

*A mi hermano Beni in memoriam. A mis
hijos.*

Cuqui

Para Julia y Cristina.
Javier

Prólogo

El desarrollo y el nivel de aplicación que la Bioestadística, como herramienta útil y rigurosa en el campo de la investigación en todas las Ciencias Sociales, ha experimentado en los últimos años, ha sido espectacular. Es indudable que este progreso en el conocimiento y aplicación de la Estadística ha venido estrechamente vinculado al que ha experimentado el área de la computación, que nos ha llevado a una sociedad absolutamente informatizada donde el ordenador se ha convertido en un utensilio personal de uso habitual. Este auge y progreso de la informática, a nivel de software y hardware, ha hecho posible, a su vez, la realización de pruebas estadísticas que, de forma habitual, hubiesen sido muy costosas desde el punto de vista humano así como manejar volúmenes de información que habrían resultado absolutamente impensables.

Un segundo factor asociado a este progreso del conocimiento en el ámbito estadístico, ha sido el cambio de actitud experimentado por todos los profesionales de las áreas de Ciencias Sociales y especialmente, en el ámbito de las Ciencias de la Salud. De una sociedad en la que los roles y el desempeño de la profesión estaban ajustados a la mera aplicación de los conocimientos adquiridos, hemos evolucionado a una Sociedad Científica donde la investigación ha pasado a formar parte esencial de su labor diaria. El interés por descubrir nuevos procedimientos a través de la experiencia acumulada, ha sido determinante en la necesidad de que todos estos profesionales se vean inmersos en la formación y aprendizaje de técnicas básicas de metodología de la investigación y de algunas más concretas como el análisis de datos.

Este cambio en la dimensión del ejercicio profesional, determina que los planes de estudio de todas las licenciaturas y diplomaturas incluyan la Bioestadística para el ámbito de Salud y Biología, como materia troncal con entidad propia y de auténtica necesidad. Se pretende, con ello, que un profesional de la Salud, o de cualquier ciencia Social, que se apoye en la cuantificación y en el estudio empírico de lo que observa a diario, entienda y conozca los conceptos básicos de la ciencia que le va a permitir, abandonando conductas pragmáticas, profundizar y comprender el fundamento científico de su área de trabajo.

No se trata de hacer expertos en Estadística. El principal objetivo de los docentes de esta materia se centra en generar, en los discentes, una

actitud crítica ante cualquier lectura científica, adquirir un lenguaje común con estadísticos y otros profesionales del área y conocer a priori los pasos y los elementos imprescindibles en cualquier investigación empírica que se apoye en el manejo de volúmenes grandes de datos y cuyo propósito final sea condensar dicha información para que pueda ser transmitida o extrapolar las conclusiones a las poblaciones de las que fueron tomadas las medidas. Es importante saber que no existe investigación si no existen objetivos previos: *no puede descartarse ni confirmarse lo que no se ha planteado.*

Ajena a esta transformación social se encuentran la gran mayoría de nuestros alumnos que cursan los primeros cursos de alguna de estas licenciaturas o diplomaturas de Ciencias Sociales o Ciencias de la Salud. Sus únicos objetivos se centran en llegar a ser médicos, biólogos, psicólogos... y no alcanzan a entender que utilidad les puede reportar una materia como la Bioestadística en su currículo. Es por ello que al margen de la dificultad intrínseca que genera el entendimiento de la materia, la enseñanza de la Bioestadística en estos cursos se ve agravada por la imposibilidad de usar cualquier tipo de motivación.

En muy distinta situación se encuentran los alumnos de postgrado que ya han comenzado su vida profesional y han tenido, por tanto, ocasión de darse cuenta de qué manera la Bioestadística les puede resultar útil y necesaria. Aunque no sea su deseo adentrarse en el mundo de la investigación, una parte importante en la transmisión de los nuevos hallazgos y conocimientos de otros colegas de su ámbito profesional, es el lenguaje estadístico. Es por ello que han de estar absolutamente familiarizados con dicha terminología si se pretende tener una actitud crítica y objetiva ante la lectura de cualquier literatura científica.

Fruto del trabajo realizado con estos sectores de estudiantes e investigadores es nuestra experiencia, que nos ha animado a escribir el presente libro que podría definirse como un Manual de Estadística básica aplicada al ámbito de la Salud. Su contenido abarca desde los aspectos más básicos de la Estadística descriptiva, en su función de resumir, presentar y comunicar los resultados de cualquier estudio a las diferentes técnicas de extrapolación de las conclusiones a una población, a partir de lo verificado en una muestra representativa de ésta. Obviamente, para ello, se hace necesario revisar las nociones más básicas de aspectos como probabilidad, Variable aleatoria,

Distribuciones de probabilidad, así como los elementos imprescindibles de toda la Inferencia Estadística: técnicas de muestreo, conceptos fundamentales, estimación confidencial y contrastes de hipótesis más importantes de la Estadística Univariante, abordando los test usados bajo supuesto de distribución gaussiana así como los de distribución libre. La variabilidad que han generado los nuevos planes de estudio no facilita la selección de unos contenidos que abarque la totalidad de los programas de todas las Universidades, sin embargo hay una parte troncal que constituye un porcentaje amplio del conjunto de todos ellos. Esta es la parte que hemos seleccionado, para nuestro contenido, de manera que podamos acercarnos lo máximo posible a lo que pudiera ser un libro de texto para las asignaturas de Bioestadística que se imparten en la mayoría de las Facultades de Medicina y Escuelas de Ciencias de la Salud.

En lo que concierne al modo y la forma, la experiencia acumulada a través de los años de docencia y el apoyo en el área de la investigación de los profesionales de la salud de nuestro entorno, nos condiciona a que teoría y práctica avancen de manera simultánea, en este manual, complementándose la una a la otra y apoyándose mutuamente, con numerosos ejemplos que puedan acercar al lector a situaciones más cotidianas de su entorno. Pretendemos con ello ayudarles a entender las nociones más abstractas y a relacionarlas con un futuro no lejano como profesional del mundo de la salud. No obstante, no hemos querido evitar tratar algunos temas con algo más de rigor, para que el lector que esté interesado en profundizar algo más, pueda hacerlo; siempre teniendo en cuenta que la lectura de dichas partes es algo optativo y que dependerá de las necesidades individuales.

A todos esos alumnos y compañeros queremos dedicarle nuestro más sincero agradecimiento, por su inestimable colaboración al orientarnos, a través de sus opiniones sinceras, sobre nuestra metodología docente y haber podido observar cual ha sido su evolución a lo largo de los años y de las diferentes etapas que se han ido sucediendo.

Esperamos que la ilusión puesta en la realización de este texto nos haya permitido suavizar, en la medida de lo posible, la aridez del tema que tratamos, y sólo comprobar que realmente pueda ser un elemento eficaz de ayuda, apoyo y consulta entre nuestros discípulos y compañeros, justificará todas las horas que hay detrás de estas líneas.

Índice general

1. Conceptos previos	13
1.1. Introducción	13
1.2. ¿Qué es la estadística?	14
1.3. Elementos. Población. Caracteres	15
1.4. Organización de los datos	17
1.4.1. Variables estadísticas	17
1.4.2. Tablas estadísticas	19
1.5. Representaciones Gráficas	21
1.5.1. Gráficos para variables cualitativas	22
1.5.2. Gráficos para variables cuantitativas	26
1.6. Problemas	36
2. Medidas descriptivas	39
2.1. Introducción	39
2.2. Estadísticos de tendencia central	40
2.2.1. La media	41
2.2.2. La mediana	43
2.2.3. La moda	46
2.2.4. Relación entre media, mediana y moda	47
2.3. Estadísticos de posición	48

2.4. Medidas de variabilidad o dispersión	55
2.4.1. Rango	55
2.4.2. Varianza	55
2.4.3. Desviación típica o estándar	56
2.4.4. Ejemplo de cálculo de medidas de dispersión	56
2.4.5. Coeficiente de variación	57
2.5. Asimetría y apuntamiento	59
2.5.1. Estadísticos de asimetría	60
2.5.2. Estadísticos de apuntamiento	66
2.6. Problemas	68
3. Variables bidimensionales	73
3.1. introducción	73
3.2. Tablas de doble entrada	75
3.2.1. Distribuciones condicionadas	76
3.3. Dependencia funcional e independencia	77
3.3.1. Dependencia funcional	77
3.3.2. Independencia	78
3.4. Covarianza	78
3.5. Coeficiente de correlación lineal de Pearson	81
3.6. Regresión	81
3.6.1. Bondad de un ajuste	84
3.6.2. Regresión lineal	86
3.7. Problemas	94
4. Cálculo de probabilidades y variables aleatorias	99
4.1. introducción	99
4.2. Experimentos y sucesos aleatorios	100
4.2.1. Operaciones básicas con sucesos aleatorios	101

4.3. Experimentos aleatorios y probabilidad	102
4.3.1. Noción frecuentista de probabilidad	102
4.3.2. Probabilidad de Laplace	105
4.3.3. Definición axiomática de probabilidad	105
4.4. Probabilidad condicionada e independencia de sucesos . . .	106
4.5. Teoremas fundamentales del cálculo de probabilidades . . .	109
4.5.1. Teorema de la probabilidad compuesta	110
4.5.2. Sistema exhaustivo y excluyente de sucesos	110
4.5.3. Teorema de la probabilidad total	111
4.5.4. Teorema de Bayes	112
4.6. Tests diagnósticos	115
4.7. Problemas	119
5. Variables aleatorias	123
5.1. Introducción	123
5.2. Variables aleatorias discretas	125
5.3. Variables aleatorias continuas	126
5.4. Medidas de tendencia central y dispersión de v.a.	129
5.4.1. Valor esperado o esperanza matemática	130
5.4.2. Varianza	130
6. Principales leyes de distribución de variables aleatorias	131
6.1. Introducción	131
6.2. Distribuciones discretas	132
6.2.1. Distribución de Bernoulli	132
6.2.2. Distribución binomial	133
6.2.3. Distribución geométrica (o de fracasos)	137
6.2.4. Distribución binomial negativa	139
6.2.5. Distribución hipergeométrica	141

6.2.6. Distribución de Poisson o de los sucesos raros	143
6.3. Distribuciones continuas	144
6.3.1. Distribución uniforme o rectangular	144
6.3.2. Distribución exponencial	146
6.3.3. Distribución normal o gaussiana	150
6.3.4. Distribución χ^2	153
6.3.5. Distribución t de Student	155
6.3.6. La distribución F de Snedecor	157
6.4. Problemas	159
7. Introducción a la inferencia	163
7.1. Introducción	163
7.2. Técnicas de muestreo sobre una población	164
7.2.1. Muestreo aleatorio	165
7.2.2. Muestreo aleatorio estratificado	166
7.2.3. Muestreo sistemático	168
7.2.4. Muestreo por conglomerados	169
7.3. Propiedades deseables de un estimador	169
7.3.1. Estimadores de máxima verosimilitud	170
7.3.2. Algunos estimadores fundamentales	172
8. Estimación confidencial	175
8.1. Introducción	175
8.2. Intervalos de confianza para la distribución normal	177
8.2.1. Intervalo para la media si se conoce la varianza	178
8.2.2. Intervalo para la media (caso general)	182
8.2.3. Intervalo de confianza para la varianza	186
8.2.4. Estimación del tamaño muestral	187

8.2.5. Intervalos para la diferencia de medias de dos poblaciones	189
8.3. Intervalos de confianza para variables dicotómicas	195
8.3.1. Intervalo para una proporción	195
8.3.2. Elección del tamaño muestral para una proporción	197
8.3.3. Intervalo para la diferencia de dos proporciones	198
8.4. Problemas	200
9. Contrastes de hipótesis	203
9.1. Introducción	203
9.1.1. Observaciones	206
9.2. Contrastes paramétricos en una población normal	210
9.2.1. Contrastes para la media	210
9.2.2. Contrastes para la varianza	218
9.3. Contrastes de una proporción	219
9.4. Contrastes para la diferencia de medias apareadas	224
9.5. Contrastes de dos distribuciones normales independientes	228
9.5.1. Contraste de medias con varianzas conocidas	228
9.5.2. Contraste de medias homocedáticas	231
9.5.3. Contraste de medias no homocedáticas	232
9.5.4. Contrastes de la razón de varianzas	234
9.5.5. Caso particular: Contraste de homocedasticidad	236
9.6. Contrastes sobre la diferencia de proporciones	244
9.7. Problemas	246
10. Contrastes basados en el estadístico Ji-Cuadrado	255
10.1. Introducción	255
10.2. El estadístico χ^2 y su distribución	256
10.3. Contraste de bondad de ajuste para distribuciones	264

10.3.1. Distribuciones de parámetros conocidos	265
10.3.2. Distribuciones con parámetros desconocidos	268
10.4. Contraste de homogeneidad de muestras cualitativas	269
10.5. Contraste de independencia de variables cualitativas	272
10.6. Problemas	278
11. Análisis de la varianza	283
11.1. Introducción	283
11.2. ANOVA con un factor	285
11.2.1. Especificación del modelo	287
11.2.2. Algo de notación relativa al modelo	289
11.2.3. Forma de efectuar el contraste	291
11.2.4. Método reducido para el análisis de un factor	292
11.2.5. Análisis de los resultados del ANOVA: Comparaciones múltiples	295
11.3. Consideraciones sobre las hipótesis subyacentes en el modelo factorial	297
11.3.1. Contraste de homocedasticidad de Cochran	298
11.3.2. Contraste de homocedasticidad de Bartlett	299
11.4. Problemas	301
12. Contrastes no paramétricos	305
12.1. Introducción	305
12.2. Aleatoriedad de una muestra: Test de rachas	306
12.3. Normalidad de una muestra: Test de D'Agostino	308
12.4. Equidistribución de dos poblaciones	309
12.4.1. Contraste de rachas de Wald—Wolfowitz	309
12.4.2. Contraste de Mann—Withney	310
12.5. Contraste de Wilcoxon para muestras apareadas	311

<i>ÍNDICE GENERAL</i>	11
12.6. Contraste de Kruskal–Wallis	313
12.7. Problemas	314
Bibliografía	321

Capítulo 1

Conceptos previos

1.1. Introducción

Iniciamos este capítulo con la definición de algunos conceptos elementales y básicos, y sin embargo pilares, para una comprensión intuitiva y real de lo que es la Bioestadística. Pretendemos introducir al estudiante en los primeros pasos sobre el uso y manejos de datos numéricos: distinguir y clasificar las características en estudio, enseñarle a organizar y tabular las medidas obtenidas mediante la construcción de tablas de frecuencia y por último los métodos para elaborar una imagen que sea capaz de mostrar gráficamente unos resultados.

El aserto “una imagen vale más que mil palabras” se puede aplicar al ámbito de la estadística descriptiva diciendo que “un gráfico bien elaborado vale más que mil tablas de frecuencias”. Cada vez es más habitual el uso de gráficos o imágenes para representar la información obtenida. No obstante, debemos ser prudente al confeccionar o interpretar gráficos, puesto que una misma información se puede representar de formas muy diversas, y no todas ellas son pertinentes, correctas o válidas. Nuestro objetivo, en este capítulo, consiste en establecer los criterios y normas mínimas que deben verificarse para construir y presentar adecuadamente los gráficos en el ámbito de la estadística descriptiva.

1.2. ¿Qué es la estadística?

Cuando coloquialmente se habla de estadística, se suele pensar en una relación de datos numéricos presentada de forma ordenada y sistemática. Esta idea es la consecuencia del concepto popular que existe sobre el término y que cada vez está más extendido debido a la influencia de nuestro entorno, ya que hoy día es casi imposible que cualquier medio de difusión, periódico, radio, televisión, etc, no nos aborde diariamente con cualquier tipo de información estadística sobre accidentes de tráfico, índices de crecimiento de población, turismo, tendencias políticas, etc.

Sólo cuando nos adentramos en un mundo más específico como es el campo de la investigación de las Ciencias Sociales: Medicina, Biología, Psicología, ... empezamos a percibir que la Estadística no sólo es algo más, sino que se convierte en la única herramienta que, hoy por hoy, permite dar luz y obtener resultados, y por tanto beneficios, en cualquier tipo de estudio, cuyos movimientos y relaciones, por su variabilidad intrínseca, no puedan ser abordadas desde la perspectiva de las leyes deterministas. Podríamos, desde un punto de vista más amplio, definir la estadística como la ciencia que estudia cómo debe emplearse la información y cómo dar una guía de acción en situaciones prácticas que entrañan incertidumbre.

La **Estadística** se ocupa de los métodos y procedimientos para recoger, clasificar, resumir, hallar regularidades y analizar los *datos*, siempre y cuando la variabilidad e *incertidumbre* sea una causa intrínseca de los mismos; así como de realizar *inferencias* a partir de ellos, con la finalidad de ayudar a la toma de *decisiones* y en su caso formular *predicciones*.

Podríamos por tanto clasificar la Estadística en descriptiva, cuando los resultados del análisis no pretenden ir más allá del conjunto de datos, e inferencial cuando el objetivo del estudio es derivar las conclusiones obtenidas a un conjunto de datos más amplio.

Estadística descriptiva: Describe, analiza y representa un grupo de datos utilizando métodos numéricos y gráficos que resumen y presentan la información contenida en ellos.

Estadística inferencial: Apoyándose en el cálculo de probabilidades y a partir de datos muestrales, efectúa estimaciones, decisiones, predicciones u otras generalizaciones sobre un conjunto mayor de datos.

1.3. Elementos. Población. Caracteres

Establecemos a continuación algunas definiciones de conceptos básicos y fundamentales básicas como son: elemento, población, muestra, caracteres, variables, etc., a las cuales haremos referencia continuamente a lo largo del texto

Individuos o elementos: personas u objetos que contienen cierta información que se desea estudiar.

Población: conjunto de individuos o elementos que cumplen ciertas propiedades comunes.

Muestra: subconjunto representativo de una población.

Parámetro: función definida sobre los valores numéricos de características medibles de una población.

Estadístico: función definida sobre los valores numéricos de una muestra.

En relación al tamaño de la población, ésta puede ser:

- **Finita**, como es el caso del número de personas que llegan al servicio de urgencia de un hospital en un día;
- **Infinita**, si por ejemplo estudiamos el mecanismo aleatorio que describe la secuencia de caras y cruces obtenida en el lanzamiento repetido de una moneda al aire.

Caracteres: propiedades, rasgos o cualidades de los elementos de la población. Estos caracteres pueden dividirse en cualitativos y cuantitativos.

Modalidades: diferentes situaciones posibles de un carácter. Las modalidades deben ser a la vez exhaustivas y mutuamente excluyentes —cada elemento posee una y sólo una de las modalidades posibles.

Clases: conjunto de una o más modalidades en el que se verifica que cada modalidad pertenece a una y sólo una de las clases.

1.4. Organización de los datos

1.4.1. Variables estadísticas

Cuando hablemos de **variable** haremos referencia a un símbolo (X, Y, A, B, \dots) que puede tomar cualquier **modalidad** (valor) de un conjunto determinado, que llamaremos **dominio de la variable** o **rango**. En función del tipo de dominio, las variables las clasificamos del siguiente modo:

Variables cualitativas, cuando las modalidades posibles son de tipo nominal. Por ejemplo, el grupo sanguíneo tiene por modalidades:

Grupos Sanguíneos posibles: A, B, AB, O

Variables cuasicuantitativas u ordinales son las que, aunque sus modalidades son de tipo nominal, es posible establecer un orden entre ellas. Por ejemplo, si estudiamos el grado de recuperación de un paciente al aplicarle un tratamiento, podemos tener como modalidades: Grado de recuperación: *Nada, Poco, Moderado, Bueno, Muy Bueno*.

A veces se representan este tipo de variables en escalas numéricas, por ejemplo, puntuar el dolor en una escala de 1 a 5. Debemos evitar sin embargo realizar operaciones algebraicas con estas cantidades. ¡Un dolor de intensidad 4 no duele el doble que otro de intensidad 2!

Variables cuantitativas o numéricas son las que tienen por modalidades cantidades numéricas con las que podemos hacer operaciones aritméticas. Dentro de este tipo de variables podemos distinguir dos grupos:

Discretas, cuando no admiten siempre una modalidad intermedia entre dos cualesquiera de sus modalidades. Un ejemplo es el número de hijos en una población de familias:

Número de hijos posibles: $0, 1, 2, 3, 4, 5, \dots$

Continuas, cuando admiten una modalidad intermedia entre dos cualesquiera de sus modalidades, v.g. el peso X de un niño al nacer.

Ocurre a veces que una variable cuantitativa continua por naturaleza, aparece como discreta. Este es el caso en que hay limitaciones en lo

que concierne a la precisión del aparato de medida de esa variable, v.g. si medimos la altura en metros de personas con una regla que ofrece dos decimales de precisión, podemos obtener

Alturas medidas en cm: 1.50 , 1.51 , 1.52 , $1.53, \dots$

En realidad lo que ocurre es que con cada una de esas mediciones expresamos que el verdadero valor de la misma se encuentra en un intervalo de radio $0,005$. Por tanto cada una de las observaciones de X representa más bien un intervalo que un valor concreto.

Tal como hemos citado anteriormente, las modalidades son las diferentes situaciones posibles que puede presentar la variable. A veces éstas son muy numerosas (v.g. cuando una variable es continua) y conviene reducir su número, agrupándolas en una cantidad inferior de **clases**. Estas clases deben ser construidas, tal como hemos citado anteriormente, de modo que sean *exhaustivas* y *excluyentes*, es decir, cada modalidad debe pertenecer a una y sólo una de las clases.

Variable cualitativa: Aquella cuyas modalidades son de tipo nominal.

Variable cuasicuantitativa: Modalidades de tipo nominal, en las que existe un orden.

Variable cuantitativa discreta: Sus modalidades son valores enteros.

Variable cuantitativa continua: Sus modalidades son valores reales.

1.4.2. Tablas estadísticas

Consideremos una población estadística de n individuos, descrita según un carácter o variable C cuyas modalidades han sido agrupadas en un número k de clases, que denotamos mediante c_1, c_2, \dots, c_k . Para cada una de las clases $c_i, i = 1, \dots, k$, introducimos las siguientes magnitudes:

Frecuencia absoluta de la clase c_i es el número n_i , de observaciones que presentan una modalidad perteneciente a esa clase.

Frecuencia relativa de la clase c_i es el cociente f_i , entre las frecuencias absolutas de dicha clase y el número total de observaciones, es decir

$$f_i = \frac{n_i}{n}$$

Obsérvese que f_i es el *tanto por uno* de observaciones que están en la clase c_i . Multiplicado por 100 % representa el porcentaje de la población que comprende esa clase.

Frecuencia absoluta acumulada N_i , se calcula sobre variables cuantitativas o cuasicuantitativas, y es el número de elementos de la población cuya modalidad es inferior o equivalente a la modalidad c_i :

$$N_i = n_1 + n_2 + \dots + n_i = \sum_{j=1}^i n_j$$

Frecuencia relativa acumulada , F_i , se calcula sobre variables cuantitativas o cuasicuantitativas, siendo el tanto por uno de los elementos de la población que están en alguna de las clases y que presentan una modalidad inferior o igual a la c_i , es decir,

$$F_i = \frac{N_i}{n} = \frac{n_1 + \dots + n_i}{n} = f_1 + \dots + f_i = \sum_{j=1}^i f_j$$

Llamaremos **distribución de frecuencias** al conjunto de clases junto a las frecuencias correspondientes a cada una de ellas. Una **tabla estadística**

sirve para presentar de forma ordenada las distribuciones de frecuencias. Su forma general es la siguiente:

Modali.	Frec. Abs.	Frec. Rel.	Frec. Abs. Acumu.	Frec. Rel. Acumu.
C	n_i	f_i	N_i	F_i
c_1	n_1	$f_1 = \frac{n_1}{n}$	$N_1 = n_1$	$F_1 = \frac{N_1}{n} = f_1$
...
c_j	n_j	$f_j = \frac{n_j}{n}$	$N_j = n_1 + \dots + n_j$	$F_j = \frac{N_j}{n} = f_1 + \dots + f_j$
...
c_k	n_k	$f_k = \frac{n_k}{n}$	$N_k = n$	$F_k = 1$
	n	1		

Ejemplo de cálculo con frecuencias

Calcular los datos que faltan en la siguiente tabla:

$l_{i-1} - l_i$	n_i	f_i	N_i
0 — 10	60	f_1	60
10 — 20	n_2	0,4	N_2
20 — 30	30	f_3	170
30 — 100	n_4	0,1	N_4
100 — 200	n_5	f_5	200
	n		

Solución:

Sabemos que la última frecuencia acumulada es igual al total de observaciones, luego $n = 200$.

Como $N_3 = 170$ y $n_3 = 30$, entonces

$$N_2 = N_3 - n_3 = 170 - 30 = 140.$$

Además al ser $n_1 = 60$, tenemos que

$$n_2 = N_2 - n_1 = 140 - 60 = 80.$$

Por otro lado podemos calcular n_4 teniendo en cuenta que conocemos la frecuencia relativa correspondiente:

$$f_4 = \frac{n_4}{n} \quad \Rightarrow \quad n_4 = f_4 \cdot n = 0,1 \times 200 = 20$$

Así:

$$N_4 = n_4 + N_3 = 20 + 170 = 190.$$

Este último cálculo nos permite obtener

$$n_5 = N_5 - N_4 = 200 - 190 = 10.$$

Al haber calculado todas las frecuencias absolutas, es inmediato obtener las relativas:

$$\begin{aligned} f_1 &= \frac{n_1}{n} = \frac{60}{200} = 0,3 \\ f_3 &= \frac{n_3}{n} = \frac{30}{200} = 0,15 \\ f_5 &= \frac{n_5}{n} = \frac{10}{200} = 0,05 \end{aligned}$$

Escribimos entonces la tabla completa:

$l_{i-1} - l_i$	n_i	f_i	N_i
0 — 10	60	0,3	60
10 — 20	80	0,4	140
20 — 30	30	0,15	170
30 — 100	20	0,1	190
100 — 200	10	0,05	200
	200		

1.5. Representaciones Gráficas

Hemos visto que la tabla estadística resume los datos que disponemos de una población, de forma que ésta se puede analizar de una manera más

sistemática y resumida . Para darnos cuenta *de un sólo vistazo* de las características de la población resulta aún más esclarecedor el uso de gráficos y diagramas, cuya construcción abordamos en esta sección.

1.5.1. Gráficos para variables cualitativas

Los gráficos más usuales para representar variables de tipo nominal son los siguientes:

Diagramas de barras: Siguiendo la figura 1.1, representamos en el eje de ordenadas las modalidades y en abscisas las frecuencias absolutas o bien, las frecuencias relativas. Si, mediante el gráfico, se intenta comparar varias poblaciones entre sí, existen otras modalidades, como las mostradas en la figura 1.2. Cuando los tamaños de las dos poblaciones son diferentes, es conveniente utilizar las frecuencias relativas, ya que en otro caso podrían resultar engañosas.

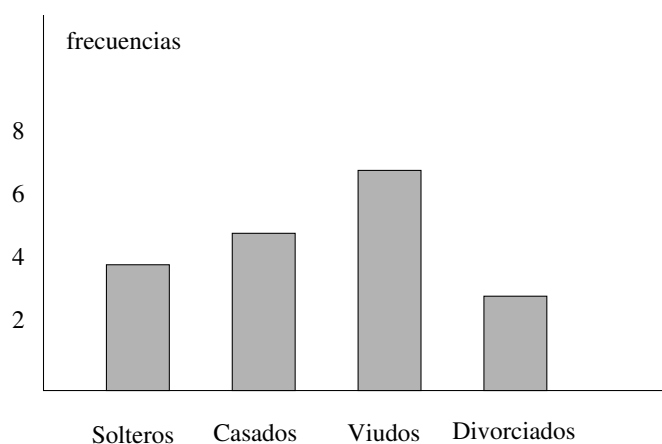


Figura 1.1: Diagrama de barras para una variable cualitativa.

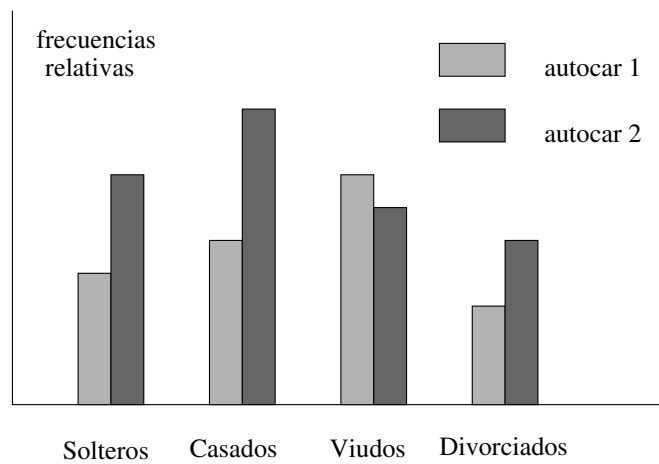


Figura 1.2: Diagramas de barras para comparar una variable cualitativa en diferentes poblaciones. Se ha de tener en cuenta que la altura de cada barra es *proporcional* al número de observaciones (frecuencias relativas).

Diagramas de sectores (también llamados *tartas*). Se divide un círculo en tantas porciones como clases existan, de modo que a cada clase le corresponde un arco de círculo proporcional a su frecuencia absoluta o relativa (figura 1.3).

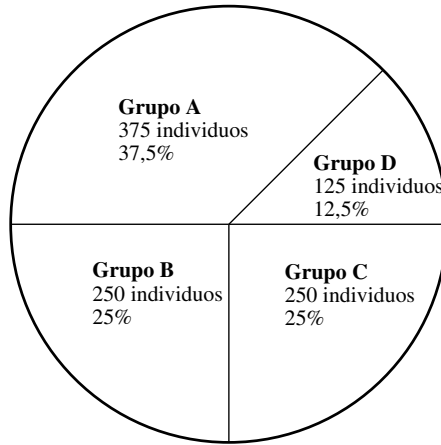


Figura 1.3: Diagrama de sectores.

El arco de cada porción se calcula usando la *regla de tres*:

$$\begin{array}{lcl} n & \longrightarrow & 360^\circ \\ n_i & \longrightarrow & x_i = \frac{360 \cdot n_i}{n} \end{array}$$

Como en la situación anterior, puede interesar comparar dos poblaciones. En este caso también es aconsejable el uso de las frecuencias relativas (porcentajes) de ambas sobre gráficos como los anteriores. Otra posibilidad es comparar las 2 poblaciones usando para cada una de ellas un diagrama semicircular, al igual que en la figura 1.4. Sean $n_1 \leq n_2$ los tamaños respectivos de las 2 poblaciones. La población más pequeña se representa con un semicírculo de radio r_1 y la mayor con otro de radio r_2 .

La relación existente entre los radios, es la que se obtiene de suponer que la relación entre las áreas de las circunferencias es igual a la de los tamaños de las poblaciones respectivas, es decir:

$$\frac{r_2^2}{r_1^2} = \frac{n_2}{n_1} \iff r_2 = r_1 \cdot \sqrt{\frac{n_2}{n_1}}$$

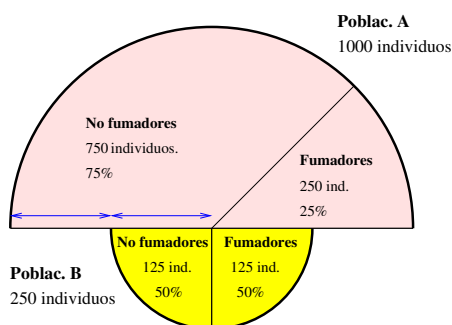


Figura 1.4: Diagrama de sectores para comparar dos poblaciones

Pictogramas Expresan con dibujos alusivo al tema de estudio las frecuencias de las modalidades de la variable. Estos gráficos se hacen representado a diferentes escalas un mismo dibujo, como vemos en la figura 1.5.

El escalamiento de los dibujos debe ser tal que el *área*¹ de cada uno de ellos sea proporcional a la frecuencia de la modalidad que representa. Este tipo de gráficos suele usarse en los medios de comunicación, para que sean comprendidos por el público no especializado, sin que sea necesaria una explicación compleja.

¹Es un error hacer la representación con una escala tal que el *perímetro* del dibujo sea proporcional a la frecuencia, ya que a frecuencia doble, correspondería un dibujo de área cuádruple, lo que da un efecto visual engañoso.

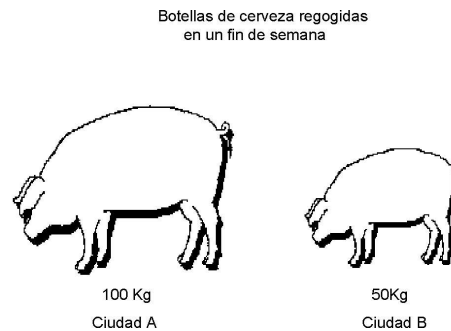


Figura 1.5: Pictograma. Las áreas son proporcionales a las frecuencias.

1.5.2. Gráficos para variables cuantitativas

Para las variables cuantitativas, consideraremos dos tipos de gráficos, en función de que para realizarlos se usen las frecuencias (absolutas o relativas) o las frecuencias acumuladas:

Diagramas diferenciales: Son aquellos en los que se representan frecuencias absolutas o relativas. En ellos se representa el número o porcentaje de elementos que presenta una modalidad dada.

Diagramas integrales: Son aquellos en los que se representan el número de elementos que presentan una modalidad inferior o igual a una dada. Se realizan a partir de las frecuencias acumuladas, lo que da lugar a gráficos crecientes, y es obvio que este tipo de gráficos no tiene sentido para variables cualitativas.

Según hemos visto existen dos tipos de variables cuantitativas: discretas y continuas. Vemos a continuación las diferentes representaciones gráficas que pueden realizarse para cada una de ellas así como los nombres específicos que reciben.

Gráficos para variables discretas

Cuando representamos una variable discreta, usamos el **diagrama de barras** cuando pretendemos hacer una gráfica diferencial. Las barras deben ser estrechas para representar el que los valores que toma la variable son discretos. El diagrama integral o acumulado tiene, por la naturaleza de la variable, forma de escalera. Un ejemplo de diagrama de barras así como su diagrama integral correspondiente están representados en la figura 1.6.

Ejemplo de variable discreta

Se lanzan tres monedas al aire en 8 ocasiones y se contabiliza el número de caras, X , obteniéndose los siguientes resultados:

$2, 1, 0, 1, 3, 2, 1, 2$

Representar gráficamente el resultado.

Solución: En primer lugar observamos que la variable X es cuantitativa discreta, presentando las modalidades: $0, 1, 2, 3$

Ordenamos a continuación los datos en una tabla estadística, y se representa la misma en la figura 1.6.

x_i	n_i	f_i	N_i	F_i
0	1	1/8	1	1/8
1	3	3/8	4	4/8
2	3	3/8	7	7/8
3	1	1/8	8	8/8
$n = 8$		1		

Ejemplo de representación gráfica

Clasificadas 12 familias por su número de hijos se obtuvo:

Número de hijos (x_i)	1	2	3	4
Frecuencias (n_i)	1	3	5	3

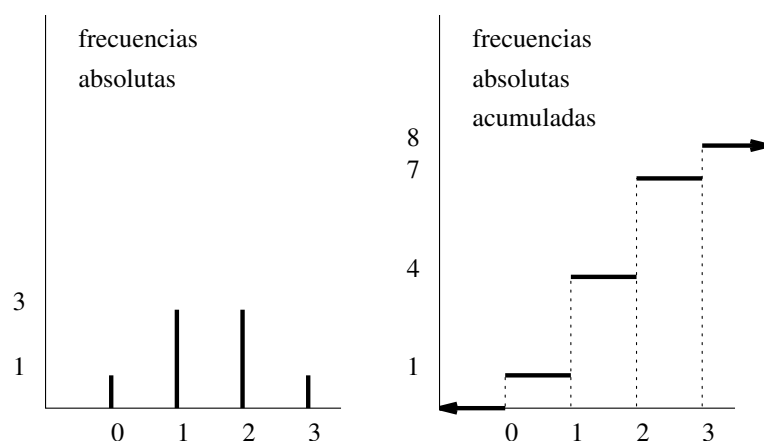


Figura 1.6: Diagrama diferencial (barras) e integral para una variable discreta. Obsérvese que el diagrama integral (creciente) contabiliza el número de observaciones de la variable inferiores o iguales a cada punto del eje de abscisas.

Comparar los diagramas de barras para frecuencias absolutas y relativas. Realizar el diagrama acumulativo creciente.

Solución: En primer lugar, escribimos la tabla de frecuencias en el modo habitual:

Variable	F. Absolutas	F. Relativas	F. Acumuladas
x_i	n_i	f_i	N_i
1	1	0,083	1
2	3	0,250	4
3	5	0,416	9
4	3	0,250	12
	12	1	

Con las columnas relativas a x_i y n_i realizamos el diagrama de barras para frecuencias absolutas, lo que se muestra en la figura 1.7. Como puede verse es idéntico (salvo un cambio de escala en el eje de ordenadas) al diagrama de barras para frecuencias relativas y que ha sido calculado

usando las columnas de x_i y f_i . El diagrama escalonado (acumulado) se ha construido con la información procedente de las columnas x_i y N_i .

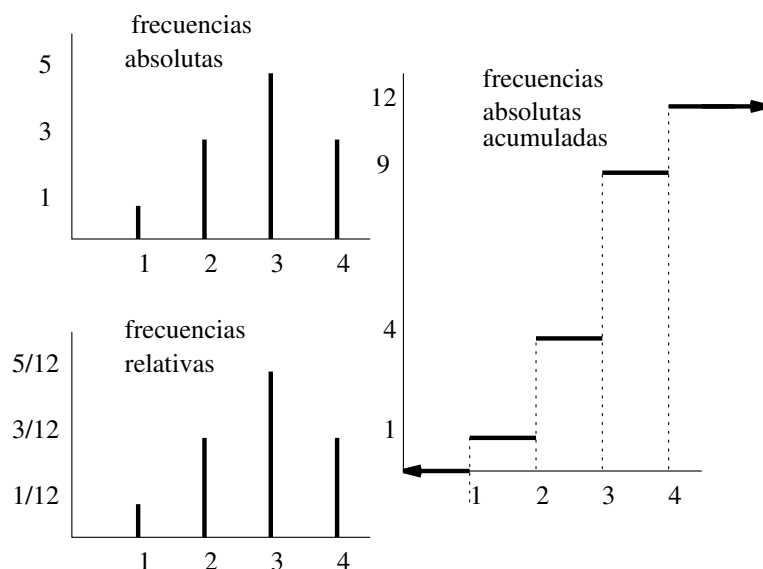


Figura 1.7: Diagramas de frecuencias para una variable discreta

Gráficos para variables continuas

Cuando las variables son continuas, utilizamos como diagramas diferenciales los *histogramas* y los *polígonos de frecuencias*.

Un *histograma* se construye a partir de la tabla estadística, representando sobre cada intervalo, un rectángulo que tiene a este segmento como base. El criterio para calcular la altura de cada rectángulo es el de mantener la proporcionalidad entre las frecuencias absolutas (o relativas) de cada intervalo y el área de los mismos. Véase la figura 1.8.

El *polígono de frecuencias* se construye fácilmente si tenemos representado previamente el histograma, ya que consiste en unir mediante líneas rectas los puntos del histograma que corresponden a las marcas de clase. Para representar el polígono de frecuencias en el primer y último interva-

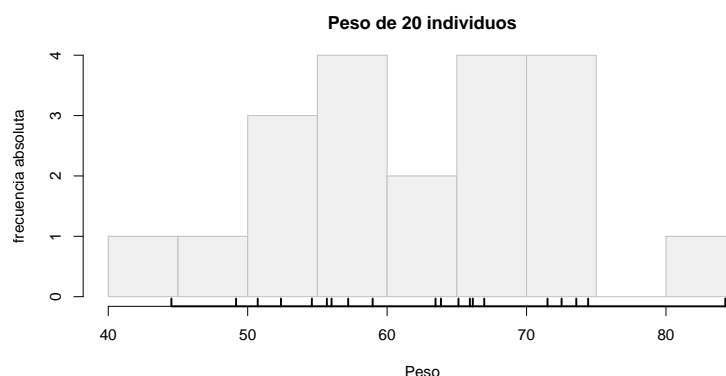


Figura 1.8: Histograma para una variable continua.

lo, suponemos que adyacentes a ellos existen otros intervalos de la misma amplitud y frecuencia nula, y se unen por una línea recta los puntos del histograma que corresponden a sus marcas de clase. Obsérvese que de este modo, el polígono de frecuencias tiene en común con el histograma el que las áreas de la gráficas sobre un intervalo son idénticas. Veanse ambas gráficas diferenciales representadas en la parte superior de la figura 1.9.

El diagrama integral para una variable continua se denomina también **polígono de frecuencias acumulado**, y se obtiene como la poligonal definida en abcisas a partir de los extremos de los intervalos en los que hemos organizado la tabla de la variable, y en ordenadas por alturas que son proporcionales a las frecuencias acumuladas. Dicho de otro modo, el polígono de frecuencias absolutas es una primitiva del histograma. Véase la parte inferior de la figura 1.9, en la que se representa a modo de ilustración los diagramas correspondientes a la variable cuantitativa continua expresada en la tabla siguiente:

Intervalos	c_i	n_i	N_i
0 — 2	1	2	2
2 — 4	3	1	3
4 — 6	5	4	7
6 — 8	7	3	10
8 — 10	9	2	12
		12	

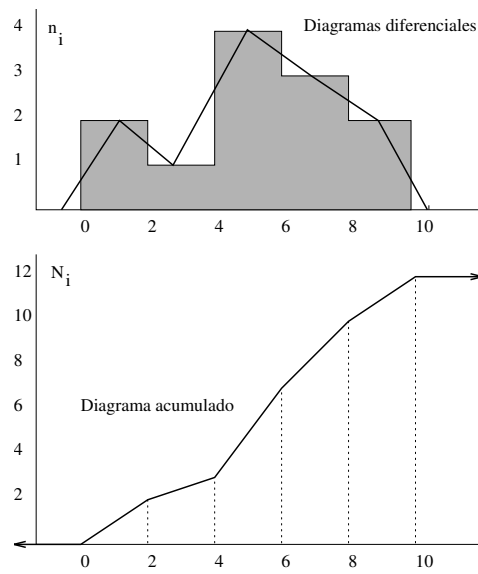


Figura 1.9: Diagramas diferenciales e integrales para una variable continua.

Ejemplo

La siguiente distribución se refiere a la duración en horas (completas) de un lote de 500 tubos:

Duración en horas	Número de tubos
300 — 500	50
500 — 700	150
700 — 1.100	275
más de 1.100	25
Total 500	

- Representar el histograma de frecuencias relativas y el polígono de frecuencias.
- Trazar la curva de frecuencias relativas acumuladas.
- Determinar el número mínimo de tubos que tienen una duración inferior a 900 horas.

Solución: En primer lugar observamos que la variable en estudio es discreta (*horas completas*), pero al tener un rango tan amplio de valores resulta más conveniente agruparla en intervalos, como si de una variable continua se tratase. La consecuencia es una ligera pérdida de precisión.

El último intervalo está abierto por el límite superior. Dado que en él hay 25 observaciones puede ser conveniente cerrarlo con una amplitud “razonable”. Todos los intervalos excepto el tercero tienen una amplitud de 200 horas, luego podríamos cerrar el último intervalo en 1.300 horas².

Antes de realizar el histograma conviene hacer una observación importante. El histograma representa las frecuencias de los intervalos mediante *áreas* y no mediante *alturas*. Sin embargo nos es mucho más fácil hacer representaciones gráficas teniendo en cuenta estas últimas. Si todos los intervalos tienen la misma amplitud no es necesario diferenciar entre los

²Cualquier otra elección para el límite superior del intervalo que sea de “sentido común” sería válida.

conceptos de área y altura, pero en este caso el tercer intervalo tiene una amplitud doble a los demás, y por tanto hay que repartir su área en un rectángulo de base doble (lo que reduce su altura a la mitad).

Así será conveniente añadir a la habitual tabla de frecuencias una columna que represente a las amplitudes a_i de cada intervalo, y otra de frecuencias relativas rectificadas, f'_i , para representar la altura del histograma. Los gráficos requeridos se representan en las figuras 1.10 y 1.11.

Intervalos	a_i	n_i	f_i	f'_i	F_i
300 — 500	200	50	0,10	0,10	0,10
500 — 700	200	150	0,30	0,30	0,40
700 — 1.100	400	275	0,55	0,275	0,95
1.100 — 1.300	200	25	0,05	0,05	1,00
n=500					

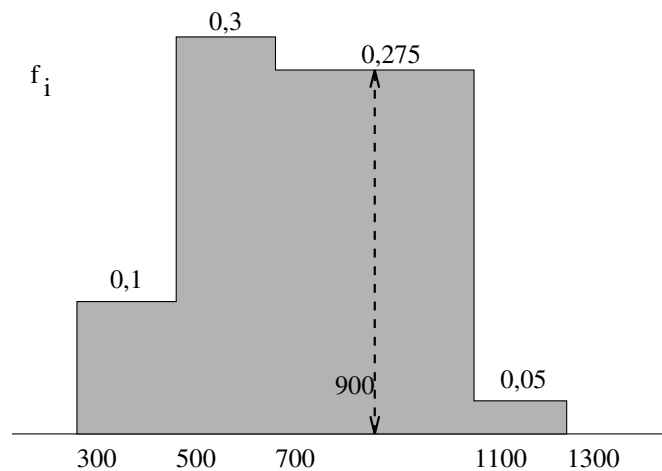


Figura 1.10: Histograma. Obsérvese que la altura del histograma en cada intervalo es f'_i que coincide en todos con f_i salvo en el intervalo 700 — 1.100 en el que $f'_i = 1/2 f_i$ ya que la amplitud de ese intervalo es doble a la de los demás.

Por otro lado, mirando la figura 1.10 se ve que sumando frecuencias relati-

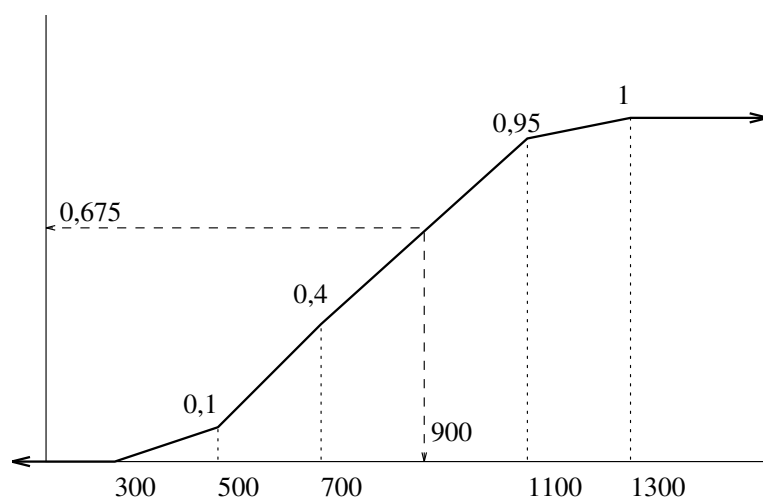


Figura 1.11: Diagrama acumulativo de frecuencias relativas

vas, hasta las 900 horas de duración hay

$$0,10 + 0,30 + 0,275 = 0,675 = 67,5 \% \text{ de los tubos.}$$

Esta cantidad se obtiene de modo más directo viendo a qué altura corresponde al valor 900 en el diagrama de frecuencias acumuladas (figura 1.11).

Como en total son 500 tubos, el número de tubos con una duración igual o menor que 900 horas es $0,675 \times 500 = 337,5$. Redondeando, 338 tubos.

Cuadro 1.1: Principales diagramas según el tipo de variable.

Tipo de variable	Diagrama
V. Cualitativa	Barras, sectores, pictogramas
V. Discreta	Diferencial (barras) Integral (en escalera)
V. Continua	Diferencial (histograma, polígono de frecuencias) Integral (diagramas acumulados)

1.6. Problemas

Ejercicio 1.1. Clasificar las siguientes variables:

1. Preferencias políticas (izquierda, derecha o centro).
2. Marcas de cerveza.
3. Velocidad en Km/h.
4. El peso en Kg.
5. Signo del zodiaco.
6. Nivel educativo (primario secundario, superior).
7. Años de estudios completados.
8. Tipo de enseñanza (privada o pública).
9. Número de empleados de una empresa.
10. La temperatura de un enfermo en grados Celsius.
11. La clase social (baja, media o alta).
12. La presión de un neumático en Nw/cm^2

Ejercicio 1.2. Clasifique las variables que aparecen en el siguiente cuestionario.

1. ¿Cuál es su edad?
2. Estado civil:
 - a) Soltero
 - b) Casado
 - c) Separado
 - d) Divorciado
 - e) Viudo

3. ¿Cuanto tiempo emplea para desplazarse a su trabajo?
4. Tamaño de su municipio de residencia:
 - a) Municipio pequeño (menos de 2.000 habitantes)
 - b) Municipio mediano (de 2.000 a 10.000 hab.)
 - c) Municipio grande (de 10.000 a 50.000 hab.)
 - d) Ciudad pequeña (de 50.000 a 100.000 hab.)
 - e) Ciudad grande (más de 100.000 hab.)
5. ¿Está afiliado a la seguridad social?

Ejercicio 1.3.

En el siguiente conjunto de datos, se proporcionan los pesos (redondeados a libras) de niños nacidos en cierto intervalo de tiempo:

4, 8, 4, 6, 8, 6, 7, 7, 7, 8, 10, 9, 7, 6, 10, 8, 5, 9, 6, 3, 7, 6, 4, 7, 6, 9, 7, 4, 7, 6, 8, 8, 9, 11, 8, 7, 10, 8, 5, 7, 7, 6, 5, 10, 8, 9, 7, 5, 6, 5.

1. Construir una distribución de frecuencia de estos pesos.
2. Encontrar las frecuencias relativas.
3. Encontrar las frecuencias acumuladas.
4. Encontrar las frecuencias relativas acumuladas.
5. Dibujar un histograma con los datos del apartado a.
6. ¿Por qué se ha utilizado un histograma para representar estos datos, en lugar de una gráfica de barras?

Capítulo 2

Medidas descriptivas

2.1. Introducción

En el capítulo anterior hemos visto cómo se pueden resumir los datos obtenidos del estudio de una muestra (o una población) en una tabla estadística o un gráfico. No obstante, tras la elaboración de la tabla y su representación gráfica, en la mayoría de las ocasiones resulta más eficaz “condensar” dicha información en algunos números que la expresen de forma clara y concisa.

Los fenómenos biológicos no suelen ser constantes, por lo que será necesario que junto a una medida que indique el valor alrededor del cual se agrupan los datos, se asocie una medida que haga referencia a la variabilidad que refleje dicha fluctuación.

Por tanto el siguiente paso y objeto de este capítulo consistirá en definir algunos tipos de medidas (estadísticos o parámetros) que los sintetizan aún más.

Es decir, dado un grupo de datos organizados en una distribución de frecuencias (o bien una serie de observaciones sin ordenar), pretendemos describirlos mediante dos o tres cantidades sintéticas.

En este sentido pueden examinarse varias características, siendo las más comunes:

- La *tendencia central* de los datos;

- La *dispersión* o *variación* con respecto a este centro;
- Los datos que ocupan ciertas *posiciones*.
- La *simetría* de los datos.
- La *forma* en la que los datos se agrupan.

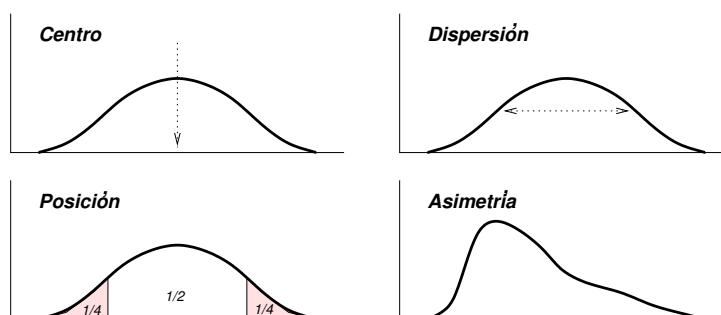


Figura 2.1: Medidas representativas de un conjunto de datos estadísticos

A lo largo de este capítulo, y siguiendo este orden, iremos estudiando los estadísticos que nos van a orientar sobre cada uno de estos niveles de información: valores alrededor de los cuales se agrupa la muestra, la mayor o menor fluctuación alrededor de esos valores, nos interesaremos en ciertos valores que marcan posiciones características de una distribución de frecuencias así como su simetría y su forma.

2.2. Estadísticos de tendencia central

Las tres medidas más usuales de tendencia central son:

- la *media*,
- la *mediana*,
- la *moda*.

En ciertas ocasiones estos tres estadísticos suelen coincidir, aunque generalmente no es así. Cada uno de ellos presenta ventajas e inconvenientes que precisaremos más adelante. En primer lugar vamos a definir los conceptos anteriores.

2.2.1. La media

La **media aritmética** de una variable estadística es la suma de todos sus posibles valores, ponderada por las frecuencias de los mismos. Es decir, si la tabla de valores de una variable X es

X	n_i	f_i
x_1	n_1	f_1
\dots	\dots	\dots
x_k	n_k	f_k

la media es el valor que podemos escribir de las siguientes formas equivalentes:

$$\begin{aligned}
 \bar{x} &= x_1 f_1 + \dots + x_k f_k \\
 &= \frac{1}{n} (x_1 n_1 + \dots + x_k n_k) \\
 &= \frac{1}{n} \sum_{i=1}^k x_i n_i
 \end{aligned}$$

Si los datos no están ordenados en una tabla, entonces

$$\boxed{\bar{x} = \frac{x_1 + \dots + x_n}{n}} \tag{2.1}$$

Algunos inconvenientes de la media

La media presenta inconvenientes en algunas situaciones:

- Uno de ellos es que es muy sensible a los valores extremos de la variable: ya que todas las observaciones intervienen en el cálculo de la media, la aparición de una observación extrema, hará que la media se desplace en esa dirección. En consecuencia,
- no es recomendable usar la media como medida central en las distribuciones muy asimétricas;
- Si consideramos una variable discreta, por ejemplo, *el número de hijos en las familias españolas* el valor de la media puede no pertenecer al conjunto de valores de la variable; Por ejemplo $\bar{x} = 1,2$ hijos.

Otras medias: Medias generalizadas

En función del tipo de problema varias generalizaciones de la media pueden ser consideradas. He aquí algunas de ellas aplicadas a unas observaciones x_1, \dots, x_n :

La media geométrica \bar{x}_g , es la media de los logaritmos de los valores de la variable:

$$\log \bar{x}_g = \frac{\log x_1 + \dots + \log x_n}{n}$$

Luego

$$\bar{x}_g = \sqrt[n]{x_1 x_2 \dots x_n}$$

Si los datos están agrupados en una tabla, entonces se tiene:

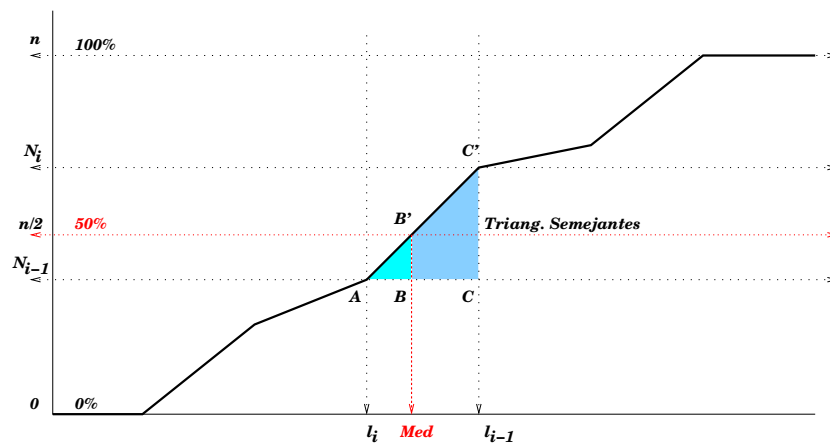
$$\bar{x}_g = \sqrt[n]{x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}}$$

La media armónica \bar{x}_a , se define como el recíproco de la media aritmética de los recíprocos, es decir,

$$\frac{1}{\bar{x}_a} = \frac{\frac{1}{x_1} + \dots + \frac{1}{x_n}}{n}$$

$$\overline{x}_a = \frac{n}{\frac{1}{x_1} + \dots + \frac{1}{x_n}}$$
$$\overline{x}_c = \sqrt{\frac{x_1^2 + \dots + x_n^2}{n}}$$

Consideramos una variable discreta X cuyas observaciones en una tabla estadística han sido ordenadas de menor a mayor. Llamaremos **mediana**, M_{ed} al primer valor de la variable que deja por debajo de sí al 50 % de las observaciones.



En el caso de variables continuas, las clases vienen dadas por intervalos, y aquí la fórmula de la mediana se complica un poco más (pero no demasiado): Sea $(l_{i-1}, l_i]$ el intervalo donde hemos encontrado que por debajo están

el 50 % de las observaciones. Entonces se obtiene la mediana a partir de las frecuencias absolutas acumuladas, mediante interpolación lineal (teorema de Thales) como sigue (figura 2.2):

$$\begin{aligned} \frac{CC'}{AC} = \frac{BB'}{AB} &\implies \frac{n_i}{a_i} = \frac{\frac{n}{2} - N_{i-1}}{M_{ed} - l_{i-1}} \\ &\implies \boxed{M_{ed} = l_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot a_i} \quad (2.2) \end{aligned}$$

Esto equivale a decir que *la mediana divide al histograma en dos partes de áreas iguales a $\frac{1}{2}$* .

Propiedades de la mediana

Entre las propiedades de la mediana, vamos a destacar las siguientes:

- Como medida descriptiva, tiene la ventaja de no estar afectada por las observaciones extremas, ya que no depende de los valores que toma la variable, sino del orden de las mismas. Por ello es adecuado su uso en distribuciones asimétricas.
- Es de cálculo rápido y de interpretación sencilla.
- A diferencia de la media, la mediana de una variable discreta es siempre un valor de la variable que estudiamos (ej. La mediana de una variable *número de hijos* toma siempre valores enteros).

Un ejemplo de cálculo de mediana

Sea X una variable discreta que ha presentado sobre una muestra las modalidades

$$X \rightsquigarrow 2, 5, 7, 9, 12 \implies \bar{x} = 7, \quad M_{ed} = 7$$

Si cambiamos la última observación por otra anormalmente grande, esto no afecta a la mediana, pero sí a la media:

$$X \rightsquigarrow 2, 5, 7, 9, 125 \implies \bar{x} = 29,6; \quad M_{ed} = 7$$

En este caso la media no es un posible valor de la variable (discreta), y se ha visto muy afectada por la observación extrema. Este no ha sido el caso para la mediana.

Un ejemplo de cálculo de media y mediana

Obtener la media aritmética y la mediana en la distribución adjunta. Determinar gráficamente cuál de los dos promedios es más significativo.

$l_{i-1} - l_i$	n_i
0 - 10	60
10 - 20	80
20 - 30	30
30 - 100	20
100 - 500	10

Solución:

$l_{i-1} - l_i$	n_i	a_i	x_i	$x_i n_i$	N_i	n_i'
0 - 10	60	10	5	300	60	60
10 - 20	80	10	15	1.200	140	80
20 - 30	30	10	25	750	170	30
30 - 100	20	70	65	1.300	190	2,9
100 - 500	10	400	300	3.000	200	0,25
$n = 200$		$\sum x_i n_i = 6,550$				

La media aritmética es:

$$\bar{x} = \frac{1}{n} \sum x_i = \frac{6,550}{200} = 32,75$$

La primera frecuencia absoluta acumulada que supera el valor $n/2 = 100$ es $N_i = 140$. Por ello el intervalo mediano es $[10; 20)$. Así:

$$M_{ed} = l_{i-1} + \frac{n/2 - N_{i-1}}{n_i} \cdot a_i = 10 + \frac{100 - 60}{80} \times 10 = 15$$

Para ver la representatividad de ambos promedios, realizamos el histograma de la figura 2.3, y observamos que dada la forma de la distribución, la mediana es más representativa que la media.

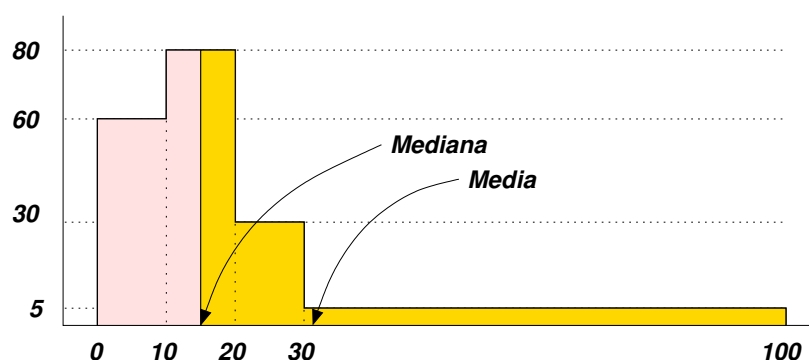


Figura 2.3: Para esta distribución de frecuencias es más representativo usar como estadístico de tendencia central la mediana que la media.

2.2.3. La moda

Llamaremos **moda** a cualquier máximo relativo de la distribución de frecuencias, es decir, cualquier valor de la variable que posea una frecuencia mayor que su anterior y su posterior.

Observación

De la moda destacamos las siguientes propiedades:

- Es muy fácil de calcular.
- Puede no ser única.

Cuadro 2.1: Resumen de las medidas de posición centrales.

MEDIDAS DE TENDENCIA CENTRAL																											
	<table> <tr> <th>DATOS SIN AGRUPAR</th> <th>DATOS AGRUPADOS</th> </tr> <tr> <td>(ordenados)</td> <td> <table> <tr> <th>Interv.</th> <th>x_i</th> <th>n_i</th> <th>N_i</th> </tr> <tr> <td>l_0-l_1</td> <td>x_1</td> <td>n_1</td> <td>N_1</td> </tr> <tr> <td>l_1-l_2</td> <td>x_2</td> <td>n_2</td> <td>N_2</td> </tr> <tr> <td>\dots</td> <td>\dots</td> <td>\dots</td> <td>\dots</td> </tr> <tr> <td>$l_{k-1}-l_k$</td> <td>x_k</td> <td>n_k</td> <td>N_k</td> </tr> </table> </td> </tr> <tr> <td>x_1, x_2, \dots, x_N</td> <td></td> </tr> </table>	DATOS SIN AGRUPAR	DATOS AGRUPADOS	(ordenados)	<table> <tr> <th>Interv.</th> <th>x_i</th> <th>n_i</th> <th>N_i</th> </tr> <tr> <td>l_0-l_1</td> <td>x_1</td> <td>n_1</td> <td>N_1</td> </tr> <tr> <td>l_1-l_2</td> <td>x_2</td> <td>n_2</td> <td>N_2</td> </tr> <tr> <td>\dots</td> <td>\dots</td> <td>\dots</td> <td>\dots</td> </tr> <tr> <td>$l_{k-1}-l_k$</td> <td>x_k</td> <td>n_k</td> <td>N_k</td> </tr> </table>	Interv.	x_i	n_i	N_i	l_0-l_1	x_1	n_1	N_1	l_1-l_2	x_2	n_2	N_2	\dots	\dots	\dots	\dots	$l_{k-1}-l_k$	x_k	n_k	N_k	x_1, x_2, \dots, x_N	
DATOS SIN AGRUPAR	DATOS AGRUPADOS																										
(ordenados)	<table> <tr> <th>Interv.</th> <th>x_i</th> <th>n_i</th> <th>N_i</th> </tr> <tr> <td>l_0-l_1</td> <td>x_1</td> <td>n_1</td> <td>N_1</td> </tr> <tr> <td>l_1-l_2</td> <td>x_2</td> <td>n_2</td> <td>N_2</td> </tr> <tr> <td>\dots</td> <td>\dots</td> <td>\dots</td> <td>\dots</td> </tr> <tr> <td>$l_{k-1}-l_k$</td> <td>x_k</td> <td>n_k</td> <td>N_k</td> </tr> </table>	Interv.	x_i	n_i	N_i	l_0-l_1	x_1	n_1	N_1	l_1-l_2	x_2	n_2	N_2	\dots	\dots	\dots	\dots	$l_{k-1}-l_k$	x_k	n_k	N_k						
Interv.	x_i	n_i	N_i																								
l_0-l_1	x_1	n_1	N_1																								
l_1-l_2	x_2	n_2	N_2																								
\dots	\dots	\dots	\dots																								
$l_{k-1}-l_k$	x_k	n_k	N_k																								
x_1, x_2, \dots, x_N																											
MEDIA	<table> <tr> <td>$\bar{x} = \frac{x_1 + \dots + x_n}{N}$</td> <td>$\bar{x} = \frac{n_1 x_1 + \dots + n_k x_k}{N}$</td> </tr> </table>	$\bar{x} = \frac{x_1 + \dots + x_n}{N}$	$\bar{x} = \frac{n_1 x_1 + \dots + n_k x_k}{N}$																								
$\bar{x} = \frac{x_1 + \dots + x_n}{N}$	$\bar{x} = \frac{n_1 x_1 + \dots + n_k x_k}{N}$																										
MEDIANA	<table> <tr> <td>Primera observación que deja debajo de sí estrictamente a las $[N/2]$ observaciones menores: $x_{[N/2]+1}$</td> <td>$M_{ed} = l_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \cdot a_i$</td> </tr> </table>	Primera observación que deja debajo de sí estrictamente a las $[N/2]$ observaciones menores: $x_{[N/2]+1}$	$M_{ed} = l_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \cdot a_i$																								
Primera observación que deja debajo de sí estrictamente a las $[N/2]$ observaciones menores: $x_{[N/2]+1}$	$M_{ed} = l_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \cdot a_i$																										
MODA	<table> <tr> <td>$M_{oda} = x_i$ de mayor frecuencia</td> <td>$M_{oda} = l_{i-1} + \frac{n'_i - n'_{i-1}}{(n'_i - n'_{i-1}) + (n'_i - n'_{i+1})} a_i$</td> </tr> </table>	$M_{oda} = x_i$ de mayor frecuencia	$M_{oda} = l_{i-1} + \frac{n'_i - n'_{i-1}}{(n'_i - n'_{i-1}) + (n'_i - n'_{i+1})} a_i$																								
$M_{oda} = x_i$ de mayor frecuencia	$M_{oda} = l_{i-1} + \frac{n'_i - n'_{i-1}}{(n'_i - n'_{i-1}) + (n'_i - n'_{i+1})} a_i$																										

2.2.4. Relación entre media, mediana y moda

En el caso de distribuciones unimodales, la mediana está con frecuencia comprendida entre la media y la moda (incluso más cerca de la media).

En distribuciones que presentan cierta inclinación, es más aconsejable el uso de la mediana. Sin embargo en estudios relacionados con propósitos estadísticos y de inferencia suele ser más apta la media.

2.3. Estadísticos de posición

Los estadísticos de posición van a ser valores de la variable caracterizados por superar a cierto porcentaje de observaciones en la población (o muestra). Tenemos fundamentalmente a los *percentiles* como medidas de posición, y asociados a ellos veremos también los *cuartiles*, *deciles* y *cuartiles*.

Percentiles

Para una variable discreta, se define el **percentil de orden k** , como la observación, P_k , que deja por debajo de si el $k\%$ de la población. Véase la figura 2.4. Esta definición nos recuerda a la mediana, pues como consecuencia de la definición es evidente que

$$M_{ed} = P_{50}$$

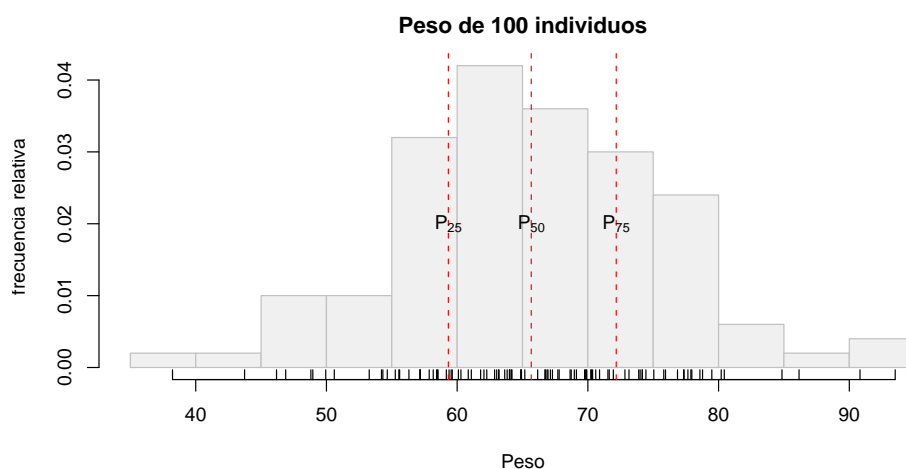


Figura 2.4: Percentiles 25, 50 y 75 de una variable. Los que se muestran dividen a la muestra en cuatro intervalos con similar número de individuos y reciben también el nombre de cuartiles.

En el caso de una variable continua, el intervalo donde se encuentra $P_k \in (l_{i-1}, l_i]$, se calcula buscando el que deja debajo de sí al $k\%$ de las observaciones. Dentro de él, P_k se obtiene según la relación:

$$P_k = l_{i-1} + \frac{n \frac{k}{100} - N_{i-1}}{n_i} \cdot a_i \quad (2.3)$$

Cuartiles

Los **cuartiles**, Q_l , son un caso particular de los percentiles. Hay 3, y se definen como:

$$Q_1 = P_{25} \quad (2.4)$$

$$Q_2 = P_{50} = M_{ed} \quad (2.5)$$

$$Q_3 = P_{75} \quad (2.6)$$

Deciles

Se definen los **deciles** como los valores de la variable que dividen a las observaciones en 10 grupos de igual tamaño. Más precisamente, definimos D_1, D_2, \dots, D_9 como:

$$D_i = P_{10i} \quad i = 1, \dots, 9$$

Ejemplo de cálculo de cuartiles con una variable discreta

Dada la siguiente distribución en el número de hijos de cien familias, calcular sus cuartiles.

x_i	n_i	N_i
0	14	14
1	10	24
2	15	39
3	26	65
4	20	85
5	15	100
n=100		

Solución:

1. Primer cuartil:

$$\frac{n}{4} = 25; \text{ Primera } N_i > n/4 = 39; \text{ luego } Q_1 = 2.$$

2. Segundo cuartil:

$$\frac{2n}{4} = 50; \text{ Primera } N_i > 2n/4 = 65; \text{ luego } Q_2 = 3.$$

3. Tercer cuartil:

$$\frac{3n}{4} = 75; \text{ Primera } N_i > 3n/4 = 85; \text{ luego } Q_3 = 4.$$

Ejemplo

Calcular los cuartiles en la siguiente distribución de una variable continua:

$l_{i-1} - l_i$	n_i	N_i
0 - 1	10	10
1 - 2	12	22
2 - 3	12	34
3 - 4	10	44
4 - 5	7	51
$n = 51$		

Solución:

1. Primer cuartil

$\frac{N}{4} = 12,75$; Primera $N_i > n/4 = 22$; La línea i es la del intervalo $[1; 2)$

$$Q_1 = l_{i-1} + \frac{\frac{n}{4} - N_{i-1}}{n_i} a_i = 1 + \frac{12,75 - 10}{12} \times 1 = 1,23$$

2. Segundo cuartil:

$\frac{2n}{4} = 25,5$; Primera $N_i > 2n/4 = 34$; La línea i es la del intervalo $[2; 3)$

$$Q_2 = l_{i-1} + \frac{\frac{2n}{4} - N_{i-1}}{n_i} a_i = 2 + \frac{25,5 - 22}{12} \times 1 = 2,29$$

3. Tercer cuartil

$\frac{3n}{4} = 38,25$; Primera $N_i > 3n/4 = 44$; La línea i es la del intervalo $[3; 4)$

$$Q_3 = l_{i-1} + \frac{\frac{3n}{4} - N_{i-1}}{n_i} a_i = 3 + \frac{38,25 - 34}{10} \times 1 = 3,445$$

Ejemplo de cálculo de cuartiles con una variable continua

Han sido ordenados los pesos de 21 personas en la siguiente tabla:

Intervalos	f.a.
$l_{i-1} - l_i$	n_i
38 — 45	3
45 — 52	2
52 — 59	7
59 — 66	3
66 — 73	6
	21

Encontrar aquellos valores que dividen a los datos en 4 partes con el mismo número de observaciones.

Solución: Las cantidades que buscamos son los tres cuartiles: Q_1 , Q_2 y Q_3 . Para calcularlos, le añadimos a la tabla las columnas con las frecuencias acumuladas, para localizar qué intervalos son los que contienen a los cuartiles buscados:

$l_{i-1} - l_i$	n_i	N_i	
38 — 45	3	3	Q_1 y Q_2 se encuentran en el intervalo
45 — 52	2	5	52—59, ya que $N_3 = 12$ es la primera
52 — 59	7	12	$\ni Q_1, Q_2$ f.a.a. que supera a $21 \cdot 1/4$ y $21 \cdot 2/4$.
59 — 66	3	15	Q_3 está en 66—73, pues $N_5 = 21$ es
66 — 73	6	21	el primer N_i mayor que $21 \cdot 3/4$.
	21		

Así se tiene que:

$$\begin{aligned} \frac{1}{4} \cdot 21 = 5,25 \Rightarrow i = 3 \Rightarrow Q_1 &= l_{i-1} + \frac{\frac{1}{4}n - N_{i-1}}{n_i} \cdot a_i \\ &= 52 + \frac{5,25 - 5}{7} \cdot 7 = 52,25 \end{aligned}$$

$$\begin{aligned} \frac{2}{4} \cdot 21 = 10,5 \Rightarrow i = 3 \Rightarrow Q_2 &= l_{i-1} + \frac{\frac{2}{4}n - N_{i-1}}{n_i} \cdot a_i \\ &= 52 + \frac{10,5 - 5}{7} \cdot 7 = 57,5 \end{aligned}$$

$$\begin{aligned}
\frac{3}{4} \cdot 21 = 15,75 \Rightarrow i = 5 \Rightarrow Q_3 &= l_{i-1} + \frac{\frac{3}{4}n - N_{i-1}}{n_i} \cdot a_i \\
&= 66 + \frac{15,75 - 15}{6} \cdot 7 = 66,875
\end{aligned}$$

Obsérvese que $Q_2 = M_{ed}$. Esto es lógico, ya que la mediana divide a la distribución en dos partes con el mismo número de observaciones, y Q_2 , hace lo mismo, pues es deja a dos cuartos de los datos por arriba y otros dos cuartos por abajo.

Ejemplo

La distribución de una variable tiene por polígono acumulativo de frecuencias el de la figura 2.5. Si el número total de observaciones es 50:

1. Elaborar una tabla estadística con los siguientes elementos: intervalos, marcas de clase, frecuencia absoluta, frecuencia absoluta acumulada, frecuencias relativa y frecuencias relativa acumulada.
2. Cuántas observaciones tuvieron un valor inferior a 10, cuántas inferior a 8 y cuántas fueron superior a 11.
3. Determine los cuartiles.

Solución:

1. En la siguiente tabla se proporciona la información pedida y algunos cálculos auxiliares que nos permitirán responder a otras cuestiones.

Intervalos	n_i	N_i	f_i	F_i	x_i	a_i	n_i'
0 – 5	10	10	0,2	0,3	2,5	5	2
5 – 7	25	35	0,5	0,7	6	2	12,5
7 – 12	5	40	0,1	0,8	9,5	5	1
12 – 15	10	50	0,2	1	13,5	7	3,33

2. Calculemos el número de observaciones pedido:

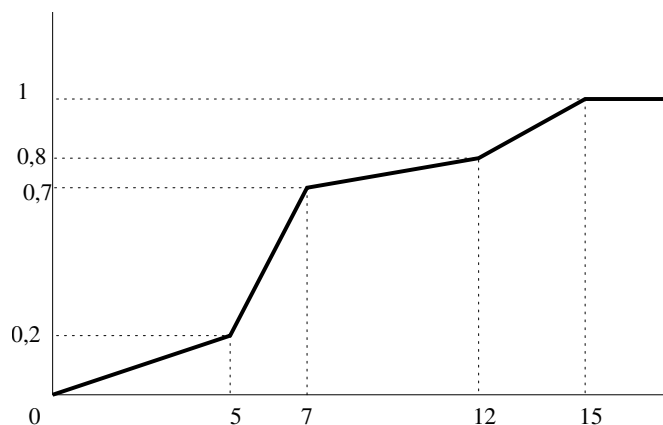


Figura 2.5: Diagrama acumulado de frecuencias relativas.

$$\begin{array}{l} 7 \text{ a } 12 \quad \text{——} \quad 5 \\ 7 \text{ a } 10 \quad \text{——} \quad x \end{array} \Leftrightarrow \begin{array}{l} 5 \quad \text{——} \quad 5 \\ 3 \quad \text{——} \quad x \end{array} \Rightarrow x = \frac{3 \times 5}{5} = 3$$

10 + 25 + 3 = 38 observaciones tomaron un valor inferior a 10

$$\begin{array}{l} 7 \text{ a } 12 \quad \text{——} \quad 5 \\ 7 \text{ a } 8 \quad \text{——} \quad x \end{array} \Leftrightarrow \begin{array}{l} 5 \quad \text{——} \quad 5 \\ 1 \quad \text{——} \quad x \end{array} \Rightarrow x = \frac{1 \times 5}{5} = 1$$

10 + 25 + 1 = 36 observaciones tomaron un valor inferior a 8

$$\begin{array}{l} 7 \text{ a } 12 \quad \text{——} \quad 5 \\ 7 \text{ a } 11 \quad \text{——} \quad x \end{array} \Leftrightarrow \begin{array}{l} 5 \quad \text{——} \quad 5 \\ 4 \quad \text{——} \quad x \end{array} \Rightarrow x = \frac{4 \times 5}{5} = 4$$

50 - (10 + 25 + 4) = 50 - 39 = 11 observaciones tomaron un valor superior a 11

3. Cuartiles:

$$Q_1 = l_{i-1} + \frac{n/4 - N_{i-1}}{n_i} \cdot a_i = 5 + \frac{12,5 - 10}{25} \cdot 2 = 5,2$$

$$Q_2 = l_{i-1} + \frac{2n/4 - N_{i-1}}{n_i} \cdot a_i = 5 + \frac{25 - 10}{25} \cdot 2 = 6,2$$

$$Q_3 = l_{i-1} + \frac{3n/4 - N_{i-1}}{n_i} \cdot a_i = 7 + \frac{37,5 - 35}{5} \cdot 5 = 9,5$$

2.4. Medidas de variabilidad o dispersión

Los estadísticos de *tendencia central* o *posición* nos indican donde se sitúa un grupo de puntuaciones. Los de *variabilidad* o *dispersión* nos indican si esas puntuaciones o valores están próximas entre sí o si por el contrario están o muy dispersas.

2.4.1. Rango

Una medida razonable de la variabilidad podría ser la **amplitud** o **rango**, que se obtiene restando el valor más bajo de un conjunto de observaciones del valor más alto.

Propiedades del rango

- Es fácil de calcular y sus unidades son las mismas que las de la variable.
- No utiliza todas las observaciones (sólo dos de ellas);
- Se puede ver muy afectada por alguna observación extrema;
- El rango aumenta con el número de observaciones, o bien se queda igual. En cualquier caso nunca disminuye.

2.4.2. Varianza

La **varianza**, S^2 , se define como la media de las diferencias cuadráticas de n puntuaciones con respecto a su media aritmética, es decir

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2.7)$$

Esta medida es siempre una cantidad positiva, con propiedades interesante para la realización de inferencia estadística. Como sus unidades son las del cuadrado de la variable, es más sencillo usar su raíz cuadrada, que es la que vemos en la siguiente sección.

2.4.3. Desviación típica o estándar

La varianza no tiene la misma magnitud que las observaciones (ej. si las observaciones se miden en metros, la varianza lo hace en metros cuadrados. Si queremos que la medida de dispersión sea de la misma dimensionalidad que las observaciones bastará con tomar su raíz cuadrada. Por ello se define la **desviación típica**, \mathcal{S} , como

$$\mathcal{S} = \sqrt{S^2}$$

2.4.4. Ejemplo de cálculo de medidas de dispersión

Calcular el rango, varianza y desviación típica de las siguientes cantidades medidas en metros:

$$3, 3, 4, 4, 5$$

Solución: El rango de esas observaciones es la diferencia entre la mayor y menor de ellas, es decir, $5 - 3 = 2$. Para calcular las restantes medidas de dispersión es necesario calcular previamente el valor con respecto al cual vamos a medir las diferencias. Éste es la media:

$$\bar{x} = (3 + 3 + 4 + 4 + 5)/5 = 3,8 \text{ metros}$$

La varianza es:

$$S^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{1}{5} (3^2 + 3^2 + 4^2 + 4^2 + 5^2) - 3,8^2 = 0,56 \text{ metros}^2$$

siendo la desviación típica su raíz cuadrada:

$$\mathcal{S} = \sqrt{S^2} = \sqrt{0,56} = 0,748 \text{ metros}$$

Propiedades de la varianza y desviación típica

- Ambas son sensibles a la variación de cada una de las puntuaciones, es decir, si una puntuación cambia, cambia con ella la varianza. La razón es que si miramos su definición, la varianza es función de *cada una de las puntuaciones*.
- La desviación típica tiene la propiedad de que en el intervalo

$$(\bar{x} - 2S, \bar{x} + 2S) \stackrel{\text{def}}{\sim} \bar{x} \pm 2S$$

se encuentra, al menos, el 75 % de las observaciones Incluso si tenemos muchos datos y estos provienen de una distribución normal (se definirá este concepto más adelante), podremos llegar al 95 %.

- No es recomendable el uso de ellas, cuando tampoco lo sea el de la media como medida de tendencia central.

2.4.5. Coeficiente de variación

Hemos visto que las medidas de centralización y dispersión nos dan información sobre una muestra. Nos podemos preguntar si tiene sentido usar estas magnitudes para comparar dos poblaciones. Por ejemplo, si nos piden comparar la dispersión de los pesos de las poblaciones de elefantes de dos circos diferentes, S nos dará información útil.

¿Pero qué ocurre si lo que comparamos es la altura de unos elefantes con respecto a su peso? Tanto la media como la desviación típica, \bar{x} y S , se expresan en las mismas unidades que la variable. Por ejemplo, en la variable altura podemos usar como unidad de longitud el metro y en la variable peso, el kilogramo. Comparar una desviación (con respecto a la media) medida en metros con otra en kilogramos no tiene ningún sentido.

El problema no deriva sólo de que una de las medidas sea de longitud y la otra sea de masa. El mismo problema se plantea si medimos cierta cantidad, por ejemplo la masa, de dos poblaciones, pero con distintas unidades. Este es el caso en que comparamos el peso en *toneladas* de una población de 100 elefantes con el correspondiente en *miligramos* de una población de 50 hormigas.

El problema no se resuelve tomando las mismas escalas para ambas poblaciones. Por ejemplo, se nos puede ocurrir medir a las hormigas con las mismas unidades que los elefantes (toneladas). Si la ingeriería genética no nos sorprende con alguna barbaridad, lo lógico es que la dispersión de la variable *peso de las hormigas* sea practicamente nula (¡Aunque haya algunas que sean 1.000 veces mayores que otras!)

En los dos primeros casos mencionados anteriormente, el problema viene de la *dimensionalidad* de las variables, y en el tercero de la diferencia enorme entre las medias de ambas poblaciones. El *coeficiente de variación* es lo que nos permite evitar estos problemas, pues elimina la dimensionalidad de las variables y tiene en cuenta la proporción existente entre medias y desviación típica. Se define del siguiente modo:

$$\mathcal{CV} = \frac{S_X}{\bar{x}} \quad (2.8)$$

Propiedades del coeficiente de variación

- Sólo se debe calcular para variables con todos los valores positivos. Todo índice de variabilidad es esencialmente no negativo. Las observaciones pueden ser positivas o nulas, pero su variabilidad debe ser siempre positiva. De ahí que sólo debemos trabajar con variables positivas, para la que tenemos con seguridad que $\bar{x} > 0$.
- No es invariante ante cambios de origen. Es decir, si a los resultados de una medida le sumamos una cantidad positiva, $b > 0$, para tener $Y = X + b$, entonces $\mathcal{CV}_Y < \mathcal{CV}_X$.
- Es invariante a cambios de escala. Así por ejemplo el coeficiente de variación de una variable medida en metros es una cantidad adimensional que no cambia si la medición se realiza en centímetros.

Tipificación

Se conoce por **tipificación** al proceso de restar la media y dividir por su desviación típica a una variable X . De este modo se obtiene una nueva

variable

$$Z = \frac{X - \bar{x}}{S} \quad (2.9)$$

de media $\bar{z} = 0$ y desviación típica $\mathcal{S}_Z = 1$, que denominamos **variable tipificada**.

Esta nueva variable carece de unidades y permite hacer comparables dos medidas que en un principio no lo son. Así por ejemplo nos podemos preguntar si un elefante es más grueso que una hormiga determinada, cada uno en relación a su población. También es aplicable al caso en que se quieran comparar individuos semejantes de poblaciones diferentes. Por ejemplo si deseamos comparar el nivel académico de dos estudiantes de diferentes Universidades para la concesión de una beca de estudios, en principio sería injusto concederla directamente al que posea una nota media más elevada, ya que la dificultad para conseguir una buena calificación puede ser mucho mayor en un centro que en el otro, lo que limita las posibilidades de uno de los estudiante y favorece al otro. En este caso, lo más correcto es comparar las calificaciones de ambos estudiantes, pero tipificadas cada una de ellas por las medias y desviaciones típicas respectivas de las notas de los alumnos de cada Universidad.

No confundir coeficiente de variación y tipificación

Los *coeficientes de variación* sirven para comparar las variabilidades de dos conjuntos de valores (muestras o poblaciones), mientras que si deseamos comparar a dos *individuos* de cada uno de esos conjuntos, es necesario usar los *valores tipificados*. Ninguno de ellos posee unidades y es un error frecuente entre estudiantes de bioestadística confundirlos.

2.5. Asimetría y apuntamiento

Sabemos cómo calcular valores alrededor de los cuales se distribuyen las observaciones de una variable sobre una muestra y sabemos cómo calcular la dispersión que ofrecen los mismos con respecto al valor de central. Nos

proponemos dar un paso más allá en el análisis de la variable. En primer lugar, nos vamos a plantear el saber si los datos se distribuyen de forma simétrica con respecto a un valor central, o si bien la gráfica que representa la distribución de frecuencias es *de una forma diferente del lado derecho que del lado izquierdo*.

Si la simetría ha sido determinada, podemos preguntarnos si la curva es más o menos *apuntada* (larga y estrecha). Este apuntamiento habrá que medirlo comparado a cierta distribución de frecuencias que consideramos *normal* (no por casualidad es éste el nombre que recibe la distribución de referencia).

Estas ideas son las que vamos a desarrollar en lo que resta del capítulo.

2.5.1. Estadísticos de asimetría

Para saber si una distribución de frecuencias es simétrica, hay que precisar con respecto a qué. Un buen candidato es la mediana, ya que para variables continuas, divide al histograma de frecuencias en dos partes de igual área. Podemos basarnos en ella para, de forma natural, decir que **una distribución de frecuencias es simétrica** si el lado derecho de la gráfica (a partir de la mediana) es la imagen por un espejo del lado izquierdo (figura 2.6).

Cuando la variable es discreta, decimos que es simétrica, si lo es con respecto a la media.

Dentro de los tipos de asimetría posible, vamos a destacar los dos fundamentales:

Asimetría positiva: Si las frecuencias más altas se encuentran en el lado izquierdo de la media, mientras que en derecho hay frecuencias más pequeñas (*cola*).

Asimetría negativa: Cuando la cola está en el lado izquierdo.

Cuando realizamos un estudio descriptivo es altamente improbable que la distribución de frecuencias sea totalmente simétrica. En la práctica diremos que la distribución de frecuencias es simétrica si lo es de un modo

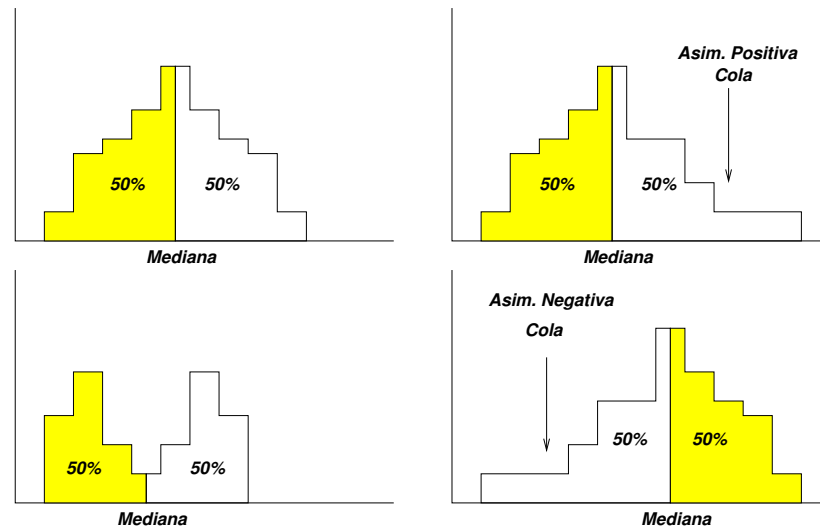


Figura 2.6: Distribuciones de frecuencias simétricas y asimétricas

aproximado. Por otro lado, aún observando cuidadosamente la gráfica, podemos no ver claro de qué lado están las frecuencias más altas. Se definen entonces toda una familia de estadísticos que ayuden a interpretar la asimetría, denominados **índices de asimetría**. El principal de ellos es el *momento central de tercer orden* que definimos a continuación.

Momento central de tercer orden

Sea X una variable cuantitativa y $p \in \mathbb{N}$. Llamamos **momento** de orden p a:

$$\mu_p = \frac{1}{n} \sum_{i=1}^n x_i^p \quad (2.10)$$

Se denomina **momento central** de orden p a la cantidad

$$m_p = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^p \quad (2.11)$$

Los momentos de orden p impar, son siempre nulos en el caso de variables simétricas, ya que para cada i que esté a un lado de la media, con $(x_i - \bar{x}) < 0$, le corresponde una observación j del otro lado de la media tal que $(x_j - \bar{x}) = -(x_i - \bar{x})$. Elevando cada una de esas cantidades a p impar, y sumando se tiene que

$$m_p = 0 \quad \text{si la distribución es simétrica.}$$

Si la distribución fuese asimétrica positiva, las cantidades $(x_i - \bar{x})^p$, con $p \geq 3$ impar positivas estarían muy aumentadas al elevarse a p . Esta propiedad nos indica que un índice de asimetría posible consiste en tomar $p = 3$ y elegir como estadístico de asimetría al momento central de tercer orden.

Apoyandonos en este índice, diremos que hay asimetría positiva si $a_3 > 0$, y que la asimetría es negativa si $a_3 < 0$.

Índice basado en los tres cuartiles (Yule–Bowley)

Si una distribución es simétrica, es claro que deben haber tantas observaciones entre la que deja por debajo de sí las tres cuartas partes de la distribución y la mediana, como entre la mediana y la que deja por debajo de sí un cuarto de todas las observaciones. De forma abreviada esto es,

$$Q_3 - Q_2 = Q_2 - Q_1$$

Una pista para saber si una distribución de frecuencias es asimétrica positiva la descubrimos observando la figura 2.7):

$$Q_3 - Q_2 > Q_2 - Q_1$$

Por analogía, si es asimétrica negativa, se tendrá

$$Q_3 - Q_2 < Q_2 - Q_1$$

Para quitar dimensionalidad al problema, utilizamos como *índice de asimetría* la cantidad:

$$\mathcal{A}_s = \frac{(\mathcal{Q}_3 - \mathcal{Q}_2) - (\mathcal{Q}_2 - \mathcal{Q}_1)}{\mathcal{Q}_3 - \mathcal{Q}_1} \quad (2.12)$$

Es claro que

$$-1 \leq \mathcal{A}_s = \frac{(\mathcal{Q}_3 - \mathcal{Q}_2) - (\mathcal{Q}_2 - \mathcal{Q}_1)}{(\mathcal{Q}_3 - \mathcal{Q}_2) + (\mathcal{Q}_2 - \mathcal{Q}_1)} \leq 1 \quad (2.13)$$

El número obtenido, \mathcal{A}_s , es invariante ante cambios de origen de referencia y de escala.

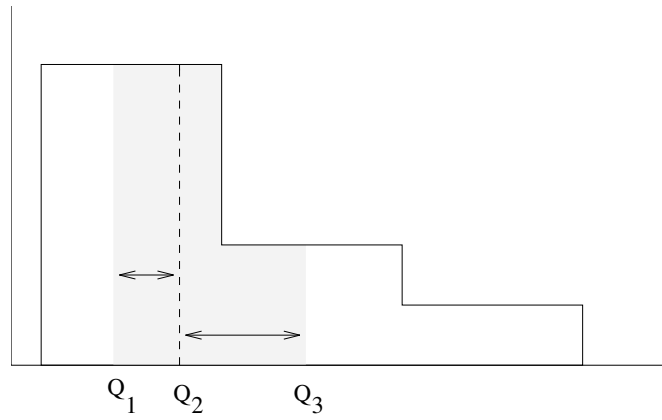


Figura 2.7: Uso de los cuartiles para medir la asimetría

Otros índices de asimetría

Basándonos en que si una distribución de frecuencias es simétrica y unimodal, entonces la media, la mediana y la moda coinciden, podemos definir otras medidas de asimetría, como son:

$$\mathcal{A}_s = \frac{\bar{x} - M_{oda}}{\mathcal{S}} \quad (2.14)$$

o bien,

$$\mathcal{A}_s = \frac{3(\bar{x} - M_{ed})}{\mathcal{S}} \quad (2.15)$$

Diremos que hay asimetría positiva si $\mathcal{A}_s > 0$ y negativa si $\mathcal{A}_s < 0$

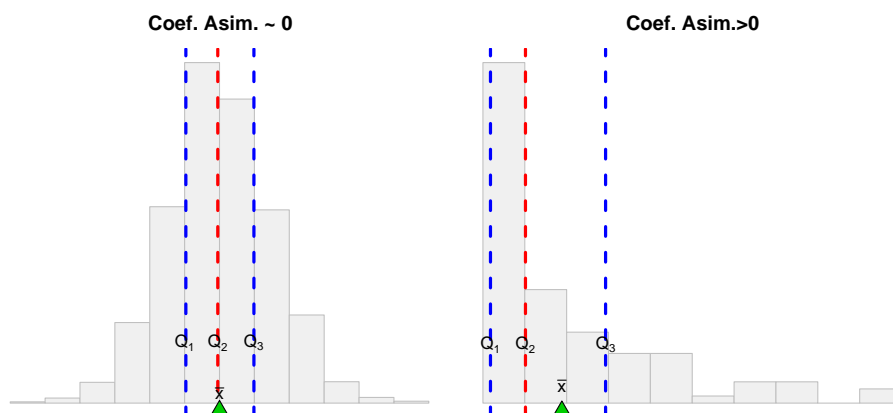


Figura 2.8: Diferencias entre las medidas de tendencia central, o bien entre las distancias entre cuartiles consecutivos indican asimetría.

Ejemplo

Las edades de un grupo de personas se reflejan en la tabla siguiente:

Intervalos	n_i
7 — 9	4
9 — 11	18
11 — 12	14
12 — 13	27
13 — 14	42
14 — 15	31
15 — 17	20
17 — 19	1

Determinar la variabilidad de la edad mediante los estadísticos varianza, desviación típica, coeficiente de variación y rango intercuartílico. Estudie la simetría de la variable.

Solución:

En primer lugar realizamos los cálculos necesarios a partir de la tabla de frecuencias:

Intervalos	n_i	x_i	N_i	$x_i n_i$	$x_i^2 n_i$
7 — 9	4	8	4	32	256
9 — 11	18	10	22	180	1.800
11 — 12	14	11,5	36	161	1.851,5
12 — 13	27	12,5	63	337,5	4.218,75
13 — 14	42	13,5	105	567	7.654,5
14 — 15	31	14,5	136	449,5	6.517,75
15 — 17	20	16	156	320	5.120
17 — 19	1	18	157	18	324
	157			2.065	27.742,25

La media es $\bar{x} = 2,065/157 = 13,15$ años. La varianza la calculamos a partir de la columna de la $x_i^2 n_i$ como sigue:

$$S^2 = 27,742,25/157 - 13,15^2 = 3,78 \text{ años}^2 \Rightarrow S = \sqrt{3,78} = 1,94 \text{ años}$$

El coeficiente de variación no posee unidades y es:

$$CV = \frac{1,94}{13,15} = 0,15 = 15 \% \text{ de variabilidad.}$$

En lo que concierne a la simetría podemos utilizar el coeficiente de asimetría de Yule–Bowley, para el cual es preciso el cálculo de los cuartiles:

$$Q_1 = 12 + \frac{39,25 - 36}{27} \times 1 = 12,12$$

$$M_{ed} = Q_2 = 13 + \frac{78,5 - 63}{42} \times 1 = 13,37$$

$$Q_3 = 14 + \frac{117,75 - 105}{31} \times 1 = 14,41$$

Lo que nos dice que aproximadamente en un rango de $Q_3 - Q_1 = 2,29$ años se encuentra el 50 % central del total de observaciones¹ Además:

$$= \mathcal{A}_s = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{(14,41 - 13,37) - (13,37 - 12,12)}{14,41 - 12,12} = -0,09$$

Este resultado nos indica que existe una ligera asimetría a la izquierda (negativa). Un resultado similar se obtiene si observamos (Figura 2.9) que la distribución de frecuencias es unimodal, siendo la moda:

$$M_{oda} = 13 + \frac{42 - 27}{(42 - 27) + (42 - 31)} \times 1 = 13,57$$

en cuyo caso podemos usar como medida del sesgo:

$$\mathcal{A}_s = \frac{\bar{x} - M_{oda}}{S} = \frac{13,15 - 13,57}{1,94} = -0,21$$

2.5.2. Estadísticos de apuntamiento

Se define el **coeficiente de aplastamiento de Fisher (curtosis)** como:

$$\gamma_2 = \frac{m_4}{\sigma^4} - 3$$

donde m_4 es el momento empírico de cuarto orden. Es éste un coeficiente adimensional, invariante ante cambios de escala y de origen. Sirve para medir si una distribución de frecuencias es muy apuntada o no. Para decir si la distribución es larga y estrecha, hay que tener un patrón de referencia. El patrón de referencia es la *distribución normal o gaussiana*² para la que se tiene

¹Eso hace que dicha cantidad sea usada como medida de dispersión, denominándose **rango intercuartílico**.

²Será introducida posteriormente.

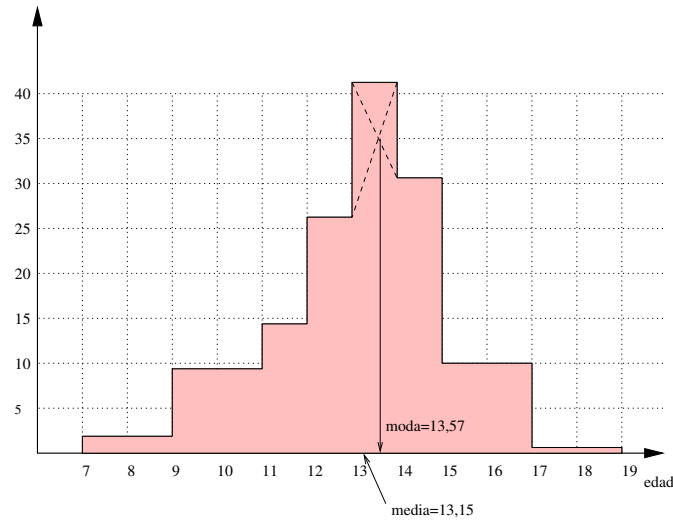


Figura 2.9: La distribución de frecuencias de la edad presenta una ligera asimetría negativa.

$$\frac{m_4}{\sigma^4} = 3 \implies \gamma_2 = 0$$

De este modo, atendiendo a γ_2 , se clasifican las distribuciones de frecuencias en

Leptocúrtica: Cuando $\gamma_2 > 0$, o sea, si la distribución de frecuencias es más apuntada que la normal;

Mesocúrtica: Cuando $\gamma_2 = 0$, es decir, cuando la distribución de frecuencias es tan apuntada como la normal;

Platicúrtica: Cuando $\gamma_2 < 0$, o sea, si la distribución de frecuencias es menos apuntada que la normal;

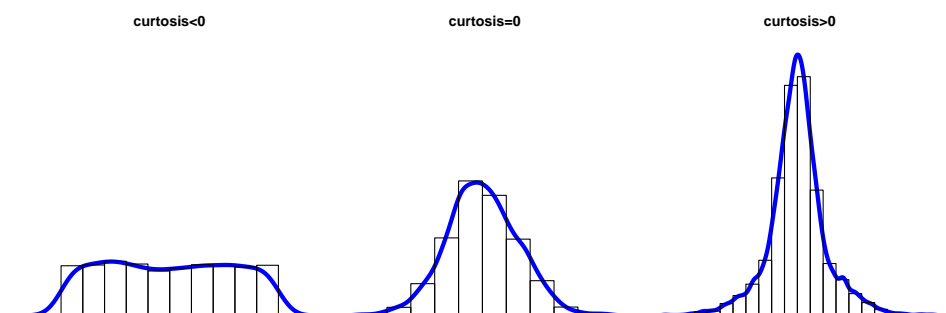


Figura 2.10: Apuntamiento de distribuciones de frecuencias

2.6. Problemas

Ejercicio 2.1. En el siguiente conjunto de números, se proporcionan los pesos (redondeados a la libra más próxima) de los bebés nacidos durante un cierto intervalo de tiempo en un hospital:

4, 8, 4, 6, 8, 6, 7, 7, 7, 8, 10, 9, 7, 6, 10, 8, 5, 9, 6, 3, 7, 6, 4, 7, 6, 9, 7, 4, 7, 6, 8, 8, 9, 11, 8, 7, 10, 8, 5, 7, 7, 6, 5, 10, 8, 9, 7, 5, 6, 5.

1. Construir una distribución de frecuencias de estos pesos.
2. Encontrar las frecuencias relativas.
3. Encontrar las frecuencias acumuladas.
4. Encontrar las frecuencias relativas acumuladas.
5. Dibujar un histograma con los datos de la parte **a**.
6. ¿Por qué se ha utilizado un histograma para representar estos datos, en lugar de una gráfica de barras?
7. Calcular las medidas de tendencia central.

8. Calcular las medidas de dispersión.
9. Calcular las medidas de forma.
10. ¿Es esta una distribución sesgada? De ser así, ¿en qué dirección?
11. Encontrar el percentil 24.

Ejercicio 2.2. A continuación se dan los resultados obtenidos con una muestra de 50 universitarios. la característica es el tiempo de reacción ante un estímulo auditivo:

0,110	0,110	0,126	0,112	0,117	0,113	0,135	0,107	0,122
0,113	0,098	0,122	0,105	0,103	0,119	0,100	0,117	0,113
0,124	0,118	0,132	0,108	0,115	0,120	0,107	0,123	0,109
0,117	0,111	0,112	0,101	0,112	0,111	0,119	0,103	0,100
0,108	0,120	0,099	0,102	0,129	0,115	0,121	0,130	0,134
0,118	0,106	0,128	0,094	0,1114				

1. ¿Cuál es la amplitud total de la distribución de los datos?
2. Obtenga la distribución de frecuencias absolutas y relativas.
3. Obtenga la distribución de frecuencias acumuladas, absolutas y relativas, con los intervalos anteriores.
4. Calcular la media y la varianza con los intervalos del apartado **b** y después calculense las mismas magnitudes sin ordenar los datos en una tabla estadística. ¿Con qué método se obtiene mayor precisión?
5. Dibuje el polígono de frecuencias relativas.
6. Dibuje el polígono de frecuencias relativas acumuladas.

Ejercicio 2.3. Con el fin de observar la relación entre la inteligencia y el nivel socioeconómico (medido por el salario mensual familiar) se tomaron dos grupos, uno formado con sujetos de cociente intelectual inferior a 95

y otro formado por los demás; De cada sujeto se anotó el salario mensual familiar. Teniendo en cuenta los resultados que se indican en la tabla:

Nivel socioeconómico	Sujetos con $CI < 95$	Sujetos con $CI \geq 95$
Intervalos	Frecuencia	Frecuencia
10 o menos $\equiv (4,10]$	75	19
10 – 16	35	26
16 – 22	20	25
22 – 28	30	30
28 – 34	25	54
más de 34 $\equiv (34,40]$	15	46

1. Dibuje un gráfico que permita comparar ambos grupos.
2. Calcule las medidas de tendencia central para aquellos sujetos con $CI < 95$.
3. Calcular las medidas de dispersión para aquellos sujetos con $CI \geq 95$.

Ejercicio 2.4. Un estudio consistió en anotar el número de palabras leídas en 15 segundos por un grupo de 120 sujetos disléxicos y 120 individuos normales. Teniendo en cuenta los resultados de la tabla

N° de palabras leídas	Disléxicos n_D	Normales n_N
25 o menos $\equiv 25$	56	1
26	24	9
27	16	21
28	12	29
29	10	28
30 o más $\equiv 30$	2	32

calcule:

1. Las medias aritméticas de ambos grupos.
2. Las medianas de ambos grupos.

3. El porcentaje de sujetos disléxicos que superaron la mediana de los normales.
4. Compare la variabilidad relativa de ambos grupos.

Ejercicio 2.5. La tabla siguiente muestra la composición por edad, sexo y trabajo de un grupo de personas con tuberculosis pulmonar en la provincia de Vizcaya en el año 1979:

Edad	Trabajadores			No trabajadores			Totales		
	Varón	Mujer	Total	Varón	Mujer	Total	Varón	Mujer	Total
14–19	2	1	3	25	40	65	27	41	68
19–24	10	4	14	20	36	56	30	40	70
24–29	32	10	42	15	50	65	47	60	107
29–34	47	12	59	13	34	47	60	46	106
34–39	38	8	46	10	25	35	48	33	81
39–44	22	4	26	7	18	25	29	22	51

1. Representar gráficamente la distribución de frecuencias de aquellas personas trabajadoras que padecen tuberculosis.
2. Representar gráficamente la distribución de frecuencias de los varones no trabajadores que padecen tuberculosis.
3. Representar gráficamente la distribución de frecuencias del número total de mujeres que padecen tuberculosis.
4. ¿Cuál es la edad en la que se observa con mayor frecuencia que no trabajan los varones? ¿Y las mujeres? Determinar asimismo la edad más frecuente (sin distinción de sexos ni ocupación).
5. ¿Por debajo de qué edad está el 50 % de los varones?
6. ¿Por encima de qué edad se encuentra el 80 % de las mujeres?
7. Obtener la media, mediana y desviación típica de la distribución de las edades de la muestra total.
8. Estudiar la asimetría de las tres distribuciones.

Ejercicio 2.6. En una epidemia de escarlatina, se ha recogido el número de muertos en 40 ciudades de un país, obteniéndose la siguiente tabla:

N° de muertos	0	1	2	3	4	5	6	7
Ciudades	7	11	10	7	1	2	1	1

1. Representar gráficamente estos datos.
2. Obtener la distribución acumulada y representarla.
3. Calcular media, mediana y moda.
4. Calcular la varianza y la desviación típica.
5. Porcentaje de ciudades con al menos 2 muertos.
6. Porcentaje de ciudades con más de 3 muertos.
7. Porcentaje de ciudades con a lo sumo 5 muertos.

Capítulo 3

Variables bidimensionales

3.1. introducción

En lo estudiado anteriormente hemos podido aprender cómo a partir de la gran cantidad de datos que describen una muestra mediante una variable, X , se representan gráficamente los mismos de modo que resulta más intuitivo hacerse una idea de como se distribuyen las observaciones.

Otros conceptos que según hemos visto, también nos ayudan en el análisis, son los estadísticos de tendencia central, que nos indican hacia donde tienden a agruparse los datos (en el caso en que lo hagan), y los estadísticos de dispersión, que nos indican si las diferentes modalidades que presenta la variable están muy agrupadas alrededor de cierto valor central, o si por el contrario las variaciones que presentan las modalidades con respecto al valor central son grandes.

También sabemos determinar ya si los datos se distribuyen de forma simétrica a un lado y a otro de un valor central.

En este capítulo pretendemos estudiar una situación muy usual y por tanto de gran interés en la práctica:

Si Y es otra variable definida sobre la misma población que X , ¿será posible determinar si existe alguna relación entre las modalidades de X y de Y ?

Un ejemplo trivial consiste en considerar una población formada por alumnos de primero de Medicina y definir sobre ella las variables

$$\begin{aligned} X &\equiv \text{altura medida en centímetros,} \\ Y &\equiv \text{altura medida en metros,} \end{aligned}$$

ya que la relación es determinista y clara: $Y = X/100$. Obsérvese que aunque la variable Y , como tal puede tener cierta dispersión, vista *como función* de X , su dispersión es nula.

Un ejemplo más parecido a lo que nos interesa realmente lo tenemos cuando sobre la misma población definimos las variables

$$\begin{aligned} X &\equiv \text{altura medida en centímetros,} \\ Y &\equiv \text{peso medida en kilogramos.} \end{aligned}$$

Intuitivamente esperamos que exista cierta relación entre ambas variables, por ejemplo,

$$Y = X - 110 \pm \textbf{dispersión}$$

que nos expresa que (en media) a mayor altura se espera mayor peso. La relación no es exacta y por ello será necesario introducir algún término que exprese la dispersión de Y con respecto a la variable X .

Es fundamental de cara a realizar un trabajo de investigación experimental, conocer muy bien las técnicas de estudio de variables bidimensionales (y n -dimensionales en general). Baste para ello pensar que normalmente las relaciones entre las variables no son tan evidentes como se mencionó arriba. Por ejemplo:

¿Se puede decir que en un grupo de personas existe alguna relación entre $X = \text{tensión arterial}$ e $Y = \text{edad}$?

Aunque en un principio la notación pueda resultar a veces algo desagradable, el lector podrá comprobar, al final del capítulo, que es bastante

accesible. Por ello le pedimos que no se asuste. Al final verá que no son para tanto.

3.2. Tablas de doble entrada

Consideramos una población de n individuos, donde cada uno de ellos presenta dos caracteres que representamos mediante las variables X e Y . Representamos mediante

$$X \rightsquigarrow x_1, x_2, \dots, x_i, \dots, x_k$$

las k modalidades que presenta la variable X , y mediante

$$Y \rightsquigarrow y_1, y_2, \dots, y_j, \dots, y_p$$

las p modalidades de Y .

Con la intención de reunir en una sólo estructura toda la información disponible, creamos una tabla formada por $k \cdot p$ casillas, organizadas de forma que se tengan k filas y p columnas. La casilla denotada de forma general mediante el subíndice $_{ij}$ hará referencia a los elementos de la muestra que presentan simultáneamente las modalidades x_i e y_j .

Y X	y_1	y_2	\dots	y_j	\dots	y_p	
x_1	n_{11}	n_{12}	\dots	n_{1j}	\dots	n_{1p}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	\dots	n_{2j}	\dots	n_{2p}	$n_{2\bullet}$
\dots	\dots	\dots	\dots	\dots	\dots	\dots	\dots
x_i	n_{i1}	n_{i2}	\dots	n_{ij}	\dots	n_{ip}	$n_{i\bullet}$
\dots	\dots	\dots	\dots	\dots	\dots	\dots	\dots
x_k	n_{k1}	n_{k2}	\dots	n_{kj}	\dots	n_{kp}	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	\dots	$n_{\bullet j}$	\dots	$n_{\bullet p}$	$n_{\bullet \bullet}$

De este modo, para $i = 1, \dots, k$, $j = 1, \dots, p$, se tiene que n_{ij} es el número de individuos o **frecuencia absoluta**, que presentan a la vez las modalidades x_i e y_j .

El número de individuos que presentan la modalidad x_i , es lo que llamamos **frecuencia absoluta marginal** de x_i y se representa como $n_{i\bullet}$. Es evidente la igualdad

$$n_{i\bullet} = n_{i1} + n_{i2} + \cdots + n_{ip} = \sum_{j=1}^p n_{ij}$$

Obsérvese que hemos escrito un símbolo “ \bullet ” en la “*parte de las jotas*” que simboliza que estamos considerando los elementos que presentan la modalidad x_i , independientemente de las modalidades que presente la variable Y . De forma análoga se define la frecuencia absoluta marginal de la modalidad y_j como

$$n_{\bullet j} = n_{1j} + n_{2j} + \cdots + n_{kj} = \sum_{i=1}^k n_{ij}$$

Estas dos distribuciones de frecuencias $n_{i\bullet}$ para $i = 1, \dots, k$, y $n_{\bullet j}$ para $j = 1, \dots, p$ reciben el nombre de **distribuciones marginales** de X e Y respectivamente.

El número total de elementos de la población (o de la muestra), n lo obtenemos de cualquiera de las siguientes formas, que son equivalentes:

$$n = n_{\bullet\bullet} = \sum_{i=1}^k n_{i\bullet} = \sum_{j=1}^p n_{\bullet j} = \sum_{i=1}^k \sum_{j=1}^p n_{ij}$$

3.2.1. Distribuciones condicionadas

De todos los elementos de la población, n , podemos estar interesados, en un momento dado, en un conjunto más pequeño y que está formado por aquellos elementos que han presentado la modalidad y_j , para algún $j = 1, \dots, p$. El número de elementos de este conjunto sabemos que es $n_{\bullet j}$. La variable X definida sobre este conjunto se denomina **variable condicionada** y se suele denotar mediante $X_{|y_j}$ o bien $X_{|Y=y_j}$. La distribución de frecuencias absolutas de esta nueva variable es exactamente la columna j de la tabla.

De la misma forma, es posible dividir la población inicial en k subconjuntos, cada uno de ellos caracterizados por la propiedad de que el i -ésimo conjunto todos los elementos verifican la propiedad de presentar la modalidad x_i . Sobre cada uno de estos conjuntos tenemos la variable condicionada $Y_{|x_i} \equiv Y_{|X=x_i}$, cuya distribución de frecuencias relativas condicionadas es:

$$f_j^i = \frac{n_{ij}}{n_{i\bullet}} \quad \forall j = 1, \dots, p$$

3.3. Dependencia funcional e independencia

La relación entre las variables X e Y , parte del objetivo de este capítulo y en general de un número importante de los estudios de las Ciencias Sociales, puede ser más o menos acentuada, pudiendo llegar ésta desde la dependencia total o *dependencia funcional* hasta la *independencia*.

3.3.1. Dependencia funcional

La dependencia funcional, que nos refleja cualquier fórmula matemática o física, es a la que estamos normalmente más habituados. Al principio del capítulo consideramos un ejemplo en el que sobre una población de alumnos definíamos las variables

$$\begin{aligned} X &\equiv \text{altura medida en centímetros,} \\ Y &\equiv \text{altura medida en metros,} \end{aligned}$$

Al tomar a uno de los alumnos, hasta que no se realice una medida sobre el mismo, no tendremos claro cual será su altura. Podemos tener cierta intuición sobre qué valor es más probable que tome (alrededor de la media, con cierta dispersión). Sin embargo, si la medida X ha sido realizada, no es necesario practicar la de Y , pues la relación entre ambas es exacta (dependencia funcional):

$$Y = X/100$$

3.3.2. Independencia

Existe un concepto que es radicalmente opuesto a la dependencia funcional, que es el de *independencia*. Se dice que dos variables X e Y son **independientes** si la distribución marginal de una de ellas es la misma que la condicionada por cualquier valor de la otra.

Esta es una de entre muchas maneras de expresar el concepto de independencia, y va a implicar una estructura muy particular de la tabla bidimensional, en el que todas las filas y todas las columnas van a ser proporcionales entre sí.

3.4. Covarianza

La **covarianza** \mathcal{S}_{XY} , es una medida que nos hablará de la variabilidad conjunta de dos variables numéricas (cuantitativas). Se define como:

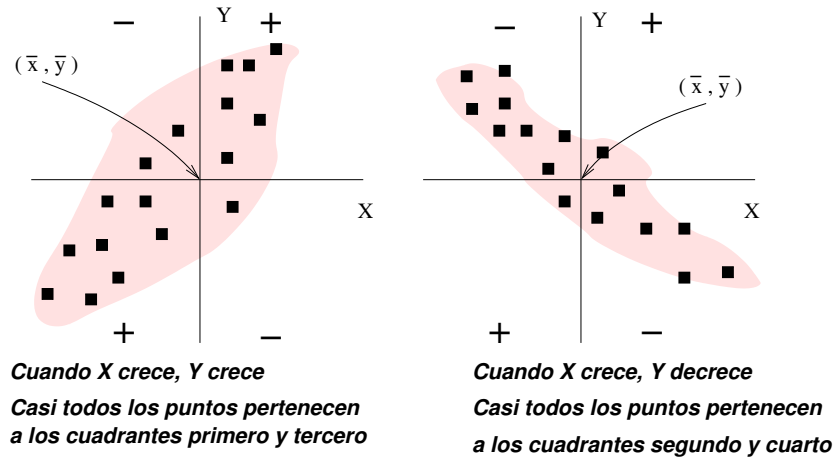
$$\mathcal{S}_{XY} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Una interpretación geométrica de la covarianza

Consideremos la *nube de puntos* formadas por las n parejas de datos (x_i, y_i) . El centro de gravedad de esta nube de puntos es (\bar{x}, \bar{y}) , o bien podemos escribir simplemente (\bar{x}, \bar{y}) si los datos no están ordenados en una tabla de doble entrada. Trasladamos los ejes XY al nuevo centro de coordenadas (\bar{x}, \bar{y}) . Queda así dividida la nube de puntos en cuatro cuadrantes como se observa en la figura 3.1. Los puntos que se encuentran en el primer y tercer cuadrante contribuyen positivamente al valor de \mathcal{S}_{XY} , y los que se encuentran en el segundo y el cuarto lo hacen negativamente.

De este modo:

- Si hay mayoría de puntos en el tercer y primer cuadrante, ocurrirá que $\mathcal{S}_{XY} \geq 0$, lo que se puede interpretar como que la variable Y tiende a aumentar cuando lo hace X ;

Figura 3.1: Interpretación geométrica de \mathcal{S}_{XY}

- Si la mayoría de puntos están repartidos entre el segundo y cuarto cuadrante entonces $\mathcal{S}_{XY} \leq 0$, es decir, las observaciones Y tienen tendencia a disminuir cuando las de X aumentan;
- Si los puntos se reparten con igual intensidad alrededor de (\bar{x}, \bar{y}) , entonces se tendrá que $\mathcal{S}_{XY} = 0$. Véase la figura 3.2 como ilustración.

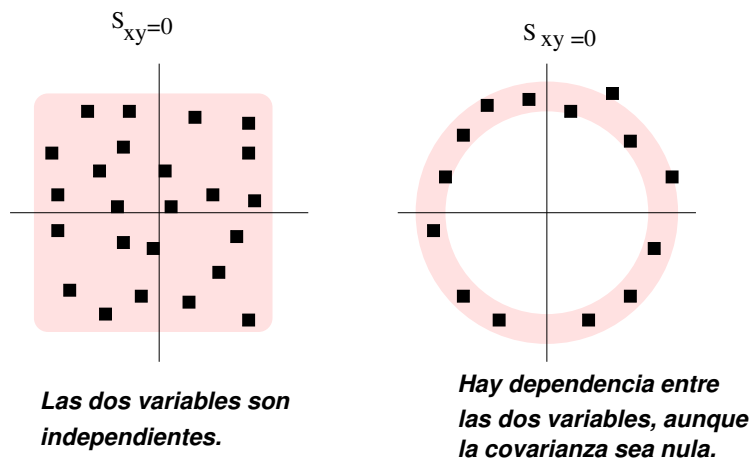


Figura 3.2: Cuando los puntos se reparte de modo más o menos homogéneo entre los cuadrantes primero y tercero, y segundo y cuarto, se tiene que $S_{XY} \approx 0$. Eso no quiere decir de ningún modo que no pueda existir ninguna relación entre las dos variables, ya que ésta puede existir como se aprecia en la figura de la derecha.

LA COVARIANZA

- Si $S_{XY} > 0$ las dos variables crecen o decrecen a la vez (nube de puntos creciente).
- Si $S_{XY} < 0$ cuando una variable crece, la otra tiene tendencia a decrecer (nube de puntos decreciente).
- Si los puntos se reparten con igual intensidad alrededor de (\bar{x}, \bar{y}) , $S_{XY} = 0$ (no hay relación lineal).

3.5. Coeficiente de correlación lineal de Pearson

La covarianza es una medida de la variabilidad común de dos variables (crecimiento de ambas al tiempo o crecimiento de una y decrecimiento de la otra), pero está afectada por las unidades en las que cada variable se mide. Así pues, es necesario definir una medida de la relación entre dos variables, y que no esté afectada por los cambios de unidad de medida. Una forma de conseguir este objetivo es dividir la covarianza por el producto de las desviaciones típicas de cada variable, ya que así se obtiene un coeficiente adimensional, r , que se denomina **coeficiente de correlación lineal de Pearson**

$$r = \frac{\mathcal{S}_{XY}}{\mathcal{S}_X \mathcal{S}_Y} \quad (3.1)$$

Propiedades del coeficiente de correlación lineal

- Carece de unidades de medida (adimensional).
- Es invariante para transformaciones lineales (cambio de origen y escala) de las variables.
- Sólo toma valores comprendidos entre -1 y 1 ,
- Cuando $|r|$ esté próximo a uno, se tiene que existe una *relación lineal* muy fuerte entre las variables.
- Cuando $r \approx 0$, puede afirmarse que no existe relación lineal entre ambas variables. Se dice en este caso que las variables son **incorreladas**.

3.6. Regresión

Las técnicas de regresión permiten hacer predicciones sobre los valores de cierta variable Y (*dependiente*), a partir de los de otra X (*independiente*), entre las que intuimos que existe una relación. Para ilustrarlo retomemos

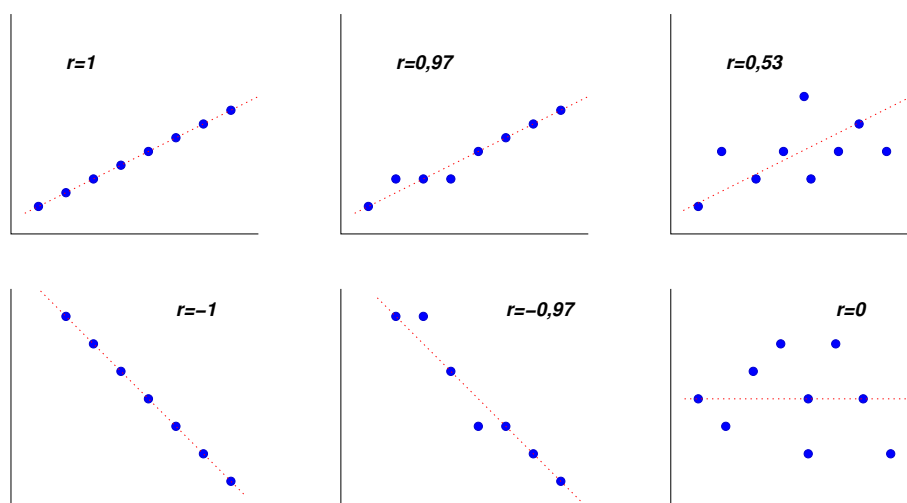


Figura 3.3: $r = \pm 1$ es lo mismo que decir que las observaciones de ambas variables están perfectamente alineadas. El signo de r , es el mismo que el de S_{XY} , por tanto nos indica el crecimiento o decrecimiento de la recta. La relación lineal es tanto más perfecta cuanto r está cercano a ± 1 .

los ejemplos mencionados al principio del capítulo. Si sobre un grupo de personas observamos los valores que toman las variables

$$X \equiv \text{altura medida en centímetros}, \quad (3.2)$$

$$Y \equiv \text{altura medida en metros}, \quad (3.3)$$

no es necesario hacer grandes esfuerzos para *intuir* que la relación que hay entre ambas es:

$$Y = \frac{X}{100}.$$

Obtener esta relación es menos evidente cuando lo que medimos sobre el mismo grupo de personas es

$$\begin{aligned} X &\equiv \text{altura medida en centímetros,} \\ Y &\equiv \text{peso en kilogramos.} \end{aligned}$$

La razón es que no es cierto que conocida la altura x_i de un individuo, podamos determinar de modo exacto su peso y_i (v.g. dos personas que miden $1,70m$ pueden tener pesos de 60 y 65 kilos). Sin embargo, alguna relación entre ellas debe existir, pues parece mucho más probable que un individuo de $2m$ pese más que otro que mida $1,20m$. Es más, nos puede parecer más o menos aproximada una relación entre ambas variables como la siguiente

$$Y = X - 110 \pm \text{error.}$$

A la deducción, a partir de una serie de datos, de este tipo de relaciones entre variables, es lo que denominamos **regresión**.

Mediante las técnicas de regresión inventamos una variable \hat{Y} como función de otra variable X (o viceversa),

$$\hat{Y} = f(X).$$

Esto es lo que denominamos **relación funcional**. El criterio para construir \hat{Y} , tal como citamos anteriormente, es que la diferencia entre Y e \hat{Y} sea pequeña.

$$\hat{Y} = f(X), \quad Y - \hat{Y} = \text{error},$$

El término que hemos denominado **error** debe ser tan pequeño como sea posible (figura 3.4). El objetivo será buscar la función (también denominada **modelo de regresión**) $\hat{Y} = f(X)$ que lo minimice. Véase la figura 3.5.

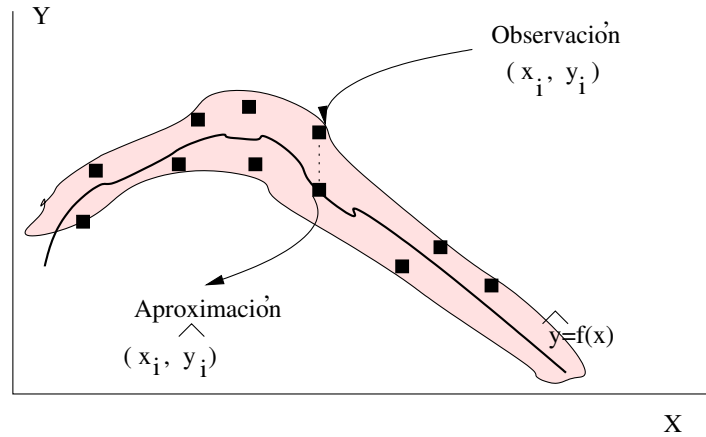


Figura 3.4: Mediante las técnicas de regresión de una variable Y sobre una variable X , buscamos una función que sea una buena aproximación de una nube de puntos (x_i, y_i) , mediante una curva del tipo $\hat{Y} = f(X)$. Para ello hemos de asegurarnos de que la diferencia entre los valores y_i e \hat{y}_i sea tan pequeña como sea posible.

3.6.1. Bondad de un ajuste

Consideremos un conjunto de observaciones sobre n individuos de una población, en los que se miden ciertas variables X e Y :

$$\begin{aligned} X &\rightsquigarrow x_1, x_2, \dots, x_n \\ Y &\rightsquigarrow y_1, y_2, \dots, y_n \end{aligned}$$

Estamos interesados en hacer regresión para determinar, de modo aproximado, los valores de Y conocidos los de X , debemos definir cierta variable $\hat{Y} = f(X)$, que debe tomar los valores

$$\hat{Y} \rightsquigarrow \hat{y}_1 = f(x_1), \hat{y}_2 = f(x_2), \dots, \hat{y}_n = f(x_n)$$

de modo que:

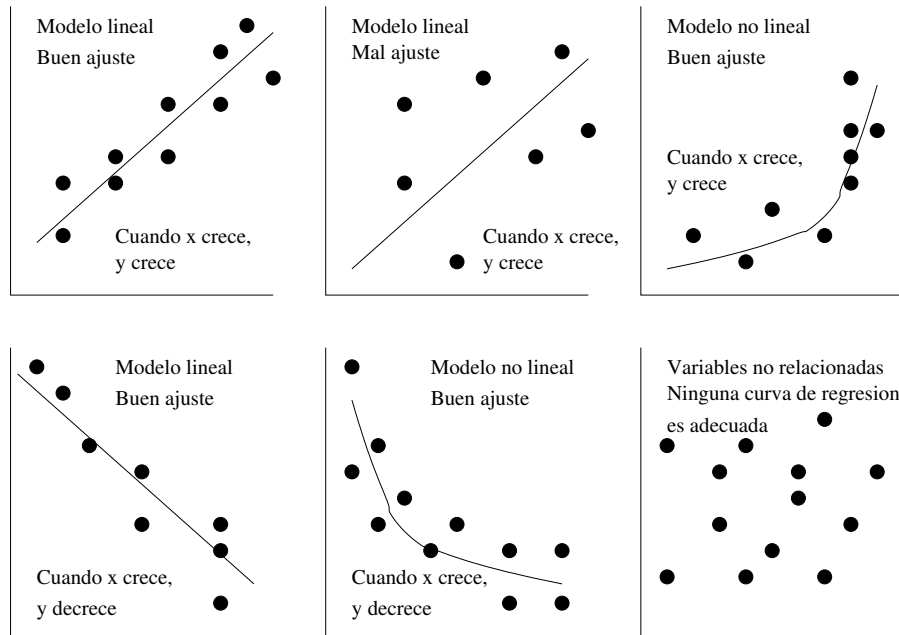


Figura 3.5: Diferentes nubes de puntos y modelos de regresión para ellas.

$$Y - \hat{Y} \rightsquigarrow y_1 - \hat{y}_1 \approx 0, y_2 - \hat{y}_2 \approx 0, \dots, y_n - \hat{y}_n \approx 0$$

Ello se puede expresar definiendo una nueva variable E que mida las diferencias entre los auténticos valores de Y y los teóricos suministrados por la regresión,

$$E = Y - \hat{Y} \rightsquigarrow e_1 = y_1 - \hat{y}_1, e_2 = y_2 - \hat{y}_2, \dots, e_n = y_n - \hat{y}_n$$

y calculando \hat{Y} de modo que E tome valores cercanos a 0. Dicho de otro modo, E debe ser una variable cuya media debe ser 0, y cuya varianza S_E^2 debe ser pequeña (en comparación con la de Y). Por ello se define el

coeficiente de determinación de la regresión de Y sobre X , $R^2_{Y|X}$, como

$$R^2_{Y|X} = 1 - \frac{S^2_E}{S^2_Y} \quad (3.4)$$

Si el ajuste de Y mediante la curva de regresión $\hat{Y} = f(X)$ es bueno, cabe esperar que la cantidad $R^2_{Y|X}$ tome un valor próximo a 1.

La cantidad $R^2_{Y|X}$ sirve entonces para medir de qué modo las diferencias entre los verdaderos valores de una variable y los de su aproximación mediante una curva de regresión son pequeños en relación con los de la variabilidad de la variable que intentamos aproximar. Por esta razón estas cantidades miden el **grado de bondad del ajuste**.

3.6.2. Regresión lineal

La **regresión lineal** consiste en encontrar aproximar los valores de una variable a partir de los de otra, usando una relación funcional de tipo lineal, es decir, buscamos cantidades a y b tales que se pueda escribir

$$\hat{Y} = a + b \cdot X \quad (3.5)$$

con el menor error posible entre \hat{Y} e Y .

Las cantidades a y b que minimizan dicho error son los llamados *coeficientes de regresión*:

$$a = \bar{y} - b \bar{x}$$

$$b = \frac{S_{XY}}{S^2_X}$$

La cantidad b se denomina *coeficiente de regresión de Y sobre X* .

En el modelo lineal de regresión la *bondad del ajuste* es simplemente r^2 . Con lo cual el modelo lineal dará mejores predicciones cuando r sea próximo a 1 ó -1.

Interpretación de los coeficientes de regresión

Obsérvese que la relación 3.5 explica cosas como que si X varía en 1 unidad, \hat{Y} varía la cantidad b . Por tanto:

- Si $b > 0$, las dos variables aumentan o disminuyen a la vez;
- Si $b < 0$, cuando una variable aumenta, la otra disminuye.

Ejemplo de cálculo con un modelo de regresión lineal

En una muestra de 1.500 individuos se recogen datos sobre dos medidas antropométricas X e Y . Los resultados se muestran resumidos en los siguientes estadísticos:

$$\begin{aligned}\bar{x} &= 14 & \mathcal{S}_X &= 2 \\ \bar{y} &= 100 & \mathcal{S}_Y &= 25 \\ & & \mathcal{S}_{XY} &= 45\end{aligned}$$

Obtener el modelo de regresión lineal que mejor aproxima Y en función de X . Utilizando este modelo, calcular de modo aproximado la cantidad Y esperada cuando $X = 15$.

Solución:

Lo que se busca es la recta, $\hat{Y} = a + b \cdot X$, que mejor aproxima los valores de Y (según el criterio de los mínimos cuadrados) en la nube de puntos que resulta de representar en un plano (X, Y) las 1.500 observaciones. Los coeficientes de esta recta son:

$$\begin{aligned}b &= \frac{\mathcal{S}_{XY}}{\mathcal{S}_X^2} = \frac{45}{4} = 11,25 \\ a &= \bar{y} - b \cdot \bar{x} = 100 - 11,25 \times 14 = -57,5\end{aligned}$$

Así, el modelo lineal consiste en:

$$\hat{Y} = -57,5 + 11,25 \cdot X$$

Por tanto, si $x = 15$, el modelo lineal predice un valor de Y de:

$$\hat{y} = -57,5 + 11,25 \cdot x = -57,5 + 11,25 \times 15 = 111,25$$

Propiedades de la regresión lineal

Una vez que ya tenemos perfectamente definida \hat{Y} , (o bien \hat{X}) nos preguntamos las relaciones que hay entre la media y la varianza de esta y la de Y (o la de X). La respuesta nos la ofrece la siguiente proposición:

Proposición

En los ajustes lineales se conservan las medias, es decir

$$\bar{\hat{y}} = \bar{y} \quad (3.6)$$

$$\bar{\hat{x}} = \bar{x} \quad (3.7)$$

En cuanto a la varianza, no necesariamente son las mismas para los verdaderos valores de las variables X e Y y sus aproximaciones \hat{X} y \hat{Y} , pues sólo se mantienen en un factor de r^2 , es decir,

$$\mathcal{S}_{\hat{Y}}^2 = r^2 \mathcal{S}_Y^2 \quad (3.8)$$

$$\mathcal{S}_{\hat{X}}^2 = r^2 \mathcal{S}_X^2 \quad (3.9)$$

Observación

Como consecuencia de este resultado, podemos decir que *la proporción de varianza explicada por la regresión lineal es del $r^2 \cdot 100\%$* .

Nos gustaría tener que $r = 1$, pues en ese caso ambas variables tendrían la misma varianza, pero esto no es cierto en general. Todo lo que se puede afirmar, como sabemos, es que

$$-1 \leq r \leq 1$$

y por tanto

$$0 \leq \mathcal{S}_{\hat{Y}}^2 \leq \mathcal{S}_Y^2$$

La cantidad que le falta a la **varianza de regresión**, $\mathcal{S}_{\hat{Y}}^2$, para llegar hasta la varianza total de Y , \mathcal{S}_Y^2 , es lo que se denomina **varianza residual**,

Proposición

La varianza residual del modelo de regresión es de Y sobre X es la varianza de la variable $E = Y - \hat{Y}$.

Obsérvese que entonces La *bondad del ajuste* es

$$R_{Y|X}^2 = 1 - \frac{\mathcal{S}_E^2}{\mathcal{S}_Y^2} = 1 - (1 - r^2) = r^2$$

Para el ajuste contrario se define el error como $E = X - \hat{X}$, y análogamente su varianza residual es también proporcional a $1 - r^2$. Todo esto se puede resumir como sigue:

Proposición

Para los ajustes de tipo lineal se tiene que los dos coeficientes de determinación son iguales a r^2 , y por tanto representan además la proporción de varianza explicada por la regresión lineal:

$$\boxed{R_{X|Y}^2 = r^2 = R_{Y|X}^2}$$

Por ello:

- Si $|r| \approx 1$ el ajuste es bueno (Y se puede calcular de modo bastante aproximado a partir de X y viceversa).
- Si $|r| \approx 0$ las variables X e Y no están relacionadas (linealmente al menos), por tanto no tiene sentido hacer un ajuste lineal. Sin embargo

no es seguro que las dos variables no posean ninguna relación en el caso $r = 0$, ya que si bien el ajuste lineal puede no ser procente, tal vez otro tipo de ajuste sí lo sea.

Ejemplo

De una muestra de ocho observaciones conjuntas de valores de dos variables X e Y , se obtiene la siguiente información:

$$\sum x_i = 24; \quad \sum x_i y_i = 64; \quad \sum y_i = 40;$$

$$S_Y^2 = 12; \quad S_X^2 = 6.$$

Calcule:

1. La recta de regresión de Y sobre X . Explique el significado de los parámetros.
2. El coeficiente de determinación. Comente el resultado e indique el tanto por ciento de la variación de Y que no está explicada por el modelo lineal de regresión.
3. Si el modelo es adecuado, ¿cuál es la predicción \hat{y} para $x = 4$.

Solución:

1. En primer lugar calculamos las medias y las covarianza entre ambas variables:

$$\begin{aligned} \bar{x} &= \sum x_i / n = 24/8 = 3 \\ \bar{y} &= \sum y_i / n = 40/8 = 5 \\ S_{XY} &= (\sum x_i y_i) / n - \bar{x} \bar{y} = 64/8 - 3 \times 5 = -7 \end{aligned} \quad (3.10)$$

Con estas cantidades podemos determinar los parámetros a y b de la recta. La pendiente de la misma es b , y mide la variación de Y cuando X aumenta en una unidad:

$$b = \frac{S_{XY}}{S_X^2} = \frac{-7}{6} = -1,667$$

Al ser esta cantidad negativa, tenemos que la pendiente de la recta es negativa, es decir, a medida que X aumenta, la tendencia es a la disminución de Y . En cuanto al valor de la ordenada en el origen, a , tenemos:

$$a = \bar{y} - b \cdot \bar{x} = 5 - \left(\frac{-7}{6}\right) \times 3 = 8,5$$

Así, la recta de regresión de Y como función de X es:

$$\hat{Y} = 8,5 - 1,667 \cdot X$$

2. El grado de bondad del ajuste lo obtenemos a partir del coeficiente de determinación:

$$R_{Y/X}^2 = r^2 = \left(\frac{S_{XY}}{S_X \cdot S_Y}\right)^2 = \frac{(-7)^2}{6 \times 12} = 0,6805 = 68,05\%$$

Es decir, el modelo de regresión lineal explica el 68 % de la variabilidad de Y en función de la de X . Por tanto queda un 32 % de variabilidad no explicada.

3. La predicción que realiza el modelo lineal de regresión para $x = 4$ es:

$$\hat{y} = 8,5 - 1,667 \cdot x = 8,5 - 1,667 \times 4 = 3,833$$

la cual hay que considerar con ciertas reservas, pues como hemos visto en el apartado anterior, hay una razonable cantidad de variabilidad que no es explicada por el modelo.

Ejemplo de cálculo en regresión lineal

En un grupo de 8 pacientes se miden las cantidades antropométricas *peso* y *edad*, obteniéndose los siguientes resultados:

Resultado de las mediciones								
$X \equiv \text{edad}$	12	8	10	11	7	7	10	14
$Y \equiv \text{peso}$	58	42	51	54	40	39	49	56

¿Existe una relación lineal importante entre ambas variables? Calcular la recta de regresión de la edad en función del peso y la del peso en función de la edad. Calcular la bondad del ajuste ¿En qué medida, por término medio, varía el peso cada año? ¿En cuánto aumenta la edad por cada kilo de peso?

Solución:

Para saber si existe una relación lineal entre ambas variables se calcula el coeficiente de correlación lineal, que vale:

$$r = \frac{\mathcal{S}_{XY}}{\mathcal{S}_X \mathcal{S}_Y} = \frac{15,2031}{2,3150 \times 6,9631} = 0,9431$$

ya que

$$\sum_{i=1}^8 x_i = 79 \implies \bar{x} = \frac{79}{8} = 9,875 \text{ años}$$

$$\sum_{i=1}^8 y_i = 389 \implies \bar{y} = \frac{389}{8} = 48,625 \text{ Kg}$$

$$\sum_{i=1}^8 x_i^2 = 823 \implies \mathcal{S}_X^2 = \frac{823}{8} - 9,875^2 = 5,3594 \text{ años}^2$$

$$\implies \mathcal{S}_X = 2,3150 \text{ años}$$

$$\sum_{i=1}^8 y_i^2 = 19,303 \implies \mathcal{S}_Y^2 = \frac{19,303}{8} - 48,625^2 = 48,4844 \text{ Kg}^2$$

$$\implies \mathcal{S}_Y = 6,9631 \text{ Kg}$$

$$\sum_{i=1}^8 x_i y_i = 3,963 \implies \mathcal{S}_{XY} = \frac{3,963}{8} - 9,875 \times 48,625 = 15,2031 \text{ Kg} \cdot \text{año}$$

Por tanto el ajuste lineal es muy bueno. Se puede decir que el ángulo entre el vector formado por las desviaciones del peso con respecto a su valor medio y el de la edad con respecto a su valor medio, θ , es:

$$r = \cos \theta \quad \implies \quad \theta = \arccos r \approx 19^\circ$$

es decir, entre esos vectores hay un buen grado de paralelismo (sólo unos 19 grados de desviación).

La recta de regresión del peso en función de la edad es

$$\begin{aligned} \hat{Y} &= a_1 + b_1 X = 20,6126 + 2,8367 \cdot X \\ a_1 &= \bar{y} - b_1 \bar{x} = 20,6126 \text{ Kg} \\ b_1 &= \frac{\mathcal{S}_{XY}}{\mathcal{S}_X^2} = 2,8367 \text{ Kg/año} \end{aligned} \quad (3.11)$$

La recta de regresión de la edad como función del peso es

$$\begin{aligned} \hat{X} &= a_2 + b_2 Y = -5,3738 + 0,3136 \cdot Y \\ a_2 &= \bar{x} - b_2 \bar{y} = -5,3738 \text{ años} \\ b_2 &= \frac{\mathcal{S}_{XY}}{\mathcal{S}_Y^2} = 0,3136 \text{ años/Kg} \end{aligned}$$

que como se puede comprobar, no resulta de despejar en la recta de regresión de Y sobre X .

La bondad del ajuste es

$$R_{X|Y}^2 = R_{Y|X}^2 = r^2 = 0,8894$$

por tanto podemos decir que el 88,94% de la variabilidad del peso en función de la edad es explicada mediante la recta de regresión correspondiente. Lo mismo podemos decir en cuanto a la variabilidad de la edad en función del peso. Del mismo modo puede decirse que hay un $100 - 88,94\% = 11,06\%$ de varianza que no es explicada por las rectas

de regresión. Por tanto la varianza residual de la regresión del peso en función de la edad es

$$\mathcal{S}_E^2 = (1 - r^2) \cdot \mathcal{S}_Y^2 = 0,1106 \times 48,4844 = 5,33 \text{ Kg}^2$$

y la de la edad en función del peso:

$$\mathcal{S}_E^2 = (1 - r^2) \cdot \mathcal{S}_X^2 = 0,1106 \times 5,3594 = 0,59 \text{ años}^2$$

Por último la cantidad en que varía el peso de un paciente cada año es, según la recta de regresión del peso en función de la edad, la pendiente de esta recta, es decir, $b_1 = 2,8367 \text{ Kg/año}$. Cuando dos personas difieren en peso, en promedio la diferencia de edad entre ambas se rige por la cantidad $b_2 = 0,3136 \text{ años/Kg}$ de diferencia.

3.7. Problemas

Ejercicio 3.1. Se realiza un estudio para establecer una ecuación mediante la cual se pueda utilizar la *concentración de estrona en saliva* (X) para predecir la *concentración del esteroide en plasma libre* (Y). Se extrajeron los siguientes datos de 14 varones sanos:

X	1,4	7,5	8,5	9	9	11	13	14	14,5	16	17	18	20	23
Y	30	25	31,5	27,5	39,5	38	43	49	55	48,5	51	64,5	63	68

1. Estúdiese la posible relación lineal entre ambas variables.
2. Obtener la ecuación que se menciona en el enunciado del problema.
3. Determinar la variación de la concentración de estrona en plasma por unidad de estrona en saliva.

Ejercicio 3.2. Los investigadores están estudiando la correlación entre *obesidad* y la *respuesta individual al dolor*. La obesidad se mide como porcentaje sobre el peso ideal (X). La respuesta al dolor se mide utilizando el

umbral de reflejo de flexión nociceptiva (Y), que es una medida de sensación de punzada. Se obtienen los siguientes datos:

X	89	90	75	30	51	75	62	45	90	20
Y	2	3	4	4,5	5,5	7	9	13	15	14

1. ¿Qué porcentaje de la varianza del peso es explicada mediante un modelo de regeseión lineal por la variación del umbral de reflejo?
2. Estúdiese la posible relación lineal entre ambas variables, obteniendo su grado de ajuste.
3. ¿Qué porcentaje de sobrepeso podemos esperar para un umbral de reflejo de 10?

Ejercicio 3.3. Se lleva a cabo un estudio, por medio de detectores radioactivos, de la *capacidad corporal para absorber hierro y plomo*. Participan en el estudio 10 sujetos. A cada uno se le da una dosis oral idéntica de hierro y plomo. Después de 12 días se mide la cantidad de cada componente retenida en el sistema corporal y, a partir de ésta, se determina el porcentaje absorbido por el cuerpo. Se obtuvieron los siguientes datos:

Porcentaje de hierro $\equiv X$	17	22	35	43	80	85	91	92	96	100
Porcentaje de plomo $\equiv Y$	8	17	18	25	58	59	41	30	43	58

1. Comprobar la idoneidad del modelo lineal de regresión.
2. Obtener la recta de regresión, si el modelo lineal es adecuado.
3. Predecir el porcentaje de hierro absorbido por un individuo cuyo sistema corporal absorbe el 15 % del plomo ingerido.

Ejercicio 3.4. Para estudiar el efecto de las aguas residuales de las alcantarillas que afluyen a un lago, se toman medidas de la concentración de nitrato en el agua. Para monitorizar la variable se ha utilizado un antiguo *método manual*. Se idea un nuevo *método automático*. Si se pone de manifiesto una alta correlación positiva entre las medidas tomadas empleando los dos métodos, entonces se hará uso habitual del método automático. Los datos obtenidos son los siguientes:

Manual $\equiv X$	25	40	120	75	150	300	270	400	450	575
Automático $\equiv Y$	30	80	150	80	200	350	240	320	470	583

1. Hallar el coeficiente de determinación para ambas variables.
2. Comprobar la idoneidad del modelo lineal de regresión. Si el modelo es apropiado, hallar la recta de regresión de Y sobre X y utilizarla para predecir la lectura que se obtendría empleando la técnica automática con una muestra de agua cuya lectura manual es de 100.
3. Para cada una de las observaciones, halle las predicciones que ofrece el modelo lineal de regresión para X en función de Y , e Y en función de X , es decir, \hat{X} e \hat{Y} .
4. Calcule los errores para cada una de dichas predicciones, es decir, las variables $X - \hat{X}$ e $Y - \hat{Y}$.
5. ¿Que relación hay entre las medias de X y \hat{X} ? ¿Y entre las de Y e \hat{Y} ?
6. Calcule las medias de $X - \hat{X}$ e $Y - \hat{Y}$. ¿Era de esperar el valor obtenido?
7. Calcule las varianzas de X , \hat{X} , Y , \hat{Y} , $X - \hat{X}$ e $Y - \hat{Y}$.
8. ¿Qué relación existe entre S_X^2 y $S_{\hat{X}}^2$? ¿Y entre S_Y^2 y $S_{\hat{Y}}^2$?
9. ¿Que relación encuentra entre S_X^2 y $S_{X-\hat{X}}^2$? ¿También es válida para S_Y^2 y $S_{Y-\hat{Y}}^2$?

10. Justifique a partir de todo lo anterior porqué se denomina r^2 como **grado de bondad del ajuste lineal**.

Ejercicio 3.5. Se ha medido el aclaramiento de creatinina en pacientes tratados con Captopril tras la suspensión del tratamiento con diálisis, resultando la siguiente tabla:

Días tras la diálisis $\equiv X$	1	5	10	15	20	25	35
Creatinina (mg/dl) $\equiv Y$	5,7	5,2	4,8	4,5	4,2	4	3,8

- Hállese la expresión de la ecuación lineal que mejor exprese la variación de la creatinina, en función de los días transcurridos tras la diálisis, así como el grado de bondad de ajuste y la varianza residual.
- ¿En qué porcentaje la variación de la creatinina es explicada por el tiempo transcurrido desde la diálisis?
- Si un individuo presenta 4'1 mg/dl de creatinina, ¿cuánto tiempo es de esperar que haya transcurrido desde la suspensión de la diálisis?

Ejercicio 3.6. En un ensayo clínico realizado tras el posible efecto hipotensor de un fármaco, se evalúa la tensión arterial diastólica (TAD) en condiciones basales (X), y tras 4 semanas de tratamiento (Y), en un total de 14 pacientes hipertensos. Se obtienen los siguiente valores de TAD:

X	95	100	102	104	100	95	95	98	102	96	100	96	110	99
Y	85	94	84	88	85	80	80	92	90	76	90	87	102	89

- ¿Existe relación lineal entre la TAD basal y la que se observa tras el tratamiento?
- ¿Cuál es el valor de TAD esperado tras el tratamiento, en un paciente que presentó una TAD basal de 95 mm de Hg?

Ejercicio 3.7. Se han realizado 9 tomas de presión intracraneal en animales de laboratorio, por un *método estándar directo* y por una nueva *técnica experimental indirecta*, obteniéndose los resultados siguientes en mm de Hg:

Método estándar $\equiv X$	9	12	28	72	30	38	76	26	52
Método experimental $\equiv Y$	6	10	27	67	25	35	75	27	53

1. Hallar la ecuación lineal que exprese la relación existente entre las presiones intracraneales, determinadas por los dos métodos.
2. ¿Qué tanto por ciento de la variabilidad de Y es explicada por la regresión? Hállese el grado de dependencia entre las dos variables y la varianza residual del mismo.

Capítulo 4

Cálculo de probabilidades y variables aleatorias

4.1. introducción

Si el único propósito del investigador es describir los resultados de un experimento concreto, los métodos analizados en los capítulos anteriores pueden considerarse suficientes. No obstante, si lo que se pretende es utilizar la información obtenida para extraer conclusiones generales sobre todos aquellos objetos del tipo de los que han sido estudiados, entonces estos métodos constituyen sólo el principio del análisis, y debe recurrirse a métodos de inferencia estadística, los cuales implican el uso inteligente de la teoría de la probabilidad.

Comenzamos este bloque interpretando la noción de probabilidad y la terminología subyacente a esta área de las matemáticas, ya que la probabilidad constituye por sí misma un concepto básico que refleja su relación con la faceta del mundo exterior que pretende estudiar: los fenómenos aleatorios, los cuales obedecen unas ciertas reglas de comportamiento. De alguna manera, el concepto de probabilidad, se relaciona o nos recuerda las propiedades de la frecuencia relativa.

A partir de ella, y junto con las definiciones de probabilidad condicionada y la de sucesos independientes, se deducen los teoremas fundamentales

del Cálculo de Probabilidades.

Nos centraremos posteriormente en el eslabón que une la teoría de la probabilidad y la estadística aplicada: la noción de variable aleatoria, mostrando de esta manera, como puede emplearse la teoría de la probabilidad para sacar conclusiones precisas acerca de una población en base a una muestra extraída de ella, y que muchos de los estudios estadísticos son de hecho, estudio de las propiedades de una o más variables aleatorias.

Tal como hemos citado anteriormente, en las aplicaciones prácticas es importante poder describir los rasgos principales de una distribución, es decir, caracterizar los resultados del experimento aleatorio mediante unos parámetros. Llegamos así al estudio de las características asociadas a una variable aleatoria introduciendo los conceptos de esperanza y varianza matemática, relacionándolos con los conceptos de media y varianza de una variable estadística.

El **cálculo de probabilidades** nos suministra las reglas para el estudio de los experimentos aleatorios o de azar, constituyendo la base para la estadística inductiva o inferencial.

Para trabajar con el cálculo de probabilidades es necesario fijar previamente cierta terminología. Vamos a introducir parte de ella en las próximas líneas.

4.2. Experimentos y sucesos aleatorios

Diremos que *un experimento es aleatorio* si se verifican las siguientes condiciones:

1. Se puede repetir indefinidamente, siempre en las mismas condiciones;
2. Antes de realizarlo, no se puede predecir el resultado que se va a obtener;
3. El resultado que se obtenga, e , pertenece a un conjunto conocido

previamente de resultados posibles. A este conjunto, de resultados posibles, lo denominaremos **espacio muestral** y lo denotaremos normalmente mediante la letra E . Los elementos del espacio muestral se denominan **sucesos elementales**.

$$e_1, e_2 \in E \quad \implies \quad e_1, e_2 \text{ son sucesos elementales.}$$

Cualquier subconjunto de E será denominado **suceso aleatorio**, y se denotará normalmente con las letras A, B, \dots

$$A, B \subset E \quad \implies \quad A, B \text{ son sucesos aleatorios.}$$

4.2.1. Operaciones básicas con sucesos aleatorios

Al ser los sucesos aleatorios nada más que subconjuntos de un conjunto E —espacio muestral—, podemos aplicarles las conocidas operaciones con conjuntos, como son la unión, intersección y diferencia:

Unión:

Dados dos sucesos aleatorios $A, B \subset E$, se denomina *suceso unión* de A y B al conjunto formado por todos los sucesos elementales que pertenecen a A o bien que pertenecen a B (incluyendo los que están en ambos simultáneamente), es decir

$$A \cup B = \{e \in E : e \in A \text{ ó } e \in B\} \quad (4.1)$$

Intersección:

Dados dos sucesos aleatorios $A, B \subset E$, se denomina *suceso intersección* de A y B al conjunto formado por todos los sucesos elementales que pertenecen a A y B a la vez, es decir,

$$A \cap B = \{e \in E : e \in A \text{ y además } e \in B\} \quad (4.2)$$

Diferencia:

Dados dos sucesos aleatorios $A, B \subset E$, se llama *suceso diferencia* de A y B , y se representa mediante $A \setminus B$, o bien $A - B$, al suceso aleatorio formado por todos los sucesos elementales que pertenecen a A , pero no a B :

$$A \setminus B \equiv A - B = \{e \in E : e \in A \text{ y además } e \notin B\} = A \cap \overline{B} \quad (4.3)$$

Diferencia simétrica:

Si $A, B \subset E$, se denomina suceso diferencia simétrica de A y B , y se representa mediante $A \triangle B$, al suceso aleatorio formado por todos los sucesos elementales que pertenecen a A y no a B , y los que están en B y no en A :

$$A \triangle B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B) \quad (4.4)$$

4.3. Experimentos aleatorios y probabilidad

Se denominan **experimentos deterministas** aquellos que realizados de una misma forma y con las mismas condiciones iniciales, ofrecen siempre el mismo resultado. Como ejemplo, tenemos que un objeto de cualquier masa partiendo de un estado inicial de reposo, y dejado caer al vacío desde una torre, llega siempre al suelo con la misma velocidad: $v = \sqrt{2gh}$.

Cuando en un experimento no se puede predecir el resultado final, hablamos de **experimento aleatorio**. Este es el caso cuando lanzamos un dado y observamos su resultado.

4.3.1. Noción frecuentista de probabilidad

En los experimentos aleatorios se observa que cuando el número de experimentos aumenta, las frecuencias relativas con las que ocurre cierto suceso e , $f_n(e)$,

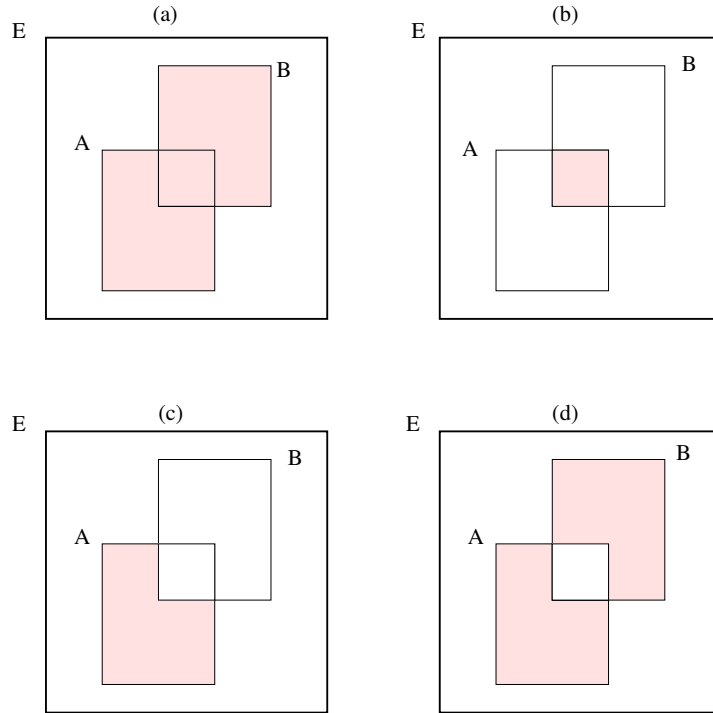


Figura 4.1: Dados dos sucesos aleatorios $A, B \subset E$ se representa: en (a) $A \cup B$; en (b) $A \cap B$; en (c) $A - B$; en (d) $A \Delta B$.

$$f_n(e) = \frac{\text{número de ocurrencias de } e}{n}$$

tiende a converger hacia cierta cantidad que denominamos **probabilidad** de e . Esta es la **noción frecuentista de probabilidad**.

$$\mathcal{P}_{rob}[e] = \lim_{n \rightarrow \infty} f_n(e)$$

En la Figura 4.2 se presenta la evolución de la frecuencia relativa del número de caras obtenido en el lanzamiento de una moneda en 100 ocasiones

(simulado por un ordenador). En principio la evolución de las frecuencias relativas es errática, pero a medida que el número de tiradas aumenta, tiende a lo que entendemos por probabilidad de cara.

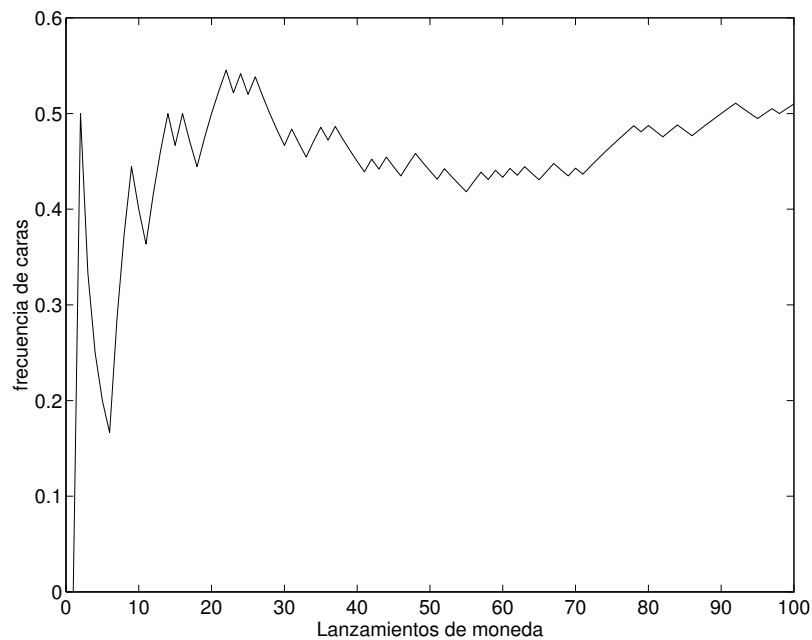


Figura 4.2: Convergencia a $1/2$ de la frecuencia relativa del número de caras obtenido en lanzamientos sucesivos de una moneda (simulación en ordenador).

Problemas de la noción frecuentista de probabilidad

La noción frecuentista de probabilidad no puede usarse en la práctica como definición de la probabilidad por que::

- se requiere realizar un número infinito de veces un experimento para calcular una probabilidad. Por ejemplo, lanzar infinitas veces un

dado para ver que las frecuencias relativas de la aparición de cada cara convergen a $1/6$. Esto puede suplirse en la práctica realizando el experimento un número suficientemente elevado de veces, hasta que tengamos la precisión que requieran nuestros cálculos. Sin embargo,

- los experimentos aleatorios a veces no pueden ser realizados, como es el caso de calcular la probabilidad de morir jugando a la ruleta rusa con un revolver: no es posible (o no se debe) calcular esta probabilidad repitiendo el experimento un número indefinidamente alto de veces para aproximarla mediante la frecuencia relativa). Para ello existen métodos mucho más seguros, como los que mencionaremos a continuación.

4.3.2. Probabilidad de Laplace

Si un experimento cualquiera puede dar lugar a un número finito de resultados posibles, y no existe ninguna razón que privilegie unos resultados en contra de otros, se calcula la probabilidad de un suceso aleatorio A , según la **regla de Laplace** como el cociente entre el número de casos favorables a A , y el de todos los posibles resultados del experimento:

$$\mathcal{P}[A] = \frac{\text{número de casos favorables a } A}{\text{número de casos posibles}}$$

4.3.3. Definición axiomática de probabilidad

Para hacer una definición rigurosa de la probabilidad, necesitamos precisar ciertas leyes o axiomas que deba cumplir una función de probabilidad. Con la **definición axiomática de la probabilidad** pretendemos dar el menor conjunto posible de estas reglas, para que las demás se deduzcan como una simple consecuencia de ellas.

Concepto axiomático de probabilidad

Dado un espacio muestral E , diremos que \mathcal{P} es una **probabilidad** sobre \mathcal{A} si las siguientes propiedades (*axiomas*) son verificadas:

Ax-1. La probabilidad es una función definida sobre \mathcal{A} y que sólo toma valores positivos comprendidos entre 0 y 1

$$\mathcal{P} : \mathcal{A} \longrightarrow [0, 1] \subset \mathbb{R}$$

$$A \subset E, A \in \mathcal{A} \longmapsto 0 \leq \mathcal{P}[A] \leq 1$$

Ax-2. La probabilidad del suceso seguro es 1

$$\mathcal{P}[E] = 1$$

Ax-3. La probabilidad de la unión numerable de sucesos disjuntos es la suma de sus probabilidades (figura 4.3):

$$A_1, A_2, \dots, A_n, \dots \in \mathcal{A} \implies \mathcal{P}\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathcal{P}[A_i]$$

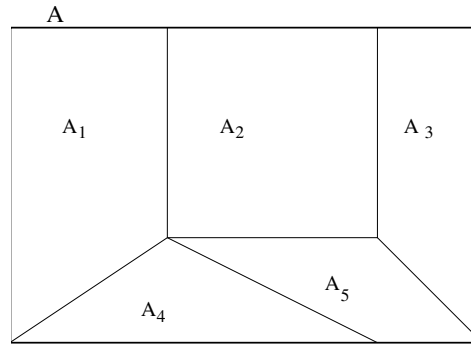


Figura 4.3: El tercer axioma de probabilidad indica que si $A = A_1 \cup A_2 \cup \dots$ con $A_i \cap A_j = \emptyset$, entonces $\mathcal{P}[A] = \mathcal{P}[A_1] + \mathcal{P}[A_2] + \dots$

4.4. Probabilidad condicionada e independencia de sucesos

Sea $B \subset E$ un suceso aleatorio de probabilidad no nula, $\mathcal{P}[B] > 0$. Para cualquier otro suceso $A \subset E$, llamamos **probabilidad condicionada** de

A a B a la cantidad que representamos mediante $\mathcal{P}[A|B]$ o bien $\mathcal{P}_B[A]$ y que se calcula como:

$$\mathcal{P}[A|B] = \frac{\mathcal{P}[A \cap B]}{\mathcal{P}[B]}$$

Ejemplo de cálculo de probabilidades condicionadas

Se lanza un dado al aire ¿Cuál es la probabilidad de que salga el número 4? Si sabemos que el resultado ha sido un número par, ¿se ha modificado esta probabilidad?

Solución:

El espacio muestral que corresponde a este experimento es

$$E = \{1, 2, 3, 4, 5, 6\}$$

y se ha de calcular la probabilidad del suceso $A = \{4\}$. Si el dado no está trucado, todos los números tienen la misma probabilidad de salir, y siguiendo la definición de probabilidad de Laplace,

$$\begin{aligned} \mathcal{P}[A] &= \frac{\text{casos favorables}}{\text{casos posibles}} \\ &= \frac{\text{número de elementos en } \{4\}}{\text{número de elementos en } \{1, 2, 3, 4, 5, 6\}} \\ &= \frac{1}{6} \end{aligned} \tag{4.5}$$

Obsérvese que para calcular la probabilidad de A según la definición de Laplace hemos tenido que suponer previamente que todos los elementos del espacio muestral tienen la misma probabilidad de salir, es decir:

$$\mathcal{P}[1] = \mathcal{P}[2] = \mathcal{P}[3] = \mathcal{P}[4] = \mathcal{P}[5] = \mathcal{P}[6]$$

Por otro lado, si ha salido un número par, de nuevo por la definición de probabilidad de Laplace tendríamos

$$\begin{aligned}
\mathcal{P}_{\text{par}}[4] &= \frac{\text{casos favorables}}{\text{casos posibles}} \\
&= \frac{\text{número de elementos en } \{4\}}{\text{número de elementos en } \{2, 4, 6\}} \\
&= \frac{1}{3}
\end{aligned}$$

Esta misma probabilidad se podría haber calculado siguiendo la definición de la probabilidad condicionada, ya que si escribimos

$$\begin{aligned}
A = \{4\} &\Rightarrow \mathcal{P}[A] = \frac{1}{6} \\
B = \{2, 4, 6\} &\Rightarrow \mathcal{P}[B] = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2} \\
A \cap B = \{4\} &\Rightarrow \mathcal{P}[A \cap B] = \frac{1}{6}
\end{aligned} \tag{4.6}$$

y entonces

$$\mathcal{P}_{\text{par}}[4] = \mathcal{P}_B[A] = \mathcal{P}[A|_B] = \frac{\mathcal{P}[A \cap B]}{\mathcal{P}[B]} = \frac{1/6}{1/2} = \frac{1}{3}$$

que por supuesto coincide con el mismo valor que calculamos usando la definición de probabilidad de Laplace.

Independencia

Obsérvese que según la definición de probabilidad condicionada, se puede escribir la probabilidad de la intersección de dos sucesos de probabilidad no nula como

$$\mathcal{P}[A \cap B] = \begin{cases} \mathcal{P}[A] \cdot \mathcal{P}[B|_A] \\ \mathcal{P}[B] \cdot \mathcal{P}[A|_B] \end{cases}$$

O sea, la probabilidad de la intersección de dos sucesos, es la probabilidad de uno cualquiera de ellos, multiplicada por la probabilidad del segundo *sabiendo que* ha ocurrido el primero.

Si entre dos sucesos no existe ninguna relación cabe esperar que la expresión “*sabiendo que*” no aporte ninguna información. De este modo introducimos el concepto de **independencia de dos sucesos A y B** como:

$$A \text{ es independiente de } B \iff \mathcal{P}[A \cap B] = \mathcal{P}[A] \cdot \mathcal{P}[B]$$

4.5. Teoremas fundamentales del cálculo de probabilidades

Hay algunos resultados importantes del cálculo de probabilidades que son conocidos bajo los nombres de *teorema de la probabilidad compuesta*, *teorema de la probabilidad total* y *teorema de Bayes*. Veamos cuales son estos teoremas, pero previamente vamos a enunciar a modo de recopilación, una serie de resultados elementales.

Reglas de cálculo de probabilidades básicas

Sean $A, B \subset E$ no necesariamente disjuntos. Se verifican entonces las siguientes propiedades:

1. Probabilidad de la unión de sucesos:

$$\mathcal{P}[A \cup B] = \mathcal{P}[A] + \mathcal{P}[B] - \mathcal{P}[A \cap B] \quad (4.7)$$

2. Probabilidad de la intersección de sucesos:

$$\mathcal{P}[A \cap B] = \begin{cases} \mathcal{P}[A] \cdot \mathcal{P}[B|A] \\ \mathcal{P}[B] \cdot \mathcal{P}[A|B] \end{cases} \quad (4.8)$$

3. Probabilidad del suceso contrario:

$$\boxed{\mathcal{P}[\bar{A}] = 1 - \mathcal{P}[A]} \quad (4.9)$$

4. Probabilidad condicionada del suceso contrario:

$$\boxed{\mathcal{P}[\bar{A}|B] = 1 - \mathcal{P}[A|B]} \quad (4.10)$$

Ejemplo de cálculo de probabilidades con intersecciones

En una universidad el 50 % de los alumnos habla inglés, el 20 % francés y el 5 % los dos idiomas ¿Cuál es la probabilidad de encontrar alumnos que hablen alguna lengua extranjera?

Solución:

Sea A el suceso *hablar inglés*: $\mathcal{P}[A] = 0,5$.

Sea B el suceso *hablar francés*: $\mathcal{P}[B] = 0,2$.

El suceso *hablar francés e inglés* es $A \cap B$: $\mathcal{P}[A \cap B] = 0,05$.

Así:

$$\mathcal{P}[A \cup B] = \mathcal{P}[A] + \mathcal{P}[B] - \mathcal{P}[A \cap B] = 0,5 + 0,2 - 0,05 = 0,65$$

4.5.1. Teorema de la probabilidad compuesta

Sea $A_1, A_2, \dots, A_n \subset E$ una colección de sucesos aleatorios. Entonces:

$$\mathcal{P}[A_1 A_2 \cdots A_n] = \mathcal{P}[A_1] \cdot \mathcal{P}[A_2 | A_1] \cdot \mathcal{P}[A_3 | A_1 A_2] \cdots \mathcal{P}[A_n | A_1 A_2 \cdots A_{n-1}]$$

4.5.2. Sistema exhaustivo y excluyente de sucesos

Los teoremas que restan nos dicen como calcular las probabilidades de sucesos cuando tenemos que el suceso seguro está descompuesto en una serie de sucesos incompatibles de los que conocemos su probabilidad. Para ello necesitamos introducir un nuevo concepto: Se dice que la colección

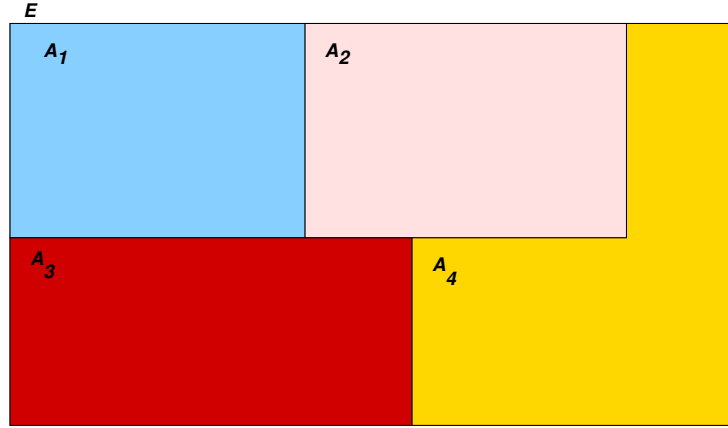


Figura 4.4: A_1, A_2, A_3, A_4 forman un sistema exhaustivo y excluyente de sucesos.

$A_1, A_2, \dots, A_n \subset E$ es un **sistema exhaustivo y excluyente de sucesos** si se verifican las relaciones (véase la figura 4.4):

$$\bigcup_{i=1}^n A_i = E$$

$$A_i \cap A_j = \emptyset \quad \forall i \neq j$$

4.5.3. Teorema de la probabilidad total

Sea $A_1, A_2, \dots, A_n \subset E$ un sistema exhaustivo y excluyente de sucesos. Entonces

$$\boxed{\forall B \subset E, \Rightarrow \mathcal{P}[B] = \sum_{i=1}^n \mathcal{P}[B|A_i] \cdot \mathcal{P}[A_i]} \quad (4.11)$$

Ejemplo de cálculo usando el teorema de la probabilidad total

Se tienen dos urnas, y cada una de ellas contiene un número diferente de bolas blancas y rojas:

- Primera urna, U_1 : 3 bolas blancas y 2 rojas;
- Segunda urna, U_2 : 4 bolas blancas y 2 rojas.

Se realiza el siguiente experimento aleatorio:

Se tira una moneda al aire y si sale cara se elige una bola de la primera urna, y si sale cruz de la segunda.

¿Cuál es la probabilidad de que salga una bola blanca?

Solución: La situación que tenemos puede ser esquematizada como

$$\begin{array}{cc}
 \boxed{\begin{array}{c} 3 \text{ } B \\ 2 \text{ } R \end{array}} & \boxed{\begin{array}{c} 4 \text{ } B \\ 2 \text{ } R \end{array}} \\
 U_1 & U_2 \\
 \mathcal{P}[U_1] = 1/2 & \mathcal{P}[U_2] = 1/2 \\
 \mathcal{P}[B|U_1] = 3/5 & \mathcal{P}[B|U_2] = 4/6
 \end{array}$$

Como U_1 y U_2 forman un sistema incompatible y excluyente de sucesos (la bola resultado debe provenir de una de esas dos urnas y de una sólo de ellas), el teorema de la probabilidad total nos permite afirmar entonces que

$$\mathcal{P}[B] = \mathcal{P}[B|U_1] \cdot \mathcal{P}[U_1] + \mathcal{P}[B|U_2] \cdot \mathcal{P}[U_2] = \frac{3}{5} \cdot \frac{1}{2} + \frac{4}{6} \cdot \frac{1}{2} = \frac{19}{30}$$

4.5.4. Teorema de Bayes

Sea $A_1, A_2, \dots, A_n \subset E$ un sistema exhaustivo y excluyente de sucesos. Sea $B \subset E$ un suceso del que conocemos todas las cantidades $\mathcal{P}[B|A_i]$, $i = 1, \dots, n$, a las que denominamos **verosimilitudes**. entonces se verifica:

$$\forall j = 1, \dots, n, \quad \mathcal{P}[A_j|B] = \frac{\mathcal{P}[B|A_j] \cdot \mathcal{P}[A_j]}{\sum_{i=1}^n \mathcal{P}[B|A_i] \cdot \mathcal{P}[A_i]} \quad (4.12)$$

Ejemplo de cálculo con el teorema de Bayes

Se tienen tres urnas. Cada una de ellas contiene un número diferente de bolas blancas y rojas:

- Primera urna, U_1 : 3 bolas blancas y 2 rojas;
- Segunda urna, U_2 : 4 bolas blancas y 2 rojas;
- Tercera urna, U_3 : 3 bolas rojas.

Se realiza el siguiente experimento aleatorio:

Alguien elije al azar y con la misma probabilidad una de las tres urnas, y saca una bola.

Si el resultado del experimento es que ha salido una bola blanca, ¿cuál es la probabilidad de que provenga de la primera urna? Calcular lo mismo para las otras dos urnas.

Solución:

Vamos a representar en un esquema los datos de que disponemos:

$\begin{array}{c} 3 \text{ } B \\ 2 \text{ } R \end{array}$	$\begin{array}{c} 4 \text{ } B \\ 2 \text{ } R \end{array}$	$\begin{array}{c} 0 \text{ } B \\ 3 \text{ } R \end{array}$
U_1	U_2	U_3
$\mathcal{P}[U_1] = 1/3$ $\mathcal{P}[B U_1] = 3/5$	$\mathcal{P}[U_2] = 1/3$ $\mathcal{P}[B U_2] = 4/6$	$\mathcal{P}[U_3] = 1/3$ $\mathcal{P}[B U_3] = 0$

En este caso U_1 , U_2 y U_3 forman un sistema incompatible y excluyente de sucesos (la bola resultado debe provenir de una de esas tres urnas y de una

sólo de ellas), por tanto es posible aplicar el teorema de Bayes:

$$\begin{aligned}
 \mathcal{P}[U_1|B] &= \frac{\mathcal{P}[B|U_1] \cdot \mathcal{P}[U_1]}{\mathcal{P}[B|U_1] \cdot \mathcal{P}[U_1] + \mathcal{P}[B|U_2] \cdot \mathcal{P}[U_2] + \mathcal{P}[B|U_3] \cdot \mathcal{P}[U_3]} \\
 &= \frac{\frac{3}{5} \cdot \frac{1}{3}}{\frac{3}{5} \cdot \frac{1}{3} + \frac{4}{6} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} \\
 &= \frac{9}{19}
 \end{aligned}$$

Con respecto a las demás urnas hacemos lo mismo:

$$\begin{aligned}
 \mathcal{P}[U_2|B] &= \frac{\mathcal{P}[B|U_2] \cdot \mathcal{P}[U_2]}{\mathcal{P}[B|U_1] \cdot \mathcal{P}[U_1] + \mathcal{P}[B|U_2] \cdot \mathcal{P}[U_2] + \mathcal{P}[B|U_3] \cdot \mathcal{P}[U_3]} \\
 &= \frac{\frac{4}{6} \cdot \frac{1}{3}}{\frac{3}{5} \cdot \frac{1}{3} + \frac{4}{6} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} \\
 &= \frac{10}{19}
 \end{aligned}$$

$$\begin{aligned}
 \mathcal{P}[U_3|B] &= \frac{\mathcal{P}[B|U_3] \cdot \mathcal{P}[U_3]}{\mathcal{P}[B|U_1] \cdot \mathcal{P}[U_1] + \mathcal{P}[B|U_2] \cdot \mathcal{P}[U_2] + \mathcal{P}[B|U_3] \cdot \mathcal{P}[U_3]} \\
 &= \frac{0 \cdot \frac{1}{3}}{\frac{3}{5} \cdot \frac{1}{3} + \frac{4}{6} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} \\
 &= 0
 \end{aligned}$$

Comentario sobre el teorema de Bayes

Obsérvese que en el ejemplo anterior, antes de realizar el experimento aleatorio de extraer una bola para ver su resultado, teníamos que la probabilidad de elegir una urna i cualquiera es $\mathcal{P}[U_i]$. Estas probabilidades se

denominan **probabilidades a priori**. Sin embargo, después de realizar el experimento, y observar que el resultado del mismo ha sido la extracción de una bola blanca, las probabilidades de cada urna han cambiado a $\mathcal{P}[U_i|B]$. Estas cantidades se denominan **probabilidades a posteriori**. Vamos a representar en una tabla la diferencia entre ambas:

a priori	a posteriori		
$\mathcal{P}[U_1] = 1/3$	$\mathcal{P}[U_1 B] = 9/19$	\implies	Las probabilidades <i>a priori</i> cambian de tal modo de las <i>a posteriori</i> que una vez observado el resultado del experimento aleatorio, se puede afirmar con certeza que no fue elegida la tercera urna.
$\mathcal{P}[U_2] = 1/3$	$\mathcal{P}[U_2 B] = 10/19$		
$\mathcal{P}[U_3] = 1/3$	$\mathcal{P}[U_3 B] = 0$		
1	1		

Este fenómeno tiene aplicaciones fundamentales en Ciencia: Cuando se tienen dos teorías científicas diferentes, T_1 y T_2 , que pretenden explicar cierto fenómeno, y a las que asociamos unas probabilidades a priori de ser ciertas,

$$\mathcal{P}[T_1], \mathcal{P}[T_2]$$

podemos llevar a cabo la experimentación que se considere más conveniente, para una vez obtenido el cuerpo de evidencia, B , calcular como se modifican las probabilidades de verosimilitud de cada teoría mediante el teorema de Bayes:

$$\mathcal{P}[T_1|B], \mathcal{P}[T_2|B]$$

Así la experimentación puede hacer que una teoría sea descartada si $\mathcal{P}[T_i|B] \approx 0$ o reforzada si $\mathcal{P}[T_i|B] \approx 1$. Una aplicación básica de esta técnica la tenemos en Medicina para decidir si un paciente padece cierta enfermedad o no, en función de los resultados de un *test diagnóstico*.

4.6. Tests diagnósticos

Los tests diagnósticos son una aplicación del teorema de Bayes a la Medicina, y se basan en lo siguiente:

1. Se sospecha que un paciente puede padecer cierta enfermedad, que tiene una **incidencia de la enfermedad en la población** (probabilidad de que la enfermedad la padezca una persona elegida al azar) de $\mathcal{P}[E]$;
2. Como ayuda al diagnóstico de la enfermedad, se le hace pasar una serie de pruebas (tests), que dan como resultado:
 - Positivo, T^+ , si la evidencia a favor de que el paciente esté enfermo es alta en función de estas pruebas;
 - Negativo, T^- , en caso contrario.

Previamente, sobre el test diagnóstico a utilizar, han debido ser *estimadas* las cantidades:

Sensibilidad: Es la probabilidad de el test de positivo sobre una persona que sabemos que padece la enfermedad, $\mathcal{P}[T^+|E]$.

Especificidad: Es la probabilidad que el test de negativo sobre una persona que no la padece, $\mathcal{P}[T^-|\bar{E}]$.

Lo que interesa saber en la práctica es, predecir si una persona está sana o enferma, a partir del resultado del test diagnóstico, es decir, las cantidades:

Índice predictivo positivo: Es la probabilidad de que un individuo esté enfermo si el test dió resultado positivo, $\mathcal{P}[E|T^+]$.

Especificidad: Es la probabilidad que el test de negativo sobre una persona que no la padece, $\bar{E}|_{\mathcal{P}[T^-]}$.

La sensibilidad y especificidad se denominan también respectivamente **tasa de verdaderos positivos** y **tasa de verdaderos negativos**. Estas cantidades son calculadas de modo aproximado, antes de utilizar el test diagnóstico, considerando grupos suficientemente numerosos de personas de las que sabemos si padecen la enfermedad o no, y estimando los porcentajes correspondientes. Típicamente esta labor es realizada por un laboratorio que quiere probar la eficacia de un test diagnóstico. Los índices predictivos son interesantes sobre todo para el clínico que efectivamente desea evaluar la probabilidad de

que un individuo esté o no enfermo, en función de los resultados de las pruebas que se realizan sobre el mismo.

Ejemplo de cálculo en tests diagnósticos

Se toman 100 personas sanas y 100 enfermas, y se observa que

	E	\bar{E}	
T^+	89	3	Tasa de verdaderos positivos: 89 %
			Tasa de falsos positivos: 3 %
T^-	11	97	Tasa de verdaderos negativos: 97 %
			Tasa de falsos negativos: 11 %
	100	100	

3. teniendo en cuenta el resultado del test diagnóstico, se utiliza el teorema de Bayes para ver cual es, a la vista de los resultados obtenidos, la probabilidad de que realmente esté enfermo si le dio positivo (**índice predictivo de verdaderos positivos**),

$$\mathcal{P}[E|T^+] = \frac{\mathcal{P}[T^+|E] \cdot \mathcal{P}[E]}{\mathcal{P}[T^+|E] \cdot \mathcal{P}[E] + \mathcal{P}[T^+|\bar{E}] \cdot \mathcal{P}[\bar{E}]},$$

o la de que esté sano si le dio negativo (**índice predictivo de verdaderos negativos**):

$$\mathcal{P}[\bar{E}|T^-] = \frac{\mathcal{P}[T^-|\bar{E}] \cdot \mathcal{P}[\bar{E}]}{\mathcal{P}[T^-|\bar{E}] \cdot \mathcal{P}[\bar{E}] + \mathcal{P}[T^-|E] \cdot \mathcal{P}[E]}$$

Otro ejemplo de cálculo con tests diagnósticos

Con el objeto de diagnosticar la colelitiasis se usan los ultrasonidos. Tal técnica tiene una sensibilidad del 91 % y una especificidad del 98 %. En la población que nos ocupa, la probabilidad de colelitiasis es de 0,2.

1. Si a un individuo de tal población se le aplican los ultrasonidos y dan positivos, ¿cuál es la probabilidad de que sufra la colelietasis?
2. Si el resultado fuese negativo, ¿cuál sería la probabilidad de que no tenga la enfermedad?

Solución:

Vamos a utilizar la siguiente notación:

- $E \equiv$ Padecer la *enfermedad* (colelietasis);
- $\bar{E} \equiv$ No padecer la enfermedad;
- $T^+ \equiv$ El resultado del test es positivo;
- $T^- \equiv$ El resultado del test es negativo;

Los datos de que disponemos son las probabilidades condicionadas

$$\text{Sensibilidad o Tasa de Verdaderos Positivos} \equiv \mathcal{P}[T^+|E] = 0,91,$$

$$\text{Especificidad o Tasa de verdaderos Negativos} \equiv \mathcal{P}[T^-|\bar{E}] = 0,98$$

y la incidencia de la enfermedad en la población

$$\mathcal{P}[E] = 0,20$$

En el primer apartado se pide calcular el “Índice Predictivo de Verdaderos Positivos”, $\mathcal{P}[E|T^+]$, que por el teorema de Bayes es:

$$\mathcal{P}[E|T^+] = \frac{\mathcal{P}[T^+|E] \cdot \mathcal{P}[E]}{\mathcal{P}[T^+|E] \cdot \mathcal{P}[E] + \underbrace{\mathcal{P}[T^+|\bar{E}]}_{1-\mathcal{P}[T^-|\bar{E}]} \cdot \underbrace{\mathcal{P}[\bar{E}]}_{1-\mathcal{P}[E]}} = \frac{0,91 \cdot 0,2}{0,91 \cdot 0,2 + 0,02 \cdot 0,8} = 0,9192$$

En el segundo apartado, se ha de calcular el “Índice Predictivo de Verdaderos Negativos”, $\mathcal{P}[\bar{E}|T^-]$,

$$\mathcal{P}[\bar{E}|T^-] = \frac{\mathcal{P}[T^-|\bar{E}] \cdot \mathcal{P}[\bar{E}]}{\mathcal{P}[T^-|\bar{E}] \cdot \mathcal{P}[\bar{E}] + \underbrace{\mathcal{P}[T^-|E] \cdot \mathcal{P}[E]}_{1-\mathcal{P}[T^+|E]}} = \frac{0,98 \cdot 0,8}{0,98 \cdot 0,8 + 0,09 \cdot 0,2} = 0,9775$$

4.7. Problemas

Ejercicio 4.1. Una mujer portadora de hemofilia clásica da a luz tres hijos.

1. ¿Cual es la probabilidad de que de los tres hijos, ninguno esté afectado por la enfermedad?
2. ¿Cual es la probabilidad de que exactamente dos de los tres niños esté afectado?

Ejercicio 4.2. El 60 % de los individuos de una población están vacunados contra una cierta enfermedad. Durante una epidemia se sabe que el 20 % la ha contraído y que 2 de cada 100 individuos están vacunados y son enfermos. Calcular el porcentaje de vacunados que enferma y el de vacunados entre los que están enfermos..

Ejercicio 4.3. La proporción de alcoholicos que existe en la población de Málaga es, aproximadamente, un 10 %; no obstante, en las bajas que dan los médicos de la Seguridad Social difícilmente se encuentra el diagnóstico de alcoholismo. Aparecen sin embargo diagnosticados de hepatopatías, lumbalgias, etc., que pueden hacer sospechar alcoholismo subyacente. Se realizó un estudio que puso de manifiesto que el 85 % de los individuos alcoholicos y el 7 % de los no alcoholicos sufrían tales patologías. Se desea saber cuál es la probabilidad de que un individuo con esas patologías sea realmente alcoholico.

Ejercicio 4.4. Dos tratamientos A y B curan una determinada enfermedad en el 20 % y 30 % de los casos, respectivamente. Suponiendo que ambos actúan de modo independiente, cuál de las dos siguientes estrategias utilizaría para curar a un individuo con tal enfermedad:

1. Aplicar ambos tratamientos a la vez.
2. Aplicar primero el tratamiento B y, si no surte efecto, aplicar el A.

Ejercicio 4.5. Se eligen al azar 3 deportistas de un equipo de 10 integrantes para realizar un control antidopaje; Se sabe que 2 de los jugadores del equipo han tomado sustancias prohibidas. ¿Cuál es la probabilidad de elegir para el análisis a alguno de los infractores?

Ejercicio 4.6. Estamos interesados en saber cuál de dos análisis A y B es mejor para el diagnóstico de una determinada enfermedad, de la cual sabemos que la presentan un 10 % de individuos de la población. El porcentaje de resultados falsos positivos del análisis A es del 15 % y el de B es del 22 %. El porcentaje de falsos negativos de A es del 7 % y de B es del 3 %. ¿Cuál es la probabilidad de acertar en el diagnóstico con cada método?

Ejercicio 4.7. Con objeto de diagnosticar la colelitiasis se usan los ultrasonidos. Tal técnica tiene una sensibilidad del 91 % y una especificidad del 98 %. En la población que nos ocupa la probabilidad de colelitiasis es del 20 %.

1. Si a un individuo de tal población se le aplican los ultrasonidos y dan positivos, ¿cuál es la probabilidad de que sufra la colelitiasis?
2. Si el resultado fuese negativo, ¿cuál es la probabilidad de que no tenga la enfermedad?

Ejercicio 4.8. Entre los estudiantes de una Facultad de Filosofía y Letras se dan las siguientes proporciones: el 40 % son hombres. El 70 % de los

varones fuman, mientras que entre las mujeres sólo fuman el 20 %. Escogido un estudiante al azar, calcúlese la probabilidad de que fume.

Ejercicio 4.9. Los estudios epidemiológicos indican que el 20 % de los ancianos sufren un deterioro neuropsicológico. Sabemos que la tomografía axial computerizada (TAC) es capaz de detectar este trastorno en el 80 % de los que lo sufren, pero que también da un 3 % de falsos positivos entre personas sanas. Si tomamos un anciano al azar y da positivo en el TAC, ¿cuál es la probabilidad de que esté realmente enfermo?

Ejercicio 4.10. Sabemos que tiene estudios superiores el 15 % de la población española, estudios medios el 40 %, estudios primarios el 35 % y no tiene estudios el 10 %. Los desempleados no se distribuyen proporcionalmente entre esas categorías, dado que de entre los de estudios superiores están sin trabajo el 10 %, entre los de estudios medios el 35 %, entre los de estudios primarios el 18 %, y entre los que no tienen estudios el 37 %. Obtenga las probabilidades de que extraído uno al azar, éste sea:

1. Titulado superior, sabiendo que está parado.
2. Un sujeto sin estudios que está en paro.
3. Un sujeto con estudios primarios o que está trabajando.

Ejercicio 4.11. Una enfermedad puede estar producida por tres virus A, B, y C. En el laboratorio hay 3 tubos de ensayo con el virus A, 2 tubos con el virus B y 5 tubos con el virus C. La probabilidad de que el virus A produzca la enfermedad es de $1/3$, que la produzca B es de $2/3$ y que la produzca el virus C es de $1/7$. Se inocula un virus a un animal y contrae la enfermedad. ¿Cuál es la probabilidad de que el virus que se inocule sea el C?

Ejercicio 4.12. El 70 % de los estudiantes aprueba una asignatura A y un 60 % aprueba otra asignatura B. Sabemos, además, que un 35 % del total

aprueba ambas. Elegido un estudiante al azar, calcular las probabilidades de las siguientes situaciones:

1. Haya aprobado la asignatura B, sabiendo que ha aprobado la A.
2. Haya aprobado la asignatura B, sabiendo que no ha aprobado la A.
3. No haya aprobado la asignatura B, sabiendo que ha aprobado la A.
4. No haya aprobado la asignatura B, sabiendo que no ha aprobado la A.

Ejercicio 4.13. La cuarta parte de los conductores de coche son mujeres. La probabilidad de que una mujer sufra un accidente en un año es de $5/10.000$, y para los hombres es de $1/10.000$. Calcúlese la probabilidad de que si acaece un accidente, el accidentado sea hombre.

Ejercicio 4.14. En un campus universitario existen 3 carreras sanitarias. Se sabe que el 50 % cursan estudios de Enfermería, el 30 % Medicina y el 20 % Veterinaria. Los que finalizaron sus estudios son el 20, 10 y 5 % respectivamente. Elegido un estudiante al azar, hállese la probabilidad de que haya acabado la carrera.

Capítulo 5

Variables aleatorias

5.1. Introducción

Normalmente, los resultados posibles (*espacio muestral* E) de un experimento aleatorio no son valores numéricos. Por ejemplo, si el experimento consiste en lanzar de modo ordenado tres monedas al aire, para observar el número de caras (\mathcal{C}) y cruces (\mathcal{R}) que se obtienen, el espacio muestral asociado a dicho experimento aleatorio sería:

$$E = \{CCC, CCR, CRC, CRR, RCC, RCR, RRC, RRR\}$$

En estadística resulta más fácil utilizar valores numéricos en lugar de trabajar directamente con los elementos de un espacio muestral como el anterior. Así preferimos identificar los sucesos $\{CRR, RCR, RRC\}$ con el valor numérico 1 que representa el *número de caras obtenidas al realizar el experimento*. De este modo aparece el concepto de **variable aleatoria unidimensional** como el de toda función

$$\begin{aligned} X : E &\longrightarrow \mathbb{R} \\ e &\longmapsto X(e) = x_e \end{aligned}$$

que atribuye un único número real x_e , a cada suceso elemental e , del espacio muestral E

Por ejemplo, en el ejemplo anterior, se define la variable aleatoria (v.a. en adelante)

$$X \equiv \text{número de caras}$$

del siguiente modo:

$$X : E \longrightarrow \mathbb{R}$$

$$X(CCC) = 3$$

$$X(CCR) = X(CRC) = X(RCC) = 2$$

$$X(RRC) = X(RCR) = X(CRR) = 1$$

$$X(RRR) = 0$$

En función de los valores que tome la variable, esta puede ser clasificada en discreta o continua del siguiente modo:

v.a. discreta es aquella que sólo puede tomar un número finito o infinito numerable de valores. Por ejemplo,

$$X : E \longrightarrow \mathbb{N}$$

v.a. continua es la que puede tomar un número infinito no numerable de valores.

$$X : E \longrightarrow \mathbb{R}$$

Vamos a estudiar los conceptos más importantes relacionados con la distribución de probabilidad de una v.a., diferenciando entre los casos de v.a. discreta y v.a. continua.

5.2. Variables aleatorias discretas

Dada una v.a. discreta $X : E \longrightarrow \mathbb{N}$, su **función de probabilidad** f , se define de modo que $f(x_i)$ es la probabilidad de que X tome ese valor:

$$f : \mathbb{N} \longrightarrow [0, 1]$$

$$x_i \longmapsto f(x_i) = \mathcal{P}[X = x_i] = \mathcal{P}[\{e, \text{ t.q. } X(e) = x_i\}]$$

Si x_i no es uno de los valores que puede tomar X , entonces $f(x_i) = 0$. La representación gráfica de la función de probabilidad se realiza mediante un diagrama de barras análogo al de distribución de frecuencias relativas para variables discretas. Por ejemplo, si retomamos el caso del lanzamiento de 3 monedas de forma que cada una de ellas tenga probabilidad $1/2$ de dar como resultado cara o cruz, se tiene que (véase la figura 5.1):

$$f(3) = \mathcal{P}[X = 3] = \mathcal{P}[\{CCC\}] = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

$$f(2) = \mathcal{P}[X = 2] = \mathcal{P}[\{RCC, CCR, CRC\}] = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

$$f(1) = \mathcal{P}[X = 1] = \mathcal{P}[\{RRC, RCR, CRR\}] = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

$$f(0) = \mathcal{P}[X = 0] = \mathcal{P}[\{RRR\}] = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

Otro concepto importante es el de **función de distribución** de una variable aleatoria discreta, F , que se define de modo que si $x_i \in \mathbb{R}$, $F(x_i)$ es igual a la probabilidad de que X tome un valor inferior o igual a x_i :

$$F : \mathbb{N} \longrightarrow [0, 1]$$

$$x_i \longmapsto F(x_i) = \mathcal{P}[X \leq x_i] = \mathcal{P}[\{e, \text{ t.q. } X(e) \leq x_i\}]$$

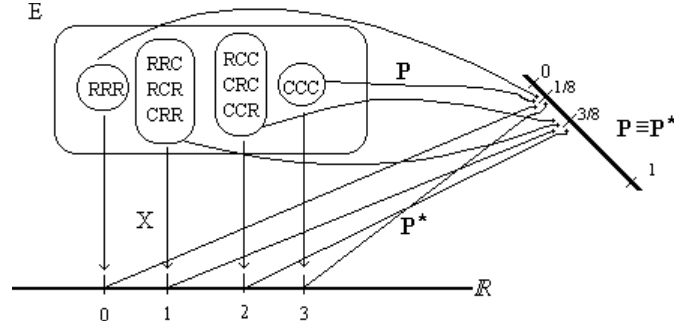


Figura 5.1: Equivalencia entre las probabilidades calculadas directamente sobre el espacio muestral E de resultados del experimento aleatorio, y las calculadas sobre el subconjunto $\{0, 1, 2, 3\} \subset \mathbb{N} \subset \mathbb{R}$ mediante la v.a. X .

Esta función se representa gráficamente del mismo modo que la distribución de frecuencias relativas acumuladas (figura 5.2). Volviendo al ejemplo de las tres monedas, se tiene que

$$F(0) = \mathcal{P}[X \leq 0] = \mathcal{P}[X = 0] = f(0) = \frac{1}{8}$$

$$F(1) = \mathcal{P}[X \leq 1] = f(0) + f(1) = \frac{1}{8} + \frac{3}{8} = \frac{4}{8}$$

$$F(2) = \mathcal{P}[X \leq 2] = f(0) + f(1) + f(2) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}$$

$$F(3) = \mathcal{P}[X \leq 3] = f(0) + f(1) + f(2) + f(3) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = \frac{8}{8} = 1$$

5.3. Variables aleatorias continuas

Si una variable discreta toma los valores x_1, \dots, x_k , la probabilidad de que al hacer un experimento, X tome uno de esos valores es 1, de modo que cada posible valor x_i contribuye con una cantidad $f(x_i)$ al total:

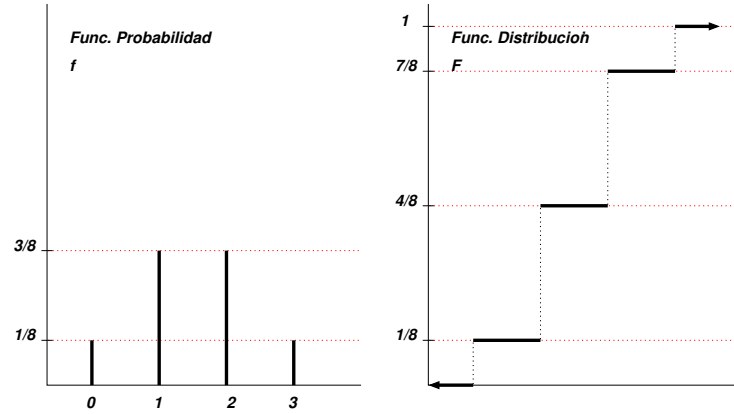


Figura 5.2: Función de probabilidad a la izquierda, y función de distribución a la derecha de una v.a. discreta

$$\sum_{i=1}^k f(x_i) = \sum_{i=1}^k \mathcal{P}[X = x_i] = 1$$

Aun cuando la variable tomase un número infinito de valores, x_1, x_2, \dots , no hay ningún problema en comprobar que cada x_i contribuye con una cantidad $f(x_i)$ al total de modo que

$$\sum_{i=1}^{\infty} f(x_i) = \sum_{i=1}^{\infty} \mathcal{P}[X = x_i] = 1$$

Cuando la variable es continua, no tiene sentido hacer una suma de las probabilidades de cada uno de los términos en el sentido anterior, ya que el conjunto de valores que puede tomar la variable es *no numerable*. En este caso, lo que generaliza de modo natural el concepto de suma (\sum) es el de integral (\int). Por otro lado, para variables continuas no tiene interés hablar de la probabilidad de que $X = x \in \mathbb{R}$, ya que esta debe de valer siempre 0, para que la *suma infinita no numerable* de las probabilidades de todos los valores de la variable no sea infinita.

De este modo es necesario introducir un nuevo concepto que sustituya en v.a. continuas, al de función de probabilidad de una v.a. discreta. Este concepto es el de **función de densidad de una v.a. continua**, que se define como una función $f : \mathbb{R} \longrightarrow \mathbb{R}$ integrable, que verifica las dos propiedades siguientes:

$$\begin{cases} f(x) \geq 0 \\ \int_{-\infty}^{+\infty} f(x) dx = 1 \end{cases} \quad (5.1)$$

y que además verifica que dado $a < b$, se tiene que

$$\mathcal{P}[a \leq X \leq b] = \int_a^b f(x) dx \quad (5.2)$$

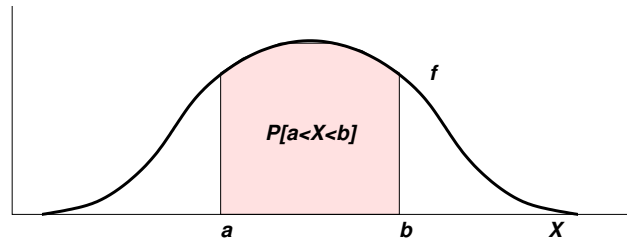


Figura 5.3: Función de densidad f . La probabilidad de un intervalo, es el área que existe entre la función y el eje de abscisas.

La **función de distribución de la v.a. continua**, F , se define de modo que dado $x \in \mathbb{R}$, $F(x)$ es la probabilidad de que X sea menor o igual que x , es decir

$$\begin{aligned} F : \mathbb{R} &\longrightarrow [0, 1] \\ x &\longmapsto F(x) = \mathcal{P}[X \leq x] = \int_{-\infty}^x f(t) dt \end{aligned} \quad (5.3)$$

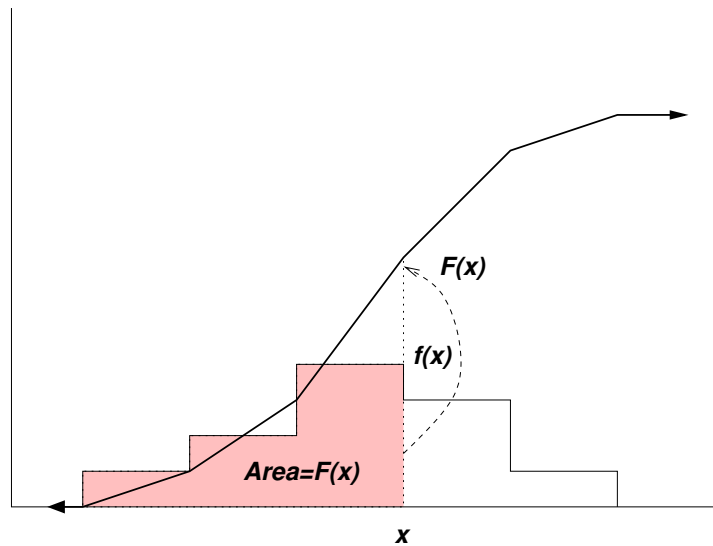


Figura 5.4: Función de distribución F , calculada a partir de la función de densidad f .

5.4. Medidas de tendencia central y dispersión de v.a.

De forma análoga a lo que se se hizo en el capítulo 2 sobre *estadística descriptiva* podemos definir para *variables aleatorias* medidas de centralización, dispersión, simetría y forma. Por su interés nos vamos a centrar en dos medidas sobre v.a. que son la esperanza matemática que desempeña un papel equivalente al de la *media* y el *momento central de segundo orden*, también denominado *varianza*.

5.4.1. Valor esperado o esperanza matemática

La **esperanza matemática** o **valor esperado** de una variable aleatoria es el concepto equivalente al de media aritmética.

Como las variables aleatorias se expresan de modo diferente en el caso discreto que en el continuo, tratemos a cada una de ellas por separado.

Sea X una v.a. *discreta*. Se denomina **esperanza matemática** de X o **valor esperado**, y se denota bien $\mathbf{E}[X]$ o bien μ , a la cantidad que se expresa como:

$$\mathbf{E}[X] = \sum_{i \in \mathcal{I}} x_i f(x_i) \quad (5.4)$$

donde \mathcal{I} es el conjunto numerable de índices de los valores que puede tomar la variable (por ejemplo $\mathcal{I} = \{1, 2, \dots, k\}$ para un número *finito* de valores de la v.a. o bien $\mathcal{I} = \mathcal{N}$ para una cantidad *infinita numerable* de los mismos.

Si X es una v.a. *continua*, se define su esperanza a partir de la función de densidad como sigue:

$$\mathbf{E}[X] = \int_{-\infty}^{+\infty} x \cdot f(x) dx \quad (5.5)$$

5.4.2. Varianza

La **varianza** la denotamos mediante $\mathbf{Var}[X]$ o bien σ^2 :

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \begin{cases} \sum_{i \in \mathcal{I}} (x_i - \mathbf{E}[X])^2 f(x_i) & \text{si } X \text{ disc.} \\ \int_{-\infty}^{+\infty} (x - \mathbf{E}[X])^2 \cdot f(x) dx & \text{si } X \text{ cont.} \end{cases}$$

Capítulo 6

Principales leyes de distribución de variables aleatorias

6.1. Introducción

Como complemento al capítulo anterior en el que definimos todos los conceptos relativos a variables aleatorias, describimos en éste las principales leyes de probabilidad que encontramos en las aplicaciones del cálculo de probabilidades. Atendiendo a la clasificación de las v.a. en discretas y continuas describiremos las principales *leyes de probabilidad* de cada una de ellas, las cuales constituirán el soporte subyacente de la inferencia estadística y a las que será necesario hacer referencia en el estudio de dicho bloque. Iniciamos este capítulo con el estudio de las distribuciones para v.a. discretas.

6.2. Distribuciones discretas

6.2.1. Distribución de Bernoulli

Consiste en realizar un experimento aleatorio una sólo vez y observar si cierto suceso ocurre o no, siendo p la probabilidad de que esto sea así (*éxito*) y $q = 1 - p$ el que no lo sea (*fracaso*). En realidad no se trata más que de una variable *dicotómica*, es decir que únicamente puede tomar dos modalidades, es por ello que el hecho de llamar éxito o fracaso a los posibles resultados de las pruebas obedece más una tradición literaria o histórica, en el estudio de las v.a., que a la situación real que pueda derivarse del resultado. Podríamos por tanto definir este experimento mediante una v.a. discreta X que toma los valores $X = 0$ si el suceso no ocurre, y $X = 1$ en caso contrario, y que se denota $X \sim \mathbf{Ber}(p)$

$$X \sim \mathbf{Ber}(p) \iff X = \begin{cases} 0 & \longrightarrow q = 1 - p = \mathcal{P}[X = 0] \\ 1 & \longrightarrow p = \mathcal{P}[X = 1] \end{cases} \quad (6.1)$$

Un ejemplo típico de este tipo de variables aleatorias consiste en lanzar una moneda al aire y considerar la v.a.

$$X \equiv \text{número de caras obtenidas} = \begin{cases} 0 & \longrightarrow q = \frac{1}{2} \\ 1 & \longrightarrow p = \frac{1}{2} \end{cases}$$

Para una v.a. de Bernoulli, tenemos que su función de probabilidad es:

$$f(x) = \begin{cases} q & \text{si } x = 0 \\ p & \text{si } x = 1 \\ 0 & \text{en cualquier otro caso;} \end{cases}$$

Los principales momentos de X son:

$$\mathbf{E}[X] = p \quad (6.2)$$

$$\mathbf{Var}[X] = p \cdot q \quad (6.3)$$

6.2.2. Distribución binomial

Se dice que una v.a. X sigue una **ley binomial** de parámetros n y p , $X \rightsquigarrow \mathbf{B}(n, p)$, si es la suma de n v.a. independientes de Bernoulli con el mismo parámetro, p :

$$X \rightsquigarrow \mathbf{B}(n, p) \iff X = X_1 + \dots + X_n, \quad \text{donde } X_i \rightsquigarrow \mathbf{Ber}(p), \forall i = 1, \dots, n \quad (6.4)$$

Esta definición puede interpretarse en el siguiente sentido: Supongamos que realizamos n pruebas de Bernoulli, X_i , donde en todas ellas, la probabilidad de éxito es la misma (p), y queremos calcular el número de éxitos, X , obtenidos el el total de las n pruebas. Su ley de probabilidad es¹ En la Figura 6.1 se representa la función de probabilidad de una variable binomial.

$$f(k) = P[X = k] = \binom{n}{k} p^k q^{n-k} \quad \forall k = 0, 1, \dots, n \quad (6.5)$$

El valor esperado y la varianza de esta variable son:

$$\mathbf{E}[X] = np$$

$$\mathbf{Var}[X] = npq$$

Ejemplo de uso de la distribución binomial

Un médico aplica un test a 10 alumnos de un colegio para detectar una enfermedad cuya incidencia sobre una población de niños es del 10 %.

¹Los valores $f(k)$ los podemos encontrar tabulados para ciertos valores pequeños de n , y ciertos valores usuales de p en la tabla 1 (al final del libro).

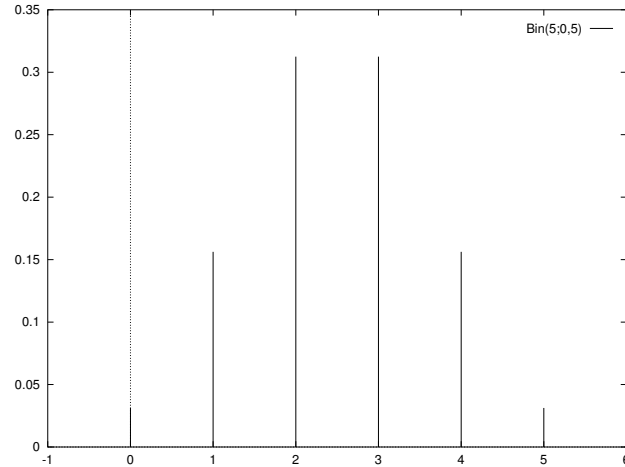


Figura 6.1: Función de probabilidad de una variable binomial cuando n es pequeño.

La sensibilidad del test es del 80 % y la especificidad del 75 %. ¿Cuál es la probabilidad de que exactamente a cuatro personas le de un resultado positivo? Si en la muestra hay cuatro personas a las que el test le da positivo, ¿cuál es la probabilidad de que entre estas, exactamente dos estén sanas? Calcular la probabilidad de que el test suministre un resultado incorrecto para dos personas. Calcular la probabilidad de que el resultado sea correcto para más de 7 personas.

Solución:

Los datos de que disponemos son:

$$\begin{aligned}
 \mathcal{P}[E] &= 0,1 && \underbrace{\text{prevalencia de la enfermedad en la población}}_{\text{Probabilidad } a \text{ priori de estar enfermo}} \\
 \mathcal{P}[T^+|E] &= 0,8 && \text{sensibilidad (verdaderos positivos)} \\
 \mathcal{P}[T^-|\bar{E}] &= 0,75 && \text{especificidad (verdaderos negativos)} \quad (6.6)
 \end{aligned}$$

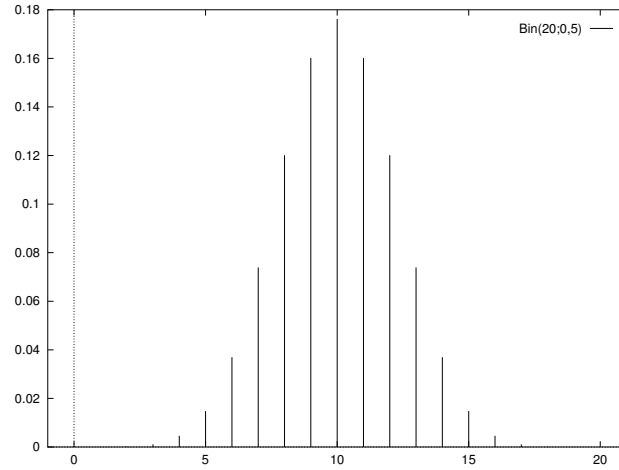


Figura 6.2: Función de probabilidad de una variable binomial cuando n es grande.

donde E , T^+ , y T^- tienen el sentido que es obvio. Si queremos saber a cuantas personas el test le dará un resultado positivo, tendremos que calcular $\mathcal{P}[T^+]$, para lo que podemos usar el teorema de la probabilidad total (estar enfermo y no estarlo forman una colección exhaustiva y excluyente de sucesos):

$$\begin{aligned}
 \mathcal{P}[T^+] &= \mathcal{P}[T^+|E] \cdot \mathcal{P}[E] + \underbrace{\mathcal{P}[T^+|\bar{E}]}_{1-\mathcal{P}[T^-|\bar{E}]} \cdot \underbrace{\mathcal{P}[\bar{E}]}_{1-\mathcal{P}[E]} \\
 &= 0,8 \times 0,1 + 0,25 \times 0,9 = 0,305
 \end{aligned}$$

Sea X_1 la v.a. que contabiliza el número de resultados positivos. Es claro que llamando $p_1 = \mathcal{P}[T^+]$, se tiene que X sigue una distribución binomial

$$X_1 \sim \mathbf{B}(n_1 = 10, p_1 = 0,305) \iff \mathcal{P}[X_1 = k] = \binom{n_1}{k} p_1^k q_1^{n_1-k}$$

Por ello la probabilidad de que a cuatro personas le de el resultado del test positivo es:

$$\mathcal{P}[X_1 = 4] = \binom{10}{4} 0,305^4 \cdot 0,695^6 = 0,2048$$

Si queremos calcular a cuantas personas les dará el test un resultado positivo aunque en realidad estén sanas, hemos de calcular previamente $\mathcal{P}[\bar{E}|T^+]$, o sea, el índice predictivo de falsos positivos:

$$\mathcal{P}[\bar{E}|T^+] = \frac{\mathcal{P}[\bar{E} \cap T^+]}{\mathcal{P}[T^+]} = \frac{\overbrace{\mathcal{P}[T^+|\bar{E}]}^{1-\mathcal{P}[T^-|\bar{E}]} \cdot \overbrace{\mathcal{P}[\bar{E}]}^{1-\mathcal{P}[E]}}{\mathcal{P}[T^+]} = 0,7377$$

Es importante observar este resultado. Antes de hacer los cálculos no era previsible que si a una persona el test le da positivo, en realidad tiene una probabilidad aproximadamente del 74 % de estar sana. Sea X_2 la variable aleatoria que contabiliza al número de personas al que el test le da positivo, pero que están sanas en realidad. Entonces

$$X_2 \rightsquigarrow \mathbf{B}(n_2 = 4, p_2 = 0,7377) \iff \mathcal{P}[X_2 = k] = \binom{n_2}{k} p_2^k q_2^{n_2-k}$$

y

$$\mathcal{P}[X_2 = 2] = \binom{4}{2} 0,7377^2 \cdot 0,2623^2 = 0,22465$$

Por último vamos a calcular la probabilidad p_3 de que el test de un resultado erróneo, que es:

$$p_3 = \underbrace{\mathcal{P}[(T^+ \cap \bar{E}) \cup (T^- \cap E)]}_{\text{incompatibles}}$$

$$\begin{aligned}
&= \mathcal{P}[T^+ \cap \overline{E}] + \mathcal{P}[T^- \cap E] \\
&= \mathcal{P}[T^+_{|E}] \cdot \mathcal{P}[\overline{E}] + \mathcal{P}[T^-_{|E}] \cdot \mathcal{P}[E] \\
&= 0,25 \times 0,9 + 0,2 \times 0,1 = 0,245
\end{aligned}$$

La variable aleatoria que contabiliza el número de resultados erróneos del test es

$$X_3 \rightsquigarrow \mathbf{B}(n_3 = 10, p_3 = 0,245) \iff \mathcal{P}[X_3 = k] = \binom{n_3}{k} p_3^k q_3^{n_3-k}$$

Como la probabilidad de que el test sea correcto para más de siete personas, es la de que sea incorrecto para menos de 3, se tiene

$$\begin{aligned}
\mathcal{P}[X_3 < 3] &= \underbrace{\mathcal{P}[X_3 \leq 2]}_{F_{X_3}(2)} \\
&= \sum_{i=0}^2 \binom{n_3}{i} p_3^i q_3^{n_3-i} \\
&= \binom{10}{0} 0,245^0 \cdot 0,755^{10} + \binom{10}{1} 0,245^1 \times 0,755^9 + \binom{10}{2} 0,245^2 \times 0,755^8 \\
&= 0,5407
\end{aligned}$$

6.2.3. Distribución geométrica (o de fracasos)

Consideramos una sucesión de v.a. independientes de Bernoulli,

$$X_1, X_2, \dots, X_i, \dots \quad \text{donde } X_i \rightsquigarrow \mathbf{Ber}(p), \quad i = 1, 2, \dots, \infty$$

Una v.a. X sigue posee una **distribución geométrica**, $X \rightsquigarrow \mathbf{Geo}(p)$, si esta es la *suma del número de fracasos obtenidos hasta la aparición del primer éxito* en la sucesión $\{X_i\}_{i=1}^{\infty}$. Por ejemplo

X_1	X_2	X_3	X_4	X_5	\dots		X
\downarrow	\downarrow	\downarrow	\downarrow	\downarrow			\downarrow
1	0	0	1	1	$\dots \implies$	$X = 0$	$f(0) = p$
<u>0</u>	1	0	1	1	$\dots \implies$	$X = 1$	$f(1) = qp$
<u>0</u>	<u>0</u>	1	0	1	$\dots \implies$	$X = 2$	$f(2) = qqp$
<u>0</u>	<u>0</u>	<u>0</u>	1	1	$\dots \implies$	$X = 3$	$f(3) = qqqp$
					\dots		

De este modo tenemos que la ley de probabilidad de X es

$$\boxed{f(k) = \mathcal{P}[X = k] = pq^k, \quad k = 0, 1, 2, \dots, \infty} \quad (6.7)$$

La media y varianza de esta variable aleatoria son:

$$\mathbf{E}[X] = \frac{q}{p}$$

$$\mathbf{Var}[X] = \frac{q}{p^2}$$

Ejemplo de uso de la distribución geométrica

Un matrimonio quiere tener una hija, y por ello deciden tener hijos hasta el nacimiento de una hija. Calcular el número esperado de hijos (entre varones y hembras) que tendrá el matrimonio. Calcular la probabilidad de que la pareja acabe teniendo tres hijos o más.

Solución: Este es un ejemplo de variable geométrica. Vamos a suponer que la probabilidad de tener un hijo varón es la misma que la de tener una hija hembra. Sea X la v.a.

$X =$ número de hijos varones antes de nacer la niña

Es claro que

$$X \rightsquigarrow \mathbf{Geo} \left(p = \frac{1}{2} \right) \iff \mathcal{P}[X = k] = q^{k-1} \cdot p = \frac{1}{2^k}$$

Sabemos que el número esperado de hijos varones es $\mathbf{E}[X] = \frac{q}{p} = 1$, por tanto el número esperado en total entre hijos varones y la niña es 2.

La probabilidad de que la pareja acabe teniendo tres o más hijos, es la de que tenga 2 o más hijos varones (la niña está del tercer lugar en adelante), es decir,

$$\begin{aligned} \mathcal{P}[X \geq 2] &= 1 - \overbrace{\mathcal{P}[X < 2]}^{X \text{ discr.}} \\ &= 1 - \mathcal{P}[X \leq 1] \\ &= 1 - \mathcal{P}[X = 0] - \mathcal{P}[X = 1] = 1 - p - qp = \frac{1}{4} \end{aligned}$$

Hemos preferido calcular la probabilidad pedida mediante el suceso complementario, ya que sería más complicado hacerlo mediante la suma infinita

$$\mathcal{P}[X \geq 2] = \sum_{i=2}^{\infty} q^i p.$$

6.2.4. Distribución binomial negativa

Sobre una sucesión de v.a. de Bernoulli independientes,

$$X_1, X_2, \dots, X_i, \dots \quad \text{donde } X_i \rightsquigarrow \mathbf{Ber}(p), \quad i = 1, 2, \dots, \infty$$

se define la v.a. X como el *número de fracasos obtenidos hasta la aparición de r éxitos* en la sucesión $\{X_i\}_{i=1}^{\infty}$. En este caso se dice que X sigue una **ley de distribución binomial negativa** de parámetros r y p y se denota del modo: $X \rightsquigarrow \mathbf{Bn}(r, p)$. Su ley de probabilidad es:

$$f(k) = \mathcal{P}[X = k] = \underbrace{\binom{k+r-1}{r-1} p^{r-1} q^k}_{\substack{k+r-1 \\ \text{primeros experimentos}}} \cdot \underbrace{p}_{\text{éxito final}} = \binom{k+r-1}{k} p^r q^k \quad (6.8)$$

$$\mathbf{E}[X] = \frac{r q}{p} \quad (6.9)$$

$$\mathbf{Var}[X] = \frac{r q}{p^2} \quad (6.10)$$

Ejemplo de variable binomial negativa

Para tratar a un paciente de una afección de pulmón han de ser operados en operaciones independientes sus 5 lóbulos pulmonares. La técnica a utilizar es tal que si todo va bien, lo que ocurre con probabilidad de 7/11, el lóbulo queda definitivamente sano, pero si no es así se deberá esperar el tiempo suficiente para intentarlo posteriormente de nuevo. Se practicará la cirugía hasta que 4 de sus 5 lóbulos funcionen correctamente. ¿Cuál es el valor esperado de intervenciones que se espera que deba padecer el paciente? ¿Cuál es la probabilidad de que se necesiten 10 intervenciones?

Solución: Este es un ejemplo claro de experimento aleatorio regido por una ley binomial negativa, ya que se realizan intervenciones hasta que se obtengan 4 lóbulos sanos, y éste es el criterio que se utiliza para detener el proceso. Identificando los parámetros se tiene:

$X =$ número de operaciones hasta obtener $r = 4$ con resultado positivo

$$X \rightsquigarrow \mathbf{Bn} \left(r = 4, p = \frac{7}{11} \right) \iff \mathcal{P}[X = k] = \binom{k+r-1}{k} q^k p^r$$

Lo que nos interesa es medir el número de intervenciones, Y , más que el número de éxitos hasta el r -ésimo fracaso. La relación entre ambas v.a.

es muy simple:

$$Y = X + r$$

Luego

$$\mathbf{E}[Y] = \mathbf{E}[X + r] = \mathbf{E}[X] + r = \frac{rp}{q} + r = \frac{4 \cdot 7/11}{4/11} + 4 = 11$$

Luego el número esperado de intervenciones que deberá sufrir el paciente es de 11. La probabilidad de que el número de intervenciones sea $Y = 10$, es la de que $X = 10 - 4 = 6$. Por tanto:

$$\mathcal{P}[Y = 10] = \mathcal{P}[X = 6] = \binom{6+4-1}{6} q^6 p^4 = 84 \cdot \left(\frac{4}{11}\right)^6 \left(\frac{7}{11}\right)^4 = 0,03185$$

6.2.5. Distribución hipergeométrica

Por claridad, consideremos el siguiente ejemplo: Tenemos una baraja de cartas españolas ($N = 40$ naipes), de las cuales nos vamos a interesar en el *palo deoros* ($D = 10$ naipes de un mismo tipo). Supongamos que de esa baraja extraemos $n = 8$ cartas de una vez (*sin reemplazamiento*) y se nos plantea el problema de calcular la probabilidad de que hayan $k = 2$ oros (*exactamente*) en esa extracción. La respuesta a este problema es

$$\begin{aligned} \mathcal{P}_{rob}[2 \text{ oros en un grupo de } 8 \text{ cartas}] &= \frac{\text{casos favorables}}{\text{casos posibles}} \\ &= \frac{\begin{array}{c} 2 \text{ naipes} \\ \text{entre los oros} \end{array} \times \begin{array}{c} 6 \text{ naipes} \\ \text{de otros palos} \end{array}}{\begin{array}{c} 8 \text{ naipes} \\ \text{cualesquiera} \end{array}} \\ &= \frac{\binom{10}{2} \cdot \binom{30}{6}}{\binom{40}{8}} = \frac{\binom{D}{k} \cdot \binom{N-D}{n-k}}{\binom{N}{n}} \end{aligned}$$

En lugar de usar como dato D es posible que tengamos la proporción existente, p , entre el número total de oros y el número de cartas de la baraja

$$p = \frac{D}{N} = \frac{10}{40} = \frac{1}{4} \implies \begin{cases} D = N \cdot p \\ N - D = N \cdot q \end{cases} \quad (q = 1 - p)$$

de modo que podemos decir que

$$\mathcal{P}_{rob}[k \text{ oros en un grupo de } n \text{ cartas}] = \frac{\binom{N \cdot p}{k} \cdot \binom{N \cdot q}{n - k}}{\binom{N}{n}}$$

Este ejemplo sirve para representar el tipo de fenómenos que siguen una ley de distribución hipergeométrica. Diremos en general que una v.a. X sigue una **distribución hipergeométrica** de parámetros, N , n y p , lo que representamos del modo $X \sim \mathbf{HGeo}(N, n, p)$, si su función de probabilidad es

$$\mathcal{P}[X = k] = \frac{\binom{N \cdot p}{k} \cdot \binom{N \cdot q}{n - k}}{\binom{N}{n}} \quad \text{si } \max\{0, n - Nq\} \leq k \leq \min\{n, Np\}$$

(6.11)

Cuando el tamaño de la población (N) es muy grande, la ley hipergeométrica tiende a aproximarse a la binomial:

$$\mathbf{HGeo}(N, n, p) \xrightarrow{N \rightarrow \infty} \mathbf{B}(n, p)$$

El valor esperado de la hipergeométrica es el mismo que el de la binomial,

$$\mathbf{E}[X] = np$$

sin embargo su varianza

$$\mathbf{Var}[X] = npq \cdot \frac{N-n}{N-1}$$

no es exactamente la de la binomial, pues está corregida por un factor, $\frac{N-n}{N-1}$, que tiende a 1 cuando $N \rightarrow \infty$. A este factor se le denomina *factor de corrección para población finita*.

6.2.6. Distribución de Poisson o de los sucesos raros

Una v.a. X posee una **ley de distribución de probabilidades del tipo Poisson** cuando

$$f(k) = \mathcal{P}[X = k] = \frac{e^{-\lambda} \lambda^k}{k!}, \quad k = 0, 1, 2, \dots \quad (6.12)$$

Este tipo de leyes se aplican a sucesos con probabilidad muy baja de ocurrir, obteniéndose como la distribución límite de una sucesión de variable binomiales, $\mathbf{B}(n, p)$, donde $n \cdot p = \lambda$, y $n \rightarrow \infty$ (por tanto $p \rightarrow 0^+$).

En general utilizaremos la distribución de Poisson como aproximación de experimentos binomiales donde el número de pruebas es muy alto, pero la probabilidad de éxito muy baja. A veces se suele utilizar como criterio de aproximación:

$$n > 30, p \leq 0,1 \Rightarrow \mathbf{B}(n, p) \cong \mathbf{Poi}(n \cdot p)$$

Su valor esperado y varianza coinciden:

$$\mathbf{E}[X] = \mathbf{Var}[X] = \lambda \quad (6.13)$$

Ejemplo de distribución de Poisson

Cierta enfermedad tiene una probabilidad muy baja de ocurrir, $p = 1/100,000$. Calcular la probabilidad de que en una ciudad con 500,000 habitantes haya más de 3 personas con dicha enfermedad. Calcular el número esperado de habitantes que la padecen.

Solución: Si consideramos la v.a. X que contabiliza el número de personas que padecen la enfermedad, es claro que sigue un modelo binomial, pero que puede ser muy bien aproximado por un modelo de Poisson, de modo que

$$X \rightsquigarrow \mathbf{B} \left(n = 500,000, p = \frac{1}{100,000} \right) \implies X \overset{\sim}{\rightsquigarrow} \mathbf{Poi}(\lambda = 5)$$

Así el número esperado de personas que padecen la enfermedad es $\mathbf{E}[X] = 5$. Como $\mathbf{Var}[X] = 5$, existe una gran dispersión, y no sería extraño encontrar que en realidad hay muchas más personas o menos que están enfermas. La probabilidad de que haya más de tres personas enfermas es:

$$\begin{aligned} \mathcal{P}[X > 3] &= 1 - \mathcal{P}[X \leq 3] \\ &= 1 - \mathcal{P}[X = 0] - \mathcal{P}[X = 1] - \mathcal{P}[X = 2] - \mathcal{P}[X = 3] \\ &= 1 - \frac{e^{-5 \cdot 0}}{0!} - \frac{e^{-5 \cdot 1}}{1!} - \frac{e^{-5 \cdot 2}}{2!} - \frac{e^{-5 \cdot 3}}{3!} \\ &= 0,735 \end{aligned}$$

6.3. Distribuciones continuas

En esta sección estudiaremos las distribuciones más importantes de v.a. continuas unidimensionales. El **soporte** de una v.a. continua se define como aquella región de \mathbb{R} donde su densidad es no nula, $f(x) \neq 0$. Para las distribuciones que enunciaremos, podrá ser bien todo \mathbb{R} , $\mathbb{R}^+ = (0, +\infty)$ o bien un segmento de la forma $[a, b] \subset \mathbb{R}$.

6.3.1. Distribución uniforme o rectangular

Se dice que una v.a. X posee una **distribución uniforme** en el intervalo $[a, b]$,

$$X \rightsquigarrow \mathbf{U}(a, b)$$

si su función de densidad es la siguiente:

$$f(x) = \frac{1}{b-a} \quad \text{si } a \leq x \leq b \quad (6.14)$$

Con esta ley de probabilidad, la probabilidad de que al hacer un experimento aleatorio, el valor de X este comprendido en cierto subintervalo de $[a, b]$ depende únicamente de la longitud del mismo, no de su posición. Cometiendo un pequeño abuso en el lenguaje, podemos decir que *en una distribución uniforme la probabilidad de todos los puntos del soporte es la misma*².

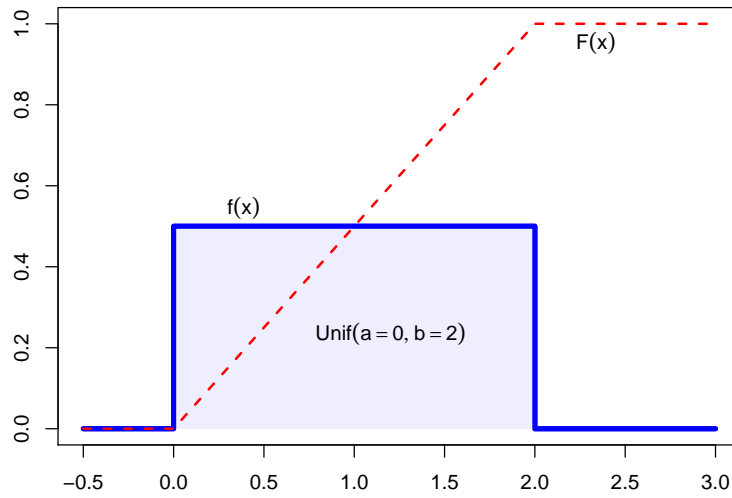


Figura 6.3: Función de densidad y de distribución de $U(a, b)$

$$\mathbf{E}[X] = \frac{b+a}{2}$$

$$\mathbf{Var}[X] = \frac{(b-a)^2}{12}$$

²Hay que observar que en principio esa afirmación es cierta para cualquier v.a. continua, ya que para ellas la probabilidad de cualquier punto es nula. Sería más preciso decir que la densidad de todos los puntos es constante en $[a, b]$.

6.3.2. Distribución exponencial

La distribución exponencial es el equivalente continuo de la distribución geométrica discreta. Esta ley de distribución describe procesos en los que:

- Nos interesa saber el tiempo hasta que ocurre determinado evento, sabiendo que,
- el tiempo que pueda ocurrir desde cualquier instante dado t , hasta que ello ocurra en un instante t_f , no depende del tiempo transcurrido anteriormente en el que no ha pasado nada.

Ejemplos de este tipo de distribuciones son:

- El tiempo que tarda una partícula radiactiva en desintegrarse. El conocimiento de la ley que sigue este evento se utiliza en Ciencia para, por ejemplo, la datación de fósiles o cualquier materia orgánica mediante la técnica del carbono 14, C^{14} ;
- El tiempo que puede transcurrir en un servicio de urgencias, para la llegada de un paciente;
- En un proceso de Poisson donde se repite sucesivamente un experimento a intervalos de tiempo iguales, el tiempo que transcurre entre la ocurrencia de dos sucesos consecutivos sigue un modelo probabilístico exponencial. Por ejemplo, el tiempo que transcurre entre que sufrimos dos veces una herida importante.

Concretando, si una v.a. continua X distribuida a lo largo de \mathbb{R}^+ , es tal que su función de densidad es

$$\boxed{f(x) = \lambda e^{-\lambda x} \text{ si } 0 < x} \quad (6.15)$$

se dice que sigue una **distribución exponencial** de parámetro λ , $X \rightsquigarrow \mathbf{Exp}(\lambda)$.

Un cálculo inmediato nos dice que si $x > 0$,

$$\int_0^x \lambda e^{-\lambda t} dt = -e^{-\lambda t} \Big|_0^x = 1 - e^{-\lambda x}$$

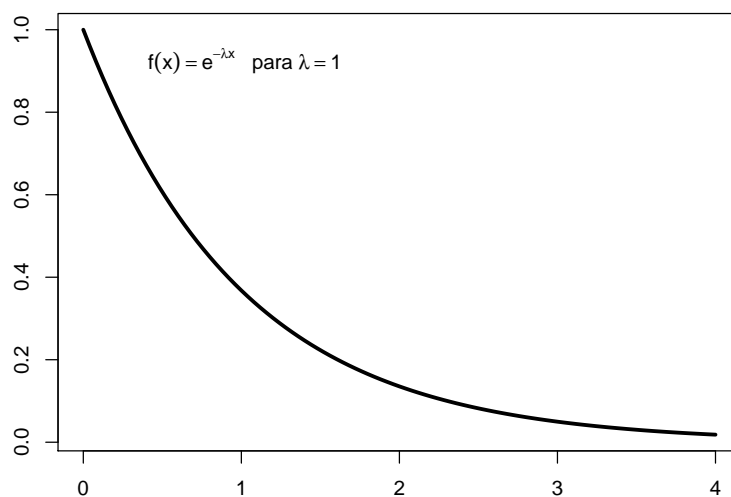


Figura 6.4: Función de densidad, f , de una **Exp**(λ).

luego la función de distribución es:

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{si } 0 < x \\ 0 & \text{en otro caso.} \end{cases}$$

$$\mathbf{E}[X] = \frac{1}{\lambda}$$

$$\mathbf{Var}[X] = \frac{1}{\lambda^2}$$

Ejemplo de variable exponencial

En un experimento de laboratorio se utilizan 10 gramos de ${}^{210}_{84}\text{Po}$. Sabiendo que la duración media de un átomo de esta materia es de 140 días,

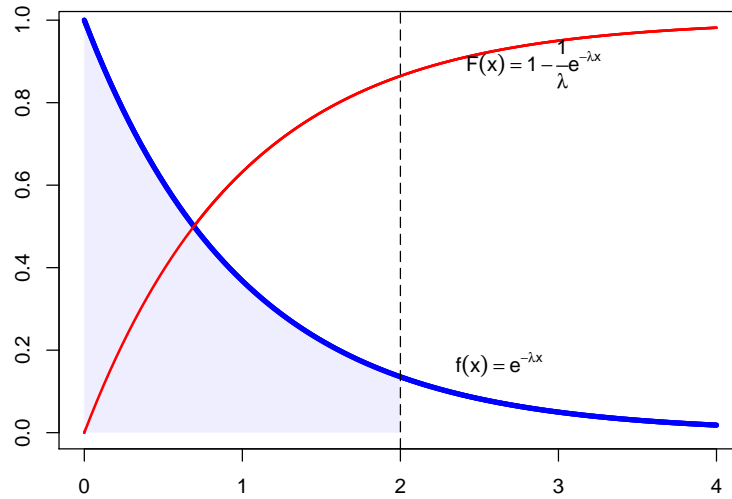


Figura 6.5: Función de distribución, F , de $\mathbf{Exp}(\lambda)$, calculada como el área que deja por debajo de sí la función de densidad.

¿cuántos idas transcurrirán hasta que haya desaparecido el 90 % de este material?

Solución: El tiempo T de desintegración de un átomo de ${}^{210}_{84}\text{Po}$ es una v.a. de distribución exponencial:

$$\begin{aligned}
 T \sim \mathbf{Exp}\left(\lambda = \frac{1}{140}\right) &\iff f(t) = \lambda e^{-\lambda t} \text{ si } \forall t \geq 0 \\
 &\iff F(t) = 1 - e^{-\lambda t}
 \end{aligned}$$

Como el número de átomos de ${}^{210}_{84}\text{Po}$ existentes en una muestra de 10 gramos es enorme, el histograma de frecuencias relativas formado por los tiempos de desintegración de cada uno de estos átomos debe ser extremadamente aproximado a la curva de densidad, f . Del mismo modo, el polígono de frecuencias relativas acumuladas debe ser muy aproximado a la curva de su función de distribución F . Entonces el tiempo que transcurre hasta

que el 90 % del material radiactivo se desintegra es el percentil 90, t_{90} , de la distribución exponencial, es decir

$$F(t_{90}) = 0,9 \Leftrightarrow e^{-\lambda t_{90}} = 1 - 0,9 \Leftrightarrow t_{90} = -\frac{1}{\lambda} \ln 0,1 \approx 322 \text{ días}$$

Otro ejemplo de variable exponencial

Se ha comprobado que el tiempo de vida de cierto tipo de marcapasos sigue una distribución exponencial con media de 16 años. ¿Cuál es la probabilidad de que a una persona a la que se le ha implantado este marcapasos se le deba reimplantar otro antes de 20 años? Si el marcapasos lleva funcionando correctamente 5 años en un paciente, ¿cuál es la probabilidad de que haya que cambiarlo antes de 25 % años?

Solución: Sea T la variable aleatoria que mide la duración de un marcapasos en una persona. Tenemos que

$$\begin{aligned} T \rightsquigarrow \mathbf{Exp} \left(\lambda = \frac{1}{16} \right) &\iff f(t) = \lambda e^{-\lambda t} \text{ si } \forall t \geq 0 \\ &\iff F(t) = 1 - e^{-\lambda t} \end{aligned}$$

Entonces

$$\mathcal{P}[T \leq 20] = \int_0^{20} f(t) dt = F(20) = 1 - e^{-\frac{20}{16}} = 0,7135$$

En segundo lugar

$$\mathcal{P}[T \leq 25 | T \geq 5] = \frac{\mathcal{P}[5 \leq T \leq 25]}{\mathcal{P}[T \geq 5]} = \frac{0,522}{0,7316} = 0,7135 \quad (6.16)$$

$$\mathcal{P}[5 \leq T \leq 25] = \int_5^{25} f(t) dt = F(25) - F(5) = 1 - e^{-\frac{25}{16}} - 1 + e^{-\frac{5}{16}} = 0,522$$

$$\mathcal{P}[T \geq 5] = \int_5^{+\infty} f(t) dt = F(+\infty) - F(5) = 1 - 1 + e^{-\frac{5}{16}} = 0,7316$$

Luego como era de esperar, por ser propio a un mecanismo exponencial,

$$\mathcal{P}[T \leq 25 | T \geq 5] = \mathcal{P}[T \leq 20]$$

o sea, en la duración que se espera que tenga el objeto, no influye en nada el tiempo que en la actualidad lleva funcionando. Es por ello que se dice que “la distribución exponencial no tiene memoria”.

6.3.3. Distribución normal o gaussiana

La *distribución gaussiana*, recibe también el nombre de *distribución normal*, ya que una gran mayoría de las v.a continuas³ de la naturaleza siguen esta distribución. Se dice que una v.a. X sigue una **distribución normal** de parámetros μ y σ^2 , lo que representamos del modo $X \sim \mathbf{N}(\mu, \sigma^2)$ si su función de densidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad \forall x \in \mathbb{R} \quad (6.17)$$

Observación

Estos dos parámetros μ y σ^2 coinciden además con la media (esperanza) y la varianza respectivamente de la distribución como se demostrará más adelante⁴:

$$\mathbf{E}[X] = \mu \quad (6.18)$$

$$\mathbf{Var}[X] = \sigma^2 \quad (6.19)$$

La forma de la función de densidad es la llamada *campana de Gauss*.

Para el lector es un ejercicio interesante comprobar que ésta alcanza un único máximo (*moda*) en μ , que es simétrica con respecto al mismo, y por

³Incluso v.a discretas pueden ser aproximadas por la ley gaussiana.

⁴Hemos adelantado al lector el significado de μ y σ^2 pues esta es una distribución que queda definida en primera instancia por su media y varianza.

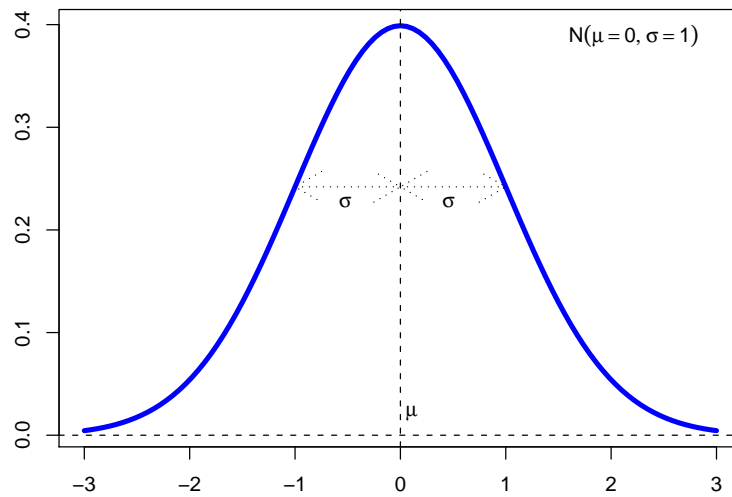


Figura 6.6: Campana de Gauss o función de densidad de una v.a. de distribución normal. EL parámetro μ indica el centro y σ la dispersión. La distancia del centro a los puntos de inflexión es precisamente σ .

tanto $\mathcal{P}[X \leq \mu] = \mathcal{P}[X \geq \mu] = 1/2$, con lo cual en μ coinciden la media, la mediana y la moda, y por último, calcular sus puntos de inflexión.

El soporte de la distribución es todo \mathbb{R} , de modo que la mayor parte de la *masa de probabilidad* (área comprendida entre la curva y el eje de abscisas) se encuentra concentrado alrededor de la media, y las ramas de la curva se extienden asintóticamente a los ejes, de modo que cualquier valor “muy alejado” de la media es posible (aunque poco probable).

La forma de la campana de Gauss depende de los parámetros μ y σ :

- μ indica la posición de la campana (*parámetro de centralización*);
- σ^2 (o equivalentemente, σ) será el parámetro de dispersión. Cuanto menor sea, mayor cantidad de masa de probabilidad habrá concentrada alrededor de la media (gráfico de f muy apuntado cerca de μ) y cuanto mayor sea “más aplastado” será.

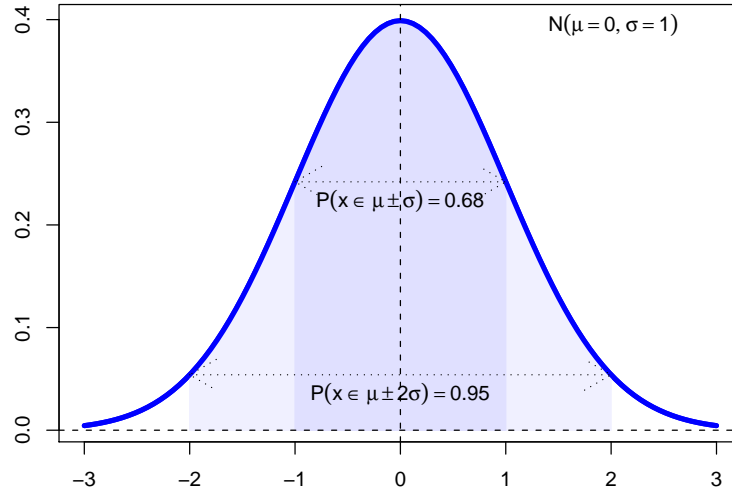


Figura 6.7: A una distancia que no supera en una desviación de la media tenemos una probabilidad del 68 %. A dos desviaciones tenemos el 95 %.

Aproximación a la normal de la ley binomial

Se demuestra que una v.a. discreta con distribución binomial, $X \sim \mathbf{B}(n, p)$ se puede aproximar mediante una distribución normal si n es suficientemente grande y p no está ni muy próximo a 0 ni a 1. Como el valor esperado y la varianza de X son respectivamente np y npq , la aproximación consiste en decir que $X \approx \mathbf{N}(np, npq)$. El convenio que se suele utilizar para poder realizar esta aproximación es:

$$X \sim \mathbf{B}(n, p) \quad \text{donde} \quad \begin{cases} n > 30 \\ np > 4 \\ nq > 4 \end{cases} \implies X \approx \mathbf{N}(np, npq)$$

aunque en realidad esta no da resultados muy precisos a menos que realmente n sea un valor muy grande o $p \approx q \approx 1/2$. Como ilustración obsérvense las figuras 6.10 y 6.11.

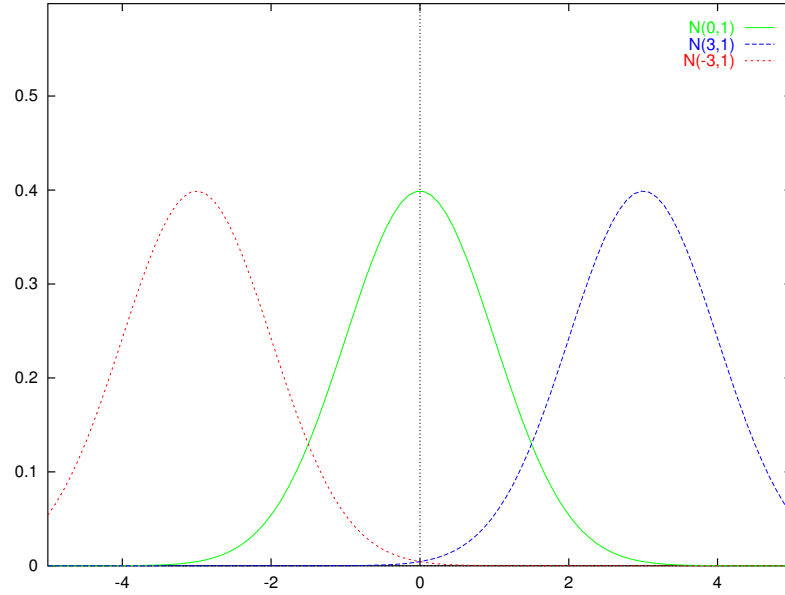


Figura 6.8: Distribuciones gaussianas con diferentes medias e igual dispersión.

6.3.4. Distribución χ^2

Si consideramos una v.a. $Z \sim \mathbf{N}(0, 1)$, la v.a. $X = Z^2$ se distribuye según una ley de probabilidad **distribución χ^2 con un grado de libertad**, lo que se representa como

$$X \sim \chi_1^2$$

Si tenemos n v.a. independientes $Z_i \sim \mathbf{N}(0, 1)$, la suma de sus cuadrados respectivos es una distribución que denominaremos **ley de distribución χ^2 con n grados de libertad**, χ_n^2 .

$$\boxed{\{Z_i\}_{i=1}^n \sim \mathbf{N}(0, 1) \implies \sum_{i=1}^n Z_i^2 \sim \chi_n^2} \quad (6.20)$$

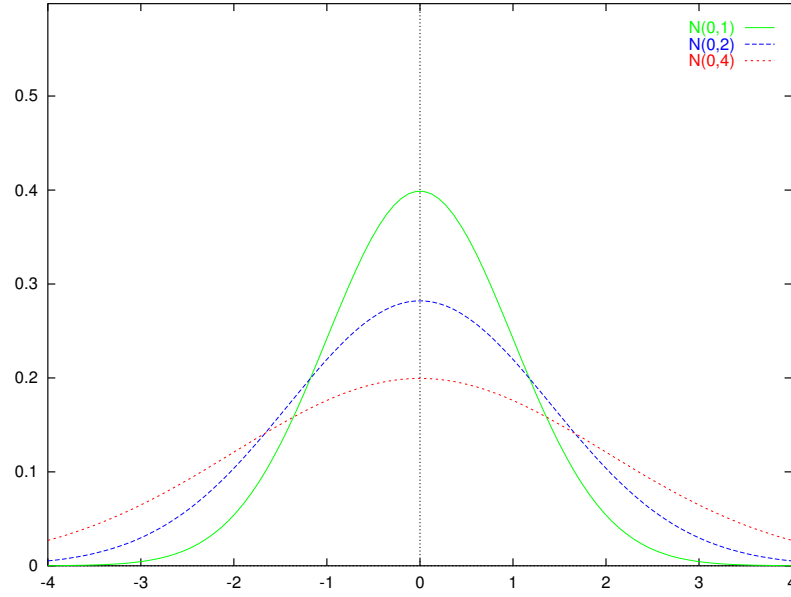


Figura 6.9: Distribuciones gaussianas con igual media pero varianza diferente.

La media y varianza de esta variable son respectivamente:

$$\mathbf{E}[X] = n \quad (6.21)$$

$$\mathbf{Var}[X] = 2n \quad (6.22)$$

En consecuencia, si tenemos X_1, \dots, X_n , v.a. independientes, donde cada $X_i \sim \mathbf{N}(\mu_i, \sigma_i^2)$, se tiene

$$\sum_{i=1}^n \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 \sim \chi_n^2$$

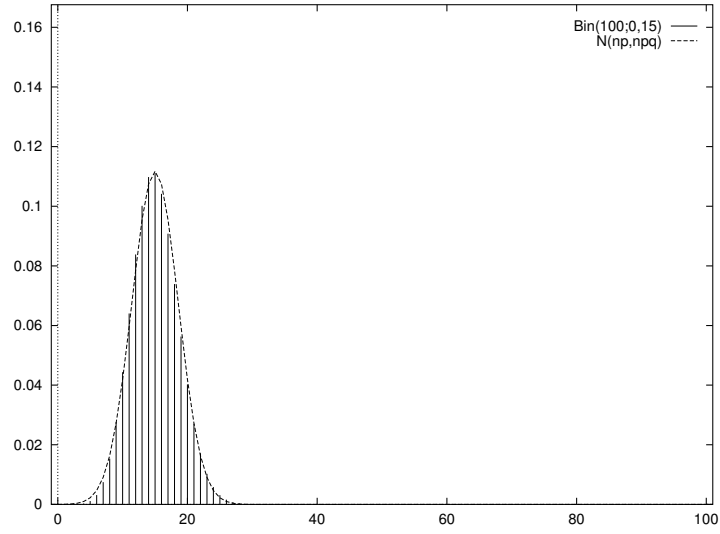


Figura 6.10: Comparación entre la función de densidad de una v.a. continua con distribución $\mathbf{N}(np, npq)$ y el diagrama de barras de una v.a. discreta de distribución $\mathbf{B}(n, p)$ para casos en que la aproximación normal de la binomial es válida. Es peor esta aproximación cuando p está próximo a los bordes del intervalo $[0, 1]$.

6.3.5. Distribución t de Student

La distribución t -Student se construye como un cociente entre una normal y la raíz de una χ^2 independientes. De modo preciso, llamamos **distribución t -Student con n grados de libertad**, t_n a la de una v.a. T ,

$$T = \frac{Z}{\sqrt{\frac{1}{n}\chi_n^2}} \rightsquigarrow t_n \quad (6.23)$$

donde $Z \rightsquigarrow \mathbf{N}(0, 1)$, $\chi_n^2 \rightsquigarrow \chi_n^2$. Este tipo de distribuciones aparece cuando tenemos $n + 1$ v.a. independientes

$$X \rightsquigarrow \mathbf{N}(\mu, \sigma^2)$$

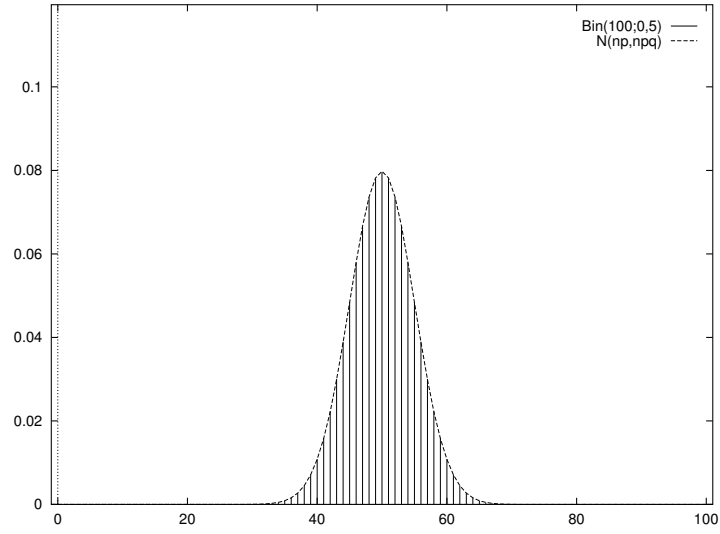


Figura 6.11: La misma comparación que en la figura anterior, pero realizada con parámetros con los que damos la aproximación normal de la binomial es mejor.

$$X_i \rightsquigarrow \mathbf{N}(\mu_i, \sigma_i^2) \quad i = 1, \dots, n$$

y nos interesa la distribución de

$$T = \frac{\frac{X - \mu}{\sigma}}{\sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2}} \rightsquigarrow \mathbf{t}_n$$

La distribución \mathbf{t} de Student tiene propiedades parecidas a $\mathbf{N}(0, 1)$:

- Es de media cero, y simétrica con respecto a la misma;
- Es algo más dispersa que la normal, pero la varianza decrece hasta 1 cuando el número de grados de libertad aumenta;

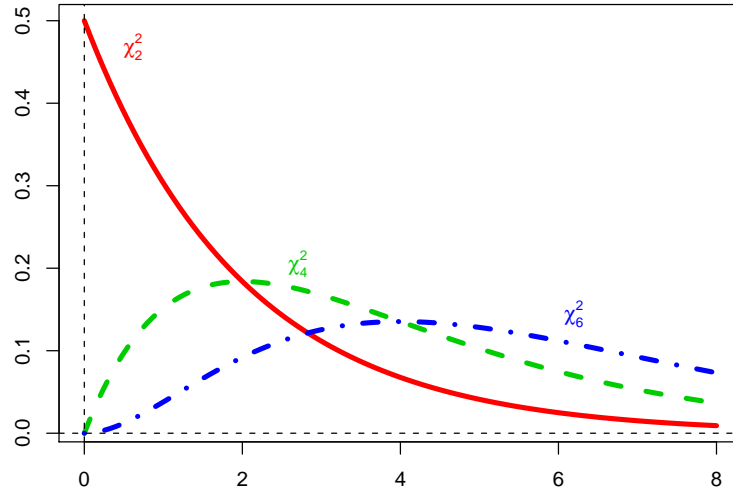


Figura 6.12: Función de densidad de χ_n^2 para valores pequeños de n .

- Para un número alto de grados de libertad se puede aproximar la distribución de Student por la normal, es decir,

$$t_n \xrightarrow{n \rightarrow \infty} N(0, 1)$$

6.3.6. La distribución F de Snedecor

Otra de las distribuciones importantes asociadas a la normal es la que se define como cociente de distribuciones χ^2 independientes. Sean $X \sim \chi_n^2$ e $Y \sim \chi_m^2$ v.a. independientes. Decimos entonces que la variable

$$F = \frac{\frac{1}{n}X}{\frac{1}{m}Y} = \frac{m}{n} \frac{X}{Y} \sim F_{n,m} \quad (6.24)$$

sigue una **distribución de probabilidad de Snedecor, con (n, m) grados de libertad**. Obsérvese que $F_{n,m} \neq F_{m,n}$.

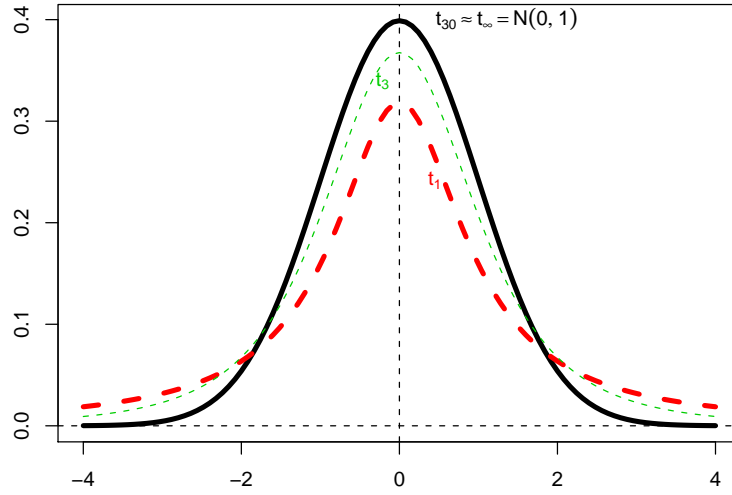


Figura 6.13: Cuando aumentan los grados de libertad, la distribución de Student se aproxima a la distribución normal tipificada.

La forma más habitual en que nos encontraremos esta distribución será en el caso en que tengamos $n + m$ v.a. independientes

$$X_i \sim \mathbf{N}(\mu_i, \sigma_i^2) \quad i = 1, \dots, n$$

$$Y_j \sim \mathbf{N}(m_j, s_j^2) \quad j = 1, \dots, m$$

y así

$$F = \frac{\frac{1}{n} \sum_{i=1}^n \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2}{\frac{1}{m} \sum_{j=1}^m \left(\frac{Y_j - m_j}{s_j} \right)^2} \sim \mathbf{F}_{n,m}$$

Es claro que la distribución de Snedecor no es simétrica, pues sólo tienen densidad de probabilidad distinta de cero, los puntos de \mathbb{R}^+ . Otra propiedad interesante de la distribución de Snedecor es:

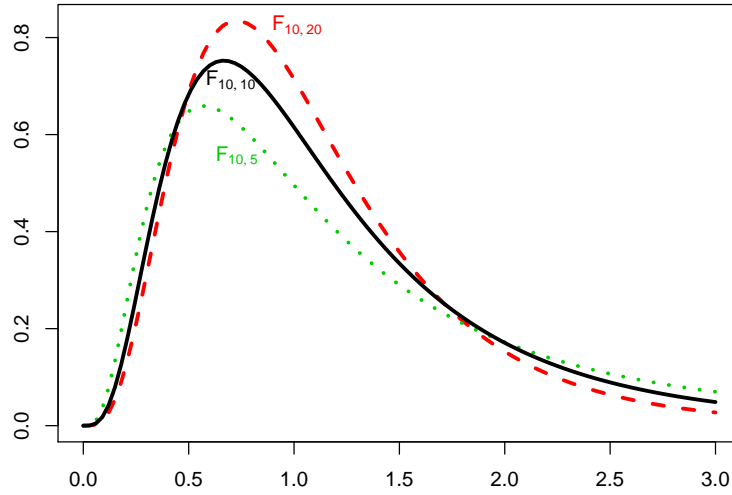


Figura 6.14: Funciones de densidad para la distribución F de Snedecor.

$$F \rightsquigarrow \mathbf{F}_{n,m} \iff \frac{1}{F} \rightsquigarrow \mathbf{F}_{m,n}$$

6.4. Problemas

Ejercicio 6.1. Para estudiar la regulación hormonal de una línea metabólica se inyectan ratas albinas con un fármaco que inhibe la síntesis de proteínas del organismo. En general, 4 de cada 20 ratas mueren a causa del fármaco antes de que el experimento haya concluido. Si se trata a 10 animales con el fármaco, ¿cuál es la probabilidad de que al menos 8 lleguen vivos al final del experimento?

Ejercicio 6.2. En una cierta población se ha observado un número medio anual de muertes por cáncer de pulmón de 12. Si el número de muertes causadas por la enfermedad sigue una distribución de Poisson, ¿cuál es la probabilidad de que durante el año en curso:

1. Haya exactamente 10 muertes por cáncer de pulmón?
2. 15 o más personas mueran a causa de la enfermedad?
3. 10 o menos personas mueran a causa de la enfermedad?

Ejercicio 6.3. Dañando los cromosomas del óvulo o del espermatozoide, pueden causarse mutaciones que conducen a abortos, defectos de nacimiento, u otras deficiencias genéticas. La probabilidad de que tal mutación se produzca por radiación es del 10 %. De las siguientes 150 mutaciones causadas por cromosomas dañados, ¿cuántas se esperaría que se debiesen a radiaciones? ¿Cuál es la probabilidad de que solamente 10 se debiesen a radiaciones?

Ejercicio 6.4. Entre los diabéticos, el nivel de glucosa en sangre X , en ayunas, puede suponerse de distribución aproximadamente normal, con media 106 mg/100 ml y desviación típica 8 mg/100 ml, es decir

$$X \sim \mathbf{N}(\mu = 106, \sigma^2 = 64)$$

1. Hallar $\mathcal{P}[X \leq 120]$
2. ¿Qué porcentaje de diabéticos tienen niveles comprendidos entre 90 y 120 ?
3. Hallar $\mathcal{P}[106 \leq X \leq 110]$.
4. Hallar $\mathcal{P}[X \leq 121]$.
5. Hallar el punto x caracterizado por la propiedad de que el 25 % de todos los diabéticos tiene un nivel de glucosa en ayunas inferior o igual a x .

Ejercicio 6.5. Una prueba de laboratorio para detectar heroína en sangre tiene un 92 % de precisión. Si se analizan 72 muestras en un mes, ¿cuál es la probabilidad de que:

1. 60 o menos estén correctamente evaluadas?
2. menos de 60 estén correctamente evaluadas?
3. exactamente 60 estén correctamente evaluadas?

Ejercicio 6.6. El 10 % de las personas tiene algún tipo de alergia. Se seleccionan aleatoriamente 100 individuos y se les entrevista. Hallar la probabilidad de que, al menos, 12 tengan algún tipo de alergia. Hallar la probabilidad de que, como máximo, 8 sean alérgicos a algo.

Ejercicio 6.7. La probabilidad de muerte resultante del uso de píldoras anticonceptivas es de $3/100,000$. De 1,000,000 de mujeres que utilizan este medio de control de natalidad:

1. ¿Cuántas muertes debidas a esta causa se esperan?
2. ¿Cuál es la probabilidad de que haya, como máximo, 25 de estas muertes?
3. ¿Cuál es la probabilidad de que el número de muertes debidas a esta causa esté entre 25 y 35, inclusive?

Ejercicio 6.8. La probabilidad de presentar una característica genética es de $1/20$.

1. Tomando una muestra de 8 individuos, calcular la probabilidad de que 3 individuos presenten la característica.
2. Tomando una muestra de 80 personas, ¿cuál será la probabilidad de que aparezcan más de 5 individuos con la característica?

Ejercicio 6.9. Se supone que en una cierta población humana el *índice cefálico* i , (cociente entre el diámetro transversal y el longitudinal expresado en tanto por ciento), se distribuye según una Normal. El 58 % de los

habitantes son dolicocefalos ($i \leq 75$), el 38 % son mesocéfalos ($75 < i \leq 80$) y el 4 % son braquicéfalos ($i > 80$). Hállese la media y la desviación típica del índice cefálico en esa población.

Ejercicio 6.10. Se supone que la glucemia basal en individuos sanos, X_s sigue una distribución

$$X_s \sim \mathbf{N}(\mu = 80, \sigma = 10),$$

mientras que en los diabéticos X_d , sigue una distribución

$$X_d \sim \mathbf{N}(\mu = 160, \sigma = 31, 4).$$

Si se conviene en clasificar como sanos al 2 % de los diabéticos:

1. ¿Por debajo de qué valor se considera sano a un individuo? ¿Cuántos sanos serán clasificados como diabéticos?
2. Se sabe que en la población en general el 10 % de los individuos son diabéticos ¿cuál es la probabilidad de que un individuo elegido al azar y diagnosticado como diabético, realmente lo sea?

Ejercicio 6.11. Supóngase que se van a utilizar 20 ratas en un estudio de agentes coagulantes de la sangre. Como primera experiencia, se dio un anticoagulante a 10 de ellos, pero por inadvertencia se pusieron todas sin marcas en el mismo recinto. Se necesitaron 12 ratas para la segunda fase del estudio y se les tomó al azar sin reemplazamiento. ¿Cuál es la probabilidad de que de las 12 elegidas 6 tengan la droga y 6 no la tengan?

Capítulo 7

Introducción a la inferencia

7.1. Introducción

El propósito de un estudio estadístico suele ser, como hemos venido citando, extraer conclusiones acerca de la naturaleza de una población. Al ser la población grande y no poder ser estudiada en su integridad en la mayoría de los casos, las conclusiones obtenidas deben basarse en el examen de solamente una parte de ésta, lo que nos lleva, en primer lugar a la justificación, necesidad y definición de las diferentes técnicas de muestreo.

Los primeros términos obligados a los que debemos hacer referencia, definidos en el primer capítulo, serán los de estadístico y estimador.

Dentro de este contexto, será necesario asumir un estadístico o estimador como una variable aleatoria con una determinada distribución, y que será la pieza clave en las dos amplias categorías de la inferencia estadística: la estimación y el contraste de hipótesis.

El concepto de estimador, como herramienta fundamental, lo caracterizamos mediante una serie de propiedades que nos servirán para elegir el “mejor” para un determinado parámetro de una población, así como algunos métodos para la obtención de ellos, tanto en la estimación puntual como por intervalos.

En el capítulo anterior dedujimos ciertas leyes de probabilidad mediante un método deductivo a partir del conocimiento del mecanismo generador

de los sucesos aleatorios. De este modo pudimos deducir las leyes de probabilidad binomial o hipergeométrica por ejemplo. Así una vez precisamente determinada la ley probabilística que subyace en el experimento aleatorio, podemos obtener muestras de la v.a. siguiendo esa ley de probabilidad. En este momento nos interesamos por el proceso contrario, es decir:

¿Cómo deducir la ley de probabilidad sobre determinado carácter de una población cuando sólo conocemos una muestra?

Este es un problema al que nos enfrentamos cuando por ejemplo tratamos de estudiar la relación entre el *fumar* y el *cáncer de pulmón* e intentamos extender las conclusiones obtenidas sobre una muestra al resto de individuos de la población.

La tarea fundamental de la **estadística inferencial**, es hacer inferencias acerca de la población a partir de una muestra extraída de la misma.

7.2. Técnicas de muestreo sobre una población

La *teoría del muestreo* tiene por objetivo, el estudio de las relaciones existentes entre la distribución de un carácter en dicha población y las distribuciones de dicho carácter en todas sus muestras.

Las ventajas de estudiar una población a partir de sus muestras son principalmente:

Coste reducido: Si los datos que buscamos los podemos obtener a partir de una pequeña parte del total de la población, los gastos de recogida y tratamiento de los datos serán menores. Por ejemplo, cuando se realizan encuestas previas a un referéndum, es más barato preguntar a 4,000 personas su intención de voto, que a 30,000,000;

Mayor rapidez: Estamos acostumbrados a ver cómo con los resultados del escrutinio de las primeras mesas electorales, se obtiene una aproximación bastante buena del resultado final de unas elecciones, muchas horas antes de que el recuento final de votos haya finalizado;

Más posibilidades: Para hacer cierto tipo de estudios, por ejemplo el de duración de cierto tipo de bombillas, no es posible en la práctica

destruirlas todas para conocer su vida media, ya que no quedaría nada que vender. Es mejor destruir sólo una pequeña parte de ellas y sacar conclusiones sobre las demás.

De este modo se ve que al hacer estadística inferencial debemos enfrentarnos con dos problemas:

- Elección de la muestra (*muestreo*), que es a lo que nos dedicaremos en este capítulo.
- Extrapolación de las conclusiones obtenidas sobre la muestra, al resto de la población (*inferencia*).

El tipo de muestreo más importante es el *muestreo aleatorio*, en el que todos los elementos de la población tienen la misma probabilidad de ser extraídos; Aunque dependiendo del problema y con el objetivo de reducir los costes o aumentar la precisión, otros tipos de muestreo pueden ser considerados como veremos más adelante: *muestreo sistemático, estratificado y por conglomerados*.

7.2.1. Muestreo aleatorio

Consideremos una población finita, de la que deseamos extraer una muestra. Cuando el proceso de extracción es tal que garantiza a cada uno de los elementos de la población la misma oportunidad de ser incluidos en dicha muestra, denominamos al proceso de selección **muestreo aleatorio**.

El muestreo aleatorio se puede plantear bajo dos puntos de vista:

- Sin reposición de los elementos;
- Con reposición.

Muestreo aleatorio sin reposición

Consideremos una población E formada por N elementos. Si observamos un elemento particular, $e \in E$, en un muestreo aleatorio sin reposición se da la siguiente circunstancia:

- La probabilidad de que e sea elegido en primer lugar es $\frac{1}{N}$;
- Si no ha sido elegido en primer lugar (lo que ocurre con una probabilidad de $\frac{N-1}{N}$), la probabilidad de que sea elegido en el segundo intento es de $\frac{1}{N-1}$.
- en el $(i + 1)$ -ésimo intento, la población consta de $N - i$ elementos, con lo cual si e no ha sido seleccionado previamente, la probabilidad de que lo sea en este momento es de $\frac{1}{N-i}$.

Muestreo aleatorio con reposición

Sobre una población E de tamaño N podemos realizar extracciones de n elementos, pero de modo que cada vez el elemento extraído es repuesto al total de la población. De esta forma un elemento puede ser extraído varias veces.

El muestreo aleatorio con reposición es también denominado **muestreo aleatorio simple**, y se caracteriza porque cada elemento de la población tiene la misma probabilidad de ser elegido, y las observaciones se realizan con reemplazamiento. De este modo, cada observación es realizada sobre la misma población (que no disminuye con las extracciones sucesivas).

7.2.2. Muestreo aleatorio estratificado

Un **muestreo aleatorio estratificado** es aquel en el que se divide la población de N individuos, en k subpoblaciones o **estratos**, atendiendo a criterios que puedan ser importantes en el estudio, de tamaños respectivos N_1, \dots, N_k ,

$$N = N_1 + N_2 + \dots + N_k$$

y realizando en cada una de estas subpoblaciones muestreos aleatorios simples de tamaño n_i $i = 1, \dots, k$.

A continuación nos planteamos el problema de cuantos elementos de muestra se han de elegir de cada uno de los estratos. Para ello tenemos

fundamentalmente dos técnicas: la asignación proporcional y la asignación óptima.

Asignación proporcional

Sea n el número de individuos de la población total que forman parte de alguna muestra:

$$n = n_1 + n_2 + \cdots + n_k$$

Cuando la asignación es **proporcional** el tamaño de la muestra de cada estrato es proporcional al tamaño del estrato correspondiente con respecto a la población total:

$$n_i = n \cdot \frac{N_i}{N}$$

Asignación óptima

Cuando se realiza un muestreo estratificado, los tamaños muestrales en cada uno de los estratos, n_i , los elige quien hace el muestreo, y para ello puede basarse en alguno de los siguientes criterios:

- Elegir los n_i de tal modo que se minimice la varianza del *estimador*, para un coste especificado, o bien,
- habiendo fijado la varianza que podemos admitir para el estimador, minimizar el coste en la obtención de las muestras.

Así en un estrato dado, se tiende a tomar una muestra más grande cuando:

- El estrato es más grande;
- El estrato posee mayor variabilidad interna (varianza);
- El muestreo es más barato en ese estrato.

7.2.3. Muestreo sistemático

Cuando los elementos de la población están ordenados en fichas o en una lista, una manera de *muestrear* consiste en

- Sea $k = \left\lceil \frac{N}{n} \right\rceil$;
- Elegir aleatoriamente un número m , entre 1 y k ;
- Tomar como muestra los elementos de la lista:

$$\left\{ e_m, e_{m+k}, e_{m+2k}, \dots, e_{m+(n-1)k} \right\}$$

Esto es lo que se denomina **muestreo sistemático**. Cuando el criterio de ordenación de los elementos en la lista es tal que los elementos más parecidos tienden a estar más cercanos, el muestreo sistemático suele ser más preciso que el aleatorio simple, ya que recorre la población de un modo más uniforme. Por otro lado, es a menudo más fácil no cometer errores con un muestreo sistemático que con este último.

El método tal como se ha definido anteriormente es sesgado si $\frac{N}{n}$ no es entero, ya que los últimos elementos de la lista nunca pueden ser escogidos. Un modo de evitar este problema consiste en considerar la lista como si fuese *circular* (el elemento $N + 1$ coincide con el primero) y:

- Sea k el entero más cercano a $\frac{N}{n}$;
- Se selecciona un número al azar m , entre 1 y N ;
- Se toma como muestra los elementos de la lista que consisten en ir saltando de k elementos en k , a partir de m , teniendo en cuenta que la lista es circular.

Se puede comprobar que con este método todos los elementos de la lista tienen la misma probabilidad de selección.

7.2.4. Muestreo por conglomerados

Si intentamos hacer un estudio sobre los habitantes de una ciudad, el muestreo aleatorio simple puede resultar muy costoso, ya que estudiar una muestra de tamaño n implica enviar a los encuestadores a n puntos distintos de la misma, de modo que en cada uno de ellos sólo se realiza una entrevista. En esta situación es más económico realizar el denominado **muestreo por conglomerados**, que consiste en elegir aleatoriamente ciertos barrios dentro de la ciudad, para después elegir calles y edificios. Una vez elegido el edificio, se entrevista a todos los vecinos.

7.3. Propiedades deseables de un estimador

Sea X una v.a. cuya función de probabilidad (o densidad de probabilidad si es continua) depende de unos parámetros $\theta_1, \dots, \theta_k$ desconocidos.

$$f(x; \theta_1, \theta_2, \dots, \theta_k)$$

Representamos mediante X_1, \dots, X_n una muestra aleatoria simple de la variable. Denotamos mediante f_c a la función de densidad conjunta de la muestra, que por estar formada por observaciones independientes, puede factorizarse del siguiente modo:

$$f_c(x_1, x_2, \dots, x_n; \theta_1, \dots, \theta_k) = f(x_1; \theta_1, \dots, \theta_k) \cdot f(x_2; \theta_1, \dots, \theta_k) \cdots f(x_n; \theta_1, \dots, \theta_k)$$

Se denomina **estimador de un parámetro** θ_i , a cualquier v.a. $\hat{\theta}_i$ que se exprese en función de la muestra aleatoria y que tenga por objetivo aproximar el valor de θ_i ,

$$\hat{\theta}_i(X_1, \dots, X_n) \quad \longleftarrow \quad \text{estimador de } \theta_i. \quad (7.1)$$

Obsérvese que el estimador *no es un valor concreto* sino una variable aleatoria, ya que aunque depende unívocamente de los valores de la muestra observados ($X_i = x_i$), la elección de la muestra es un proceso aleatorio. Una vez que la muestra ha sido elegida, se denomina **estimación** el valor numérico que toma el estimador sobre esa muestra.

Intuitivamente, las características que serían deseables para esta nueva variable aleatoria (que usaremos para estimar el parámetro desconocido) deben ser:

- Consistencia: Cuando el tamaño de la muestra crece arbitrariamente, el valor estimado se aproxima al parámetro desconocido.
- Carencia de sesgo: El valor medio que se obtiene de la estimación para diferentes muestras debe ser el valor del parámetro.
- Eficiencia: Al estimador, al ser v.a., no puede exigírsele que para una muestra cualquiera se obtenga como estimación el valor exacto del parámetro. Sin embargo podemos pedirle que su dispersión con respecto al valor central (varianza) sea tan pequeña como sea posible.
- Suficiencia: El estimador debería aprovechar toda la información existente en la muestra.

7.3.1. Estimadores de máxima verosimilitud

Sea X una v.a. con función de probabilidad

$$f(x; \theta)$$

Las muestras aleatorias simples de tamaño n , X_1, X_2, \dots, X_n tienen por distribución de probabilidad conjunta

$$f_c(x_1, x_2, \dots, x_n; \theta) = f(x_1, x_2, \dots, x_n; \theta) f(x_1; \theta) \cdot f(x_2; \theta) \cdots f(x_n; \theta)$$

Esta función que depende de $n + 1$ cantidades podemos considerarla de dos maneras:

- Fijando θ , es una función de las n cantidades x_i . Esto es la función de probabilidad o densidad.
- Fijados los x_i como consecuencia de los resultados de elegir una muestra mediante un experimento aleatorio, es únicamente función de θ . A esta función de θ la denominamos **función de verosimilitud**.

En este punto podemos plantearnos el que dado una muestra sobre la que se ha observado los valores x_i , una posible estimación del parámetro es aquella que maximiza la función de verosimilitud. (cf. figura 7.1)

$$x_1, \dots, x_n \text{ fijados} \implies \text{Verosimilitud} \equiv V(\theta) = f(x_1, x_2, \dots, x_n; \theta)$$

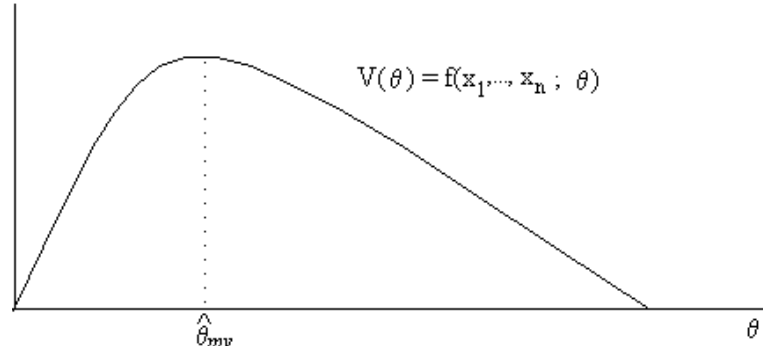


Figura 7.1: La función de verosimilitud se obtiene a partir de la función de densidad, intercambiando los papeles entre parámetro y estimador. En una función de verosimilitud consideramos que las observaciones x_1, \dots, x_n están fijadas, y se representa la gráfica con el valor de los valores que tomaría la función de densidad para todos los posibles valores del parámetro θ . El estimador máximo verosímil del parámetro buscado, $\hat{\theta}_{MV}$, es aquel que maximiza su función de verosimilitud, $V(\theta)$.

Como es lo mismo maximizar una función que su logaritmo (al ser este una función estrictamente creciente), este máximo puede calcularse derivando con respecto a θ la función de verosimilitud (bien su logaritmo) y tomando como estimador máximo verosímil al que haga la derivada nula:

$$\frac{\partial \log V}{\partial \theta} (\hat{\theta}_{MV}) = 0.$$

De modo más preciso, se define el **estimador máximo verosímil** como la v.a.

$$\hat{\theta}_{MV} = \max_{\theta \in \mathbb{R}} f(X_1, X_2, \dots, X_n; \tilde{\theta})$$

Los estimadores de máxima verosimilitud tienen ciertas propiedades en general que a continuación enunciamos:

1. Son consistentes;
2. Son invariantes frente a transformaciones biunívocas, es decir, si $\hat{\theta}_{\mathcal{MV}}$ es el estimador máximo verosímil de θ y $g(\tilde{\theta})$ es una función biunívoca de $\tilde{\theta}$, entonces $g(\hat{\theta}_{\mathcal{MV}})$ es el estimador máximo verosímil de $g(\theta)$.
3. Si $\hat{\theta}$ es un estimador suficiente de θ , su estimador máximo verosímil, $\hat{\theta}_{\mathcal{MV}}$ es función de la muestra a través de $\hat{\theta}$;
4. Son asintóticamente normales;
5. Son asintóticamente eficientes, es decir, entre todos los estimadores consistentes de un parámetro θ , los de máxima verosimilitud son los de varianza mínima.
6. No siempre son insesgados.

7.3.2. Algunos estimadores fundamentales

Vamos a estudiar las propiedades de ciertos estimadores que por su importancia en las aplicaciones resultan fundamentales: estimadores de la esperanza matemática y varianza de una distribución de probabilidad.

Estimador de la esperanza matemática

Consideremos las muestras de tamaño n , X_1, X_2, \dots, X_n , de un carácter sobre una población que viene expresado a través de una v.a. X que posee momentos de primer y segundo orden, es decir, existen $\mathbf{E}[X]$ y $\mathbf{Var}[X]$:

$$X_1, X_2, \dots, X_n, \quad \left\{ \begin{array}{l} \mathbf{E}[X_i] = \mu \\ \mathbf{Var}[X_i] = \sigma^2 \end{array} \right.$$

El estimador *media muestral* que denotaremos normalmente como \bar{X} (en lugar de $\hat{\mu}$ es

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

verifica:

$$\mathbf{E} [\bar{X}] = \mu$$

$$\mathbf{Var} [\bar{X}] = \frac{\sigma^2}{n}$$

Por tanto es un estimador insesgado. Si además sabemos que X se distribuye según una ley gaussiana, se puede comprobar que coincide con el estimador de máxima verosimilitud:

Proposición

$$X_i \rightsquigarrow \mathbf{N}(\mu, \sigma) \implies \bar{X} \equiv \hat{\mu}_{\mathcal{MV}} \rightsquigarrow \mathbf{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Estimador de la varianza

Al elegir un estimador de $\sigma^2 = \mathbf{Var} [X]$, podemos comenzar con el estimador más natural (que es el estimador máximo verosimil) sin embargo éste no es insesgado, ya que el valor esperado del estimador

$$\mathcal{S}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

se demuestra que es $(n-1)/n \cdot \sigma^2$. De esta manera, para conseguir un estimador insesgado de la varianza se introduce la cuasivarianza muestral:

$$\hat{\mathcal{S}}^2 = \frac{n}{n-1} \mathcal{S}^2 \tag{7.2}$$

la cual presenta como valor esperado σ^2 . Se puede comprobar además que

$$\frac{(n-1)\hat{\mathcal{S}}^2}{\sigma^2} \rightsquigarrow \chi_{n-1}$$

Capítulo 8

Estimación confidencial

8.1. Introducción

En el capítulo anterior establecimos toda la teoría que concierne a la definición y concepto de un estimador puntual, así como las propiedades deseables que debe verificar para considerar el producto una “buena” estimación del parámetro.

Existen, no obstante, multitud de circunstancias en las que el interés de un estudio no estriba tanto en obtener una estimación puntual para un parámetro, como determinar un posible “rango” de valores o “intervalo” en los que pueda precisarse, con una determinada probabilidad, que el verdadero valor del parámetro se encuentra dentro de esos límites.

Las técnicas que abordan este tipo de situaciones, se encuadran dentro de la estadística Inferencial bajo el título de “Estimación Confidencial.” “Estimación por Intervalos de Confianza”. El desarrollo teórico de como llega a constituirse un intervalo, realizado en el caso más intuitivo y sencillo, así como los intervalos de confianza para los parámetros más usuales: medias, varianzas y proporciones, para una y dos poblaciones, son el objetivo de este capítulo. Para ello empezamos bajo el supuesto de que nuestra variable en estudio es una variable aleatoria que sigue una distribución cualquiera. Nuestro objetivo será determinar los límites del intervalo de confianza para éstos.

Sea $X \sim \mathbf{Fam}(\theta)$ una v.a. de cierta familia, que se distribuye según un parámetro θ que desconocemos. Para estimar dicho parámetro a partir de una muestra aleatoria simple

$$\vec{X} \stackrel{\text{def}}{=} X_1, X_2, \dots, X_n$$

hemos definido lo que es un estimador $\hat{\theta}(\vec{X})$ y hemos enunciado las buenas propiedades que es deseable que posea. Cuando se realiza el experimento aleatorio de extraer una muestra concreta de la población, el estimador (que a veces denominaremos *estimador puntual*) nos da una aproximación de θ .

$$\left. \begin{array}{l} X_1 = x_1 \\ X_2 = x_2 \\ \dots \\ X_n = x_n \end{array} \right\} \Rightarrow \hat{\theta}(\underbrace{x_1, x_2, \dots, x_n}_{\vec{x} = (x_1, x_2, \dots, x_n)}) \approx \theta$$

Esto es lo que se denomina *estimación puntual*, pues se asigna un punto como estimación del valor del parámetro.

La **estimación confidencial** o **estimación por intervalos de confianza** asigna un conjunto de valores como estimación del parámetro, que generalmente tiene forma de intervalo: $I(\vec{X})$.

Diremos que $I(\vec{X})$ es un **intervalo aleatorio al nivel de significación** α , o equivalentemente, **intervalo aleatorio al nivel de confianza** $1 - \alpha$ si

$$\mathcal{P} [\theta \in I(\vec{X})] \geq 1 - \alpha,$$

o lo que es lo mismo

$$\mathcal{P} [\theta \notin I(\vec{X})] < \alpha.$$

Cuando un intervalo aleatorio $I(\vec{X})$ tiene una probabilidad menor del $100 \cdot \alpha \%$ de que el parámetro no esté en el intervalo decimos que el intervalo es de confianza $1 - \alpha$, o de significación α .

Es importante comprender correctamente esta idea: $I(\vec{X})$ es un conjunto aleatorio que depende de la muestra elegida. Por tanto para cada muestra tenemos un intervalo de confianza diferente. Si elegimos un nivel de confianza por ejemplo de $\alpha = 95 \%$, y encontramos (mediante la técnica que sea) intervalos de confianza al 95% que se correspondan con cada una de las muestras, lo que sabemos es que en el 95% de los casos los intervalos de confianza dieron una respuesta correcta. En el 5% restante se obtuvo una respuesta incorrecta.

Cuando una muestra ha sido elegida mediante un muestreo aleatorio simple, no tiene sentido decir $\theta \in I(\vec{x})$ con probabilidad $1 - \alpha$, pues sólo puede ocurrir que (fijada la muestra) el parámetro esté o que no esté dentro del intervalo. Sin embargo por comodidad a veces se utiliza esa expresión, donde lo que queremos con esa frase es expresar la idea de que “si hubiésemos tomado muestras del mismo tamaño en una gran cantidad de ocasiones, hubiésemos acertado por lo menos en un $100 \cdot (1 - \alpha) \%$ de las ocasiones al decir que el parámetro estaba en el intervalo que cada muestra suministra”.

8.2. Intervalos de confianza para la distribución normal

Dada una variable aleatoria de distribución gaussiana $X \sim \mathbf{N}(\mu, \sigma^2)$, nos interesamos en primer lugar, en calcular intervalos de confianza para sus dos parámetros, μ y σ^2 .

He aquí un resumen de las situaciones que consideraremos:

la media si se conoce la varianza: Este no es un caso práctico (no se puede conocer σ^2 sin conocer previamente μ), pero sirve para introducirnos en el problema de la estimación confidencial de la media;

anza para la media (caso general): Este se trata del caso con verdadero interés práctico. Por ejemplo sirve para estimar intervalos que contenga la media del colesterol en sangre en una población, la altura, el peso, etc, cuando disponemos de una muestra de la variable.

Intervalo de confianza para la varianza: Éste es otro caso de interés en las aplicaciones. El objetivo es calcular un intervalo de confianza para σ^2 , cuando sólo se dispone de una muestra.

Estimación de tamaño muestral La utilidad consiste en decidir cuál deberá ser el tamaño necesario de una muestra para obtener intervalos de confianza para una media, con precisión y significación dadas de antemano. Para que esto sea posible es necesario poseer cierta información previa, que se obtiene a partir de las denominadas **muestras piloto**.

Más adelante, consideramos el caso en que tenemos dos poblaciones donde cada una sigue su propia ley de distribución $\mathbf{N}(\mu_1, \sigma_1^2)$ y $\mathbf{N}(\mu_2, \sigma_2^2)$. Los problemas asociados a este caso son

Diferencia de medias homocedásticas Se realiza el cálculo del intervalo de confianza suponiendo que ambas variables tienen la misma varianza, es decir **son homocedásticas**. En la práctica se usa este cálculo, cuando ambas variables tienen parecida dispersión.

Diferencia de medias (caso general) Es el mismo caso que el anterior, pero se realiza cuando se observa que hay diferencia notable en la dispersión de ambas variables.

8.2.1. Intervalo para la media si se conoce la varianza

Este caso que planteamos es más a nivel teórico que práctico: difícilmente vamos a poder conocer con exactitud σ^2 mientras que μ es desconocido. Sin embargo nos aproxima del modo más simple a la estimación confidencial de medias.

Para estimar μ , el estadístico que mejor nos va a ayudar es \bar{X} , del que

conocemos su ley de distribución:

$$\bar{X} \rightsquigarrow \underbrace{\mathbf{N}\left(\mu, \frac{\sigma^2}{n}\right)}_{\substack{\text{un parámetro} \\ \text{desconocido}}}$$

Esa ley de distribución depende de μ (desconocida). Lo más conveniente es hacer que la ley de distribución no dependa de ningún parámetro desconocido, para ello tipificamos:

$$Z = \underbrace{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}_{\substack{\text{par. desconocido} \\ + \\ \text{estimador} \\ + \\ \text{cosas conocidas}}} \rightsquigarrow \underbrace{\mathbf{N}(0, 1)}_{\text{tabulada}}$$

Este es el modo en que haremos siempre la estimación puntual: *buscaremos una relación en la que intervengan el parámetro desconocido junto con su estimador y de modo que estos se distribuyan según una ley de probabilidad que es bien conocida y a ser posible tabulada.*

De este modo, fijado $\alpha \in (0, 1)$, consideramos la v.a. $Z \rightsquigarrow \mathbf{N}(0, 1)$ y tomamos un intervalo que contenga una masa de probabilidad de $1 - \alpha$. Este intervalo lo queremos tan pequeño como sea posible. Por ello lo mejor es tomarlo simétrico con respecto a la media (0), ya que allí es donde se acumula más masa (véase la figura 8.1). Así las dos colas de la distribución (zonas más alejadas de la media) se repartirán a partes iguales el resto de la masa de probabilidad, α .

Vamos a precisar cómo calcular el intervalo de confianza:

- Sea $z_{\alpha/2}$ el percentil $100 \cdot \frac{\alpha}{2}$ de Z , es decir, aquel valor de \mathcal{R} que deja por debajo de sí la cantidad $\frac{\alpha}{2}$ de la masa de probabilidad de Z , es decir:

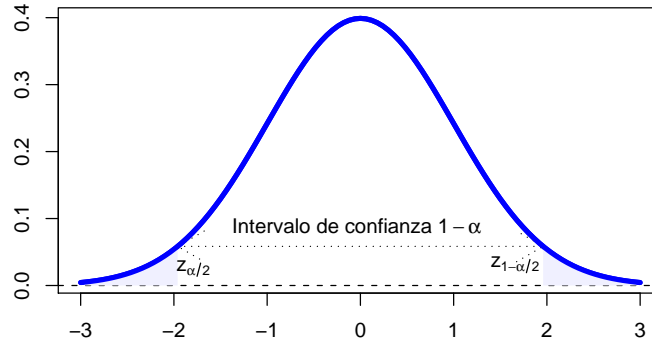


Figura 8.1: La distribución $N(0, 1)$ y el intervalo más pequeño posible cuya probabilidad es $1 - \alpha$. Por simetría, los cuantiles $z_{\alpha/2}$ y $z_{1-\alpha/2}$ sólo difieren en el signo.

$$\mathcal{P}[Z \leq z_{\alpha/2}] = \frac{\alpha}{2}$$

- Sea $z_{1-\alpha/2}$ el percentil $100 \cdot \frac{1-\alpha}{2}$, es decir,

$$\mathcal{P}[Z \leq z_{1-\alpha/2}] = 1 - \frac{\alpha}{2}$$

Es útil considerar en este punto la simetría de la distribución normal, y observar que los percentiles anteriores son los mismos aunque con el signo cambiado:

$$z_{\alpha/2} = -z_{1-\alpha/2}$$

- El intervalo alrededor del origen que contiene la mayor parte de la masa de probabilidad $(1 - \alpha)$ es el intervalo siguiente (cf. Figura 8.1):

$$\left[z_{\alpha/2}, z_{1-\alpha/2} \right] = \left[-z_{1-\alpha/2}, z_{1-\alpha/2} \right]$$

lo que habitualmente escribiremos como:

$$|Z| \leq z_{1-\alpha/2}$$

- De este modo podemos afirmar que existe una probabilidad de $1 - \alpha$ de que al extraer una muestra aleatoria de la variable en estudio, ocurra:

$$\begin{aligned} |Z| \leq z_{1-\alpha/2} &\Rightarrow \\ &\Rightarrow \frac{|\bar{X} - \mu|}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\alpha/2} \\ &\Rightarrow |\bar{X} - \mu| \leq z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \end{aligned}$$

De este modo un intervalo de confianza al nivel $1 - \alpha$ para la esperanza de una normal de varianza conocida es el comprendido entre los valores

$$x_{\alpha/2} = \bar{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$x_{1-\alpha/2} = \bar{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\boxed{\mu = \bar{X} \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}} \quad (8.1)$$

Ejemplo

Se sabe que el peso de los recién nacidos sigue una distribución normal con una desviación típica de 0,75 kg. Si en una muestra aleatoria simple de

100 de ellos se obtiene una media muestral de 3 kg, y una desviación típica de 0,5 kg, calcular un intervalo de confianza para la media poblacional que presente una confianza del 95 %.

Solución: En primer lugar hay que mencionar que la situación planteada no es habitual, ya que si somos capaces de obtener $\sigma = 0,75$, es natural que hayamos podido calcular también μ , y no necesitaríamos una muestra aleatoria para estimar μ confidencialmente. Esto ocurre porque el ejemplo tiene utilidad puramente académica.

Para calcular μ usamos el estadístico:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathbf{N}(0, 1)$$

que como se observa no depende de la dispersión de la muestra, ya que tenemos la “fortuna” de disponer de la dispersión exacta de la población. Esto no es lo habitual en una situación práctica, y como veremos más adelante, el papel de la dispersión exacta de la población (desconocido) será sustituido por el de la dispersión de la muestra.

Un intervalo de confianza al 95 % se calcula teniendo en cuenta que $Z \sim \mathbf{N}(0, 1)$, y dicha distribución presenta un 95 % de probabilidad de ocurrir entre sus cuantiles $z_{0,025} = -1,96$ y $z_{0,975} = 1,96$ (son de signo opuesto por simetría de la distribución normal). Luego con una confianza del 95 % ocurre:

$$-1,96 \leq Z \leq +1,96 \Leftrightarrow |Z| \leq +1,96 \Leftrightarrow |\bar{x} - \mu| \leq +1,96 \frac{\sigma}{\sqrt{n}} \Leftrightarrow |\mu - 3| \leq 0,147$$

Es decir con una confianza del 95 % tenemos que $\mu = 3 \pm 0,147 \text{ kg}$. Esto debe ser interpretado como que la técnica que se usa para el calcular el intervalo de confianza da una respuesta correcta en 95 de cada 100 estudios basados en una muestra aleatoria simple diferente sobre la misma población.

8.2.2. Intervalo para la media (caso general)

El intervalo de confianza al nivel $1 - \alpha$ para la esperanza de una distribución gaussiana cuando sus parámetros son desconocidos es:

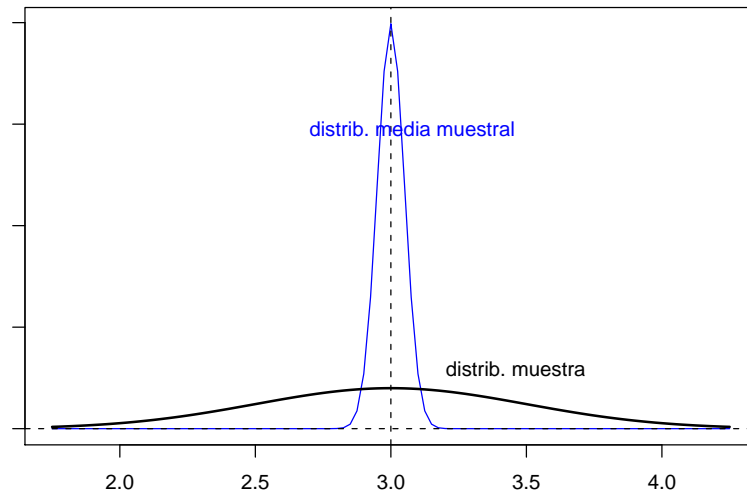


Figura 8.2: Un intervalo de confianza para la media podemos visualizarlo como el que correspondería a una distribución normal con el mismo centro que la de la población, pero cuya desviación está reducida en \sqrt{n} .

$$\mu = \bar{X} \pm t_{n-1, 1-\alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}}$$

Ejemplo

Se sabe que el peso de los recién nacidos sigue una distribución normal. Si en una muestra aleatoria simple de 100 de ellos se obtiene una media muestral de 3 kg, y una desviación típica de 0,5 kg, calcular un intervalo de confianza para la media poblacional que presente una confianza del 95 %.

Solución: Para calcular μ usamos el estadístico:

$$T = \frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \rightsquigarrow t_{n-1}$$

que a diferencia del ejemplo mencionado anteriormente, no depende de σ

(desconocido) si no de su estimación puntual insesgada:

$$\hat{S} = \sqrt{n/(n-1)} S = \sqrt{100/99} 0,5 = 0,503$$

Un intervalo de confianza al 95 % se calcula teniendo en cuenta que $T \sim t_{n-1}$, y dicha distribución presenta un 95 % de probabilidad de ocurrir entre sus cuantiles $T_{n-1;0,025} = -1,98$ y $T_{n-1;0,975} = 1,98$ (son de signo opuesto por simetría de la distribución de Student). Luego con una confianza del 95 % ocurre:

$$|\bar{x} - \mu| \leq +1,98 \frac{\hat{S}}{\sqrt{n}} \Leftrightarrow |\mu - 3| \leq 0,1$$

Es decir con una confianza del 95 % tenemos que $\mu = 3 \pm 0,1 kg$.

Ejemplo

Se quiere estimar un intervalo de confianza al nivel de significación $\alpha = 0,05$ para la altura media μ de los individuos de una ciudad. En principio sólo sabemos que la distribución de las alturas es una v.a. X de distribución normal. Para ello se toma una muestra de $n = 25$ personas y se obtiene

$$\begin{aligned}\bar{x} &= 170 \text{ cm} \\ S &= 10 \text{ cm}\end{aligned}$$

Solución:

Este ejemplo es similar al anterior, pero vamos a resolverlo de una manera más detallada.

En primer lugar, en estadística inferencial, los estadísticos para medir la dispersión más convenientes son los insesgados. Por ello vamos a dejar de lado la desviación típica muestral, para utilizar la cuasidesviación típica:

$$S = 10 \implies \hat{S} = S \sqrt{\frac{n}{n-1}} = 10 \sqrt{\frac{25}{24}} = 10'206$$

$$\mu = 170 \pm 2,06 \cdot \frac{10,206}{5} = 170 \pm 4,204$$

o dicho de forma más precisa: Con un nivel de confianza del 95 % podemos decir que la media poblacional está en el intervalo siguiente:

$$\mu \in [165,796; 174,204]$$

Ejemplo

Este ejemplo se puede considerar como una introducción a los contrastes de hipótesis. La variable IL se presenta en los niños recién nacidos con una distribución normal de media 2,5. En un grupo de 31 niños con sepsis neonatal se encuentra que el valor medio de IL es de $\bar{x} = 1,8$ y $\hat{S} = 0,2$. ¿Cree que presenta la presencia de sepsis neonatal afecta el valor de IL?

Solución: Si no hubiese relación entre la sepsis neonatal y el valor de IL debería ocurrir que el valor de IL en niños nacidos con sepsis se comporte del mismo modo que en los niños normales. Por tanto debería seguir una distribución normal. Además un intervalo de confianza al 95 % para la media de la población de niños sépticos, calculado a partir de los datos de la muestra debería contener (con una confianza del 95 %) a la media de la población de niños normales. Si no fuese así habría que pensar que la variable IL está relacionada con la presencia de sepsis.

Calculemos el intervalo de confianza para la media de los niños con sepsis. Para ello elegimos el estadístico más adecuado a los datos que poseemos:

$$T = \frac{\bar{x} - \mu}{\hat{S}/\sqrt{31}} \rightsquigarrow t_{30}$$

Un intervalo de confianza al 95 % se calcula teniendo en cuenta que $T \rightsquigarrow t_{30}$, y dicha distribución presenta un 95 % de probabilidad de ocurrir entre sus cuantiles $T_{30;0,025} = -2,04$ y $T_{30;0,975} = 2,04$ (son de signo opuesto por simetría de la distribución de Student). Luego con una confianza del 95 % ocurre:

$$|1,8 - \mu| \leq +2,04 \frac{0,2}{\sqrt{31}} \Leftrightarrow |\mu - 1,8| \leq 0,07$$

Por tanto podemos afirmar (con una confianza del 95 %) que la media poblacional de los niños con sepsis estaría comprendida entre los valores 1,73 y 1,87, que están muy alejados de 2,5 (media de los niños normales). Por tanto, podemos afirmar con una confianza del 95 % que están relacionados la IL y la sépsis en niños recién nacidos.

8.2.3. Intervalo de confianza para la varianza

Un intervalo de confianza al nivel $1 - \alpha$ para la varianza de una distribución gaussiana (cuyos parámetros desconocemos) lo obtenemos como

$$\sigma^2 \in \left[\frac{(n-1)\hat{\mathcal{S}}^2}{\chi_{n-1,1-\alpha/2}^2}, \frac{(n-1)\hat{\mathcal{S}}^2}{\chi_{n-1,\alpha/2}^2} \right]$$

Ejemplo

Se estudia la altura de los individuos de una ciudad, obteniéndose en una muestra de tamaño 25 los siguientes valores:

$$\bar{x} = 170 \text{ cm}$$

$$\mathcal{S} = 10 \text{ cm}$$

Calcular un intervalo de confianza con $\alpha = 0,05$ para la varianza σ^2 de la altura de los individuos de la ciudad.

Solución:

$$\sigma^2 \in [63,45; 201,60]$$

Por tanto, para el valor poblacional de la desviación típica tenemos que

$$7,96 \leq \sigma \leq 14,199$$

con una confianza del 95 %, que por supuesto contiene a las estimaciones puntuales $\mathcal{S} = 10$ y $\hat{\mathcal{S}} = 10,206$ calculados sobre la muestra.

8.2.4. Estimación del tamaño muestral

Antes de realizar un estudio de inferencia estadística sobre una variable, lo primero es decidir el número de elementos, n , a elegir en la muestra aleatoria. Para ello consideremos que el estudio se basara en una variable de distribución normal, y nos interesa obtener para un nivel de significación α dado, una precisión (error) d .

Para ello, recordemos que un intervalo de confianza para una media en el caso general se escribe como:

$$\mu = \bar{X} \pm \underbrace{t_{n-1, 1-\alpha/2} \cdot \frac{\hat{\mathcal{S}}}{\sqrt{n}}}_{\text{precisión } d}$$

Si n es suficientemente grande, la distribución **t** de Student se aproxima a la distribución normal. Luego una manera de obtener la precisión buscada consiste en elegir n con el siguiente criterio:

$$n \geq \frac{z_{1-\alpha/2}^2}{d^2} \hat{\mathcal{S}}^2$$

Donde $\hat{\mathcal{S}}^2$ es una estimación puntual *a priori* de la varianza de la muestra. Para obtenerla nos podemos basar en una cota superior conocida por nuestra experiencia previa, o simplemente, tomando una **muestra piloto** que sirve para dar una idea previa de los parámetros que describen una población.

Ejemplo

En los últimos ejemplos se ha estudiado la variable *altura de los individuos de una población*, considerando que ésta es una variable que se

$$\left. \begin{array}{l}
 X \sim \overbrace{\mathbf{N}(\mu, \sigma^2)}^? \leftarrow \text{población normal} \\
 \bar{X} \leftarrow \text{media de la muestra} \\
 \hat{S}^2 \leftarrow \text{cuasivarianza de la muestra} \\
 n \leftarrow \text{tamaño de la muestra}
 \end{array} \right\}$$

Intervalos de confianza

Para μ cuando σ^2 se conoce	$\mu \in \bar{X} \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
Para μ cuando σ^2 no se conoce	$\mu \in \bar{X} \pm t_{n-1, 1-\alpha/2} \cdot \frac{\hat{S}}{\sqrt{n}}$
Para σ^2 con μ desconocido	$\sigma^2 \in \left[\frac{(n-1)\hat{S}^2}{\chi_{n-1, 1-\alpha/2}^2}, \frac{(n-1)\hat{S}^2}{\chi_{n-1, \alpha/2}^2} \right]$

Cuadro 8.1: Intervalos de confianza para los parámetros de una población normal, a partir de una muestra aleatoria simple de la misma.

distribuye de modo gaussiana.

Para ello se tomó una muestra de 25 individuos (que podemos considerar piloto), que ofreció los siguientes resultados:

$$\begin{aligned}\bar{x} &= 170 \text{ cm} \\ \mathcal{S} &= 10 \text{ cm}\end{aligned}$$

Calcular el tamaño que debería tener una muestra para que se obtuviese un intervalo de confianza para la media poblacional con un nivel de significación $\alpha = 0,01$ (al 99 %) y con una precisión de $d = 1$ cm.

Solución:

Obsérvese que sobre la muestra piloto, el error cometido al estimar el intervalo al 95 % fue aproximadamente de 4'2 cm por lo que si buscamos un intervalo de confianza tan preciso, el tamaño de la muestra, n , deberá ser bastante mayor. En este caso se obtiene:

$$n \approx \frac{z_{0,995}^2 \cdot 10,206^2}{1^2} = 2,58^2 \cdot 10,206^2 \approx 694$$

Por tanto, si queremos realizar un estudio con toda la precisión requerida en el enunciado se debería tomar una muestra de 694 individuos. Esto es una indicación de gran utilidad antes de comenzar el estudio. Una vez que el muestreo haya sido realizado, debemos confirmar que el error para el nivel de significación dado es inferior o igual a 1 cm, utilizando la muestra obtenida.

8.2.5. Intervalos para la diferencia de medias de dos poblaciones

Consideremos el caso en que tenemos dos poblaciones de modo que el carácter que estudiamos en ambas (X_1 y X_2) son v.a. distribuidas según leyes gaussianas

$$X_1 \rightsquigarrow \mathbf{N}(\mu_1, \sigma_1^2)$$

$$X_2 \rightsquigarrow \mathbf{N}(\mu_2, \sigma_2^2)$$

En cada una de estas poblaciones se extrae mediante muestreo aleatorio simple, muestras que no tienen por que ser necesariamente del mismo tamaño (respectivamente n_1 y n_2)

$$\begin{aligned}\vec{X}_1 &\equiv X_{11}, X_{12}, \dots, X_{1n_1} \\ \vec{X}_2 &\equiv X_{21}, X_{22}, \dots, X_{2n_2}\end{aligned}$$

Podemos plantearnos a partir de las muestras el saber qué diferencias existen entre las medias de ambas poblaciones, o por ejemplo estudiar las relación existente entre sus dispersiones respectivas. A ello vamos a dedicar los siguientes puntos.

Intervalo para la diferencia de medias homocedáticas

Supongamos que dos poblaciones tengan varianzas idénticas (**homocedasticidad**), σ^2 . Es decir

$$\sigma^2 = \sigma_1^2 = \sigma_2^2.$$

Por razones análogas a las expuestas en el caso de una población una población, se tiene que

$$\left. \begin{aligned} \chi_{n_1-1}^2 &= \frac{(n_1-1)\hat{S}_1^2}{\sigma} \rightsquigarrow \chi_{n_1-1}^2 \\ \chi_{n_2-1}^2 &= \frac{(n_2-1)\hat{S}_2^2}{\sigma} \rightsquigarrow \chi_{n_2-1}^2 \end{aligned} \right\} \chi^2_{\text{reprod.}} \Rightarrow \chi_{n_1+n_2-2}^2 = \chi_{n_1-1}^2 + \chi_{n_2-1}^2 \rightsquigarrow \chi_{n_1+n_2-2}^2$$

De manera similar al caso de la media de una población, si las varianzas fuesen conocidas, podemos definir la v.a.

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \rightsquigarrow \mathbf{N}(0, 1)$$

Cuando las varianzas de las poblaciones son desconocidas, pero podemos asumir que al menos son iguales, el siguiente estadístico se distribuye como una t de Student con $n_1 + n_2 - 2$ grados de libertad:

$$T_{n_1+n_2-2} = \frac{Z}{\sqrt{\frac{1}{n_1+n_2-2} \chi_{n_1+n_2-2}^2}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{S} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightsquigarrow t_{n_1+n_2-2} \quad (8.2)$$

donde se ha definido a \hat{S}^2 como la **cuasivarianza muestral ponderada** de \hat{S}_1^2 y \hat{S}_2^2

$$\hat{S}^2 = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}$$

Si $1 - \alpha$ es el nivel de significación con el que deseamos establecer el intervalo para la diferencia de las dos medias, calculamos el valor $t_{n_1+n_2-2, 1-\alpha/2}$ que deja por encima de si $\alpha/2$ de la masa de probabilidad de $T_{n_1+n_2-2}$

$$\mathcal{P}[T_{n_1+n_2-2} > t_{n_1+n_2-2, 1-\alpha/2}] = \frac{\alpha}{2} \Leftrightarrow \mathcal{P}[|T_{n_1+n_2-2}| \leq t_{n_1+n_2-2, 1-\alpha/2}] = 1 - \alpha$$

Repetiendo un proceso que ya hemos realizado en ocasiones anteriores, tenemos una probabilidad de $1 - \alpha$ de que a extraer una muestra aleatoria simple ocurra:

$$\begin{aligned} |T_{n_1+n_2-2}| \leq t_{n_1+n_2-2, 1-\alpha/2} &\Leftrightarrow \\ \Leftrightarrow \frac{|(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)|}{\hat{S} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} &\leq t_{n_1+n_2-2, 1-\alpha/2} \\ \Leftrightarrow |\mu_1 - \mu_2| \leq (\bar{X}_1 - \bar{X}_2) + t_{n_1+n_2-2, 1-\alpha/2} \cdot \hat{S} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \end{aligned}$$

Luego el intervalo de confianza al nivel $1 - \alpha$ para la diferencia de esperanzas de dos poblaciones con la misma varianza (aunque esta sea desconocida) es:

$$\mu_1 - \mu_2 = (\bar{X}_1 - \bar{X}_2) \pm t_{n_1+n_2-2, 1-\alpha/2} \cdot \hat{S} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Ejemplo

Queremos estudiar la influencia que puede tener el tabaco con el peso de los niños al nacer. Para ello se consideran dos grupos de mujeres embarazadas (unas que fuman y otras que no) y se obtienen los siguientes datos sobre el peso X , de sus hijos:

$$\left\{ \begin{array}{ll} \text{Madres fumadoras} & \rightarrow n_1 = 35 \text{ mujeres, } \bar{x}_1 = 3,6 \text{ Kg } \hat{S}_1 = 0,5 \text{ Kg} \\ \text{Madres no fumadoras} & \rightarrow n_2 = 27 \text{ mujeres, } \bar{x}_2 = 3,2 \text{ Kg } \hat{S}_2 = 0,8 \text{ Kg} \end{array} \right.$$

En ambos grupos los pesos de los recién nacidos provienen de sendas distribuciones normales de medias desconocidas, y con varianzas que si bien son desconocidas, podemos suponer que son las mismas. Calcular en cuanto influye el que la madre sea fumadora en el peso de su hijo.

Solución:

Si X_1 es la v.a. que describe el peso de un niño que nace de madre no fumadora, y X_2 el de un hijo de madre fumadora, se tiene por hipótesis que

$$\exists \mu_1, \mu_2, \sigma^2, \text{ tales que } \left\{ \begin{array}{l} X_1 \rightsquigarrow \mathbf{N}(\mu_1, \sigma^2) \\ X_2 \rightsquigarrow \mathbf{N}(\mu_2, \sigma^2) \end{array} \right.$$

Si queremos estimar en cuanto influye el que la madre sea fumadora en el peso de su hijo, podemos estimar un intervalo de confianza para $\mu_1 - \mu_2$, lo que nos dará la diferencia de peso esperado entre un niño del primer grupo y otro del segundo. El estadístico que se ha de aplicar para esta cuestión es:

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\hat{S} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightsquigarrow t_{n_1+n_2-2} = t_{35+27-2} = t_{60}$$

donde

$$\hat{S}^2 = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2} = \frac{34 \cdot 0,5^2 + 26 \cdot 0,8^2}{60} = 0,419 \implies \hat{S} = 0,6473$$

Consideramos un nivel de significación que nos parezca aceptable, por ejemplo $\alpha = 0,05$, y el intervalo buscado se obtiene a partir de:

$$\frac{|\overbrace{(3,6 - 3,2)}^{0,4} - (\mu_1 - \mu_2)|}{\underbrace{0,6473 \sqrt{\frac{1}{35} + \frac{1}{27}}}_{0,1658}} \leq t_{60;1-0,05/2} = t_{60;0,975} = 2$$

$$\implies \mu_1 - \mu_2 = 0,4 \pm 2 \cdot 0,1658 \implies \mu_1 - \mu_2 = 0,4 \pm 0,3316$$

con lo cual se puede decir que un intervalo de confianza para el peso esperado en que supera un hijo de madre no fumadora al de otro de madre fumadora está comprendido con un nivel de confianza del 95% entre los 0,068 Kg y los 0,731 Kg.

$$\left\{ \begin{array}{l} \overbrace{X_1 \rightsquigarrow \mathbf{N}(\mu_1, \sigma_1^2)}^? \\ X_2 \rightsquigarrow \underbrace{\mathbf{N}(\mu_2, \sigma_2^2)}_? \end{array} \right\} \leftarrow \begin{array}{l} \text{poblaciones normales} \\ \\ \overline{X}_1, \overline{X}_2 \leftarrow \text{medias de las muestras} \\ \hat{\mathcal{S}}_1^2, \hat{\mathcal{S}}_2^2 \leftarrow \text{cuasivarianzas de las muestras} \\ n_1, n_2 \leftarrow \text{tamaños de las muestras} \end{array}$$

Intervalos de confianza para $\mu_1 - \mu_2$

Si $\sigma_1^2 = \sigma_2^2$ (desconocidos)	$\mu_1 - \mu_2 \in (\overline{X}_1 - \overline{X}_2) \pm t_{n_1+n_2-2, 1-\alpha/2} \cdot \hat{\mathcal{S}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
Si $\sigma_1^2 \neq \sigma_2^2$ (desconocidos)	$\mu_1 - \mu_2 \in (\overline{X}_1 - \overline{X}_2) \pm t_{f, 1-\alpha/2} \cdot \sqrt{\frac{\hat{\mathcal{S}}_1^2}{n_1} + \frac{\hat{\mathcal{S}}_2^2}{n_2}}$

$$\text{donde} \left\{ \begin{array}{l} \hat{\mathcal{S}}^2 = \frac{(n_1 - 1)\hat{\mathcal{S}}_1^2 + (n_2 - 1)\hat{\mathcal{S}}_2^2}{n_1 + n_2 - 2} \\ \\ f = \frac{\left(\frac{\hat{\mathcal{S}}_1^2}{n_1} + \frac{\hat{\mathcal{S}}_2^2}{n_2}\right)^2}{\frac{1}{n_1 + 1} \left(\frac{\hat{\mathcal{S}}_1^2}{n_1}\right)^2 + \frac{1}{n_2 + 1} \left(\frac{\hat{\mathcal{S}}_2^2}{n_2}\right)^2} - 2 \leftarrow \text{Welch.} \end{array} \right.$$

Cuadro 8.2: Intervalos de confianza para la diferencia de las medias de dos poblaciones normales, calculados a partir de sendas muestras independientes de cada una de ellas. Los resultados dependen de que podamos suponer cierta o no la condición de homocedasticidad.

8.3. Intervalos de confianza para variables dicotómicas

Cuando tenemos una variable dicotómica (o de Bernoulli) a menudo interesa saber en qué proporción de casos, p ocurre el éxito en la realización de un experimento. También nos puede interesar el comparar la diferencia existente entre las proporciones en distintas poblaciones. También es de interés calcular para un nivel de significación dado, el tamaño muestral necesario para calcular un intervalo de confianza de cuyo radio sea menor que cierta cantidad.

8.3.1. Intervalo para una proporción

Sean $X_1, \dots, X_n \sim \mathbf{Ber}(p)$. Si queremos estimar el parámetro p , la manera más natural de hacerlo consiste en definir la suma de estas —lo que nos proporciona una distribución Binomial

$$X = X_1 + \dots + X_n \sim \mathbf{B}(n, p)$$

y tomar como estimador suyo la v.a.

$$\hat{p} = \frac{X}{n}.$$

Es decir, tomamos como estimación de p la proporción de éxitos obtenidos en las n pruebas. \hat{p} .

La distribución del número de éxitos es binomial, y puede ser aproximada a la normal cuando el tamaño de la muestra n es grande, y p no es una cantidad muy cercana a cero o uno:

$$X \sim \mathbf{B}(n, p) \Rightarrow X \overset{\sim}{\approx} \mathbf{N}(np, npq)$$

El estimador \hat{p} no es más que un cambio de escala de X , por tanto

$$\hat{p} = \frac{X}{n} \overset{\sim}{\approx} \mathbf{N}\left(p, \frac{pq}{n}\right) \quad \Rightarrow \quad \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \approx Z \sim \mathbf{N}(0, 1)$$

Esta expresión presenta dificultades para el cálculo, siendo más cómodo sustituirla por la siguiente aproximación:

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \approx Z \sim \mathbf{N}(0, 1)$$

Para encontrar el intervalo de confianza al nivel de significación α para p se considera el intervalo que hace que la distribución de $Z \sim \mathbf{N}(0, 1)$ deje la probabilidad α fuera del mismo. Es decir, se considera el intervalo cuyos extremos son los cuantiles $\alpha/2$ y $1 - \alpha/2$. Así se puede afirmar con una confianza de $1 - \alpha$ que:

$$p = \hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \text{ con una confianza de } 1 - \alpha$$

Ejemplo

Se quiere estimar el resultado de un referéndum mediante un sondeo. Para ello se realiza un muestreo aleatorio simple con $n = 100$ personas y se obtienen 35 % que votarán a favor y 65 % que votarán en contra (suponemos que no hay indecisos para simplificar el problema a una variable dicotómica). Con un nivel de significación del 5 %, calcule un intervalo de confianza para el verdadero resultado de las elecciones.

Solución: Dada una persona cualquiera (i) de la población, el resultado de su voto es una variable dicotómica:

$$X_i \sim \mathbf{Ber}(p)$$

El parámetro a estimar en un intervalo de confianza con $\alpha = 0,05$ es p , y tenemos sobre una muestra de tamaño $n = 100$, la siguiente estimación puntual de p :

$$\hat{p} = \frac{35}{100} = 0,35 \implies \hat{q} = 0,65$$

El intervalo de confianza buscado es:

$$p = 0,65 \pm 0,0935$$

Por tanto, tenemos con esa muestra un error aproximado de 9,3 puntos al nivel de confianza del 95 %.

8.3.2. Elección del tamaño muestral para una proporción

En un ejemplo previo con una muestra de 100 individuos se realizó una estimación confidencial, con un 95 % de confianza, del porcentaje de votantes a una cuestión en un referéndum, obteniéndose un margen de error de 9,3 puntos.

Si pretendemos reducir el error a 1 punto y queremos aumentar el nivel de confianza hasta el 97 % ($\alpha = 0'03$) hemos de tomar una muestra lógicamente de mayor tamaño, N .

Un valor de N que satisface nuestros requerimientos con respecto al error es:

$$N \geq \hat{p}\hat{q} \frac{z_{1-\alpha/2}^2}{\text{error}^2}$$

Si en un principio no tenemos una idea sobre que valores puede tomar p , debemos considerar el peor caso posible, que es en el que se ha de estimar el tamaño muestral cuando $p = q = 1/2$. Así:

$$N \geq \frac{1}{4} \frac{z_{1-\alpha/2}^2}{\text{error}^2} \text{ cuando no se tiene estimación de } p$$

Ejemplo

Se quiere estimar el resultado de un referéndum mediante un sondeo, y sin tener una idea sobre el posible resultado del mismo, se desea conocer el tamaño de muestra que se ha de tomar para obtener un intervalo al 97 % de confianza, con un error del 1

Solución:

Como no se tiene una idea previa del posible resultado del referéndum, hay que tomar un tamaño de muestra, N , que se calcula mediante:

$$N \geq \frac{1}{4} \frac{z_{0,985}^2}{0,01^2} = \frac{0,25 \cdot 2,17^2}{0,01^2} = 11,773$$

Así para tener un resultado tan fiable, el número de personas a entrevistar debe ser muy elevado —lo que puede volver excesivamente costoso

el sondeo.

8.3.3. Intervalo para la diferencia de dos proporciones

Vamos a considerar que tenemos dos poblaciones de modo que en cada una de ellas estudiamos una v.a. dicotómica (Bernoulli) de parámetros respectivos p_1 y p_2 . De cada población vamos a extraer muestras de tamaño n_1 y n_2

$$\begin{aligned}\vec{X}_1 &\equiv X_{11}, X_{12}, \dots, X_{1n_1} \\ \vec{X}_2 &\equiv X_{21}, X_{22}, \dots, X_{2n_2}\end{aligned}$$

Entonces

$$\begin{aligned}X_1 &= \sum_{i=1}^{n_1} X_{1i} \rightsquigarrow \mathbf{B}(n_1, p_1) \\ X_2 &= \sum_{i=1}^{n_2} X_{2i} \rightsquigarrow \mathbf{B}(n_2, p_2)\end{aligned}$$

Si las muestras son suficientemente grandes ocurre que una aproximación para un intervalo de confianza al nivel $1 - \alpha$ para la diferencia de proporciones de dos poblaciones es:

$$p_1 - p_2 \in (\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Ejemplo

Se cree que la osteoporosis está relacionada con el sexo. Para ello se elige una muestra de 100 hombres de más de 50 años y una muestra de 200 mujeres en las mismas condiciones. Se obtiene que 10 hombres y 40 mujeres con algún grado de osteoporosis. ¿Qué podemos concluir con una confianza del 95 %? **Solución:**

Llamamos p_1 a la incidencia de la osteoporosis en las mujeres de más de 50 años y p_2 a la de los hombres. Calculemos un intervalo de confianza para la diferencia $(p_1 - p_2)$. Si 0 no forma parte de dicho intervalo con una confianza del 95 % podemos decir que p_1 es diferente a p_2 (con tal grado de confianza, por supuesto).

La estimación puntual insesgada que podemos hacer de ambos parámetros a partir de los datos muestrales son:

$$\begin{aligned}\hat{p}_1 &= 40/200 = 0,2 \\ \hat{p}_2 &= 10/100 = 0,1\end{aligned}$$

$$(p_1 - p_2) = (0,2 - 0,1) \pm \sqrt{\frac{0,2 \times 0,8}{200} + \frac{0,1 \times 0,9}{100}} = 0,08$$

Es decir, tenemos una confianza del 95 % en la afirmación de que la diferencia entre la incidencia de osteoporosis en mujeres y hombres está entre 0,02 (2 %) y 0,18 (18 %).

Obsérvese que como 0 % no es un valor de dicho intervalo puede concluirse con una confianza del 95 % que hay diferente incidencia de osteoporosis en hombres que en mujeres para las personas de más de 50 años. Esta conclusión es algo más pobre de lo que hemos obtenido con el intervalo de confianza, pero visto de esta manera, este ejemplo puede considerarse como una introducción a los contrastes de hipótesis.

8.4. Problemas

Ejercicio 8.1. Se ha medido el volumen diario de bilis, expresado en litros, en 10 individuos sanos, obteniéndose

0,98; 0,85; 0,77; 0,92; 1,12; 1,06; 0,89; 1,01; 1,21; 0,77.

¿Cuanto vale la producción diaria media de bilis en individuos sanos suponiendo que la muestra ha sido obtenida por muestreo aleatorio simple sobre una población normal?

Ejercicio 8.2. La cantidad mínima requerida para que un anestésico surta efecto en una intervención quirúrgica fue por término medio de 50 mg, con una desviación típica de 10,2 mg, en una muestra de 60 pacientes. Obtener un intervalo de confianza para la media al 99 %, suponiendo que la muestra fue extraída mediante muestreo aleatorio simple sobre una población normal.

Ejercicio 8.3. Un investigador está interesado en estimar la proporción de muertes debidas a cáncer de estómago en relación con el número de defunciones por cualquier tipo de neoplasia. Su experiencia le indica que sería sorprendente que tal proporción supere el valor de $1/3$. ¿Qué tamaño de muestra debe tomar para estimar la anterior proporción, con una confianza del 99 %, para que el valor estimado no difiera del valor real en más de 0,03?

Ejercicio 8.4. Se desea realizar una estimación confidencial de la varianza de la estatura de los niños varones de 10 años de una ciudad con una confianza del 95 %. ¿Cuál será dicho intervalo si se toma una muestra de 101 niños al azar, entre todos los que reúnen las características deseadas, y medimos sus estaturas, y se obtienen las siguientes estimaciones puntuales: $\bar{x} = 138,6 \text{ cm}$, $S^2 = 29,16 \text{ cm}^2$?

Ejercicio 8.5. Un cardiólogo se encuentra interesado en encontrar límites de confianza al 90 %, para la presión sistólica tras un cierto ejercicio físico. Obtenerlos si en 50 individuos se obtuvo $\bar{x} = 13$, $S = 3$ y suponemos que el comportamiento de la v.a. es normal.

Ejercicio 8.6. En una muestra de 25 bebés varones de 12 semanas de vida, se obtuvo un peso medio de 5.900 gr y una desviación típica de 94 gr.

1. Obtener un intervalo de confianza (al 95 %) para el peso medio poblacional.
2. ¿Cuántos niños habría que tomar para estimar dicha media con una precisión de 15 gr?

Ejercicio 8.7. En un determinado servicio de odontología se sabe que el 22 % de las visitas llevan consigo una extracción dentaria inmediata. En cierto año, de 2.366 visitas, 498 dieron lugar a una extracción inmediata. ¿Entran en contradicción las cifras de ese año con el porcentaje establecido de siempre?

Ejercicio 8.8. Sólo una parte de los pacientes que sufren un determinado síndrome neurológico consiguen una curación completa; Si de 64 pacientes observados se han curado 41, dar una estimaciones puntual y un intervalos de la proporción de los que sanan. ¿Qué número de enfermos habría que observar para estimar la proporción de curados con un error inferior a 0,05 y una confianza del 95 %?

Ejercicio 8.9. Se desea estimar el tiempo medio de sangría en fumadores de más de 20 cigarrillos diarios, con edades comprendidas entre 35 y 40 años, con una precisión de 5 segundos. Ante la ausencia de cualquier información acerca de la variabilidad del tiempo de sangría es este tipo de individuos, se tomó una muestra preliminar de 5 individuos, en los que se obtuvieron los siguientes tiempos (en segundos):

97, 80, 67, 91, 73.

Determinar el tamaño mínimo de muestra, al 95 %, para cumplir el objetivo anterior.

Ejercicio 8.10. En una determinada región se tomó una muestra aleatoria de 125 individuos, de los cuales 12 padecían afecciones pulmonares.

1. Estímese la proporción de afecciones pulmonares en dicha región.
2. Si queremos estimar dicha proporción con un error máximo del 4 %, para una confianza del 95 %, ¿qué tamaño de muestra debemos tomar?

Ejercicio 8.11. En una muestra de tabletas de aspirinas, de las cuales observamos su peso expresado en gramos, obtenemos:

1,19; 1,23; 1,18; 1,21; 1,27; 1,17; 1,15; 1,14; 1,19; 1,2

Suponiendo la Normalidad para esta distribución de pesos, determinar un intervalo al 80 % de confianza para la varianza.

Ejercicio 8.12. Se quiere estimar la incidencia de la hipertensión arterial en el embarazo. ¿Cuántas embarazadas tenemos que observar para, con una confianza del 95 %, estimar dicha incidencia con un error del 2 % en los siguientes casos:

1. Sabiendo que un sondeo previo se ha observado un 9 % de hipertensas.
2. Sin ninguna información previa.

Capítulo 9

Contrastes de hipótesis

9.1. Introducción

Hasta ahora hemos estudiado cómo a partir de una muestra de una población podemos obtener una estimación puntual o bien establecer un intervalo más o menos aproximado para encontrar los parámetros que rigen la ley de probabilidad de una v.a. definida sobre la población. Es lo que denominábamos *estimación puntual* y *estimación confidencial* respectivamente.

Pueden presentarse en la práctica, situaciones en las que exista una teoría preconcebida relativa a la característica de la población sometida a estudio. Tal sería el caso, por ejemplo si pensamos que un tratamiento nuevo puede tener un porcentaje de mejoría mayor que otro estándar, o cuando nos planteamos si los niños de las distintas comunidades españolas tienen la misma altura. Este tipo de circunstancias son las que nos llevan al estudio de la parcela de la Estadística Inferencial que se recoge bajo el título genérico de **Contraste de Hipótesis**. Implica, en cualquier investigación, la existencia de dos teorías o hipótesis implícitas, que denominaremos hipótesis nula e hipótesis alternativa, que de alguna manera reflejarán esa idea a priori que tenemos y que pretendemos contrastar con la “realidad”. De la misma manera aparecen, implícitamente, diferentes tipos de errores que podemos cometer durante el procedimiento. No podemos olvi-

dar que, habitualmente, el estudio y las conclusiones que obtengamos para una población cualquiera, se habrán apoyado exclusivamente en el análisis de sólo una parte de ésta. De la probabilidad con la que estemos dispuestos a asumir estos errores, dependerá, por ejemplo, el tamaño de la muestra requerida. Desarrollamos en este capítulo los contrastes de hipótesis para los parámetros más usuales que venimos estudiando en los capítulos anteriores: medias, varianzas y proporciones, para una o dos poblaciones. Los contrastes desarrollados en este capítulo se apoyan en que los datos de partida siguen una distribución normal.

Los **contrastos de significación** se realizan:

- suponiendo *a priori* que la ley de distribución de la población es conocida.
- Se extrae una muestra aleatoria de dicha población.
- Si la distribución de la muestra es “diferente” de la distribución de probabilidad que hemos asignado *a priori* a la población, concluimos que probablemente sea errónea la suposición inicial.

Ejemplo

Supongamos que debemos realizar un estudio sobre la altura media de los habitantes de cierto pueblo de España. Antes de tomar una muestra, lo lógico es hacer la siguiente suposición *a priori*, (hipótesis que se desea contrastar y que denotamos H_0):

H_0 : La altura media no difiere de la del resto del país.

Al obtener una muestra de tamaño $n = 8$, podríamos encontrarnos ante uno de los siguientes casos:

1. Muestra = {1,50 ;1,52; 1,48; 1,55; 1,60; 1,49; 1,55; 1,63}

2. Muestra = {1,65; 1,80; 1,73; 1,52; 1,75; 1,65; 1,75; 1,78}

Intuitivamente, en el caso **a** sería lógico suponer que salvo que la muestra obtenida sobre los habitantes del pueblo sea muy poco representativa¹, la hipótesis H_0 debe ser rechazada. En el caso **b** tal vez no podamos afirmar con rotundidad que la hipótesis H_0 sea cierta, sin embargo no podríamos descartarla y la admitimos por una cuestión de simplicidad.

Este ejemplo sirve como introducción de los siguientes conceptos: En un contraste de hipótesis (también denominado *test de hipótesis* o *Contraste de significación*) se decide si cierta hipótesis H_0 que denominamos **hipótesis nula** puede ser rechazada o no a la vista de los datos suministrados por una muestra de la población. Para realizar el contraste es necesario establecer previamente una **hipótesis alternativa** (H_1) que será admitida cuando H_0 sea rechazada. Normalmente H_1 es la negación de H_0 , aunque esto no es necesariamente así.

El procedimiento general consiste en definir un estadístico T relacionado con la hipótesis que deseamos contrastar. A éste lo denominamos **estadístico del contraste**. A continuación suponiendo que H_0 es verdadera se calcula un intervalo de denominado intervalo de aceptación² de la hipótesis nula, (T_i, T_s) de manera que al calcular sobre la muestra $T = T_{exp}$ el criterio a seguir sea:

$$\left\{ \begin{array}{ll} \text{Si } T_{exp} \in (T_i, T_s) & \implies \text{no rechazamos } H_0 \quad (\nRightarrow \text{rechazamos } H_1); \\ \text{Si } T_{exp} \notin (T_i, T_s) & \implies \text{rechazamos } H_0 \text{ y aceptamos } H_1 \end{array} \right.$$

El intervalo de aceptación o más precisamente, de no rechazo de la hipótesis nula, se establece fijando una cantidad α suficientemente pequeña denominada **nivel de significación**, de modo que la probabilidad de que el estadístico del contraste tome un valor fuera del mismo — **región crítica** —

$$\text{región crítica} \equiv \mathcal{C} = \mathbb{R} \setminus (T_i, T_s)$$

¹Esto ocurre con muy baja probabilidad en un muestreo aleatorio simple cuando el número de observaciones es alto

²Se entiende la palabra “aceptación” como en el sentido de “no rechazo”.

cuando la hipótesis nula es cierta sea inferior o al $100 \cdot \alpha \%$; Esto se ha de entender como sigue:

Si H_0 es correcta el criterio de rechazo sólo se equivoca con probabilidad α , que es la probabilidad de que una muestra ofrezca un valor del estadístico del contraste extraño (en la región crítica).

La decisión de rechazar o no la hipótesis nula está al fin y al cabo basado en la elección de una muestra tomada al azar, y por tanto es posible cometer decisiones erróneas. Los errores que se pueden cometer se clasifican como sigue:

Error de tipo *I*: Es el error que consiste en rechazar H_0 cuando es cierta. La probabilidad de cometer este error es lo que anteriormente hemos denominado nivel de significación. Es una costumbre establecida el denotarlo siempre con la letra α

$$\alpha = \mathcal{P} \left[\text{rechazar } H_0 | H_0 \text{ es cierta} \right] = \mathcal{P} \left[\text{aceptar } H_1 | H_0 \text{ es cierta} \right].$$

Error de tipo *II*: Es el error que consiste en no rechazar H_0 cuando es falsa. La probabilidad de cometer este error la denotamos con la letra β

$$\beta = \mathcal{P} \left[\text{no rechazar } H_0 | H_0 \text{ es falsa} \right] \left(\neq \mathcal{P} \left[\text{no rechazar } H_0 | H_1 \text{ es cierta} \right] \right)$$

9.1.1. Observaciones

1. Los errores de tipo *I* y *II* no están relacionados más que del siguiente modo: Cuando α decrece β crece. Por tanto no es posible encontrar tests que hagan tan pequeños como queramos ambos errores simultáneamente. De este modo es siempre necesario *privilegiar* a una de las hipótesis, de manera que no será rechazada, a menos que su falsedad se haga muy evidente. En los contrastes, la hipótesis privilegiada es H_0 que sólo será rechazada cuando la evidencia de su falsedad supere el umbral del $100 \cdot (1 - \alpha) \%$.
2. Al tomar α muy pequeño tendremos que β se puede aproximar a uno. Lo ideal a la hora de definir un test es encontrar un compromiso sa-

tisfactorio entre α y β (aunque siempre a favor de H_0). Denominamos **potencia de un contraste** a la cantidad $1 - \beta$, es decir

$$\text{potencia del contraste} \equiv 1 - \beta = \mathcal{P} \left[\text{rechazar } H_0 | H_0 \text{ es falsa} \right]$$

	no rechazar H_0	rechazar H_0
H_0 es cierta	Correcto Probabilidad $1 - \alpha$	Error tipo I Probabilidad α
H_0 es falsa	Error tipo II Probabilidad β	Correcto Probabilidad $1 - \beta$

3. En el momento de elegir una hipótesis privilegiada podemos en principio dudar entre si elegir una dada o bien su contraria. Criterios a tener en cuenta en estos casos son los siguientes:

- **Simplicidad científica:** A la hora de elegir entre dos hipótesis científicamente razonables, tomaremos como H_0 aquella que sea más simple.
- **Las consecuencias de equivocarnos:** Por ejemplo al juzgar el efecto que puede causar cierto tratamiento médico que está en fase de experimentación, en principio se ha de tomar como hipótesis nula aquella cuyas consecuencias por no rechazarla siendo falsa son menos graves, y como hipótesis alternativa aquella en la que el aceptarla siendo falsa trae peores consecuencias. Es decir,

$$\left\{ \begin{array}{l} H_0 : \text{ el paciente empeora o queda igual ante el tratamiento} \\ H_1 : \text{ el paciente mejora con el tratamiento} \end{array} \right.$$

Otro ejemplo claro es cuando acaban de instalar un nuevo ascensor en el edificio que habitamos y queremos saber si el ascensor caerá o no al vacío cuando nosotros estemos dentro. Una persona prudente es la que espera a que un número suficiente de vecinos suyos hayan usado el ascensor (muestra aleatoria) y realiza un test del tipo

$$\begin{cases} H_0 : & \text{el ascensor se caerá} \\ H_1 : & \text{el ascensor no se caerá} \end{cases}$$

y sólo aceptará la hipótesis alternativa para $\alpha \approx 0$ aunque para ello tenga que ocurrir que $\beta \approx 1$, ya que las consecuencias del error de tipo *I* (ir al hospital) son mucho más graves que las del error del tipo *II* (subir a pie varios pisos).

Es decir a la hora de decidirse por una de las dos hipótesis no basta con elegir la más probable (nadie diría “voy a tomar el ascensor pues la probabilidad de que no se caiga es del 60 %”). Hay que elegir siempre la hipótesis H_0 a menos que la evidencia a favor de H_1 sea muy *significativa*.

Volviendo al ejemplo de la estatura de los habitantes de un pueblo, un estadístico de contraste adecuado es \bar{X} . Si la hipótesis H_0 fuese cierta se tendría que

$$\bar{X} \rightsquigarrow \mathbf{N} \left(\mu, \frac{\sigma^2}{n} \right)$$

(suponiendo claro está que la distribución de las alturas de los españoles siga una distribución normal de parámetros conocidos, por ejemplo³)

$$\mathbf{N}(\mu = 1,74, \sigma^2 = 10^2)$$

Denotemos mediante μ_0 el verdadero valor de la media en el pueblo que estudiamos. Como la varianza de \bar{X} es pequeña para grandes valores de n , lo lógico es pensar que si el valor obtenido con la muestra $\bar{X} = \bar{x}$ está muy alejado de $\mu = 1,74$ (región crítica), entonces

- o bien la muestra es muy extraña si H_0 es cierta (probabilidad α);

³Estos valores de la media y la desviación típica no han sido tomados de ningún estudio.

- o bien la hipótesis H_0 no es cierta.

Concretamente en el caso **a**, donde la muestra es

$$\text{Muestra} = \{1, 50; 1, 52; 1, 48; 1, 55; 1, 60; 1, 49; 1, 55; 1, 63\}$$

el contraste de hipótesis conveniente es:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

En este caso H_1 no es estrictamente la negación de H_0 . Esto dará lugar a un **contraste unilateral**, que son aquellos en los que la región crítica está formada por un sólo intervalo:

$$\begin{aligned} \text{Intervalo de no rechazo de } H_0 &\equiv (T_i, +\infty) \\ \text{Región crítica} &\equiv (-\infty, T_i] \end{aligned}$$

En el caso **b**, donde la muestra es

$$\text{Muestra} = \{1, 65; 1, 80; 1, 73; 1, 52; 1, 75; 1, 65; 1, 75; 1, 78\}$$

el contraste de hipótesis que deberíamos realizar es:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

Como vemos, ahora sí se puede decir que H_1 es la negación de H_0 . Esto es un **contraste bilateral**, que son aquellos en los que la región crítica está formada por dos intervalos separados:

$$\begin{aligned} \text{Intervalo donde no se rechaza } H_0 &\equiv (T_i, T_s) \\ \text{Región crítica} &\equiv (-\infty, T_i] \cup [T_s, +\infty) \end{aligned}$$

Los últimos conceptos que introducimos son:

Hipótesis simple: Aquella en la que se especifica un único valor del parámetro. Este es el caso de las hipótesis nulas en los dos últimos contrastes mencionados.

Hipótesis compuesta: Aquella en la que se especifica más de un posible valor del parámetro. Por ejemplo tenemos que son compuestas las hipótesis alternativas de esos mismos contrastes.

9.2. Contrastes paramétricos en una población normal

Supongamos que la característica X que estudiamos sobre la población sigue una distribución normal y tomamos una muestra de tamaño n

$$\vec{X} \equiv X_1, \dots, X_n$$

mediante muestreo aleatorio simple. Vamos a ver cuales son las técnicas para contrastar hipótesis sobre los parámetros que rigen X . Vamos a comenzar haciendo diferentes tipos de contrastes para medias y después sobre las varianzas y desviaciones típicas.

9.2.1. Contrastes para la media

Test de dos colas con varianza desconocida

Sea $X \sim \mathbf{N}(\mu, \sigma^2)$ donde ni μ ni σ^2 son conocidos y queremos realizar el contraste

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

Al no conocer σ^2 va a ser necesario estimarlo a partir de su estimador insesgado: la cuasivarianza muestral, \hat{S}^2 . Por ello la distribución del estimador del contraste será una t de Student, que ha perdido un grado de libertad:

$$H_0 \text{ cierta} \iff T_{exp} = \frac{\bar{X} - \mu_0}{\frac{\hat{S}}{\sqrt{n}}} \sim t_{n-1}$$

Consideramos como región crítica \mathcal{C} , a las observaciones de T_{exp} extremas

$$\mathcal{C} = \left\{ T_{exp} < -t_{n-1, 1-\alpha/2} \quad \text{ó} \quad t_{n-1, 1-\alpha/2} < T_{exp} \right\}$$

Observación

Para dar una forma homogénea a todos los contrastes de hipótesis es costumbre denominar al valor del estadístico del contraste calculado sobre la muestra como **valor experimental** y a los extremos de la región crítica, como **valores teóricos**. Definiendo entonces

$$T_{exp} = \frac{\bar{X} - \mu_0}{\frac{\hat{S}}{\sqrt{n}}}$$

$$T_{teo} = t_{n-1, 1-\alpha/2}$$

el resultado del contraste es el siguiente:

$$\begin{cases} \text{si } |T_{exp}| \leq T_{teo} \implies & \text{no rechazamos } H_0; \\ \text{si } |T_{exp}| > T_{teo} \implies & \text{rechazamos } H_0 \text{ y aceptamos } H_1. \end{cases}$$

Tests de una cola con varianza desconocida

Si realizamos el contraste

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} \quad \left(\text{o bien} \quad \begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} \right)$$

por analogía con el contraste bilateral, definiremos

$$T_{exp} = \frac{\bar{X} - \mu_0}{\frac{\hat{S}}{\sqrt{n}}}$$

$$T_{teo} = t_{n-1, 1-\alpha}$$

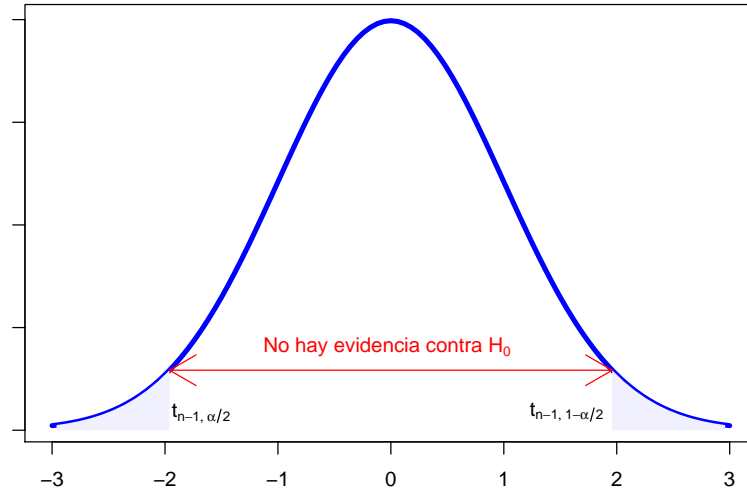


Figura 9.1: Sombreada apreciamos la región crítica sombreada para el contraste bilateral de una media.

y el criterio para contrastar al nivel de significación α es

$$\begin{cases} \text{si } T_{exp} \geq -T_{teo} \implies & \text{no rechazamos } H_0; \\ \text{si } T_{exp} \leq -T_{teo} \implies & \text{rechazamos } H_0 \text{ y aceptamos } H_1. \end{cases}$$

Para el contraste contrario,

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases} \quad \left(\text{o bien} \quad \begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases} \right)$$

definimos T_{exp} y T_{teo} como anteriormente y el criterio a aplicar es (véase la figura 9.3):

$$\begin{cases} \text{si } T_{exp} \leq T_{teo} \implies & \text{no rechazamos } H_0; \\ \text{si } T_{exp} > T_{teo} \implies & \text{rechazamos } H_0 \text{ y aceptamos } H_1. \end{cases}$$

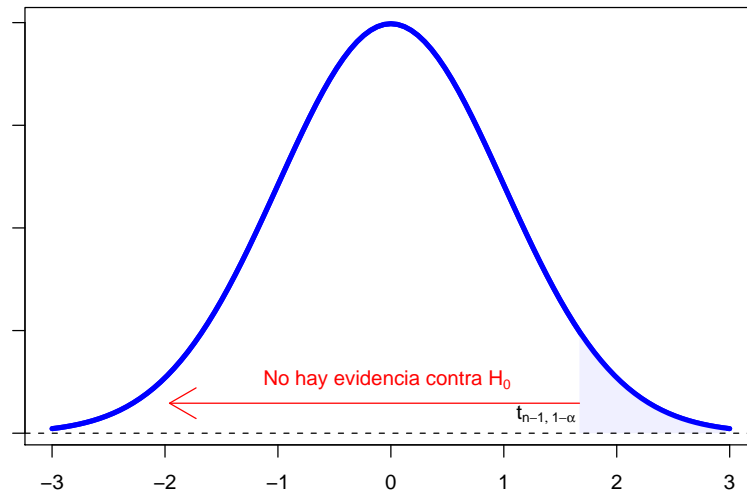


Figura 9.2: Región crítica a la derecha para el contrastes unilaterales de una media.

Ejemplo

Conocemos que las alturas X de los individuos de una ciudad, se distribuyen de modo gaussiano. Deseamos contrastar con un nivel de significación de $\alpha = 0,05$ si la altura media es diferente de 174 cm. Para ello nos basamos en un estudio en el que con una muestra de $n = 25$ personas se obtuvo:

$$\begin{aligned}\bar{x} &= 170 \text{ cm} \\ \mathcal{S} &= 10 \text{ cm}\end{aligned}$$

Solución:

El contraste que se plantea es:

$$\begin{cases} H_0 : \mu = 174 \text{ cm} \\ H_1 : \mu \neq 174 \text{ cm} \end{cases}$$

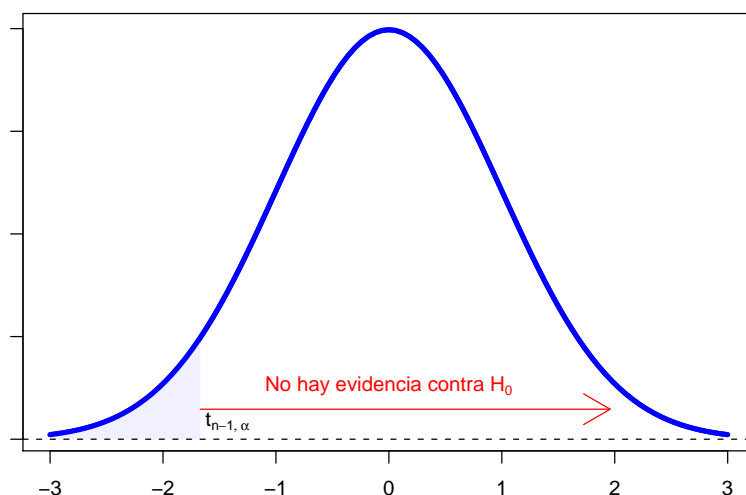


Figura 9.3: Región crítica a la izquierda para el contrastes unilateral de una media.

La técnica a utilizar consiste en suponer que H_0 es cierta y ver si el valor que toma el estadístico

$$T_{exp} = \frac{\bar{x} - 174}{\frac{\hat{S}}{\sqrt{n}}} \sim t_{n-1} = t_{24}$$

es “razonable.” no bajo esta hipótesis, para el nivel de significación dado. Aceptaremos la hipótesis alternativa (y en consecuencia se rechazará la hipótesis nula) si no lo es, es decir, si

$$|T_{exp}| \geq t_{24;1-\alpha/2} = t_{24,0,975} = 2,06$$

Para ello procedemos al cálculo de T_{exp} :

$$S = 10 \implies \hat{S} = S \sqrt{\frac{n}{n-1}} = 10 \sqrt{\frac{25}{24}} = 10'206$$

$$|T_{exp}| = \frac{|170 - 174|}{\frac{10,206}{\sqrt{25}}} = |-1,959| \leq t_{24;0,975} = 2,06$$

Luego, aunque podamos pensar que ciertamente el verdadero valor de μ no es 174, no hay una evidencia suficiente para rechazar esta hipótesis al nivel de confianza del 95 %. Es decir, no se rechaza H_0 .

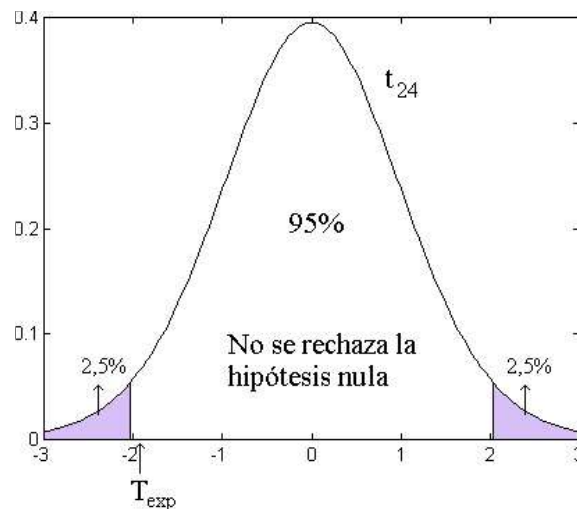


Figura 9.4: El valor de T_{exp} no está en la región crítica (aunque ha quedado muy cerca), por tanto al no ser la evidencia en contra de H_0 suficientemente significativa, ésta hipótesis no se rechaza.

Ejemplo

Consideramos el mismo ejemplo de antes. Visto que no hemos podido rechazar el que la altura media de la población sea igual a 174 cm, deseamos realizar el contraste sobre si la altura media es menor de 174 cm.

Solución:

Ahora el contraste es

$$\begin{cases} H_0 : \mu \geq 174 \text{ cm} \\ H_1 : \mu < 174 \text{ cm} \end{cases}$$

Para realizar este contraste, consideramos el caso límite y observamos si la hipótesis nula debe ser rechazada o no. Este es:

$$\begin{cases} H'_0 : \mu = 174 \text{ cm} \\ H_1 : \mu < 174 \text{ cm} \end{cases}$$

De nuevo la técnica a utilizar consiste en suponer que H'_0 es cierta y ver si el valor que toma el estadístico

$$T_{exp} = \frac{\bar{x} - 174}{\frac{\hat{S}}{\sqrt{n}}} \sim t_{n-1} = t_{24}$$

es aceptable bajo esta hipótesis, con un nivel de confianza del 95 %. Se aceptará la hipótesis alternativa (y en consecuencia se rechazará la hipótesis nula) si

$$T_{exp} \leq t_{24;\alpha} = -t_{24;1-\alpha} = -t_{24;0,95} = -1,71$$

Recordamos que el valor de T_{exp} obtenido fue de

$$T_{exp} = -1,959 < t_{24;0,05} = -t_{24;0,95} = -1,71$$

Por ello hemos de aceptar la hipótesis alternativa

Es importante observar este hecho curioso: Mientras que en el ejemplo anterior no existía una evidencia significativa para decir que $\mu \neq 174$ cm, el “simple hecho” de plantearnos un contraste que parece el mismo pero en versión unilateral nos conduce a rechazar de modo significativo que $\mu = 174$ y aceptamos que $\mu < 174$ cm. Es por ello que podemos decir que no sólo

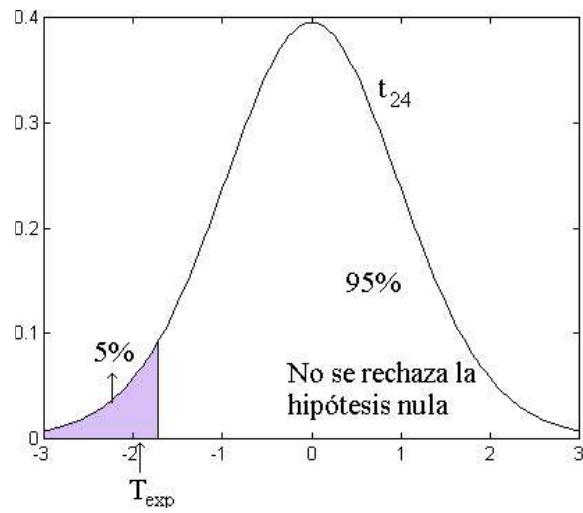


Figura 9.5: El valor t_{exp} está en la región crítica, por tanto existe una evidencia significativa en contra de H_0 , y a favor de H_1 .

H'_0 es rechazada, sino también H_0 . Es en este sentido en el que los tests con H_0 y H'_0 los consideramos equivalentes:

$$\left\{ \begin{array}{l} H'_0 : \mu = 174 \text{ cm} \\ H_1 : \mu < 174 \text{ cm} \end{array} \right. \iff \left\{ \begin{array}{l} H_0 : \mu \geq 174 \text{ cm} \\ H_1 : \mu < 174 \text{ cm} \end{array} \right.$$

9.2.2. Contrastes para la varianza

Consideremos que el carácter que estudiamos sobre la población sea una v.a. normal cuya media y varianza son desconocidas. Vamos a contrastar la hipótesis

$$H_0 : \sigma^2 = \sigma_0^2, \quad \text{donde } \sigma_0^2 \text{ es un valor prefijado}$$

frente a otras hipótesis alternativas que podrán dar lugar a contrastes bilaterales o unilaterales. La técnica consiste en utilizar el teorema de Cochran, para observar que el siguiente estadístico experimental que utiliza el estimador insesgado de la varianza, posee una distribución χ^2 , con $n-1$ grados de libertad:

$$H_0 \text{ cierta} \implies \chi_{exp}^2 = (n-1) \cdot \frac{\hat{S}^2}{\sigma_0^2} \rightsquigarrow \chi_{n-1}^2$$

Entonces construimos las regiones críticas que correspondan a las hipótesis alternativas que se formulen en cada caso atendiendo a la ley de distribución χ^2 .

Contraste bilateral

Cuando el contraste a realizar es

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases}$$

definimos

$$\chi_{exp}^2 = (n-1) \cdot \frac{\hat{S}^2}{\sigma_0^2}$$

$$a_{teo} = \chi_{n-1, \alpha/2}^2$$

$$b_{teo} = \chi_{n-1, 1-\alpha/2}^2$$

y el criterio que suministra el contraste es

$$\left\{ \begin{array}{ll} \text{si } a_{teo} \leq \chi_{exp}^2 \leq b_{teo} & \implies \text{no rechazamos } H_0; \\ \text{si } \chi_{exp}^2 < a_{teo} \text{ ó } \chi_{exp}^2 > b_{teo} & \implies \text{rechazamos } H_0 \text{ y aceptamos } H_1. \end{array} \right.$$

Contrastes unilaterales

Para un contraste de significación al nivel α del tipo

$$\left\{ \begin{array}{l} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{array} \right. \quad \left(\text{o bien } \left\{ \begin{array}{l} H_0 : \sigma^2 \geq \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{array} \right. \right)$$

se tiene que el resultado del mismo es:

$$a_{teo} = \chi_{n-1, \alpha}^2 \longrightarrow \left\{ \begin{array}{ll} \text{si } a_{teo} \leq \chi_{exp}^2 & \implies \text{no rechazamos } H_0; \\ \text{si } \chi_{exp}^2 < a_{teo} & \implies \text{rechazamos } H_0 \text{ y aceptamos } H_1. \end{array} \right.$$

Para el contraste contrario tenemos la formulación análoga

$$\left\{ \begin{array}{l} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{array} \right. \quad \left(\text{o bien } \left\{ \begin{array}{l} H_0 : \sigma^2 \leq \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{array} \right. \right)$$

calculamos el extremo inferior de la región crítica en una tabla de la distribución χ_{n-1}^2

$$b_{teo} = \chi_{n-1, 1-\alpha}^2 \longrightarrow \left\{ \begin{array}{ll} \text{si } \chi_{exp}^2 \leq b_{teo} & \implies \text{no rechazamos } H_0; \\ \text{si } b_{teo} < \chi_{exp}^2 & \implies \text{rechazamos } H_0 \text{ y aceptamos } H_1. \end{array} \right.$$

9.3. Contrastes de una proporción

Supongamos que poseemos una sucesión de observaciones independientes, de modo que cada una de ellas se comporta como una distribución de Bernoulli de parámetro p :

$$\vec{X} \equiv X_1, \dots, X_i, \dots, X_n, \quad \text{donde } X_i \rightsquigarrow \mathbf{Ber}(p)$$

La v.a. X , definida como el número de éxitos obtenidos en una muestra de tamaño n es por definición una v.a. de distribución binomial:

$$X = \sum_{i=1}^n X_i \rightsquigarrow \mathbf{B}(n, p)$$

La proporción muestral (estimador del verdadero parámetro p a partir de la muestra) es

$$\hat{P} = \frac{X}{n}$$

Nos interesamos en el contraste de significación de

$$H_0 : p = p_0, \quad \text{donde } p_0 \text{ es un valor prefijado}$$

frente a otras hipótesis alternativas. Para ello nos basamos en un estadístico (de contraste) que ya fue considerado anteriormente en la construcción de intervalos de confianza para proporciones y que sigue una distribución aproximadamente normal para tamaños muestrales suficientemente grandes:

$$\hat{P} = \frac{X}{n} \rightsquigarrow \mathbf{N}\left(p, \frac{pq}{n}\right)$$

Si la hipótesis H_0 es cierta se tiene

$$\boxed{\hat{P} = \frac{X}{n} \rightsquigarrow \mathbf{N}\left(p_0, \frac{p_0 q_0}{n}\right) \iff \frac{\hat{P} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = Z_{exp} \rightsquigarrow \mathbf{N}(0, 1)}$$

Contraste bilateral

Para el contraste

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$$

extraemos una muestra y observamos el valor $X = x \Rightarrow \hat{p} = \frac{x}{n}$. Entonces se define

$$Z_{exp} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

$$Z_{teo} = z_{1-\alpha/2}$$

siendo el criterio de aceptación o rechazo de la hipótesis nula el que refleja la figura 9.6:

$$\begin{cases} \text{si } |Z_{exp}| \leq Z_{teo} \implies \text{aceptamos } H_0; \\ \text{si } |Z_{exp}| > Z_{teo} \implies \text{rechazamos } H_0 \text{ y aceptamos } H_1. \end{cases}$$

Contrastes unilaterales

Consideremos un contraste del tipo

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p < p_0 \end{cases} \quad \left(\text{o bien } \begin{cases} H_0 : p \geq p_0 \\ H_1 : p < p_0 \end{cases} \right)$$

$$\begin{cases} Z_{exp} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \\ Z_{teo} = z_\alpha \end{cases} \rightarrow \begin{cases} \text{si } Z_{exp} \leq Z_{teo} \implies \text{rechazamos } H_0 \text{ y aceptamos } H_1; \\ \text{si } Z_{exp} > Z_{teo} \implies \text{no rechazamos } H_0. \end{cases}$$

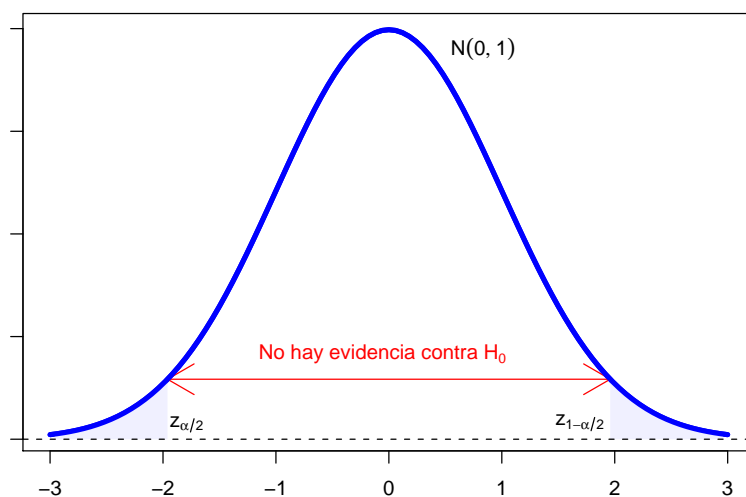


Figura 9.6: Contraste bilateral de una proporción.

Para el test unilateral contrario, se tiene la expresión simétrica

$$\left\{ \begin{array}{l} H_0 : p = p_0 \\ H_1 : p > p_0 \end{array} \right. \quad \left(\text{o bien} \left\{ \begin{array}{l} H_0 : p \leq p_0 \\ H_1 : p > p_0 \end{array} \right. \right)$$

Luego

$$\left\{ \begin{array}{l} Z_{exp} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \\ Z_{teo} = z_{1-\alpha} \end{array} \right. \rightarrow \left\{ \begin{array}{l} \text{si } Z_{exp} \leq Z_{teo} \implies \text{no rechazamos } H_0; \\ \text{si } Z_{exp} > Z_{teo} \implies \text{rechazamos } H_0 \text{ y aceptamos } H_1. \end{array} \right.$$

Ejemplo

Se cree que determinada enfermedad se presenta en mayor medida en hombres que en mujeres. Para ello se elige una muestra aleatoria de 100 de

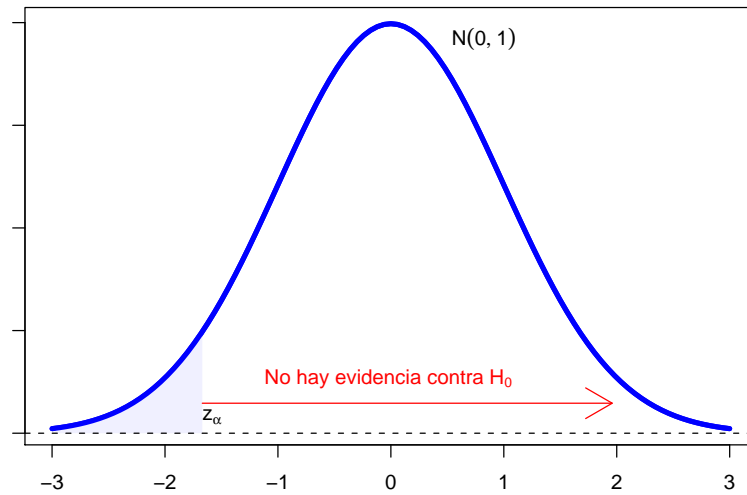


Figura 9.7: Contraste unilateral cuando se tiene $H_0 : p \geq p_0$

estos enfermos y se observa que 70 son hombres. ¿Qué podemos concluir?

Solución: Sea p la proporción de hombres que existen entre los enfermos. Queremos encontrar evidencia a favor (H_1) de que $p > 1/2$, pero nuestra hipótesis de partida (mientras no tengamos evidencia en contra) es que $p = 1/2$ (H_0). Es decir, planteemos el siguiente contraste unilateral para una proporción:

$$\begin{cases} H_0 : p = 1/2 \\ H_1 : p > 1/2 \end{cases}$$

La estimación puntual de p es $\hat{p} = 70/100 = 0,7$. El estadístico que usamos para el contraste es:

$$Z = \frac{\hat{p} - p}{\sqrt{pq/n}} \rightsquigarrow \mathbf{N}(0, 1)$$

Está claro que se obtiene mayor evidencia a favor de H_1 cuando los valores de \hat{p} se acercan a 1, o lo que es lo mismo, cuando Z se hace “suficientemente grande”. Dicho de otro modo, los valores críticos de Z (los que nos conducen

a rechazar H_0 y aceptar H_1 son los de la cola de la derecha de la distribución $N(0, 1)$.

Si elegimos $\alpha = 5\%$, los valores críticos son los que están situados a la derecha del percentil 95 de esta distribución, es decir, los valores superiores a $z_{teo} = z_{1-\alpha} = 1,96$.

Veamos si el valor experimental del estadístico (el calculado a partir de la muestra si suponemos cierta H_0) supera o no dicho valor:

$$Z_{exp} = \frac{\hat{p} - p}{\sqrt{pq/n}} = \frac{0,7 - 0,5}{\sqrt{0,5 \times 0,5/100}} = 4$$

Como se aprecia, Z_{exp} entra ampliamente dentro de la región crítica, por tanto hemos de concluir con el rechazo de la hipótesis nula y la aceptación de la hipótesis alternativa.

Resumamos el ejemplo con otras palabras: Si la hipótesis nula fuese cierta, deberíamos esperar que el valor del estadístico Z no fuese “demasiado grande”. Por tanto como hemos obtenido un valor “grande” del mismo, debemos concluir que la hipótesis de partida (H_0) ha de ser rechazada. El valor z_{teo} se calcula exclusivamente a partir de α , y nos sirve para saber a que nos referimos por un valor “demasiado grande” para Z .

9.4. Contrastes para la diferencia de medias apareadas

Las muestras apareadas aparecen como distintas observaciones realizadas sobre los mismos individuos. Un ejemplo de observaciones apareadas consiste en considerar a un conjunto de n personas a las que se le aplica un tratamiento médico y se mide por ejemplo el nivel de insulina en la sangre antes (X) y después del mismo (Y)

Paciente	x_i	y_i	d_i
1	150	120	30
2	180	130	50
...
n	140	90	50

No es posible considerar a X e Y como variables independientes ya que va a existir una dependencia clara entre las dos variables. Si queremos contrastar el que los pacientes han experimentado o no una mejoría con el tratamiento, llamemos d_i a la diferencia entre las observaciones antes y después del tratamiento

$$d_i = x_i - y_i$$

Supongamos que la v.a. que define la diferencia entre el antes y después del tratamiento es una v.a. d que se distribuye normalmente, pero cuyas media y varianza son desconocidas

$$d \rightsquigarrow \mathbf{N}(\mu_d, \sigma_d^2)$$

Si queremos contrastar la hipótesis de que el tratamiento ha producido cierto efecto Δ

$$H_0 : \mu_d = \Delta,$$

en el caso en que H_0 fuese cierta tendríamos que el estadístico de contraste que nos conviene es

$$T_{exp} = \frac{\bar{d} - \Delta}{\frac{1}{\sqrt{n}} \hat{S}_d} \rightsquigarrow \mathbf{t}_{n-1}$$

donde \bar{d} es la media muestral de las diferencias d_i y \hat{S}_d es la cuasivarianza muestral de las mismas. El tipo de contraste sería entonces del mismo tipo que el realizado para la media con varianza desconocida.

Contraste bilateral

Consideramos el contraste de tipo

$$\begin{cases} H_0 : \mu_d = \Delta \\ H_1 : \mu_d \neq \Delta \end{cases}$$

Entonces se define

$$T_{exp} = \frac{\bar{d} - \Delta}{\frac{1}{\sqrt{n}} \hat{S}_d}$$

y se rechaza la hipótesis nula cuando $T_{exp} < -t_{n-1, 1-\alpha/2}$ ó $T_{exp} > t_{n-1, 1-\alpha/2}$.

Contrastes unilaterales

Si el contraste es

$$\begin{cases} H_0 : \mu_d = \Delta \\ H_1 : \mu_d < \Delta \end{cases} \quad \left(\text{o bien} \begin{cases} H_0 : \mu_d \geq \Delta \\ H_1 : \mu_d < \Delta \end{cases} \right)$$

entonces se rechaza H_0 si $T_{exp} < -t_{n-1, 1-\alpha}$. Para el test contrario

$$\begin{cases} H_0 : \mu_d = \Delta \\ H_1 : \mu_d > \Delta \end{cases} \quad \left(\text{o bien} \begin{cases} H_0 : \mu_d \leq \Delta \\ H_1 : \mu_d > \Delta \end{cases} \right)$$

se rechaza H_0 si $T_{exp} > t_{n-1, 1-\alpha}$.

Ejemplo

Se pretende demostrar que cierto tratamiento practicado durante un mes, ayuda a reducir el colesterol. Para ello se reliza un estudio con una muestra aleatoria simple de 10 personas. Los resultados se muestran a continuación.

Antes	200	210	330	240	260	300	245	210	190	225
Después	150	200	275	250	200	250	200	180	190	205

¿Que podemos concluir de estos datos.

Solución: Obsérvese que las mediciones se realizan sobre las mismas personas, por tanto no tenemos dos muestras aleatorias independientes, sino una sola, en la cual lo que nos interesa es la diferencia producida entre el colesterol antes del tratamiento y después del mismo. Para ello

introducimos una nueva variable que expresa la diferencia existente entre el colesterol antes del tratamiento y después del mismo:

$$d = X_{ant} - X_{des}$$

Antes	200	210	330	240	260	300	245	210	190	225
Después	150	200	275	250	200	250	200	180	190	205
Diferencia	50	10	55	-10	60	50	45	30	0	20

Encontrar evidencia a favor de que el tratamiento surge el efecto deseado (baja el colesterol) es lo mismo que encontrar evidencia estadísticamente significativa en el contraste:

$$\begin{cases} H_0 : \mu_d = 0 \\ H_1 : \mu_d > 0 \end{cases}$$

Esto es de nuevo un contraste para una media, que se realiza sobre la variable *diferencia*. El estadístico que usamos es:

$$T_{exp} = \frac{\bar{d} - \mu_d}{\frac{\hat{S}_d}{\sqrt{n}}} \sim \mathbf{t}_{n-1} = \mathbf{t}_9$$

Si \bar{d} es “*muy grande*” deberemos concluir que la hipótesis H_1 es correcta, lo que equivale a decir que la región crítica del contraste está en la cola de la derecha de la distribución \mathbf{t}_9 . Si elegimos un nivel de significación $\alpha = 0,05$, los valores críticos del contraste son los que superan al percentil 95 de la distribución mencionada, es decir, son los que superan la cantidad $T_{teo} = T_{9;0,95} = 1,8331$.

Para ver si T_{exp} supera el valor teórico hemos de calcular previamente a partir de la muestra las estimaciones insesgadas de la media y la desviación típica:

$$\begin{aligned} \bar{d} &= 31 \\ \hat{S}_d &= 7,43 \end{aligned}$$

Luego si suponemos que la hipótesis nula es cierta y que la variable diferencia sigue una distribución normal de parámetros desconocidos, tenemos:

$$T_{exp} = \frac{31 - 0}{7,43/\sqrt{10}} = 13,19$$

El valor experimental se encuentra claramente en la región crítica del contraste ($T_{exp} > T_{teo}$) por tanto concluimos que existe evidencia estadísticamente significativa en contra de la hipótesis nula y a favor de la hipótesis alternativa (al menos con un nivel de significación del 5 %).

9.5. Contrastes de dos distribuciones normales independientes

Consideramos a lo largo de toda esta sección a dos poblaciones normales que representamos mediante

$$X_1 \sim \mathbf{N}(\mu_1, \sigma_1^2)$$

$$X_2 \sim \mathbf{N}(\mu_2, \sigma_2^2)$$

De las que *de modo independiente* se extraen muestras de tamaño respectivo n_1 y n_2 . Los tests que vamos a realizar están relacionados con la diferencias existentes entre ambas medias o los cocientes de sus varianzas.

9.5.1. Contraste de medias con varianzas conocidas

De manera similar al caso del contraste para una media, queremos en esta ocasión contrastar la hipótesis de que las dos poblaciones (cuyas varianzas suponemos conocidas) sólo difieren en una cantidad Δ

$$H_0 : \mu_1 - \mu_2 = \Delta$$

frente a hipótesis alternativas que darán lugar a contrastes unilaterales o bilaterales como veremos más tarde. Para ello nos basamos en la distribución del siguiente estadístico de contraste:

$$\begin{aligned}
H_0 \text{ cierta} &\implies \begin{cases} \bar{X}_1 \rightsquigarrow \mathbf{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \\ \bar{X}_2 \rightsquigarrow \mathbf{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right) \end{cases} \\
&\implies \bar{X}_1 - \bar{X}_2 \rightsquigarrow \mathbf{N}\left(\Delta, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \\
&\iff \boxed{Z = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightsquigarrow \mathbf{N}(0, 1)}
\end{aligned}$$

Contraste bilateral

Consideremos en primer lugar el contraste de dos colas

$$\begin{cases} H_0 : \mu_1 - \mu_2 = \Delta \\ H_1 : \mu_1 - \mu_2 \neq \Delta \end{cases}$$

Se define entonces

$$\begin{aligned}
Z_{exp} &= \frac{(\bar{X}_1 - \bar{X}_2) - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \\
Z_{teo} &= z_{1-\alpha/2}
\end{aligned}$$

y el test consiste en

$$\begin{cases} \text{si } |Z_{exp}| \leq Z_{teo} \implies & \text{no rechazamos } H_0; \\ \text{si } |Z_{exp}| > Z_{teo} \implies & \text{rechazamos } H_0 \text{ y aceptamos } H_1. \end{cases}$$

Contrastes unilaterales

Para el test

$$\left\{ \begin{array}{l} H_0 : \mu_1 - \mu_2 = \Delta \\ H_1 : \mu_1 - \mu_2 < \Delta \end{array} \right. \quad \left(\text{o bien} \left\{ \begin{array}{l} H_0 : \mu_1 - \mu_2 \geq \Delta \\ H_1 : \mu_1 - \mu_2 < \Delta \end{array} \right. \right)$$

el contraste consiste en

$$Z_{teo} = z_\alpha = -z_{1-\alpha} \rightarrow \left\{ \begin{array}{ll} \text{si } Z_{exp} \geq Z_{teo} & \implies \text{no rechazamos } H_0; \\ \text{si } Z_{exp} < Z_{teo} & \implies \text{rechazamos } H_0 \text{ y aceptamos } H_1. \end{array} \right.$$

y para el contraste de significación contrario

$$\left\{ \begin{array}{l} H_0 : \mu_1 - \mu_2 = \Delta \\ H_1 : \mu_1 - \mu_2 > \Delta \end{array} \right. \quad \left(\text{o bien} \left\{ \begin{array}{l} H_0 : \mu_1 - \mu_2 \leq \Delta \\ H_1 : \mu_1 - \mu_2 > \Delta \end{array} \right. \right)$$

se tiene

$$Z_{teo} = z_{1-\alpha} \rightarrow \left\{ \begin{array}{ll} \text{si } Z_{exp} \leq Z_{teo} & \implies \text{no rechazamos } H_0; \\ \text{si } Z_{exp} > Z_{teo} & \implies \text{rechazamos } H_0 \text{ y aceptamos } H_1. \end{array} \right.$$

9.5.2. Contraste de medias homocedáticas

Ahora consideramos el problema de contrastar

$$H_0 : \mu_1 - \mu_2 = \Delta$$

cuando sólo conocemos que las varianzas de ambas poblaciones son iguales, pero desconocidas. El estadístico que usaremos para el contraste fue ya introducido en la relación (8.2), pues si suponemos que H_0 es cierta se tiene

$$T_{exp} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{S} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathbf{t}_{n_1+n_2-2}$$

donde \hat{S}^2 es la cuasivarianza muestral ponderada de \hat{S}_1^2 y \hat{S}_2^2

$$\hat{S}^2 = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}$$

Obsérvese que se han perdido dos grados de libertad a causa de la estimación de $\sigma_1^2 = \sigma_2^2$ mediante \hat{S}_1^2 y \hat{S}_2^2 .

Contraste bilateral

Para el contraste de significación

$$\begin{cases} H_0 : \mu_1 - \mu_2 = \Delta \\ H_1 : \mu_1 - \mu_2 \neq \Delta \end{cases}$$

se tiene como en casos anteriores que el contraste adecuado consiste en definir

$$T_{exp} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{S} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$T_{teo} = t_{n_1+n_2-2, 1-\alpha/2}$$

y rechazar o admitir la hipótesis nula siguiendo el criterio

$$\begin{cases} \text{si } |T_{exp}| \leq T_{teo} \implies & \text{no rechazamos } H_0; \\ \text{si } |T_{exp}| > T_{teo} \implies & \text{rechazamos } H_0 \text{ y aceptamos } H_1. \end{cases}$$

Contrastes unilaterales

Cuando el contraste es unilateral del modo

$$\begin{cases} H_0 : \mu_1 - \mu_2 = \Delta \\ H_1 : \mu_1 - \mu_2 < \Delta \end{cases} \quad \left(\text{o bien } \begin{cases} H_0 : \mu_1 - \mu_2 \geq \Delta \\ H_1 : \mu_1 - \mu_2 < \Delta \end{cases} \right)$$

el contraste se realiza siguiendo el mismo proceso que en otros realizados anteriormente, lo que nos lleva a

$$T_{teo} = -t_{n_1+n_2-2, 1-\alpha} \rightarrow \begin{cases} \text{si } T_{exp} \geq T_{teo} \implies & \text{no rechazamos } H_0; \\ \text{si } T_{exp} < T_{teo} \implies & \text{rechazamos } H_0 \text{ y aceptamos } H_1. \end{cases}$$

y cuando el contraste de significación es el contrario

$$\begin{cases} H_0 : \mu_1 - \mu_2 = \Delta \\ H_1 : \mu_1 - \mu_2 > \Delta \end{cases} \quad \left(\text{o bien } \begin{cases} H_0 : \mu_1 - \mu_2 \leq \Delta \\ H_1 : \mu_1 - \mu_2 > \Delta \end{cases} \right)$$

del mismo modo

$$T_{teo} = t_{n_1+n_2-2, 1-\alpha} \rightarrow \begin{cases} \text{si } T_{exp} \leq T_{teo} \implies & \text{no rechazamos } H_0; \\ \text{si } T_{exp} > T_{teo} \implies & \text{rechazamos } H_0 \text{ y aceptamos } H_1. \end{cases}$$

9.5.3. Contraste de medias no homocedáticas

Consideramos el contraste

$$H_0 : \mu_1 - \mu_2 = \Delta$$

en el caso más problemático, es decir cuando sólo conocemos de las dos poblaciones que su distribución es normal, y que sus varianzas no son conocidas y *significativamente* diferentes. En este caso el estadístico de contraste tendrá una ley de distribución muy particular. Consistirá en una distribución **t** de Student, con un número de grados de libertad que en lugar de depender de modo determinista de la muestra (a través de su tamaño), depende de un modo aleatorio mediante las varianzas muestrales. Concretamente, el estadístico que nos interesa es

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \rightsquigarrow \mathbf{t}_f$$

donde f es el *número de grados de libertad* que se calcula mediante la **fórmula de Welch**

$$f = \frac{\left(\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}\right)^2}{\frac{1}{n_1 + 1} \left(\frac{\hat{S}_1^2}{n_1}\right)^2 + \frac{1}{n_2 + 1} \left(\frac{\hat{S}_2^2}{n_2}\right)^2} - 2$$

No desarrollamos en detalle los cálculos a realizar, pues la técnica para efectuar los contrastes son análogos a los vistos anteriormente cuando las varianzas son desconocidas e iguales.

Observación

Si lo que pretendemos contrastar es si las medias poblacionales de dos muestras independientes obtenidas de poblaciones normales son idénticas, esto se reduce a los casos anteriores tomando $\Delta = 0$, es decir, realizando el contraste:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

9.5.4. Contrastes de la razón de varianzas

Consideramos dos muestras independientes de dos poblaciones que se distribuyen normalmente (cuyas medias y varianzas son desconocidas). Vamos a abordar cuestiones relacionadas con saber si las varianzas de ambas poblaciones son las mismas, o si la razón (cociente) entre ambas es una cantidad conocida, R . La igualdad entre las dos varianzas puede escribirse $\sigma_1^2 - \sigma_2^2 = 0$ o bien, la existencia de una diferencia entre ambas (Δ), del modo $\sigma_1^2 - \sigma_2^2 = \Delta$. Este modo de escribir la diferencia entre varianzas (que era el adecuado para las medias) no es sin embargo fácil de utilizar para las varianzas, de modo que nos será más fácil sacarle partido a las expresiones de las relaciones entre varianzas como

$$\frac{\sigma_1^2}{\sigma_2^2} = R.$$

Por ejemplo, si $R = 1$ tenemos que ambas varianzas son iguales.

Consideramos entonces la hipótesis nula

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = R$$

la cual vamos a contrastar teniendo en cuenta que:

$$\left. \begin{array}{l} \frac{(n_1 - 1) \hat{S}_1^2}{\sigma_1^2} \rightsquigarrow \chi_{n_1-1}^2 \\ \frac{(n_2 - 1) \hat{S}_2^2}{\sigma_2^2} \rightsquigarrow \chi_{n_2-1}^2 \end{array} \right\} \Rightarrow \frac{\frac{1}{(n_1 - 1)} \frac{(n_1 - 1) \hat{S}_1^2}{\sigma_1^2}}{\frac{1}{(n_2 - 1)} \frac{(n_2 - 1) \hat{S}_2^2}{\sigma_2^2}} = \frac{\sigma_2^2}{\sigma_1^2} \frac{\hat{S}_1^2}{\hat{S}_2^2} \rightsquigarrow \mathbf{F}_{n_1-1, n_2-1}$$

Por tanto el estadístico del contraste que nos conviene tiene una distribución conocida cuando H_0 es cierta —véase la definición de la distribución de Snedecor:

$$F = \frac{1}{R} \frac{\hat{S}_1^2}{\hat{S}_2^2} \sim \mathbf{F}_{n_1-1, n_2-1}$$

Contraste bilateral

El contraste bilateral para el cociente de varianzas se escribe como:

$$\left\{ \begin{array}{l} H_0 : \frac{\sigma_1^2}{\sigma_2^2} = R \\ H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq R \end{array} \right.$$

Habida cuenta que la distribución \mathbf{F} de Snedecor no es simétrica sino que sólo toma valores positivos, se rechazará la hipótesis nula cuando el el valor que tome el estadístico del contraste al aplicarlo sobre una muestra sea muy cercano a cero, o bien, muy grande. Es decir, se define el estadístico experimental y los límites de la región crítica como:

$$F_{exp} = \frac{1}{R} \frac{\hat{S}_1^2}{\hat{S}_2^2}$$

$$a_{teo} = F_{n_1-1, n_2-1, \alpha/2}$$

$$b_{teo} = F_{n_1-1, n_2-1, 1-\alpha/2}$$

y el criterio de aceptación o rechazo es:

$$\left\{ \begin{array}{ll} \text{si } a_{teo} \leq F_{exp} \leq b_{teo} & \implies \text{no rechazamos } H_0; \\ \text{si } F_{exp} < a_{teo} \text{ ó } F_{exp} > b_{teo} & \implies \text{rechazamos } H_0. \end{array} \right.$$

9.5.5. Caso particular: Contraste de homocedasticidad

En la práctica un contraste de gran interés es el de la **homocedasticidad** o igualdad de varianzas. Decimos que dos poblaciones son *homocedáticas* si tienen la misma varianza. El test de homocedasticidad sería entonces el mismo que el de un cociente de varianzas, donde $R = 1$, es decir:

$$\left\{ \begin{array}{l} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{array} \right\} \Longleftrightarrow \left\{ \begin{array}{l} H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \end{array} \right.$$

Observación

Una de las razones de la importancia de este contraste es la siguiente: Si queremos estudiar la diferencia entre las medias de dos poblaciones normales, el caso más realista es considerar un contraste donde las varianzas de las poblaciones son desconocidas. Ante esta situación podemos encontrarnos dos situaciones:

1. Las dos varianzas son iguales. Este es el caso más favorable pues utilizamos la distribución de Student para el contraste con un número de grados de libertad que sólo depende del tamaño de la muestra.
2. Las varianzas son distintas. En este caso el número de grados de libertad es una v.a. (fórmula de Welch) y por tanto al realizar el contraste se pierde cierta precisión.

En esta situación lo recomendable es

- En primer lugar realizar un test de homocedasticidad.
- Si la igualdad de varianzas no puede ser rechazada de modo significativo, aplicamos un test de diferencia de medias suponiendo que las varianzas son desconocidas pero iguales. En otro caso se utiliza la aproximación de Welch.

Observación

Al realizar el contraste bilateral sobre la igualdad de varianzas podemos también economizar parte de trabajo definiendo F_{exp} como el cociente entre la mayor varianza muestral y la menor

$$F_{exp} = \begin{cases} \frac{\hat{S}_1^2}{\hat{S}_2^2} & \text{si } \hat{S}_1^2 \geq \hat{S}_2^2 \\ \frac{\hat{S}_2^2}{\hat{S}_1^2} & \text{si } \hat{S}_2^2 > \hat{S}_1^2 \end{cases} \implies F_{exp} \geq 1$$

ya que así no es necesario calcular el extremo inferior para la región donde no se rechaza H_0 , pues F_{exp} nunca estará próxima a 0. Con esta definición de F_{exp} el criterio a seguir frente al contraste de significación para un valor α dado es:

$$\begin{aligned} F_{teo} &= \begin{cases} F_{n_1-1, n_2-1, 1-\alpha} & \text{si } \hat{S}_1^2 \geq \hat{S}_2^2 \\ F_{n_2-1, n_1-1, 1-\alpha} & \text{si } \hat{S}_2^2 > \hat{S}_1^2 \end{cases} \\ \implies &\begin{cases} \text{si } F_{exp} \leq b_{teo} \implies & \text{no rechazamos } H_0; \\ \text{si } F_{exp} > b_{teo} \implies & \text{rechazamos } H_0. \end{cases} \end{aligned}$$

Ejemplo

Se desea comparar la actividad motora espontánea de un grupo de 25 ratas control y otro de 36 ratas desnutridas. Se midió el número de veces que pasaban delante de una célula fotoeléctrica durante 24 horas. Los datos obtenidos fueron los siguientes:

Ratas de control	$n_1 = 25$	$\bar{x}_1 = 869,8$	$S_1 = 106,7$
Ratas desnutridas	$n_2 = 36$	$\bar{x}_2 = 465$	$S_2 = 153,7$

¿Se observan diferencias significativas entre el grupo control y el grupo desnutrido?

Solución:

En primer lugar, por tratarse de un problema de inferencia estadística, nos serán más útiles las cuasivarianzas que las varianzas. Por ello calculamos:

$$\begin{aligned}\hat{S}_1^2 &= \frac{n_1}{n_1 - 1} S_1^2 = \frac{25}{24} 106,7^2 = 11,859,26 \\ \hat{S}_2^2 &= \frac{n_2}{n_2 - 1} S_2^2 = \frac{36}{35} 153,7^2 = 24,298,653\end{aligned}$$

El contraste que debemos realizar está basado en el de la t de Student para la diferencia de medias de dos poblaciones. Para ello conocemos dos estadísticos posibles, según que las varianzas poblacionales de ambos grupos de ratas puedan ser supuestas iguales (homocedasticidad) o distintas (heterocedasticidad). Para ello realizamos previamente el contraste:

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases} \iff \begin{cases} H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \end{cases}$$

Suponiendo H_0 cierta, tenemos que el estadístico del contraste conveniente es

$$F_{exp} = \begin{cases} \frac{\hat{S}_1^2}{\hat{S}_2^2} & \text{si } \hat{S}_1^2 \geq \hat{S}_2^2 \\ \frac{\hat{S}_2^2}{\hat{S}_1^2} & \text{si } \hat{S}_2^2 > \hat{S}_1^2 \end{cases} \implies F_{exp} \geq 1$$

ya que así no es necesario calcular el extremo inferior para la región donde no se rechaza H_0 . En este caso:

$$\begin{aligned}F_{exp} &= \frac{\hat{S}_2^2}{\hat{S}_1^2} = 2'0489 \rightsquigarrow \mathbf{F}_{n_2-1, n_1-1} \\ F_{teo} &= F_{35, 24, 0'95} \approx 2'97\end{aligned}$$

Como $F_{exp} \leq F_{teo}$, no podemos concluir (al menos al nivel de significación $\alpha = 0'05$) que H_0 deba ser rechazada (figura 9.8).

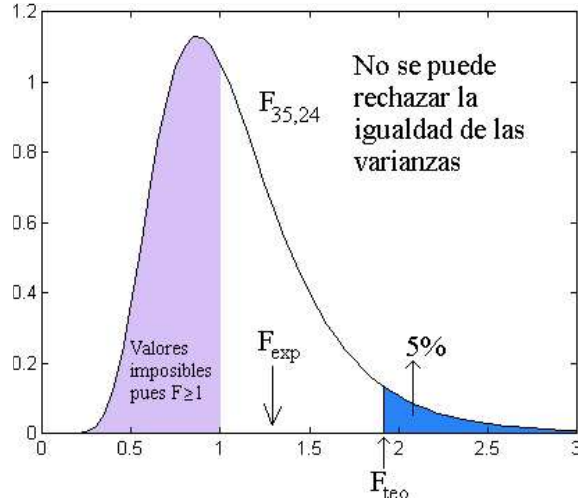


Figura 9.8: No hay evidencia significativa para rechazar la homocedasticidad. El estadístico del contraste ha sido elegido modo que el numerador de F_{exp} sea mayor que el denominador, es decir, $F_{exp} > 1$.

Por lo tanto no rechazamos la hipótesis de homocedasticidad de ambas poblaciones, y pasamos a contrastar la igualdad de las medias

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

utilizando el estadístico más sencillo (el que no necesita aproximar los grados de libertad mediante la fórmula de Welch). Para ello calculamos en primer lugar la cuasivarianza muestral ponderada:

$$\hat{S}^2 = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2} = 19,238'6$$

y posteriormente

$$T_{exp} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{S} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 11'2101 \rightsquigarrow t_{n_1+n_2-2} = t_{59}$$

$$T_{teo} = t_{n_1+n_2-2, 1-\alpha/2} = t_{59, 0'975} \approx 2$$

Como $|T_{teo}| \leq T_{exp}$ concluimos que se ha de rechazar la hipótesis de igualdad de las medias, y por tanto aceptamos que las medias son diferentes.

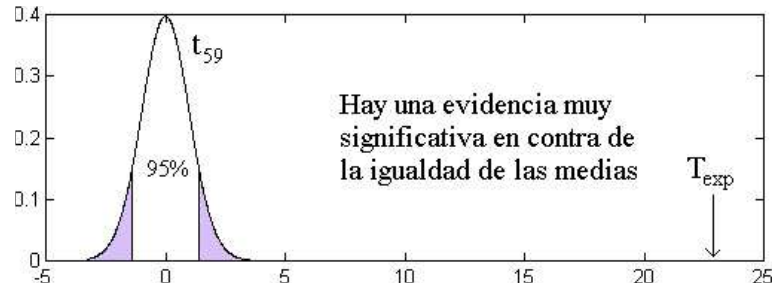


Figura 9.9: Hay una gran evidencia en contra de la hipótesis de que ambas medias poblacionales coincidan, y a favor de que la de la primera población es mayor que la de la segunda.

Ejemplo

Supongamos que cierta variable numérica se comporta de modo gaussiano sobre dos poblaciones, de las que se han extraído respectivamente una muestra aleatoria simple. Los resultados se muestran a continuación:

Muestra 1	10	30	32	23	23	24	20	18	19	45		
Muestra 2	32	39	35	30	37	28	34	33	25	30	37	33

¿Cree que las distribuciones normales que describen a ambas poblaciones, poseen los mismos parámetros?

Solución: La distribución normal está descrita por dos parámetros: La media y la varianza. Vamos a realizar entonces el contraste adecuado para

cada uno de estos parámetros. Como el contraste de igualdad de medias depende de que las varianzas sean iguales o distintas, vamos a comenzar por el contraste de homocedasticidad (igualdad de varianzas).

Previamente, resumimos la información existente en las muestras con los estimadores insesgados de los parámetros:

Primera muestra	Segunda muestra
$X_1 \rightsquigarrow \mathbf{N}(\mu_1, \sigma_1^2)$	$X_2 \rightsquigarrow \mathbf{N}(\mu_2, \sigma_2^2)$
$n_1 = 10$	$n_2 = 12$
$\bar{x}_1 = 22,182$	$\bar{x}_2 = 32,75$
$\hat{S}_1 = 9,513$	$\hat{S}_2 = 4,048$

El contraste de homocedasticidad es el siguiente:

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

El estadístico del contraste lo elegimos de tal modo que la varianza mayor esté en el numerador, pues de este modo tenemos que la región crítica no es nada más que la cola de la derecha de la distribución de Snedecor:

$$F_{exp} = \frac{\hat{S}_{mayor}^2}{\hat{S}_{menor}^2} = \frac{\hat{S}_1^2}{\hat{S}_2^2} = \frac{9,513^2}{4,048^2} = 5,5222$$

Si elegimos un nivel de significación $\alpha = 5\%$, el valor crítico para dicho estadístico (aquel a partir del cual rechazamos la homocedasticidad) es

$$F_{teo} = F_{10;12;0,95} = 2,8962$$

Por tanto se rechaza la hipótesis de igualdad de varianzas.

El contraste de igualdad de medias es:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Desconocemos el valor de las varianzas poblacionales, pero al menos sabemos que hemos rechazado la igualdad de las mismas, por tanto el estadístico del contraste es:

$$T = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} = \frac{24,4 - 32,75}{\sqrt{9,513^2/10 + 4,048^2/12}} = -2,5874$$

La región crítica en este caso está dividida en dos zonas (contraste bilateral). Por tanto hemos de observar si el estadístico del contraste es un valor inferior al percentil 2,5 o superior al 97,5 de la distribución teórica (la que seguiría el estadístico del contraste si la hipótesis nula fuese cierta). Como T_{exp} es un valor negativo, basta con que nos preocupemos nada más que de la cola de la izquierda:

$$T_{teo} = T_{f;0,025} = -T_{f;0,975} = -T_{12,29;0,975} = -2,173$$

donde f es el *número de grados de libertad* que se calcula mediante la **fórmula de Welch**

$$f = \frac{\left(\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}\right)^2}{\frac{1}{n_1 + 1} \left(\frac{\hat{S}_1^2}{n_1}\right)^2 + \frac{1}{n_2 + 1} \left(\frac{\hat{S}_2^2}{n_2}\right)^2} - 2 = 12,29$$

Como T_{exp} es un valor de la región crítica del contraste de igualdad de medias de poblaciones normales con varianzas diferentes, hemos de rechazar (al menos para una significación del 5 %) que las medias de ambas poblaciones coincidan.

Ejemplo

Supongamos que cierta variable numérica se comporta de modo gaussiano sobre dos poblaciones.

Muestra 1	10	30	32	23	23	24	20	18	19	35		
Muestra 2	12	28	30	30	20	25	31	15	12	22	24	40

¿Se puede decir que la media de la primera población es menor que la de la segunda? Usar un nivel de significación del 10 % **Solución:** Hemos de realizar un contraste de medias, pero para decidir el estadístico del contraste a elegir, debemos contrastar la similitud entre las dispersiones de ambas poblaciones.

Para empezar resumimos la información existente en las muestras:

Primera muestra	Segunda muestra
$X_1 \rightsquigarrow \mathbf{N}(\mu_1, \sigma_1^2)$	$X_2 \rightsquigarrow \mathbf{N}(\mu_2, \sigma_2^2)$
$n_1 = 10$	$n_2 = 12$
$\bar{x}_1 = 22,4$	$\bar{x}_2 = 23,08$
$\hat{S}_1 = 9,721$	$\hat{S}_2 = 10,466$

El contraste de homocedasticidad se escribe:

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

El estadístico del contraste lo elegimos de tal modo que la varianza mayor esté en el numerador, pues de este modo tenemos que la región crítica no es nada más que la cola de la derecha de la distribución de Snedecor:

$$F_{exp} = \frac{\hat{S}_{mayor}^2}{\hat{S}_{menor}^2} = \frac{\hat{S}_2^2}{\hat{S}_1^2} = \frac{10,466^2}{9,721^2} = 1,1593$$

Si elegimos un nivel de significación $\alpha = 10\%$, el valor crítico para dicho estadístico (aquel a partir del cual rechazamos la homocedasticidad) es

$$F_{teo} = F_{12;10;0,90} = 2,3961$$

Por tanto no encontramos diferencia que sea estadísticamente significativa entre ambas varianzas, es decir, no rechazamos la hipótesis de homocedasticidad.

El contraste de medias es:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases}$$

Desconocemos el valor de las varianzas poblacionales, pero las diferencias entre ellas (sean cuales sean) no son estadísticamente significativas. Por tanto vamos a elegir como estadístico del contraste al que se usa cuando podemos asumir que las varianzas son iguales:

$$T_{exp} = \frac{(\bar{x}_1 - \bar{x}_2)}{\hat{S} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = -0,1574$$

Esta claro que rechazaremos la hipótesis nula y aceptaremos la alternativa, cuando los datos muestrales de la primera muestra sean significativamente menores que los de la segunda, en cuyo caso el estadístico del contraste tomaría valores muy pequeños. Dicho de otro modo, la región crítica es la región comprendida a la izquierda del percentil 10 de la distribución $t_{n_1+n_2-2}$.

$$T_{teo} = T_{10+12-2;0,10} = -T_{20;0,90} = -1,3253$$

Como T_{exp} no es un valor de la región crítica del contraste, concluimos que no hay evidencia estadísticamente significativa en contra de la hipótesis nula y a favor de la alternativa.

9.6. Contrastes sobre la diferencia de proporciones

Supongamos que tenemos dos muestras independientes tomadas sobre dos poblaciones, en la que estudiamos una variable de tipo dicotómico (Bernoulli):

$$\begin{aligned} \vec{X}_1 &\equiv X_{11}, X_{12}, \dots, X_{1n_1} \\ \vec{X}_2 &\equiv X_{21}, X_{22}, \dots, X_{2n_2} \end{aligned}$$

Si X_1 y X_2 contabilizan en cada caso el número de éxitos en cada muestra se tiene que cada una de ellas se distribuye como una variable aleatoria binomial:

$$\begin{aligned} X_1 &= \sum_{i=1}^{n_1} X_{1i} \rightsquigarrow \mathbf{B}(n_1, p_1) \\ X_2 &= \sum_{i=1}^{n_2} X_{2i} \rightsquigarrow \mathbf{B}(n_2, p_2) \end{aligned}$$

de modo que los estimadores de las proporciones en cada población tienen distribuciones que de un modo aproximado son normales (cuando n_1 y n_2 son bastante grandes)

$$\begin{aligned} \hat{P}_1 &= \frac{X_1}{n_1} \rightsquigarrow \mathbf{N}\left(p_1, \frac{p_1 q_1}{n_1}\right) \\ \hat{P}_2 &= \frac{X_2}{n_2} \rightsquigarrow \mathbf{N}\left(p_2, \frac{p_2 q_2}{n_2}\right) \end{aligned}$$

El contraste que nos interesa realizar es el de si la diferencia entre las proporciones en cada población es una cantidad conocida Δ

$$H_0 : p_1 - p_2 = \Delta$$

Si H_0 fuese cierta se tendría que

$$\hat{P}_1 - \hat{P}_2 \rightsquigarrow \mathbf{N}\left(\underbrace{p_1 - p_2}_{\Delta}, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)$$

Desafortunadamente ni p_1 ni p_2 son conocidos de antemano y utilizamos sus estimadores, lo que da lugar a un error que es pequeño cuando los tamaños muestrales son importantes:

$$\boxed{\frac{(\hat{p}_1 - \hat{p}_2) - \Delta}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}} = Z_{exp} \rightsquigarrow \mathbf{N}(0, 1)}$$

Contraste bilateral

El contraste bilateral sobre la diferencia de proporciones es

$$\begin{cases} H_0 : p_1 - p_2 = \Delta \\ H_1 : p_1 - p_2 \neq \Delta \end{cases}$$

Entonces se define

$$Z_{exp} = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}$$

y se rechaza la hipótesis nula si $Z_{exp} < -z_{1-\alpha/2}$ o si $Z_{exp} > z_{1-\alpha/2}$

Contrastes unilaterales

En el contraste

$$\begin{cases} H_0 : p_1 - p_2 = \Delta \\ H_1 : p_1 - p_2 < \Delta \end{cases} \quad \left(\text{o bien} \begin{cases} H_0 : p_1 - p_2 \geq \Delta \\ H_1 : p_1 - p_2 < \Delta \end{cases} \right)$$

se rechazará H_0 si $Z_{exp} < -z_{1-\alpha}$. Para el test contrario

$$\begin{cases} H_0 : p_1 - p_2 = \Delta \\ H_1 : p_1 - p_2 > \Delta \end{cases} \quad \left(\text{o bien} \begin{cases} H_0 : p_1 - p_2 \leq \Delta \\ H_1 : p_1 - p_2 > \Delta \end{cases} \right)$$

se rechaza H_0 si $Z_{exp} > z_{1-\alpha}$.

9.7. Problemas

En todos los problemas que siguen a continuación, se supone que las muestras han sido elegidas de modo independiente, y que las cantidades

cuantitativas que se miden, se distribuyen de modo gaussiano. En temas posteriores se verá cómo contrastar si estas premisas pueden ser aceptadas o no al examinar las muestras.

Ejercicio 9.1. El calcio se presenta normalmente en la sangre de los mamíferos en concentraciones de alrededor de 6 mg por cada 100 ml del total de sangre. La desviación típica normal de ésta variable es 1 mg de calcio por cada 100 ml del volumen total de sangre. Una variabilidad mayor a ésta puede ocasionar graves trastornos en la coagulación de la sangre. Una serie de nueve pruebas sobre un paciente revelaron una media muestral de 6,2 mg de calcio por 100 ml del volumen total de sangre, y una desviación típica muestral de 2 mg de calcio por cada 100 ml de sangre. ¿Hay alguna evidencia, para un nivel $\alpha = 0,05$, de que el nivel medio de calcio para este paciente sea más alto del normal?

Ejercicio 9.2. El número de accidentes mortales en una ciudad es, en promedio, de 12 mensuales. Tras una campaña de señalización y adecentamiento de las vías urbanas se contabilizaron en 6 meses sucesivos

8, 11, 9, 7, 10, 9

accidentes mortales. ¿Fue efectiva la campaña?

Ejercicio 9.3. El promedio de las puntuaciones de un número elevado de alumnos de Bioestadística es de 6,50. Un determinado año se examinaron 50 alumnos con resultados promedio de 7,25 y desviación típica de 1. ¿Variaron las calificaciones?

Ejercicio 9.4. El peso medio de mujeres de 30 a 40 años es de 53 kg. Un estudio realizado en 16 mujeres de tales edades que siguen una dieta vegetariana da $\bar{x} = 50$ y $S = 5$. ¿Modifica la dieta el peso medio?

Ejercicio 9.5. Una población infantil se dice que es susceptible de recibir una campaña de educación e higiene si su porcentaje de niños con dientes

cariados es superior al 15 %. Una población con 12.637 niños, ¿debe hacerse la campaña si de 387 de ellos 70 tenían algún diente cariado?

Ejercicio 9.6. Un 8 % de los individuos que acuden a un servicio sanitario son hiperutilizadores del mismo (más de 11 visitas al año) y, de entre ellos, un 70 % son mujeres. De entre los no hiperutilizadores, son mujeres el 51 %. ¿Puede afirmarse que han variado los hábitos de estas si, tras una campaña de información y control de visitas, de 90 mujeres elegidas al azar 6 resultaron hiperutilizadoras?

Ejercicio 9.7. Se conoce que un 20 % de los individuos tratados crónicamente con digoxina sufren una reacción adversa por causa de ella. A 10 pacientes se les administró durante largo tiempo digoxina mas otros medicamentos, y de ellos 5 desarrollaron la reacción adversa. ¿Puede afirmarse que la asociación entre la digoxina y los otros medicamentos hace variar el número de reacciones adversas?

Ejercicio 9.8. Para comprobar si un tratamiento con ácidos grasos es eficaz en pacientes con eczema atípico, se tomaron 10 pacientes con eczema de más de 9 meses y se les sometió durante 3 semanas a un tratamiento ficticio (placebo) y durante las tres siguientes a un tratamiento con ácidos grasos. Tras cada periodo, un médico ajeno al proyecto evaluó la importancia del eczema en una escala de 0 (no eczema) a 10 (tamaño máximo de eczema). Los datos fueron los siguientes:

Placebo	6	8	4	8	5	6	5	6	4	5
Tratamiento	5	6	4	5	3	6	6	2	2	6

¿Es eficaz el tratamiento?

Ejercicio 9.9. En un programa de Control de Enfermedades Crónicas, la hipertensión está incluida como la primera patología a controlar. 15 pacientes hipertensos son sometidos al programa y controlados en su tensión

asistólica antes y después de 6 meses de tratamiento. Los datos son los siguientes:

Inic.	180	200	160	170	180	190	190	180	190	160	170	190	200	210	220
Fin.	140	170	160	140	130	150	140	150	190	170	120	160	170	160	150

¿Es efectivo el tratamiento?

10.- Muchos autores afirman que los pacientes con depresión tienen una función cortical por debajo de lo normal debido a un riego sanguíneo cerebral por debajo de lo normal. A dos muestras de individuos, unos con depresión y otros normales, se les midió un índice que indica el flujo sanguíneo en la materia gris (dado en $\text{mg}/(100\text{g}/\text{min})$) obteniéndose:

Depresivos	$n_1 = 19$	$\bar{x}_1 = 47$	$\hat{S}_1 = 7'8$
Normales	$n_2 = 22$	$\bar{x}_2 = 53'8$	$\hat{S}_2 = 6'1$

¿Hay evidencia significativa a favor de la afirmación de los autores?

Ejercicio 9.10. Por fistulización se obtuvo el pH de 6 muestras de bilis hepática con los siguientes resultados:

7,83; 8,52; 7,32; 7,79; 7,57; 6,58

Se desea saber al nivel de significación del 0,05 si la bilis hepática puede considerarse neutra. Si se conociera $\sigma = 0,5$, ¿qué decisión tomaríamos?

Ejercicio 9.11. La prueba de la d-xilosa permite la diferenciación entre una esteatorrea originada por una mala absorción intestinal y la debida a una insuficiencia pancreática, de modo que cifras inferiores a 4 grs. de d-xilosa, indican una mala absorción intestinal. Se realiza dicha prueba a 10 individuos, obteniéndose una media de 3,5 grs. y una desviación típica de 0'5 grs. ¿Sepuede decir que esos pacientes padecen una mala absorción intestinal?

Ejercicio 9.12. La eliminación por orina de aldosterona está valorada en individuos normales en 12 mgs/24 h. por término medio. En 50 individuos con insuficiencia cardíaca se observó una eliminación media de aldosterona de 13 mgs/24 h., con una desviación típica de 2,5 mgs/24 h.

1. ¿Son compatibles estos resultados con los de los individuos normales?
2. ¿La insuficiencia cardíaca aumenta la eliminación por orina de aldosterona?

Ejercicio 9.13. La tabla siguiente muestra los efectos de un placebo y de la hidroclorotiacida sobre la presión sanguínea sistólica de 11 pacientes.

Placebo	211	210	210	203	196	190	191	177	173	170	163
H-cloro	181	172	196	191	167	161	178	160	149	119	156

Según estos datos experimentales, ¿podemos afirmar que existe diferencia en la presión sistólica media durante la utilización de estos dos fármacos?

Ejercicio 9.14. Se sabe que el 70 % de los pacientes internados en un hospital traumatológico requieren algún tipo de intervención quirúrgica. Para determinar si un nuevo método de fisioterapia reduce el porcentaje de intervenciones, se aplica éste a 30 pacientes de los cuales 17 requieren alguna intervención quirúrgica. Comprobar que no hay razones suficientes para afirmar la eficacia del método con un nivel de confianza del 95 %.

Ejercicio 9.15. De un estudio sobre la incidencia de la hipertensión en la provincia de Málaga, se sabe que en la zona rural el porcentaje de hipertensos es del 27,7 %. Tras una encuesta a 400 personas de una zona urbana, se obtuvo un 24 % de hipertensos.

1. ¿Se puede decir que el porcentaje de hipertensos en la zona urbana es distinto que en la zona rural?

2. ¿Es menor el porcentaje de hipertensos en la zona urbana que en la zona rural?

Ejercicio 9.16. Con cierto método de enseñanza para niños subnormales se obtiene una desviación típica de 8, en las puntuaciones de los tests finales. Se pone a prueba un nuevo método y se ensaya en 51 niños. Las calificaciones obtenidas en los tests finales dan una desviación típica de 10. ¿Puede asegurarse que el nuevo método produce distinta variación en las puntuaciones?

Ejercicio 9.17. Se desea comparar la actividad motora espontánea de un grupo de 25 ratas control y otro de 36 ratas desnutridas. Se midió el número de veces que pasaban delante de una célula fotoeléctrica durante 24 horas. Los datos obtenidos fueron los siguientes:

Ratas de control	$n_1 = 25$	$\bar{x}_1 = 869,8$	$S_1 = 106,7$
Ratas desnutridas	$n_2 = 36$	$\bar{x}_2 = 465$	$S_2 = 153,7$

¿Se observan diferencias significativas entre el grupo control y el grupo desnutrido?

Ejercicio 9.18. Se pretende comprobar la hipótesis expuesta en algunos trabajos de investigación acerca de que la presencia del antígeno AG-4 está relacionada con un desenlace. Con éste fin, se hizo una revisión sobre las historias clínicas de 21 mujeres muertas por carcinoma de cuello uterino, observando que 6 de ellas presentaban el citado antígeno. Por otro lado y con fines de comparación se tomó otra muestra de 42 personas, con edades similares a las del grupo anterior y que reaccionaron bien al tratamiento del carcinoma de cuello uterino, en 28 de las cuales se observó la presencia del citado antígeno. ¿Está relacionada la presencia del antígeno con una efectividad del tratamiento?

Ejercicio 9.19. Se quiso probar si la cirrosis de hígado hacia variar el

índice de actividad de la colinesterasa en suero. Se eligieron dos muestras aleatorias e independientes de individuos. Los resultados fueron:

Individuos normales	$n_1 = 20$	$\bar{x}_1 = 1,8$	$S_1 = 0,4$
Individuos cirróticos	$n_2 = 25$	$\bar{x}_2 = 0,66$	$S_2 = 0,2$

La cirrosis de hígado, ¿hace variar el índice de la colinesterasa en suero?

Ejercicio 9.20. Un investigador ha realizado el siguiente experimento: Tomó una *primera muestra* de 25 pacientes que padecían cierto síntoma y otra *segunda muestra* de 30 pacientes con el mismo síntoma. A los de la primera muestra les aplicó un tratamiento específico y a los de la segunda les dio un placebo. Anotó el tiempo en horas en que cada uno dijo que el síntoma había desaparecido y obtuvo los siguientes resultados:

Muestra 1 ^a	$n_1 = 25$	$\sum_i x_{i1} = 85$	$\sum_i x_{i1}^2 = 343$
Muestra 2 ^a	$n_2 = 30$	$\sum_i x_{i2} = 216$	$\sum_i x_{i2}^2 = 1,650$

¿Puede concluir el investigador que el tratamiento es realmente efectivo?

Ejercicio 9.21. Para comprobar si la tolerancia a la glucosa en sujetos sanos tiende a decrecer con la edad se realizó un test oral de glucosa a dos muestras de pacientes sanos, unos jóvenes y otros adultos. El test consistió en medir el nivel de glucosa en sangre en el momento de la ingestión (nivel basal) de 100 grs. de glucosa y a los 60 minutos de la toma. Los resultados fueron los siguientes:

Jóvenes:	Basal	81	89	80	75	74	97	76	89	83	77
	60 minutos	136	150	149	141	138	154	141	155	145	147
Adultos:	Basal	98	94	93	88	79	90	86	89	81	90
	60 minutos	196	190	191	189	159	185	182	190	170	197

1. ¿Se detecta una variación significativa del nivel de glucosa en sangre en cada grupo?
2. ¿Es mayor la concentración de glucosa en sangre a los 60 minutos, en adultos que en jóvenes?
3. El contenido basal de glucosa en sangre, ¿es menor en jóvenes que en adultos?
4. ¿Se detecta a los 60 minutos una variación del nivel de glucosa en sangre diferente de los adultos, en los jóvenes?

Capítulo 10

Contrastes basados en el estadístico Ji–Cuadrado

10.1. Introducción

Existen multitud de situaciones en el ámbito de la salud en el que las variables de interés, las cuales no pueden cuantificarse mediante cantidades numéricas, entre las que el investigador esté interesado en determinar posibles relaciones. Ejemplos de este tipo de variables pueden ser las complicaciones tras una intervención quirúrgica, el sexo, el nivel socio-cultural, etc. En este caso tendríamos, a lo sumo, las observaciones agrupadas en forma de frecuencia, dependiendo de las modalidades que presente cada paciente en cada una de las variables, por los que los métodos estudiados en los capítulos anteriores no serían aplicables.

El objetivo de este tema es el estudio de este tipo de cuestiones en relación con las variables cualitativas (y también v.a. discretas o continuas agrupadas en intervalo). Estos son los contrastes asociados con el estadístico χ^2 . En general este tipo de tests consisten en tomar una muestra y observar si hay diferencia significativa entre las frecuencias observadas y las especificadas por la ley teórica del modelo que se contrasta, también denominadas “frecuencias esperadas”.

Sin embargo, aunque éste sea el aspecto más conocido, el uso del test

χ^2 no se limita al estudio de variables cualitativas. Podríamos decir que existen tres aplicaciones básicas en el uso de este test, y cuyo desarrollo veremos en el transcurso de este capítulo:

Tres son los temas que abordaremos de esta manera:

- Test de ajuste de distribuciones: Es un contraste de significación para saber si los datos de una muestra son conformes a una ley de distribución teórica que sospechamos que es la correcta.
- Test de varias muestras cualitativas: Sirve para contrastar la igualdad de procedencia de un conjunto de muestras de tipo cualitativo.
- Test para tablas de contingencia: Es un contraste para determinar la dependencia o independencia de caracteres cualitativos.

10.2. El estadístico χ^2 y su distribución

Sea X una v.a. cuyo rango son los valores $i = 1, 2, \dots, k$, de modo que p_i es la probabilidad de cada valor;

$$X \rightsquigarrow \begin{cases} 1 \rightarrow \mathcal{P}[X = 1] = p_1 \\ 2 \rightarrow \mathcal{P}[X = 2] = p_2 \\ \dots \\ i \rightarrow \mathcal{P}[X = i] = p_i \\ \dots \\ k \rightarrow \mathcal{P}[X = k] = p_k \end{cases}$$

Supongamos que el resultado de un experimento aleatorio es una clase c_1, c_2, \dots, c_k ($c_i, i = 1, \dots, k$), que puede representar valores cualitativos, discretos o bien intervalos para variables continuas. Sea p_i la probabilidad de que el resultado del experimento sea la clase c_i . Vamos a considerar contrastes cuyo objetivo es comprobar si ciertos valores p_i^0 , propuestos para las cantidades p_i son correctas o no, en función de los resultados experimentales

$$\left\{ \begin{array}{l} H_0 : \text{Los } p_i^0 \text{ son correctos} \\ H_1 : \text{Alguno de los } p_i^0 \text{ es falso} \end{array} \right. \iff \left\{ \begin{array}{l} H_0 : \left\{ \begin{array}{ll} p_1 = p_1^0 & \text{y} \\ p_2 = p_2^0 & \text{y} \\ \dots & \\ p_k = p_k^0 & \end{array} \right. \\ H_1 : \left\{ \begin{array}{ll} p_1 \neq p_1^0 & \text{o bien} \\ p_2 \neq p_2^0 & \text{o bien} \\ \dots & \\ p_k \neq p_k^0 & \end{array} \right. \end{array} \right. \quad (10.1)$$

Mediante muestreo aleatorio simple, se toma una muestra de tamaño n y se obtienen a partir de ella unas *frecuencias observadas* de cada clase que representamos mediante $\mathcal{O}_1, \mathcal{O}_1, \dots, \mathcal{O}_k$

Clase	Frec. Abs.
c_i	\mathcal{O}_i
c_1	\mathcal{O}_1
c_2	\mathcal{O}_2
\dots	\dots
c_k	\mathcal{O}_k
$\sum_{i=1}^k \mathcal{O}_i = n$	

Supongamos que la hipótesis nula es cierta. Al ser $p_i = p_i^0$ la proporción de elementos de la clase c_i en la población, el número de individuos de que presentan esta modalidad al tomar una muestra de tamaño n , es una v.a. de distribución binomial, $\mathbf{B}(n, p_i^0)$. Por tanto la *frecuencia esperada* de individuos de esa clase es

$$\mathcal{E}_i = n \cdot p_i^0 \quad \forall i = 1, 2, \dots, k$$

$$\sum_{i=1}^k \mathcal{E}_i = n \cdot \sum_{i=1}^k p_i^0 = n$$

Obsérvese que a diferencia de las cantidades \mathcal{O}_i , que son las frecuencias que realmente se obtienen en una muestra, las frecuencias esperadas no tienen por que ser números enteros. De cualquier modo, bajo la suposición de que H_0 es cierta cabe esperar que las diferencias entre las cantidades \mathcal{E}_i y \mathcal{O}_i sea pequeña.

Pearson propuso el estadístico

$$\chi^2 = \sum_{i=1}^k \frac{(\mathcal{O}_i - \mathcal{E}_i)^2}{\mathcal{E}_i}$$

el cual, siguiendo la línea de razonamiento anterior debe tomar valores pequeños si H_0 es cierta. Si al tomar una muestra, su valor es grande eso pone en evidencia que la hipótesis inicial es *probablemente* falsa. Para decidir cuando los valores de χ^2 son grandes es necesario conocer su ley de probabilidad. Se tiene entonces el siguiente resultado

Teorema

[Ley asintótica para χ^2] Si la hipótesis H_0 es cierta, entonces χ^2 se distribuye aproximadamente como:

$$\chi^2 = \sum_{i=1}^k \frac{(\mathcal{O}_i - \mathcal{E}_i)^2}{\mathcal{E}_i} \approx \chi_{k-p-h}^2$$

donde el número de grados de libertad depende de

- El número k , de clases usadas;
- El número p de parámetros estimados a partir de la muestra para calcular los \mathcal{E}_i . Por ejemplo si todas las cantidades p_i^0 son especificadas entonces $p = 0$.
- El número de relaciones o condiciones impuestas a los \mathcal{E}_i . Por ejemplo, si la única condición sobre los \mathcal{E}_i es que $\sum_{i=1}^k \mathcal{E}_i = n$ entonces $h = 1$.

La aproximación mejora cuando n es grande y los p_i son cercanos a $\frac{1}{2}$.

Como sólo son los valores grandes de χ^2 los que nos llevan a rechazar H_0 , la región crítica es

$$\mathcal{C} = (\chi_{k-p-h,1-\alpha}^2, \infty)$$

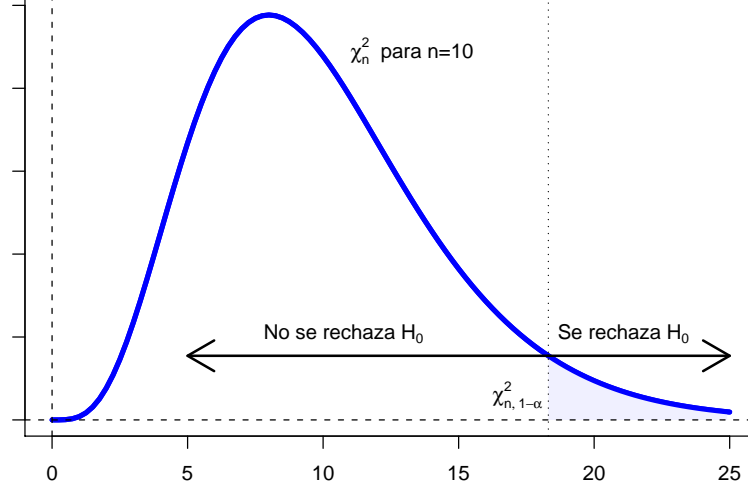


Figura 10.1: Región crítica (sombreada) para un contraste con el estadístico χ^2 .

es decir,

$$\text{sean } \begin{cases} \chi_{exp}^2 = \sum_{i=1}^k \frac{(\mathcal{O}_i - \mathcal{E}_i)^2}{\mathcal{E}_i} \\ \chi_{teo}^2 = \chi_{k-p-h,1-\alpha}^2 \end{cases} \longrightarrow \begin{cases} \text{Si } \chi_{exp}^2 \leq \chi_{teo}^2 \text{ no rechazamos } H_0; \\ \text{Si } \chi_{exp}^2 > \chi_{teo}^2 \text{ se rechaza } H_0 \text{ y se acepta } H_1. \end{cases}$$

Observación

A pesar de que el contraste parece ser bilateral al ver la expresión de la relación (10.1), la forma de \mathcal{C} , nos indica que el contraste es unilateral:

Sólo podemos saber si existe desajuste entre lo esperado y lo observado, pero no podemos contrastar hipótesis alternativas del tipo “ p_i mayor que cierto valor”.

Observación

Obsérvese que en realidad χ^2 no es una variable aleatoria continua: Los posibles resultados de la muestra se resumen en las cantidades $\mathcal{O}_1, \mathcal{O}_2, \dots, \mathcal{O}_k$, que *únicamente* toman valores discretos. Luego las cantidades

$$\chi_{exp}^2(\mathcal{O}_1, \mathcal{O}_2, \dots, \mathcal{O}_k)$$

sólo puede tomar un número finito de valores distintos (aunque sean cantidades con decimales). Por tanto su distribución *no es continua*. Luego al realizar la aproximación mencionada hay que precisar en **qué condiciones el error cometido es pequeño**. De modo aproximado podemos enunciar el siguiente criterio que recuerda al de la aproximación binomial por la distribución normal:

1. $n > 30$;
2. $\mathcal{E}_i = n \cdot p_i > 5$ para todo $i = 1, \dots, k$.

Sin embargo esta regla resulta demasiado estricta a la hora de aplicarla en la práctica. Se utiliza entonces una regla más flexible y que no sacrifica demasiada precisión con respecto a la anterior:

1. Para ninguna clase ocurre que $\mathcal{E}_i = n \cdot p_i < 1$
2. $\mathcal{E}_i = n \cdot p_i > 5$ para casi todos los $i = 1, \dots, k$, salvo a lo sumo un 20 % de ellos.

Si a pesar de todo, estas condiciones no son verificadas, es necesario agrupar las clases que tengan menos elementos con sus adyacentes.

Observación

El lector puede considerar los contrastes con el estadístico χ^2 como una generalización del contraste de proporciones. Para ello le invitamos a estudiar el siguiente ejemplo.

Ejemplo

Se desea saber si cierta enfermedad afecta del mismo modo a los hombres que a las mujeres. Para ello se considera una muestra de $n = 618$ individuos que padecen la enfermedad, y se observa que 341 son hombres y el resto son mujeres. ¿Qué conclusiones se obtiene de ello?

Solución:

El contraste a realizar se puede plantear de dos formas que después veremos que son equivalentes:

Contraste de una proporción: Si p es el porcentaje de hombres en la población de enfermos, podemos considerar el contraste:

$$\begin{cases} H_0 : p = 1/2 \\ H_1 : p \neq 1/2 \end{cases}$$

De la muestra obtenemos la siguiente estimación puntual del porcentaje de enfermos de sexo masculino:

$$\hat{p} = 341/618 = 0,55178$$

Para ver si esto es un valor “coherente” con la hipótesis nula, calculemos la significatividad del contraste:

$$Z_{exp} = \frac{\hat{p} - p}{\sqrt{p * q/n}} \rightsquigarrow \mathbf{N}(0, 1).$$

Por otro lado,

$$Z_{exp} = \frac{0,55178 - 0,5}{\sqrt{0,5 \times 0,5/60}} = 2,574$$

Como el contraste es de tipo bilateral, la significatividad del contraste es (buscando en la tabla de la distribución normal):

$$\mathcal{P}[|Z| > 2,574] = 2 \cdot \mathcal{P}[Z > 2,574] = 2 * 0,005 = 1 \% < 5 \%$$

Lo que nos indica que se ha de rechazar la hipótesis nula y aceptar la hipótesis alternativa, es decir, afirmamos que existe una evidencia significativa a favor de la hipótesis de que la enfermedad no afecta por igual a hombres y mujeres.

Contraste con el estadístico χ^2 : En este caso planteamos el contraste:

$$\left\{ \begin{array}{l} H_0 : \left| \begin{array}{l} p_{hombres} = 1/2 \\ p_{mujeres} = 1/2 \end{array} \right. \quad y \\ H_1 : \left| \begin{array}{l} p_{hombres} \neq 1/2 \\ p_{mujeres} \neq 1/2 \end{array} \right. \quad \text{o bien} \end{array} \right.$$

Para resolverlo escribimos en una tabla los frecuencias muestrales observadas de hombres y mujeres, junto a los valores esperados en el caso de que la hipótesis nula fuese cierta:

	frecuencias observadas \mathcal{O}_i	frecuencias esperadas \mathcal{E}_i	diferencia $\mathcal{O}_i - \mathcal{E}_i$	$(\mathcal{O}_i - \mathcal{E}_i)^2 / \mathcal{E}_i$
Hombres	341	$618 \times 1/2 = 309$	9	$32^2/309$
Mujeres	277	$618 \times 1/2 = 309$	-9	$(-32)^2/309$
	618	618	0	6,63

Consideremos entonces el estadístico

$$\chi^2 = \sum_{i=1}^k \frac{(\mathcal{O}_i - \mathcal{E}_i)^2}{\mathcal{E}_i} \rightsquigarrow \chi_{k-p-h}^2 = \chi_{2-0-1}^2 = \chi_1^2$$

donde:

- $k = 2$ es el número de modalidades posibles que toma la variable sexo: *hombres* y *mujeres*;
- $p = 0$ es el número de parámetros estimados;
- $h = 1$ es el número de restricciones impuestas a los valores esperados. Sólo hay una (que es habitual), que consiste en que el número esperado de enfermos entre hombres y mujeres es 60.

El estadístico calculado sobre la muestra ofrece el valor experimental:

$$\chi_{exp}^2 = 6,63$$

que es el percentil 99 de la distribución χ_1^2 . De nuevo se obtiene que la significatividad del contraste es del $1\% < 5\%$.

En conclusión, con los dos métodos llegamos a que hay una fuerte evidencia en contra de que hay el mismo porcentaje de hombres y mujeres que padecen la enfermedad. La ventaja de la última forma de plantear el contraste (diferencia entre frecuencias observadas y esperadas) es que la técnica se puede aplicar a casos más generales que variables dicotómicas, como se verá más adelante.

Observación

Hay una fórmula alternativa para el cálculo de χ^2 cuya expresión es más fácil de utilizar cuando realizamos cálculos:

Proposición

$$\chi^2 = \sum_{i=1}^k \frac{\mathcal{O}_i^2}{\mathcal{E}_i} - n$$

Demostración

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^k \frac{(\mathcal{O}_i - \mathcal{E}_i)^2}{\mathcal{E}_i} \\
&= \sum_{i=1}^k \frac{\mathcal{O}_i^2 - 2\mathcal{O}_i\mathcal{E}_i + \mathcal{E}_i^2}{\mathcal{E}_i} \\
&= \sum_{i=1}^k \frac{\mathcal{O}_i^2}{\mathcal{E}_i} - 2 \sum_{i=1}^k \mathcal{O}_i + \sum_{i=1}^k \mathcal{E}_i \\
&= \sum_{i=1}^k \frac{\mathcal{O}_i^2}{\mathcal{E}_i} - 2n + n \\
&= \sum_{i=1}^k \frac{\mathcal{O}_i^2}{\mathcal{E}_i} - n
\end{aligned}$$

10.3. Contraste de bondad de ajuste para distribuciones

Vamos a aplicar el contraste χ^2 para determinar a través de una muestra si una v.a. X sigue o no cierta distribución. Podemos encontrarnos entonces con dos casos:

La ley de la v.a. X que deseamos contrastar está completamente determinada.

La ley de la v.a. X no es totalmente conocida y es necesario estimar algunos de sus parámetros.

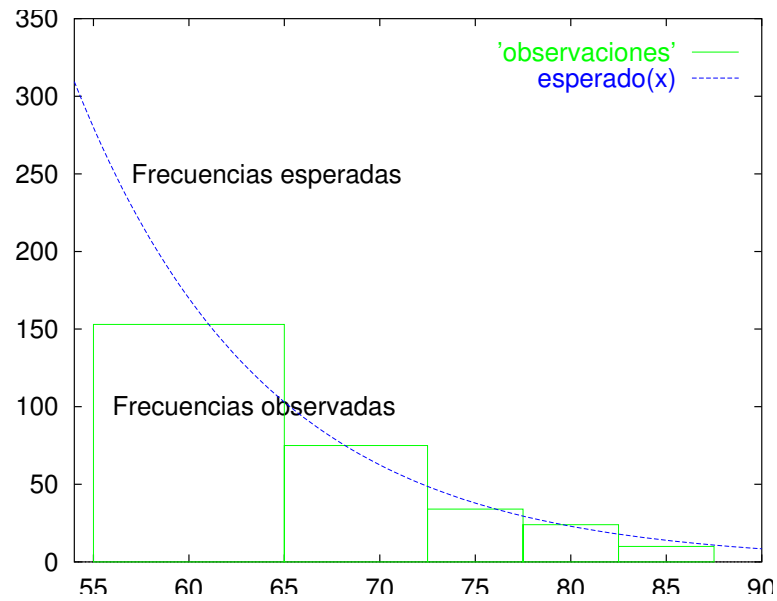


Figura 10.2: En los contrastes de distribuciones, se compara si las observaciones (histograma) se distribuye según una ley de probabilidad conocida.

10.3.1. Distribuciones de parámetros conocidos

Deseamos contrastar si la v.a. X sigue una ley de distribución

$$X \sim \begin{cases} 1 \rightarrow \mathcal{P}[X = 1] = p_1 \\ 2 \rightarrow \mathcal{P}[X = 2] = p_2 \\ \dots \\ i \rightarrow \mathcal{P}[X = i] = p_i \\ \dots \\ k \rightarrow \mathcal{P}[X = k] = p_k \end{cases}$$

donde todos los p_i están fijados (hipótesis H_0). Entonces por lo mencionado anteriormente, el contraste consiste en:

$$\left\{ \begin{array}{l} \chi_{exp}^2 = \sum_{i=1}^k \frac{(\mathcal{O}_i - n p_i)^2}{n p_i} \\ \chi_{teo}^2 = \chi_{k-1, 1-\alpha}^2 \end{array} \right. \longrightarrow \left\{ \begin{array}{l} \text{Si } \chi_{exp}^2 \leq \chi_{teo}^2 \text{ no rechazamos } H_0; \\ \text{Si } \chi_{exp}^2 > \chi_{teo}^2 \text{ se rechaza } H_0 \end{array} \right.$$

En este contraste se comete cierto error de aproximación y por tanto será tanto mejor cuanto mayor sea n .

Ejemplo

Dadas dos parejas de genes Aa y Bb , la descendencia del cruce efectuado según las leyes de Mendel, debe estar compuesto del siguiente modo:

Leyes de Mendel	\longrightarrow	Frecuencias	
		Fenotipo	relativas
		AB	9/16
		Ab	3/16
		aB	3/16
		ab	1/16

Elegidos 300 individuos al azar de cierta población se observa la siguiente distribución de frecuencias:

Frecuencias	
Fenotipo	observadas
AB	165
Ab	47
aB	67
ab	21
Total	300

¿Se puede aceptar que se cumplen las leyes de Mendel sobre los individuos de dicha población?

Solución:

El contraste a realizar es:

$$\left\{ \begin{array}{l} H_0 : \text{Se cumplen las leyes de Mendel} \\ H_1 : \text{No se cumplen} \end{array} \right. \iff \left\{ \begin{array}{l} H_0 : \left\{ \begin{array}{ll} p_{AB} = 9/16 & \text{y} \\ p_{Ab} = 3/16 & \text{y} \\ p_{aB} = 3/16 & \text{y} \\ p_{ab} = 1/16 & \end{array} \right. \\ H_1 : \left\{ \begin{array}{ll} p_{AB} \neq 9/16 & \text{o bien} \\ p_{Ab} \neq 3/16 & \text{o bien} \\ p_{aB} \neq 3/16 & \text{o bien} \\ p_{ab} \neq 1/16 & \end{array} \right. \end{array} \right.$$

Para ello vamos a representar en una sólo tabla las frecuencias observadas, junto con las que serían de esperar en el caso de que H_0 fuese cierta:

Fenotipo	\mathcal{O}_i	\mathcal{E}_i	$\mathcal{O}_i^2/\mathcal{E}_i$
AB	165	$300 \times 9/16 = 168,75$	161,33
Ab	47	$300 \times 3/16 = 52,25$	42,27
aB	67	$300 \times 3/16 = 52,25$	85,91
ab	21	$300 \times 1/16 = 18,75$	23,52
Total	300	300	313,03

Bajo la hipótesis de que H_0 sea cierta, se tiene que:

$$\chi_{exp}^2 = \sum_i \mathcal{O}_i^2/\mathcal{E}_i - n \rightsquigarrow \chi_{4-0-1}^2$$

ya que 4 son los posibles fenotipos, no se ha estimado ningún parámetro (la distribución según las leyes de Mendel es conocida), y sobre las cantidades \mathcal{E}_i existe solamente una restricción, que es: $\sum_i \mathcal{E}_i = 300$.

Por otro lado,

$$\chi_{exp}^2 = \sum_i \mathcal{O}_i^2/\mathcal{E}_i - n = 313,03 - 300 = 13,03$$

que según la tabla de la distribución χ^2 es aproximadamente el percentil 99,5 de la distribución χ_3^2 . Por tanto la significatividad del contraste es del

$0,5\% < 5\%$, lo que nos conduce a rechazar la hipótesis de que la población de la que la muestra ha sido extraída sigue las leyes de Mendel.

Al mismo resultado llegamos sin calcular con precisión la significatividad del contraste, sino considerando que el valor teórico máximo que admitimos para el estadístico experimental con un nivel de significación del 5% es el percentil 95 de χ^2_3 , es decir,

$$\chi^2_{teo} = \chi^2_{3;0,95} = 7,815$$

y claramente ocurre que $\chi^2_{exp} > \chi^2_{teo}$, por lo que se rechaza la hipótesis nula.

Obsérvese también que el que se haya rechazado la hipótesis nula significa que hay diferencia **estadísticamente significativa** entre las frecuencias observadas y las esperadas.

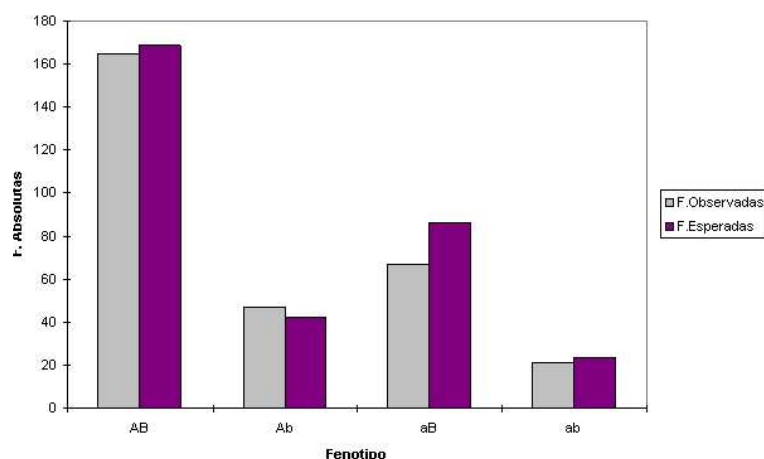


Figura 10.3: Aunque aparentan ser aproximadamente iguales las frecuencias observadas y esperadas, existe diferencia estadísticamente significativa entre ellas.

10.3.2. Distribuciones con parámetros desconocidos

Supongamos que la distribución de X que queremos contrastar no especifica ciertos valores de r parámetros

$$X \rightsquigarrow \mathbf{Fam}(\theta_1, \dots, \theta_r) \implies X \rightsquigarrow \begin{cases} 1 \rightarrow \mathcal{P}[X = 1] = p_1(\theta_1, \dots, \theta_r) \\ 2 \rightarrow \mathcal{P}[X = 2] = p_2(\theta_1, \dots, \theta_r) \\ \dots \\ i \rightarrow \mathcal{P}[X = i] = p_i(\theta_1, \dots, \theta_r) \\ \dots \\ k \rightarrow \mathcal{P}[X = k] = p_k(\theta_1, \dots, \theta_r) \end{cases}$$

Estimemoslos a partir de la muestra, y consideremos las cantidades

$$p_i = p_i(\hat{\theta}_1, \dots, \hat{\theta}_r)$$

Entonces el contraste consiste en

$$\begin{cases} \chi_{exp}^2 = \sum_{i=1}^k \frac{(\mathcal{O}_i - n p_i)^2}{n p_i} \\ \chi_{teo}^2 = \chi_{k-r-1, 1-\alpha}^2 \end{cases} \longrightarrow \begin{cases} \text{Si } \chi_{exp}^2 \leq \chi_{teo}^2 \text{ no rechazamos } H_0; \\ \text{Si } \chi_{exp}^2 > \chi_{teo}^2 \text{ se rechaza } H_0 \end{cases}$$

10.4. Contraste de homogeneidad de muestras cualitativas

Vamos a generalizar el contraste de comparación de dos proporciones (página 244). Consideremos una variable cualitativa (o cuantitativa agrupada en intervalos) que puede tomar valores en diferentes clases. Se toman r muestras diferentes y se desea contrastar:

$$\begin{cases} H_0 : \text{Las } r \text{ muestras son homogéneas con respecto a la variable} \\ H_1 : \text{Alguna muestra es diferente} \end{cases}$$

La manera de proceder consiste en representar las r muestras en una tabla del tipo

	Muestra ₁	Muestra ₂	...	Muestra _r	Frec. clases ↓
Clase ₁	\mathcal{O}_{11}	\mathcal{O}_{12}	...	\mathcal{O}_{1r}	F_1
Clase ₂	\mathcal{O}_{21}	\mathcal{O}_{22}	...	\mathcal{O}_{2r}	F_2
...
Clase _k	\mathcal{O}_{k1}	\mathcal{O}_{k2}	...	\mathcal{O}_{kr}	F_k
Tamaño muestras →	C_1	C_2	...	C_r	T

donde

$\mathcal{O}_{ij} \rightarrow$ frecuencia observada de la clase i en la muestra j

$F_i = \sum_{j=1}^k \mathcal{O}_{ij} \rightarrow$ número de individuos de la clase i

$C_j = \sum_{i=1}^r \mathcal{O}_{ij} \rightarrow$ total de individuos de la muestra j

$T = \sum_{i=1}^r F_i = \sum_{j=1}^k C_j \rightarrow$ total de individuos muestreados

Bajo la hipótesis H_0 , la frecuencia esperada para la clase i en la muestra j es —compárese con la condición de independencia en tablas de doble entrada, relación (??):

$$\mathcal{E}_{ij} = \frac{F_i \cdot C_j}{T}$$

La diferencia entre lo esperado y lo observado la mide el estadístico χ^2

$$\chi_{exp}^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(\mathcal{O}_{ij} - \mathcal{E}_{ij})^2}{\mathcal{E}_{ij}} = \sum_{i=1}^r \sum_{j=1}^k \frac{\mathcal{O}_{ij}^2}{\mathcal{E}_{ij}} - T$$

Su distribución es aproximadamente χ_{gl}^2 , donde los grados de libertad, $gl = a - b - c$, se calculan teniendo en cuenta que

$$a = k \cdot r \rightarrow \text{número de casillas}$$

$$b = k - 1 \rightarrow \text{número de parámetros estimados}$$

$$c = r \rightarrow \text{relaciones impuestas sobre los } \mathcal{E}_{ij} \quad (10.2)$$

Por tanto

$$\chi^2 \rightsquigarrow \chi_{(k-1) \times (r-1)}^2$$

y rechazamos H_0 si $\chi_{exp}^2 > \chi_{(k-1) \times (r-1), 1-\alpha}^2$.

Ejemplo

Se desea saber si la distribución de los grupos sanguíneos es similar en los individuos de dos poblaciones. Para ello se elige una muestra aleatoria simple de cada una de ellas, obteniéndose los datos reflejados en la tabla:

Frec. Obs.	A	B	AB	0
Muestra 1	90	80	110	20
Muestra 2	200	180	240	30

¿Qué conclusiones pueden obtenerse de estos datos si se usa un nivel de significación del 5%?

Solución: Poseemos una variable cualitativa X , que es el grupo sanguíneo, y debemos contrastar si la distribución es la misma en la primera población y la segunda. Para ello planteamos el contraste de homogeneidad conveniente:

$$\left\{ \begin{array}{l} H_0 : \text{La variable } X \text{ se distribuye igualmente en ambas poblaciones} \\ H_1 : \text{La distribución no es homogénea} \end{array} \right.$$

Para ello escribimos la que sería la distribución de frecuencias esperadas. Éstas se calculan a partir de las frecuencias marginales de la distribución de frecuencias esperadas:

Frec. Esp.	A	B	AB	0	
Muestra 1	91,58	82,11	110,53	15,79	300
Muestra 2	198,42	177,89	239,47	34,21	650
	290	260	350	50	950

El estadístico del contraste mide las discrepancia entre las observaciones observadas y esperadas:

$$\chi_{exp}^2 = \sum_{i=1}^2 \sum_{j=1}^4 \frac{\mathcal{O}_{ij}^2}{\mathcal{E}_{ij}} - 950 = \frac{90^2}{91,58} + \dots + \frac{30^2}{34,21} - 950 = 1,76$$

Los valores críticos están a la derecha del percentil 95 de la distribución $\chi_{(2-1) \times (4-1)}^2 = \chi_3^2$, que es $\chi_{teo}^2 = \chi_{3;0,95}^2 = 2,35$. Por tanto de dichas muestras no se obtiene evidencia estadística suficiente en contra de que exista una distribución homogénea del grupo sanguíneo en ambas poblaciones.

10.5. Contraste de independencia de variables cualitativas

A partir de una población se toma mediante muestreo aleatorio simple una muestra de tamaño n . En cada observación se analizan dos características cualitativas A y B (o cuantitativas agrupadas en intervalos), las cuales presentan r y s modalidades respectivamente. Deseamos contrastar si las dos variables son independientes, o sea, queremos realizar un test de significación para las hipótesis:

$$\begin{cases} H_0 : \text{Las características } A \text{ y } B \text{ son independientes} \\ H_1 : \text{Las características } A \text{ y } B \text{ están asociadas} \end{cases}$$

Este test puede ser enunciado de forma equivalente ordenando la muestra en una tabla de doble entrada denominada **tabla de contingencia**, muy parecida a la de la sección anterior:

B	B_1	B_2	\dots	B_j	\dots	B_s	
A							
A_1	n_{11}	n_{12}	\dots	n_{1j}	\dots	n_{1p}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	\dots	n_{2j}	\dots	n_{2p}	$n_{2\bullet}$
\dots	\dots	\dots	\dots	\dots	\dots	\dots	\dots
A_i	n_{i1}	n_{i2}	\dots	n_{ij}	\dots	n_{ip}	$n_{i\bullet}$
\dots	\dots	\dots	\dots	\dots	\dots	\dots	\dots
A_r	n_{r1}	n_{r2}	\dots	n_{rj}	\dots	n_{rp}	$n_{r\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	\dots	$n_{\bullet j}$	\dots	$n_{\bullet p}$	$n_{\bullet \bullet}$

Aunque sobre la población las siguientes probabilidades sean desconocidas, introducimos la siguiente notación

$p_{ij} \rightarrow$ Probabilidad de una observación del tipo (A_i, B_j) ;

$p_{i\bullet} \rightarrow$ Probabilidad de una observación de A_i ;

$p_{\bullet j} \rightarrow$ Probabilidad de una observación de B_j ; (10.3)

Recordando el concepto de independencia entre variables bidimensionales cualitativas, otro modo de escribir el contraste a realizar lo obtenemos basándonos en la relación (??):

$$\begin{cases} H_0 : \forall i = 1, \dots, r \forall j = 1, \dots, s & p_{ij} = p_{i\bullet} p_{\bullet j} \\ H_1 : \exists i = 1, \dots, r \exists j = 1, \dots, s & p_{ij} \neq p_{i\bullet} p_{\bullet j} \end{cases}$$

La idea para realizar este contraste consiste en comparar como en los casos anteriores las frecuencias esperadas bajo la hipótesis H_0 , $\mathcal{E}_{ij} = n_{\bullet\bullet} p_{i\bullet} p_{\bullet j}$, con las obtenidas en la muestra, $\mathcal{O}_{ij} = n_{ij}$. Como las cantidades p_i y p_j no son en principio conocidas, han de ser estimadas a partir de las frecuencias observadas

$$\begin{cases} \hat{p}_{i\bullet} = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \\ \hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n_{\bullet\bullet}} \end{cases} \implies \mathcal{E}_{ij} = n_{\bullet\bullet} \hat{p}_{i\bullet} \hat{p}_{\bullet j} = \frac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}}$$

lo que nos hace perder $(r-1) + (s-1)$ grados de libertad adicionales al estadístico del contraste:

$$\chi_{exp}^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - \mathcal{E}_{ij})^2}{\mathcal{E}_{ij}} \approx \chi_{(r-1) \times (s-1)}^2$$

Luego rechazamos H_0 si $\chi_{exp}^2 > \chi_{(r-1) \times (s-1), 1-\alpha}^2$.

Observación

Aunque el contraste de homogeneidad de muestras es conceptualmente diferente al de independencia de variables cualitativas, obsérvese la analogía existente entre los criterios de aceptación o rechazo de ambas hipótesis.

Ejemplo

500 niños de escuela primaria se clasificaron de acuerdo con el grupo socioeconómico y la presencia o ausencia de cierto defecto en la pronunciación, los resultados son los siguientes:

	Grupo socioeconómico				
	Superior	Medio-Superior	Medio-Inferior	Inferior	Total
Con defecto	8	24	32	27	91
Sin defecto	42	121	138	108	409
Total	50	145	170	135	500

¿Son compatibles estos datos con la hipótesis de que el defecto en la pronunciación, no está relacionado con el grupo socioeconómico?

Solución: En forma de contraste de hipótesis, se ha de realizar el siguiente:

$$\begin{cases} H_0 : \text{Son independientes el nivel socioeconómico y el defecto de pronunciación} \\ H_1 : \text{No son independientes ambas cuestiones.} \end{cases}$$

Para ver si H_0 puede considerarse cierta, o si por el contrario hay una fuerte evidencia a favor de H_1 , fijamos un nivel de significación $\alpha = 0,05$, y analizamos gracias al estadístico χ^2 , las diferencias existentes entre los valores esperados y los observados, de suponer H_0 cierta, es decir, las diferencias entre las cantidades

$$\begin{aligned} \mathcal{O}_{ij} &= n_{ij} \\ \mathcal{E}_{ij} &= \frac{n_{i\bullet} \cdot n_{\bullet j}}{n_{\bullet\bullet}} \end{aligned}$$

Defecto	Grupo socioeconómico				Total
	Superior	Medio superior	Medio inferior	Inferior	
Si	$\mathcal{O}_{11} = 8$	$\mathcal{O}_{12} = 24$	$\mathcal{O}_{13} = 32$	$\mathcal{O}_{14} = 27$	$n_{1\bullet} = 91$
	$\mathcal{E}_{11} = 9,1$	$\mathcal{E}_{12} = 26,39$	$\mathcal{E}_{13} = 30,94$	$\mathcal{E}_{14} = 24,57$	
	$\frac{\mathcal{O}_{11}^2}{\mathcal{E}_{11}} = 7,033$	$\frac{\mathcal{O}_{12}^2}{\mathcal{E}_{12}} = 21,82$	$\frac{\mathcal{O}_{13}^2}{\mathcal{E}_{13}} = 33,096$	$\frac{\mathcal{O}_{14}^2}{\mathcal{E}_{14}} = 29,67$	
No	$\mathcal{O}_{21} = 42$	$\mathcal{O}_{22} = 121$	$\mathcal{O}_{23} = 138$	$\mathcal{O}_{24} = 108$	$n_{2\bullet} = 409$
	$\mathcal{E}_{21} = 40,9$	$\mathcal{E}_{22} = 118,61$	$\mathcal{E}_{23} = 139,06$	$\mathcal{E}_{24} = 110,43$	
	$\frac{\mathcal{O}_{21}^2}{\mathcal{E}_{21}} = 43,130$	$\frac{\mathcal{O}_{22}^2}{\mathcal{E}_{22}} = 123,438$	$\frac{\mathcal{O}_{23}^2}{\mathcal{E}_{23}} = 136,948$	$\frac{\mathcal{O}_{24}^2}{\mathcal{E}_{24}} = 105,623$	
Total	$n_{\bullet 1} = 50$	$n_{\bullet 2} = 145$	$n_{\bullet 3} = 170$	$n_{\bullet 4} = 135$	$n_{\bullet\bullet} = 500$

El número de grados de libertad del estadístico del contraste es $gl = (2 - 1) \times (4 - 1) = 3$. Luego de ser H_0 cierta, la cantidad χ_{exp}^2 no debería superar el valor teórico. que se muestra en la Figura 10.4:

$$\chi_{teo}^2 = \chi_{gl,1-\alpha} = \chi_{3,0'95} = 7'81.$$

Calculemos χ_{exp}^2 :

$$\chi_{exp}^2 = \sum_{i,j} \frac{(\mathcal{O}_{ij} - \mathcal{E}_{ij})^2}{\mathcal{E}_{ij}} = \sum_{i,j} \frac{\mathcal{O}_{ij}^2}{\mathcal{E}_{ij}} - n_{\bullet\bullet} = 500,758 - 500 = 0,758$$

En consecuencia, no existe evidencia significativa a favor de la hipótesis alternativa, o sea, no se rechaza la independencia entre el defecto de pronunciación de los niños de la población y el nivel socioeconómico de su familia.

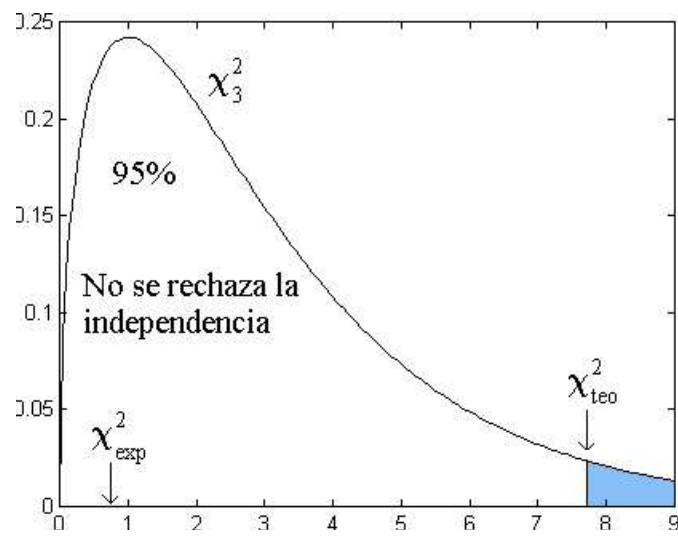


Figura 10.4: Comparación del valor teórico con el experimental.

10.6. Problemas

Ejercicio 10.1. Ante la sospecha de que el hábito de fumar de una embarazada puede influir en el peso de su hijo al nacer, se tomaron dos muestras, una de fumadoras y otra de no fumadoras, y se clasificó a sus hijos en tres categorías en función de su peso en relación con los percentiles \mathcal{P}_{10} y \mathcal{P}_{90} de la población. El resultado se expresa en la tabla siguiente:

¿Madre fumadora?	Peso del niño		
	Menor de \mathcal{P}_{10}	Entre \mathcal{P}_{10} y \mathcal{P}_{90}	Mayor de \mathcal{P}_{90}
Si	117	529	19
No	124	1147	117

¿Hay una evidencia significativa a favor de la sospecha a la vista de los resultados de la muestra?

Ejercicio 10.2. Varios libros de Medicina Interna recomiendan al médico la palpación de la arteria radial con el fin de evaluar el estado de la pared arterial. Se tomaron 215 pacientes y se les clasificó según la *palpabilidad* de dicha arteria (grados 0, 1 y 2 para no palpable, palpable y muy palpable o dura, respectivamente) y según una puntuación de 0 a 4 en orden creciente de *degeneración arterial* (evaluada tras la muerte del paciente y su análisis anatómo-patológico). Los datos son los de la tabla siguiente:

Degeneración	Palpabilidad		
	0	1	2
0	20	5	5
1	60	20	10
2	45	15	15
3	10	5	5

¿Existe relación entre el grado de palpabilidad y el análisis anatomopatológico?

Ejercicio 10.3. Se realizó una encuesta a 2979 andaluces para evaluar su opinión acerca de la atención recibida en los Ambulatorios de la Seguridad Social, clasificándolos también en relación a sus estudios. Analizar los datos

de la siguiente tabla:

Nivel de estudios	Opinión		
	Buena	Regular	Mala
Ninguno	800	144	32
Primarios	905	312	67
Bachiller	287	157	44
Medios	95	48	11
Superiores	38	32	7

Ejercicio 10.4. Con el fin de conocer si un cierto tipo de bacterias se distribuyen al azar en un determinado cultivo o si, por el contrario, lo hacen con algún tipo de preferencia (el centro, los extremos, etc...), se divide un cultivo en 576 áreas iguales y se cuenta el número de bacterias en cada área. Los resultados son los siguientes:

n^o de bacterias	0	1	2	3	4	≥ 5
n^o de áreas	229	211	93	35	7	1

¿Obedecen los datos a una distribución de Poisson?

Ejercicio 10.5. La siguiente tabla recoge la distribución de los triglicéridos en suero, expresados en mg/dl en 90 niños de 6 años:

Nivel de triglicéridos	Frecuencias
10 – 20	5
20 – 30	11
30 – 40	15
40 – 50	24
50 – 60	18
60 – 70	12
70 – 80	4
80 – 90	1

Contrastar la hipótesis de que el nivel de triglicéridos en niños de 6 años

sigue una distribución Normal.

Ejercicio 10.6. La distribución en Andalucía del grupo sanguíneo es de un 35 %, 10 %, 6 % y un 49 % para los grupos A, B, AB y O respectivamente. En Málaga, se realizó el estudio en una muestra de 200 individuos obteniéndose una distribución del 50 %, 30 %, 18 %, y 10 % para los grupos A, B, AB y O respectivamente.

Se desea saber si la distribución del grupo sanguíneo en dicha provincia es igual que en Andalucía.

Ejercicio 10.7. En un estudio diseñado para determinar la aceptación por una parte de los pacientes de un nuevo analgésico, 100 médicos seleccionaron cada uno de ellos una muestra de 25 pacientes para participar en el estudio. Cada paciente después de haber tomado el nuevo analgésico durante un periodo de tiempo determinado, fue interrogado para saber si prefería éste o el que había tomado anteriormente con regularidad, obteniendo los siguientes resultados:

n° de pacientes que prefieren el nuevo analgésico	n° de médicos que obtienen estos resultados	n° total de pacientes que prefieren el nuevo analgésico
0	5	0
1	6	6
2	8	16
3	10	30
4	10	40
5	15	75
6	17	102
7	10	70
8	10	80
9	9	81
10 o más	0	0
Total	100	500

Queremos saber si estos datos se ajustan a una distribución binomial.

Ejercicio 10.8. Disponemos de una muestra de 250 mujeres mayores de 18 años, cuyos pesos son los presentados en la tabla adjunta, y queremos saber si los datos de esta muestra provienen de una distribución Normal.

Pesos	n° de mujeres
30 – 40	16
40 – 50	18
50 – 60	22
60 – 70	51
70 – 80	62
80 – 90	55
90 – 100	22
100 – 110	4

Ejercicio 10.9. Deseamos conocer, si las distribuciones atendiendo al grupo sanguíneo, en tres muestras referidas atendiendo al tipo de tensión arterial, se distribuyen de igual manera. Para lo cual, se reunió una muestra de 1500 sujetos a los que se les determinó su grupo sanguíneo y se les tomó la tensión arterial, clasificándose ésta en baja, normal, y alta. Obteniéndose los siguientes resultados:

Tensión arterial	Grupo sanguíneo				Total
	A	B	AB	O	
Baja	28	9	7	31	75
Normal	543	211	90	476	1.320
Alta	44	22	8	31	105
Total	615	242	105	538	1.500

Ejercicio 10.10. La recuperación producida por dos tratamientos distintos A y B se clasifican en tres categorías: muy buena, buena y mala. Se administra el tratamiento A a 30 pacientes y B a otros 30: De las 22 recuperaciones muy buenas, 10 corresponden al tratamiento A; de las 24 recuperaciones buenas, 14 corresponden al tratamiento A y de los 14 que tienen una mala recuperación corresponden al tratamiento A. ¿Son igualmente efectivos

ambos tratamientos para la recuperación de los pacientes?

Capítulo 11

Análisis de la varianza

11.1. Introducción

Del mismo modo que el contraste χ^2 generalizaba el contraste de dos proporciones, es necesario definir un nuevo contraste de hipótesis que sea aplicable en aquellas situaciones en las que el número de medias que queremos comparar sea superior a dos. Es por ello por lo que el **análisis de la varianza**, **ANOVA**¹ surge como una generalización del contraste para dos medias de la **t** de Student, cuando el *número de muestras a contrastar es mayor que dos*.

Por ejemplo, supongamos que tenemos 3 muestras de diferentes tamaños que suponemos que provienen de tres poblaciones normales con la misma varianza:

$$\begin{array}{ll} \vec{x}_1 \in \mathbb{R}^{n_1} & X_1 \sim \mathbf{N}(\mu_1, \sigma^2) \\ \vec{x}_2 \in \mathbb{R}^{n_2} & X_2 \sim \mathbf{N}(\mu_2, \sigma^2) \\ \vec{x}_3 \in \mathbb{R}^{n_3} & X_3 \sim \mathbf{N}(\mu_3, \sigma^2) \end{array}$$

Si queremos realizar el contraste

¹Del término inglés “*Analysis of variance*”.

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 \\ H_1 : \mu_1 \neq \mu_2 \text{ ó } \mu_1 \neq \mu_3 \text{ ó } \mu_2 \neq \mu_3 \end{cases}$$

podríamos en plantearnos como primer método el fijar una cantidad α próxima a cero y realizar los $\binom{3}{2} = 3$ contrastes siguientes con α como nivel de significación:

$$\begin{cases} H'_0 : \mu_1 = \mu_2 \\ H'_1 : \mu_1 \neq \mu_2 \end{cases} \quad \text{nivel de significación } \alpha$$

$$\begin{cases} H''_0 : \mu_1 = \mu_3 \\ H''_1 : \mu_1 \neq \mu_3 \end{cases} \quad \text{nivel de significación } \alpha$$

$$\begin{cases} H'''_0 : \mu_2 = \mu_3 \\ H'''_1 : \mu_2 \neq \mu_3 \end{cases} \quad \text{nivel de significación } \alpha$$

de modo que se aceptaría H_1 y se rechazaría H_0 sólo si alguna de las hipótesis alternativas H'_1 , H''_1 ó H'''_1 es aceptada y rechazada su correspondiente hipótesis nula. El error de tipo I para este contraste es:

$$\begin{aligned} & \mathcal{P}_{rob} [\text{Rechazar } H_0 | H_0 \text{ es cierta}] \\ &= 1 - \mathcal{P}_{rob} [\text{No rechazar } H_0 | H_0 \text{ es cierta}] \\ &= 1 - \mathcal{P}_{rob} [\text{No rechazar } H'_0 \text{ ni } H''_0 \text{ ni } H'''_0 | H'_0 \text{ y } H''_0 \text{ y } H'''_0 \text{ son ciertas}] \\ &= 1 - (1 - \alpha)^3 \end{aligned}$$

Por ello el nivel de significación obtenido para este contraste sobre la igualdad de medias de tres muestras no es α como hubiésemos esperado obtener inicialmente, sino $1 - (1 - \alpha)^3$. Por ejemplo, si tomamos un nivel de significación $\alpha = 0.1$ para cada uno de los contrastes de igualdad de dos medias, se obtendría que el nivel de significación (error de tipo I) para el contraste de las tres medias es de $1 - 0.9^3 = 0.27$, lo que es una cantidad muy alta para lo que acostumbramos a usar.

En consecuencia, *no es adecuado realizar el contraste de igualdad de medias de varias muestras mediante una multitud de contrastes de igualdad de medias de dos muestras.*

Una técnica que nos permite realizar el contraste de modo conveniente es la que exponemos en este capítulo y que se denomina **análisis de la varianza**.

11.2. ANOVA con un factor

Se denomina **modelo factorial con un factor** o **ANOVA con un factor** al modelo (lineal) en el que la variable analizada la hacemos depender de un sólo factor de tal manera que las causas de su variabilidad son englobadas en una componente aleatoria que se denomina **error experimental**:

$$X = \text{factor} \pm \text{error}$$

Vamos a exponer esto con más claridad. Consideremos una variable sobre la que actúa un factor que puede presentarse bajo un determinado número de niveles, t . Por ejemplo podemos considerar un fármaco que se administra a $t = 3$ grupos de personas y se les realiza cierta medición del efecto causado:

	Resultado de la medición								
Gripe (nivel 1)	5	3	2	5	4	3			$\rightarrow n_1 = 6$
Apendicitis (nivel 2)	8	9	6	7	8	9	10	8	$\rightarrow n_2 = 9$
Sanos (nivel 3)	2	3	2	1	2	3	2		$\rightarrow n_3 = 7$

En este caso los factores que influyen en las observaciones son tres: el que la persona padezca la gripe, apendicitis, o que esté sana.

De modo general podemos representar las t muestras (o niveles) del siguiente modo:

Niveles	Observaciones de X				tamaños muestrales
Nivel 1 $\equiv N_1$	x_{11}	x_{12}	\cdots	x_{1n_1}	n_1
Nivel 2 $\equiv N_2$	x_{21}	x_{22}	\cdots	x_{2n_2}	n_2
\cdots			\cdots		\cdots
Nivel $t \equiv N_t$	x_{t1}	x_{t2}	\cdots	x_{tn_t}	n_t

donde por supuesto, los tamaños de cada muestra n_i , no tienen por que ser iguales. En este caso decimos que se trata del **modelo no equilibrado**.

Observación

De ahora en adelante asumiremos que las siguientes condiciones son verificadas por las t muestras:

- Las observaciones proceden de poblaciones normales;
- Las t muestras son aleatorias e independientes. Además, dentro de cada nivel las observaciones son independientes entre sí.
- En el modelo de un factor suponemos que las observaciones del nivel i , x_{ij} , provienen de una variable X_{ij} de forma que todas tienen la misma varianza —hipótesis de homocedasticidad:

$$X_{ij} \sim \mathbf{N}(\mu_i, \sigma^2) \quad j = 1, \dots, n_i$$

o lo que es lo mismo,

$$X_{ij} = \mu_i + \epsilon_{ij}, \quad \text{donde } \epsilon_{ij} \sim \mathbf{N}(0, \sigma^2)$$

De este modo μ_i es el valor esperado para las observaciones del nivel i , y los errores ϵ_{ij} son variables aleatorias independientes, con valor

esperado nulo, y con el mismo grado de dispersión para todas las observaciones.

Otro modo de escribir lo mismo consiste en introducir una cantidad μ que sea el valor esperado para una persona cualquiera de la población (sin tener en cuenta los diferentes niveles), y considerar los efectos α_i introducidos por los niveles, de modo que

$$\begin{aligned}\mu_i &= \mu + \alpha_i & i = 1, \dots, t \\ \sum_{i=1}^t n_i \alpha_i &= 0\end{aligned}$$

11.2.1. Especificación del modelo

Con todo lo anterior, el modelo ANOVA de un factor puede escribirse como

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \text{donde } \epsilon_{ij} \rightsquigarrow \mathbf{N}(0, \sigma^2)$$

y con la siguiente interpretación:

- μ es una constante común a todos los niveles;
- α_i es el efecto producido por el i -ésimo nivel. Al sumarlos todos deben compensarse los efectos negativos con los positivos para que la media común a todos los niveles sea realmente μ . Esto implica en particular que los efectos, α_i , de los niveles no son independientes;
- ϵ_{ij} es la parte de la variable X_{ij} no explicada por μ ni α_i , y que se distribuye del mismo modo (aunque independientemente) para cada observación, según la ley gaussiana:

$$\epsilon_{ij} \rightsquigarrow \mathbf{N}(0, \sigma^2)$$

Ésta es la condición de **homocedasticidad**, y es fundamental en el análisis de la varianza.

Obsérvese que ahora podemos escribir el contraste de que los diferentes niveles no tienen influencia sobre la observación de la variable como:

$$\begin{cases} H_0 & : \mu_1 = \mu_2 = \cdots = \mu_t \\ H_1 & : \text{Al menos dos son distintos} \end{cases}$$

o bien

$$\begin{cases} H_0 & : \alpha_1 = \alpha_2 = \cdots = \alpha_t = 0 \\ H_1 & : \text{Algún } \alpha_i \neq 0 \end{cases}$$

Observación

Se utiliza el nombre de *análisis de la varianza* ya que el elemento básico del análisis estadístico será precisamente el estudio de la variabilidad. Teóricamente es posible dividir la variabilidad de la variable que se estudia en dos partes:

- La originada por el factor en cuestión;
- La producida por los restantes factores que entran en juego, conocidos o no, controlables o no, que se conocen con el nombre de error experimental.

Si mediante los contrastes estadísticos adecuados la variación producida por cierto factor es significativamente mayor que la producida por el error experimental podemos aceptar la hipótesis de que los distintos niveles del factor actúan de forma distinta.

Ejemplo

Consideremos dos muestras tomadas en diferentes niveles de una variable, de forma que ambas tengan la misma varianza muestral (lo que indica que no se puede rechazar la igualdad de varianzas poblacionales) y medias muestrales bastante diferentes. Por ejemplo:

$$\left. \begin{array}{l} \text{nivel 1} \\ \overbrace{1, 2, 3} \rightsquigarrow \left\{ \begin{array}{l} n_1 = 3 \\ \bar{x}_1 = 2 \\ \hat{S}_1^2 = 1 \end{array} \right\} \\ \\ \text{nivel 2} \\ \overbrace{11, 12, 13} \rightsquigarrow \left\{ \begin{array}{l} n_2 = 3 \\ \bar{x}_2 = 12 \\ \hat{S}_2^2 = 1 \end{array} \right\} \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} n = n_1 + n_2 = 6 \\ \bar{x} = 7 \\ \hat{S}^2 \approx 5,55 \end{array} \right.$$

La dispersión calculada al medir la de los dos niveles conjuntamente es mucho mayor que la de cada uno de ellos por separado. Por tanto puede deducirse que ambos niveles no tienen el mismo valor esperado.

11.2.2. Algo de notación relativa al modelo

Este apartado está dedicado a introducir alguna notación para escribir los términos que serán más importantes a la hora de realizar un contraste por el método ANOVA. En primer lugar tenemos:

$$\begin{aligned} N &= \sum_{i=1}^t n_i && \text{número total de observaciones (entre todos los niveles)} \\ x_{i\bullet} &= \sum_{j=1}^{n_i} x_{ij} && \text{suma de las observaciones del nivel } i \\ \bar{x}_{i\bullet} &= \frac{x_{i\bullet}}{n_i} && \text{media muestral del nivel } i \\ x_{\bullet\bullet} &= \sum_{i=1}^t \sum_{j=1}^{n_i} x_{ij} = \sum_{i=1}^t n_i \bar{x}_{i\bullet} && \text{suma de todas las observaciones} \\ \bar{x}_{\bullet\bullet} &= \frac{x_{\bullet\bullet}}{N} && \text{media muestral de todas las observaciones} \end{aligned}$$

Usando estos términos vamos a desglosar la variación total de la muestra en variación total dentro de cada nivel (intravariación) más la variación entre los distintos niveles (intervariación). Para ello utilizamos la proposición ?? (página ??):

$$SCT = SCD + SCE$$

donde

$$\begin{aligned} SCT &= \sum_{i=1}^t \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2 && \text{Suma de Cuadrados Totales} \\ SCD &= \sum_{i=1}^t \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2 && \text{SC Dentro de cada nivel} \\ SCE &= \sum_{i=1}^t n_i (\bar{x}_{i.} - \bar{x}_{..})^2 && \text{SC Entre todos los niveles} \end{aligned}$$

Observación

En el cálculo del estadístico SCT intervienen N cantidades, ligadas por una relación:

$$x_{..} = \sum_{i=1}^t \sum_{j=1}^{n_i} x_{ij}$$

de este modo el número de grados de libertad de este estadístico es $N - 1$ (recuérdese la noción de grados de libertad de un estadístico, página ??). Por razones análogas tenemos que el número de grados de libertad de SCD es $N - t$ y el de SCE es $t - 1$. Así introducimos los siguientes estadísticos:

$$\hat{S}_T^2 = \frac{SCT}{N - 1} \quad \text{Cuasivarianza total} \quad (11.1)$$

$$\hat{S}_E^2 = \frac{SCE}{t - 1} \quad \text{Intervarianza} \quad (11.2)$$

$$\hat{S}_D^2 = \frac{SCD}{N - t} \quad \text{Intravarianza} \quad (11.3)$$

Estos son los estadísticos que realmente nos interesan a la hora de realizar el contraste de igualdad de medias. Cuando la diferencia entre los efectos de los diferentes niveles sea muy baja, es de esperar que la cuasi-varianza total sea próxima a la intravarianza, o lo que es lo mismo, que la intervianza sea pequeña en relación con la intravarianza.

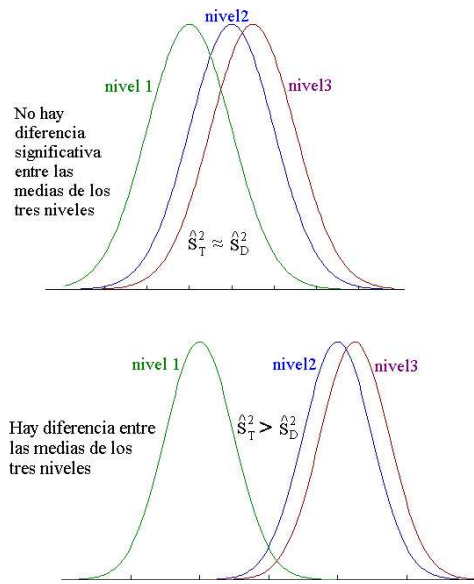


Figura 11.1: En la figura de superior no existe una evidencia significativa en contra de que las medias de los tres grupos de observaciones coinciden. En la figura inferior sí.

11.2.3. Forma de efectuar el contraste

Consideramos el contraste

$$\begin{cases} H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_t = 0 \\ H_1 : \text{Algún } \alpha_i \neq 0 \end{cases}$$

y suponemos que estamos en las condiciones del modelo factorial de un

factor. Si H_0 es cierta se puede demostrar que el siguiente estadístico se distribuye como una \mathbf{F} de Snedecor:

$$F_{exp} = \frac{\hat{S}_E^2}{\hat{S}_D^2} \sim \mathbf{F}_{t-1, N-t}$$

Luego si al calcular F_{exp} obtenemos que $F_{exp} > F_{t-1, N-t, 1-\alpha}$ donde α es un nivel de significación dado, deberemos de rechazar la hipótesis nula (ya que si H_0 fuese cierta, era de esperar que \hat{S}_E^2 fuese pequeño en relación con \hat{S}_D^2).

11.2.4. Método reducido para el análisis de un factor

En este apartado vamos a resumir lo más importante de lo visto hasta ahora, indicando la forma más sencilla de realizar el contraste. En primer lugar calculamos los siguientes estadísticos a partir de la tabla de las observaciones en cada nivel:

$$\begin{aligned} A &= \sum_{i=1}^t \sum_{j=1}^{n_i} x_{ij}^2 \\ B &= \sum_{i=1}^t \frac{x_{i\bullet}^2}{n_i} \\ C &= \frac{x_{\bullet\bullet}^2}{N} \end{aligned}$$

Niveles	Observaciones de X				Cálculos al margen			
Nivel 1	x_{11}	x_{12}	\cdots	x_{1n_1}	n_1	$x_{1\bullet}$	$\frac{x_{1\bullet}^2}{n_1}$	$\sum_{j=1}^{n_1} x_{1j}^2$
Nivel 2	x_{21}	x_{22}	\cdots	x_{2n_2}	n_2	$x_{2\bullet}$	$\frac{x_{2\bullet}^2}{n_2}$	$\sum_{j=1}^{n_2} x_{2j}^2$
\dots	\dots				\dots	\dots		
Nivel t	x_{t1}	x_{t2}	\cdots	x_{tn_t}	n_t	$x_{t\bullet}$	$\frac{x_{t\bullet}^2}{n_t}$	$\sum_{j=1}^{n_t} x_{tj}^2$
					N	$x_{\bullet\bullet}$	B	A

Entonces las siguientes cantidades admiten una expresión muy sencilla:

$$\begin{aligned}
 SCE &= B - C & \implies & \hat{S}_E^2 = \frac{SCE}{t-1} \\
 SCT &= A - C \\
 SCD &= A - B & \implies & \hat{S}_D^2 = \frac{SCD}{N-t}
 \end{aligned}$$

Calculamos

$$F_{exp} = \frac{\hat{S}_E^2}{\hat{S}_D^2}$$

y dado el nivel de significación α buscamos en una tabla de la distribución **F** de Snedecor el valor

$$F_{teo} = F_{t-1, N-t, 1-\alpha}$$

rechazando H_0 si $F_{exp} > F_{teo}$. como se aprecia en la Figura 11.2.

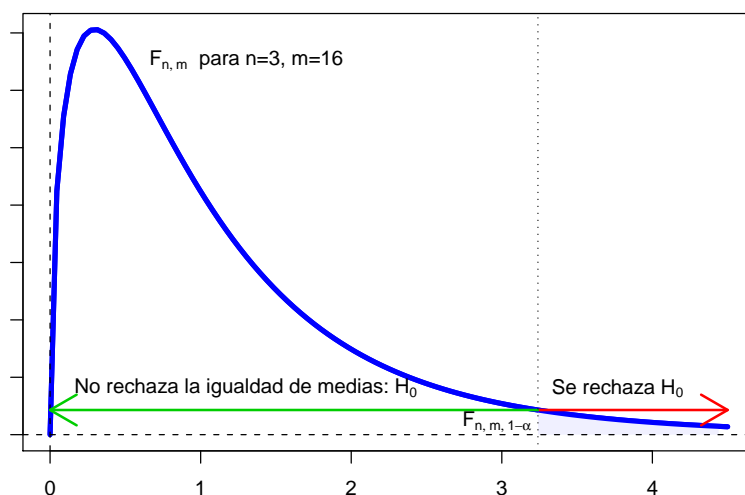


Figura 11.2: Región crítica en un contraste ANOVA.

Ejemplo

Se aplican 4 tratamientos distintos a 4 grupos de 5 pacientes, obteniéndose los resultados de la tabla que se adjunta. Queremos saber si se puede concluir que todos los tratamientos tienen el mismo efecto. Para ello vamos a suponer que estamos en condiciones de aplicar el modelo de un factor².

²Esto es algo que debe ser contrastado previamente. En principio la independencia entre las observaciones es algo bastante natural a la hora de realizar un estudio, pero no lo es tanto la condición de homocedasticidad. Más adelante veremos ciertos contrastes de homocedasticidad que deben ser siempre realizados antes de aplicar esta técnica: test de Cochran y test de Bartlett.

Tratamientos	Observaciones	n_i	$x_{i\bullet}$	$\frac{x_{i\bullet}^2}{n_i}$	$\sum_{j=1}^{n_i} x_{ij}^2$
Tratamiento 1	-1 1 2 0 -1	5	1	1/5	7
Tratamiento 2	-2 -4 -5 -4 -7	5	-22	484/5	110
Tratamiento 3	0 -1 -2 -4 -1	5	-8	64/5	22
Tratamiento 4	1 4 6 3 8	5	22	484/5	126
		$N = 20$	$x_{\bullet\bullet} = 7$	$B = \frac{1,033}{5}$	$A = 265$
		\Downarrow $C = \frac{49}{20}$			

Fuente de variación	grados de libertad	Suma cuadrados	Cuasivarianzas	Estadístico
Entre tratamientos	$t - 1 = 3$	$SC\mathcal{E} = B - C$ $= 204,15$	$\hat{S}_E^2 = \frac{SC\mathcal{E}}{t-1}$ $= 68,167$	$F_{exp} = \frac{\hat{S}_E^2}{\hat{S}_D^2}$ $= 18,676$
Dentro de los tratamientos	$N - t = 16$	$SC\mathcal{D} = A - B$ $= 58,4$	$\hat{S}_D^2 = \frac{SC\mathcal{D}}{N-t}$ $= 3,65$	$F_{teo} = F_{t-1, N-t}$ $= 3,24$

En conclusión, $F_{exp} > F_{teo}$, por tanto se ha de rechazar la igualdad de efectos de los tratamientos.

En la Figura 11.4 se representan las observaciones de cada nivel de tratamiento mediante una curva normal cuyos parámetros se han estimado puntualmente a partir de las observaciones. Obsérvese que las diferencias más importantes se encuentran entre Los tratamientos 2 y 4. Esto motiva los contrastes de comparaciones múltiples (dos a dos), para que, en el caso en que la igualdad de medias sea rechazada, se pueda establecer qué niveles tuvieron mayor influencia en esta decisión.

11.2.5. Análisis de los resultados del ANOVA: Comparaciones múltiples

Una vez contrastado el que existen diferencias significativas mediante el análisis de la varianza, nos interesa conocer que niveles del factor son los que han influido más para que se de este resultado. Como ilustración, en

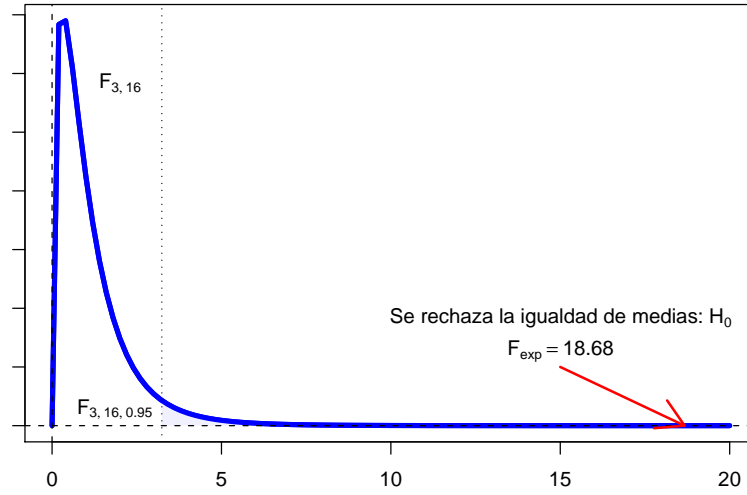


Figura 11.3: Se rechaza la hipótesis de que los tratamientos tienen el mismo efecto en los diferentes grupos. Hay gran evidencia estadística en contra.

el último ejemplo se ve claramente que los tratamientos segundo y cuarto dan resultados muy diferentes, y probablemente de ahí venga el que se haya rechazado la igualdad de todos los efectos.

El método más simple es el de *Bonferroni*, que consiste en realizar todas las comparaciones por parejas:

$$\left\{ \begin{array}{l} H_0 : \mu_i = \mu_j \\ H_1 : \mu_i \neq \mu_j \end{array} \right. \quad i, j = 1, \dots, t \quad i \neq j \quad \Rightarrow \binom{t}{2} \text{ contrastes}$$

lo que corresponde a los ya conocidos contrastes de la t de Student, que tienen en este caso como estadístico experimental a (de nuevo suponiendo la homocedasticidad en todas las muestras):

$$T_{exp} = \frac{\bar{x}_i - \bar{x}_j}{\hat{S}_D \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \rightsquigarrow t_{N-t}$$

11.3. CONSIDERACIONES SOBRE LAS HIPÓTESIS SUBYACENTES EN EL MODELO FAC

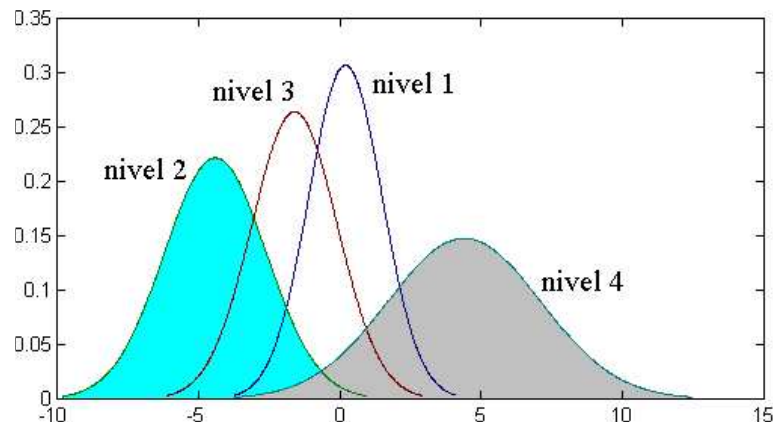


Figura 11.4: Las diferencias más importantes se encuentran entre los niveles 2 y 4.

ya que la intravarianza \hat{S}_D , es un estimador de σ^2 con $N - t$ grados de libertad.

Sin embargo el nivel de significación de los contrastes debe ser disminuido para tener en cuenta que ahora al hacer multitud de contrastes aumenta la probabilidad del error de tipo I. Para una probabilidad de error de tipo I (nivel de significación) α , el procedimiento de comparaciones múltiples de Bonferroni nos indica que declaremos significativas las diferencias entre muestras cuando estas sean significativas en contrastes bilaterales para el estadístico anterior para el nivel de significación

$$\alpha' = \frac{\alpha}{\binom{t}{2}}$$

11.3. Consideraciones sobre las hipótesis subyacentes en el modelo factorial

Para aplicar el modelo de un factor hemos hecho, entre otras, las siguientes suposiciones:

- Las observaciones de cada muestra han de ser independientes y también la de las muestras entre sí. Para ello podemos aplicar cualquiera de los contrastes no paramétricos de aleatoriedad. En principio esta aleatoriedad es algo que es bastante razonable admitir si la metodología para elegir los datos (muestreo) ha sido realizada siguiendo técnicas adecuadas.
- Los datos han de ser normales en cada una de las muestras. Esto es algo que debería ser contrastado previamente antes de utilizar el ANOVA de un factor mediante, por ejemplo, el test de ajuste a la distribución normal mediante el estadístico χ^2 que ya conocemos, o bien el test de d'Agostino, que veremos más adelante en la página 308, y que es mucho más cómodo de utilizar;
- Las varianzas de cada muestra son todas iguales, es decir:

$$\begin{cases} H_0 & : \sigma_1 = \sigma_2 = \cdots = \sigma_t \\ H_1 & : \text{Algún } \sigma_i \neq \sigma_j \end{cases}$$

Para esto podemos utilizar un par de contrastes que exponemos brevemente a continuación: contraste de Cochran y contraste de Bartlett.

11.3.1. Contraste de homocedasticidad de Cochran

Este test se aplica cuando $n = n_1 = n_2 = \cdots = n_t$ y si ha sido verificada previamente la aleatoriedad y la normalidad de las observaciones. En este caso $N = t \cdot n$. El estadístico del contraste es:

$$R_{exp} = \frac{\max \left\{ \hat{S}_i^2 \right\}_{i=1}^t}{\sum_{i=1}^t \hat{S}_i^2}$$

donde se define \hat{S}_i^2 como la cuasivarianza de la muestra del nivel i , es decir

$$\hat{S}_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\bullet})^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} x_{ij}^2 - \frac{n_i}{n_i - 1} \bar{x}_{i\bullet}^2$$

11.3. CONSIDERACIONES SOBRE LAS HIPÓTESIS SUBYACENTES EN EL MODELO FAC

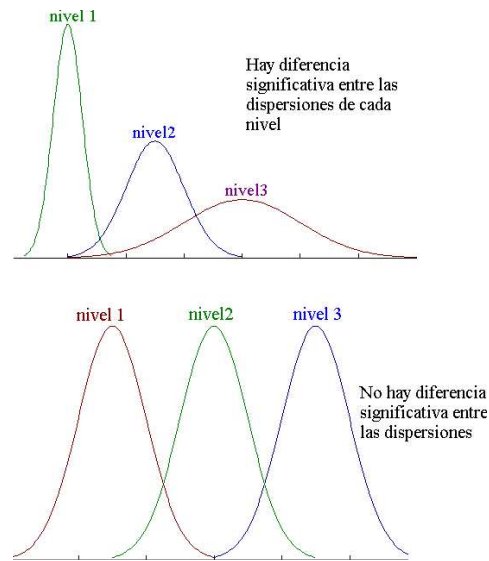


Figura 11.5: En la figura superior hay evidencia en contra de la homocedasticidad en las tres muestras. En la inferior, no.

Fijado un nivel de significación α se busca en la tabla de la distribución de Cochran el valor

$$R_{teo} = R_{n-1,t,1-\alpha}$$

y se rechaza H_0 si $R_{exp} > R_{teo}$.

11.3.2. Contraste de homocedasticidad de Bartlett

Este test se aplica si estamos en la misma situación que en el de Cochran, pero en este caso no es necesario el que todas las muestras sean del mismo tamaño. El estadístico del contraste es:

$$\chi_{exp}^2 = \frac{1}{k} \left[(N - t) \ln \hat{S}_D^2 - \sum_{i=1}^t \ln \hat{S}_i^2 \right]$$

siendo

$$k = 1 + \frac{1}{3(t-1)} \left(\sum_{i=1}^t \frac{1}{n_i - 1} - \frac{1}{N - t} \right)$$

Se rechaza H_0 si $\chi_{exp}^2 > \chi_{t-1, 1-\alpha}^2$

11.4. Problemas

1.- Para evaluar la influencia del tipo de acidosis del recién nacido en los niveles de glucemia medidos en el cordón umbilical del mismo, se obtuvieron los datos de la siguiente tabla:

Niveles de glucemia										
Controles	51	56	58	60	62	63	65	68	72	73
Acid. Respiratoria	60	65	66	68	68	69	73	75	78	80
Acid. Metabólica	69	73	74	78	79	79	82	85	87	88
Acid. Mixta	70	75	76	77	79	80	82	86	88	89

Obtener conclusiones a partir de los resultados de esas muestras.

2.- Se desea saber si el grado de ansiedad es el mismo, por término medio, en tres enfermedades distintas. Para ello se tomaron tres muestras de 10, 12 y 8 personas, respectivamente, con esas enfermedades, pasándoles a cada una de ellas un test que mide el grado de ansiedad del individuo. Los resultados se dan en la tabla adjunta.

Enfermedad	Grado de ansiedad										
A	4	6	5	5	6	3	3	2	6	5	
B	2	1	5	5	4	6	4	4	4	3	2
C	7	5	8	7	9	3	5	5			

¿Que puede concluirse de los datos?.

3.- En una experiencia para comparar la eficacia de diversas técnicas en el tratamiento del dolor producido por una intervención quirúrgica superficial, 28 pacientes se agruparon al azar en 4 grupos de 7, tratando al primero con placebo, y a los siguientes con dos tipos de analgésicos (A y B) y acupuntura. Los datos se dan en la siguiente tabla:

Tratamiento	Minutos para la remisión del dolor						
Placebo	35	22	5	14	38	42	65
Analgésico A	85	80	46	61	99	114	110
Analgésico B	100	107	142	88	63	94	70
Acupuntura	86	125	103	99	154	75	160

¿Que conclusiones pueden obtenerse de esta experiencia?.

4.- Se está llevando a cabo un estudio para comprobar el efecto de tres dietas diferentes en el nivel de colesterolina de pacientes hipercolesterinémicos. Para ello se han seleccionado al azar 3 grupos de pacientes, de tamaños 12, 8 y 10. Los niveles de colesterolina medidos después de 2 semanas de dieta se representan a continuación:

Dieta	Nivel de colesterolina											
A	2'9	3'35	3'25	3	3'3	3'1	3'25	3'25	3'1	3'05	3'25	3
B	3'15	2'95	2'8	3'1	2'75	2'6	2'8	3'05				
C	3	2'6	2'65	2'2	2'55	2'3	2'35	2'6	2'35	2'6		

Analice los resultados obtenidos.

5.- En un colectivo de 5 individuos se aplican 3 fármacos para estudiar su influencia sobre sus movimientos respiratorios (número de inspiraciones por minuto). Los valores obtenidos para cada individuo vienen expresados en la tabla:

	Individuos				
	1	2	3	4	5
Antes de los tratamientos	14	16	18	15	20
Después de <i>I</i>	16	17	21	16	24
Después de <i>II</i>	15	14	18	15	22
Después de <i>III</i>	17	16	20	13	18

Estudie si el efecto de estos fármacos en la variación respiratoria producida

puede considerarse o no el mismo.

Capítulo 12

Contrastes no paramétricos

12.1. Introducción

Hasta ahora todas las técnicas utilizadas para realizar algún tipo de inferencia exigían:

- bien asumir de ciertas hipótesis como la *aleatoriedad* en las observaciones que componen la muestra, o la *normalidad* de la población, o la igualdad de varianzas de dos poblaciones, etc;
- o bien, la estimación de cualquier parámetro como la *media*, *varianza*, *proporción*, etc, de la población.

El conjunto de estas técnicas de inferencia se denominan **técnicas paramétricas**. Existen sin embargo otros métodos paralelos cuyos procedimientos no precisan la estimación de parámetros ni suponer conocida ninguna ley de probabilidad subyacente en la población de la que se extrae la muestra. Estas son las denominadas **técnicas no paramétricas** o **contrastes de distribuciones libres**, algunos de los cuales desarrollamos en este capítulo. Sus mayores atractivos residen en que:

- Son más fáciles de aplicar que las alternativas paramétricas;

- Al no exigir ninguna condición suplementaria a la muestra sobre su proveniencia de una población con cierto tipo de distribución, son más generales que las paramétricas, pudiéndose aplicar en los mismos casos en que estas son válidas.

Por otro lado, esta liberación en los supuestos sobre la población tiene inconvenientes. El principal es la falta de sensibilidad que poseen para detectar efectos importantes. En las técnicas no paramétricas juega un papel fundamental la ordenación de los datos, hasta el punto de que en gran cantidad de casos ni siquiera es necesario hacer intervenir en los cálculos las magnitudes observadas, más que para establecer una relación de menor a mayor entre las mismas, denominadas **rangos**.

12.2. Aleatoriedad de una muestra: Test de rachas

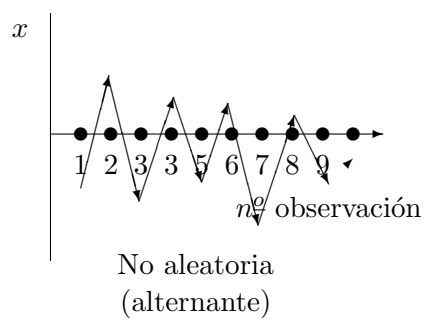
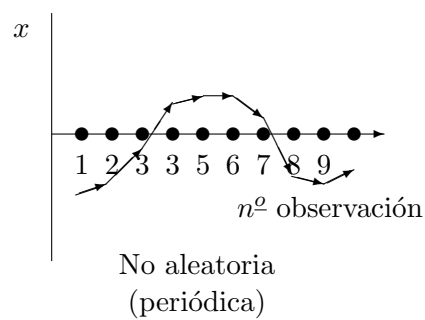
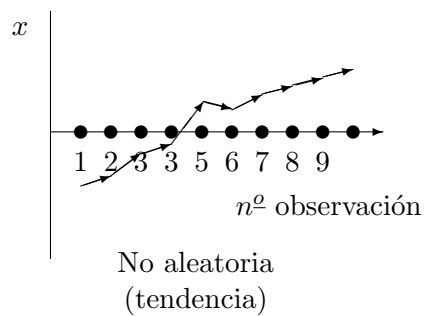
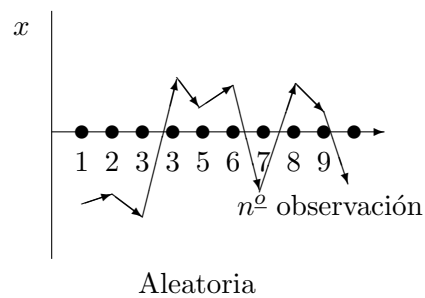
A veces al realizar un muestreo, puede llegar a influir el orden temporal o espacial en que las muestras han sido elegidas, con lo cual no estamos en las condiciones de un muestreo aleatorio simple, ya que la ley de probabilidad varía de una observación a otra. Como ilustración obsérvese la figura adjunta. También podemos denominar a este contraste como **test de independencia de las observaciones de una muestra**.

Consideremos una muestra de tamaño n que ha sido dividida en dos categorías \ominus y \oplus con n_1 y n_2 observaciones cada una. Se denomina **racha** a una sucesión de valores de la misma categoría. Por ejemplo si estudiamos una población de personas podemos considerar como categoría el sexo

$\ominus \equiv$ ser hombre

$\oplus \equiv$ ser mujer

$$\overbrace{\underbrace{\oplus \oplus \oplus}_3 \underbrace{\ominus \ominus}_2 \underbrace{\oplus}_1 \underbrace{\ominus \ominus}_3}_{4 \text{ rachas}} \quad \left\{ \begin{array}{l} n_1 = 5 \\ n_2 = 4 \\ n = n_1 + n_2 = 9 \end{array} \right.$$



En función de las cantidades n_1 y n_2 se espera que el número de rachas no sea ni muy pequeño ni muy grande.

Si las *observaciones son cantidades numéricas* estas pueden ser divididas en dos categorías que poseen aproximadamente el mismo tamaño ($n_1 = n_2 \pm 1$), si consideramos la mediana de las observaciones como el valor que sirve para dividir a la muestra:

$$\begin{aligned}\ominus &\equiv \text{observación inferior a la mediana} \\ \oplus &\equiv \text{observación superior a la mediana}\end{aligned}$$

Se define la v.a. R como el *número de rachas*. Su distribución está tabulada para los casos $n_1 \leq 20$ y $n_2 \leq 20$ (tabla 7 de Downie). La aleatoriedad en la extracción de la muestra se rechaza cuando $R \leq R_{n_1, n_2, \alpha/2}$ ó $R \geq R_{n_1, n_2, 1-\alpha/2}$.

12.3. Normalidad de una muestra: Test de D'Agostino

Consideremos n observaciones, las cuales ordenamos de menor a mayor y les asignamos su rango en función de este orden

$$\begin{array}{c} \boxed{\begin{array}{c} \text{Observaciones} \\ \text{ordenadas} \end{array}} \rightarrow x_1 \quad x_2 \quad x_3 \quad \cdots \quad x_i \quad \cdots \quad x_n \\[10pt] \boxed{\begin{array}{c} \text{Rango} \end{array}} \rightarrow 1 \quad 2 \quad 3 \quad \cdots \quad i \quad \cdots \quad n \end{array}$$

Se calculan sobre la muestra la media, la desviación típica un estadístico T y por último el estadístico del contraste D cuya distribución está tabulada

$$T = \sum_{i=1}^n \left(i - \frac{n+1}{2} \right) x_i = \sum_{i=1}^n i x_i - \frac{n(n+1)}{2} \bar{x} \quad (12.1)$$

$$D = \frac{T}{n^2 \mathcal{S}} \quad (12.2)$$

En la tabla de la distribución del estadístico de D'Agostino, (tabla 8) D , para un nivel de significación α , se busca un intervalo $(D_{n,\alpha}, D^{n,\alpha})$ de modo

que si $D \notin (D_{n,\alpha}, D^{n,\alpha})$ se rechaza la normalidad y en otro caso se asume. Para realizar este test es necesario que al menos $n \geq 10$.

12.4. Equidistribución de dos poblaciones

Estas son las alternativas no paramétricas del contraste de la t de Student para poblaciones normales (sección §9.5, página 228). Están concebidas para contrastar la hipótesis de que dos muestras aleatorias independientes

$$\begin{aligned}\vec{x} &= x_1, x_2, \dots, x_{n_1} \\ \vec{y} &= y_1, y_2, \dots, y_{n_2}\end{aligned}$$

proviene de poblaciones que tienen idénticas distribuciones. Para aplicar estos contrastes será en primer lugar necesario contrastar si cada una de las muestras se ha obtenido mediante un mecanismo aleatorio. Esto puede realizarse mediante un test de rachas.

Supongamos que el contraste de aleatoriedad de ambas muestras (*cuantitativas*) no permite que ésta se rechace a un nivel de significación α . Entonces aplicaremos el **contraste de Mann—Withney** o el de **rachas de Wald—Wolfowitz**, que exponemos a continuación.

12.4.1. Contraste de rachas de Wald—Wolfowitz

Si combinamos las dos muestras y disponemos el conjunto completo de todas las observaciones, ordenadas de menor a mayor, cabe esperar que bajo la hipótesis

H_0 : Las poblaciones de las que provienen las muestras están equidistribuidas

las dos muestras estén muy entremezcladas, y por tanto el número de rachas, R_{exp} , formadas por las categorías

$$\begin{aligned}\ominus &\equiv \text{Observación de la muestra } \vec{x} \\ \oplus &\equiv \text{Observación de la muestra } \vec{y}\end{aligned}$$

debe ser muy alto.

Cuando $n_1, n_2 \leq 20$ el valor teórico del número de rachas por debajo del cual se rechaza H_0 ,

$$R_{teo} = R_{n_1, n_2, \alpha}$$

se busca en la tabla 7 (de Downie) y entonces no se rechaza H_0 si $R_{exp} \geq R_{teo}$ y se rechaza en otro caso.

12.4.2. Contraste de Mann—Withney

El objetivo es el mismo que el del test anterior: contrastar la hipótesis

$$\begin{cases} H_0 & : \text{ Las poblaciones de las que provienen las muestras están equidistribuidas} \\ H_1 & : \text{ Las poblaciones no están equidistribuidas} \end{cases}$$

para dos muestras \vec{x}, \vec{y} cuantitativas independientes, tomadas de modo aleatorio. El contraste se efectúa combinando las dos muestras y disponiendo el conjunto completo de las observaciones, ordenado de menor a mayor. Se asignan después números de rango a cada observación

Observaciones unidas y ordenadas	$\vec{z} = \vec{x} \cup \vec{y}$	\rightarrow	z_1	z_2	z_3	\cdots	z_i	\cdots	$z_{n_1+n_2}$
Rango		\rightarrow	1	2	3	\cdots	i	\cdots	$n_1 + n_2$

Se calcula después la suma de los rangos de las observaciones pertenecientes a la primera muestra y a la segunda, obteniéndose respectivamente R_1 y R_2 , para después calcular los estadísticos

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \quad (12.3)$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2 = n_1 n_2 - U_1 \quad (12.4)$$

Entonces si la hipótesis H_0 es cierta, U_1 y U_2 tienen una distribución de Mann—Withney de parámetros n_1 y n_2 que está tabulada (tabla 9) para

valores en que

$$\begin{cases} \text{máx}\{n_1, n_2\} \leq 40 \\ \text{mín}\{n_1, n_2\} \geq 20 \end{cases}$$

Para el **contraste bilateral**, se define

$$U_{exp} = \text{mín}\{U_1, U_2\} \quad (12.5)$$

y se rechaza H_0 si $U_{exp} < U_{n_1, n_2, \alpha}$.

Si el contraste que pretendemos realizar es **unilateral**, como por ejemplo,

$$\begin{cases} H_0 & : \text{La primera población toma valores menores o iguales a la segunda} \\ H_1 & : \text{Los de la segunda son menores} \end{cases}$$

rechazaremos la hipótesis nula si $U_1 < U_{n_1, n_2, \alpha}$. Si el test es el contrario

$$\begin{cases} H_0 & : \text{La segunda población toma valores menores o iguales a la primera} \\ H_1 & : \text{Los de la primera son menores} \end{cases}$$

se rechaza H_0 si $U_2 < U_{n_1, n_2, \alpha}$.

12.5. Contraste de Wilcoxon para muestras apareadas

El **contraste de Wilcoxon** es la técnica no paramétrica paralela a el de la **t** de Student para muestras apareadas (sección §9.4, página 224). Igualmente tendríamos de n parejas de valores (x_i, y_i) que podemos considerar como una variable medida en cada sujeto en dos momentos diferentes.

$\forall i = 1, \dots, n$, i -ésima observación $\equiv (x_i, y_i) \rightarrow$ diferencia $\equiv d_i = x_i - y_i$

El test de Wilcoxon, al igual que los otros contrastes no paramétricos puede realizarse siempre que lo sea su homólogo paramétrico, con el inconveniente

de que este último detecta diferencias significativas en un 95 % de casos que el de la t de Student.

Sin embargo a veces las hipótesis necesarias para el test paramétrico (normalidad de las diferencias apareadas, d_i) no se verifican y es estrictamente necesario realizar el contraste que presentamos aquí. Un caso muy claro de no normalidad es cuando los datos pertenecen a una escala ordinal.

El procedimiento consiste en:

1. Ordenar las cantidades $|d_i|$ de menor a mayor y obtener sus rangos.
2. Consideramos las diferencias d_i cuyo signo (positivo o negativo) tiene menor frecuencia (no consideramos las cantidades $d_i = 0$) y calculamos su suma, T

$$T = \begin{cases} \sum_{d_i > 0} i & \text{si los signos positivos de } d_i \text{ son menos frecuentes;} \\ \sum_{d_i < 0} i & \text{si los signos negativos de } d_i \text{ son menos frecuentes.} \end{cases}$$

Del mismo modo es necesario calcular la cantidad T' , suma de los rangos de las observaciones con signo de d_i de mayor frecuencia, pero si hemos ya calculado T la siguiente expresión de T' es más sencilla de usar

$$T' = m(n + 1) - T$$

donde m es el número de rangos con signo de d_i de menor frecuencia.

3. Si T ó T' es menor o igual que las cantidades que aparecen en la tabla de Wilcoxon (tabla número 10), se rechaza la hipótesis nula del contraste

$$\begin{cases} H_0 & : \text{ No hay diferencia entre las observaciones apareadas} \\ H_1 & : \text{ Si la hay} \end{cases}$$

12.6. Contraste de Kruskal–Wallis

El contraste de **Kruskal–Wallis** es la alternativa no paramétrica del método *ANOVA*, es decir, sirve para contrastar la hipótesis de que k muestras cuantitativas han sido obtenidas de la misma población. La única exigencia versa sobre la aleatoriedad en la extracción de las muestras, no haciendo referencia a ninguna de las otras condiciones adicionales de homocedasticidad y normalidad necesarias para la aplicación del test paramétrico ANOVA.

De este modo, este contraste es el que debemos aplicar necesariamente cuando no se cumple algunas de las condiciones que se necesitan para aplicar dicho método.

Al igual que las demás técnicas no paramétricas, ésta se apoya en el uso de los rangos asignados a las observaciones.

Para la exposición de este contraste, supongamos que tenemos k muestras representadas en una tabla como sigue,

Niveles	Observaciones de X			
Nivel 1 $\equiv N_1$	x_{11}	x_{12}	\cdots	x_{1n_1}
Nivel 2 $\equiv N_2$	x_{21}	x_{22}	\cdots	x_{2n_2}
\cdots			\cdots	
Nivel $k \equiv N_k$	x_{k1}	x_{k2}	\cdots	x_{kn_k}

El número total de elementos en todas las muestras es:

$$N = n_1 + n_2 + \cdots + n_k \quad (12.6)$$

La hipótesis a contrastar es:

$$\left\{ \begin{array}{l} H_0 : \text{Las } k \text{ muestras provienen de la misma población} \\ H_1 : \text{Alguna proviene de una población con mediana diferente a las demás} \end{array} \right.$$

El modo de realizar el contraste es el siguiente:

- Se ordenan las observaciones de menor a mayor, asignando a cada una de ellas su rango (1 para la menor, 2 para la siguiente, \dots , N para la mayor).

- Para cada una de las muestras, se calcula R_i , $i = 1, \dots, k$, como la suma de los rangos de las observaciones que les corresponden. Si H_0 es falsa, cabe esperar que esas cantidades sean muy diferentes.
- Se calcula el estadístico:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1) \quad (12.7)$$

La regla para decidir si se ha de rechazar o no la hipótesis nula es la siguiente:

- Si el número de muestras es $k = 3$ y el número de observaciones en cada una de ellas no pasa de 5 se rechaza H_0 si el valor de H supera el valor teórico que encontramos en la tabla de Kruskal–Wallis —tabla número 11.
- En cualquier otro caso, se compara el valor de H con el de la tabla de la χ^2_{k-1} con $k - 1$ grados de libertad. Se rechaza H_0 si el valor del estadístico supera el valor teórico $\chi^2_{k-1, 1-\alpha}$.

12.7. Problemas

1.- Recientes estudios sobre el ejercicio de la Medicina en centros en los que no actúan estudiantes, indican que la duración media de la visita por paciente es de 22 minutos. Se cree que en centros donde con un elevado número de estudiantes en prácticas esta cifra es menor. Se obtuvieron los siguientes datos sobre las visitas de 20 pacientes aleatoriamente seleccionados:

Duración en minutos de la visita							
21'6	13'4	20'4	16'4	23'5	26'8	24'8	19'3
23'4	9'4	16'8	21'9	24'9	15'6	20'1	16'2
18'7	18'1	19'1	18'9				

1. ¿Constituyen estos datos una muestra aleatoria?

2. ¿Podemos concluir en base a estos datos que la población de la cual fue extraída esta muestra sigue una distribución Normal?

2.- Se realiza un estudio para determinar los efectos de poner fin a un bloqueo renal en pacientes cuya función renal está deteriorada a causa de una metástasis maligna avanzada de causa no urológica. Se mide la tensión arterial de cada paciente antes y después de la operación. Se obtienen los siguientes resultados:

Tensión arterial

Antes	150	132	130	116	107	100	101	96	90	78
Después	90	102	80	82	90	94	84	93	89	87

¿Se puede concluir que la intervención quirúrgica tiende a disminuir la tensión arterial?

3.- Se ensayaron dos tratamientos antirreumáticos administrados al azar, sobre dos grupos de 10 pacientes, con referencia a una escala convencional (a mayor puntuación, mayor eficacia), valorada después del tratamiento. Los resultados fueron:

Nivel de eficacia del tratamiento

Tratamiento primero	12	15	21	17	38	42	10	23	35	28
Tratamiento segundo	21	18	25	14	52	65	40	43	35	42

Decidir si existe diferencia entre los tratamientos.

4.- Puesto que el hígado es el principal lugar para el metabolismo de los fármacos, se espera que los pacientes con enfermedades de hígado tengan dificultades en la eliminación de fármacos. Uno de tales fármacos es la fenilbutazona. Se realiza un estudio de la respuesta del sistema a este fármaco. Se estudian tres grupos: controles normales, pacientes con cirrosis hepática,

pacientes con hepatitis activa crónica. A cada individuo se les suministra oralmente 19 mg de fenilbutazona/Kg. de peso. Basándose en los análisis de sangre se determina para cada uno el tiempo de máxima concentración en plasma (en horas). Se obtienen estos datos:

Normal	Cirrosis	Hepatitis
4	22'6	16'6
30'6	14'4	12'1
26'8	26'3	7'2
37'9	13'8	6'6
13'7	17'4	12'5
49		15'1
		6'7
		20

¿Se puede concluir que las tres poblaciones difieren respecto del tiempo de máxima concentración en plasma de fenilbutazona?

5.- El administrador de un laboratorio está considerando la compra de un aparato para analizar muestras de sangre. En el mercado hay 5 de tales aparatos. Se le pide a cada uno de los 7 técnicos médicos que después de probar los aparatos, les asignen un rango de acuerdo con el orden de preferencia, dándole el rango 1 al preferido. Se obtienen los siguientes datos:

Técnico	Analizador de sangre				
	<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>	<i>V</i>
1	1	3	4	2	5
2	4	5	1	2	3
3	4	1	3	5	2
4	1	3	2	5	4
5	1	2	3	4	5
6	5	1	3	2	4
7	5	1	4	3	2

Utilizar el contraste adecuado para determinar si los técnicos perciben diferencias entre los aparatos.

6.- Los efectos de tres drogas con respecto al tiempo de reacción a cierto estímulo fueron estudiados en 4 grupos de animales experimentales. El grupo *IV* sirvió de grupo control, mientras que a los grupos *I*, *II* y *III* les fueron aplicadas las drogas A, B y C respectivamente, con anterioridad a la aplicación del estímulo:

$I \leftarrow A$	$II \leftarrow B$	$III \leftarrow C$	$IV \leftarrow \text{Control}$
17	8	3	2
20	7	5	5
40	9	2	4
31	8	9	3
35			

¿Puede afirmarse que los tres grupos difieren en cuanto al tiempo de reacción?

7.- La tabla siguiente muestra los niveles de residuo pesticida (PPB) en muestras de sangre de 4 grupos de personas. Usar el test de Kruskal–Wallis para contrastar a un nivel de confianza de 0'05, la hipótesis nula de que no existe diferencia en los niveles de PPB en los cuatro grupos considerados.

Niveles de PPB

Grupo <i>I</i>	10	37	12	31	11	9	23
Grupo <i>II</i>	4	35	32	19	33	18	8
Grupo <i>III</i>	15	5	10	12	6	6	15
Grupo <i>IV</i>	7	11	1	08	2	5	3

8.- La cantidad de aminoácidos libres fue determinada para 4 especies de ratas sobre 1 muestra de tamaño 6 para cada especie. Comprobar si el contenido de aminoácidos libres es el mismo para las 4 especies.

Especies de ratas

<i>I</i>	<i>II</i>	<i>III</i>	<i>IV</i>
431'1	477'1	385'5	366'8
440'2	479'0	387'9	369'9
443'2	481'3	389'6	371'4
445'5	487'8	391'4	373'2
448'6	489'6	399'1	377'2
451'2	403'6	379'4	381'3

9.- Los siguientes datos nos dan el peso de comida (en Kg.) consumidos por adulto y día en diferentes momentos en un año. Usar un contraste no paramétrico para comprobar si el consumo de comida es el mismo en los 4 meses considerados.

Febrero	Mayo	Agosto	Noviembre
4'7	4'7	4'8	4'9
4'9	4'4	4'7	5'2
5'0	4'3	4'6	5'4
4'8	4'4	4'4	5'1
4'7	4'1	4'7	5'6

10.- Se hizo un estudio neurofisiológico sobre la conducción motora tibial posterior en dos grupos de pacientes embarazadas con las siguientes determinaciones:

Conducción motora tibial posterior

Primer grupo	51	40	41	53	48	50	45	58	45	44
Segundo grupo	58	43	40	45	41	42	44	52	56	48

Comprobar la igualdad o no de ambas muestras.

11.- En un experimento diseñado para estimar los efectos de la inhalación prolongada de óxido de cadmio, 15 animales de laboratorio sirvieron de su-

jetos para el experimento, mientras que 10 animales similares sirvieron de controles. La variable de interés fue el nivel de hemoglobina después del experimento. Se desea saber si puede concluirse que la inhalación prolongada de óxido de cadmio disminuye el nivel de hemoglobina según los siguientes datos que presentamos:

Nivel de hemoglobina

Expuestos	14'4	14'2	13'8	16'5	14'1	16'6	15'9	15'6	14'1	15'3
	15'7	16'7	13'7	15'3	14'0					
No expuestos	17'4	16'2	17'1	17'5	15'0	16'0	16'9	15'0	16'3	16'8

12.- A 11 ratas tratadas crónicamente con alcohol se les midió la presión sanguínea sistólica antes y después de 30 minutos de administrarles a todas ellas una cantidad fija de etanol, obteniéndose los datos siguientes:

Presión sanguínea sistólica

Antes	126	120	124	122	130	129	114	116	119	112	118
Después	119	116	117	122	127	122	110	120	112	110	111

¿Hay un descenso significativo de la presión sanguínea sistólica tras la ingestión de etanol?

13.- Un test de personalidad, tiene dos formas de determinar su valoración suponiendo inicialmente que ambos métodos miden igualmente la extraversión. Para ello se estudia en 12 personas obteniéndose los siguientes resultados:

Medida de la extraversión

Forma A	12	18	21	10	15	27	31	6	15	13	8	10
Forma B	10	17	20	5	21	24	29	7	11	13	8	11

¿Hay diferencia entre los dos métodos?

Bibliografía

- [AB 92] P. ARMITAGE, G. BERRY, *Estadística para la Investigación Biomédica*. Doyma, Barcelona, 1992.
- [Cal 74] G. CALOT, *Curso de Estadística Descriptiva*. Paraninfo, Madrid, 1974.
- [Car 82] J.L. CARRASCO DE LA PEÑA, *El Método Estadístico en la Investigación Médica*. Karpus, Madrid, 1982.
- [Dan 90] W.W. DANIEL, *Applied Nonparemetric Statistics*. PWS-Kent Publishing Company, Boston, 1990.
- [Ham 90] L.C. HAMILTON, *Modern Data Analysis*. Brooks/Cole Publishing Company, Pacific Grove, 1990.
- [Mar 94] A. MARTÍN ANDRÉS, J.D. LUNA DEL CASTILLO, *Bioestadística para las Ciencias de la salud*. Norma, Granada, 1994.
- [MS 88] L.A. MARASCUILO, R.C. SERLIN, *Statistical Methods for the Social and Behavioral Sciences*. W.H. Freeman and Company, Nueva York, 1988.
- [Peñ 94] D. PEÑA SÁNCHEZ DE RIVERA, *Estadística: Modelos y Métodos*, 1. Alianza Universidad Textos, Madrid, 1994.
- [RMR 91] T. RIVAS MOYA, M.A. MATEO, F. RÍUS DÍAZ, M. RUIZ, *Estadística Aplicada a las Ciencias Sociales: Teoría y Ejercicios (EAC)*. Secretariado de Publicaciones de la Universidad de Málaga, Málaga, 1991.

- [**RM 92**] E. RUBIO CALVO, T. MARTÍNEZ TERRER Y OTROS, *Bioestadística*. Colección Monografías Didácticas, Universidad de Zaragoza, Zaragoza, 1992.
- [**RS 79**] R.D. REMINGTON, M.A. SCHORK, *Estadística Biométrica y Sanitaria*. Prentice Hall International, Madrid, 1979.
- [**Rum 77**] L. RUIZ-MAYA, *Métodos Estadísticos de investigación (Introducción al Análisis de la Varianza)*. I.N.E. Artes Gráficas, Madrid, 1977.
- [**SR 90**] E. SÁNCHEZ FONT, F. RÍUS DÍAZ, *Guía para la Asignatura de Bioestadística (EAC)*. Secretariado de Publicaciones de la Universidad de Málaga, Málaga, 1990.
- [**ST 85**] STEEL, TORRIE, *Bioestadística (Principios y Procedimientos)*. Mac Graw-Hill, Bogotá, 1985.
- [**Tso 89**] M. TSOKOS, *Estadística para Psicología y Ciencias de la Salud*. Interamericana Mac Graw-Hill, Madrid, 1989.
- [**WG 82**] S.L. WEINBERG, K.P. GOLDBERG, *Estadística Básica para las Ciencias Sociales*. Nueva Editorial Interamericana, Mexico, 1982.
- [**Zar 74**] J.H. ZAR, *Biostatistical Analysis*. Prentice Hall Inc., Englewood Cliffs, 1974.