiente*.

Para llevar a cabo un análisis de covarianza con las especificaciones que el programa tiene establecidas por defecto:

- En el cuadro de diálogo principal (ver Figura 15.1) seleccionar una variable *cuantitativa* (de intervalo o razón) y trasladarla al cuadro **Dependiente**.
- Seleccionar una o más variables *categorías* (nominales u ordinales) y trasladarlas a las listas **Factores fijos** o **Factores aleatorios**.
- Seleccionar la(s) variable(s) cuyo efecto se desea controlar y trasladarla(s) a la lista **Covariables** (sólo variables con formato numérico).

En un análisis de covarianza, los efectos de interés siguen siendo los referidos a cada factor y a las interacciones entre factores. Estos efectos se interpretan en los términos ya conocidos: un nivel crítico (*Sig.*) menor que 0,05 delata la presencia de un efecto significativo.

No obstante, el análisis de covarianza también permite evaluar el efecto individual de cada una de las covariables incluidas en el modelo: el procedimiento **Univariante** ofrece, para cada covariable, un estadístico *F* con su correspondiente nivel crítico. Este estadístico *F* permite contrastar la hipótesis nula de que el coeficiente de regresión correspondiente a una covariable vale cero en la población. Como consecuencia del contraste de esa hipótesis, se puede llegar, como en todo contraste, a dos conclusiones distintas: (1) que una covariable no posee efecto significativo, es decir, que no está linealmente relacionada con la variable dependiente; o (2) que una covariable posee efecto significativo, es decir, que está linealmente relacionada con la variable dependiente.

Una covariable que no posee efecto significativo es una covariable que puede ser eliminada del análisis. De hecho, si todas las covariables utilizadas poseen efectos no significativos, cabe esperar que los resultados del ANCOVA sean similares a los del ANOVA, indicando esto que no es necesario ejercer control sobre las covariables incluidas en el análisis.

Si una o más covariables poseen efectos significativos, pueden ocurrir dos cosas: que los resultados del ANOVA y los del ANCOVA sean los mismos, o que sean distintos. En primer lugar, si los resultados del ANOVA y los del ANCOVA son los mismos, esto significa que, a pesar de que una o más covariables correlacionan con la variable dependiente, y a pesar de que el efecto atribuible a esas covariables ha sido eliminado de la variación de la variable dependiente, el efecto de las variables independientes sobre la dependiente permanece inalterado; lo que significa que la relación entre la(s) covariable(s) y la variable dependiente no afecta a la relación entre los factores y la variable dependiente.

En segundo lugar, si los resultados del ANOVA y los del ANCOVA son distintos, puede ocurrir que lo sean por dos motivos: porque un efecto significativo del ANOVA ha pasado a ser no significativo en el ANCOVA, o porque un efecto no significativo del ANOVA ha pasado a ser significativo en el ANCOVA. En el primer caso (efecto significativo que deja de serlo), se puede interpretar que la relación detectada en el ANOVA entre ese efecto y la variable dependiente era espúrea, artificial y, probablemente, debida a las covariables incluidas en el análisis; y, en el segundo caso (efecto no significativo que pasa a serlo), se puede interpretar que la variable independiente en cuestión, aun no estando relacionada con la variable dependiente globalmente considerada, sí correlaciona con la parte de la variable dependiente que no está explicada o que no es atribuible a la(s) covariable(s).

Según se desprende de los párrafos anteriores, la interpretación apropiada de los resultados de un ANCOVA requiere utilizar como punto de referencia los resultados obtenidos en el correspondiente ANOVA.

### Ejemplo: MLG Univariante > Covariables

Este ejemplo muestra cómo llevar a cabo un análisis de covarianza con el procedimiento Univariante y cómo interpretar sus resultados.

En concreto, y siguiendo con el archivo *Datos de empleados*, se intenta averiguar si las diferencias observadas en *salario* (salario actual) entre los distintos subgrupos definidos por las variables *catlab* (categoría laboral) y *minoría* (clasificación de minorías) se mantienen al controlar el efecto (al introducir como *covariables*) de las variables *expprev* (experiencia previa) y *tiempemp* (meses desde el contrato). Para ello:

- En el cuadro de diálogo principal (ver Figura 15.1), trasladar la variable *salario* al cuadro **Dependiente** y las variables *catlab* y *minoría* a la lista **Factores fijos**.
- Trasladar las variables *expprev* y *tiempemp* a la lista **Covariables**.

Aceptando estas selecciones, el *Visor de resultados* ofrece la información que recoge la Tabla 15.5. La información referida a las covariables se encuentra en las filas encabezadas *tiempemp* y *expprev*. El resto de efectos son los ya vistos en la Tabla 15.2. Ambas covariables poseen estadísticos con niveles críticos (*Sig.*) menores que 0,05, por lo que puede afirmarse que ambas se encuentran linealmente relacionadas con la variable dependiente *salario*. Por tanto, parece que, en principio, tiene sentido haber incluido estas covariables en el análisis.

**Tabla 15.5.** Resumen del ANOVA (con covariables). Contrastes de los efectos inter-sujetos

Variable dependiente: Salario actual					
Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	92.046.574.962,81 <sup>a</sup>	7	13.149.510.708,97	133,59	,000
Intersección	6.558.230.692,62	1	6.558.230.692,62	66,63	,000
tiempemp	1.210.038.517,57	1	1.210.038.517,57	12,29	,000
expprev	492.908.058,49	1	492.908.058,49	5,01	,026
catlab	27.203.135.908,84	2	13.601.567.954,42	138,18	,000
minoría	299.670.835,00	1	299.670.835,00	3,04	,082
catlab * minoría	1.091.353.984,82	2	545.676.992,41	5,54	,004
Error	45.869.920.473,53	466	98.433.305,74		
Total	699.467.436.925,00	474			
Total corregida	137.916.495.436,34	473			

a. R cuadrado = ,667 (R cuadrado corregida = ,662)

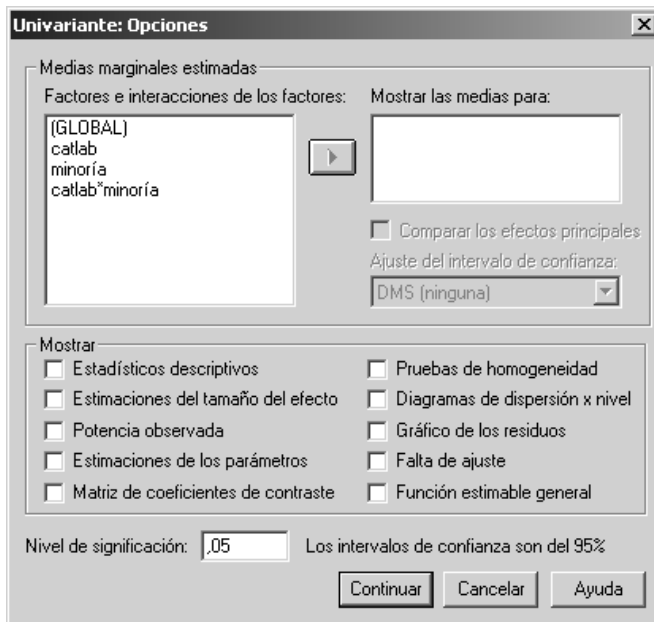
No obstante, puede comprobarse que, tras controlar el efecto de estas dos covariables, los tres efectos presentes en el modelo (*catlab*, *minoría* y la interacción *catlab\*minoría*) mantienen la misma significación que ya tenían en el ANOVA (ver Tabla 15.2) antes de controlar el efecto de las covariables: los efectos de *catlab* y de la interacción *catlab\*minoría* siguen siendo significativos y el efecto de *minoría* sigue siendo no significativo. Por tanto, aunque puede tener sentido controlar el efecto de las variables *tiempemp* y *expprev* (pues ambas están linealmente relacionadas con la variable dependiente *salario*), lo cierto es que los efectos de *catlab*, *minoría* y *catlab\*minoría*, que son los efectos que realmente interesa evaluar en un diseño de dos factores, no se alteran por ejercer tal control. Podría decirse que la relación existente entre las covariables y la variable dependiente no afecta (no altera) a la relación existente entre las variables independientes y la dependiente.

## Opciones

El subcuadro de diálogo *Opciones* permite obtener información relacionada con varios aspectos complementarios del análisis: estadísticos descriptivos, estimaciones del tamaño de los efectos, pruebas de homogeneidad, etc. Para obtener esta información:

- Pulsar el botón **Opciones...** del cuadro de diálogo principal (ver Figura 15.1) para acceder al subcuadro de diálogo *Univariante: Opciones* que muestra la Figura 15.7.

Figura 15.7. Subcuadro de diálogo *Univariante: Opciones*



**Medias marginales estimadas.** La lista **Factores e interacción de los factores** muestra un listado con todos los factores incluidos en el diseño y todas las posibles interacciones entre ellos.

Trasladando el efecto deseado a la lista **Mostrar las medias para**, el SPSS ofrece una estimación de las medias correspondientes a todos los niveles de ese efecto. Estas medias no son las observadas o empíricas, sino las estimadas a partir de los parámetros del modelo (ver más adelante, dentro de este mismo apartado, la opción *Estimaciones de los parámetros*). Las medias *observadas* son medias *ponderadas*:

$$\bar{Y}_j = \sum_k n_{jk} \bar{Y}_{jk} / \sum_k n_{jk}, \quad \bar{Y}_k = \sum_j n_{jk} \bar{Y}_{jk} / \sum_j n_{jk}, \quad \bar{Y}_{jk} = \sum_i Y_{ijk} / n_{jk}$$

Las medias *estimadas* son medias *no ponderadas*. Se estiman como si en cada casilla existiera una sola observación (ver Searle, Speed y Milliken, 1980):

$$\hat{\mu}_j = \sum_k \bar{Y}_{jk} / K, \quad \hat{\mu}_k = \sum_j \bar{Y}_{jk} / J, \quad \hat{\mu}_{jk} = \bar{Y}_{jk}$$

“ Comparar los efectos principales. Esta opción permite obtener todas las comparaciones dos a dos entre las medias correspondientes a los factores que han sido previamente trasladados a la lista **Mostrar medias para**. Estas comparaciones por pares se llevan a cabo con la prueba *T* para dos muestras independientes (ver Capítulo 13). Las opciones del menú desplegable **Ajuste del intervalo de confianza** permiten decidir si se desea o no ejercer control sobre la tasa de error (es decir, si se desea o no controlar la probabilidad de cometer errores de tipo I en el conjunto total de comparaciones). La opción por defecto, *DMS (ninguna)*, no ejerce control sobre la tasa de error. La opción *Bonferroni* controla la tasa de error multiplicando el nivel crítico concreto de cada comparación por el número de comparaciones que se están llevando a cabo entre las medias correspondientes a un mismo efecto. La opción *Sidak* corrige la tasa de error mediante  $1 - (1 - p_c)^k$ , donde  $p_c$  se refiere al nivel crítico de una comparación concreta y  $k$  al número de comparaciones.

El procedimiento **Univariante** también permite analizar los efectos *simples*. Es decir, permite comparar entre sí los niveles de un factor dentro de cada nivel del otro factor, lo cual es especialmente útil para interpretar el efecto de la interacción (si bien debe tenerse en cuenta que los efectos simples no agotan el significado de la interacción). Para analizar los efectos simples es necesario recurrir a la sintaxis SPSS. Para ello:

- En el cuadro de diálogo *Univariante: Opciones* (ver Figura 15.7), seleccionar el efecto que contiene la interacción (en el ejemplo, *catlab\*minoría*), junto con algún efecto principal, y trasladarlo a la lista **Mostrar las medias para**.
- Marcar la opción **Comparar los efectos principales** y pulsar el botón **Continuar** para volver al cuadro de diálogo *Univariante* (ver Figura 15.1).
- Pulsar el botón **Pegar** para escribir en el *Editor de sintaxis* la sintaxis SPSS correspondiente a las elecciones hechas y modificar la línea «EMMEANS = TABLES(catlab\*minoría)» añadiendo lo siguiente: «COMPARE(minoría) ADJ(BONFERRONI)».

Ejecutando la sintaxis se obtienen los resultados que muestran las Tablas 15.6 y 15.7. La Tabla 15.6 contiene las estimaciones de las medias de cada casilla (es decir, de cada combinación *catlab\*minoría*).

**Tabla 15.6.** Medias estimadas

Variable dependiente: Salario actual

Categoría laboral	Clasificación de minorías	Media	Error típ.	Intervalo de confianza al 95%.	
				Límite inferior	Límite superior
Administrativo	No	28.341,09	606,89	27.148,52	29.533,65
	Sí	26.244,25	1.080,95	24.120,14	28.368,37
Seguridad	No	31.178,57	2.694,64	25.883,48	36.473,67
	Sí	30.680,77	2.796,36	25.185,79	36.175,75
Directivo	No	63.374,81	1.127,25	61.159,72	65.589,91
	Sí	76.037,50	5.041,21	66.131,29	85.943,71

La Tabla 15.7 ofrece las comparaciones entre cada nivel de *minoría* (clasificación de minorías) dentro de cada nivel de *catlab* (categoría laboral). Los resultados incluyen el nivel crítico y el intervalo de confianza asociado a cada comparación (lógicamente, la corrección de Bonferroni no tiene efecto cuando, como en el ejemplo, únicamente se comparan los dos niveles de un factor dentro de cada nivel del otro factor). Los niveles críticos

obtenidos permiten afirmar que el salario medio de las dos minorías comparadas únicamente es distinto ( $p = 0,015$ ) en el grupo de directivos (ver Figura 15.6).

**Tabla 15.7.** Contrastes de los efectos simples

Variable dependiente: Salario actual

Categoría laboral	(I) Clasificación de minorías	(J) Clasificación de minorías	Diferencia entre medias (I-J)	Error típ.	Sig. <sup>a</sup>	Intervalo de confianza al 95 % para diferencia <sup>a</sup>	
						Límite inferior	Límite superior
Administrativo	No	Sí	2096,834	1239,664	,091	-339,163	4532,831
	Sí	No	-2096,834	1239,664	,091	-4532,831	339,163
Seguridad	No	Sí	497,802	3883,391	,898	-7133,240	8128,844
	Sí	No	-497,802	3883,391	,898	-8128,844	7133,240
Directivo	No	Sí	-12662,687*	5165,705	,015	-22813,54	-2511,840
	Sí	No	12662,687*	5165,705	,015	2511,840	22813,535

Basadas en las medias marginales estimadas.

\*. La diferencia de las medias es significativa al nivel ,05.

a. Ajuste para comparaciones múltiples: Bonferroni.

**Mostrar.** Este recuadro contiene información adicional:

- " **Estadísticos descriptivos.** Media, desviación típica y tamaño de cada nivel y de cada combinación de niveles. Esta tabla no se ofrece si se seleccionan más de 18 factores.
- " **Estimaciones del tamaño del efecto.** Estimaciones del grado en que cada factor o combinación de factores está afectando a la variable dependiente. El SPSS ofrece el estadístico *eta cuadrado parcial* (ver Tabla 15.8), que se obtiene, para un efecto concreto  $E$ , de la siguiente manera:  $(F_E \times gl_E) / (F_E \times gl_E + gl_{error})$ ; es decir, dividiendo el producto del estadístico  $F$  y de los grados de libertad de ese efecto entre ese mismo producto más los grados de libertad del error. Este estadístico se interpreta como *proporción de varianza explicada*: es una estimación de la proporción de varianza de la variable dependiente que está explicada o que es atribuible a cada efecto. El *Visor* muestra las estimaciones del tamaño de los efectos (*eta al cuadrado parcial*) en la propia tabla resumen del ANOVA.

**Tabla 15.8.** Resumen del ANOVA incluyendo las estimaciones del tamaño de los efectos

Variable dependiente: Salario actual

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta al cuadrado parcial	Potencia observada <sup>a</sup>
Modelo corregido	90.341.824.425,07 <sup>b</sup>	5	18.068.364.885,01	177,74	,000	,66	1,00
Intersección	153.679.697.553,40	1	153.679.697.553,40	1.511,77	,000	,76	1,00
catlab	25.962.509.537,76	2	12.981.254.768,88	127,70	,000	,35	1,00
minoría	237.964.814,40	1	237.964.814,40	2,34	,127	,00	,33
catlab * minoría	788.578.413,07	2	394.289.206,54	3,88	,021	,02	,70
Error	47.574.671.011,27	468	101.655.279,94				
Total	699.467.436.925,00	474					
Total corregida	137.916.495.436,34	473					

a. Calculado con alfa = ,05

b. R cuadrado = ,655 (R cuadrado corregida = ,651)

- " **Potencia observada.** Estimaciones de la potencia asociada al contraste de cada efecto. La potencia observada de un contraste se refiere a la capacidad de ese contraste para detectar una diferencia poblacional tan grande como la diferencia muestral de hecho observada. El SPSS calcula el valor de la potencia utilizando un nivel de significación de 0,05, pero este valor puede cambiarse mediante la opción **Nivel de significación** que se encuentra dentro de este mismo cuadro de diálogo. La potencia de cada efecto aparece en la tabla resumen del ANOVA (ver Tabla 15.8).
- " **Estimaciones de los parámetros.** Los modelos de ANOVA contienen una serie de parámetros a partir de los cuales se obtienen las medias que el modelo estima para cada nivel o combinación de niveles. Al marcar esta opción, el SPSS ofrece la información que recoge la Tabla 15.9.

**Tabla 15.9.** Estimaciones de los parámetros del modelo

Variable dependiente: Salario actual

Parámetro	B	Error típ.	t	Sig.	Intervalo de confianza al 95%.	
					Límite inferior	Límite superior
Intersección	76.037,50	5.041,21	15,08	,000	66.131,29	85.943,71
[catlab=1]	-49.793,25	5.155,80	-9,66	,000	-59.924,63	-39.661,86
[catlab=2]	-45.356,73	5.764,85	-7,87	,000	-56.684,92	-34.028,54
[catlab=3]	0 <sup>b</sup>	.	.	.	.	.
[minoría=0]	-12.662,69	5.165,71	-2,45	,015	-22.813,54	-2.511,84
[minoría=1]	0 <sup>b</sup>	.	.	.	.	.
[catlab=1] * [minoría=0]	14.759,52	5.312,37	2,78	,006	4.320,47	25.198,57
[catlab=1] * [minoría=1]	0 <sup>b</sup>	.	.	.	.	.
[catlab=2] * [minoría=0]	13.160,49	6.462,60	2,04	,042	461,18	25.859,80
[catlab=2] * [minoría=1]	0 <sup>b</sup>	.	.	.	.	.
[catlab=3] * [minoría=0]	0 <sup>b</sup>	.	.	.	.	.
[catlab=3] * [minoría=1]	0 <sup>b</sup>	.	.	.	.	.

a. Calculado con alfa = ,05

b. Al parámetro se le ha asignado el valor cero porque es redundante.

Las estimaciones de las medias se obtienen combinando los parámetros involucrados en la obtención de cada media. Por ejemplo, la estimación de la media de los *administrativos blancos* se obtiene sumando: el valor de la constante o *Intersección* (76.037,500), el valor correspondiente a *administrativos* ( $[CATLAB=1] = -49.793,247$ ), el valor correspondiente a *blancos* ( $[MINORÍA=0] = -12.662,687$ ) y el valor correspondiente a los *administrativos blancos* ( $[CATLAB=1] * [MINORÍA=0] = 14.759,522$ ). Se obtiene así un salario medio estimado de 28.341,09, que es el valor que muestra la Tabla 15.6 para los *administrativos blancos*. Puesto que las estimaciones de los parámetros suman cero para cada efecto, la tabla no recoge las estimaciones de los parámetros redundantes.

La Tabla 15.9 también muestra el error típico asociado a cada estimación y un estadístico *t* que permite contrastar la hipótesis de que un determinado parámetro vale cero en la población. Además, cada estimación aparece acompañada de su intervalo de confianza calculado al 95 %.

- " **Matriz de coeficientes de contraste.** Permite obtener la matriz **L** con los coeficientes asociados a cada efecto. Ver más adelante el apartado *Contrastes personalizados: La sentencia LMATRIX*.

- " **Pruebas de homogeneidad.** Ofrece el estadístico de Levene sobre homogeneidad de varianzas, el cual permite contrastar la hipótesis de que la varianza de la variable dependiente es la misma en el conjunto de poblaciones definidas por la combinación de factores. En el ejemplo, la Tabla 15.10 muestra, para el estadístico de Levene, un valor de 24,72, con un nivel crítico asociado menor que 0,0005. Puesto que el nivel crítico es muy pequeño, debe rechazarse la hipótesis de homogeneidad o igualdad de varianzas.

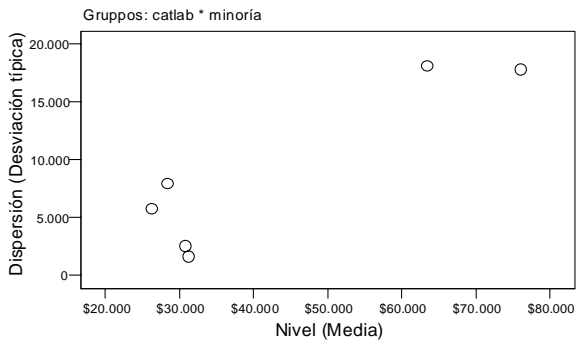
**Tabla 15.10.** Contraste de *Levene* sobre igualdad de varianzas

Variable dependiente: Salario actual

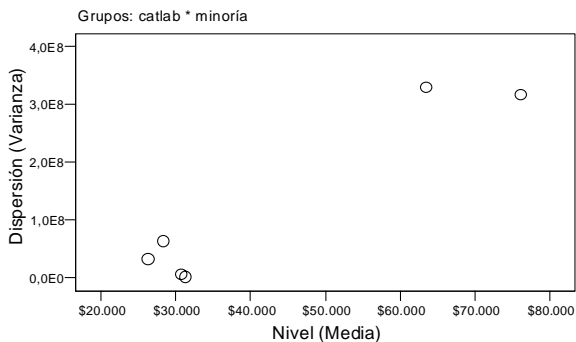
F	gl1	gl2	Significación
24,720	5	468	,000

- " **Diagramas de dispersión por nivel.** Los diagramas de *dispersión por nivel* ofrecen información gráfica sobre el grado de parecido existente entre las varianzas muestrales. Ayudan a detectar la posible existencia de algún tipo de relación entre el tamaño de las medias y el de las varianzas. Cuando las varianzas son iguales, los puntos del gráfico se encuentran a la misma altura, es decir, alineados horizontalmente. Las Figuras 15.8.a y 15.8.b muestran estos gráficos referidos a las variables *salario actual* (dispersión) y *categoría laboral por minoría* (nivel).

**Figura 15.8.a.** Diagrama de dispersión (desviación típica) por nivel. Variable dependiente: *salario*



**Figura 15.8.b.** Diagrama de dispersión (varianza) por nivel. Variable dependiente: *salario*

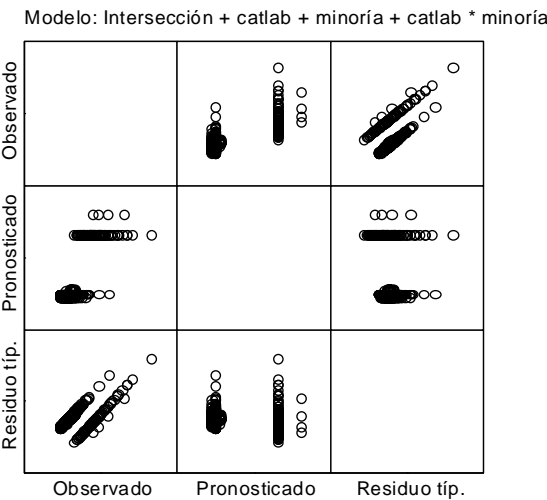




Los gráficos muestran 6 puntos, uno por cada nivel resultante de combinar *catlab* y *minoría*. El hecho de que los puntos no se encuentren horizontalmente alineados indica que las varianzas no son homogéneas (lo cual coincide con la información ofrecida por el estadístico de Levene). Aunque la tendencia no es del todo clara, parece que las casillas-niveles con medias más grandes son también las casillas-niveles que muestran mayor variación (las dos casillas-niveles con medias más altas muestran mayor variación que el resto de las casillas-niveles).

" **Gráfico de los residuos.** En el contexto del modelo lineal general, los residuos son las diferencias existentes entre los valores observados (es decir, las puntuaciones obtenidas en la variable dependiente) y los valores pronosticados por el modelo (existe un pronóstico por cada combinación de niveles, es decir, por cada casilla). En los modelos de ANOVA se asume que los residuos son independientes entre sí y que se distribuyen de forma aproximadamente normal. Pero además se asume, según se ha señalado ya, que las varianzas poblacionales de cada combinación de niveles son homogéneas. El gráfico de los residuos (ver Figura 15.9) ayuda a formarse una idea sobre el cumplimiento de algunos de estos supuestos (independencia, homogeneidad de varianzas) y sobre la bondad del modelo.

**Figura 15.9.** Gráfico de los residuos



Si los residuos son independientes, el gráfico correspondiente a la relación entre los valores *pronosticados* y los *residuos tipificados* no debe mostrar ninguna pauta de variación sistemática (una línea, una curva, etc.). Y si las varianzas son homogéneas, la dispersión de la nube de puntos correspondiente a los *residuos tipificados* debe ser homogénea a lo largo de todos los valores *pronosticados*. Del gráfico de la Figura 15.9 se desprende que, aunque los residuos parecen independientes (pues no muestran una pauta de variación sistemática), la dispersión de los mismos no es la misma a lo largo de todos los valores *pronosticados*.

Cuando el modelo aplicado ofrece un buen ajuste a los datos, la nube de puntos referida a la relación entre los valores *observados* y los *pronosticados* muestra una pauta de

relación claramente lineal. Lógicamente, la pauta es tanto más lineal cuanto mejor se ajusta el modelo a los datos.

**Nivel de significación.** Esta opción permite modificar el nivel de significación con el que se construyen los intervalos de confianza y con el que se calcula la potencia observada (el valor por defecto es 0,05).

## Contrastes personalizados

Además de las comparaciones *post hoc* (que sirven para interpretar los efectos *principales* comparando dos a dos los niveles de un factor) y de las comparaciones referidas a los efectos *simples* (que sirven para interpretar el efecto de la interacción comparando los niveles de un factor dentro de cada nivel del otro factor), el procedimiento **Univariante** también ofrece la posibilidad de llevar a cabo contrastes planeados o *a priori*, incluyendo comparaciones de *tendencia*.

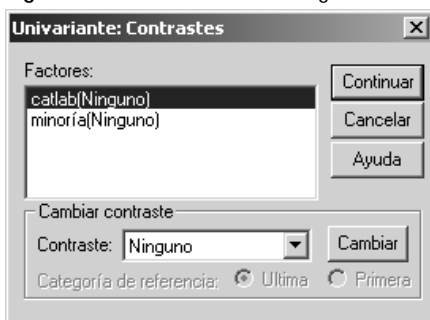
Algunos de estos contrastes pueden obtenerse a partir de las opciones del subcuadro de diálogo *Univariante: Contrastes*. El resto de contrastes pueden obtenerse mediante sintaxis con la sentencia LMATRIX.

## Contrastes predefinidos

Para llevar a cabo los contrastes que el procedimiento tiene predefinidos:

- Pulsar el botón **Contrastes...** del cuadro de diálogo principal (ver Figura 15.1) para acceder al subcuadro de diálogo *Univariante: Contrastes* que muestra la Figura 15.10.

Figura 15.10. Subcuadro de diálogo *Univariante: Contrastes*



La lista **Factores** contiene un listado con los factores previamente seleccionados. Por defecto, los factores no tienen asignado ningún tipo de contraste. Para asignar un tipo de contraste se debe utilizar el menú desplegable **Contraste** para seleccionar el contraste deseado y pulsar el botón **Cambiar** para validar la selección hecha. Con cualquiera de los contrastes de la lista se obtienen  $k-1$  comparaciones entre los  $k$  niveles de un factor; pero cada contraste define un tipo particular de comparaciones:

- **Desviación.** Todas las categorías (niveles) del factor, excepto una (la última, por defecto), se comparan con la media total, es decir, con la media de todas las categorías. En lugar de la última categoría, puede omitirse la primera seleccionando la opción **Primera** en **Categoría de referencia**. Para omitir una categoría distinta de la primera o la última hay que pegar la sintaxis después de marcar las opciones deseadas y añadir en la línea **CONTRAST**, detrás del nombre del contraste, entre paréntesis, el número de orden (1, 2, 3, etc.) correspondiente a la categoría que se desea omitir. Los números de orden de las categorías deben ser 1, 2, 3, etc., aunque los valores originales de las categorías sean, por ejemplo, 2, 4, 7, etc.). Si se desea omitir, por ejemplo, la segunda categoría del factor *catlab*, la línea **CONTRAST** debe quedar de esta manera: «Contrast (catlab) = deviation(2)».
- **Simple.** Cada categoría se compara con la categoría de referencia. La categoría de referencia puede ser la primera o la última. Para seleccionar una categoría de referencia distinta de la primera o la última es necesario utilizar la sintaxis en la forma descrita en el párrafo anterior.
- **Diferencia.** Cada categoría, excepto la primera, se compara con la media de las categorías anteriores. En los diseños equilibrados, las  $k-1$  comparaciones de este contraste son ortogonales.
- **Helmert.** Cada categoría, excepto la última, se compara con la media de las categorías posteriores. En los diseños equilibrados, las  $k-1$  comparaciones de este contraste son ortogonales.
- **Repetido.** Cada categoría, excepto la primera, se compara con la categoría anterior.
- **Polinómico.** Comparaciones de tendencia. La primera comparación corresponde a la tendencia lineal; la segunda, a la tendencia cuadrática; etc. En un diseño equilibrado, los contrastes polinómicos son ortogonales. Para más detalles sobre este tipo de comparaciones puede consultarse el párrafo *Polinómico* del apartado *Comparaciones planeadas o a priori*, en el capítulo anterior sobre *ANOVA de un factor*.
- **Especial** (sólo disponible mediante sintaxis). Además de las opciones disponibles en el cuadro de diálogo *Univariante: Contrastes*, también existe la posibilidad de definir cualquier otra comparación entre categorías que pueda resultar de interés. Para ello hay que utilizar la instrucción **CONTRAST** seguida de la especificación **SPECIAL**.  
Para comparar, por ejemplo, la primera categoría de la variable *catlab* con la segunda, y la primera con la tercera, debe utilizarse la siguiente línea de sintaxis: «Contrast (catlab) = special (1 -1 0 1 0 -1)». Para más detalles sobre cómo asignar coeficientes puede consultarse el párrafo *Coeficientes* del apartado *Comparaciones planeadas o a priori*, en el capítulo anterior sobre *ANOVA de un factor*.

## La sentencia **LMATRIX**

Al margen de las comparaciones que es posible llevar a cabo desde los cuadros de diálogo, la sentencia **LMATRIX** permite efectuar cualquier tipo de comparación mediante sintaxis. En el ejemplo que se viene utilizando en este capítulo se ha encontrado un efecto significativo

de la interacción *catlab\*minoría* (ver Tabla 15.3). Comparar el *salario* de las dos *minorías* en cada *categoría laboral* (o el de las distintas *categorías laborales* en cada *minoría*) puede ayudar a entender este efecto. Para efectuar estas comparaciones puede utilizarse la sentencia LMATRIX, la cual permite llevar a cabo contrastes personalizados asignando valores a los coeficientes de la matriz **L** en la hipótesis general  $\mathbf{LB} = \mathbf{0}$  (**B** es el vector de parámetros).

El modelo matemático correspondiente a un diseño de dos factores de efectos fijos, completamente al azar, adopta la forma (ver capítulo anterior, apartado *Modelos de ANOVA*):

$$\mu_{jk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk}$$

(*j* se refiere a los niveles del primer factor y *k* a los niveles del segundo factor). En el ejemplo sobre *salario*, *catlab*, y *minoría*, el modelo puede representarse mediante:

$$\text{salario}_{jk} = \text{constante} + \text{catlab}_j + \text{minoría}_k + \text{catlab*minoría}_{jk}$$

(*j* = 1, 2, 3; *k* = 1, 2,). La parte izquierda de la ecuación recoge los pronósticos del modelo, es decir, el salario medio que el modelo pronostica para cada combinación entre los niveles de los factores. La parte derecha de la ecuación recoge las dos variables independientes, la interacción entre ambas y los parámetros del modelo. El modelo incluye 12 parámetros: la constante, los tres correspondientes a los niveles de *catlab*, los dos correspondientes a los niveles de *minoría* y los seis correspondientes a las combinaciones entre los tres niveles de *catlab* y los dos de *minoría*. Es decir, el vector de parámetros **B** incluye los siguientes parámetros:

$$\mathbf{B}' = (\text{constante}, \text{catlab}_1, \text{catlab}_2, \text{catlab}_3, \text{minoría}_1, \text{minoría}_2, \\ \text{catlab*minoría}_{11}, \text{catlab*minoría}_{12}, \text{catlab*minoría}_{13}, \\ \text{catlab*minoría}_{21}, \text{catlab*minoría}_{22}, \text{catlab*minoría}_{23})$$

Y la matriz de coeficientes **L** incluye el peso o coeficiente asignado a cada parámetro del modelo:

$$\mathbf{L} = (l_1, l_2, l_3, l_4, l_5, l_6, l_7, l_8, l_9, l_{10}, l_{11}, l_{12})$$

Para definir contrastes personalizados basta con especificar los valores que deben tomar los coeficientes de la matriz **L** en la expresión  $\mathbf{LB}$ :

$$\mathbf{LB} = l_1 \text{constante} + l_2 \text{catlab}_1 + l_3 \text{catlab}_2 + l_4 \text{catlab}_3 + l_5 \text{minoría}_1 + l_6 \text{minoría}_2 + \\ l_7 \text{catlab*minoría}_{11} + l_8 \text{catlab*minoría}_{12} + l_9 \text{catlab*minoría}_{13} + \\ l_{10} \text{catlab*minoría}_{21} + l_{11} \text{catlab*minoría}_{22} + l_{12} \text{catlab*minoría}_{23}$$

La sentencia LMATRIX permite definir contrastes personalizados asignando a cada parámetro los coeficientes apropiados. Así, por ejemplo, para comparar las dos *minorías* en la primera categoría laboral, a los coeficientes  $l_5$  y  $l_7$  asociados a los parámetros correspondientes a la primera categoría de *minoría* ( $\text{minoría}_1$ ) y a la combinación de la primera categoría de *minoría* con la primera categoría laboral ( $\text{catlab*minoría}_{11}$ ) se les asigna un valor de 1; y a los coeficientes  $l_6$  y  $l_8$  asociados a los parámetros correspondientes a la segunda categoría de *minoría* ( $\text{minoría}_2$ ) y a la combinación de la segunda categoría de *minoría* con la primera categoría laboral ( $\text{catlab*minoría}_{12}$ ) se les asigna un valor de -1. Al resto de coeficientes se les asignan ceros para excluir del contraste los efectos que no intervienen en la comparación que se está llevando a cabo. Por tanto, la expresión  $\mathbf{LB}$  correspondiente a la comparación de las dos *minorías* en la primera categoría laboral queda de la siguiente manera:

$$\begin{aligned}\mathbf{LB} &= (1)minoría_1 + (1)catlab*minoría_{11} + (-1)minoría_2 + (-1)catlab*minoría_{12} \\ &= (minoría_1 - minoría_2) + (catlab*minoría_{11} - catlab*minoría_{12})\end{aligned}$$

En la primera parte de la expresión se están comparando las dos minorías; en la segunda parte se indica que esta comparación entre las dos minorías se refiere a la primera categoría laboral. De modo similar, la expresión **LB** correspondiente a la comparación de las dos minorías en la segunda categoría laboral adopta la forma:

$$\begin{aligned}\mathbf{LB} &= (1)minoría_1 + (1)catlab*minoría_{21} + (-1)minoría_2 + (-1)catlab*minoría_{22} \\ &= (minoría_1 - minoría_2) + (catlab*minoría_{21} - catlab*minoría_{22})\end{aligned}$$

Por último, la expresión **LB** correspondiente a la comparación de las dos minorías en la tercera categoría laboral adopta la forma:

$$\begin{aligned}\mathbf{LB} &= (1)minoría_1 + (1)catlab*minoría_{31} + (-1)minoría_2 + (-1)catlab*minoría_{32} \\ &= (minoría_1 - minoría_2) + (catlab*minoría_{31} - catlab*minoría_{32})\end{aligned}$$

Para llevar a cabo estos contrastes personalizados con la sentencia **LMATRIX**:

- En el cuadro de diálogo principal (ver Figura 15.1), seleccionar la variable *salario* (salario actual) y trasladarla al cuadro **Dependiente**.
- Seleccionar las variables *catlab* (categoría laboral) y *minoría* (clasificación de minorías) y trasladarlas a la lista **Factores fijos**.
- Pulsar el botón **Pegar** para obtener la sintaxis correspondiente a las elecciones hechas.

El *Editor de sintaxis* muestra el siguiente resultado:

```
UNIANOVA
salario BY catlab minoría
/METHOD = SSTYPE(3)
/INTERCEPT = INCLUDE
/CRITERIA = ALPHA(.05)
/DESIGN = catlab minoría catlab*minoría.
```

**METHOD** indica que se van a utilizar las sumas de cuadrados Tipo III; **INTERCEPT** recuerda que el modelo solicitado incluye la constante; **CRITERIA** establece el nivel de significación que se utilizará para construir los intervalos de confianza; y **DESIGN** recoge los efectos incluidos en el modelo. Los valores asignados a estas cuatro sentencias son los que el procedimiento **UNIANOVA** utiliza por defecto; por tanto, no es necesario incluirlos. Para poder efectuar contrastes personalizados es necesario incluir la sentencia **LMATRIX**:

```
UNIANOVA
salario BY catlab minoría
/LMATRIX = 'Comparaciones entre las dos minorías en cada categoría laboral'
minoría 1 -1 catlab*minoría 1 -1 0 0 0 0;
minoría 1 -1 catlab*minoría 0 0 1 -1 0 0;
minoría 1 -1 catlab*minoría 0 0 0 0 1 -1.
```

La expresión entre apóstrofes de la sentencia **LMATRIX** es una etiqueta descriptiva que servirá para identificar los resultados en el *Visor*. A continuación aparecen definidas las tres compa-

raciones entre las dos minorías dentro de cada categoría laboral. Los coeficientes asignados permiten definir cada comparación. En la primera línea, los coeficientes de la primera parte (minoría 1, -1) comparan las dos minorías (estos coeficientes son los que en la expresión **LB** están asociados a los efectos *minoría<sub>1</sub>* y *minoría<sub>2</sub>*); y los coeficientes de la segunda parte (*catlab\*minoría 1, -1, 0, 0, 0, 0*) indican que esa comparación entre las dos minorías debe hacerse dentro de la primera categoría laboral, pues los coeficientes 1 y -1 se han asignado a los efectos *catlab\*minoría<sub>11</sub>* y *catlab\*minoría<sub>12</sub>*. En la segunda línea, los coeficientes asignados (*catlab\*minoría 0, 0, 1, -1, 0, 0*) indican que la comparación entre las dos minorías debe hacerse dentro de la segunda categoría laboral, pues los coeficientes 1 y -1 se han asignado a los efectos *catlab\*minoría<sub>21</sub>* y *catlab\*minoría<sub>22</sub>*. En la tercera línea, los coeficientes asignados (*catlab\*minoría 0, 0, 0, 0, 1, -1*) indican que la comparación entre las dos minorías debe hacerse dentro de la tercera categoría laboral, pues los coeficientes 1 y -1 se han asignado a los efectos *catlab\*minoría<sub>31</sub>* y *catlab\*minoría<sub>32</sub>*.

Ejecutando esta sintaxis se obtienen, entre otros, los resultados que muestran las Tablas 15.11 y 15.12. Los resultados de la Tabla 15.11 son idénticos a los ya obtenidos en este mismo capítulo al estudiar los *efectos simples* con otra estrategia diferente (ver Tabla 15.7), con la diferencia de que ahora no se está aplicando la corrección de Bonferroni al calcular los niveles críticos ni al construir los intervalos de confianza. Los niveles críticos (*Sig.*) permiten concluir que el salario medio de las dos minorías únicamente es distinto entre los directivos (tercera categoría laboral, L3).

**Tabla 15.11.** Contrastes personalizados basados en la matriz L

Variable dependiente: Salario actual

Contraste <sup>a</sup>	Estimación del contraste	Valor hipotetizado	Diferencia (Estim - Hipotet)	Error típ.	Sig.	Intervalo de confianza al 95 %	
						Límite inferior	Límite superior
L1	2.096,83	0	2.096,83	1.239,66	,091	-339,16	4.532,83
L2	497,80	0	497,80	3.883,39	,898	-7.133,24	8.128,84
L3	-12.662,69	0	-12.662,69	5.165,71	,015	-22.813,54	-2.511,84

a. Basados en la matriz de coeficientes de contraste (L') definida por el usuario. Comparaciones entre las dos minorías en cada categoría laboral.

La Tabla 15.12 ofrece un contraste global de los tres contrastes personalizados definidos mediante la sentencia LMATRIX. El tamaño del nivel crítico obtenido (*Sig.* = 0,032) indica que, de los tres contrastes definidos, los grupos comparados difieren significativamente en al menos uno.

**Tabla 15.12.** Valoración global de los contrastes personalizados

Variable dependiente: Salario actual

Fuente	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Contraste	903.340.499,12	3	301.113.499,71	2,96	,032
Error	47.574.671.011,27	468	101.655.279,94		

Los resultados de la Tabla 15.11 son exactamente los mismos tanto si las tres comparaciones se definen utilizando una sola sentencia LMATRIX (como en el ejemplo) como si se utilizan tres sentencias LMATRIX distintas, una para cada comparación. Sin embargo, los resultados de la Tabla 15.12 dependen del número de sentencias LMATRIX que se incluyan en la sintaxis.

xis: el SPSS genera un contraste por cada sentencia LMATRIX. Por tanto, si se utiliza una sola sentencia LMATRIX para definir todas las comparaciones, el contraste de la Tabla 15.11 ofrece una valoración de todas las comparaciones planteadas. Por el contrario, si dentro de cada sentencia LMATRIX se comparan los niveles de un factor dentro de cada nivel del otro factor (se utilizan, por tanto, tantas sentencias LMATRIX como niveles tiene el segundo factor), el SPSS ofrece contrastes como el de la Tabla 15.12 en número igual al de niveles del segundo factor; cada uno de estos contrastes permite valorar el efecto del primer factor dentro de cada nivel del segundo (es decir, cada uno de estos contrastes permite valorar un *efecto simple*).

Puesto que la variable *minoría* sólo tiene dos niveles, sólo es necesario hacer una comparación entre minorías por cada categoría laboral (tres comparaciones en total); cada una de esas comparaciones capta el efecto de *minoría* en cada *categoría laboral*; si se utilizan tres sentencias LMATRIX para definir esas tres comparaciones, el SPSS genera tres contrastes como el de la Tabla 15.12 equivalentes a los tres contrastes de la Tabla 15.11. Sin embargo, puesto que la variable *categoría laboral* tiene tres niveles, para comparar cada nivel con cada otro es necesario hacer tres comparaciones. Por tanto, para contrastar el efecto de la *categoría laboral* en cada *minoría* es necesario hacer tres comparaciones por cada *minoría* (seis comparaciones en total). Estas comparaciones pueden efectuarse utilizando dos sentencias LMATRIX: una con las comparaciones referidas a la primera *minoría* («minoría=0») y otra con las comparaciones referidas a la segunda *minoría* («minoría=1»). La sintaxis correspondiente quedará de esta manera:

#### UNIANOVA

salario BY catlab minoría

/LMATRIX = 'Comparaciones entre las categorías laborales en minoría = 0'

catlab 1 -1 0 catlab\*minoría 1 0 -1 0 0 0;

catlab 1 0 -1 catlab\*minoría 1 0 0 0 -1 0;

catlab 0 1 -1 catlab\*minoría 0 0 1 0 -1 0

/LMATRIX = 'Comparaciones entre las categorías laborales en minoría = 1'

catlab 1 -1 0 catlab\*minoría 0 1 0 -1 0 0;

catlab 1 0 -1 catlab\*minoría 0 1 0 0 0 -1;

catlab 0 1 -1 catlab\*minoría 0 0 0 1 0 -1.

Ejecutando esta sintaxis se obtienen los resultados que muestran las Tablas 15.13 a la 15.16. La Tabla 15.13 ofrece las comparaciones entre las tres *categorías laborales* dentro de la primera categoría de *minoría* («minoría=0»). La nota a pie de tabla recoge la etiqueta incluida en la sintaxis. *L1* se refiere a la comparación entre las categorías 1 y 2 (administrativos con agentes de seguridad); *L2* se refiere a la comparación entre las categorías 1 y 3 (administrativos con directivos); y *L3* se refiere a la comparación entre las categorías 2 y 3 (agentes de seguridad con administrativos).

Los niveles críticos (*Sig.*) asociados a cada comparación indican que el grupo de directivos difiere tanto del de administrativos como del de agentes de seguridad (*Sig.* < 0,0005 en ambos casos), mientras que estos dos grupos no difieren entre sí (*Sig.* = 0,305). Por tanto, el salario medio del grupo «minoría=0» no es el mismo en las tres categorías laborales (no debe olvidarse que al efectuar estas comparaciones no se está utilizando ningún tipo de control sobre la tasa de error).

La Tabla 15.14 ofrece una valoración del efecto global de *categoría laboral* en la primera categoría de *minoría* («minoría=0»). El nivel crítico asociado al estadístico *F* (*Sig.* < 0,0005)

confirma que existe un efecto significativo de la *categoría laboral* dentro de la primera categoría de *minoría*.

**Tabla 15.13.** Contrastes personalizados basados en la matriz L («minoría = 0»)

Variable dependiente: Salario actual

Contraste <sup>a</sup>	Estimación del contraste	Valor hipotetizado	Diferencia (Estim. - Hipotet.)	Error típ.	Sig.	Intervalo de confianza al 95 %	
						Límite inferior	Límite superior
L1	-2.837,48	0	-2.837,48	2.762,14	,305	-8.265,21	2.590,24
L2	-35.033,73	0	-35.033,73	1.280,24	,000	-37.549,45	-32.518,00
L3	-32.196,24	0	-32.196,24	2.920,92	,000	-37.935,99	-26.456,50

a. Basada en la matriz de coeficientes de contraste (L') definida por el usuario: Comparaciones entre las categorías laborales cuando *minoría* = 0

**Tabla 15.14.** Valoración global de los contrastes personalizados («minoría = 0»)

Variable dependiente: Salario actual

Fuente	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Contraste	76.465.544.051,73	2	38.232.772.025,86	376,10	,000
Error	47.574.671.011,27	468	101.655.279,94		

La Tabla 15.15 ofrece las comparaciones entre las tres *categorías laborales* dentro de la segunda *minoría* («*minoría* = 1»). Al igual que en el caso anterior, los niveles críticos (*Sig.*) indican que el grupo de directivos difiere tanto del de administrativos como del de agentes de seguridad (*Sig.* < 0,0005 en ambos casos), mientras que estos dos grupos no difieren entre sí (*Sig.* = 0,140).

**Tabla 15.15.** Contrastes personalizados basados en la matriz L («minoría = 1»)

Variable dependiente: Salario actual

Contraste <sup>b</sup>	Estimación del contraste	Valor hipotetizado	Diferencia (Estim. - Hipotet.)	Error típ.	Sig.	Intervalo de confianza al 95 %	
						Límite inferior	Límite superior
L1	-4.436,52	0	-4.436,52	2.998,01	,140	-10.327,75	1.454,72
L2	-49.793,25	0	-49.793,25	5.155,80	,000	-59.924,63	-39.661,86
L3	-45.356,73	0	-45.356,73	5.764,85	,000	-56.684,92	-34.028,54

b. Basada en la matriz de coeficientes de contraste (L') definida por el usuario: Comparaciones entre las categorías laborales en *minoría* = 1

Por último, la Tabla 15.16 ofrece una valoración del efecto global de la *categoría laboral* en la segunda categoría de *minoría* («*minoría* = 1»). El valor del nivel crítico asociado al estadístico *F* (*Sig.* < 0,0005) confirma que, dentro del grupo «*minoría* = 1», existe al menos una categoría laboral que difiere de al menos otra; por tanto, puede afirmarse que existe un efecto significativo de la *categoría laboral* dentro de la segunda categoría de *minoría*.

**Tabla 15.16.** Valoración global de los contrastes personalizados («minoría = 1»)

Variable dependiente: Salario actual

Fuente	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Contraste	9539010589,909	2	4.769.505.294,95	46,92	,000
Error	47.574.671.011,27	468	101.655.279,94		



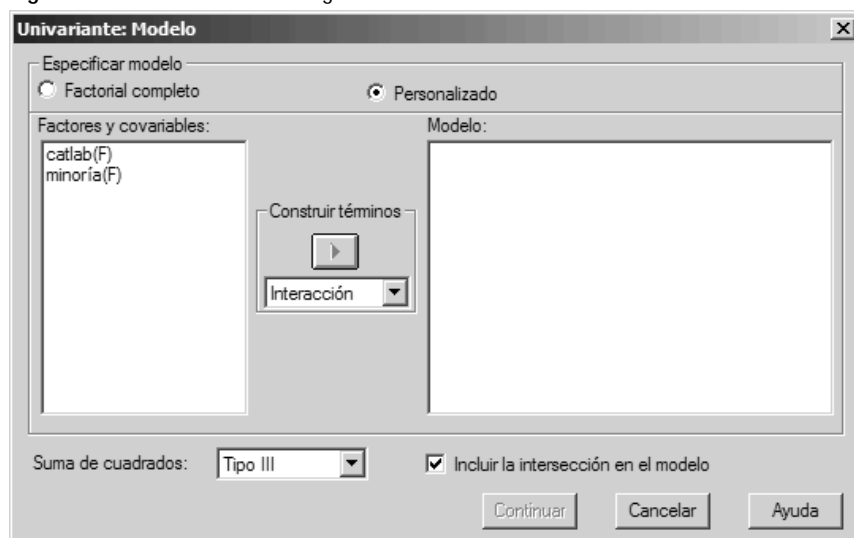
## Modelos personalizados

Además del modelo completamente aleatorizado, el procedimiento **Univariante** permite ajustar modelos de bloques aleatorios, modelos jerárquicos o anidados, análisis de regresión, etc. En este apartado se explica cómo proceder para obtener algunos de estos modelos.

Para ajustar un modelo personalizado:

- Pulsar el botón **Modelo...** del cuadro de diálogo principal (ver Figura 15.1) para acceder al subcuadro de diálogo *Univariante: Modelo* que muestra la Figura 15.11.

Figura 15.11. Subcuadro de diálogo *Univariante: Modelo*



**Factorial completo.** Opción válida para ajustar modelos completamente aleatorizados. Es decir, modelos en los que intervienen todos los efectos principales previamente definidos como factores y todas las posibles interacciones entre ellos. Es la opción que se encuentra activa por defecto.

**Personalizado.** Esta opción permite definir modelos distintos del completamente aleatorizado. Para definir un modelo concreto hay que seleccionar en la lista **Factores y covariables** los efectos deseados y trasladarlos a la lista **Modelo** utilizando el botón flecha y las opciones del menú desplegable del recuadro **Construir términos**.

**Sumas de cuadrados.** Este menú desplegable permite elegir entre cuatro métodos distintos de cálculo de las sumas de cuadrados: I, II, III y IV. Las sumas de cuadrados Tipo III son las más utilizadas y las que se obtienen por defecto:

- En las sumas de cuadrados **Tipo I** (método conocido como *descomposición jerárquica*), cada término se corrige sólo respecto al término que le precede en el modelo. Se utiliza normalmente en los modelos equilibrados en los que cualquier efecto principal

se evalúa antes que cualquier efecto de interacción de primer orden, cualquier efecto de interacción de primer orden se evalúa antes que cualquier efecto de interacción de segundo orden, y así sucesivamente. También se utiliza en los modelos anidados en los que el primer efecto está anidado dentro del segundo efecto el segundo efecto dentro del tercero, y así sucesivamente (esta forma de anidado sólo puede especificarse mediante sintaxis).

- Las sumas de cuadrados **Tipo II** se obtienen teniendo en cuenta únicamente los efectos pertinentes. Un efecto pertinente es un efecto que no está contenido en el efecto que se está evaluando. Para cualesquiera dos efectos  $E_1$  y  $E_2$ , se dice que  $E_1$  está contenido en  $E_2$  si se da alguna de estas tres condiciones: (1) ambos efectos tienen la misma covariable, (2)  $E_2$  consta de más términos que  $E_1$ , y (3) todos los términos presentes en  $E_1$  también lo están en  $E_2$ . Si sólo se evalúan efectos principales (es decir, si el modelo no incluye ningún término de interacción), cada efecto se ajusta teniendo en cuenta el resto de efectos presentes en el modelo (lo que equivale al método *regresión* de obtención de sumas de cuadrados). Las sumas de cuadrados Tipo II se utilizan normalmente en los modelos equilibrados, en los modelos que sólo contienen efectos principales (no interacciones), y también en los diseños anidados en los que cada efecto especificado está anidado sobre el anterior.
  - Las sumas de cuadrados **Tipo III** se calculan ajustando cada efecto teniendo en cuenta cualquier otro efecto que no lo contenga y de forma independiente de cualquier efecto que lo contenga, si existe. Estas sumas de cuadrados no se alteran por las variaciones del tamaño muestral de las casillas, de modo que son útiles para los modelos no equilibrados sin casillas vacías. También son apropiadas para cualquier modelo que lo sean las sumas de cuadrados Tipo I y Tipo II.
  - Las sumas de cuadrados **Tipo IV** son apropiadas para analizar tanto modelos equilibrados como no equilibrados cuando existen casillas vacías. Los niveles de un factor se comparan promediando uno o más niveles de otros factores. Estas comparaciones se llevan a cabo a partir de las combinaciones disponibles entre tratamientos.
- “ Incluir la intersección en el modelo. Esta opción permite decidir si el modelo que se desea ajustar debe o no incluir el término constante (la media total).

## Modelos con bloques aleatorios

Para construir, por ejemplo, un modelo con dos factores (*catlab* y *minoría*) aleatorizado en bloques (definidos por *sexo*), se debe tener en cuenta que, en un modelo de este tipo, el factor no interactúa con los bloques, y que, por tanto, los efectos presentes en el modelo son sólo los efectos principales de *catlab* y *minoría*.

Para definir un modelo con bloques aleatorios hay que trasladar a la lista **Modelo** las variables individuales pero no la interacción entre ambas. En general, para definir un modelo con bloques aleatorios:

- En el cuadro de diálogo principal (ver Figura 15.1), seleccionar la variable dependiente y trasladarla al cuadro **Dependiente**.

- Seleccionar las *variables-factores* (las variables independientes) y las *variables-bloques* (las variables que se desea utilizar para definir los bloques) y trasladarlas a la lista **Factores fijos**.
- Pulsar el botón **Modelo...** para acceder al subcuadro de diálogo *Univariante: Modelo* (ver Figura 15.11) y marcar la opción **Personalizado**.
- Seleccionar las *variables-factores* y las *variables-bloques* dentro de la lista **Factores y covariables**.
- Seleccionar **Efectos principales** dentro del menú desplegable **Construir términos** y pulsar el botón flecha para trasladar a la lista **Modelo** las variables seleccionadas.
- Seleccionar ahora, en la lista **Factores y covariables**, sólo las *variables-factores*.
- Seleccionar la opción **Todas de 2** dentro del menú desplegable **Construir términos** y pulsar el botón flecha para trasladar a la lista **Modelo** todas las posibles combinaciones entre cada dos *variables-factores* (dejando fuera las combinaciones entre las *variables-factores* y las *variables-bloques*, y dejando fuera también las combinaciones de las *variables-bloques* entre sí).
- Si se trata de un modelo con más de dos factores, seleccionar de nuevo en la lista **Factores y covariables** sólo las *variables-factores*. Después, seleccionar la opción **Todas de 3** dentro del menú desplegable **Construir términos** y pulsar el botón flecha para trasladar a la lista **Modelo** todas las posibles combinaciones entre cada tres *variables-factores*. Etc.

Lo característico de la tabla resumen del ANOVA que se obtiene al ajustar un modelo con bloques aleatorios es que la(s) variable(s) utilizada(s) para definir bloques no interactúa(n) con el resto de efectos incluidos en el modelo.

## Modelos jerárquicos o anidados

En los diseños jerárquicos uno de los factores está anidado en el otro factor. Esto significa que los niveles de uno de los factores son distintos en cada nivel del otro factor. Si, por ejemplo, las dos minorías del ejemplo anterior tuvieran categorías laborales distintas, el factor *categoría laboral* estaría anidado en el factor *minoría*. En este tipo de diseños no es posible evaluar el efecto de la interacción, pero sí los efectos principales. Para ello:

- En el cuadro de diálogo principal (ver Figura 15.1), seleccionar la variable dependiente y trasladarla al cuadro **Dependiente**.
- Seleccionar las variables independientes y trasladarlas a la lista **Factores fijos**.
- Pulsar el botón **Modelo...** para acceder al subcuadro de diálogo *Univariante: Modelo* (ver Figura 15.11) y marcar la opción **Personalizado**.
- Seleccionar las variables que definen tanto el factor no anidado y como el factor anidado dentro de la lista **Factores y covariables**.
- Seleccionar **Efectos principales** dentro del menú desplegable **Construir términos** y pulsar el botón flecha para trasladar a la lista **Modelo** esas variables. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

- Pulsar el botón **Pegar** para generar la sintaxis correspondiente a las selecciones hechas e ir al *Editor de sintaxis* para editar la sintaxis recién pegada. Si, por ejemplo, el nombre del factor no anidado es *minoría* y el del anidado *catlab*, la última línea de la sintaxis pegada quedará de esta manera: «Design *minoría catlab*».
- Modificar la última línea de la sintaxis añadiendo, a continuación del nombre del factor anidado, *catlab*, el nombre del factor no anidado, entre paréntesis. La última línea de la sintaxis debe quedar, por tanto, de la siguiente manera: «Design *minoría catlab(minoría)*».

Lo característico de la tabla resumen del ANOVA al ajustar un modelo jerárquico o anidado es, por un lado, que el factor anidado no interactúa con el no anidado y, por otro, que el efecto correspondiente al factor anidado aparece con su nombre seguido del nombre del factor no anidado (entre paréntesis).

## Análisis de regresión lineal

El procedimiento **Univariante** puede utilizarse para ajustar un modelo de regresión lineal y para contrastar las hipótesis referidas a los efectos presentes en un modelo de regresión (aunque debe tenerse en cuenta que para esto es más cómodo utilizar el procedimiento **Regresión lineal**; ver Capítulo 18). Para ajustar un modelo de regresión:

- En el cuadro de diálogo principal (ver Figura 15.1), seleccionar la variable dependiente y trasladarla al cuadro **Dependiente**.
- Seleccionar las variables independientes y trasladarlas a la lista **Covariables**.
- Pulsar el botón **Modelo...** para acceder al subcuadro de diálogo *Univariante: Modelo* (ver Figura 15.11) y seleccionar la opción **Personalizado**.
- Seleccionar todas las variables de la lista **Factores y covariables**.
- Seleccionar la opción **Efectos principales** dentro del menú desplegable **Construir términos** y pulsar el botón flecha para trasladar a la lista **Modelo** las variables seleccionadas.

Con estas especificaciones mínimas se obtiene la tabla resumen del ANOVA con información sobre la significación estadística de cada variable independiente. Para obtener, además, los coeficientes de regresión:

- En el subcuadro de diálogo *Univariante: Opciones* (ver Figura 15.7), marcar la opción **Estimaciones de los parámetros**.

Utilizando la variable *salario* como variable dependiente y las variables *salini*, *educ*, *tiemp-emp* y *expprev* como covariables se obtienen, entre otros, los resultados que muestra la Tabla 15.17.

La tabla resumen del ANOVA (no se incluye aquí) ofrece la información necesaria para contrastar, en los términos ya conocidos, el efecto individual (sin interacciones) de las cuatro variables independientes incluidas en el modelo y el efecto del modelo globalmente considerado.

La tabla de estimaciones (ver Tabla 15.17) ofrece las estimaciones de los coeficientes de regresión, es decir, las estimaciones de los parámetros del modelo. Con estos coeficientes es posible obtener los valores que el modelo pronostica para cada combinación de variables independientes:

$$\text{salario}' = -16.149,67 + 1,77 \text{ salini} + 669,91 \text{ educ} + 161,49 \text{ tiempemp} - 17,30 \text{ expprev}$$

La tabla incluye, además, pruebas *t* e intervalos de confianza para contrastar la significación de cada coeficiente. La hipótesis nula que se pone a prueba es que el parámetro en cuestión vale cero. En el ejemplo, todos los niveles críticos (*Sig.*) son menores que 0,0005, por lo que puede afirmarse que todos los parámetros estimados son distintos de cero.

**Tabla 15.17.** Estimaciones de los parámetros (coeficientes de regresión)

Variable dependiente: Salario actual						
Parámetro	B	Error típ.	t	Sig.	Intervalo de confianza al 95%.	
					Límite inferior	Límite superior
Intersección	-16.149,67	3.255,47	-4,96	,000	-22.546,78	-9.752,56
salini	1,77	,06	30,11	,000	1,65	1,88
educ	669,91	165,60	4,05	,000	344,51	995,32
tiempemp	161,49	34,25	4,72	,000	94,19	228,78
expprev	-17,30	3,53	-4,90	,000	-24,24	-10,37

## Homogeneidad de las pendientes de regresión

En el análisis de covarianza (ANCOVA) ya estudiado en este mismo capítulo, dentro de cada nivel de la variable independiente (o dentro de cada combinación entre los niveles de las variables independientes), existe una ecuación de regresión referida a la relación entre la variable dependiente y la covariable.

Uno de los supuestos del análisis es que las pendientes (los coeficientes de regresión) de esas ecuaciones de regresión son homogéneas. Es decir, en el análisis de covarianza se asume que la relación entre la variable dependiente y la covariable es constante a lo largo de todos los grupos definidos por la variable independiente (o todos los subgrupos definidos por la combinación de niveles de las variables independientes). Para contrastar ese supuesto:

- En el cuadro de diálogo principal (ver Figura 15.1), seleccionar la variable dependiente y trasladarla al cuadro **Dependiente**.
- Seleccionar la variable independiente y trasladarla a la lista **Factores fijos**.
- Seleccionar la covariable y trasladarla a la lista **Covariables**.
- Pulsar el botón **Modelo...** para acceder al subcuadro de diálogo *Univariante: Modelo* (ver Figura 15.11) y marcar la opción **Personalizado**.
- Seleccionar, dentro de la lista **Factores y covariables**, tanto la variable independiente como la covariable.
- Seleccionar **Efectos principales** dentro del menú desplegable **Construir términos** y pulsar el botón flecha para trasladar a la lista **Modelo** las variables seleccionadas.

- Seleccionar de nuevo la variable independiente y la covariable en la lista **Factores y covariables**.
- Seleccionar **Interacción** dentro del menú desplegable **Construir términos** y pulsar el botón flecha para trasladar a la lista **Modelo** el efecto referido a la interacción entre la variable independiente y la covariable.

Procediendo de esta manera es posible contrastar la hipótesis de homogeneidad de las pendientes mediante el estadístico  $F$  referido a la interacción (que se interpreta en los términos ya conocidos): si el efecto de la interacción es significativo ( $Sig. < 0,05$ ), las pendientes no son homogéneas; en caso contrario, puede asumirse que las pendientes son homogéneas.

Si puede asumirse homogeneidad de las pendientes, puede llevarse a cabo un ANCOVA tal como está descrito en el apartado *Análisis de covarianza* de este mismo capítulo, es decir, utilizando un modelo basado en una única estimación de la pendiente de la recta de regresión (lo cual tiene sentido porque se asume que las pendientes son homogéneas).

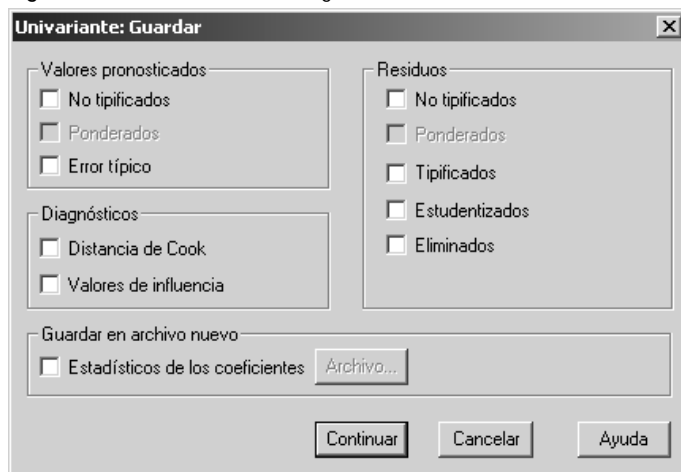
Pero si se rechaza la hipótesis de homogeneidad de las pendientes, el análisis de covarianza debe llevarse a cabo construyendo un modelo personalizado que incorpore estimaciones separadas para cada pendiente de regresión. Para ello:

- En el cuadro de diálogo principal (ver Figura 15.1), seleccionar la variable dependiente y trasladarla al cuadro **Dependiente**.
- Seleccionar la variable independiente y trasladarla a la lista **Factores fijos**.
- Seleccionar la covariable del diseño y trasladarla a la lista **Covariables**.
- Pulsar el botón **Modelo...** para acceder al subcuadro de diálogo *Univariante: Modelo* (ver Figura 15.11) y marcar la opción **Personalizado**.
- Dentro de la lista **Factores y covariables**, seleccionar la variable independiente.
- Seleccionar **Efectos principales** dentro del menú desplegable **Construir términos** y pulsar el botón flecha para trasladar a la lista **Modelo** la variable previamente seleccionada.
- En la lista **Factores y covariables**, seleccionar tanto la variable independiente como la covariable.
- Seleccionar **Interacción** dentro del menú desplegable **Construir términos** y pulsar el botón flecha para trasladar a la lista **Modelo** el efecto referido a la interacción entre la variable independiente y la covariable.
- Desmarcar la opción **Incluir la intersección en el modelo**.

## Guardar pronósticos y residuos

Entre las posibilidades del procedimiento **Univariante** está la de crear nuevas variables en el archivo de datos con los pronósticos y los residuos que se derivan del modelo ajustado. Para crear estas variables:

- Pulsar el botón **Guardar...** del cuadro de diálogo principal (ver Figura 15.1) para acceder al subcuadro de diálogo *Univariante: Guardar* que muestra la Figura 15.12.

Figura 15.12. Subcuadro de diálogo *Univariante: Guardar*

**Valores pronosticados.** Las opciones de este recuadro generan tres nuevas variables en el *Editor de datos*. Estas nuevas variables reciben automáticamente un nombre seguido de un número de serie: *NOMBRE\_#*. Por ejemplo, la primera vez que se solicitan durante una sesión los *pronósticos no tipificados*, la nueva variable con los pronósticos recibe el nombre *PRE\_1*; si se vuelven a solicitar los pronósticos no tipificados durante la misma sesión, la nueva variable recibe el nombre *PRE\_2*; etc.

- " **No tipificados:** valores pronosticados por el modelo para cada combinación entre los niveles de los factores. Estos valores pronosticados no son más que las medias marginales estimadas que se obtienen mediante el subcuadro de diálogo *Univariante: Opciones* (ver Figura 15.7). Nombre de la nueva variable: *PRE\_#*.
- " **Ponderados:** pronósticos ponderados. Son los pronósticos que se derivan del modelo cuando se está utilizando una variable de ponderación (ver Figura 15.1). Nombre de la nueva variable: *WPR\_#*.
- " **Error típico:** error típico (desviación típica) de los pronósticos. Nombre de la nueva variable: *SPE\_#*.

**Diagnósticos.** Este recuadro recoge dos medidas que expresan el grado en que cada caso se aleja de los demás. Estas medidas permiten identificar casos con un peso excesivo en el modelo que se está evaluando:

- " **Cook.** La distancia de Cook (1977, 1979) mide el cambio que se produce en las estimaciones de cada casilla al ir eliminando cada caso del modelo. Una distancia de Cook grande indica que ese caso tiene un peso considerable en el resultado de la estimación. En general, un caso con una distancia de Cook superior a 1 debe ser revisado: normalmente se trata de casos con demasiado peso en el modelo. Nombre de la nueva variable: *COO\_#*.
- " **Valores de influencia.** Representan una medida de la influencia potencial de cada caso. Referido a las variables independientes, un valor de influencia es una medida norma-

lizada del grado de distanciamiento de un caso respecto del centro de su distribución. Los casos muy alejados pueden influir de forma muy importante en las estimaciones del modelo, pero no necesariamente tienen por qué hacerlo. Los valores de influencia pueden interpretarse utilizando la siguiente regla general: los valores menores que 0,2 son poco problemáticos; los valores comprendidos entre 0,2 y 0,5 son arriesgados; y los valores mayores que 0,5 deberían evitarse. Nombre de la nueva variable: *LEV\_#*.

**Residuos.** Un residuo es la diferencia entre el valor pronosticado por el modelo de ANOVA (un pronóstico por casilla) y el valor observado en esa casilla. El procedimiento permite obtener distintos tipos de residuos:

- " **No tipificados.** Diferencia entre el valor pronosticado para cada casilla y la media observada en esa casilla. Nombre de la nueva variable: *RES\_#*.
- " **Ponderados.** Diferencia entre el valor pronosticado ponderado para cada casilla y la media observada en esa casilla. Sólo están disponibles si se ha seleccionado una variable de ponderación (ver Figura 15.1). Nombre de la nueva variable: *WRE\_#*.
- " **Tipificados.** Residuos divididos por su desviación típica. Son residuos transformados en puntuaciones *Z* (es decir, en una variable tipificada con media 0 y desviación típica 1). Nombre de la nueva variable: *ZRE\_#*.
- " **Estudentizados.** Residuos divididos por su desviación típica, basada ésta en la distancia existente entre cada caso y la media de su correspondiente casilla. Al igual que ocurre con los residuos tipificados, los estudentizados están escalados en unidades de desviación típica. Se distribuyen según el modelo de probabilidad *t* de *Student* con  $n-p-1$  grados de libertad ( $p$  se refiere al número de variables independientes). Con muestras grandes, aproximadamente el 95 % de estos residuos debería encontrarse en el rango  $(-2, +2)$ . Nombre de la nueva variable: *SRE\_#*.
- " **Eliminados.** Residuos obtenidos al efectuar los pronósticos eliminando del modelo el caso sobre el que se efectúa el pronóstico. Son residuos muy útiles para detectar puntos de influencia (casos con gran peso en el modelo). Nombre de la nueva variable: *DRE\_#*.

**Guardar en un archivo nuevo.** El procedimiento permite crear un archivo de datos nuevo con información sobre la matriz de varianzas-covarianzas de las estimaciones de los parámetros:

- " **Estadísticos de los coeficientes.** Guarda en un archivo de datos la matriz de varianzas-covarianzas de los parámetros del modelo. También se incluyen, para cada variable dependiente, las estimaciones de los parámetros, los errores típicos de esas estimaciones, y los niveles críticos y grados de libertad asociados a cada estimación.





## Análisis de varianza (III)

### El procedimiento *Modelo lineal general: Medidas repetidas*

#### Medidas repetidas

Los modelos de análisis de varianza (ANOVA) con *medidas repetidas* (MR) sirven para estudiar el efecto de uno o más factores cuando al menos uno de ellos es un factor *intra-sujetos*. En los factores *inter-sujetos* o *completamente aleatorizados* (los estudiados en los Capítulos 14 y 15), a cada nivel del factor se le asigna o le corresponde un grupo diferente de sujetos. Por el contrario, un factor *intra-sujetos* o con *medidas repetidas* se caracteriza porque todos los niveles del factor se aplican a los mismos los sujetos.

El diseño más simple de medidas repetidas consiste en medir dos variables en una misma muestra de sujetos. Los datos de este diseño se analizan con la prueba *T* para *muestras relacionadas* ya estudiada en el Capítulo 13. Pero los diseños de medidas repetidas pueden tener más de dos medidas y más de un factor.

Consideremos una investigación diseñada para conocer la opinión de los consumidores sobre cinco productos alternativos o rivales. Puede optarse por seleccionar tantos grupos de sujetos como productos disponibles (cinco) y hacer que cada grupo opine sobre un solo producto. De esta manera, se tendrá un diseño con un factor (tipo de producto, con cinco niveles) y tantos grupos de sujetos como niveles tiene el factor (cinco). Para analizar los datos de este diseño se puede utilizar un *ANOVA de un factor completamente aleatorizado* (ver Capítulo 14).

En lugar de esto, puede seleccionarse un único grupo de sujetos y pedir a cada sujeto que exprese su preferencia por cada uno de los cinco productos rivales. En ese caso, se seguirá teniendo un diseño de un factor (el tipo de producto, con cinco niveles), pero un solo grupo de sujetos que pasa por las cinco condiciones definidas por los niveles del factor (todos los sujetos opinan sobre todos los productos). Para analizar los datos de este diseño se puede utilizar un *ANOVA de un factor con medidas repetidas*.

Las ventajas de los diseños de medidas repetidas son evidentes: requieren menos sujetos que un diseño completamente aleatorizado y permiten eliminar la variación residual debida a las diferencias entre los sujetos (pues se utilizan los mismos). Como contrapartida, es necesario vigilar algunos efectos atribuibles precisamente a la utilización de los mismos sujetos, tales como el efecto de *arrastre*, que ocurre cuando se administra una condición antes de que haya finalizado el efecto de otra administrada previamente; o el efecto del *aprendizaje por la práctica*, que ocurre cuando las respuestas de los sujetos pueden mejorar con la repetición

y, como consecuencia de ello, los tratamientos administrados en último lugar parecen más efectivos que los administrados en primer lugar, sin que haya diferencias reales entre ellos (en estos casos es importante controlar el orden de presentación de las condiciones). Obviamente, conviene conocer las ventajas e inconvenientes de estos diseños para decidir correctamente cuándo es apropiado utilizarlos.

La opción **Medidas repetidas** del procedimiento **Modelo lineal general** permite ajustar modelos de ANOVA unifactoriales y factoriales con medidas repetidas en todos los factores o sólo en algunos, es decir, modelos con todos los factores intra-sujetos y modelos con factores inter-sujetos e intra-sujetos combinados. También permite incluir covariables para ajustar modelos de análisis de covarianza.

## Modelo de un factor

Conviene comenzar el estudio del ANOVA de medidas repetidas con el caso más simple: el modelo de un factor. Este modelo permite analizar los datos procedentes de un diseño con un solo grupo de sujetos y un único factor cuyos niveles se aplican a todos los sujetos.

### Datos

Para ilustrar la aplicación de este modelo se va a utilizar un ejemplo tomado de Pardo y San Martín (1998, págs. 263-265). En un experimento diseñado para estudiar el efecto del *paso del tiempo* sobre la *calidad del recuerdo*, a un grupo de 9 sujetos se les hace memorizar una historia durante 20 minutos. Más tarde, al cabo de una *hora*, de un *día*, de una *semana* y de un *mes*, se les pide que intenten repetir la historia escribiendo todo lo que recuerden. Un grupo de expertos evalúa la calidad del recuerdo de cada sujeto hasta elaborar los datos que muestra la Tabla 16.1. Se trata de un diseño de un factor (al que puede llamarse *tiempo*) con cuatro niveles (los cuatro momentos en los que se registra el recuerdo: al cabo de una *hora*, de un *día*, de una *semana* y de un *mes*) y una variable dependiente (la *calidad del recuerdo*).

**Tabla 16.1.** Datos de un diseño de un factor (*tiempo*) con medidas repetidas

<i>Sujetos</i>	<i>Hora</i>	<i>Día</i>	<i>Semana</i>	<i>Mes</i>
1	16	8	8	12
2	12	9	9	10
3	12	10	10	8
4	15	13	7	11
5	18	12	12	12
6	13	13	8	10
7	18	16	10	13
8	15	9	6	6
9	16	9	11	8

Desde el punto de vista de la disposición de los datos, la diferencia más evidente entre un factor completamente aleatorizado (CA) y un factor con medidas repetidas (MR) se encuentra

en la correspondencia existente entre el factor y el número de variables del archivo de datos. Mientras que un factor CA se corresponde con una única variable del archivo (una variable que toma distintos valores, cada uno de los cuales define un nivel del factor), un factor MR se corresponde con tantas variables del archivo de datos como niveles tiene el factor (cada una de esas variables define un nivel del factor MR).

## Análisis básico

Para llevar a cabo un ANOVA de un factor con medidas repetidas:

- Seleccionar la opción **Modelo lineal general > Medidas repetidas...** del menú **Analizar** para acceder al cuadro de diálogo *Definición de factor(es) de medidas repetidas* que muestra la Figura 16.1.

Figura 16.1. Cuadro de diálogo *Definición de factor(es) de medidas repetidas*

Este primer cuadro de diálogo, previo al principal, permite empezar a definir el factor (o factores) MR asignándole un nombre y especificando el número de niveles de que consta.

**Nombre del factor intra-sujetos** El primer paso para definir un factor MR o intra-sujetos consiste en asignarle un nombre. Puesto que un factor MR se corresponde con más de una variable del archivo de datos, es un factor que todavía no existe en ninguna parte. Debe crearse asignándole un nombre en este cuadro de texto. El nombre debe ajustarse a los nombres de variable de los archivos de datos SPSS, con la excepción de que no puede exceder de 8 caracteres. Por supuesto, puede elegirse cualquier nombre, pero conviene utilizar uno que tenga sentido para el usuario. En el ejemplo del cuadro de diálogo de la Figura 16.1 se ha elegido el nombre *tiempo* para identificar al factor definido por las variables *hora*, *día*, *semana* y *mes*.

**Número de niveles.** Este cuadro de texto permite introducir el número de niveles (*niveles = variables*) de que consta el factor recién nombrado (4 en el ejemplo de la Figura 16.1).

Tras asignar nombre y número de niveles al factor MR:

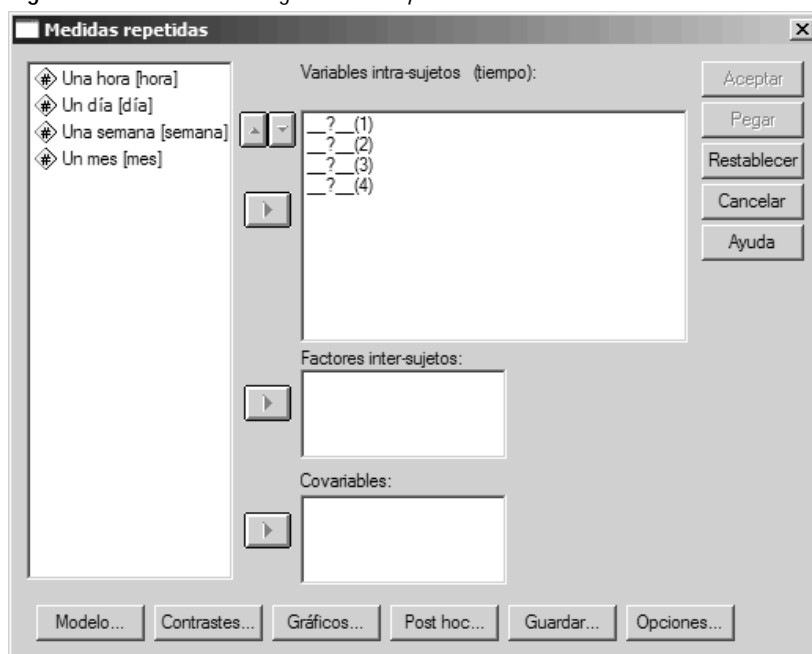
- Pulsar el botón **Añadir** para trasladar a la lista central y hacer efectivos tanto el nombre del factor como el número de niveles asignado. La lista mostrará entonces el nombre elegido y, entre paréntesis, el número de niveles.
- Utilizar los botones **Cambiar** y **Borrar** para modificar o eliminar, respectivamente, factores previamente añadidos.

**Nombre para la medida.** La mitad inferior del cuadro de diálogo permite definir más de una variable dependiente. El significado y la utilidad de incluir más de una variable dependiente se tratan más adelante, en este mismo capítulo, en el apartado *Modelo de un factor: Más de una variable dependiente*.

Una vez *añadidos* el nombre y el número de niveles del factor MR:

- Pulsar el botón **Definir** para acceder al cuadro de diálogo *Medidas repetidas* que muestra la Figura 16.2.

Figura 16.2. Cuadro de diálogo *Medidas repetidas*

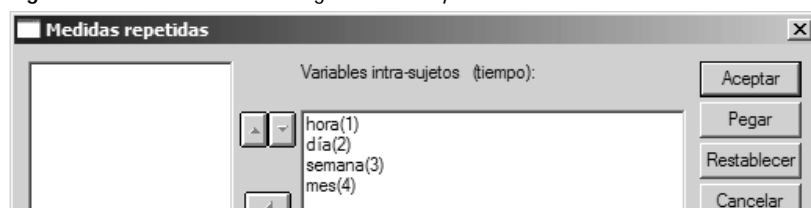


**Variables intra-sujetos.** Esta lista está preparada para recibir los nombres de las variables del archivo de datos que definen el factor intra-sujetos. Puesto que en el cuadro de diálogo previo (ver Figura 16.1) se ha indicado que el factor MR (al que se ha llamado *tiempo*) tiene 4 niveles, el SPSS está esperando que se le indique cuáles son las 4 variables que definen esos 4 niveles. Para ello:

- Seleccionar las variables *hora*, *día*, *semana* y *mes* en la lista del archivo de datos y trasladarlas a la lista **Variables intra-sujetos** utilizando el correspondiente botón flecha (utilizar los botones de desplazamiento vertical ▲ ▼ para modificar, si fuera necesario, el orden de las variables seleccionadas).

La Figura 16.2.bis muestra el cuadro de diálogo *Medidas repetidas* con las variables *hora*, *día*, *semana* y *mes* ya trasladadas a la lista **Variables intra-sujetos** (el cuadro se ha truncado para mostrar únicamente la parte relevante).

Figura 16.2.bis. Cuadro de diálogo *Medidas repetidas*



**Factores inter-sujetos.** En el caso de que el diseño incluya uno o más factores inter-sujetos, las variables que los definen deben trasladarse a esta lista (ver más adelante, en este mismo capítulo, el apartado *Modelo de dos factores: Medidas repetidas en un solo factor*).

**Covariables.** Si el diseño incluye una o más covariables, deben trasladarse a esta lista (ver, en el capítulo anterior sobre *ANOVA factorial*, el apartado *Análisis de covarianza*).

### **Ejemplo: MLG > ANOVA de un factor con medidas repetidas**

Este ejemplo muestra cómo aplicar un ANOVA de un factor con medidas repetidas a los datos de la Tabla 16.1. El diseño consta de un factor MR o intra-sujetos (al que se ha decidido llamar *tiempo*; con 4 niveles: *hora*, *día*, *semana* y *mes*), y una variable dependiente (la *calidad del recuerdo* medida por un grupo de expertos). Se trata de evaluar el posible efecto del paso del tiempo sobre la calidad del recuerdo.

- Reproducir en el *Editor de datos* los datos de la Tabla 16.1 (o abrir el archivo *ANOVA 1 repetidas* que se encuentra en la página web del manual).
- Seleccionar la opción **Modelo lineal general > Medidas repetidas...** del menú **Analizar** para acceder al cuadro de diálogo *Definición de factor(es) de medidas repetidas* (ver Figura 16.1).
- Introducir el nombre del factor MR (*tiempo*) en el cuadro de texto **Nombre del factor intra-sujetos** y el número de niveles de que consta el factor (4) en el cuadro de texto **Número de niveles**. Pulsar el botón **Añadir** para validar y el botón **Definir** para acceder al cuadro de diálogo *Medidas repetidas* (ver Figura 16.2).
- Seleccionar las variables *hora*, *día*, *semana* y *mes* y trasladarlas a la lista **Variables intra-sujetos**.

Aceptando estas elecciones, el *Visor* ofrece varias tablas de resultados basadas en las especificaciones que el programa tiene establecidas por defecto.

Las Tablas 16.2 a la 16.4 ofrecen varios estadísticos para poner a prueba la hipótesis nula referida al efecto del factor *tiempo*. La Tabla 16.2 contiene cuatro estadísticos multivariados: la *traza de Pillai*, la *lambda de Wilks*, la *traza de Hotelling* y la *raíz mayor de Roy*. Para una descripción de estos estadísticos puede consultarse Bock (1975) o Tabachnik y Fidel (2001). Se interpretan de la misma manera que el resto de estadísticos ya estudiados: puesto que el nivel crítico (*Sig.*) asociado a cada uno de ellos (en el ejemplo es el mismo para todos: 0,003) es menor que 0,05, se puede rechazar la hipótesis nula de igualdad de medias y concluir que la calidad del recuerdo no es la misma en los cuatro momentos temporales definidos por el factor *tiempo*.

**Tabla 16.2.** Contrastes multivariados

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Sig.
tiempo	Traza de Pillai	,894	16,844	3,000	6,000	,003
	Lambda de Wilks	,106	16,844	3,000	6,000	,003
	Traza de Hotelling	8,422	16,844	3,000	6,000	,003
	Raíz mayor de Roy	8,422	16,844	3,000	6,000	,003

Los modelos de medidas repetidas asumen que las varianzas de las diferencias entre cada dos niveles del factor MR son iguales. Por ejemplo, con 4 niveles, pueden hacerse 6 pares de combinaciones dos a dos entre niveles: 1-2, 1-3, 1-4, 2-3, 2-4 y 3-4. Calculando las diferencias entre las puntuaciones de esos 6 pares se obtienen 6 nuevas variables. En el modelo de un factor MR se asume que las varianzas de esas 6 variables son iguales. Este supuesto equivale a afirmar que la matriz de varianzas-covarianzas es *circular* (Huynd y Mandeville, 1979; para una aclaración de este supuesto, ver Kirk, 1982, págs. 256-261; o Winer, Brown y Michels, 1991, págs. 239-273). Y el procedimiento *Medidas repetidas* ofrece (ver Tabla 16.3), para evaluarlo, el *contraste de esfericidad de Mauchly* (1940). Puesto que el nivel crítico asociado al estadístico *W* (*Sig.* = 0,96) es mayor que 0,05, no puede rechazarse la hipótesis de esfericidad (puede, por tanto, asumirse que la matriz de varianzas-covarianzas es esférica).

**Tabla 16.3.** Contraste de esfericidad de *Mauchly*

Medida: MEASURE_1							
Efecto intra-sujetos	W de Mauchly	Chi-cuadrado aprox.	gl	Sig.	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	Limite-inferior
tiempo	,857	1,040	5	,960	,902	1,000	,333

Contrasta la hipótesis nula de que la matriz de covarianza error de las variables dependientes transformadas es proporcional a una matriz identidad.

En el caso de que el estadístico *W* lleve al rechazo de la hipótesis de esfericidad es posible optar por dos soluciones alternativas: (1) basar la decisión sobre la hipótesis de igualdad de medias en los estadísticos multivariados de la Tabla 16.2 (pues no les afecta el incumplimiento del supuesto de esfericidad), o (2) utilizar el estadístico *F* univariado que ofrece la Tabla 16.4 aplicando un índice corrector llamado *epsilon* (Box, 1954a, 1954b). Este índice corrector (ver Tabla 16.3, mitad derecha) expresa el grado en que la matriz de varianzas-covarianzas se aleja

de la esfericidad: en condiciones de esfericidad perfecta *epsilon* vale 1. La tabla ofrece dos estimaciones de *epsilon*: *Greenhouse-Geisser* (1959; Geisser y Greenhouse, 1958) y *Huynh-Feldt* (1976); la primera de ellas es algo más conservadora. Un tercer valor, *Límite inferior*, expresa el valor que adoptaría *epsilon* en el caso de incumplimiento extremo del supuesto de esfericidad. Para poder utilizar el estadístico *F* univariado en condiciones de no-esfericidad es necesario corregir los grados de libertad de *F* (tanto los del numerador como los del denominador) multiplicándolos por el valor estimado de *epsilon*. La Tabla 16.4 ofrece estos valores corregidos y sus correspondientes niveles críticos.

Si no se incumple el supuesto de esfericidad es preferible utilizar la aproximación univariada (en su versión *Esfericidad asumida*; ver Tabla 16.4), pues, en condiciones de esfericidad, el estadístico univariado *F* es más potente que los estadísticos multivariados de la Tabla 16.2, sobre todo con muestras pequeñas (aunque, por supuesto, si ambas estrategias conducen a la misma decisión, es irrelevante utilizar una u otra).

Los resultados de la Tabla 16.4 indican que las cuatro versiones del estadístico *F* (la no corregida: *Esfericidad asumida*; y las tres corregidas: *Greenhouse-Geisser*, *Huynh-Feldt* y *Límite inferior*) conducen a la misma conclusión, que a su vez coincide con la ya alcanzada con la aproximación multivariada: puesto que el nivel crítico (*Sig.*) es, en todos los casos, menor que 0,05, se puede rechazar la hipótesis de igualdad de medias y concluir que la calidad del recuerdo no es la misma en las cuatro medidas obtenidas.

**Tabla 16.4.** Efectos intra-sujetos

Medida: MEASURE\_1

Fuente		Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
tiempo	Esfericidad asumida	186,750	3	62,250	18,675	,000
	Greenhouse-Geisser	186,750	2,707	68,981	18,675	,000
	Huynh-Feldt	186,750	3,000	62,250	18,675	,000
	Límite-inferior	186,750	1,000	186,750	18,675	,003
Error(tiempo)	Esfericidad asumida	80,000	24	3,333		
	Greenhouse-Geisser	80,000	21,658	3,694		
	Huynh-Feldt	80,000	24,000	3,333		
	Límite-inferior	80,000	8,000	10,000		

## Aspectos complementarios del análisis

El cuadro de diálogo *Medidas repetidas* (ver Figura 16.2) contiene una serie de botones específicos que permiten personalizar los resultados del análisis. Estos botones ya se han descrito en el capítulo anterior sobre *ANOVA factorial*, pero conviene señalar algunos detalles.

### Modelo...

El único modelo con sentido en un diseño con un solo factor es justamente el que incluye ese factor. Por tanto, con el diseño de un factor carece de sentido utilizar este botón (excepto para cambiar el tipo de suma de cuadrados que el SPSS utiliza por defecto; cosa, por otro lado, completamente innecesaria y poco recomendable con el modelo de un factor).



**Contrastes...**

El procedimiento **Medidas repetidas** asigna, por defecto, contrastes de tipo **Polinómico** a los factores MR (ver, en el capítulo anterior sobre *ANOVA factorial*, el apartado *Contrastes personalizados*). Estos contrastes polinómicos permiten estudiar el tipo de relación existente entre el factor y la variable dependiente: lineal, cuadrática, cúbica, etc. Pero podrían no tener sentido dependiendo del factor MR que se esté utilizando. Si fuera ese el caso, puede optarse por asignar como contraste para el factor MR la opción **Ninguno** o cualquier otra de las disponibles (si tuviera sentido), o puede, simplemente, ignorarse la información de la tabla de resultados correspondiente a los contrastes polinómicos.

Si no se modifica la opción por defecto del botón **Contrastes...**, el *Visor* ofrece los contrastes polinómicos que muestra la Tabla 16.5. Puesto que se trata de contrastes ortogonales, la tabla muestra tantos contrastes como niveles tiene el factor, menos uno; dado que el factor *tiempo* del ejemplo tiene cuatro niveles, aparecen tres contrastes: lineal, cuadrático y cúbico.

La tabla recoge la información necesaria para contrastar la hipótesis nula de que el polinomio o componente evaluado vale cero en la población; es decir, la hipótesis nula de que no existe relación lineal, cuadrática, etc.

En el ejemplo, los valores de los niveles críticos (*Sig.*) asociados a cada estadístico *F* permiten rechazar las hipótesis referidas a los componentes lineal y cuadrático, pero no la referida al componente cúbico. Esto significa que las medias de la calidad del recuerdo en cada momento temporal se ajustan tanto a una línea recta (componente lineal) como a una curva (componente cuadrático).

Cuando existe más de un componente significativo, es probable que el de mayor orden se ajuste mejor, pero las funciones más parsimoniosas (más simples) son más fáciles de interpretar y, generalmente, más útiles. No obstante, decidir qué polinomio de los significativos se interpreta depende, fundamentalmente, de las hipótesis previas que haya establecido el investigador. Por otra parte, un gráfico de perfil (ver siguiente apartado) puede ayudar a comprender lo que está ocurriendo.

**Tabla 16.5.** Contrastes intra-sujetos

Medida: MEASURE_1						
Fuente		Suma de	gl	Media	F	Sig.
tiempo		cuadrados tipo III		cuadrática		
tiempo	Lineal	130,050	1	130,050	53,082	,000
	Cuadrático	56,250	1	56,250	20,455	,002
	Cúbico	,450	1	,450	,094	,767
Error(tiempo)	Lineal	19,600	8	2,450		
	Cuadrático	22,000	8	2,750		
	Cúbico	38,400	8	4,800		

La última tabla de resultados (Tabla 16.6.) ofrece el contraste de los efectos inter-sujetos. En un diseño de un factor intra-sujetos el único efecto inter-sujetos es el que se refiere a la media total (la constante del modelo). El estadístico *F* de la Tabla 16.6 permite contrastar la hipótesis de que la media poblacional total vale cero. Puesto que el nivel crítico (*Sig.* < 0,0005) es menor que 0,05, se puede rechazar esa hipótesis y concluir que la media total es distinta de cero. Generalmente, este contraste carece de utilidad, excepto si se desea efectuar pronósticos, en cuyo caso conviene valorar la conveniencia de incluir la constante en el modelo.

Tabla 16.6. Efectos inter-sujetos

Medida: MEASURE\_1

Variable transformada: Promedio

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Intersección	4556,250	1	4556,250	405,000	,000
Error	90,000	8	11,250		

## Gráficos...

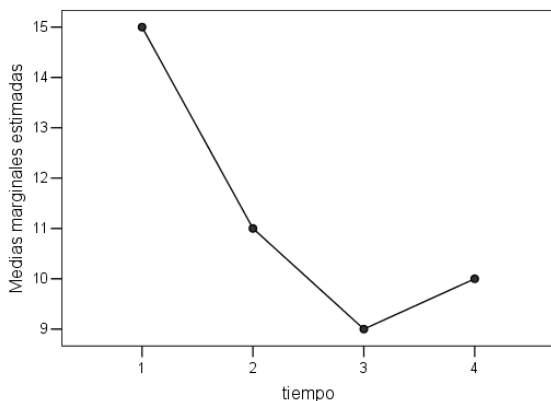
Esta opción permite obtener un gráfico de líneas o de perfil representando el efecto (evolución, tendencia) de los niveles del factor MR. Para obtener el gráfico de perfil:

- Pulsar el botón Gráficos... del cuadro de diálogo *Medidas repetidas* (ver Figura 16.2) para acceder al subcuadro de diálogo *Medidas repetidas: Gráficos de perfil* que muestra la Figura 16.3.

Figura 16.3. Subcuadro de diálogo *Medidas repetidas: Gráficos de perfil*

- Seleccionar el factor MR (*tiempo* en el ejemplo) en la lista Factores y trasladarlo al cuadro Eje horizontal con el correspondiente botón flecha.
- Pulsar el botón **Añadir** para trasladar la variable seleccionada a la lista inferior y hacer efectiva la selección.

Aceptando estas elecciones se obtiene el gráfico de perfil que muestra la Figura 16.4. Puede observarse en él que la calidad del recuerdo (variable dependiente) va disminuyendo con el paso del tiempo, pero sólo hasta el momento 3 (una *semana*), a partir del cual se aprecia una ligera recuperación. Parece, por tanto, que, tal como se acaba de constatar con los contrastes polinómicos en el apartado anterior, el comportamiento de las medias se ajusta a una función cuadrática.

Figura 16.4. Gráfico de perfil representando el efecto del factor *tiempo*

### Post hoc...

Las comparaciones *post hoc* no están disponibles para los factores MR. Sólo pueden utilizarse para comparar los distintos niveles de un factor inter-sujetos (cuando existe) en cada uno de los niveles del factor MR (ver más adelante, en este mismo capítulo, el apartado *Modelo de dos factores con medidas repetidas en un solo factor*). Para comparar dos a dos los niveles de un factor MR puede utilizarse la opción **Comparar los efectos principales** que se describe más adelante, en este mismo capítulo, en el apartado *Opciones...*

### Guardar...

Todas las opciones de este subcuadro de diálogo se explican en el Capítulo 18 sobre *Análisis de regresión lineal*. Es en los modelos de regresión donde quizá tiene más utilidad detenerse a inspeccionar los residuos y los pronósticos.

### Opciones...

Este cuadro de diálogo también se ha explicado en el capítulo anterior sobre *ANOVA factorial*, pero ahora contiene algunas variantes que conviene explicar. En primer lugar, las opciones **Pruebas de homogeneidad** y **Diagramas de dispersión por nivel** no están disponibles porque, dado que en un diseño con un factor MR no existen grupos, no tiene sentido establecer supuestos sobre las varianzas poblacionales de los grupos.

Por otro lado, aunque, según se ha señalado ya, el procedimiento **Medidas repetidas** no permite efectuar comparaciones *post hoc* entre los niveles de un factor MR, la opción **Comparar los efectos principales** sirve para comparar dos a dos los distintos niveles del factor. Para ello obtener estas comparaciones:

- Pulsar el botón **Opciones...** del cuadro de diálogo principal (ver Figura 16.2) para acceder al subcuadro de diálogo *Medidas repetidas: Opciones* que muestra la Figura 16.5.

Figura 16.5. Subcuadro de diálogo *Medidas repetidas: Opciones*

- Seleccionar la variable *tiempo* en la lista **Factores e interacciones de los factores** y trasladarla, con el botón flecha, a la lista **Mostrar las medias para**.
- Marcar la opción **Comparar los efectos principales**.
- Seleccionar la opción **Bonferroni** dentro del menú desplegable **Ajuste del intervalo de confianza**.

Estas elecciones permiten obtener dos tablas de resultados. La primera de ellas (Tabla 16.7) es la tabla de *Medias estimadas*: ofrece, para cada nivel del factor *tiempo*, la media estimada, el error típico de la media y el intervalo de confianza para la media (calculado al 95 %).

Tabla 16.7. Medias estimadas

Medida: MEASURE\_1

tiempo	Media	Error típ.	Intervalo de confianza al 95%.	
			Límite inferior	Límite superior
1	15,000	,764	13,239	16,761
2	11,000	,882	8,966	13,034
3	9,000	,645	7,511	10,489
4	10,000	,764	8,239	11,761

La Tabla 16.8 ofrece las comparaciones dos a dos entre los niveles del factor. Los niveles críticos (*Significación*) de esta tabla están ajustados mediante la corrección de Bonferroni (para controlar la tasa de error). Puede comprobarse en la tabla que únicamente existen diferencias significativas entre el momento o nivel 1 (*hora*) y el resto de momentos o niveles.

**Tabla 16.8.** Comparaciones por pares

Medida: MEASURE\_1

(I) tiempo	(J) tiempo	Diferencia entre medias (I-J)	Error típ.	Sig. <sup>a</sup>	Intervalo de confianza al 95 % para diferencia <sup>a</sup>	
					Límite inferior	Límite superior
1	2	4,000	,928	,015	,772	7,228
	3	6,000	,816	,000	3,160	8,840
	4	5,000	,764	,001	2,343	7,657
2	3	2,000	1,014	,504	-1,527	5,527
	4	1,000	,782	1,000	-1,720	3,720
3	4	-1,000	,833	1,000	-3,899	1,899

Basadas en las medias marginales estimadas.

a. Ajuste para comparaciones múltiples: Bonferroni.

El cuadro de diálogo *Opciones* del procedimiento **Medidas repetidas** contiene tres opciones nuevas que no están incluidas en el cuadro de diálogo *Opciones* del procedimiento **Univariante** estudiado en el capítulo anterior: *matrices SCPC*, *matriz SCPC residual* y *matriz de transformación*.

**Matriz de transformación.** Ofrece los coeficientes normalizados que el SPSS asigna a cada nivel del factor MR en cada uno de los contrastes definidos en el cuadro de diálogo *Contrastes*. Así, por ejemplo, la Tabla 16.9 muestra los coeficientes asignados a los niveles del factor *tiempo* en cada uno de los posibles contrastes *polinómicos* (que son los que el programa aplica por defecto a los factores MR). Para comprender mejor el significado de estos coeficientes, puede consultarse el apartado *Comparaciones planeadas o a priori*, en el Capítulo 14 sobre *ANOVA de un factor*.

**Tabla 16.9.** Matriz de transformación para el factor *tiempo*

Medida: MEASURE\_1

Variable dependiente	tiempo		
	Lineal	Cuadrático	Cúbico
Una hora	-,671	,500	-,224
Un día	-,224	-,500	,671
Una semana	,224	-,500	-,671
Un mes	,671	,500	,224

**Matrices SCPC** (matrices de *sumas de cuadrados y de productos cruzados*). Genera una matriz diferente para cada efecto inter-sujetos, para cada efecto intra-sujetos y para cada término error. Para un efecto dado, la matriz SCPC muestra, en la diagonal, la suma de cuadrados correspondiente a ese efecto descompuesta en tantas partes como grados de libertad tiene ese efecto (ver Tabla 16.10). Fuera de la diagonal de la matriz se ofrecen las covarianzas entre cada contraste.

La descomposición de la suma de cuadrados se efectúa a partir de los contrastes definidos en el cuadro de diálogo *Contrastes*. Si para un efecto concreto se han definido, por ejemplo, contrastes *polinómicos*, las sumas de cuadrados de la diagonal de la matriz SCPC correspondiente a ese efecto mostrará la suma de cuadrados asociada a cada polinomio o tendencia. En la Tabla 16.10 puede comprobarse que, al sumar las sumas de cuadrados de cada contraste

(polinomio en este caso), se obtiene la suma de cuadrados del factor *tiempo*:  $130,05 + 56,25 + 0,45 = 186,75$  (ver Tabla 16.4).

Tabla 16.10. Matriz SCPC correspondiente al efecto del factor *tiempo*

			tiempo : columna		
			Lineal	Cuadrático	Cúbico
Hipótesis	Intersección	Lineal	130,050	-85,530	-7,650
		Cuadrático	-85,530	56,250	5,031
		Cúbico	-7,650	5,031	,450
Error		Lineal	19,600	-2,236	-,800
		Cuadrático	-2,236	22,000	-2,236
		Cúbico	-,800	-2,236	38,400

Basada en la suma de cuadrados tipo III

**Matriz SCPC residual** (matriz de *sumas de cuadrados y productos cruzados residual*). Esta matriz contiene tres subtablas (ver Tabla 16.11), todas ellas con información sobre los residuos (los residuos de un modelo son las diferencias entre los valores observados y los valores pronosticados por el modelo).

La primera subtabla ofrece las sumas de cuadrados de cada nivel del factor (en la diagonal principal) y las sumas de productos cruzados (fuera de la diagonal). La segunda contiene las varianzas (en la diagonal principal) y las covarianzas (fuera de la diagonal). La tercera incluye la misma información que la segunda, pero tipificada; es decir, ofrece las correlaciones entre los residuos de cada dos niveles del factor.

Tabla 16.11. Matriz SCPC residual

		Una hora	Un día	Una semana	Un mes
Suma de cuadrados y productos cruzados	Una hora	42,000	18,000	12,000	21,000
	Un día	18,000	56,000	6,000	27,000
	Una semana	12,000	6,000	30,000	11,000
	Un mes	21,000	27,000	11,000	42,000
Covarianza	Una hora	5,250	2,250	1,500	2,625
	Un día	2,250	7,000	,750	3,375
	Una semana	1,500	,750	3,750	1,375
	Un mes	2,625	3,375	1,375	5,250
Correlación	Una hora	1,000	,371	,338	,500
	Un día	,371	1,000	,146	,557
	Una semana	,338	,146	1,000	,310
	Un mes	,500	,557	,310	1,000

Basada en la suma de cuadrados tipo III

Al solicitar la matriz SCPC residual, el SPSS ofrece también el contraste de esfericidad de Bartlett (ver Tabla 16.12), similar al contraste de Mauchly ya estudiado. Si no se incumple el supuesto de normalidad, el contraste de Bartlett permite evaluar la hipótesis de que la matriz de varianzas-covarianzas residual es proporcional a una matriz identidad. Para ello, ofrece dos estadísticos asintóticamente equivalentes: la *razón de verosimilitudes* y el estadístico *chi-cuadrado*, los cuales se interpretan de igual forma que cualquier otro estadístico de contraste: si el nivel crítico (*Significación*) es menor que 0,05, se rechaza la hipótesis de esfericidad.

Tabla 16.12. Contraste de esfericidad de *Bartlett*

Razón de verosimilitud	,015
Chi-cuadrado aprox.	5,951
gl	9
Significación	,754

## Más de una variable dependiente

Siguiendo con el ejemplo sobre memoria, consideremos una situación en la que para medir la calidad del recuerdo se ha utilizado, en cada momento temporal, una medida de *reconocimiento* y otra de *recuerdo libre*. En la tabla de datos (ver Tabla 16.1) habrá dos variables por cada momento temporal: una variable con las puntuaciones de la medida *reconocimiento* y otra variable con las puntuaciones de la medida *recuerdo libre* (es decir, habrá 8 variables en lugar de 4). Por supuesto, las medidas de *reconocimiento* y de *recuerdo libre* podrían tratarse como niveles de un segundo factor MR. Pero si no interesa evaluar el efecto de ese factor ni el efecto de la interacción entre ese factor y el factor *tiempo*, lo apropiado es utilizar ambas medidas como variables dependientes.

Para ello puede utilizarse la mitad inferior del cuadro de diálogo *Definición de factor(es) de medidas repetidas* (ver Figura 16.1). Cuando se utilizan, por ejemplo, dos medidas tales como *reconocimiento* y *recuerdo*, el SPSS ofrece varios estadísticos multivariados para contrastar el efecto del factor *tiempo* teniendo en cuenta ambas medidas simultáneamente, y varios estadísticos univariados para contrastar el efecto del factor *tiempo* en cada medida por separado. Para definir más de una medida por cada nivel del factor MR:

- Asignar nombre a la primera medida en el cuadro de texto **Nombre para la medida** y pulsar el botón **Añadir** (el nombre no puede tener más de 8 caracteres, debe ajustarse a las reglas de los nombres de variable del SPSS y no puede duplicar el nombre de una variable ya existente en el archivo de datos).
- Repetir la operación para cada una de las medidas restantes.
- Pulsar el botón **Definir** para continuar con el análisis.

Si se repitiera ahora el ejemplo sobre memoria incluyendo dos medidas (por ejemplo, *reconocimiento* y *recuerdo*), se obtendrían los resultados ya estudiados en el ejemplo anterior, pero referidos a ambas variables.

## Modelo de dos factores, ambos con medidas repetidas

En un diseño de dos factores, ambos con medidas repetidas, los sujetos que participan en el experimento pasan por *todas* las condiciones experimentales, es decir, por todas las condiciones definidas por las posibles combinaciones entre los niveles de ambos factores.

Continuando con el ejemplo sobre memoria del apartado anterior, imaginemos que se ha repetido el experimento añadiendo un nuevo factor MR. Además del factor *tiempo* ya considerado (con cuatro niveles: *hora*, *día*, *semana*, *mes*), se ha incluido un nuevo factor al que se

le ha llamado *contenido* (con dos niveles: *números* y *letras*). Es decir, los sujetos han realizado dos tareas de memorización: una con números y otra con letras.

## Datos

A seis sujetos aleatoriamente seleccionados se les ha hecho memorizar durante 20 minutos dos listas distintas: una de *letras* y otra de *números*. Más tarde, al cabo de una *hora*, de un *día*, de una *semana* y de un *mes*, se les ha pedido que intenten repetir ambas listas. Un grupo de expertos ha evaluado la calidad del recuerdo de cada sujeto y les ha asignado las puntuaciones que muestra la Tabla 16.13.

**Tabla 16.13.** Datos de un diseño de dos factores (*tiempo* × *contenido*) con medidas repetidas en ambos

<i>Sujetos</i>	<i>Hora</i>		<i>Día</i>		<i>Semana</i>		<i>Mes</i>	
	<i>Números</i>	<i>Letras</i>	<i>Números</i>	<i>Letras</i>	<i>Números</i>	<i>Letras</i>	<i>Números</i>	<i>Letras</i>
1	6	8	6	6	3	4	2	3
2	7	10	5	8	5	5	5	2
3	4	7	2	7	1	2	3	2
4	7	11	5	9	3	3	4	6
5	6	10	4	6	4	4	5	3
6	5	9	2	4	1	3	1	5

El propósito de este experimento consiste averiguar si existen diferencias en la *calidad del recuerdo* en función de dos variables: el paso del *tiempo* (una *hora*, un *día*, una *semana*, un *mes*) y el *contenido* del material memorizado (*números* o *letras*). Se trata, por tanto, de un diseño con dos factores MR (*tiempo*, con cuatro niveles, y *contenido*, con dos niveles) y una variable dependiente cuantitativa (la *calidad del recuerdo* evaluada y cuantificada por un grupo de expertos). Conviene recordar en este momento que, el hecho de que se encuentren diferencias en una variable en función de los niveles de otra no significa que la relación detectada sea de tipo causal. Un contraste estadístico nunca es garantía, por sí sólo, de relación causal. Para poder concluir causalmente es necesario atender a aspectos de diseño y/o a aspectos teóricos.

Para reproducir los datos de la Tabla 16.13 en el *Editor de datos* del SPSS es necesario crear tantas variables como el número de condiciones resultantes de combinar los niveles de ambos factores. Puesto que el experimento del ejemplo incluye un factor con 4 niveles y otro con 2, es necesario crear  $4 \times 2 = 8$  variables.

Para nombrar estas variables puede utilizarse cualquier nombre válido, pero, obviamente, conviene asignarles nombres que permitan identificarlas fácilmente. En el ejemplo que reproduce la Figura 16.7 se han asignado los siguientes nombres:

*hora\_n* = una hora, lista de números (nivel: 1, 1)

*hora\_l* = una hora, lista de letras (nivel: 1, 2)

*día\_n* = un día, lista de números (nivel: 2, 1)

*día\_l* = un día, lista de letras (nivel: 2, 2)



*semana\_n* = una semana, lista de números (nivel: 3, 1)

*semana\_l* = una semana, lista de letras (nivel: 3, 2)

*mes\_n* = un mes, lista de números (nivel: 4, 1)

*mes\_l* = un mes, lista de letras (nivel: 4,2)

La Figura 16.6 muestra el aspecto del *Editor de datos* después de reproducir en él las variables y los datos de la Tabla 16.13. Por supuesto, puede optarse por utilizar nombres más sencillos, como, por ejemplo,  $x_1, x_2, x_3, \dots, x_8$ ; y también puede optarse por asignar o no etiquetas de variable a esos nombres; pero estos detalles corresponde decidirlos al propio usuario. Aquí, por claridad, se ha optado por asignar los nombres ya mencionados (los cuales permiten identificar rápidamente cada combinación de niveles) y las etiquetas de variable que aparecen en el cuadro de diálogo principal que muestra la Figura 16.7.

**Figura 16.6.** Datos de la Tabla 16.13 reproducidos en el *Editor de datos*

	id	hora_n	hora_l	día_n	día_l	semana_n	semana_l	mes_n	mes_l
1	1	6	8	6	6	3	4	2	3
2	2	7	10	5	8	5	5	5	2
3	3	4	7	2	7	1	2	3	2
4	4	7	11	5	9	3	3	4	6
5	5	6	10	4	6	4	4	5	3
6	6	5	9	2	4	1	3	1	5

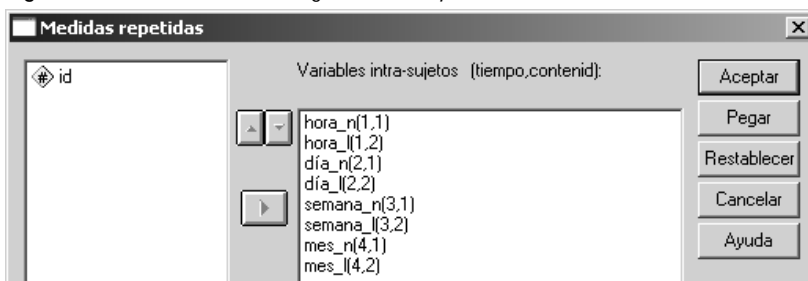
## Análisis básico

Para entender fácilmente los pasos que se siguen en este apartado, es recomendable haber asimilado bien los pasos seguidos en el apartado *Modelo de un factor: Análisis básico* de este mismo capítulo. La mayor parte de las acciones que es necesario llevar a cabo para ajustar un modelo factorial son prácticamente idénticas a las explicadas a propósito del modelo de un factor. Para llevar a cabo un ANOVA de dos factores, ambos con medidas repetidas:

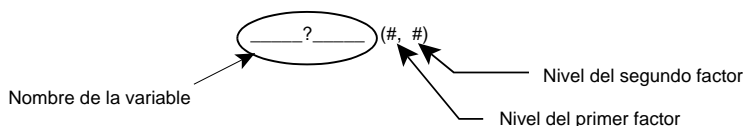
- Reproducir en el *Editor de datos* los datos de la Tabla 16.13 tal como muestra la Figura 16.6 (o abrir el archivo *ANOVA 2 repetidas en ambos* que se encuentra en la página web del manual).
- Seleccionar la opción **Modelo lineal general > Medidas repetidas...** del menú **Analizar** para acceder al cuadro de diálogo *Definición de factor(es) de medidas repetidas* (ver Figura 16.1).
- Asignar nombre (*tiempo*) y número de niveles (4) al primer factor MR, y pulsar el botón **Añadir**. Asignar nombre (*contenido*) y número de niveles (2) al segundo factor MR, y pulsar el botón **Añadir** (el nombre asignado a un factor no puede exceder de 8 caracteres, de ahí el nombre asignado al segundo factor).
- Utilizar los botones **Cambiar** y **Borrar** para modificar o eliminar, respectivamente, factores previamente añadidos.
- Pulsar el botón **Definir** para acceder al cuadro de diálogo *Medidas repetidas* que muestra la Figura 16.7.

Figura 16.7. Cuadro de diálogo *Medidas repetidas*

La lista **Variables intra-sujetos** está preparada para recibir los nombres de las variables que corresponden a los niveles de los factores definidos. Trasladando las variables a esta lista, con el correspondiente botón flecha, se obtiene el resultado que muestra la Figura 16.7.bis (se ha truncado la figura para mostrar únicamente la parte relevante).

Figura 16.7.bis. Cuadro de diálogo *Medidas repetidas*

Es muy importante vigilar que cada variable se traslada al lugar correcto. Para ello, debe tenerse en cuenta que el orden en el que aparecen listadas las condiciones experimentales en el cuadro **Variables intra-sujetos** depende del orden en el que se han definido previamente los factores MR en el cuadro de diálogo *Definición de factor(es) de medidas repetidas* (ver Figura 16.1). El esquema de la Figura 16.8 puede ayudar a comprender cómo deben trasladarse las variables.

Figura 16.8. Correspondencia entre *variables* y *niveles* en los factores MR

### Ejemplo: *MLG > ANOVA de dos factores, ambos con medidas repetidas*

Este ejemplo muestra cómo aplicar un ANOVA de dos factores, con medidas repetidas en ambos, a los datos de la Tabla 16.13. El diseño incluye dos factores MR: *tiempo* (con 4 niveles: *hora*, *día*, *semana* y *mes*) y *contenido* (con 2 niveles: *números* y *letras*). Se sigue utilizando como variable dependiente una medida de la *calidad del recuerdo*. La Figura 16.6 muestra cómo reproducir los datos de la Tabla 16.13 en el *Editor de datos* del SPSS.

Ya se ha explicado (ver capítulo anterior sobre *ANOVA factorial*) que en un diseño de estas características (dos factores) existen tres efectos de interés: el efecto individual del primer factor, el efecto individual del segundo factor y el efecto conjunto de la interacción entre los dos factores. Para obtener un ANOVA de dos factores, con medidas repetidas en ambos:

- Seleccionar la opción **Modelo lineal general > Medidas repetidas...** del menú **Analizar** para acceder al cuadro de diálogo *Definición de factor(es) de medidas repetidas* (ver Figura 16.1).
- Introducir el nombre del primer factor MR (*tiempo*) en el cuadro de texto **Nombre del factor intra-sujetos** y el número de niveles de que consta ese factor (4) en el cuadro de texto **Número de niveles**. Pulsar el botón **Añadir** para hacer efectivos el nombre y el número de niveles del factor.
- Introducir el nombre del segundo factor MR (*contenid*—sólo se permiten 8 caracteres) en el cuadro de texto **Nombre del factor intra-sujetos** y el número de niveles de que consta ese factor (2) en el cuadro de texto **Número de niveles**. Pulsar el botón **Añadir** para hacer efectivos el nombre y el número de niveles del segundo factor.
- Pulsar el botón **Definir** para acceder al cuadro de diálogo *Medidas repetidas* (ver Figura 16.7).
- Seleccionar las 8 variables de la lista de variables y trasladarlas, en el orden correcto, a la lista **Variables intra-sujetos**.

Aceptando estas elecciones, el *Visor* ofrece varias tablas de resultados basadas en las especificaciones que el programa tiene establecidas por defecto.

La Tabla 16.14 ofrece cuatro estadísticos multivariados para poner a prueba cada una de las hipótesis nulas de interés en este diseño. Estos estadísticos multivariados se interpretan de la misma manera que el resto de estadísticos ya estudiados. En primer lugar, puesto que el nivel crítico (*Sig.* = 0,002) asociado al efecto del factor *tiempo* es menor que 0,05, se puede rechazar la hipótesis nula de igualdad de medias referida a ese factor y, por tanto, concluir que la calidad del recuerdo no es la misma en los cuatro momentos temporales considerados. En segundo lugar, puesto que el nivel crítico (*Sig.* = 0,006) asociado al efecto del factor *contenid* es menor que 0,05, se puede rechazar la hipótesis nula de igualdad de medias referida al factor *contenid* y, por tanto, concluir que la calidad del recuerdo no es la misma en las dos listas uti-

lizadas. Por último, puesto que el nivel crítico ( $Sig. = 0,083$ ) asociado al efecto de la interacción *tiempo-contenido* es mayor que 0,05, no se puede rechazar la hipótesis nula referida al efecto de la interacción (no se puede afirmar que exista efecto significativo de la interacción).

Tabla 16.14. Contrastes multivariados

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Sig.
tiempo	Traza de Pillai	,990	97,676 <sup>a</sup>	3,000	3,000	,002
	Lambda de Wilks	,010	97,676 <sup>a</sup>	3,000	3,000	,002
	Traza de Hotelling	97,676	97,676 <sup>a</sup>	3,000	3,000	,002
	Raíz mayor de Roy	97,676	97,676 <sup>a</sup>	3,000	3,000	,002
contenid	Traza de Pillai	,803	20,351 <sup>a</sup>	1,000	5,000	,006
	Lambda de Wilks	,197	20,351 <sup>a</sup>	1,000	5,000	,006
	Traza de Hotelling	4,070	20,351 <sup>a</sup>	1,000	5,000	,006
	Raíz mayor de Roy	4,070	20,351 <sup>a</sup>	1,000	5,000	,006
tiempo * contenid	Traza de Pillai	,863	6,277 <sup>a</sup>	3,000	3,000	,083
	Lambda de Wilks	,137	6,277 <sup>a</sup>	3,000	3,000	,083
	Traza de Hotelling	6,277	6,277 <sup>a</sup>	3,000	3,000	,083
	Raíz mayor de Roy	6,277	6,277 <sup>a</sup>	3,000	3,000	,083

a. Estadístico exacto

La Tabla 16.15 ofrece el estadístico *W* de Mauchly para contrastar la hipótesis de esfericidad (ver, en este mismo capítulo, el ejemplo del apartado *Modelo de un factor*). La tabla ofrece un estadístico para cada uno de los efectos presentes en el modelo. Puesto que el nivel crítico ( $Sig.$ ) asociado al estadístico *W* es mayor que 0,05 en los tres casos, puede asumirse que las tres matrices de varianzas-covarianzas son esféricas. La significación referida al factor *contenido* no aparece porque con dos niveles no tiene sentido hablar de esfericidad (con dos niveles sólo existe una covarianza que, obviamente, es igual a sí misma).

Tabla 16.15. Contraste de esfericidad de *Mauchly*

Medida: MEASURE\_1

Efecto intra-sujetos	W de Mauchly	Chi-cuadrado aprox.	gl	Sig.	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	Límite inferior
tiempo	,418	3,246	5	,672	,753	1,000	,333
contenid	1,000	,000	0	.	1,000	1,000	1,000
tiempo * contenid	,219	5,654	5	,356	,521	,715	,333

Contrasta la hipótesis nula de que la matriz de covarianza error de las variables dependientes transformadas es proporcional a una matriz identidad.

La Tabla 16.16 muestra los estadísticos *F* univariados asociados a cada efecto. Al igual que ocurría con la aproximación multivariada (ver Tabla 16.14), la univariada también lleva al rechazo de las hipótesis nulas referidas a los efectos de los factores *tiempo* ( $Sig. < 0,0005$ ) y *contenido* ( $Sig. = 0,006$ ). Pero, a diferencia de lo que ocurría con la aproximación multivariada, la univariada lleva al rechazo de la hipótesis referida al efecto de la interacción.

Al producirse esta incongruencia entre ambas aproximaciones, es necesario optar por una de las dos. Según se ha señalado ya, la aproximación multivariada no exige esfericidad y, por

tanto, es una elección apropiada en condiciones de no esfericidad. Pero los datos del ejemplo no incumplen el supuesto de esfericidad (ver Tabla 16.15). Y en condiciones de esfericidad, la aproximación univariada es, según se ha dicho, más potente que la multivariada (particularmente con tamaños muestrales pequeños).

Por tanto, en el ejemplo, debe optarse por la aproximación univariada y concluir que el efecto de la interacción es significativo; y esto, tanto si se asume esfericidad ( $Sig. = 0,011$ ) como si se aplica el corrector *épsilon* en cualquiera de sus dos versiones (con *Greenhouse-Geisser*:  $Sig. = 0,040$ ; y con *Huynh-Feldt*:  $Sig. = 0,023$ ).

**Tabla 16.16.** Efectos intra-sujetos

Medida: MEASURE_1		Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
tiempo	Esfericidad asumida	145,729	3	48,576	38,058	,000
	Greenhouse-Geisser	145,729	2,260	64,495	38,058	,000
	Huynh-Feldt	145,729	3,000	48,576	38,058	,000
	Límite-inferior	145,729	1,000	145,729	38,058	,002
Error(tiempo)	Esfericidad asumida	19,146	15	1,276		
	Greenhouse-Geisser	19,146	11,298	1,695		
	Huynh-Feldt	19,146	15,000	1,276		
	Límite-inferior	19,146	5,000	3,829		
contenid	Esfericidad asumida	35,021	1	35,021	20,351	,006
	Greenhouse-Geisser	35,021	1,000	35,021	20,351	,006
	Huynh-Feldt	35,021	1,000	35,021	20,351	,006
	Límite-inferior	35,021	1,000	35,021	20,351	,006
Error(contenid)	Esfericidad asumida	8,604	5	1,721		
	Greenhouse-Geisser	8,604	5,000	1,721		
	Huynh-Feldt	8,604	5,000	1,721		
	Límite-inferior	8,604	5,000	1,721		
tiempo * contenid	Esfericidad asumida	21,062	3	7,021	5,315	,011
	Greenhouse-Geisser	21,062	1,562	13,483	5,315	,040
	Huynh-Feldt	21,062	2,145	9,821	5,315	,023
	Límite-inferior	21,062	1,000	21,062	5,315	,069
Error(tiempo*contenid)	Esfericidad asumida	19,812	15	1,321		
	Greenhouse-Geisser	19,812	7,811	2,537		
	Huynh-Feldt	19,812	10,723	1,848		
	Límite-inferior	19,812	5,000	3,962		

## Aspectos complementarios del análisis

Ya se ha señalado repetidamente que contrastar las hipótesis referidas a los tres efectos presentes en un modelo de dos factores es sólo el primer paso del análisis. El procedimiento *Medidas repetidas* permite obtener información adicional basada en algunos aspectos complementarios. Aunque estos aspectos complementarios se han descrito ya en el apartado sobre el modelo de un factor, en un análisis de varianza de dos factores hay al menos dos acciones que es preciso abordar: (1) obtener un *gráfico de perfil* representando las medias de las casillas (para poder interpretar el efecto de la interacción en el caso de que sea significativo), y (2) efectuar comparaciones múltiples entre las medias de los efectos que resulten significativos.

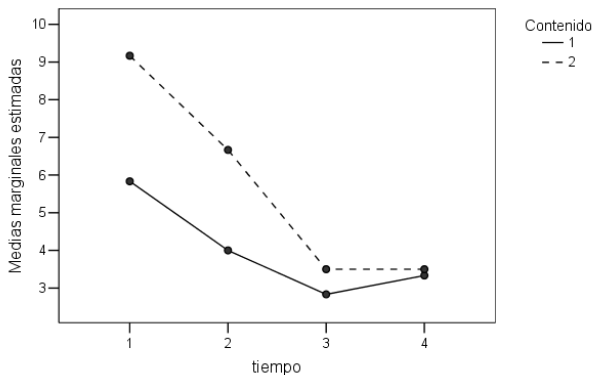
## Gráfico de perfil

Para obtener un gráfico de perfil representando el efecto de la interacción:

- Pulsar el botón **Gráficos...** del cuadro de diálogo principal (ver Figura 16.2) para acceder al subcuadro de diálogo *Medidas repetidas: Gráficos de perfil* (ver Figura 16.3).
- Seleccionar uno de los factores MR (por ejemplo, *tiempo*) en la lista **Factores** y trasladarlo al cuadro **Eje horizontal**.
- Seleccionar el otro factor MR (*contenido*) en la lista **Factores** y trasladarlo al cuadro **Líneas distintas**.
- Pulsar el botón **Añadir** para trasladar las variables seleccionadas a la lista inferior y, con ello, hacer efectiva la selección.

Aceptando estas elecciones, el *Visor de resultados* construye el gráfico de perfil que muestra la Figura 16.9.

**Figura 16.9.** Gráfico de perfil representando el efecto de la interacción *tiempo-contenido*



Las líneas del gráfico indican que la calidad del recuerdo va decreciendo con el paso del tiempo, pero sólo hasta el tercer nivel del factor (una *semana*); en el cuarto nivel (un *mes*) se aprecia un estancamiento o, incluso, una ligera mejora. Esto ocurre tanto con la lista de *números* como con la de *letras*. Sin embargo, la diferencia entre ambas listas es más evidente al principio (*hora* y *día*) que al final (*semana* y *mes*). No obstante, para poder afirmar esto último es necesario efectuar comparaciones múltiples que ayuden a descifrar el significado de la interacción. El siguiente apartado explica cómo llevar a cabo algunas de estas comparaciones: las correspondientes a los *efectos simples*.

## Comparaciones múltiples

Ya se ha señalado que, aunque las comparaciones *post hoc* no están disponibles para los factores MR, es posible efectuar comparaciones múltiples utilizando la opción **Comparar efectos principales** del cuadro de diálogo *Opciones* (ver, en este mismo capítulo, el apartado *Modelo de un factor: Aspectos complementarios del análisis: Opciones*). Para ello:

- Pulsar e botón **Opciones...** del cuadro de diálogo *Medidas repetidas* (ver Figura 16.2) para acceder al subcuadro de diálogo *Medidas repetidas: Opciones*.
- Seleccionar la variable *tiempo* en la lista Factores e interacciones de los factores y trasladarla a la lista **Mostrar las medias para**.
- Marcar la opción **Comparar los efectos principales**.
- Seleccionar la opción **Bonferroni** dentro del menú desplegable **Ajuste del intervalo de confianza** (para controlar la tasa de error).

Con estas especificaciones se obtienen los resultados que muestran las Tablas 16.17 y 16.18. La Tabla 16.17 ofrece las medias marginales que el modelo estima para cada nivel del factor *tiempo*, además del error típico y del intervalo de confianza correspondiente a cada media.

**Tabla 16.17.** Medias estimadas

Medida: MEASURE\_1

tiempo	Media	Error típ.	Intervalo de confianza al 95%.	
			Límite inferior	Límite superior
1	7,500	,516	6,173	8,827
2	5,333	,601	3,789	6,878
3	3,167	,527	1,812	4,521
4	3,417	,396	2,398	4,435

La Tabla 16.18 muestra las comparaciones por pares entre los niveles del factor *tiempo*. Para controlar la tasa de error, tanto los niveles críticos (*Sig.*) como los intervalos de confianza están ajustados mediante la corrección de Bonferroni (ver, en el capítulo anterior sobre *ANOVA factorial*, el párrafo *Comparar los efectos principales* del apartado *Opciones*). El resultado de las comparaciones indica que la calidad del recuerdo en el nivel 1 (*hora*) es significativamente mejor (*Sig.* < 0,05) que en el resto de niveles; y mejor también en el nivel 2 (*día*) que en el nivel 3 (*semana*). Los intervalos de confianza permiten llegar a la misma conclusión.

**Tabla 16.18.** Comparaciones por pares entre niveles del factor *tiempo*

Medida: MEASURE\_1

(I) tiempo	(J) tiempo	Diferencia entre medias (I-J)	Error típ.	Sig. <sup>a</sup>	Intervalo de confianza al 95 % <sup>a</sup>	
					Límite inferior	Límite superior
1	2	2,167	,477	,037	,153	4,180
	3	4,333	,401	,001	2,640	6,027
	4	4,083	,271	,000	2,939	5,228
2	1	-2,167	,477	,037	-4,180	-,153
	3	2,167	,494	,043	,081	4,253
	4	1,917	,523	,087	-,290	4,124
3	1	-4,333	,401	,001	-6,027	-2,640
	2	-2,167	,494	,043	-4,253	-,081
	4	-,250	,544	1,000	-2,545	2,045
4	1	-4,083	,271	,000	-5,228	-2,939
	2	-1,917	,523	,087	-4,124	,290
	3	,250	,544	1,000	-2,045	2,545

Basadas en las medias marginales estimadas.

a. Ajuste para comparaciones múltiples: Bonferroni.

Además de las comparaciones referidas a los efectos principales, el procedimiento **Medidas repetidas** también permite obtener información sobre los *efectos simples*, es decir, también permite comparar entre sí los niveles de un factor dentro de cada nivel del otro factor, lo cual es útil para interpretar el efecto de la interacción (aunque no agote su significado). No obstante, para poder efectuar estas comparaciones, es necesario recurrir a la sintaxis SPSS. Para evaluar los efectos *simples* mediante sintaxis:

- En el subcuadro de diálogo *Medidas repetidas: Opciones* seleccionar el efecto que contiene la interacción (en el ejemplo, *tiempo\*contenid*) y trasladarlo a la lista **Mostrar las medias para**. Para que se active la opción **Comparar los efectos principales** es necesario, además, trasladar algún efecto principal como, por ejemplo, *tiempo*.
- Marcar la opción **Comparar los efectos principales** y pulsar el botón **Continuar** para volver al cuadro de diálogo *Medidas repetidas*.
- Pulsar el botón **Pegar** para pegar en el *Editor de sintaxis* la sintaxis SPSS asociada a las elecciones hechas y modificar la línea «/EMMEANS=TABLES (tiempo\* contenid)» añadiendo: «COMPARE(contenid) ADJ(BONFERRONI)». La línea completa debe quedar así: «/EMMEANS = TABLES (tiempo\*contenid) COMPARE(contenid) ADJ(BONFERRONI)».

Estas especificaciones permiten obtener dos tablas. La primera de ellas (Tabla 16.19) ofrece las medias que el modelo estima para cada casilla (para cada nivel del factor *contenid* en cada nivel del factor *tiempo*), y el error típico y el intervalo de confianza asociado a cada media.

**Tabla 16.19.** Medias estimadas

Medida: MEASURE\_1

tiempo	contenid	Media	Error típ.	Intervalo de confianza al 95%.	
				Límite inferior	Límite superior
1	1	5,833	,477	4,606	7,060
	2	9,167	,601	7,622	10,711
2	1	4,000	,683	2,244	5,756
	2	6,667	,715	4,829	8,504
3	1	2,833	,654	1,152	4,515
	2	3,500	,428	2,399	4,601
4	1	3,333	,667	1,620	5,047
	2	3,500	,671	1,776	5,224

La Tabla 16.20 contiene el resultado de las comparaciones entre cada nivel del factor *contenid* dentro cada nivel del factor *tiempo*, es decir, contiene la información referida a los *efectos simples*. Con el fin de controlar la tasa de error (la probabilidad de cometer errores de tipo I), tanto los niveles críticos (*Sig.*) como los intervalos de confianza se ajustan mediante la corrección de Bonferroni (se indica en una nota a pie de tabla).

Los resultados de la Tabla 16.20 muestran que el recuerdo medio de *números* y *letras* difiere significativamente en los momentos temporales 1 y 2 (al cabo de una *hora* y de un *día*) pero no en los momentos 3 y 4 (al cabo de una *semana* y de un *mes*). El gráfico de perfil de la interacción (ver Figura 16.9) ayuda a comprender el significado de estas comparaciones. En él se aprecia con claridad que, a medida que va pasando el tiempo, van desapareciendo las diferencias iniciales entre el recuerdo medio de *números* y el de *letras*.



**Tabla 16.20.** Comparaciones por pares entre las medias de las casillas

Medida: MEASURE\_1

tiempo	(I) contenido	(J) contenido	Diferencia entre medias (I-J)	Error tip.	Sig. <sup>a</sup>	Intervalo de confianza al 95 % <sup>a</sup>	
						Límite inferior	Límite superior
1	1	2	-3,333	,333	,000	-4,190	-2,476
	2	1	3,333	,333	,000	2,476	4,190
2	1	2	-2,667	,715	,014	-4,504	-,829
	2	1	2,667	,715	,014	,829	4,504
3	1	2	-,667	,333	,102	-1,524	,190
	2	1	,667	,333	,102	-,190	1,524
4	1	2	-,167	1,078	,883	-2,937	2,603
	2	1	,167	1,078	,883	-2,603	2,937

Basadas en las medias marginales estimadas.

a. Ajuste para comparaciones múltiples: Bonferroni.

## Modelo de dos factores con medidas repetidas en un factor

Este modelo de ANOVA permite analizar datos provenientes de un diseño de dos factores con medidas repetidas en uno de ellos. Se trata, por tanto, de un modelo que incluye un factor inter-sujetos (con un grupo de sujetos distinto en cada nivel) y un factor intra-sujetos (por cuyos niveles pasan todos los sujetos).

### Datos

En un experimento sobre memoria se ha registrado la calidad del recuerdo en 15 sujetos al intentar evocar un texto previamente aprendido: 5 sujetos han intentado evocar el texto en condiciones de *reconocimiento*, otros 5 en condiciones de *recuerdo asistido* (con una pequeña ayuda) y otros 5 en condiciones de *recuerdo libre*. Los registros se han efectuado en cuatro momentos temporales distintos: al cabo de una *hora*, de un *día*, de una *semana* y de un *mes*.

Se trata, por tanto, de un diseño de dos factores: un factor inter-sujetos (al que puede llamarse *memoria*) con tres niveles (*reconocimiento*, *recuerdo asistido* y *recuerdo libre*) y un factor intra-sujetos (al que puede llamarse *tiempo*) con cuatro niveles (*hora*, *día*, *semana* y *mes*). Como variable dependiente se utiliza la *calidad del recuerdo*. La Tabla 16.21 muestra los datos obtenidos.

**Tabla 16.21.** Datos de un diseño de dos factores (*memoria* y *tiempo*) con medidas repetidas en un factor

	Reconocimiento					Recuerdo asistido					Recuerdo libre			
	Hora	Día	Sem.	Mes		Hora	Día	Sem.	Mes		Hora	Día	Sem.	Mes
S <sub>1</sub>	10	8	7	8	S <sub>6</sub>	8	6	5	3	S <sub>11</sub>	7	5	4	3
S <sub>2</sub>	9	8	7	6	S <sub>7</sub>	8	7	6	5	S <sub>12</sub>	8	6	4	4
S <sub>3</sub>	8	6	6	7	S <sub>8</sub>	9	7	5	6	S <sub>13</sub>	8	6	5	6
S <sub>4</sub>	7	7	6	6	S <sub>9</sub>	8	6	4	4	S <sub>14</sub>	8	5	3	4
S <sub>5</sub>	9	9	8	8	S <sub>10</sub>	7	5	4	5	S <sub>15</sub>	7	5	4	3

Para reproducir los datos de la Tabla 16.21 en el *Editor de datos* del SPSS es necesario crear cinco variables: una para definir el factor inter-sujetos y cuatro para definir los cuatro niveles del factor intra-sujetos. La Figura 16.10 muestra el aspecto del *Editor de datos* después de introducir en él los datos de la Tabla 16.21. Se ha creado la variable *memoria* haciéndole tomar los valores 1, 2 y 3 (con etiquetas: 1 = «reconocimiento», 2 = «recuerdo asistido» y 3 = «recuerdo libre»). Y para definir los cuatro niveles del factor intra-sujetos se han creado cuatro variables: *hora*, *día*, *semana* y *mes*.

**Figura 16.10.** Datos de la Tabla 16.21 reproducidos en el *Editor de datos*

	memoria	hora	día	semana	mes
1	1	10	8	7	8
2	1	9	8	7	6
3	1	8	6	6	7
4	1	7	7	6	6
5	1	9	9	8	8
6	2	8	6	5	3
7	2	8	7	6	5
8	2	9	7	5	6
9	2	8	6	4	4
10	2	7	5	4	5
11	3	7	5	4	3
12	3	8	6	4	4
13	3	8	6	5	6
14	3	8	5	3	4
15	3	7	5	4	3

## Análisis básico

Para llevar a cabo un ANOVA de dos factores con medidas repetidas en un solo factor:

- Seleccionar la opción **Modelo lineal general > Medidas repetidas...** del menú **Analizar** para acceder al cuadro de diálogo *Definición de factor(es) de medidas repetidas* (ver Figura 16.1) y asignar nombre y número de niveles al factor MR; pulsar el botón **Añadir**.
- Pulsar el botón **Definir** para acceder al cuadro de diálogo *Medidas repetidas* (ver Figura 16.2).
- Seleccionar la variables que definen los niveles del factor intra-sujetos y trasladarlas a la lista **Variables intra-sujetos** utilizando el correspondiente botón flecha.
- Seleccionar la variable que define el factor inter-sujetos y trasladarla a la lista **Factores inter-sujetos**.

### *Ejemplo: MLG > ANOVA de dos factores con medidas repetidas en un factor*

Para aplicar un ANOVA de dos factores con medidas repetidas en un factor a los datos de la Tabla 16.21 (reproducidos en la Figura 16.10 y disponibles en el archivo *ANOVA 2 repetidas en uno* que se encuentra en la página web del manual):

- Seleccionar la opción **Modelo lineal general > Medidas repetidas...** del menú **Analizar** para acceder al cuadro de diálogo *Definición de factor(es) de medidas repetidas* (ver Figura 16.1).
- Introducir el nombre del primer factor MR (*tiempo*) en el cuadro de texto **Nombre del factor intra-sujetos** y el número de niveles de que consta ese factor (4) en el cuadro de texto **Número de niveles**. Pulsar el botón **Añadir**.
- Pulsar el botón **Definir** para acceder al cuadro de diálogo *Medidas repetidas* (ver Figura 16.2), seleccionar las 4 variables (*hora*, *día*, *semana* y *mes*) que definen los cuatro niveles del factor intra-sujetos y trasladarlas a la lista **Variables intra-sujetos**.
- Seleccionar la variable que define el factor inter-sujetos (*memoria*) y trasladarla a la lista **Factores inter-sujetos**.

Aceptando estas elecciones, el *Visor* ofrece varias tablas de resultados basadas en las especificaciones que el programa tiene establecidas por defecto. Muchas de estas tablas son idénticas a las ya estudiadas en los apartados anteriores, pero ahora existe información nueva referida al efecto del factor inter-sujetos. La Tabla 16.22 muestra cuatro estadísticos multivariados, todos los cuales permiten contrastar las hipótesis nulas referidas a los efectos en los que se encuentra involucrado el factor intra-sujetos *tiempo* (sin necesidad de asumir esfericidad). De los tres efectos relevantes en este modelo (*tiempo*, *memoria* y *tiempo\*memoria*), dos de ellos tienen que ver con el factor intra-sujetos *tiempo*: el propio factor y la interacción *memoria\*tiempo*. Los cuatro estadísticos coinciden en señalar que el efecto del factor *tiempo* es significativo ( $Sig. < 0,00005$ ). Pero no ocurre lo mismo con la interacción *tiempo\*memoria*: sólo con la *raíz mayor de Roy* se considera significativo ese efecto ( $Sig. = 0,014$ ).

En relación con este resultado hay que recordar que la aproximación multivariada es más conservadora que la univariada, sobre todo con muestras pequeñas, como es el caso. Si se cumple el supuesto de esfericidad, es preferible basar las decisiones en la aproximación univariada. Y si se incumple, pero el tamaño muestral es pequeño, también es preferible utilizar la aproximación univariada acompañada del corrector *epsilon*.

Tabla.16.22. Contrastes multivariados

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Sig.
tiempo	Traza de Pillai	,95	61,96 <sup>a</sup>	3,00	10,00	,000
	Lambda de Wilks	,05	61,96 <sup>a</sup>	3,00	10,00	,000
	Traza de Hotelling	18,59	61,96 <sup>a</sup>	3,00	10,00	,000
	Raíz mayor de Roy	18,59	61,96 <sup>a</sup>	3,00	10,00	,000
tiempo * memoria	Traza de Pillai	,64	1,73	6,00	22,00	,161
	Lambda de Wilks	,38	2,07 <sup>a</sup>	6,00	20,00	,102
	Traza de Hotelling	1,58	2,36	6,00	18,00	,073
	Raíz mayor de Roy	1,54	5,65 <sup>b</sup>	3,00	11,00	,014

a. Estadístico exacto.

b. El estadístico es un límite superior para la F, el cual ofrece un límite inferior para el nivel de significación.

La Tabla 16.23 ofrece el *contraste de esfericidad de Mauchly*. Puesto que el nivel crítico asociado al estadístico *W* ( $Sig. = 0,120$ ) es mayor que 0,05, puede asumirse esfericidad y, consecuentemente, las decisiones sobre los efectos intra-sujetos pueden basarse en la aproximación univariada (estadísticos *F* de la Tabla 16.24).

Tabla 16.23. Contraste de esfericidad de *Mauchly*

Medida: MEASURE\_1

Efecto intra-sujetos	W de Mauchly	Chi-cuadrado aprox.	gl	Sig.	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	Límite-inferior
tiempo	,441	8,777	5	,120	,711	1,000	,333

Contrasta la hipótesis nula de que la matriz de covarianza error de las variables dependientes transformadas es proporcional a una matriz identidad.

La Tabla 16.24 muestra los estadísticos *F* univariados referidos a los efectos intra-sujetos. La información relativa al efecto individual del factor *tiempo* es consistente con la obtenida con la aproximación multivariada (ver Tabla 16.22). Tanto si se asume esfericidad como si, no asumiéndola, se utiliza uno cualquiera de los correctores *épsilon*, se obtiene un nivel crítico tan pequeño (*Sig.* < 0,0005) que lo razonable es rechazar la hipótesis nula referida al efecto del factor *tiempo*: puede afirmarse, por tanto, que el efecto del factor *tiempo* es significativo y, en consecuencia, que la calidad del recuerdo no es la misma en los cuatro medidas efectuadas.

Tabla 16.24. Efectos intra-sujetos

Medida: MEASURE\_1

Fuente		Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
tiempo	Esfericidad asumida	82,850	3	27,617	70,511	,000
	Greenhouse-Geisser	82,850	2,132	38,860	70,511	,000
	Huynh-Feldt	82,850	3,000	27,617	70,511	,000
	Límite-inferior	82,850	1,000	82,850	70,511	,000
tiempo * memoria	Esfericidad asumida	7,300	6	1,217	3,106	,015
	Greenhouse-Geisser	7,300	4,264	1,712	3,106	,030
	Huynh-Feldt	7,300	6,000	1,217	3,106	,015
	Límite-inferior	7,300	2,000	3,650	3,106	,082
Error(tiempo)	Esfericidad asumida	14,100	36	,392		
	Greenhouse-Geisser	14,100	25,584	,551		
	Huynh-Feldt	14,100	36,000	,392		
	Límite-inferior	14,100	12,000	1,175		

Y en lo referente al efecto de la interacción *tiempo\*memoria*, tanto si se asume esfericidad como si se aplica cualquiera de las estimaciones del corrector *épsilon*, se obtienen niveles críticos por debajo de 0,05; por lo que puede concluirse que existe efecto significativo de la interacción. Nótese que los niveles críticos referidos al efecto de la interacción están próximos a 0,05; esto, unido al pequeño tamaño de la muestra, podría estar explicando que la aproximación multivariada no haya encontrado significativo el efecto de la interacción.

Por supuesto, para poder interpretar el efecto de la interacción es necesario llevar a cabo comparaciones múltiples para los *efectos simples* y obtener un gráfico de perfil. Se harán ambas cosas más adelante.

La Tabla 16.25 contiene información referente al factor inter-sujetos *memoria*. El nivel crítico asociado al estadístico *F* (*Sig.* = 0,001) permite rechazar la hipótesis nula y afirmar que el efecto del factor *memoria* es significativo: puede concluirse que la calidad del recuerdo no es la misma en las tres condiciones de memorización establecidas.

Tabla 16.25. Efectos inter-sujetos

Medida: MEASURE\_1

Variable transformada: Promedio

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Intersección	2318,817	1	2318,817	1058,015	,000
memoria	53,633	2	26,817	12,236	,001
Error	26,300	12	2,192		

## Aspectos complementarios del análisis

Ya se ha señalado repetidamente que contrastar las hipótesis nulas referidas a los tres efectos presentes en un modelo de dos factores es sólo el primer paso del análisis. Hecho esto, debe buscarse información adicional basada en ciertos aspectos complementarios que, aunque ya se han descrito en los apartados anteriores, conviene abordar de nuevo aquí.

Para poder interpretar correctamente los efectos involucrados en un ANOVA factorial es preciso: (1) efectuar comparaciones múltiples entre las medias de los efectos significativos y (2) obtener un *gráfico de perfil* representando el efecto de la interacción (para poder interpretarlo en el caso de que resulte significativo). Además, según se verá enseguida, cuando el diseño incluye algún factor inter-sujetos es necesario establecer un nuevo supuesto relacionado con las matrices de varianzas-covarianzas.

### Igualdad entre las matrices de varianzas-covarianzas

En una ANOVA factorial mixto (un factor inter-sujetos y un factor intra-sujetos), además del supuesto de esfericidad que afecta al estadístico  $F$  de la solución univariada (ver apartado *Modelo de un factor*), debe establecerse un supuesto adicional que afecta tanto al estadístico  $F$  como a los estadísticos de la solución multivariada. El estadístico  $F$  asume *igualdad de varianzas*, es decir, que las varianzas de las medidas repetidas (las varianzas de cada nivel del factor intra-sujetos) son iguales en todas las poblaciones definidas por los niveles del factor inter-sujetos. Los estadísticos multivariados asumen *igualdad de las matrices de varianzas-covarianzas*, es decir, que las matrices de varianzas-covarianzas de las medidas repetidas son iguales en todas las poblaciones definidas por los niveles del factor inter-sujetos. Para contrastar estos supuestos:

- Pulsar el botón **Opciones...** del cuadro de diálogo principal (ver Figura 16.2) para acceder al subcuadro de diálogo *Medidas repetidas: Opciones*.
- Marcar la opción **Pruebas de homogeneidad**.

Al solicitar las pruebas de homogeneidad, el SPSS ofrece dos estadísticos: el de *Box* y el de *Levene*. La Tabla 16.26 muestra el estadístico de Box. El estadístico  $M$  (y su transformación en  $F$ ) permite contrastar la hipótesis de igualdad de las matrices de varianzas-covarianzas. En el ejemplo, el estadístico  $F$  toma un valor de 0,459 y tiene asociado un nivel crítico (*Significación*) de 0,915, lo cual permite asumir que las matrices de varianzas-covarianzas entre las 4 medidas repetidas (*hora*, *día*, *semana* y *mes*) son iguales en las tres poblaciones definidas por el factor *memoria*.

Tabla 16.26. Contraste de *Box* sobre la igualdad de las matrices de varianzas-covarianzas

M de Box	10,680
F	,459
gl1	10
gl2	305,976
Significación	,915

La Tabla 16.27 ofrece el estadístico *F* de Levene. Este estadístico contrasta la hipótesis de igualdad de varianzas. Esta hipótesis se contrasta para cada nivel del factor intra-sujetos. En el ejemplo, puesto que todos los niveles críticos son mayores que 0,05, puede asumirse que, en las cuatro medidas utilizadas, las varianzas de las tres poblacionales definidas por el factor *memoria* son iguales.

Tabla 16.27. Contraste de Levene sobre la igualdad de las varianzas error

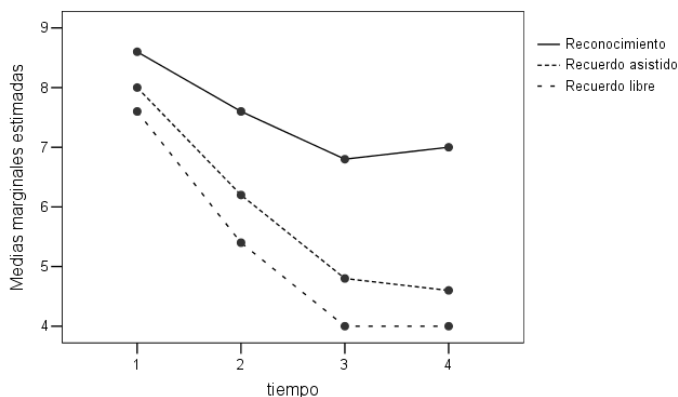
	F	gl1	gl2	Sig.
Una hora	1,540	2	12	,254
Un día	1,143	2	12	,351
Una semana	,426	2	12	,663
Un mes	,026	2	12	,974

## Gráfico de perfil

Para obtener un gráfico de perfil representando el efecto de la interacción:

- Pulsar el botón **Gráficos...** del cuadro de diálogo principal (ver Figura 16.2) para acceder al subcuadro de diálogo *Medidas repetidas: Gráficos de perfil* (ver Figura 16.3).
- Seleccionar el factor intra-sujetos *tiempo* en la lista **Factores** y trasladarlo al cuadro **Eje horizontal**.
- Seleccionar el factor inter-sujetos *memoria* en la lista **Factores** y trasladarlo al cuadro **Líneas distintas**.
- Pulsar el botón **Añadir** para trasladar las variables seleccionadas a la lista inferior y, con ello, hacer efectiva la selección.

Aceptando estas elecciones, el *Visor de resultados* construye el gráfico de perfil que muestra la Figura 16.11. Las líneas del gráfico permiten apreciar que la calidad del recuerdo va disminuyendo paulatinamente con el paso del tiempo hasta llegar al nivel 3 (una *semana*), momento a partir del cual se produce un estancamiento final. También puede observarse en el gráfico que la disminución de la calidad del recuerdo no es igualmente intensa en las tres condiciones definidas por el factor *memoria*. Parece claro que bajo la condición de *reconocimiento* se produce una pérdida de la calidad del recuerdo sensiblemente menor que en las otras dos condiciones. No obstante, para poder interpretar correctamente el efecto de la interacción es necesario llevar a cabo comparaciones múltiples. El siguiente apartado explica cómo realizar algunas de estas comparaciones; entre ellas, las que permiten evaluar los efectos *simples*, los cuales, aunque no agotan el significado de la interacción, permiten formarse una idea bastante precisa sobre lo que está ocurriendo.

Figura 16.11. Gráfico de perfil representando el efecto de la interacción *tiempo-memoria*

### Comparaciones múltiples

Al efectuar comparaciones múltiples es necesario seguir estrategias distintas dependiendo de que se intente evaluar un efecto inter-sujetos o un efecto intra-sujetos. Para evaluar un efecto inter-sujetos pueden utilizarse los métodos del cuadro de diálogo *Post hoc*. Por el contrario, las comparaciones múltiples referidas a los efectos intra-sujetos deben obtenerse utilizando la opción **Comparar efectos principales** del cuadro de diálogo *Opciones*. Para obtener comparaciones múltiples referidas a los *efectos intra-sujetos*:

- Pulsar el botón **Opciones...** del cuadro de diálogo principal (ver Figura 16.2) para acceder al subcuadro de diálogo *Medidas repetidas: Opciones*.
- Seleccionar los efectos *tiempo* y *tiempo\*memoria* en la lista **Factores e interacciones de los factores** y trasladarla a la lista **Mostrar las medias para**.

*Nota:* aunque sólo interesara evaluar el efecto de la interacción, también habría que trasladar el factor *tiempo* (o cualquier otro factor principal) a la lista **Mostrar las medias para** a fin de activar la opción **Comparar los efectos principales**.

- Marcar la opción **Comparar los efectos principales**.
- Seleccionar la opción **Bonferroni** dentro del menú desplegable **Ajuste del intervalo de confianza** (la corrección de Bonferroni –también la de Sidak– se utiliza para controlar la tasa de error cuando se efectúan varias comparaciones). Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Pegar** para escribir en el *Editor de sintaxis* la sintaxis SPSS correspondiente a las elecciones hechas y modificar la línea «/EMMEANS=TABLES (tiempo\* memoria)» añadiendo lo siguiente: «COMPARE(memoria) ADJ(BONFERRONI)». La línea completa debe quedar así: «/EMMEANS=TABLES (tiempo\*memoria) COMPARE(memoria) ADJ(BONFERRONI)».

Con estas especificaciones se obtienen, además de las dos tablas de medias estimadas para cada nivel y para combinación entre niveles (ver, por ejemplo, las Tablas 16.17 y 16.18), dos tablas de comparaciones. La primera de estas tablas ofrece las comparaciones por pares entre

los niveles del factor *tiempo* (Tabla 16.28). Para controlar la tasa de error (es decir, la probabilidad de cometer errores de tipo I), tanto los niveles críticos (*Sig.*) como los intervalos de confianza están ajustados mediante la corrección de Bonferroni.

El resultado de estas comparaciones por pares indica que la calidad del recuerdo en el nivel 1 (una *hora*) es significativamente mejor (*Sig.* < 0,0005) que en el resto de niveles; y mejor también en el nivel 2 (un *día*) que en el nivel 3 (una *semana*; *Sig.* < 0,0005) y 4 (un *mes*; *Sig.* = 0,006). No existen, sin embargo, diferencias significativas (*Sig.* = 1,00) entre los niveles 3 (una *semana*) y 4 (un *mes*).

**Tabla 16.28.** Comparaciones por pares entre los niveles del factor *tiempo*

Medida: MEASURE\_1

(I) tiempo	(J) tiempo	Diferencia entre medias (I-J)	Error típ.	Sig. <sup>a</sup>	Intervalo de confianza al 95 % <sup>a</sup>	
					Límite inferior	Límite superior
1	2	1,667	,176	,000	1,111	2,223
	3	2,867	,221	,000	2,170	3,564
	4	2,867	,254	,000	2,066	3,667
2	1	-1,667	,176	,000	-2,223	-1,111
	3	1,200	,133	,000	,780	1,620
	4	1,200	,275	,006	,333	2,067
3	1	-2,867	,221	,000	-3,564	-2,170
	2	-1,200	,133	,000	-1,620	-,780
	4	,000	,275	1,000	-,867	,867
4	1	-2,867	,254	,000	-3,667	-2,066
	2	-1,200	,275	,006	-2,067	-,333
	3	,000	,275	1,000	-,867	,867

Basadas en las medias marginales estimadas.

a. Ajuste para comparaciones múltiples: Bonferroni.

La segunda tabla de comparaciones recoge el resultado de comparar por pares los niveles del factor *memoria* dentro cada nivel del factor *tiempo* (Tabla 16.29). Es decir, recoge la información relativa a los *efectos simples*. Al igual que antes, y con el fin de controlar la tasa de error, tanto los niveles críticos (*Sig.*) como los intervalos de confianza se han ajustado mediante la corrección de Bonferroni.

Los resultados de estas comparaciones, junto con el gráfico de perfil (ver Figura 16.11), ayudan a comprender el significado del efecto de la interacción. En primer lugar, en el momento temporal 1 (una *hora*), la calidad del recuerdo es similar en las tres condiciones de *memoria* (*Sig.* > 0,05 en todos los casos). Pero a partir de ese momento la calidad del recuerdo se muestra significativamente mejor en la condición *reconocimiento* que en las condiciones *recuerdo asistido* y *recuerdo libre* (*Sig.* < 0,05), con la única excepción del momento 2 (un *día*), en el que la diferencia entre *reconocimiento* y *recuerdo asistido* no alcanza a ser significativa (*Sig.* = 0,080). Las condiciones *recuerdo asistido* y *recuerdo libre* no difieren en ninguno de los momentos temporales considerados. El gráfico de perfil sobre el efecto de la interacción (ver Figura 16.11) ayuda a comprender más fácilmente el significado de las diferencias encontradas.

En este momento es importante señalar que el estudio de los *efectos simples* no agota el significado de la interacción. De hecho, un efecto simple incluye información tanto de la interacción como del efecto principal involucrado. En el ejemplo, se ha encontrado que la diferen-



cia entre las tres condiciones de *memoria* no es la misma en los cuatro momentos medidos; pero no se ha hecho ninguna comparación entre las diferencias existentes en cada momento. Por ejemplo, se ha encontrado que la diferencia entre *reconocimiento* y *recuerdo asistido* al cabo de *una hora* (0,60) no es significativa ( $p=0,250$ ); y se ha encontrado que esa diferencia al cabo de *una semana* (2,00) sí es significativa ( $p=0,006$ ); pero no se ha realizado ninguna comparación entre esas dos diferencias (0,60–2,00) para valorar su significación.

**Tabla 16.29.** Comparaciones entre los niveles del factor *memoria* en cada nivel del factor *tiempo*

Medida: MEASURE\_1

Tiempo	(I) Memoria	(J) Memoria	Diferencia entre medias (I-J)	Error típ.	Sig. <sup>a</sup>	Intervalo de confianza al 95 % <sup>a</sup>	
						Límite inferior	Límite superior
1	Reconocimiento	Recuerdo asistido	,600	,529	,837	-,871	2,071
		Recuerdo libre	1,000	,529	,250	-,471	2,471
	Recuerdo asistido	Reconocimiento	-,600	,529	,837	-2,071	,871
		Recuerdo libre	,400	,529	1,000	-1,071	1,871
	Recuerdo libre	Reconocimiento	-1,000	,529	,250	-2,471	,471
		Recuerdo asistido	-,400	,529	1,000	-1,871	1,071
2	Reconocimiento	Recuerdo asistido	1,400	,554	,080	-,139	2,939
		Recuerdo libre	2,200	,554	,006	,661	3,739
	Recuerdo asistido	Reconocimiento	-1,400	,554	,080	-2,939	,139
		Recuerdo libre	,800	,554	,522	-,739	2,339
	Recuerdo libre	Reconocimiento	-2,200	,554	,006	-3,739	-,661
		Recuerdo asistido	-,800	,554	,522	-2,339	,739
3	Reconocimiento	Recuerdo asistido	2,000	,503	,006	,601	3,399
		Recuerdo libre	2,800	,503	,000	1,401	4,199
	Recuerdo asistido	Reconocimiento	-2,000	,503	,006	-3,399	-,601
		Recuerdo libre	,800	,503	,414	-,599	2,199
	Recuerdo libre	Reconocimiento	-2,800	,503	,000	-4,199	-1,401
		Recuerdo asistido	-,800	,503	,414	-2,199	,599
4	Reconocimiento	Recuerdo asistido	2,400	,712	,017	,422	4,378
		Recuerdo libre	3,000	,712	,004	1,022	4,978
	Recuerdo asistido	Reconocimiento	-2,400	,712	,017	-4,378	-,422
		Recuerdo libre	,600	,712	1,000	-1,378	2,578
	Recuerdo libre	Reconocimiento	-3,000	,712	,004	-4,978	-1,022
		Recuerdo asistido	-,600	,712	1,000	-2,578	1,378

Basadas en las medias marginales estimadas.

a. Ajuste para comparaciones múltiples: Bonferroni.

Para efectuar comparaciones múltiples entre los niveles del **factor inter-sujetos**:

- Pulsar el botón **Post hoc...** del cuadro de diálogo *Medidas repetidas* (ver Figura 16.2) para acceder al subcuadro de diálogo *Medidas repetidas: Comparaciones post hoc* (las opciones de este cuadro de diálogo ya se han descrito en el Capítulo 14 sobre *ANOVA de un factor*).
- Marcar la opción **Tukey** del recuadro **Asumiendo varianzas iguales**. Puesto que, de acuerdo con el resultado del contraste de Levene –ver Tabla 16.27–, puede asumirse que las varianzas poblacionales de los niveles del factor inter-sujetos son iguales, en

este momento se ha optado por seleccionar un método del recuadro **Asumiendo varianzas iguales**. Si todavía no se conociera el resultado del contraste de Levene, lo apropiado sería seleccionar, además de la prueba de Tukey, alguno de los métodos del recuadro **No asumiendo varianzas iguales** para, más tarde, decidir en cuál de los dos métodos basar la decisión.

Aceptando estas elecciones, el *Visor de resultados* ofrece las comparaciones por pares que muestra la Tabla 16.30. Los resultados obtenidos con la prueba de Tukey indican que la calidad del recuerdo obtenida en la condición *reconocimiento* difiere significativamente de la obtenida en las condiciones *recuerdo asistido* ( $Sig. = 0,013$ ) y *recuerdo libre* ( $Sig. = 0,001$ ). Y que entre las condiciones *recuerdo asistido* y *recuerdo libre* no existen diferencias ( $Sig. = 0,377$ ).

**Tabla 16.30.** Comparaciones *post-hoc* (Tukey) entre los niveles del factor inter-sujetos *memoria*

Medida: MEASURE\_1

DHS de Tukey

(I) Memoria	(J) Memoria	Diferencia entre medias (I-J)	Error típ.	Sig.	Intervalo de confianza al 95%.	
					Límite inferior	Límite superior
Reconocimiento	Recuerdo asistido	1,60	,468	,013	,35	2,85
	Recuerdo libre	2,25	,468	,001	1,00	3,50
Recuerdo asistido	Reconocimiento	-1,60	,468	,013	-2,85	-,35
	Recuerdo libre	,65	,468	,377	-,60	1,90
Recuerdo libre	Reconocimiento	-2,25	,468	,001	-3,50	-1,00
	Recuerdo asistido	-,65	,468	,377	-1,90	,60

Basado en las medias observadas.

Por supuesto, cuando el efecto de la interacción es significativo, la interpretación de los efectos principales pierde importancia. Aunque es cierto que la calidad del recuerdo obtenida en la condición *reconocimiento* es, globalmente considerada, mejor que la obtenida en las condiciones *recuerdo asistido* y *recuerdo libre*, las comparaciones efectuadas a propósito de los efectos simples indican que esto no es cierto al cabo de *una hora* ni al cabo de *un día*.



## Relación entre variables

### El procedimiento *Correlaciones*

Cuando se analizan datos, el interés del analista suele centrarse en dos grandes objetivos: comparar grupos y estudiar relaciones. En el Capítulo 12 se han descrito ambos aspectos del análisis referidos a variables categóricas. En los Capítulos 13, 14, 15 y 16 se ha estudiado una serie de técnicas de análisis diseñadas para comparar grupos en una variable cuantitativa. Falta saber cómo estudiar la relación entre variables cuantitativas.

Suele decirse que los sujetos más frustrados son también más agresivos; que cuanto mayor es el nivel educativo, mayor es el nivel de renta; que los niveles altos de colesterol en sangre suelen ir acompañados de dietas alimenticias ricas en grasas; que el consumo de cigarrillos está asociado a problemas de tipo vascular; que los sujetos muestran más interés por una tarea cuanto mayor es el tamaño de la recompensa que reciben; que la ingestión de alcohol produce problemas perceptivos; etc.

En todos estos ejemplos se habla de *relación entre dos variables*. En este capítulo se estudian algunos estadísticos que permiten cuantificar el grado de relación existente entre dos variables. El procedimiento *Correlaciones* incluye tres opciones: (1) *Bivariadas*, para el estudio de la relación entre dos variables cuantitativas, (2) *Parciales*, para el estudio de la relación entre dos variables cuantitativas cuando se controla o elimina el efecto de terceras variables y (3) *Distancias*, para el estudio de la relación entre dos variables cualquiera que sea su nivel de medida. En este capítulo se describen los tres procedimientos.

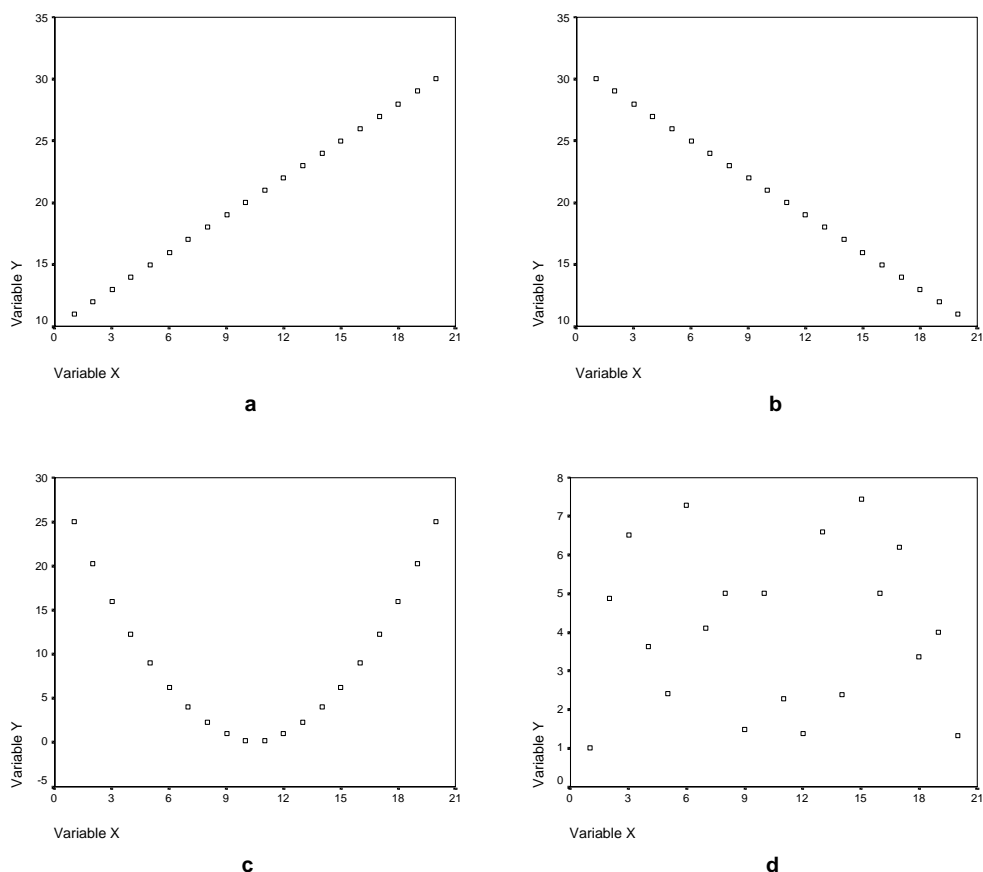
### Correlaciones bivariadas

El concepto de *relación* o *correlación* se refiere al grado de parecido o variación conjunta existente entre dos o más variables. Este apartado aborda el estudio de un tipo particular de relación llamada *lineal* y se limita a considerar únicamente el caso de dos variables cuantitativas (*correlación simple*).

Una relación lineal *positiva* entre dos variables *X* e *Y* significa que los valores de las dos variables varían de forma parecida: los sujetos que puntúan alto en *X* tienden a puntuar alto en *Y* y los sujetos que puntúan bajo en *X* tienden a puntuar bajo en *Y*. Una relación lineal *negativa* significa que los valores de ambas variables varían justamente al revés: los sujetos que puntúan alto en *X* tienden a puntuar bajo en *Y* y los sujetos que puntúan bajo en *X* tienden a puntuar alto en *Y*.

La forma más directa e intuitiva de formarse una primera impresión sobre el tipo de relación existente entre dos variables cuantitativas es a través de un *diagrama de dispersión* (este tipo de diagramas puede obtenerse mediante la opción **Dispersión...** del menú **Gráficos**). Un diagrama de dispersión es un gráfico en el que una de las variables ( $X$ ) se coloca en el eje de abscisas, la otra variable ( $Y$ ) en el de ordenadas y los pares de puntuaciones de cada sujeto ( $x, y$ ) se representan como una nube de puntos. La forma de la nube de puntos informa sobre el tipo de relación existente entre las variables. La Figura 17.1 muestra cuatro diagramas de dispersión que reflejan cuatro tipos de diferentes de relación.

**Figura 17.1.** Diagramas de dispersión expresando diferentes tipos de relación



La Figura 17.1.a muestra una situación en la que cuanto mayores son las puntuaciones en una de las variables, mayores son también las puntuaciones en la otra; cuando ocurre esto, los puntos se sitúan en una línea recta ascendente y se habla de relación *lineal positiva*. La Figura 17.1.b representa una situación en la que cuanto mayores son las puntuaciones en una de las variables, menores son las puntuaciones en la otra; en este caso, los puntos se sitúan en una línea recta descendente y se habla de relación *lineal negativa*. En la situación representada en la Figura 17.1.c también existe una pauta de variación clara, pero no es lineal: los puntos

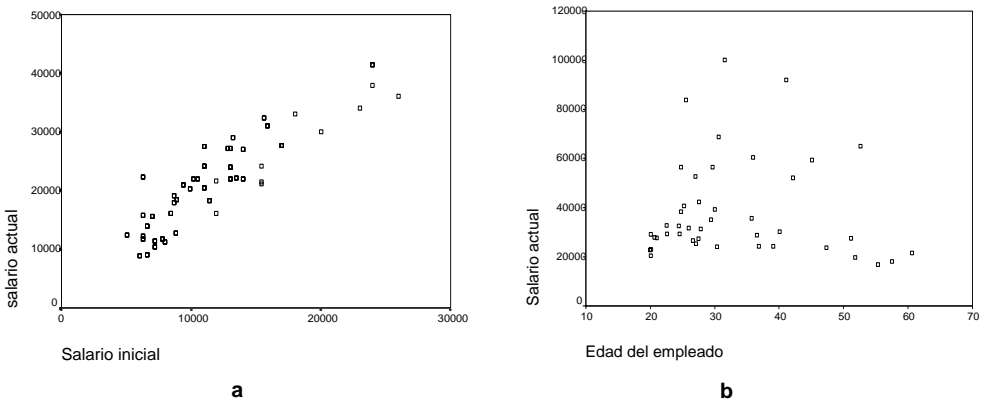
no dibujan una línea recta. Y en la Figura 17.1.d no parece existir ninguna pauta de variación clara, lo cual queda reflejado en una nube de puntos dispersa, muy lejos de lo que podría ser una línea recta.

Parece claro, por tanto, que un diagrama de dispersión permite formarse una idea bastante aproximada sobre el *tipo de relación* existente entre dos variables. Pero, además, observando los diagramas de la Figura 17.1, se desprende que un diagrama de dispersión también puede utilizarse como una forma de *cuantificar* el grado de relación lineal existente entre dos variables: basta con observar el grado en el que la nube de puntos se ajusta a una línea recta.

Sin embargo, utilizar un diagrama de dispersión como una forma de cuantificar la relación entre dos variables no es, en la práctica, tan útil como puede parecer a primera vista. Esto es debido a que la relación entre dos variables no siempre es perfecta o nula: habitualmente no es ni lo uno ni lo otro. Consideremos los diagramas de dispersión de la Figura 17.2. En el diagrama de la Figura 17.2.a, los puntos, aun no estando situados todos ellos en una línea recta, se aproximan bastante a ella. Podría encontrarse una línea recta ascendente que representara de forma bastante aproximada el conjunto total de los puntos del diagrama, lo cual indica que la relación entre las variables *salario inicial* y *salario actual* es lineal y positiva: a mayor *salario inicial*, mayor *salario actual*.

En el diagrama de la Figura 17.2.b, por el contrario, da la impresión de que no hay forma de encontrar una recta a la que poder aproximar los puntos. Al margen de que entre las variables *edad* y *salario actual* pueda existir algún tipo de relación, parece claro que la relación no es de tipo lineal.

Figura 17.2. Diagramas de dispersión representando *relación lineal* (a) e *independencia lineal* (b)



Estas consideraciones sugieren que hay nubes de puntos a las que es posible ajustar una línea recta mejor de lo que es posible hacerlo a otras. Por lo que el ajuste de una recta a una nube de puntos no parece una cuestión de todo o nada, sino más bien de *grado* (más o menos ajuste). Lo cual advierte sobre la necesidad de utilizar algún índice numérico capaz de cuantificar ese grado de ajuste con mayor precisión de lo que permite hacerlo una simple inspección del diagrama de dispersión.

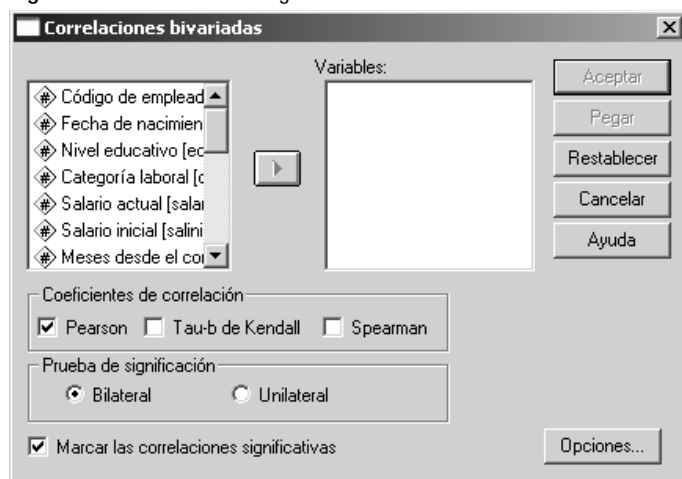
Estos índices numéricos suelen denominarse *coeficientes de correlación* y poseen la importante propiedad de permitir cuantificar el grado de relación lineal existente entre dos variables cuantitativas. Por supuesto, al mismo tiempo que permiten cuantificar el grado de

relación lineal existente entre dos variables, también sirven para valorar el grado de ajuste de la nube de puntos a una línea recta.

El procedimiento **Correlaciones bivariadas** ofrece tres de estos coeficientes:  $r_{xy}$  de Pearson,  $\tau_{b}$  de Kendall y  $\rho$  de Spearman. Para obtener estos coeficientes de correlación:

- Seleccionar la opción **Correlaciones > Bivariadas...** del menú **Analizar** para acceder al cuadro de diálogo *Correlaciones bivariadas* que muestra la Figura 17.3.

Figura 17.3. Cuadro de diálogo *Correlaciones bivariadas*



La lista de variables muestra únicamente las variables del archivo de datos que poseen formato numérico. Desde este cuadro de diálogo es posible obtener varios coeficientes de correlación y algunos estadísticos descriptivos básicos. Para ello:

- Seleccionar las variables cuantitativas (al menos ordinales) cuyo grado de relación se desea cuantificar y trasladarlas a la lista **Variables**. Es necesario trasladar al menos dos variables.

**Coeficientes de correlación.** Pueden seleccionarse uno o más de los siguientes tres coeficientes de correlación:

- “ **Pearson.** El coeficiente de correlación de Pearson (1896) es, quizá, el más conocido y utilizado para estudiar el grado de relación lineal existente entre dos variables cuantitativas. Se suele representar por  $r$  y se obtiene tipificando el promedio de los productos de las puntuaciones diferenciales (desviaciones de la media) de cada caso en las dos variables correlacionadas:

$$r_{xy} = \frac{\sum xy}{n S_x S_y}$$

( $x$  y  $y$  se refieren a las puntuaciones diferenciales de cada par;  $n$  al número de casos;  $S_x$  y  $S_y$  a las desviaciones típicas de cada variable).

El coeficiente de correlación de Pearson toma valores entre  $-1$  y  $1$ : un valor de  $1$  indica relación lineal perfecta positiva; un valor de  $-1$  indica relación lineal perfecta negativa (en ambos casos los puntos del correspondiente diagrama de dispersión se encuentran dispuestos en una línea recta); un valor de  $0$  indica relación lineal nula (lo que ocurre, por ejemplo, en los ejemplos de las Figuras 17.1.c y 17.1.d). El coeficiente  $r_{xy}$  es una medida simétrica: la correlación entre  $X$  e  $Y$  es la misma que entre  $Y_i$  y  $X_i$ .

Es muy importante tener presente que un coeficiente de correlación alto no implica *causalidad*. Dos variables pueden estar linealmente relacionadas (incluso muy relacionadas) sin que una sea causa de la otra.

Al marcar la opción **Pearson** el *Visor* ofrece una matriz de correlaciones cuadrada, con *unos* en la diagonal (pues la relación entre una variable y ella misma es perfecta –si bien esos *unos* son el resultado de tipificar la varianza de cada variable) y con los coeficientes de correlación entre cada dos variables duplicados en los triángulos superior e inferior de la matriz. Cada coeficiente aparece acompañado del número de casos sobre el que ha sido calculado y del nivel crítico que le corresponde bajo la hipótesis nula de que su verdadero valor poblacional es cero.

- " **Tau-b de Kendall.** Este coeficiente de correlación es apropiado para estudiar la relación entre variables ordinales. Se basa en el número de inversiones y no inversiones entre casos y ya ha sido descrito en el Capítulo 12, en el apartado *Estadísticos: Datos ordinales*. Toma valores entre  $-1$  y  $1$ , y se interpreta exactamente igual que el coeficiente de correlación de Pearson. La utilización de este coeficiente tiene sentido si las variables no alcanzan el nivel de medida de intervalo y/o no puede asumirse que la distribución poblacional conjunta de las variables sea normal.
- " **Spearman.** El coeficiente de correlación *rho* de Spearman (1904) es el coeficiente de correlación de Pearson, pero aplicado después de transformar las puntuaciones originales en rangos. Toma valores entre  $-1$  y  $1$ , y se interpreta exactamente igual que el coeficiente de correlación de Pearson. Al igual que ocurre con el coeficiente *tau-b* de Kendall, el de Spearman puede utilizarse como una alternativa al de Pearson cuando las variables estudiadas son ordinales y/o se incumple el supuesto de normalidad.

**Prueba de significación.** Junto con cada coeficiente de correlación, el *Visor* ofrece la información necesaria para contrastar la hipótesis nula de que el valor poblacional del coeficiente es cero. Esta hipótesis se contrasta mediante un valor tipificado que, en el caso del coeficiente de correlación de Pearson, adopta la siguiente forma:

$$T = \frac{r_{xy} \sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

Si se asume que la muestra utilizada ha sido aleatoriamente extraída de una población en la que las dos variables correlacionadas se distribuyen normalmente, el estadístico  $T$  se distribuye según el modelo de probabilidad  $t$  de Student con  $n-2$  grados de libertad. El SPSS permite seleccionar el nivel crítico deseado:



**Bilateral.** Opción apropiada para cuando no existen expectativas sobre la *dirección* de la relación. Indica la probabilidad de obtener coeficientes tan alejados de cero o más que el valor obtenido.

**Unilateral.** Opción apropiada para cuando existen expectativas sobre la *dirección* de la relación. Indica la probabilidad de obtener coeficientes iguales o mayores que el obtenido si el coeficiente es positivo, o iguales o menores que el obtenido si el coeficiente es negativo.

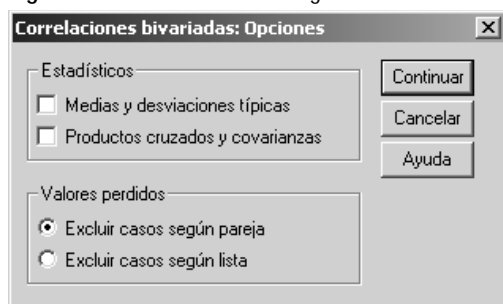
- " **Marcar las correlaciones significativas.** Esta opción, que se encuentra activa por defecto, permite obtener el nivel crítico exacto asociado a cada coeficiente de correlación. Si se desactiva esta opción, en lugar del nivel crítico, el *Visor* muestra un asterisco al lado de los coeficientes con nivel crítico menor que 0,05 y dos asteriscos al lado de los coeficientes con nivel crítico menor que 0,01.

## Opciones

El subcuadro de diálogo *Opciones* permite obtener alguna información adicional (algunos estadísticos descriptivos, la covarianza, etc.) y controlar el tratamiento que se desea dar a los valores perdidos:

- ' Pulsar el botón **Opciones...** del cuadro de diálogo principal (ver Figura 17.3) para acceder al subcuadro de diálogo *Correlaciones bivariadas: Opciones* que muestra la Figura 17.4.

Figura 17.4. Subcuadro de diálogo *Correlaciones bivariadas: Opciones*



**Estadísticos.** Si se ha elegido el coeficiente de correlación de Pearson (ver Figura 17.3), las opciones de este recuadro permiten seleccionar una o más de las siguientes opciones:

- " **Medias y desviaciones típicas.** Ofrece, para cada variable, la media, la desviación típica insesgada y el número de casos válidos.
- " **Productos cruzados y covarianzas.** Ofrece, para cada par de variables, el numerador del coeficiente de correlación de Pearson (es decir, los productos cruzados de las desviaciones de cada puntuación respecto de su media) y ese mismo numerador dividido por  $n - 1$  (es decir, la covarianza).

**Valores perdidos.** Las dos opciones de este recuadro permiten seleccionar el tratamiento que se desea dar a los valores perdidos.

**Excluir casos según pareja.** Se excluyen del cálculo de cada coeficiente de correlación los casos que poseen valor perdido en alguna de las dos variables que se están correlacionando.

**Excluir casos según lista.** Se excluyen del cálculo de todos los coeficientes de correlación solicitados los casos que poseen valor perdido en cualquiera de las variables seleccionadas en la lista **Variables**.

### **Ejemplo: Correlaciones > Bivariadas**

Este ejemplo muestra cómo obtener los coeficientes de correlación y los estadísticos del procedimiento **Correlaciones bivariadas**. Se utiliza el archivo *Datos de empleados* (ubicado en la misma carpeta en la que está instalado el SPSS).

- En el cuadro de diálogo principal (ver Figura 17.3), seleccionar las variables *salini* (salario inicial), *salario* (salario actual) y *tiempemp* (meses desde el contrato) y trasladarlas a la lista **Variables**.
- Marcar las opciones **Pearson**, **Tau-b de Kendall** y **Spearman** del recuadro **Coefficientes de correlación**.
- Pulsar el botón **Opciones...** para acceder al cuadro de diálogo *Correlaciones bivariadas: Opciones* (ver Figura 17.4) y, en el recuadro **Estadísticos**, marcar las opciones **Medias y desviaciones típicas** y **Productos cruzados y covarianzas**.

Aceptando estas elecciones, el *Visor de resultados* ofrece la información que recogen las Tablas 17.1 a la 17.3. La Tabla 17.1 contiene información descriptiva: la media aritmética, la desviación típica insesgada y el número de casos válidos.

**Tabla 17.1.** Estadísticos descriptivos

	Media	Desviación típica	N
Salario actual	\$34,419.57	\$17,075.661	474
Salario inicial	\$17,016.09	\$7,870.638	474
Meses desde el contrato	81,11	10,061	474

La Tabla 17.2 ofrece información sobre el *coeficiente de correlación de Pearson* y su significación estadística. Cada celda contiene cinco valores referidos al cruce entre cada par de variables: (1) el valor del coeficiente de correlación de Pearson; (2) el nivel crítico bilateral que corresponde a ese coeficiente (*Sig. bilateral*; el nivel crítico unilateral puede obtenerse dividiendo por 2 el bilateral); (3) la suma de los valores elevados al cuadrado (para el cruce de una variable consigo misma) o la suma de productos cruzados (para el cruce de dos variables distintas); (4) la covarianza (que se obtiene dividiendo la suma de productos cruzados entre el número de casos válidos); y (5) el número de casos válidos (*N*) sobre el que se han efectuado los cálculos.

El nivel crítico permite decidir sobre la hipótesis nula de independencia lineal (o lo que es lo mismo, sobre la hipótesis de que el coeficiente de correlación vale cero en la población). Se rechazará la hipótesis nula de independencia (y se concluirá que existe relación lineal significativa) cuando el nivel crítico sea menor que el nivel de significación establecido (generalmente, 0,05). Así, observando los niveles críticos de la Tabla 17.2, puede afirmarse que las variables *salario inicial* y *salario actual* correlacionan significativamente ( $Sig. < 0,0005$ ) y que la variable *meses desde el contrato* no correlaciona ni con la variable *salario inicial* ( $Sig. = 0,668$ ) ni con la variable *salario actual* ( $Sig. = 0,067$ ).

El SPSS no puede calcular un coeficiente de correlación cuando todos los casos de una de las variables (o de las dos) son casos con valor perdido, o cuando todos los casos tienen el mismo valor en una o en las dos variables correlacionadas (si todos los valores son iguales la desviación típica de esa variable vale cero). Cuando ocurre esto, el SPSS sustituye el coeficiente de correlación por una coma. También muestra una coma en lugar del nivel crítico ( $Sig.$ ) correspondiente al cruce de una variable consigo misma.

Tabla 17.2. Coeficientes de correlación de *Pearson* y covarianzas

		Salario actual	Salario inicial	Meses desde el contrato
Salario actual	Correlación de Pearson	1.000	.880**	.084
	Sig. (bilateral)	.	.000	.067
	Suma de cuad. y prod. cruzados	137916495436.340	55948605047.732	6833347.489
	Covarianza	291578214.453	118284577.268	14446.823
	N	474	474	474
Salario inicial	Correlación de Pearson	.880**	1.000	-.020
	Sig. (bilateral)	.000	.	.668
	Suma de cuad. y prod. cruzados	55948605047.732	29300904965.454	-739866.498
	Covarianza	118284577.268	61946944.959	-1564.200
	N	474	474	474
Meses desde el contrato	Correlación de Pearson	.084	-.020	1.000
	Sig. (bilateral)	.067	.668	.
	Suma de cuad. y prod. cruzados	6833347.489	-739866.498	47878.295
	Covarianza	14446.823	-1564.200	101.223
	N	474	474	474

\*\* La correlación es significativa al nivel 0,01 (bilateral).

La Tabla 17.3 recoge la información referida a los coeficientes *tau-b de Kendall* (primera mitad) y *rho de Spearman* (segunda mitad). Esta tabla ofrece tres valores para cada cruce de variables: (1) el valor del coeficiente de correlación; (2) el nivel crítico asociado a cada coeficiente ( $Sig.$ ); y (3) el número de casos sobre el que se ha calculado cada coeficiente.

Puesto que tanto el coeficiente de Kendall como el de Spearman se basan en las propiedades ordinales de los datos, sus valores y niveles críticos no tienen por qué coincidir con los obtenidos mediante el coeficiente de correlación de Pearson. De hecho, por ejemplo, la relación entre las variables *salario* (salario actual) y *tiempemp* (meses desde el contrato), que con el coeficiente de correlación de Pearson no alcanzaba a ser significativa ( $Sig. = 0,067$ ), ahora ha pasado a ser significativa tanto con el coeficiente *tau-b* ( $Sig. = 0,022$ ) como con el coeficiente *rho* ( $Sig. = 0,023$ ). Y la relación entre *salario* y *salini*, aunque resulta significativa con los tres coeficientes utilizados, su valor ha bajado de 0,88 con el coeficiente de correlación de Pearson a 0,656 con el de Kendall.

Tabla 17.3. Coeficientes de correlación *tau-b* de Kendall y *rho* de Spearman

		Tau b de Kendall			Rho de Spearman		
		Salario actual	Salario inicial	Meses desde el contrato	Salario actual	Salario inicial	Meses desde el contrato
Salario actual	Coef. correlación	1.000	.656**	.071*	1.000	.826**	.105*
	Sig. (bilateral)	.	.000	.022	.	.000	.023
	N	474	474	474	474	474	474
Salario inicial	Coef. correlación	.656**	1.000	-.046	.826**	1.000	-.063
	Sig. (bilateral)	.000	.	.146	.000	.	.168
	N	474	474	474	474	474	474
Meses desde el contrato	Coef. correlación	.071*	-.046	1.000	.105*	-.063	1.000
	Sig. (bilateral)	.022	.146	.	.023	.168	.
	N	474	474	474	474	474	474

\*\* La correlación es significativa al nivel 0,01 (bilateral).

\* La correlación es significativa al nivel 0,05 (bilateral).

## Correlaciones parciales

El procedimiento **Correlaciones > parciales** permite estudiar la relación lineal existente entre dos variables cuantitativas controlando el posible efecto de una o más variables cuantitativas extrañas. Un coeficiente de *correlación parcial* es una técnica de control estadístico que expresa el grado de relación lineal *neta* existente entre dos variables, es decir, el grado de relación lineal existente entre dos variables tras eliminar de ambas el efecto atribuible a terceras variables. La lógica de una correlación parcial es similar a la ya estudiada a propósito del análisis de covarianza.

Por ejemplo, se sabe que la correlación entre las variables *inteligencia* y *rendimiento escolar* es alta y positiva. Sin embargo, cuando se controla el efecto de terceras variables como el *número de horas de estudio* o el *nivel educativo de los padres*, la correlación entre *inteligencia* y *rendimiento* desciende sensiblemente, lo cual indica que la relación entre *inteligencia* y *rendimiento* está condicionada, depende o está modulada por el *número de horas de estudio* y el *nivel educativo de los padres*.

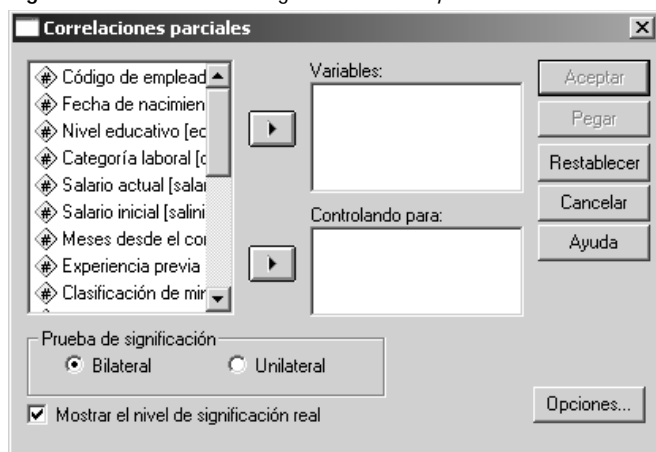
Para obtener coeficientes de correlación parcial:

- Seleccionar la opción **Correlaciones > Parciales...** del menú **Analizar** para acceder al cuadro de diálogo *Correlaciones parciales* que muestra la Figura 17.5.

La lista de variables muestra un listado de todas las variables del archivo de datos que poseen formato numérico. Para obtener un coeficiente de correlación parcial:

- Trasladar a la lista **Variables** las variables que se desea correlacionar.
- Trasladar a la lista **Controlando para** las variables cuyo efecto se desea controlar.

El procedimiento **Correlaciones parciales** puede manipular un total de 400 variables, de las cuales hasta un máximo de 100 pueden ser variables de control (es decir, variables trasladadas a la lista **Controlando para**).

Figura 17.5. Cuadro de diálogo *Correlaciones parciales*

La ecuación para obtener el coeficiente de correlación parcial depende del número de variables que se estén controlando:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \quad (\text{correlación parcial de primer orden})$$

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}} \quad (\text{correlación parcial de segundo orden})$$

Los coeficientes de mayor orden se obtienen siguiendo la misma lógica. Se habla de correlación de *primer orden* para indicar que se está controlando el efecto de una variable; de *segundo orden*, para indicar que se está controlando el efecto de dos variables; etc. Siguiendo esta lógica, cuando no se ejerce control sobre ninguna variable (es decir, cuando se utiliza el coeficiente de correlación de Pearson descrito en el apartado anterior) se habla de correlación de *orden cero*.

**Prueba de significación.** Junto con cada coeficiente de correlación parcial, el *Visor* ofrece la información necesaria para contrastar la hipótesis nula de que el valor poblacional del coeficiente de correlación parcial vale cero. Esta hipótesis se contrasta mediante un valor tipificado del coeficiente de correlación parcial que adopta la siguiente forma:

$$T = \frac{r_{12.p} \sqrt{m - p - 2}}{\sqrt{1 - r_{12.p}^2}}$$

donde  $m$  se refiere al número mínimo de casos con puntuación válida en el conjunto de posibles correlaciones de orden cero entre cada par de variables seleccionadas (es decir, el número de casos de la correlación de orden cero con menor número de casos válidos) y  $p$  es el número de variables controladas.

El estadístico  $T$  permite contrastar la hipótesis nula de que el valor poblacional del coeficiente de correlación parcial es cero. Este estadístico se distribuye según el modelo de probabilidad  $t$  de Student con  $m - p - 2$  grados de libertad. El SPSS permite seleccionar el tipo de nivel crítico deseado:

**Bilateral.** Opción apropiada para cuando no existen expectativas sobre la *dirección* de la relación. Indica la probabilidad de obtener coeficientes tan alejados de cero o más que el valor absoluto del coeficiente obtenido.

**Unilateral.** Opción apropiada para cuando existen expectativas sobre la *dirección* de la relación. Indica la probabilidad de obtener coeficientes tan grandes o más grandes que el obtenido si el coeficiente es positivo, o tan pequeños o más pequeños que el obtenido si el coeficiente de correlación es negativo.

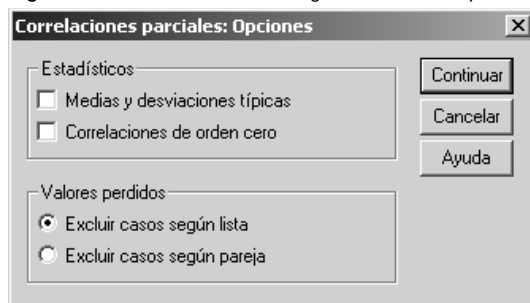
“ **Mostrar el nivel de significación real.** Esta opción, que se encuentra activa por defecto, permite obtener el nivel crítico exacto y los grados de libertad asociados a cada coeficiente de correlación parcial. Al desactivar esta opción, en lugar del nivel crítico exacto, el *Visor* muestra un asterisco al lado de los coeficientes cuyo nivel crítico es menor o igual que 0,05 y dos asteriscos al lado de los coeficientes cuyo nivel crítico es menor o igual que 0,01.

## Opciones

Para obtener alguna información adicional (algunos estadísticos descriptivos y los coeficientes de correlación de orden cero) y para controlar el tratamiento que se desea dar a los valores perdidos:

• Pulsar el botón **Opciones...** del cuadro de diálogo principal (ver Figura 17.5) para acceder al subcuadro de diálogo *Correlaciones parciales: Opciones* que muestra la Figura 17.6.

Figura 17.6. Subcuadro de diálogo *Correlaciones parciales: Opciones*



**Estadísticos.** Este recuadro contiene las siguientes opciones:

“ **Medias y desviaciones típicas.** Permite obtener la media aritmética, la desviación típica insesgada y el número de casos válidos de cada variable individualmente considerada.

- “ **Correlaciones de orden cero.** Permite obtener los coeficientes de correlación de orden cero entre cada par de variables; es decir, permite obtener los coeficiente de correlación de Pearson entre cada par de variables sin ejercer ningún control sobre terceras variables. Estas correlaciones son exactamente las mismas que se obtienen con el procedimiento **Correlaciones bivariadas**.

**Valores perdidos.** Las dos opciones de este recuadro permiten decidir qué tratamiento se desea dar a los valores perdidos.

**Excluir casos según pareja.** Se excluyen del cálculo de cada coeficiente de correlación los casos con valor perdido en alguna de las variables que están interviniendo en el coeficiente de correlación parcial.

**Excluir casos según lista.** Se excluyen del cálculo de todos los coeficientes de correlación solicitados los casos con valor perdido en cualquiera de las variables seleccionadas en la lista **Variables**.

### **Ejemplo: Correlaciones > Parciales**

Este ejemplo muestra cómo utilizar el procedimiento **Correlaciones parciales** para cuantificar e interpretar la relación entre dos variables cuando se controla el efecto de terceras variables. Se sigue utilizando el archivo *Datos de empleados*.

- En el cuadro de diálogo principal (ver Figura 17.5), seleccionar las variables *salini* (salario inicial) y *salario* (salario actual) y trasladarlas a la lista **Variables**. Estas son las dos variables que interesa correlacionar.
- Seleccionar las variables *educ* (nivel educativo), *tiempemp* (meses desde el contrato) y *expprev* (experiencia previa) y trasladarlas a la lista **Controlando para**. Estas son las tres variables cuyo efecto se desea controlar.
- Pulsar el botón **Opciones...** para acceder al cuadro de diálogo *Correlaciones parciales: Opciones* (ver Figura 17.6) y, en el recuadro **Estadísticos**, marcar las opciones **Medias** y **desviaciones típicas** y **Correlaciones de orden cero**.

Aceptando estas selecciones, el *Visor de resultados* ofrece la información que muestran las Tablas 17.4 y 17.5.

La Tabla 17.4 contiene la media aritmética, la desviación típica insesgada y el número de casos válidos; todo ello, para cada variable individualmente considerada.

**Tabla 17.4.** Estadísticos descriptivos

	Media	Desv. típica	N
Salario actual	34419.57	17075.661	474
Salario inicial	17016.09	7870.638	474
Nivel educativo	13.49	2.885	474
Meses desde el contrato	81.11	10.061	474
Experiencia previa (meses)	95.86	104.586	474

La Tabla 17.5 contiene las correlaciones bivariadas y las parciales. La mitad superior de la tabla (*Variables controladas = ninguna*) ofrece una matriz cuadrada con los coeficientes de correlación de orden cero (es decir, con los coeficientes de correlación bivariados sin parcializar efectos) entre todas las variables seleccionadas. La matriz ofrece, para cada par de variables, el coeficiente de correlación de Pearson (*Correlación*), los grados de libertad asociados al estadístico de contraste  $T$  ( $gl$  = número de casos válidos menos dos), y el nivel crítico bilateral asociado al estadístico  $T$  (*Sig. bilateral*). El nivel crítico unilateral se obtiene dividiendo el bilateral por dos.

La información de esta matriz es doblemente útil: por un lado, informa sobre el grado de relación existente entre las dos variables que interesa estudiar (en el ejemplo, *salario inicial* y *salario actual*); por otro, permite averiguar si las variables cuyo efecto se desea controlar (*nivel educativo*, *meses de contrato* y *experiencia previa*) están o no relacionadas con las dos variables que interesa correlacionar.

Así, puede verse que el coeficiente de correlación entre *salario inicial* y *salario actual* vale 0,88, con un nivel crítico *Sig.* < 0,0005 que permite rechazar la hipótesis nula de *no relación* y afirmar que el coeficiente es significativamente distinto de cero (o que el valor poblacional del coeficiente es distinto de cero). También se observa que, de las tres variables incluidas en el análisis para controlar su efecto, *nivel educativo* correlaciona significativamente tanto con *salario inicial* como con *salario actual* (*Sig.* < 0,0005 en ambos casos), *meses de contrato* no correlaciona ni con *salario inicial* (*Sig.* = 0,668) ni con *salario actual* (*Sig.* = 0,067), y *experiencia previa* correlaciona con *salario actual* (*Sig.* = 0,034) pero no con *salario inicial* (*Sig.* = 0,327).

**Tabla 17.5.** Correlaciones de orden cero (correlaciones bivariadas) y correlaciones parciales

Variables controladas			salario actual	salario inicial	nivel educativo	Meses de contrato	Experien. previa
-ninguno- <sup>a</sup>	Salario actual	Correlación	1,000	,880	,661	,084	-,097
		Sig. (bilateral)	.	,000	,000	,067	,034
		gl	0	472	472	472	472
	Salario inicial	Correlación	,880	1,000	,633	-,020	,045
		Sig. (bilateral)	,000	.	,000	,668	,327
		gl	472	0	472	472	472
	Nivel educativo	Correlación	,661	,633	1,000	,047	-,252
		Sig. (bilateral)	,000	,000	.	,303	,000
		gl	472	472	0	472	472
	Meses de contrato	Correlación	,084	-,020	,047	1,000	,003
		Sig. (bilateral)	,067	,668	,303	.	,948
		gl	472	472	472	0	472
	Experien. previa	Correlación	-,097	,045	-,252	,003	1,000
		Sig. (bilateral)	,034	,327	,000	,948	.
		gl	472	472	472	472	0
educ & tiempemp & expprev	Salario actual	Correlación	1,000	,812			
		Sig. (bilateral)	.	,000			
	Salario inicial	Correlación	,812	1,000			
		Sig. (bilateral)	,000	.			
		gl	469	0			

a. Las casillas contienen correlaciones de orden cero (de Pearson).



La mitad inferior de la tabla ofrece el coeficiente de correlación parcial entre las variables *salario inicial* y *salario actual*. El coeficiente de correlación parcial entre esas variables (es decir, el coeficiente de correlación obtenido tras eliminar de la relación entre esas dos variables el efecto atribuible a las variables *nivel educativo*, *meses de contrato* y *experiencia previa*) vale 0,812, con un nivel crítico *Sig.* < 0,0005. Puesto que el valor del nivel crítico es menor que 0,05, puede afirmarse que el valor poblacional del coeficiente de correlación parcial entre *salario inicial* y *salario actual* es distinto de cero.

Puesto que el coeficiente de correlación parcial sigue siendo significativo y su diferencia con el coeficiente de orden cero es más bien escasa (ha bajado de 0,88 a 0,81), puede afirmarse: (1) que entre las variables *salario inicial* y *salario actual* existe relación lineal significativa, y (2) que tal relación no se ve sustancialmente alterada tras controlar el efecto de las variables *nivel educativo*, *meses de contrato* y *experiencia previa*.

## Distancias

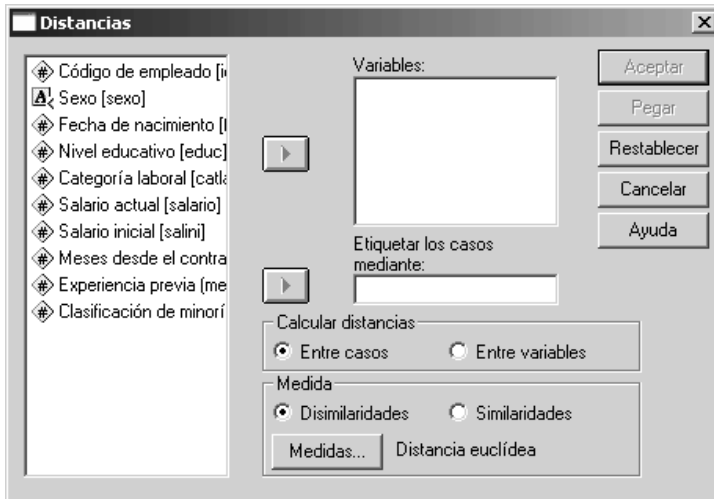
Los coeficientes de correlación estudiados en los apartados anteriores representan una forma particular de cuantificar la *distancia* existente entre dos variables, pero no la única. Existen otras muchas formas de cuantificar esa distancia. El procedimiento **Distancias** incluye un gran número de medidas que se diferencian, básicamente, por el tipo de datos para el que han sido diseñadas: cuantitativos, categóricos o dicotómicos. Estas medidas pueden utilizarse tanto para obtener la distancia entre dos *variables* como para obtener la distancia entre dos *casos*. A partir de ahora se utilizará el término *elemento* para hablar indistintamente de casos y variables.

Las medidas del procedimiento **Distancias** también se diferencian por la forma de abordar el estudio de la distancia entre dos elementos: *similaridad* o *disimilaridad*. Las medidas de **similaridad** evalúan el grado de parecido o proximidad existente entre dos elementos. Los valores más altos en la medida indican mayor parecido o proximidad: cuando dos elementos se encuentran juntos (se parecen mucho), el valor de las medidas de similaridad es máximo. El coeficiente de correlación de Pearson es, quizá, la medida de similaridad más ampliamente conocida y utilizada.

Las medidas de **disimilaridad** ponen el énfasis sobre el grado de diferencia o lejanía existente entre dos elementos. Los valores más altos en la medida indican mayor diferencia o lejanía: cuando dos elementos se encuentran juntos, el valor de las medidas de disimilaridad es nulo. El estadístico *chi*-cuadrado de Pearson y la distancia euclídea (la longitud del segmento lineal que une dos elementos) son, quizá, las medidas de disimilaridad más conocidas y utilizadas. Conviene señalar que son las medidas de disimilaridad las que han terminado acaparrando la acepción de *medidas de distancia*.

Además de una gran cantidad de medidas de distancia, el procedimiento **Distancias** incluye otras prestaciones de especial interés: ofrece la posibilidad de utilizar diferentes métodos para estandarizar los valores originales antes de calcular las distancias, y permite crear y guardar una matriz de distancias que más tarde puede usarse con otros procedimientos SPSS como el análisis factorial, el escalamiento o el análisis de conglomerados. Para obtener *medidas de distancia*:

- Seleccionar la opción **Correlaciones > Distancias...** del menú **Analizar** para acceder al cuadro de diálogo *Distancias* que muestra la Figura 17.7.

Figura 17.7. Cuadro de diálogo *Distancias*

La lista de variables del archivo de datos ofrece un listado de todas las variables numéricas y de cadena del archivo (las variables de cadena sólo pueden utilizarse para identificar casos). Para obtener la distancia entre dos o más elementos:

- Seleccionar las variables cuya distancia se desea evaluar (o las variables en las que debe basarse la distancia entre casos) y trasladarlas a la lista **Variables**.

**Etiquetar los casos mediante.** En los resultados, los casos individuales se identifican por el número de registro (fila) que ocupan en el *Editor de datos*. Si se desea utilizar otro criterio, esta opción permite seleccionar una *variable de cadena* para identificar los casos (si la cadena tiene más de ocho caracteres, sólo se utilizan los ocho primeros).

**Calcular distancias.** El procedimiento permite calcular distancias entre casos y entre variables. En ambos casos las distancias se calculan a partir de las puntuaciones de los casos en el conjunto de variables seleccionadas.

**Medida.** Las medidas de distancia están agrupadas en dos bloques: similaridad y disimilaridad. El botón **Medidas...** (ver más abajo) conduce a un subcuadro de diálogo que permite elegir la medida de distancia que se desea utilizar. Este subcuadro de diálogo tiene dos versiones que se diferencian por el tipo de medidas que ofrecen. La versión a la que se accede mediante el botón **Medidas...** depende de la opción marcada en este recuadro:

**Disimilaridades.** Medidas de *diferencia* o *lejanía*. Los valores más altos indican que los elementos son muy distintos o que se encuentran muy alejados.

**Similaridades.** Medidas de *parecido* o *cercanía*. Los valores más altos indican que los elementos son muy parecidos o que se encuentran muy próximos.

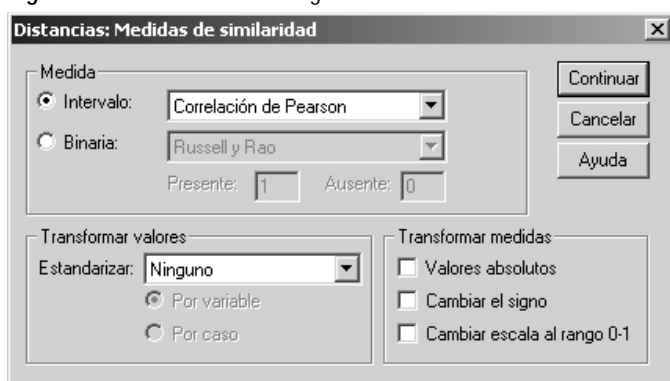
El botón **Medidas...** conduce a un subcuadro de diálogo que permite seleccionar la medida de distancia que se desea utilizar. Este subcuadro de diálogo varía dependiendo de que en el recuadro **Medida** se haya elegido **Disimilaridades** o **Similaridades**.

## Medidas de similaridad

Para obtener una medida de similaridad:

- En el cuadro de diálogo principal (ver Figura 17.7), seleccionar la opción **Similaridad** del recuadro **Medida**.
- Pulsar el botón **Medidas...** para acceder al subcuadro de diálogo *Distancias: Medidas de similaridad* que muestra la Figura 17.8.

Figura 17.8. Subcuadro de diálogo *Distancias: Medidas de similaridad*



### Intervalo

Las medidas de similaridad disponibles están agrupadas en dos bloques: *intervalo* y *binaria*. Las medidas de intervalo son útiles para estudiar la similaridad entre variables cuantitativas medidas con una escala de intervalo o razón:

- **Correlación de Pearson.** Coeficiente de correlación de Pearson entre dos vectores de datos cuantitativos. Toma valores entre  $-1$  y  $1$ :

$$CORRELATION(X, Y) = \frac{\sum_i^n z_{x_i} z_{y_i}}{n - 1}$$

- **Coseno.** Medida estrechamente relacionada con el coeficiente de correlación de Pearson. Es el coseno del ángulo formado por dos vectores de puntuaciones. Toma valores entre  $-1$  y  $1$ :

$$COSINE(X, Y) = \frac{\sum_i^n X_i Y_i}{\sqrt{\left(\sum_i^n X_i^2\right) \left(\sum_i^n Y_i^2\right)}}$$

## Binaria

Las medidas para datos *binarios* se utilizan para medir distancias entre variables dicotómicas, es decir, entre variables cuyos valores únicamente reflejan la presencia o ausencia de la característica medida: «tratados-no tratados», «recuperados-no recuperados», «a favor-en contra», «acierto-error», etc.

Generalmente, la *presencia* de la característica se codifica con el valor 1 y la *ausencia* con el valor 0 (ver Tabla 17.6). No obstante, el SPSS admite cualquier tipo de codificación numérica. Las opciones **Presente** y **Ausente** (ver Figura 17.8) permiten indicar los códigos con los que se han registrado en el *Editor de datos* las presencias y las ausencias de las características medidas. Una vez que se han introducido los códigos que corresponden a la presencia de la característica (opción **Presente**) y a la ausencia de la característica (opción **Ausente**), el SPSS únicamente tiene en cuenta esos valores, ignorando el resto, si existen.

La Tabla 17.6 muestra una tabla de contingencias 2×2 con la notación habitualmente utilizada al resumir los datos referidos a dos variables dicotómicas.

**Tabla 17.6.** Tabla de contingencias con dos variables dicotómicas

		<i>Variable <math>Y_i</math></i>		
		1	0	
<i>Variable <math>X_i</math></i>	1	<i>a</i>	<i>b</i>	<i>a + b</i>
	0	<i>c</i>	<i>d</i>	<i>c + d</i>
		<i>a + c</i>	<i>b + d</i>	<i>n</i>

En la tabla, *n* se refiere al número total de casos, *a* se refiere al número de casos que comparten la presencia de ambas características; *d* se refiere al número de casos que comparten la ausencia de ambas características (*a* y *d* son las *concordancias*); *b* y *c* se refieren el número de casos que presentan una característica y no la otra (las *discordancias*).

Existe un gran número de medidas para calcular la distancia entre los elementos de una tabla de contingencias de estas características. Estas medidas difieren, básicamente, en la importancia que conceden a cada casilla de la tabla. Se considera que dos elementos son tanto más similares entre sí cuanto mayor número de presencias o ausencias comparten. Pero las presencias y las ausencias no tienen por qué tener la misma importancia al valorar la similitud. Si dos sujetos responden sí a la pregunta «¿ha padecido alguna enfermedad grave en los últimos tres meses?», esa concordancia posee mucho mayor valor informativo que si ambos sujetos responden no. Sin embargo, si dos sujetos responden sí a la pregunta «¿ha ido alguna vez a la playa en verano?», esa concordancia posee mucho menor valor informativo que si ambos sujetos responden no.

Por esta razón, algunas medidas no tienen en cuenta las ausencias conjuntas; otras conceden más importancia a las concordancias que a las discordancias, o al revés; otras sólo tienen en cuenta las presencias conjuntas; otras, las ausencias; etc. Puesto que cada una de ellas pone el énfasis en un aspecto concreto de las frecuencias de la tabla, la decisión sobre qué medida conviene utilizar no es una cuestión trivial. Sobre todo si se tiene en cuenta que muchas de ellas ofrecen resultados que no son equivalentes (no son medidas monótonas entre sí, pudien-

do darse inversiones de valores en los elementos comparados) y que el cambio de codificación de las presencias-ausencias (el cambio de ceros por unos y de unos por ceros) también puede hacer variar el resultado.

Las fórmulas que se ofrecen a continuación están diseñadas para evaluar la distancia entre *dos variables a partir de un cierto número de casos*. No obstante, intercambiando en la Tabla 17.6 las variables  $X_i$  e  $Y_i$  por los casos  $i$  e  $i'$ , todas las fórmulas que se ofrecen pueden utilizarse para calcular la distancia entre *dos casos a partir de un cierto número de variables*.

- **Russel y Rao.** Es el producto escalar binario:

$$RR(X, Y) = \frac{a}{n}$$

- **Concordancia simple** o *emparejamiento simple*. Es el cociente entre el número de concordancias y el número total de características:

$$SM(X, Y) = \frac{a+d}{n}$$

- **Jaccard.** Medida conocida también como *tasa de similitud*. No tiene en cuenta las ausencias conjuntas ( $d$ ) y pondera por igual las concordancias y las discordancias:

$$JACCARD(X, Y) = \frac{a}{a+b+c}$$

- **Dice.** También conocida como medida de Czekanowski o de Sorenson. No tiene en cuenta las ausencias conjuntas, pero concede valor doble a las presencias conjuntas:

$$DICE(X, Y) = \frac{2a}{2a+b+c}$$

- **Rogers y Tanimoto.** Incluye las ausencias conjuntas tanto en el numerador como en el denominador y concede doble valor a las discordancias:

$$RT(X, Y) = \frac{a+d}{a+d+2(b+c)}$$

- **Sokal y Sneath 1.** Incluye las ausencias conjuntas tanto en el numerador como en el denominador y concede doble valor a las concordancias:

$$SSI(X, Y) = \frac{2(a+d)}{2(a+d)+b+c}$$

- **Sokal y Sneath 2.** Excluye las ausencias conjuntas y concede doble valor a las discordancias:

$$SS2(X, Y) = \frac{a}{a+2(b+c)}$$

- **Sokal y Sneath 3.** Excluye las concordancias del denominador. Esta medida tiene un límite inferior de 0, pero no tiene límite superior. Y no es posible calcularla si no existen discordancias (es decir, si  $b=c=0$ ). En ese caso, el programa asigna un valor arbitrario de 9999,999 como límite superior tanto si no hay discordancias como si el valor de la medida excede de ese valor:

$$SS3(X, Y) = \frac{a + d}{b + c}$$

- **Kulczynski 1.** Excluye las ausencias conjuntas del numerador y las concordancias del denominador. Tiene un límite inferior de 0, pero no tiene límite superior. Y no es posible calcularla si no existen discordancias (es decir, si  $b=c=0$ ). En ese caso, el procedimiento asigna un valor arbitrario de 9999,999 como límite superior tanto si no hay discordancias como si el valor de la medida excede de ese valor:

$$K1(X, Y) = \frac{a}{b + c}$$

- **Kulczynski 2.** Probabilidad condicional de que la característica medida esté presente en una variable dado que lo está en la otra. La medida final es el promedio de las dos medidas posibles:  $P(X|Y)$  y  $P(Y|X)$ . Toma valores entre 0 y 1:

$$K2(X, Y) = \frac{a/(a + b) + a/(a + c)}{2}$$

- **Sokal y Sneath 4.** Probabilidad condicional de que la característica medida se encuentre en el mismo estado (presente o ausente) en las dos variables. La medida final es el promedio de las dos medidas posibles:  $P(X|Y)$  y  $P(Y|X)$ . Toma valores entre 0 y 1:

$$SS4(X, Y) = \frac{a/(a + b) + a/(a + c) + d/(b + d) + d/(c + d)}{4}$$

- **Hamann.** Probabilidad de que la característica medida se encuentre en el mismo estado en las dos variables (presente o ausente en ambas), menos la probabilidad de que la característica se encuentre en distinto estado en ambas variables (presente en una y ausente en otra). Toma valores entre -1 y 1:

$$HAMANN(X, Y) = \frac{(a + d) - (b + c)}{a + b + c + d}$$

- **Lambda de Goodman y Kruskal.** Evalúa el grado en que el estado (presente o ausente) de una característica en una variable puede predecirse a partir del estado de esa característica en la otra variable. En concreto, *lambda* mide la reducción proporcional del error de predicción que se consigue al utilizar una variable como predictora de la otra cuando las direcciones de la predicción son de igual importancia. *Lambda* toma valores entre 0 y 1:

$$LAMBDA(X, Y) = \frac{t_1 - t_2}{2(a + b + c + d)}$$

donde:  $t_1 = \text{máx}(a, b) + \text{máx}(c, d) + \text{máx}(a, c) + \text{máx}(b, d)$   
 $t_2 = \text{máx}(a+c, b+d) + \text{máx}(a+b, c+d)$

- **D de Andemberg.** Al igual que *lambda*, evalúa la capacidad predictiva de una variable sobre otra. Y, al igual que *lambda*, mide la reducción en la probabilidad del error de predicción cuando una de las variables es utilizada para predecir la otra. Toma valores entre 0 y 1:

$$D(X, Y) = \frac{t_1 + t_2}{2(a + b + c + d)}$$

donde  $t_1$  y  $t_2$  se definen de la misma manera que en la medida *lambda* de Goodman y Kruskal.

- **Y de Yule.** El coeficiente de coligación *Y* de Yule es una función de los productos cruzados en una tabla 2×2. Toma valores entre -1 y 1:

$$Y(X, Y) = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

- **Q de Yule.** Versión para tablas 2×2 de la medida ordinal *gamma* de Goodman y Kruskal. También es una función de los productos cruzados. Toma valores entre -1 y 1:

$$Q(X, Y) = \frac{ad - bc}{ad + bc}$$

- **Ochiai.** Medida de similaridad. Versión binaria del coseno. Toma valores entre 0 y 1:

$$OCHIAI(X, Y) = \sqrt{\left(\frac{a}{a+b}\right)\left(\frac{a}{a+c}\right)}$$

- **Sokal y Sneath 5.** Medida de similaridad. Toma valores entre 0 y 1:

$$SS5(X, Y) = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

- **Correlación Phi (de cuatro puntos).** Versión binaria de coeficiente de correlación de Pearson. Es la medida de asociación más utilizada para datos binarios. Toma valores entre 0 y 1:

$$PHI(X, Y) = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

- **Dispersión.** Medida de similaridad. Toma valores entre 0 y 1:

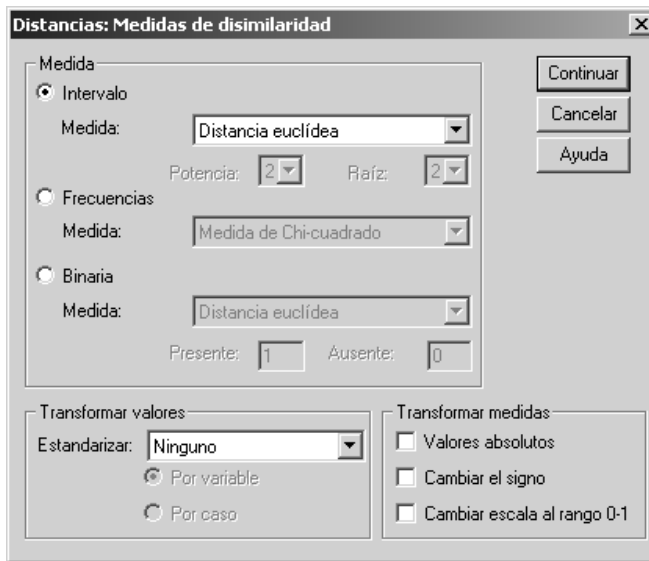
$$DISPER(X, Y) = \frac{ad - bc}{(a + b + c + d)^2}$$

## Medidas de disimilaridad

Para obtener una medida de disimilaridad:

- En el cuadro de diálogo principal (ver Figura 17.7), seleccionar la opción **Disimilaridad** del recuadro **Medida**.
- Pulsar el botón **Medidas...** para acceder al subcuadro de diálogo *Distancias: Medidas de disimilaridad* que muestra la Figura 17.9.

Figura 17.9. Subcuadro de diálogo *Distancias: Medidas de disimilaridad*



### Intervalo

Las medidas de disimilaridad disponibles están agrupadas en tres bloques: *intervalo*, *frecuencias* y *binaria*. Las medidas de intervalo son útiles para estudiar la similaridad entre variables cuantitativas obtenidas con una escala de intervalo o razón:

- Distancia euclídea.** Medida de disimilaridad utilizada por defecto para datos de intervalo. Raíz cuadrada de la suma de los cuadrados de las diferencias entre los valores de las variables:

$$EUCLID(X, Y) = \sqrt{\sum_i (X_i - Y_i)^2}$$

- Distancia euclídea al cuadrado.** Suma de los cuadrados de las diferencias entre los valores de las variables:

$$SEUCLID(X, Y) = \sum_i (X_i - Y_i)^2$$



- **Chebychev.** Diferencia más grande en valor absoluto entre los valores de dos variables:

$$CHEBYCHEV(X, Y) = \max_i |X_i - Y_i|$$

- **Bloques.** También llamada distancia *absoluta*, distancia de *ciudad*, de *Manhatan*, y del *taxista*. Es la suma de los valores absolutos de las diferencias entre los valores de dos variables:

$$BLOCK(X, Y) = \sum_i |X_i - Y_i|$$

- **Minkowsky.** Medida de disimilaridad basada en la distancia euclídea. Raíz de orden  $p$  de la suma de las potencias de orden  $p$  de los valores absolutos de las diferencias entre los valores de dos variables:

$$MINKOWSKI(X, Y) = \left( \sum_i |X_i - Y_i|^p \right)^{\frac{1}{p}}$$

donde  $p$  es cualquier número entero positivo.

- **Personalizada.** Medida de disimilaridad basada en la distancia euclídea. Raíz de orden  $r$  de la suma de las potencias de orden  $p$  de los valores absolutos de las diferencias entre los valores de dos variables:

$$POWER(X, Y) = \left( \sum_i |X_i - Y_i|^p \right)^{\frac{1}{r}}$$

donde  $p$  y  $r$  son dos números enteros positivos cualesquiera.

## Frecuencias

Esta opción incluye dos medidas de disimilaridad para datos categóricos. Ambas se basan en el estadístico *chi-cuadrado* de Pearson para el estudio de tablas de contingencias bidimensionales (ver Capítulo 12).

- **Chi-cuadrado.** Medida de disimilaridad utilizada por defecto para datos categóricos. Se basa en las divergencias existentes entre las frecuencias observadas y las esperadas bajo la hipótesis de independencia. Su magnitud depende del tamaño muestral:

$$CHISQ(X, Y) = \sqrt{\sum_i [X_i - E(X_i)]^2 / E(X_i) + \sum_i [Y_i - E(Y_i)]^2 / E(Y_i)}$$

- **Phi-cuadrado.** La medida *chi-cuadrado* tipificada por la raíz cuadrada del número de casos. Su valor no depende del tamaño muestral:

$$PH2(X, Y) = CHISQ(X, Y) / \sqrt{n}$$

**Binaria**

- **Distancia euclídea**. Versión binaria de la distancia euclídea. Su valor mínimo es 0, pero no tiene máximo:

$$BEUCLID(X, Y) = \sqrt{b + c}$$

- **Distancia euclídea al cuadrado**. Su valor mínimo es 0, pero no tiene máximo:

$$BSEUCLID(X, Y) = b + c$$

- **Diferencia de tamaño**. Su valor mínimo es 0, pero no tiene máximo:

$$SIZE(X, Y) = \frac{(b - c)^2}{(a + b + c + d)^2}$$

- **Diferencia de configuración**. Toma valores entre 0 y 1:

$$PATTERN(X, Y) = \frac{bc}{(a + b + c + d)^2}$$

- **Varianza**. Su valor mínimo es 0, pero no tiene máximo:

$$VARIANCE(X, Y) = \frac{b + c}{4(a + b + c + d)^2}$$

- **Forma**. No tiene límite inferior ni superior:

$$BSHAPE(X, Y) = \frac{(a + b + c + d)(b + c) - (b - c)^2}{(a + b + c + d)^2}$$

- **Lance y Williams**. También se conoce como el *coeficiente no métrico de Bray-Curtis*. Toma valores entre 0 y 1:

$$BLWMN(X, Y) = \frac{b + c}{2a + b + c}$$

Todas las variables seleccionadas deben compartir el mismo nivel de medida: no tiene sentido mezclar variables cuantitativas con variables categóricas. Y conviene recordar que el procedimiento no permite trabajar con variables de cadena.

Para una descripción más detallada de estas medidas de distancia puede consultarse Anderberg (1973) o Romesburg (1984).

## Transformación de los valores

Muchas de las medidas de distancia (por ejemplo, la distancia euclídea y el resto de medidas derivadas de ella) no son invariantes respecto a la métrica de los datos, ya que las diferencias existentes entre las variables con puntuaciones muy altas pueden anular las diferencias existentes entre las variables con puntuaciones muy bajas. Por ejemplo, en el archivo *Datos de empleados* utilizado en los capítulos previos, la variable *salario* (salario actual) toma valores comprendidos entre 15.150,00 y 135.000,00; por el contrario, la variable *tiempemp* (meses desde el contrato) toma valores comprendidos entre 63 y 98. Lógicamente, al obtener una medida de la distancia entre dos empleados, la diferencia en los salarios tendrá mucha más presencia en la medida utilizada que la diferencia en los meses de contrato.

Para resolver este problema suele recomendarse no analizar las puntuaciones directas de las variables (los datos en bruto) sino las puntuaciones transformadas a escalas del mismo rango (escala 0–1, escala típica, etc.). Las opciones del recuadro **Transformar valores** (ver Figuras 17.8 y 17.9) permiten elegir entre distintos tipos de transformación y decidir si la transformación se desea aplicar tomando como referencia los casos o las variables. La transformación elegida se aplica a todos los elementos del análisis. Estas opciones no están disponibles cuando se selecciona una medida de distancia binaria. En todos los casos es posible seleccionar los *elementos* (casos o variables) que se desea transformar. Las opciones de transformación son las siguientes:

- **Ninguno.** No se aplica ningún método de transformación.
- **Puntuaciones Z.** A cada valor se le resta la media del elemento y esa diferencia se divide por la desviación típica del elemento. Se obtienen de este modo valores estandarizados con media 0 y desviación típica 1. Si la desviación típica vale 0, se asigna un 0 a todos los valores.
- **Rango -1 a 1.** Cada valor se divide por el rango o amplitud del elemento. Se obtienen de este modo valores estandarizados con amplitud 1 en una escala cuya unidad de medida es el rango o amplitud del elemento. Si el rango o amplitud vale cero, no se efectúa la transformación.
- **Rango 0 a 1.** A cada valor se le resta el valor más pequeño del elemento y esa diferencia se divide entre el rango o amplitud del elemento. Se obtienen de esta manera valores estandarizados comprendidos entre 0 y 1. Si el rango vale 0, se asigna un 0,5 a todos los valores.
- **Magnitud máxima de 1.** Cada valor se divide por el valor más grande del elemento. Se obtienen de este modo valores estandarizados con un máximo de 1 y un mínimo variable. Si el valor más grande vale 0, se divide por el valor absoluto del valor más pequeño y se suma 1.
- **Media 1.** Divide cada valor por la media del elemento. Se obtienen de este modo valores estandarizados con media igual a 1, y en una escala cuya unidad de medida es la media del elemento. Si la media vale 0, se suma un 1 a todos los valores.
- **Desviación típica 1.** Divide cada valor por la desviación típica del elemento. Se obtienen de este modo valores estandarizados con desviación típica igual a 1 y en una escala cuya unidad de medida es la desviación típica del elemento. Si la desviación típica vale 0, no se efectúa la transformación.

## Transformación de las medidas

Las opciones del recuadro **Transformar medidas** (ver Figuras 17.8 y 17.9) permiten transformar los valores de la matriz de distancias. Si se selecciona más de una transformación, el procedimiento las realiza en el siguiente orden:

- " **Valores absolutos.** Obtiene el valor absoluto de las distancias calculadas.
- " **Cambiar el signo.** Cambia el signo de las distancias calculadas, transformando así las medidas de similaridad en medidas de disimilaridad y viceversa.
- " **Cambiar escala al rango 0–1.** Se resta a todos los valores de la matriz de distancias la distancia más pequeña y cada nueva distancia se divide por el rango o amplitud de todas las distancias. Se obtienen así valores que oscilan entre 0 y 1.



## Análisis de regresión lineal

### El procedimiento *Regresión lineal*

El análisis de regresión lineal es una técnica estadística utilizada para estudiar la relación entre variables cuantitativas. Se adapta a una amplia variedad de situaciones. En la investigación social, puede utilizarse para predecir un amplio rango de fenómenos, desde medidas económicas hasta diferentes aspectos del comportamiento humano. En el contexto de la investigación de mercados puede utilizarse para determinar en cuál de diferentes medios de comunicación puede resultar más eficaz invertir; o para predecir cuál será el número de ventas de un determinado producto. En áreas como la física puede utilizarse para caracterizar la relación entre variables o para calibrar medidas. Etc.

Tanto en el caso de dos variables (regresión *simple*) como en el de más de dos variables (regresión *múltiple*), el análisis de regresión lineal puede utilizarse para explorar y cuantificar la relación entre una variable llamada dependiente o criterio ( $Y$ ) y una o más variables llamadas independientes o predictoras ( $X_1, X_2, \dots, X_p$ ), así como para desarrollar una ecuación lineal con fines predictivos. Además, el análisis de regresión lleva asociada una serie de estrategias de diagnóstico (análisis de los residuos, puntos de influencia) que informan sobre la estabilidad e idoneidad del análisis y que proporcionan pistas sobre cómo perfeccionarlo.

El objetivo de este capítulo es el de proporcionar los fundamentos del análisis de regresión. Al igual que en los capítulos precedentes, no se hará hincapié en los aspectos más técnicos del análisis, sino que se intentará fomentar la comprensión de cuándo y cómo utilizar la técnica y cómo interpretar los resultados. También se prestará atención a cuestiones como el chequeo de los supuestos del análisis de regresión y la forma de proceder cuando se incumplen. Para profundizar en estos y otros aspectos del análisis de regresión lineal, puede consultarse Montgomery, Peck y Vining (2001).

### La recta de regresión

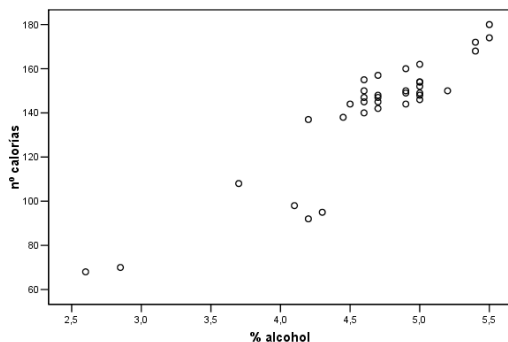
En el capítulo anterior sobre *Correlación lineal* se ha tenido ocasión de comprobar que un diagrama de dispersión ofrece una idea bastante aproximada sobre el *tipo de relación* existente entre dos variables cuantitativas. Pero, además, un diagrama de dispersión también puede utilizarse como una forma de *cuantificar* el grado de relación lineal existente entre dos variables: basta con observar el grado en el que la nube de puntos se ajusta a una línea recta.

Ahora bien, aunque un diagrama de dispersión permite formarse una primera impresión muy rápida sobre el tipo de relación existente entre dos variables, utilizarlo como una forma

de *cuantificar* esa relación tiene un serio inconveniente: la relación entre dos variables no siempre es perfecta o nula; de hecho, habitualmente no es ni lo uno ni lo otro.

Consideremos, como ejemplo, un pequeño conjunto de datos con información sobre 35 marcas de cerveza. Se desea estudiar si existe relación entre el grado de alcohol de las cervezas y su contenido calórico. Un buen punto de partida para obtener una primera impresión sobre esta relación podría ser representar su nube de puntos, tal como muestra el diagrama de dispersión de la Figura 18.1.

**Figura 18.1.** Diagrama de dispersión: *porcentaje de alcohol por n° de calorías*



El eje vertical muestra el número de calorías (por cada tercio de litro) y el horizontal el contenido de alcohol (expresado en porcentaje). A simple vista, parece existir una relación positiva entre ambas variables: conforme aumenta el porcentaje de alcohol, también aumenta el número de calorías. En esta muestra no hay cervezas que teniendo alto contenido de alcohol tengan pocas calorías y tampoco hay cervezas que teniendo muchas calorías tengan poco alcohol. La mayor parte de las cervezas de la muestra se agrupan entre el 4,5% y el 5% de alcohol, siendo relativamente pocas las cervezas que tienen un contenido de alcohol inferior a ése. Se podría haber extendido el rango de la muestra incluyendo cervezas sin alcohol, pero el rango de calorías y alcohol considerados parece bastante apropiado: no hay, por ejemplo, cervezas con un contenido de alcohol del 50%, o cervezas sin calorías.

¿Cómo describir apropiadamente estos datos? Quizá bastaría con decir, simplemente, que un aumento en el porcentaje de alcohol va acompañado de un aumento en el número de calorías; pero esto, aunque correcto, es poco específico. ¿Qué hacer para obtener una descripción más concreta? Sin duda, ofrecer un listado de los datos concretos de que se dispone; pero esto, aunque preciso, no resulta demasiado informativo.

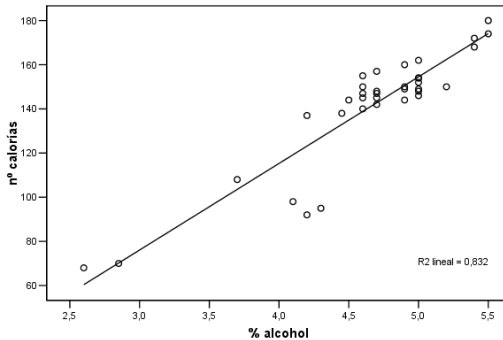
Es posible hacer algo mucho más útil como, por ejemplo, describir la pauta observada en la nube de puntos mediante una función matemática simple, tal como una línea recta. A primera vista, una línea recta podría ser un buen punto de partida para resumir una nube de puntos como la de la Figura 18.1. Puesto que una línea recta posee una fórmula muy simple,

$$Y'_i = B_0 + B_1 X_i,$$

puede comenzarse obteniendo los coeficientes  $B_0$  y  $B_1$  que definen la recta. El coeficiente  $B_1$  es la pendiente de la recta: el cambio medio que se pronostica en el número de calorías ( $Y'_i$ ) por cada unidad de cambio que se produce en el porcentaje de alcohol ( $X_i$ ). El coeficiente  $B_0$  es el punto en el que la recta corta el eje vertical: el número medio de calorías que correspon-

de a una cerveza con porcentaje de alcohol cero. Conociendo los valores de estos dos coeficientes, ya es posible reproducir la recta y describir con ella la relación existente entre el contenido de alcohol y el número de calorías. Aunque todavía no se entre en los detalles de cómo obtener los valores de  $B_0$  y  $B_1$ , nada impide poder visualizar esa recta (ver Figura 18.2).

**Figura 18.2.** Diagrama de dispersión y recta de regresión: % de alcohol por n° de calorías



$$Y_i = -3377 + 37,65 X_i$$

$$\text{n° de calorías} = -33,77 + 37,65 (\% \text{ de alcohol})$$

Puede observarse que, en general, la recta hace un seguimiento bastante bueno de los datos. La fórmula de la recta aparece a la derecha del diagrama. La pendiente de la recta ( $B_1$ ) indica que, en promedio, a cada incremento de una unidad en el porcentaje de alcohol ( $X_i$ ) se le pronostica un incremento de 37,65 calorías ( $Y_i$ ). El origen de la recta ( $B_0$ ) sugiere que una cerveza sin alcohol (grado de alcohol cero) podría contener -33,77 calorías. Y esto, obviamente, no parece posible. Al examinar la nube de puntos se ve que la muestra no contiene cervezas con menos de un 2% de alcohol. Así, aunque el origen de la recta aporta información sobre lo que podría ocurrir si se extrapola hacia abajo la pauta observada en los datos hasta llegar a una cerveza con grado de alcohol cero, al hacer esto se estarían efectuando pronósticos en un rango de valores que va más allá de lo que abarcan los datos disponibles, y eso es algo extremadamente arriesgado en el contexto del análisis de regresión\*.

## La mejor recta de regresión

En una situación ideal (e irreal) en la que todos los puntos de un diagrama de dispersión se encontraran en una línea recta, no habría que preocuparse por encontrar la recta que mejor resume los puntos del diagrama: simplemente uniendo los puntos entre sí se obtendría la recta con el mejor ajuste posible. Pero en una nube de puntos más realista (como la de las Figuras 18.1 y 18.2) es posible trazar muchas rectas diferentes. Y, obviamente, no todas ellas se ajustarán igualmente bien a la nube de puntos. De lo que se trata es de encontrar la recta capaz de convertirse en el mejor representante del conjunto total de puntos.

Existen diferentes procedimientos para ajustar una función simple, cada uno de los cuales intenta minimizar una medida diferente del grado de ajuste (ver Rousseeuw y Leroy, 1987).

\* Debe aprenderse una lección de esto: la primera cosa razonable que podría hacerse es añadir al estudio alguna cerveza con porcentaje de alcohol cero; probablemente así se obtendría una recta con un origen más realista.



La elección tradicionalmente preferida ha sido la recta que hace *mínima la suma de los cuadrados de las distancias verticales entre cada punto y la recta*. Esto significa que, de todas las rectas posibles, existe una y sólo una que consigue que las distancias verticales entre cada punto y la recta sean mínimas (las distancias se elevan al cuadrado porque, de lo contrario, al ser unas positivas y otras negativas, se anularían unas con otras al sumarlas).

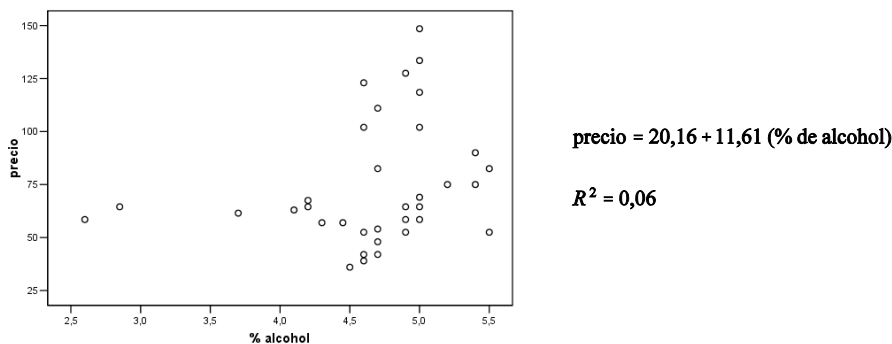
## Bondad de ajuste

Además de acompañar la recta con su fórmula, podría resultar útil disponer de alguna indicación precisa del grado en el que la recta se ajusta a la nube de puntos. De hecho, la mejor recta posible no tiene por qué ser buena.

En una situación como la representada en la Figura 18.3, la recta consigue un ajuste bastante más pobre que en el caso de la Figura 18.2. Ahora el diagrama de dispersión representa el porcentaje de alcohol de las cervezas (eje horizontal) y el precio de las mismas (eje vertical). Y no parece existir la misma pauta de asociación detectada entre las variables de la situación anterior. No parece que en esta nueva nube de puntos la disposición de los puntos invite a elegir una recta como resumen o representante de lo que está ocurriendo.

Así pues, aunque siempre resulta posible, cualquiera que sea la nube de puntos, obtener la recta mínimo-cuadrática, no siempre esa recta es un buen resumen o representante de los puntos de la nube. Además de la recta, es necesario obtener alguna información adicional que permita determinar el grado de fidelidad con que esa recta describe la pauta de relación existente en los datos, es decir, el grado de ajuste de la recta a la nube de puntos.

Figura 18.3. Diagrama de dispersión, recta de regresión y ajuste: % de alcohol por precio



¿Cómo puede cuantificarse ese *mejor o peor* ajuste de la recta? Hay diferentes formas de responder a esta pregunta. Podría utilizarse la media de las distancias existentes entre todos los puntos y la recta (residuos), o la media de los residuos en valor absoluto, o las medianas de alguna de esas medidas, o alguna función ponderada de esas medidas, etc.

Una medida de ajuste que ha recibido gran aceptación en el contexto del análisis de regresión es el **coeficiente de determinación  $R^2$** : el cuadrado del coeficiente de correlación múltiple. Se trata de una medida estandarizada que toma valores entre 0 y 1 (0 cuando las variables son independientes y 1 cuando entre ellas existe relación perfecta).

Este coeficiente posee una interpretación muy intuitiva: representa el grado de ganancia que se obtiene al predecir una variable a partir del conocimiento que se tiene de otra u otras variables. Si se quiere, por ejemplo, pronosticar el número de calorías de una cerveza sin el conocimiento de otras variables, se utilizaría la media del número de calorías. Pero si se dispone de información sobre otra variable y del grado de relación entre ambas, es posible mejorar ese pronóstico. El valor  $R^2$  del diagrama de la Figura 18.2 vale 0,83, lo que indica que si se conoce el porcentaje de alcohol de una cerveza, los pronósticos sobre su contenido calórico pueden mejorarse en un 83 % si, en lugar de utilizar como pronóstico el número medio de calorías, el pronóstico se basa el porcentaje de alcohol.

Comparando este resultado con el del diagrama de la Figura 18.3 (donde  $R^2$  vale 0,06) se comprenderá fácilmente el valor informativo de  $R^2$ : en este segundo caso, el conocimiento del contenido de alcohol de una cerveza sólo permite mejorar los pronósticos sobre su precio en un 6 %, lo cual está indicando, además de que tales pronósticos no mejoran de forma importante, que existe un mal ajuste de la recta a la nube de puntos. Parece evidente, sin tener todavía otro tipo de información, que el porcentaje de alcohol de las cervezas está más relacionado con el número de calorías que con su precio.

## Resumen

En este primer apartado se ha ofrecido una primera aproximación al análisis de regresión lineal como una técnica estadística que permite estudiar la relación entre una variable dependiente (VD) y una o más variables independientes (VI) con el doble propósito de: (1) averiguar en qué medida la VD puede estar explicada por la(s) VI y (2) obtener predicciones en la VD a partir de la(s) VI. El procedimiento implica, básicamente, obtener la ecuación mínimo-cuadrática que mejor expresa la relación entre la VD y la(s) VI y estimar mediante el coeficiente de determinación la calidad de la ecuación de regresión obtenida. Estos dos pasos deben ir acompañados de un chequeo del cumplimiento de las condiciones o supuestos que garantizan la validez del procedimiento.

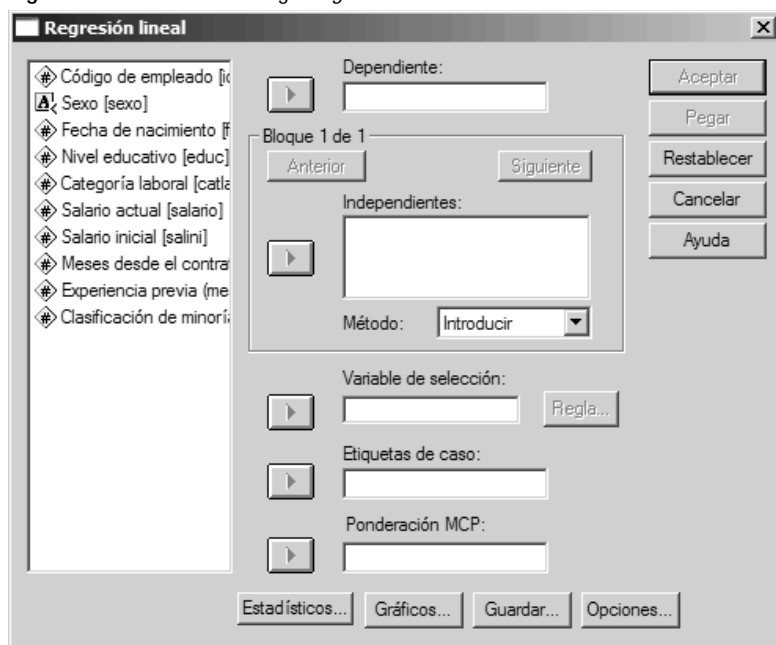
## Análisis de regresión lineal simple

Este apartado aborda un estudio algo más formal del análisis de regresión comenzando con el *modelo de regresión lineal simple* (*simple* = una variable independiente), pero conviene no perder de vista que, puesto que generalmente interesará estudiar simultáneamente más de una variable predictora, este modelo es sólo un punto de partida en el estudio del análisis de regresión.

Se sigue utilizando en todo momento el archivo *Datos de empleados* que, como ya se ha dicho, se encuentra en la misma carpeta en la que está instalado el SPSS. De momento, se utiliza la variable *salario* (salario actual) como variable dependiente y la variable *salini* (salario inicial) como variable independiente o predictora.

Para llevar a cabo un análisis de regresión simple con las especificaciones que el programa tiene establecidas por defecto:

- Seleccionar la opción **Regresión > Lineal...** del menú **Analizar** para acceder al cuadro de diálogo *Regresión lineal* que muestra la Figura 18.4.

Figura 18.4. Cuadro de diálogo *Regresión lineal*

- Seleccionar la variable *salario* (salario actual) en la lista de variables del archivo de datos y trasladarla al cuadro **Dependiente**.
- Seleccionar la variable *salini* (salario inicial) y trasladarla a la lista **Independientes**.

Aceptando estas selecciones, el *Visor* de resultados ofrece la información que muestran las Tablas 18.1 a la 18.3.

## Bondad de ajuste

La primera información que se obtiene (Tabla 18.1) se refiere al *coeficiente de correlación múltiple* ( $R$ ) y a su cuadrado. Puesto que el modelo de regresión del ejemplo sólo incluye dos variables, el coeficiente de correlación múltiple no es otra cosa que el valor absoluto del coeficiente de correlación de Pearson entre esas dos variables (ver capítulo anterior). Su cuadrado ( $R$  *cuadrado*) es el coeficiente de determinación:

$$R^2 = 1 - \frac{\text{Suma de cuadrados de los residuos}}{\text{Suma de cuadrados total}}$$

(los residuos son las diferencias existentes entre las puntuaciones observadas y los pronósticos obtenidos con la recta).

Además de la proporción de mejora en los pronósticos,  $R^2$  expresa la proporción de varianza de la variable dependiente que está explicada por la variable independiente. En el ejemplo (ver Tabla 18.1),  $R$  toma un valor muy alto (su máximo es 1); y  $R^2$  indica que el 77,5 %

de la variabilidad del *salario actual* está explicada por, depende de, o está asociada al *salario inicial*. Es importante señalar en este momento que el análisis de regresión no permite afirmar que las relaciones detectadas sean de tipo causal: únicamente es posible hablar de relación y de grado de relación. Debe quedar muy claro desde el principio que una relación, por sí sola, nunca implica causalidad.

$R^2$  *cuadrado corregida* es una corrección a la baja de  $R^2$  que se basa en el número de casos y de variables independientes:

$$R^2_{\text{corregida}} = R^2 - [p(1 - R^2)/(n - p - 1)]$$

( $p$  se refiere al número de variables independientes). En una situación con pocos casos y muchas variables independientes,  $R^2$  es un estimador algo optimista (artificialmente alto) del verdadero coeficiente de correlación poblacional. En tal caso, el valor de  $R^2$  *corregida* será sensiblemente más bajo que el de  $R^2$ . En el ejemplo, como hay 474 casos y una sola variable independiente, los dos valores de  $R^2$  (el corregido y el no corregido) son prácticamente iguales.

El *error típico de la estimación* ( $S_e$ ) es la desviación típica de los residuos, es decir, la desviación típica de las distancias existentes entre las puntuaciones en la variable dependiente ( $Y_i$ ) y los pronósticos efectuados con la recta de regresión ( $\hat{Y}_i$ ), aunque no exactamente, pues la suma de las distancias al cuadrado están divididas por  $n-2$ :

$$S_e = \sqrt{\sum (Y_i - \hat{Y}_i)^2 / (n - 2)}$$

En realidad, este error típico es la raíz cuadrada de la *media cuadrática residual* que recoge la Tabla 18.2. Representa, por tanto, una medida de la parte de variabilidad de la variable dependiente que no está explicada por la recta de regresión. Cuanto mayor es  $R^2$ , menor es  $S_e$ .

Tabla 18.1. Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,880	,775	,774	8.115,36

La tabla *resumen del ANOVA* (Tabla 18.2) informa sobre si existe o no relación significativa entre la variable independiente y la dependiente. El estadístico  $F$  permite contrastar la hipótesis nula de que el valor poblacional de  $R$  es cero (que en el modelo de regresión simple equivale a contrastar la hipótesis de que la pendiente de la recta de regresión vale cero). El nivel crítico (*Sig.*) indica que, si se supone que el valor poblacional de  $R$  es cero, es improbable (probabilidad  $< 0,0005$ ) que  $R$ , en esta muestra, tome el valor 0,88. Lo cual implica que el valor poblacional de  $R$  es mayor que cero y que, en consecuencia, puede afirmarse que ambas variables están linealmente relacionadas.

Tabla 18.2. Resumen del ANOVA

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	106.831.048.750,12	1	106.831.048.750,12	1622,12	,000
	Residual	31.085.446.686,22	472	65.858.997,22		
	Total	137.916.495.436,34	473			

# Ecuación de regresión

La Tabla 18.3 muestra los coeficientes de la recta de regresión. La columna etiquetada *Coeficientes no estandarizados* (no tipificados) contiene los coeficientes de regresión parcial que definen la ecuación de regresión en puntuaciones directas.

Tabla 18.3. Coeficientes de regresión parcial

Modelo	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
	B	Error típ.	Beta		
1 (Constante)	1.928,21	888,68		2,17	,031
Salario inicial	1,91	,05	,88	40,28	,000

El coeficiente no tipificado (no estandarizado) correspondiente a la *constante* es el *origen* de la recta de regresión ( $B_0$ ). Recibe el nombre de *constante* porque, según se verá, es la constante del modelo de regresión:

$$B_0 = \bar{Y} - B_1 \bar{X}$$

Y el coeficiente no tipificado (no estandarizado) correspondiente a *salario inicial* es la *pendiente* de la recta de regresión ( $B_1$ ):

$$B_1 = \frac{\sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

$B_1$  indica el cambio medio que corresponde a la variable dependiente (*salario*) por cada unidad de cambio de la variable independiente (*salini*). Según esto, la ecuación de regresión queda de la siguiente manera:

$$\text{Pronóstico en salario} = 1928,21 + 1,91 \text{ salini}$$

Es decir, a cada valor de *salini* le corresponde un pronóstico en *salario* basado en un incremento constante (1928,206) más 1,909 veces el valor de *salini*.

## Coeficientes de regresión tipificados

Los coeficientes *Beta* (coeficientes de regresión parcial tipificados o estandarizados) son los coeficientes que definen la ecuación de regresión cuando ésta se obtiene tras tipificar las variables originales, es decir, tras convertir las puntuaciones directas en típicas. Se obtiene de la siguiente manera:  $\beta_1 = B_1 (S_x / S_y)$ .

En el análisis de regresión simple, el coeficiente de regresión tipificado correspondiente a la única variable independiente presente en la ecuación coincide exactamente con el coeficiente de correlación de Pearson. En regresión múltiple no ocurre esto pero, según se verá enseguida, los coeficientes de regresión tipificados ayudan a valorar la importancia relativa de cada variable independiente dentro de la ecuación.

## Pruebas de significación

Finalmente, los estadísticos  $t$  y sus niveles críticos (*Sig.*) permiten contrastar las hipótesis nulas de que los coeficientes de regresión valen cero en la población. Estos estadísticos  $t$  se obtienen dividiendo los coeficientes de regresión  $B_0$  y  $B_1$  entre sus correspondientes errores típicos:

$$t_{B_0} = \frac{B_0}{S_{B_0}} \quad \text{y} \quad t_{B_1} = \frac{B_1}{S_{B_1}}$$

Siendo:

$$S_{B_0} = S_e \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2}} \quad \text{y} \quad S_{B_1} = \frac{S_e}{\sqrt{\sum (X_i - \bar{X})^2}}$$

Estos estadísticos  $t$  se distribuyen según el modelo de probabilidad  $t$  de Student con  $n-2$  grados de libertad. Por tanto, pueden utilizarse para decidir si un determinado coeficiente de regresión es significativamente distinto de cero y, en consecuencia, en el caso de  $B_1$ , si la variable independiente está significativamente relacionada con la dependiente.

Puesto que en regresión simple se trabaja con una única variable independiente, el resultado del estadístico  $t$  (Tabla 18.3) es equivalente al del estadístico  $F$  de la tabla *resumen del ANOVA* (Tabla 18.2). De hecho, en regresión simple,  $t^2 = F$ .

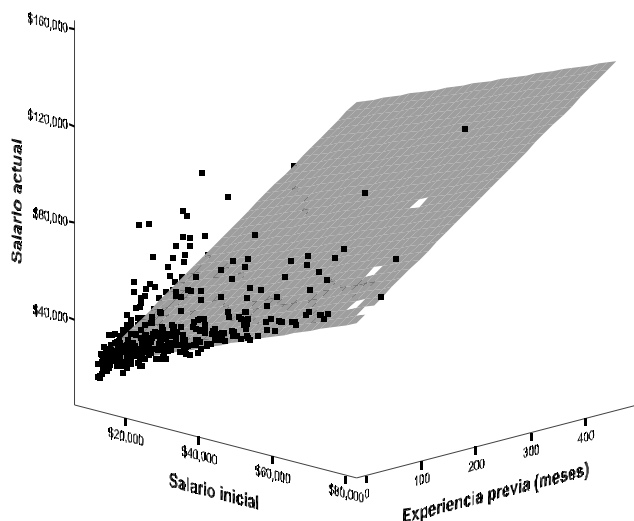
Los resultados de la Tabla 18.3 permiten (a la espera de contrastar los supuestos en los que se basa el análisis de regresión), establecer las siguientes conclusiones:

1. El *origen* poblacional de la recta de regresión ( $\beta_0$ ) es significativamente distinto de cero (generalmente, contrastar la hipótesis « $\beta_0 = 0$ » carece de utilidad, pues no contiene información sobre la relación entre  $X_i$  e  $Y_i$ ).
2. La *pendiente* poblacional de la recta de regresión (el coeficiente de regresión  $\beta_1$  correspondiente a *salini*) es significativamente distinta de cero, lo cual permite afirmar que entre *salario* y *salini* existe relación lineal significativa.

## Análisis de regresión lineal múltiple

El procedimiento **Regresión lineal** permite utilizar más de una variable independiente y, por tanto, permite ajustar modelos de regresión lineal múltiple (*múltiple* = más de una variable independiente).

Pero, en un análisis de regresión múltiple, la ecuación de regresión ya no define una recta en un plano, sino un hiperplano en un espacio multidimensional. En un modelo con, por ejemplo, dos variables independientes, el diagrama de dispersión adopta la forma de un plano en un espacio tridimensional. Así, con *salario* como variable dependiente y *salini* (salario inicial) y *expprev* (experiencia previa) como variables independientes, el diagrama de dispersión adopta el formato que muestra la Figura 18.5.

Figura 18.5. Diagrama de dispersión: *salario actual* sobre *salario inicial* y *experiencia previa*

Es decir, con dos variables independientes es necesario utilizar tres ejes para poder representar el correspondiente diagrama de dispersión. Y si en lugar de dos variables independientes se utilizaran tres, sería necesario un espacio de cuatro dimensiones para poder construir el diagrama de dispersión. Y un espacio de cinco dimensiones para poder construir el diagrama correspondiente a un modelo con cuatro variables independientes. Etc. Por tanto, con dos variables independientes, la representación gráfica de un modelo de regresión resulta poco intuitiva y, por tanto, poco útil. Y con más de dos variables independientes, la representación gráfica resulta simplemente imposible.

Sin embargo, la complejidad de la representación gráfica de una ecuación de regresión múltiple contrasta con la simplicidad de su la expresión algebraica:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon_i$$

En un modelo de estas características, la variable dependiente ( $Y_i$ ) se interpreta como la combinación lineal de un conjunto de  $p$  variables independientes ( $X_j$ ), cada una de las cuales va acompañada de un coeficiente ( $\beta_j$ ) que indica el peso relativo de esa variable en la ecuación. El modelo incluye además una constante ( $\beta_0$ ) y un componente aleatorio (los residuos:  $\epsilon_i$ ) que recoge todo lo que las variables independientes no explican.

Este modelo, en cuanto modelo estadístico que es, se basa en una serie de supuestos que se estudiarán con detalle en el siguiente apartado.

Los términos del modelo de regresión, al igual que los de cualquier otro modelo estadístico, son valores poblacionales. Para poder trabajar con él es necesario estimarlos. Y las estimaciones mínimo-cuadráticas se obtienen, según se ha señalado ya, intentando minimizar la suma de las diferencias al cuadrado entre los valores observados ( $Y_i$ ) y los pronosticados ( $\hat{Y}_i$ ):

$$\hat{Y}_i = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_p X_p$$

Al igual que en el análisis de regresión simple descrito en el apartado anterior, se seguirá utilizando la variable *salario* (salario actual) como variable dependiente. Pero ahora se van a incluir en el modelo tres variables independientes: *salini* (salario inicial), *expprev* (experiencia previa) y *educ* (nivel educativo).

Para llevar a cabo un análisis de regresión múltiple con las especificaciones que el programa tiene establecidas por defecto:

- Seleccionar la opción **Regresión > Lineal...** del menú **Analizar** para acceder al cuadro de diálogo *Regresión lineal* (ver Figura 18.4).
- Seleccionar la variable *salario* en la lista de variables del archivo de datos y trasladarla al cuadro **Dependiente**.
- Seleccionar las variables *salini*, *expprev* y *educ* en la lista de variables del archivo de datos y trasladarlas a la lista **Independientes**.

Aceptando estas selecciones, el *Visor de resultados* ofrece la información que muestran las Tablas 18.4 a la 18.6.

Bondad de ajuste

La Tabla 18.4 ofrece un resumen del modelo. Este resumen se refiere, básicamente, a la calidad del modelo de regresión, es decir, al grado de *ajuste* (de parecido) entre los pronósticos de la ecuación de regresión y el salario de los sujetos: tomadas juntas, las tres variables independientes incluidas en el análisis explican un 80 % de la varianza de la variable dependiente ( $R^2$  corregida = 0,80). Además, el error típico de los residuos (8.115,36 en el análisis de regresión simple) ha disminuido algo (7.631,92 en el análisis de regresión múltiple), lo que indica una ligera mejora en el ajuste. De nuevo, como el número de variables es pequeño en relación al número de casos, el valor corregido de  $R^2$  es casi idéntico al valor no corregido.

Tabla 18.4. Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,895	,802	,800	7.631,92

El estadístico *F* (ver Tabla 18.5) contrasta la hipótesis nula de que el valor poblacional de *R* es cero y, por tanto, permite decidir si existe relación lineal significativa entre la variable dependiente y el conjunto de variables independientes tomadas juntas. El valor del nivel crítico (*Sig.* < 0,0005), puesto que es menor que 0,05, indica que sí existe relación lineal significativa. Puede afirmarse, por tanto, que el hiperplano definido por la ecuación de regresión ofrece un buen ajuste a la nube de puntos.

Tabla 18.5. Resumen del ANOVA

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	110.540.801.465,35	3	36.846.933.821,78	632,61	,000
	Residual	27.375.693.970,99	470	58.246.157,39		
	Total	137.916.495.436,34	473			



# Ecuación de regresión

La tabla de *coeficientes de regresión parcial* (ver Tabla 18.6) contiene toda la información necesaria para construir la ecuación de regresión mínimo-cuadrática. En la columna encabezada *Coeficientes no estandarizados* (no tipificados) se encuentran los coeficientes ( $B_j$ ) que forman parte de la ecuación en puntuaciones directas:

$$\text{Pronóstico en salario} = -3.661,52 + 1,75 \text{ salini} - 16,73 \text{ expprev} + 735,96 \text{ educ}$$

Estos coeficientes no tipificados se interpretan en los términos ya conocidos. Por ejemplo, el coeficiente correspondiente a la variable *salini*, que vale 1,75, indica que, si el resto de términos de la ecuación se mantienen constantes, a un aumento de una unidad (un dólar) en *salini* le corresponde un aumento de 1,75 unidades (dólares) en *salario*.

Tabla 18.6. Coeficientes de regresión parcial

Modelo	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
	B	Error típ.	Beta		
1 (Constante)	-3.661,52	1.935,49		-1,89	,059
Salario inicial	1,75	,06	,81	29,20	,000
Experiencia previa	-16,73	3,61	-,10	-4,64	,000
Nivel educativo	735,96	168,69	,12	4,36	,000

Conviene señalar que estos coeficientes no son independientes entre sí. De hecho, reciben el nombre de coeficientes de regresión *parcial* porque el valor concreto estimado para cada coeficiente se ajusta o corrige teniendo en cuenta la presencia en el modelo del resto de variables independientes. Conviene, por tanto, interpretarlos con cautela.

El signo del coeficiente de regresión parcial de una variable puede no ser el mismo que el del coeficiente de correlación simple entre esa variable y la dependiente. Esto es debido a los ajustes que se llevan a cabo para poder obtener la mejor ecuación posible. Aunque existen diferentes explicaciones para justificar el cambio de signo de un coeficiente de regresión, una de las que deben ser más seriamente consideradas es la que se refiere a la presencia de un alto grado de asociación entre algunas de las variables independientes (colinealidad). Se tratará esta cuestión más adelante.

## Coeficientes de regresión tipificados

Los coeficientes *Beta* están basados en las puntuaciones típicas y, por tanto, son directamente comparables entre sí. Indican la cantidad de cambio, en puntuaciones típicas, que se producirá en la variable dependiente por cada cambio de una unidad en la correspondiente variable independiente (manteniendo constantes el resto de variables independientes).

Estos coeficientes proporcionan una pista muy útil (aunque no definitiva) sobre la importancia relativa de cada variable independiente en la ecuación de regresión. En principio, una variable tiene tanto más peso (importancia) en la ecuación de regresión cuanto mayor (en valor absoluto) es su coeficiente de regresión tipificado. Observando los coeficientes *Beta* de

la Tabla 18.6 puede comprobarse que la variable *salini* es la más importante (la que más peso tiene en la ecuación); después, *educ*; por último, *expprev*. No obstante, esta interpretación debe ser asumida con cautela (ver Cooley y Lohnes, 1971; o Darlington, 1968, 1990): lo ya dicho sobre la interdependencia de los coeficientes de regresión parcial no tipificados también vale para los tipificados.

## Pruebas de significación

Las pruebas *t* y sus niveles críticos (últimas dos columnas de la Tabla 18.6: *t* y *Sig.*) sirven para contrastar la hipótesis nula de que un coeficiente de regresión vale cero en la población. Niveles críticos (*Sig.*) pequeños (menores que 0,05) indican que esa hipótesis debe ser rechazada. Un coeficiente de regresión de cero indica ausencia de relación lineal. Un coeficiente significativamente distinto de cero indica que existe relación lineal. Consecuentemente, los coeficientes significativamente distintos de cero informan sobre qué variables poseen un peso significativo en la ecuación de regresión, es decir, qué variables contribuyen significativamente al ajuste del modelo.

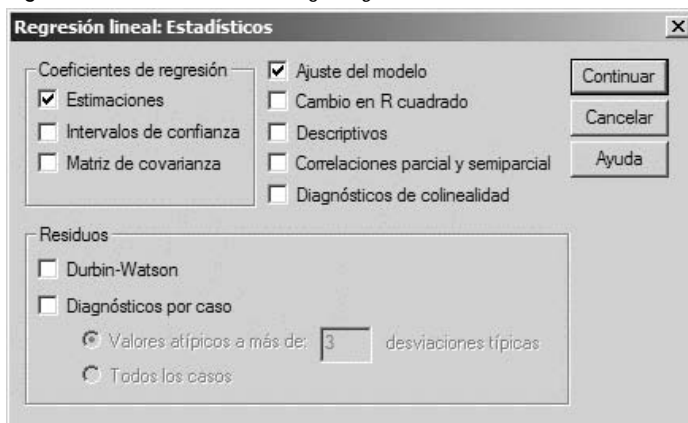
El nivel crítico asociado a los estadísticos *t* (ver Tabla 18.6) indica que las tres variables poseen coeficientes significativamente distintos de cero: en todos ellos, *Sig.* < 0,0005. Las tres variables, por tanto, contribuyen de forma significativa al ajuste del modelo.

## Información complementaria

Además de obtener la ecuación de regresión y valorar la calidad de su ajuste, un análisis de regresión no debe renunciar a obtener algunos estadísticos descriptivos elementales como la matriz de correlaciones, la media y la desviación típica de cada variable, el número de casos con el que se está trabajando, etc. Para obtener estos estadísticos:

- Pulsar el botón Estadísticos... del cuadro de diálogo principal (ver Figura 18.4) para acceder al subcuadro de diálogo *Regresión lineal: Estadísticos* que muestra la Figura 18.6.

Figura 18.6. Subcuadro de diálogo *Regresión lineal: Estadísticos*



Entre las opciones que ofrece este subcuadro de diálogo, existen dos que se encuentran marcadas por defecto. Estas dos opciones ya marcadas son precisamente las que permiten obtener la información que recogen las Tablas 18.1 a la 18.6 cuando se pulsa el botón **Aceptar** del cuadro de diálogo *Regresión lineal* (ver Figura 18.4) sin hacer otra cosa que seleccionar la variable dependiente y la(s) independiente(s):

- " **Estimaciones.** Ofrece las estimaciones de los coeficientes de regresión parcial no tipificados (*B*) y tipificados (*Beta*), junto con las pruebas de significación *t* individuales para contrastar las hipótesis de que el valor poblacional de esos coeficientes es cero (ver Tablas 18.3 y 18.6).
- " **Ajuste del modelo.** Muestra el coeficiente de correlación múltiple, su cuadrado corregido y no corregido, y el error típico de los residuos (ver Tablas 18.1 y 18.4: *R*, *R*<sup>2</sup>, *R*<sup>2</sup> corregida y error típico de la estimación). Esta opción también permite obtener la tabla *resumen del ANOVA*, la cual contiene el estadístico *F* para contrastar la hipótesis «*R*=0» (ver Tablas 18.2 y 18.4).

Al margen de estas dos opciones, que se encuentran activas por defecto, el subcuadro de diálogo *Regresión lineal: Estadísticos* (Figura 18.6) contiene varias opciones que ofrecen información muy útil en un análisis de regresión:

- " **Intervalos de confianza.** Esta opción, situada en el recuadro **Coefficientes de regresión**, hace que, además de las estimaciones puntuales de los coeficientes de regresión parcial (las cuales ya se obtienen con la opción **Estimaciones**), puedan obtenerse también los intervalos de confianza para esos coeficientes (ver Tabla 18.7).

Estos intervalos informan de los límites entre los que cabe esperar que se encuentre el valor poblacional de cada coeficiente de regresión. Los límites se obtienen sumando y restando 1,96 errores típicos al valor del correspondiente coeficiente de regresión (1,96 porque el SPSS trabaja, por defecto, con un nivel de confianza de 0,95).

Intervalos de confianza muy amplios indican que las estimaciones obtenidas son poco precisas y, probablemente, inestables (cosa que suele ocurrir, por ejemplo, cuando existen problemas de colinealidad; se estudiará esta cuestión más adelante, en el apartado dedicado a los supuestos del modelo de regresión).

**Tabla 18.7.** Coeficientes de regresión parcial, incluyendo los intervalos de confianza

Modelo: 1

	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza para B al 95%	
	B	Error típ.	Beta			Límite inferior	Límite superior
(Constante)	-3.661,52	1.935,49		-1,89	,059	-7.464,80	141,77
Salario inicial	1,75	,06	,81	29,20	,000	1,63	1,87
Experiencia previa	-16,73	3,61	-,10	-4,64	,000	-23,81	-9,65
Nivel educativo	735,96	168,69	,12	4,36	,000	404,48	1.067,43

- " **Matriz de covarianza.** Muestra una matriz con las covarianzas y correlaciones existentes entre los coeficientes de regresión parcial. Los valores obtenidos (ver Tabla 18.8) indican que, efectivamente, los coeficientes de regresión parcial no son independientes entre sí.

Tabla 18.8. Correlaciones entre los coeficientes de regresión

Modelo: 1

		Nivel educativo	Experiencia previa	Salario inicial
Correlaciones	Nivel educativo	1,00	,36	-,67
	Experiencia previa	,36	1,00	-,27
	Salario inicial	-,67	-,27	1,00
Covarianzas	Nivel educativo	28.456,06	220,96	-6,74
	Experiencia previa	220,96	13,00	-,06
	Salario inicial	-6,74	-,06	,00

“ **Descriptivos.** Esta opción permite obtener la media y la desviación típica insesgada de todas las variables incluidas en el análisis, y el número de casos válidos (ver Tabla 18.9). También permite obtener la matriz de correlaciones bivariadas entre el conjunto de variables incluidas en el análisis (ver Tabla 18.10). En esta matriz de correlaciones, cada coeficiente de correlación aparece acompañado de su correspondiente nivel crítico (que permite decidir sobre la hipótesis nula de que el coeficiente vale cero en la población), y del número de casos sobre el que se ha calculado cada coeficiente (que coincidirá o no con el número de casos válidos del análisis de regresión dependiendo de la opción elegida para el tratamiento de valores perdidos en el cuadro de diálogo *Opciones*).

Lógicamente, en la diagonal principal de la matriz de correlaciones aparecen unos, pues la relación entre una variable y ella misma es perfecta (en realidad, la matriz de correlaciones es la matriz de varianzas-covarianzas tipificada: en la diagonal principal aparecen las varianzas tipificadas; y, fuera de la diagonal, las covarianzas tipificadas).

Tabla 18.9. Estadísticos descriptivos

	Media	Desviación típ.	N
Salario actual	34.419,57	17.075,66	474
Salario inicial	17.016,09	7.870,64	474
Experiencia previa	95,86	104,59	474
Nivel educativo	13,49	2,88	474

Tabla 18.10. Correlaciones entre las variables incluidas en el modelo

		Salario actual	Salario inicial	Experiencia previa	Nivel educativo
Salario actual	Correlación de Pearson	1,000	,880	-,097	,661
	Sig. (unilateral)	.	,000	,017	,000
	N	474	474	474	474
Salario inicial	Correlación de Pearson	,880	1,000	,045	,633
	Sig. (unilateral)	,000	.	,163	,000
	N	474	474	474	474
Experiencia previa	Correlación de Pearson	-,097	,045	1,000	-,252
	Sig. (unilateral)	,017	,163	.	,000
	N	474	474	474	474
Nivel educativo	Correlación de Pearson	,661	,633	-,252	1,000
	Sig. (unilateral)	,000	,000	,000	.
	N	474	474	474	474

“ **Correlaciones parcial y semiparcial.** Esta opción permite obtener los coeficientes de correlación parcial y semiparcial entre la variable dependiente y cada una de las variables independientes.

Un coeficiente de *correlación parcial* expresa el grado de relación existente entre dos variables tras eliminar de ambas el efecto debido a terceras variables (ver capítulo anterior). En el contexto del análisis de regresión, los coeficientes de correlación parcial expresan el grado de relación existente entre cada variable independiente y la variable dependiente tras eliminar de ambas el efecto debido al resto de variables independientes incluidas en la ecuación de regresión.

Un coeficiente de *correlación semiparcial* expresa el grado de relación existente entre dos variables tras eliminar de una de ellas el efecto debido a terceras variables. En el contexto del análisis de regresión, los coeficientes de correlación semiparcial expresan el grado de relación existente entre la variable dependiente y la parte de cada variable independiente que no está explicada por el resto de variables independientes incluidas en la ecuación de regresión.

Seleccionando la opción **Correlaciones parcial y semiparcial**, la tabla de *coeficientes de regresión parcial* (Tabla 18.6, ya vista) incluye la información adicional que muestra la Tabla 18.11.

Junto con los coeficientes de correlación parcial y semiparcial, aparecen las correlaciones de *orden cero*, es decir, los coeficientes de correlación calculados sin tener en cuenta la presencia de terceras variables (se trata de los mismos coeficientes que aparecen en la Tabla 18.10). Comparando entre sí estos coeficientes (de orden cero, parcial y semiparcial) pueden encontrarse pautas de relación interesantes. En la información que ofrece la Tabla 18.11 se observa, por ejemplo, que la correlación entre la variable dependiente *salario actual* y la variable independiente *nivel educativo* vale 0,66. Sin embargo, al eliminar de *salario actual* y de *nivel educativo* el efecto atribuible al resto de variables independientes incluidas en el modelo (*salario inicial* y *experiencia previa*), la correlación baja hasta 0,20 (parcial); y cuando el efecto atribuible a *salario inicial* y *experiencia previa* se elimina sólo de *nivel educativo*, la correlación baja hasta 0,09 (semiparcial). Lo cual está indicando que la relación entre *salario actual* y *nivel educativo* podría explicarse casi por completo recurriendo a las otras dos variables independientes.

**Tabla 18.11.** Coeficientes de regresión parcial y coeficientes de correlación parcial y semiparcial

Modelo: 1

	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Correlaciones		
	B	Error típ.	Beta			Orden cero	Parcial	Semi-parcial
(Constante)	-3.661,52	1.935,49		-1,89	,059			
Salario inicial	1,75	,06	,81	29,20	,000	,88	,80	,60
Exp. previa	-16,73	3,61	-,10	-4,64	,000	-,10	-,21	-,10
Nivel educativo	735,96	168,69	,12	4,36	,000	,66	,20	,09

El resto de opciones del subcuadro de diálogo *Regresión lineal: Estadísticos* (ver Figura 18.6) tienen que ver con algunos supuestos del modelo de regresión lineal (*estadísticos de colinealidad, residuos*) y con el análisis de regresión por pasos (*cambio en R cuadrado*). Todas estas opciones se tratan más adelante.

## Supuestos del modelo de regresión lineal

Los supuestos de un modelo estadístico se refieren a una serie de condiciones que deben darse para garantizar la validez del modelo. Al efectuar aplicaciones prácticas del modelo de regresión, es necesario vigilar el cumplimiento de estos supuestos:

1. *Linealidad.* La ecuación de regresión adopta una forma particular; en concreto, la variable dependiente es la suma de un conjunto de elementos: el origen de la recta, una combinación lineal de variables independientes o predictoras y los residuos. El incumplimiento del supuesto de linealidad suele denominarse error de especificación. Algunos ejemplos son: omisión de variables independientes importantes, inclusión de variables independientes irrelevantes, no linealidad (la relación entre las variables independientes y la dependiente no es lineal), parámetros cambiantes (los parámetros no permanecen constantes durante el tiempo que dura la recogida de datos), no aditividad (el efecto de alguna variable independiente es sensible a los niveles de alguna otra variable independiente), etc.
2. *Independencia.* Los residuos son independientes entre sí, es decir, los residuos constituyen una variable aleatoria (los residuos son las diferencias entre los valores observados y los pronosticados). Es frecuente encontrarse con residuos autocorrelacionados cuando se trabaja con series temporales.
3. *Homocedasticidad.* Para cada valor de la variable independiente (o combinación de valores de las variables independientes), la varianza de los residuos es constante.
4. *Normalidad.* Para cada valor de la variable independiente (o combinación de valores de las variables independientes), los residuos se distribuyen normalmente con media cero.
5. *No-colinealidad.* No existe relación lineal exacta entre ninguna de las variables independientes. El incumplimiento de este supuesto da origen a colinealidad o multicolinealidad.

Sobre el cumplimiento del primer supuesto puede obtenerse información a partir de una inspección del diagrama de dispersión: si se tiene intención de utilizar el modelo de regresión lineal, lo razonable es que la relación entre la variable dependiente y las independientes sea de tipo lineal (existen *gráficos parciales* que permiten obtener una representación de la relación *neta* existente entre dos variables; se estudiarán más adelante). El quinto supuesto, *no-colinealidad*, no tiene sentido en regresión simple, pues es imprescindible la presencia de más de una variable independiente (se estudiarán diferentes formas de diagnosticar la presencia de colinealidad). El resto de los supuestos, *independencia*, *homocedasticidad* y *normalidad*, están estrechamente asociados al comportamiento de los residuos; por tanto, un análisis cuidadoso de los residuos puede informar sobre el cumplimiento de los mismos.

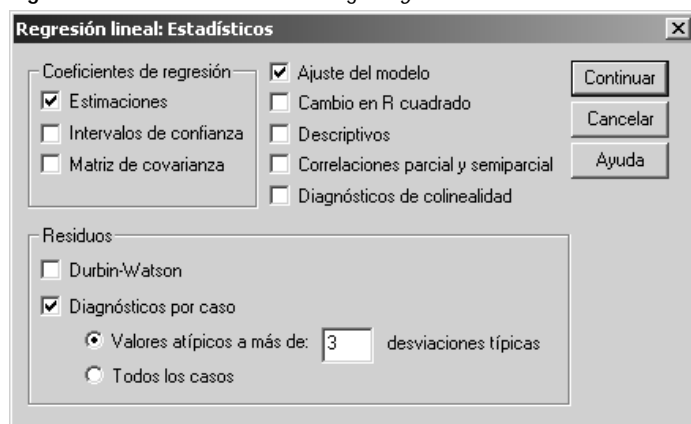
## Análisis de los residuos

Los residuos de un modelo estadístico son las diferencias existentes entre los valores observados y los valores pronosticados:  $(Y_i - \hat{Y}_i)$ . Pueden obtenerse marcando la opción **No tipificados** dentro del recuadro **Residuos** en el subcuadro de diálogo *Regresión lineal: Guardar nuevas variables* (ver Figura 18.12, más adelante). Los residuos son muy importantes en el análisis

de regresión. En primer lugar, informan sobre el grado de exactitud de los pronósticos: cuanto más pequeño es el error típico de los residuos (ver Tabla 18.1: *error típico de la estimación*), mejores son los pronósticos, o lo que es lo mismo, mejor se ajusta la recta de regresión a los puntos del diagrama de dispersión. En segundo lugar, el análisis de las características de los casos con residuos grandes (sean positivos o negativos; es decir, *grandes en valor absoluto*) puede ayudar a detectar casos atípicos y, consecuentemente, a perfeccionar la ecuación de regresión a través de un estudio detallado de los mismos (ver Cook, 1993).

La opción **Diagnósticos por caso** del cuadro de diálogo *Regresión lineal: Estadísticos* (ver Figura 18.6.bis) ofrece un listado de todos los residuos o, alternativamente (y esto es más interesante), un listado de los residuos que se alejan de cero (el valor esperado de los residuos) en más de un determinado número de desviaciones típicas.

Figura 18.6.bis. Subcuadro de diálogo *Regresión lineal: Estadísticos*



Por defecto, el SPSS ofrece un listado de los residuos que se alejan de cero más de 3 desviaciones típicas, pero esto puede cambiarse introduciendo el valor deseado. Para obtener un listado de los residuos que se alejan de cero más de 3 desviaciones típicas:

- Marcar la opción **Diagnósticos por caso** y seleccionar **Valores atípicos a más de [3] desviaciones típicas**.

Acceptando estas selecciones, el *Visor* ofrece los resultados que muestran las Tablas 18.12 y 18.13. La Tabla 18.12 contiene los *residuos tipificados* (residuos divididos por su error típico; una variable tipificada, con media 0 y desviación típica 1). La tabla recoge los casos con residuos que se alejan de su media (cero) más de 3 desviaciones típicas. Si estos residuos están normalmente distribuidos (cosa que se asume en el análisis de regresión), cabe esperar que el 95 % de ellos se encuentre en el rango  $[-1,96, +1,96]$ . Y el 99,9 %, en el rango  $[-3, +3]$ . Es fácil, por tanto, identificar los casos que poseen residuos grandes.

En la práctica, los casos con residuos muy grandes o muy pequeños deben ser examinados para averiguar si las puntuaciones que tienen asignadas son o no correctas. Si, a pesar de tener asociados residuos muy grandes o muy pequeños, las puntuaciones asignadas son correctas, conviene estudiar esos casos detenidamente para averiguar si difieren de algún modo y de forma sistemática del resto de los casos. Esto último es fácil de hacer con el SPSS pues,

según se verá más adelante, el programa permite salvar los residuos correspondientes a cada caso como una variable más del archivo de datos y, a partir de ahí, utilizarlos en los procedimientos SPSS que se considere pertinente.

Tabla 18.12. Diagnósticos por caso (listado de los residuos más grandes en valor absoluto)

Número de caso	Residuo tipificado	Salario actual	Valor pronosticado	Residuo
18	6,381	103.750	55.048,80	48.701,20
32	3,095	110.625	87.004,54	23.620,46
103	3,485	97.000	70.405,22	26.594,78
106	3,897	91.250	61.505,37	29.744,63
205	-3,781	66.750	95.602,99	-28.852,99
218	5,981	80.000	34.350,68	45.649,32
274	4,953	83.750	45.946,77	37.803,23
449	3,167	70.000	45.829,66	24.170,34
454	3,401	90.625	64.666,70	25.958,30

Además de la tabla de *diagnósticos por caso*, el *Visor* ofrece una tabla resumen con información sobre el valor máximo y mínimo, y la media y la desviación típica insesgada de los pronósticos, de los residuos, de los pronósticos tipificados y de los residuos tipificados (ver Tabla 18.13). Especialmente importante es advertir que la media de los residuos vale cero.

Tabla 18.13. Estadísticos sobre los residuos

	Mínimo	Máximo	Media	Desviación típ.	N
Valor pronosticado	12.382,90	146.851,63	34.419,57	15.287,30	474
Residuo	-28.852,99	48.701,20	,00	7.607,68	474
Valor pronosticado tipificado	-1,44	7,35	,00	1,00	474
Residuo tipificado	-3,78	6,38	,00	1,00	474

## Independencia

El verdadero interés de los residuos reside en su capacidad para ofrecer información crucial sobre el cumplimiento de varios supuestos del modelo de regresión lineal. En concreto, un análisis detallado de los residuos permite obtener información sobre los supuestos de independencia, homocedasticidad, normalidad y linealidad.

Uno de los supuestos básicos del modelo de regresión lineal es el de independencia entre los residuos (supuesto éste que es necesario vigilar cuando los datos se han recogido siguiendo una secuencia temporal). El estadístico de *Durbin-Watson* (1950, 1951, 1971) proporciona información sobre el grado de independencia (o, si se prefiere, del grado de autocorrelación) existente entre ellos:

$$DW = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$



(donde  $e_i$  se refiere a los residuos:  $e_i = Y_i - \hat{Y}_i$ ). El estadístico *DW* oscila entre 0 y 4, y toma el valor 2 cuando los residuos son completamente independientes. Los valores menores que 2 indican autocorrelación positiva; los valores mayores que 2 indican autocorrelación negativa. Suele asumirse que los residuos son independientes cuando el estadístico *DW* toma valores comprendidos entre 1,5 y 2,5 (existen tablas para valorar este estadístico; ver por ejemplo, SPSS, 1999). Para obtener el estadístico de *Durbin-Watson*:

- En el subcuadro de diálogo *Regresión lineal: Estadísticos* (ver Figura 18.6.bis), seleccionar la opción de **Durbin-Watson**.

Esta opción permite obtener la tabla *resumen del modelo* (ya vista) con información adicional referida al estadístico de Durbin-Watson (ver Tabla 18.4). Puesto que el estadístico *DW* vale 1,832 (se encuentra entre 1,5 y 2,5), puede asumirse que los residuos son independientes (es decir, no hay razones para pensar que se incumpla el supuesto de independencia).

Debe tenerse en cuenta que la independencia entre residuos sólo es necesario evaluarla cuando el orden de los casos en el archivo de datos se ajusta a algún tipo de secuencia. Si los casos constituyen una muestra aleatoria, su posición en el archivo será irrelevante y la relación entre los residuos cambiará al cambiar las posiciones de los casos.

Tabla 18.14. Resumen del modelo

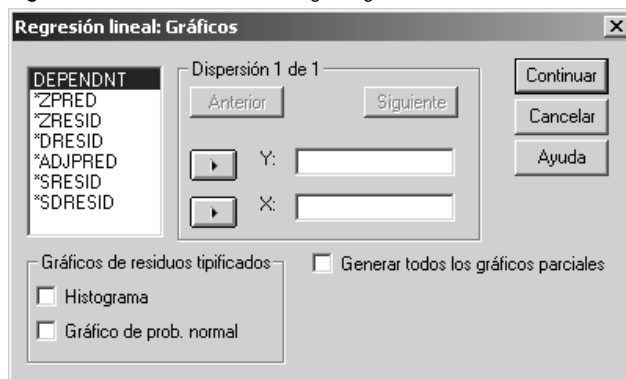
Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,895	,802	,800	7.631,92	1,832

## Homocedasticidad

El procedimiento *Regresión lineal* dispone de una serie de gráficos que proporcionan, entre otras cosas, información sobre el grado de cumplimiento de los supuestos de homocedasticidad y normalidad de los residuos. Para obtener estos gráficos:

- Pulsar el botón **Gráficos...** del cuadro de diálogo principal (ver Figura 18.4) para acceder al subcuadro de diálogo *Regresión lineal: Gráficos* que muestra la Figura 18.7.

Figura 18.7. Subcuadro de diálogo *Regresión lineal: Gráficos*



Las variables listadas permiten obtener diferentes gráficos de dispersión. Las variables precedidas por un asterisco son variables creadas por el SPSS; todas ellas pueden crearse en el *Editor de datos* marcando las opciones pertinentes del recuadro **Residuos** del subcuadro de diálogo *Regresión lineal: Guardar nuevas variables* (ver Figura 18.12):

**DEPENDENT:** variable dependiente de la ecuación de regresión.

**ZPRED** (pronósticos tipificados): pronósticos divididos por su desviación típica. Son pronósticos transformados en puntuaciones  $Z$  (es decir, en una variable tipificada con media 0 y desviación típica 1).

**ZRESID** (residuos tipificados): residuos divididos por su desviación típica. El tamaño de cada residuo tipificado indica el número de desviaciones típicas que un residuo se aleja de su media (la cual vale cero), de modo que, si están normalmente distribuidos (y esto es algo que asume el modelo de regresión), el 95 % de estos residuos se encontrará en el rango  $(-1.96, +1.96)$ , lo cual permite identificar fácilmente casos con residuos grandes. En muchos contextos es habitual valorar estos residuos utilizando 3 desviaciones típicas y, por tanto, considerar *grandes* los residuos cuyo valor absoluto es mayor que 3.

**DRESID** (residuos eliminados o corregidos): residuos obtenidos al efectuar los pronósticos eliminando de la ecuación de regresión el caso sobre el que se efectúa el pronóstico. El residuo correspondiente a cada caso se obtiene a partir del pronóstico efectuado con una ecuación de regresión en la que no se ha incluido ese caso. Son residuos muy útiles para detectar puntos de influencia (casos con gran peso en la ecuación de regresión; ver más adelante).

**ADJPRED** (pronósticos corregidos): pronósticos efectuados con una ecuación de regresión en la que no se incluye el caso pronosticado (ver residuos eliminados o corregidos). Diferencias importantes entre **PRED** y **ADJPRED** delatan la presencia de puntos de influencia (casos con gran peso en la ecuación de regresión; ver más adelante).

**SRESID** (residuos estudentizados): residuos divididos por su desviación típica, basada ésta en cómo de próximo se encuentra un caso a su(s) media(s) en la(s) variable(s) independiente(s). Al igual que ocurre con los residuos tipificados (a los que se parecen mucho), los estudentizados están escalados en unidades de desviación típica. Se distribuyen según el modelo de probabilidad  $t$  de *Student* con  $n-p-1$  grados de libertad ( $p$  se refiere al número de variables independientes incluidas en el análisis). Con muestras grandes, aproximadamente el 95 % de estos residuos debería encontrarse en el rango  $(-2, +2)$ .

**SDRESID** (residuos corregidos estudentizados): residuos corregidos divididos por su desviación típica. Útiles también para detectar puntos de influencia.

Algunas de estas variables permiten identificar *puntos de influencia* (se estudian más adelante), pero hay, entre otras, dos variables cuyo diagrama de dispersión informa sobre el supuesto de homocedasticidad o igualdad de varianzas: **ZPRED** y **ZRESID**. El supuesto de igualdad de varianzas implica que la variación de los residuos debe ser uniforme en todo el rango de valores pronosticados. O, lo que es lo mismo, que el tamaño de los residuos es independiente del tamaño de los pronósticos, de donde se desprende que el diagrama de dispersión de los pronósticos tipificados (**ZPRED**) y los residuos tipificados (**ZRESID**) no debe mostrar ninguna pauta de asociación.

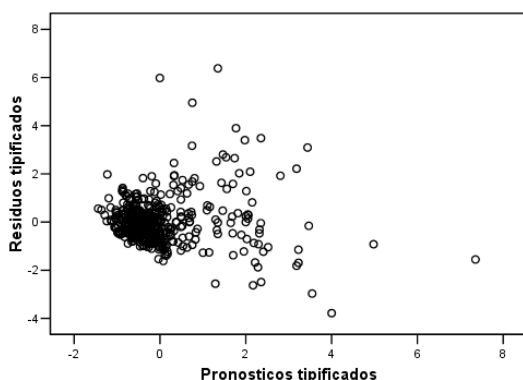
Para obtener un diagrama de dispersión con los pronósticos tipificados (ZPRED) y los residuos tipificados (ZRESID):

- Trasladar la variable ZRESID al cuadro Y: del recuadro Dispersión 1 de 1.
- Trasladar la variable ZPRED al cuadro X: del recuadro Dispersión 1 de 1.

Aceptando estas elecciones, el *Visor* de resultados ofrece el diagrama de dispersión que muestra la Figura 18.8.

En él puede observarse, por un lado, que los residuos y los pronósticos parecen ser independientes, pues la nube de puntos no sigue ninguna pauta de asociación clara, ni lineal ni de otro tipo. Sin embargo, no parece claro que las varianzas sean homogéneas. Más bien parece que conforme va aumentando el valor de los pronósticos también lo va haciendo la dispersión de los residuos: los pronósticos menores que la media (los que en el diagrama tienen puntuación típica por debajo de cero) parecen estar algo más concentrados que los pronósticos mayores que la media (los que en el diagrama tienen puntuación típica mayor que cero). Es relativamente frecuente encontrar esta pauta de comportamiento en los residuos.

Figura 18.8. Diagrama de dispersión: *pronósticos tipificados* por *residuos tipificados*



Cuando un diagrama de dispersión delata la presencia de varianzas heterogéneas, puede utilizarse una transformación de la variable dependiente para resolver el problema (tal como una transformación *logarítmica* o una transformación *raíz cuadrada*). No obstante, al utilizar una transformación de la variable dependiente, no debe descuidarse el problema de interpretación que añade el cambio de escala.

El diagrama de dispersión de las variables ZPRED y ZRESID posee la utilidad adicional de permitir detectar relaciones de tipo no lineal entre las variables. Si la relación es, de hecho, no lineal, el diagrama puede contener indicios sobre otro tipo de función de ajuste: por ejemplo, los residuos tipificados podrían, en lugar de estar homogéneamente dispersos a lo largo del diagrama, seguir un trazado curvilíneo, lo cual estaría reflejando la presencia de una relación cuadrática.

Al margen de esta aproximación visual al problema de la homocedasticidad de los residuos, conviene recordar que el estadístico de *Levene* (disponible en el procedimiento **Explorar**), permite contrastar la hipótesis de igualdad de varianzas.

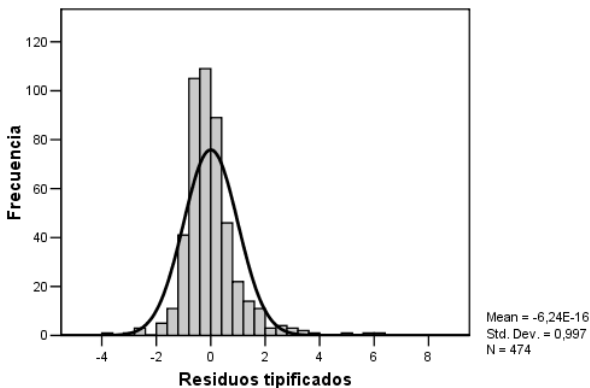
## Normalidad

El recuadro **Gráficos de los residuos tipificados** (ver Figura 18.7) contiene dos opciones gráficas que informan sobre el grado en el que los residuos tipificados se aproximan a una distribución normal:

- " **Histograma.** Esta opción ofrece un histograma de los residuos tipificados con una curva normal superpuesta (ver Figura 18.9). La curva se construye tomando una media de 0 y una desviación típica de 1, es decir, la misma media y la misma desviación típica que los residuos tipificados.

En el histograma de la Figura 18.9 se puede observar, en primer lugar, que la parte central de la distribución acumula muchos más casos de los que acumula una curva normal. En segundo lugar, la distribución es algo asimétrica: en la cola positiva de la distribución existen valores más extremos que en la negativa (esto ocurre cuando uno o varios errores muy grandes, correspondientes por lo general a valores atípicos, son contrarrestados con muchos residuos pequeños de signo opuesto). La distribución de los residuos, por tanto, no parece seguir el modelo de probabilidad normal, de modo que los resultados del análisis deben ser interpretados con cautela.

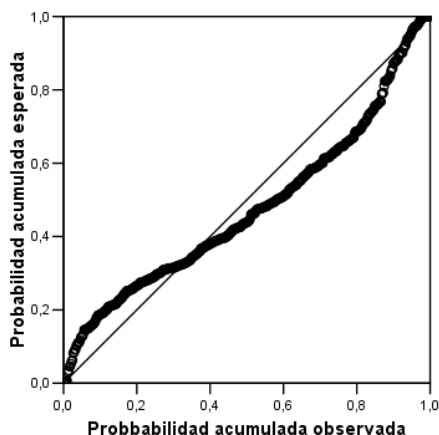
**Figura 18.9.** Histograma de los residuos tipificados



- " **Gráfico de probabilidad normal.** Permite obtener un diagrama de *probabilidad normal*. En este diagrama, en el eje de abscisas está representada la probabilidad acumulada que corresponde a cada residuo tipificado (calculada como la proporción de casos que quedan por debajo de cada residuo tras ordenarlos de menor a mayor); y el eje de ordenadas representa la probabilidad acumulada teórica que corresponde a cada puntuación típica en una curva normal con media 0 y desviación típica 1. Cuando los residuos se distribuyen normalmente, la nube de puntos se encuentra alineada sobre la diagonal del gráfico.

La Figura 18.10 muestra el gráfico de probabilidad normal. La información que ofrece es similar a la ya obtenida con el histograma de la Figura 18.9: puesto que los puntos no se encuentran alineados sobre la diagonal del gráfico, no parece que los residuos se distribuyan normalmente.

Figura 18.10. Gráfico de probabilidad normal de los residuos



Al margen de esta aproximación gráfica al problema de la normalidad de los residuos, conviene recordar que el procedimiento **Explorar** (dentro del menú **Analizar > Estadísticos descriptivos**) contiene estadísticos (*Kolmogorov-Smirnov*, *Shapiro-Wilk*) que permiten contrastar la hipótesis de normalidad.

## Linealidad

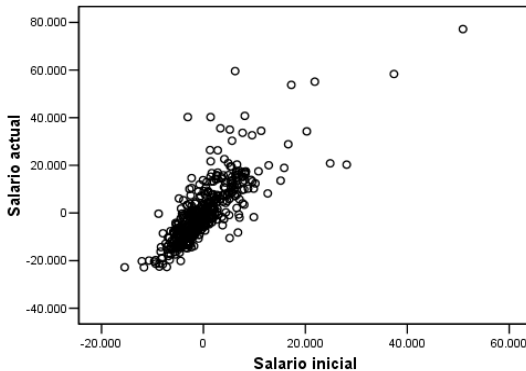
El cuadro de diálogo *Regresión lineal: Gráficos* contiene una opción que permite generar un tipo particular de diagramas de dispersión llamados *diagramas de regresión parcial*. Estos diagramas ayudan a formarse una idea rápida sobre la forma que adopta una relación. En el contexto del análisis de regresión son muy útiles porque permiten examinar la relación existente entre la variable dependiente y cada una de las independientes por separado, tras eliminar de cada relación el efecto atribuible al resto de variables incluidas en el análisis.

Estos diagramas son similares a los de dispersión ya estudiados, pero no se basan en las puntuaciones originales de las dos variables representadas, sino en los residuos obtenidos al llevar a cabo un análisis de regresión con el resto de las variables independientes. Por ejemplo, en el diagrama de regresión parcial de *salario actual* y *salario inicial* están representados los residuos que resultan de efectuar un análisis de regresión sobre *salario actual* incluyendo todas las variables independientes excepto *salario inicial*, y los residuos que resultan de efectuar un análisis de regresión sobre *salario inicial* incluyendo el resto de variables independientes. La utilidad de estos diagramas está en que, puesto que se controla el efecto del resto de las variables, muestran la relación *neta* entre las variables representadas. Además, las rectas que mejor se ajustan a la nube de puntos de estos diagramas son las definidas por los correspondientes coeficientes de regresión (es justamente en esa nube de puntos en la que se basan los coeficientes de regresión parcial). Para obtener estos diagramas de regresión parcial:

- En el subcuadro de diálogo *Regresión lineal: Gráficos* (ver Figura 18.7), marcar la opción **Generar todos los gráficos parciales**.

Esta opción genera tantos gráficos parciales como variables independientes incluya el análisis. En el ejemplo, puesto que el análisis incluye tres variables independientes, se obtienen tres de estos gráficos. La Figura 18.11 muestra uno de ellos. Puede observarse que la relación entre *salario inicial* (una de las variables independientes) y *salario actual* (la variable dependiente), tras eliminar el efecto del resto de variables independientes, es claramente lineal y positiva.

Figura 18.11. Diagrama de dispersión: regresión parcial de *salario actual* sobre *salario inicial*



Los diagramas de regresión parcial permiten formarse una rápida idea sobre el tamaño y el signo de los coeficientes de regresión parcial (los coeficientes de la ecuación de regresión). En estos diagramas, los valores extremos pueden resultar muy informativos.

## Colinealidad

Existe colinealidad perfecta cuando una de las variables independientes se relaciona de forma perfectamente lineal con una o más del resto de variables independientes de la ecuación. Esto ocurre, por ejemplo, cuando se utilizan como variables independientes en la misma ecuación las puntuaciones de las subescalas de un test y la puntuación total en el test (que es la suma de las subescalas y, por tanto, una combinación lineal perfecta de las mismas). Se habla de colinealidad parcial o, simplemente, colinealidad, cuando entre las variables independientes de una ecuación existen correlaciones altas. Se puede dar, por ejemplo, en una investigación de mercados al tomar registros de muchos atributos de un mismo producto; o al utilizar muchos indicadores económicos para construir una ecuación de regresión. En términos generales, cuantas más variables hay en una ecuación, más fácil es que exista colinealidad (aunque, en principio, bastan dos variables).

La colinealidad es un problema porque, en el caso de colinealidad perfecta, no es posible estimar los coeficientes de la ecuación de regresión; y en el caso de colinealidad parcial, aumenta el tamaño de los residuos tipificados y esto produce coeficientes de regresión muy inestables: pequeños cambios en los datos (añadir o quitar un caso, por ejemplo) produce cambios muy grandes en los coeficientes de regresión. Esta es una de las razones por las que es posible encontrar coeficientes con signo cambiado: correlaciones positivas pueden transformarse en coeficientes de regresión negativos (incluso significativamente negativos). Curiosamente, la

medida de ajuste  $R^2$  no se altera por la presencia de colinealidad; pero los efectos atribuidos a las variables independientes pueden ser engañosos.

Al evaluar la existencia o no de colinealidad, la dificultad estriba precisamente en determinar cuál es el grado máximo de relación permisible entre las variables independientes. No existe un consenso generalizado sobre esta cuestión, pero puede servir de guía la presencia de ciertos indicios que pueden encontrarse en los resultados de un análisis de regresión (estos indicios, no obstante, pueden tener su origen en otras causas):

- El estadístico  $F$  que evalúa el ajuste general de la ecuación de regresión es significativo, pero no lo es ninguno de los coeficientes de regresión parcial.
- Los coeficientes de regresión parcial tipificados (los coeficientes *beta*) están inflados tanto en positivo como en negativo (adoptan, al mismo tiempo, valores mayores que 1 y menores que -1).
- Existen valores de tolerancia pequeños (próximos a 0,01). La tolerancia de una variable independiente es la proporción de varianza de esa variable que no está asociada (que no depende) del resto de variables independientes incluidas en la ecuación. Una variable con una tolerancia de, por ejemplo, 0,01 (muy poca tolerancia) es una variable que comparte el 99 % de su varianza con el resto de variables independientes, lo cual significa que se trata de una variable redundante casi por completo.
- Los coeficientes de correlación estimados son muy grandes (por encima de 0,90 en valor absoluto).

Las afirmaciones del tipo *inflados*, *próximos a cero*, *muy grandes*, etc., se deben al hecho de que no existe un criterio estadístico formal en el que basar las decisiones. Sólo existen recomendaciones basadas en trabajos de simulación.

Al margen de estos indicios, el SPSS ofrece la posibilidad de obtener algunos estadísticos que pueden ayudar a diagnosticar la presencia de colinealidad. Se trata de estadísticos orientativos que, aunque pueden ayudar a determinar si existe mayor o menor grado de colinealidad, no permiten tomar una decisión clara sobre si el grado de colinealidad existente constituye un problema. Para obtener estos estadísticos:

- En el subcuadro de diálogo *Regresión lineal: Estadísticos* (ver Figura 18.6.bis), seleccionar la opción **Diagnósticos de colinealidad**.

Esta opción permite obtener los estadísticos de colinealidad que recogen las Tablas 18.15 y 18.16. La Tabla 18.15 es la tabla de *coeficientes de regresión parcial* ya vista, pero ahora contiene información adicional sobre los niveles de tolerancia y sus inversos (FIV).

El nivel de *tolerancia* de una variable se obtiene restando a 1 el coeficiente de determinación resultante de regresar esa variable sobre el resto de variables independientes ( $1 - R_x^2$ ). Valores de tolerancia muy pequeños indican que esa variable puede ser explicada por una combinación lineal del resto de variables, lo cual significa que existe colinealidad.

Los *factores de inflación de la varianza* (FIV) son los inversos de los niveles de tolerancia ( $FIV = 1/(1 - R_x^2)$ ). Reciben ese nombre porque se utilizan para calcular las varianzas de los coeficientes de regresión. Cuanto mayor es el FIV de una variable, mayor es la varianza del correspondiente coeficiente de regresión. De ahí que uno de los problemas de la presencia de colinealidad (tolerancias pequeñas, valores FIV grandes) sea la inestabilidad de las estimaciones de los coeficientes de regresión. Los valores mayores que 10 se consideran grandes.

Tabla 18.15. Coeficientes de regresión parcial y niveles de tolerancia

Modelo: 1

	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Estadísticos de colinealidad	
	B	Error típ.	Beta			Tolerancia	FIV
(Constante)	-3.661,52	1.935,49		-1,89	,059		
Salario inicial	1,75	,06	,81	29,20	,000	,55	1,80
Experiencia previa	-16,73	3,61	-,10	-4,64	,000	,87	1,15
Nivel educativo	735,96	168,69	,12	4,36	,000	,52	1,92

La Tabla 18.16 muestra la solución resultante de aplicar un análisis de componentes principales a la matriz tipificada no centrada de productos cruzados de las variables independientes. Los *autovalores* informan sobre el número de dimensiones o factores diferentes que subyacen en el conjunto de variables independientes utilizadas. La presencia de varios autovalores próximos a cero indica que las variables independientes están muy relacionadas entre sí (colinealidad).

Los *índices de condición* son la raíz cuadrada del cociente entre el autovalor más grande y cada uno del resto de los autovalores. En condiciones de no-colinealidad, estos índices no deben superar el valor 15. Índices mayores que 15 indican un posible problema. Índices mayores que 30 delatan un serio problema de colinealidad.

Las *proporciones de varianza* recogen la proporción de varianza de cada coeficiente de regresión parcial que está explicada por cada dimensión o factor. En condiciones de no-colinealidad, cada dimensión suele explicar gran cantidad de varianza de un solo coeficiente (excepto en lo que se refiere al coeficiente  $B_0$  o *constante*, que siempre aparece asociado a uno de los otros coeficientes; en el ejemplo, el término constante aparece asociado al coeficiente de *Nivel educativo*). La colinealidad es un problema cuando una dimensión o factor con un *índice de condición* alto, contribuye a explicar gran cantidad de la varianza de los coeficientes de dos o más variables.

Si se detecta la presencia de colinealidad en un conjunto de datos, hay que aplicar algún tipo de remedio. Por ejemplo: aumentar el tamaño de la muestra (esta solución puede resultar útil si existen pocos casos en relación con el número de variables); crear indicadores múltiples combinando variables (por ejemplo, promediando variables; o efectuando un análisis de componentes principales para reducir las variables a un conjunto de componentes independientes, y utilizando esos componentes en el análisis de regresión); excluir variables redundantes (excluir variables que correlacionan muy alto con otras, dejando únicamente las que se consideran más importantes); utilizar una técnica de estimación sesgada, tal como la regresión *rige*.

Tabla 18.16. Diagnósticos de colinealidad

Modelo: 1

Dimensión	Autovalor	Índice de condición	Proporciones de la varianza			
			(Constante)	Salario inicial	Experiencia previa	Nivel educativo
1	3,40	1,00	,00	,01	,02	,00
2	,49	2,64	,00	,01	,79	,00
3	,10	5,93	,11	,62	,01	,01
4	,01	15,89	,88	,36	,18	,98



## Puntos de influencia

Todos los casos contribuyen a la obtención de la ecuación de regresión, pero no todos lo hacen con la misma fuerza. Los puntos de influencia son casos que afectan de forma importante al valor de la ecuación de regresión. La presencia de puntos de influencia no tiene por qué constituir un problema en regresión: de hecho, lo normal es que en un análisis de regresión no todos los casos tengan la misma importancia (desde el punto de vista estadístico). Sin embargo, el analista debe ser consciente de la presencia de tales puntos pues, entre otras cosas, podría tratarse de casos con valores erróneos. Sólo siendo conscientes de si existen o no puntos de influencia es posible corregir el análisis.

El procedimiento **Regresión lineal** ofrece varias medidas para detectar la presencia de puntos de influencia. Para obtenerlas:

- Pulsar el botón **Guardar...** del cuadro de diálogo principal (ver Figura 18.4) para acceder al subcuadro de diálogo *Regresión lineal: Guardar nuevas variables* que muestra la Figura 18.12.
- Marcar todas las opciones de los recuadros **Distancias** y **Estadísticos de influencia** (todas estas opciones crean variables nuevas en el archivo de datos).

Figura 18.12. Subcuadro de diálogo *Regresión lineal: Guardar nuevas variables*

**Regresión lineal: Guardar nuevas variables**

**Valores pronosticados**

- ☐ No tipificados
- ☐ Tipificados
- ☐ Corregidos
- ☐ E.T. del pronóstico promedio

**Residuos**

- ☐ No tipificados
- ☐ Tipificados
- ☐ Estudentizados
- ☐ Eliminados
- ☐ Eliminados estudentizados

**Distancias**

- ☐ Mahalanobis
- ☐ De Cook
- ☐ Valores de influencia

**Estadísticos de influencia**

- ☐ DfBetas
- ☐ DfBetas tipificadas
- ☐ DfAjuste
- ☐ DfAjuste tipificado
- ☐ Razón entre covarianzas

**Intervalos de pronóstico**

- ☐ Media ☐ Individuos
- Intervalo de confianza: 95 %

**Guardar en archivo nuevo**

- ☐ Estadísticos de los coeficientes: Archivo...

**Exportar información del modelo al archivo XML**

Examinar

Continuar  
Cancelar  
Ayuda

**Distancias.** Este recuadro recoge tres medidas que expresan el grado en que cada caso se aleja de los demás:

- " **Mahalanobis.** La distancia de Mahalanobis (1936) mide el grado de distanciamiento de cada caso respecto de los promedios del conjunto de variables independientes. En regresión simple, esta distancia se obtiene simplemente elevando al cuadrado la puntuación típica de cada caso en la variable independiente. En regresión múltiple se obtiene multiplicando por  $n-1$  el valor de influencia de cada caso (ver más abajo).
- " **Cook.** La distancia de Cook (1977, 1979) mide el cambio que se produce en las estimaciones de los coeficientes de regresión al ir eliminando cada caso de la ecuación de regresión (se obtiene a partir de la suma al cuadrado de los cambios que se producen en los pronósticos al ir eliminando casos de la ecuación). Una distancia de Cook grande indica que ese caso tiene un peso considerable en la ecuación. Para evaluar estas distancias puede utilizarse la distribución  $F$  con  $p+1$  y  $n-p-1$  grados de libertad ( $p$  se refiere al número de variables independientes y  $n$  al tamaño de la muestra). En general, un caso con una distancia de Cook superior a 1 debe ser revisado.
- " **Valores de influencia.** Representan una medida de la influencia potencial de cada caso. Referido a las variables independientes, un valor de influencia es una medida normalizada del grado de distanciamiento de un punto respecto del centro de su distribución. Los puntos muy alejados pueden influir de forma muy importante en la ecuación de regresión (aunque no necesariamente tienen por qué hacerlo).  
Los valores de influencia oscilan entre  $1/n$  y  $(n-1)/n$ , con media  $p/n$  ( $p$  se refiere al número de variables independientes y  $n$  al tamaño de la muestra). Y pueden interpretarse utilizando la siguiente regla general: los valores menores que 0,2 son poco problemáticos; los valores comprendidos entre 0,2 y 0,5 son arriesgados; y los valores mayores que 0,5 deberían evitarse. Con más de 6 variables y al menos 20 casos, se considera que un valor de influencia debe ser revisado si es mayor que  $2p/n$ .

**Estadísticos de influencia.** Este recuadro contiene varios estadísticos que contribuyen a precisar la posible presencia de puntos de influencia (ver Belsley, Kuh y Welsch, 1980):

- " **DfBetas** (diferencia en las betas). Mide el cambio que se produce en los coeficientes de regresión tipificados (betas) como consecuencia de ir eliminando cada caso de la ecuación de regresión. El SPSS crea en el *Editor de datos* tantas variables nuevas como coeficientes beta tiene la ecuación de regresión (es decir, tantos como variables independientes más uno, el correspondiente al término constante).
- " **DfBetas tipificadas.** Es el cociente entre  $DfBetas$  (párrafo anterior) y su error típico. Generalmente, un valor mayor que  $2/\sqrt{n}$  delata la presencia de un posible punto de influencia. El SPSS crea en el *Editor de datos* tantas variables nuevas como coeficientes beta tiene la ecuación de regresión.
- " **DfAjuste** (diferencia en el ajuste). Mide el cambio que se produce en el pronóstico de un caso cuando ese caso es eliminado de la ecuación de regresión.
- " **DfAjuste tipificado.** Es el cociente entre  $DfAjuste$  (párrafo anterior) y su error típico. En general, se consideran puntos de influencia los casos en los que  $DfAjuste$  tipificado es mayor que  $2/\sqrt{(p/n)}$ , siendo  $p$  el número de variables independientes y  $n$  el tamaño de la muestra.
- " **Razón entre las covarianzas (RV).** Indica en qué medida la matriz de productos cruzados (base del análisis de regresión) cambia con la eliminación de cada caso. Con mues-

tras grandes, se considera que un caso es un punto de influencia si le corresponde un valor RV mayor que  $1+3p/n$  o menor que  $1-3p/n$ .

Además de crear las variables correspondientes a cada una de estas opciones, el SPSS ofrece una tabla resumen (ver Tabla 18.17) que incluye, para todos los estadísticos del recuadro Distancias (ver Figura 18.12), el valor mínimo, el máximo, la media, la desviación típica insesgada y el número de casos válidos. La tabla también recoge información sobre los pronósticos y los residuos.

Tabla 18.17. Estadísticos sobre los residuos

	Mínimo	Máximo	Media	Desviación típ.	N
Valor pronosticado	12.382,90	146.851,63	34.419,57	15.287,30	474
Valor pronosticado tipificado	-1,44	7,35	,00	1,00	474
Error típico del valor pronosticado	375,76	3.186,70	645,15	274,71	474
Valor pronosticado corregido	12.275,05	149.354,23	34.425,49	15.356,14	474
Residuo bruto	-28.852,99	48.701,20	,00	7.607,68	474
Residuo tip.	-3,78	6,38	,00	1,00	474
Residuo estud.	-3,90	6,40	,00	1,00	474
Residuo eliminado	-30.675,46	48.999,29	-5,92	7.704,67	474
Residuo eliminado estudentizado	-3,96	6,69	,00	1,01	474
Dist. de Mahalanobis	,15	81,47	2,99	5,17	474
Distancia de Cook	,00	,24	,00	,02	474
Valor de influencia centrado	,00	,17	,01	,01	474

Conviene señalar que los puntos de influencia no tienen por qué tener asociados residuos particularmente grandes; de hecho, un punto de influencia no sólo puede provocar una pérdida de ajuste, sino que puede hacer que el ajuste global mejore sustancialmente (por ejemplo, cuando todos los puntos están agrupados en una esquina del diagrama y un punto se encuentra muy alejado de ellos en la esquina opuesta). Por tanto, el problema que plantean los puntos de influencia no es precisamente de falta de ajuste. No obstante, es muy aconsejable examinarlos por su desproporcionada influencia sobre la ecuación de regresión. Puesto que estos puntos son distintos de los de demás, conviene precisar en qué son distintos.

Una vez identificados y examinados, podrían ser eliminados del análisis simplemente porque entorpecen el ajuste, o porque su presencia está haciendo obtener medidas de ajuste infladas. También se podrían eliminar los casos muy atípicos simplemente argumentando que el objetivo del análisis es construir una ecuación para entender lo que ocurre con los casos típicos, no con los atípicos. Este argumento es más convincente si los casos atípicos representan a una subpoblación especial que se sale del rango de variación normal. Por otro lado, si existe un conjunto de casos que parece formar un subgrupo separado del resto, podría considerarse la posibilidad de incorporar este hecho al modelo de regresión mediante una variable *dummy* (con unos y ceros para diferenciar ambos subgrupos) o desarrollando diferentes ecuaciones de regresión para los diferentes subgrupos.

Entre los expertos estadísticos no existe un acuerdo completo sobre la conveniencia o no de eliminar un determinado caso. No existe, por tanto una regla en la que basar estas decisiones. Pero al usuario puede ayudarle a decidir sobre este particular el pensar que, cuando se decide eliminar un caso, tal acción debe ser justificada ante quien pregunte por las razones que han llevado a eliminarlo.

## Análisis de regresión por pasos (regresión *stepwise*)

En los apartados previos se ha descrito un método de regresión en el que el control sobre las variables utilizadas para construir el modelo de regresión recae sobre el propio analista. Es el analista quien *decide* qué variables independientes desea incluir en la ecuación de regresión trasladándolas a la lista **Independientes**.

Sin embargo, no es infrecuente encontrarse con situaciones en las que, existiendo un elevado número de posibles variables independientes, no existe una teoría o un trabajo previo que oriente al analista en la elección de las variables relevantes. Este tipo de situaciones pueden afrontarse utilizando procedimientos diseñados para seleccionar, entre una gran cantidad de variables, sólo un conjunto reducido de las mismas: aquellas que permiten obtener el mejor ajuste posible. Con estos procedimientos de selección, el control sobre las variables que han de formar parte de la ecuación de regresión pasa de las manos o el criterio del analista a una regla de decisión basada en criterios estadísticos.

### Criterios de selección de variables

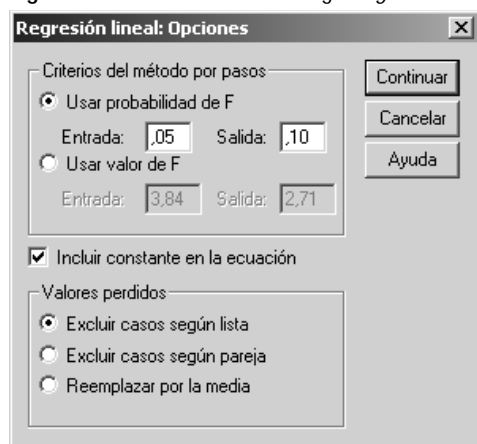
Existen diferentes criterios estadísticos para seleccionar variables en un modelo de regresión. Algunos de estos criterios son: el valor del coeficiente de correlación múltiple  $R^2$  (corregido o sin corregir), el valor del coeficiente de correlación parcial entre cada variable independiente y la dependiente, el grado de reducción que se obtiene en el error típico de los residuos al incorporar una variable, etc. De una u otra forma, todos ellos coinciden en intentar maximizar el ajuste del modelo de regresión utilizando el mínimo número posible de variables. Los métodos por pasos que incluye el SPSS (ver siguiente apartado) basan la selección de variables en dos criterios estadísticos:

**1. Criterio de significación.** De acuerdo con este criterio, sólo se incorporan al modelo de regresión aquellas variables que contribuyen al ajuste del modelo de forma significativa. La contribución individual de una variable al ajuste del modelo se establece contrastando, a partir del coeficiente de correlación parcial, la hipótesis de independencia entre esa variable y la variable dependiente. Para decidir si se mantiene o rechaza esa hipótesis de independencia, el SPSS incluye dos criterios de significación:

- *Probabilidad de F.* Una variable pasa a formar parte del modelo de regresión si el nivel crítico asociado a su coeficiente de correlación parcial al contrastar la hipótesis de independencia es menor que 0,05 (probabilidad de *entrada*). Y queda fuera del modelo de regresión si ese nivel crítico es mayor que 0,10 (probabilidad de *salida*).
- *Valor de F.* Una variable pasa a formar parte del modelo de regresión si el valor del estadístico  $F$  utilizado para contrastar la hipótesis de independencia es mayor que 3,84 (valor de *entrada*). Y queda fuera del modelo si el valor del estadístico  $F$  es menor que 2,71 (valor de *salida*).

Para elegir entre estos criterios de significación:

- Pulsar el botón **Opciones...** del cuadro de diálogo principal (ver Figura 18.4) para acceder al subcuadro de diálogo *Regresión lineal: Opciones* que muestra la Figura 18.13.

Figura 18.13. Subcuadro de diálogo *Regresión lineal: Opciones*

Las opciones del recuadro **Criterios del método por pasos** permiten seleccionar uno de los criterios de significación disponibles y modificar las probabilidades de entrada y salida.

2. **Criterio de tolerancia.** Superado el criterio de *significación*, una variable sólo pasa a formar parte del modelo si su nivel de tolerancia es mayor que el nivel establecido por defecto (este nivel es 0,0001, pero puede cambiarse mediante sintaxis) y si, además, aun correspondiéndole un coeficiente de correlación parcial significativamente distinto de cero, su incorporación al modelo no hace que alguna de las variables previamente seleccionadas pase a tener un nivel de tolerancia por debajo del nivel establecido por defecto (aunque esto último depende también del *método de selección de variables* elegido; ver siguiente apartado). El concepto de tolerancia tiene que ver con la parte que una variable tiene de propia, de diferente de las demás, y ya se ha explicado en el apartado *Supuestos del modelo de regresión: Colinealidad*.

Una forma intuitiva de comprender y valorar el efecto resultante de aplicar estos criterios de selección consiste en observar el cambio que se va produciendo en el coeficiente de determinación  $R^2$  a medida que se van incorporando (o eliminando) variables al modelo. Este cambio se define como  $R^2_{\text{cambio}} = R^2 - R_i^2$ , donde  $R_i^2$  se refiere al coeficiente de determinación obtenido con todas las variables independientes excepto la  $i$ -ésima ( $R^2_{\text{cambio}}$  es, en realidad, el cuadrado del coeficiente de correlación semiparcial entre la variable dependiente y la variable incluida o eliminada). Un cambio grande en  $R^2$  indica que esa variable contribuye de forma importante al ajuste del modelo. Para obtener los valores de  $R^2_{\text{cambio}}$  y su significación (grado en que el cambio observado en  $R^2$  difiere de cero):

- Marcar la opción **Cambio en R cuadrado** del cuadro de diálogo *Regresión lineal: Estadísticos* (ver Figura 18.6).

Según se verá enseguida, esta opción permite obtener: el valor de  $R^2_{\text{cambio}}$  que se va produciendo con la incorporación de cada nueva variable independiente, el valor del estadístico  $F$  resultante de contrastar la hipótesis de que el valor poblacional de  $R^2_{\text{cambio}}$  es cero y, por último, el nivel crítico asociado a ese estadístico  $F$ .

## Métodos de selección de variables

Existen diferentes métodos para seleccionar las variables independientes que debe incluir un modelo de regresión, pero los que mayor aceptación han recibido (sin que esto signifique que son los mejores) son los métodos de selección por pasos (*stepwise*). Con estos métodos, se selecciona en primer lugar la *mejor* variable (siempre de acuerdo con algún criterio estadístico); a continuación, la mejor de las restantes; y así sucesivamente hasta que ya no quedan variables que cumplan los criterios de selección.

El procedimiento **Regresión lineal** del SPSS incluye varios de estos métodos de selección de variables. Todos ellos se encuentran disponibles en el botón de menú desplegable de la opción **Método** del cuadro de diálogo *Regresión lineal* (ver Figura 18.4). Dos de estos métodos permiten incluir o excluir, en un solo paso, todas las variables independientes seleccionadas (no son, por tanto, métodos de selección por pasos):

- ▼ **Introducir.** Este método construye la ecuación de regresión utilizando todas las variables seleccionadas en la lista **Independientes** (ver Figura 18.4). Es el método utilizado por defecto.
- ▼ **Eliminar.** Elimina en un solo paso todas las variables de la lista **Independientes** y ofrece los coeficientes de regresión que corresponderían a cada variable independiente en el caso de que se utilizara cada una de ellas individualmente para construir la ecuación de regresión.

El resto de métodos de selección de variables son métodos por pasos, es decir, métodos que van incorporando o eliminando variables paso a paso dependiendo de que éstas cumplan o no los criterios de selección:

- ▼ **Hacia adelante.** Las variables se incorporan al modelo de regresión una a una. En el primer paso se selecciona la variable independiente que, además de superar los criterios de *entrada*, más alto correlaciona (en valor absoluto) con la dependiente. En los siguientes pasos se utiliza como criterio de selección el coeficiente de correlación parcial: van siendo seleccionadas una a una las variables que, además de superar los criterios de *entrada*, poseen el coeficiente de correlación parcial más alto en valor absoluto. La relación entre la variable dependiente y cada una de las variables independientes se parcializa controlando el efecto de la(s) variable(s) independiente(s) previamente seleccionada(s).

La selección de variables se detiene cuando no quedan variables que superen el criterio de *entrada*. Utilizar como criterio de *entrada* el tamaño, en valor absoluto, del coeficiente de correlación parcial, es equivalente a seleccionar la variable con menor *probabilidad de F* o mayor *valor de F*.

- ▼ **Hacia atrás.** Este método comienza incluyendo en el modelo todas las variables seleccionadas en la lista **Independientes** (ver Figura 18.4) y luego procede a eliminarlas una a una. La primera variable eliminada es aquella que, además de cumplir los criterios de *salida*, lleva asociado el menor cambio (disminución) en  $R^2$ . En cada paso sucesivo se van eliminando una a una las variables con coeficientes de regresión no significativos y siempre en orden inverso al tamaño del cambio en  $R^2$  asociado a la eliminación de cada variable.

La eliminación de variables se detiene cuando no quedan variables en el modelo que cumplan los criterios de salida.

- ▼ **Pasos sucesivos.** Este método es una especie de mezcla de los métodos *hacia adelante* y *hacia atrás*. Comienza, al igual que el método *hacia adelante*, seleccionando, en el primer paso, la variable independiente que, además de superar los criterios de *entrada*, más alto correlaciona (en valor absoluto) con la variable dependiente. A continuación, selecciona la variable independiente que, además de superar los criterios de *entrada*, posee el coeficiente de correlación parcial más alto (en valor absoluto). Cada vez que se incorpora una nueva variable al modelo, las variables previamente seleccionadas son, al igual que en el método *hacia atrás*, evaluadas nuevamente para determinar si siguen cumpliendo o no los criterios de *salida*. Si alguna variable seleccionada cumple los criterios de salida, es expulsada del modelo.

El proceso se detiene cuando no quedan variables que superen los criterios de *entrada* y las variables seleccionadas no cumplen los criterios de *salida*.

## Regresión por pasos

Para ilustrar el funcionamiento del análisis de regresión por pasos se presenta a continuación un ejemplo con el método *pasos sucesivos*. Se utiliza el salario actual (*salario*) como variable dependiente y, como variables independientes, la fecha de nacimiento (*fechnac*), el nivel educativo (*educ*), el salario inicial (*salini*), la experiencia previa (*expprev*) y la clasificación de minorías (*minoría*). El objetivo del análisis es encontrar un modelo de regresión que explique, con el mínimo número posible de variables independientes, la mayor cantidad posible de la varianza de la variable *salario*. Para llevar a cabo el análisis:

- En el cuadro de diálogo principal (ver Figura 18.4), seleccionar la variable *salario* y trasladarla al cuadro **Dependiente**.
- Seleccionar las variables *fechnac*, *educ*, *salini*, *expprev* y *minoría*, y trasladarlas a la lista **Independientes**.
- Pulsar el botón de menú desplegable del cuadro **Método** y seleccionar la opción **Pasos sucesivos**.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Regresión lineal: Estadísticos* (ver Figura 18.6) y marcar la opción **Cambio en  $R$  cuadrado**.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 18.18 a la 18.22.

La Tabla 18.18 ofrece un resumen del modelo final al que se ha llegado. En la columna *Modelo* se indica en número de pasos dados para construir el modelo de regresión: tres pasos. En el primer paso se ha seleccionado la variable *salario inicial*, en el segundo, *experiencia previa* y, en el tercero, *nivel educativo*. También se indica si en alguno de los pasos se ha eliminado alguna variable previamente seleccionada; en este ejemplo no se han eliminado variables. Por último, se informa sobre el método de selección aplicado (*Por pasos*) y sobre los criterios de *entrada* y *salida* utilizados: una variable es incorporada al modelo si su coeficiente de regresión parcial es significativamente distinto de cero con un nivel de significación del 5 % y, una vez seleccionada, sólo es eliminada del modelo si con la incorporación de otra u otras variables en un paso posterior su coeficiente de regresión parcial deja de ser distinto de cero con un nivel de significación del 10 %.

Tabla 18.18. Variables introducidas/eliminadas

Modelo	Variables introducidas	Variables eliminadas	Método
1	Salario inicial	.	Por pasos (criterio: Prob. de F para entrar $\leq$ .050, Prob. de F para salir $\geq$ .100).
2	Experiencia previa	.	Por pasos (criterio: Prob. de F para entrar $\leq$ .050, Prob. de F para salir $\geq$ .100).
3	Nivel educativo	.	Por pasos (criterio: Prob. de F para entrar $\leq$ .050, Prob. de F para salir $\geq$ .100).

La Tabla 18.19 recoge, para cada paso, el valor de  $R^2$ , el cambio experimentado por  $R^2$ , y el estadístico  $F$  y su significación. El estadístico  $F$  permite contrastar la hipótesis de que el cambio en  $R^2$  vale cero en la población. Al incorporar la primera variable (*Modelo 1*), el valor de  $R^2$  es 0,775. Lógicamente, en el primer paso,  $R^2_{\text{cambio}} = R^2$ . Al contrastar la hipótesis de que el valor poblacional de  $R^2_{\text{cambio}}$  es cero se obtiene un estadístico  $F$  de 1.620,83 que, con 1 y 471 grados de libertad, tiene asociado un nivel crítico  $\text{Sig.} < 0,0005$ . Puesto que este valor es menor que 0,05, puede afirmarse que la proporción de varianza explicada por la variable *salario inicial* (variable seleccionada en el primer paso) es significativamente distinta de cero.

En el segundo paso (*Modelo 2*), el valor de  $R^2$  aumenta hasta 0,794. Esto supone un cambio de 0,019 (aproximadamente un 2 %). La tabla muestra el valor del estadístico  $F$  (43,18) obtenido al contrastar la hipótesis de que el valor poblacional de  $R^2_{\text{cambio}}$  es cero, y su significación ( $\text{Sig.} < 0,0005$ ). Aunque se trata de un incremento muy pequeño (un 2 %), el valor del nivel crítico permite afirmar que la variable *experiencia previa* (la variable incorporada al modelo en el segundo paso) contribuye significativamente a explicar lo que ocurre con la variable dependiente; o, lo que es lo mismo, a mejorar el ajuste.

En el tercer y último paso (*Modelo 3*),  $R^2$  toma un valor de 0,802, lo cual supone un incremento de 0,008 (aproximadamente un 1 por ciento). De nuevo se trata de un incremento muy pequeño, pero al evaluar su significación se obtiene un estadístico  $F$  de 19,29 y un nivel crítico asociado menor que 0,0005, lo cual está indicando que la variable *nivel educativo* (la variable incorporada en el tercer paso), también contribuye de forma significativa a explicar el comportamiento de la variable dependiente (o, lo que es lo mismo, a mejorar el ajuste). Las tres variables seleccionadas en el modelo final consiguen explicar un 80 % ( $R^2 = 0,802$ ) de la variabilidad observada en el *salario actual*.

Tabla 18.19. Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Estadísticos de cambio				
					Cambio en R cuadrado	F del cambio	gl1	gl2	Sig.
1	,880 <sup>a</sup>	,775	,774	8.119,79	,775	1.620,83	1	471	,000
2	,891 <sup>b</sup>	,794	,793	7.778,94	,019	43,18	1	470	,000
3	,896 <sup>c</sup>	,802	,801	7.631,84	,008	19,29	1	469	,000

a. Variables predictoras: (Constante), Salario inicial

b. Variables predictoras: (Constante), Salario inicial, Experiencia previa (meses)

c. Variables predictoras: (Constante), Salario inicial, Experiencia previa (meses), Nivel educativo

La tabla *resumen del ANOVA* (Tabla 18.20) muestra el valor del estadístico  $F$  obtenido al contrastar la hipótesis de que el valor poblacional de  $R^2$  en cada paso es cero. Ahora no se evalúa el cambio que se va produciendo en el valor de  $R^2$  de un paso a otro, sino el valor de  $R^2$  en



cada paso. Lógicamente, si  $R^2$  es significativamente distinta de cero en el primer paso, también lo será en los pasos sucesivos.

**Tabla 18.20.** Resumen del ANOVA

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	106.862.706.669,34	1	106.862.706.669,34	1.620,83	,000 <sup>a</sup>
	Residual	31.053.506.813,54	471	65.931.012,34		
	Total	137.916.213.482,88	472			
2	Regresión	109.475.617.434,36	2	54.737.808.717,18	904,58	,000 <sup>b</sup>
	Residual	28.440.596.048,52	470	60.511.906,49		
	Total	137.916.213.482,88	472			
3	Regresión	110.599.332.927,33	3	36.866.444.309,11	632,96	,000 <sup>c</sup>
	Residual	27.316.880.555,54	469	58.244.947,88		
	Total	137.916.213.482,88	472			

a. Variables predictoras: (Constante), Salario inicial

b. Variables predictoras: (Constante), Salario inicial, Experiencia previa

c. Variables predictoras: (Constante), Salario inicial, Experiencia previa, Nivel educativo

La Tabla 18.21 contiene los *coeficientes de regresión parcial* correspondientes a cada una de las variables incluidas en el modelo de regresión; estos coeficientes sirven para construir la ecuación de regresión en cada paso (incluyendo el término constante). Las primeras columnas recogen el valor de los coeficientes de regresión parcial ( $B$ ) y su error típico. A continuación aparecen los coeficientes de regresión parcial tipificados ( $Beta$ ), que ayudan a precisar la importancia relativa de cada variable dentro de la ecuación. Las dos últimas columnas muestran los estadísticos  $t$  y los niveles críticos ( $Sig$ ) obtenidos al contrastar las hipótesis de que los coeficientes de regresión parcial valen cero en la población. Un nivel crítico por debajo de 0,05 indica que la variable contribuye significativamente a mejorar el ajuste del modelo.

Utilizar el estadístico  $t$  para contrastar la hipótesis de que un coeficiente de regresión parcial vale cero es exactamente lo mismo que utilizar el estadístico  $F$  para contrastar la hipótesis de que el valor poblacional del cambio observado en  $R^2$  vale cero. De hecho, elevando al cuadrado los valores del estadístico  $t$  de la Tabla 18.21 se obtienen los valores del estadístico  $F$  de la Tabla 18.19. Ambos estadísticos permiten valorar la significación estadística de la contribución individual de una variable a la proporción de varianza explicada por el conjunto de variables independientes previamente incluidas en el modelo.

**Tabla 18.21.** Coeficientes de regresión parcial

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	1.929,52	889,17		2,17	,031
	Salario inicial	1,91	,05	,88	40,26	,000
2	(Constante)	3.856,96	900,93		4,28	,000
	Salario inicial	1,92	,05	,89	42,28	,000
	Experiencia previa	-22,50	3,42	-,14	-6,57	,000
3	(Constante)	-3.708,90	1.936,04		-1,92	,056
	Salario inicial	1,75	,06	,81	29,19	,000
	Experiencia previa	-16,75	3,61	-,10	-4,65	,000
	Nivel educativo	741,31	168,77	,13	4,39	,000

Por último, la Tabla 18.22 muestra los coeficientes de regresión parcial de las variables no seleccionadas para formar parte de la ecuación de regresión en cada paso. La información que contiene esta tabla permite conocer en detalle por qué unas variables han sido seleccionadas y otras no.

En el primer paso se ha seleccionado la variable *salario inicial* porque es la que más alto correlaciona, en valor absoluto, con la variable dependiente (esta información se encuentra en la Tabla 18.10). En este primer paso, todavía están fuera del modelo el resto de variables independientes. La columna *Beta dentro* contiene el valor que tomaría el coeficiente de regresión parcial tipificado de cada variable en el caso de ser seleccionada en el siguiente paso. Y las dos columnas siguientes (*t* y *Sig.*) informan sobre si el valor que adoptaría el coeficiente de regresión parcial de una variable en el caso de ser incorporada al modelo sería o no significativamente distinto de cero.

Tabla 18.22. Variables excluidas

Modelo		Beta dentro	t	Sig.	Correlación parcial	Estadísticos de colinealidad
						Tolerancia
1	Fecha de nacimiento	,14 <sup>a</sup>	6,47	,000	,29	1,00
	Nivel educativo	,17 <sup>a</sup>	6,39	,000	,28	,60
	Experiencia previa	-,14 <sup>a</sup>	-6,57	,000	-,29	1,00
	Clasificación de minorías	-,04 <sup>a</sup>	-1,81	,071	-,08	,98
2	Fecha de nacimiento	,07 <sup>b</sup>	2,02	,044	,09	,35
	Nivel educativo	,13 <sup>b</sup>	4,39	,000	,20	,52
	Clasificación de minorías	-,02 <sup>b</sup>	-,88	,377	-,04	,95
3	Fecha de nacimiento	,05 <sup>c</sup>	1,56	,120	,07	,35
	Clasificación de minorías	-,02 <sup>c</sup>	-,97	,335	-,04	,95

a. Variables predictoras en el modelo: (Constante), Salario inicial

b. Variables predictoras en el modelo: (Constante), Salario inicial, Experiencia previa

c. Variables predictoras en el modelo: (Constante), Salario inicial, Experiencia previa, Nivel educativo

Puede comprobarse que, en el primer paso, hay tres variables todavía no seleccionadas (*nivel educativo*, *experiencia previa* y *fecha de nacimiento*) cuyos coeficientes de regresión poseen niveles críticos (*Sig.*) por debajo de 0,05 (criterio de *entrada*). Entre ellas, la variable cuyo coeficiente de correlación parcial es el mayor en valor absoluto (*experiencia previa* = -0,290) y, además, posee un nivel de tolerancia por encima de 0,001 (*nivel de tolerancia mínimo* establecido por defecto), es justamente la variable seleccionada en el segundo paso. Es decir, la variable seleccionada en el segundo paso es aquella que, cumpliendo todos los criterios de entrada, posee el coeficiente de correlación parcial más alto.

En el segundo paso todavía quedan fuera de la ecuación dos variables cuyos coeficientes de regresión serían significativos en caso de ser seleccionadas en el siguiente paso: *nivel educativo* y *fecha de nacimiento*. De esas dos variables, se ha seleccionado en el tercer paso la variable *nivel educativo* porque, teniendo un nivel de tolerancia por encima de 0,001, es la que posee el coeficiente de correlación parcial más alto.

Después del tercer paso todavía quedan dos variables fuera de la ecuación: *fecha de nacimiento* y *clasificación étnica*. Pero, puesto que ninguna de las dos supera el criterio de *entrada* (*Sig.* < 0,05), es decir, puesto que a ninguna de ellas le corresponde un coeficiente de regresión parcial significativamente distinto de cero, el proceso se detiene y ambas variables quedan fuera del modelo.

# Qué variables debe incluir la ecuación de regresión

El método de selección por pasos ha llevado a construir una ecuación de regresión con tres variables. Esas tres variables han sido incluidas en el modelo porque poseen coeficientes de regresión parcial significativos. Sin embargo, la primera variable explica el 78 % de la varianza de la variable dependiente, la segunda el 2 %, y la tercera el 2 %. Si en lugar del método *pasos sucesivos* se utiliza el método *introducir*, se obtienen los resultados que muestran las Tablas 18.23 y 18.24.

Por un lado, la ganancia que se obtiene en  $R^2$  utilizando las cinco variables en lugar de las tres seleccionadas con el método por pasos es extremadamente pequeña:  $0,803 - 0,802 = 0,001$  (y el valor de  $R^2$  corregida ni siquiera cambia: vale 0,801 en ambos casos). No parece que tenga mucho sentido añadir dos variables a un modelo para obtener una mejora de una milésima en la proporción de varianza explicada. Aunque es cierto que  $R^2$  nunca disminuye cuando se van incorporando nuevas variables al modelo de regresión, sino que aumenta o se queda como está, esto no significa, necesariamente, que la ecuación con más variables se ajuste mejor a los datos poblacionales. Generalmente, conforme va aumentando la calidad del modelo, va disminuyendo el error típico de los residuos (*Error típ. de la estimación*). Pero el incremento que se va produciendo en  $R^2$  al ir añadiendo variables no se corresponde necesariamente con una disminución del error típico de los residuos. Con cada variable nueva que se incorpora al modelo, la suma de cuadrados de la regresión gana un grado de libertad y la suma de cuadrados de los residuos lo pierde. Por tanto, el error típico de los residuos puede aumentar cuando el descenso de la variación residual es demasiado pequeño para compensar el grado de libertad que pierde la suma de cuadrados de los residuos. Estas consideraciones sugieren la conveniencia de utilizar modelos de regresión parsimoniosos, es decir, modelos con un número reducido de variables independientes (con el mínimo número posible de variables).

Tabla 18.23. Resumen del modelo

Modelo: 1

R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
,896	,803	,801	7.620,33

Por otro lado, las variables que tienen pesos significativos en la ecuación de regresión previamente obtenida con el método *pasos sucesivos* no son las mismas que las que tienen pesos significativos en la ecuación obtenida con el método *introducir* (ver Tablas 18.21 y 18.24). Esta diferencia entre métodos de selección de variables debe ser tenida muy en cuenta. ¿Cuáles son las variables *buenas*?

Atendiendo a criterios puramente estadísticos, la ecuación de regresión con las tres variables seleccionadas por el método *pasos sucesivos*, es la mejor de las posibles con el mínimo número de variables. Pero en la práctica, la decisión sobre cuántas variables debe incluir la ecuación de regresión puede tomarse teniendo en cuenta, además de los criterios estadísticos, otro tipo de consideraciones. Si, por ejemplo, resulta muy costoso (tiempo, dinero, etc.) obtener las unidades de análisis, un modelo con una única variable independiente podría resultar lo bastante apropiado. Si las consecuencias de los residuos de los pronósticos fueran muy graves, debería intentarse minimizar el tamaño de los residuos incluyendo en el modelo las tres

variables del método *pasos sucesivos* o las cuatro con pesos significativos del método *introducir*. Así pues, para decidir con qué modelo de regresión quedarse, casi siempre es conveniente tomar en consideración criterios adicionales a los puramente estadísticos.

Por otro lado, puesto que los métodos de selección por pasos construyen la ecuación de regresión basándose exclusivamente en criterios estadísticos, podría ocurrir que alguna variable realmente relevante desde el punto de vista teórico quedara fuera del modelo de regresión final. Y esto es algo que hay que vigilar con especial cuidado.

Tabla 18.24. Coeficientes de regresión parcial

Modelo: 1

	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
	B	Error típ.	Beta		
(Constante)	-33.438,68	19.113,33		-1,75	,081
Fecha de nacimiento	2,52E-06	,00	,05	1,58	,116
Nivel educativo	712,74	169,67	,12	4,20	,000
Salario inicial	1,74	,06	,80	28,87	,000
Experiencia previa	-9,27	5,72	-,06	-1,62	,106
Clasificación de minorías	-865,29	867,21	-,02	-1,00	,319

Por supuesto, los contrastes estadísticos sirven de apoyo para tomar decisiones. Pero, dado que la potencia de un contraste se incrementa conforme lo hace el tamaño de la muestra, hay que ser muy cautelosos con las conclusiones a las que se llega. Esto significa que, con muestras grandes, efectos muy pequeños desde el punto de vista de su importancia teórica o práctica pueden resultar estadísticamente significativos. Por el contrario, con muestras pequeñas, para que un efecto resulte significativo, debe tratarse de un efecto importante (con muestras pequeñas, existe mayor grado de coincidencia entre la significación estadística y la importancia práctica). Por esta razón, en la determinación de la ecuación de regresión final, debe tenerse en cuenta, cuando se trabaja con muestras grandes, la conveniencia de considerar elementos de decisión adicionales a la pura significación estadística.

Puesto que la utilización de los métodos de selección por pasos está bastante generalizada, conviene también alertar sobre el peligro de alcanzar un resultado falsamente positivo (un error de tipo I). Es decir, si se examina un número de variables lo bastante grande, tarde o temprano una o más pueden resultar significativas sólo por azar. Este riesgo es tanto mayor cuanto más variables se incluyen en el análisis. Para evitar este problema, si la muestra es lo bastante grande, puede dividirse en dos, aplicar el análisis a una mitad y verificar en la otra mitad si se confirma el resultado obtenido. Si la muestra es pequeña, esta solución es inviable y, por tanto, el riesgo de cometer un error de tipo I permanece.

## Cómo efectuar pronósticos

Si el objetivo del análisis de regresión es el de evaluar la capacidad de un conjunto de variables independientes para dar cuenta del comportamiento de una variable dependiente, no es necesario añadir nada más a lo ya estudiado. Sin embargo, si el objetivo principal del análisis es el de poder efectuar pronósticos en casos nuevos, todavía falta saber algunas cosas.

Ya se ha explicado cómo utilizar los coeficientes de regresión parcial ( $B$ ) para construir la ecuación de regresión (ver Tabla 18.6). En el apartado *Regresión múltiple* se ha llegado a la siguiente ecuación de regresión:

$$\text{Pronóstico (salario)} = -3.661,52 + 1,75 \text{ salini} - 16,73 \text{ expprev} + 735,96 \text{ educ}$$

Puesto que se conocen los pesos de la ecuación de regresión, podría utilizarse la opción **Calcular** del menú **Transformar** para obtener los valores que la ecuación pronostica a cada caso. Pero esto es completamente innecesario. El subcuadro de diálogo *Regresión lineal: Guardar nuevas variables* (ver Figura 18.12) contiene opciones que permiten generar diferentes tipos de variables relacionadas con los pronósticos.

**Valores pronosticados.** Las opciones de este recuadro generan, en el *Editor de datos*, cuatro nuevas variables. Estas nuevas variables reciben automáticamente un nombre seguido de un número de serie: *NOMBRE\_#*. Por ejemplo, la primera vez que se solicitan durante una sesión los *pronósticos tipificados*, la nueva variable con los pronósticos tipificados recibe el nombre *ZPR\_1*. Si se vuelven a solicitar los pronósticos tipificados durante la misma sesión, la nueva variable recibe el nombre *ZPR\_2*. Etc.

- " **No tipificados:** pronósticos de la ecuación de regresión en puntuaciones directas. Nombre: *PRE\_#*.
- " **Tipificados:** pronósticos convertidos en puntuaciones típicas (restando a cada pronóstico la media de los pronósticos y dividiendo la diferencia por la desviación típica de los pronósticos). Nombre: *ZPR\_#*.
- " **Corregidos:** pronóstico que corresponde a cada caso cuando la ecuación de regresión se obtiene sin incluir ese caso. Nombre: *ADJ\_#*.
- " **E.T. del pronóstico promedio:** error típico de los pronósticos correspondientes a los casos que tienen el mismo valor en las variables independientes. Nombre: *SEP\_#*.

Al efectuar pronósticos es posible optar entre: (1) efectuar un pronóstico individual  $Y_i'$  para cada caso concreto  $X_i$ , o (2) pronosticar para cada caso la media de los pronósticos ( $Y_0'$ ) correspondientes a todos los casos con el mismo valor  $X_0$  en la(s) variable(s) independiente(s); a esta media es a la que se llama *pronóstico promedio*. En ambos casos se obtiene el mismo pronóstico ( $Y_i' = Y_0'$ ), pero cada tipo de pronóstico (ambos son variables aleatorias) tiene un error típico distinto. La Figura 18.14 puede ayudar a comprender la diferencia entre estos dos errores típicos. Al efectuar un pronóstico individual para un determinado valor de  $X_i$ , el error de estimación o variación residual ( $Y_i - Y_i'$ ) puede contener dos fuentes de error (identificadas en la Figura 18.14 con los números 1 y 2):

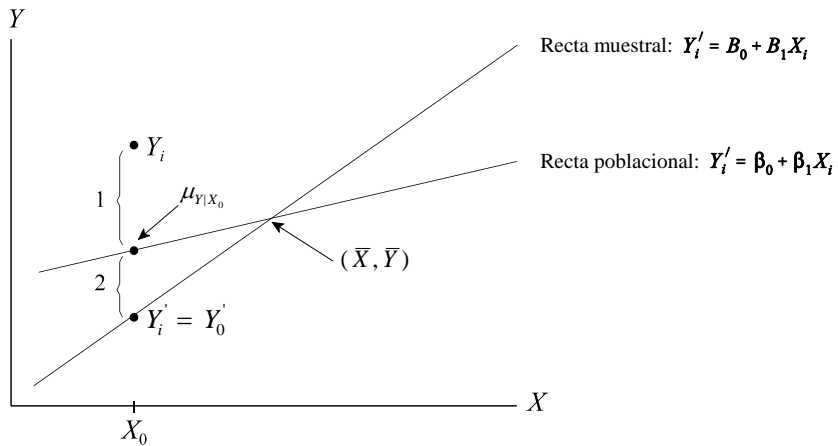
- (1) La diferencia entre el valor observado en la variable dependiente ( $Y_i$ ) y la media poblacional correspondiente a  $X_0$  ( $\mu_{Y|X_0}$ ).
- (2) La diferencia entre el pronóstico para ese caso ( $Y_i'$  o  $Y_0'$ ) y la media poblacional correspondiente a  $X_0$  ( $\mu_{Y|X_0}$ ).

En un pronóstico *individual* entran en juego ambas fuentes de error. Pero en un pronóstico *promedio* sólo entra en juego la segunda fuente de error. Por tanto, para un valor dado de  $X_0$ , el error típico del pronóstico *promedio* siempre será menor o igual que el error típico del pronóstico *individual*. En consecuencia, al construir intervalos

de confianza para los pronósticos, la amplitud del intervalo cambiará dependiendo del error típico que se tome como referencia.

Además, observando la Figura 18.14, puede intuirse fácilmente que los errores típicos del pronóstico promedio (que ya se ha dicho que están basados en las distancias entre  $Y_0'$  y  $\mu_{Y|X_0}$ ) serán tanto menores cuanto más se parezcan  $X_0$  y  $\bar{X}$ , pues cuanto más se parezcan, más cerca estará la recta muestral de la poblacional y, consecuentemente, más cerca estarán  $Y_0'$  y  $\mu_{Y|X_0}$ .

Figura 18.14. Tipos de error en los pronósticos de la regresión



**Intervalos de pronóstico** (parte inferior del subcuadro de diálogo *Regresión lineal: Guardar nuevas variables*; ver Figura 18.12). Las opciones de este recuadro permiten obtener los intervalos de confianza para los pronósticos:

- " **Media.** Intervalo de confianza basado en los errores típicos de los pronósticos promedio.
- " **Individuos.** Intervalo de confianza basado en los errores típicos de los pronósticos individuales.

La opción **Intervalo de confianza** \_\_\_ % permite establecer el nivel de confianza con el que se construyen los intervalos de confianza.

Lógicamente, estos dos intervalos son distintos. Para un valor dado de  $X$ , el primer intervalo (media) es más estrecho que el segundo (individuos). Recuérdese lo dicho en este mismo apartado sobre los errores típicos de los pronósticos. Cada una de estas dos opciones (media e individuos) genera en el *Editor de datos* dos nuevas variables con el límite inferior y superior del intervalo. Estas nuevas variables reciben los siguientes nombres:

- **LMCI\_#:** límite inferior del intervalo de confianza para el pronóstico medio.
- **UMCI\_#:** límite superior del intervalo de confianza para el pronóstico medio.
- **LICI\_#:** límite inferior del intervalo de confianza para el pronóstico individual.
- **UICI\_#:** límite superior del intervalo de confianza para el pronóstico individual.

## Validez del modelo de regresión

El modelo de regresión puede ser validado utilizando casos nuevos. Para ello, basta con obtener los pronósticos para esos casos nuevos y, a continuación, calcular el coeficiente de correlación entre los valores observados en la variable dependiente y los valores pronosticados para esos casos nuevos. En teoría, el coeficiente de correlación así obtenido debería ser igual al coeficiente de correlación múltiple del análisis de regresión ( $R$ ). En la práctica, si el modelo es lo bastante bueno, se observarán pequeñas diferencias entre esos coeficientes atribuibles únicamente al azar muestral. Es muy importante que los nuevos casos representen a las mismas poblaciones que los casos originalmente utilizados para obtener la ecuación de regresión.

En ocasiones, es posible que no se tenga acceso a nuevos datos o que sea muy difícil obtenerlos. En esos casos, todavía es posible validar el modelo de regresión si la muestra es lo bastante grande. Basta con utilizar la mitad de los casos de la muestra (aleatoriamente seleccionados) para obtener la ecuación de regresión y la otra mitad de la muestra para efectuar los pronósticos. Un modelo fiable debería llevar a obtener una correlación similar entre los valores observados y pronosticados de ambas mitades.

## Análisis de regresión curvilínea

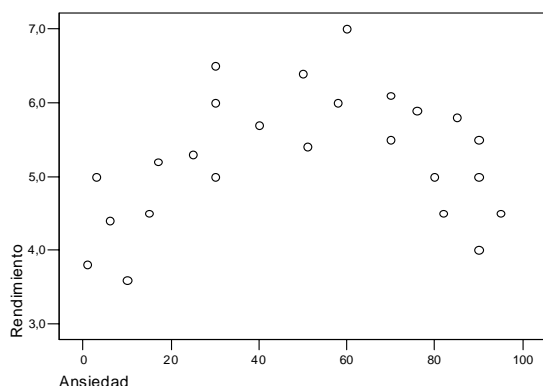
### El procedimiento *Estimación curvilínea*

El *análisis de regresión lineal* estudiado en el capítulo anterior únicamente refleja una forma particular de regresión. Y, aunque la regresión lineal es, quizá, la forma de regresión más útil y extendida, existen funciones no lineales que pueden resultar más apropiadas para estudiar determinados problemas.

Se sabe, por ejemplo, que al intentar resolver una tarea compleja, los sujetos excesivamente ansiosos y los muy relajados rinden peor que los sujetos que mantienen niveles de ansiedad intermedios; es decir, la relación existente entre la ansiedad y el rendimiento no es lineal, sino cuadrática. Por tanto, si se desea pronosticar el rendimiento de los sujetos a partir de su nivel de ansiedad, una ecuación de regresión lineal ofrecerá peores pronósticos y ajuste que una función cuadrática.

El archivo *Ansiedad rendimiento* (puede encontrarse en la página *web* del manual) contiene datos de 25 sujetos en los que se han medido las variables *rendimiento* (en una escala de 0 a 7) y *ansiedad* (en una escala de 0 a 100). El diagrama de dispersión de la Figura 19.1 ofrece una representación de la relación entre la *ansiedad* y el *rendimiento*.

Figura 19.1. Diagrama de dispersión de *ansiedad* por *rendimiento*



La nube de puntos revela con claridad el tipo de relación existente. De hecho, la variable *ansiedad* no se relaciona linealmente con la variable *rendimiento* ( $r=0,21$ ;  $Sig.=0,306$ ). Pero,



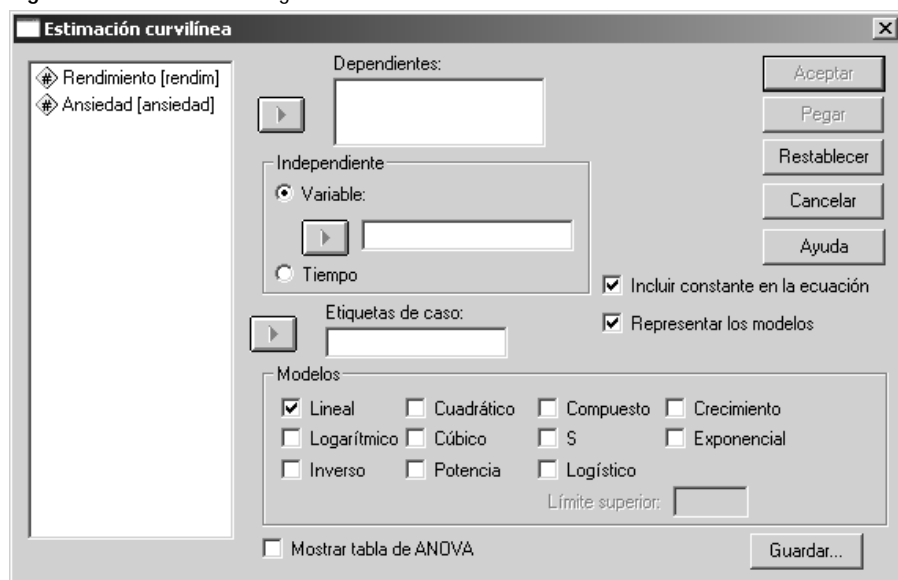
al ajustar un modelo cuadrático a la relación entre esas dos variables, ocurre que las puntuaciones en *rendimiento* correlacionan 0,76 ( $Sig. < 0,0005$ ) con los pronósticos basados en *ansiedad*.

## Estimación curvilínea

El procedimiento Estimación curvilínea ofrece once modelos diferentes de estimación para dos variables (una independiente y una dependiente), con los correspondientes estadísticos y gráficos asociados a cada uno de ellos. Para obtener cualquiera de estos modelos de estimación:

- Seleccionar la opción **Regresión > Estimación curvilínea...** del menú **Analizar** para acceder al cuadro de diálogo *Estimación curvilínea* que muestra la Figura 19.2.

Figura 19.2. Cuadro de diálogo *Estimación curvilínea*



Aunque la lista de variables del archivo de datos muestra todas las variables del archivo, lo cierto es que la estimación curvilínea sólo tiene sentido con variables cuantitativas. Para obtener las estimaciones que el procedimiento tiene establecidas por defecto:

- Seleccionar una variable cuantitativa y trasladarla a la lista **Dependientes**.
- Seleccionar una variable cuantitativa y trasladarla al cuadro **Independiente**.

**Dependientes.** Trasladar a esta lista la(s) variable(s) dependiente(s) que se desea utilizar en el análisis. El procedimiento genera un modelo de regresión (con las estimaciones correspondientes) para cada variable dependiente seleccionada.

**Independientes.** Las opciones de este recuadro permiten decidir qué se va a utilizar como variable independiente del análisis:

**Variable.** Permite seleccionar como variable independiente una variable del archivo de datos.

**Tiempo.** Cuando la variable dependiente es una serie temporal, esta opción permite utilizar el *tiempo* como variable independiente del análisis. En ese caso, el procedimiento genera una variable temporal (una serie temporal) en la que la distancia temporal entre cada caso es uniforme. Esta variable temporal es la que se utiliza como variable independiente en el análisis.

**Etiquetas de caso.** En los diagramas de dispersión que ofrece el procedimiento, los casos individuales se identifican por el número de registro (fila) que ocupan en el *Editor de datos*. Si se desea utilizar como identificadores de caso los valores de alguna variable del archivo de datos, este cuadro permite seleccionar esa variable.

" **Incluir constante en la ecuación.** Esta opción, que se encuentra activa por defecto, permite decidir si se desea o no incluir el término constante en el modelo o modelos de regresión solicitados.

" **Representar los modelos.** Permite obtener un diagrama de dispersión representando la relación entre las variables independiente y dependiente. En el eje de abscisas se representan los valores de la variable independiente y en el de ordenadas los de la dependiente. Los puntos del diagrama se muestran unidos por líneas. El procedimiento genera un diagrama de dispersión distinto para cada variable dependiente, pero las curvas correspondientes a cada modelo solicitado aparecen en el mismo diagrama.

**Modelos.** El procedimiento Estimación curvilínea permite obtener estimaciones para 11 modelos de regresión distintos ( $Y$  = variable dependiente;  $X$  = variable independiente):

" Lineal:  $Y = B_0 + B_1 X$

" Logarítmico:  $Y = B_0 + B_1 \ln(X)$

" Inverso:  $Y = B_0 + B_1 (1/X)$

" Cuadrático:  $Y = B_0 + B_1 X + B_2 X^2$

" Cúbico:  $Y = B_0 + B_1 X + B_2 X^2 + B_3 X^3$

" Potencia:  $Y = B_0 (X^{B_1})$

" Compuesto:  $Y = B_0 (B_1^X)$

" S:  $Y = e^{B_0 + B_1 (1/X)}$

" Logístico:  $Y = 1/(1/c + B_0 B_1^X)$ , donde  $c$  se refiere al límite superior de la función, el cual debe ser un valor mayor que el valor máximo de la variable independiente. El valor de  $c$  debe incluirse en el cuadro Límite superior.

" Crecimiento:  $Y = e^{B_0 + B_1 X}$

" Exponencial:  $Y = B_0 e^{B_1 X}$

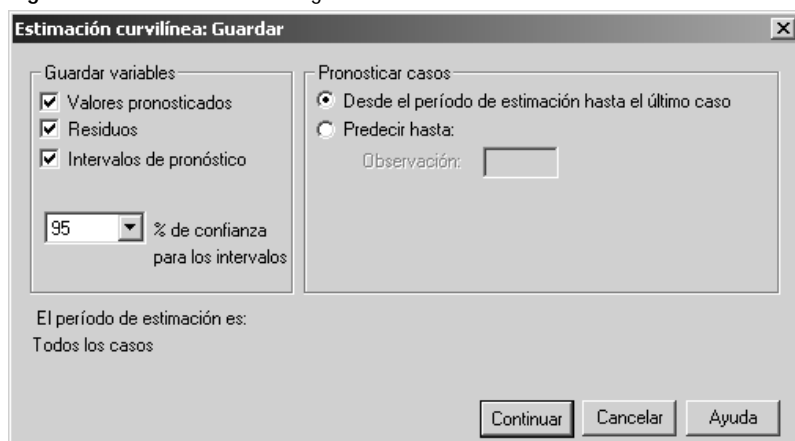
**Mostrar tabla de ANOVA.** Esta opción permite obtener una tabla resumen del ANOVA con información sobre la significación del ajuste de cada modelo. El estadístico  $F$  de la tabla sirve para contrastar la hipótesis nula de que la correlación entre los valores observados y los pronosticados por el modelo vale cero en la población.

## Pronósticos y residuos

El procedimiento **Estimación curvilínea** permite crear nuevas variables (en el *Editor de datos*) con los pronósticos y residuos asociados a cada modelo seleccionado. Para crear estas variables:

- Pulsar el botón **Guardar...** del cuadro de diálogo principal (ver Figura 19.2) para acceder al subcuadro de diálogo *Estimación curvilínea: Guardar* que muestra la Figura 19.3.

Figura 19.3. Subcuadro de diálogo *Estimación curvilínea: Guardar*



**Guardar variables.** Las opciones de este recuadro permiten decidir qué variables se desea crear en el *Editor de datos*:

- **Valores pronosticados.** Crea una variable en el *Editor de datos* con los pronósticos derivados del modelo de regresión solicitado. La nueva variable recibe el nombre  $FIT\_#$ . Se crean tantas variables como modelos solicitados.
- **Residuos.** Crea una variable en el *Editor de datos* con los residuos derivados del modelo de regresión solicitado. Los residuos son las diferencias entre los valores observados en la variable dependiente y los valores pronosticados por el modelo de regresión. La nueva variable recibe el nombre  $ERR\_#$ . Se crean tantas variables como modelos solicitados.
- **Intervalos de pronóstico.** Crea dos variables en el *Editor de datos* con los límites superior e inferior del intervalo de confianza para los pronósticos. Las nuevas variables reciben los nombres  $LCL\_#$  (límite inferior) y  $UCL\_#$  (límite superior). El intervalo

se construye, por defecto, utilizando una confianza del 95 por ciento, pero este valor puede cambiarse utilizando el cuadro de texto **% de confianza para los intervalos**.

**Pronosticar casos.** En el caso de que se haya elegido el *tiempo* como variable independiente, es posible elegir el rango temporal en el que se desea efectuar pronósticos:

- " Desde el periodo de estimación hasta el último caso. Se obtienen pronósticos dentro del rango temporal correspondiente a los casos del archivo de datos. Si se ha reducido el rango original del archivo mediante la opción **Seleccionar casos > Basándose en el rango del tiempo o de los casos** del menú **Datos**, los pronósticos únicamente se calculan para el rango reducido.
- " **Predecir hasta.** Se obtienen pronósticos dentro del rango temporal comprendido entre el valor más pequeño de la serie temporal y el valor especificado en el cuadro de texto **Observación**. Esta opción es útil para obtener pronósticos que vayan más allá del rango temporal original (para hacer esto posible, el SPSS crea casos nuevos en el *Editor de datos*).

### **Ejemplo: Estimación curvilínea**

Este ejemplo muestra cómo utilizar el procedimiento **Estimación curvilínea** y cómo interpretar los resultados que ofrece. Se basa en los datos del archivo *Ansiedad rendimiento*, el cual puede obtenerse en la página *web* del manual. Para generar distintos modelos de regresión con la variable *rendimiento* como variable dependiente y la variable *ansiedad* como variable independiente:

- ' En el cuadro de diálogo principal (ver Figura 19.2), seleccionar la variable *rendimiento* y trasladarla a la lista **Dependientes**.
- ' Seleccionar la variable *ansiedad* y trasladarla al cuadro **Independiente (Variable)**.
- ' Marcar las opciones **Lineal**, **Logarítmico** y **Cuadrático** del recuadro **Modelos**.
- ' Pulsar el botón **Guardar...** para acceder al subcuadro de diálogo *Estimación curvilínea: Guardar* (ver Figura 19.3) y marcar las opciones **Valores pronosticados** y **Residuos** del recuadro **Guardar variables**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor de resultados* ofrece la información que muestran las Tablas 19.1 a la 19.4 y la Figura 19.4. Adicionalmente, el *Editor de datos* muestra nuevas variables con los pronósticos y residuos correspondientes a cada modelo estimado (los nombres de estas variables van acompañados de un número de orden que se corresponde con el número de orden asignado a cada modelo en la Tabla 19.1).

Las tres primeras tablas ofrecen información descriptiva de los datos, de las variables y de los modelos solicitados. La Tabla 19.1 contiene algunos detalles de los modelos que se van a ajustar: el nombre de la variable dependiente (*rendimiento*), el de la variable independiente (*ansiedad*), los modelos o ecuaciones que se van a ajustar (*lineal*, *logarítmico* y *cuadrático*), si esos modelos incluirán o no el término constante, si se ha utilizado alguna variable para etiquetar los casos en los gráficos de dispersión y el nivel mínimo de tolerancia con el que se va a trabajar.

**Tabla 19.1.** Descripción del modelo

Nombre del modelo		MOD_2
Variable dependiente	1	Rendimiento
Ecuación	1	Lineal
	2	Logarítmica
	3	Cuadrático
Variable independiente		Ansiedad
Constante		Incluidos
Variable cuyos valores etiquetan las observaciones en los gráficos		Sin especificar
Tolerancia para la entrada de términos en ecuaciones		,0001

La Tabla 19.2 ofrece un resumen de los casos procesados. En el ejemplo se está utilizando un *total* de 25 casos y no se ha *excluido* ningún caso por tener valor perdido en una o en ambas variables. Los *casos pronosticados* y de los *casos creados nuevos* sólo se dan cuando se utiliza el *tiempo* como variable independiente. Los *casos pronosticados* se refieren a los pronósticos asignados a casos no filtrados cuando previamente al análisis se aplica un filtro basado en el rango de casos (opción **Seleccionar casos...** > **Basándose en el rango del tiempo o de los casos** del menú **Datos**). Los *casos creados nuevos* se obtienen cuando se solicitan pronósticos por encima del rango de valores de la variable independiente (ver Figura 19.3).

**Tabla 19.2.** Resumen de los casos procesados

	N
Total de casos	25
Casos excluidos <sup>a</sup>	0
Casos pronosticados	0
Casos creados nuevos	0

a. Los casos con un valor perdido en cualquier variable se excluyen del análisis.

La Tabla 19.3 ofrece un resumen que incluye, para cada variable, el número de valores positivos, ceros y negativos, y el número de valores perdidos (distinguiendo entre los definidos por el sistema y los definidos por el usuario).

**Tabla 19.3.** Resumen de las variables procesadas

		Variables	
		Dependiente	Independiente
		Rendimiento	Ansiedad
Número de valores positivos		25	25
Número de ceros		0	0
Número de valores negativos		0	0
Número de valores perdidos	Perdidos definidos por el usuario	0	0
	Perdidos del sistema	0	0

La Tabla 19.4 contiene los resultados del análisis: el modelo estimado (*Ecuación*), el cuadrado del coeficiente de correlación (coeficiente de determinación) entre los valores de la variable dependiente y los valores pronosticados por cada modelo (*R cuadrado*), el estadístico *F*,

los grados de libertad del numerador ( $gl1$ ) y del denominador ( $gl2$ ) del estadístico  $F$ , la significación del estadístico  $F$  y las estimaciones correspondientes a los coeficientes o parámetros de cada modelo ( $b_0$ ,  $b_1$ , etc.).

Los tres modelos solicitados ofrecen diferente grado de ajuste a los datos. Con el modelo *lineal* se obtiene un coeficiente de determinación ( $R^2$ ) de 0,045. Al contrastar con el estadístico  $F$  que el coeficiente  $R$  vale cero en la población, se obtiene un nivel crítico mayor que 0,05 ( $Sig=0,306$ ), por lo que no puede rechazarse la hipótesis de relación lineal nula. Puede concluirse, por tanto, que el modelo de regresión lineal no permite obtener una representación apropiada de la relación entre *ansiedad* y *rendimiento* (el modelo lineal no permite obtener un buen ajuste a los datos).

Con el modelo *logarítmico* se obtiene mejor ajuste a los datos que con el modelo lineal. De hecho, el estadístico  $F$  tiene asociado un nivel crítico ( $Sig.=0,030$ ) que permite rechazar la hipótesis de relación nula. No obstante, debe tenerse en cuenta que el coeficiente de determinación toma un valor más bien pequeño (0,189).

Por último, con el modelo *cuadrático*, no sólo se obtiene un estadístico  $F$  cuyo nivel crítico ( $Sig.<0,0005$ ) lleva al rechazo de la hipótesis nula de independencia, sino que el coeficiente de determinación toma un valor de 0,573. Por tanto, puede concluirse que en la relación entre *ansiedad* y *rendimiento* existe un componente cuadrático estadísticamente significativo; además, esta relación cuadrática puede cuantificarse de la siguiente manera: el 57,3 % de la variabilidad del *rendimiento* está explicado (depende o puede ser anticipado) por la variable *ansiedad*.

**Tabla 19.4.** Resumen del modelo y estimaciones de los parámetros

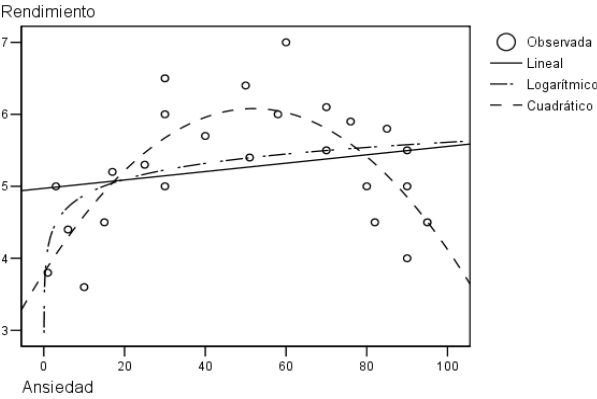
Variable dependiente: Rendimiento								
Ecuación	Resumen del modelo					Estimaciones de los parámetros		
	R cuadrado	F	gl1	gl2	Sig.	Constante	b1	b2
Lineal	,045	1,096	1	23	,306	4,972	,006	
Logarítmica	,189	5,347	1	23	,030	4,138	,320	
Cuadrática	,573	14,744	2	22	,000	3,798	,088	-,001

La variable independiente es Ansiedad.

Cuanto mayor es el número de términos que incluye un modelo, mayor es también el valor del coeficiente de determinación  $R^2$ . Por tanto, en una situación dada, el valor de  $R^2$  será menor en un modelo lineal que en un modelo cuadrático, y menor en éste que en un modelo cúbico. Sin embargo, esto no significa que un modelo con más términos sea mejor. Lo ideal es encontrar un modelo capaz de explicar la mayor cantidad de varianza con el menor número de términos; y para encontrar ese modelo es muy útil contrastar la significación de cada término por separado (esta cuestión se discute más adelante; ver Tabla 19.5). En el ejemplo, de los tres modelos ajustados, dos de ellos ofrecen un ajuste significativo a los datos: el logarítmico y el cuadrático. El modelo logarítmico tiene menos términos que el cuadrático; sin embargo, la calidad del ajuste (medida con el coeficiente de determinación) es sensiblemente mejor con el modelo cuadrático.

Por otro lado, la nube de puntos suele ayudar a elegir el mejor modelo. El *Visor* ofrece el diagrama de dispersión de las variables *ansiedad* y *rendimiento* con las curvas correspondientes a los tres modelos evaluados (ver Figura 19.4). El diagrama muestra con claridad que el modelo cuadrático hace un seguimiento de la nube de puntos sensiblemente mejor que los modelos lineal y logarítmico.

Figura 19.4. Diagrama de dispersión: *ansiedad por rendimiento* (curvas lineal y cuadrática)



Marcando la opción **Mostrar tabla de ANOVA** (ver Figura 19.2, parte inferior del cuadro de diálogo) y seleccionando únicamente el modelo **Cuadrático**, el *Visor de resultados* ofrece información adicional que incluye, entre otras cosas, pruebas de significación individuales para cada parámetro estimado.

La Tabla 19.5 muestra el valor del coeficiente de correlación ( $R$ ), su cuadrado o coeficiente de determinación ( $R$  cuadrado), su valor corregido ( $R$  cuadrado corregida; ver, en el capítulo anterior sobre *Análisis de regresión lineal*, el apartado *Análisis de regresión simple: bondad de ajuste*) y el error típico de los residuos (*Error típico de la estimación*).

Tabla 19.5. Resumen del modelo cuadrático

R	R cuadrado	R cuadrado corregida	Error típico de la estimación
,757	,573	,534	,590

La variable independiente es Ansiedad.

La información de la tabla resumen del ANOVA (ver Tabla 19.6) incluye las sumas de cuadrados, los grados de libertad y las medias cuadráticas utilizadas para obtener el estadístico  $F$ . Este estadístico y su nivel crítico (que ya se habían obtenido anteriormente en los resultados que el procedimiento ofrece por defecto) permiten contrastar la hipótesis nula de que la relación estudiada (en este caso, la cuadrática) vale cero en la población. Puesto que el valor del nivel crítico es muy pequeño ( $\text{Sig.} < 0,0005$ ), se puede rechazar la hipótesis nula y concluir que en la relación entre *ansiedad* y *rendimiento* existe un componente cuadrático significativo.

Tabla 19.6. Resumen del ANOVA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Regresión	10,262	2	5,131	14,744	,000
Residual	7,656	22	,348		
Total	17,918	24			

La variable independiente es Ansiedad.

Por último, la Tabla 19.7 contiene las estimaciones de los coeficientes no estandarizados del modelo cuadrático ( $B$ ), su error típico, su valor tipificado ( $Beta$ ) y un estadístico  $T$  junto con su nivel crítico ( $Sig.$ ) que permite contrastar la hipótesis nula de que el valor poblacional del coeficiente estimado es cero. Los resultados indican que tanto el término constante como los dos términos asociados a la variable *ansiedad* (*Ansiedad* y *Ansiedad\*\*2*), son significativamente distintos de cero: en todos ellos  $Sig. < 0,0005$ .

**Tabla 19.7.** Coeficientes del modelo cuadrático

	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
	B	Error típico	Beta		
Ansiedad	,088	,016	3,216	5,424	,000
Ansiedad ** 2	-,001	,000	-3,090	-5,210	,000
(Constante)	3,798	,318		11,939	,000

Para un estudio más detallado de todos estos conceptos puede revisarse el capítulo anterior sobre *Análisis de regresión lineal*.





## Fiabilidad de las escalas

### El procedimiento *Análisis de fiabilidad*

El procedimiento **Análisis de fiabilidad** ofrece un conjunto de estadísticos diseñados para valorar algunas propiedades métricas de los instrumentos de medida.

En las ciencias sociales, del comportamiento y de la salud es frecuente utilizar pruebas de rendimiento, escalas de actitudes, encuestas de opinión, cuestionarios de diversa índole, tests psicológicos, etc. Utilizaremos el nombre genérico de *escala* para referirnos a este tipo de instrumentos de medida. Una escala no es simplemente un conjunto de preguntas, sino un conjunto de preguntas que poseen una serie de propiedades métricas. La teoría de la medida o de la medición se ocupa de estudiar y cuantificar estas propiedades, las cuales forman parte indispensable e inseparable de la escala. La información sobre estas propiedades debe acompañar a la escala junto con el método de corrección, las instrucciones de aplicación, los baremos normativos y la guía para la interpretación de las puntuaciones.

Así pues, para que un instrumento de medida pueda utilizarse con confianza debe ir acompañado de una serie de propiedades que aseguren su capacidad para medir. Estas propiedades pueden agruparse en tres grandes apartados: fiabilidad, validez y factibilidad.

La *fiabilidad* es la capacidad de la escala para medir de forma consistente, precisa y sin error la característica que se desea medir. Aspectos importantes de la fiabilidad son: la capacidad de la escala para obtener, cuando se aplica a los mismos sujetos, la misma medición en dos situaciones diferentes (es decir, la estabilidad de la medición cuando no ha existido cambio alguno); la consistencia de sus elementos para medir la misma característica (es decir, el grado de homogeneidad de sus elementos); la ausencia de error en las mediciones.

La *validez* es la capacidad de la escala para medir lo que pretende medir y no otros aspectos distintos de los pretendidos. Aspectos importantes de la validez de una escala son: su capacidad para discriminar entre sujetos de los que se sabe que difieren en la característica medida (validez discriminativa); la concordancia existente entre las mediciones obtenidas con la escala y las obtenidas con otras estrategias (otras escalas) que miden la misma característica (validez concurrente); la articulación de las puntuaciones en dimensiones similares a las propuestas teóricamente (validez de constructo); la concordancia entre las puntuaciones obtenidas con la escala y las valoraciones hechas por expertos (validez de contenido). En el caso de los instrumentos de medida utilizados en el ámbito clínico debe prestarse especial atención a la sensibilidad de la escala para detectar los cambios producidos en la evolución clínica de los sujetos.

La *factibilidad* se refiere a la facilidad para aplicar la escala en diversas situaciones y a distintos grupos de sujetos. Aspectos destacables de la factibilidad de una escala son: el grado

de dificultad en la comprensión de las instrucciones, preguntas, dibujos, etc; el tiempo necesario para aplicar la escala; su capacidad para captar la falta de sinceridad de los sujetos; etc.

Este capítulo se centra en el estudio de las propiedades métricas englobadas bajo el concepto de *fiabilidad*. En concreto, en la *fiabilidad de las escalas*. Una escala está formada por un conjunto de elementos (preguntas, ítem, etc.) cada uno de los cuales mide de manera individual la característica que se intenta medir. Los estadísticos que ofrece el procedimiento **Análisis de fiabilidad** asumen que los elementos de la escala se combinan aditivamente, es decir, que la puntuación global de la escala se obtiene sumando las puntuaciones de sus elementos (existen escalas no aditivas –por ejemplo, aquellas en las que la puntuación total se obtiene multiplicando las puntuaciones de los elementos– pero no serán tratadas aquí). El procedimiento también asume que todos los elementos de la escala miden la característica deseada en la misma dirección (es decir, el procedimiento no admite utilizar simultáneamente elementos formulados de manera positiva y negativa: las puntuaciones altas deben tener el mismo significado en todos los elementos).

Por otro lado, las escalas pueden dividirse en dos grandes grupos. Las escalas *unidimensionales* miden una única característica o dimensión; por tanto, todas las preguntas o elementos miden la misma dimensión. Las escalas *multidimensionales* miden más de una característica o dimensión; por tanto, las preguntas están agrupadas por dimensiones de tal forma que unas preguntas miden una dimensión y otras preguntas miden otra dimensión distinta. Cuando la escala es de tipo multidimensional, el cálculo de la fiabilidad se realiza por separado para cada una de las subescalas o dimensiones de la escala; en principio, no deben mezclarse elementos de las distintas subescalas o dimensiones en el cálculo de la fiabilidad.

## Concepto de fiabilidad

Ya se ha señalado que la fiabilidad de una escala se refiere a la capacidad de la escala para medir de forma consistente y precisa la característica que pretende medir. Una balanza es fiable si cada vez que se pesa el mismo objeto se obtiene el mismo resultado. Una escala es fiable si cada vez que se mide a los mismos sujetos se obtiene el mismo resultado. Pero al medir *sujetos* surge un problema que no es tan evidente cuando se miden *objetos*: los sujetos son cambiantes; de modo que no es fácil saber si la variabilidad en las mediciones obtenidas se debe a la imprecisión de la escala o a los cambios operados en los sujetos. El análisis de fiabilidad se ocupa de la precisión del instrumento, es decir, de los errores incontrolables, inevitables e impredecibles asociados a todo proceso de medida.

El concepto de fiabilidad incluye dos aspectos complementarios: la *consistencia interna* y la *estabilidad temporal*. La consistencia interna recoge el grado de coincidencia o parecido (homogeneidad) existente entre los elementos que componen la escala. La estabilidad en el tiempo se refiere a la capacidad del instrumento para arrojar las mismas mediciones cuando se aplica en momentos distintos a los mismos sujetos.

La teoría clásica de los test asume que las puntuaciones de los sujetos en una escala (puntuaciones *observadas* o *empíricas*:  $X_i$ ) pueden interpretarse como la suma de dos componentes independientes: las *puntuaciones verdaderas* de los sujetos en la característica medida ( $V_i$ ) y el conjunto de fuentes de *error* que concurren en la medición ( $E_i$ ). Esta relación entre las puntuaciones observadas, las verdaderas y el error de medida puede representarse mediante la siguiente ecuación ( $i = 1, 2, \dots, n$ ):

$$X_i = V_i + E_i$$

Las medidas de fiabilidad intentan cuantificar qué cantidad de la variabilidad de las mediciones obtenidas en una escala (puntuaciones observadas) se debe a las puntuaciones verdaderas y qué cantidad se debe a los errores de medida. Como se asume que los errores de medida son independientes de las puntuaciones verdaderas (pues son atribuibles a factores externos que afectan a los sujetos de forma no controlable ni predecible, y se consideran aleatorios), la varianza de las puntuaciones observadas puede descomponerse de esta manera:

$$\sigma_x^2 = \sigma_v^2 + \sigma_e^2$$

El **coeficiente de fiabilidad** ( $\rho_{xx}$ ) se define como la proporción de varianza de las puntuaciones observadas que es atribuible a la variabilidad de las puntuaciones verdaderas:

$$\rho_{xx} = \sigma_v^2 / \sigma_x^2$$

Este coeficiente puede interpretarse de modo similar a como se hace con el coeficiente  $R^2$  en un análisis de regresión lineal. No obstante, el coeficiente de fiabilidad se suele expresar, no tanto en términos de la varianza de las puntuaciones verdaderas, sino de la varianza error:

$$\rho_{xx} = 1 - (\sigma_e^2 / \sigma_x^2)$$

Por tanto, el coeficiente de fiabilidad es un valor que oscila entre 0 y 1, y se encuentra tanto más próximo a 1 cuanto menor es la variabilidad error de las mediciones. De hecho, el coeficiente de fiabilidad suele interpretarse como un indicador de la precisión o ausencia de error de las mediciones de la escala.

Por su parte, el **índice de fiabilidad** ( $\rho_{xy}$ ) se define como la correlación existente entre las puntuaciones observadas y las puntuaciones verdaderas, y puede demostrarse que está íntimamente relacionado con el coeficiente de fiabilidad:

$$\rho_{xy} = \frac{\sigma_v}{\sigma_x} = \sqrt{\rho_{xx}}$$

Puesto que tanto la varianza de las puntuaciones verdaderas como la varianza de los errores son valores desconocidos, el valor del coeficiente de fiabilidad debe ser estimado. Es decir,  $\rho_{xx}$  es un valor poblacional que debe estimarse a partir de la información muestral.

## Análisis de fiabilidad

La estimación del coeficiente de fiabilidad puede hacerse tomando como referencia distintos escenarios o modelos. El primero de estos escenarios se da cuando se dispone de una escala formada por un conjunto de elementos que se considera que son representativos de todos los elementos que hubiera sido posible utilizar. Las puntuaciones de cada elemento se obtienen administrando la escala a una muestra representativa de sujetos en una única ocasión y la puntuación total de la escala se obtiene sumando las puntuaciones de los elementos. En este escenario, la fiabilidad de la escala puede estimarse a partir del grado de homogeneidad existente

entre los elementos de la escala o a partir del grado de relación existente entre *dos mitades* de la escala.

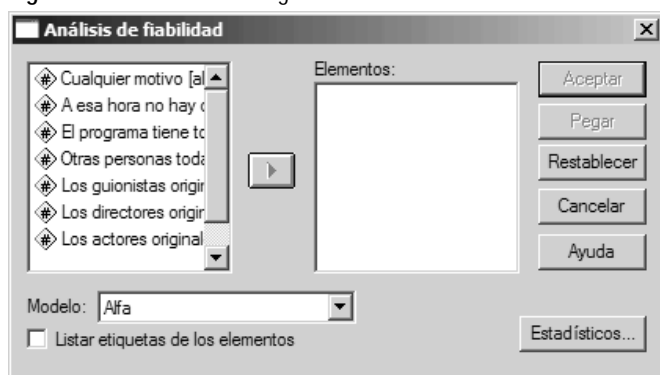
Un segundo escenario se da cuando se dispone de dos *formas paralelas* (equivalentes, pero formadas por elementos o preguntas distintas) de una misma escala y ambas formas se administran a los mismos sujetos. La fiabilidad de la escala se evalúa correlacionando los resultados obtenidos con ambas formas. Este escenario se utiliza cuando, por las características de los elementos de la escala, existe la posibilidad de que el recuerdo de las respuestas pueda influir en la puntuación final de la escala y se desea evitar que esto ocurra; o cuando hay que comparar grupos medidos de manera consecutiva y existe la posibilidad de que haya intercambio de información entre los grupos.

El tercer escenario se da cuando se llevan a cabo dos mediciones consecutivas con la misma escala (*test y retest*), es decir, cuando la escala se administra en dos ocasiones sucesivas a los mismos sujetos. En este escenario, la fiabilidad de la escala se establece a partir de la correlación existente entre los resultados obtenidos en ambas aplicaciones. Este escenario se utiliza para valorar la estabilidad de las puntuaciones obtenidas con la escala a lo largo del tiempo.

El procedimiento **Análisis de fiabilidad** ofrece estimaciones de la fiabilidad para estos diferentes escenarios. Por tanto, para elegir la estrategia de estimación apropiada, es necesario tener en cuenta el escenario en el que se está trabajando. Para llevar a cabo un análisis de fiabilidad:

- Seleccionar la opción **Escala > Análisis de fiabilidad...** del menú **Analizar** para acceder al cuadro de diálogo *Análisis de fiabilidad* que se muestra en la Figura 20.1.

Figura 20.1. Cuadro de diálogo *Análisis de fiabilidad*



La lista de variables contiene un listado de todas las variables *numéricas* del archivo de datos (las variables de *cadena* no están incluidas en este listado porque el procedimiento no permite utilizar este tipo de variables). Para obtener un análisis de fiabilidad con las especificaciones que el procedimiento tiene establecidas por defecto:

- Seleccionar el conjunto de variables que se desea analizar y trasladarlas a la lista **Elementos** y pulsar el botón **Aceptar**.
- **Listar etiquetas de los elementos.** Esta opción permite solicitar un listado de las *etiquetas de variable* de todos los elementos (variables) analizados.

### Ejemplo: Análisis de fiabilidad

Este ejemplo muestra cómo obtener e interpretar los resultados que el procedimiento **Análisis de fiabilidad** ofrece por defecto.

Los datos en los que se basa corresponden a una encuesta realizada a 906 espectadores de televisión sobre los motivos por los que estarían dispuestos a seguir viendo un determinado programa en la siguiente temporada. Estos datos están disponibles en el archivo *tv-survey*, el cual se encuentra en la misma carpeta en la que está instalado el SPSS. Las siete variables del archivo son dicotómicas (1 = «sí», 0 = «no»). Para llevar a cabo el análisis:

- En el cuadro de diálogo *Análisis de fiabilidad* (ver Figura 20.1), seleccionar todas las variables del archivo (siete variables en total) y trasladarlas a la lista **Elementos**.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestran las Tablas 20.1 y 20.2. La Tabla 20.1 incluye información relacionada con el número de casos procesados: se están utilizando 906 casos *válidos* del *total* de 906 casos que contiene el archivo; por tanto, no se ha *excluido* ningún caso del archivo por tener valor perdido (por defecto, el procedimiento deja fuera del análisis cualquier caso con valor perdido en alguna de las variables incluidas en el análisis).

**Tabla 20.1.** Resumen de los casos procesados

		N	%
Casos	Válidos	906	100.0
	Excluidos <sup>a</sup>	0	.0
	Total	906	100.0

a. Eliminación por lista basada en todas las variables del procedimiento.

La Tabla 20.2 informa sobre el número de elementos (variables) incluidos en el análisis (7 en el ejemplo) y sobre el valor del coeficiente de fiabilidad *alfa* de Cronbach. El coeficiente de fiabilidad *alfa* puede interpretarse de modo similar a como se hace con un coeficiente de correlación al cuadrado ( $R^2$ ); mide el grado de homogeneidad o parecido existente entre los elementos (en realidad es una especie de promedio de las correlaciones entre los elementos). En el siguiente apartado se explica con detalle este coeficiente; de momento, basta con saber que, desde el punto de vista de la fiabilidad de la escala, el valor obtenido (0,898) es bastante bueno (valores por encima de 0,90 suelen considerarse excelentes).

**Tabla 20.2.** Coeficiente de fiabilidad

Alfa de Cronbach	N de elementos
.898	7

Conviene señalar que, aunque el valor del coeficiente de fiabilidad (único resultado que el procedimiento ofrece por defecto) permite conocer el comportamiento global de la escala, no permite valorar el comportamiento individual de los elementos que la componen. Para esto último es necesario, según se verá más adelante, seleccionar algunas opciones.

## Modelos de fiabilidad

La lista desplegable **Modelo** (ver Figura 20.1) permite seleccionar el modelo que servirá de referente para estimar la fiabilidad. El modelo seleccionado no altera el valor de los estadísticos seleccionados, pero permite obtener diferentes coeficientes de fiabilidad. Los modelos disponibles en este menú desplegable son (los siguientes apartados ofrecen una descripción detallada de cada modelo):

- ▼ **Alfa.** Es el modelo por defecto. Valora la consistencia interna de la escala a partir de la correlación inter-elementos promedio.
- ▼ **Dos mitades.** Divide la escala en dos partes y calcula la fiabilidad a partir de la correlación entre ambas partes.
- ▼ **Guttman.** Calcula los límites inferiores de Guttman.
- ▼ **Paralelo.** Asume que todos los elementos tienen iguales varianzas observada y error.
- ▼ **Paralelo estricto.** Asume los supuestos del modelo paralelo y, además, que las medias de todos los elementos son iguales.

### Modelo alfa

El modelo *alfa* (Cronbach, 1951) asume que la escala está compuesta por elementos homogéneos aleatoriamente seleccionados de la población de los posibles elementos que miden la misma característica. También asume que la *consistencia interna* de la escala puede evaluarse mediante la correlación existente entre sus elementos.

El coeficiente *alfa* es una estimación del límite inferior de la fiabilidad poblacional (coincide con el límite  $L_3$  de Guttman; ver más adelante). Depende del número de elementos de la escala ( $k$ ) y del cociente entre la covarianza promedio de los elementos y su varianza promedio. Llamando  $j$  a un elemento cualquiera de la escala ( $j = 1, 2, \dots, k$ ) y  $j'$  a otro elemento cualquiera distinto de  $j$ , el coeficiente *alfa* se define de la siguiente manera:

$$\alpha = \frac{k \bar{S}_{jj'}^2 / \bar{S}_j^2}{1 + (k-1) \bar{S}_{jj'}^2 / \bar{S}_j^2}$$

Existe otra fórmula equivalente del coeficiente *alfa* cuya utilización en la literatura psicométrica está más generalizada:

$$\alpha = \frac{k}{k-1} \left( 1 - \frac{\sum_j S_j^2}{S_x^2} \right)$$

Y, si se asume que las varianzas de los elementos son iguales (o si los elementos originales se estandarizan), puede obtenerse una versión estandarizada de *alfa* a partir de las correlaciones entre los elementos:

$$\alpha = \frac{k \bar{r}_{jj'}}{1 + (k-1) \bar{r}_{jj'}}$$

Esta versión estandarizada sólo se obtiene en el SPSS al seleccionar las opciones relacionadas con los estadísticos *entre-elementos* (ver Figura 20.2). Cuando la escala está formada por elementos dicotómicos, el coeficiente *alfa* coincide con la fórmula  $KR_{20}$  de Kuder-Richardson:

$$\alpha = \frac{k}{k-1} \left( 1 - \frac{\sum_j p_j q_j}{S_x^2} \right)$$

### Ejemplo: Análisis de fiabilidad > Modelo alfa

Este ejemplo muestra cómo obtener e interpretar el coeficiente de fiabilidad *alfa* de Cronbach con los datos del archivo *tv-survey*. El modelo *alfa* de Cronbach es el que utiliza por defecto el procedimiento **Análisis de fiabilidad** del SPSS, de modo que no es necesario marcar ninguna opción concreta para obtenerlo:

- Seleccionar todas las variables de la escala y trasladarlas a la lista **Elementos**.
- Comprobar en la lista desplegable **Modelo** que está seleccionado el modelo **Alfa**.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Análisis de fiabilidad: Estadísticos* y seleccionar la opción **Correlaciones** del recuadro **Entre-elementos** (para obtener el coeficiente *alfa* estandarizado).

Aceptando estas selecciones, el *Visor* ofrece, entre otros, los resultados que muestran las Tablas 20.3 y 20.4.

La Tabla 20.3, además de informar del número de elementos analizados (7 variables), ofrece el coeficiente de fiabilidad *alfa*. Los valores por encima de 0,8 se suelen considerar buenos y los valores por encima de 0,9 excelentes. El valor del ejemplo, 0,898, es alto, lo que indica gran consistencia interna entre los elementos de la escala.

Una interpretación bastante extendida de un coeficiente *alfa* alto es que la escala está midiendo una única dimensión, sin embargo, esto es algo que no puede deducirse de forma directa. Según señalan Green, Lissitz y Mulaik (1977), una elevada consistencia interna no necesariamente implica unidimensionalidad.

La tabla también incluye el valor del coeficiente *alfa* estandarizado. Puesto que las varianzas de los elementos son muy parecidas, el valor del coeficiente estandarizado es similar al del coeficiente no estandarizado.

**Tabla 20.3.** Coeficiente de fiabilidad (modelo alfa)

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
.898	.894	7



La Tabla 20.4 ofrece una matriz con los coeficientes de correlación entre cada par de elementos de la escala. Puede observarse en la tabla que no todos los coeficientes de correlación obtenidos son altos: oscilan entre aproximadamente 0,30 y 0,80.

Tabla 20.4. Matriz de correlaciones entre los elementos de la escala

	Alguna	Aburrido	Críticas	Iguals	Guión	Director	Reparto
Alguna	1.000	.815	.813	.782	.408	.421	.303
Aburrido	.815	1.000	.826	.807	.422	.423	.307
Críticas	.813	.826	1.000	.804	.458	.453	.336
Iguals	.782	.807	.804	1.000	.443	.460	.340
Guión	.408	.422	.458	.443	1.000	.632	.625
Director	.421	.423	.453	.460	.632	1.000	.600
Reparto	.303	.307	.336	.340	.625	.600	1.000

## Modelo de dos mitades

El modelo de *dos mitades* (*split*) asume que la escala está constituida por dos partes de igual longitud. Ambas mitades pueden sumarse para obtener la puntuación total en la escala.

Esta opción es útil cuando se dispone de dos mediciones consecutivas (*test-retest*) y se desea valorar la estabilidad de las medidas entre ambas mediciones; o cuando se dispone de dos formas paralelas de la misma escala y se desea valorar si realmente son equivalentes. En ocasiones, aunque no sea posible administrar la escala en dos ocasiones diferentes, puede interesar conocer la fiabilidad de la escala total a partir de la fiabilidad de sus partes (como ocurre, por ejemplo, con las escalas que valoran conocimientos).

El procedimiento asume que la primera medición ( $X_1$ ) está representada por la primera mitad de las variables seleccionadas en el cuadro de diálogo (en la lista **Elementos**), y la segunda medición ( $X_2$ ) por la segunda mitad de las variables seleccionadas. Los resultados que se obtienen dependen, por tanto, de cómo se seleccionen los elementos en el cuadro de diálogo. Si los elementos de la escala están ordenados por su nivel de dificultad, es recomendable seleccionar primero los elementos impares y luego los pares.

Dado que el coeficiente *alfa* es proporcional a la longitud de la escala, el procedimiento ofrece varios coeficientes para poder elegir el más apropiado de acuerdo con el escenario en el que se han obtenido las mediciones. El primero de estos coeficientes es el coeficiente de correlación entre las puntuaciones totales de las dos subescalas o mitades ( $X_1$  y  $X_2$ ):

$$r_{x_1 x_2} = \frac{(S_x^2 - S_{x_1}^2 - S_{x_2}^2) / 2}{S_{x_1} S_{x_2}}$$

El procedimiento también ofrece el *estadístico de dos mitades Guttman*, que sirve para evaluar la fiabilidad de la escala total (es decir, de la escala formada por la suma de las dos mitades) cuando puede asumirse que la varianza de todos los elementos son iguales:

$$r_{\text{Guttman}} = \frac{2 (S_x^2 - S_{x_1}^2 - S_{x_2}^2)}{S_x^2}$$

Por último, dentro del modelo de dos mitades, el procedimiento ofrece las dos versiones de la *profecía de Spearman-Brown* para la fiabilidad de la escala total (la escala de doble longitud). En la primera versión se asume que las dos subescalas son de igual longitud:

$$r_{\text{Spearman-Brown-iguales}} = \frac{2r_{x_1x_2}}{1+r_{x_1x_2}}$$

En la segunda versión de la profecía de Spearman-Brown se asume que las subescalas son de distinta longitud:

$$r_{\text{Spearman-Brown-distintas}} = \frac{-r_{x_1x_2}^2 + \sqrt{r_{x_1x_2}^4 + 4r_{x_1x_2}^2(1-r_{x_1x_2}^2)k_1k_2/k^2}}{2(1-r_{x_1x_2}^2)k_1k_2/k^2}$$

En ambas versiones se asume que las dos subescalas tienen la misma fiabilidad y que sus varianzas son iguales. El estadístico de Spearman-Brown establece que es posible calcular la fiabilidad de la escala de longitud doble a partir de las escalas de longitud simple.

Al seleccionar el modelo de dos mitades, tanto los estadísticos descriptivos como el coeficiente *alfa* se calculan por separado para cada una de las mitades.

### Ejemplo: Análisis de fiabilidad > Modelo de dos mitades

Este ejemplo muestra cómo obtener e interpretar los estadísticos del modelo de *dos mitades*. Se siguen utilizando los datos del archivo *tv-survey*. Para obtener los estadísticos asociados a este modelo:

- Seleccionar todas las variables de la escala y desplazarlas a la lista **Elementos**.
- En la lista desplegable **Modelo** seleccionar la opción **Dos mitades**.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestra la Tabla 20.5. Puesto que se ha seleccionado el modelo de *dos mitades*, los elementos se dividen en dos subescalas: los primeros cuatro elementos seleccionados forman la primera subescala (*Parte 1*); los siguientes tres elementos forman la segunda subescala (*Parte 2*). En dos notas a pie de tabla se indica qué elementos en concreto pertenecen a cada subescala. En la primera subescala, el valor de *alfa* es 0,944; en la segunda, 0,826. De acuerdo con estos resultados y teniendo en cuenta que el coeficiente *alfa* vale 0,894 cuando se calcula sobre toda la escala, parece evidente que la primera mitad o subescala es más fiable que la segunda (aunque no debe olvidarse que el valor del coeficiente *alfa* depende del número de elementos).

Después de informar del número total de elementos analizados (7), se ofrece el coeficiente de correlación entre las dos mitades o subescalas (*Correlación entre formas* = 0,503). Este coeficiente sirve como una estimación de la estabilidad temporal de las mediciones; por tanto, resulta apropiado cuando se llevan a cabo dos mediciones consecutivas con la misma escala (*test-retest*) y se desea conocer el grado de parecido existente entre ellas.

A continuación aparecen las dos versiones del coeficiente de fiabilidad de Spearman-Brown: la versión que asume que las subescalas son de igual longitud (0,669) y la versión que

asume que las subescalas son de distinta longitud (0,673). Y, por último, se ofrece el coeficiente de fiabilidad de dos mitades de Guttman (0,577). Así pues, si puede asumirse que todos los elementos tienen la misma varianza, la estimación del coeficiente de fiabilidad para la escala de 7 elementos vale 0,577; si se asume que ambas partes tienen la misma longitud, el coeficiente de fiabilidad vale 0,669; y se asume que tienen distinta longitud, el coeficiente de fiabilidad vale 0,673.

**Tabla 20.5.** Coeficientes de fiabilidad (modelo de dos mitades)

Alfa de Cronbach	Parte 1	Valor	,944
		N de elementos	4 <sup>a</sup>
	Parte 2	Valor	,826
		N de elementos	3 <sup>b</sup>
	N total de elementos		7
Correlación entre formas			,503
Coeficiente de Spearman-Brown	Longitud igual		,669
	Longitud desigual		,673
Dos mitades de Guttman			,577

a. Los elementos son: Cualquier motivo, A esa hora no hay otros programas populares, El programa tiene todavía buenas críticas, Otras personas todavía ven el programa.

b. Los elementos son: Otras personas todavía ven el programa, Los guionistas originales permanecen en el programa, Los directores originales permanecen en el programa, Los actores originales siguen en el programa.

## Modelo de Guttman

El modelo de Guttman permite obtener varias estimaciones del límite inferior de la fiabilidad (Guttman, 1945). Guttman ha propuesto 6 de estas estimaciones, todas las cuales están incluidas en el procedimiento *Análisis de fiabilidad*. El *Visor de resultados* ofrece estas estimaciones con los nombres *Lambda 1* a *Lambda 6* (en todos los casos,  $j \neq j'$ ):

$$L_1 = \frac{\sum_j S_j^2}{S_x^2} \quad L_2 = L_1 + \frac{\sqrt{\frac{k}{k-1} \sum_j \sum_{j'} S_{jj'}^2}}{S_x^2} \quad L_3 = \frac{k}{k-1} L_1$$

$$L_4 = \frac{2 \sum_j \sum_{j'} S_{jj'}}{S_x^2} \quad L_5 = L_1 + \frac{2 \sqrt{\max_j \sum_j S_{jj'}^2}}{S_x^2} \quad L_6 = 1 - \frac{\sum_j [S^{-1}]_{jj'}^{-1}}{S_x^2}$$

$L_1$  es una estimación simple en la que se basan otros límites.  $L_3$  es mejor que  $L_1$ ; es mayor y coincide con el coeficiente *alfa* de Cronbach.  $L_2$  es preferible a los dos anteriores, aunque su cálculo es más complejo.  $L_5$  es preferible a  $L_2$  cuando existe un elemento cuyas covarianzas con el resto de los elementos son muy altas y el resto de los elementos no presentan grandes

covarianzas entre ellos.  $L_6$  es preferible a  $L_2$  cuando las correlaciones entre elementos son bajas en comparación con la correlación múltiple al cuadrado entre cada elemento y los restantes.  $L_4$  es el coeficiente de Guttman del modelo de dos mitades y es un límite inferior de la fiabilidad de cualquiera de las dos partes de la escala.

### Ejemplo: Análisis de fiabilidad > Modelo de Guttman

Este ejemplo muestra cómo obtener e interpretar los estadísticos asociados al modelo de Guttman con los datos del archivo *tv-survey*. Para obtener estos estadísticos:

- Seleccionar todas las variables de la escala y trasladarlas a la lista **Elementos**.
- En la lista desplegable **Modelo** seleccionar la opción **Guttman**.

Aceptando estas selecciones, el *Visor de resultados* ofrece la información que muestra la Tabla 20.6. Puede comprobarse en estos resultados que *Lambda 3* es igual al coeficiente de fiabilidad *alfa* de Cronbach (ver Tabla 20.3), y que *Lambda 4* es igual al coeficiente de dos mitades de Guttman (ver Tabla 20.5). También puede comprobarse que el valor más alto es el de *Lambda 6*, aunque exceptuando los valores de *Lambda 1* y *Lambda 4*, todos los límites ofrecen un valor similar.

Tabla 20.6. Coeficientes de fiabilidad (modelo de Guttman)

Lambda	1	,769
	2	,915
	3	,898
	4	,577
	5	,894
	6	,927
N de elementos		7

Se ha calculado la matriz de covarianzas y se utiliza en el análisis.

### Modelo de medidas paralelas

Los modelos de *medidas paralelas* y *estrictamente paralelas* (Kristof 1963, 1969) asumen que los elementos de la escala son versiones paralelas (equivalentes) de una población de elementos que miden la característica que se intenta medir. El modelo de *medidas paralelas* asume que las puntuaciones verdaderas de todos los elementos tienen la misma varianza. El modelo de *medidas estrictamente paralelas* asume que, además de las varianzas, también son iguales las medias.

Estos dos modelos permiten obtener estimaciones de la varianza de las puntuaciones verdaderas y de la varianza error. En el modelo de *medidas paralelas*, por ejemplo, estas estimaciones se obtienen de la siguiente manera:

$$S_{\text{verdadera}}^2 = \frac{2}{k(k-1)} \sum_j \sum_{j'} S_{jj'}$$

$$S_{\text{error}}^2 = \frac{1}{k} \sum_j S_j^2 - \frac{2}{k(k-1)} \sum_j \sum_{j'} S_{jj'}$$

Los dos modelos ofrecen también una estimación de la varianza común (que en el modelo de *medidas paralelas* se obtiene como el promedio de las varianzas de los elementos); y el modelo *estrictamente paralelo* ofrece, además, una estimación de la media común (que se obtiene como el promedio de las medias de todos los elementos).

Para la fiabilidad, se ofrecen dos estimaciones. Una sesgada, que en el modelo de *medidas paralelas* no es otra cosa que el coeficiente de fiabilidad *alfa* de Cronbach; y otra insesgada, que consiste en aplicar al valor de *alfa* una corrección basada en el número de casos:

$$\alpha' = \frac{2 + \alpha(n-3)}{n-1}$$

### Ejemplo: Análisis de fiabilidad > Modelo de medidas paralelas

Este ejemplo muestra cómo obtener e interpretar los resultados del modelo de *medidas paralelas* con los datos del archivo *tv-survey*. Para ello:

- Seleccionar todas las variables de la escala y trasladarlas a la lista **Elementos**.
- En la lista desplegable **Modelo** seleccionar la opción **Paralelo**.

Aceptando estas selecciones, el *Visor de resultados* ofrece la información que muestran las Tablas 20.7 y 20.8. La primera tabla ofrece un contraste sobre bondad de ajuste que permite poner a prueba la hipótesis nula de que las varianzas poblacionales de los elementos son iguales (recuérdese que éste es un supuesto básico del modelo de medidas paralelas). Puesto que el estadístico de contraste (*Chi-cuadrado* = 1.968,281) tiene asociado un nivel crítico muy pequeño (*Sig.* < 0,0005), debe rechazarse la hipótesis de igualdad de varianzas. En consecuencia, el modelo de medidas paralelas no parece apropiado para estos datos.

**Tabla 20.7.** Contraste de la bondad de ajuste del modelo

Chi-cuadrado	Valor	1968,281
	gl	26
	Sig.	,000
Log del determinante de	Matriz no restringida	-16,885
	Matriz restringida	-14,704

Bajo el supuesto del modelo paralelo

La Tabla 20.8 ofrece las estimaciones de los parámetros relevantes del modelo: la varianza de las puntuaciones observadas (*Varianza común* = 0,199), la varianza de las puntuaciones verdaderas (*Varianza verdadera* = 0,111), la varianza de los errores (*Varianza error* = 0,088) y la correlación promedio entre los elementos (*Correlación inter-elementos común* = 0,556).

Las dos últimas filas recogen las estimaciones de la fiabilidad de la escala bajo el modelo de medidas paralelas. La estimación sesgada es el coeficiente de fiabilidad *alfa* (*Fiabilidad de la escala* = 0,898); la estimación insesgada se obtiene aplicando a *alfa* una corrección basada en el número de sujetos (*Fiabilidad de la escala insesgada* = 0,898).

Puesto que no puede asumirse como válido el modelo de *medidas paralelas* (se ha rechazado la hipótesis de igualdad entre las varianzas de los elementos), tampoco podrá asumirse como válido el modelo de *medidas estrictamente paralelas*.

Tabla 20.8. Coeficientes de fiabilidad (modelo de medidas paralelas)

Varianza común	,199
Varianza verdadera	,111
Varianza error	,088
Correlación inter-elementos común	,556
Fiabilidad de la escala	,898
Fiabilidad de la escala (insesgada)	,898

Se ha calculado la matriz de covarianzas y se utiliza en el análisis.

## Estadísticos

La mayor parte de la información que es posible obtener con el procedimiento **Análisis de fiabilidad** es opcional. El subcuadro de diálogo **Estadísticos** permite seleccionar toda esta información opcional, que incluye desde varios estadísticos descriptivos (para la escala y para cada elemento) hasta el coeficiente de correlación intraclase, pasando por varios contrastes sobre igualdad de medias. Para obtener estos estadísticos:

- Pulsar el botón **Estadísticos...** del cuadro de diálogo principal (ver Figura 20.1) para acceder al subcuadro de diálogo *Análisis de fiabilidad: Estadísticos* que muestra la Figura 20.2.

Figura 20.2. Subcuadro de diálogo *Análisis de fiabilidad: Estadísticos*

**Análisis de fiabilidad: Estadísticos**

Descriptivos para:

- ☐ Elemento
- ☐ Escala
- ☐ Escala si se elimina elemento

Entre-elementos:

- ☐ Correlaciones
- ☐ Covarianzas

Resúmenes:

- ☐ Medias
- ☐ Varianzas
- ☐ Covarianzas
- ☐ Correlaciones

Tabla de ANOVA:

- ☒ Ninguno
- ☐ Prueba F
- ☐ Chi-cuadrado de Friedman
- ☐ Chi-cuadrado de Cochran

☐ I-cuadrado de Hotelling ☐ Prueba de aditividad de Tukey

☒ Coeficiente de correlación intraclase

Modelo: **Dos factores, efectos mixtos** Tipo: **Consistencia**

Intervalo de confianza: **95** % Valor de prueba: **0**

Continuar Cancelar Ayuda

Todos los estadísticos incluidos en este subcuadro de diálogo se describen a continuación siguiendo la organización por recuadros del propio subcuadro, e incluyendo un ejemplo por recuadro.

## Descriptivos

El recuadro **Descriptivos** genera estadísticos descriptivos para cada uno de los elementos y para la escala total. La opciones disponibles son las siguientes:

- " **Elemento.** Media, desviación típica y número de casos válidos de cada elemento (variable) de la escala.
- " **Escala.** Media, varianza, desviación típica y número de variables (elementos) de la escala total. Se entiende que la escala total se obtiene sumando todos los elementos. En el caso de que se haya elegido el modelo de *dos mitades* o el de *Guttman*, el procedimiento también ofrece estos estadísticos para cada una de las mitades o subescalas.
- " **Escala si se elimina el elemento.** Estadísticos que valoran el comportamiento de la escala total cuando se van eliminando uno a uno los elementos de la escala: media de la escala cuando se elimina cada elemento, varianza de la escala cuando se elimina cada elemento, correlación entre cada elemento y la escala sin incluir ese elemento, y coeficiente de fiabilidad *alfa* de Cronbach cuando se elimina cada elemento.

### **Ejemplo: Análisis de fiabilidad > Estadísticos > Descriptivos**

Este ejemplo muestra cómo obtener e interpretar los estadísticos del recuadro **Descriptivos** del subcuadro de diálogo *Análisis de fiabilidad: Estadísticos* (Figura 20.2). Se sigue utilizando el archivo *tv-survey*. Para obtener estos estadísticos:

- ' Seleccionar las tres opciones del recuadro **Descriptivos** (ver Figura 20.2): **Elemento**, **Escala** y **Escala si se elimina el elemento**.

Aceptando estas selecciones se obtienen los resultados que muestran las Tablas 20.9 a la 20.11. La Tabla 20.9 ofrece, *para cada elemento*, la media, la desviación típica y el número de casos válidos. Puede apreciarse que no hay valores perdidos ( $N = 906$  en todos los elementos). Esta información es relevante para valorar la *factibilidad* de la escala: debe prestarse especial atención a la presencia de elementos con un elevado número de valores perdidos.

En cuanto a las medias, dado que los elementos son dicotómicos (1 = «sí», 0 = «no»), las medias reflejan la proporción de casos de la muestra que han respondido *sí* a cada pregunta. Puede apreciarse que las medias difieren entre sí. En los tres últimos elementos existe una elevada proporción de sujetos que han respondido *sí*; son elementos que se refieren a la permanencia de los equipos originales en el programa. Otro grupo de elementos han sido contestados afirmativamente por, aproximadamente, la mitad de la muestra; son elementos que se refieren a las valoraciones de los críticos y allegados. El elemento con media más baja es el referido a cualquier motivo inespecífico.

La variabilidad de los elementos también es informativa. Los elementos que tienen menor variabilidad son los que poseen menor capacidad para discriminar (diferenciar) entre los sujetos. En el ejemplo, las desviaciones típicas no son especialmente informativas, dado que la varianza de un elemento dicotómico está relacionada con su promedio.

La Tabla 20.10 ofrece los estadísticos descriptivos referidos a la *escala total*. La media total (4,55) indica que, por término medio, los encuestados han elegido de 4 a 5 motivos para seguir el programa en la próxima temporada.

Tabla 20.9. Descriptivos de cada elemento

	Media	Desviación típica	N
Cualquier motivo	.49	.500	906
A esa hora no hay otros programas populares	.50	.500	906
El programa tiene todavía buenas críticas	.50	.500	906
Otras personas todavía ven el programa	.53	.499	906
Los guionistas originales permanecen en el programa	.81	.389	906
Los directores originales permanecen en el programa	.83	.378	906
Los actores originales siguen en el programa	.89	.315	906

Tabla 20.10. Descriptivos de la escala total

Media	Varianza	Desviación típica	N de elementos
4.55	6.040	2.458	7

La Tabla 20.11 muestra varios estadísticos que permiten valorar el comportamiento de la escala total cuando se va eliminando uno a uno cada elemento. Las dos primeras columnas contienen la media y la varianza de la escala cuando se va eliminando cada elemento; fuertes cambios en estos valores podrían estar delatando elementos cuyo comportamiento (en términos de media o varianza) está muy alejado del de los restantes elementos.

La tercera columna recoge las correlaciones entre cada elemento y el total de la escala excluido el elemento (*índice de homogeneidad corregido*); si todos los elementos miden la misma dimensión, estas correlaciones serán altas; una correlación baja estaría indicando que el elemento en cuestión no apunta en la misma dirección que el resto de los elementos.

La cuarta columna ofrece el coeficiente de determinación  $R^2$  resultante de incluir cada elemento como variable dependiente en un análisis de regresión con los restantes elementos como variables independientes. Estos coeficientes constituyen un buen indicador del grado de parecido de cada elemento con los restantes (para obtener esta columna es necesario marcar la opción **Correlaciones** del recuadro **Resúmenes**; ver Figura 20.2).

Tabla 20.11. Descriptivos obtenidos al eliminar cada elemento

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-total corregida	Correlación múltiple al cuadrado	Alfa de Cronbach si se elimina el elemento
Cualquier motivo	4,07	4,171	,792	,740	,871
A esa hora no hay otros programas populares	4,05	4,144	,808	,768	,869
El programa tiene todavía buenas críticas	4,05	4,113	,827	,770	,867
Otras personas todavía ven el programa	4,02	4,142	,811	,732	,869
Los guionistas permanecen en el programa	3,74	4,877	,589	,523	,894
Los directores permanecen en el programa	3,72	4,905	,593	,503	,894
Los actores originales siguen en el programa	3,66	5,240	,486	,461	,904

La última columna ofrece los valores del coeficiente *alfa* cuando se elimina cada elemento. Comparando estos valores con el valor del coeficiente *alfa* para toda la escala, es posible determinar si existe algún elemento que se diferencia de los demás; si el coeficiente de fiabilidad



aumenta al eliminar un elemento, probablemente ese elemento no está midiendo la misma dimensión que el resto de elementos y no debería formar parte de la escala. En el ejemplo, al eliminar el último elemento, la fiabilidad de la escala sube a 0,904, es decir, se produce un ligero incremento sobre el valor de *alfa* para la escala total (0,898). Además, este elemento es el que menos correlaciona con los restantes (0,486). Dado que este elemento ha sido elegido por muchas personas, eliminarlo de la escala haría disminuir la media total (3,66), pero haría mejorar la capacidad discriminativa de la escala, ya que aumentaría su varianza (5,24).

## Resúmenes

Las opciones del recuadro **Resúmenes** (ver Figura 20.2) permiten obtener varios estadísticos descriptivos sobre algunas distribuciones. Las distribuciones disponibles son:

- " **Medias.** Distribución de las medias de los elementos.
- " **Varianzas.** Distribución de las varianzas de los elementos.
- " **Covarianzas.** Distribución de las covarianzas entre los elementos.
- " **Correlaciones.** Distribución de las correlaciones entre los elementos.

Para todas estas distribuciones, el procedimiento ofrece la media, el valor mínimo y máximo, la amplitud o rango, el cociente entre el valor máximo y el mínimo, y la varianza. En el caso de que se haya elegido el modelo de *dos mitades* o el de *Guttman*, el procedimiento también ofrece estos estadísticos para las distribuciones de cada una de las mitades.

### *Ejemplo: Análisis de fiabilidad > Estadísticos > Resúmenes*

Este ejemplo muestra cómo obtener e interpretar los estadísticos del recuadro **Resúmenes** del subcuadro de diálogo *Análisis de fiabilidad: Estadísticos* (Figura 20.2):

- ' Seleccionar las cuatro opciones del recuadro **Resúmenes** (ver Figura 20.2): **Medias**, **Varianzas**, **Covarianzas** y **Correlaciones**.

Aceptando estas selecciones se obtienen los resultados que muestra la Tabla 20.12. Estos resultados se refieren a las distribuciones de las medias y de las varianzas de los elementos, y a las distribuciones de las covarianzas y de las correlaciones entre los elementos.

**Tabla 20.12.** Descriptivos de las distribuciones

	Media	Mínimo	Máximo	Rango	Máximo/ mínimo	Varianza	N de elementos
Medias de los elementos	.650	.487	.889	.402	1.825	.033	7
Varianzas de los elementos	.199	.099	.250	.151	2.524	.004	7
Covarianzas inter-elementos	.111	.048	.207	.159	4.332	.004	7
Correlaciones inter-elementos	.547	.303	.826	.523	2.726	.036	7

Los estadísticos referidos a estas distribuciones son útiles para diagnosticar la presencia de elementos que se comportan de manera anómala o que no se comportan como el resto de los

elementos de la escala. Así, por ejemplo, puede observarse que la correlación promedio entre los elementos es 0,547, siendo 0,303 la menor y 0,826 la mayor. No parece que las correlaciones entre los elementos sean muy homogéneas. La razón *Max/Min* es un buen indicador de la existencia de elementos anómalos en la escala, pero debe tenerse en cuenta que depende de la métrica del estadístico valorado.

## Entre elementos

Las opciones del recuadro **Entre-elementos** permiten obtener información sobre el grado de relación existente entre los elementos. Las opciones disponibles son:

- " **Correlaciones.** Matriz de orden  $k$  que contiene unos en la diagonal principal y las correlaciones entre cada par de elementos fuera de la diagonal principal.
- " **Covarianzas.** Matriz de orden  $k$  que contiene las varianzas de cada elemento en la diagonal principal y las covarianzas entre cada par de elementos fuera de la diagonal principal.

### *Ejemplo: Análisis de fiabilidad > Estadísticos > Entre-elementos*

Este ejemplo muestra cómo obtener la matriz de varianzas-covarianzas y la matriz de correlaciones del recuadro **Entre-elementos** del subcuadro de diálogo *Análisis de fiabilidad: Estadísticos* (Figura 20.2). Se sigue utilizando el archivo *tv-survey*. Para obtener estas matrices:

- Seleccionar las dos opciones del recuadro **Entre-elementos** (ver Figura 20.2): **Correlaciones** y **Covarianzas**.

Aceptando estas selecciones el *Visor* ofrece dos matrices de datos. La primera de ellas (ver Tabla 20.13) es la matriz de correlaciones: contiene unos en la diagonal y las correlaciones entre cada par de elementos fuera de la diagonal. Aunque todas las correlaciones del ejemplo presentan un tamaño muy aceptable, parece que existen grupos con distinto grado de relación entre sus elementos. En general, éste no es un buen síntoma, ya que sugiere la posibilidad de que la escala se esté comportando de manera multidimensional (para comprobar si esta sospecha es cierta puede aplicarse un análisis factorial).

**Tabla 20.13.** Matriz de correlaciones

	Alguna	Aburrido	Críticas	Iguals	Guión	Director	Reparto
Alguna	1.000	.815	.813	.782	.408	.421	.303
Aburrido	.815	1.000	.826	.807	.422	.423	.307
Críticas	.813	.826	1.000	.804	.458	.453	.336
Iguals	.782	.807	.804	1.000	.443	.460	.340
Guión	.408	.422	.458	.443	1.000	.632	.625
Director	.421	.423	.453	.460	.632	1.000	.600
Reparto	.303	.307	.336	.340	.625	.600	1.000

La segunda matriz (ver Tabla 20.14) es la matriz de varianzas-covarianzas: contiene las varianzas de cada elemento en la diagonal principal y las covarianzas entre cada par de elemen-

tos fuera de la diagonal principal. La información de esta matriz es difícil de interpretar cuando los elementos de la escala no tienen la misma métrica; por este motivo, es preferible solicitar e interpretar su versión estandarizada, es decir, la matriz de correlaciones (recuérdese que el coeficiente de correlación de Pearson se obtiene dividiendo la covarianza entre el producto de desviaciones típicas).

**Tabla 20.14.** Matriz de varianzas-covarianzas

	Alguna	Aburrido	Críticas	Iguals	Guión	Director	Reparto
Alguna	.250	.204	.203	.195	.079	.079	.048
Aburrido	.204	.250	.207	.202	.082	.080	.048
Críticas	.203	.207	.250	.201	.089	.086	.053
Iguals	.195	.202	.201	.249	.086	.087	.053
Guión	.079	.082	.089	.086	.151	.093	.077
Director	.079	.080	.086	.087	.093	.143	.071
Reparto	.048	.048	.053	.053	.077	.071	.099

En estas matrices es muy importante vigilar la presencia de valores negativos. Dado que la mayoría de los estadísticos de fiabilidad se basan en el supuesto de que los elementos de la escala se combinan aditivamente, la presencia de correlaciones negativas (y en especial la presencia de filas o columnas enteras con signo negativo) podría estar indicando que existen elementos codificados en sentido inverso al de los demás elementos de la escala; si fuera ése el caso, sería necesario recodificar esos elementos para poder estimar correctamente la fiabilidad de la escala.

## Tabla de ANOVA

Las opciones del recuadro **Tabla de ANOVA** ofrecen varios estadísticos que permiten contrastar la hipótesis nula de igualdad entre las medias de los elementos.

La igualdad de medias es, entre otras cosas, uno de los supuestos del modelo de *medidas estrictamente paralelas*. Además, contrastar esta hipótesis de igualdad de medias resulta especialmente interesante cuando los elementos de una escala son valoraciones emitidas por jueces, pues, en ese caso, la fuente de variabilidad inter-medidas refleja la variabilidad inter-jueces. Las opciones disponibles son:

“ **Ninguna.** No se ofrece la tabla de ANOVA. Es la opción por defecto.

“ **Prueba F.** Ofrece el estadístico *F* del análisis de varianza para contrastar la hipótesis nula de que todos los elementos de la escala tienen la misma media. Es una opción válida cuando los elementos están medidos en escala de intervalo o razón.

En los resultados que ofrece el procedimiento, la tabla de ANOVA descompone la variabilidad total de la escala en dos fuentes de variabilidad: la atribuible a las diferencias entre los sujetos (variación inter-sujetos) y la atribuible a las diferencias entre las medidas (variación intra-sujetos).

Si se considera que cada sujeto constituye su propio grupo, estas variabilidades pueden interpretarse exactamente igual que en un diseño experimental sin interacción (ver, por ejemplo, Winer, Brown y Michels, 1991). Los sujetos constituyen un *factor inter-su-*

*jetos*; la variabilidad entre los niveles de este factor representa las diferencias existentes entre los sujetos, la cual recoge el hecho de que la muestra contiene sujetos con distinto nivel en la característica que se está midiendo (habilidad, actitud, conocimientos). Y los distintos elementos de la escala constituyen los niveles de un *factor intra-sujetos* o de medidas repetidas; la variabilidad entre los niveles de este factor representa las diferencias observadas entre el conjunto de medidas efectuadas a los mismos sujetos.

La variabilidad asociada al factor intra-sujetos puede, a su vez, descomponerse en dos fuentes de variabilidad: una atribuible a las diferencias existentes entre las puntuaciones de un mismo sujeto en los distintos elementos (variabilidad inter-medidas) y otra atribuible a los errores aleatorios que se cometen en la medición (variabilidad residual).

Si todos los elementos de la escala asignaran la misma puntuación a cada sujeto (es decir, si cada sujeto puntuara de la misma forma en todos los elementos), la variabilidad inter-medidas no sería distinta de la variabilidad residual, en cuyo caso las medias de todos los elementos serían iguales. La información de la tabla de ANOVA permite contrastar la hipótesis de igualdad de medias a través del estadístico  $F$ , el cual se obtiene dividiendo la media cuadrática asociada a la variabilidad entre las medidas ( $MC_{\text{inter-medidas}}$ ) entre la media cuadrática asociada a los errores de medida ( $MC_{\text{residual}}$ ):

$$F = \frac{MC_{\text{inter-medidas}}}{MC_{\text{residual}}}$$

Si puede asumirse *esfericidad* (ver Capítulo 16), este estadístico se distribuye según el modelo de probabilidad  $F$  con  $k-1$  y  $(n-1)(k-1)$  grados de libertad. Por tanto, es posible utilizar el estadístico  $F$  y su distribución de probabilidad para tomar decisiones sobre la hipótesis de igualdad de medias entre los elementos.

- " **Chi-cuadrado de Friedman.** Opción válida para contrastar la hipótesis nula de que todos los elementos tienen la misma media cuando los elementos están medidos en escala ordinal. El SPSS genera la misma tabla de ANOVA que con la opción **Prueba  $F$** , pero en lugar del estadístico  $F$  ofrece el estadístico  $X_r^2$  de Friedman y el coeficiente de concordancia  $W$  de Kendall (ver Capítulo 21 sobre *Análisis no paramétrico*):

$$X_r^2 = \frac{SC_{\text{inter-medidas}}}{MC_{\text{intra-sujetos}}} \quad \text{y} \quad W = \frac{SC_{\text{inter-medidas}}}{SC_{\text{total}}}$$

Ambos estadísticos se distribuyen según el modelo de probabilidad  $\chi^2$  con  $k-1$  grados de libertad. De modo que es posible utilizar estos estadísticos y su distribución muestral para tomar decisiones sobre la hipótesis de igualdad de medias entre los elementos.

- " **Chi-cuadrado de Cochran.** Opción válida para contrastar la hipótesis nula de igualdad de medias cuando los elementos son dicotómicos (unos y ceros). El SPSS genera la misma tabla de ANOVA que con la opción **Prueba  $F$** , pero en lugar del estadístico  $F$  ofrece el estadístico  $Q$  de Cochran (ver Capítulo 21 sobre *Análisis no paramétrico*). El estadístico  $Q$  de Cochran se obtiene con la misma ecuación que el estadístico  $X_r^2$  de Friedman, pero con la diferencia de que, mientras  $X_r^2$  se aplica a datos ordinales,  $Q$  se aplica a datos dicotómicos.

**Ejemplo: Análisis de fiabilidad > Estadísticos > Tabla de ANOVA**

Este ejemplo muestra cómo obtener e interpretar los resultados del recuadro **Tabla de ANOVA** del subcuadro de diálogo *Análisis de fiabilidad: Estadísticos* (Figura 20.2) con los datos del archivo *tv-survey*. Para obtener estos resultados:

- En el recuadro **Tabla de ANOVA** (ver Figura 20.2), seleccionar la opción **Chi-cuadrado de Cochran**.

*Nota:* puesto que los datos del archivo *tv-survey* son dicotómicos, no tiene sentido utilizar ni la prueba *F* ni el estadístico de Friedman; no obstante, con todas las pruebas se obtienen tablas de resultados idénticas; únicamente cambia en ellas el valor de estadístico de contraste.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestra la Tabla 20.15. Independientemente de la prueba elegida, las tablas de resultados siempre son tablas en formato ANOVA.

La Tabla 20.15 muestra los resultados del estadístico *Q* de Cochran. El dato relevante en estas tablas se encuentra en la última columna (*Sig.*). Puesto que el nivel crítico asociado al estadístico *F* es menor que 0,0005, debe rechazarse la hipótesis de que las medias poblacionales de los elementos son iguales.

**Tabla 20.15.** Resumen del ANOVA. Estadístico *Q* de Cochran

		Suma de cuadrados	gl	Media cuadrática	Q de Cochran	Sig.
Inter-personas		780,866	905	,863	1491,561	,000
Intra-personas	Inter-elementos	181,487	6	30,248		
	Residual	479,942	5430	,088		
	Total	661,429	5436	,122		
Total		1442,295	6341	,227		

Media global = ,65

**Prueba  $T^2$  de Hotelling**

El estadístico  $T^2$  de Hotelling (ver Winer, Brown y Michels, 1991, págs. 278-281) es un estadístico multivariante que permite, al igual que el estadístico *F* del ANOVA, contrastar la hipótesis nula de igualdad de medias.

Comparado con el estadístico *F*, el estadístico  $T^2$  de Hotelling tiene la ventaja de que no necesita asumir *normalidad* en las distribuciones de donde se muestrea ni *esfericidad* en la matriz de varianzas-covarianzas (ver Capítulo 16). No obstante, en el caso de que pueda asumirse normalidad y esfericidad, el estadístico *F* es más potente que el estadístico  $T^2$ , sobre todo con muestras pequeñas.

El estadístico  $T^2$  se utiliza habitualmente para comparar dos vectores de medias multivariantes y se basa en la matriz de varianzas-covarianzas entre los elementos. En el contexto del análisis de fiabilidad se utiliza para contrastar la hipótesis nula de igualdad de medias entre los elementos de la escala.

Para obtener  $T^2$  se comienza definiendo el vector  $\mathbf{Y}$ , que contiene todas las diferencias de medias entre cada elemento y el elemento  $k$ :

$$\mathbf{Y} = \begin{bmatrix} \bar{X}_1 - \bar{X}_k \\ \bar{X}_2 - \bar{X}_k \\ \dots \\ \bar{X}_{k-1} - \bar{X}_k \end{bmatrix}$$

El estadístico  $T^2$  se calcula como:

$$T^2 = n \mathbf{Y}' \mathbf{B}^{-1} \mathbf{Y},$$

donde  $n$  es el número de sujetos de la muestra y  $\mathbf{B}$  es la matriz definida por:

$$\mathbf{B} = \mathbf{CSC}'$$

siendo  $\mathbf{C}$  una matriz identidad de rango  $k-1$  ampliada con un vector columna  $-1$  posterior, y  $\mathbf{S}$  la matriz de varianzas-covarianzas entre los elementos. El estadístico  $T^2$  se puede transformar en el estadístico  $F$ :

$$F = \frac{n-k-1}{(n-1)(k-1)} T^2$$

el cual se distribuye según el modelo de probabilidad  $F$  con  $k-1$  y  $n-k-1$  grados de libertad. Si la probabilidad asociada al estadístico  $F$  es menor 0,05, puede rechazarse la hipótesis de igualdad entre las medias de los elementos comparados.

### ***Ejemplo: Análisis de fiabilidad > Estadísticos > $T^2$ de Hotelling***

Este ejemplo muestra cómo obtener e interpretar el estadístico  $T^2$  de Hotelling con los datos del archivo *tv-survey* (aunque los datos del archivo corresponden a variables dicotómicas, se aplica el estadístico  $T^2$  para ejemplificar su uso). Para obtener el estadístico  $T^2$ :

- En el cuadro de diálogo *Análisis de fiabilidad: Estadísticos* (ver Figura 20.2), seleccionar la opción ***T-cuadrado de Hotelling***.

Aceptando estas selecciones se obtiene el resultado que muestra la Tabla 20.16. El dato relevante se encuentra en la última columna: puesto que el nivel crítico es muy pequeño ( $\text{Sig.} < 0,00005$ ), se puede afirmar que los promedios poblacionales de los elementos comparados no son iguales.

Tabla 20.16.  $T^2$  de Hotelling

T-cuadrado de Hotelling	F	gl1	gl2	Sig.
661.813	109.693	6	900	.000

## Prueba de aditividad de Tukey

La opción **Prueba de aditividad de Tukey** ofrece el contraste de *aditividad* (también llamado de *no aditividad*) de Tukey (1949), junto con una estimación de la potencia a la que habría que elevar las puntuaciones observadas para conseguir aditividad.

Esta prueba permite contrastar el supuesto de que los elementos de la escala no interactúan con los sujetos; es decir, el supuesto de que el efecto debido a los elementos y el efecto debido a los sujetos se combinan aditivamente; o, de otro modo, el supuesto de que en el modelo de ANOVA que recoge ambos efectos, el efecto de la interacción entre ellos es nulo. Para obtener el contraste de aditividad se comienza calculando:

$$d_i = \bar{X}_i - \bar{X} \quad \text{y} \quad d_j = \bar{X}_j - \bar{X}$$

y las sumas de cuadrados:

$$SC_{\text{no aditiv}} = \frac{\left( \sum_i \sum_j X_{ij} d_i d_j \right)^2}{\sum_i d_i^2 \sum_j d_j^2} \quad \text{y} \quad SC_{\text{resto}} = SC_{\text{residual}} - SC_{\text{no aditiv}}$$

La propuesta de Tukey se basa en el siguiente argumento: si se asume que el modelo es aditivo, la variabilidad no debida ni a los elementos ni a los sujetos (que son justamente los dos términos del modelo aditivo) es variabilidad residual. Esta variabilidad residual puede dividirse en dos componentes: el relacionado con la interacción elementos-sujetos o  $SC_{\text{no aditiv}}$  (componente que puede considerarse responsable de la no aditividad) y el no relacionado con la interacción elementos-sujetos o  $SC_{\text{resto}}$  (componente formado por el resto de variabilidad residual). La comparación de ambos componentes puede informar sobre la verdadera importancia del componente relacionado con la no aditividad. La prueba de aditividad de Tukey adopta el siguiente formato:

$$F_{\text{no aditiv}} = \frac{SC_{\text{no aditiv}}}{SC_{\text{resto}} / [(n-1)(k-1) - 1]}$$

Bajo la hipótesis de aditividad (no interacción elementos-sujetos), este estadístico se distribuye según el modelo de probabilidad  $F$  con 1 y  $(n-1)(k-1)-1$  grados de libertad. Cuando el nivel crítico asociado al estadístico  $F_{\text{no aditiv}}$  es menor que 0,05, no es posible asumir aditividad, lo que significa que los resultados obtenidos asumiendo aditividad (como la tabla de ANOVA o el coeficiente de correlación intraclase) deben ser interpretados con cautela.

### **Ejemplo: Análisis de fiabilidad > Estadísticos > Prueba de aditividad de Tukey**

Este ejemplo muestra cómo obtener e interpretar la prueba de aditividad de Tukey con los datos del archivo *tv-survey*. Para obtener este contraste:

- En el cuadro de diálogo *Análisis de fiabilidad: Estadísticos* (ver Figura 20.2), seleccionar la opción **Prueba de aditividad de Tukey**.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestra la Tabla 20.17. El primer bloque de información recoge la tabla resumen del ANOVA. Esta tabla es igual que las ya obtenidas con las opciones del recuadro **Tabla de ANOVA**, pero incluye información adicional. La variabilidad residual aparece ahora descompuesta en dos partes: la debida a la interacción elementos-sujetos (*No aditividad*) y la no debida a esa interacción (*Equilibrio*). El estadístico  $F_{\text{no aditiv}}$  vale 845,945 y tiene asociado un nivel crítico (*Sig.*) menor que 0,0005; puesto que este nivel crítico es muy pequeño, se puede rechazar la hipótesis de aditividad y concluir que el modelo aditivo no es apropiado.

Al final de la tabla aparece, en una nota a pie de tabla, una estimación de la potencia a la que podrían elevarse las puntuaciones de los elementos para lograr aditividad. El resultado del ejemplo indica que las puntuaciones deberían elevarse a 2,107. Pero debe tenerse en cuenta que esta estrategia sólo es apropiada para variables cuantitativas. Con elementos dicotómicos (unos y ceros) como los de este ejemplo no tiene sentido elevar los valores a una potencia.

**Tabla 20.17.** Resumen del ANOVA. Contraste de no aditividad de Tukey

			Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-personas			780,866	905	,863		
Intra-personas	Inter-elementos		181,487	6	30,248	342,219	,000
	Residual	No aditividad	64,703 <sup>a</sup>	1	64,703	845,945	,000
		Equilibrio	415,239	5429	,076		
		Total	479,942	5430	,088		
	Total		661,429	5436	,122		
Total			1442,295	6341	,227		

Media global = ,65

a. Estimación de Tukey de la potencia a la que es necesario elevar las observaciones para conseguir aditividad = 2,107.

## Coefficiente de correlación intraclass

Esta opción permite obtener una medida del grado de consistencia o acuerdo existente entre los elementos de la escala. Por tanto, es especialmente apropiada para valorar el acuerdo existente entre las valoraciones emitidas por un conjunto de jueces o expertos.

Esta medida se llama *coeficiente de correlación intraclass* (CCI) y se basa en la teoría de la generalizabilidad. El procedimiento **Análisis de fiabilidad** ofrece un intervalo de confianza para la fiabilidad estimada mediante el CCI y un estadístico  $F$  que permite contrastar la hipótesis nula de que el CCI poblacional adopta un determinado valor.

La idea subyacente en el cálculo del CCI es que las puntuaciones observadas obedecen a dos fuentes de variabilidad: una atribuible a los elementos (las valoraciones de los jueces) y otra atribuible a los sujetos. Si las valoraciones son consistentes (homogéneas), la variabilidad observada entre ellas debe ser pequeña en comparación con la variabilidad observada entre los sujetos. Por tanto, es posible utilizar los modelos de ANOVA ya conocidos para analizar la situación.

La estimación del CCI (ver McGraw y Wong, 1996) depende del tipo de aleatorización que se adopte para cada una de estas dos fuentes de variabilidad. En todos los casos se asume que los sujetos constituyen una muestra aleatoria de todos los casos posibles, por lo que este



factor de clasificación se considera de *efectos aleatorios*. Pero respecto de los elementos caben dos posibilidades: que el conjunto de elementos sea una muestra aleatoria de todos los posibles elementos, o que el conjunto de elementos sea único en algún sentido.

**Modelo.** Las opciones de menú desplegable **Modelo** (ver Figura 20.2) permiten elegir uno de estos tres modelos:

- ▼ **Dos factores, efectos mixtos.** Si los sujetos se consideran una muestra aleatoria pero los elementos (o las valoraciones) se consideran fijos, el modelo de ANOVA que representa la situación es el de *dos factores de efectos mixtos*, sin interacción (el factor *elementos* o *jueces* de efectos fijos y el factor *sujetos* de efectos aleatorios). Tal es el caso, por ejemplo, cuando los elementos son valoraciones de jueces que no pueden considerarse extrapolables a otros jueces, o cuando los elementos de la escala son los únicos elementos posibles (o los únicos elementos que se desea considerar) para evaluar la característica medida.
- ▼ **Dos factores, efectos aleatorios.** Si tanto los sujetos como los elementos (o las valoraciones) se consideran una muestra aleatoria de sus respectivas poblaciones, el modelo de ANOVA que representa la situación es el de *dos factores de efectos aleatorios*, sin interacción. Este es el escenario más habitual, pues los elementos de una escala, generalmente, no son otra cosa que una muestra de la población de posibles elementos capaces de medir una determinada característica, y los jueces seleccionados para efectuar las valoraciones no son más que una muestra aleatoria de la población de posibles jueces.
- ▼ **Un factor, efectos aleatorios.** Por último, si los elementos se han aplicado sin orden y no se tiene constancia de qué posición ocupa cada uno de ellos (es decir, si las puntuaciones de una misma columna del archivo de datos no corresponden al mismo elemento), lo apropiado es utilizar un ANOVA de *un factor de efectos aleatorios*. Sólo se tiene en cuenta el efecto de los *sujetos* y, obviamente, se considera aleatorio. Esta situación se da, por ejemplo, cuando un grupo de  $n$  sujetos es valorado por  $k$  jueces, pero no se sabe qué valoración corresponde a cada juez.

Por supuesto, la elección del modelo debe ajustarse a las características de los datos. Y esto, no sólo porque las estimaciones que se obtienen con cada modelo son diferentes, sino porque también son diferentes las conclusiones que es posible alcanzar con cada uno de ellos: mientras que las conclusiones obtenidas con el modelo de un factor de efectos aleatorios son generalizables a la población de posibles niveles del factor, las conclusiones obtenidas con un modelo de factor de efectos fijos sólo son generalizables a los niveles del factor.

El procedimiento calcula, por un lado, el CCI *individual*, que es conceptualmente similar a la fiabilidad individual de cada uno de los elementos (o fiabilidad de la escala si ésta estuviera formada por un único elemento); y, por otro lado, el CCI *promedio*, que es una estimación de la fiabilidad de la escala total. El coeficiente de correlación intraclase *individual* (para un solo elemento) se obtiene de la siguiente manera:

$$CCI_j = \frac{MC_{\text{inter-sujetos}} - MC_{\text{error}}}{MC_{\text{inter-sujetos}} + (k-1)MC_{\text{error}}}$$

Y el coeficiente de correlación intraclass *promedio* (para los  $k$  elementos de la escala) se obtiene de la siguiente manera:

$$CCI_k = \frac{MC_{\text{inter-sujetos}} - MC_{\text{error}}}{MC_{\text{inter-sujetos}}}$$

Estas dos ecuaciones sirven para los tres modelos de ANOVA disponibles en el procedimiento. Lo único que cambia en cada modelo es la forma concreta de estimar las medias cuadráticas.

**Tipo.** Al margen de cuál sea el modelo de ANOVA apropiado para representar los datos, el grado de homogeneidad de los elementos (o de las valoraciones de los jueces) puede establecerse desde dos perspectivas diferentes. Ambas perspectivas están disponibles en el menú desplegable **Tipo**:

- ▼ **Consistencia.** La consistencia se refiere al grado de parecido existente entre los elementos considerando que, para que dos elementos sean parecidos, basta con que varíen conjuntamente (correlacionen), sin necesidad de que la media y la varianza de los mismos sean iguales. Imaginemos una situación con 3 sujetos y 2 elementos en la que cada sujeto obtiene estas puntuaciones: 2, 4; 3, 5; 4, 6. Las puntuaciones no son idénticas, pero son consistentes: el sujeto que más bajo puntúa en los dos elementos es el mismo; y el que puntúa más alto en los dos elementos también. El  $CCI_k$  aplicado a estos datos vale 1. Las fórmulas anteriormente propuestas para el  $CCI_j$  y el  $CCI_k$  valoran la consistencia.
- ▼ **Acuerdo absoluto.** El acuerdo absoluto se refiere al grado de parecido existente entre los elementos considerando que, para que dos elementos sean parecidos, además de variar de la misma manera (correlacionar), la media y la varianza de los mismos deben ser iguales. Al valorar el grado de acuerdo absoluto existente entre las puntuaciones del apartado anterior, el  $CCI_k$  toma el valor 0,5. El  $CCI_k$  sólo tomaría el valor 1 si cada sujeto tuviera la misma puntuación en los dos elementos. Por tanto, cuando entre los elementos de una escala existe acuerdo absoluto, también existe consistencia; pero cuando existe consistencia, puede no existir acuerdo absoluto.

Para obtener el grado de acuerdo absoluto, las fórmulas del CCI correspondientes a los modelos de dos factores cambian. En la fórmula del CCI *individual* hay que sumar al denominador la cantidad  $(k/n)(MC_{\text{inter-medidas}} - MC_{\text{error}})$ . Y en la fórmula del CCI *promedio* hay que sumar al denominador la cantidad  $(MC_{\text{inter-medidas}} - MC_{\text{error}})/n$ .

**Intervalo de confianza.** Esta opción permite seleccionar el nivel de confianza con el que se desea obtener el intervalo de confianza para el CCI. El valor por defecto es 95 %, pero este valor puede cambiarse introduciendo un valor entre 0,0000001 y 99,99999. Las ecuaciones para obtener estos intervalos de confianza pueden encontrarse en McGraw y Wong (1996, pág. 41).

**Valor de prueba.** El procedimiento **Análisis de fiabilidad** ofrece un estadístico  $F$  (ver McGraw y Wong, 1996, pág. 42) que permite contrastar la hipótesis nula de que el  $CCI_k$  adopta, en la población, un determinado valor. Este valor es el **Valor de prueba**, el cual, por defecto, está establecido en cero (puede cambiarse introduciendo cualquier valor comprendido entre 0 y 0,9999999).

**Ejemplo: Análisis de fiabilidad > Estadísticos > Coeficiente de correlación intraclase**

Este ejemplo muestra cómo obtener e interpretar el *coeficiente de correlación intraclase* con los datos del archivo *jueces* (este archivo se encuentra en la misma carpeta en la que está instalado el SPSS). El archivo contiene las valoraciones efectuadas por 8 jueces a 300 gimnastas. Para obtener el coeficiente de correlación intraclase:

- En el cuadro de diálogo *Análisis de fiabilidad: Estadísticos* (ver Figura 20.2), seleccionar la opción **Coeficiente de correlación intraclase**.
- En el menú desplegable **Modelo**, seleccionar la opción **Dos factores, efectos aleatorios** y dejar el resto de opciones (**Tipo**, **Intervalo de confianza** y **Valor de prueba**) en su valor por defecto.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestra la Tabla 20.18. A pie de tabla se indica que se está utilizando un modelo de dos factores, ambos de efectos aleatorios. La primera nota a pie de tabla indica que se está valorando la *consistencia* (por tanto, no el *acuerdo absoluto*). Valorar la *consistencia* significa que se está interesado en averiguar si los mejores gimnastas reciben las valoraciones más altas de todos los jueces; el hecho de que haya jueces cuyas valoraciones sean sistemáticamente más altas o más bajas que las del resto de jueces no afecta a la consistencia; la consistencia únicamente se centra en el grado de relación existente entre las valoraciones. Si además del grado de parecido entre las valoraciones se desea averiguar si los promedios de las valoraciones son iguales, es necesario evaluar el *acuerdo absoluto*.

La tabla ofrece varios estadísticos para el CCI individual (*Medidas individuales*) y para el CCI promedio (*Medidas promedio*). El valor del CCI promedio es 0,898; y los límites entre los que se estima que está su verdadero valor poblacional, con una confianza del 95 % (95 % C.I.), valen 0,8870 y 0,907. El estadístico *F* y su nivel crítico (*Sig.* < 0,0005) permiten rechazar la hipótesis de que el valor poblacional del CCI promedio es cero (el valor del CCI propuesto en la hipótesis nula se indica en *Valor verdadero 0*). Este valor promedio es justamente la fiabilidad estimada para la escala mediante el coeficiente de correlación intraclase.

**Tabla 20.18.** Coeficiente de correlación intraclase

	Correlación intraclase <sup>a</sup>	Intervalo de confianza 95%		Prueba F con valor verdadero 0			
		Límite inferior	Límite superior	Valor	gl1	gl2	Sig.
Medidas individuales	.556 <sup>b</sup>	.529	.583	9.762	905	5430	.000
Medidas promedio	.898	.887	.907	9.762	905	5430	.000

Modelo de efectos aleatorios de dos factores en el que tanto los efectos de las personas como los efectos de las medidas son aleatorios.

a. Coeficientes de correlación intraclase de tipo C utilizando una definición de consistencia, la varianza inter-medidas se excluye de la varianza del denominador.

b. El estimador es el mismo, ya esté presente o no el efecto de interacción.

## Análisis no paramétrico

### El procedimiento *Pruebas no paramétricas*

En los capítulos previos se han estudiado varios procedimientos estadísticos diseñados para contrastar hipótesis sobre parámetros poblacionales tales como la media, la varianza, el coeficiente de correlación, los coeficientes de regresión, etc. Entre estos procedimientos, la prueba *t* de Student y el estadístico *F* del ANOVA son, quizá, los más representativos. Todos ellos coinciden en tres características: (1) permiten contrastar hipótesis referidas a algún parámetro ( $\mu$ ,  $\sigma^2$ ,  $\rho$ ,  $\beta$ , etc.); (2) exigen el cumplimiento de determinados supuestos sobre las poblaciones originales de las que se extraen los datos (generalmente normalidad y homocedasticidad); y (3) analizan datos obtenidos con una escala de medida de intervalo o razón.

Estas tres características combinadas permiten agrupar estos procedimientos estadísticos en una gran familia de técnicas de análisis denominada *contrastes paramétricos* o *pruebas paramétricas*. Se trata, sin duda, de las técnicas estadísticas más frecuentemente utilizadas por analistas e investigadores en todo tipo de áreas científicas. Pero su utilidad en la investigación aplicada se ve algo reducida, básicamente, por dos razones: por un lado, exigen el cumplimiento de algunos supuestos (normalidad, igualdad de varianzas, etc.) que en ocasiones pueden resultar demasiado exigentes; por otro, obligan a trabajar con unos niveles de medida (intervalo, razón) que no siempre resulta fácil alcanzar.

Afortunadamente, los contrastes paramétricos no son los únicos disponibles. Existen contrastes que permiten poner a prueba hipótesis no referidas a parámetros poblacionales; existen también contrastes que no necesitan establecer supuestos exigentes sobre las poblaciones de donde se extraen las muestras; y existen, por último, contrastes que no necesitan trabajar con datos obtenidos con una escala de medida de intervalo o razón. Esta otra familia de contrastes se conoce con el nombre de *contrastes no paramétricos* o *pruebas no paramétricas*.

Algunos autores utilizan el término *no paramétricos* para referirse únicamente a los contrastes que no plantean hipótesis sobre parámetros y que se centran en las propiedades nominales u ordinales de los datos, y añaden el término *contrastes de distribución libre* para referirse a los contrastes que no establecen supuestos (o establecen supuestos poco exigentes, como simetría o continuidad) sobre las poblaciones originales de las que se extraen las muestras. Pero lo cierto es que el incumplimiento de cualquiera de las tres características señaladas al principio puede ser considerada razón suficiente para caracterizar a un contraste como *no paramétrico*. Por tanto, aquí, utilizaremos la denominación genérica de *no paramétricos* para todos aquellos contrastes que no se ajusten a una cualquiera de las tres características de los contrastes *paramétricos* y se englobará en ese término genérico a los contrastes *de distribución libre*.

Más allá del acuerdo que pueda existir sobre esta cuestión, poner el énfasis en el *nivel de medida* de los datos contribuye a simplificar notablemente la clasificación e identificación de las distintas técnicas de análisis de datos. Por tanto, los contrastes serán clasificados tomando como referencia el *tipo de datos* que permiten analizar (independientemente del tipo de hipótesis que permitan contrastar e independientemente de los supuestos que sea necesario establecer) y serán caracterizados como *no paramétricos* siempre que no se ajusten a una cualquiera de las tres características de los contrastes *paramétricos*.

Este capítulo ofrece una descripción de las técnicas de análisis que el SPSS clasifica como *pruebas no paramétricas*. Todas ellas pueden considerarse *no paramétricas* porque no plantean hipótesis sobre parámetros o porque analizan datos obtenidos con una escala de medida débil (o mejor, datos que, aun estando medidos con una escala de intervalo o razón, se analizan aprovechando sólo sus propiedades nominales u ordinales); y muchas de ellas pueden considerarse de *distribución libre* porque no establecen supuestos demasiado exigentes sobre las poblaciones originales.

Todas estas pruebas se encuentran en la opción **Pruebas no paramétricas** del menú **Analizar**. Y aparecen ordenadas por el número de muestras que permiten analizar (el módulo SPSS *Pruebas exactas* incluye dos pruebas adicionales no incluidas en el módulo *Base*: la prueba de *Jonckheere-Terpstra* y la prueba de *homogeneidad marginal*):

- **Pruebas para una muestra:** *Chi-cuadrado* (bondad de ajuste con variables categóricas), *Binomial* (proporciones y cuantiles), *Rachas* (aleatoriedad) y *Kolmogorov-Smirnov* (bondad de ajuste con variables cuantitativas).
- **Pruebas para dos muestras independientes:** *U de Mann-Whitney*, *Kolmogorov-Smirnov*, *Reacciones extremas de Moses* y *Rachas de Wald-Wolfowitz*.
- **Pruebas para varias muestras independientes:** *H de Kruskal-Wallis* y *Mediana*.
- **Pruebas para dos muestras relacionadas:** *Wilcoxon*, *Signos* y *McNemar*.
- **Pruebas para varias muestras relacionadas:** *Friedman*, *W de Kendall* y *Q de Cochran*.

## Pruebas para una muestra

### Prueba *chi-cuadrado* para una muestra

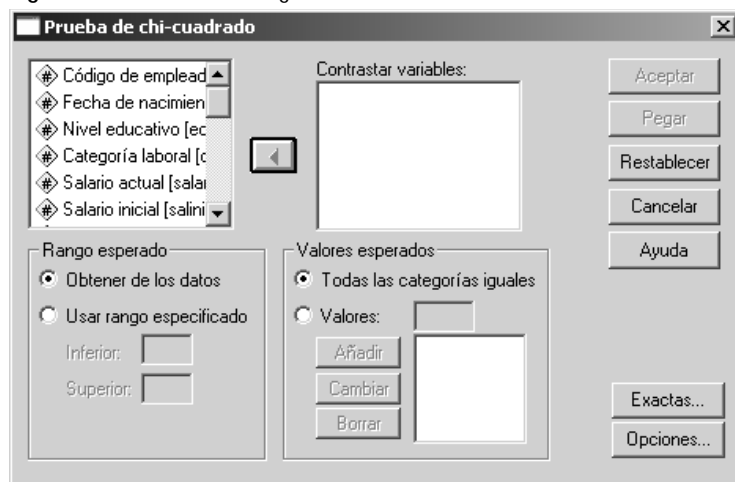
La prueba *chi-cuadrado* para una muestra permite averiguar si la distribución empírica de una variable categórica se ajusta o no (se parece o no) a una determinada distribución teórica (uniforme, binomial, multinomial, etc.). Esta hipótesis de ajuste, o mejor, de bondad de ajuste, se pone a prueba utilizando un estadístico originalmente propuesto por Pearson (1900; ver también Cochran, 1952) para comparar las frecuencias observadas o empíricas con las esperadas o teóricas de cada categoría, es decir, un estadístico diseñado para comparar las frecuencias de hecho obtenidas en una muestra concreta (frecuencias observadas:  $n_i$ ) con las frecuencias que cabría esperar encontrar si la distribución teórica de la variable fuera realmente la propuesta en la hipótesis nula (frecuencias esperadas:  $m_i$ ):

$$X^2 = \sum_i \frac{(n_i - m_i)^2}{m_i}$$

Las frecuencias esperadas  $m_i$  se obtienen multiplicando la probabilidad teórica de cada categoría  $\pi_i$  (la que corresponda de acuerdo con la hipótesis nula) por el número de casos válidos:  $m_i = n\pi_i$ . Si no existen casillas vacías y el número de frecuencias esperadas menores que 5 no superan el 20% del total de frecuencias esperadas (Cochran, 1952), el estadístico  $X^2$  se distribuye según el modelo de probabilidad *chi*-cuadrado con  $I-1$  grados de libertad (donde  $I$  se refiere al número de categorías de la variable). Para obtener la prueba *chi*-cuadrado:

- Seleccionar la opción **Pruebas no paramétricas > Chi-cuadrado...** del menú **Analizar** para acceder al cuadro de diálogo *Prueba de chi-cuadrado* que muestra la Figura 21.1.

Figura 21.1. Cuadro de diálogo *Prueba de chi-cuadrado*



La lista de variables del archivo de datos ofrece un listado de todas las variables con formato numérico. Para contrastar la hipótesis de bondad de ajuste referida a una variable categórica:

- Trasladar esa variable a la lista **Contrastar variables**. Si se selecciona más de una variable, el SPSS ofrece tantos contrastes como variables.

**Rango esperado.** Es posible decidir qué rango de valores de la variable seleccionada deben tenerse en cuenta en el análisis:

**Obtener de los datos.** Cada valor distinto de la variable se considera una categoría para el análisis.

**Usar rango especificado.** Sólo se tienen en cuenta los valores comprendidos entre los límites especificados en los cuadros de texto **Inferior** y **Superior**. Los valores no incluidos en esos límites se excluyen del análisis.

**Valores esperados.** Las opciones de este recuadro sirven para hacer explícitas las frecuencias esperadas con las que se desea comparar las observadas:

**Todas las categorías iguales.** Las frecuencias esperadas se obtienen dividiendo el número total de casos válidos entre el número de categorías de la variable. Esto equivale a efectuar el ajuste a una distribución uniforme.

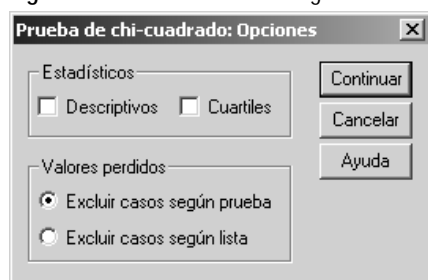
**Valores.** Esta opción permite definir frecuencias esperadas concretas. Es importante tener en cuenta que los valores que se introducen se interpretan como proporciones, no como frecuencias absolutas. Deben introducirse tantos valores como categorías: el SPSS divide cada valor por la suma de todos los valores. Así, por ejemplo, si una variable tiene dos categorías y se introducen los enteros 6 y 4, el SPSS interpreta que la frecuencia esperada de la primera categoría es 6/10 del número de casos válidos y que la frecuencia esperada de la segunda categoría es 4/10 del número de casos válidos. De esta forma, resulta fácil definir, por ejemplo, las frecuencias esperadas correspondientes a una distribución binomial o multinomial.

El orden en el que se introducen los valores es muy importante: el procedimiento hace corresponder la secuencia introducida con las categorías de la variable cuando éstas se encuentran ordenadas de forma ascendente.

**Opciones...** Este botón permite obtener algunos estadísticos descriptivos y decidir qué tratamiento se desea dar a los valores perdidos:

- Pulsar el botón **Opciones...** del cuadro de diálogo principal (ver Figura 21.1) para acceder al subcuadro de diálogo *Prueba de chi-cuadrado: Opciones* que muestra la Figura 21.2.

Figura 21.2. Subcuadro de diálogo *Prueba de chi-cuadrado: Opciones*



**Estadísticos.** Es posible obtener los siguientes estadísticos descriptivos:

- **Descriptivos.** Número de casos válidos, media, desviación típica, valor mínimo y valor máximo.
- **Cuantiles.** Centiles 25, 50 y 75.

Conviene señalar que estos estadísticos no siempre tendrán sentido, pues la prueba *chi-cuadrado* se utiliza generalmente con variables categóricas. Para contrastar la hipótesis de bondad de ajuste con variables cuantitativas es preferible utilizar la prueba de Kolmogorov-Smirnov.

**Valores perdidos.** Estas opciones permiten decidir qué tratamiento se desea dar a los valores perdidos en el caso de que se haya seleccionado más de una variable:

**Excluir casos según prueba.** Se excluyen de cada contraste los casos con valor perdido en la variable que se está contrastando. Es la opción por defecto.

**Excluir casos según pareja.** Se excluyen de todos los contrastes solicitados los casos con algún valor perdido en cualquiera de las variables seleccionadas.

### Ejemplo: Pruebas no paramétricas > Chi-cuadrado

Este ejemplo muestra cómo utilizar la prueba *chi*-cuadrado para contrastar la hipótesis de bondad de ajuste con una muestra (una variable).

El archivo *Datos de empleados ampliado* (puede obtenerse en la página *web* del manual) contiene una variable llamada *estudios* (nivel de estudios) que se ha obtenido recodificando los valores de la variable original *educ* (nivel educativo). Los años de formación académica de la variable *educ* se han recodificado en 4 categorías de estudios: primarios, secundarios, medios y superiores.

Se desea averiguar si es razonable asumir que la nueva variable *estudios* sigue una distribución uniforme. Para ello:

- En el cuadro de diálogo principal (ver Figura 21.1), seleccionar la variable *estudios* (nivel de estudios) y trasladarla a la lista **Contrastar variables**.

*Nota:* puesto que la variable utilizada es categórica, no tiene sentido solicitar la media, la desviación típica, los cuartiles, etc.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 21.1 y 21.2. La Tabla 21.1 contiene las frecuencias observadas (*N observado*) y las esperadas (*N esperado*), así como las diferencias entre ambas (*residuos*). Puesto que en el cuadro de diálogo se ha dejado marcada la opción **Todas las categorías iguales** del recuadro **Valores esperados**, todas las frecuencias esperadas (*N esperado*) son iguales. Esto es justamente lo que pronostica la hipótesis nula cuando la distribución de referencia es la uniforme.

**Tabla 21.1.** Frecuencias observadas y esperadas de la variable *estudios*

	N observado	N esperado	Residuos
Primarios	53	118,5	-65,5
Secundarios	190	118,5	71,5
Medios	181	118,5	62,5
Superiores	50	118,5	-68,5
Total	474		

La Tabla 21.2, ofrece la información necesaria para tomar una decisión sobre la hipótesis nula de bondad de ajuste: el valor del estadístico de Pearson (*Chi-cuadrado*=151,907), sus grados de libertad (*gl* = «nº de categorías menos uno» = 3) y su nivel crítico (*Sig.* < 0,0005). Puesto que el nivel crítico es menor que 0,05, se puede rechazar la hipótesis nula de bondad de ajuste y concluir que las frecuencias de las categorías de la variable *estudios* no se ajustan a una distribución uniforme.

**Tabla 21.2.** Estadístico *chi*-cuadrado

	Nivel de estudios
Chi-cuadrado <sup>a</sup>	151,907
gl	3
Sig. asintót.	,000

a. 0 casillas (,0%) tienen frecuencias esperadas menores que 5. La frecuencia de casilla esperada mínima es 118,5.



En una nota a pie de tabla se indica el número y porcentaje de casillas con frecuencias esperadas menores que 5. Puesto que el estadístico de Pearson se aproxima a la distribución *chi-cuadrado* tanto mejor cuanto mayor es el tamaño muestral, suele asumirse (siguiendo la recomendación de Cochran, 1952) que, si existen frecuencias esperadas menores que 5, éstas no deben superar el 20 % del total de frecuencias de la tabla.

## Prueba binomial

Una variable dicotómica o dicotomizada es una variable categórica que sólo toma dos valores: «éxito–fracaso», «a favor–en contra», «tratados–no tratados», «recuperados–no recuperados», «aprobados–suspensos», etc. Llamaremos, de forma genérica, *acierto* y *error* a los dos niveles de una variable de este tipo.

La prueba *binomial* permite averiguar si una variable dicotómica o dicotomizada sigue o no un determinado modelo de probabilidad. En concreto, permite contrastar la hipótesis de que la proporción observada de *aciertos* se ajusta a la proporción teórica de una distribución binomial (lo cual se traduce, según se verá enseguida, en la posibilidad de contrastar hipótesis sobre proporciones y sobre cuantiles). John Arbuthnott (1710) fue el primero en utilizar este procedimiento para demostrar que la proporción de varones nacidos en Londres en un determinado periodo de tiempo era significativamente diferente de la proporción de mujeres.

Si se extraen muestras aleatorias de tamaño  $n$  y en cada muestra se define la variable  $X$  = «número de aciertos en las  $n$  extracciones», se obtiene una variable aleatoria distribuida, si la proporción de aciertos ( $\pi$ ) permanece constante en cada extracción, según el modelo de probabilidad binomial, con parámetros  $n$  = «número de extracciones» y  $\pi$  = «proporción de aciertos». Es posible, por tanto, utilizar las probabilidades de la distribución binomial para conocer la probabilidad exacta asociada a cada uno de los valores de la variable  $X$ .

Además, a medida que  $n$  aumenta, la distribución de  $X$  se aproxima a la distribución normal con parámetros:  $E(X) = n\pi$  y  $\sigma_X = \sqrt{n\pi(1-\pi)}$ . De modo que la variable\*:

$$Z = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}}$$

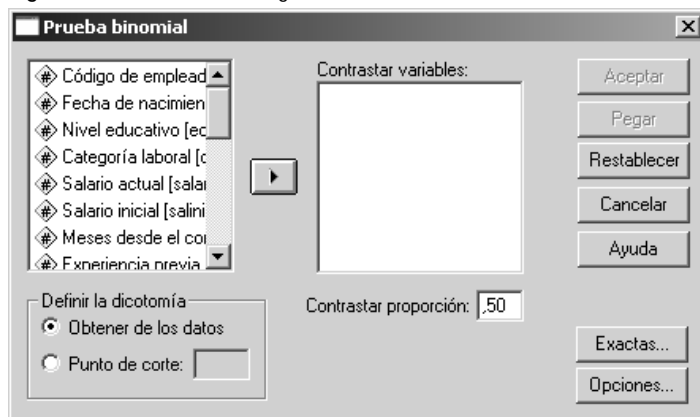
se distribuye según el modelo de probabilidad normal  $N(0, 1)$ . Es posible, también, por tanto, utilizar la distribución normal para conocer las probabilidades asociadas a los valores de  $X$ .

El SPSS utiliza ambas soluciones. Con muestras pequeñas ( $n \leq 25$ ), utiliza la distribución binomial para obtener las probabilidades exactas asociadas a los valores del estadístico  $X$ . Con muestras grandes ( $n > 25$ ) utiliza la distribución normal para obtener las probabilidades asociadas a los valores del estadístico  $Z$  (y, consecuentemente, las probabilidades aproximadas asociadas al estadístico  $X$ ).

Para obtener la prueba *binomial*:

- Seleccionar la opción **Pruebas no paramétricas > Binomial...** del menú **Analizar** para acceder al cuadro de diálogo *Prueba binomial* que muestra la Figura 21.3.

\* El SPSS utiliza la *corrección por continuidad*, que consiste en sumar (si  $X$  es menor que  $n\pi$ ) o restar (si  $X$  es mayor que  $n\pi$ ) 0,5 puntos a  $X$  para hacer el contraste algo más conservador:  $Z = [X \pm 0,5 - n\pi] / \sqrt{n\pi(1-\pi)}$ .

Figura 21.3. Cuadro de diálogo *Prueba binomial*

La lista de variables del archivo de datos ofrece un listado de todas las variables con formato numérico. Para obtener la prueba *binomial*:

- Seleccionar una o más variables y trasladarlas a la lista **Contrastar variables**. Si se traslada más de una variable, el SPSS ofrece un contraste por cada variable.

**Definir dicotomía.** Las opciones de este recuadro permiten definir qué valores de la variable seleccionada van a utilizarse como categorías:

**Obtener de los datos.** Si la variable seleccionada es dicotómica, esta opción deja que sean los propios valores de la variable los que definan la dicotomía. En ese caso, se contrasta la hipótesis de que la proporción observada en la primera categoría se ajusta (se parece) a la proporción teórica propuesta en **Contrastar proporción**.

**Punto de corte.** Si la variable seleccionada no es dicotómica es necesario dicotomizarla indicando el valor concreto que se utilizará para efectuar el corte: los valores menores o iguales que el punto de corte constituyen el primer grupo y los valores mayores el segundo. Esta opción es especialmente útil cuando lo que interesa es contrastar hipótesis sobre la mediana o sobre algún otro cuantil. Es decir, esta opción permite obtener los contrastes conocidos en la literatura estadística como *prueba de los signos* y *prueba de los cuantiles* (ver San Martín y Pardo, 1989, págs 91-97). Si se desea contrastar, por ejemplo, la hipótesis de que la mediana del *salario inicial* es 25.000 dólares (*prueba de los signos*), puede utilizarse el valor 25.000 como **Punto de corte** y 0,5 (la proporción de casos acumulados hasta la mediana) como valor del contraste en **Contrastar proporción**. Y si se desea contrastar la hipótesis de que el centil 80 del *salario inicial* vale 40.000 dólares (*prueba de los cuantiles*), puede utilizarse 40.000 como **Punto de corte** y 0,80 (la proporción de casos acumulados hasta el centil 80) como valor del contraste en el cuadro de texto **Contrastar proporción**.

Así pues, las opciones del recuadro **Definir dicotomía** permiten decidir, entre otras cosas, qué tipo de contraste se desea efectuar: sobre una *proporción* (si la variable es dicotómica) o sobre la *mediana* o cualquier otro *cuantil* (si la variable es al menos ordinal).

**Contrastar proporción.** Este cuadro de texto permite especificar el valor poblacional propuesto en la hipótesis nula. Por defecto, se asume que la variable dicotómica seleccionada sigue el modelo de distribución de probabilidad binomial con  $\pi=0,5$ . Pero este valor de prueba puede cambiarse introduciendo un valor entre 0,001 y 0,999.

De las dos categorías de la variable dicotómica, la que se toma como categoría de referencia es la que corresponde al valor del primer caso válido del archivo de datos. Teniendo esto en cuenta:

- **Si el valor de prueba es 0,5**, el SPSS interpreta que el contraste es *bilateral* y obtiene el nivel crítico multiplicando por dos la probabilidad de encontrar un número de casos igual o mayor que el de la categoría de referencia (si la proporción de casos de la categoría de referencia es mayor que 0,5), o igual o menor que el de la categoría de referencia (si la proporción de casos de la categoría de referencia es menor que 0,5).
- **Si el valor de prueba es distinto de 0,5**, el SPSS interpreta que el contraste es *unilateral* y ofrece el nivel crítico resultante de calcular la probabilidad de encontrar un número de casos igual o mayor que el de la categoría de referencia (si la proporción de casos de la categoría de referencia es mayor que el valor de prueba; contraste unilateral derecho) o igual o menor que el de la categoría de referencia (si la proporción de casos de la categoría de referencia es menor que el valor de prueba; contraste unilateral izquierdo).

El botón **Opciones...** conduce a un subcuadro de diálogo idéntico al de la Figura 21.2 que permite obtener algunos estadísticos descriptivos y decidir qué tratamiento se desea dar a los valores perdidos.

### **Ejemplo: Pruebas no paramétricas > Binomial**

Este ejemplo muestra cómo utilizar la prueba *binomial* para contrastar la hipótesis de bondad de ajuste referida a una variable dicotómica. En concreto, se va a utilizar la variable *minoría* (clasificación étnica) del archivo *Datos de empleados*. Si se asume que el 70 % de los habitantes de EE.UU. es de raza blanca, puede resultar interesante averiguar si ese porcentaje se mantiene en la entidad bancaria a la que se refiere el archivo de datos. Para ello:

- En el cuadro de diálogo principal (ver Figura 21.3), seleccionar la variable *minoría* (clasificación de minorías) y trasladarla a la lista **Contrastar variables**.
- Introducir el valor 0,70 en el cuadro de texto **Contrastar proporción** para especificar el valor de prueba.
- Puesto que la variable es dicotómica, dejar marcada la opción **Obtener de los datos del recuadro Definir dicotomía**.
- Pulsar el botón **Opciones...** para acceder al subcuadro de diálogo *Prueba binomial: Opciones* y marcar la opción **Descriptivos**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 21.3 y 21.4. La Tabla 21.3 ofrece en primer lugar el número de casos incluidos en el análisis ( $N = 474$ ). Y dado que se está analizando una variable dicotómica, la media indica la proporción de unos (*minoría* = 1 = «sí»).

Tabla 21.3. Estadísticos descriptivos

	N	Media	Desviación típica	Mínimo	Máximo
Clasificación de minorías	474	,22	,414	0	1

La Tabla 21.4 comienza identificando la variable que se está utilizando en el contraste y los dos grupos que definen la dicotomía: *Grupo 1* (*minoría* = «no») y *Grupo 2* (*minoría* = «sí»). El SPSS toma como categoría de referencia la categoría correspondiente al primer caso válido del archivo de datos: *minoría* = 0 = «no». La *Proporción observada* de casos en esa categoría es  $370/474 = 0,78$  y la *Proporción de prueba* es 0,70. Es muy importante fijarse en el valor de estas dos proporciones: puesto que el valor de prueba es distinto de 0,5 y la proporción observada de la categoría de referencia es mayor que el valor de prueba ( $0,78 > 0,70$ ), el SPSS interpreta que el contraste es unilateral derecho y ofrece, como nivel crítico, la probabilidad de obtener, con  $n = 474$  y  $\pi = 0,70$ , un número de casos igual o mayor que 370 (el número de casos de la categoría de referencia). Ese nivel crítico (*Significación asintótica unilateral*) es menor que 0,0005, por lo que se puede rechazar la hipótesis nula de bondad de ajuste ( $\pi < 0,7$ ) en favor de la alternativa ( $\pi > 0,7$ ) y concluir que la verdadera proporción poblacional es mayor que 0,70 (dado que el tamaño muestral es mayor que 25, la solución propuesta se basa en la aproximación normal, no en las probabilidades exactas de la distribución binomial).

Tabla 21.4. Prueba binomial

	Categoría	N	Proporción observada	Prop. de prueba	Sig. asintót. (unilateral)
Clasificación de minorías	Grupo 1	No	370	,78	,70
	Grupo 2	Sí	104	,22	
	Total	474	1,00		

a. Basado en la aproximación Z.

## Prueba de las rachas

La prueba de las *rachas* sirve para evaluar la *aleatoriedad* de una secuencia de observaciones, es decir, para estudiar si las observaciones de una muestra son independientes entre sí. En una serie temporal, por ejemplo, las observaciones no son aleatorias: lo que ocurre con una observación cualquiera depende, generalmente, de las características de alguna observación anterior. En una muestra aleatoria, por el contrario, debe esperarse que lo que ocurre con una observación cualquiera es independiente de las características de las observaciones anteriores.

El concepto de *racha* hace referencia a una secuencia de observaciones de un mismo tipo. Supongamos que se lanza una moneda al aire 10 veces y que se obtiene el siguiente resultado: CCCXCCXXXC. En este resultado hay 5 rachas: CCC, X, CC, XXX y C. A simple vista, el resultado obtenido parece *aleatorio*. Pero si en lugar de ese resultado se hubiera obtenido este otro: CCCCCXXXXX (sólo dos rachas) resultaría fácil ponerse de acuerdo en que la secuencia obtenida no parece aleatoria. Como tampoco parece aleatoria una secuencia con demasiadas rachas: CXCXCXCXCX (10 rachas).

Pues bien, la prueba de las rachas permite determinar si el número de rachas ( $R$ ) observado en una determinada muestra de tamaño  $n$  es lo suficientemente grande o lo suficientemente

pequeño como para poder rechazar la hipótesis de independencia (o *aleatoriedad*) entre las observaciones\*.

Para obtener el número de rachas de un conjunto de observaciones es necesario que éstas estén clasificadas en dos grupos exhaustivos y mutuamente exclusivos (variable dicotómica). Si no lo están, se deberá utilizar algún criterio (mediana, media, moda, etc.) para hacer que lo estén. Una vez clasificadas las  $n$  observaciones en dos grupos (de tamaños  $n_1$  y  $n_2$ ), el SPSS utiliza, para contrastar la hipótesis de *aleatoriedad* o independencia, una tipificación\*\* del número de rachas ( $R$ ):

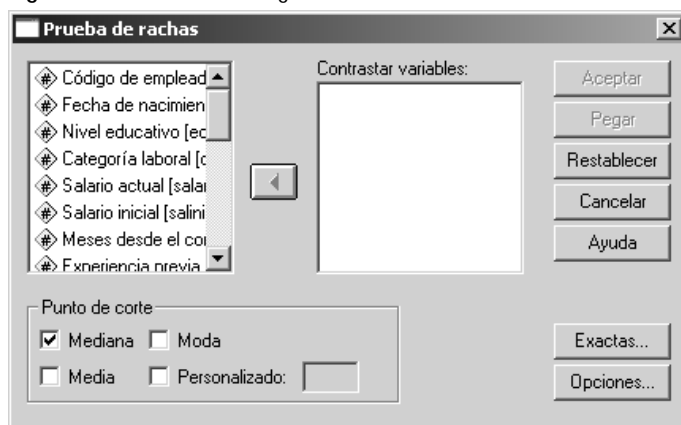
$$Z = \frac{R - E(R)}{\sigma_R}$$

donde:  $E(R) = 2n_1n_2/n + 1$  y  $\sigma_R = \sqrt{2n_1n_2(2n_1n_2 - n)/[n^2(n - 1)]}$ . El estadístico  $Z$  se distribuye según el modelo de probabilidad normal  $N(0, 1)$ . El SPSS ofrece el nivel crítico bilateral resultante de multiplicar por 2 la probabilidad de encontrar un número de rachas igual o menor que el encontrado (si  $R < E(R)$ ), o igual o mayor que el encontrado (si  $R > E(R)$ ).

Para obtener la prueba de las rachas:

- Seleccionar la opción **Pruebas no paramétricas > Rachas...** del menú **Analizar** para acceder al cuadro de diálogo *Prueba de las rachas* que muestra la Figura 21.4.

Figura 21.4. Cuadro de diálogo *Prueba de las rachas*



\* Conviene no confundir la hipótesis de *aleatoriedad* con la hipótesis de bondad de ajuste estudiada a propósito de la prueba *binomial*. Obtener 5 caras y 5 cruces al lanzar una moneda 10 veces es un resultado que se ajusta perfectamente a la hipótesis de equiprobabilidad ( $\pi_{\text{cara}} = \pi_{\text{cruz}} = 0,5$ ), pero si las 5 caras salen al principio y las 5 cruces al final, esto haría dudar de la hipótesis de independencia o *aleatoriedad*.

\*\* Si el tamaño muestral es menor que 50, el estadístico  $Z$  se obtiene utilizando la *corrección por continuidad* de la siguiente manera:

- Si  $[R - E(R)] < -0,5$ , se suma 0,5 a  $R$ . Es decir:  $Z = [R + 0,5 - E(R)] / \sigma_R$ .
- Si  $[R - E(R)] > 0,5$ , se resta 0,5 a  $R$ . Es decir:  $Z = [R - 0,5 - E(R)] / \sigma_R$ .
- Si  $|R - E(R)| \leq 0,5$ ,  $Z = 0$ .

La lista de variables del archivo de datos ofrece un listado de todas las variables con formato numérico. Para contrastar la hipótesis de *aleatoriedad* referida a una variable:

- Trasladar esa variable a la lista **Contrastar variables**. Si se traslada más de una variable, el SPSS ofrece un contraste por cada variable.

**Punto de corte.** Para poder calcular el número de rachas es necesario que las observaciones estén clasificadas en dos grupos. Si no lo están, debe utilizarse algún criterio para hacer que lo estén. Si se desea contrastar la hipótesis de independencia referida a una *variable cuantitativa*, puede utilizarse como criterio de dicotomización (como punto de corte) la **Mediana**, la **Moda** o la **Media**. En ese caso, los valores más pequeños que el punto de corte pasan a formar parte del primer grupo y los valores iguales o mayores que el punto de corte pasan a formar parte del segundo grupo. Si se desea contrastar la hipótesis de independencia referida a una *variable categórica* puede utilizarse como punto de corte la opción **Personalizado**. Si la variable es, por ejemplo, dicotómica, con códigos 0 y 1, puede utilizarse como punto de corte el valor 0,5 (o cualquier otro comprendido entre 0 y 1, incluido el 1), de modo que los casos con código 0 pasen a formar parte del primer grupo y los casos con valor 1 pasen a formar parte del segundo grupo.

El botón **Opciones...** conduce a un subcuadro de diálogo idéntico al de la Figura 21.2 que permite obtener algunos estadísticos descriptivos y decidir qué tratamiento se desea dar a los valores perdidos.

### **Ejemplo: Pruebas no paramétricas > Rachas**

Este ejemplo muestra cómo utilizar la prueba de las *rachas* para contrastar la hipótesis de *aleatoriedad* referida a la variable *salini* (salario inicial) del archivo *Datos de empleados*:

- En el cuadro de diálogo principal (ver Figura 21.4), seleccionar la variable *salini* (salario inicial) y trasladarla a la lista **Contrastar variables**.
- Dejar marcada la opción **Mediana** del recuadro **Punto de corte** para categorizar la variable utilizando la mediana (este es el criterio más comúnmente utilizado como punto de corte).

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestra la Tabla 21.5. La tabla comienza indicando el valor utilizado como punto de corte para la dicotomización: *Valor de prueba* = 15.000. Una nota a pie de tabla recuerda que ese punto de corte es la mediana.

**Tabla 21.5.** Prueba de las rachas

	Salario inicial
Valor de prueba <sup>a</sup>	\$15,000
Casos < Valor de prueba	212
Casos >= Valor de prueba	262
Casos en total	474
Número de rachas	180
Z	-5,149
Sig. asintót. (bilateral)	,000

a. Mediana

A continuación aparece el número de casos del primer grupo (*Casos* < *Valor de prueba* = 212), el número de casos del segundo grupo (*Casos* >= *Valor de prueba* = 262), el número de casos válidos (*Casos en total* = 474) y el *número de rachas* computadas (150).

La tabla ofrece, por último, el valor del estadístico de contraste ( $Z = -5,149$ ) y su nivel crítico (*Significación asintótica bilateral* < 0,0005). Puesto que el nivel crítico es muy pequeño (menor que 0,05), se puede rechazar la hipótesis de independencia y concluir que la secuencia de observaciones estudiada no es aleatoria.

## Prueba de Kolmogorov-Smirnov para una muestra

Al igual que las pruebas *chi-cuadrado* para una muestra y *binomial*, la prueba de *Kolmogorov-Smirnov* (K-S) para una muestra (Kolmogorov, 1933) también es una prueba de bondad de ajuste: sirve para contrastar la hipótesis nula de que la distribución de una variable se ajusta a una determinada distribución teórica de probabilidad. Pero a diferencia de las primeras, que han sido diseñadas más bien para evaluar el ajuste de variables categóricas, la prueba de K-S para una muestra se adapta mejor a situaciones en las que interesa evaluar el ajuste de variables cuantitativas.

Para contrastar la hipótesis nula de bondad de ajuste, la prueba de K-S se basa en la comparación de dos funciones de distribución (o funciones de probabilidad acumuladas): una función de distribución empírica  $F(X_i)$  y una función de distribución teórica  $F_0(X_i)$ .

Para obtener la función de distribución empírica  $F(X_i)$  se comienza ordenando los valores de  $X_i$  de forma ascendente, es decir, desde el valor más pequeño  $X_{[1]}$  hasta el más grande  $X_{[n]}$ . Tras esto, la función de distribución empírica para cada valor de  $X_i$  se obtiene de la siguiente manera:  $F(X_i) = i/n$  ( $i$  se refiere al rango correspondiente a cada observación).

La forma de obtener la función de distribución teórica depende de la distribución concreta propuesta en la hipótesis. Si la distribución propuesta es, por ejemplo, la *uniforme*, la función de distribución teórica para cada valor de  $X_i$  se obtiene así:  $F_0(X_i) = (X_i - X_{[1]}) / (X_{[n]} - X_{[1]})$ . Si la distribución teórica propuesta es, por ejemplo, la de Poisson, entonces la función de distribución teórica se obtiene así:  $F_0(X_i) = \sum_{l=0}^i [e^{-\lambda} \lambda^l / l!]$ . Etc.

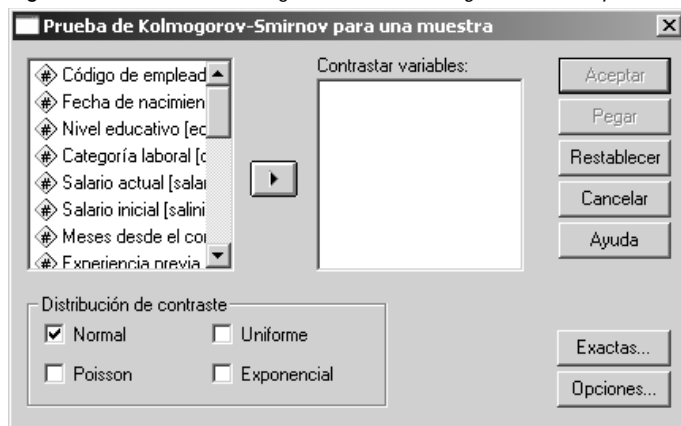
Una vez obtenidas las distribuciones empírica y teórica, el estadístico de K-S se calcula a partir de la diferencia  $D_i = F(X_i) - F_0(X_i)$  más grande existente entre ambas:

$$Z_{K-S} = \max |D_i| \sqrt{n}$$

Este estadístico  $Z$  se distribuye según el modelo de probabilidad normal  $N(0, 1)$ . El SPSS utiliza el método de Smirnov (1948) para obtener las probabilidades concretas asociadas a los valores del estadístico  $Z$ . Este método difiere del estándar (basado en las probabilidades de la curva normal estandarizada), pero es equivalente.

Para obtener la prueba de bondad de ajuste de *Kolmogorov-Smirnov* para una muestra:

- Seleccionar la opción **Pruebas no paramétricas > K-S de una muestra...** del menú **Analizar** para acceder al cuadro de diálogo *Prueba de Kolmogorov-Smirnov para una muestra* que recoge la Figura 21.5.

Figura 21.5. Cuadro de diálogo *Prueba de Kolmogorov-Smirnov para una muestra*

La lista de variables del archivo de datos ofrece un listado de todas las variables con formato numérico. Para contrastar la hipótesis de bondad de ajuste referida a una variable:

- Trasladar esa variable a la lista **Contrastar variables**. Si se traslada más de una variable, el SPSS ofrece un contraste por cada variable.

**Distribución de contraste.** Las opciones de este recuadro permiten elegir la distribución teórica a la cual se desea ajustar la distribución empírica de la variable seleccionada: **Normal**, **Uniforme**, **Poisson** y **Exponencial** (puede seleccionarse más de una). Los parámetros de las diferentes distribuciones se estiman a partir de los datos. No es posible obtener el ajuste a una distribución normal si la varianza de la variable vale cero; ni a una distribución de Poisson si la media de la variable vale cero o los valores no son, todos ellos, enteros no negativos.

El botón **Opciones...** conduce a un subcuadro de diálogo idéntico al de la Figura 21.2 que permite obtener algunos estadísticos descriptivos y decidir qué tratamiento se desea dar a los valores perdidos.

### **Ejemplo: Pruebas no paramétricas > Kolmogorov-Smirnov**

Este ejemplo muestra cómo contrastar con la prueba de *Kolmogorov-Smirnov* la hipótesis de normalidad referida a la variable *salini* (salario inicial) del archivo *Datos de empleados*:

- En el cuadro de diálogo principal (ver Figura 21.5), seleccionar la variable *salini* (salario inicial) y trasladarla a la lista **Contrastar variables**.
- Dejar marcada la opción **Normal** del recuadro **Distribución de contraste** para efectuar el ajuste a la distribución normal.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestra la Tabla 21.6. La tabla ofrece, en primer lugar, el número de casos válidos (*N*) y los parámetros de la distribución seleccionada, es decir, de la distribución normal (*Media* y *Desviación típica*). A continuación ofrece las diferencias más extremas entre las frecuencias acumuladas empíricas y teóricas (la más grande de las positivas, la más pequeña de las negativas y la más grande de las dos en



valor absoluto). Por último, ofrece el estadístico de K-S ( $Z=5,484$ ) y su nivel crítico (*Significación asintótica bilateral*  $< 0,0005$ ). Puesto que el valor del nivel crítico es muy pequeño (menor que 0,05), se puede rechazar la hipótesis de normalidad y concluir que las puntuaciones de la variable *salario inicial* no se ajustan a una distribución normal.

Tabla 21.6. Prueba de *Kolmogorov-Smirnov* para una muestra (distribución normal)

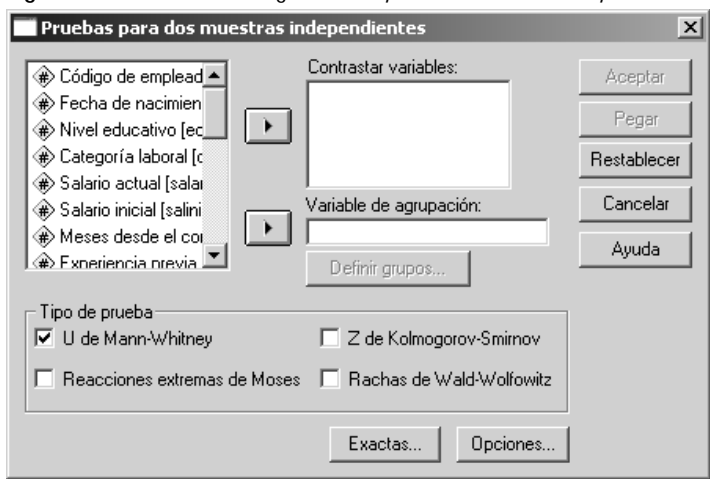
		Salario inicial
N		474
Parámetros normales	Media	\$17,016.09
	Desviación típica	\$7,870.638
Diferencias más extremas	Absoluta	,252
	Positiva	,252
	Negativa	-,170
Z de Kolmogorov-Smirnov		5,484
Sig. asintót. (bilateral)		,000

## Pruebas para dos muestras independientes

Este procedimiento contiene cuatro pruebas no paramétricas diseñadas para analizar datos provenientes de diseños con una variable independiente dicotómica (cuyos niveles definen dos grupos o muestras) y una variable dependiente cuantitativa al menos ordinal (en la cual interesa comparar los dos grupos o muestras): la prueba *U* de Mann-Whitney, la prueba de Kolmogorov-Smirnov para dos muestras, la prueba de *reacciones extremas* de Moses y la prueba de las *rachas* de Wald-Wolfowitz. Para obtener cualquiera de estas pruebas:

- Seleccionar la opción **Pruebas no paramétricas > Dos muestras independientes...** del menú **Analizar** para acceder al cuadro de diálogo *Pruebas para dos muestras independientes* que muestra la Figura 21.6.

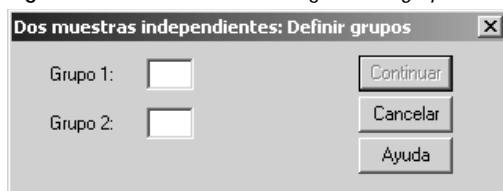
Figura 21.6. Cuadro de diálogo *Pruebas para dos muestras independientes*



La lista de variables del archivo de datos ofrece un listado de todas las variables con formato numérico. Para obtener cualquiera de las pruebas no paramétricas incluidas en el procedimiento (puede seleccionarse más de una simultáneamente):

- Seleccionar la variable en la que interesa comparar los grupos y trasladarla a la lista **Contrastar variables**. Si se traslada más de una variable, el SPSS ofrece un contraste por cada variable.
- Seleccionar la variable que define los dos grupos (muestras) que interesa comparar y trasladarla al cuadro **Variable de agrupación**.
- Pulsar el botón **Definir grupos...** para acceder al subcuadro de diálogo *Dos muestras independientes: Definir grupos* que muestra la Figura 21.7, el cual permite indicar cuáles son los dos códigos de la variable de agrupación que corresponden a los grupos (muestras) que interesa comparar.

Figura 21.7. Subcuadro de diálogo *Definir grupos*



- En el recuadro **Tipo de prueba**, marcar la opción u opciones correspondientes a las pruebas que se desea obtener. Conviene tener en cuenta que no todas ellas permiten contrastar la misma hipótesis.

El botón **Opciones...** conduce a un subcuadro de diálogo (ver Figura 21.2) que permite obtener algunos estadísticos descriptivos y controlar el tratamiento de los valores perdidos.

## Prueba *U* de Mann-Whitney

La prueba *U* de Mann-Whitney\* es una excelente alternativa a la prueba *t* sobre diferencia de medias cuando no se cumplen los supuestos en los que se basa la prueba *t* (normalidad y homocedasticidad), o cuando no es apropiado utilizar la prueba *t* porque el nivel de medida de los datos es ordinal.

Consideremos dos muestras independientes:  $Y_1$ , de tamaño  $n_1$ , e  $Y_2$ , de tamaño  $n_2$ , extraídas de la misma población o de dos poblaciones idénticas. Al mezclar las  $n_1 + n_2 = n$  observa-

---

\* La prueba *U* de Mann-Whitney fue originalmente propuesta por Wilcoxon (1945) para el caso de grupos de igual tamaño. Festinger (1946) desarrolló independientemente un procedimiento equivalente al de Wilcoxon. Pero fueron Mann y Whitney (1947) los primeros en extender el procedimiento al caso de grupos de tamaños desiguales y los primeros también en proporcionar tablas para poder utilizarlo con muestras pequeñas. Fueron precisamente las aportaciones de Mann y Whitney las que más contribuyeron a la divulgación del procedimiento; de ahí que, generalmente, sea conocido como prueba de Mann-Whitney. Sin embargo, en algunos contextos este procedimiento todavía puede encontrarse con la denominación de prueba de Wilcoxon-Mann-Whitney, o prueba de Wilcoxon para muestras independientes (la cual no debe ser confundida con la prueba de Wilcoxon para dos muestras relacionadas que se explica más adelante en este mismo capítulo).

ciones y, como si se tratara de una sola muestra, asignar rangos  $R_i$  a las  $n$  puntuaciones (un 1 a la más pequeña, un 2 a la más pequeña de las restantes, ..., un  $n$  a la más grande; resolviendo los empates asignando el rango promedio), se obtienen  $n_1$  rangos  $R_{1i}$  (los  $n_1$  rangos correspondientes a las observaciones de la muestra  $Y_1$ ) y  $n_2$  rangos  $R_{2i}$  (los  $n_2$  rangos correspondientes a las observaciones de la muestra  $Y_2$ ).

Entre los múltiples estadísticos que podrían definirse en una situación como la descrita, considérense estos dos:  $S_1$  = «suma de los rangos asignados a la muestra 1» y  $S_2$  = «suma de los rangos asignados a la muestra 2». Teniendo esto en cuenta, el estadístico  $U$  adopta la siguiente forma en cada grupo:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - S_1 \quad \text{y} \quad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - S_2$$

Puesto que se está asumiendo que las dos muestras se han extraído de dos poblaciones idénticas, cabe esperar que  $U_1$  y  $U_2$  sean aproximadamente iguales (excepto en la parte atribuible a las fluctuaciones propias del azar muestral). Si los valores de  $U_1$  y  $U_2$  son muy distintos, existirá cierta evidencia de que las muestras proceden de poblaciones distintas. Consecuentemente, la hipótesis nula de que ambos promedios poblacionales son iguales podrá rechazarse si  $U_1$  (o  $U_2$ ) es demasiado grande o demasiado pequeño.

Para determinar esto último, la decisión puede basarse en la probabilidad concreta asociada al estadístico  $U$ :

$$U = U_1 \quad \text{si } U_1 < n_1 n_2 / 2$$

$$U = U_2 \quad \text{si } U_1 > n_1 n_2 / 2$$

Con muestras pequeñas ( $n \neq 30$ ) el SPSS ofrece el nivel crítico bilateral exacto asociado al estadístico  $U$ , el cual se obtiene multiplicando por 2 la probabilidad de obtener valores menores o iguales que  $U$  (esta probabilidad se calcula utilizando el algoritmo de Dineen y Blakesley, 1973).

Con muestras grandes ( $n > 30$ ), el SPSS ofrece una tipificación\* del estadístico  $U$  (incluyendo corrección por empates) que se distribuye aproximadamente  $N(0, 1)$ :

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2}{n(n-1)} \left( \frac{n^3 - n}{12} - \sum_i \frac{t_i^3 - t_i}{12} \right)}}$$

( $k$  se refiere al número de rangos distintos en los que existen empates y  $t_i$  al número de puntuaciones empatadas en el rango  $i$ ). El nivel crítico bilateral se obtiene multiplicando por 2 la probabilidad de obtener valores menores o iguales que  $Z$ .

---

\* Existen diferentes versiones de los estadísticos  $U$  y  $Z$  (ver, por ejemplo, San Martín y Pardo, 1989, pág. 126; o Marascuilo y McSweeney, 1977, págs. 267-278), pero todas ellas son equivalentes y conducen al mismo resultado.

## Prueba de reacciones extremas de Moses

Esta prueba sirve para estudiar si existe diferencia en el grado de dispersión o variabilidad de dos distribuciones. Aunque hasta ahora sólo se ha hablado de la heterogeneidad de varianzas como de algo relacionado con los residuos del modelo lineal (ANOVA, regresión, etc.) y, por tanto, como algo poco deseable, lo cierto es que la heterogeneidad de varianzas puede constituir, ella misma, un resultado experimental relevante. Esto significa que, en ocasiones, el estudio de la variabilidad puede ser un fin en sí misma y no sólo un paso previo para la comparación de medias (ver, por ejemplo, Bryk y Raudenbush, 1988).

Supongamos que se desea evaluar el nivel de desarrollo cognitivo alcanzado por dos grupos de niños que han seguido programas educativos distintos. Si el interés del investigador se centra simplemente en constatar cuál de los dos grupos ha alcanzado, en promedio, mayor nivel de desarrollo, bastará con comparar los promedios de ambos grupos con alguno de los procedimientos (paramétricos o no paramétricos) ya estudiados. Pero esta forma de proceder pasaría por alto una cuestión de gran importancia: podría ocurrir que uno de los métodos educativos consiguiera incrementar el nivel de desarrollo de los niños de forma generalizada (todos los niños mejoran su nivel de desarrollo) y que el otro método educativo consiguiera el mismo objetivo con sólo unos pocos niños, aunque de forma más marcada, o podría ocurrir que consiguiera incrementar mucho el nivel de desarrollo de unos niños y muy poco el de otros (reacciones extremas). Estas diferencias entre métodos no quedarían reflejadas en los promedios, pero sí en la variabilidad, por lo que sólo acompañando el contraste de medias con un contraste de varianzas podría obtenerse información fiable sobre lo que está ocurriendo.

Existen diferentes procedimientos para contrastar la hipótesis de que dos varianzas poblacionales son iguales. Ya se ha estudiado (ver Capítulo 11 sobre *Análisis exploratorio*) uno de los más utilizados, debido a Levene (1960); pero se trata de un procedimiento paramétrico. Moses (1952) ha diseñado un procedimiento no paramétrico que puede utilizarse con variables ordinales.

Consideremos dos muestras ( $c = control$  y  $e = experimental$ ) extraídas aleatoriamente de la misma población o de dos poblaciones idénticas. Para obtener el estadístico de Moses se comienza ordenando las  $n = n_c + n_e$  observaciones de forma ascendente y asignándoles, como si todas las observaciones constituyeran una única muestra, rangos de 1 a  $n$ : un 1 a la más pequeña, un 2 a la más pequeña de las restantes, etc. (los empates se resuelven asignando el rango medio). A continuación se calcula la *amplitud del grupo control* ( $A_c$ ) restando los rangos correspondientes al valor más grande y más pequeño de ese grupo y sumando 1 a esa diferencia; el resultado se redondea al entero más próximo (el SPSS considera que el grupo *control* es el grupo con el código menor).

Dado que la amplitud es una medida de dispersión muy inestable, Moses sugiere utilizar al *amplitud recortada* ( $A_r$ ). Para ello, se fija un valor pequeño ( $r$ ) y se calcula la amplitud tras descartar  $r$  valores del grupo control por arriba y por abajo (en el SPSS,  $r$  es igual a la parte entera de  $0,05n_c$ , o a 1, si  $0,05n_c$  es menor que 1). La amplitud recortada se obtiene restando los rangos correspondientes al valor más grande y al más pequeño del grupo control tras eliminar del cómputo los  $r$  valores más grandes y los  $r$  valores más pequeños de ese grupo; y, por supuesto, sumando 1 a esa diferencia.

Es evidente que  $A_r$  no puede ser menor que  $n_c - 2r$  (ni mayor que  $n - 2r$ ). Además, si en el grupo experimental se han producido reacciones extremas, la amplitud del grupo control tenderá a su valor mínimo, pues habrá pocas observaciones del grupo experimental entremezcladas con las del control. Por tanto, podría resultar útil conocer la probabilidad asociada a

los valores  $A_r$  que superen en alguna cantidad el valor  $n_c - 2r$ . Llamando  $s$  a la cantidad en que un determinado valor observado de  $A_r$  supera a  $n_c - 2r$ , puede obtenerse la probabilidad de encontrar amplitudes  $A_s = n_c - 2r + s$  como la observada o menores (hasta  $n_c - 2r$ ) mediante:

$$P(A_s \leq n_c - 2r + s) = \frac{\sum_{i=0}^s \left[ \binom{i + n_c - 2r - 2}{i} \binom{n_e + 2r + 1 - i}{n_e - i} \right]}{\binom{n}{n_c}}$$

El SPSS calcula esta probabilidad tanto para  $r = 0$  como para  $r = 0,05 n_c$  (en este último caso, si  $r < 1$ , se toma 1; si  $r > 1$ , se toma la parte entera de  $r$ ). Si esta probabilidad es pequeña (menor que 0,05), se podrá rechazar la hipótesis de que ambas muestras proceden de poblaciones con la misma amplitud.

## Prueba de Kolmogorov-Smirnov para dos muestras

Esta prueba sirve para contrastar la hipótesis de que dos muestras proceden de la misma población. Para ello, compara las funciones de distribución (funciones de probabilidad acumuladas) de ambas muestras:  $F_1(X_i)$  y  $F_2(X_i)$ . A diferencia de la prueba  $U$  de Mann-Whitney (que compara dos promedios poblacionales asumiendo que ambas distribuciones tienen la misma forma y, por tanto únicamente es sensible a las diferencias entre los promedios), la prueba de Kolmogorov-Smirnov es sensible a cualquier tipo de diferencia entre las dos distribuciones: tendencia central, simetría, variabilidad, etc.

Para obtener las funciones de distribución de las dos muestras se comienza asignando rangos a los valores de  $X_i$ . Esta asignación de rangos se realiza de forma separada para cada muestra y los eventuales empates se resuelven asignando el rango promedio a las puntuaciones empatadas.

Tras asignar rangos a los valores de ambas muestras, la función de distribución empírica para cada valor de  $X_i$  se obtiene, en cada muestra, de la siguiente manera:  $F_j(X_i) = i/n_j$  (donde  $i$  se refiere al rango correspondiente a cada observación). A continuación se obtienen las diferencias  $D_i = F_1(X_i) - F_2(X_i)$ , donde  $F_1(X_i)$  se refiere a la función de distribución de la muestra de mayor tamaño. Una vez obtenidas las diferencias  $D_i$ , la hipótesis de que las dos muestras proceden de la misma población se pone a prueba utilizando una tipificación de la diferencia más grande en valor absoluto (Smirnov, 1939, 1948):

$$Z_{K-S} = \max_i |D_i| \sqrt{(n_1 n_2) / (n_1 + n_2)}$$

Este estadístico  $Z$  se distribuye según el modelo de probabilidad normal  $N(0, 1)$ . El SPSS utiliza el método de Smirnov (1948) para obtener las probabilidades concretas asociadas a los valores del estadístico  $Z$ . Este método difiere del estándar (basado en las probabilidades de la curva normal estandarizada), pero es equivalente. Si la probabilidad de obtener una diferencia tan grande como la observada es muy pequeña (generalmente, menor que 0,05), se podrá rechazar la hipótesis de que ambas muestras proceden de la misma población.

## Prueba de las rachas de Wald-Wolfowitz

La prueba de las *rachas* para dos muestras independientes (Wald y Wolfowitz, 1940) es similar a la prueba de las *rachas* para una muestra ya estudiada en este mismo capítulo. Aplicada a dos muestras independientes, permite contrastar la hipótesis de que ambas muestras proceden de la misma población. Al igual que la prueba de Kolmogorov-Smirnov para dos muestras, la de las *rachas* es sensible no sólo a diferencias entre los promedios, sino a diferencias en variabilidad, simetría, etc.

Para obtener el número de rachas, se ordenan de menor a mayor las  $n = n_1 + n_2$  observaciones de ambas muestras como si se tratara de una sola muestra. Ordenadas las puntuaciones, el número de rachas ( $R$ ) se obtiene contando el número de secuencias de observaciones pertenecientes al mismo grupo. Si existen empates entre observaciones de muestras distintas, el SPSS calcula tanto el número mínimo de rachas como el máximo.

Si las dos muestras proceden de la misma población, las observaciones ordenadas de ambas muestras estarán entremezcladas y el número de rachas será alto. Por el contrario, si las muestras proceden de poblaciones distintas, una de ellas tendrá valores más altos que la otra (las observaciones ordenadas no estarán tan entremezcladas) y el número de rachas será bajo. Por tanto, un número alto de rachas indica que las muestras proceden de la misma población y un número bajo de rachas indica que las muestras proceden de poblaciones distintas.

Para decidir cuándo el número de rachas encontrado es lo bastante pequeño como para rechazar la hipótesis de que las muestras proceden de la misma población, el SPSS utiliza dos estrategias distintas dependiendo del tamaño de las muestras. Si  $n > 30$ , utiliza la aproximación normal (ver, en este mismo capítulo, el estadístico  $Z$  descrito en el apartado *Prueba de las rachas*); pero a diferencia de lo que ocurre con el estadístico  $Z$  para una muestra, aquí se utiliza un nivel crítico unilateral: la probabilidad de obtener un número de rachas ( $R$ ) igual o menor que el obtenido ( $r$ ).

Si  $n < 30$ , el SPSS ofrece el nivel crítico unilateral exacto. Para ello, si el número observado de rachas es par, utiliza la siguiente ecuación:

$$P(R \leq r) = \frac{2}{\binom{n}{n_1}} \sum_{i=2}^r \binom{n_1-1}{\frac{i}{2}-1} \binom{n_2-1}{\frac{i}{2}-1}$$

Y si el número observado de rachas es impar:

$$P(R_i \leq r) = \frac{1}{\binom{n}{n_1}} \sum_{i=2}^r \left[ \binom{n_1-1}{k-1} \binom{n_2-1}{k-2} + \binom{n_1-1}{k-2} \binom{n_2-1}{k-1} \right]$$

(con  $i = 1, 2, \dots, r$ ; y  $k = 2r - 1$ ). En ambas ecuaciones se está calculando la probabilidad de obtener un número de rachas igual o menor que el encontrado. Se rechazará la hipótesis nula de que las muestras proceden de la misma población cuando esa probabilidad sea menor que el nivel de significación establecido (generalmente, 0,05).

**Ejemplo: Pruebas no paramétricas > Dos muestras independientes**

Este ejemplo muestra cómo obtener e interpretar todas las pruebas incluidas en el procedimiento **Pruebas no paramétricas > Dos muestras independientes...** (se sigue utilizando el archivo *Datos de empleados*):

- En el cuadro de diálogo principal (ver Figura 21.6), seleccionar la variable *salini* (salario inicial) y trasladarla a la lista **Contrastar variables**; seleccionar la variable *minoría* (clasificación de minorías) y trasladarla al cuadro **Variable de agrupación**.
- Pulsar el botón **Definir grupos...** para acceder al subcuadro de diálogo *Dos muestras independientes: Definir grupos* que muestra la Figura 21.7, e introducir los códigos 0 y 1 (que son los códigos que definen los dos grupos de la variable *minoría*). Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Marcar las cuatro opciones del recuadro **Tipo de prueba**.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 21.7 a la 21.11. La primera de ellas (Tabla 21.7) ofrece el tamaño de cada grupo (*N*), el *rango promedio* resultante de la asignación de rangos a cada grupo y la *suma* de esos rangos.

**Tabla 21.7.** Rangos

Clasificación de minorías		N	Rango promedio	Suma de rangos
Salario inicial	No	370	249,14	92180,50
	Sí	104	196,10	20394,50
	Total	474		

La Tabla 21.8 ofrece el estadístico *U* de *Mann-Whitney* (también ofrece el estadístico *W* de *Wilcoxon*, que es una versión equivalente del estadístico *U*; ver Pardo y San Martín, 1998, págs. 424-427). La tipificación de ambos vale  $Z = -3,495$ . Y el nivel crítico bilateral (*Significación asintótica bilateral*) es menor que 0,0005. Por tanto, se puede rechazar la hipótesis de igualdad de promedios y concluir que los grupos definidos por la variable *minoría* proceden de poblaciones con distinto *salario inicial* medio.

**Tabla 21.8.** Prueba de *Mann-Whitney*

	Salario inicial <sup>a</sup>
U de Mann-Whitney	14934,500
W de Wilcoxon	20394,500
Z	-3,495
Sig. asintót. (bilateral)	,000

a. Variable de agrupación: Clasificación de minorías

La Tabla 21.9 contiene la prueba de *reacciones extremas* de *Moses*. Por supuesto, la variable *minoría* no parece del todo apropiada para utilizar la prueba de *Moses*, pero, puesto que se trata de un ejemplo, sirve para poder interpretar la información que ofrece el *Visor*. La tabla recoge, en primer lugar, la *Amplitud del grupo control* (467) y la probabilidad de obtener una amplitud como esa o menor (*Significación unilateral* < 0,0005). A continuación muestra la

*Amplitud recortada del grupo control* ( $N = 434$ ) y la probabilidad de obtener una amplitud como esa o menor (*Significación unilateral* = 0,990). Puesto que el nivel crítico asociado a la amplitud recortada es mayor que 0,05, se puede considerar que no se han observado reacciones extremas. La última línea recoge el valor de  $r$ , es decir, el número de casos eliminados por arriba y por abajo para obtener la amplitud recortada.

Tabla 21.9. Prueba de *Moses*

		Salario inicial <sup>a</sup>
Amplitud observada del grupo control		467
	Sig. (unilateral)	,000
Amplitud recortada del grupo control		434
	Sig. (unilateral)	,990
Valores atípicos recortados de cada extremo		18

a. Variable de agrupación: Clasificación de minorías

La Tabla 21.10 ofrece la prueba de *Kolmogorov-Smirnov*. En primer lugar aparecen las diferencias más extremas (*Absoluta*, *Positiva* y *Negativa*) entre las funciones de distribución de ambas muestras. Y a continuación se informa del resultado de la tipificación de la diferencia más extrema en valor absoluto (*Z de Kolmogorov-Smirnov* = 2,134) acompañado de su correspondiente nivel crítico bilateral (*Significación asintótica bilateral* < 0,0005). Puesto que el nivel crítico obtenido es menor que 0,05, se puede rechazar la hipótesis de igualdad de distribuciones y concluir que los grupos comparados difieren significativamente en *salario inicial*.

Tabla 21.10. Prueba de *Kolmogorov-Smirnov*

		Salario inicial <sup>a</sup>
Diferencias más extremas	Absoluta	,237
	Positiva	,000
	Negativa	-,237
Z de Kolmogorov-Smirnov		2,134
Sig. asintót. (bilateral)		,000

a. Variable de agrupación: Clasificación de minorías

La Tabla 21.11 recoge el resultado de la prueba de las *rachas*. Contiene información sobre el número mínimo y máximo de rachas (mínimo y máximo que dependen del tratamiento que se dé a los empates), sobre el valor del estadístico  $Z$  en cada caso, y sobre el nivel crítico unilateral (*Significación asintótica unilateral*) asociado al estadístico  $Z$ . La presencia de 25 empates hace que el resultado obtenido en ambas situaciones sea muy diferente. En circunstancias como ésta es aconsejable recurrir a otras pruebas para tomar una decisión.

Tabla 21.11. Prueba de las *rachas* de *Wald-Wolfowitz*

		Número de rachas	Z	Sig. asintót. (unilateral)
Salario inicial <sup>b</sup>	Mínimo posible	40 <sup>a</sup>	-16,576	,000
	Máximo posible	200 <sup>a</sup>	4,923	1,000

a. Hay 25 empates inter-grupos que implican 348 casos.

b. Variable de agrupación: Clasificación de minorías

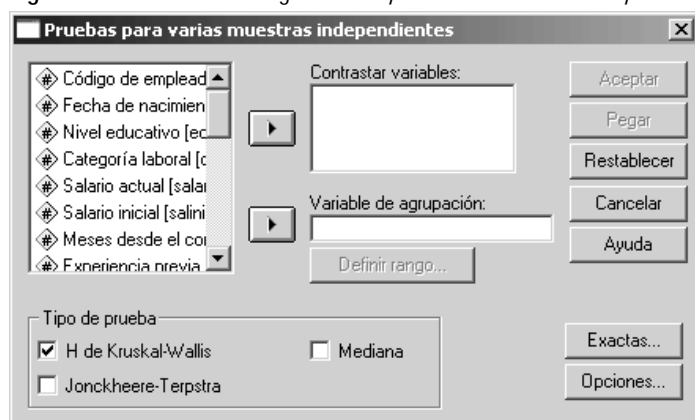


## Pruebas para varias muestras independientes

Este procedimiento contiene tres pruebas no paramétricas diseñadas para analizar datos provenientes de diseños con una variable independiente categórica (con varias categorías o niveles que definen varios grupos o muestras) y una variable dependiente cuantitativa al menos ordinal en la cual interesa comparar las muestras: la prueba  $H$  de Kruskal-Wallis, la prueba de la mediana y la prueba de Jonckheere-Terpstra (ésta última sólo se incluye en el módulo *Pruebas exactas*). Para obtener cualquiera de ellas:

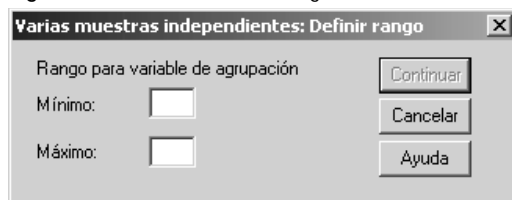
- Seleccionar la opción **Pruebas no paramétricas > Varias muestras independientes...** del menú **Analizar** para acceder al cuadro de diálogo *Pruebas para varias muestras independientes* que muestra la Figura 21.8.

Figura 21.8. Cuadro de diálogo *Pruebas para varias muestras independientes*



La lista de variables del archivo de datos ofrece un listado de todas las variables con formato numérico. Para obtener cualquiera de las pruebas no paramétricas incluidas en el procedimiento (puede seleccionarse más de una simultáneamente):

- Seleccionar la variable en la que interesa comparar los grupos y trasladarla a la lista **Contrastar variables**. Si se traslada más de una variable, el SPSS ofrece un contraste por cada variable.
- Seleccionar la variable que define los grupos (muestras) que interesa comparar y trasladarla al cuadro **Variable de agrupación**.
- Pulsar el botón **Definir grupos...** para acceder al subcuadro de diálogo *Varias muestras independientes: Definir rango* que muestra la Figura 21.9. Este subcuadro de diálogo permite indicar cuáles son los dos códigos de la variable de agrupación que corresponden a los grupos (muestras) que interesa comparar: se comparan los grupos con los códigos **Mínimo** y **Máximo** y todos los comprendidos entre ellos. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- En el recuadro **Tipo de prueba**, marcar la opción u opciones correspondientes a la(s) prueba(s) que se desea obtener.

Figura 21.9. Subcuadro de diálogo *Varias muestras independientes: Definir rango*

El botón **Opciones...** (ver Figura 21.8) conduce a un subcuadro de diálogo idéntico al de la Figura 21.2 que permite obtener algunos estadísticos descriptivos y decidir qué tratamiento se desea dar a los valores perdidos.

## Prueba $H$ de Kruskal-Wallis

La prueba de Mann-Whitney para dos muestras independientes fue extendida al caso de más de dos muestras por Kruskal y Wallis (1952). La situación experimental que permite resolver esta prueba es similar a la estudiada a propósito del ANOVA de un factor completamente aleatorizado:  $J$  muestras son aleatoria e independientemente extraídas de  $J$  poblaciones para averiguar si las  $J$  poblaciones son idénticas o alguna de ellas presenta promedios mayores que otra.

Las ventajas fundamentales de esta prueba frente al estadístico  $F$  del ANOVA de un factor completamente aleatorizado son dos: (1) no necesita establecer supuestos sobre las poblaciones originales tan exigentes como los del estadístico  $F$  (normalidad, homocedasticidad); y (2) permite trabajar con datos ordinales\*. Por contra, si se cumplen los supuestos en los que se basa el estadístico  $F$ , la potencia de éste es mayor que la que es posible alcanzar con el estadístico  $H$  de Kruskal-Wallis.

Teniendo en cuenta que en muchas situaciones reales resulta demasiado arriesgado suponer normalidad y homocedasticidad (especialmente si las muestras son pequeñas y/o los tamaños muestrales desiguales), y considerando además que en otras situaciones el nivel de medida de los datos puede no ir más allá del ordinal, la prueba de Kruskal-Wallis representa una excelente alternativa al estadístico  $F$  del ANOVA de un factor completamente aleatorizado.

Consideremos un diseño con  $J$  muestras aleatorias e independientes de tamaños  $n_1, n_2, \dots, n_J$  extraídas de la misma población o de  $J$  poblaciones idénticas, con  $n = n_1 + n_2 + \dots + n_J$  (es decir, siendo  $n$  el conjunto total de observaciones). Asignando rangos desde 1 hasta  $n$  a ese conjunto de  $n$  observaciones como si se tratara de una sola muestra (si existen empates se asigna el promedio de los rangos empatados) es posible definir los valores  $R_{ij}$  = «rangos asignados a las observaciones  $i$  de la muestra  $j$ » y  $R_j$  = «suma de los rangos asignados a las  $n_j$  observaciones de la muestra  $j$ », es decir:

$$R_j = \sum_i^{n_j} R_{ij} \quad \text{y} \quad \bar{R}_j = \frac{R_j}{n_j}$$

\* No es infrecuente encontrarse manuales de estadística en los que la prueba  $H$  de Kruskal-Wallis aparece con la denominación *análisis de varianza por rangos*.

Obviamente, si la hipótesis nula de que las  $J$  poblaciones son idénticas es verdadera, los  $R_j$  de las distintas muestras serán parecidos. Siguiendo una lógica similar a la del estadístico  $U$  de Mann-Whitney, es posible obtener, tomando como punto de partida la suma de los rangos de cada muestra, un estadístico con distribución muestral conocida capaz de ofrecer información sobre el parecido existente entre las  $J$  poblaciones (ver, por ejemplo, San Martín y Pardo, 1989, págs. 225-227)\*:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^J \frac{R_j^2}{n_j} - 3(n+1)$$

Bajo la hipótesis nula de que los  $J$  promedios poblacionales son iguales, el estadístico  $H$  se distribuye según el modelo de probabilidad *chi*-cuadrado, con  $J-1$  grados de libertad.

## Prueba de la mediana

La prueba de la mediana es similar a la prueba *chi*-cuadrado ya estudiada en el Capítulo 12 sobre *Tablas de contingencias*; la única diferencia entre ambas es que, ahora, en lugar de utilizar dos variables categóricas, una de ellas es cuantitativa y se dicotomiza utilizando la mediana (de ahí el nombre de la prueba).

Se tienen, por tanto, dos variables: una variable categórica que define  $J$  muestras de tamaño  $n_j$  (con  $n = \sum n_j$ ) y una variable al menos ordinal. El objetivo de la prueba de la mediana es contrastar la hipótesis de que las  $J$  muestras proceden de poblaciones con la misma mediana. Para ello, se comienza ordenando todas las observaciones y calculando la mediana total (la mediana de las  $n$  observaciones):

$$Mdn = (X_{[n/2]} + X_{[n/2+1]})/2 \quad \text{si } n \text{ es par}$$

$$Mdn = X_{[(n+1)/2]} \quad \text{si } n \text{ es impar}$$

(donde  $X_{[n]}$  se refiere al valor más grande y  $X_{[1]}$  al más pequeño). A continuación se registra, dentro de cada muestra, el número de casos con puntuación igual o menor que la mediana (grupo 1) y el número de casos con valor mayor que la mediana (grupo = 2). Tras esto, se construye una tabla de contingencias bidimensional de tamaño  $2 \times J$ , con las 2 filas correspondientes a los dos grupos obtenidos al dicotomizar por la mediana y las  $J$  columnas correspondientes a las  $J$  categorías de la variable categórica (es decir, a las  $J$  muestras independientes). A las frecuencias de la tabla resultante se aplica el estadístico  $X^2$  de Pearson ya estudiado en el Capítulo 12, en el apartado *Estadísticos: Chi-cuadrado*.

---

\* En el caso de que existan empates, el SPSS utiliza una corrección que hace el contraste algo más conservador:

$$H' = \frac{H}{1 - \sum_{i=1}^k (t_i^3 - t_i) / (n^3 - n)}$$

donde  $k$  se refiere al número de rangos distintos en los que existen empates y  $t_i$  al número de valores empatados en cada rango.

### Ejemplo: Pruebas no paramétricas > Varias muestras independientes

Este ejemplo muestra cómo obtener e interpretar las pruebas incluidas en el procedimiento **Pruebas no paramétricas > Varias muestras independientes...** Se sigue utilizando el archivo *Datos de empleados*:

- En el cuadro de diálogo principal (ver Figura 21.8), seleccionar la variable *salario* (salario actual) y trasladarla a la lista **Contrastar variables**; seleccionar la variable *cat-lab* (categoría laboral) y trasladarla al cuadro **Variable de agrupación**.
- Pulsar el botón **Definir grupos...** para acceder al subcuadro de diálogo *Varias muestras independientes: Definir grupos* que muestra la Figura 21.9, e introducir los códigos 1 y 3 para indicar que se desea comparar los grupos 1 y 3 y todos los comprendidos entre ellos (es decir, los grupos 1, 2 y 3). Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Marcar las opciones **H de Kruskal-Wallis** y **Mediana** del recuadro **Tipo de prueba**.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 21.12 a la 21.15. La primera de ellas (Tabla 21.12) ofrece el tamaño de cada grupo (*N*) y los *Rangos promedio* resultantes de la asignación de rangos a las puntuaciones de los tres grupos.

Tabla 21.12. Rangos

	Categoría laboral	N	Rango promedio
Salario actual	Administrativo	363	190,37
	Seguridad	27	278,98
	Directivo	84	427,85
	Total	474	

La Tabla 21.13 contiene el estadístico de Kruskal-Wallis ( $\text{Chi-cuadrado}=207,679$ ), sus grados de libertad ( $gl=2$ ) y su nivel crítico (*Significación asintótica*  $< 0,0005$ ). Puesto que el nivel crítico es menor que 0,05, se puede rechazar la hipótesis de igualdad de promedios poblacionales y concluir que las poblaciones comparadas difieren en *salario actual*.

Tabla. 21.13. Prueba de *Kruskal-Wallis*

	Salario actual <sup>a</sup>
Chi-cuadrado	207,679
gl	2
Sig. asintót.	,000

a. Variable de agrupación: Categoría laboral

Aunque la literatura estadística ofrece procedimientos para efectuar comparaciones múltiples tras obtener un estadístico *H* significativo (ver, por ejemplo, Pardo y San Martín, 1998, págs. 437-441), para analizar con el SPSS qué categorías laborales difieren entre sí se puede utilizar la prueba de Mann-Whitney para dos muestras independientes acompañada de la corrección de Bonferroni para controlar la tasa de error (probabilidad de cometer errores de tipo I). Esta corrección consiste en utilizar un nivel de significación igual a 0,05 dividido por el número de comparaciones por pares que se desea realizar. Con tres categorías laborales hay que hacer

tres comparaciones por pares (1-2, 1-3 y 2-3). Por tanto, la aplicación de la corrección de Bonferroni llevará a tomar decisiones con un nivel de significación de  $0,05/3=0,017$ . Es decir, se considerará que dos grupos difieren significativamente cuando el nivel crítico obtenido sea menor que 0,017.

Las Tablas 21.14 y 21.15 contienen la información relacionada con la prueba de la *mediana*. La Tabla 21.14 muestra el resultado de la dicotomización, es decir, el resultado de clasificar a los sujetos de cada *categoría laboral* por debajo y por encima de la mediana de la variable *salario*. Como dato interesante puede observarse que no existen *directivos* que estén por debajo de la mediana.

Tabla 21.14. Frecuencias

	Categoría laboral		
	Administrativo	Seguridad	Directivo
Salario actual > Mediana	128	25	84
<= Mediana	235	2	0

La Tabla 21.15 ofrece el tamaño de la muestra total (*N*), el valor de la mediana (*Mediana*), el estadístico  $X^2$  de Pearson (*Chi-cuadrado*), sus grados de libertad (*gl*) y el nivel crítico asintótico asociado al estadístico (*Sig. asintót.*). Puesto que el nivel crítico es menor que 0,05, se puede rechazar la hipótesis de igualdad de promedios poblacionales y concluir que las poblaciones comparadas difieren en *salario actual*.

Tabla 21.15. Prueba de la mediana

	Salario actual <sup>a</sup>
N	474
Mediana	\$28,875.00
Chi-cuadrado	135,133 <sup>b</sup>
gl	2
Sig. asintót.	,000

a. Variable de agrupación: Categoría laboral

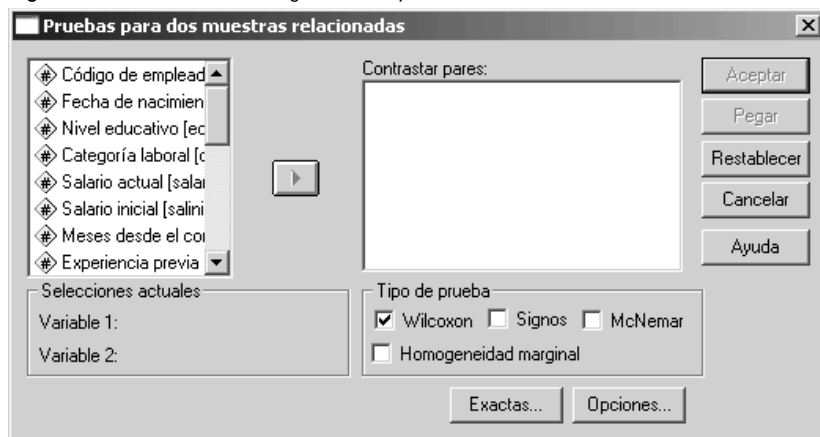
b. 0 casillas (,0%) tienen frecuencias esperadas menores que 5. La frecuencia de casilla esperada mínima es 13,5.

## Pruebas para dos muestras relacionadas

Las pruebas de este apartado permiten analizar datos provenientes de diseños con medidas repetidas. Las pruebas de *Wilcoxon* y de los *Signos* sirven para contrastar hipótesis sobre igualdad de promedios; la prueba de *McNemar* sirve para contrastar hipótesis sobre igualdad de proporciones (ver el Capítulo 12 sobre *Tablas de contingencias* para una descripción de esta prueba). Todas ellas se ajustan a diseños del tipo *antes-después*, pero difieren en el tipo de variables que permiten analizar. Para obtener cualquiera de estas pruebas:

- Seleccionar la opción **Pruebas no paramétricas > Dos muestras relacionadas...** del menú **Anali-**  
**zar** para acceder al cuadro de diálogo *Pruebas para dos muestras relacionadas* que reco-

ge la Figura 21.10.

Figura 21.10. Cuadro de diálogo *Pruebas para dos muestras relacionadas*

La lista de variables del archivo de datos ofrece un listado de todas las variables con formato numérico. Para obtener cualquiera de ellas (la prueba de *homogeneidad marginal* sólo está disponible si se ha instalado el módulo *Pruebas exactas*):

- Seleccionar las variables cuyas medianas o proporciones interesa comparar y trasladarlas a la lista **Contrastar pares**. Puede seleccionarse más de un par de variables: el SPSS ofrece un contraste por cada par seleccionado.

El recuadro **Selecciones actuales** muestra las variables seleccionadas antes de pulsar el botón flecha. Y el botón **Opciones...** conduce a un subcuadro de diálogo, idéntico al de la Figura 21.2, que permite obtener algunos estadísticos descriptivos y decidir qué tratamiento se desea dar a los valores perdidos.

## Prueba de Wilcoxon

Se toman dos medidas ( $X_i$  e  $Y_i$ ) a un grupo de  $m$  sujetos y se calculan las diferencias en valor absoluto entre las dos puntuaciones de cada par:

$$D_i = |X_i - Y_i| \quad (i = 1, 2, \dots, m)$$

Se desechan las  $D_i$  nulas y únicamente se consideran las  $n$  diferencias  $D_i$  no nulas ( $n < m$ ). Se asignan *rangos* ( $R_i$ ) desde 1 hasta  $n$  a esas  $D_i$  no nulas: el rango 1 a la  $D_i$  más pequeña, el rango 2 a la  $D_i$  más pequeña de las restantes, ..., el rango  $n$  a la  $D_i$  más grande (si existen empates, se resuelven asignando el promedio de los rangos). Se suman, por un lado, los  $R_i^+$ , es decir, los rangos correspondientes a las  $D_i$  con  $X_i > Y_i$ , y se llama  $S_+$  a esta suma; se suman, por otro lado, los  $R_i^-$ , es decir, los rangos correspondientes a las  $D_i$  con  $X_i < Y_i$ , y se llama  $S_-$  a esta otra suma. Hecho esto, si se asume que las puntuaciones  $X_i$  e  $Y_i$  proceden de poblaciones con la misma mediana ( $Mdn_x = Mdn_y$ ), cabe esperar que:

$$P(X_i < Y_i) = P(X_i > Y_i)$$

por lo que, si la hipótesis  $H_0: Mdn_x = Mdn_y$  es verdadera, en una muestra aleatoria de  $n$  observaciones cabe encontrar aproximadamente tantos valores  $X_i > Y_i$  como valores  $X_i < Y_i$  (salvando, por supuesto, las fluctuaciones atribuibles al azar muestral). Pero, además, si la distribución de las diferencias es *simétrica* (lo cual exige escala de intervalo o razón), las  $D_i$  positivas se alejarán de cero *en igual medida* que las  $D_i$  negativas, por lo que es fácil deducir que:

$$S_+ = \sum R_i^+ \approx S_- = \sum R_i^-$$

Es decir, si  $Mdn_x = Mdn_y$  y la distribución de las diferencias  $D_i$  es simétrica,  $S_+$  y  $S_-$  tomarán valores parecidos. De modo que una fuerte discrepancia entre  $S_+$  y  $S_-$  hará dudar de la veracidad de  $H_0$ . Por tanto, los valores  $S_+$  y  $S_-$  pueden utilizarse para obtener información sobre la hipótesis  $H_0: Mdn_x = Mdn_y$  (Wilcoxon, 1945, 1949).

Con tamaños muestrales pequeños no resulta complicado obtener la distribución exacta de  $S_+$  o  $S_-$  (ver, por ejemplo, Pardo y San Martín, 1998, págs. 420-422). Pero es más rápido obtener una tipificación de  $S$  ( $S$  se refiere al menor de  $S_+$  y  $S_-$ ) cuya distribución se aproxima, conforme el tamaño muestral va aumentando, al modelo de probabilidad normal  $N(0, 1)$ :

$$Z = \frac{S - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \sum_{i=1}^k \frac{t_i^3 - t_i}{48}}}$$

( $k$  se refiere al número rangos distintos en los que existen empates y  $t_i$  al número de puntuaciones empatadas en el rango  $i$ ). El SPSS ofrece el nivel crítico bilateral resultante de multiplicar por 2 la probabilidad de obtener valores menores o iguales que  $Z$ .

## Prueba de los signos

La prueba de los signos guarda una muy estrecha relación con la prueba *binomial* ya estudiada anteriormente en este mismo capítulo. Al igual que la prueba de Wilcoxon, la prueba de los signos permite contrastar la hipótesis de igualdad entre dos medianas poblacionales. Pero, mientras la prueba de Wilcoxon aprovecha la información ordinal de los datos (aunque exige nivel de medida de intervalo o razón), la prueba de los signos sólo aprovecha de los datos sus propiedades nominales (aunque exige nivel de medida al menos ordinal).

La situación es similar a la estudiada a propósito de la prueba de Wilcoxon. Se toman dos medidas ( $X_i$  e  $Y_i$ ) a un grupo de  $m$  sujetos y se calculan las diferencias:

$$D_i = |X_i - Y_i| \quad (i = 1, 2, \dots, m)$$

entre las dos puntuaciones de cada par. Se desechan las  $D_i$  nulas y únicamente se consideran las  $n$  diferencias  $D_i$  no nulas ( $n \leq m$ ). Si se asume que las puntuaciones  $X_i$  e  $Y_i$  proceden de poblaciones con la misma mediana ( $Mdn_x = Mdn_y$ ), debe verificarse que:

$$P(X_i < Y_i) = P(X_i > Y_i) = 0,5$$

de modo que, si la hipótesis  $H_0: Mdn_x = Mdn_y$  es verdadera, al seleccionar una muestra aleatoria de  $n$  observaciones y medir en ella las variables  $X_i$  e  $Y_i$  cabe esperar encontrar aproximadamente tantos valores  $X_i > Y_i$  como valores  $X_i < Y_i$ , es decir, aproximadamente tantas diferencias  $D_i$  positivas como negativas (salvando, por supuesto, las fluctuaciones atribuibles al azar propio del proceso de muestreo). Bajo estas condiciones, las variables:

$n_+$  = número de signos positivos

$n_-$  = número de signos negativos

se distribuyen según el modelo binomial con parámetros  $n$  y  $\pi = 0,50$ . De modo que puede utilizarse la distribución binomial para conocer las probabilidades asociadas a  $n_+$  y  $n_-$  y, a partir de ellas, contrastar la hipótesis  $H_0: Mdn_x = Mdn_y$ .

Si  $n < 25$ , el SPSS toma el valor  $r = \min(n_+, n_-)$  y, utilizando las probabilidades de la distribución binomial, calcula el nivel crítico bilateral resultante de multiplicar por 2 la probabilidad de obtener valores iguales o menores que  $r$ .

Si  $n > 25$ , el SPSS tipifica el valor de  $r$  (utilizando *corrección por continuidad*) y ofrece el nivel crítico resultante de multiplicar por 2 la probabilidad de encontrar valores iguales o menores que  $Z$ :

$$Z = \frac{r + 0,5 - n/2}{0,5 \sqrt{n}}$$

### Ejemplo: Pruebas no paramétricas > Dos muestras relacionadas

Este ejemplo muestra cómo obtener e interpretar las pruebas incluidas en el procedimiento **Pruebas no paramétricas > Dos muestras relacionadas...** Se sigue utilizando el archivo *Datos de empleados*:

- En el cuadro de diálogo principal (ver Figura 21.10), seleccionar las variables *exp-prev* (experiencia previa) y *tiempemp* (meses desde el contrato) y trasladarlas a la lista **Contrastar pares**.
- En el recuadro **Tipo de prueba**, marcar las opciones **Signos** y **Wilcoxon** (para una descripción de cómo aplicar e interpretar la prueba de McNemar, consultar el Capítulo 12 sobre *Tablas de contingencias*).
- Pulsar el botón **Opciones** para acceder al subcuadro de diálogo *Pruebas para dos muestras relacionadas: Estadísticos* y marcar las opciones **Descriptivos** y **Cuartiles**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 21.16 a la 21.20.

La Tabla 21.16 comienza ofreciendo algunos estadísticos descriptivos para las dos variables incluidas en el análisis: el número de casos válidos en ambas variables ( $N$ ), la media, la desviación típica, el valor más pequeño (*Mínimo*), el más grande (*Máximo*) y los tres cuartiles (*Percentiles* 25, 50 y 75). La media de meses de experiencia previa es mayor que la media de meses desde el contrato. Se trata de averiguar si esa diferencia muestral es lo bastante grande como para pensar que las medias de las respectivas poblaciones son distintas.



Tabla 21.16. Estadísticos descriptivos

	N	Media	Desv. típ.	Mínimo	Máximo	Percentiles		
						25	50 (Mediana)	75
Meses desde el contrato	474	81,11	10,061	63	98	72,00	81,00	90,00
Experiencia previa (meses)	474	95,86	104,586	0	476	19,00	55,00	140,00

Las dos tablas siguientes contienen información relacionada con la prueba de *Wilcoxon*. La Tabla 21.17 ofrece el número, media y suma de los rangos negativos y de los rangos positivos. Las notas a pie de tabla permiten conocer el significado de los rangos positivos y negativos. También ofrece el número de empates (casos que no son incluidos en el análisis) y el número total de sujetos. La Tabla 21.18 muestra el estadístico de *Wilcoxon* (*Z*) y su nivel crítico bilateral (*Sig. asintót. bilateral*). Puesto que el valor del nivel crítico (0,450) es mayor que 0,05, no se puede rechazar la hipótesis de igualdad entre los promedios comparados.

Tabla 21.17. Rangos

		N	Rango promedio	Suma de rangos
Experiencia previa - Meses desde el contrato	Rangos negativos	290 <sup>a</sup>	199,35	57810,50
	Rangos positivos	181 <sup>b</sup>	294,73	53345,50
	Empates	3 <sup>c</sup>		
	Total	474		

a. Experiencia previa (meses) < Meses desde el contrato

b. Experiencia previa (meses) > Meses desde el contrato

c. Experiencia previa (meses) = Meses desde el contrato

Tabla 21.18. Prueba de *Wilcoxon*

	Experiencia previa - Meses desde el contrato
Z	-,755 <sup>a</sup>
Sig. asintót. (bilateral)	,450

a. Basado en los rangos positivos.

Las Tablas 21.19 y 21.20 contienen la información relacionada con la prueba de los *signos*. La Tabla 21.19 muestra las diferencias negativas, las positivas y los empates entre cada par de puntuaciones; las notas a pie de tabla permiten saber qué diferencias se están considerando negativas y cuáles positivas. Dado que el tamaño muestral es mayor que 25, la Tabla 21.20 ofrece el estadístico *Z* (−4,976) y su correspondiente nivel crítico bilateral (*Sig. asintót. bilateral* < 0,0005). Puesto que el valor del nivel crítico es menor que 0,05, puede rechazarse la hipótesis de igualdad de promedios y concluir que los promedios de las variables comparadas, *experiencia previa* y *meses desde el contrato*, difieren significativamente.

El resultado tan distinto al que se llega con ambas pruebas (*Wilcoxon* y *signos*) se debe al tipo de información en que se basa cada una. Mientras que la prueba de *Wilcoxon* tiene en cuenta el tamaño de cada diferencia (aprovecha información cuantitativa), la prueba de los *signos* únicamente considera el signo de la diferencia (sólo aprovecha información cualitativa). Por tanto, la decisión sobre qué prueba utilizar debe basarse en la naturaleza de las variables analizadas.

Tabla 21.19. Frecuencias

	N
Experiencia previa - Diferencias negativas <sup>a</sup>	290
Meses desde el contrato Diferencias positivas <sup>b</sup>	181
Empates <sup>c</sup>	3
Total	474

a. Experiencia previa (meses) < Meses desde el contrato

b. Experiencia previa (meses) > Meses desde el contrato

c. Experiencia previa (meses) = Meses desde el contrato

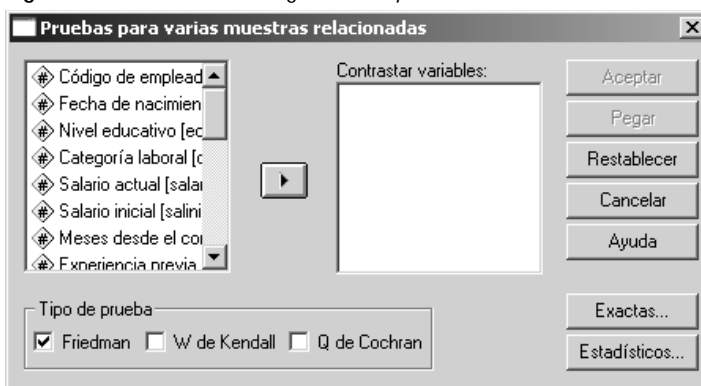
Tabla 21.20. Prueba de los signos

	Experiencia previa - Meses desde el contrato
Z	-4,976
Sig. asintót. (bilateral)	,000

## Pruebas para varias muestras relacionadas

Las pruebas agrupadas en este apartado permiten analizar datos provenientes de diseños con medidas repetidas. La prueba de Friedman y el coeficiente de concordancia  $W$  de Kendall sirven para estudiar  $J$  medidas ordinales; la prueba de Cochran permite contrastar la hipótesis de igualdad de proporciones con  $J$  variables dicotómicas. Para obtener cualquiera de ellas:

- Seleccionar la opción **Pruebas no paramétricas > Varias muestras relacionadas...** del menú **Analizar** para acceder al cuadro de diálogo *Pruebas para varias muestras relacionadas* que recoge la Figura 21.11.

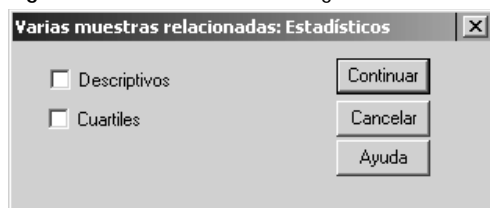
Figura 21.11. Cuadro de diálogo *Pruebas para varias muestras relacionadas*

La lista de variables del archivo de datos ofrece un listado de todas las variables con formato numérico. Para obtener cualquiera de las pruebas no paramétricas incluidas en el procedimiento (puede seleccionarse más de una simultáneamente, pero debe tenerse presente que no tiene sentido mezclar la prueba de Cochran con las de Friedman y Kendal):

- Seleccionar las variables cuyas medianas o proporciones interesa comparar y trasladarlas a la lista **Contrastar variables**.

El botón **Estadísticos...** da acceso al subcuadro de diálogo *Varias muestras relacionadas: Estadísticos* que muestra la Figura 21.12. Este subcuadro de diálogo permite obtener varios estadísticos descriptivos (tamaño muestral, media, desviación típica, valor mínimo y máximo) y los cuartiles.

Figura 21.12. Subcuadro de diálogo *Varias muestras relacionadas: Estadísticos*



## Prueba de Friedman

La prueba de Friedman (1937) sirve para comparar  $J$  promedios poblacionales cuando se trabaja con muestras relacionadas. La situación experimental que permite resolver esta prueba es similar a la estudiada a propósito del ANOVA de un factor con *medidas repetidas*: a  $n$  sujetos (o a  $n$  bloques, cada uno de tamaño  $J$ ) se le aplican  $J$  tratamientos o se le toman  $J$  medidas con intención de averiguar si los promedios de esos  $J$  tratamientos o medidas son diferentes. Las ventajas de esta prueba frente al estadístico  $F$  del ANOVA son las mismas que se han señalado a propósito del estadístico  $H$  de Kruskal-Wallis: no es necesario establecer los supuestos tan exigentes del estadístico  $F$  del ANOVA (normalidad, igualdad de varianzas) y permite trabajar con datos ordinales. La prueba de Friedman, por tanto, constituye una alternativa al estadístico  $F$  cuando no se cumplen los supuestos paramétricos del ANOVA o el nivel de medida de los datos es ordinal.

El diseño está formado por  $J$  muestras o tratamientos relacionados y por una muestra aleatoria de  $n$  sujetos o bloques independientes entre sí e independientes de los tratamientos. Las puntuaciones originales deben ser transformadas en rangos  $R_{ij}$ . Esos rangos se asignan independientemente para cada sujeto o bloque; es decir, se asignan rangos de 1 a  $J$  a las observaciones del sujeto o bloque 1; lo mismo con el bloque 2; y lo mismo con el resto de los bloques por separado.

Los rangos asignados a cada sujeto o bloque suman, en todos los casos,  $J(J+1)/2$  (pues en cada sujeto o bloque se están asignando rangos desde 1 a  $J$ ). Llamando  $R_{ij}$  al rango asignado al sujeto o bloque  $i$  en el tratamiento o muestra  $j$ , y  $R_j$  a la suma de los rangos asignados a las  $n$  observaciones de la muestra  $j$ :

$$R_j = \sum_i^n R_{ij} \quad \rightarrow \quad \bar{R}_j = \frac{R_j}{n}$$

Obviamente, si los promedios poblacionales son iguales, los  $R_j$  serán parecidos. Tomando como punto de partida estas sumas de rangos, Friedman (1937) ha diseñado un estadístico con

distribución muestral conocida que ofrece información sobre el parecido existente entre las  $J$  poblaciones (ver, por ejemplo, Pardo y San Martín, 1998, págs. 441-445):

$$X_r^2 = \frac{12}{nJ(J+1)} \sum_j R_j^2 - 3n(J+1)$$

El estadístico  $X_r^2$  de Friedman se distribuye según el modelo de probabilidad *chi*-cuadrado con  $J-1$  grados de libertad.

## Coeficiente de concordancia $W$ de Kendall

El coeficiente de concordancia  $W$  (obtenido independientemente por Kendall y Babington-Smith, 1939, y por Wallis, 1939) sirve para estudiar la relación (acuerdo, concordancia) entre  $J > 2$  conjuntos de rangos. La necesidad de estudiar la relación entre  $J$  conjuntos de rangos se presenta con cierta frecuencia en diferentes áreas de conocimiento. Tales situaciones se producen, por ejemplo, cuando una muestra aleatoria de  $n$  sujetos u objetos es clasificada según  $J$  características; o cuando  $J$  jueces evalúan, ordenan o clasifican una muestra de  $n$  sujetos u objetos según una característica. Cualquiera que sea la forma de obtener ese conjunto de  $J$  rangos, llamaremos  $R_{ij}$  al rango correspondiente al sujeto u objeto  $i$  en la característica  $j$ , o al rango asignado al sujeto u objeto  $i$  por el juez  $j$ ; y  $R_i$  a la suma de los rangos correspondientes al sujeto u objeto  $i$ :

$$R_i = \sum_{j=1}^J R_{ij}$$

Existe concordancia perfecta entre  $J$  conjuntos de rangos cuando todos los jueces valoran o clasifican a los  $n$  sujetos u objetos del mismo modo (es decir, cuando los jueces coinciden plenamente en sus juicios) o cuando los  $n$  sujetos u objetos son clasificados de idéntica manera en las  $J$  características consideradas. Cuando esto ocurre, todos los jueces coinciden en asignar el rango 1 a uno de los sujetos u objetos, el rango 2 a otro de los sujetos u objetos, ..., el rango  $n$  a otro de los sujetos u objetos. Esto significa que los totales  $R_i$  correspondientes a los diferentes sujetos u objetos serán:  $1J, 2J, 3J, \dots, iJ, \dots, nJ$ .

Por el contrario, no existe concordancia entre  $J$  conjuntos de rangos cuando los  $n$  sujetos u objetos son valorados o clasificados de diferente forma por los  $J$  jueces (es decir, cuando los jueces no coinciden en sus juicios) o cuando los  $n$  sujetos u objetos son clasificados de diferente manera en las  $J$  características consideradas. Cuando esto ocurre, a uno de los sujetos u objetos le corresponden rangos de 1 a  $n$ , a otro de los sujetos u objetos le corresponden igualmente rangos de 1 a  $n$ , y lo mismo con el resto de los sujetos u objetos. Lo cual implica que, en el caso de concordancia nula, los totales  $R_i$  correspondientes a los diferentes sujetos u objetos serán iguales:

$$R_1 = R_2 = \dots = R_i = \dots = R_n = \frac{J(n+1)}{2}$$

(pues la suma de los  $J$  conjuntos de rangos vale  $Jn(n+1)/2$ ). Así pues, el grado de concordancia existente queda reflejado en la variabilidad entre los totales  $R_i$  de los diferentes sujetos u

objetos. Cuando la concordancia entre  $J$  conjuntos de rangos es perfecta, la variabilidad entre los  $R_i$  es máxima; cuando la concordancia es nula, la variabilidad entre los  $R_i$  es mínima. Teniendo esto en cuenta, puede definirse el estadístico:

$$S = \sum_{i=1}^n \left( R_i - \frac{J(n+1)}{2} \right)^2$$

el cual representa la variabilidad observada entre cada total  $R_i$  y el total que cabría esperar si la concordancia fuera nula.  $S$  valdrá cero cuando la concordancia existente sea nula (pues, en ese caso, todos los totales  $R_i$  serán iguales entre sí e iguales a  $J(n+1)/2$ ) y alcanzará su valor máximo en el caso de concordancia perfecta, es decir, cuando entre los totales  $R_i$  exista la máxima variabilidad:

$$S_{\text{máx}} = \frac{J^2 n(n^2 - 1)}{12}$$

Ahora bien, si se desea obtener un coeficiente que valga 0 en el caso de concordancia nula y 1 en el caso de concordancia perfecta\* puede resultar útil una transformación consistente en dividir  $S$  entre su valor máximo posible. Esta solución es justamente lo que se conoce como coeficiente de concordancia  $\hat{W}$  de Kendall\*\*:

$$\hat{W} = \frac{12 \sum_i R_i^2}{J^2 n(n^2 - 1)} - \frac{3(n+1)}{n-1}$$

Cuando entre  $J$  conjuntos de rangos existe concordancia máxima,  $\hat{W}$  vale 1; cuando se da concordancia nula,  $\hat{W}$  vale 0.

Para poder afirmar que existe concordancia significativa entre  $J$  conjuntos de rangos es necesario hacer inferencias sobre el parámetro  $W$ . Esto, en realidad, tiene fácil solución pues  $\hat{W}$  es fácilmente transformable en el estadístico  $X_r^2$  de Friedman (ver apartado anterior):

$$X_r^2 = J(n-1) \hat{W}$$

\* Con  $J$  conjuntos de rangos no tiene sentido un coeficiente con valores negativos, pues no es posible encontrar un desacuerdo total. Si entre dos conjuntos de rangos existe relación perfecta negativa, el tercer conjunto de rangos necesariamente estará relacionado con uno de los dos anteriores o con ninguno de ellos; y lo mismo vale decir del cuarto, y del quinto, etc.; y eso es algo de lo que no tiene sentido hablar en términos negativos.

\*\* La presencia de empates dentro de un mismo conjunto de rangos hace que  $\hat{W}$  tome un valor más pequeño del que le corresponde. El SPSS utiliza el coeficiente de Kendall aplicando una corrección por empates:

$$\hat{W} = \frac{12 \sum_i R_i^2 - 3J^2 n(n+1)^2}{J^2 n(n^2 - 1) - J \sum_1^k (t_i^3 - t_i)}$$

donde  $t_i$  se refiere al número de puntuaciones empatadas en un rango dado y  $k$  al número de rangos distintos en los que se produce empate.

De hecho, el coeficiente  $\hat{W}$  de Kendall y el estadístico  $X_r^2$  de Friedman son aplicables al mismo tipo de situaciones. Debido a que los rangos se asignan independientemente dentro de cada sujeto o bloque, mantener la hipótesis de que las distribuciones poblacionales son idénticas dentro de cada sujeto o bloque utilizando el estadístico de Friedman (es decir, mantener la hipótesis de igualdad de tratamientos) es exactamente la misma cosa que mantener mediante el coeficiente de concordancia de Kendall la hipótesis de que las sumas (los totales  $R_i$ ) de los  $J$  rangos asignados a cada sujeto u objeto son iguales (es decir, la hipótesis de concordancia nula).

## Prueba de Cochran

El estudio de más de dos proporciones relacionadas es una generalización del procedimiento expuesto en el Capítulo 12 para el caso de dos proporciones relacionadas. Se sigue trabajando con una variable que sólo puede tomar dos valores (variable dicotómica o dicotomizada), pero con más de dos ( $J > 2$ ) muestras relacionadas.

El diseño es bastante simple: a  $n$  sujetos se le toman  $J$  medidas de una variable dicotómica, o  $J$  variables dicotómicas son medidas en una muestra de  $n$  sujetos. Se trata de un diseño idéntico al presentado a propósito del ANOVA de un factor con medidas repetidas (o bloques con un sujeto por nivel y bloque), pero con la diferencia de que, aquí, la variable medida (la variable dependiente) es una variable dicotómica, es decir, una variable que sólo puede tomar dos valores.

Las proporciones marginales  $P_{+j}$  representan las proporciones de *aciertos* de cada muestra o tratamiento:  $P_{+j} = T_{+j}/n$  (siendo  $T_{+j}$  la suma de *aciertos* de cada muestra). Si las  $J$  muestras proceden de poblaciones idénticas, cabe esperar que las proporciones marginales  $P_{+j}$  sean iguales, excepto en la parte atribuible a las fluctuaciones propias del azar muestral. Basándose en este hecho, Cochran (1950) ha diseñado un procedimiento\* que permite poner a prueba la hipótesis de igualdad entre  $J$  proporciones poblacionales ( $H_0: \pi_{+1} = \pi_{+2} = \dots = \pi_{+J}$ ):

$$Q = \frac{J(J-1) \sum T_{+j}^2 - (J-1)T^2}{JT - \sum T_{i+}^2}$$

El estadístico  $Q$  de Cochran se distribuye según  $\chi^2$  con  $J-1$  grados de libertad. El SPSS ofrece como nivel crítico la probabilidad de obtener valores iguales mayores que el encontrado.

## Ejemplo: Pruebas no paramétricas > Varias muestras relacionadas

Este ejemplo muestra cómo obtener e interpretar las pruebas incluidas en el procedimiento **Pruebas no paramétricas > Varias muestras relacionadas**... Se ofrecen dos ejemplos distintos: uno para la prueba de Friedman y para el coeficiente de concordancia  $W$  (con datos al menos ordinales) y otro para la prueba de Cochran (con datos nominales).

\* Este procedimiento es generalización del de McNemar para dos proporciones relacionadas. De hecho, si  $J = 2$ , el estadístico de McNemar y el de Cochran son exactamente el mismo (ver, por ejemplo, Conover, 1980, pág. 204).

### Prueba de Friedman y coeficiente de concordancia *W* de Kendall

Este ejemplo muestra cómo aplicar la prueba de *Friedman* y el *coeficiente de concordancia W de Kendall* a las valoraciones emitidas por 8 jueces a una muestra de 300 gimnastas. Los datos se encuentran en el archivo *Jueces (Judges)*, en la misma carpeta en la que está instalado el SPSS.

Los datos se ajustan a un diseño de un factor con ocho niveles (los ocho jueces) y una variable dependiente (las valoraciones de los jueces). Puesto que todos los sujetos son calificados por todos los jueces, se trata de un factor de medidas repetidas. Y hablar de *medidas repetidas* es equivalente a hablar de *muestras relacionadas*.

Dadas las características del diseño, tanto la prueba de Friedman como el coeficiente de concordancia *W* son estadísticos apropiados para obtener información de estos datos. No obstante, las hipótesis que permiten contrastar, aunque equivalentes, son distintas. El estadístico de Friedman contrasta la hipótesis de que los ocho promedios poblacionales comparados son iguales; el coeficiente de concordancia *W* contrasta la hipótesis de concordancia nula, es decir, la hipótesis de que los ocho conjuntos de puntuaciones comparados son independientes entre sí.

Para aplicar la prueba de *Friedman* y el *coeficiente de concordancia W de Kendall* a estos datos:

- En el cuadro de diálogo principal (ver Figura 21.11), seleccionar todas las variables del archivo (ocho en total) y trasladarlas a la lista **Contrastar variables**.
- En el recuadro **Tipo de prueba**, marcar las opciones **Friedman** y ***W* de Kendall**.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Varias muestras relacionadas: Estadísticos* (ver Figura 21.12) y marcar las opciones **Descriptivos** y **Cuartiles**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 21.21 a la 21.24.

La Tabla 21.21 ofrece algunos estadísticos descriptivos de las ocho variables seleccionadas: el número de casos válidos en todas ellas, la media, la desviación típica, el valor más pequeño, el más grande y los tres cuartiles.

**Tabla 21.21.** Estadísticos descriptivos

	N	Media	Desv. típ.	Mínimo	Máximo	Percentiles		
						25	50 (Mediana)	75
Italia	300	8,4960	,86742	7,00	10,00	7,8000	8,5000	9,2000
Corea del Sur	300	8,9183	,81992	7,10	10,00	8,3000	9,0000	9,7000
Rumanía	300	8,0853	,81732	7,00	9,80	7,4000	8,0000	8,7000
Francia	300	8,9703	,67726	7,20	9,90	8,5000	9,1000	9,5750
China	300	8,0380	,67360	7,00	9,50	7,5000	7,9000	8,5000
Estados Unidos	300	8,8763	,95931	7,00	10,00	8,1000	9,1000	9,8000
Rusia	300	8,1813	,97893	7,00	10,00	7,3000	8,0000	9,2000
Entusiasta de butaca	300	8,4697	1,03761	7,00	10,00	7,5000	8,4000	9,5000

La Tabla 21.22 recoge, para cada variable, los rangos medios resultantes del proceso de asignación de rangos.

Tabla 21.22. Rangos

	Rango promedio
Italia	4,39
Corea del Sur	6,59
Rumanía	2,44
Francia	6,53
China	2,48
Estados Unidos	6,41
Rusia	2,75
Entusiasta de butaca	4,41

La Tabla 21.23 ofrece los resultados de la *prueba de Friedman*. Incluye el número de casos válidos ( $N=300$ ), el valor del estadístico de Friedman (*Chi-cuadrado*=1212,907), sus grados de libertad ( $gl=7$ ) y el nivel crítico asociado al estadístico (*Sig. asintót.* < 0,0005). Puesto que el nivel crítico obtenido es menor que 0,05, se puede rechazar la hipótesis de igualdad de promedios poblacionales y concluir que la valoraciones promedio efectuadas por los distintos jueces no son iguales; o lo que es lo mismo, hay al menos un juez cuya valoración promedio difiere de la de al menos otro juez.

Aunque la literatura estadística recoge procedimientos para efectuar comparaciones múltiples cuando el estadístico de Friedman resulta significativo (ver, por ejemplo, Pardo y San Martín, 1998, pág. 447), para analizar con el SPSS qué variables difieren entre sí puede utilizarse la prueba de Wilcoxon para dos muestras relacionadas, pero aplicando la corrección de Bonferroni para controlar la tasa de error (ver más arriba el ejemplo de la prueba de *Kruskal-Wallis*).

Tabla 21.23. Prueba de *Friedman*

N	300
Chi-cuadrado	1212,907
gl	7
Sig. asintót.	,000

La Tabla 21.24 contiene la información relacionada con el *coeficiente de concordancia W de Kendall*. La tabla muestra el número de casos válidos ( $N = 300$ ), el valor del estadístico *W* (0,578), su tipificación (*Chi-cuadrado* = 1212,907, la cual toma exactamente el mismo valor que el estadístico de Friedman; ver Tabla 21.23), sus grados de libertad ( $gl = 7$ ) y el nivel crítico asociado a esa tipificación (*Sig. asintót.* < 0,0005). Puesto que el valor del nivel crítico es menor que 0,05, se puede rechazar la hipótesis de concordancia nula y concluir que entre las valoraciones de los jueces existe acuerdo significativo.

Tabla 21.24. Coeficiente de concordancia *W* de *Kendall*

N	300
W de Kendall	,578
Chi-cuadrado	1212,907
gl	7
Sig. asintót.	,000



Conviene recordar en este momento que, aunque la prueba de Friedman permite comparar los promedios de  $J$  variables ordinales y el estadístico  $W$  de Kendall permite contrastar la presencia de asociación entre  $J$  variables ordinales, lo cierto es que tratándose de datos ordinales ambas cosas son equivalentes. Dado que los rangos se asignan independientemente para cada sujeto, sólo es posible encontrar asociación entre  $J$  conjuntos de rangos si existen al menos dos promedios que difieren entre sí, y viceversa. De hecho, según se ha explicado ya, el estadístico  $X_r^2$  de Friedman y el del coeficiente de concordancia  $W$  de Kendall toman exactamente el mismo valor (uno es transformación del otro) y se distribuyen de la misma manera.

### Prueba de Cochran

Para ilustrar la prueba de Cochran se va a utilizar una encuesta realizada a 906 espectadores de televisión sobre los motivos por los que estarían dispuestos a seguir viendo un determinado programa en la siguiente temporada. Estos datos están disponibles en el archivo *tv-survey*, el cual se encuentra en la misma carpeta en la que está instalado el SPSS. Las siete variables del archivo (siete motivos) son dicotómicas: 1 = «sí», 0 = «no».

Puesto que todos los sujetos responden a las siete preguntas, se trata de un diseño de medidas repetidas o, lo que es lo mismo, de muestras relacionadas. Y dado que la variable dependiente o *respuesta* es dicotómica, lo relevante es analizar *proporciones*. En consecuencia, lo apropiado en una situación de este tipo será utilizar la prueba de Cochran para el contraste de  $J$  proporciones relacionadas. Para aplicar la prueba de Cochran a estos datos:

- En el cuadro de diálogo principal (ver Figura 21.11), seleccionar todas las variables (siete en total) y trasladarlas a la lista **Contrastar variables**.
- En el recuadro **Tipo de prueba**, marcar la opción Cochran.
- Pulsar el botón **Estadísticos...** para acceder al subcuadro de diálogo *Varias muestras relacionadas: Estadísticos* y marcar la opción **Descriptivos**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 21.25 a la 21.27.

La Tabla 21.25 ofrece, para cada variable seleccionada, algunos descriptivos básicos: el número de casos válidos (no hay casos con valor perdido, la media (que al tratarse de variables dicotómicas no es otra cosa que la proporción de «unos»), la desviación típica insesgada, y los valores mínimo y máximo.

**Tabla 21.25.** Estadísticos descriptivos

	N	Media	Desviación típica	Mínimo	Máximo
Cualquier motivo	906	,49	,500	0	1
A esa hora no hay otros programas populares	906	,50	,500	0	1
El programa tiene todavía buenas críticas	906	,50	,500	0	1
Otras personas todavía ven el programa	906	,53	,499	0	1
Los guionistas originales permanecen en el programa	906	,81	,389	0	1
Los directores originales permanecen en el programa	906	,83	,378	0	1
Los actores originales siguen en el programa	906	,89	,315	0	1

La Tabla 21.26 muestra el número (frecuencia) de respuestas de cada tipo observadas en cada pregunta (recuérdese que: 1 = «sí», 0 = «no»).

Tabla 21.26. Frecuencias

	Valor	
	0	1
Cualquier motivo	465	441
A esa hora no hay otros programas populares	451	455
El programa tiene todavía buenas críticas	450	456
Otras personas todavía ven el programa	427	479
Los guionistas originales permanecen en el programa	168	738
Los directores originales permanecen en el programa	156	750
Los actores originales siguen en el programa	101	805

La Tabla 21.27 ofrece el número de casos válidos ( $N=10$ ), el estadístico de Cochran ( $Q$  de Cochran = 1491,561), sus grados de libertad ( $gl=6$ ) y su nivel crítico (*Sig. asintót.* < 0,000). Puesto que el nivel crítico es menor que 0,05, se puede rechazar la hipótesis de igualdad de proporciones y concluir que la proporción de televidentes que elige cada motivo no es la misma.

Tabla 21.27. Prueba de Cochran

N	906
Q de Cochran	1491,561 <sup>a</sup>
gl	6
Sig. asintót.	,000

a. 0 se trata como un éxito.

Aunque la literatura estadística recoge procedimientos para efectuar comparaciones múltiples cuando el estadístico de Cochran resulta significativo (ver, por ejemplo, Pardo y San Martín, 1998, págs. 508-510), para contrastar con el SPSS qué proporciones difieren entre sí puede utilizarse la prueba de *McNemar* para dos muestras relacionadas (ya estudiada en el Capítulo 12), pero acompañada de la corrección de Bonferroni para controlar la tasa de error (ver, en este mismo capítulo, el ejemplo sobre la prueba de *Kruskal-Wallis*).



## Análisis de variables de respuesta múltiple

### El procedimiento *Respuestas múltiples*

#### Variables de respuesta múltiple

La expresión *variables de respuesta múltiple* se utiliza para identificar variables en las que los sujetos pueden dar más de una respuesta, es decir, variables en las que un mismo sujeto puede tener valores distintos.

En una investigación de tipo social, por ejemplo, podría pedirse a los encuestados: *Señale cuál de los siguientes tipos de transporte público urbano ha utilizado durante el último mes: autobús, metro, tren, taxi*. Obviamente, un mismo sujeto podría marcar más de una respuesta: por esta razón, a la cuestión planteada se le llama variable de respuesta múltiple. Otro ejemplo; en una investigación médica podría pedirse a los pacientes que indicaran cuál de una serie de síntomas han padecido durante la última semana; puesto que los pacientes pueden marcar más de un síntoma, de nuevo la cuestión planteada es una variable de respuesta múltiple.

Al intentar codificar variables de respuesta múltiple surge un problema: el SPSS sólo permite utilizar variables con un único código para cada caso. Por esta razón, para trabajar con variables de respuesta múltiple es necesario utilizar más de una variable. Esto puede hacerse siguiendo dos estrategias distintas.

La primera estrategia consiste en crear tantas variables *dicotómicas* como alternativas de respuesta tiene la pregunta. En el ejemplo sobre transporte público habría que crear cuatro variables: *autobús, metro, tren y taxi*. Si un encuestado marca la respuesta *autobús*, tendrá en la primera variable un «uno»; si no la marca, un «cero». Cada encuestado tendrá unos en las variables que haya marcado y ceros en las que no haya marcado. Esta forma de codificar las variables de respuesta múltiple se llama método o estrategia de *dicotomías múltiples*.

La segunda estrategia de codificación de variables de respuesta múltiple consiste en crear tantas variables *categorías* como respuestas distintas hayan dado los sujetos. Si hay algún sujeto que ha marcado las cuatro respuestas, hay que crear cuatro variables; por ejemplo: *resp1, resp2, resp3 y resp4*. Si ningún sujeto ha marcado más de tres respuestas, bastará con crear tres variables; etc. Todas las variables categóricas se codifican ahora con cuatro valores: 1 = «autobús», 2 = «metro», 3 = «tren» y 4 = «taxi». En la primera variable, a cada sujeto se le asigna el código correspondiente a su primera respuesta; en la segunda variable, el código correspondiente a su segunda respuesta (si existe); etc. Si un sujeto responde, por ejemplo, *autobús y taxi*, en la variable *resp1* tendrá un 1 (código correspondiente a *autobús*) y en la variable *resp2* tendrá un 4 (código correspondiente a *taxi*); en el resto de variables tendrá códigos de valor perdido; si un sujeto únicamente responde *metro*, en la variable *resp1* tendrá un

2 (código correspondiente a *metro*) y en el resto de variables códigos de valor perdido; etc. Esta forma de codificar las variables de respuesta múltiple se denomina método o estrategia de *categorías múltiples*.

Aunque toda variable de respuesta múltiple puede codificarse utilizando cualquiera de las dos estrategias (*dicotomías* o *categorías múltiples*), las características de la propia variable pueden hacer más recomendable una u otra. En la pregunta sobre transportes públicos, puesto que las posibles respuestas son sólo cuatro, el método de dicotomías múltiples puede resultar tan válido como el de categorías múltiples, y es más rápido y sencillo. Pero cuando el número de posibles respuestas es muy alto (un listado de, por ejemplo, 25 síntomas) y los sujetos sólo marcan unas pocas respuestas, es más apropiado el método de categorías múltiples.

La Figura 22.1 muestra la codificación asignada a las respuestas de una muestra de 20 encuestados a los que se les ha preguntado por el tipo de transporte público urbano que utilizan (los datos están disponibles en el archivo *Transportes públicos*, el cual puede encontrarse en la página web del manual). El archivo recoge, en primer lugar, las variables *id* (número de identificación) y *sexo* (1=«varones» y 2=«mujeres»). A continuación aparecen cuatro variables dicotómicas (*autobús*, *metro*, *tren* y *taxi*) en las que el valor 1 indica que se ha utilizado ese transporte y el valor 0 que no se ha utilizado (método de dicotomías múltiples). Las últimas tres variables ofrecen la misma información que las cuatro variables dicotómicas, pero en formato de categorías múltiples. En este segundo formato, puesto que ningún sujeto ha marcado los cuatro transportes (el que más ha marcado ha marcado tres), sólo es necesario crear tres variables categóricas. El primer sujeto, por ejemplo, ha utilizado el *autobús* y el *tren*, luego en la variable *resp\_1* tiene un 1 (código correspondiente a *autobús*) y en la variable *resp\_2* tiene un 3 (código correspondiente a *tren*); y como ya no ha marcado ninguna respuesta más, en la variable *resp\_3* tiene un 0 (que funciona como código de valor perdido).

Figura 22.1. Datos correspondientes a una muestra de 20 encuestados

	id	genero	autobus	metro	tren	taxi	resp_1	resp_2	resp_3
1	1	1	1	0	1	0	1	3	0
2	2	1	1	1	0	0	1	2	0
3	3	1	1	1	1	0	1	2	3
4	4	1	1	0	1	0	1	3	0
5	5	1	0	1	1	0	2	3	0
6	6	1	0	0	0	1	4	0	0
7	7	1	1	0	1	0	1	3	0
8	8	1	0	1	1	0	2	3	0
9	9	1	0	1	0	1	2	4	0
10	10	1	1	1	1	0	1	2	3
11	11	2	1	1	0	0	1	2	0
12	12	2	0	1	1	0	2	3	0
13	13	2	0	1	0	0	2	0	0
14	14	2	1	1	1	0	1	2	3
15	15	2	0	1	1	0	2	3	0
16	16	2	1	0	1	0	1	3	0
17	17	2	0	1	0	1	2	4	0
18	18	2	0	1	1	0	2	3	0
19	19	2	1	0	0	1	1	4	0
20	20	2	0	1	1	1	2	3	4

Si ahora se desea conocer, utilizando un procedimiento SPSS convencional, cuántos sujetos utilizan cada medio de transporte, se debe describir cada variable de forma individual, por se-

parado. Así, por ejemplo, con el procedimiento **Estadísticos descriptivos > Frecuencias...** del menú **Analizar** se obtienen los resultados que muestra la Tabla 22.1. La tabla informa sobre la frecuencia de uso de cada medio de transporte, pero por separado, es decir, utilizando una tabla distinta para cada medio de transporte. Por el contrario, el procedimiento **Respuestas múltiples** permite ordenar las frecuencias como muestra la Tabla 22.2, donde cada variable dicotómica constituye una categoría de la variable de respuesta múltiple.

**Tabla 22.1.** Frecuencia de uso del transporte público urbano: variables tabuladas por separado

**Autobús**

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	No utilizado	10	50,0	50,0	50,0
	Utilizado	10	50,0	50,0	100,0
	Total	20	100,0	100,0	

**Metro**

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	No utilizado	6	30,0	30,0	30,0
	Utilizado	14	70,0	70,0	100,0
	Total	20	100,0	100,0	

**Tren**

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	No utilizado	7	35,0	35,0	35,0
	Utilizado	13	65,0	65,0	100,0
	Total	20	100,0	100,0	

**Taxi**

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	No utilizado	15	75,0	75,0	75,0
	Utilizado	5	25,0	25,0	100,0
	Total	20	100,0	100,0	

**Tabla 22.2.** Frecuencia de uso del transporte público urbano: variables tabuladas juntas

		Respuestas		Porcentaje de casos
		Nº	Porcentaje	
Transporte público	Autobús	10	23,8%	50,0%
	Metro	14	33,3%	70,0%
	Tren	13	31,0%	65,0%
	Taxi	5	11,9%	25,0%
Total		42	100,0%	210,0%

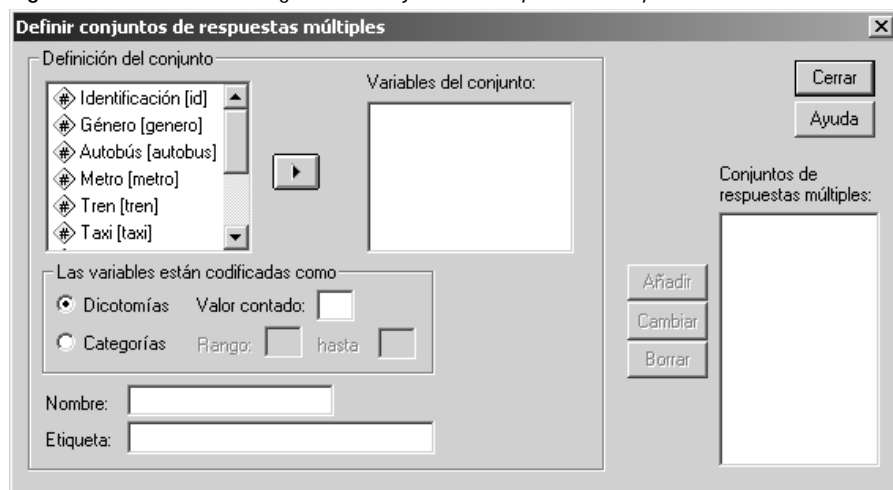
Por supuesto, para obtener los resultados de la Tabla 22.2 es necesario haber definido previamente las variables de respuesta múltiple; en los apartados que siguen se explica cómo definir variables de respuesta múltiple y cómo obtener tablas de frecuencias. También se explica cómo obtener tablas de contingencias combinando dos o tres variables de respuesta múltiple y combinando variables individuales con variables de respuesta múltiple.

## Definir conjuntos de respuestas múltiples

Antes de poder obtener distribuciones de frecuencias y tablas de contingencias con variables de respuesta múltiple es necesario definir unas nuevas variables llamadas *conjuntos* de respuestas múltiples. El procedimiento **Respuestas múltiples** permite definir tanto *conjuntos de categorías múltiples* (agrupando variables categóricas) como *conjuntos de dicotomías múltiples* (agrupando variables dicotómicas). Se pueden definir hasta 20 conjuntos de respuestas múltiples. Para definir un conjunto:

- Seleccionar la opción **Respuestas múltiples > Definir conjuntos...** del menú **Analizar** para acceder al cuadro de diálogo *Definir conjuntos de respuestas múltiples* que muestra la Figura 22.2.

Figura 22.2. Cuadro de diálogo *Definir conjuntos de respuestas múltiples*



La lista de variables del archivo de datos muestra un listado de todas las variables *numéricas*. Para crear un conjunto se debe comenzar seleccionando las variables que se desea incluir en el conjunto (al menos dos) y trasladándolas a la lista **Variables del conjunto**.

**Las variables están codificadas como.** Las variables individuales que van a formar parte del conjunto pueden estar codificadas de dos maneras: como dicotomías o como categorías (ver apartado anterior). Esta circunstancia hay que indicársela al SPSS seleccionando una de estas dos opciones:

**Dicotomías.** Debe utilizarse esta opción cuando las variables individuales que van a formar parte del conjunto están codificadas como variables dicotómicas (como ocurre con las variables *autobús*, *metro*, *tren* y *taxi* en el archivo de datos de la Figura 22.1). En el cuadro de texto **Valor contado** hay que introducir el valor de la variable que debe computarse. Generalmente, las variables dicotómicas se codifican con «unos» (para la presencia de la característica observada) y «ceros» (para la ausencia de la característica observada); por tanto, generalmente, el valor que habrá que intro-

ducir en el cuadro de texto **Valor contado** será un «uno». Cada variable que contenga al menos una aparición del valor contado se convierte en una categoría del nuevo conjunto de dicotomías múltiples.

**Categorías.** Esta opción debe utilizarse cuando las variables individuales que van a formar parte del conjunto están codificadas como variables categóricas (como ocurre con las variables *resp1*, *resp2* y *resp3* en el archivo de la Figura 22.1). El nuevo conjunto de respuestas múltiples tendrá el mismo rango de valores (categorías) que las variables individuales que lo componen; las cuales, a su vez, deben estar categorizadas de idéntica manera. Para definir ese rango de categorías es necesario introducir, en los cuadros de texto **Rango** y **hasta**, los números enteros que identifican los valores mínimo y máximo entre los que están incluidos los códigos correspondientes a las categorías de las variables individuales.

**Nombre.** Cada nuevo conjunto de variables debe tener un nombre único y debe ajustarse a las reglas propias de los nombres de variable del SPSS. Conviene tener en cuenta que existen algunas palabras reservadas que no se pueden utilizar como nombres de variable: *casenum*, *sysmis*, *jdate*, *date*, *time*, *length* y *width*.

**Etiqueta.** Este cuadro de texto permite introducir, si se desea, una etiqueta descriptiva del conjunto recién nombrado. La etiqueta puede tener hasta 40 caracteres.

**Conjuntos de respuestas múltiples.** Una vez que se han seleccionado las variables que formarán parte del conjunto, que se ha indicado si las variables están codificadas como dicotomías o como categorías y que se ha asignado un nombre al nuevo conjunto, el botón **Añadir** permite hacer efectivas todas estas definiciones incluyendo el nuevo conjunto en la lista **Conjuntos de respuestas múltiples**. Seleccionando un conjunto previamente añadido a la lista de **Conjuntos de respuestas múltiples**, los botones **Borrar** y **Cambiar** permiten eliminarlo o modificarlo.

Los conjuntos definidos sólo pueden utilizarse con los procedimientos **Frecuencias** y **Tablas de contingencias** del menú **Respuestas múltiples**. No están disponibles en el resto de procedimientos SPSS.

### ***Ejemplo: Respuestas múltiples > Definir conjuntos de respuestas múltiples***

Este ejemplo muestra cómo definir conjuntos de respuestas múltiples a partir de las variables dicotómicas y categóricas del archivo de la Figura 22.1 (estos datos están disponibles en el archivo *Transportes públicos*, el cual puede encontrarse en la página web del manual). Para definir un conjunto de **dicotomías múltiples**:

- En el cuadro de diálogo principal (ver Figura 22.2), seleccionar las variables *autobús*, *metro*, *tren* y *taxi* y trasladarlas a la lista **Variables del conjunto**.
- Marcar la opción **Dicotomías** del recuadro **Las variables están codificadas como** e introducir el valor 1 en el cuadro de texto **Valor contado**.
- Para asignar nombre y etiqueta al nuevo conjunto, escribir *trans\_d* en el cuadro de texto **Nombre** y *Transporte público (dicotomías)* en el cuadro de texto **Etiqueta**.
- Pulsar el botón **Añadir** para trasladar (y con ello definir) el conjunto a la lista **Conjuntos de respuestas múltiples**.



Para definir un conjunto de *categorías múltiples*:

- En el cuadro de diálogo *Definir conjuntos de respuestas múltiples* (ver Figura 22.2), seleccionar las variables *resp1*, *resp2* y *resp3* y trasladarlas a la lista **Variables del conjunto**.
- Marcar la opción **Categorías** del recuadro **Las variables están codificadas como** e introducir los valores 1 y 4 en los cuadros de texto **Rango** y **hasta**.
- Para asignar nombre y etiqueta al nuevo conjunto, escribir *trans\_c* en el cuadro de texto **Nombre** y *Transporte público (categorías)* en el cuadro de texto **Etiqueta**.
- Pulsar el botón **Añadir** para trasladar (y con ello definir) el conjunto a la lista **Conjuntos de respuestas múltiples**.

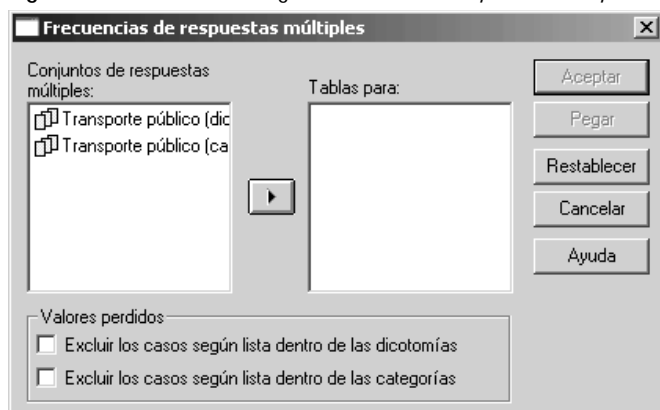
A partir de este momento, los conjuntos recién creados (*trans\_d* y *trans\_c*) estarán disponibles en los cuadros de diálogo asociados al menú **Respuestas múltiples** y podrán utilizarse para obtener tablas de frecuencias y de contingencias.

## Tablas de frecuencias

Para obtener tablas de frecuencias con variables de respuesta múltiple:

- Seleccionar la opción **Respuestas múltiples > Frecuencias...** del menú **Analizar** para acceder al cuadro de diálogo *Frecuencias de respuestas múltiples* que muestra la Figura 22.3.

Figura 22.3. Cuadro de diálogo *Frecuencias de respuestas múltiples*



**Conjuntos de respuestas múltiples.** Este cuadro ofrece un listado de todos los conjuntos previamente definidos en el cuadro de diálogo *Definir conjuntos de respuestas múltiples* (ver Figura 22.2). Para obtener distribuciones de frecuencias:

- Seleccionar el conjunto o conjuntos que se desea describir y trasladarlos a la lista **Tablas para**.

**Valores perdidos.** Las opciones de este recuadro permiten controlar el tipo de tratamiento que se desea dar a los valores perdidos:

- " **Excluir casos según lista dentro de las dicotomías.** Se excluyen del análisis los casos con valor perdido en cualquiera de los conjuntos dicotómicos seleccionados. Con esta opción desactivada se eliminan del análisis de cada conjunto únicamente los casos con valor perdido en ese conjunto. Un caso se considera un valor perdido cuando no puntúa (valor contado) en ninguna de las variables dicotómicas del conjunto.
- " **Excluir casos según lista dentro de las categorías.** Se excluyen del análisis los casos con valor perdido en cualquiera de los conjuntos categóricos seleccionados. Con esta opción desactivada se eliminan del análisis de cada conjunto únicamente los casos con valor perdido en ese conjunto. Un caso se considera un valor perdido cuando no contiene valores dentro del rango definido en ninguna de las variables categóricas del conjunto.

### **Ejemplo: Respuestas múltiples > Frecuencias**

Este ejemplo muestra cómo obtener tablas de frecuencias con variables de respuesta múltiple a partir de los datos del archivo de la Figura 22.1 (los datos están disponibles en el archivo *Transportes públicos*, el cual puede encontrarse en la página web del manual). Para obtener tablas de frecuencias basadas en los dos conjuntos de respuestas múltiples definidos en el ejemplo anterior:

- ' En el cuadro de diálogo principal (ver Figura 22.3), seleccionar los dos conjuntos previamente definidos, *\$trans\_d* y *\$trans\_c*, y trasladarlos a la lista **Tablas para**.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestran las Tablas 22.3 y 22.4. Los resultados de la Tabla 22.3 corresponden al conjunto de dicotomías múltiples *\$trans\_d*, el cual se basa en las variables *autobús*, *metro*, *tren* y *taxi*.

**Tabla 22.3.** Frecuencia de uso del transporte público urbano (dicotomías múltiples)

		Respuestas		Porcentaje de casos
		Nº	Porcentaje	
Transporte público (dicotomías)	Autobús	10	23,8%	50,0%
	Metro	14	33,3%	70,0%
	Tren	13	31,0%	65,0%
	Taxi	5	11,9%	25,0%
Total		42	100,0%	210,0%

a. Conjunto de dicotomías múltiples tabulado en el valor 1.

Los resultados de la Tabla 22.4 corresponden al conjunto de categorías múltiples *\$trans\_c*, el cual se basa en las variables *resp1*, *resp2* y *resp3*. Aunque el conjunto *\$trans\_d* se basa en dicotomías múltiples y el conjunto *\$trans\_c* en categorías múltiples, lo cierto es que ambos contienen la misma información; por este motivo los resultados de las dos tablas que ofrece el *Visor* (22.3 y 22.4) son idénticos.

**Tabla 22.4.** Frecuencia de uso del transporte público urbano (categorías múltiples)

		Respuestas		Porcentaje de casos
		Nº	Porcentaje	
Transporte público (categorías)	Autobús	10	23,8%	50,0%
	Metro	14	33,3%	70,0%
	Tren	13	31,0%	65,0%
	Taxi	5	11,9%	25,0%
Total		42	100,0%	210,0%

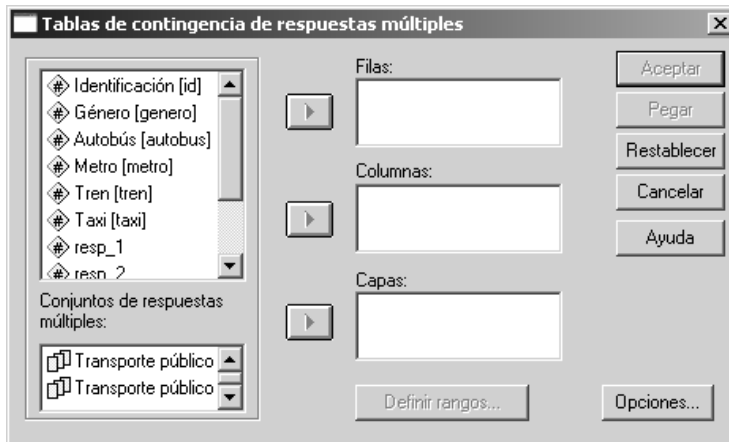
a. Conjunto de categorías múltiples

## Tablas de contingencias

Además de tablas de frecuencias, el procedimiento **Respuestas múltiples** permite obtener tablas de contingencias combinando conjuntos de respuestas múltiples. Para obtener tablas de contingencias:

- Seleccionar la opción **Respuestas múltiples > Tablas de contingencias...** del menú **Analizar** para acceder al cuadro de diálogo *Tablas de contingencias de respuestas múltiples* que muestra la Figura 22.4.

**Figura 22.4.** Cuadro de diálogo *Tablas de contingencias de respuestas múltiples*



La lista de variables del archivo de datos ofrece un listado de todas las variables con formato numérico. La lista **Conjuntos de respuestas múltiples** ofrece un listado de los conjuntos previamente definidos en el cuadro de diálogo *Definir conjuntos de respuestas múltiples* (ver Figura 22.2). El procedimiento permite obtener tablas de contingencias de dos dimensiones (combinando dos conjuntos o variables) y de tres dimensiones (combinando tres conjuntos o variables).

Para obtener una *tabla de contingencias de dos dimensiones*:

- Seleccionar un conjunto (o variable) y trasladarlo a la lista **Filas**.
- Seleccionar un segundo conjunto (o variable) y trasladarlo a la lista **Columnas**.

Es posible combinar tanto conjuntos como variables individuales: dos conjuntos, dos variables, o un conjunto y una variable. Ahora bien, si se desea combinar únicamente variables individuales, existe otro procedimiento SPSS mucho más completo que éste para obtener tablas de contingencias (ver Capítulo 12).

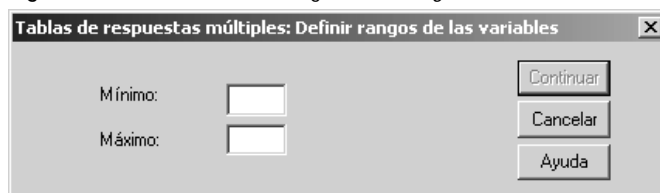
Para obtener una *tabla de contingencias de tres dimensiones*:

- Seleccionar un tercer conjunto (o variable) y trasladarlo a la lista **Capas**.

Cuando se traslada una variable individual (no un conjunto) a cualquiera de las dimensiones del cuadro de diálogo (filas, columnas, capas), el SPSS coloca detrás de la variable dos signos de interrogación entre paréntesis. Esto significa que el procedimiento está esperando que se defina, para esa variable, el rango de valores que se desea incluir en la tabla de contingencias. Para definir el rango de valores de una variable:

- Seleccionar la variable cuyo rango se desea definir y pulsar el botón **Definir rangos...** (ver Figura 22.4) para acceder al subcuadro de diálogo *Tablas de respuestas múltiples: Definir rango de las variables* que muestra la Figura 22.5.

Figura 22.5. Subcuadro de diálogo *Definir rangos de las variables*

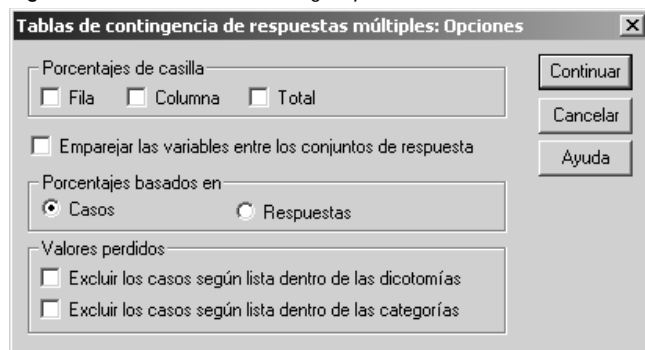


- Introducir, en los cuadros de texto **Mínimo** y **Máximo**, los códigos más pequeño y más grande del rango de códigos correspondientes a las categorías que se desea tabular.
- Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Si no se indica otra cosa, el procedimiento ofrece las frecuencias de respuesta conjunta únicamente en valor absoluto, no en valor porcentual. Y los totales marginales (en valor absoluto y porcentual) los calcula tomando como referencia el número de casos (no en el número de respuestas). No obstante, el botón **Opciones...** conduce a un subcuadro de diálogo que permite decidir si se desea o no obtener frecuencias porcentuales y si éstas deben basarse en el número de casos válidos del archivo o en el número de respuestas.

Para controlar el tipo de porcentajes que se desea obtener y el tratamiento que deben recibir los casos con valor perdido:

- Pulsar el botón **Opciones...** del cuadro de diálogo principal (ver Figura 22.4) para acceder al subcuadro de diálogo *Tablas de contingencias de respuestas múltiples: Opciones* que muestra la Figura 22.6.

Figura 22.6. Subcuadro de diálogo *Opciones*

**Porcentajes de casilla.** Las opciones de este recuadro permiten decidir qué tipo de porcentajes se desea incluir en las casillas de la tabla de contingencias. Puede elegirse una o más de las siguientes opciones:

- " **Fila.** Muestra el porcentaje que la frecuencia de respuesta de cada casilla representa sobre el total de casos o respuestas de su fila.
- " **Columna.** Muestra el porcentaje que la frecuencia de respuesta de cada casilla representa sobre el total de casos o respuestas de su columna.
- " **Total.** Muestra el porcentaje que la frecuencia de respuesta de cada casilla representa sobre el total de casos o respuestas de la tabla.
- " **Emparejar las variables entre los conjuntos de respuesta.** Al combinar dos conjuntos de respuestas múltiples, el SPSS cruza, por defecto, cada variable del primer conjunto con cada variable del segundo conjunto y suma las respuestas. Cuando se están combinando conjuntos de *categorías múltiples*, esta opción de emparejamiento hace que la primera variable del primer conjunto se cruce con la primera variable del segundo conjunto, la segunda variable del primer conjunto con la segunda variable del segundo conjunto, etc. De esta manera, el número total de respuestas computadas es menor. Al marcar esta opción, los porcentajes de respuesta únicamente es posible obtenerlos tomando como referencia el número de respuestas (no el número de casos); de hecho, al marcar esta opción se desactiva la opción **Casos** del recuadro **Porcentajes basados en**.

**Porcentajes basados en.** Los porcentajes de las casillas y los porcentajes de las frecuencias marginales pueden calcularse tomado como referencia:

**Casos.** El número de casos válidos. Es la opción que actúa por defecto.

**Respuestas.** El número de respuestas. Debe tenerse en cuenta que, al utilizar variables de respuesta múltiple, el número de respuestas es mayor que el número de casos. En los conjuntos de dicotomías múltiples, el número de respuestas se obtiene a partir del número de «unos» (valor contado) de cada dicotomía. En los conjuntos de categorías múltiples, el número de respuestas se obtiene a partir del número de respuestas con valor comprendido en el rango establecido al definir el conjunto (ver Figura 22.2).

**Valores perdidos.** Las opciones de este recuadro permiten controlar el tipo de tratamiento que se desea dar a los valores perdidos.

Es posible excluir del análisis los casos con valor perdido en cualquiera de los conjuntos seleccionados o excluir únicamente los casos con valor perdido en cada conjunto analizado. Un caso se considera un valor perdido cuando no puntúa (valor contado) en ninguna de las variables del conjunto. Las dos opciones disponibles son idénticas a las ya descritas anteriormente, en este mismo capítulo, en el párrafo *Valores perdidos* del apartado *Tablas de frecuencias*.

### **Ejemplo: Respuestas múltiples > Tablas de contingencias**

Este ejemplo muestra cómo obtener tablas de contingencias con conjuntos de respuestas múltiples. Se va a utilizar la variable *género* (variable individual) y el conjunto de dicotomías múltiples *trans\_d* definido en el primer ejemplo de este mismo capítulo (los datos están disponibles en el archivo *Transportes públicos*, en la página web del manual). Para cruzar estas dos variables en una tabla de contingencias:

- En el cuadro de diálogo principal (ver Figura 22.4) seleccionar la variable *género* y trasladarla a la lista *Filas*.
- Manteniendo seleccionada la variable *género* dentro de la lista *Filas*, pulsar el botón **Definir rangos...** (ver Figura 22.4) para acceder al subcuadro de diálogo *Definir rangos de las variables* (ver Figura 22.5).
- Introducir los códigos 1 y 2 en los cuadros de texto **Mínimo** y **Máximo**, y pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Seleccionar el conjunto *\$trans\_d* y trasladarlo a la lista *Columnas*.
- Pulsar el botón **Opciones...** para acceder al subcuadro de diálogo *Opciones* (ver Figura 22.6) y marcar las opciones **Fila**, **Columna** y **Total** del recuadro **Porcentajes de casilla**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestra la Tabla 22.5. El contenido de las casillas (*Recuento*) refleja el *número de respuestas* de cada tipo: 6 *varones* utilizan el *autobús*, 8 *mujeres* utilizan el *metro*, etc. Los totales de las columnas (que son los totales que corresponden al conjunto *transporte público*) también reflejan el número de respuestas: de los 20 encuestados, 10 manifiestan utilizar el *autobús*, 14 el *metro*, etc.

El resto de valores de la tabla está referido al *número de encuestados* o número de casos válidos (esta circunstancia queda aclarada en una nota a pie de tabla). Así, los totales marginales de las filas (los totales marginales de la variable individual *género*) reflejan el número de hombres y mujeres. Y los porcentajes de las casillas están calculados sobre el número de casos válidos: los 6 *varones* que utilizan el *metro* representan el 60,0% de los 10 varones encuestados, el 42,9% de los 14 encuestados que utilizan el metro y el 30,0% de los 20 encuestados de la muestra.

El total de la tabla indica que la muestra está formada por 20 encuestados o casos válidos. Lógicamente, este total coincide con la suma de los totales de las filas (que recogen el número de encuestados), pero no con la suma de los totales de las columnas (que recogen el número de respuestas).

**Tabla 22.5.** Tabla de contingencias de *género* por *transporte público* (porcentajes basados en casos)

			Transporte público (dicotomías)				Total
			Autobús	Metro	Tren	Taxi	
Género	Varones	Recuento	6	6	7	2	10
		% dentro de genero	60,0%	60,0%	70,0%	20,0%	
		% dentro de \$trans_d	60,0%	42,9%	53,8%	40,0%	
		% del total	30,0%	30,0%	35,0%	10,0%	50,0%
	Mujeres	Recuento	4	8	6	3	10
		% dentro de genero	40,0%	80,0%	60,0%	30,0%	
		% dentro de \$trans_d	40,0%	57,1%	46,2%	60,0%	
		% del total	20,0%	40,0%	30,0%	15,0%	50,0%
Total	Recuento	10	14	13	5	20	
	% del total	50,0%	70,0%	65,0%	25,0%	100,0%	

Los porcentajes y los totales se basan en los encuestados.

Marcando la opción **Respuestas** del recuadro **Porcentajes basados en** (ver Figura 22.6), todos los valores de la tabla (frecuencias y porcentajes) quedan referidos, no al *número encuestados válidos*, sino al *número de respuestas* (ver Tabla 22.6). Una nota a pie de tabla indica esta circunstancia cambiando de *encuestados* (casos válidos) a *respuestas*. Ahora, los 6 *varones* que utilizan el *metro* representan un 28,6% de las 21 respuestas dadas por los 10 varones encuestados, un 42,9% de los 14 encuestados que utilizan el metro y un 14,3% de las 42 respuestas dadas por los 20 encuestados de la muestra.

El total de la tabla indica que los 20 encuestados han dado 42 respuestas. Puesto que ahora todos los valores están basados en el número de respuestas, este total coincide tanto con la suma de los totales de las filas como con la suma de los totales de las columnas.

**Tabla 22.6.** Tabla de contingencias de *género* por *transporte público* (porcentajes basados en respuestas)

			Transporte público (dicotomías)				Total
			Autobús	Metro	Tren	Taxi	
Género	Varones	Recuento	6	6	7	2	21
		% dentro de genero	28,6%	28,6%	33,3%	9,5%	
		% dentro de \$Trans_d	60,0%	42,9%	53,8%	40,0%	
		% del total	14,3%	14,3%	16,7%	4,8%	50,0%
	Mujeres	Recuento	4	8	6	3	21
		% dentro de genero	19,0%	38,1%	28,6%	14,3%	
		% dentro de \$Trans_d	40,0%	57,1%	46,2%	60,0%	
		% del total	9,5%	19,0%	14,3%	7,1%	50,0%
Total	Recuento	10	14	13	5	42	
	% del total	23,8%	33,3%	31,0%	11,9%	100,0%	

Los porcentajes y los totales se basan en las respuestas.

# Referencias bibliográficas

- Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.
- Amón, J. (1979). *Estadística para psicólogos*. Vol 1: *Estadística descriptiva* (2ª ed.). Madrid: Pirámide.
- Amón, J. (1984). *Estadística para psicólogos*. Vol 2: *Probabilidad y estadística inferencial* (3ª ed.). Madrid: Pirámide.
- Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press.
- Arbuthnott, J. (1710). An argument for Divine Providence taken from the constant regularity observed in the birth of both sexes. *Philosophical Transactions*, 27, 186-190.
- Beherens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, 2, 131-160.
- Belsley, D. A., Kuh, E. y Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: Wiley.
- Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.
- Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, 43, 572-574.
- Box, G. E. P. (1954a). Some theorems on quadratic forms applied in the study of analysis of variance problems. I: Effects of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, 25, 290-302.
- Box, G. E. P. (1954b). Some theorems on quadratic forms applied in the study of analysis of variance problems. II: Effects of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics*, 25, 484-498.
- Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, 91, 14-28.
- Breslow, N. E. y Day, N. E. (1980). *Statistical methods in cancer research*. Vol 1: *The analysis of case-control studies*. Lyon: IARC.
- Breslow, N. E. y Day, N. E. (1987). *Statistical methods in cancer research*. Vol 2: *The design and analysis of cohort studies*. Lyon: IARC.
- Brown, M. B. y Forsythe, A. B. (1974). The ANOVA and multiple comparisons for data with heterogeneous variances. *Biometrics*, 30, 719-724.
- Bryk, A. S. y Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, 104, 396-404.
- Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility*. Thousand Oaks, CA: Sage.
- Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37, 256-266.
- Cochran, W. G. (1952). The  $\chi^2$  test of goodness of fit. *Annals of Mathematical Statistics*, 23, 315-345.



- Cochran, W. G. (1954). Some methods for strengthening the common  $\chi^2$  tests. *Biometrics*, 10, 417-451.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- Conover, W. J. (1980). *Practical nonparametric statistics* (2ª ed.). New York: Wiley.
- Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, 19, 15-18.
- Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, 74, 169-174.
- Cook, R. D. (1993). Exploring partial residual plots. *Technometrics*, 35, 351-362.
- Cook, R. D. y Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.
- Cooley, W. W. y Lohnes, P. R. (1971). *Multivariate data analysis*. New York: Wiley.
- Cramer, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, 69, 161-182.
- Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.
- Dineen, L. C. y Blakesley, B. C. (1973). Algorithm AS 62: Generator for the sampling distribution of the Mann-Whitney  $U$  statistic. *Applied Statistics*, 22, 269-273.
- Dixon, W. J. (ed.) (1983). *BMDP statistical software manual*. Los Angeles: University of California Press.
- Duncan, D. B. (1955). Multiple range and multiple  $F$  tests. *Biometrics*, 11, 1-42.
- Dunn, C. W. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56, 52-64.
- Dunnnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50, 1096-1121.
- Dunnnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association*, 75, 795-800.
- Durbin, J. y Watson, G. S. (1950). Testing for serial correlation in least-squares regression I. *Biometrika*, 37, 409-438.
- Durbin, J. y Watson, G. S. (1951). Testing for serial correlation in least-squares regression II. *Biometrika*, 38, 159-178.
- Durbin, J. y Watson, G. S. (1971). Testing for serial correlation in least-squares regression III. *Biometrika*, 58, 1-19.
- Einot, I. y Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association*, 70, 574-583.
- Festinger, L. (1946). The significance of difference between means without reference to the frequency distribution function. *Psychometrika*, 11, 97-105.
- Fisher, R. A. (1922). On the interpretation of chi-square from contingency tables, and the calculation of  $P$ . *Journal of the Royal Statistical Society*, 85, 87-94.
- Fisher, R. A. (1924). The conditions under which  $X^2$  measures the discrepancy between observation and hypothesis. *Journal of the Royal Statistical Society*, 87, 442-450.
- Fisher, R. A. (1935). *Statistical methods for research workers* (5ª ed.). Edinburgh: Oliver and Boyd (existe 14ª ed. en 1973, New York: Hafner).

- Fleiss, J. L., Cohen, J. y Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72, 323-327.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 61, 1081-1096.
- Gabriel, K. R. (1969). Simultaneous test procedures: Some theory of multiple comparisons. *Annals of Mathematical Statistics*, 40, 224-240.
- Games, P. A. y Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal  $n$ 's and/or variances: A Monte Carlo study. *Journal of Educational Statistics*, 1, 113-125.
- Geisser, S. y Greenhouse, S. W. (1958). An extension of Box' results on the use of  $F$  distribution in multivariate analysis. *Annals of Mathematical Statistics*, 29, 885-891.
- Goodman, L. A. y Kruskal, W. H. (1979). *Measures of association for cross classifications*. New York: Springer.
- Green, S. B., Lissitz, R. W. y Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, 37, 827-838.
- Greenhouse, S. W. y Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, 24, 95-112.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255-282.
- Haberman, S. J. (1973). The analysis of residuals in cross-classification tables. *Biometrics*, 29, 205-220.
- Hagen, R. L. (1997). In praise of the hypothesis statistical test. *American Psychologist*, 52, 15-24.
- Harlow, L. L., Mulaik, S. A. y Steiger, J. H. (1997). *What if there were no significance test*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Haviland, M. G. (1990). Yates' correction for continuity and the analysis of 2x2 contingency tables (with comments). *Statistics in Medicine*, 9, 363-383.
- Hays, W. L. (1995). *Statistics* (5ª ed.). New York: Holt, Rinehart and Winston.
- Hochberg, Y. (1974). Some generalizations of the T-method in simultaneous inference, *Journal of Multivariate Analysis*, 4, 224-234.
- Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics*, 2, 360-378.
- Howitt, D. y Cramer, D. (2000). *First steps in research and statistics*. London: Routledge.
- Huynh, H. y Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot design. *Journal of Educational Statistics*, 1, 69-82.
- Huynh, H. y Mandeville, G. K. (1979). Validity conditions in repeated measures design. *Psychological Bulletin*, 86, 964-973.
- Kendall, M. G. (1963). *Rank correlation methods* (3ª ed.). London: Griffin (existe 4ª ed. en 1970).
- Kendall, M. G. y Babington-Smith, B. (1939). The problem of  $m$  rankings. *The Annals of Mathematical Statistics*, 10, 275-287.
- Keuls, M. (1952). The use of studentized range in connection with an analysis of variance. *Euphytica*, 1, 112-122.
- Kirk, R. E. (1982). *Experimental design. Procedures for the behavioral sciences* (2ª ed.). Belmont, CA: Brooks/Cole (existe 3ª ed. en 1995).
- Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell' Istituto Italiano degli Attuari*, 4, 83-91.

- Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika*, 28, 221-238.
- Kristof, W. (1969). Estimation of true score and error variance for tests under various equivalence assumptions. *Psychometrika*, 34, 489-507.
- Kruskal, W. H. y Wallis, W. A. (1952). Use of ranks on one-criterion variance analysis. *Journal of the American Statistical Association*, 47, 583-621 (aparecen correcciones en el volumen 48, pp. 907-911).
- Landis, J. R. y Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- Levene, H. (1960). Robust tests for the equality of variances. En J. Olkin (ed.): *Contributions to probability and statistics*. Palo Alto, CA: Stanford University Press.
- Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62, 399-402.
- Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings National Science India*, 2, 49-55.
- Mann, H. B. y Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, 50-60.
- Mantel, N. y Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.
- Marascuilo, L. A. y McSweeney, M. (1977). *Nonparametric and distribution-free methods*. Monterrey, CA: Brooks/Cole.
- Mauchly, J. W. (1940). Significance test for sphericity of a normal  $n$ -variate distribution. *Annals of Mathematical Statistics*, 11, 204-209.
- McGraw, K. O. y S. P. Wong (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30-46.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12, 153-157.
- Montgomery, D. C., Peck, E. A. y Vining, G. G. (2001). *Introduction to linear regression analysis* (3ª ed.). New York: Wiley.
- Moses, (1952). A two sample test. *Psychometrika*, 17, 239-247.
- Newman, D. (1939). The distribution of the range in samples of a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika*, 31, 20-30.
- Neyman, J. y Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference (2ª parte). *Biometrika*, 20, 263-294.
- Norušis, M. J. (2003). *SPSS 12.0. Statistical procedures companion*. Upper Saddle River, NJ: Prentice Hall.
- Norušis, M. J. y SPSS, Inc. (1993). *SPSS for Windows. Base system user's guide release 6.0*. Chicago, IL: SPSS Inc.
- Norušis, M. J. y SPSS, Inc. (1994). *SPSS advanced statistics 6.1*. Chicago, IL: SPSS Inc.
- Nunnally, J. C. (1987). *Teoría psicométrica*. México: Trillas.
- Palmer, A. L. (1999). *Análisis de datos. Etapa exploratoria*. Madrid: Pirámide.
- Pardo, A. y San Martín, R. (1994). *Análisis de datos en psicología II*. Madrid: Pirámide.
- Pardo, A. y San Martín, R. (1998). *Análisis de datos en psicología II* (2ª ed.). Madrid: Pirámide.
- Pardo, A. (2002). *Análisis de datos categóricos*. Madrid: UNED.
- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophical Magazine*, 50, 157-175.

- Pearson, K. (1911). On the probability that two independent distributions of frequency are really samples from the same population. *Biometrika*, 8, 250-254.
- Pearson, K. (1913). On the probable error of a correlation coefficient as found from a fourfold table. *Biometrika*, 9, 22-27.
- Rousseeuw, P. J. (1998). Robust estimation and identifying outliers. En H. M. Wadsworth (ed.), *Handbook of statistical methods for engineers and scientists*. New York: McGraw-Hill.
- Rousseeuw, P. J. y Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.
- Romesburg, H. C. (1984). *Cluster analysis for researchers*. Belmont, CA: Lifetime Learning Publications.
- Ryan, T. A. (1960). Significance tests for multiple comparisons of proportions, variances and other statistics. *Psychological Bulletin*, 57, 318-328.
- Ryan, T. P. (1997). *Modern regression methods*. New York: Wiley.
- San Martín, R. y Pardo, A. (1989). *Psicoestadística. Contrastes paramétricos y no paramétricos*. Madrid: Pirámide.
- Scheffé, H. A. (1953). A method for judging all possible contrasts in the analysis of variance. *Biometrika*, 40, 87-104.
- Scheffé, H. A. (1959). *The analysis of variance*. New York: Wiley.
- Searle, S. R., Speed, F. M. y Milliken, G. A. (1980). Population marginal means in the linear model: An alternative to least squares means. *The American Statistician*, 34, 216-221.
- Shapiro, S. S. y Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591-611.
- Sidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62, 626-633.
- Smirnov, N. V. (1939). Estimate of deviation between empirical distribution functions in two independent samples. *Bulletin Moscow University*, 2, 3-16 [ruso].
- Smirnov, N. V. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19, 279-281.
- Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, 27, 799-811.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72-101.
- SPSS (1991). *SPSS statistical algorithms* (2ª ed.). Chicago, IL: SPSS Inc.
- SPSS (1999). *SPSS Base 10.0. Applications guide*. Chicago, IL: SPSS Inc.
- Tabachnik, B. G. y Fidel, L. S. (2001). *Using multivariate statistics* (2ª ed.). Boston: Allyn and Bacon.
- Tamhane, A. C. (1977). Multiple comparisons in model I one-way ANOVA with unequal variances. *Communications in Statistics*, A6(1), 5-32.
- Tamhane, A. C. (1979). A comparison of procedures for multiple comparisons of means with unequal variances. *Journal of the American Statistical Association*, 74, 471-480.
- Tarone, R. E. (1985). On heterogeneity tests based on efficient scores. *Biometrika*, 72, 91-95.
- Tarone, R. E., Gart, J. J. y Hauck, W. W. (1983). On the asymptotic relative efficiency of certain noniterative estimators of a common relative risk or odds ratio. *Biometrika*, 70, 519-522.
- Theil, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology*, 76, 103-154.

- Toothaker, L. E. (1991). *Multiple comparison for researchers*. London: Sage.
- Tukey, J. W. (1949). One degree of freedom for nonadditivity. *Biometrics*, 5, 232-234.
- Tukey, J. W. (1953). *The problem of multiple comparisons*. Princeton University (manuscrito mimeografiado).
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison Wesley.
- Wald, A. y Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Annals of Mathematical Statistics*, 11, 147-162.
- Waller, R. A. y Duncan, D. B. (1969). A Bayes rule for the symmetric multiple comparison problem. *Journal of the American Statistical Association*, 64, 1484-1503.
- Wallis, W. A. (1939). The correlation ratio for ranked data. *Journal of the American Statistical Association*, 34, 533-538.
- Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29, 350-362.
- Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, 38, 330-336.
- Welsch, R. E. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association*, 72, 566-575.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, 80-83.
- Wilcoxon, F. (1949). *Some rapid approximate statistical procedures*. American Cyanamid Co., Stanford Research Laboratories.
- Winer, B. J., Brown, D. R. y Michels, K. M. (1991). *Statistical principles in experimental design* (3ª ed.). New York: McGraw-Hill.
- Yates, F. (1934). Contingency tables involving small numbers and the  $\chi^2$  test. *Journal of the Royal Statistical Society*, 1, 217-235.

# Índice de materias

## A

*A posteriori*, comparaciones (ver *Comparaciones múltiples*)

*A priori*, comparaciones (ver *Comparaciones múltiples*)

Abrir archivos:

de datos (SPSS, SAS, Excel, Lotus, dBase, etc.), 35-38

de resultados, 162

de sintaxis, 200

Abrir archivos con formato texto, 45-50

Abrir archivos nuevos (de datos, de resultados, de sintaxis, de procesos), 33

Abrir bases de datos (Acces, dBase, Excel, Fox-Pro, etc.), 38-44

Acuerdo, índices de:

coeficiente de correlación intraclase, 529-532

$\kappa$  de Cohen, 294-296

*W* de Kendall, 525, 565-571

Aditividad, prueba de Tukey sobre, 528-529

Afijación (ver *Muestreo aleatorio estratificado*)

Afirmación del consecuente, falacia de la, 231

Agregar datos, 140-142

Aleatoria, muestra, 215

Aleatorio, muestreo, 215

Aleatorios, bloques, efectos (ver *ANOVA*)

Aleatorización, 119-120

generadores de números aleatorios, 119-120

semilla de aleatorización, 120

*Alfa*, nivel de significación o riesgo, 228-229

*Alfa* de Cronbach (ver *Fiabilidad, análisis de*)

Alfanumérica, variable (ver *Cadena*)

Alternativa, hipótesis, 224-225

Amplitud o rango, 236-238, 245, 252

Amplitud intercuartílica, 265, 269, 276-278

Análisis estadístico, 207-208

ANCOVA (Análisis de covarianza), 370-372, 390-391

ANOVA (Análisis de varianza):

bloques aleatorios, 387-388

efectos fijos y aleatorios, 338-339, 346, 348, 362-363

efectos principales, 361, 364, 374, 404-405, 415-416

efectos simples, 374-375, 383, 417, 424-426 en condiciones de varianzas desiguales o heterocedasticidad (soluciones de Brown-Forsythe y Welch), 347-349

estimaciones de los parámetros, 376, 389-390

factor inter-sujetos (completamente aleatorizado), 395-396, 399, 402-403, 418-422

factor intra-sujetos (medidas repetidas), 395-399, 401-402, 406, 411-414, 418-423

homocedasticidad (homogeneidad o igualdad de varianzas), 323, 326-327, 347-348, 377, 422-423

interacción entre factores, 336-337, 361, 364-365, 368-370, 374-375, 381-385, 413-414, 415, 417, 423-426

lógica del ANOVA, 339-342

matriz de coeficientes del contraste,

matriz de sumas de cuadrados y productos cruzados, 406-407

matriz de transformación, 406

matriz de varianzas-covarianzas, 422-423

matriz *L*, 381

matriz residual de sumas de cuadrados y productos cruzados, 407-408

medias estimadas, 373-374, 405

modelo aleatorizado en bloques, 387-388

modelo de dos factores, ambos con medidas repetidas, 408-418

modelo de dos factores completamente aleatorizados, 361-393

modelo de dos factores, con medidas repetidas en un factor (mixto o *split-plot*), 418-427

modelo de un factor completamente aleatorizado, 342-359

modelo de un factor con medidas repetidas, 396-408

modelos de ANOVA (aleatorizados, medidas repetidas, bloques), 335-339

modelos jerárquicos o anidados, 388-389

muestreo de niveles, 338-339

- potencia observada, 376, 379
- tamaño del efecto, 375
- tipo de aleatorización, 337-338
- Aproximación normal a la binomial, 538, 541
- Archivos de datos, 35-58
  - abrir archivos de datos, 35-38
  - abrir bases de datos, 38-44
  - archivos usados recientemente, 58
  - copiar propiedades de datos, 71
  - guardar archivos de datos, 50-53
    - marcar como de sólo lectura, 53
  - hacer una caché del archivo de datos, 54
  - imprimir archivos de datos, 55-57
    - presentación preliminar, 55
  - información del archivo de datos, 46-47
  - leer datos con formato texto, 45-50
    - archivos de ancho fijo, 46
    - archivos delimitados (formato libre), 46
  - mostrar información de datos, 54-54
- Archivos de resultados (ver *Visor de resultados*)
- Archivos de sintaxis, 199-204
  - abrir y guardar archivos, 200
  - ejecutar sintaxis, 204
  - generar sintaxis, 201-203
  - reglas básicas, 204
- Archivos, extensiones:
  - .jnl (diario de la sesión), 203
  - .sav (datos), 22, 36, 51
  - .sbs (procesos), 13
  - .spo (resultados), 162
  - .sps (sintaxis), 200
  - .tlo (aspectos de tabla), 186
- Asimetría, índice de, 236-237, 239, 245
- Asociación (ver *Medidas de asociación*)
- Ayuda, 27-34
  - asesor de resultados, 27, 33
  - asesor estadístico, 27, 31
  - botones de ayuda, 34
  - contextual, 32-33
  - estudios de casos, 33
  - guía de sintaxis, 32
  - por temas, 27-30
  - tutorial del SPSS, 31
- B**
- Barra de estado, 210-22
- Barra de herramientas, 15-16
  - personalizar una barra de herramientas, 16-21
- Barra de menús, 9-14
  - editor de menús, 11-14
  - menús, 10-11
- Bartlett, prueba de esfericidad, 407-408
- Beta* ( $\beta$ ), coeficiente de regresión parcial tipificado, 462, 466-467, 480, 483
- Binomial, prueba, 538-541
- Bondad de ajuste:
  - chi*-cuadrado sobre, 534-538
  - en regresión lineal, 458-459, 460-461, 465
  - Kolmogorov-Smirnov (una muestra), 544-546
- Bonferroni (corrección para comparaciones múltiples), 350, 374, 383, 405, 416-417, 424-425
- Bowker (McNemar-Bowker), prueba homogeneidad marginal, 302, 304-305
- Box, prueba sobre homog. de varianzas, 422-423
- Breslow-Day, prueba sobre homog. de *odds-ratios*, 308
- Brown-Forsythe, estadístico para el contraste de la igualdad de medias, 347-348
- Buscar (datos, casos, variables), 76-78
- C**
- Caché de datos, 54
- Cadena:
  - funciones de, 89
  - variables de, 63
- Calcular nuevas variables, 85-92
  - expresiones condicionales, 90-92
  - expresiones numéricas, 87-88
  - funciones, 88-90
  - operadores (aritm., relacionales, lógicos), 88
- Caso-control, diseño de, 298-300
- Casos atípicos, 263-264, 267, 269, 472, 484
- Casos duplicados, identificar, 143-145
- Casos extremos, 263-267, 269, 271, 277
- Categorizar variables, 96-102
  - categorización automática, 100-102,
  - categorización manual 99-100 (ver también *Recodificar*)
  - categorizador visual, 97-99
- Catóricas, variables, 279
- Chi*-cuadrado de Pearson:
  - prueba de bondad de ajuste, 534-538
  - prueba de independencia y homogeneidad de proporciones en tablas de contingencias, 284-286
- Circularidad (ver *Esfericidad*)
- Cochran:
  - combinación de tablas 2x2, 306-308
  - prueba no paramétrica para varias muestras relacionadas, 525-526, 567, 570-571
- Coefficiente de concordancia W de Kendall, 525, 565-571

- Coefficiente de contingencia, 287, 291
  - Coefficiente de determinación ( $R^2$ ), 458-459
  - Coefficiente de determinación corregido ( $R^2$  corregida), 458-460, 480, 486, 502-504
  - Coefficiente de dispersión, 252
  - Coefficiente de fiabilidad, 509
  - Coefficiente de incertidumbre de Theil, 289-291
  - Coefficiente de variación, 252
  - Coefficientes de correlación (ver también *Medidas de asociación*):
    - $d$  de Somers, 292-293
    - $eta$ , 294, 317
    - $eta$ -cuadrado, 375
    - $gamma$  ( $\gamma$ ) de Goodman y Kruskal, 292-293
    - intraclase, 529-532
    - parcial, 437-442, 470, 485-488, 491
    - $phi$ , 287-288, 448
    - $rho$  de Spearman, 287, 433, 436-437
    - $r_{xy}$  Pearson, 286-287, 432-433, 435-436, 440, 444, 448, 460, 462
    - semiparcial, 470, 486
    - $tau$ -b de Kendall, 292-293, 433, 436-437
    - $tau$ -c de Kendall, 292-293
  - Coefficientes de regresión parcial:
    - no tipificados, 456-457, 462-463, 466-468
    - tipificados, 462, 466-467, 480, 483
  - Cohen, índice de acuerdo  $kappa$  ( $\kappa$ ), 294-296
  - Cohortes, diseño de, 297-298, 300-301
  - Colinealidad, 471, 479-481
  - Comparaciones múltiples:
    - a priori* o planeadas, 354-359, 379-380
    - a posteriori* o *post hoc*, 349-354, 366-368, 424-427
    - de tendencia, 355-357, 380, 406-407
  - Concentración, índice de, 252
  - Confianza, nivel de, 228, 232
  - Conjunto, análisis, 145-146
  - Conjuntos de variables (definir, usar), 81-83
  - Contar valores, 102-105
  - Contraste de hipótesis, 221-231
    - decisión, 230-231
    - estadístico de contraste, 227-228
    - falacia de la afirmación del consecuente, 231
    - hipótesis estadísticas, 224-225
    - nivel crítico (valor  $p$ ), 230
    - regla de decisión, 228-230
    - supuestos, 225-226
    - unilateral y bilateral, 225, 229
  - Contrastes para comparar los niveles de una variable categórica (desviación, simple, diferencia, Helmert, repetido, polinómico, especial), 379-380 (ver también *Comparaciones múltiples a priori o planeadas*)
  - Contrastes no paramétricos (ver *No paramétrico, análisis*)
  - Contrastes sobre medias, 319-332
    - prueba  $T$  para dos muestras independientes, 322-328
    - prueba  $T$  para dos muestras relacionadas, 328-332
    - prueba  $T$  para una muestra, 319-322
  - Cook, distancia de, 483
  - Copiar resultados en otras aplicaciones (ver *Visor de resultados*)
  - Corrección por continuidad (corrección de Yates), 286, 307, 538, 542, 561
  - Correlación lineal, 429-4327
  - Correlación parcial, 437-442, 470, 485-488, 491
    - corr. de orden cero, 437-438, 440-442, 470
  - Correlación semiparcial, 470, 486
  - Covarianza, 434-436
  - Covarianza, análisis de, (ver *ANCOVA*)
  - Cramer, coeficiente  $V$ , 287-288, 291
  - Cronbach, coeficiente de fiabilidad  $alfa$ , 512-514
  - Cuadros de diálogo SPSS, 5-8
  - Cuantiles, 235-236
  - Cuartiles, 235
  - Cubos OLAP, 255-260
  - Curtosis, índice de, 237, 239, 245
  - Curva normal, 247-249
- ## D
- Datos/archivos usados recientemente, 58
  - Deciles, 236
  - Definir variables, 60-73
    - alineación del texto, 66-67
    - anchura de la variable, 62-63
    - anchura de las columnas, 66
    - automáticamente, 67-70
    - fechas, 72-73
    - decimales, 62-63
    - etiquetas de variable y de valor, 64-65
    - nombre de variable, 61
    - nivel de medida, 67
    - tipo de variable (formato), 61-63
    - valores perdidos, 65-66
  - Descriptivos, estadísticos (sobre tendencia central, posición, dispersión, forma de la distribución), 235-239243-247
  - Desviación típica, 236-237, 244-245
  - Desviación promedio absoluta, 252



Detener el procesador SPSS, 55

Diagramas:

de barras, 239-241

de barras agrupadas, 280-282

de caja, 267-269

de dispersión, 430-431, 455-458, 463-464, 475-476, 479, 497, 499, 503-504

de dispersión parciales, 478-479

de dispersión por nivel, 275-278, 377-378

de líneas, 347, 349, 357, 415, 423-424

de normalidad (ver *Exploratorio, análisis*)

de sectores, 239-240

de tallo y hojas, 270-271

histograma, 239-241, 271-272

Diccionario de datos, 71

Diferencia estacional, 116

Diferencia mínima significativa (DMS), 350

Diferencial relativo al precio, 252

Diseño longitudinal, 240

antes-después, 301, 328

prospectivo o de cohortes, 297-298

retrospectivo o de caso-control, 298-300

Diseño ortogonal, 145-152

Diseño transversal, 296

Distancia de Mahalanobis, 483

Distancias:

medidas de similaridad, 444-448

medidas de disimilaridad, 449-451

Distribución muestral, 217-220

Distribución o tabla de frecuencias, 233-235, 242

Duncan, prueba del rango múltiple, 351, 354

Dunnett, estadísticos  $T_3$  y  $C$  para comparaciones por pares, 352

Dunnett, prueba para comparaciones por pares con un grupo control, 352

Durbin-Watson, prueba de independencia de los residuos, 473-474

## E

Editar resultados (ver *Visor de resultados*)

Editar tablas (ver *Visor de resultados*)

Editor de datos, 3-5, 63-83

*display* del editor de datos, 59, 73, 74, 79

editar datos, 74-78

borrar datos, 76

buscar datos, casos, variables, 76-77

deshacer/rehacer, 74

insertar casos, variables, 78

mover/copiar datos, 75

seleccionar datos, 75

entrar datos, 73-74

estructura de un archivo de datos, 60

modificar el aspecto de editor de datos, 79

etiquetas (ver-ocultar), 79

fuentes, 69

segmentar ventana, 80

vista de datos/vista de variables, 59-61

Editor de menús, 11-14

Editor de sintaxis, 5, 200-204

Efectos:

aleatorios, 338-339, 346, 348, 363, 530, 532

fijos, 338-339, 346, 348, 362-363, 530

mixtos, 530

principales, 361, 364, 374, 404-405, 415-416

simples, 374-375, 383, 417, 424-426

*Épsilon* ( $\epsilon$ ), corrector (Box, Greenhouse-Geisser, Huynh-Feldt), 400-401

Error muestral máximo, 232

Error típico, 220

Escalas de medida (nominal, ordinal, de intervalo, de razón), 209-212

Esfericidad, 400-401, 407-408, 413-414, 420-421

Estadística descriptiva, 207

Estadística inferencial o inductiva, 207

Estadística no paramétrica (ver *No paramétrico, análisis*)

Estadístico, 214

Estimación por intervalos, 232

Estimación puntual, 231

Estimador, 231-232

Estimadores robustos centrales (estimadores  $M$ ), 263-265

*Eta* ( $\eta$ ), coeficiente de correlación, 294, 317

*Eta-cuadrado* ( $\eta^2$ ), 375

Etiquetas de variable y de valor (ver *Definir variable*)

Exploratorio, análisis, 261-278

casos atípicos y extremos, 263, 266-267, 269

diagramas de caja, 267-269

diagrama de tallo y hojas, 270-271

estadísticos descriptivos, 263-265

estimadores robustos centrales (estimadores  $M$ ), 263-265

histograma, 271-272

homogeneidad de varianzas:

dispersión por nivel, 275-278, 377-378

prueba de Levene, 275-278

normalidad:

gráficos de normalidad, 272-275

pruebas de normalidad, 272-273

Exportar resultados, 194-198

**F**

*F*, estadístico de Fisher (ANOVA), 342-343  
 Factor, variable independiente del ANOVA, 336-337 (ver también ANOVA)  
 Fecha/hora, operaciones con fechas y horas, 89, 111-115  
 Fecha, definir variables tipo fecha, 72-73  
 Fiabilidad, análisis de, 507-532  
   acuerdo, 529  
   acuerdo absoluto, 531-532  
    $\alpha$  de Cronbach, 512-514  
   coeficiente de fiabilidad, 509  
   concepto de fiabilidad, 507-509  
   consistencia, 512-513, 531-532  
   consistencia interna, 508, 512-513  
   descriptivos, 519-524  
     elemento, 520-521  
     entre elementos, 523-524  
     escala, 520-523  
     escala si se elimina elemento, 520-522  
   error de medida, 508-509, 525  
   estabilidad, 507-508, 510, 514, 515  
   Guttman:  
     coeficiente de fiabilidad, 514, 516  
     estimaciones del límite inferior de la fiabilidad, 516-517  
   índice de fiabilidad, 509  
   índice de homogeneidad corregido, 521  
   intraclase, coeficiente de correlación, 529-532  
   Kuder-Richardson (KR<sub>20</sub>), 513  
   modelo  $\alpha$ , 512-514  
   modelo de dos mitades, 514-516  
   modelo de Guttman, 516-517  
   modelo de medidas paralelas y estrictamente paralelas, 517-519  
   profecía de Spearman-Brown, 515  
   prueba de aditividad de Tukey, 528-529  
   prueba no paramétrica de Cochran, 525-526  
   prueba no paramétrica de Friedman, 525  
   prueba *F* (ANOVA), 524-525  
   puntuaciones observadas, 508-509, 518, 519  
   puntuaciones verdaderas, 508-509  
     varianza de las puntuaciones verdaderas, 509, 517, 518  
    $T^2$  de Hotelling, 526-527  
   varianza de los errores, 509, 518, 525  
 Fijos, efectos, 338-339, 346, 348, 362-363, 530  
 Filtrar casos (ve *Seleccionar casos*)  
 Fisher:  
   estadístico *F* (ANOVA), 342-343, 345  
   prueba exacta de, 286

## Frecuencias:

distribución de, 233-235, 242  
 esperadas o teóricas, 284-285, 309  
 observadas o empíricas, 284-285, 309, 311  
 Friedman, prueba no paramétrica, 525, 564-570  
 Funciones (ver *Calcular nuevas variables*)  
 Fundir archivos (añadir casos, añadir variables), 135-139

**G**

Gabriel, prueba de, 351-352, 354  
 Games-Howell, prueba de, 352-354, 367  
 Goodman y Kruskal:  
    $\gamma$  ( $\gamma$ ), 292-293, 448  
    $\lambda$  ( $\lambda$ ), 288-289, 447-448  
   reducción proporcional del error, 288-289, 448  
    $\tau$ , 289  
 Greenhouse-Geisser, corrector  $\epsilon$  ( $\epsilon$ ), 400-401, 414, 421  
 Guardar archivos, 50-53  
   guardar como, 53  
   guardar con contraseña, 163  
   marcar como de sólo lectura, 53  
 Guía de sintaxis, 27, 32  
 Guttman:  
   coeficiente de fiabilidad, 514, 516  
   límite inferior de la fiabilidad, 516-517

**H**

*H*, prueba no paramétrica de Kruskal-Wallis, 555-557  
 Hipótesis estadísticas, 222, 224-225  
 Histograma, 239-241, 271-272  
 Hochberg, GT2, 351, 354  
 Homogeneidad de las pendientes de regresión, 390-391  
 Homogeneidad de proporciones (ver *Chi-cuadrado*, *McNemar* y *McNemar-Bowker*)  
 Homogeneidad o igualdad de varianzas, 323, 326-327, 347-348, 377, 422-423, 471, 474-476  
   dispersión por nivel, 275-278, 377-378  
   prueba de Levene, 275-278, 347-348, 358, 377, 422-423  
   prueba *M* de Box, 422-423  
 Hotelling:  
    $T^2$  de, 526-527  
   traza de, 400, 413, 420  
 Huynh-Feldt, corrector  $\epsilon$  ( $\epsilon$ ), 401, 414, 421

**I**

- Imprimir:
  - archivos de datos, 55-57
  - archivos de resultados, 189-193
- Incertidumbre, 208
- Inducción, 208
- Inferencia estadística, 221
- Insertar casos y variables, 78
- Interacción entre factores, 336-337, 361, 364-365, 368-370, 374-375, 381-385, 413-414, 415, 417, 423-426
- Intervalo, escala de medida de, 210-211
- Intervalo de confianza, 232
  - coeficiente de correlación intraclase, 529-532
  - coeficiente de regresión lineal, 390, 468
  - dos medias independientes, 325-327, 353, 379, 382-383, 405, 416-417, 426-427
  - dos medias relacionadas, 330-331, 425
  - índice de riesgo relativo, 297-301
  - media, 232, 251, 320-321, 344-345, 348, 365, 374
  - mediana, 251
  - odds ratio* (razón de las ventajas), 298-301
  - odds ratio* común, 308
  - parámetros de un modelo lineal, 376
  - pronósticos de la regresión lineal, 494-495
  - pronósticos promedio de la regresión lineal, 494-495
  - pronósticos de la regresión curvilínea, 500-501
- Inter-sujetos, factor, 395-396, 399, 402-403, 418-422
- Intra-sujetos, factor, 395-399, 401-402, 406, 411-414, 418-423
- Intraclase, coeficiente de correlación, 529-532

**K**

- Kappa* ( $\kappa$ ) de Cohen, 294-296
- Kendall:
  - coeficiente de concordancia *W*, 525, 565-571
  - coeficientes de correlación *tau-b* ( $\tau$ -b) y *tau-c* ( $\tau$ -c), 292-293, 433, 436-437
- Kolmogorov-Smirnov:
  - prueba de normalidad, 272-273
  - prueba de bondad de ajuste (prueba no paramétrica para una muestra) 544-546
  - prueba no paramétrica para dos muestras independientes, 550, 553
- Kruskal-Wallis, prueba no paramét. *H* de, 555-557
- Kuder-Richardson (KR20), 513

**L**

- Lambda* ( $\lambda$ ) de Goodman y Kruskal, 288-289, 291
- Lambda* ( $\lambda$ ) de Wilks, 400, 413, 420
- Leer datos de con formato texto, 45-50
- Levene, prueba sobre igualdad de varianzas, 275-278, 347-348, 358, 377, 422-423
- Lillieffors, prueba de normalidad, 272

**M**

- Mahalanobis, distancia, 483
- Mann-Whitney, prueba *U*, 547-548, 552, 557
- Mantel-Haenszel, prueba de independencia condicional, 306-308
- Mauchly, prueba *W* de esfericidad, 400, 413, 421
- McNemar, prueba para dos proporciones relacionadas, 301-306, 558, 567, 571
- McNemar-Bowker, 302, 304-305
- Media:
  - aritmética, 236-238, 244-245, 251, 259
  - error típico de la, 236, 245
  - móvil, 117
  - ponderada, 251, 252
  - recortada o truncada, 263, 265
- Mediana, 236-238, 251-253, 264-265, 267-269, 275-278, 539, 542-543, 556-560
- Mediana, prueba no paramétrica de la, 556-558
- Mediana móvil, 117
- Medias estimadas, 373-374, 405
- Medición, 209
- Medidas de asociación (ver también *Distancias*):
  - acuerdo (*kappa* de Cohen), 294-296
  - odds-ratio* (razón de ventajas), 298-301
  - odds-ratio* común, 308
  - para datos nominales, 287-291
    - basadas en *chi*-cuadrado, 287-291
    - basadas en la reducción proporcional del error, 288-291
  - para datos nominales-de intervalo, 294
  - para datos ordinales, 291-293
  - riesgo relativo, índice de, 297-301
  - simétricas y asimétricas, 288-289, 291-293
- Medidas repetidas, 338, 395-396, 398-399, 408-412, 418-419, 422-423
- Mínimos cuadrados, 363, 458, 461
- Mínimos cuadrados ponderados, 363
- Mixtos, efectos, 530
- Moda, 236-237, 246,
- Modelo lineal general, 333-335
- Moses, prueba no paramétrica de reacciones extremas, 549-550, 552-553

**Muestra, 213**

aleatoria, 154-155, 215

**Muestreo, 215-220**

afijación, 216

aleatorio estratificado, 216

aleatorio por conglomerados, 216-217

polietápico, 217

aleatorio sistemático, 215-216

distribución muestral, 217-220

probabilístico y no probabilístico, 215

**N**Nivel crítico (valor  $p$ ), 230

Nivel de confianza, 228, 232

Nivel de significación o riesgo, 228-229

Niveles de medida (nominal, ordinal, de intervalo, de razón), 209-212

No paramétrico, análisis, 533-571

caracterización y clasificación de los contrastes no paramétricos, 534-544

binomial, prueba, 538-541

*chi*-cuadrado sobre bondad de ajuste, 534-538

Cochran, prueba de, 525-526, 567, 570-571

Friedman, prueba de, 525, 564-570

Kendall, coeficiente de concordancia  $W$ , 525, 565-571

Kolmogorov-Smirnov, prueba para una muestra (bondad de ajuste), 522-523, 544-546

Kolmogorov-Smirnov, prueba para dos muestras independientes, 550, 553

Kruskal-Wallis, estadístico  $H$ , 555-557Mann-Whitney, estadístico  $U$ , 547-548, 552

mediana, prueba de la, 556-558

Moses, prueba de reacciones extremas, 549-550, 552-553

rachas, prueba para una muestra, 541-544

rachas, prueba de Wald-Wolfowitz para dos muestras independientes, 551-553

signos, prueba de los, 560-563

Wilcoxon (dos muestras independ.), 547, 552

Wilcoxon (dos muestras relacionadas) 559-562, 569

Nominal, escala o nivel de medida, 210

Normal, curva, 247-249

Normalidad:

gráficos de normalidad, 272-275, 477-478

pruebas de normalidad, 272-273

*Ntiles*, 101, 107

Nula, hipótesis, 224-226

Números aleatorios (ver *Semilla de aleatorización*)**O***Odds ratio* (razón de las ventajas), 298-301,*Odds ratio* común, 308

Ordenar casos, 123-124

Ordinal, escala o nivel de medida, 210-212

Ortogonal, diseño, 145-152

Ortogonales:

comparaciones, 355

polinomios, 355-357, 380, 406

**P** $p$ , nivel crítico, 230

Parámetro, 213-214, 335, 381

Parámetros, estimaciones de los, 376, 390, 393

Pearson:

coeficiente de correlación  $r_{xy}$ , 286-287, 432-433, 435-436, 440, 444, 448, 460, 462*chi*-cuadrado sobre bondad de ajuste, 534-538*chi*-cuadrado sobre independencia y homogeneidad de proporciones, 228-230

Pegar sintaxis, 5, 8, 201

Pegar variables, 68

Percentiles, 235-238, 265, 267

métodos de cálculo, 263-264

*Phi* ( $\phi$ ), coeficiente de correlación, 287-288, 448

Pillai, traza de, 400, 413, 420

Pivotar tablas (ver *Visor de resultados*)

Población, 212-213

Ponderar casos, 156-159

*Post hoc*, comparaciones (ver *Comparaciones múltiples post hoc*)

Potencia observada, 376, 379

Principales, efectos, 361, 364, 374, 404-405, 415-416

Proporciones relacionadas, 301-306, 525-526, 558, 567, 570-571

Prueba de significación 221 (ver *Contraste de hipótesis*)

Puntos medios de los intervalos, 237

Puntuaciones típicas ( $z$ ), 243, 247-249, 273-274**R**

Rachas, prueba no paramétrica de las, 541-544

Rachas, prueba de Wald-Wolfowitz, 551-553

Rango o amplitud, 236-238, 245, 252

Rangos:

asignar rangos, 101, 105-108

rangos empatados, 108

tipos de rangos, 106-107

- Razón (estadísticos para el cociente entre dos variables), 249-255
- Razón, escala de medida de, 211
- Razón de verosimilitudes, 285
- Recodificar:
- en distintas variables, 94-96
  - en las mismas variables, 92-94
  - recodificación automática, 109-110
- Reestructurar el archivo de datos, 125-134
- convertir casos en variables, 132-134
  - convertir variables en casos, 126-131
- Región crítica (ver *Zona crítica*)
- Regresión curvilínea, 497-505
- intervalos para los pronósticos, 500-501
  - pronósticos, 500-501
  - residuos, 500-501
  - modelos de estimación curvilínea, 499
- Regresión lineal, 355-396
- bondad de ajuste, 458-459, 460-461, 465
  - coeficiente de determinación ( $R^2$ ), 458-459
  - coeficiente de determinación corregido ( $R^2$  corregida), 458-460, 480, 486, 502-504
  - coeficientes de regresión parcial, 456-457, 462-463, 466-468, 479-481, 490-491, 494
  - intervalo de confianza, 468
  - coeficientes de regresión parcial tipificados, 462, 466-467, 480, 483
  - correlación parcial, 437-442, 470, 485-488
  - correlación semiparcial, 470, 486
  - diagramas de dispersión, 455-458, 463-464, 475-476
  - diagramas de dispersión parciales, 478-479
  - ecuación o recta de regresión, 455-458, 462-464, 466, 494-495, 499
  - estadísticos de influencia, 483-484
  - índices de condición, 481
  - pronósticos, 456-457, 493-495
  - corregidos, 475, 494
  - error típico de los, 494-495
  - intervalos de confianza para los, 494-495
  - no tipificados, 459, 471, 484, 494
  - tipificados, 473, 475-476, 484, 494
  - puntos de influencia, 482-484
  - regresión en formato ANOVA, 461, 465, 489-490,
  - regresión por pasos, 485-493
  - residuos, 458, 360-378
  - eliminados o corregidos, 375, 384
  - error típico de los, 461, 465, 468, 472, 492
  - estudentizados, 475
  - estudentizados corregidos, 475
  - histograma con curva normal, 477
  - no tipificados, 458, 460-461, 464, 471-474
  - tipificados, 472-478
  - supuestos del modelo de regresión, 371-381
  - homogeneidad o igualdad de varianzas (homocedasticidad), 471, 474-476
  - independencia, 471, 473-474
  - linealidad, 471, 478-479
  - no-colinealidad, 471, 479-481
  - normalidad, 471, 477-478
  - tolerancia, 480-481, 486, 491
  - variable dependiente o criterio, 455
  - variables independientes o predictoras, 455
- Residuos (errores):
- en el análisis de regresión (ver *Regresión lineal: residuos*)
  - en el análisis de varianza, 336, 378, 393, 407
  - en tablas de contingencia, 310-313
  - gráfico de los residuos, 301
- Respuesta múltiple, variables de, 573-584
- categorías múltiples, 573-574, 578-580
  - conjuntos de respuestas múltiples, 576-578
  - dicotomías múltiples, 573-574, 576-579
  - tablas de frecuencias, 578-580
  - tablas de contingencias, 580-584
- Riesgo, índices de:
- Cochran, 306-308
  - índice de riesgo relativo, 297-301
  - Mantel-Haenszel, 306-308
  - odds ratio* (razón de las ventajas), 298-301
  - odds ratio* común, 308
- Riesgo, nivel de, 228-229
- Roy, raíz mayor de, 400, 413, 420
- Ryan-Einot-Gabriel-Welsch ( $F$  y  $Q$ ), 351
- ## S
- Scheffé, prueba de, 350-351, 354
- Segmentar el archivo de datos, 152-153
- Segmentar ventana, 80
- Seleccionar casos, 153-156
- eliminar casos, 156
  - filtrar casos, 156
  - muestra aleatoria, 154-155
  - rango de casos, 155
  - definir filtro o condición, 154
  - variable de filtro, 156
- Semilla de aleatorización, 119-120
- Serie temporal, crear, 103-105

Shapiro-Wilk, prueba de normalidad, 272-273  
 Sidak, prueba de, 350  
 Significación, nivel de, 228-229  
 Signos, prueba no paramétrica de los, 560-563  
 Simétricas, medidas de asociación, 288-289, 291-293  
 Simples, efectos, 374-375, 383, 417, 424-426  
 Syntaxis SPSS, 5, 8, 32, 200-204  
 Somers, coeficiente *d*, 292-293  
 Spearman, coeficiente de correlación *rho*, 287, 433, 436-437  
 Spearman-Brown, profecía de, 515  
 Student:  
   prueba *T* para dos muestras independientes, 322-328  
   prueba *T* para dos muestras relacionadas, 328-332  
   prueba *T* para una muestra, 319-322  
 Student-Newman-Keuls, 351  
 Suavizado, 117  
 Suma de cuadrados, 345, 355, 364, 406-407  
 Sumas de cuadrados, tipos de, 386-387  
 Sumas de cuadrados y productos cruzados, 406-407, 434-436  
 Supuestos de un contraste:  
   aditividad, 528-529  
   esfericidad, 400-401, 407-408, 413-414, 420-421  
   homocedasticidad (homogeneidad o igualdad de varianzas), 275-278, 347-348, 358, 377-378, 323, 326-327, 347-348, 377, 422-423, 471, 474-476  
   independencia, 471, 473-474  
   linealidad, 471, 478-479  
   no-colinealidad, 471, 479-481  
   normalidad, 272-275, 471, 477-478  
   simetría, 560

## T

*t* de Student (ver *Student*)  
 $T^2$  de Hotelling, 400, 413, 420  
 Tablas de contingencias, 279-283  
   gráfico de barras agrupadas, 280-281  
   porcentajes de fila, de columna y totales, 309-313  
   tablas de contingencias segmentadas, 282-283  
 Tamaño del efecto, 375  
 Tamhane, estadístico  $T^2$ , 352  
 Tarone, prueba de homogeneidad de *odds-ratios*, 308

Tasa de error, 349, 366, 374,  
*Tau* ( $\tau$ ) de Goodman y Kruskal, 289, 291  
*Tau-b* ( $\tau$ -b) y *tau-c* ( $\tau$ -c) de Kendall, 292-293, 433, 436-437  
 Tendencia, comparaciones de (contrastes polinómicos), 355-357, 380, 406  
 Theil, coeficiente de incertidumbre de, 233-235  
 Típicas, puntuaciones *z*, 198, 203-204  
 Tolerancia, nivel de, 380-381, 386, 391  
 Transponer archivos, 124  
 Tukey:  
   prueba de aditividad, 528-529  
   pruebas para comparaciones múltiples (*HDS*, Tukey-b), 351, 353-354, 367-368  
 Tutorial del SPSS, 27, 31

## U

*U*, prueba no paramétrica de Mann-Whitney, 547-548, 552, 557  
 Universo (ver *Población*)

## V

Valor *p* (nivel crítico), 230  
 Valores atípicos, 263-264, 267, 269, 472, 484  
 Valores de influencia, 482-484  
 Valores extremos, 263-267, 269, 271, 277  
 Valores perdidos:  
   definir, 65-66  
   reemplazar, 118-119  
 Variables (ver *Definir variables*)  
 Varianza, 236-237, 244  
 Varianza, análisis de (ver *ANOVA*)  
 Varianzas-covarianzas, matriz de, 422-423, 407, 413, 469  
 Ventanas SPSS, 3-6  
 Visor de resultados, 161-198  
   anchura de las casillas, 188  
   aspectos de tabla, 186-187  
   características del texto, 187-188  
   contenido del Visor, 161-162  
   copiar resultados en otras aplicaciones, 193-194  
   editar resultados, 163-169  
   editar tablas, 169-173  
   enviar mensaje con los resultados, 162  
   esquema del Visor, 161-162  
   exportar resultados., 194-198  
   guardar con contraseña, 163  
   imprimir resultados, 189-193

- pivotar tablas, 173-177
  - paneles de pivotado, 174-176
  - señalizadores, 176-177
- propiedades de casilla, 183-186
- propiedades de tabla, 177-182

## W

- W, coeficiente de concordancia de Kendall, 525, 565-571
- W de Mauchly, 400, 413, 421
- Wald-Wolfowitz, prueba no paramétrica para dos muestras independientes, 551-553
- Waller-Duncan, prueba de, 352-354
- Welch, corrección para los grados de libertad, 323
- Welch, estadístico para el contraste de la igualdad de medias, 347-349

Wilcoxon:

- prueba no paramétrica para dos muestras independientes, 547, 552
- prueba no paramétrica para dos muestras relacionadas, 559-562, 569

Wilks, *lambda* ( $\lambda$ ) de, 400, 413, 420

## Y

Yates, corrección por continuidad de, 286, 307, 538, 542, 561

## Z

- z, puntuaciones típicas, 243, 247-249, 273-274
- Zona crítica o de rechazo, 228-229
- Zona de aceptación, 228-229

# Análisis de datos con SPSS 13 Base

Los profesores Antonio Pardo y Miguel Ángel Ruiz cuentan con una dilatada experiencia en análisis de datos, tanto desde un punto de vista docente (en la Universidad Autónoma de Madrid y en el departamento de formación de SPSS Ibérica) como desde el punto de vista aplicado (en numerosos centros de investigación públicos y privados).

Esta obra es fruto de esa experiencia. No se trata de un material diseñado exclusivamente para prestar ayuda al usuario más básico en el manejo de aplicaciones informáticas y en el conocimiento de herramientas estadísticas, sino también para servir de ayuda al usuario más avanzado. Todo ello prestando más atención a los aspectos prácticos o aplicados que a los teóricos o formales, aunque sin descuidar estos últimos. Por tanto, el propósito de este manual es doble: pretende servir de apoyo al usuario más novato en el manejo del paquete estadístico SPSS en su versión para Windows y, al mismo tiempo, ayudar al usuario más experimentado a comprender e interpretar los detalles asociados a cada técnica estadística.

Este libro contiene dos partes bien diferenciadas. La primera incluye, fundamentalmente, los procedimientos que permiten utilizar SPSS como gestor de datos. Describe cómo construir un archivo de datos y cómo preparar los datos para el análisis, y contiene una descripción pormenorizada de las tres ventanas principales de SPSS: el Editor de datos, el Visor de resultados y el Editor de sintaxis. La segunda parte incluye varios procedimientos que permiten utilizar SPSS como programa de análisis estadístico. En esta segunda parte se incluyen las técnicas de análisis estadístico más comúnmente ofertadas en los planes de estudio de las licenciaturas en las que se utiliza la estadística como herramienta de apoyo: estadística descriptiva, estadística exploratoria, contrastes sobre medias, análisis de varianza, análisis de correlación y regresión, estadística no paramétrica y fiabilidad de las escalas.

Sin duda, quienes se vean en la necesidad de utilizar el análisis estadístico de datos encontrarán en este manual una herramienta de gran ayuda para abordar todo tipo de tareas con el programa SPSS.

[www.mhe.es/pardo\\_spss](http://www.mhe.es/pardo_spss)

[www.mcgraw-hill.es](http://www.mcgraw-hill.es)

The McGraw-Hill Companies