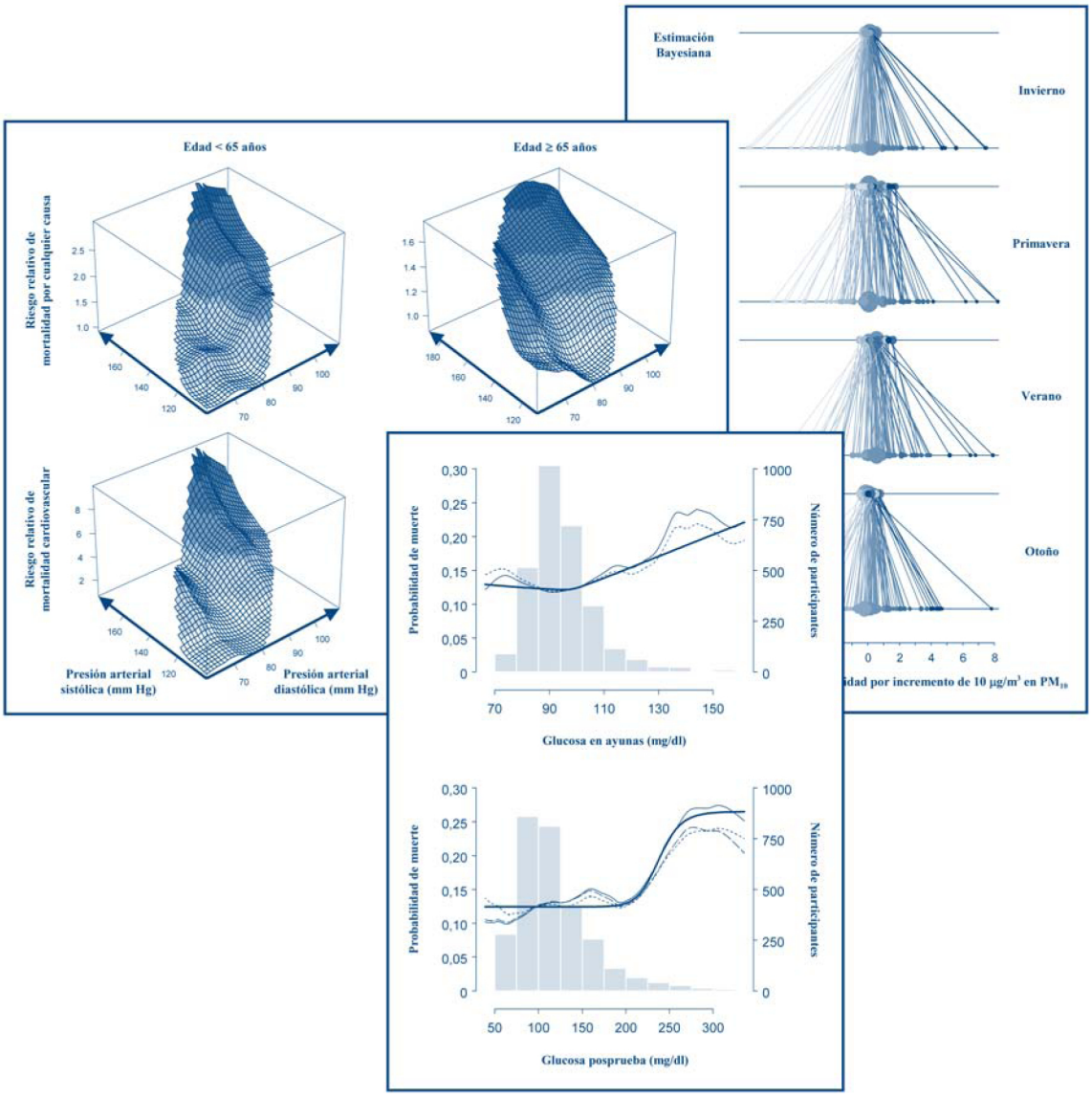
Instituto de Salud Carlos III
Escuela Nacional de Sanidad
Centro Nacional de Epidemiología

Roberto Pastor-Barriuso

BIOESTADÍSTICA



GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
Instituto de Salud Carlos III
Centro Nacional de Epidemiología
ENS Escuela Nacional de Sanidad

Centro Nacional de Epidemiología
Instituto de Salud Carlos III
Monforte de Lemos, 5
28029 MADRID (ESPAÑA)
Tel.: 91 822 20 00
Fax: 91 387 78 15
http://www.isciii.es

Escuela Nacional de Sanidad
Instituto de Salud Carlos III
Sinesio Delgado, 8
28029 MADRID (ESPAÑA)
Tel.: 91 822 22 74
Fax: 91 387 78 56
http://www.isciii.es

Catálogo general de publicaciones oficiales:
http://publicacionesoficiales.boe.es/

Para obtener este libro de forma gratuita en internet (formato pdf):
http://publicaciones.isciii.es/

EDITA: ESCUELA NACIONAL DE SANIDAD y CENTRO NACIONAL
DE EPIDEMIOLOGÍA - Instituto de Salud Carlos III
Madrid, diciembre de 2012

N.I.P.O. (en línea): 477-11-083-3
I.S.B.N.: 978-84-695-3775-6

Imprime: Agencia Estatal Boletín Oficial del Estado.
Avda. de Manoteras, 54. 28050 – MADRID

BIOESTADÍSTICA

Roberto Pastor-Barriuso

Científico Titular

Centro Nacional de Epidemiología,
Instituto de Salud Carlos III,
Madrid

Para citar este libro

Pastor-Barriuso R. Bioestadística. Madrid: Escuela Nacional de Sanidad y Centro Nacional de Epidemiología, Instituto de Salud Carlos III, 2012.

A la memoria de Carmen

A Marta, Pablo, Miguel y Antonio

ÍNDICE

TEMA 1

ESTADÍSTICA DESCRIPTIVA

1.1 INTRODUCCIÓN

La estadística es la rama de las matemáticas aplicadas que permite estudiar fenómenos cuyos resultados son en parte inciertos. Al estudiar sistemas biológicos, esta incertidumbre se debe al desconocimiento de muchos de los mecanismos fisiológicos y fisiopatológicos, a la incapacidad de medir todos los determinantes de la enfermedad y a los errores de medida que inevitablemente se producen. Así, al realizar observaciones en clínica o en salud pública, los resultados obtenidos contienen una parte sistemática o estructural, que aporta información sobre las relaciones entre las variables estudiadas, y una parte de "ruido" aleatorio. El objeto de la estadística consiste en extraer la máxima información sobre estas relaciones estructurales a partir de los datos recogidos.

En estadística se distinguen dos grandes grupos de técnicas:

- La **estadística descriptiva**, en la que se estudian las técnicas necesarias para la organización, presentación y resumen de los datos obtenidos.
- La **estadística inferencial**, en la que se estudian las bases lógicas y las técnicas mediante las cuales pueden establecerse conclusiones sobre la población a estudio a partir de los resultados obtenidos en una muestra.

El análisis de una base de datos siempre partirá de técnicas simples de resumen de los datos y presentación de los resultados. A partir de estos resultados iniciales, y en función del diseño del estudio y de las hipótesis preestablecidas, se aplicarán las técnicas de inferencia estadística que permitirán obtener conclusiones acerca de las relaciones estructurales entre las variables estudiadas. Las técnicas de estadística descriptiva no precisan de asunciones para su interpretación, pero en contrapartida la información que proporcionan no es fácilmente generalizable. La estadística inferencial permite esta generalización, pero requiere ciertas asunciones que deben verificarse para tener un grado razonable de seguridad en las inferencias.

A continuación se definen algunos conceptos generales que aparecen repetidamente a lo largo de la exposición:

- **Población** es el conjunto de todos los elementos que cumplen ciertas propiedades y entre los cuales se desea estudiar un determinado fenómeno.
- **Muestra** es un subconjunto de la población seleccionado mediante un mecanismo más o menos explícito. En general, rara vez se dispone de los recursos necesarios para estudiar a toda la población y, en consecuencia, suelen emplearse muestras obtenidas a partir de estas poblaciones.

Ejemplo 1.1 Algunos ejemplos de poblaciones son:

— Las personas residentes en Washington D.C. a 1 de enero de 2010.

— Las personas infectadas con el virus de inmunodeficiencia humana en Brasil a día de hoy.

Para estas poblaciones, algunas muestras podrían ser:

— 500 residentes en Washington D.C. a 1 de enero de 2010 seleccionados mediante llamadas telefónicas aleatorias.

— Todas las personas que acuden a un hospital de Río de Janeiro durante el presente año para realizarse un test del virus de inmunodeficiencia humana y que resultan ser positivas.

- **Variables** son propiedades o cualidades que presentan los elementos de una población. Las variables pueden clasificarse en:
  - **Variables cualitativas o atributos** son aquellas que no pueden medirse numéricamente y que, a su vez, pueden ser:
    — **Nominales**, en las que no pueden ordenarse las diferentes categorías.
    — **Ordinales**, en las que pueden ordenarse las categorías, pero no puede establecerse la distancia relativa entre las mismas.
  - **Variables cuantitativas** son aquellas que tienen una interpretación numérica y que se subdividen en:
    — **Discretas**, sólo pueden tomar unos valores concretos dentro de un intervalo.
    — **Continuas**, pueden tomar cualquier valor dentro de un intervalo.

    En la práctica, todas las variables continuas que medimos son discretas en el sentido de que, debido a las limitaciones de los sistemas de medida, las variables continuas no pueden adoptar todos los valores dentro de un intervalo. De cara a los análisis posteriores, la principal distinción se establece, por tanto, entre variables con relativamente pocas categorías (como número de hijos) frente a variables con muchas categorías (como niveles de colesterol en sangre).

Ejemplo 1.2 Algunos ejemplos de variables son:

— Variables cualitativas nominales: sexo, raza, estado civil (soltero, casado, viudo, separado, divorciado), religión (católico, protestante, otros), nacionalidad.

— Variables cualitativas ordinales: salud auto-percibida (buena, regular, mala), severidad de la enfermedad (leve, moderada, grave). Por ejemplo, para esta última variable ordinal, podemos establecer un orden de severidad, pero no podemos decir que la diferencia de severidad entre un paciente moderado y uno leve sea la misma que entre uno grave y uno moderado.

— Variables cuantitativas discretas: número de hijos, número de dientes cariados.

— Variables cuantitativas continuas: edad, peso, altura, presión arterial, niveles de colesterol en sangre.

- **Estadístico** es cualquier operación realizada sobre los valores de una variable.
- **Parámetro** es un valor de la población sobre el que se desea realizar inferencias a partir de estadísticos obtenidos de la muestra, que en este caso se denominan **estimadores**. Por convención, los parámetros poblacionales se denotan con letras del alfabeto griego, mientras que los estimadores muestrales se denotan con letras de nuestro alfabeto.

Ejemplo 1.3 Algunos ejemplos de estadísticos incluyen:

— La media de los valores de colesterol de una muestra.

— El valor más alto de colesterol de una muestra.

— La suma de los valores de colesterol de una muestra elevados al cuadrado.

Así, por ejemplo, la media del colesterol en una población, que se denotaría por μ, es un parámetro que se estima a partir de la media de los valores de colesterol en una muestra obtenida de esa población, que se representaría por $\bar{x}$.

En el presente tema, se revisan las herramientas fundamentales para la realización de un análisis descriptivo de las variables recogidas en una muestra, tanto mediante estimadores de la tendencia central, posición y dispersión como mediante la utilización de representaciones gráficas.

1.2 MEDIDAS DE TENDENCIA CENTRAL

Las medidas de tendencia central informan acerca de cuál es el valor más representativo de una determinada variable o, dicho de forma equivalente, estos estimadores indican alrededor de qué valor se agrupan los datos observados. Las medidas de tendencia central de la muestra sirven tanto para resumir los resultados observados como para realizar inferencias acerca de los parámetros poblacionales correspondientes. A continuación se describen los principales estimadores de la tendencia central de una variable.

1.2.1 Media aritmética

La media aritmética, denotada por $\bar{x}$, se define como la suma de cada uno de los valores muestrales dividida por el número de observaciones realizadas. Si denotamos por n el tamaño muestral y por x_i el valor observado para el sujeto i-ésimo, $i = 1, ..., n$, la media vendría dada por

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + ... + x_n}{n}.$$

La media es la medida de tendencia central más utilizada y de más fácil interpretación. Corresponde al "centro de gravedad" de los datos de la muestra. Su principal limitación es que está muy influenciada por los valores extremos y, en este caso, puede no ser un fiel reflejo de la tendencia central de la distribución.

Ejemplo 1.4 En este y en los sucesivos ejemplos sobre estimadores muestrales, se utilizarán los valores del colesterol HDL obtenidos en los 10 primeros sujetos del estudio *"European Study on Antioxidants, Myocardial Infarction and Cancer of the Breast"* (EURAMIC), un estudio multicéntrico de casos y controles realizado entre 1991 y 1992 en ocho países Europeos e Israel para evaluar el efecto de los antioxidantes en el riesgo de desarrollar un primer infarto agudo de miocardio en hombres adultos. Los valores obtenidos fueron 0,89, 1,58, 0,79, 1,29, 1,42, 0,84, 1,06, 0,87, 1,96 y 1,53 mmol/l. La media de los niveles del colesterol HDL en estos 10 participantes es

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{0{,}89 + 1{,}58 + ... + 1{,}53}{10} = 1{,}223 \text{ mmol/l}.$$

La media aritmética presenta las siguientes propiedades:

- Cambio de origen (traslación). Si se suma una constante a cada uno de los datos de una muestra, la media de la muestra resultante es igual a la media inicial más la constante utilizada; si $y_i = x_i + c$, entonces $\bar{y} = \bar{x} + c$. Un cambio de origen que se realiza con frecuencia es el centrado de la variable, que consiste en restar a cada valor de la muestra su media. La media de una variable centrada será, por tanto, igual a 0.
- Cambio de escala (unidades). Si se multiplica cada uno de los datos de una muestra por una constante, la media de la muestra resultante es igual a la media inicial por la constante utilizada; si $y_i = cx_i$, entonces $\bar{y} = c\bar{x}$.
- Cambio simultáneo de origen y escala. Si se multiplica cada uno de los datos de una muestra por una constante y al resultado se le suma otra constante, la media de la muestra resultante es igual a la media inicial por la primera constante, más la segunda constante; si $y_i = c_1x_i + c_2$, entonces $\bar{y} = c_1\bar{x} + c_2$.

Ejemplo 1.5 Para transformar los valores del colesterol HDL de mmol/l a mg/dl se multiplica por el factor de conversión 38,8. Así, utilizando la propiedad del cambio de escala, la media del colesterol HDL en mg/dl se calcularía directamente a partir de su media en mmol/l como $1{,}223 \cdot 38{,}8 = 47{,}45$ mg/dl.

1.2.2 Mediana

La mediana es el valor de un variable que deja por encima el 50% de los datos de la muestra y por debajo el otro 50%. Para calcular la mediana, es necesario ordenar los valores de la muestra de menor a mayor. Si el tamaño muestral n es impar, la mediana viene dada por el valor $(n + 1)/2$-ésimo. Si n es par, la mediana viene dada por la media aritmética de los valores $(n/2)$ y $(n/2 + 1)$-ésimos. La principal ventaja de la mediana es que no está influenciada por los valores extremos. No obstante, se utiliza menos que la media como medida de tendencia central porque su tratamiento estadístico es más complejo.

Ejemplo 1.6 Para obtener la mediana del colesterol HDL en la muestra del estudio EURAMIC, se ordena en primer lugar los valores de menor a mayor; esto es, 0,79, 0,84, 0,87, 0,89, 1,06, 1,29, 1,42, 1,53, 1,58 y 1,96 mmol/l. Como el tamaño muestral es par ($n = 10$), la mediana será la media de los dos valores centrales (en este caso, el 5º y el 6º), que corresponde a $(1{,}06 + 1{,}29)/2 = 1{,}175$ mmol/l.

Comparación de la media aritmética y la mediana. En las distribuciones simétricas (ambas colas de la distribución son semejantes), la media es aproximadamente igual a la mediana. En distribuciones sesgadas positivamente (la cola superior de la distribución es mayor que la inferior), la media tiende a ser mayor que la mediana; mientras que en distribuciones sesgadas negativamente (la cola inferior de la distribución es mayor que la superior), la media tiende a ser menor que la mediana. La comparación de la media y la mediana permite evaluar, por tanto, la asimetría de una distribución.

Ejemplo 1.7 En la muestra del estudio EURAMIC la media del colesterol HDL es ligeramente superior a la mediana (1,223 y 1,175 mmol/l, respectivamente). En consecuencia, la distribución de estos 10 valores del colesterol HDL es aproximadamente simétrica con un leve sesgo positivo.

1.2.3 Media geométrica

La media geométrica, denotada por $\bar{x}_G$, se define como la raíz n-ésima del producto de los valores de una muestra de tamaño n,

$$\bar{x}_G = \left(\prod_{i=1}^{n} x_i \right)^{1/n} = \sqrt[n]{x_1 x_2 \cdot \ldots \cdot x_n} \ .$$

En la práctica, la forma más sencilla de calcular la media geométrica consiste en calcular primero el logaritmo de cada valor muestral, hallar a continuación la media de los logaritmos y deshacer finalmente la transformación logarítmica. Para calcular los logaritmos se puede usar cualquier base, siempre y cuando el logaritmo y el antilogaritmo estén en la misma base. Notar que la media geométrica sólo puede emplearse como medida de tendencia central en variables que toman valores positivos.

> ***Ejemplo 1.8*** Para calcular la media geométrica del colesterol HDL en la muestra del estudio EURAMIC, se halla primero el logaritmo natural de cada uno de los valores y a continuación se calcula su media aritmética,
>
> $$\log \bar{x}_G = \frac{1}{10} \sum_{i=1}^{10} \log x_i = \frac{\log(0{,}89) + \ldots + \log(1{,}53)}{10}$$
> $$= \frac{-0{,}117 + \ldots + 0{,}425}{10} = 0{,}155.$$
>
> La media geométrica es, por tanto, $\bar{x}_G = \exp(0{,}155) = 1{,}168$ mmol/l.

Al igual que la mediana, la media geométrica es útil como medida de tendencia central para variables muy asimétricas, en las que un pequeño grupo de observaciones extremas tienen una excesiva influencia sobre la media aritmética. La media geométrica tiene la ventaja adicional de presentar un tratamiento estadístico más sencillo que la mediana.

1.3 MEDIDAS DE POSICIÓN: CUANTILES

Los cuantiles indican la posición relativa de una observación con respecto al resto de la muestra. A continuación se describen los cuantiles más utilizados:

- **Percentiles** son los valores de una variable que dejan un determinado porcentaje de los datos por debajo de ellos. Así, por ejemplo, el percentil 10 es el valor superior al 10% de las observaciones, pero inferior al 90% restante. La mediana corresponde, por tanto, al percentil 50. En una muestra de tamaño n, previamente ordenada de menor a mayor, el percentil p-ésimo se define como:
  - Si $np/100$ es un número entero, la media de las observaciones $(np/100)$ y $(np/100+1)$-ésimas.
  - Si $np/100$ no es un número entero, el valor k-ésimo de la muestra, siendo k el menor entero superior a $np/100$.
- **Deciles**, corresponden a los percentiles 10, 20, ..., 90. Los deciles se utilizan para dividir la muestra en 10 grupos de igual tamaño.
- **Quintiles**, corresponden a los percentiles 20, 40, 60 y 80, y dividen la muestra en 5 grupos de igual tamaño.

- **Cuartiles**, corresponden a los percentiles 25, 50 y 75, y dividen la muestra en 4 grupos de igual tamaño.
- **Terciles**, corresponden a los percentiles 33,3 y 66,7, y dividen la muestra en 3 grupos de igual tamaño.

Ejemplo 1.9 Los 10 valores del colesterol HDL ordenados de menor a mayor son 0,79, 0,84, 0,87, 0,89, 1,06, 1,29, 1,42, 1,53, 1,58 y 1,96 mmol/l. Dado que $10p/100 = 1$ es un número entero para $p = 10$, el percentil 10 es la media de la primera y segunda observación, que corresponde a $(0{,}79 + 0{,}84)/2 = 0{,}815$ mmol/l. De igual forma, como $10p/100 = 2{,}5$ no es un número entero para $p = 25$, el percentil 25 es el tercer valor de la muestra, que corresponde a 0,87 mmol/l.

Es importante recordar que, para calcular cuantiles, los valores de la muestra deben estar previamente ordenados. Si el tamaño muestral es grande, la forma más rápida de obtener los cuantiles manualmente es realizando un gráfico de tallo y hojas (ver más adelante).

1.4 MEDIDAS DE DISPERSIÓN

Las medidas de dispersión indican el grado de variabilidad de los datos y se complementan con las medidas de tendencia central en la descripción de una muestra. En este apartado se presentan las principales medidas de dispersión.

1.4.1 Varianza y desviación típica

La varianza muestral, denotada por s^2, se define como la suma de los cuadrados de las diferencias entre cada valor de la muestra y su media, dividida por el tamaño muestral menos 1,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right).$$

Como puede apreciarse, cuanto más dispersos estén los datos, mayores serán los cuadrados de las desviaciones $(x_i - \bar{x})^2$ y mayor será la varianza s^2. Notar que las desviaciones de cada valor respecto de la media se elevan al cuadrado para evitar que se compensen las desviaciones positivas (valores superiores a la media) con las negativas (valores inferiores a la media). Cabe destacar también que, en la fórmula de la varianza muestral, el denominador es $n - 1$ en lugar de n. Esto se debe a que, una vez calculada la media, el número de valores independientes de la muestra (denominado "grados de libertad") para el cálculo de la varianza es $n - 1$ (conocida la media y $n - 1$ valores, el valor restante se deduciría automáticamente). Una justificación más formal para esta definición de la varianza se aporta en el Tema 5.

La varianza muestral es difícil de interpretar como medida de dispersión, ya que sus unidades son las de la variable original al cuadrado. La medida de dispersión más utilizada es la desviación típica o desviación estándar s, que se define como la raíz cuadrada de la varianza

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

y, en consecuencia, presenta las mismas unidades que la variable original. Al igual que la media, la desviación típica está influenciada por valores muy extremos (gran desviación respecto de la

media), que inflarían la estimación resultante, no siendo un buen reflejo de la dispersión global de los datos.

Ejemplo 1.10 Conocida la media del colesterol HDL en los 10 primeros participantes del estudio EURAMIC, $\bar{x}$ = 1,223 mmol/l, la varianza vendría dada por

$$s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2 = \frac{(0{,}89 - 1{,}223)^2 + \ldots + (1{,}53 - 1{,}223)^2}{9}$$
$$= \frac{0{,}111 + \ldots + 0{,}094}{9} = 0{,}156\ (\text{mmol/l})^2$$

y la desviación típica por $s = \sqrt{0{,}156} = 0{,}395$ mmol/l.

Algunas propiedades de la varianza y la desviación típica son:

- Cambio de origen (traslación). Si se suma una constante a cada uno de los datos de una muestra, la varianza y la desviación típica no cambian; si $y_i = x_i + c$, entonces $s_y^2 = s_x^2$ y $s_y = s_x$.
- Cambio de escala (unidades). Si se multiplica cada uno de los datos de una muestra por una constante, la varianza resultante es igual a la varianza inicial por la constante al cuadrado y la desviación típica es igual a la desviación típica inicial por dicha constante; si $y_i = cx_i$, entonces $s_y^2 = c^2 s_x^2$ y $s_y = cs_x$. Un cambio de escala que se realiza con frecuencia es la división de todos los valores de una muestra por su desviación típica. La desviación típica de la variable resultante será, por tanto, igual a 1.

Las propiedades del cambio de origen y escala se emplean para la estandarización de variables, que consiste en restarle a los valores de una variable su media y dividirlos por su desviación típica. La variable estandarizada resultante tiene media 0 y desviación típica 1; es decir, si $z_i = (x_i - \bar{x})/s_x$, entonces $\bar{z} = 0$ y $s_z = 1$.

1.4.2 Rango intercuartílico

El rango intercuartílico se define como la diferencia entre el tercer y el primer cuartil (percentiles 75 y 25, respectivamente). El rango intercuartílico indica la amplitud del 50% central de la muestra y se usa como medida de dispersión cuando la variable presenta valores extremos. En tal caso, suele ir acompañado de la mediana como medida de tendencia central.

Ejemplo 1.11 A partir de los 10 valores del colesterol HDL ordenados de menor a mayor, los percentiles 25 y 75 vienen determinados por la tercera (0,87 mmol/l) y octava observación (1,53 mmol/l), respectivamente. El rango intercuartílico se calcula entonces como la diferencia entre ambos percentiles, 1,53 – 0,87 = 0,66 mmol/l.

1.4.3 Coeficiente de variación

El coeficiente de variación se define como el cociente entre la desviación típica y la media aritmética, expresado como porcentaje, $100s/\bar{x}$. Este estimador no está afectado por cambios de escala ya que, al multiplicar los valores de una variable por un mismo factor, tanto la media como la desviación típica cambian por dicho factor y su cociente permanece inalterable. El coeficiente de variación relaciona la desviación típica con la media y es útil para comparar la variabilidad de diferentes variables con distintas medias. Así, por ejemplo, una desviación típica de 10 kg en una muestra de adultos con un peso medio de 70 kg indicaría un mismo grado de dispersión que una desviación

típica de 0,5 kg en una muestra de recién nacidos con un peso medio de 3,5 kg (ambos coeficientes de variación son 100·10/70 = 100·0,5/3,5 = 14,3%).

Ejemplo 1.12 El coeficiente de variación de los 10 primeros valores del colesterol HDL en el estudio EURAMIC sería $100s/\bar{x}$ = 100·0,395/1,223 = 32,3%; es decir, la desviación típica es aproximadamente un tercio de la media.

1.5 REPRESENTACIONES GRÁFICAS

En el análisis e interpretación de los datos de un estudio, es importante no limitarse a realizar medidas de resumen numéricas. Las medidas de tendencia central y dispersión deben completarse con gráficos que permitan observar directamente las características y relaciones de las variables estudiadas. En esta sección se revisan los principales métodos gráficos para presentar y resumir una variable.

1.5.1 Diagrama de barras

Los diagramas de barras son adecuados para representar variables cualitativas y cuantitativas discretas. En estos diagramas se representan las categorías de la variable en el eje horizontal y sus frecuencias (absolutas o relativas) en el eje vertical. Para cada categoría de la variable se construye un rectángulo de anchura constante y altura proporcional a la frecuencia. Los rectángulos están separados unos de otros por la misma distancia para reflejar la discontinuidad de la variable.

Ejemplo 1.13 La representación del diagrama de barras del hábito tabáquico en el grupo control del estudio EURAMIC se ilustra en la Figura 1.1. De los 700 controles del estudio que no habían padecido un infarto agudo de miocardio, todos salvo uno presentaban información sobre el consumo de tabaco. De éstos, un 27,2% (190/699) eran nunca fumadores, un 35,3% (247/699) eran ex fumadores, y el restante 37,5% (262/699) eran fumadores actuales.

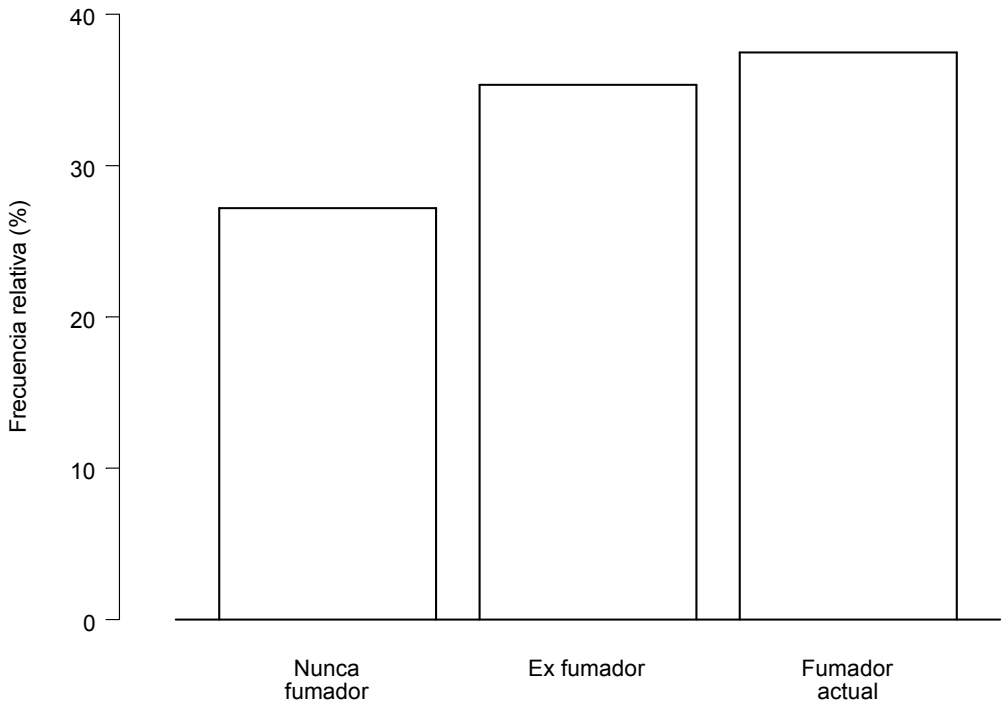


Figura 1.1 Diagrama de barras del hábito tabáquico en el grupo control del estudio EURAMIC.

1.5.2 Histograma y polígono de frecuencias

El histograma es el principal método gráfico para la representación de variables cuantitativas continuas. En primer lugar, los valores de la variable continua se agrupan en categorías exhaustivas (cubren todo el rango de la variable) y mutuamente excluyentes (no se solapan). En el eje horizontal del histograma se representan las categorías o intervalos y en el eje vertical las frecuencias (absolutas o relativas) de cada intervalo. Posteriormente, se construye un rectángulo para cada categoría, cuya anchura es igual a la longitud del intervalo y cuyo área es proporcional a la frecuencia (si los intervalos tienen distinta longitud, las alturas de los rectángulos del histograma no serán proporcionales a las frecuencias).

El polígono de frecuencias se construye uniendo con líneas rectas los puntos medios de las bases superiores de los rectángulos que conforman un histograma. Tanto el histograma como el polígono de frecuencias sirven para representar gráficamente la distribución de una variable continua.

Ejemplo 1.14 El histograma de la distribución del colesterol HDL en el grupo control del estudio EURAMIC se presenta en la Figura 1.2. En este caso, se representa la frecuencia absoluta en el eje vertical e intervalos de distinta longitud en el eje horizontal. Para los intervalos de menor longitud (0,2 mmol/l), la altura de los rectángulos es igual a la frecuencia; así, por ejemplo, la altura del rectángulo en el intervalo 1,2-1,4 mmol/l es igual a los 86 sujetos con niveles del colesterol HDL dentro de este rango. Sin embargo, para los intervalos de mayor longitud, la altura de la barra es igual a la frecuencia dividida por el incremento relativo de la longitud del intervalo; así, por ejemplo, para el intervalo 1,4-1,7 mmol/l, cuya frecuencia es 55 y su longitud es 1,5 veces la longitud mínima, la altura de la barra es 55/1,5 = 36,7. La Figura 1.2 se completa con el polígono de frecuencias, que muestra una distribución del colesterol HDL aproximadamente simétrica con la cola superior ligeramente mayor que la inferior.

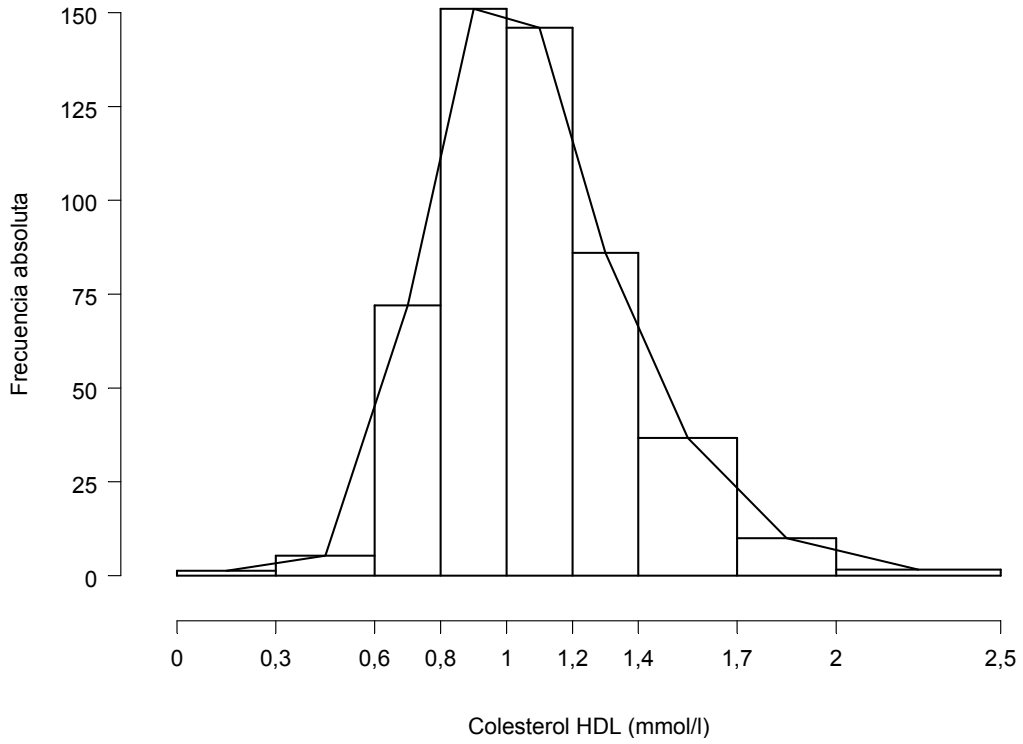


Figura 1.2 Histograma y polígono de frecuencias del colesterol HDL en el grupo control del estudio EURAMIC.

1.5.3 Gráfico de tallo y hojas

Este gráfico tiene la ventaja de reflejar los datos originales de la muestra, a la vez que permite visualizar la distribución de frecuencias. En primer lugar, para cada observación de la variable, se separa el último dígito significativo (hoja) de los restantes dígitos del valor de la variable (tallo). A continuación, todos los posibles tallos se colocan ordenados en una misma columna. Finalmente, para cada valor de la variable, se coloca su hoja a la derecha del tallo correspondiente. Las hojas de un mismo tallo suelen colocarse en orden creciente. El resultado se conoce con el nombre de gráfico de tallo y hojas.

Ejemplo 1.15 La Figura 1.3 muestra el gráfico de tallo y hojas del colesterol HDL en los 100 primeros controles del estudio EURAMIC con datos para esta variable. Los 2 valores más bajos del colesterol HDL son 0,21 y 0,26 mmol/l, cuyo tallo común es 0,2 y sus respectivas hojas son 1 y 6, que aparecen a la derecha de la primera línea del gráfico. El siguiente tallo es 0,3, que no tiene ninguna hoja ya que no hay valores entre 0,30 y 0,39 mmol/l, y lo mismo sucede con el tallo 0,4. En el tallo 0,5 hay una hoja igual a 7, que corresponde al valor 0,57 mmol/l. En el tallo 0,6 hay 5 hojas (35558), que corresponden a los 5 valores del colesterol HDL entre 0,60 y 0,69 mmol/l y que son 0,63, 0,65, 0,65, 0,65 y 0,68 mmol/l. El resto de los tallos se interpreta de la misma manera. A partir de este gráfico resulta sencillo calcular los cuantiles; así, por ejemplo, la mediana se obtendría como la media de los valores ordenados en las posiciones 50 y 51, (1,10 + 1,12)/2 = 1,11 mmol/l.

Frecuencia	Tallo	Hoja
2	0,2	16
0	0,3	
0	0,4	
1	0,5	7
5	0,6	35558
3	0,7	467
12	0,8	002344455579
13	0,9	0013334566779
13	1,0	0111123455559
9	1,1	023456789
15	1,2	000023356689999
7	1,3	1223778
6	1,4	345789
6	1,5	133689
2	1,6	44
2	1,7	34
2	1,8	36
1	1,9	0
1	2,0	9

Figura 1.3 Gráfico de tallo y hojas del colesterol HDL en los 100 primeros controles del estudio EURAMIC.

1.5.4 Diagrama de caja

El diagrama de caja permite evaluar la tendencia central, la dispersión y la simetría de la distribución de una variable, así como identificar valores extremos. Los límites inferior y superior de la caja corresponden a los percentiles 25 y 75; es decir, la altura de la caja representa el rango intercuartílico e indica la dispersión de la muestra. La línea horizontal dentro de la caja corresponde a la mediana y representa la tendencia central de la muestra. El gráfico se completa con barras verticales a ambos lados de la caja de longitud 1,5 veces el rango intercuartílico. Los valores extremos, aquellos distanciados de los límites de la caja entre 1,5 y 3 veces el rango intercuartílico, se representan con un círculo y los valores muy extremos, aquellos alejados de la caja más de 3 veces el rango intercuartílico, se denotan mediante un asterisco.

En este gráfico, si la distribución es simétrica, los límites superior e inferior de la caja estarán aproximadamente a la misma distancia de la mediana, mientras que si la distribución está sesgada positivamente, el límite superior estará más alejado de la mediana que el inferior y si la distribución está sesgada negativamente, el límite inferior estará más alejado de la mediana que el superior.

Ejemplo 1.16 La Figura 1.4 muestra el diagrama de caja del colesterol HDL en el grupo control del estudio EURAMIC. Como puede observarse, esta distribución presenta un leve sesgo positivo ya que el límite superior de la caja está ligeramente más alejado de la mediana que el límite inferior.

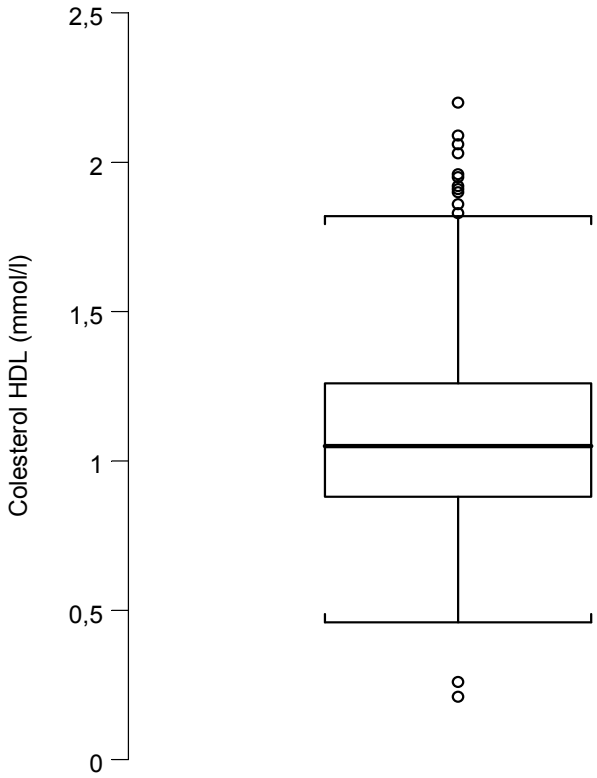


Figura 1.4 **Diagrama de caja del colesterol HDL en el grupo control del estudio EURAMIC.**

1.6 REFERENCIAS

1. Colton T. *Estadística en Medicina*. Barcelona: Salvat, 1979.
2. Glantz SA. *Primer of Biostatistics, Fifth Edition*. New York: McGraw-Hill/Appleton & Lange, 2001.
3. Pagano M, Gauvreau K. *Principles of Biostatistics, Second Edition*. Belmont, CA: Duxbury Press, 2000.
4. Rosner B. *Fundamentals of Biostatistics, Sixth Edition*. Belmont, CA: Duxbury Press, 2006.

TEMA 2

PROBABILIDAD

2.1 INTRODUCCIÓN

Se denominan **experimentos estocásticos, aleatorios o no determinísticos** a aquellos en los que pueden obtenerse resultados distintos cuando se repiten en idénticas circunstancias. Los fenómenos biológicos tienen en este sentido una componente aleatoria importante. La herramienta matemática que constituye la base para el estudio de fenómenos con una componente aleatoria es la teoría de la **probabilidad**, que proporciona modelos teóricos aplicables a la frecuencia de los distintos resultados de un experimento.

A continuación, se revisan algunos conceptos previos que van a ser necesarios para sistematizar la noción de probabilidad.

- **Espacio muestral**, denotado por Ω, es el conjunto de los posibles resultados de un experimento aleatorio.
- Se denomina **suceso** a cualquier subconjunto del espacio muestral Ω. Los sucesos pueden ser elementos simples de Ω o conjuntos de elementos. Dos sucesos particulares son el suceso seguro Ω, que contiene todos los elementos del espacio muestral, y el suceso imposible o conjunto vacío $\varnothing$, que no contiene ningún elemento.

  Ejemplo 2.1 Si el experimento consiste en observar el número de supervivientes a los 6 meses de 4 pacientes con cáncer sometidos a tratamiento, el espacio muestral será $\Omega = \{0, 1, 2, 3, 4\}$. Si el experimento consiste en medir los niveles de colesterol HDL de una persona, el espacio muestral será $\Omega = (0, \infty)$.

  En el primer experimento, algunos sucesos podrían ser: no observar ningún superviviente $A = \{0\}$, observar 1 ó 2 supervivientes $B = \{1, 2\}$ u observar al menos 2 supervivientes $C = \{2, 3, 4\}$. En el segundo experimento, algunos de los posibles sucesos incluirían: tener un colesterol HDL $\leq$ 1 mmol/l $A = (0, 1]$ o tener un colesterol HDL > 1,5 mmol/l $B = (1{,}5, \infty)$.

- El **suceso unión** $A \cup B$ es el evento constituido por los elementos que pertenecen a A o B, o a ambos a la vez.
- El **suceso intersección** $A \cap B$ es el evento formado por los elementos que pertenecen simultáneamente a A y B.
- **Sucesos disjuntos, incompatibles o mutuamente excluyentes** son aquellos que no pueden ocurrir simultáneamente; es decir, su intersección es el conjunto vacío, $A \cap B = \varnothing$.
- El **suceso complementario** del suceso A, denotado por A^c, es el evento que ocurre cuando no se realiza A.

Estos sucesos están representados en los diagramas de la Figura 2.1. En general, las operaciones entre sucesos se rigen por la teoría de conjuntos, de la cual pueden derivarse algunas propiedades importantes como $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$, $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $(A \cup B)^c = A^c \cap B^c$ y $(A \cap B)^c = A^c \cup B^c$.

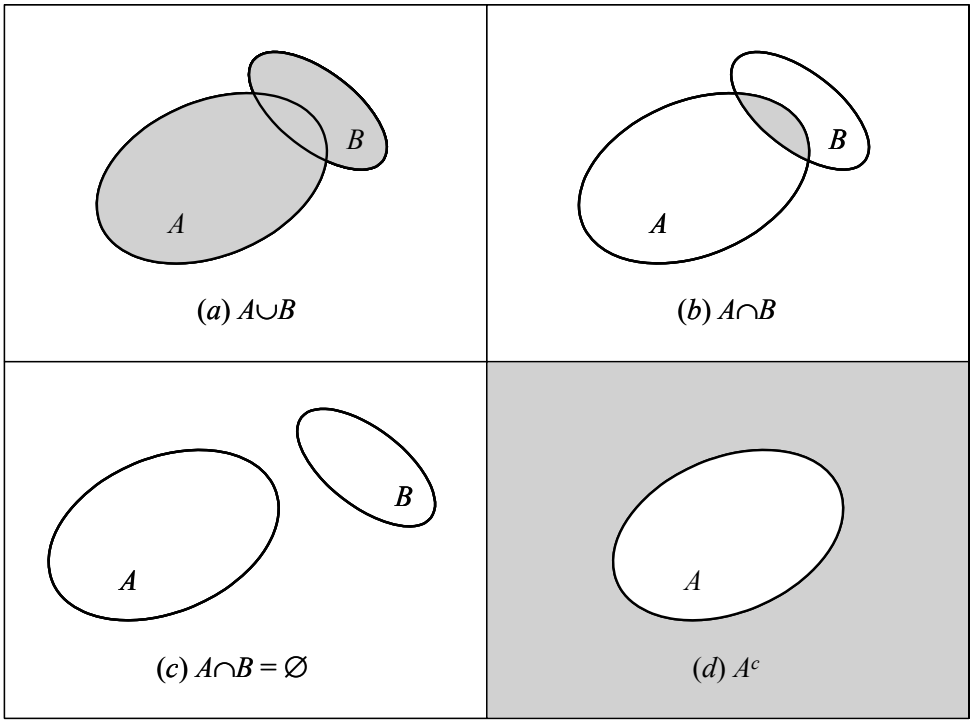


Figura 2.1 Diagramas de los sucesos unión (*a*), intersección (*b*), sucesos mutuamente excluyentes (*c*) y suceso complementario (*d*).

Ejemplo 2.2 En el experimento de supervivencia a los 6 meses de 4 pacientes con cáncer, la unión de los sucesos $B = \{1, 2\}$ y $C = \{2, 3, 4\}$ es $B \cup C = \{1, 2, 3, 4\}$ y su intersección es $B \cap C = \{2\}$. Al medir los niveles de colesterol HDL de una persona, los sucesos $A = (0, 1]$ y $B = (1{,}5, \infty)$ son mutuamente excluyentes ya que $A \cap B = \varnothing$. Asimismo, en este experimento el complementario de A es el suceso $A^c = (1, \infty)$.

En este tema se define el concepto de probabilidad y se introducen las reglas básicas para operar con probabilidades. Estas reglas constituyen la base para el cálculo e interpretación de los procedimientos de inferencia estadística (por ejemplo, el valor P de un contraste de hipótesis –véase Tema 5–) y permiten también evaluar la sensibilidad, la especificidad y los valores predictivos de las pruebas diagnósticas.

2.2 CONCEPTO Y DEFINICIONES DE PROBABILIDAD

El concepto de probabilidad es intuitivo, tal y como se refleja en el lenguaje cotidiano: la probabilidad de un suceso refleja la verosimilitud de que éste ocurra, de forma que los sucesos más probables se darán con mayor frecuencia que los menos probables. Sin embargo, para abordar la probabilidad de forma sistemática, es necesaria una definición rigurosa, a la vez que compatible con nuestra intuición. Dos definiciones de probabilidad de uso común son:

- **Definición frecuentista** (von Mises). Al repetir un experimento indefinidamente, la probabilidad de un suceso es el límite del cociente entre el número de veces que ocurre dicho suceso y el número de experimentos realizados,

$$P(A) = \lim_{n \to \infty} \frac{\#A}{n},$$

donde $\#A$ es el número de veces que se realiza A en los n experimentos.

Ejemplo 2.3 Supongamos que se desea conocer la probabilidad de ser mujer entre todos los recién nacidos vivos en España. Según los datos del Instituto Nacional de Estadística, se registraron 226.170 niñas de 466.371 nacimientos en 2005, 233.773 de 482.957 en 2006 y 238.632 de 492.527 en 2007. La proporción acumulada de niñas es 226.170/466.371 = 0,4850 en 2005, 459.943/949.328 = 0,4845 en 2005-2006 y 698.575/1.441.855 = 0,4845 en 2005-2007. Aumentando indefinidamente los registros anuales, el límite de estos cocientes 0,4850, 0,4845, 0,4845, ... determinaría la probabilidad de ser mujer. En la práctica, sin embargo, no es posible realizar infinitos experimentos y las probabilidades teóricas se estiman mediante probabilidades empíricas obtenidas a partir de un número finito de experimentos. Así, utilizando los datos disponibles de nacimientos en 2005-2007, se estimaría una probabilidad de ser mujer de 0,4845.

- **Definición axiomática** (Kolmogorov). La probabilidad es una función que asigna a cada posible suceso de un experimento un valor numérico, de tal forma que se cumplan los siguientes axiomas:

  (i) No negatividad: $P(A) \geq 0$,

  (ii) Normatividad: $P(\Omega) = 1$,

  (iii) Aditividad: Si $A_1, A_2, \ldots$ son sucesos mutuamente excluyentes, entonces

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots = \sum_{i=1}^{\infty} P(A_i).$$

  Notar que esta definición de probabilidad tan sólo especifica las propiedades generales que debe tener una función de probabilidad, pero no permite la asignación de probabilidades a un suceso concreto. No obstante, de la definición axiomática se derivan algunas propiedades importantes de la función de probabilidad:

  (iv) $P(\varnothing) = 0$,

  (v) $P(A^c) = 1 - P(A)$,

  (vi) Si A está incluido en B, $A \subset B$, entonces $P(A) \leq P(B)$,

  (vii) $0 \leq P(A) \leq 1$,

  (viii) Sub-aditividad: Para cualquier colección de sucesos $A_1, A_2, \ldots$,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i),$$

  (ix) Principio de inclusión-exclusión: Sean $A_1, A_2, \ldots, A_k$ sucesos cualesquiera,

$$P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i) - \sum_{1 \leq i < j \leq k} P(A_i \cap A_j) + \ldots + (-1)^{k+1} P(A_1 \cap A_2 \cap \ldots \cap A_k).$$

Del tercer axioma de la probabilidad se deduce que, si dos sucesos son mutuamente excluyentes, la probabilidad de la unión es la suma de sus probabilidades por separado. El principio de inclusión-exclusión generaliza este resultado para sucesos no necesariamente

excluyentes: la probabilidad de la unión de dos sucesos cualesquiera es la suma de sus probabilidades por separado, menos la probabilidad de la intersección,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Este principio puede aplicarse a colecciones con más de dos sucesos. Así, por ejemplo, para tres sucesos cualesquiera, se cumple que

$$\begin{aligned} P(A \cup B \cup C) = {} & P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ & + P(A \cap B \cap C). \end{aligned}$$

Ejemplo 2.4 Supongamos que la probabilidad de ser bebedor en una determinada población de adultos es 0,20, la probabilidad de ser diabético es 0,03 y la probabilidad de ser simultáneamente bebedor y diabético es 0,01. Si se denota por B al suceso ser bebedor y por D al suceso ser diabético, la probabilidad de que un individuo de esta población sea bebedor, diabético o ambos a la vez viene determinada por

$$P(B \cup D) = P(B) + P(D) - P(B \cap D) = 0{,}20 + 0{,}03 - 0{,}01 = 0{,}22.$$

2.3 PROBABILIDAD CONDICIONAL E INDEPENDENCIA DE SUCESOS

La probabilidad de un suceso puede depender de la realización de otro suceso. Así, por ejemplo, la probabilidad de tener un infarto de miocardio es diferente en los hombres que en las mujeres; es decir, la probabilidad del suceso tener un infarto de miocardio depende del suceso ser hombre o ser mujer. El concepto matemático que permite formalizar cómo se modifica la probabilidad de un suceso en función de otro es la probabilidad condicional. En general, la probabilidad del suceso B condicionada al suceso A se define como

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

De forma intuitiva, condicionar por el suceso A es equivalente a seleccionar por este suceso. Así, P(infarto|hombre) es equivalente a seleccionar en primer lugar a los hombres y posteriormente determinar su probabilidad de tener un infarto de miocardio.

El concepto de probabilidad condicional tiene numerosas aplicaciones en epidemiología y salud pública. Por ejemplo, si D es el suceso tener una enfermedad y E es el suceso estar expuesto a un factor de riesgo, $P(D|E)$ es la probabilidad de la enfermedad entre los expuestos, $P(D|E^c)$ es la probabilidad de la enfermedad entre los no expuestos y $\psi = P(D|E)/P(D|E^c)$ es el riesgo relativo de la enfermedad entre los expuestos y los no expuestos.

Ejemplo 2.5 Continuando con el ejemplo anterior, la probabilidad de que un bebedor sea diabético se calcula como

$$P(D|B) = \frac{P(B \cap D)}{P(B)} = \frac{0{,}01}{0{,}20} = 0{,}05$$

y la probabilidad de que un no bebedor sea diabético como

$$P(D|B^c) = \frac{P(B^c \cap D)}{P(B^c)} = \frac{P(D) - P(B \cap D)}{1 - P(B)} = \frac{0{,}03 - 0{,}01}{1 - 0{,}20} = 0{,}025.$$

Así, el riesgo de diabetes es el doble en los sujetos bebedores que en los no bebedores, $\psi = P(D|B)/P(D|B^c) = 0{,}05/0{,}025 = 2$.

Se dice que dos **sucesos** son **independientes** si la ocurrencia de uno no afecta a la probabilidad del otro; es decir, A y B son independientes si $P(B|A) = P(B|A^c) = P(B)$ o, de forma equivalente, si $P(A|B) = P(A|B^c) = P(A)$. En consecuencia, si dos sucesos son independientes, puede probarse que

$$P(A \cap B) = P(A)P(B|A) = P(A)P(B).$$

Por tanto, dos sucesos también pueden definirse como independientes si la probabilidad de su intersección es igual al producto de la probabilidad de cada suceso por separado.

Ejemplo 2.6 A partir de los resultados del ejemplo anterior, puede concluirse que los sucesos padecer diabetes y ser bebedor no son independientes dado que la probabilidad de ser diabético es diferente en bebedores que en no bebedores,

$$P(D|B) = 0{,}05 \neq 0{,}025 = P(D|B^c);$$

es decir, el riesgo relativo es distinto de la unidad, $\psi = 2 \neq 1$. Esta dependencia se refleja también en el hecho de que la probabilidad de ser simultáneamente bebedor y diabético no es el producto de sus probabilidades,

$$P(B \cap D) = 0{,}01 \neq 0{,}20 \cdot 0{,}03 = P(B)P(D).$$

Notar que la probabilidad de la intersección de dos sucesos cualesquiera

$$P(A \cap B) = P(A)P(B|A)$$

no equivale al producto de sus probabilidades, salvo que ambos sucesos sean independientes. En general, para cualquier conjunto de sucesos $A_1, A_2, ..., A_k$, la probabilidad de su intersección es

$$\begin{aligned} P(A_1 \cap A_2 \cap ... \cap A_k) &= P(A_1)P(A_2 \cap ... \cap A_k|A_1) \\ &= P(A_1)P(A_2|A_1)P(A_3 \cap ... \cap A_k|A_1 \cap A_2) = ... \\ &= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdot \ldots \cdot P(A_k|A_1 \cap A_2 \cap ... \cap A_{k-1}). \end{aligned}$$

En el caso de que estos sucesos sean mutuamente independientes, las probabilidades condicionales de la fórmula anterior se reducen a probabilidades no condicionales y, en consecuencia, la probabilidad de la intersección es igual al producto de sus probabilidades,

$$P\left(\bigcap_{i=1}^{k} A_i\right) = P(A_1 \cap A_2 \cap ... \cap A_k) = P(A_1)P(A_2) \cdot \ldots \cdot P(A_k) = \prod_{i=1}^{k} P(A_i).$$

2.4 REGLA DE LA PROBABILIDAD TOTAL

La probabilidad no condicional de un suceso B se relaciona con su probabilidad condicionada en la ocurrencia o no de otro suceso A mediante la fórmula

$$P(B) = P(A \cap B) + P(A^c \cap B) = P(A)P(B|A) + P(A^c)P(B|A^c).$$

Así, la probabilidad no condicional de B es la media ponderada de las probabilidades condicionales de B dado A y A^c. Esta descomposición de la probabilidad del suceso B en términos de A y A^c es aplicable porque estos sucesos constituyen una partición del espacio muestral; es decir, A y A^c son sucesos exhaustivos $A \cup A^c = \Omega$ y mutuamente excluyentes $A \cap A^c = \varnothing$.

En general, para un conjunto de sucesos $A_1, A_2, ..., A_k$ globalmente exhaustivos y mutuamente excluyentes que formen una partición del espacio muestral, se verifica que

$$P(B) = \sum_{i=1}^{k} P(A_i \cap B) = \sum_{i=1}^{k} P(A_i)P(B \mid A_i),$$

conocida como regla de la probabilidad total. Esta fórmula es particularmente útil en epidemiología, donde se emplean con frecuencia las particiones. Por ejemplo, al dividir la población en grupos de edad y sexo se están empleando categorías globalmente exhaustivas y mutuamente excluyentes. En general, siempre que se divide la población en estratos se aplica una partición a esa población.

Ejemplo 2.7 En una población de mayores de 65 años, los individuos con edades entre 65-74, 75-84 y ≥ 85 años constituyen el 60, 30 y 10% de la población. La prevalencia de la enfermedad de Alzheimer en estos grupos de edad es respectivamente de 20, 75 y 300 casos por 1000. La prevalencia global de la enfermedad de Alzheimer en esta población de mayores de 65 años se calcularía

$$\begin{aligned} P(A) &= \sum_{i=1}^{3} P(E_i)P(A \mid E_i) \\ &= 0{,}60 \cdot 0{,}020 + 0{,}30 \cdot 0{,}075 + 0{,}10 \cdot 0{,}300 = 0{,}0645, \end{aligned}$$

resultando 64,5 casos por 1000 personas.

2.5 TEOREMA DE BAYES

El teorema de Bayes permite obtener la probabilidad condicional de A dado B a partir de la probabilidad de A y de las probabilidades condicionales inversas de B dado A y A^c. Aplicando la definición de probabilidad condicional y la regla de la probabilidad total, se obtiene que

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(A^c)P(B \mid A^c)}.$$

El teorema de Bayes se usa con frecuencia en la evaluación de pruebas diagnósticas. Cuando se desarrolla una prueba diagnóstica y se comparan sus resultados con los de un patrón oro (método de referencia en el diagnóstico de la enfermedad), suelen determinarse los siguientes parámetros o características propias de la prueba diagnóstica:

- **Sensibilidad** es la probabilidad de obtener un resultado positivo de la prueba diagnóstica entre los sujetos realmente enfermos, $S = P(+|D)$.
- **Especificidad** es la probabilidad de obtener un resultado negativo entre los sujetos realmente sanos, $E = P(-|D^c)$.

En la aplicación clínica de una prueba diagnóstica a una determinada población interesa conocer, sin embargo, los siguientes parámetros:

- **Valor predictivo positivo** es la probabilidad de tener la enfermedad entre las personas que tienen un resultado positivo, $VP+ = P(D|+)$.
- **Valor predictivo negativo** es la probabilidad de no tener la enfermedad entre las personas que tienen un resultado negativo, $VP- = P(D^c|-)$.

Aplicando el teorema de Bayes, pueden calcularse los valores predictivos en función de la prevalencia de la enfermedad en la población y de la sensibilidad y especificidad de la prueba diagnóstica,

$$VP+ = P(D\,|\,+) = \frac{P(D)P(+\,|\,D)}{P(D)P(+\,|\,D) + P(D^c)P(+\,|\,D^c)} = \frac{PS}{PS + (1-P)(1-E)},$$

$$VP- = P(D^c\,|\,-) = \frac{P(D^c)P(-\,|\,D^c)}{P(D)P(-\,|\,D) + P(D^c)P(-\,|\,D^c)} = \frac{(1-P)E}{P(1-S) + (1-P)E}.$$

Ejemplo 2.8 La sensibilidad de la prueba ELISA para detectar seropositividad frente al virus de inmunodeficiencia humana es del 99% y su especificidad es del 96%. En una población con una prevalencia de infección por el virus de inmunodeficiencia humana del 0,3%, únicamente el 6,9% de las personas con un resultado positivo del test ELISA estarán realmente infectadas,

$$VP+ = \frac{PS}{PS + (1-P)(1-E)} = \frac{0{,}003 \cdot 0{,}99}{0{,}003 \cdot 0{,}99 + 0{,}997 \cdot 0{,}04} = 0{,}069,$$

mientras que prácticamente todas las personas con resultado negativo estarán libres de la infección,

$$VP- = \frac{(1-P)E}{P(1-S) + (1-P)E} = \frac{0{,}997 \cdot 0{,}96}{0{,}003 \cdot 0{,}01 + 0{,}997 \cdot 0{,}96} = 1{,}000.$$

Sin embargo, en una población de alto riesgo con una prevalencia del virus de inmunodeficiencia humana del 10%, el 73,3% de los sujetos con resultado positivo estarán realmente infectados,

$$VP+ = \frac{PS}{PS + (1-P)(1-E)} = \frac{0{,}10 \cdot 0{,}99}{0{,}10 \cdot 0{,}99 + 0{,}90 \cdot 0{,}04} = 0{,}733,$$

siendo muy improbable la infección entre aquellos sujetos con resultado negativo,

$$VP- = \frac{(1-P)E}{P(1-S) + (1-P)E} = \frac{0{,}90 \cdot 0{,}96}{0{,}10 \cdot 0{,}01 + 0{,}90 \cdot 0{,}96} = 0{,}999.$$

Como puede apreciarse, el valor predictivo positivo de esta prueba varía enormemente en función de la prevalencia poblacional de la infección.

En general, si $A_1, A_2, ..., A_k$ son sucesos globalmente exhaustivos y mutuamente excluyentes, el teorema de Bayes puede generalizarse como

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B \mid A_i)}{\sum_{j=1}^{k} P(A_j)P(B \mid A_j)}.$$

Ejemplo 2.9 Continuando con el Ejemplo 2.7, la distribución de los casos de la enfermedad de Alzheimer por grupo de edad viene dada por

$$P(E_1|A) = \frac{P(E_1)P(A \mid E_1)}{\sum_{i=1}^{3} P(E_i)P(A \mid E_i)} = \frac{0{,}60 \cdot 0{,}020}{0{,}0645} = 0{,}186,$$

$$P(E_2|A) = \frac{P(E_2)P(A \mid E_2)}{\sum_{i=1}^{3} P(E_i)P(A \mid E_i)} = \frac{0{,}30 \cdot 0{,}075}{0{,}0645} = 0{,}349,$$

$$P(E_3|A) = \frac{P(E_3)P(A \mid E_3)}{\sum_{i=1}^{3} P(E_i)P(A \mid E_i)} = \frac{0{,}10 \cdot 0{,}300}{0{,}0645} = 0{,}465.$$

Esto es, el 18,6, 34,9 y 46,5% de los casos de la enfermedad de Alzheimer tienen edades entre 65-74, 75-84 y ≥ 85 años, respectivamente.

2.6 REFERENCIAS

1. Billingsley P. *Probability and Measure, Third Edition.* New York: John Wiley & Sons, 1995.
2. Casella G, Berger RL. *Statistical Inference, Second Edition.* Belmont, CA: Duxbury Press, 2002.
3. Feller W. *An Introduction to Probability Theory and Its Applications, Volume 1, Third Edition.* New York: John Wiley & Sons, 1968.
4. Rosner B. *Fundamentals of Biostatistics, Sixth Edition.* Belmont, CA: Duxbury Press, 2006.

TEMA 3

VARIABLES ALEATORIAS Y DISTRIBUCIONES DE PROBABILIDAD

3.1 INTRODUCCIÓN

En el tema de estadística descriptiva se revisaron las técnicas necesarias para la realización de un análisis descriptivo de las variables recogidas en una muestra. El presente tema se centra en describir algunos modelos teóricos de probabilidad que permiten caracterizar la distribución poblacional de determinadas variables y que, a su vez, son aplicables a múltiples situaciones prácticas.

Cuando se realiza un estudio o un experimento aleatorio, es frecuente asignar a los resultados del mismo una cantidad numérica. A la función que asocia un número real a cada resultado de un experimento se le denomina variable aleatoria. Aunque el concepto de variable se ha introducido con anterioridad, una definición más formal de **variable aleatoria** es, por tanto, la de una función definida sobre el espacio muestral Ω que asigna a cada posible resultado de un experimento un valor numérico. Aunque en general pueden definirse múltiples variables aleatorias para un mismo experimento, es aconsejable seleccionar en cada caso aquellas variables que recojan las características fundamentales del experimento. Las variables aleatorias suelen denotarse por letras mayúsculas del final del alfabeto, tales como X, Y o Z, mientras que los valores que pueden tomar se representan por sus correspondientes letras minúsculas, x, y o z.

> ***Ejemplo 3.1*** A continuación se definen algunas variables aleatorias para los experimentos del Ejemplo 2.1 del tema anterior. En el experimento consistente en observar la supervivencia a los 6 meses de 4 pacientes con cáncer sometidos a tratamiento, una variable aleatoria X podría ser el número de supervivientes, que tomaría los valores $X = 0, 1, 2, 3$ ó 4 en función del número de pacientes que hayan sobrevivido a los 6 meses. Alternativamente, podría definirse otra variable aleatoria Y como el número de muertes, cuyos valores serían $Y = 0, 1, 2, 3$ ó 4 en función del número de muertes observadas. Para el experimento de medir el colesterol HDL de una persona, la variable aleatoria X más natural sería el nivel de colesterol HDL en mmol/l, que podría tomar cualquier valor positivo. Si el interés se centra en saber si los niveles de colesterol HDL son superiores o inferiores al umbral de 0,90 mmol/l, otra variable aleatoria Y podría definirse como $Y = 0$ si el nivel observado es inferior a 0,90 mmol/l y 1 en caso contrario. La elección de los valores 0 y 1 es arbitraria, bastaría con asignar dos valores distintos para diferenciar ambos tipos de resultados.

Como las variables aleatorias son funciones definidas sobre el espacio muestral, sus posibles valores tendrán asociada una probabilidad, que corresponderá a la probabilidad del suceso constituido por aquellos resultados del experimento que toman dichos valores. Los diferentes valores de una variable aleatoria y las probabilidades asociadas constituyen la **distribución de probabilidad** de la variable.

> ***Ejemplo 3.2*** En el primer experimento del ejemplo anterior, el número de supervivientes es una variable aleatoria que toma los valores $X = 0, 1, 2, 3$ ó 4. La probabilidad asociada al valor 0 $P(X = 0)$ sería la probabilidad del suceso "ninguno de los 4 pacientes sobrevive

a los 6 meses", la probabilidad asociada al valor 1 $P(X = 1)$ sería la probabilidad del suceso "sólo 1 de los 4 pacientes sobrevive a los 6 meses", y así sucesivamente. En el segundo experimento, el nivel de colesterol HDL es una variable aleatoria X que puede tomar cualquier valor en el intervalo $(0, \infty)$. En este caso no tiene sentido preguntarse, por ejemplo, cuál es la probabilidad de tener exactamente un nivel de colesterol HDL de 1 mmol/l, ya que si esta variable se pudiera determinar con una precisión infinita, la probabilidad $P(X = 1) = 0$. En tal caso, deberíamos preguntarnos por la probabilidad de un determinado intervalo de valores. Así, por ejemplo, la probabilidad $P(X \leq 1)$ sería la probabilidad del suceso "tener niveles de colesterol HDL menores o iguales a 1 mmol/l".

En general, se distinguen dos grandes grupos de variables aleatorias:

- **Variables aleatorias discretas** son aquellas que tan sólo puede tomar un número discreto (finito o infinito) de valores. Cada uno de estos valores lleva asociada una probabilidad positiva, mientras que la probabilidad de los restantes valores es 0.
- **Variables aleatorias continuas** son aquellas que pueden tomar cualquier valor dentro de un intervalo. En este caso, la probabilidad de obtener un valor concreto es 0, por lo que las probabilidades se asignan a intervalos de valores.

A continuación se describen las principales características de las variables aleatorias discretas y continuas, así como algunas distribuciones teóricas de probabilidad que serán aplicables a muchas de las variables aleatorias utilizadas en la práctica.

3.2 DISTRIBUCIONES DE PROBABILIDAD DISCRETAS

Las variables aleatorias discretas toman un número discreto de valores con probabilidad no nula y, en consecuencia, estarán completamente caracterizadas si se conoce la probabilidad asociada a cada uno de estos valores. La función que asigna a cada posible valor x_i, $i = 1, 2, ...$, de la variable discreta X su probabilidad $P(X = x_i)$ se conoce como **función de masa de probabilidad**. Esta función debe cumplir las siguientes propiedades: la probabilidad de cada valor ha de estar entre 0 y 1, $0 < P(X = x_i) \leq 1$, y la suma de las probabilidades para todos los valores debe ser igual a 1,

$$\sum_{i \geq 1} P(X = x_i) = 1.$$

Una vez conocida la función de masa de probabilidad, la probabilidad de que una variable aleatoria discreta X esté comprendida en cualquier subconjunto A se calcula como la suma de las probabilidades de aquellos valores x_i incluidos dentro de ese subconjunto,

$$P(X \in A) = \sum_{x_i \in A} P(X = x_i).$$

En particular, la **función de distribución** $F(x)$ de una variable aleatoria X se define como la probabilidad de observar un valor menor o igual a x,

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i).$$

La función de distribución de una variable discreta será una función escalonada creciente con saltos en los valores x_i con probabilidad no nula.

Ejemplo 3.3 Supongamos que por estudios previos se estima que, después de 6 meses de tratamiento en 4 pacientes con cáncer, la probabilidad de que sobrevivan 0, 1, 2, 3 ó 4 pacientes viene determinada por la segunda columna de la Tabla 3.1. Estos valores y sus probabilidades constituyen la función de masa de probabilidad de la variable número de supervivientes, que se muestra en la Figura 3.1(*a*). Los valores de la función de distribución en 0, 1, 2, 3 y 4 aparecen en la tercera columna de la Tabla 3.1; así, por ejemplo, la función de distribución en 1 es $F(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = 0{,}1296 + 0{,}3456 = 0{,}4752$. La función de distribución de esta variable se representa en la Figura 3.1(*b*). Notar que $F(x)$ está definida sobre cualquier número real, aun cuando la variable tome sólo los valores 0, 1, 2, 3 y 4 con probabilidad no nula.

En el primer tema de estadística descriptiva, se definieron la media y la varianza muestral como medidas de tendencia central y dispersión de una variable en una muestra. A continuación, se definen medidas análogas para la distribución poblacional de una variable aleatoria. La **esperanza** o **media poblacional** de una variable aleatoria discreta X, denotada por μ o $E(X)$, se define como la suma de los productos de cada valor x_i por su probabilidad $P(X = x_i)$,

$$\mu = E(X) = \sum_{i \geq 1} x_i P(X = x_i)\,.$$

Tabla 3.1 Función de masa de probabilidad y función de distribución del número de supervivientes a los 6 meses de 4 pacientes con cáncer sometidos a tratamiento.

Número de supervivientes (x)	Función de masa $P(X = x)$	Función de distribución $F(x) = P(X \leq x)$
0	0,1296	0,1296
1	0,3456	0,4752
2	0,3456	0,8208
3	0,1536	0,9744
4	0,0256	1,0000

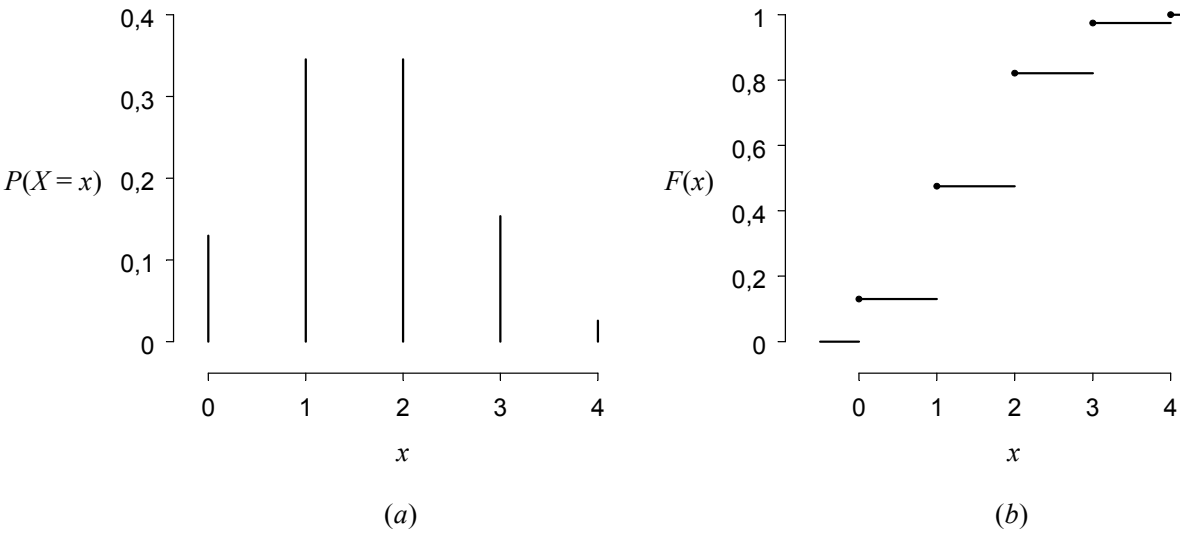


Figura 3.1 Función de masa de probabilidad (*a*) y función de distribución (*b*) del número de supervivientes a los 6 meses de 4 pacientes con cáncer sometidos a tratamiento.

La esperanza es la media de los valores x_i ponderados por su probabilidad y representa así el valor promedio de la variable aleatoria. Notar que la media muestral se puede calcular de forma similar, multiplicando cada valor observado de la variable por su frecuencia relativa. La **varianza poblacional** de una variable aleatoria discreta X, abreviada por σ^2 o var(X), se define como la esperanza del cuadrado de la desviación de la variable respecto de su media,

$$\sigma^2 = \text{var}(X) = E(X-\mu)^2 = \sum_{i\geq 1}(x_i - \mu)^2 P(X = x_i)$$
$$= \sum_{i\geq 1} x_i^2 P(X = x_i) - \mu^2 = E(X^2) - \mu^2.$$

Así, la varianza resulta ser la media ponderada del cuadrado de las desviaciones en los valores x_i. La raíz cuadrada de la varianza es la desviación típica poblacional σ, que representa la dispersión de la variable aleatoria respecto de su media poblacional.

Ejemplo 3.4 A partir de los datos del ejemplo anterior, el valor esperado del número de supervivientes a los 6 meses de 4 pacientes con cáncer sometidos a tratamiento sería

$$\mu = \sum_{k=0}^{4} kP(X = k) = 0{\cdot}0{,}1296 + 1{\cdot}0{,}3456 + ... + 4{\cdot}0{,}0256 = 1{,}60,$$

y la varianza

$$\sigma^2 = \sum_{k=0}^{4}(k - \mu)^2 P(X = k)$$
$$= (0-1{,}60)^2 0{,}1296 + ... + (4-1{,}60)^2 0{,}0256 = 0{,}96.$$

Es decir, el número esperado de supervivientes a los 6 meses es 1,60 y la desviación típica $\sigma = \sqrt{0{,}96} = 0{,}98$.

3.2.1 Distribución binomial

La distribución binomial es un modelo teórico de distribución de probabilidad discreta aplicable a aquellos experimentos en los que se realizan n pruebas independientes, cada una de ellas con sólo dos resultados posibles (éxito o fracaso) y la misma probabilidad de éxito π. En tal caso, se dice que la variable aleatoria X "número de éxitos en las n pruebas" sigue una distribución binomial con parámetros n y π. A partir de los resultados del tema de probabilidad (véase Ejemplo 3.5), puede probarse que la distribución binomial toma valores en $k = 0, 1, ..., n$ con probabilidad

$$P(X = k) = \binom{n}{k}\pi^k (1-\pi)^{n-k},$$

donde $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ es el número de combinaciones de n elementos tomados de k en k, con $n! = n(n-1)\cdot\ldots\cdot 1$ y $0! = 1$. Por supuesto, estas probabilidades constituyen una función de masa de probabilidad ya que, para cualquier n y π, su suma es exactamente igual a 1. En la práctica, resulta tedioso calcular las probabilidades de una distribución binomial mediante la

fórmula anterior. Por ello, en la Tabla 1 del Apéndice se facilitan las probabilidades binomiales para $n = 2, 3, ..., 20$ y $\pi = 0{,}05, 0{,}10, ..., 0{,}50$.

En general, la distribución binomial se aplica al estudio de observaciones repetidas e independientes de una misma variable dicotómica (con sólo dos resultados posibles), tal como el resultado de un tratamiento (éxito o fracaso) en pacientes de similares características sometidos a una misma terapia.

Ejemplo 3.5 En los ejemplos anteriores, se ha considerado el experimento de observar la supervivencia (o muerte) en pacientes con un determinado cáncer sometidos al mismo tratamiento. Si por estudios previos se sabe que la supervivencia a los 6 meses en dichos pacientes es del 40%, el número de supervivientes a los 6 meses en una muestra de 4 pacientes seguirá una distribución binomial X de parámetros $n = 4$ y $\pi = 0{,}4$.

Utilizando las leyes de la probabilidad, si denotamos por S_i al suceso de que sobreviva el i-ésimo paciente, la probabilidad de que sobrevivan únicamente los dos primeros pacientes vendría dada por

$$P(S_1 \cap S_2 \cap S_3^c \cap S_4^c) = P(S_1)P(S_2)P(S_3^c)P(S_4^c) = 0{,}4^2(1 - 0{,}4)^2,$$

dado que el resultado en cada paciente es independiente y todos tienen una misma probabilidad de supervivencia del 0,4. En general, la probabilidad de que sobrevivan 2 pacientes cualesquiera puede descomponerse, en función de qué pacientes sobrevivan, como

$$\begin{aligned} P(X = 2) = P\{&(S_1 \cap S_2 \cap S_3^c \cap S_4^c) \cup (S_1 \cap S_2^c \cap S_3 \cap S_4^c) \\ &\cup (S_1 \cap S_2^c \cap S_3^c \cap S_4) \cup (S_1^c \cap S_2 \cap S_3 \cap S_4^c) \\ &\cup (S_1^c \cap S_2 \cap S_3^c \cap S_4) \cup (S_1^c \cap S_2^c \cap S_3 \cap S_4)\}. \end{aligned}$$

Esta probabilidad está constituida por la unión de tantos sucesos como posibles combinaciones de 4 pacientes tomados de 2 en 2; es decir, $\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{24}{4} = 6$ sucesos. Además, estos sucesos son mutuamente excluyentes y todos ellos tienen una misma probabilidad de ocurrir de $0{,}4^2(1 - 0{,}4)^2$. En consecuencia, la probabilidad de que sobrevivan 2 pacientes cualesquiera es

$$P(X = 2) = \binom{4}{2} 0{,}4^2 (1 - 0{,}4)^2 = 0{,}3456,$$

que corresponde a la probabilidad de la distribución binomial de parámetros $n = 4$ y $\pi = 0{,}4$ para $k = 2$. Aplicando esta fórmula, las probabilidades para $k = 0, 1, 2, 3$ ó 4 supervivientes aparecen en la Tabla 3.1 y en la Figura 3.1(a). Estas probabilidades también pueden obtenerse directamente de la Tabla 1 del Apéndice.

A partir de las fórmulas generales para la esperanza y la varianza de una variable aleatoria discreta, puede probarse que la esperanza de una distribución binomial de parámetros n y π es

$$E(X) = \sum_{k=0}^{n} kP(X = k) = \sum_{k=0}^{n} k\binom{n}{k}\pi^k (1 - \pi)^{n-k} = n\pi$$

y su varianza es

$$\begin{aligned}\text{var}(X) &= \sum_{k=0}^{n}(k-n\pi)^2 P(X=k)\\ &= \sum_{k=0}^{n}(k-n\pi)^2\binom{n}{k}\pi^k(1-\pi)^{n-k} = n\pi(1-\pi).\end{aligned}$$

Así, el número esperado de éxitos es igual al número de pruebas realizadas por la probabilidad individual de éxito. La varianza $n\pi(1-\pi)$ disminuye cuanto menor sea el número de pruebas y más extrema sea la probabilidad de éxito. En el caso particular de que $\pi = 0$ ó 1, la varianza será 0 ya que todas las pruebas serán respectivamente fracasos o éxitos.

> ***Ejemplo 3.6*** Continuando con el ejemplo anterior, el número esperado de supervivientes a los 6 meses de 4 pacientes con cáncer sometidos a tratamiento es $n\pi = 4{\cdot}0{,}4 = 1{,}60$, la varianza $n\pi(1-\pi) = 4{\cdot}0{,}4{\cdot}0{,}6 = 0{,}96$ y la desviación típica $\sqrt{n\pi(1-\pi)} = 0{,}98$. Estos resultados coinciden con los obtenidos en el Ejemplo 3.4, donde la media y la varianza se calculaban a partir de las fórmulas generales para variables discretas.

3.2.2 Distribución de Poisson

La distribución de Poisson es otro modelo teórico de distribución discreta particularmente útil para el estudio epidemiológico de la ocurrencia de determinadas enfermedades. Se dice que la variable aleatoria X "número de casos de una determinada enfermedad a lo largo de un periodo de tiempo t", donde t es un intervalo de tiempo arbitrariamente largo, tal como 1 ó 10 años, sigue una distribución de Poisson si se cumplen las siguientes hipótesis respecto a la incidencia acumulada IA de la enfermedad (esto es, la probabilidad de desarrollar un nuevo caso en un periodo de tiempo determinado):

- **Proporcionalidad**: La probabilidad de observar un caso es aproximadamente proporcional al tiempo transcurrido, de tal forma que en un intervalo de tiempo arbitrariamente corto, la probabilidad de observar un caso es muy pequeña y la probabilidad de observar más de un caso es esencialmente nula.
- **Estacionaridad**: El número de casos por unidad de tiempo permanece aproximadamente constante a lo largo de todo el periodo de tiempo t. Notar que, si se produjera un cambio substancial de la incidencia de la enfermedad en el tiempo, esta asunción no sería aplicable.
- **Independencia**: La ocurrencia de un caso en un determinado instante no afecta a la probabilidad de observar nuevos casos en periodos posteriores. Así, por ejemplo, esta hipótesis de independencia no se cumplirá en brotes epidémicos.

Aunque la distribución de Poisson se emplea habitualmente en el estudio de la morbi-mortalidad debida a determinadas enfermedades, esta distribución es en general aplicable a la ocurrencia en el tiempo de aquellos sucesos aleatorios que satisfagan las hipótesis anteriores (por ejemplo, los accidentes de tráfico).

Bajo estas asunciones, se establece que la probabilidad de que ocurran k sucesos, $k = 0, 1, 2, \ldots$, en un periodo de tiempo t para una variable aleatoria X que sigue una distribución de Poisson es

$$P(X=k) = \frac{e^{-\mu}\mu^k}{k!},$$

donde el parámetro μ es el número esperado de sucesos en el periodo de tiempo t. A diferencia de la distribución binomial, donde el número de éxitos k no puede exceder el número finito de pruebas realizadas, en la distribución de Poisson el número de pruebas se considera infinito y el número de sucesos k puede ser arbitrariamente grande, aunque la probabilidad $P(X=k)$ decrecerá al aumentar k hasta hacerse esencialmente nula. Para cualquier parámetro $\mu > 0$, estas probabilidades son positivas y suman 1, constituyendo una función de masa de probabilidad. En la Tabla 2 del Apéndice se presentan las probabilidades de Poisson para μ de 0,5 a 20 en intervalos de 0,5.

Una característica importante de la distribución de Poisson es que tanto su media como su varianza son iguales al parámetro μ,

$$E(X) = \sum_{k\geq 0} kP(X=k) = \sum_{k\geq 0} k\frac{e^{-\mu}\mu^k}{k!} = \mu,$$

$$\text{var}(X) = \sum_{k\geq 0} (k-\mu)^2 P(X=k) = \sum_{k\geq 0} (k-\mu)^2 \frac{e^{-\mu}\mu^k}{k!} = \mu.$$

Ejemplo 3.7 Según el último Atlas de Mortalidad por Cáncer en España, la tasa de mortalidad por cáncer de vesícula en hombres es de $I = 1{,}80$ casos por 100.000 personas-año. Partiendo de esta información, se pretende determinar la distribución del número de muertes por cáncer de vesícula en un periodo de 1 ó 2 años en una población de 140.000 hombres. Las asunciones de estacionaridad e independencia parecen razonables por tratarse de casos de mortalidad por cáncer en periodos cortos de tiempo. Además, como la tasa de mortalidad I es baja y se asume constante en el tiempo, puede probarse que la incidencia acumulada en un periodo de tiempo t es

$$IA_t = 1 - \exp(-It) \approx It;$$

es decir, la probabilidad de que un individuo de esta población muera por cáncer de vesícula es aproximadamente proporcional al tiempo transcurrido, cumpliéndose así la hipótesis de proporcionalidad. La incidencia acumulada en 1 año es $IA_1 = 0{,}000018$ y en 2 años $IA_2 = 0{,}000018 \cdot 2 = 0{,}000036$. En consecuencia, el número de muertes por cáncer de vesícula en un periodo de tiempo t seguirá una distribución de Poisson con un número esperado de casos igual al producto del tamaño poblacional por la probabilidad individual de muerte en dicho periodo, $\mu = 140.000 \cdot 0{,}000018 = 2{,}52$ muertes esperadas en 1 año y $140.000 \cdot 0{,}000036 = 5{,}04$ en 2 años.

Estas distribuciones de probabilidad se muestran en la Tabla 3.2 y en la Figura 3.2. Por ejemplo, la probabilidad de que no se produzca ninguna muerte por cáncer de vesícula durante 1 año en esta población se calcula a partir de la distribución de Poisson de parámetro $\mu = 2{,}52$ como $P(X=0) = e^{-\mu}\mu^0/0! = e^{-2{,}52} = 0{,}0805$. Estas distribuciones también pueden aproximarse mediante las probabilidades de Poisson de la Tabla 2 del Apéndice para $\mu = 2{,}5$ y 5. En la Figura 3.2 puede observarse como, al aumentar el número esperado de muertes, la distribución tiende a ser más simétrica alrededor del valor esperado y su varianza aumenta.

Tabla 3.2 **Distribución de probabilidad del número de muertes por cáncer de vesícula en periodos de 1 y 2 años en una población de 140.000 hombres.**

Número de muertes (k)	$P(X = k)$	
	1 año	2 años
0	0,0805	0,0065
1	0,2028	0,0326
2	0,2555	0,0822
3	0,2146	0,1381
4	0,1352	0,1740
5	0,0681	0,1754
6	0,0286	0,1474
7	0,0103	0,1061
8	0,0032	0,0668
9	0,0009	0,0374
10	0,0002	0,0189
11	0,0001	0,0086
12	0,0000	0,0036
13	0,0000	0,0014
14	0,0000	0,0005
15	0,0000	0,0002
16	0,0000	0,0001
17	0,0000	0,0000

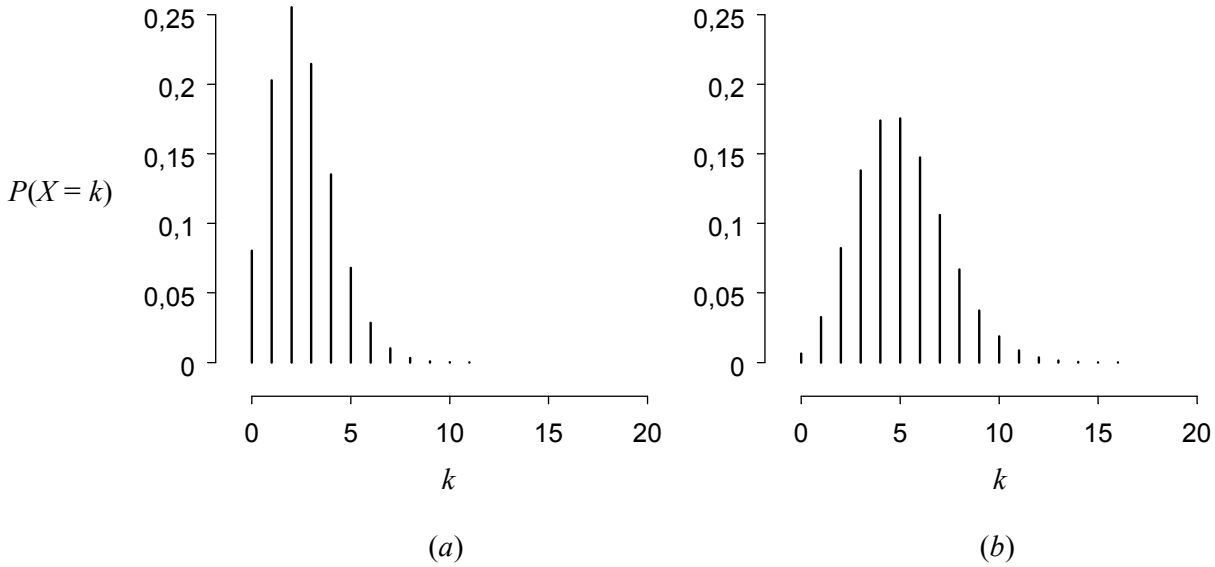


Figura 3.2 **Distribución de probabilidad del número de muertes por cáncer de vesícula en un periodo de 1 año (*a*) y de 2 años (*b*) en una población de 140.000 hombres.**

3.2.3 Aproximación de Poisson a la distribución binomial

Bajo determinadas circunstancias, la distribución de Poisson puede utilizarse como aproximación a la distribución binomial. Supongamos que, en una distribución binomial, el número de pruebas n es grande y la probabilidad individual de éxito π es pequeña. En tal caso, el número de éxitos de la distribución binomial puede ser muy grande y su varianza será aproximadamente igual al valor esperado, $n\pi(1-\pi) \approx n\pi$. Como se vio en el apartado anterior, estas dos características son propias de una distribución de Poisson, lo que sugiere la validez del siguiente resultado: si el número de pruebas n es grande y la probabilidad de éxito π es pequeña, la distribución binomial se aproxima a una distribución de Poisson con parámetro $\mu = n\pi$. Por regla general, esta aproximación se considera suficientemente precisa cuando $n \geq 100$ y $\pi \leq 0{,}01$.

Este resultado es particularmente útil en la práctica, ya que el cálculo de las probabilidades binomiales para n grande y π pequeña es muy laborioso, en cuyo caso las probabilidades de Poisson son más fáciles de manejar y facilitan resultados virtualmente idénticos.

Ejemplo 3.8 Retomemos del ejemplo anterior la variable aleatoria X correspondiente al número de muertes por cáncer de vesícula en un periodo de 2 años en una población de 140.000 hombres. El experimento subyacente consistiría en observar, para cada uno de los n = 140.000 hombres, la ocurrencia o no de una muerte por cáncer de vesícula durante un periodo de 2 años. El resultado en cada sujeto es independiente y la probabilidad de que un individuo promedio de esta población muera por cáncer de vesícula en 2 años es $\pi = IA_2 = 0{,}000036$. Por tanto, el número de muertes por cáncer de vesícula en esta población a lo largo de 2 años seguirá una distribución binomial con parámetros $n = 140.000$ y $\pi = 0{,}000036$. Así, por ejemplo, la probabilidad de que ocurran exactamente 2 muertes es

$$P(X=2) = \binom{140.000}{2} 0{,}000036^2 0{,}999964^{139.998} = 0{,}082220.$$

Utilizando la aproximación de Poisson a la distribución binomial, el número de muertes por cáncer de vesícula en un periodo de 2 años seguirá aproximadamente una distribución de Poisson con parámetro $\mu = n\pi = 140.000 \cdot 0{,}000036 = 5{,}04$. En consecuencia, la probabilidad de observar 2 muertes puede aproximarse por

$$P(X=2) \approx \frac{e^{-5{,}04}\, 5{,}04^2}{2!} = 0{,}082222,$$

que coincide casi perfectamente con la probabilidad binomial exacta.

3.3 DISTRIBUCIONES DE PROBABILIDAD CONTINUAS

Las variables aleatorias continuas son aquellas que pueden tomar cualquier valor dentro de un intervalo. La probabilidad de que estas variables tomen exactamente un valor determinado es 0 y, en consecuencia, carece de sentido definir una función de masa de probabilidad. Para las variables aleatorias continuas, las probabilidades se asignan a intervalos de valores mediante una **función de densidad de probabilidad**, denotada por $f(x)$. Esta función ha de ser no negativa para cualquier valor x, $f(x) \geq 0$, y el área total bajo la curva definida por esta función de densidad debe ser igual a 1,

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

A partir de la función de densidad, la probabilidad de que una variable aleatoria continua X tome valores dentro de cualquier intervalo (a, b) puede calcularse como el área bajo la función de densidad entre los puntos a y b,

$$P(a < X < b) = \int_a^b f(x)\, dx\,.$$

Así, aun cuando la probabilidad de obtener un valor concreto es 0, la función de densidad tomará valores elevados en regiones de alta probabilidad y valores pequeños en regiones de baja probabilidad. La **función de distribución** $F(x)$ corresponde a la probabilidad de que la variable tome un valor igual o inferior a x y, en el caso de una variable aleatoria continua, se calcula como el área bajo de la curva de la función de densidad a la izquierda de x,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt\,.$$

La función de distribución de una variable aleatoria continua es una función que, partiendo de 0, crece de forma continua hasta alcanzar el valor 1.

> ***Ejemplo 3.9*** La función de densidad para el colesterol HDL en hombres adultos se representa en la Figura 3.3(*a*). Notar que, aunque el área bajo la curva ha de ser igual a 1, la función de densidad puede tomar valores superiores a 1. Los niveles de colesterol HDL próximos a 1 mmol/l son los que tienen mayor probabilidad de ocurrir, mientras que para niveles inferiores y superiores esta probabilidad decrece. Así, por ejemplo, la probabilidad de que un hombre adulto tenga un nivel de colesterol HDL inferior a 0,90 mmol/l (niveles bajos según las recomendaciones del "*National Cholesterol Education Program*") corresponde al área sombreada bajo la curva a la izquierda de 0,90 mmol/l y es igual a $P(X \leq 0{,}90) = 0{,}3274$. Esta probabilidad también puede obtenerse a partir de la función de distribución del colesterol HDL, que se representa en la Figura 3.3(*b*). Esta función presenta el aspecto característico de las funciones de distribución para variables continuas aproximadamente simétricas.

Al igual que para variables discretas, la **esperanza** o **media poblacional** de una variable aleatoria continua representa el valor promedio de esa variable, y se define como

$$\mu = E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx\,.$$

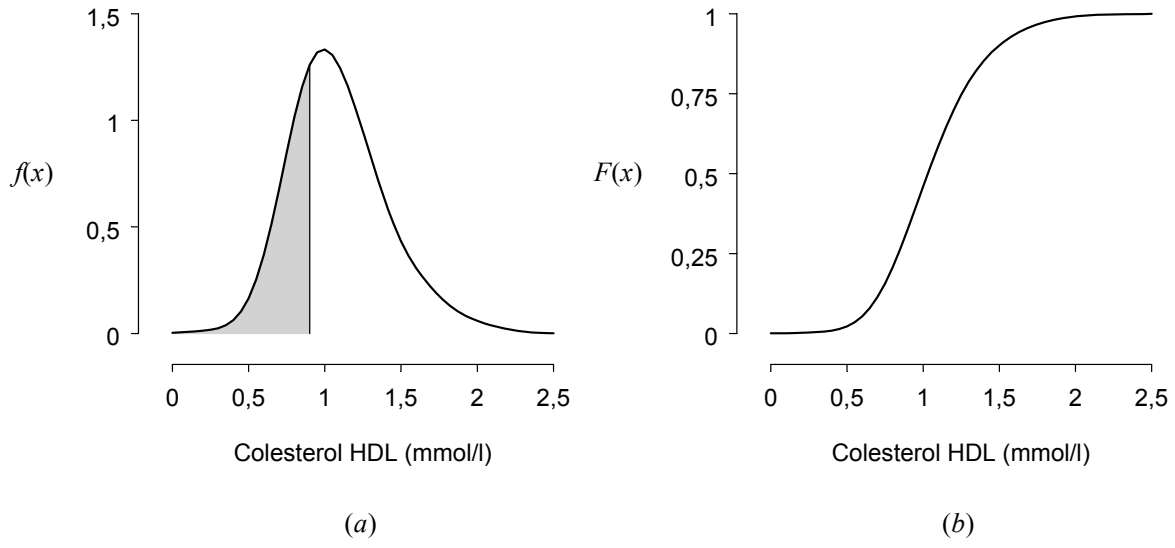


Figura 3.3 **Función de densidad de probabilidad (*a*) y función de distribución (*b*) del colesterol HDL en hombres adultos.**

La **varianza poblacional** de una variable aleatoria continua es la esperanza de las desviaciones al cuadrado de los valores de la variable respecto de su media, y se calcula como

$$\sigma^2 = \text{var}(X) = E(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

$$= \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu^2 = E(X^2) - \mu^2.$$

La raíz cuadrada de la varianza es la desviación típica poblacional σ, que representa la dispersión de la variable aleatoria respecto de su media poblacional. Estas expresiones para la media y la varianza poblacional de una variable continua son similares a las facilitadas para variables discretas, salvo que la suma sobre el número discreto de valores con probabilidad no nula se reemplaza por la integral sobre todos los posibles valores de la variable continua.

Ejemplo 3.10 Utilizando la función de densidad del ejemplo anterior, el valor esperado del colesterol HDL en una población de hombres adultos sería

$$\mu = \int_0^{\infty} x\, f(x)\, dx = 1{,}10 \text{ mmol/l},$$

y la desviación típica

$$\sigma = \left(\int_0^{\infty} (x - 1{,}10)^2 f(x)\, dx \right)^{1/2} = 0{,}30 \text{ mmol/l}.$$

Existen muchos modelos teóricos de distribuciones continuas, cada una de ellas caracterizada por una fórmula o expresión concreta para la función de densidad. A continuación se revisa en detalle la distribución normal, que es la utilizada con mayor frecuencia en estadística. Otras distribuciones continuas, como la *t* de Student, chi-cuadrado o *F* de Fisher, se discutirán según vayan surgiendo a lo largo del texto.

3.3.1 Distribución normal

La distribución normal, también denominada distribución Gaussiana, es el modelo teórico de distribución continua más utilizado en la práctica. Muchas mediciones epidemiológicas y clínicas presentan distribuciones similares al modelo teórico normal (presión arterial, colesterol sérico, índice de masa corporal) o bien pueden transformarse para conseguir distribuciones aproximadamente normales (típicamente mediante transformaciones logarítmicas de los datos originales). No obstante, como se verá en los temas posteriores, la utilidad fundamental de la distribución normal surge dentro de las técnicas de inferencia estadística: incluso cuando la distribución poblacional de una variable diste mucho de ser normal, puede probarse que, bajo ciertas condiciones, la distribución de los valores medios de dicha variable seguirá un modelo aproximadamente normal.

Una variable aleatoria continua *X* sigue una distribución normal si su función de densidad es

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right),$$

para cualquier valor x en la recta real, $-\infty < x < \infty$. Esta función de densidad depende de los parámetros μ y σ, donde

- μ representa la esperanza o media poblacional de la distribución y
- σ corresponde a su desviación típica poblacional.

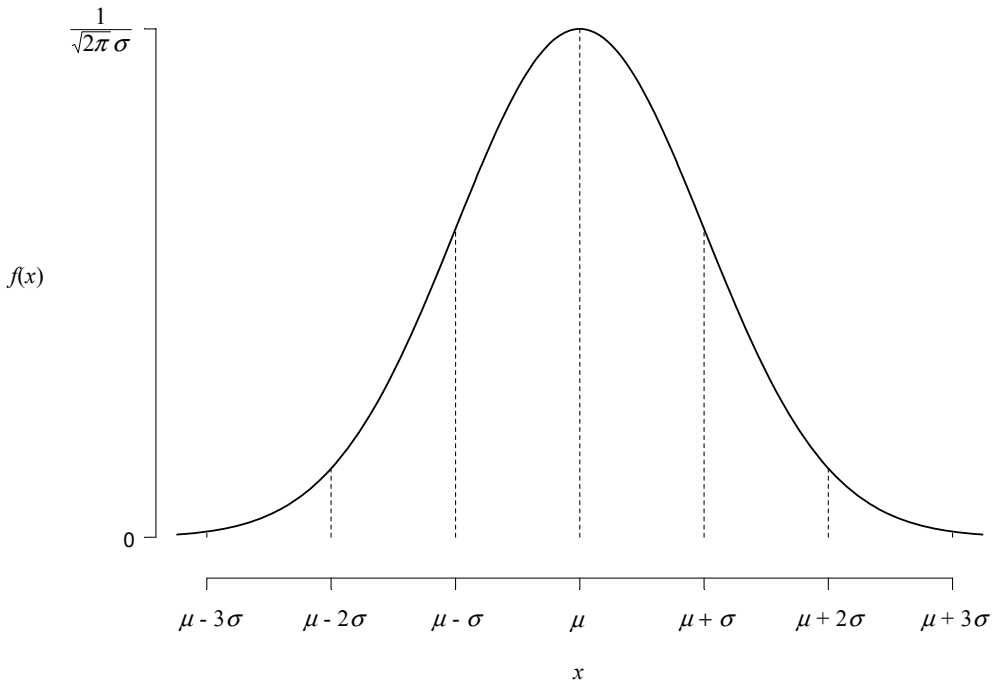


Figura 3.4 **Función de densidad de una distribución normal con media μ y desviación típica σ.**

La distribución normal o Gaussiana con media μ y varianza σ^2 se denota abreviadamente por $N(\mu, \sigma^2)$. Para cualquier μ y $\sigma > 0$, la función de densidad normal es positiva y el área total bajo la curva es igual a 1. Esta función de densidad, que aparece representada en la Figura 3.4, tiene forma de campana, es simétrica alrededor de la media μ y tiene dos puntos de inflexión en $\mu + \sigma$ y $\mu - \sigma$. Al tratarse de una distribución simétrica, la media y la mediana coinciden. El valor más frecuente $1/(\sqrt{2\pi}\,\sigma)$ se alcanza en la media μ y su dispersión alrededor del valor medio aumenta al aumentar la desviación típica σ. Así, puede probarse que el 68,27% del área bajo una función de densidad normal está comprendido entre $\mu \pm \sigma$, el 95,45% entre $\mu \pm 2\sigma$ y el 99,73% entre $\mu \pm 3\sigma$.

La distribución normal con media 0 y desviación típica 1 se denomina **distribución normal estandarizada**, y suele denotarse por Z o $N(0, 1)$. La función de densidad de una distribución normal estandarizada se reduce a

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right),$$

para cualquier $-\infty < z < \infty$, que se representa en la Figura 3.5(a). Como puede observarse, se trata de una función simétrica alrededor de 0. Para obtener las probabilidades bajo la función de densidad normal estandarizada, no se recurre al cálculo integral, ya que estas probabilidades están tabuladas y son fácilmente accesibles. En general, estas tablas facilitan la función de distribución; es decir, la probabilidad de que la variable normal estandarizada tome un valor igual o inferior a z. La función de distribución normal estandarizada se denota por $\Phi(z) = P(Z \leq z)$, y se ilustra en la Figura 3.5(b). En la Tabla 3 del Apéndice se facilita la función de distribución $\Phi(z)$ para valores de z no negativos.

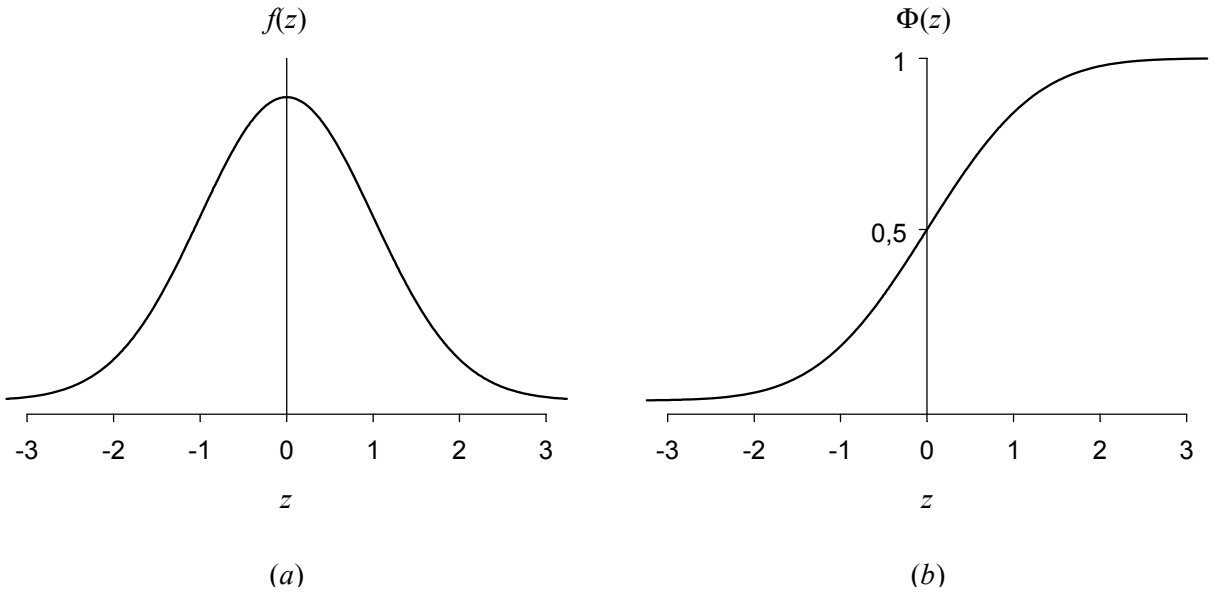


Figura 3.5 **Función de densidad (*a*) y función de distribución (*b*) de una variable aleatoria normal estandarizada.**

Ejemplo 3.11 La probabilidad de obtener un valor inferior a 0,50 en una distribución normal estandarizada se obtiene directamente de la Tabla 3 del Apéndice como el valor de la función de distribución en 0,50; es decir, $P(Z \leq 0{,}50) = \Phi(0{,}50) = 0{,}6915$. Asimismo, aunque en la Tabla 3 del Apéndice no aparecen las probabilidades acumuladas para valores negativos, la probabilidad de obtener un valor inferior a $-0{,}25$ en una distribución normal estandarizada puede calcularse fácilmente a partir de dicha tabla. Como la distribución normal estandarizada es simétrica alrededor de 0, la probabilidad a la izquierda de $-0{,}25$ es igual a la probabilidad a la derecha de 0,25 y, en consecuencia, $P(Z \leq -0{,}25) = P(Z \geq 0{,}25) = 1 - P(Z \leq 0{,}25) = 1 - \Phi(0{,}25) = 1 - 0{,}5987 = 0{,}4013$. A partir de los resultados anteriores, la probabilidad de que un valor de la distribución normal estandarizada se encuentre entre $-0{,}25$ y 0,50 viene dada por $P(-0{,}25 \leq Z \leq 0{,}50) = P(Z \leq 0{,}50) - P(Z \leq -0{,}25) = 0{,}6915 - 0{,}4013 = 0{,}2902$.

El percentil 97,5 de una distribución normal estandarizada se denota por $z_{0,975}$ y corresponde al valor z que deja por debajo una probabilidad del 0,975. De la Tabla 3 del Apéndice, se tiene que $\Phi(1{,}96) = 0{,}9750$ y, por tanto, $z_{0,975} = 1{,}96$. Por tratarse de una distribución simétrica en 0, el percentil 2,5 corresponde al percentil 97,5 con signo opuesto; es decir, el percentil 2,5 es $z_{0,025} = -z_{0,975} = -1{,}96$. Así, los valores $\pm 1{,}96$ abarcan el 95% central de la distribución normal estandarizada. Este resultado será particularmente útil en los temas de inferencia estadística.

El cálculo de probabilidades para cualquier distribución normal con media μ y varianza σ^2 no requiere de tablas específicas, sino que puede realizarse a partir de las tablas de la distribución normal estandarizada. Para ello, se hace uso del siguiente resultado sobre la estandarización de una distribución normal: si una variable aleatoria X sigue una distribución normal con media μ y varianza σ^2, $X \sim N(\mu, \sigma^2)$, entonces la variable aleatoria $Z = (X - \mu)/\sigma$ sigue una distribución normal estandarizada,

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1),$$

donde el símbolo ~ significa "estar distribuido como". Como ya se comentó en el Tema 1, al restar a los valores de una variable su media y dividirlos por su desviación típica, la variable resultante tiene media 0 y desviación típica 1. El resultado anterior garantiza además que la variable estandarizada conserva la distribución normal. Este procedimiento de estandarización de variables normales permite utilizar las tablas correspondientes a la distribución normal estandarizada.

Ejemplo 3.12 Supongamos que el colesterol HDL en una población de hombres adultos sigue una distribución normal X con media $\mu = 1{,}10$ mmol/l y desviación típica $\sigma = 0{,}30$ mmol/l. Utilizando la estandarización de variables normales, el porcentaje de hombres de esta población que tienen niveles de colesterol HDL entre 0,90 y 1,20 mmol/l corresponde a

$$P(0{,}90 \le X \le 1{,}20) = P\left(\frac{0{,}90-1{,}10}{0{,}30} \le \frac{X-1{,}10}{0{,}30} \le \frac{1{,}20-1{,}10}{0{,}30}\right)$$

$$= P(-0{,}67 \le Z \le 0{,}33) = P(Z \le 0{,}33) - P(Z \le -0{,}67).$$

Utilizando la Tabla 3 del Apéndice, se obtiene que $P(Z \le 0{,}33) = \Phi(0{,}33) = 0{,}6293$ y $P(Z \le -0{,}67) = \Phi(-0{,}67) = 1 - \Phi(0{,}67) = 1 - 0{,}7486 = 0{,}2514$. Así, resulta que $P(0{,}90 \le X \le 1{,}20) = 0{,}6293 - 0{,}2514 = 0{,}3779$; es decir, el 37,79% de los hombres de esta población tienen niveles de colesterol HDL entre 0,90 y 1,20 mmol/l.

Para obtener el percentil 90 de la distribución del colesterol HDL en esta población, se calcula primero el percentil 90 en la distribución normal estandarizada, que corresponde a $z_{0,90} = 1{,}28$, ya que $\Phi(1{,}28) \approx 0{,}90$. Para pasar este percentil estandarizado al correspondiente percentil del colesterol HDL basta resolver $z_{0,90} = (x_{0,90} - \mu)/\sigma$. Por tanto, el percentil 90 del colesterol HDL es $x_{0,90} = \mu + z_{0,90}\sigma = 1{,}10 + 1{,}28 \cdot 0{,}30 = 1{,}484$ mmol/l.

3.3.2 Aproximación normal a la distribución binomial

El cálculo de las probabilidades binomiales es muy laborioso cuando el número de pruebas n en muy elevado. Como se vio anteriormente, si n es grande y la probabilidad de éxito π es muy pequeña, la distribución binomial puede aproximarse mediante una distribución de Poisson. En este apartado, se revisa el comportamiento de una distribución binomial para un número de pruebas n grande y una probabilidad individual de éxito π no excesivamente extrema. En la Figura 3.6 se muestran las distribuciones binomiales para los parámetros $\pi = 0{,}10$ y $n = 10$, 25, 50 y 100. Al aumentar el número de pruebas, la distribución binomial tiende a ser más simétrica y se aproxima progresivamente a una distribución normal con la misma media $n\pi$ y varianza $n\pi(1-\pi)$ que la distribución binomial (Figura 3.6(d)). En general, puede probarse que si el número de pruebas n es elevado y la probabilidad de éxito π no es excesivamente extrema, de forma que $n\pi(1-\pi) \ge 5$, la distribución binomial con parámetros n y π se aproxima a una distribución normal con media $n\pi$ y varianza $n\pi(1-\pi)$.

Este resultado es un caso particular del llamado teorema central del límite, que se presentará más adelante (véase Tema 4), y se utiliza para aproximar las probabilidades binomiales mediante la distribución normal. Así, para una variable binomial X con parámetros n y π que cumpla las condiciones anteriores, la probabilidad $P(k_1 \le X \le k_2)$ se aproxima mediante el área bajo la curva de la distribución normal $N(n\pi, n\pi(1-\pi))$ entre $k_1 - 1/2$ y $k_2 + 1/2$, donde $k_1 \le k_2$ son números enteros cualesquiera. Notar que, al utilizar la aproximación normal, los límites del intervalo se amplían en 1/2 para incluir las probabilidades de obtener exactamente k_1 o k_2 éxitos. Este ajuste se conoce como **corrección por continuidad** y se deriva del hecho de aproximar una distribución binomial discreta mediante una distribución normal continua.

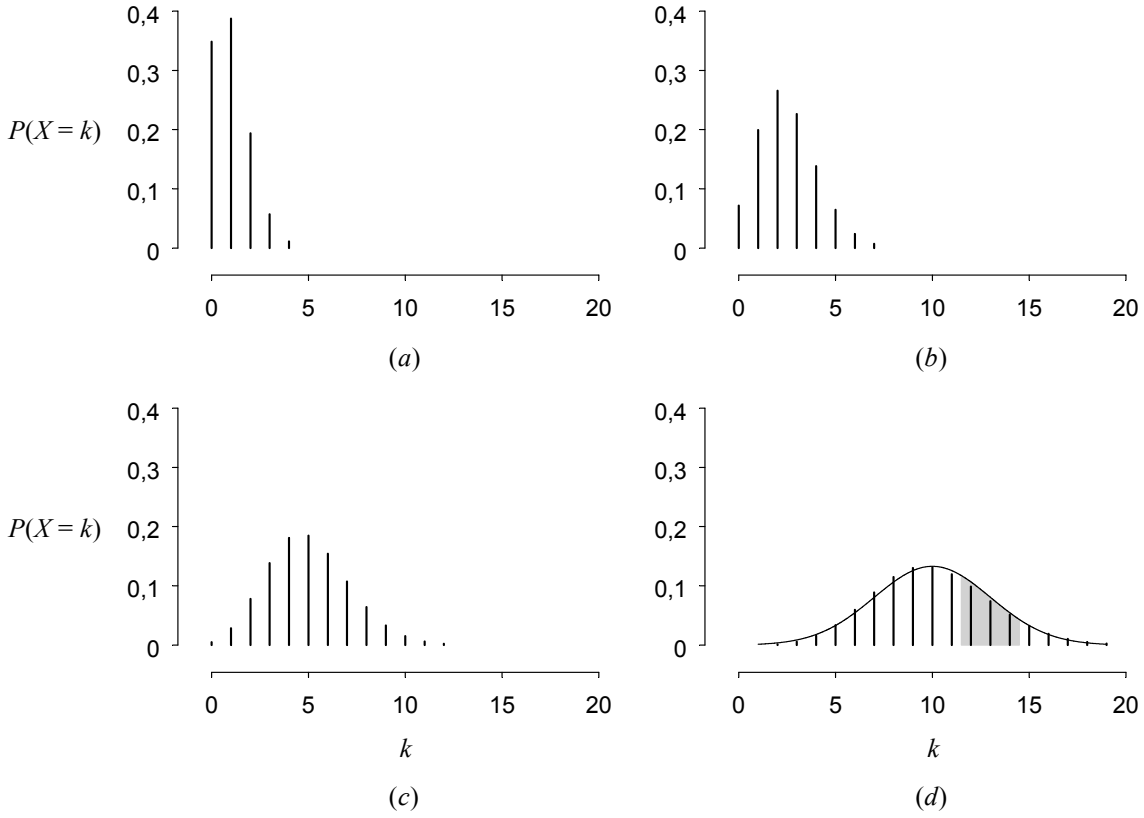


Figura 3.6 **Distribuciones binomiales con parámetros $\pi = 0{,}10$ y $n = 10$ (*a*), 25 (*b*), 50 (*c*) y 100 (*d*). En el panel *d*, se representa además la función de densidad de una distribución normal con media $n\pi = 100 \cdot 0{,}10 = 10$ y varianza $n\pi(1-\pi) = 100 \cdot 0{,}10 \cdot 0{,}90 = 9$.**

Ejemplo 3.13 La probabilidad de obtener entre 12 y 14 éxitos sobre un total de 100 pruebas con una probabilidad individual de éxito del 0,10 se obtiene a partir de la distribución binomial X con parámetros $n = 100$ y $\pi = 0{,}10$ como

$$P(12 \le X \le 14) = \sum_{k=12}^{14} \binom{100}{k} 0{,}10^k (1-0{,}10)^{100-k}$$

$$= 0{,}0988 + 0{,}0743 + 0{,}0513 = 0{,}2244,$$

cuyo cálculo es bastante laborioso. Sin embargo, como $n\pi(1-\pi) = 100 \cdot 0{,}10 \cdot 0{,}90 = 9 \ge 5$, una aproximación razonable a esta probabilidad puede obtenerse a partir de la distribución normal Y con media $n\pi = 100 \cdot 0{,}10 = 10$ y varianza $n\pi(1-\pi) = 9$ mediante

$$P(11{,}5 < Y < 14{,}5) = P\left(\frac{11{,}5-10}{3} < \frac{Y-10}{3} < \frac{14{,}5-10}{3}\right)$$

$$= P(0{,}5 < Z < 1{,}5) = \Phi(1{,}5) - \Phi(0{,}5)$$

$$= 0{,}9332 - 0{,}6915 = 0{,}2417.$$

Esta probabilidad corresponde al área sombreada en la Figura 3.6(*d*).

3.3.3 Aproximación normal a la distribución de Poisson

La distribución normal también puede emplearse como aproximación a la distribución de Poisson cuando el número esperado de casos μ es moderadamente grande. En la Figura 3.7 se representan las distribuciones de Poisson con parámetros $\mu = 1$, 2,5, 5 y 10, donde puede apreciarse que, al aumentar el número esperado de casos, las probabilidades de Poisson tienden a distribuirse de forma normal. En términos generales, una distribución de Poisson con parámetro μ se aproxima a una distribución normal con media y varianza iguales a μ, cuando el número esperado de casos es moderadamente elevado, típicamente $\mu \geq 10$. Así, para una variable aleatoria X que siga a una distribución de Poisson con parámetro μ moderadamente grande, la probabilidad $P(k_1 \leq X \leq k_2)$ puede aproximarse mediante el área bajo la curva de la distribución normal $N(\mu, \mu)$ entre $k_1 - 1/2$ y $k_2 + 1/2$.

Ejemplo 3.14 Si el número de casos de una enfermedad a lo largo de un año en una determinada población sigue una distribución de Poisson X de parámetro $\mu = 10$, la probabilidad de tener 15 o más casos en un mismo año es exactamente

$$P(X \geq 15) = \sum_{k \geq 15} \frac{e^{-10} 10^k}{k!} = 0{,}0835,$$

que puede aproximarse mediante la distribución normal $Y \sim N(10, 10)$ como

$$P(X \geq 15) \approx P(Y > 14{,}5) = P\left(\frac{Y - 10}{\sqrt{10}} > \frac{14{,}5 - 10}{\sqrt{10}} \right)$$

$$= P(Z > 1{,}42) = 1 - \Phi(1{,}42) = 1 - 0{,}9222 = 0{,}0778.$$

Esta aproximación corresponde al área sombreada bajo la curva normal en la Figura 3.7(*d*).

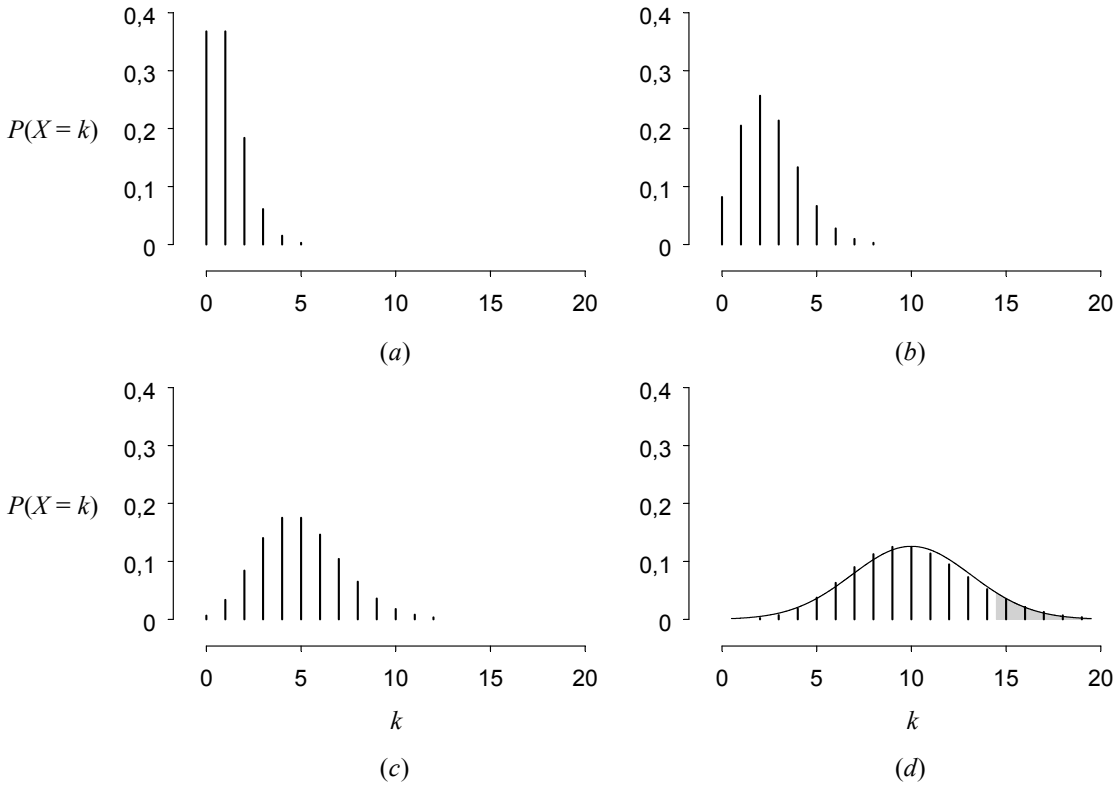


Figura 3.7 Distribuciones de Poisson con parámetros $\mu = 1$ (*a*), 2,5 (*b*), 5 (*c*) y 10 (*d*). En el panel *d*, se representa además la función de densidad de una distribución normal con media y varianza iguales a $\mu = 10$.

3.4 COMBINACIÓN LINEAL DE VARIABLES ALEATORIAS

En este apartado se introducen algunas propiedades de la combinación lineal de variables aleatorias (discretas o continuas) que serán útiles para la estimación e inferencia estadística. En particular, se pretende derivar el valor esperado y la varianza de la combinación lineal $c_1X_1 + ... + c_kX_k$, donde c_1, ..., c_k son constantes arbitrarias y X_1, ..., X_k son variables aleatorias con esperanzas μ_1, ..., μ_k y varianzas σ_1^2 , ..., σ_k^2 . Como el valor esperado de la suma de variables aleatorias es igual a la suma de sus respectivas esperanzas, se tiene que

$$E\left(\sum_{i=1}^{k} c_i X_i\right) = \sum_{i=1}^{k} E(c_i X_i) = \sum_{i=1}^{k} c_i E(X_i) = \sum_{i=1}^{k} c_i \mu_i,$$

ya que $E(c_iX_i) = c_iE(X_i)$. Es decir, la esperanza de una combinación lineal de variables aleatorias es la combinación lineal de sus esperanzas.

A partir de este resultado, y recordando que $\text{var}(X) = E(X^2) - \mu^2$, puede calcularse la varianza de una combinación lineal de variables aleatorias como

$$\begin{aligned}
\text{var}\left(\sum_{i=1}^{k} c_i X_i\right) &= E\left(\sum_{i=1}^{k} c_i X_i\right)^2 - \left(\sum_{i=1}^{k} c_i \mu_i\right)^2 \\
&= \sum_{i=1}^{k} c_i^2 E(X_i^2) + 2 \sum_{1 \le i < j \le k} c_i c_j E(X_i X_j) - \left(\sum_{i=1}^{k} c_i^2 \mu_i^2 + 2 \sum_{1 \le i < j \le k} c_i c_j \mu_i \mu_j\right) \\
&= \sum_{i=1}^{k} c_i^2 \{E(X_i^2) - \mu_i^2\} + 2 \sum_{1 \le i < j \le k} c_i c_j \{E(X_i X_j) - \mu_i \mu_j\} \\
&= \sum_{i=1}^{k} c_i^2 \sigma_i^2 + 2 \sum_{1 \le i < j \le k} c_i c_j \{E(X_i X_j) - \mu_i \mu_j\}.
\end{aligned}$$

Así, la varianza de una combinación lineal no depende sólo de la varianza específica de cada variable σ_i^2 , sino también de los términos $E(X_iX_j) - \mu_i\mu_j$, que se conocen como covarianzas entre las variables X_i y X_j. En general, la **covarianza poblacional** entre dos variables aleatorias X e Y con esperanzas μ_x y μ_y se define como

$$\text{cov}(X, Y) = E\{(X - \mu_x)(Y - \mu_y)\} = E(XY) - \mu_x\mu_y,$$

y es una medida de la relación lineal entre ambas variables. Si valores altos (o bajos) de X tienden a asociarse con valores altos (o bajos) de Y, la covarianza será positiva; mientras que si valores altos de una variable se relacionan con valores bajos de la otra variable, la covarianza será negativa. No obstante, resulta complicado determinar el grado de relación lineal entre dos variables a partir de la magnitud de la covarianza, ya que ésta depende de las unidades de medida de las variables. Una medida alternativa del grado de asociación lineal entre dos variables aleatorias X e Y es el **coeficiente de correlación poblacional** ρ_{xy}, que se define como

$$\rho_{xy} = \frac{\text{cov}(X, Y)}{\sigma_x \, \sigma_y},$$

donde σ_x y σ_y son las desviaciones típicas de X e Y. El coeficiente de correlación carece de unidades y toma valores entre -1 y 1; de tal forma que si $\rho_{xy} = 1$, las variables presentan una relación lineal positiva perfecta, y si $\rho_{xy} = -1$, las variables presentan una relación lineal negativa perfecta. Cuando $\rho_{xy} = 0$, se dice que las variables están incorrelacionadas. Notar que si dos variables son independientes, en el sentido de que el conocimiento del valor que toma una

variable no aporta ninguna información sobre el valor de la otra variable, entonces están incorrelacionadas; pero que la incorrelación no implica necesariamente independencia, ya que las variables podrían presentar una dependencia no lineal aun cuando $\rho_{xy} = 0$. Este y otros aspectos sobre el coeficiente de correlación se discutirán en mayor detalle en el Tema 10.

La varianza de una combinación lineal de variables aleatorias queda entonces determinada por

$$\text{var}\left(\sum_{i=1}^{k} c_i X_i\right) = \sum_{i=1}^{k} c_i^2 \sigma_i^2 + 2 \sum_{1 \le i < j \le k} c_i c_j \, \text{cov}(X_i, X_j)$$
$$= \sum_{i=1}^{k} c_i^2 \sigma_i^2 + 2 \sum_{1 \le i < j \le k} c_i c_j \sigma_i \, \sigma_j \rho_{ij},$$

donde ρ_{ij} es el coeficiente de correlación entre X_i y X_j. En el caso de que las variables sean mutuamente independientes (bastaría la condición menos restrictiva de que estuvieran incorrelacionadas), la varianza de la combinación lineal es

$$\text{var}\left(\sum_{i=1}^{k} c_i X_i\right) = \sum_{i=1}^{k} c_i^2 \sigma_i^2 .$$

Ejemplo 3.15 Supongamos que la media y la desviación típica de la presión arterial sistólica X_1 en una determinada población son $\mu_1 = 130$ mm Hg y $\sigma_1 = 20$ mm Hg, y la media y la desviación típica de la presión arterial diastólica X_2 son $\mu_2 = 80$ mm Hg y $\sigma_2 = 10$ mm Hg. Supongamos además que el coeficiente de correlación entre la presión arterial sistólica y diastólica de los sujetos de esta población es $\rho_{12} = 0{,}60$. El valor esperado de la presión del pulso, definida como la diferencia entre la presión arterial sistólica y diastólica, sería

$$E(X_1 - X_2) = \mu_1 - \mu_2 = 130 - 80 = 50 \text{ mm Hg}$$

y, teniendo en cuenta la correlación entre ambas variables, la varianza de la presión del pulso vendría dada por

$$\text{var}(X_1 - X_2) = \sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}$$
$$= 20^2 + 10^2 - 2 \cdot 20 \cdot 10 \cdot 0{,}60 = 260 \text{ (mm Hg)}^2,$$

para una desviación típica $\sqrt{260} = 16{,}1$ mm Hg.

Los resultados anteriores son válidos para cualquier variable aleatoria. No obstante, si las variables $X_1, \ldots, X_k$ siguen una distribución normal, puede probarse que la combinación lineal $c_1X_1 + \ldots + c_kX_k$ también seguirá una distribución normal con la media y varianza descritas anteriormente. Este resultado se utilizará en los temas de inferencia.

Ejemplo 3.16 El colesterol HDL en las mujeres adultas de una población sigue una distribución normal X_1 con media $\mu_1 = 1{,}25$ mmol/l y desviación típica $\sigma_1 = 0{,}35$ mmol/l, y en los hombres adultos de dicha población sigue una distribución normal X_2 con media $\mu_2 = 1{,}10$ mmol/l y desviación típica $\sigma_2 = 0{,}30$ mmol/l. Así, la diferencia del colesterol HDL entre las mujeres y los hombres de esta población se distribuirá según una normal con media

$$E(X_1 - X_2) = \mu_1 - \mu_2 = 1{,}25 - 1{,}10 = 0{,}15 \text{ mmol/l}$$

y varianza

$$\text{var}(X_1 - X_2) = \sigma_1^2 + \sigma_2^2 = 0{,}35^2 + 0{,}30^2 = 0{,}213 \text{ (mmol/l)}^2,$$

o desviación típica $\sqrt{0{,}213} = 0{,}46$ mmol/l, ya que los valores para distintos sujetos son independientes y, en consecuencia, $\rho_{12} = 0$.

3.5 REFERENCIAS

1. Bickel PJ, Doksum KA. *Mathematical Statistics: Basic Ideas and Selected Topics, Volume 1, Second Edition.* Upper Saddle River, NJ: Prentice Hall, 2001.
2. Casella G, Berger RL. *Statistical Inference, Second Edition.* Belmont, CA: Duxbury Press, 2002.
3. Colton T. *Estadística en Medicina.* Barcelona: Salvat, 1979.
4. Feller W. *An Introduction to Probability Theory and Its Applications, Volume 1, Third Edition*. New York: John Wiley & Sons, 1968.
5. Rosner B. *Fundamentals of Biostatistics, Sixth Edition.* Belmont, CA: Duxbury Press, 2006.
6. Stuart A, Ord JK. *Kendall's Advanced Theory of Statistics, Volume 1, Distribution Theory, Sixth Edition.* London: Edward Arnold, 1994.

TEMA 4

PRINCIPIOS DE MUESTREO Y ESTIMACIÓN

4.1 INTRODUCCIÓN

Un primer paso en la realización de un estudio o proyecto de investigación es definir la población de la cual se desea conocer una determinada característica o parámetro. Ocasionalmente, resulta factible obtener información para todos los elementos de la población mediante registros o censos. Sin embargo, en la mayoría de los estudios no es posible obtener información de toda la población, por lo que debemos limitarnos a la recogida de datos en una pequeña fracción del total o muestra.

La utilización de muestras presenta varias ventajas con respecto a la enumeración completa de la población:

- Coste reducido. Si los datos se obtienen de una pequeña fracción del total, los gastos se reducen. Incluso si la obtención de información en toda la población es factible, suele ser mucho más eficiente la utilización de técnicas de muestreo.
- Mayor rapidez. Los datos pueden ser más fácilmente recolectados y estudiados si se utiliza una muestra que si se emplean todos los elementos de la población. Por tanto, el uso de técnicas de muestreo es especialmente importante cuando se necesita la información con carácter urgente.
- Mayor flexibilidad y mayores posibilidades de estudio. La disponibilidad de registros completos es limitada. Muy a menudo, la única alternativa posible para la realización de un estudio es la obtención de datos por muestreo.
- Mayor control de calidad del proceso de recogida de datos. Al recoger datos en un número menor de efectivos, resulta más fácil recoger un número mayor de variables por individuo, así como tener un mejor control de la calidad del proceso de recogida de datos.

Si se dispone de información para todas las unidades de la población, el parámetro poblacional de interés quedará determinado con total precisión. Sin embargo, si se emplea únicamente una fracción del total, el parámetro poblacional desconocido ha de estimarse a partir de la muestra, con el consiguiente error derivado tanto por el carácter parcial de la muestra como por su posible falta de representatividad poblacional. La **teoría de muestreo** persigue un doble objetivo. Por un lado, estudia las técnicas que permiten obtener muestras representativas de la población de forma eficiente. Por otro lado, la teoría de muestreo indica cómo utilizar los resultados del muestreo para estimar los parámetros poblacionales, conociendo a la vez el grado de incertidumbre de las estimaciones. Así, la teoría de muestreo pretende dar respuesta a varias preguntas de interés:

- ¿Cómo se eligen a los individuos que componen la muestra?
- ¿Cuántos individuos formarán parte de la muestra?
- ¿Cómo se cuantifican las diferencias existentes entre los resultados obtenidos en la muestra y los que hubiéramos obtenido si el estudio se hubiera llevado a cabo en toda la población?

Estas cuestiones están estrechamente relacionadas entre sí. Así, por ejemplo, al aumentar el tamaño muestral aumenta la exactitud en las estimaciones. La determinación del tamaño muestral se tratará más adelante (véase Tema 9). En el presente tema, se discuten los principales tipos de muestreo probabilístico, así como la estimación en el muestreo aleatorio simple. Antes de ello, es conveniente revisar la definición de algunos conceptos que se utilizan de forma repetida a lo largo del capítulo:

- **Población** o **universo muestral** es la colección de elementos o **unidades de análisis** acerca de los cuales se desea información. Con frecuencia, no se puede obtener información de toda la población, sino tan sólo de unidades que cumplen una serie de características (criterios de inclusión/exclusión). La **población marco** es aquella sobre la que es posible obtener información. La muestra se obtiene de la población marco, por lo que debe recordarse que las conclusiones extraídas de la muestra son generalizables a la población marco y no necesariamente a la población de inicio o universo.
- Dentro del proceso de selección de una muestra, la población suele dividirse en **unidades de muestreo**, que deben constituir una partición de toda la población. Estas unidades de muestreo pueden coincidir con las unidades de análisis, pero también pueden estar constituidas por un conjunto de distintas unidades de análisis.

  Ejemplo 4.1 Supongamos que se desea estudiar la capacidad funcional de una población de ancianos institucionalizados. Para ello, se dispone de un lista de residencias, algunas de las cuales se seleccionan para el estudio. Dentro de cada residencia seleccionada, se eligen a su vez algunos ancianos que formarán parte de la muestra definitiva. En tal caso, la selección de la muestra se habría realizado en dos etapas: las residencias constituirían las unidades de muestreo de primera etapa y los ancianos (unidades de análisis) serían las unidades de muestreo de segunda etapa.

- **Muestreo probabilístico** es aquel en que todas las unidades de la población tienen una probabilidad conocida y no nula de ser seleccionadas para la muestra. El muestreo probabilístico minimiza la probabilidad de sesgos (si el tamaño muestral no es muy limitado, la muestra será muy probablemente representativa de la población) y permite cuantificar el error cometido en las estimaciones como consecuencia de la variabilidad aleatoria. La teoría del muestreo se basa fundamentalmente en el muestreo probabilístico, ya que otros tipos de muestreo (de conveniencia, por cuotas) están sujetos a una mayor probabilidad de sesgos y es más difícil extrapolar los resultados a la población.
- En el **muestreo con reposición**, cada vez que se elige un nuevo elemento muestral se dispone de toda la población para realizar la selección, mientras que en el **muestreo sin reposición** los elementos que ya han aparecido en la muestra no están disponibles para ser elegidos de nuevo. En el muestreo con reposición, por tanto, una unidad poblacional puede aparecer más de una vez en la muestra. En la práctica, el muestreo suele realizarse sin reposición. No obstante, si el tamaño de la población es muy grande con respecto al tamaño muestral, la probabilidad de que un elemento de la población sea elegido más de una vez en la muestra es tan pequeña que ambos tipos de muestreo son similares.

4.2 PRINCIPALES TIPOS DE MUESTREO PROBABILÍSTICO

En este apartado se describen brevemente los principales procedimientos probabilísticos de selección de muestras, tales como los muestreos aleatorio simple, sistemático, estratificado, por

conglomerados y polietápico. Un tratamiento más extenso de estos procedimientos puede encontrarse en los libros de muestreo referenciados al final del tema.

4.2.1 Muestreo aleatorio simple

El muestreo aleatorio simple es el más sencillo y conocido de los distintos tipos de muestreo probabilístico. Supongamos que se pretende seleccionar una muestra de tamaño n a partir de una población de N unidades. Un muestreo aleatorio simple es aquel en el que cualquier subconjunto de tamaño n tiene la misma probabilidad de ser seleccionado. Puede probarse que el muestreo aleatorio simple es un procedimiento equiprobabilístico; es decir, todas las unidades de la población tienen la misma probabilidad n/N de ser elegidas en la muestra.

Para la selección de una muestra aleatoria simple, se enumeran previamente las unidades del universo o población de 1 a N y a continuación se seleccionan n números distintos entre 1 y N utilizando algún procedimiento aleatorio, típicamente mediante una tabla de números aleatorios o un generador de números aleatorios por ordenador.

- Las tablas de números aleatorios son tablas con los dígitos 0, 1, 2, ..., 9, donde cada dígito tiene la misma probabilidad de ocurrir y el valor de un dígito concreto es independiente del valor de cualquier otro dígito de la tabla. En la Tabla 4 del Apéndice se facilitan 1000 dígitos aleatorios.
- La mayoría de los programas de análisis estadístico contienen generadores de números aleatorios. Estos generadores producen grandes secuencias de dígitos pseudoaleatorios, que satisfacen aproximadamente las mismas propiedades de aleatoriedad enunciadas anteriormente.

> ***Ejemplo 4.2*** Supongamos que, en el ejemplo anterior, se dispone de una lista completa de los $N = 875$ ancianos institucionalizados en dicha población, de los cuales se desean seleccionar $n = 10$. La selección de una muestra aleatoria simple de este tamaño puede realizarse a partir de la Tabla 4 del Apéndice como sigue. Comenzando en cualquier lugar de esta tabla y leyendo grupos de 3 dígitos en cualquier dirección, seleccionar los 10 primeros números distintos entre 1 y 875. Por ejemplo, empezando en el primer dígito de la tercera fila y de izquierda a derecha, estos números son: 339, 117, 619, 68, 440, 788, 696, 716, 183 y 546. Notar que los números 897 y 898 han sido descartados por ser superiores a $N = 875$. La muestra aleatoria simple estaría así constituida por aquellos ancianos de la población numerados previamente por estos 10 valores.

Puede probarse que, como el muestreo aleatorio simple es un procedimiento equiprobabilístico, una media o una proporción poblacional se estiman simplemente mediante la media o proporción muestral. La estimación de parámetros poblacionales a partir de una muestra aleatoria simple, así como la varianza o error de las estimaciones, se discutirá en detalle al final de este tema.

4.2.2 Muestreo sistemático

En ocasiones, la numeración consecutiva de las unidades de la población y la posterior selección de una muestra aleatoria simple resultan muy laboriosas. En tales circunstancias, un procedimiento alternativo más sencillo es el llamado muestreo sistemático. Bajo este procedimiento, no siempre es necesario numerar previamente los elementos de la población, sino que basta con disponer de alguna ordenación explícita (por ejemplo, orden de archivo de historias clínicas o visitas sucesivas de pacientes a una consulta médica).

Para la selección de una muestra sistemática de tamaño n de una población de N unidades, se elige aleatoriamente un número de arranque r entre 1 y k, donde k es la parte entera de N/n, y a partir del elemento que ocupa el lugar r, se toman los restantes elementos en intervalos de amplitud k hasta completar la muestra deseada. Así, la muestra estará constituida por los elementos ordenados en los lugares r, $r + k$, $r + 2k$, ..., $r + (n - 1)k$. Como en general N no es múltiplo de n, este método de selección no es necesariamente equiprobabilístico (si N/n no es un número entero, las unidades comprendidas entre los lugares $nk + 1$ y N nunca podrán formar parte de la muestra). Una modificación a este procedimiento, que garantiza la obtención de una muestra equiprobabilística, consiste en seleccionar el número aleatorio de arranque r entre 1 y N, y tomar cada k-ésima unidad a partir de ahí, continuando en el primer elemento al alcanzar el final de la lista.

> ***Ejemplo 4.3*** Para seleccionar una muestra sistemática de tamaño $n = 10$ de la población de $N = 875$ ancianos institucionalizados, se calcula primero la amplitud del intervalo de selección como la parte entera de $N/n = 875/10 = 87{,}5$; es decir, $k = 87$. Si se seleccionara el número de arranque r entre 1 y 87, el último anciano seleccionado ocuparía en el lugar $r + (n - 1)k = r + (10 - 1)87 = r + 783$, que sería siempre inferior o igual a 870 (dado que $r \leq 87$). En consecuencia, los ancianos en los lugares 871 a 875 nunca podrían formar parte de la muestra. Para asegurar un muestreo equiprobabilístico, el número de arranque se selecciona aleatoriamente entre 1 y 875. Suponiendo que este número de arranque fue $r = 427$ y tomando intervalos de amplitud $k = 87$, la muestra sistemática quedaría integrada por aquellos ancianos en los lugares 427, 514, 601, 688, 775, 862, 74, 161, 248 y 335.

En el muestreo sistemático, la ordenación de los elementos de la población determinará las posibles muestras. En consecuencia, este orden ha de estar exento de cualquier periodicidad relacionada con las variables a estudio. Así, por ejemplo, si para estimar el nivel de contaminación atmosférica en una ciudad se toma una muestra sistemática de días con $k = 7$, la muestra estará formada por los mismos días de la semana y presentará un claro sesgo por falta de representatividad. No obstante, estas periodicidades son muy infrecuentes en la práctica y pueden solventarse con facilidad (en el ejemplo anterior, bastaría con utilizar un intervalo de selección distinto de 7). En general, si la ordenación de las unidades de la población es esencialmente aleatoria, la estimación de parámetros y sus correspondientes errores en un muestreo sistemático se realiza igual que en un muestreo aleatorio simple.

4.2.3 Muestreo estratificado

En los muestreos anteriores, las muestras se seleccionan por procedimientos puramente aleatorios. Así, si el tamaño muestral es suficientemente grande, la muestra será muy probablemente representativa de la población. Sin embargo, no existe una garantía absoluta de que la muestra finalmente seleccionada sea representativa para cualquier variable de interés. Cuando se desea asegurar la representatividad de determinados subgrupos o estratos de la población, la alternativa más sencilla es seleccionar por separado distintas submuestras dentro de cada estrato. Este procedimiento de selección se conoce como muestreo estratificado. Los **estratos** han de definir subgrupos de población que sean internamente homogéneos con respecto a la característica o parámetro de interés y, por tanto, heterogéneos entre sí. En la práctica, los estratos se definen en función de variables fáciles de medir previamente y relevantes para el tema objeto de estudio (por ejemplo, edad, sexo, raza o área geográfica de residencia). En general, el número de estratos ha de ser reducido (rara vez resulta eficiente utilizar más de 5 estratos) y el tamaño por estrato no debe ser muy pequeño.

Para la selección de una muestra estratificada de tamaño n, la población de N unidades se divide en K estratos de tamaños N_1, N_2, ..., N_K, cuya suma es igual a N. Los estratos son mutuamente excluyentes y exhaustivos, de tal forma que cada elemento de la población pertenece a uno y sólo a uno de los estratos. Una vez determinados estos estratos, se selecciona por separado una muestra de cada estrato de tamaño n_1, n_2, ..., n_K, respectivamente, cuya suma será igual al tamaño total n de la muestra. La selección dentro de cada estrato suele realizarse por muestreo aleatorio simple o sistemático, y el procedimiento se denomina entonces muestreo aleatorio estratificado.

En el muestreo estratificado, es necesario determinar cómo se distribuye el tamaño muestral total n entre los distintos estratos; es decir, la asignación de los tamaños muestrales n_1, n_2, ..., n_K. Aunque existen distintos tipos de asignación en función del tamaño y varianza por estrato (véase referencias al final del tema), nos limitaremos aquí a la asignación proporcional, que es el procedimiento utilizado con mayor frecuencia. En la **asignación proporcional**, la muestra total se reparte entre los estratos de forma proporcional al tamaño de cada estrato en la población. Así, como la proporción poblacional en cada estrato es N_k/N, el tamaño muestral del estrato k-ésimo será

$$n_k = n \frac{N_k}{N}.$$

Resulta inmediato probar que esta asignación da lugar a una muestra equiprobabilística.

Ejemplo 4.4 La capacidad funcional de los ancianos disminuye en gran medida con la edad. Supongamos que, de los $N = 875$ ancianos institucionalizados, se sabe que el 60% tienen menos de 75 años ($N_1 = 525$) y el restante 40% tienen 75 o más años ($N_2 = 350$). Para simplificar la exposición, supongamos además que los ancianos menores de 75 años corresponden a los primeros 525 números de la lista. Así, de los $n = 10$ ancianos seleccionados por muestreo aleatorio simple en el Ejemplo 4.2, la mitad resultaron ser mayores de 75 años. Esto es, por simple variabilidad aleatoria, los mayores de 75 años están ligeramente sobrerrepresentados en la muestra y, en consecuencia, la capacidad funcional media obtenida de esta muestra podría infraestimar la verdadera capacidad funcional de los ancianos institucionalizados. Para asegurar una mejor representatividad muestral por edad, podría realizarse un muestreo estratificado con asignación proporcional a ambos estratos de edad. Es decir, de la muestra de tamaño $n = 10$, seleccionaríamos 6 ancianos menores de 75 años ($n_1 = nN_1/N = 10 \cdot 0{,}6 = 6$) y 4 mayores de 75 años ($n_2 = nN_2/N = 10 \cdot 0{,}4 = 4$). Utilizando un muestreo aleatorio simple dentro de cada estrato, los 6 números seleccionados entre 1 y 525 fueron 505, 493, 24, 402, 371 y 265, y los 4 números seleccionados entre 526 y 875 fueron 851, 820, 717 y 696. La muestra estratificada proporcional estaría formada por los 10 ancianos correspondientes a dichos números.

Cabe reseñar aquí dos características importantes del muestreo estratificado. Por un lado, la asignación proporcional es la única que produce muestras equiprobabilísticas y, en consecuencia, la media y proporción poblacional se estiman mediante la media y la proporción muestral. Para cualquier otra asignación, la estimación de parámetros poblacionales requiere de la inclusión de pesos para cada observación muestral (típicamente, el inverso de la probabilidad de selección). Por otra parte, para un mismo tamaño muestral, el muestreo estratificado facilita estimaciones ligeramente más precisas (con menor error) que el muestreo aleatorio simple. Este resultado es debido a que, cuanto más homogéneos sean los estratos, más precisas serán las estimaciones en dichos estratos y esto redundará en una mayor precisión de las estimaciones para toda la población.

4.2.4 Muestreo por conglomerados

La aplicación de los diseños muestrales anteriores requiere de la enumeración u ordenación de todos los elementos de la población. Sin embargo, a menudo no se dispone de una lista completa o, aun disponiendo de tal lista, resulta muy costoso obtener información de las unidades muestreadas. Por ejemplo, si se seleccionara una muestra aleatoria simple de 1000 individuos de una gran ciudad, los individuos seleccionados estarían muy dispersos y la recogida de información sería extraordinariamente laboriosa. En tales circunstancias, una alternativa consiste en clasificar a la población en grupos o conglomerados, para así seleccionar una muestra de estos conglomerados y después tomar a todas o a una parte de las unidades incluidas dentro de los conglomerados seleccionados. Este método de selección se denomina muestreo por conglomerados y presenta dos ventajas fundamentales:

- Este muestreo es la única alternativa posible cuando no se dispone de una lista con todas las unidades de la población. En el muestreo por conglomerados, únicamente es necesario contar con listas de las unidades que integran los conglomerados seleccionados.
- Aun cuando otras técnicas de muestreo sean posibles, con frecuencia el muestreo por conglomerados resulta más económico, ya que las unidades muestrales están concentradas en los conglomerados seleccionados.

Notar que, a diferencia de la estratificación, donde interesa que los estratos sean lo más homogéneos posible, los **conglomerados** deben ser heterogéneos: en cada conglomerado debe haber unidades representativas de toda la población, de lo contrario se perdería información al seleccionar únicamente algunos de ellos. El número de conglomerados es típicamente elevado, de los cuales suele seleccionarse un número relativamente pequeño para resolver el problema de la dispersión muestral.

Supongamos que se pretende extraer una muestra de tamaño n a partir de una población de N unidades agrupadas en M conglomerados de tamaños N_1, N_2, ..., N_M. Entre los distintos métodos de selección por conglomerados, el **muestreo por conglomerados con probabilidad proporcional a su tamaño** resulta particularmente útil en la práctica. Para llevar a cabo este muestreo, se procede como sigue:

1. Ordenar arbitrariamente los conglomerados y calcular los tamaños acumulados. Estos tamaños acumulados delimitarán, para cada conglomerado, un rango de valores de amplitud igual a su tamaño poblacional.
2. Si se pretende seleccionar m conglomerados, extraer una muestra sistemática de tamaño m entre 1 y N. Los conglomerados seleccionados serán aquellos cuyo rango incluya alguno de los valores muestreados.
3. Dentro de cada conglomerado seleccionado, obtener una muestra aleatoria simple o sistemática de tamaño n/m.

Ejemplo 4.5 Con cualquiera de las técnicas de muestreo utilizadas en los ejemplos anteriores, la muestra incluiría muy probablemente ancianos institucionalizados en múltiples residencias, con el consiguiente inconveniente en la recogida de información. Supongamos que los $N = 875$ ancianos institucionalizados se encuentran distribuidos en $M = 15$ residencias con los tamaños especificados en la Tabla 4.1. Para optimizar el trabajo de campo, se decide extraer la muestra de tamaño $n = 10$ a partir de $m = 2$ residencias (conglomerados) seleccionadas con probabilidades proporcionales a sus tamaños.

Tabla 4.1 Distribución del número de ancianos institucionalizados por residencia.

Residencia (i)	Tamaño (N_i)	Tamaño acumulado	Rango asignado
1	50	50	1 – 50
2	30	80	51 – 80
3	35	115	81 – 115
4	70	185	116 – 185
5	55	240	186 – 240
6	45	285	241 – 285
7	125	410	286 – 410
8	80	490	411 – 490
9	20	510	491 – 510
10	100	610	511 – 610
11	65	675	611 – 675
12	35	710	676 – 710
13	40	750	711 – 750
14	75	825	751 – 825
15	50	875	826 – 875

En primer lugar, se asigna a cada residencia un rango de valores de amplitud igual a su tamaño (Tabla 4.1). A continuación, se extrae una muestra sistemática de tamaño 2 entre 1 y 875: si el número de arranque resultó ser 316, los valores muestreados son 316 y 753 (ver apartado de muestreo sistemático). Así, como el valor 316 está incluido dentro del rango asignado a la residencia 7 y el valor 753 en el rango de la residencia 14, resultan seleccionadas las residencias 7 y 14.

Para completar la muestra de n = 10 ancianos, se extraen finalmente muestras aleatorias simples de tamaño $n/m = 10/2 = 5$ de las residencias 7 y 14. De los 125 ancianos institucionalizados en la residencia 7, se seleccionaron los números 74, 23, 104, 111 y 57; y de los 75 ancianos de la residencia 14, los números 38, 51, 25, 34 y 41. En conclusión, la muestra total estará formada por los ancianos listados en los lugares 74, 23, 104, 111 y 57 de la residencia número 7, más aquellos que ocupan los lugares 38, 51, 25, 34 y 41 de la residencia número 14.

El muestreo por conglomerados con probabilidades proporcionales a sus tamaños facilita muestras equiprobabilísticas, así la media y la proporción poblacional pueden estimarse mediante sus correspondientes funciones muestrales. En general, para un tamaño muestral constante, la precisión de las estimaciones en un muestreo por conglomerados es menor que en un muestreo aleatorio simple. Las unidades de un mismo conglomerado suelen estar correlacionadas y, en consecuencia, aportan menos información que los elementos seleccionados de forma más dispersa mediante un muestreo aleatorio simple.

4.2.5 Muestreo polietápico

Los diseños muestrales empleados en la práctica se realizan combinando las técnicas descritas anteriormente. En muchas situaciones, resulta más apropiado obtener la muestra final en diferentes **etapas** o pasos. En un muestreo polietápico, la población se divide en grupos exhaustivos y mutuamente excluyentes, que constituyen las llamadas unidades de primera etapa; cada una de ellas se desagrega a su vez en subgrupos o unidades de segunda etapa, y así sucesivamente, hasta llegar en una última etapa a los elementos o unidades de análisis. La selección de unidades en cada una de las etapas se realiza mediante una técnica de muestreo diferente y la muestra final será la resultante de aplicar sucesivamente cada una de estas técnicas.

Ejemplo 4.6 En el ejemplo anterior se seleccionaron 2 de las 15 residencias y, dentro de cada residencia seleccionada, se eligieron a su vez 5 ancianos para formar la muestra definitiva. Este procedimiento de selección es, de hecho, un muestreo bietápico: las residencias constituirían las unidades de muestreo de primera etapa y los ancianos serían las unidades de muestreo de segunda etapa.

Una técnica de muestreo en etapas que se emplea con cierta frecuencia es el **muestreo estratificado polietápico**. Bajo esta técnica, las unidades de primera etapa se clasifican en distintos estratos y, dentro de cada estrato, se selecciona al menos una de sus unidades de primera etapa. La muestra final resultará de aplicar sucesivas etapas de muestreo dentro de las unidades de primera etapa seleccionadas en cada estrato. Este muestreo permite obtener una mayor representatividad muestral al seleccionar unidades dentro de todos los estratos.

Ejemplo 4.7 Supongamos que, de las 15 residencias listadas en la Tabla 4.1, las residencias 4, 7, 8, 10 y 14 son públicas, con un total de 450 ancianos (51,4%), y las restantes 10 residencias son privadas, con un total de 425 ancianos (48,6%). En el Ejemplo 4.5, las 2 residencias seleccionadas (7 y 14) fueron públicas; es decir, la muestra final no incluyó a ningún anciano institucionalizado en residencias privadas. Para garantizar la representatividad de los ancianos institucionalizados tanto en residencias públicas como privadas, bastaría con seleccionar una residencia de cada uno de estos estratos. En la Tabla 4.2, se muestran las 15 residencias reorganizadas según su carácter público o privado. Para las residencias públicas, se escogió aleatoriamente el número 20 entre 1 y 450, resultando así seleccionada la residencia 4, cuyo rango incluye dicho número. Para las residencias privadas, se extrajo aleatoriamente el número 326 entre 1 y 425, resultando seleccionada la residencia 12. A continuación, se procedería a escoger aleatoriamente 5 ancianos de estas 2 residencias. Notar que, como ambos estratos tienen aproximadamente el mismo tamaño, la muestra resultante sería equiprobabilística.

Apuntar, por último, que en la mayoría de los muestreos polietápicos el error muestral es sensiblemente superior al de un muestreo aleatorio simple, debido principalmente a la correlación entre los elementos que integran las unidades de primera etapa.

Tabla 4.2 Distribución del número de ancianos institucionalizados en residencias públicas y privadas.

Residencia (i)	Tamaño (N_i)	Tamaño acumulado	Rango asignado
Pública			
4	70	70	1 – 70
7	125	195	71 – 195
8	80	275	196 – 275
10	100	375	276 – 375
14	75	450	376 – 450
Privada			
1	50	50	1 – 50
2	30	80	51 – 80
3	35	115	81 – 115
5	55	170	116 – 170
6	45	215	171 – 215
9	20	235	216 – 235
11	65	300	236 – 300
12	35	335	301 – 335
13	40	375	336 – 375
15	50	425	376 – 425

4.3 ESTIMACIÓN EN EL MUESTREO ALEATORIO SIMPLE

Una vez descritas las principales técnicas de muestreo probabilístico, nos ocuparemos a continuación de la estimación de parámetros poblacionales. En adelante, se asume que la muestra se ha obtenido mediante un muestreo aleatorio simple a partir de una población de tamaño esencialmente infinito.

El cálculo del valor exacto de un **parámetro** poblacional requiere del conocimiento del valor de la variable objeto de estudio para todos y cada uno de los elementos de la población. Como se ha comentado anteriormente, en la mayoría de las ocasiones no se dispone de esta información, sino que se cuenta tan sólo con una muestra. A la función de los valores de una muestra que permite hacerse una idea acerca del valor del parámetro poblacional se le denomina **estimador**, y al resultado de aplicar dicha función a una determinada muestra se le llama **estimación**. Aún cuando el muestreo puede realizarse con múltiples propósitos, nos centraremos aquí en la estimación de una media y de una proporción poblacional.

4.3.1 Estimación puntual de una media poblacional

Supongamos que $x_1, x_2, ..., x_n$ son los valores obtenidos en una muestra de tamaño n para una variable con media poblacional μ y varianza σ^2 desconocidas. Un estimador natural de la media poblacional μ es la media muestral

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

Esta media muestral quedará completamente determinada una vez obtenida la muestra, pero el valor de la estimación variará en función de la muestra seleccionada. Así, la media muestral puede considerarse como una variable aleatoria, cuyo valor dependerá de la muestra finalmente seleccionada de entre todas las posibles muestras de tamaño n de la población de referencia. A la distribución de los valores de $\bar{x}$ sobre todas las posibles muestras del mismo tamaño se le denomina **distribución muestral** de $\bar{x}$. Las razones teóricas que justifican la utilización de la media muestral como estimador de la media poblacional, frente a otros posibles estimadores, se basan en esta distribución muestral.

A partir de los resultados del Apartado 3.4, el valor esperado de la distribución muestral de $\bar{x}$ es

$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \mu;$$

es decir, las medias muestrales de cualquier variable aleatoria están centradas alrededor de su verdadera media poblacional o, dicho de forma equivalente, las medias muestrales no sobreestiman ni infraestiman sistemáticamente la media poblacional. En términos estadísticos, se dice entonces que $\bar{x}$ es un **estimador centrado** o **insesgado** de μ. La conveniencia de utilizar estimadores insesgados parece clara ya que, en caso contrario, las estimaciones del parámetro poblacional estarían sistemáticamente sesgadas respecto a su verdadero valor. Otras medidas muestrales de tendencia central, como la mediana o la media geométrica, son en general estimadores sesgados de la media poblacional.

Ejemplo 4.8 Supongamos que el grupo control del estudio EURAMIC constituye toda la población o universo a estudio, cuya media poblacional del colesterol HDL es $\mu = 1{,}09$ mmol/l.

A partir de esta población, se obtienen 1000 muestras aleatorias simples de tamaño $n = 10$ y, en cada una de ellas, se calcula la media muestral $\bar{x}$ del colesterol HDL. El histograma de estas medias muestrales se representa en la Figura 4.1(*a*), que constituye una aproximación a la distribución muestral de $\bar{x}$. Como puede apreciarse, los valores de $\bar{x}$ difieren entre las distintas muestras, pero su distribución conjunta está centrada alrededor de la verdadera media poblacional $\mu = 1{,}09$ mmol/l (línea vertical en trazo discontinuo). En las Figuras 4.1(*b*) y (*c*) se presentan las distribuciones muestrales de la mediana y la media geométrica para estas mismas muestras. Ambas distribuciones muestrales presentan un claro sesgo respecto a la media poblacional, tendiendo a infraestimar su verdadero valor de 1,09 mmol/l.

Notar que el interés de este ejemplo es meramente académico ya que, en la práctica, se desconoce la verdadera media poblacional y se dispone de una única muestra.

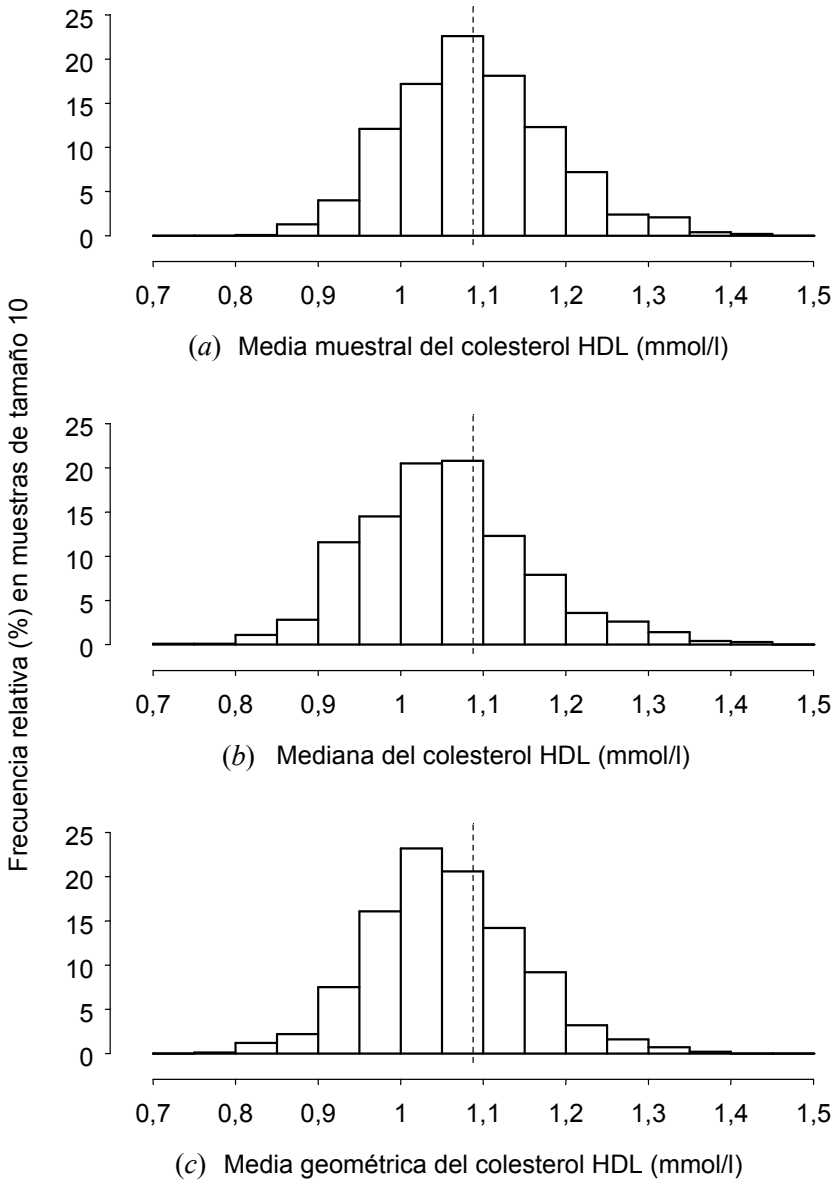


Figura 4.1 Distribución muestral de la media aritmética (*a*), la mediana (*b*) y la media geométrica (*c*) del colesterol HDL en 1000 muestras aleatorias simples de tamaño $n = 10$ obtenidas a partir del grupo control del estudio EURAMIC. La línea vertical en trazo discontinuo corresponde a la media poblacional $\mu = 1{,}09$ mmol/l de colesterol HDL.

4.3.2 Error estándar de la media muestral

Dado que la media muestral es un estimador insesgado de la media poblacional, todas las posibles medias muestrales estarán distribuidas alrededor de la media poblacional. No obstante, queda por determinar el grado de variabilidad o dispersión de estas medias muestrales alrededor de μ. La dispersión de las medias muestrales $\bar{x}$ de tamaño n vendrá determinada por la varianza de su distribución muestral, que es igual a

$$\text{var}(\bar{x}) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(x_i) = \frac{\sigma^2}{n},$$

dado que los distintos valores de la muestra son independientes (véase Apartado 3.4). Puede observarse que la variabilidad de las medias muestrales será tanto mayor cuanto mayor sea la varianza poblacional σ^2 de la variable a estudio. Por otra parte, esta variabilidad disminuye conforme aumenta el tamaño n de la muestra; es decir, al aumentar el tamaño muestral, las medias de las distintas muestras estarán más próximas a la verdadera media poblacional.

> ***Ejemplo 4.9*** En las Figuras 4.2(*a*), (*b*) y (*c*) se presentan las medias del colesterol HDL en 1000 muestras aleatorias simples de tamaño n = 10, 25 y 100, respectivamente, obtenidas a partir de los controles del estudio EURAMIC. En estas gráficas se puede apreciar que, independientemente del tamaño muestral, las medias muestrales están centradas alrededor de la media poblacional de 1,09 mmol/l. Sin embargo, al aumentar el tamaño muestral, se observa una disminución substancial de la variabilidad de las medias muestrales. Así, por ejemplo, la proporción de muestras con un nivel medio de colesterol HDL entre 1,03 y 1,15 mmol/l es del 48,7% para n = 10, 69,1% para n = 25 y 95,4% para n = 100.

Aun cuando en la práctica carece de sentido tomar repetidas muestras, las propiedades de la distribución muestral de $\bar{x}$ pueden utilizarse para cuantificar el error cometido en la estimación a partir de una única muestra de tamaño n. La desviación estándar de la distribución muestral de $\bar{x}$ es

$$SE(\bar{x}) = \sqrt{\text{var}(\bar{x})} = \frac{\sigma}{\sqrt{n}},$$

que facilita un valor promedio de la distancia de las distintas medias muestrales de tamaño n respecto de la medida poblacional. Esta cantidad $SE(\bar{x})$ se conoce como **error estándar** de la media muestral y permite cuantificar el grado de incertidumbre en la estimación de una media a partir de una muestra de tamaño n.

En la práctica, para poder calcular el error estándar, es necesario obtener previamente una estimación de la varianza poblacional σ^2 de la variable a estudio, dado que este parámetro es típicamente desconocido. La varianza poblacional σ^2 puede estimarse a partir de la propia muestra mediante la varianza muestral

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

Puede probarse que la varianza muestral es un estimador insesgado de la varianza poblacional; es decir, el valor esperado de s^2 sobre todas las posibles muestras es $E(s^2) = \sigma^2$. El error estándar de la media muestral se estima entonces como $s/\sqrt{n}$. Así, una vez seleccionada una muestra concreta, la media muestral $\bar{x}$ facilitará una estimación insesgada de la media poblacional y el error de dicha estimación vendrá determinado por $s/\sqrt{n}$.

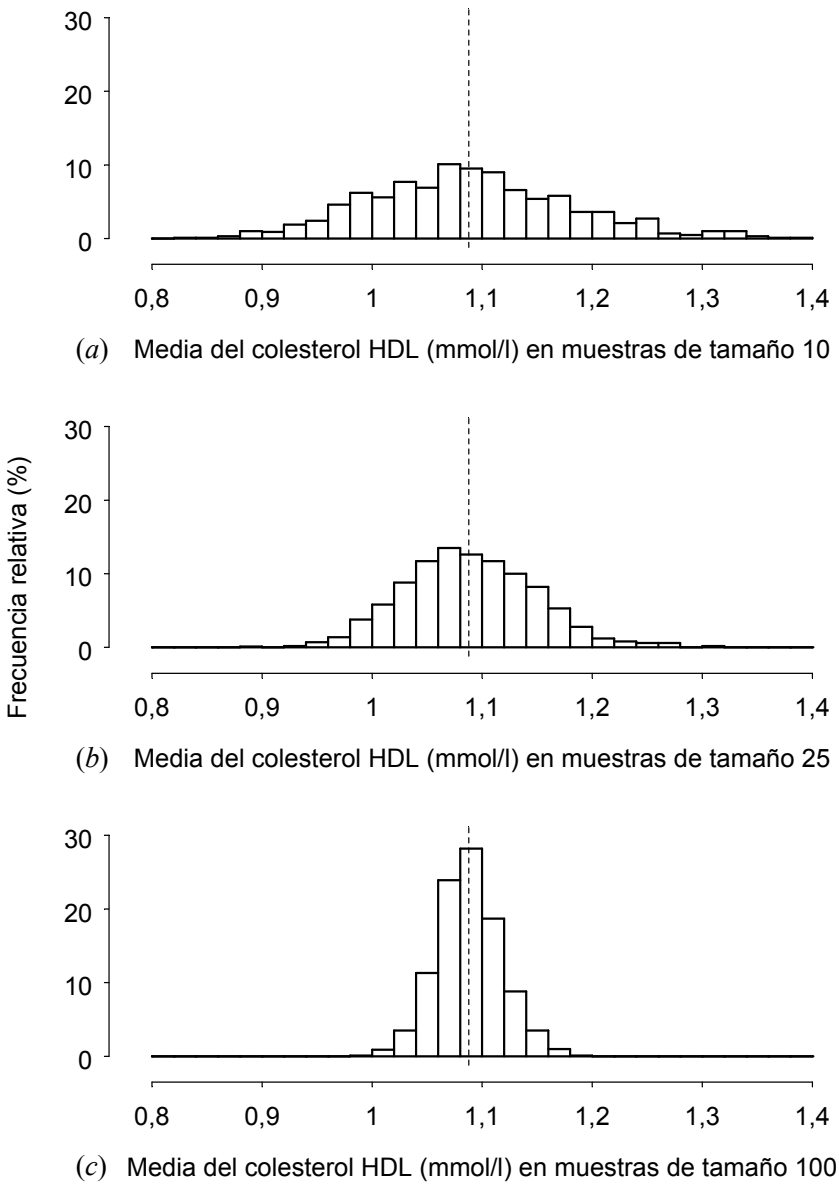


Figura 4.2 **Distribución muestral de la media del colesterol HDL en 1000 muestras aleatorias simples de tamaño $n = 10$ (*a*), 25 (*b*) y 100 (*c*) obtenidas a partir del grupo control del estudio EURAMIC. La línea vertical en trazo discontinuo corresponde a la media poblacional $\mu = 1{,}09$ mmol/l de colesterol HDL.**

Ejemplo 4.10 A partir de los controles del estudio EURAMIC, se ha obtenido una muestra aleatoria simple de tamaño $n = 10$, cuyos valores de colesterol HDL son 1,45, 1,32, 1,74, 0,82, 0,92, 1,46, 1,10, 0,88, 0,97 y 0,63 mmol/l. La media muestral es

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1{,}45 + 1{,}32 + \ldots + 0{,}63}{10} = 1{,}13 \text{ mmol/l}$$

y la varianza muestral

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$= \frac{(1{,}45 - 1{,}13)^2 + \ldots + (0{,}63 - 1{,}13)^2}{9} = 0{,}12\,(\text{mmol/l})^2.$$

Por tanto, la estimación puntual de la media poblacional del colesterol HDL es $\bar{x}$ = 1,13 mmol/l y su error estándar es

$$SE(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{0{,}35}{\sqrt{10}} = 0{,}11 \text{ mmol/l.}$$

Notar que, en este ejemplo ilustrativo, el error de la estimación muestral es exactamente $\bar{x} - \mu$ = 1,13 – 1,09 = 0,04 mmol/l. En la práctica, sin embargo, el error exacto no puede calcularse ya que μ es desconocido y, en consecuencia, se emplea $SE(\bar{x})$ como estimación del error promedio que cabría esperar en similares circunstancias (esto es, en todas las posibles muestras del mismo tamaño obtenidas de la población de referencia).

4.3.3 Teorema central del límite

En los apartados anteriores se ha probado que, para cualquier variable aleatoria, el valor esperado y la varianza de la distribución de las medias muestrales son μ y σ^2/n, respectivamente. No se ha analizado, sin embargo, el aspecto global de la distribución muestral de $\bar{x}$. Retomando el ejemplo de la distribución muestral de las medias de colesterol HDL (Figura 4.2), puede observarse que la forma de esta distribución tiende a aproximarse a una distribución normal conforme aumenta el tamaño muestral. Esta característica puede resultar intuitivamente lógica, ya que la distribución subyacente del colesterol HDL en la población presenta un aspecto aproximadamente normal (ver Figura 1.2 del Tema 1). Dado que muchas de las variables utilizadas en la práctica no presentan una distribución poblacional normal, cabría preguntarse si esta tendencia a la normalidad de la distribución muestral de $\bar{x}$ se mantiene para cualquier tipo de variable aleatoria.

Ejemplo 4.11 En la Figura 4.3 se muestra la distribución de los niveles de β-caroteno en tejido adiposo en el grupo control del estudio EURAMIC, que presenta una distribución marcadamente asimétrica con una media de μ = 0,37 µg/g. Las Figuras 4.4(*a*), (*b*) y (*c*)

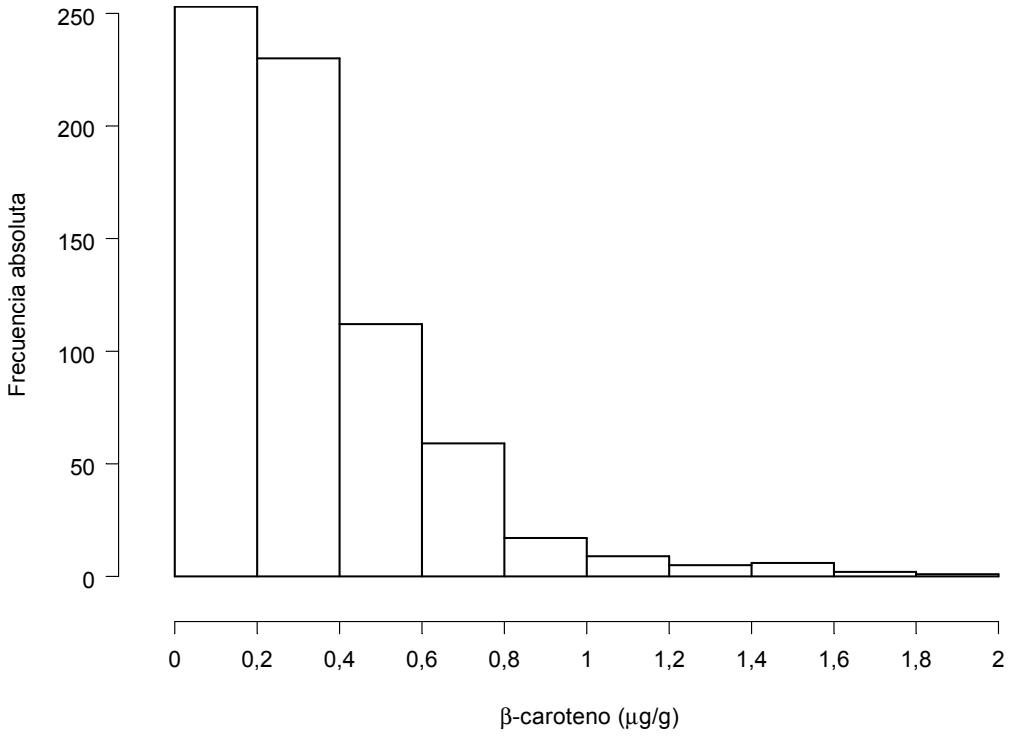


Figura 4.3 Distribución de frecuencias del nivel de β-caroteno en el grupo control del estudio EURAMIC.

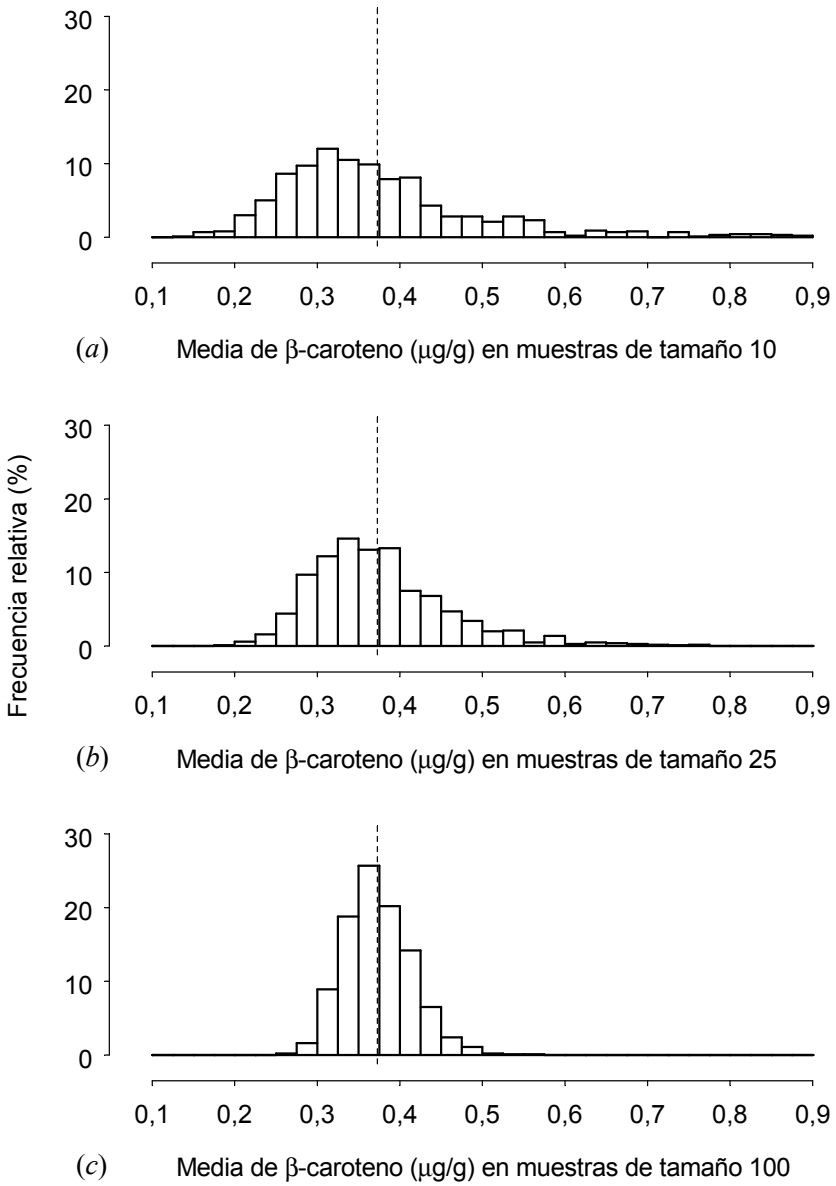


Figura 4.4 **Distribución muestral de la media de β-caroteno en 1000 muestras aleatorias simples de tamaño n = 10 (*a*), 25 (*b*) y 100 (*c*) obtenidas a partir del grupo control del estudio EURAMIC. La línea vertical en trazo discontinuo corresponde a la media poblacional μ = 0,37 µg/g de β-caroteno.**

representan las medias de β-caroteno en 1000 muestras aleatorias simples de tamaño n = 10, 25 y 100, respectivamente, obtenidas a partir de los controles del estudio EURAMIC. En estas gráficas puede observarse, de forma empírica, las siguientes propiedades:

— Ausencia de sesgo: para cualquier tamaño muestral, el promedio de las medias muestrales es similar a la media poblacional.

— Disminución del error estándar: al aumentar el tamaño muestral, disminuye la variabilidad en la distribución de las medias.

— Aproximación a la distribución normal: al aumentar el tamaño muestral, la distribución de las medias se aproxima a una distribución normal centrada en la media poblacional.

En los ejemplos anteriores, se ha comprobado de forma empírica que, independientemente de la forma de la variable aleatoria en la población, la distribución de las medias muestrales tiende a

seguir una distribución normal, particularmente cuando aumenta el tamaño de la muestra. Uno de los principales resultados en estadística, conocido como **teorema central del límite**, formaliza esta intuición: para cualquier variable aleatoria X con media μ y varianza σ^2, la distribución de las medias en muestras aleatorias simples de tamaño n se aproxima, al aumentar el tamaño muestral, a una distribución normal con media μ y varianza σ^2/n; es decir, al aumentar n,

$$\bar{x} \overset{\sim}{\to} N\left(\mu, \frac{\sigma^2}{n}\right),$$

donde el símbolo $\overset{\sim}{\to}$ significa "distribuirse aproximadamente como". Así, aun cuando la distribución de una variable en la población diste mucho de ser normal, el teorema central del límite permite utilizar la distribución normal como aproximación a la distribución de $\bar{x}$ si el tamaño muestral es suficientemente grande. Aunque el tamaño muestral necesario variará en función de la variable objeto de estudio, esta aproximación será razonablemente precisa siempre que n sea superior a 50.

Ejemplo 4.12 La media y la varianza del colesterol HDL en los controles del estudio EURAMIC son $\mu = 1{,}09$ mmol/l y $\sigma^2 = 0{,}086$ (mmol/l)^2. Por el teorema central del límite, la distribución de las medias en muestras de tamaño $n = 100$ será aproximadamente normal con media $\mu = 1{,}09$ mmol/l y varianza $\sigma^2/n = 0{,}086/100 = 0{,}00086$ (mmol/l)^2,

$$\bar{x} \overset{\sim}{\to} N(1{,}09,\ 0{,}00086).$$

Así, por ejemplo, la probabilidad de que la media de colesterol HDL en una muestra de tamaño $n = 100$ esté comprendida entre 1,03 y 1,15 mmol/l puede calcularse como

$$\begin{aligned} P(1{,}03 \le \bar{x} \le 1{,}15) &= P\left(\frac{1{,}03-1{,}09}{0{,}029} \le \frac{\bar{x}-1{,}09}{0{,}029} \le \frac{1{,}15-1{,}09}{0{,}029}\right) \\ &= P(-2{,}05 \le Z \le 2{,}05) \\ &= 2\Phi(2{,}05) - 1 = 0{,}9596. \end{aligned}$$

En el Ejemplo 4.9 se comprobó empíricamente que la proporción de muestras de tamaño $n = 100$ con un nivel medio de colesterol HDL entre 1,03 y 1,15 mmol/l es del 95,4%, que coincide casi perfectamente con el resultado obtenido bajo la aproximación normal.

Como se mostrará en los siguientes temas, el teorema central del límite constituye la base fundamental del proceso de inferencia estadística, dado que posibilita tanto la construcción de intervalos de confianza como el contraste de hipótesis acerca de la media poblacional μ.

4.3.4 Estimación de una proporción poblacional

Supongamos que el interés del estudio se centra en estimar la proporción π de individuos o elementos de la población que cumplen una determinada característica. En tal caso, resulta conveniente definir una variable aleatoria X que toma el valor 1 en los individuos que presentan dicha característica y 0 en quienes no la presentan. La media poblacional de esta variable aleatoria discreta es

$$\mu = \sum_{k=0}^{1} k\, P(X = k) = \pi$$

y su varianza

$$\sigma^2 = \sum_{k=0}^{1} (k-\pi)^2 P(X=k)$$
$$= \pi^2(1-\pi) + (1-\pi)^2\pi = \pi(1-\pi).$$

Si se selecciona una muestra aleatoria simple de tamaño n, en la cual k individuos presentan la característica de interés ($x_i = 1$) y los restantes $n - k$ individuos no la presentan ($x_i = 0$), el estimador natural de la proporción poblacional es la proporción muestral

$$p = \frac{k}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}\,.$$

A partir de esta notación, es evidente que una proporción muestral es un caso particular de una media muestral para una variable dicotómica con la codificación arriba indicada. Así, el teorema central del límite puede aplicarse a la forma particular de esta variable X para obtener el siguiente resultado: la distribución muestral de una proporción p se aproxima, al aumentar el tamaño muestral, a una distribución normal con media π y varianza $\pi(1-\pi)/n$,

$$p \overset{\sim}{\to} N\left(\pi, \frac{\pi(1-\pi)}{n}\right).$$

En consecuencia, pueden extraerse las siguientes propiedades de una proporción muestral:

- La proporción muestral p es un estimador insesgado de la proporción poblacional π; es decir, $E(p) = \pi$.
- La varianza muestral de p viene determinada por $\pi(1-\pi)/n$; así, al aumentar el tamaño muestral, las proporciones muestrales estarán más próximas a la verdadera proporción poblacional.
- Al aumentar el tamaño muestral, la distribución de las proporciones muestrales tiende a aproximarse a una distribución normal. Esta aproximación es suficientemente precisa si $n\pi(1-\pi) \geq 5$.

Ejemplo 4.13 En las Figuras 4.5(*a*), (*b*) y (*c*) se presentan las proporciones de fumadores actuales en 1000 muestras aleatorias simples de tamaño n = 10, 25 y 100, respectivamente, obtenidas a partir del grupo control del estudio EURAMIC, donde la proporción de fumadores actuales es $\pi = 0{,}37$. Para cualquier tamaño n de la muestra, las proporciones muestrales están distribuidas alrededor de la proporción poblacional (ausencia de sesgo). Al aumentar n, la distribución muestral de la proporción de fumadores actuales presenta una menor variabilidad y se aproxima a una distribución normal centrada en la proporción poblacional $\pi = 0{,}37$.

A partir de las propiedades anteriores se deduce que, para una muestra aleatoria de tamaño n, la proporción muestral p es un estimador insesgado de la proporción poblacional π y su error estándar viene determinado por la raíz cuadrada de la varianza muestral de p,

$$SE(p) = \sqrt{\text{var}(p)} = \sqrt{\frac{\pi(1-\pi)}{n}}\,,$$

que puede estimarse a partir de la propia muestra mediante $\sqrt{p(1-p)/n}$.

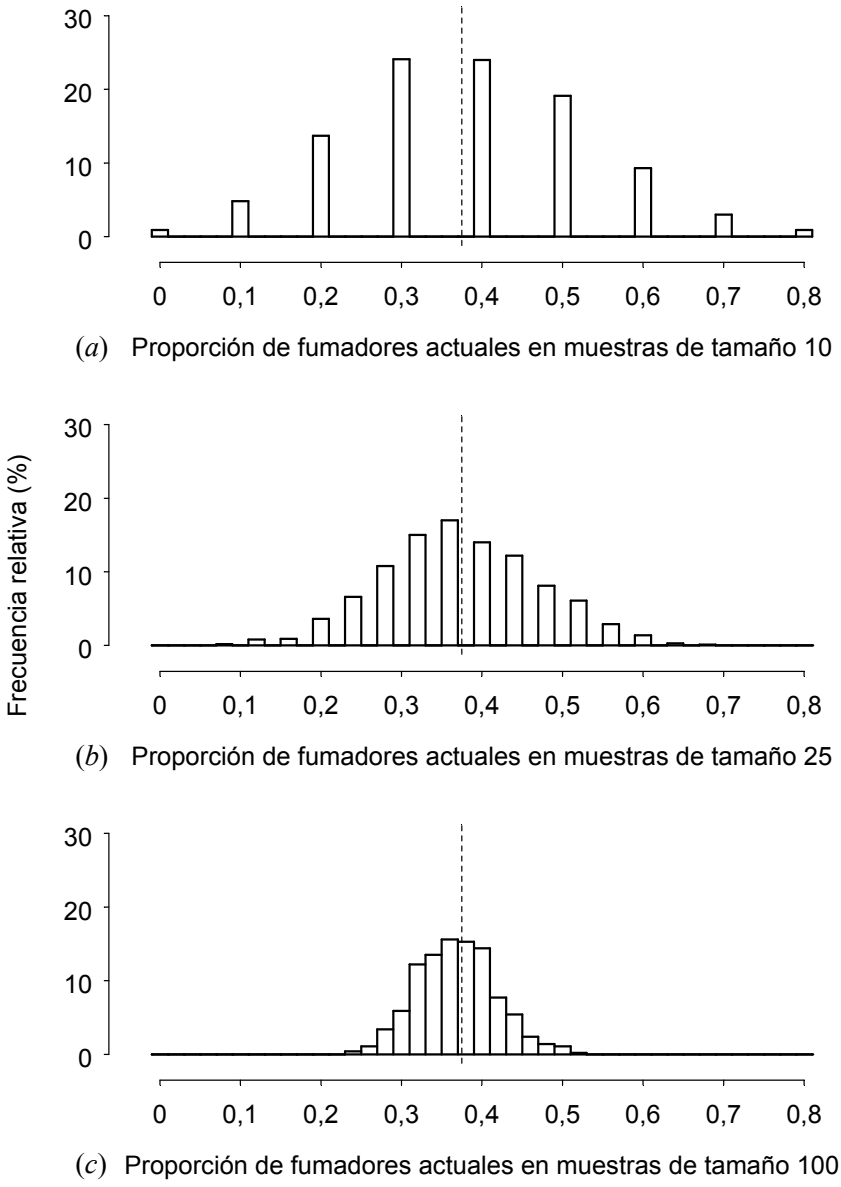


Figura 4.5 **Distribución muestral de la proporción de fumadores actuales en 1000 muestras aleatorias simples de tamaño $n = 10$ (*a*), 25 (*b*) y 100 (*c*) obtenidas a partir del grupo control del estudio EURAMIC. La línea vertical en trazo discontinuo corresponde a la proporción poblacional de fumadores actuales $\pi = 0{,}37$.**

Ejemplo 4.14 A partir de una muestra aleatoria simple de $n = 100$ controles del estudio EURAMIC, se obtuvieron $k = 35$ fumadores actuales. La estimación puntual de la proporción de fumadores actuales es

$$p = \frac{k}{n} = \frac{35}{100} = 0{,}35,$$

y su error estándar es

$$SE(p) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}35(1-0{,}35)}{100}} = 0{,}05,$$

que corresponde al error promedio que cabría esperar entre todas las posibles muestras de tamaño 100 de la población a estudio.

En este apartado se ha discutido la estimación puntual de una proporción poblacional π y su correspondiente error estándar. No obstante, no se ha hecho un uso práctico de la aproximación normal a la distribución muestral de p. Esta aproximación se retomará más adelante para obtener intervalos de confianza y pruebas de hipótesis sobre la proporción poblacional π (véase Tema 7).

4.4 REFERENCIAS

1. Bickel PJ, Doksum KA. *Mathematical Statistics: Basic Ideas and Selected Topics.* Englewood Cliffs, NJ: Prentice Hall, 1977.
2. Casella G, Berger RL. *Statistical Inference, Second Edition.* Belmont, CA: Brooks/Cole, 2001.
3. Cochran WG. *Sampling Techniques, Third Edition.* New York: John Wiley & Sons, 1977.
4. Kish L. *Survey Sampling.* New York: John Wiley & Sons, 1995.
5. Lehmann EL, Casella G. *Theory of Point Estimation, Second Edition.* New York: Springer Verlag, 1998.
6. Levy PS, Lemeshow S. *Sampling of Populations: Methods and Applications, Third Edition.* New York: John Wiley & Sons, 1999.
7. Rosner B. *Fundamentals of Biostatistics, Fifth Edition.* Belmont, CA: Duxbury Press, 1999.
8. Serfling RJ. *Approximation Theorems of Mathematical Statistics.* New York: John Wiley & Sons, 1980.
9. Silva LC. *Diseño Razonado de Muestras y Captación de Datos para la Investigación Sanitaria.* Madrid: Díaz de Santos, 2000.
10. Stuart A, Ord JK. *Kendall's Advanced Theory of Statistics, Volume 1, Distribution Theory, Sixth Edition.* London: Edward Arnold, 1994.

TEMA 5

INFERENCIA ESTADÍSTICA

5.1 INTRODUCCIÓN

La teoría del muestreo aporta diversos métodos formales para seleccionar muestras a partir de una determinada población. La información obtenida de dichas muestras puede resumirse utilizando técnicas de estadística descriptiva. Sin embargo, cuando se trabaja con una muestra, rara vez nos interesa la muestra como tal, sino que ésta interesa por su capacidad para aportar información con respecto a otros sujetos o a otras situaciones.

En los estudios descriptivos, el interés radica en seleccionar una muestra representativa de la población de referencia, o dicho más concretamente, la muestra ha de presentar el mismo grado de diversidad que la población respecto al parámetro o característica objeto de estudio. Las técnicas de muestreo probabilístico descritas en el tema anterior facilitan muestras que serán muy probablemente representativas de la población si el tamaño muestral es suficientemente grande. De esta forma, los resultados de la muestra podrán inferirse a toda población con un grado razonable de certidumbre.

> ***Ejemplo 5.1*** En las Encuestas Nacionales de Salud, se obtiene información de una muestra representativa a nivel provincial o nacional. Esta muestra interesa por la información que aporta sobre toda la población. En este caso, la representatividad de la muestra es determinante para la validez de las conclusiones derivadas del proceso inferencial.

En los estudios epidemiológicos analíticos, los resultados son interesantes porque pueden aplicarse a situaciones de salud semejantes. En este caso, el objetivo principal del diseño es asegurar la comparabilidad o semejanza de los grupos de estudio, más que la representatividad poblacional de la muestra. En los ensayos clínicos randomizados, los sujetos se asignan a los distintos grupos de tratamiento mediante algún mecanismo aleatorio (por ejemplo, mediante un muestreo aleatorio simple). Así, si el tamaño muestral es grande, las características basales de los sujetos asignados a los distintos grupos serán muy similares. En consecuencia, las diferencias observadas entre estos grupos a lo largo del seguimiento podrán atribuirse al tratamiento objeto de estudio.

> ***Ejemplo 5.2*** El primer ensayo clínico publicado sobre el papel de la aspirina en la prevención primaria de enfermedades cardiovasculares se realizó en médicos americanos participantes en el "*Physicians' Health Study*", seleccionados además por otras características de salud. En este caso, los sujetos a estudio no son representativos de la población a la que se aplicarán posteriormente los resultados (población general de hombres adultos a riesgo de padecer un primer evento cardiovascular), pero en cambio se garantizó la comparabilidad de las personas que tomaban aspirina y quienes no la tomaban mediante la asignación aleatoria del tratamiento y el uso de la técnica del doble ciego (tanto el investigador como el paciente desconocían el tratamiento asignado).

La estadística inferencial aporta las técnicas necesarias para extraer conclusiones sobre el valor poblacional de un determinado parámetro a partir de la evaluación de una única muestra.

Como se discutió en el tema anterior, las conclusiones derivadas de este proceso inferencial siempre estarán sujetas a error como consecuencia de la variabilidad aleatoria inherente al propio procedimiento de selección muestral. Por ello, resulta necesario disponer no sólo de una estimación puntual, sino también de un intervalo de confianza, que facilite un rango de valores verosímiles para el parámetro poblacional, así como de una prueba de significación estadística, que permita determinar el grado de compatibilidad de los datos muestrales con una hipótesis predeterminada. En este tema, se revisan los fundamentos y la interpretación de las técnicas estadísticas de inferencia: la estimación puntual, el intervalo de confianza y el contraste de hipótesis. Para simplificar la exposición, se asume que la muestra se obtiene por muestreo aleatorio simple y que la población de referencia es de tamaño muy superior a la muestra.

5.2 ESTIMACIÓN PUNTUAL

Una forma natural de estimar muchos parámetros poblacionales consiste en utilizar el estadístico muestral correspondiente. Así, la media muestral es un estimador puntual de la media poblacional y la proporción de casos de una enfermedad en la muestra es un estimador puntual de la probabilidad de tener la enfermedad en la población. No obstante, para un determinado parámetro poblacional, pueden contemplarse distintos estimadores alternativos. Algunos estimadores de la media poblacional distintos de la media muestral podrían ser, por ejemplo, la mediana, la media del 50% central de la muestra o la media de los valores máximo y mínimo. En este apartado se presentan algunos criterios estadísticos que justifican la elección de un determinado estimador frente a otras posibles alternativas.

Los méritos de un estimador no se juzgan por la estimación resultante en una muestra concreta, sino por la distribución de todos los posibles valores o estimaciones a que pueda dar lugar; esto es, por las propiedades de su distribución muestral. Entre las principales propiedades estadísticas que ha de satisfacer un buen estimador muestral cabe destacar las siguientes:

- **Ausencia de sesgo**. Un estimador es insesgado si su valor medio sobre todas las posibles muestras de tamaño n coincide con el parámetro poblacional. La insesgadez de un estimador es una propiedad deseable ya que sus estimaciones no diferirán sistemáticamente del parámetro poblacional.

Ejemplo 5.3 Como se probó en el tema anterior, la media y la proporción muestral son estimadores insesgados de la media y la proporción poblacional, respectivamente, $E(\bar{x}) = \mu$ y $E(p) = \pi$. Sin embargo, la varianza muestral definida por $\Sigma(x_i - \bar{x})^2/n$ es un estimador sesgado de la varianza poblacional, ya que

$$E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2\right)=\frac{1}{n}\sum_{i=1}^{n}E(x_i^2)-E\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}E(x_i^2)-\frac{1}{n^2}\left(\sum_{i=1}^{n}E(x_i^2)+2\sum_{1\le i<j\le n}E(x_i)E(x_j)\right)$$

$$=\frac{n-1}{n^2}\sum_{i=1}^{n}E(x_i^2)-\frac{2}{n^2}\sum_{1\le i<j\le n}E(x_i)E(x_j)$$

$$=\frac{n-1}{n}(\sigma^2+\mu^2)-\frac{n-1}{n}\mu^2=\frac{n-1}{n}\sigma^2;$$

es decir, este estadístico tiende a infraestimar la varianza poblacional σ^2 por un factor de $(n-1)/n$. Notar que este sesgo será tanto mayor cuanto menor sea el tamaño muestral. En consecuencia, es preferible utilizar la varianza muestral definida por $s^2 = \Sigma(x_i - \bar{x})^2/(n-1)$ como estimador insesgado de la varianza poblacional,

$$E(s^2) = E\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = \sigma^2.$$

- **Mínima varianza**. Además de la insesgadez de un estimador, que garantiza que las estimaciones estarán centradas alrededor del parámetro poblacional, interesa también que las distintas estimaciones difieran lo menos posible de dicho parámetro; es decir, que la varianza muestral del estimador sea mínima. De esta forma, se tendrá una mayor confianza en que la estimación resultante de la muestra finalmente seleccionada esté próxima al parámetro poblacional. Por ello, entre los distintos estimadores insesgados de un determinado parámetro, es conveniente seleccionar aquel que presente una menor varianza (o, de forma equivalente, un menor error estándar). En general, puede demostrarse que, si la distribución poblacional subyacente es normal, la media $\bar{x}$ y la varianza muestral s^2 son respectivamente los estimadores insesgados de μ y σ^2 con menor varianza. De la misma forma, la proporción muestral p es el estimador insesgado de π con menor error estándar.

  Ejemplo 5.4 Para cualquier distribución poblacional, la media muestral es un estimador insesgado de la media poblacional y su error estándar es

  $$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}.$$

  En el caso de que la distribución subyacente sea normal, puede probarse que la mediana también es un estimador insesgado de la media poblacional y que su error estándar es aproximadamente

  $$SE(\text{mediana}) \cong 1{,}25\frac{\sigma}{\sqrt{n}}.$$

  Así, aunque ambos estimadores son insesgados, el error estándar de la mediana es un 25% mayor que el de la media muestral y, por tanto, la mediana tenderá a facilitar estimaciones menos precisas que la media muestral.

- **Consistencia**. Las propiedades de insesgadez y mínima varianza se refieren a la distribución muestral del estimador para un tamaño n fijo de la muestra. La consistencia, sin embargo, hace referencia al comportamiento del estimador al aumentar n. Se dice que un estimador es consistente si, al aumentar el tamaño de la muestra, la probabilidad de que el estimador difiera del verdadero parámetro poblacional se reduce progresivamente. La consistencia es, por tanto, un requerimiento básico para un buen estimador ya que bastará con aumentar el tamaño muestral para obtener estimaciones arbitrariamente próximas al verdadero parámetro. Por supuesto, la media, la varianza y la proporción muestral son estimadores consistentes de sus respectivos parámetros poblacionales.

  Ejemplo 5.5 En el Ejemplo 4.9 se evaluó empíricamente el comportamiento de la media muestral de colesterol HDL en muestras de tamaño n = 10, 25 y 100 obtenidas a partir de los controles del estudio EURAMIC, donde la media poblacional del colesterol HDL

es μ = 1,09 mmol/l. La proporción de muestras con niveles medios de colesterol HDL próximos a μ = 1,09 mmol/l, pongamos por ejemplo entre 1,03 y 1,15 mmol/l, aumentó de un 48,7% para n = 10 a un 69,1% para n = 25 y a un 95,4% para n = 100. Este resultado corrobora empíricamente la consistencia de la media muestral como estimador de la media poblacional: la probabilidad de obtener estimaciones próximas al verdadero nivel medio aumenta progresivamente conforme aumenta el tamaño muestral.

En los problemas de estimación más simples, como es el caso de una media o una proporción poblacional, se dispone de un estimador natural que cumple las propiedades descritas anteriormente. En otros problemas más complejos, como por ejemplo en la estimación de parámetros en modelos de regresión, la elección de un estimador razonable no es tan directa. En general, existen diversos métodos formales para obtener estimadores con buenas propiedades estadísticas, entre los que destacan el método de máxima verosimilitud, el método de mínimos cuadrados y el método de los momentos. Los métodos de mínimos cuadrados y máxima verosimilitud se presentarán en el contexto particular de los modelos de regresión lineal (Temas 10 y 11) y logística (Tema 12), respectivamente. No obstante, los principios generales de estos procesos de estimación y la evaluación de los estimadores resultantes pueden consultarse en los textos de estadística matemática referenciados al final del tema.

5.3 ESTIMACIÓN POR INTERVALO

Como ya se ha comentado previamente, las estimaciones puntuales obtenidas a partir de una muestra diferirán del parámetro poblacional y, en consecuencia, quedará un margen de incertidumbre que se expresa en términos del error estándar del estimador. Así, resulta natural la pretensión de disponer de una medida del parámetro poblacional que incorpore tanto la estimación puntual como su error estándar. Esta medida es el intervalo de confianza, que facilita un rango de valores dentro del cual se encontrará el verdadero valor del parámetro poblacional con un cierto grado de confianza. En este apartado se describe detenidamente el procedimiento para la construcción de un intervalo de confianza para la media poblacional. Los principios básicos del cálculo e interpretación de intervalos de confianza para otros parámetros son similares y se discutirán en los siguientes temas.

5.3.1 Distribución *t* de Student

El método más extendido para el cálculo de intervalos de confianza se basa en las propiedades de la distribución muestral del estimador. Por el teorema central del límite sabemos que, para cualquier variable aleatoria con media μ y varianza σ^2, la distribución de las medias muestrales $\bar{x}$ es aproximadamente normal con media μ y varianza σ^2/n si el tamaño muestral es suficientemente grande; es decir,

$$\bar{x} \overset{\sim}{\to} N\left(\mu, \frac{\sigma^2}{n}\right)$$

o, de forma equivalente, aplicando la estandarización de una distribución normal

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \overset{\sim}{\to} N(0,\ 1).$$

Esta cantidad estandarizada depende de dos parámetros desconocidos: la media poblacional μ, que es el parámetro objeto de inferencia, y la desviación típica poblacional σ, que es un parámetro auxiliar necesario para conocer el error estándar en la estimación de μ. Parece entonces lógico sustituir en la expresión anterior el valor desconocido de σ por la desviación típica muestral s. Sin embargo, como s es un estimador de σ que conlleva a su vez un error de muestreo, el estadístico resultante $(\bar{x}-\mu)/(s/\sqrt{n})$ presentará una mayor imprecisión. Puede probarse que la distribución de este estadístico ya no será normal, sino que seguirá aproximadamente una distribución conocida como t de Student con $n-1$ grados de libertad y denotada por t_{n-1},

$$\frac{\bar{x}-\mu}{s/\sqrt{n}} \overset{\sim}{\rightarrow} t_{n-1}\,.$$

La distribución t de Student es una distribución simétrica alrededor de 0 y de aspecto parecido al de una distribución normal estandarizada, aunque menos apuntada en el centro y con más probabilidad en los extremos (Figura 5.1). Los grados de libertad de una distribución t de Student determinan su dispersión: al aumentar los grados de libertad, disminuye la variabilidad y la distribución t de Student se aproxima a una distribución normal estandarizada. Cuanto menor sea el tamaño muestral n, mayor será el error de la desviación típica muestral s y, en consecuencia, la distribución t de Student otorgará una mayor dispersión al estadístico $(\bar{x}-\mu)/(s/\sqrt{n})$. Por el contrario, si el tamaño muestral es grande, s facilitará una estimación precisa de σ, de tal forma que la distribución de dicho estadístico será aproximadamente normal. En la Tabla 5 del Apéndice se presentan los percentiles de la distribución t de Student para distintos grados de libertad.

> ***Ejemplo 5.6*** De la Tabla 5 del Apéndice se obtiene que el percentil 97,5 en una distribución t de Student con 2, 5, 10 y 30 grados de libertad es respectivamente $t_{2;0,975} = 4{,}303$, $t_{5;0,975} = 2{,}571$, $t_{10;0,975} = 2{,}228$ y $t_{30;0,975} = 2{,}042$. Por tratarse de distribuciones simétricas en 0, el percentil 2,5 coincide con el correspondiente percentil 97,5 con signo opuesto; es decir, $t_{2;0,025} = -4{,}303$, $t_{5;0,025} = -2{,}571$, $t_{10;0,025} = -2{,}228$ y $t_{30;0,025} = -2{,}042$. Por tanto, el 95% central de la distribución t de Student con 2, 5, 10 y 30 grados de libertad está comprendido entre ± 4,303, ± 2,571, ± 2,228 y ± 2,042, respectivamente. Así, puede observarse que la dispersión de la distribución t de Student disminuye al aumentar los grados de libertad, aproximándose a una distribución normal estandarizada (95% de los valores entre ± 1,96, Ejemplo 3.11).

5.3.2 Intervalo de confianza para una media poblacional

A partir de los resultados anteriores puede construirse un intervalo de confianza para la media poblacional. En general, la estimación por intervalo lleva asociada una probabilidad o **nivel de confianza**, denotada en términos porcentuales por $100(1-\alpha)\%$, que indica la cobertura del parámetro poblacional. Aunque en la práctica se utilizan casi exclusivamente los intervalos de confianza al 95% ($\alpha = 0{,}05$), nos referiremos aquí de forma genérica al intervalo de confianza al $100(1-\alpha)\%$ para la media poblacional. Utilizando la aproximación t de Student al estadístico $(\bar{x}-\mu)/(s/\sqrt{n})$, se sigue que hay una probabilidad $1-\alpha$ de que dicho estadístico esté

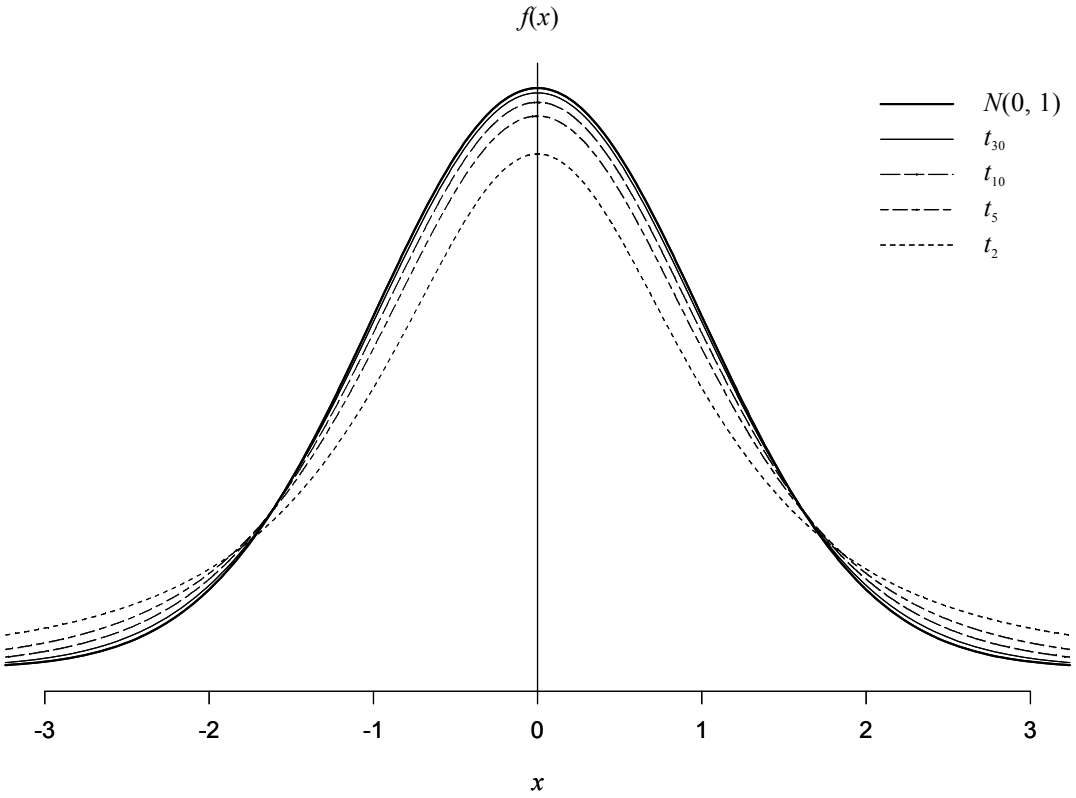


Figura 5.1 **Función de densidad de la distribución *t* de Student con 2, 5, 10 y 30 grados de libertad, y función de densidad normal estandarizada.**

comprendido entre los percentiles $\alpha/2$ y $1-\alpha/2$ de una distribución *t* de Student con $n-1$ grados de libertad, denotados respectivamente por $t_{n-1,\alpha/2}$ y $t_{n-1,1-\alpha/2}$; esto es,

$$P\left(t_{n-1,\alpha/2} < \frac{\bar{x}-\mu}{s/\sqrt{n}} < t_{n-1,1-\alpha/2} \right) = 1-\alpha \, .$$

Este resultado se representa gráficamente en la Figura 5.2. Por la simetría de la distribución *t* de Student, $t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}$ y la expresión anterior puede rescribirse como

$$P\left(-t_{n-1,1-\alpha/2} < \frac{\bar{x}-\mu}{s/\sqrt{n}} < t_{n-1,1-\alpha/2} \right) = 1-\alpha \, .$$

Para despejar la media poblacional, se multiplica cada término de la desigualdad por el error estándar $s/\sqrt{n}$ y a continuación se resta la media muestral $\bar{x}$, resultando que

$$P\left(\bar{x} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \right) = 1-\alpha \, .$$

Así, el intervalo de confianza (IC) al $100(1-\alpha)\%$ para la media poblacional viene determinado por

$$\bar{x} \pm t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}},$$

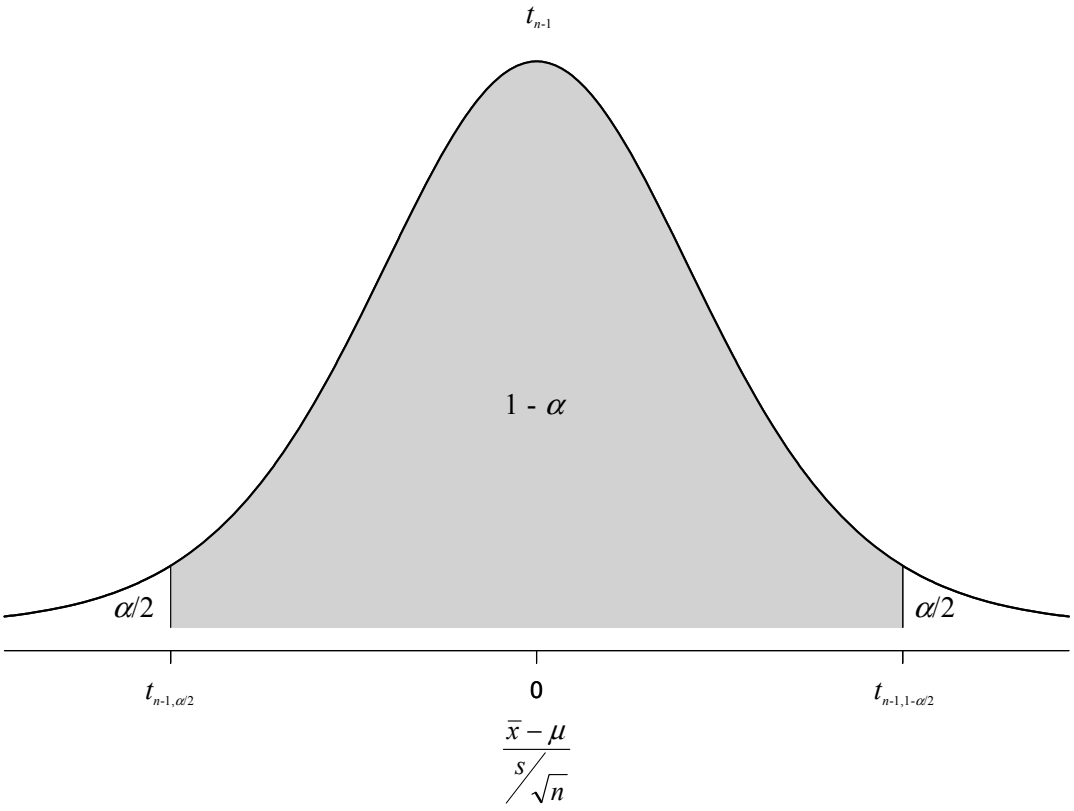


Figura 5.2 **Distribución muestral del estadístico $(\bar{x}-\mu)/(s/\sqrt{n})$.**

que depende tanto de la estimación puntual $\bar{x}$ (valor central del intervalo) como de su error estándar $s/\sqrt{n}$.

Los límites del intervalo están determinados por datos muestrales y, en consecuencia, el intervalo de confianza variará en función de la muestra seleccionada. El principio fundamental de la estimación por intervalo radica en que, de todas las posibles muestras del mismo tamaño de la población de referencia, el $100(1-\alpha)\%$ de los intervalos resultantes incluirá el parámetro poblacional. Así, aunque no es posible saber si efectivamente un intervalo concreto incluye o no el parámetro desconocido, se tendrá una confianza del $100(1-\alpha)\%$ en que el único intervalo disponible esté entre aquellos que contienen dicho parámetro. En otras palabras, el nivel de confianza de un intervalo hace referencia a la frecuencia con la cual el método produce intervalos certeros y no a la probabilidad de que el intervalo obtenido en una muestra concreta incluya el parámetro poblacional.

Ejemplo 5.7 En la Figura 5.3 se presentan los IC al 95% para la media poblacional del colesterol HDL en 100 muestras aleatorias de tamaño $n = 10$ obtenidas a partir de los controles del estudio EURAMIC. En cada una de las muestras, el IC al 95% se calculó como

$$\bar{x} \pm t_{9;0,975}\frac{s}{\sqrt{10}} = \bar{x} \pm 2{,}262\frac{s}{\sqrt{10}},$$

donde $\bar{x}$ y s son las correspondientes medias y desviaciones típicas muestrales. Así, por ejemplo, en la primera muestra se obtuvo $\bar{x} = 1{,}20$ y $s = 0{,}30$, de tal forma que la estimación puntual de la media poblacional de colesterol HDL resultó ser 1,20 mmol/l y su IC al 95% $1{,}20 \pm 2{,}262 \cdot 0{,}30/\sqrt{10} = (0{,}99; 1{,}41)$; es decir, a partir de esta muestra puede afirmarse con una confianza del 95% que la media poblacional del colesterol HDL se encuentra entre 0,99 y 1,41 mmol/l.

En este ejemplo ilustrativo, donde se conoce el verdadero valor de la media poblacional μ = 1,09 mmol/l, puede comprobarse empíricamente el significado del nivel de confianza al 95%: 94 de los 100 intervalos calculados contienen efectivamente la media poblacional, mientras que los 6 restantes no la contienen. Un IC particular puede o no incluir el parámetro y, por tanto, carece de sentido decir que hay una probabilidad del 95% de que μ se encuentre dentro de un intervalo concreto.

La estimación por intervalo facilita un rango de valores verosímiles o compatibles con la media poblacional μ, cuya amplitud depende de:

- El nivel de confianza 100(1 – α)%. Cuanto mayor sea la confianza deseada para un intervalo, mayor será la amplitud del mismo.

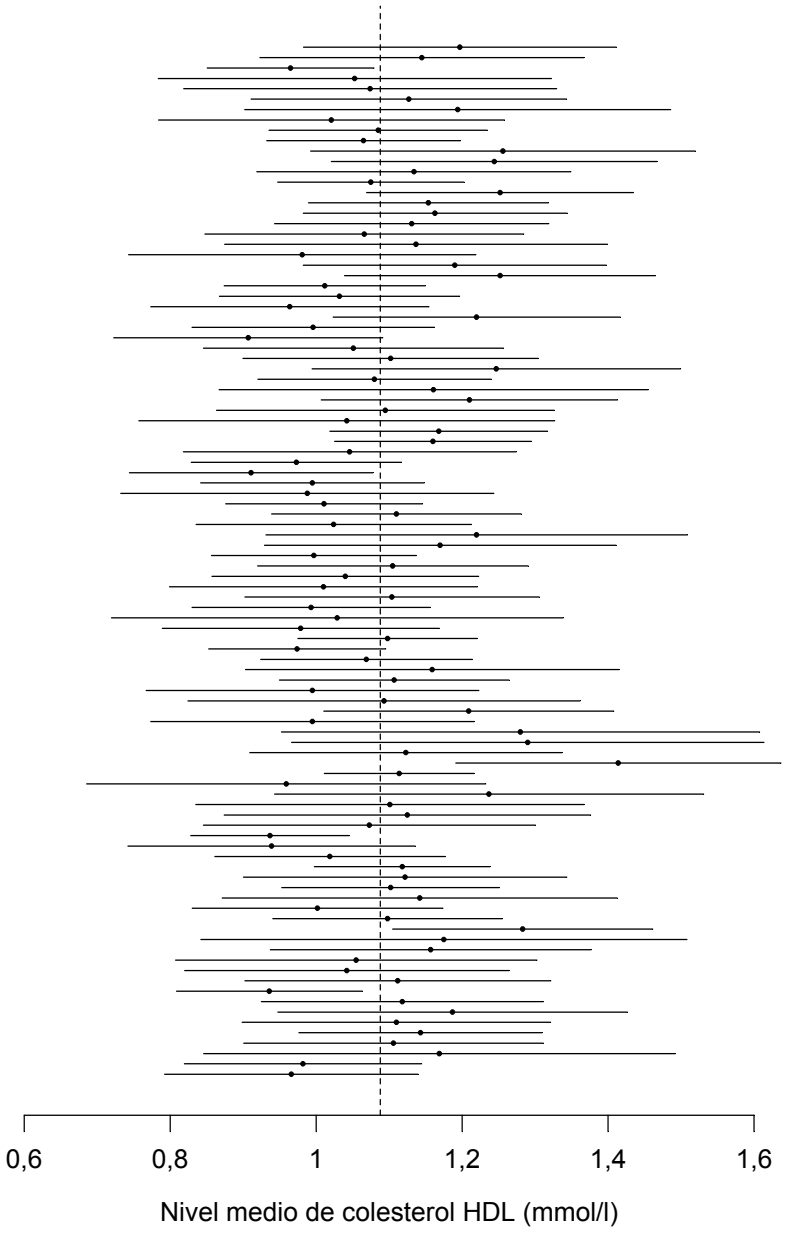


Figura 5.3 **Estimaciones puntuales (círculos) e intervalos de confianza al 95% (líneas horizontales) para la media poblacional del colesterol HDL en 100 muestras aleatorias de tamaño n = 10 obtenidas a partir de los controles del estudio EURAMIC. La línea vertical en trazo discontinuo corresponde al verdadero nivel medio μ = 1,09 mmol/l de colesterol HDL.**

Ejemplo 5.8 En la primera muestra del ejemplo anterior, el IC al 99% (α = 0,01) se calcularía como

$$\bar{x} \pm t_{9;0,995} \frac{s}{\sqrt{10}} = 1{,}20 \pm 3{,}250 \frac{0{,}30}{\sqrt{10}} = (0{,}89;\ 1{,}51);$$

esto es, la media poblacional del colesterol HDL se encuentra entre 0,89 y 1,51 mmol/l con una confianza del 99%. Notar que este intervalo es más amplio que el correspondiente intervalo al 95% (0,99; 1,41).

- El error estándar de la estimación $SE(\bar{x}) = s/\sqrt{n}$. Cuanto mayor sea el error de la estimación, mayor será la amplitud del intervalo. Es decir, la amplitud de un intervalo de confianza aporta una medida de la precisión de la estimación.

Ejemplo 5.9 En una muestra aleatoria de tamaño n = 100 de los controles del EURAMIC se obtuvo $\bar{x}$ = 1,09 y s = 0,31, resultando un IC al 95% para la media poblacional de

$$\bar{x} \pm t_{99;0,975} \frac{s}{\sqrt{100}} = 1{,}09 \pm 1{,}984 \frac{0{,}31}{10} = (1{,}03;\ 1{,}15).$$

Así, a partir de esta muestra de mayor tamaño, se concluye que la media poblacional del colesterol HDL se encuentra entre 1,03 y 1,15 mmol/l con un nivel de confianza del 95%. Este intervalo es mucho más preciso que los intervalos representados en la Figura 5.3 para muestras de tamaño n = 10.

Como se verá más adelante, el cálculo de los intervalos de confianza es similar para todos los parámetros. En general, el intervalo de confianza al 100(1 – α)% para un determinado parámetro poblacional se construye como

$$\text{estimador puntual} \pm x_{1-\alpha/2} SE,$$

donde $x_{1-\alpha/2}$ denota el percentil 1 – α/2 de la distribución muestral del estimador.

5.4 CONTRASTE DE HIPÓTESIS

En ocasiones, el interés de la investigación se centra no tanto en estimar un parámetro desconocido, sino en dilucidar si dicho parámetro es compatible con un valor predeterminado. A partir de conocimientos previos o mediante un razonamiento lógico, se pueden elaborar hipótesis o conjeturas sobre el fenómeno o parámetro objeto de estudio (por ejemplo, establecer la hipótesis de que la media de una población toma un valor determinado). La validez de estas hipótesis poblacionales ha de ser contrastada estadísticamente a partir de la información disponible en la muestra. Las técnicas que permiten evaluar el grado de compatibilidad de los datos muestrales con una hipótesis predeterminada se conocen genéricamente con el nombre de tests (pruebas o contrastes) de hipótesis.

5.4.1 Formulación de hipótesis

Los tests de hipótesis parten del planteamiento de una hipótesis nula, denotada por H_0, que representa el valor preestablecido del parámetro poblacional. Esta hipótesis nula se aceptará si los datos muestrales no aportan suficiente evidencia en contra de la misma. Por el contrario, si se cuenta con pruebas suficientes para contradecir la hipótesis nula, ésta se rechazará en favor de una hipótesis alternativa, denotada por H_1, que corresponde generalmente a la negación de la

hipótesis nula. En este punto, cabe incidir en que el término "aceptar" la hipótesis nula no implica que dicha hipótesis sea efectivamente cierta, sino que se carece de evidencia suficiente para rechazarla. Como se verá más adelante, las hipótesis nunca pueden ser corroboradas completamente, quedando siempre un margen o probabilidad de error.

Ejemplo 5.10 En un estudio para determinar la eficacia de un fármaco antihipertensivo, se compara la presión arterial de un grupo de pacientes tratados con dicho fármaco con la de un grupo de pacientes tratados con placebo. La hipótesis nula más natural, en este caso, es la hipótesis de no efecto del tratamiento; es decir, la presión arterial media de la población tratada con el fármaco μ_T es igual a la media de la población no tratada μ_P. La hipótesis alternativa sería, por el contrario, que las presiones arteriales medias de ambas poblaciones son distintas. Así, el contraste de hipótesis quedaría formulado como

$$H_0: \mu_T = \mu_P,$$
$$H_1: \mu_T \neq \mu_P.$$

La hipótesis nula se aceptará a no ser que los resultados del ensayo clínico muestren una gran diferencia entre los grupos que resulte poco compatible con una ausencia de efecto del tratamiento.

Supongamos hipotéticamente que el grupo control del estudio EURAMIC constituye la población a estudio. Para contrastar si la media poblacional del colesterol HDL μ es igual a un determinado valor, pongamos por ejemplo 1 mmol/l, el test de hipótesis se formularía como

$$H_0: \mu = 1,$$
$$H_1: \mu \neq 1.$$

La elección entre ambas hipótesis dependerá de los resultados obtenidos en una muestra de los controles del estudio EURAMIC.

En los ejemplos anteriores, se ha planteado una hipótesis alternativa **bilateral**; es decir, se aceptan como evidencia contra la hipótesis nula las diferencias en ambos sentidos. En algunas circunstancias, donde las desviaciones de la hipótesis nula en algún sentido carecen de importancia o son simplemente inconcebibles, es posible formular un contraste **unilateral**, aceptando como evidencia contra H_0 únicamente las diferencias en un sentido.

Ejemplo 5.11 En el estudio de la eficacia del fármaco antihipertensivo, se formuló una hipótesis alternativa bilateral $H_1: \mu_T \neq \mu_P$. En este caso, se admite que la evidencia en contra de la hipótesis nula puede provenir tanto por un efecto nocivo del tratamiento ($\mu_T > \mu_P$) como por la eficacia del mismo ($\mu_T < \mu_P$). Si en fases previas del ensayo clínico se ha comprobado la ausencia de efectos secundarios del tratamiento, la posibilidad de que la presión arterial media de los tratados sea superior a la media de los no tratados ($\mu_T > \mu_P$) carecería de sentido y sólo podría explicarse por variabilidad aleatoria. En tal caso, cabría plantearse el siguiente contraste de hipótesis unilateral

$$H_0: \mu_T = \mu_P,$$
$$H_1: \mu_T < \mu_P,$$

donde sólo se considera como alternativa a H_0 la posibilidad de que el tratamiento antihipertensivo sea eficaz.

Los contrastes bilaterales son más conservadores que sus correspondientes contrastes unilaterales, dado que aquellos contemplan desviaciones de H_0 en cualquier sentido. En la mayor parte de las aplicaciones prácticas se utilizan hipótesis alternativas bilaterales, ya que resulta imposible excluir con absoluta certeza diferencias en alguno de los dos sentidos. Así, todos los contrastes de hipótesis planteados a lo largo de este texto están basados en hipótesis alternativas bilaterales.

5.4.2 Contraste estadístico para la media de una población

En este apartado se discuten los conceptos básicos para la realización e interpretación de un contraste de hipótesis bilateral sobre la media de una población. Esto es, se pretende contrastar la hipótesis nula H_0: $\mu = \mu_0$ frente a la hipótesis alternativa bilateral H_1: $\mu \neq \mu_0$, donde μ_0 es un valor predeterminado de la media poblacional. El contraste de otros parámetros, así como la comparación de parámetros entre distintas poblaciones, se presentará en temas posteriores.

La elección entre las hipótesis nula y alternativa dependerá de los resultados obtenidos en la muestra o, más concretamente, de la compatibilidad de la media muestral $\bar{x}$ con el valor predeterminado μ_0. Como la media muestral es un estimador sujeto a error, el objetivo es determinar si la variabilidad inherente al muestreo constituye una explicación probable para la diferencia observada entre la media muestral $\bar{x}$ y el valor predeterminado μ_0 de la media poblacional. Para ello, se calcula la probabilidad de que bajo la hipótesis nula, una media muestral difiera tanto o más de μ_0 que el valor observado de $\bar{x}$. Esta probabilidad se conoce como **valor *P*** del contraste de hipótesis y determina el grado de compatibilidad de los datos muestrales con la hipótesis nula. Si este valor P es elevado, los datos muestrales serán compatibles con el valor μ_0 de la media poblacional, careciendo así de evidencia para rechazar la hipótesis nula. Por el contrario, si el valor P es pequeño, la media muestral resultará poco compatible con el valor preestablecido μ_0, concluyendo entonces que los datos aportan suficiente evidencia para rechazar dicha hipótesis. En general, cuanto menor sea el valor P, menos compatibles serán los datos con la hipótesis nula.

La decisión de rechazar la hipótesis nula se basa en la definición de un umbral preestablecido o **nivel de significación** α, tradicionalmente $\alpha = 0{,}05$. Si el valor P es inferior o igual que α se rechaza la hipótesis nula o, de forma equivalente, se afirma que los resultados son estadísticamente significativos; en caso contrario, si P es superior a α se acepta la hipótesis nula, concluyendo que los resultados del test no son estadísticamente significativos.

Para conocer el valor P del contraste es por tanto necesario calcular la probabilidad de que las medias de todas las posibles muestras de tamaño n difieran tanto o más de μ_0 que el valor observado de $\bar{x}$, asumiendo que la media poblacional es μ_0. Bajo la hipótesis nula H_0: $\mu = \mu_0$, las medias muestrales se distribuirán alrededor de μ_0, de tal forma que sus desviaciones estandarizadas

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

seguirán aproximadamente una distribución t de Student con $n - 1$ grados de libertad (Apartado 5.3.1). Una vez calculado el valor de este estadístico t a partir de los datos observados en la muestra, el valor P del contraste vendrá determinado por el área bajo la curva de la distribución t_{n-1} para aquellos valores tanto o más distantes de 0 que el valor observado de t (esto es, desviaciones de μ_0 mayores o iguales que la observada en cualquiera de los dos sentidos). En la Figura 5.4 se representa gráficamente el cálculo del valor P para este contraste de hipótesis.

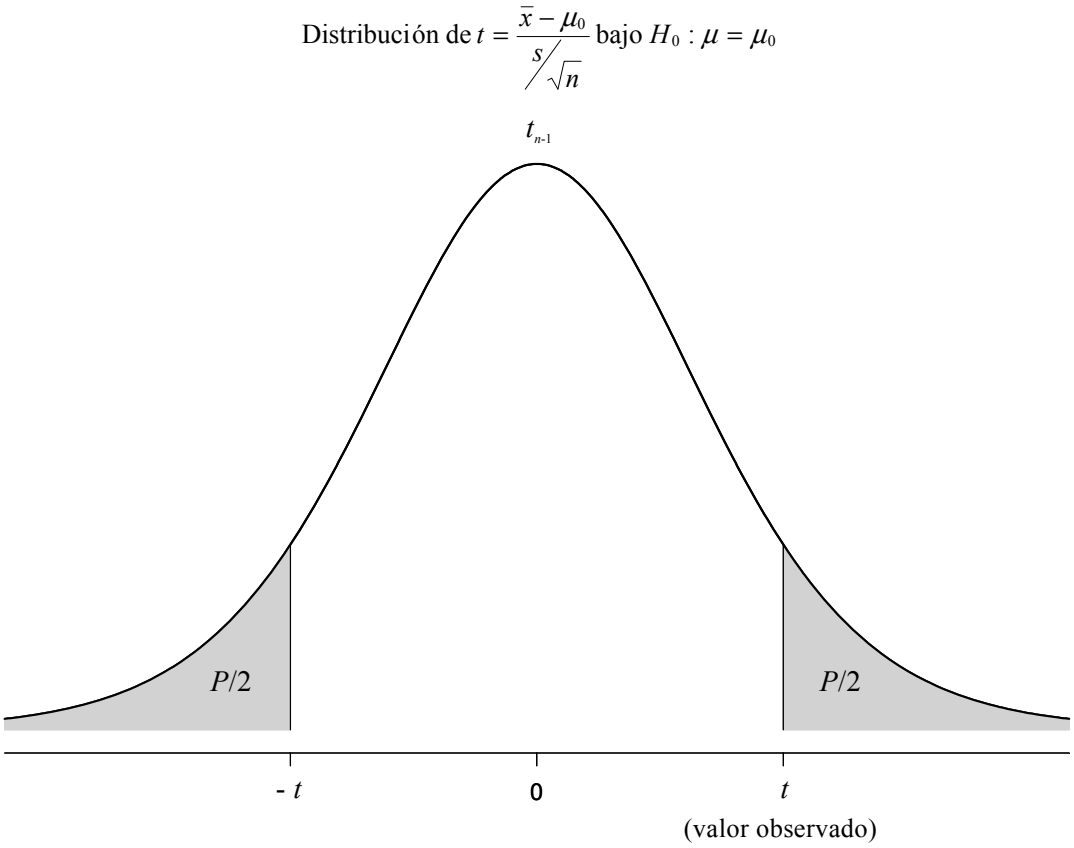


Figura 5.4 **Valor *P* para el contraste bilateral de la media de una población.**

Ejemplo 5.12 Supongamos que se pretende contrastar si la media poblacional del colesterol HDL en los controles del EURAMIC es igual a 1 mmol/l mediante el test de hipótesis bilateral

$$H_0: \mu = 1,$$
$$H_1: \mu \neq 1.$$

Para ello, se obtiene una muestra de tamaño $n = 10$ donde la media y desviación típica resultaron ser $\bar{x} = 1{,}20$ y $s = 0{,}30$ mmol/l. A partir de estos datos se calcula el estadístico del contraste

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1{,}20 - 1}{0{,}30/\sqrt{10}} = 2{,}11,$$

que determina la diferencia estandarizada (dividida por el error estándar) entre la media muestral $\bar{x}$ y el valor predeterminado μ_0. La distribución muestral de este estadístico bajo la hipótesis nula H_0: $\mu = 1$ seguirá aproximadamente una t de Student con 9 grados de libertad ($n - 1 = 10 - 1 = 9$). Así, si la hipótesis nula fuera cierta (esto es, si la verdadera media poblacional fuera 1 mmol/l), la probabilidad de obtener una muestra de 10 sujetos con una media de colesterol superior o igual a 1,20 mmol/l (mayor o igual desviación que la observada por la derecha) o inferior o igual a 0,80 mmol/l (mayor o igual desviación que la observada por la izquierda) sería

$$P = P(\bar{x} \geq 1{,}20 \mid H_0) + P(\bar{x} \leq 0{,}80 \mid H_0)$$

$$= P\left(\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \geq \frac{1{,}20 - \mu_0}{s/\sqrt{n}} \Big| H_0\right) + P\left(\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \leq \frac{0{,}80 - \mu_0}{s/\sqrt{n}} \Big| H_0\right)$$

$$\approx P(t_9 \geq 2{,}11) + P(t_9 \leq -2{,}11) = 2P(t_9 \geq 2{,}11) = 0{,}064,$$

que corresponde al área bajo la curva de la distribución t_9 para valores superiores a 2,11 (valor observado del estadístico) o inferiores a $-2{,}11$. Notar que el valor exacto de P se ha obtenido por ordenador. No obstante, utilizando la Tabla 5 del Apéndice, puede comprobarse que el estadístico $t = 2{,}11$ está comprendido entre los percentiles $t_{9;0,95} = 1{,}833$ y $t_{9;0,975} = 2{,}262$, de lo cual se deduce la desigualdad $0{,}025 < P(t_9 \geq 2{,}11) < 0{,}05$, que equivale a un valor P bilateral comprendido entre $0{,}05 < P < 0{,}10$.

Si se adopta el nivel de significación $\alpha = 0{,}05$ como regla de decisión, los resultados de esta muestra no aportan suficiente evidencia para rechazar la hipótesis nula ($P = 0{,}064 > 0{,}05$), concluyendo que la verdadera media poblacional del colesterol HDL no resulta significativamente distinta de 1 mmol/l.

El valor P determina la significación estadística de los resultados de un contraste de hipótesis, y depende tanto de la magnitud de la diferencia entre el verdadero valor del parámetro y su valor predeterminado bajo H_0, como del tamaño muestral. Así, una pequeña diferencia puede resultar estadísticamente significativa si el tamaño muestral es suficientemente grande y, por el contrario, una gran diferencia puede no alcanzar la significación estadística si la muestra es insuficiente. En consecuencia, el valor P no debe interpretarse como una medida de la magnitud de la diferencia o asociación objeto de estudio.

Ejemplo 5.13 En el ejemplo anterior se observó una diferencia en el colesterol HDL de 0,20 mmol/l entre el valor determinado bajo la hipótesis nula $\mu_0 = 1$ mmol/l y la media $\bar{x} = 1{,}20$ mmol/l en una muestra de tamaño $n = 10$. Los resultados del test no fueron estadísticamente significativos ($P = 0{,}064$) pero la magnitud de la diferencia podría ser clínicamente importante de confirmarse en estudios con mayor tamaño muestral.

Supongamos que se plantea el mismo contraste bilateral de la hipótesis nula H_0: $\mu = 1$ a partir de una muestra de tamaño $n = 100$ con media $\bar{x} = 1{,}09$ mmol/l y desviación típica $s = 0{,}31$ mmol/l. El estadístico del contraste es

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1{,}09 - 1}{0{,}31/10} = 2{,}90$$

y, por tanto, el valor P vendría determinado por

$$P = P(t_{99} \geq 2{,}90) + P(t_{99} \leq -2{,}90) = 2P(t_{99} \geq 2{,}90) = 0{,}005.$$

Utilizando la aproximación normal a la distribución t de Student con 99 grados de libertad, el valor P también puede aproximarse a partir de la Tabla 3 del Apéndice como

$$P = 2P(t_{99} \geq 2{,}90) \approx 2\{1 - \Phi(2{,}90)\} = 0{,}004.$$

En este caso, aunque la diferencia entre el valor predeterminado y la media muestral resultó ser sensiblemente menor (0,09 mmol/l), los resultados del test fueron

estadísticamente significativos ($P = 0{,}005$), aportando suficiente evidencia para rechazar la hipótesis nula.

La realización de una prueba de hipótesis presenta la misma estructura básica para todos los parámetros. En general, se calcula primero un estadístico del contraste, cuyo numerador corresponde a la diferencia entre el valor observado en la muestra y el valor esperado bajo la hipótesis nula, y cuyo denominador representa la variabilidad o error estándar de la estimación. El valor *P* se obtiene entonces como la probabilidad de obtener un valor del estadístico tanto o más extremo que el observado en el estudio, asumiendo que la hipótesis nula es cierta.

El contraste de hipótesis para un determinado parámetro está relacionado con su correspondiente intervalo de confianza. Si se contrasta la hipótesis nula H_0: $\mu = \mu_0$ frente a la hipótesis alternativa bilateral H_1: $\mu \neq \mu_0$, el resultado será estadísticamente significativo para un nivel $\alpha = 0{,}05$ si el IC al 95% para μ no incluye el valor μ_0. Por el contrario, este contraste no resultará estadísticamente significativo si el IC al 95% para μ contiene al valor μ_0. No obstante, ambos métodos facilitan información complementaria. El intervalo de confianza aporta una medida de la magnitud y precisión en la estimación del parámetro, aunque no facilita el valor exacto de *P* o el grado de compatibilidad con una hipótesis nula de interés. El valor *P* sí determina la compatibilidad de los datos con una determinada hipótesis, pero no facilita una medida de la magnitud del parámetro o asociación objeto de estudio. En general, el uso de los contrastes de hipótesis como forma exclusiva de presentar los resultados de un estudio está siendo ampliamente cuestionado en la actualidad. La presentación de los resultados de un estudio ha de consistir fundamentalmente en el estimador puntual y el intervalo de confianza, que pueden completarse con el valor *P* de la hipótesis correspondiente.

Ejemplo 5.14 En la primera muestra de tamaño $n = 10$ del Ejemplo 5.7 se obtuvo una media de 1,20 mmol/l y una desviación típica de 0,30 mmol/l, de tal forma que el IC al 95% para la media poblacional del colesterol HDL resultó ser (0,99; 1,41). Estos mismos datos muestrales se emplearon en el Ejemplo 5.12 para el contraste bilateral de la hipótesis nula H_0: $\mu = 1$, obteniendo un valor *P* de 0,064. Ambos resultados son consistentes dado que el IC al 95% incluye el valor preestablecido de 1 mmol/l para la hipótesis nula y, por tanto, el contraste no resulta estadísticamente significativo para un nivel $\alpha = 0{,}05$.

En el Ejemplo 5.9, a partir de una muestra de tamaño $n = 100$ con $\bar{x} = 1{,}09$ mmol/l y $s = 0{,}31$ mmol/l, se obtuvo un IC al 95% para la media poblacional del colesterol HDL de (1,03; 1,15). El correspondiente contraste de H_0: $\mu = 1$ frente a H_1: $\mu \neq 1$ se realizó en el Ejemplo 5.13, resultando un valor *P* de 0,005. En este caso, el valor 1 mmol/l queda fuera de los límites de confianza al 95% y, en consecuencia, los resultados del test son estadísticamente significativos.

5.4.3 Errores y potencia de un contraste de hipótesis

Como se comentó anteriormente, las hipótesis nunca pueden ser corroboradas completamente, quedando siempre un margen o probabilidad de error. La elección entre las hipótesis nula y alternativa conlleva a alguna de las situaciones presentadas en la Tabla 5.1. Si se acepta la hipótesis nula cuando ésta es cierta, o si se rechaza la hipótesis nula cuando la alternativa es cierta, se habrá tomado una decisión correcta. Sin embargo, es posible cometer alguno de los siguientes tipos de error en un contraste de hipótesis:

Tabla 5.1 Resultados posibles en un contraste de hipótesis.

	Realidad	
Decisión	**H_0 cierta**	**H_1 cierta**
Aceptar H_0	Correcto	Error de tipo II
Rechazar H_0	Error de tipo I	Correcto

- El **error de tipo I** consiste en rechazar la hipótesis nula cuando ésta es, en realidad, cierta. Como se comentó anteriormente, el nivel de significación α se utiliza para clasificar los resultados obtenidos en un test como significativos si el valor $P \leq \alpha$, en cuyo caso se rechaza la hipótesis nula, o como no significativos si $P > \alpha$, en cuyo caso se acepta la hipótesis nula. Con esta regla de decisión, puede comprobarse a partir de la Figura 5.4 que

$$\begin{aligned} P(\text{error de tipo I}) &= P(\text{rechazar } H_0 \mid H_0 \text{ cierta}) \\ &= P(t \geq t_{n-1,1-\alpha/2} \mid H_0 \text{ cierta}) + P(t \leq t_{n-1,\alpha/2} \mid H_0 \text{ cierta}) \\ &= P(t_{n-1} \geq t_{n-1,1-\alpha/2}) + P(t_{n-1} \leq t_{n-1,\alpha/2}) = \alpha/2 + \alpha/2 = \alpha; \end{aligned}$$

es decir, la probabilidad de cometer un error de tipo I viene determinada de antemano por el nivel de significación α. Así, por ejemplo, para un test con un nivel de significación $\alpha = 0{,}05$, la probabilidad de incurrir en un error de tipo I será del 0,05; esto es, si la hipótesis nula es cierta, ésta se rechazará erróneamente en un 5% de los contrastes de hipótesis realizados sobre todas las posibles muestras del mismo tamaño.

Ejemplo 5.15 A partir de los controles del EURAMIC se obtienen 1000 muestras aleatorias de tamaño $n = 10$ y, en cada una de ellas, se realiza el contraste de hipótesis bilateral para la media poblacional del colesterol HDL

$$H_0: \mu = 1{,}09,$$
$$H_1: \mu \neq 1{,}09,$$

mediante el estadístico

$$t = \frac{\bar{x} - 1{,}09}{s/\sqrt{10}},$$

donde $\bar{x}$ y s son las correspondientes medias y desviaciones típicas muestrales. En cada muestra, se calcula el valor P como el área bajo la curva de la distribución t_9 para valores tanto o más distantes de 0 que el valor observado de t, y se decide rechazar la hipótesis nula si $P \leq 0{,}05$. Así, la hipótesis nula se aceptó en un 94,4% de las muestras (944 de 1000) y se rechazó en un 5,6% (56 de 1000).

En este ejemplo ilustrativo, la hipótesis nula es cierta ya que la media poblacional del colesterol HDL en el grupo control del EURAMIC es efectivamente $\mu = 1{,}09$ mmol/l. Por lo tanto, se tomó la decisión correcta de aceptar H_0 en el 94,4% de las muestras y se rechazó erróneamente H_0 (error de tipo I) en el restante 5,6%, que concuerda casi perfectamente con el nivel de significación $\alpha = 0{,}05$ preestablecido para el contraste.

- El **error de tipo II** consiste en aceptar la hipótesis nula cuando, en realidad, es cierta la hipótesis alternativa. La probabilidad de cometer un error de tipo II se denota por β,

$$P(\text{error de tipo II}) = P(\text{aceptar } H_0 \mid H_1 \text{ cierta}) = \beta.$$

Si la hipótesis alternativa es cierta, la probabilidad de tomar la decisión correcta y, por tanto, rechazar la hipótesis nula se conoce como **potencia** del test,

$$\text{Potencia} = P(\text{rechazar } H_0 \mid H_1 \text{ cierta})$$
$$= 1 - P(\text{error de tipo II}) = 1 - \beta.$$

La probabilidad de error de tipo II β y la potencia de un contraste $1 - \beta$ no están predeterminadas de antemano y, como se comprobará a continuación, dependen de distintos factores, como el nivel de significación α, la desviación del verdadero valor del parámetro respecto al valor nulo $\mu - \mu_0$, la dispersión de los datos σ y el tamaño muestral n.

Supongamos, para simplificar la exposición, que una variable aleatoria tiene media desconocida μ y varianza conocida σ^2, y que se pretende contrastar la hipótesis nula H_0: $\mu = \mu_0$ frente a la hipótesis alternativa H_1: $\mu = \mu_1$, donde $\mu_1 \neq \mu_0$. Por el teorema central del límite, se sabe que la distribución muestral de $\bar{x}$ en muestras de tamaño n será aproximadamente $N(\mu_0, \sigma^2/n)$ si H_0 es cierta o, en caso contrario, $N(\mu_1, \sigma^2/n)$ si H_1 es cierta. La distribución muestral de $\bar{x}$ bajo las hipótesis nula y alternativa se representa en la Figura 5.5. Para un nivel de significación α, el contraste de hipótesis no resultará significativo ($P > \alpha$) si el estadístico

$$-z_{1-\alpha/2} < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha/2}$$

o, de forma equivalente, si

$$\mu_0 - z_{1-\alpha/2}\sigma/\sqrt{n} < \bar{x} < \mu_0 + z_{1-\alpha/2}\sigma/\sqrt{n};$$

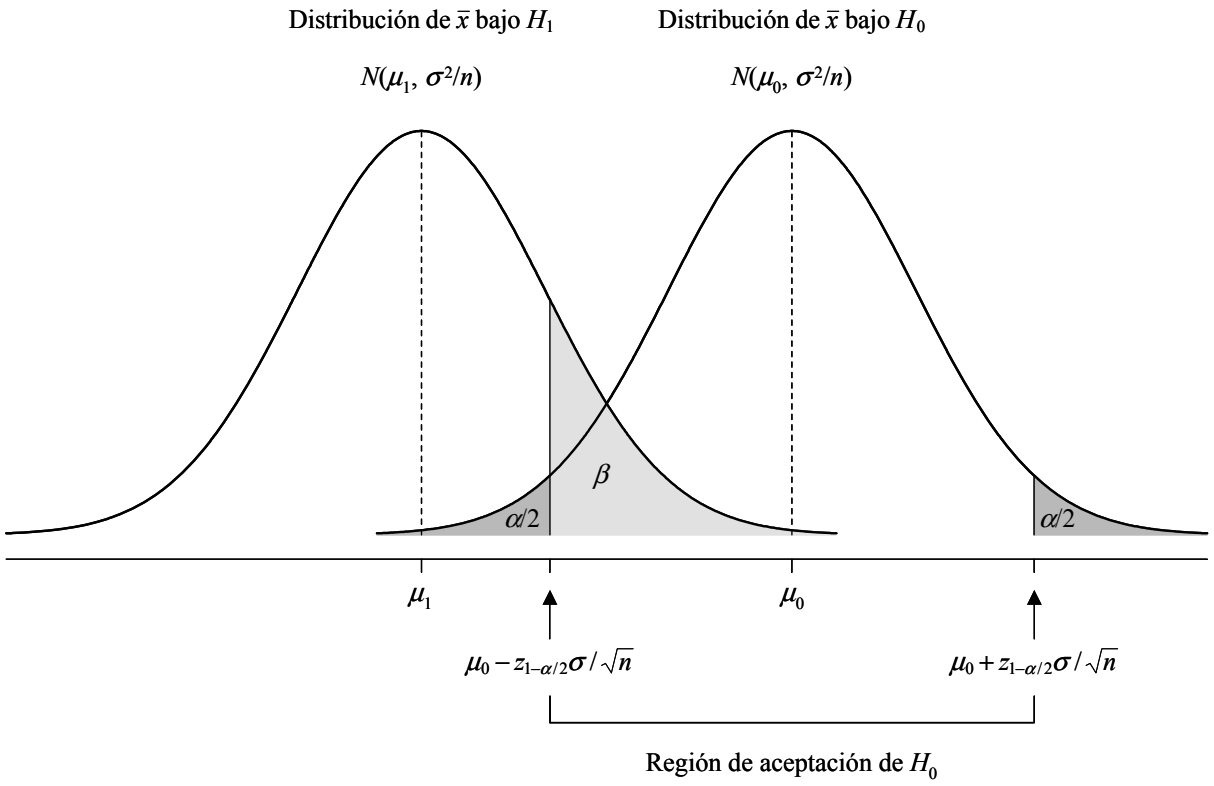
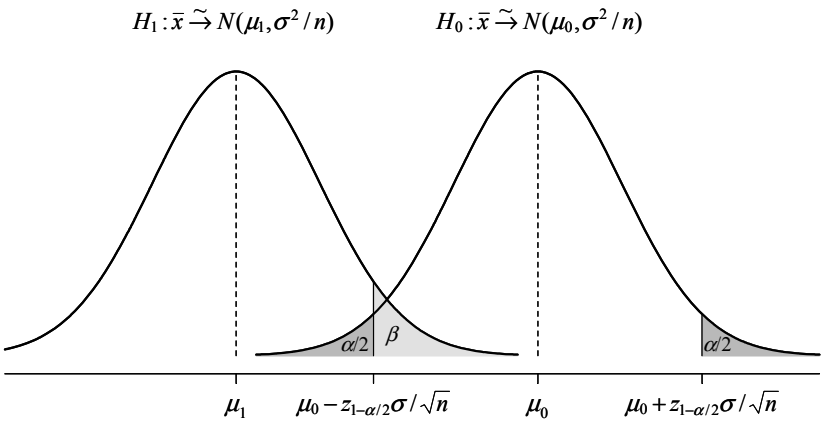


Figura 5.5 **Errores de tipo I y II para el contraste bilateral de la hipótesis nula H_0: $\mu = \mu_0$ frente a la hipótesis alternativa H_1: $\mu = \mu_1$ en una distribución con varianza conocida.**

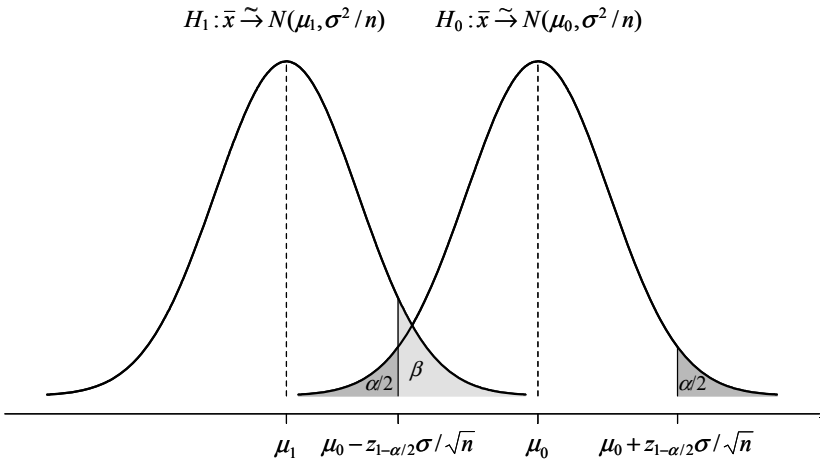
es decir, la hipotesis nula se aceptará en todas aquellas muestras con una media $\bar{x}$ comprendida en la región $\mu_0 \pm z_{1-\alpha/2}\sigma/\sqrt{n}$, que se denomina comúnmente como región de aceptación. Así, la probabilidad de un error de tipo I α está determinada por el área bajo la curva para H_0 situada fuera de la región de aceptación (área en gris oscuro de la Figura 5.5), y la probabilidad de error de tipo II β por el área bajo la curva para H_1 situada dentro de la región de aceptación (área en gris claro de la Figura 5.5).

El balance entre las probabilidades de un error de tipo I y tipo II puede observarse en la Figura 5.5. Si se reduce la probabilidad de error de tipo I α (esto es, se aumenta la región de aceptación), aumenta la probabilidad de error de tipo II β; mientras que si α aumenta, disminuye β. En la práctica, la estrategia habitual es fijar α en un nivel predeterminado (típicamente $\alpha = 0{,}05$) e intentar minimizar β o, de forma equivalente, maximizar la potencia $1 - \beta$ del contraste. Para α fijo, la potencia $1 - \beta$ depende de la superposición de las distribuciones nula y alternativa de $\bar{x}$, que está a su vez determinada por los siguientes factores:

- La diferencia subyacente $\mu_1 - \mu_0$. La potencia para detectar una hipótesis alternativa cierta será tanto mayor cuanto mayor sea la diferencia entre el verdadero valor del parámetro μ_1 y el valor nulo μ_0. Esta situación se ilustra en la Figura 5.6(a), donde se observa un incremento de la potencia como consecuencia de una mayor diferencia entre μ_1 y μ_0.



(a)



(b)

Figura 5.6 **Errores de tipo I y II para una mayor diferencia $\mu_0 - \mu_1$ (a) y para un mayor tamaño muestral n (b).**

Tabla 5.2 Porcentaje de muestras de tamaño n = 10, 25 y 100 con resultados significativos ($P \leq 0{,}05$) para el contraste bilateral de las hipótesis nulas H_0: μ = 1 y 1,05 mmol/l sobre la media poblacional del colesterol HDL en los controles del estudio EURAMIC.

	Hipótesis nula H_0: $\mu = \mu_0$	
Tamaño muestral (n)	$\mu_0 = 1$	$\mu_0 = 1{,}05$
10	11,2	5,0
25	26,9	8,0
100	85,7	23,0

- El error estándar $\sigma/\sqrt{n}$. Al aumentar el tamaño muestral n, disminuye el error estándar de la media muestral y, en consecuencia, la variabilidad de las distribuciones nula y alternativa de $\bar{x}$. Así, para un nivel de significación α predeterminado, la potencia del contraste aumenta conforme aumenta el tamaño de la muestra (Figura 5.6(b)). Esta relación puede utilizarse tanto para calcular la potencia de un contraste una vez determinado el tamaño muestral, como para estimar a priori el tamaño muestral necesario para una determinada potencia. Este último punto se discutirá con mayor detalle en el Tema 9 de determinación del tamaño muestral.

Ejemplo 5.16 A partir de los controles del EURAMIC se obtienen 1000 muestras aleatorias de tamaño n = 10, 25 y 100 y, en cada una de ellas, se realiza el contraste bilateral de las hipótesis nulas H_0: μ = 1 y 1,05 mmol/l para la media poblacional del colesterol HDL. Para cada muestra y contraste, el valor P se calcula según los métodos del Apartado 5.4.2 y la hipótesis nula se rechaza si $P \leq 0{,}05$. En la Tabla 5.2 se presenta el porcentaje de muestras con resultados significativos para los distintos tamaños muestrales e hipótesis nulas.

En este caso, ambas hipótesis nulas son falsas dado que la verdadera media del colesterol HDL en los controles del estudio EURAMIC es 1,09 mmol/l. Así, los porcentajes de la Tabla 5.2 representan valores empíricos de la potencia de cada contraste. Para una desviación subyacente de $\mu - \mu_0 = 1{,}09 - 1 = 0{,}09$ mmol/l entre el verdadero nivel medio de colesterol HDL y el valor nulo, la potencia resultó ser del 11,2% para n = 10, 26,9% para n = 25 y 85,7% para n = 100. Para una desviación de $\mu - \mu_0 = 1{,}09 - 1{,}05 = 0{,}04$ mmol/l, la potencia se redujo a un 5,0% para n = 10, 8,0% para n = 25 y 23,0% para n = 100. Como puede apreciarse, sólo se alcanza una potencia aceptable para detectar una diferencia de 0,09 mmol/l con un tamaño muestral de 100, mientras que sería necesaria una muestra mayor para poder detectar una diferencia de 0,04 mmol/l.

5.5 REFERENCIAS

1. Bickel PJ, Doksum KA. *Mathematical Statistics: Basic Ideas and Selected Topics.* Englewood Cliffs, NJ: Prentice Hall, 1977.
2. Casella G, Berger RL. *Statistical Inference, Second Edition.* Belmont, CA: Brooks/Cole, 2001.
3. Colton T. *Estadística en Medicina.* Barcelona: Salvat, 1979.

4. Lehmann EL. *Testing Statistical Hypotheses, Second Edition.* New York: Springer Verlag, 1997.
5. Lehmann EL, Casella G. *Theory of Point Estimation, Second Edition.* New York: Springer Verlag, 1998.
6. Rosner B. *Fundamentals of Biostatistics, Fifth Edition*. Belmont, CA: Duxbury Press, 1999.
7. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology, Third Edition.* Philadelphia: Lippincott Williams & Wilkins, 2008.
8. Snedecor GW, Cochran WG. *Statistical Methods, Eighth Edition.* Ames, IA: Iowa State University Press, 1989.
9. Stuart A, Ord JK, Arnold S. *Kendall's Advanced Theory of Statistics, Volume 2A, Classical Inference and the Linear Model, Sixth Edition.* London: Edward Arnold, 1999.

TEMA 6

INFERENCIA SOBRE MEDIAS

6.1 INTRODUCCIÓN

En el presente tema se revisan las técnicas básicas de inferencia a partir de datos de carácter cuantitativo. En la mayor parte de las ocasiones, la inferencia sobre variables cuantitativas se centra en el estudio de parámetros subyacentes tales como la media y la varianza poblacional. A partir de los datos obtenidos en muestras aleatorias y utilizando los principios de inferencia descritos en el tema anterior, se pretende dar respuesta a los siguientes tipos de problemas:

- La estimación de la media y la varianza de una población.

  Ejemplo 6.1 Supongamos que los controles del estudio EURAMIC constituyen una muestra representativa de la población de referencia del estudio. A partir de los valores de colesterol HDL obtenidos en los controles, ¿cuál es la estimación y el intervalo de confianza al 95% para la media y la varianza del colesterol HDL en la población de referencia? ¿Son estos datos muestrales compatibles con una verdadera media poblacional de 1 mmol/l?

- La comparación de medias y varianzas poblacionales a partir de dos muestras independientes.

  Ejemplo 6.2 En el estudio EURAMIC se comparan dos muestras independientes: una muestra de casos de infarto de miocardio, recogida de las unidades de cuidados intensivos, y una muestra independiente de controles, representativos de la población de la que proceden los casos. ¿Cuál es entonces la estimación y el intervalo de confianza al 95% para la diferencia en los niveles medios de colesterol HDL entre los casos de infarto y los sujetos libres de la enfermedad? ¿Es esta diferencia estadísticamente significativa?

  En un ensayo clínico para evaluar la eficacia antihipertensiva de un nuevo medicamento, se asignaron aleatoriamente 100 pacientes hipertensos a uno de los dos grupos de tratamiento: un grupo que toma la medicación a estudio y otro que toma un placebo. Después de 4 semanas de tratamiento, se compararon las medias de presión arterial sistólica entre ambos grupos como medida de la eficacia de dicho medicamento. ¿Cuál es la estimación puntual y el intervalo de confianza al 95% para la reducción en el nivel medio de presión arterial sistólica? ¿Cómo se determina si esta reducción es efecto del tratamiento o se debe a simple variabilidad aleatoria?

- La comparación de medias poblacionales a partir de dos muestras dependientes.

  Ejemplo 6.3 En un estudio de casos y controles sobre el efecto del colesterol HDL en el riesgo de desarrollar infarto de miocardio, cada caso se emparejó por grupo de edad y sexo a un control libre de la enfermedad. En este caso, las medias de colesterol HDL de los casos y de los controles no pueden analizarse como medidas procedentes de muestras independientes, ya que es esperable un cierto grado de correlación entre los valores de

colesterol HDL en cada pareja caso-control. ¿Cómo contrastar entonces si existe una asociación significativa entre el nivel de colesterol HDL y la ocurrencia de un infarto de miocardio?

Para evaluar la eficacia de un fármaco antihipertensivo, se seleccionaron 50 pacientes hipertensos y se administró a todos ellos dicho fármaco durante 4 semanas. La presión arterial sistólica de cada paciente se determinó tanto al comienzo del estudio como después de las 4 semanas de tratamiento. En tal caso, los valores medios de presión arterial antes y después del tratamiento no son independientes, ya que los datos recogidos en un mismo paciente están correlacionados. En estas circunstancias, ¿cómo estimar la reducción media de presión arterial sistólica al administrar dicho tratamiento?

Para cada uno de estos problemas, se facilitan las técnicas de inferencia apropiadas para obtener estimaciones puntuales y por intervalo del parámetro poblacional objeto de estudio, así como para el contraste de hipótesis preestablecidas. Estos procedimientos van a permitir inferir los resultados del estudio al ámbito poblacional de forma clara y sucinta.

6.2 INFERENCIA SOBRE UNA MEDIA Y VARIANZA POBLACIONAL

La media y la varianza poblacional son parámetros que representan la tendencia central y dispersión de la distribución subyacente de una variable aleatoria. Estos parámetros son típicamente desconocidos y, en consecuencia, han de ser estimados a partir de los valores observados de dicha variable en una muestra. En esta sección, se presentan los métodos de estimación y contraste para la media y la varianza de una distribución poblacional.

6.2.1 Inferencia sobre la media de una población

La estimación e inferencia de una media poblacional μ se discutió en el tema anterior. Para cualquier variable aleatoria, se ha comprobado que la media muestral $\bar{x}$ es un estimador insesgado y consistente de μ y que, en el caso de distribuciones normales, es el estimador con menor error estándar. Estas características hacen de la media muestral un buen estimador puntual de la media poblacional.

Utilizando las propiedades de la distribución muestral de la media, es posible obtener un intervalo de confianza al $100(1-\alpha)\%$ para la media poblacional μ como

$$\bar{x} \pm t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}.$$

A su vez, el contraste de la hipótesis nula H_0: $\mu = \mu_0$ frente a la hipótesis alternativa bilateral H_1: $\mu \neq \mu_0$ puede realizarse mediante el estadístico

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

Bajo la hipótesis nula, este estadístico seguirá aproximadamente una distribución t de Student con $n-1$ grados de libertad y, en consecuencia, el valor P del contraste puede calcularse como el área bajo la curva de esta distribución para aquellos valores tanto o más distantes de 0 que el valor observado de t. En general, el planteamiento de una determinada hipótesis nula puede proceder de estudios previos o de hipótesis biológicas respecto al comportamiento de las

variables, aunque en el caso de una única media poblacional los contrastes de hipótesis pueden resultar un tanto artificiales.

Ejemplo 6.4 Entre los $n = 539$ controles del estudio EURAMIC con determinaciones del colesterol HDL, la media y desviación típica fueron $\bar{x} = 1{,}09$ y $s = 0{,}29$ mmol/l. Así, el IC al 95% para la media de colesterol HDL en la población de referencia resultó ser

$$1{,}09 \pm t_{538;0{,}975}\,\frac{0{,}29}{\sqrt{539}} = 1{,}09 \pm 1{,}96 \cdot 0{,}012 = (1{,}07;\ 1{,}11).$$

Estos datos muestrales también se emplearon para el contraste bilateral de la hipótesis nula H_0: $\mu = 1$. Para ello, se calculó el estadístico del contraste

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1{,}09 - 1}{0{,}29/\sqrt{539}} = 7{,}21,$$

cuya distribución bajo la hipótesis nula será t_{538} o, de forma equivalente, normal estandarizada. De la Tabla 3 del Apéndice se desprende que la probabilidad de obtener valores superiores a 7,21 en una distribución normal estandarizada es virtualmente nula, por lo que el valor P bilateral será inferior a 0,001. En conclusión, el nivel medio de colesterol HDL en esta población difiere significativamente de 1 mmol/l ($P < 0{,}001$). De hecho, la media poblacional de colesterol HDL se estimó en 1,09 mmol/l, con un intervalo de confianza al 95% comprendido entre 1,07 y 1,11 mmol/l.

6.2.2 Inferencia sobre la varianza de una población

En ocasiones, el interés se centra en estimar no sólo la media de una variable aleatoria continua, sino también su varianza poblacional. Como se mostró en el Apartado 5.2 del tema anterior, la varianza muestral s^2 es un estimador insesgado y consistente de la varianza poblacional σ^2 de cualquier variable aleatoria, siendo además el estimador insesgado con menor error estándar para distribuciones normales.

Al igual que ocurría en el caso de una media, los intervalos de confianza y las pruebas de hipótesis sobre la varianza poblacional σ^2 se basan en la distribución muestral de s^2. Si la distribución subyacente de la variable es normal, puede probarse que el estadístico $(n - 1)s^2/\sigma^2$ sigue una distribución denominada **chi-cuadrado** con $n - 1$ grados de libertad y denotada por χ^2_{n-1},

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Como puede apreciarse en la Figura 6.1, la distribución chi-cuadrado sólo toma valores positivos y está sesgada a la derecha. Los grados de libertad de una distribución chi-cuadrado determinan su tendencia central, dispersión y asimetría: al aumentar los grados de libertad, aumenta la media y la varianza de la distribución y disminuye su sesgo a la derecha. En la Tabla 6 del Apéndice se presentan los percentiles de la distribución chi-cuadrado para distintos grados de libertad.

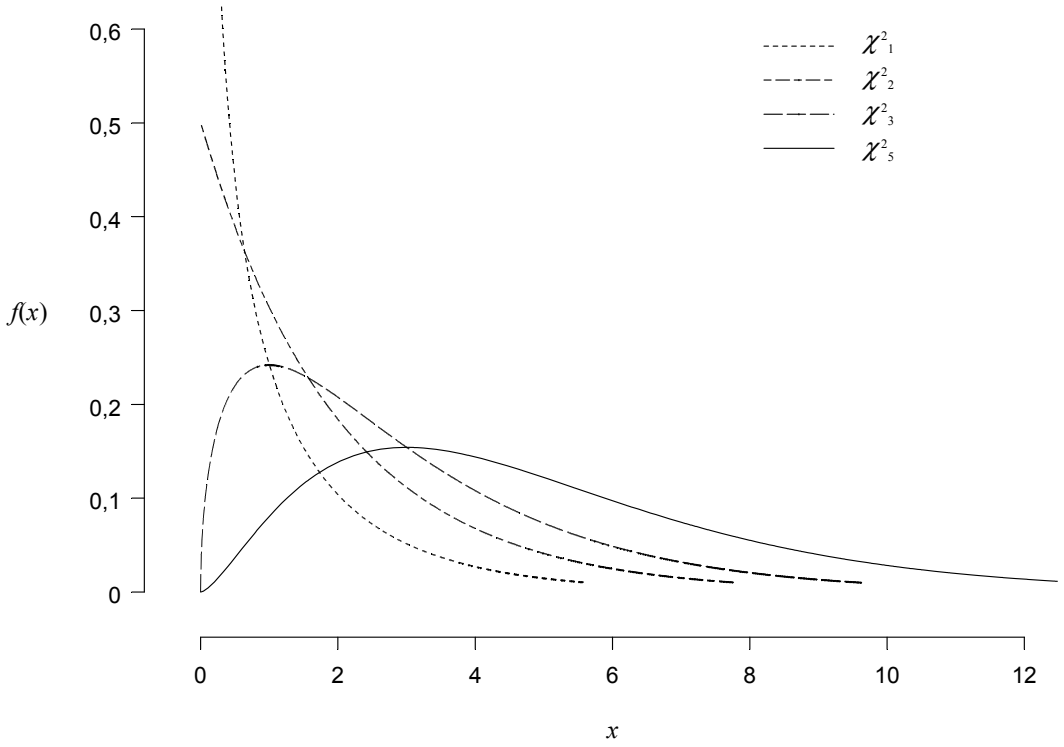


Figura 6.1 **Función de densidad de la distribución chi-cuadrado con 1, 2, 3 y 5 grados de libertad.**

A partir de la distribución χ^2_{n-1} del estadístico $(n-1)s^2/\sigma^2$ resulta sencillo calcular un intervalo de confianza para la varianza poblacional. El $100(1-\alpha)\%$ de la distribución muestral de este estadístico está comprendido entre los percentiles $\alpha/2$ y $1-\alpha/2$ de la distribución chi-cuadrado con $n-1$ grados de libertad, denotados por $\chi^2_{n-1,\alpha/2}$ y $\chi^2_{n-1,1-\alpha/2}$,

$$P\left(\chi^2_{n-1,\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{n-1,1-\alpha/2}\right) = 1-\alpha.$$

Manipulando esta desigualdad para despejar la varianza poblacional, se obtiene que

$$P\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right) = 1-\alpha;$$

es decir, el IC al $100(1-\alpha)\%$ para la varianza poblacional σ^2 viene determinado por

$$[(n-1)s^2/\chi^2_{n-1,1-\alpha/2}, (n-1)s^2/\chi^2_{n-1,\alpha/2}],$$

cuyos límites pueden calcularse a partir de los datos observados en la muestra. A diferencia de los intervalos de confianza para μ, que están centrados alrededor de $\bar{x}$, los intervalos de confianza para σ^2 no son simétricos alrededor de s^2, particularmente cuando el tamaño muestral es reducido.

De igual forma, el contraste de una determinada hipótesis nula H_0: $\sigma^2 = \sigma_0^2$ frente a la hipótesis alternativa bilateral H_1: $\sigma^2 \neq \sigma_0^2$ puede realizarse mediante el estadístico

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2},$$

que bajo H_0 sigue una distribución chi-cuadrado con $n-1$ grados de libertad. Así, el valor P del test se obtiene como el doble del área a la izquierda de este estadístico bajo la distribución χ^2_{n-1}, si $s^2 \leq \sigma_0^2$, o como el doble del área a la derecha del estadístico, si $s^2 > \sigma_0^2$. Es importante notar que, si la distribución subyacente dista mucho de ser normal, los intervalos de confianza y los contrastes para la varianza poblacional son menos fiables que para la media, en cuyo caso conviene proceder con cautela.

Ejemplo 6.5 Utilizando la desviación típica $s = 0{,}29$ mmol/l del colesterol HDL en los $n = 539$ controles del EURAMIC, el IC al 95% para la varianza poblacional viene determinado por

$$(538 \cdot 0{,}29^2 / \chi^2_{538;0{,}975}\,,\ 538 \cdot 0{,}29^2 / \chi^2_{538;0{,}025}\,)$$
$$= (45{,}25/604{,}16;\ 45{,}25/475{,}62) = (0{,}075;\ 0{,}095),$$

ya que los percentiles 2,5 y 97,5 de la distribución chi-cuadrado con 538 grados de libertad son respectivamente $\chi^2_{538;0{,}025} = 475{,}62$ y $\chi^2_{538;0{,}975} = 604{,}16$. Así, el IC al 95% para la desviación típica del colesterol HDL en la población de referencia es

$$(\sqrt{0{,}075}\,;\sqrt{0{,}095}\,) = (0{,}27;\ 0{,}31).$$

Para determinar si los niveles de colesterol HDL en los controles del EURAMIC son compatibles con una desviación típica poblacional de 0,30 mmol/l, se contrastó bilateralmente la hipótesis nula H_0: $\sigma^2 = 0{,}30^2$ mediante el estadístico

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{538 \cdot 0{,}29^2}{0{,}30^2} = 502{,}73.$$

Como $s < \sigma_0$, el valor P corresponde a $2P(\chi^2_{538} \leq 502{,}73) = 2 \cdot 0{,}140 = 0{,}280$; es decir, el contraste no resultó estadísticamente significativo, careciendo entonces de evidencia para rechazar la hipótesis nula. La conclusión de este contraste es consistente con el intervalo de confianza para σ, dado que éste incluye el valor nulo $\sigma_0 = 0{,}30$ mmol/l.

6.3 COMPARACIÓN DE MEDIAS EN DOS MUESTRAS INDEPENDIENTES

Hasta ahora se han revisado las técnicas estadísticas para realizar inferencias sobre el valor de un parámetro en una población. Sin embargo, una situación mucho más frecuente en la práctica es la comparación de un determinado parámetro entre dos poblaciones distintas. En este apartado se presentan los métodos para comparar la media poblacional de una variable cuantitativa a partir de dos muestras independientes, donde las observaciones de una muestra no están relacionadas o emparejadas con las observaciones de la otra muestra.

En adelante, la media y la varianza de la variable aleatoria en la primera población se denotan por μ_1 y σ_1^2, y en la segunda población por μ_2 y σ_2^2. El objetivo se centra en estimar la diferencia entre ambas medias poblacionales $\mu_1 - \mu_2$ a partir de dos muestras independientes de dichas poblaciones de tamaños n_1 y n_2 con medias respectivas $\bar{x}_1$ y $\bar{x}_2$ y varianzas s_1^2 y s_2^2.

Como cabría esperar, el estimador puntual es la diferencia de las medias muestrales $\bar{x}_1 - \bar{x}_2$, que representa además un estimador insesgado y consistente de la diferencia subyacente $\mu_1 - \mu_2$ en la población. Para realizar inferencias sobre esta diferencia de medias poblacionales, es necesario

conocer la distribución muestral de $\bar{x}_1 - \bar{x}_2$. Si ambos tamaños muestrales n_1 y n_2 son suficientemente grandes (recuérdese el teorema central del límite), las medias muestrales $\bar{x}_1$ y $\bar{x}_2$ seguirán aproximadamente las distribuciones normales $N(\mu_1, \sigma_1^2/n_1)$ y $N(\mu_2, \sigma_2^2/n_2)$, respectivamente. Así, al tratarse de muestras independientes (véase Apartado 3.4), la distribución muestral de la diferencia de medias también será aproximadamente normal con media

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2$$

y varianza

$$\text{var}(\bar{x}_1 - \bar{x}_2) = \text{var}(\bar{x}_1) + \text{var}(\bar{x}_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2.$$

En consecuencia, se tiene que

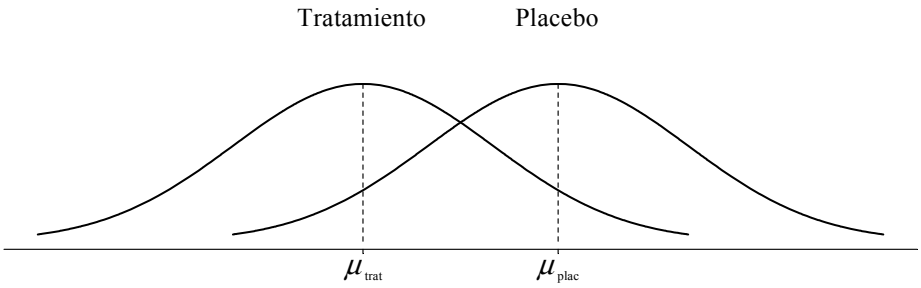
$$\bar{x}_1 - \bar{x}_2 \overset{\sim}{\to} N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

o, aplicando la estandarización de una distribución normal,

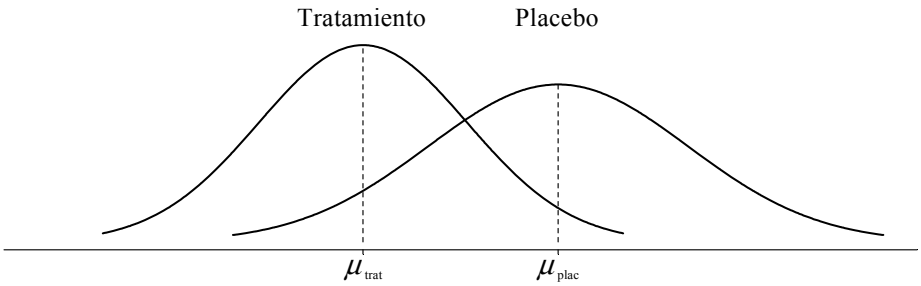
$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \overset{\sim}{\to} N(0,\ 1).$$

Esta distribución muestral constituye la base para la comparación de dos medias poblacionales a partir de muestras independientes. No obstante, para hacer uso de este resultado, es necesario estimar previamente las varianzas desconocidas σ_1^2 y σ_2^2 de ambas poblaciones. La estimación se simplifica notablemente si se asume que las dos varianzas son iguales $\sigma_1^2 = \sigma_2^2$, en cuyo caso es posible obtener una estimación combinada de la varianza común para ambas poblaciones. Por el contrario si $\sigma_1^2 \neq \sigma_2^2$, cada varianza poblacional deberá estimarse por separado, siendo entonces más impreciso el proceso de inferencia. Parece razonable pensar que la comparación de medias es más complicada en distribuciones con distinta variabilidad que en distribuciones con una misma varianza. La igualdad de varianzas no es una asunción puramente teórica, sino que tiene implicaciones prácticas como puede apreciarse en el siguiente ejemplo.

Ejemplo 6.6 En el ensayo clínico del Ejemplo 6.2 se pretende comparar las medias de presión arterial sistólica entre el grupo placebo y el grupo bajo tratamiento antihipertensivo. Si este tratamiento produjera una reducción del nivel de presión arterial aproximadamente igual en todos los pacientes, cabría esperar que la distribución de la presión arterial en los tratados presentara un nivel medio inferior que en el grupo placebo manteniendo inalterable la variabilidad. En tal caso, estaríamos ante una comparación de medias en distribuciones con igual varianza (Figura 6.2(*a*)). En caso contrario, si el tratamiento produjera una disminución de la presión arterial sistólica proporcional al nivel basal de cada paciente (esto es, mayor reducción en los sujetos con niveles más altos), la presión arterial en el grupo tratado tendría menor nivel medio y dispersión que en el grupo placebo. Bajo esta circunstancia, nos encontraríamos con una comparación de medias en distribuciones con distinta varianza (Figura 6.2(*b*)).



(*a*) Efecto constante



(*b*) Efecto proporcional

Figura 6.2 Distribución de la presión arterial sistólica en los grupos placebo y tratamiento de un hipotético ensayo clínico asumiendo un efecto constante (*a*) o proporcional (*b*) del tratamiento antihipertensivo.

6.3.1 Comparación de medias en distribuciones con igual varianza

Si se asume que las varianzas poblacionales son iguales $\sigma_1^2 = \sigma_2^2$, resulta natural estimar una única varianza combinada a partir de la información disponible en ambas muestras. Así, se obtendrá un estimador más estable de la varianza poblacional, lo que redundará en una mayor precisión de la estimación de la diferencia de medias y en una mayor potencia del contraste.

La media de las varianzas muestrales s_1^2 y s_2^2 podría utilizarse como estimador combinado de la varianza. Esta media es, sin embargo, ineficiente ya que otorga el mismo peso a ambas varianzas muestrales, aun cuando la varianza estimada a partir de una muestra mayor sea más fiable. Para dar más peso a los resultados obtenidos con mayor tamaño muestral, la estimación combinada de la varianza se obtiene como la media de s_1^2 y s_2^2 ponderada por sus correspondientes grados de libertad

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}.$$

El numerador de s^2 es simplemente la suma de las desviaciones al cuadrado respecto de la media de cada grupo, y el denominador corresponde al número de grados de libertad para el cálculo de este estimador: $n_1 - 1$ grados de libertad en la primera muestra y $n_2 - 1$ en la segunda, $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$.

En la distribución muestral de la diferencia de medias, las varianzas desconocidas σ_1^2 y σ_2^2 pueden entonces sustituirse por la estimación combinada de la varianza s^2. Sin embargo, como esta estimación s^2 está sujeta al error del muestreo, la distribución de la diferencia de medias ya no será normal, sino que seguirá aproximadamente una distribución t de Student con $n_1 + n_2 - 2$ grados de libertad,

$$\frac{x_1 - x_2 - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{\sim}{\to} t_{n_1+n_2-2}\,.$$

A partir de este resultado, y siguiendo un procedimiento análogo al utilizado para una media (Apartado 5.3.2), puede derivarse un intervalo de confianza al 100(1 – α)% para la diferencia de medias poblacionales $\mu_1 - \mu_2$ como

$$\bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2,1-\alpha/2}\; s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

que está centrado alrededor de la diferencia de medias muestrales y cuya amplitud depende de su error estándar $SE(\bar{x}_1 - \bar{x}_2) = s\sqrt{1/n_1 + 1/n_2}$. Notar que este intervalo es una generalización bastante natural del intervalo para la media de una población.

Ejemplo 6.7 En el estudio EURAMIC, la media y la desviación típica del colesterol HDL entre los $n_{ca} = 462$ casos de infarto de miocardio fueron $\bar{x}_{ca} = 0{,}98$ y $s_{ca} = 0{,}25$ mmol/l, y entre los $n_{co} = 539$ controles fueron $\bar{x}_{co} = 1{,}09$ y $s_{co} = 0{,}29$ mmol/l. De estos datos se deduce que la estimación puntual de la diferencia en el nivel medio de colesterol HDL es $\bar{x}_{ca} - \bar{x}_{co} = 0{,}98 - 1{,}09 = -0{,}11$ mmol/l. Si asumimos una misma variabilidad del colesterol HDL en casos y controles, la varianza combinada de ambas muestras vendría determinado por

$$s^2 = \frac{(n_{ca} - 1)s_{ca}^2 + (n_{co} - 1)s_{co}^2}{n_{ca} + n_{co} - 2}$$

$$= \frac{(462 - 1)\,0{,}25^2 + (539 - 1)\,0{,}29^2}{462 + 539 - 2} = 0{,}074;$$

es decir, la desviación típica combinada es $s = \sqrt{0{,}074} = 0{,}272$ mmol/l, cuyo valor está más próximo a la desviación típica observada en los controles que en los casos (mayor tamaño muestral de los primeros). Así, el error estándar de la diferencia de medias puede calcularse como

$$SE(\bar{x}_{ca} - \bar{x}_{co}) = s\sqrt{\frac{1}{n_{ca}} + \frac{1}{n_{co}}} = 0{,}272\sqrt{\frac{1}{462} + \frac{1}{539}} = 0{,}017.$$

A partir de la diferencia de medias muestrales y de su error estándar, y teniendo en cuenta que la distribución t de Student con $n_{ca} + n_{co} - 2 = 999$ grados de libertad es virtualmente idéntica a una distribución normal estandarizada, el IC al 95% para $\mu_{ca} - \mu_{co}$ viene dado por

$$\bar{x}_{ca} - \bar{x}_{co} \pm t_{999;0,975}\, SE(\bar{x}_{ca} - \bar{x}_{co})$$

$$= -0{,}11 \pm 1{,}96 \cdot 0{,}017 = (-0{,}14;\, -0{,}08).$$

De los resultados del estudio EURAMIC puede entonces concluirse que el nivel medio de colesterol HDL en los casos de infarto es inferior en 0,11 mmol/l al nivel medio de los sujetos libres de la enfermedad, estando esta diferencia comprendida entre 0,08 y 0,14 mmol/l con una confianza del 95%.

En el caso de la comparación de medias entre dos poblaciones, la hipótesis nula más natural es la igualdad de ambas medias poblacionales. Para realizar el contraste de esta hipótesis nula H_0: $\mu_1 = \mu_2$ frente a la hipótesis alternativa bilateral H_1: $\mu_1 \neq \mu_2$ a partir de dos muestras independientes de igual varianza, se emplea el siguiente test estadístico

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

que sigue aproximadamente una distribución t de Student con $n_1 + n_2 - 2$ grados de libertad si la hipótesis nula H_0: $\mu_1 = \mu_2$ es cierta. Por tanto, el valor P se obtiene como el área bajo la distribución $t_{n_1+n_2-2}$ para valores más extremos que el valor observado de t. Esta prueba de hipótesis se conoce genéricamente como el **test de la *t* de Student para muestras independientes con igual varianza**.

Ejemplo 6.8 Un nivel medio de colesterol HDL significativamente más bajo en los casos de infarto que en los sujetos libres de enfermedad sería compatible con la hipótesis de que el colesterol HDL es un factor protector para el infarto de miocardio. En este ejemplo, se pretende contrastar esta hipótesis a partir de los niveles de colesterol HDL observados en los casos y controles del estudio EURAMIC. El resultado de este contraste, junto con la estimación puntual y por intervalo obtenidas en el ejemplo anterior, permiten evaluar no sólo la significación estadística sino también la relevancia clínica y de salud pública del hallazgo.

Asumiendo igualdad de varianzas poblacionales, el contraste bilateral de la hipótesis nula H_0: $\mu_{ca} = \mu_{co}$ se realiza mediante el estadístico

$$t = \frac{\bar{x}_{ca} - \bar{x}_{co}}{SE(\bar{x}_{ca} - \bar{x}_{co})} = \frac{-0{,}11}{0{,}017} = -6{,}35.$$

Si ambas medias poblacionales fueran iguales, la distribución de este estadístico sería t_{999} o aproximadamente normal estandarizada. El valor P bilateral se obtiene entonces como el doble de la probabilidad a la izquierda de $-6{,}35$ en la distribución normal estandarizada, que corresponde a $P < 0{,}001$. Así, puede concluirse que existen diferencias muy significativas en el nivel medio de colesterol HDL entre los infartados y los sujetos libres de enfermedad. Esta diferencia significativa es perfectamente consistente con el intervalo de confianza calculado en el ejemplo anterior, puesto que éste no contenía al cero (valor nulo para la diferencia de medias).

Los métodos descritos en este apartado pueden extenderse a la comparación de tres o más medias poblacionales. Las técnicas para comparar medias en múltiples muestras independientes se conocen con el nombre de **análisis de la varianza de una vía** y pueden consultarse en los libros referenciados al final del tema. Aunque estos procedimientos no se tratan explícitamente en este texto, la comparación de múltiples medias a partir de datos independientes también puede abordarse mediante los modelos de regresión lineal que se presentarán más adelante (Temas 10 y 11).

6.3.2 Contraste para la igualdad de varianzas

La comparación de medias presentada en el apartado anterior se fundamenta en la asunción de igualdad de varianzas. Esta asunción es determinante para poder calcular una estimación combinada de la varianza. En este apartado se presentan los métodos para contrastar estadísticamente la hipótesis de homogeneidad de varianzas en dos muestras independientes.

El test para la igualdad de varianzas poblacionales se basa en la comparación de las varianzas muestrales s_1^2 y s_2^2. Como se apuntó anteriormente (Apartado 6.2.2), si la distribución subyacente de la variable es normal en ambas poblaciones, los estadísticos $(n_1 - 1)s_1^2/\sigma_1^2$ y $(n_2 - 1)s_2^2/\sigma_2^2$ se distribuyen como una chi-cuadrado con $n_1 - 1$ y $n_2 - 1$ grados de libertad, respectivamente. Combinando la distribución de estos estadísticos en ambas muestras independientes, se obtiene que

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim \frac{\chi^2_{n_1-1}/(n_1-1)}{\chi^2_{n_2-1}/(n_2-1)}.$$

A la derecha de esta expresión se tiene el cociente de dos variables independientes chi-cuadrado divididas por sus respectivos grados de libertad, que se conoce como la distribución ***F* de Fisher** con $n_1 - 1$ grados de libertad en el numerador y $n_2 - 1$ en el denominador, y se denota por F_{n_1-1,n_2-1}. Así, la razón entre s_1^2/σ_1^2 y s_2^2/σ_2^2 sigue una distribución F con $n_1 - 1$ y $n_2 - 1$ grados de libertad,

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

La distribución F de Fisher toma sólo valores positivos y está sesgada positivamente con un valor más frecuente (moda) menor de 1 y una media mayor de 1. Al aumentar los grados de libertad del numerador y denominador, tanto la media como la moda se aproximan al valor 1 (Figura 6.3). Los percentiles de la distribución F de Fisher para distintos grados de libertad del numerador y denominador se presentan en la Tabla 7 del Apéndice.

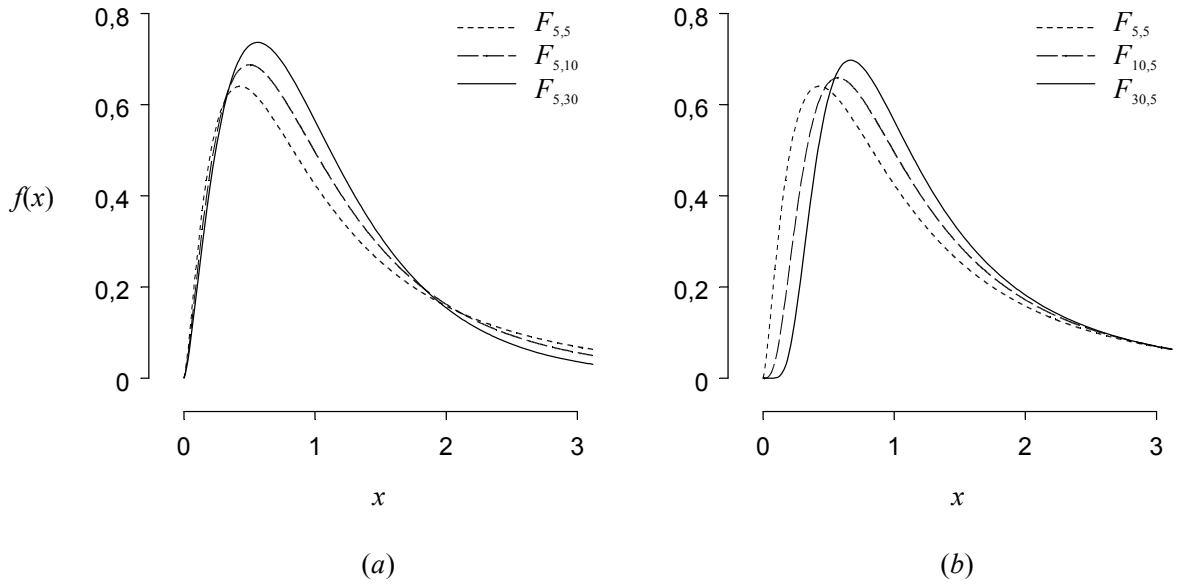


Figura 6.3 **Función de densidad de la distribución *F* de Fisher al aumentar los grados de libertad del denominador (*a*) y del numerador (*b*).**

Ejemplo 6.9 Utilizando la Tabla 7 del Apéndice, el percentil 97,5 de una distribución F de Fisher con 5 grados de libertad en el numerador y denominador es $F_{5;5;0,975} = 7,15$, y para 30 grados de libertad en ambos es $F_{30;30;0,975} = 2,07$. Aunque esta tabla no facilita percentiles inferiores, puede comprobarse que el percentil α en una distribución F con d_1 y d_2 grados de libertad es igual al inverso del percentil $1 - \alpha$ en una distribución F con d_2 y d_1 grados de libertad, $F_{d_1,d_2,\alpha} = 1/F_{d_2,d_1,1-\alpha}$. Así, el percentil 2,5 en las distribuciones anteriores es $F_{5;5;0,025} = 1/F_{5;5;0,975} = 1/7,15 = 0,14$ y $F_{30;30;0,025} = 1/F_{30;30;0,975} = 1/2,07 = 0,48$. Por tanto, el 95% central de la distribución $F_{5,5}$ está comprendido entre 0,14 y 7,15, y de la distribución $F_{30,30}$ entre 0,48 y 2,07. Puede entonces observarse que, al aumentar el número de grados de libertad del numerador y denominador, la distribución F de Fisher se hace menos dispersa y más simétrica alrededor del valor 1.

A partir de la distribución muestral F_{n_1-1,n_2-1} del cociente entre s_1^2/σ_1^2 y s_2^2/σ_2^2, resulta sencillo calcular un intervalo de confianza para la razón de dos varianzas poblacionales σ_1^2/σ_2^2. No obstante, por su mayor utilidad práctica, nos centraremos aquí en el test para la igualdad de varianzas. El contraste bilateral de la hipótesis nula H_0: $\sigma_1^2 = \sigma_2^2$ frente a la alternativa H_1: $\sigma_1^2 \neq \sigma_2^2$ se basa en la razón de las varianzas muestrales

$$F = \frac{s_1^2}{s_2^2}.$$

Si la hipótesis nula de igualdad de varianzas $\sigma_1^2 = \sigma_2^2$ es cierta, la razón $(s_1^2/\sigma_1^2)/(s_2^2/\sigma_2^2)$ se reduce a s_1^2/s_2^2, de tal forma que este estadístico se distribuirá según una F de Fisher con $n_1 - 1$ grados de libertad en el numerador y $n_2 - 1$ en el denominador. El valor P del contraste se calcula entonces como el doble de la probabilidad a la izquierda de este estadístico bajo la distribución F_{n_1-1,n_2-1}, si $s_1^2 \leq s_2^2$, o como el doble del área a la derecha del estadístico, si $s_1^2 > s_2^2$.

Ejemplo 6.10 En los Ejemplos 6.7 y 6.8 se comparó la media del colesterol HDL entre los casos y controles del EURAMIC bajo la asunción de homogeneidad de varianzas. La validez de estos resultados dependerá del cumplimiento de dicha hipótesis. Para contrastar bilateralmente la hipótesis nula H_0: $\sigma_{ca}^2 = \sigma_{co}^2$, se calcula el test estadístico

$$F = \frac{s_{ca}^2}{s_{co}^2} = \frac{0,25^2}{0,29^2} = 0,74,$$

que sigue una distribución F con $n_{ca} - 1 = 461$ y $n_{co} - 1 = 538$ grados de libertad bajo H_0. Como $s_{ca} < s_{co}$, el valor P es igual a $2P(F_{461,538} \leq 0,74) = 2 \cdot 0,0005 = 0,001$. Notar que este valor P sería idéntico si se hubiera utilizado el estadístico inverso $F = s_{co}^2 / s_{ca}^2 = 1,35$. En tal caso, el valor P se obtendría a partir de la distribución $F_{538,461}$ como $2P(F_{538,461} \geq 1,35) = 2 \cdot 0,0005 = 0,001$.

La variabilidad del colesterol HDL resulta significativamente menor entre los casos de infarto que entre los individuos libres de la enfermedad, con lo cual no puede aceptarse la hipótesis de igualdad de varianzas. En consecuencia, los procedimientos utilizados en los Ejemplos 6.7 y 6.8 son inadecuados para comparar los niveles medios de colesterol HDL entre casos y controles.

Existen otras técnicas estadísticas para la comparación de varianzas en muestras independientes, tales como el **test de Bartlett** o la **prueba de Levene**. En general, estas técnicas

permiten comparar varianzas entre dos o más grupos y, en el caso del test de Levene, la comparación no requiere que la distribución subyacente de la variable sea normal. Los lectores interesados pueden consultar estos procedimientos en las referencias incluidas al final del tema.

6.3.3 Comparación de medias en distribuciones con distinta varianza

Cuando las varianzas poblacionales son distintas, carece de sentido calcular una estimación combinada de la varianza, ya que ésta infraestimará o sobreestimará la variabilidad específica de cada población. En este caso, aun perdiendo algo de precisión, es preferible estimar por separado las varianzas poblacionales σ_1^2 y σ_2^2 mediante sus correspondientes varianzas muestrales s_1^2 y s_2^2.

Así, sustituyendo σ_1^2 por s_1^2 y σ_2^2 por s_2^2 en la distribución muestral de la diferencia de medias, se obtiene el estadístico

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Aunque resulta complicado derivar la distribución exacta de este estadístico, existen diversas aproximaciones que funcionan bien en la práctica. El método más utilizado es la aproximación de Welch, que permite aproximar la distribución de este estadístico mediante una t de Student con los siguientes grados de libertad

$$d = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$$

Puede comprobarse que d es siempre inferior o igual a $n_1 + n_2 - 2$; es decir, esta distribución t de Student será más dispersa que la empleada en el caso de igualdad de varianzas. Esto es lo que cabría esperar ya que, al estimar por separado las varianzas, la distribución resultante ha de reflejar mayor incertidumbre. Esto conllevará una disminución tanto en la precisión de los intervalos de confianza como en la potencia de los contrastes.

En el caso de distribuciones con distinta varianza, el intervalo de confianza al $100(1 - \alpha)\%$ para la diferencia de medias poblacionales $\mu_1 - \mu_2$ vendrá determinado por

$$\bar{x}_1 - \bar{x}_2 \pm t_{d,1-\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

donde d son los grados de libertad calculados según la fórmula anterior. De igual forma, para contrastar la hipótesis nula H_0: $\mu_1 = \mu_2$ frente a la alternativa H_1: $\mu_1 \neq \mu_2$ a partir de dos muestras independientes con distinta varianza, se emplea el estadístico

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

que bajo la hipótesis nula, se distribuye aproximadamente según una t de Student con d grados de libertad. Así, el valor P viene dado por la probabilidad de obtener valores más extremos que el valor observado de t bajo la distribución t_d. Este contraste se conoce con el nombre de **test de la *t* de Student para muestras independientes con distinta varianza**.

Ejemplo 6.11 En el Ejemplo 6.10 se contrastó que la variabilidad del colesterol HDL difiere significativamente entre los casos de infarto y los sujetos libres de la enfermedad. Por ello, la comparación del nivel medio de colesterol HDL entre casos y controles ha de realizarse mediante la prueba *t* de Student para muestras independientes con distinta varianza. La estimación puntual de la diferencia de medias es $\bar{x}_{ca} - \bar{x}_{co} = 0{,}98 - 1{,}09 = -0{,}11$ mmol/l, cuyo error estándar se estima directamente por

$$SE(\bar{x}_{ca} - \bar{x}_{co}) = \sqrt{\frac{s_{ca}^2}{n_{ca}} + \frac{s_{co}^2}{n_{co}}} = \sqrt{\frac{0{,}25^2}{462} + \frac{0{,}29^2}{539}} = 0{,}017.$$

En el caso de varianzas heterogéneas, los grados de libertad para la distribución de la diferencia de medias vienen determinados por la aproximación de Welch

$$d = \frac{(s_{ca}^2 / n_{ca} + s_{co}^2 / n_{co})^2}{(s_{ca}^2 / n_{ca})^2 / (n_{ca} - 1) + (s_{co}^2 / n_{co})^2 / (n_{co} - 1)}$$

$$= \frac{(0{,}25^2 / 462 + 0{,}29^2 / 539)^2}{(0{,}25^2 / 462)^2 / (462 - 1) + (0{,}29^2 / 539)^2 / (539 - 1)} = 998{,}97.$$

Notar que, en este ejemplo, los grados de libertad son casi iguales a los obtenidos bajo la asunción de igualdad de varianzas ($n_{ca} + n_{co} - 2 = 999$). A partir de estos resultados es posible calcular un IC al 95% para $\mu_{ca} - \mu_{co}$ como

$$\bar{x}_{ca} - \bar{x}_{co} \pm t_{998{,}97;0{,}975}\, SE(\bar{x}_{ca} - \bar{x}_{co})$$
$$= -0{,}11 \pm 1{,}96 \cdot 0{,}017 = (-0{,}14;\ -0{,}08),$$

y contrastar la hipótesis nula H_0: $\mu_{ca} = \mu_{co}$ mediante el estadístico

$$t = \frac{\bar{x}_{ca} - \bar{x}_{co}}{SE(\bar{x}_{ca} - \bar{x}_{co})} = \frac{-0{,}11}{0{,}017} = -6{,}44,$$

que bajo la distribución $t_{998{,}97}$ o normal estandarizada, corresponde a un valor *P* menor que 0,001. Así, se pone de manifiesto que los casos de infarto presentan un nivel medio de colesterol HDL significativamente inferior que los sujetos libres de la enfermedad ($P < 0{,}001$), con una diferencia estimada en 0,11 mmol/l (IC al 95% 0,08-0,14 mmol/l). En este caso, los resultados obtenidos asumiendo homogeneidad o heterogeneidad de varianzas son virtualmente idénticos debido, en parte, a que ambos tamaños muestrales no difieren sustancialmente.

En resumen, la comparación de medias en muestras independientes requiere contrastar en primer lugar la igualdad de varianzas, para después utilizar según proceda el test de la *t* de Student con igual o distinta varianza. Esta distinción no es meramente académica: si la variabilidad difiere entre ambas poblaciones, los procedimientos de estimación y contraste asumiendo igualdad de varianzas pueden ser muy engañosos, particularmente en muestras pequeñas o moderadas cuyos tamaños n_1 y n_2 difieran sustancialmente.

6.4 COMPARACIÓN DE MEDIAS EN DOS MUESTRAS DEPENDIENTES

Los datos dependientes surgen cuando las observaciones recogidas en el estudio están correlacionadas entre sí. A continuación se presentan algunos mecanismos y diseños epidemiológicos que generan datos dependientes:

- La obtención de dos o más determinaciones de la misma variable en un mismo sujeto da lugar a datos dependientes, que pueden presentarse como:
  - Diferentes medidas de la misma variable en un momento determinado, habitualmente para aumentar la fiabilidad del instrumento de medida.
  - Determinaciones de la misma variable en diferentes localizaciones anatómicas.
  - Medidas repetidas en el mismo sujeto a lo largo del tiempo, bien sea en comparaciones antes y después de un tratamiento, en ensayos clínicos cruzados o en estudios de medidas repetidas con visitas sucesivas.
- La selección de los participantes en un estudio emparejándolos por determinadas características pronósticas genera datos dependientes entre los sujetos emparejados. El ejemplo más habitual es el emparejamiento en el diseño de los estudios de casos y controles.
- Los datos de estudios procedentes de sujetos de una misma familia o de animales pertenecientes a la misma camada suelen ser también dependientes.

En todos estos casos, la correlación se limita a los grupos específicos donde se genera la dependencia, que suelen ser habitualmente parejas. Así, en un estudio de casos y controles emparejados, los datos de cada pareja son dependientes, pero los datos de las distintas parejas son independientes entre sí. Igualmente, en un estudio de medidas repetidas, los datos de un mismo individuo son dependientes, mientras que los resultados en diferentes individuos son independientes entre sí.

Las muestras dependientes están constituidas por observaciones en los mismos sujetos o en distintos sujetos emparejados según ciertas características pronósticas de interés. De esta forma, la distribución de dichas características será similar en ambas muestras, eliminando así la posibilidad de que estos factores influyan en la comparación objeto de estudio. En general, el emparejamiento es una técnica frecuentemente utilizada en el diseño de estudios clínicos o epidemiológicos con el propósito de controlar por determinados factores de confusión (ver textos de método epidemiológico referenciados al final del tema). Estos diseños requieren de técnicas específicas de análisis que preserven el emparejamiento. En este apartado se revisan los métodos estadísticos para el tratamiento de un caso específico de dependencia, en el que se dispone de dos determinaciones de una variable continua para cada pareja de datos dependientes.

> ***Ejemplo 6.12*** Supongamos que en el estudio EURAMIC se seleccionan aleatoriamente 50 casos de infarto de miocardio. Como la edad es un importante factor pronóstico de enfermedades coronarias, cada uno de estos casos se emparejó por grupos quinquenales de edad a un control libre de la enfermedad. Así, por ejemplo, para un caso de 62 años de edad se seleccionó aleatoriamente un control entre todos los controles disponibles con edades comprendidas entre 60 y 64 años. La muestra resultante de aplicar este procedimiento constituiría un estudio de casos y controles emparejados. En este estudio, cabría esperar un cierto grado de correlación en la información recogida para cada pareja, dado que tanto el caso como el control se encuentran en el mismo rango de edad. En la Tabla 6.1 se presentan los niveles de colesterol HDL en las 50 parejas de casos y controles.

Tabla 6.1 Colesterol HDL en 50 casos y controles del estudio EURAMIC emparejados según grupos quinquenales de edad.

	Colesterol HDL (mmol/l)				Colesterol HDL (mmol/l)		
Pareja	Caso	Control	*d**	Pareja	Caso	Control	*d**
1	0,81	0,63	0,18	26	0,96	1,29	−0,33
2	0,91	0,91	0,00	27	1,33	0,72	0,61
3	0,98	0,76	0,22	28	0,93	1,04	−0,11
4	0,91	1,19	−0,28	29	0,32	1,54	−1,22
5	0,55	0,99	−0,44	30	0,86	1,08	−0,22
6	0,62	1,14	−0,52	31	0,93	1,12	−0,19
7	0,79	0,73	0,06	32	1,40	1,75	−0,35
8	0,89	1,08	−0,19	33	1,50	1,29	0,21
9	1,24	0,87	0,37	34	0,92	1,17	−0,25
10	1,76	1,04	0,72	35	0,88	0,93	−0,05
11	1,35	1,03	0,32	36	0,82	0,88	−0,06
12	0,72	1,09	−0,37	37	1,52	0,74	0,78
13	0,94	1,12	−0,18	38	1,68	1,45	0,23
14	1,01	1,20	−0,19	39	0,81	1,02	−0,21
15	0,98	1,62	−0,64	40	0,60	1,15	−0,55
16	0,92	1,25	−0,33	41	1,16	1,49	−0,33
17	0,68	1,31	−0,63	42	0,75	0,98	−0,23
18	1,48	1,00	0,48	43	0,96	1,31	−0,35
19	1,23	0,78	0,45	44	1,46	1,15	0,31
20	0,83	0,95	−0,12	45	0,76	1,51	−0,75
21	0,92	1,13	−0,21	46	0,76	1,01	−0,25
22	0,82	0,97	−0,15	47	1,12	1,26	−0,14
23	1,21	0,74	0,47	48	1,01	0,91	0,10
24	0,78	0,88	−0,10	49	0,99	1,63	−0,64
25	0,88	1,14	−0,26	50	0,75	1,45	−0,70

* Diferencia de colesterol HDL entre caso y control.

Para concretar el problema supongamos que se dispone de n pares de observaciones de una variable aleatoria continua. En cada pareja de datos dependientes, una observación x_1 corresponde a la primera muestra y la otra observación x_2 a la segunda muestra. El objetivo se centra en comparar las medias poblacionales μ_1 y μ_2 a partir de estas dos muestras dependientes.

Los procedimientos desarrollados en el Apartado 6.3 no pueden aplicarse a esta situación, ya que las medias de ambas muestras no son independientes por provenir de observaciones correlacionadas. Sin embargo, la comparación se simplifica notablemente si se calculan las diferencias $d = x_1 - x_2$ en cada una de las n observaciones emparejadas. Por un lado, como las distintas parejas no están relacionadas entre sí, estas diferencias son independientes. Por otro lado, la media de las diferencias $\bar{d}$ coincide con la diferencia de medias muestrales,

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - x_{i2})$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_{i1} - \frac{1}{n}\sum_{i=1}^{n} x_{i2} = \bar{x}_1 - \bar{x}_2$$

y, en consecuencia, $\bar{d}$ es un estimador insesgado de la diferencia de medias poblacionales $\mu_1 - \mu_2$. Así, el problema de la comparación de medias en dos muestras dependientes queda reducido a una simple inferencia sobre la media de una única muestra de n diferencias independientes.

Los métodos del Apartado 6.2.1 para la media de una muestra pueden entonces utilizarse para calcular un intervalo de confianza al $100(1-\alpha)\%$ para $\mu_1 - \mu_2$ como

$$\bar{d} \pm t_{n-1,1-\alpha/2}\frac{s_d}{\sqrt{n}},$$

donde s_d es la desviación típica de las diferencias observadas. De igual forma, la hipótesis de igualdad de medias poblacionales H_0: $\mu_1 = \mu_2$ puede contrastarse frente a la hipótesis alternativa H_1: $\mu_1 \neq \mu_2$ mediante el estadístico

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}.$$

Bajo la hipótesis nula, las diferencias observadas se distribuirían aleatoriamente alrededor del valor 0, de tal forma que este estadístico seguiría una distribución t de Student con $n-1$ grados de libertad. El valor P corresponderá, por tanto, a la probabilidad bajo la distribución t_{n-1} para valores más extremos que el valor observado de t. Esta prueba se denomina habitualmente como el **test de la *t* de Student para muestras dependientes**.

Ejemplo 6.13 Para preservar el emparejamiento entre los casos y controles de la Tabla 6.1, se calcula la diferencia de colesterol HDL $d = x_{ca} - x_{co}$ en cada pareja. Como puede apreciarse, predominan las parejas donde el caso presenta un nivel inferior de colesterol HDL que su correspondiente control (diferencias negativas). De hecho, la media de estas diferencias

$$\bar{d} = \frac{1}{50}\sum_{i=1}^{50} d_i = \frac{0{,}18 + 0{,}00 + \ldots - 0{,}70}{50} = -0{,}12$$

es una estimación de la diferencia en el nivel medio de colesterol HDL entre los casos de infarto y los sujetos libres de la enfermedad. La varianza de las diferencias viene dada por

$$s_d^2 = \frac{1}{49}\sum_{i=1}^{50}(d_i - \bar{d})^2$$

$$= \frac{(0{,}18 + 0{,}12)^2 + \ldots + (-0{,}70 + 0{,}12)^2}{49} = 0{,}16,$$

luego el error estándar de $\bar{d}$ es

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{0{,}40}{\sqrt{50}} = 0{,}057.$$

Así, el IC al 95% para la diferencia de medias poblacionales $\mu_{ca} - \mu_{co}$ se obtiene como

$$\bar{d} \pm t_{49;0,975}\, SE(\bar{d}) = -0,12 \pm 2,01 \cdot 0,057 = (-0,23;\ -0,01),$$

y la hipótesis nula H_0: $\mu_{ca} = \mu_{co}$ se contrasta mediante el test estadístico

$$t = \frac{\bar{d}}{SE(\bar{d})} = \frac{-0,12}{0,057} = -2,13,$$

cuyo valor P asociado en la distribución t_{49} es $P = 2P(t_{49} \leq -2,13) = 2 \cdot 0,019 = 0,038$. De este estudio de casos y controles emparejados puede entonces concluirse que la media del colesterol HDL en los casos de infarto es inferior en 0,12 mmol/l al nivel medio de los controles (IC al 95% 0,01-0,23 mmol/l), siendo esta diferencia estadísticamente significativa (P = 0,038). Esta conclusión es consistente con la obtenida en el Ejemplo 6.11 para las muestras completas e independientes de casos y controles. No obstante, cabe destacar las siguientes particularidades. Por un lado, esta estimación está sujeta a mayor variabilidad aleatoria ya que tan sólo utiliza 50 parejas de casos y controles. Por otro lado, el diseño emparejado permite comparar casos con controles de similar edad y, en consecuencia, la estimación será menos propensa a posibles sesgos derivados de la diferencia de edad entre casos y controles.

Los procedimientos presentados en este apartado se limitan a la comparación de una variable continua a partir de dos muestras emparejadas sujeto a sujeto. El **análisis de la varianza de dos vías** permite extender esta comparación a casos más generales de dependencia, tales como el diseño de parejas con más de un sujeto por muestra (por ejemplo, un estudio de casos y controles donde cada caso se empareja con 2 controles) o la comparación de tres o más muestras dependientes (por ejemplo, un ensayo clínico donde cada paciente recibe diversos tratamientos alternativos). Los métodos de análisis de la varianza de dos vías pueden consultarse en los textos estadísticos citados a continuación.

6.5 REFERENCIAS

1. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research, Fourth Edition*. Oxford: Blackwell Science, 2001.
2. Bickel PJ, Doksum KA. *Mathematical Statistics: Basic Ideas and Selected Topics*. Englewood Cliffs, NJ: Prentice Hall, 1977.
3. Breslow NE, Day NE. *Statistical Methods in Cancer Research, Volume 1, The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer, 1980.
4. Casella G, Berger RL. *Statistical Inference, Second Edition*. Belmont, CA: Brooks/Cole, 2001.
5. Colton T. *Estadística en Medicina*. Barcelona: Salvat, 1979.
6. Fleiss JL. *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons, 1986.

7. Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research: Principles and Quantitative Methods*. New York: John Wiley & Sons, 1982.
8. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. *Applied Regression Analysis and Other Multivariable Methods, Third Edition*. Belmont, CA: Duxbury Press, 1998.
9. Rosner B. *Fundamentals of Biostatistics, Fifth Edition*. Belmont, CA: Duxbury Press, 1999.
10. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology, Third Edition*. Philadelphia: Lippincott Williams & Wilkins, 2008.
11. Snedecor GW, Cochran WG. *Statistical Methods, Eighth Edition*. Ames, IA: Iowa State University Press, 1989.
12. Stuart A, Ord JK, Arnold S. *Kendall's Advanced Theory of Statistics, Volume 2A, Classical Inference and the Linear Model, Sixth Edition*. London: Edward Arnold, 1999.

TEMA 7

INFERENCIA SOBRE PROPORCIONES

7.1 INTRODUCCIÓN

En el análisis de datos epidemiológicos es frecuente el estudio de variables dicotómicas, que reflejan la presencia o ausencia de una determinada característica en los miembros de una población. El interés radica fundamentalmente en estimar la proporción π de individuos o elementos de la población que presentan dicha característica.

Esta proporción poblacional π es un parámetro desconocido que se estima mediante la proporción muestral $p = k/n$, donde k es el número observado de individuos que presentan la característica de interés en una muestra aleatoria de tamaño n. La distribución muestral de una proporción ya se discutió en el Apartado 4.3.4. Brevemente, recordamos que una proporción muestral p tiende a distribuirse de forma normal con media π y varianza $\pi(1-\pi)/n$,

$$p \overset{\sim}{\to} N\left(\pi, \frac{\pi(1-\pi)}{n}\right),$$

cuando el tamaño muestral es suficientemente grande y la proporción poblacional no es excesivamente extrema, de tal forma que se cumpla la condición $n\pi(1-\pi) \geq 5$. Esta aproximación se utilizará repetidamente a lo largo de este tema de inferencia sobre datos de carácter binario o dicotómico.

Al igual que en el tema de inferencia sobre medias, este capítulo aborda la estimación de una proporción poblacional, así como la comparación de proporciones a partir de muestras dependientes e independientes. Para cada problema de inferencia sobre proporciones se presentará un estimador puntual del parámetro poblacional objeto de estudio, un intervalo de confianza y una prueba de significación.

7.2 INFERENCIA SOBRE UNA PROPORCIÓN POBLACIONAL

Con frecuencia se desea conocer la proporción π de individuos que poseen una cierta característica en la población. Como ya se apuntó en el Apartado 5.2, la proporción muestral p es un buen estimador puntual de la proporción poblacional, ya que p es el estimador insesgado y consistente de π con menor error estándar.

Utilizando la aproximación normal a la distribución muestral de p, se tiene la siguiente relación

$$P\left(-z_{1-\alpha/2} < \frac{p-\pi}{\sqrt{\pi(1-\pi)/n}} < z_{1-\alpha/2}\right) = 1-\alpha,$$

donde $z_{1-\alpha/2}$ es el percentil $1-\alpha/2$ de la distribución normal estandarizada. El método más sencillo para obtener un intervalo de confianza consiste en sustituir el error estándar de p por su estimación $\sqrt{p(1-p)/n}$ y despejar la proporción poblacional

$$P\left(p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < \pi < p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha.$$

Así, el intervalo de confianza al $100(1-\alpha)\%$ para la proporción poblacional π viene dado por

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\,.$$

Para realizar el contraste de la hipótesis nula H_0: $\pi = \pi_0$ frente a la alternativa bilateral H_1: $\pi \neq \pi_0$, puede emplearse el estadístico

$$z = \frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}},$$

cuya distribución será aproximadamente $N(0, 1)$ si la hipótesis nula H_0: $\pi = \pi_0$ es cierta. El valor P del test corresponde entonces a la probabilidad bajo la distribución normal estandarizada para valores más alejados de 0 que el valor observado de z.

Ejemplo 7.1 A partir de los controles del estudio EURAMIC, se pretende estimar la proporción de individuos en la población de referencia de dicho estudio que presentan niveles de colesterol HDL inferiores o iguales a 0,90 mmol/l (niveles bajos según el "*National Cholesterol Education Program*"). En $k = 158$ de los $n = 539$ controles se observaron valores inferiores o iguales a este umbral, obteniéndose una proporción muestral

$$p = k/n = 158/539 = 0{,}293.$$

Dado que $np(1-p) = 111{,}7 \geq 5$, puede emplearse la aproximación normal para calcular un IC al 95% para la proporción poblacional π como

$$0{,}293 \pm z_{0{,}975}\sqrt{\frac{0{,}293(1-0{,}293)}{539}}$$

$$= 0{,}293 \pm 1{,}96 \cdot 0{,}020 = (0{,}255;\ 0{,}332);$$

es decir, la proporción poblacional de sujetos con niveles bajos de colesterol HDL está comprendida entre el 25,5 y el 33,2% con una confianza del 95%. Asimismo, para determinar si los datos muestrales son compatibles con una proporción subyacente del 30%, se contrastó la hipótesis H_0: $\pi = 0{,}30$ versus H_1: $\pi \neq 0{,}30$ mediante el estadístico

$$z = \frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0{,}293-0{,}30}{\sqrt{\frac{0{,}30(1-0{,}30)}{539}}} = -0{,}35,$$

que corresponde a un valor $P = 2P(Z \leq -0{,}35) = 2\{1 - \Phi(0{,}35)\} = 0{,}726$ en las tablas de la distribución normal estandarizada (Tabla 3 del Apéndice). Por tanto, puede concluirse que la prevalencia poblacional de niveles bajos de colesterol HDL no es significativamente distinta del 30%.

Los procedimientos de inferencia presentados en este apartado asumen que el tamaño muestral es suficientemente grande para aplicar la aproximación normal; es decir, ha de cumplirse el requerimiento mínimo de que $n\pi(1-\pi) \geq 5$. No obstante, en el Apéndice de este tema (Apartado 7.8) se facilitan correcciones de estos métodos que permiten aumentar la cobertura de los intervalos de confianza y reducir la probabilidad de un error de tipo I en los contrastes, particularmente cuando el tamaño muestral es moderado o pequeño. Esta corrección

de la aproximación normal se conoce como **corrección por continuidad** y es aplicable a la mayoría de los procedimientos estadísticos descritos en este tema. En adelante, se tratarán los métodos de inferencia sin corrección por continuidad. Las correspondientes versiones con corrección se presentan en el Apéndice al final del tema.

7.3 COMPARACIÓN DE PROPORCIONES EN DOS MUESTRAS INDEPENDIENTES

Supongamos ahora que el interés radica en comparar la proporción de sujetos con una determinada característica en dos muestras independientes. Este planteamiento general es aplicable a las comparaciones realizadas en cualquiera de los siguientes diseños de un estudio:

- Un **estudio prospectivo** es aquel en el que n_1 individuos expuestos a una intervención (**ensayo clínico**) o a un potencial factor de riesgo (**estudio de cohortes**) y n_2 individuos no expuestos son seguidos a lo largo de un periodo de tiempo para determinar cuántos desarrollan la enfermedad. Los tamaños muestrales de ambos grupos n_1 y n_2 están fijados de antemano y, en el caso de un ensayo clínico, la intervención se asigna de forma aleatoria a cada sujeto. El objetivo se centra en comparar la proporción de sujetos que desarrollan la enfermedad entre los expuestos y los no expuestos.
- Un **estudio retrospectivo** (**estudio de casos y controles**) es aquel en el que m_1 sujetos con la enfermedad (casos) y m_2 sujetos libres de ella (controles) son examinados para determinar cuántos han estado previamente expuestos al potencial factor de riesgo. Bajo este diseño, el número de casos y controles está predeterminado y, en consecuencia, ha de compararse la proporción de expuestos entre los sujetos con y sin la enfermedad.
- Un **estudio transversal** es aquel en el que se selecciona un total de n individuos en un instante determinado para establecer en cada sujeto la presencia o ausencia de la exposición y la enfermedad. A diferencia de los estudios prospectivos, donde se compara la incidencia de nuevos casos de la enfermedad, los estudios transversales comparan la prevalencia de la enfermedad en un instante determinado entre expuestos y no expuestos.

Ejemplo 7.2 En el "*Second National Health and Nutrition Examination Survey*" (NHANES II), una encuesta llevada a cabo entre 1976 y 1980 en Estados Unidos, se recogieron datos del nivel de colesterol sérico total en una muestra representativa de 7.712 sujetos entre 30 y 74 años de edad sin diagnóstico previo de enfermedad cardiovascular o cáncer. Tras un seguimiento medio de 15 años, se determinó el estatus vital de cada sujeto y, en su caso, la causa de muerte. Así, en este estudio de cohortes prospectivo se registraron 254 muertes por enfermedad cardiovascular entre los 2.713 participantes con niveles de colesterol total superiores o iguales a 6,20 mmol/l (niveles altos según el "*National Cholesterol Education Program*") y 309 muertes por enfermedad cardiovascular entre los 4.999 participantes con niveles de colesterol total inferiores a 6,20 mmol/l.

Ejemplo 7.3 En el estudio de casos y controles EURAMIC, se clasificó a los sujetos según tuvieran valores superiores o inferiores al umbral de 0,90 mmol/l de colesterol HDL. De los 462 casos de infarto de miocardio con datos disponibles, 193 tuvieron valores de colesterol HDL inferiores o iguales a 0,90 mmol/l; mientras que de los 539 controles libres de la enfermedad, 158 presentaron valores de colesterol HDL inferiores a dicho umbral.

Tabla 7.1 Tabla 2×2 genérica de la asociación entre exposición y enfermedad.

Exposición	Enfermedad		Total
	Sí	No	
Sí	a	b	n_1
No	c	d	n_2
Total	m_1	m_2	n

En general, los resultados de la comparación de una variable dicotómica en dos muestras independientes suelen organizarse en una tabla 2×2 (Tabla 7.1). En este apartado suponemos que se analizan datos de un estudio prospectivo, en el que se pretende estimar la diferencia en la proporción de enfermos entre expuestos y no expuestos. Estos métodos pueden aplicarse igualmente a estudios retrospectivos, pero comparando la proporción de expuestos entre casos y controles (ver Ejemplo 7.5).

La proporción de enfermos en la muestra de sujetos expuestos viene dada por $p_1 = a/n_1$ y en la muestra de sujetos no expuestos por $p_2 = c/n_2$. Si n_1 y n_2 son suficientemente grandes, estas proporciones muestrales tenderán a distribuirse de forma normal, $p_1 \overset{\sim}{\to} N(\pi_1, \pi_1(1-\pi_1)/n_1)$ y $p_2 \overset{\sim}{\to} N(\pi_2, \pi_2(1-\pi_2)/n_2)$. Además, como ambas muestras son independientes (véase Apartado 3.4), se tiene que

$$p_1 - p_2 \overset{\sim}{\to} N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right).$$

De este resultado se desprende que $p_1 - p_2$ es un estimador puntual insesgado de la diferencia de riesgos subyacente $\pi_1 - \pi_2$ entre expuestos y no expuestos, $E(p_1 - p_2) = \pi_1 - \pi_2$. El intervalo de confianza al $100(1-\alpha)\%$ para $\pi_1 - \pi_2$ se obtiene siguiendo el mismo procedimiento utilizado para una proporción como

$$p_1 - p_2 \pm z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}},$$

que es simétrico alrededor de la diferencia de proporciones muestrales con una amplitud directamente proporcional a la estimación de su error estándar.

Para determinar si existen diferencias en la probabilidad subyacente de desarrollar la enfermedad entre los sujetos expuestos y no expuestos, se contrasta la hipótesis nula H_0: $\pi_1 = \pi_2$ frente a la hipótesis alternativa bilateral H_1: $\pi_1 \neq \pi_2$. Bajo la hipótesis nula de igualdad de proporciones H_0: $\pi_1 = \pi_2 = \pi$, se cumple que

$$p_1 - p_2 \overset{\sim}{\to} N\left(0, \pi(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right),$$

donde π corresponde a la probabilidad de enfermar común para expuestos y no expuestos. Aunque esta probabilidad π es desconocida, su valor puede estimarse mediante la proporción combinada de enfermos en ambas muestras $\bar{p} = (a+c)/(n_1+n_2) = m_1/n$. Así, el estadístico propuesto para este test es

$$z = \frac{p_1 - p_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

que bajo H_0 sigue aproximadamente una distribución normal estandarizada, lo que permitirá determinar la significación estadística de la diferencia entre proporciones.

Tabla 7.2 Muertes por enfermedad cardiovascular (ECV) durante el seguimiento del estudio NHANES II según niveles del colesterol sérico total.

Colesterol total (mmol/l)	Mortalidad por ECV		
	Sí	No	Total
≥ 6,20	254	2.459	2.713
< 6,20	309	4.690	4.999
Total	563	7.149	7.712

Ejemplo 7.4 En la Tabla 7.2 se presenta el número de muertes por enfermedad cardiovascular observadas durante el seguimiento del estudio NHANES II entre los sujetos con niveles altos y moderados-bajos de colesterol sérico total (Ejemplo 7.2). La proporción de muertes por enfermedad cardiovascular es $p_1 = 254/2.713 = 0{,}094$ en los participantes con niveles de colesterol total superiores a 6,20 mmol/l y $p_2 = 309/4.999 = 0{,}062$ en aquellos con niveles inferiores a 6,20 mmol/l. Por tanto, la estimación puntual de la diferencia de riesgos subyacente es $p_1 - p_2 = 0{,}094 - 0{,}062 = 0{,}032$ y su intervalo de confianza al 95%

$$0{,}032 \pm z_{0{,}975}\sqrt{\frac{0{,}094(1-0{,}094)}{2.713}+\frac{0{,}062(1-0{,}062)}{4.999}}$$

$$= 0{,}032 \pm 1{,}96{\cdot}0{,}007 = (0{,}019;\ 0{,}045).$$

Para el contraste bilateral de la hipótesis nula de igualdad de proporciones poblacionales H_0: $\pi_1 = \pi_2$ se emplea el estadístico

$$z = \frac{0{,}032}{\sqrt{0{,}073(1-0{,}073)\left(\frac{1}{2.713}+\frac{1}{4.999}\right)}} = 5{,}13,$$

donde $\bar{p} = 563/7.712 = 0{,}073$ es la proporción global de muertes por enfermedad cardiovascular en todos los participantes del NHANES II. El valor P del test se obtiene como $2P(Z \geq 5{,}13) = 2\{1 - \Phi(5{,}13)\} < 0{,}001$. En resumen, después de 15 años de seguimiento, la incidencia acumulada de muertes por enfermedad cardiovascular en los sujetos con niveles altos de colesterol total excedió en 32 casos por 1.000 a la de los participantes con niveles más bajos (IC al 95% entre 19 y 45 casos por 1.000), siendo esta diferencia muy significativa ($P < 0{,}001$).

Ejemplo 7.5 La Tabla 7.3 muestra los casos de infarto de miocardio y los controles del EURAMIC con valores de colesterol HDL superiores o inferiores a 0,90 mmol/l. A partir de esta tabla 2×2, se pretende comparar la proporción de sujetos con niveles bajos de colesterol HDL (≤ 0,90 mmol/l) entre casos $p_1 = c/m_1 = 193/462 = 0{,}418$ y controles $p_2 = d/m_2 = 158/539 = 0{,}293$. La diferencia de proporciones muestrales es $p_1 - p_2 = 0{,}418 - 0{,}293 = 0{,}125$ y el IC al 95% para $\pi_1 - \pi_2$ viene dado por

$$p_1 - p_2 \pm z_{0{,}975}\sqrt{\frac{p_1(1-p_1)}{m_1}+\frac{p_2(1-p_2)}{m_2}}$$

$$= 0{,}125 \pm 1{,}96\sqrt{\frac{0{,}418(1-0{,}418)}{462}+\frac{0{,}293(1-0{,}293)}{539}}$$

$$= 0{,}125 \pm 1{,}96{\cdot}0{,}030 = (0{,}065;\ 0{,}184).$$

Tabla 7.3 Colesterol HDL en los casos de infarto agudo de miocardio y los controles del estudio EURAMIC.

Colesterol HDL (mmol/l)	Infarto de miocardio		
	Caso	Control	Total
> 0,90	269	381	650
≤ 0,90	193	158	351
Total	462	539	1.001

El estadístico para el contraste bilateral de la hipótesis nula H_0: $\pi_1 = \pi_2$ se calcula como

$$z = \frac{p_1 - p_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{m_1} + \frac{1}{m_2}\right)}}$$

$$= \frac{0{,}125}{\sqrt{0{,}351(1-0{,}351)\left(\frac{1}{462} + \frac{1}{539}\right)}} = 4{,}12,$$

donde $\bar{p} = n_2/n = 351/1.001 = 0{,}351$ es la proporción total de sujetos con niveles bajos de colesterol HDL. La significación estadística del contraste es por tanto $P = 2\{1 - \Phi(4{,}12)\} < 0{,}001$. Así, los casos de infarto de miocardio son significativamente más propensos a presentar niveles bajos de colesterol HDL que los sujetos libres de la enfermedad ($P < 0{,}001$), con una diferencia de proporciones del 12,5% (IC al 95% 6,5-18,4%).

7.4 ASOCIACIÓN ESTADÍSTICA EN UNA TABLA DE CONTINGENCIA

En este apartado se presenta una prueba de significación estadística para evaluar de forma genérica la presencia o ausencia de asociación entre las variables dicotómicas representadas en una tabla 2×2. Este procedimiento no facilita estimaciones de efecto, sino únicamente valores P, y es aplicable a estudios prospectivos (marginales n_1 y n_2 fijos), retrospectivos (marginales m_1 y m_2 fijos) y transversales (tamaño muestral n fijo).

Para contrastar si las variables de una tabla 2×2 son independientes, se comparan las frecuencias observadas O_{ij} en cada celda (i, j) de la tabla con sus frecuencias esperadas E_{ij} bajo la hipótesis nula de independencia, donde $i = 1, 2$ denota la fila y $j = 1, 2$ la columna. Estas frecuencias esperadas E_{ij} se calculan como el producto de sus correspondientes marginales n_i y m_j, dividido por el tamaño muestral total n,

$$E_{ij} = \frac{n_i m_j}{n}.$$

Así, por ejemplo, si en un estudio prospectivo no hubiera asociación entre exposición y enfermedad, la frecuencia esperada de expuestos que desarrollan la enfermedad sería igual al producto del número de expuestos n_1 por la proporción combinada de enfermos m_1/n, $E_{11} = n_1 m_1/n$. Igualmente, en un estudio retrospectivo la frecuencia esperada de casos que han estado expuestos al factor de riesgo correspondería al producto del número de casos m_1 por la proporción combinada de expuestos n_1/n, $E_{11} = m_1 n_1/n$. Asimismo, en un estudio transversal la frecuencia esperada de sujetos a la vez expuestos y enfermos sería igual al producto del número total de

sujetos n por las proporciones de expuestos n_1/n y de enfermos m_1/n, $E_{11} = n(n_1/n)(m_1/n) = n_1m_1/n$. Notar, por tanto, que los valores esperados bajo la hipótesis nula de independencia coinciden en los distintos tipos de diseño.

Ejemplo 7.6 La Tabla 7.2 muestra los valores observados de la asociación entre la mortalidad por enfermedad cardiovascular y el colesterol total en el estudio prospectivo NHANES II. Si ambas variables fueran independientes, la probabilidad de morir por enfermedad cardiovascular sería igual en los sujetos con niveles altos y bajos de colesterol total. Esta probabilidad podría entonces estimarse mediante la proporción combinada de muertes en ambas muestras 563/7.712 = 0,073. Así, entre los 2.713 participantes con niveles altos de colesterol total, cabría esperar 2.713·0,073 = 198,1 muertes por enfermedad cardiovascular bajo la hipótesis nula de independencia. Aplicando este mismo razonamiento, los valores esperados en cada celda vendrían dados por

$$E_{11} = \frac{2.713 \cdot 563}{7.712} = 198{,}1,$$

$$E_{12} = \frac{2.713 \cdot 7.149}{7.712} = 2.514{,}9,$$

$$E_{21} = \frac{4.999 \cdot 563}{7.712} = 364{,}9,$$

$$E_{22} = \frac{4.999 \cdot 7.149}{7.712} = 4.634{,}1.$$

Estos valores esperados se representan en la Tabla 7.4. Notar que los marginales de la tabla de frecuencias observadas (Tabla 7.2) y esperadas (Tabla 7.4) coinciden. De hecho, una vez calculado el valor esperado en una cualquiera de las celdas, los restantes valores esperados de la tabla 2×2 quedan determinados por dichos marginales.

Para evaluar la independencia de las variables de una tabla 2×2, se comparan las frecuencias observadas y esperadas mediante el estadístico

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Tabla 7.4 Frecuencias esperadas bajo la hipótesis de independencia entre la mortalidad por enfermedad cardiovascular (ECV) y el colesterol total en el estudio NHANES II.

Colesterol total (mmol/l)	Mortalidad por ECV		
	Sí	No	Total
≥ 6,20	198,1	2.514,9	2.713
< 6,20	364,9	4.634,1	4.999
Total	563	7.149	7.712

Cuanto mayor sea la diferencia entre los valores observados y esperados, mayor será la magnitud del estadístico y, en consecuencia, se tendrá mayor evidencia en contra de la hipótesis nula de independencia. En particular, puede probarse que si las variables de la tabla 2×2 son independientes, este estadístico sigue aproximadamente una distribución chi-cuadrado con 1 grado de libertad (sólo una frecuencia esperada de la tabla 2×2 es independiente). El valor *P* del contraste corresponde entonces a la probabilidad a la derecha del estadístico χ^2 bajo la distribución χ^2_1. Esta prueba se conoce con el nombre de **test chi-cuadrado de independencia o asociación de Pearson**, y puede aplicarse siempre que los marginales de la tabla sean suficientemente grandes, de tal forma que todas las frecuencias esperadas sean superiores o iguales a 5.

Ejemplo 7.7 A partir de los valores observados y esperados bajo la hipótesis de independencia entre la mortalidad por enfermedad cardiovascular y el colesterol sérico total, se obtiene el test estadístico

$$\chi^2 = \frac{(254-198{,}1)^2}{198{,}1} + \frac{(2.459-2.514{,}9)^2}{2.514{,}9}$$
$$+ \frac{(309-364{,}9)^2}{364{,}9} + \frac{(4.690-4.634{,}1)^2}{4.634{,}1}$$
$$= 15{,}80 + 1{,}24 + 8{,}58 + 0{,}68 = 26{,}30.$$

Como las frecuencias esperadas son claramente superiores a 5, este estadístico se distribuirá aproximadamente como una chi-cuadrado con 1 grado de libertad bajo la hipótesis nula de independencia. Utilizando la Tabla 6 del Apéndice, puede comprobarse que el valor calculado del estadístico es muy superior al percentil $\chi^2_{1;0,995} = 7{,}88$, de lo cual se deduce que $P = P(\chi^2_1 \geq 26{,}30) < 0{,}005$. Así, los niveles altos de colesterol total están significativamente asociados con la mortalidad por enfermedad cardiovascular.

La hipótesis nula de independencia entre las variables de una tabla 2×2 equivale a la igualdad de dos proporciones poblacionales. De hecho, puede probarse que el estadístico χ^2 de Pearson es igual al cuadrado del estadístico *z* de la comparación de proporciones en muestras independientes, de tal forma que los valores *P* resultantes de ambos procedimientos son idénticos (la distribución chi-cuadrado con 1 grado de libertad es, por definición, igual al cuadrado de una distribución normal estandarizada). Cabría preguntarse entonces cuál es la aportación del test de independencia de Pearson. En primer lugar, los cálculos de este test no dependen del diseño utilizado para generar los datos. En segundo lugar, esta prueba puede generalizarse de forma sencilla a la comparación de múltiples proporciones en una tabla con *r* filas y *c* columnas.

Para contrastar la independencia de dos variables categóricas en una tabla *r*×*c*, se calcula el estadístico

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

donde las frecuencias esperadas $E_{ij} = n_i m_j / n$ se calculan de la misma forma que en una tabla 2×2. Bajo la hipótesis nula de independencia, dicho estadístico se distribuye aproximadamente según una chi-cuadrado con $(r-1)(c-1)$ grados de libertad. Los grados de libertad corresponden al número de frecuencias esperadas independientes para el cálculo del estadístico, una vez determinados los marginales de la tabla $r \times c$. La aproximación chi-cuadrado a la distribución del estadístico será válida si el tamaño muestral es suficientemente grande. En concreto, el criterio más aceptado para aplicar este test es que ningún valor esperado sea inferior a 1 y que no más del 20% de las celdas tengan valores esperados inferiores a 5.

Ejemplo 7.8 La Tabla 7.5 muestra las muertes por enfermedad cardiovascular entre los participantes del estudio NHANES II con un colesterol sérico total inferior a 5,20 mmol/l (nivel deseable), entre 5,20 y 6,19 mmol/l (nivel limítrofe alto) y superior o igual a 6,20 mmol/l (hipercolesterolemia). Para determinar si la incidencia de muertes por enfermedad cardiovascular difiere entre los tres grupos, se calculan en primer lugar las frecuencias esperadas mediante el producto de sus correspondientes marginales dividido por el tamaño muestral total. Estas frecuencias esperadas se presentan entre paréntesis en la Tabla 7.5. A continuación, se comparan los valores observados y esperados mediante el estadístico

$$\chi^2 = \frac{(254-198{,}1)^2}{198{,}1} + \frac{(2.459-2.514{,}9)^2}{2.514{,}9} + \frac{(174-175{,}8)^2}{175{,}8} + \frac{(2.234-2.232{,}2)^2}{2.232{,}2} + \frac{(135-189{,}1)^2}{189{,}1} + \frac{(2.456-2.401{,}9)^2}{2.401{,}9}$$

$$= 15{,}80 + 1{,}24 + 0{,}02 + 0{,}00 + 15{,}50 + 1{,}22 = 33{,}79.$$

Dado que las frecuencias esperadas son superiores a 5, puede utilizarse la distribución chi-cuadrado con $(3-1)(2-1) = 2$ grados de libertad (Tabla 6 del Apéndice) para obtener un valor $P = P(\chi^2_2 \geq 33{,}79) < 0{,}005$. Esto es, la incidencia de muertes por enfermedad cardiovascular difiere significativamente entre los tres grupos, obteniéndose una incidencia acumulada en los 15 años de seguimiento de 52, 72 y 94 muertes por cada 1.000 participantes con niveles deseables, limítrofes altos y altos de colesterol total, respectivamente.

Tabla 7.5 Frecuencias observadas (esperadas) de muertes por enfermedad cardiovascular (ECV) entre los participantes del NHANES II con niveles de colesterol total < 5,20, 5,20-6,19 y ≥ 6,20 mmol/l.

Colesterol total (mmol/l)	Mortalidad por ECV		Total
	Sí	No	
≥ 6,20	254 (198,1)	2.459 (2.514,9)	2.713
5,20-6,19	174 (175,8)	2.234 (2.232,2)	2.408
< 5,20	135 (189,1)	2.456 (2.401,9)	2.591
Total	563	7.149	7.712

7.5 TEST DE TENDENCIA EN UNA TABLA *r*×2

A partir de una tabla *r*×2, el test chi-cuadrado de Pearson permite contrastar la hipótesis nula de igualdad de proporciones H_0: $\pi_1 = \pi_2 = ... = \pi_r$ frente a la hipótesis alternativa H_1: $\pi_i \neq \pi_j$, donde i y j son 2 muestras cualesquiera. Un resultado significativo de esta prueba indicaría que al menos 2 de las r proporciones poblacionales son heterogéneas. En el caso de que los grupos o muestras estén intrínsecamente ordenados, cabría preguntarse además si estas proporciones siguen alguna tendencia determinada a lo largo de los grupos. En este apartado se presenta un test específico para detectar la existencia de un gradiente o componente lineal (creciente o decreciente) entre las proporciones de los sucesivos grupos.

En primer lugar, se asigna una puntuación s_i a cada una de las muestras ordenadas. Esta puntuación puede representar un atributo numérico del grupo (ver Ejemplo 7.9), o simplemente tomar los valores 1, 2, ..., r indicando el orden de los grupos. A continuación, se relacionan las proporciones observadas p_i con sus correspondientes puntuaciones s_i mediante el estadístico

$$\chi^2 = \frac{\left(\sum_{i=1}^{r} n_i (p_i - \bar{p})(s_i - \bar{s})\right)^2}{\bar{p}(1-\bar{p})\sum_{i=1}^{r} n_i (s_i - \bar{s})^2},$$

donde n_i es el tamaño de cada muestra, $n = \sum n_i$, $\bar{p} = \sum n_i p_i / n$ es la proporción combinada en todas las muestras y $\bar{s} = \sum n_i s_i / n$ es la puntuación media. Notar que si las proporciones observadas tienden a aumentar o disminuir con las puntuaciones, el numerador del estadístico será grande. Si, por el contrario, las proporciones no varían en función de la puntuación de cada grupo, el numerador estará próximo a 0. Bajo la hipótesis nula de ausencia de una componente lineal en la tendencia, el estadístico anterior seguirá aproximadamente una distribución chi-cuadrado con 1 grado de libertad. Esta prueba se conoce genéricamente como **test chi-cuadrado de tendencia** y, a diferencia del test de independencia o asociación, puede aplicarse incluso cuando algunas muestras tengan un tamaño reducido, basta con que la muestra total sea suficientemente grande y la proporción combinada no muy extrema, $n\bar{p}(1-\bar{p}) \geq 5$. Finalmente, cabe reseñar que el test de tendencia no permite contrastar la idoneidad de la relación lineal; este test únicamente determina la existencia de una componente lineal significativa, independientemente de cuál sea la relación subyacente.

> ***Ejemplo 7.9*** En el ejemplo anterior se detectaron diferencias significativas en el riesgo de muerte por enfermedad cardiovascular entre los participantes del NHANES II con niveles de colesterol total < 5,20, 5,20-6,19 y ≥ 6,20 mmol/l. De hecho, se observa un claro incremento en las incidencias acumuladas $p_1 = 135/2.591 = 0{,}052$, $p_2 = 174/2.408 = 0{,}072$ y $p_3 = 254/2.713 = 0{,}094$ de las sucesivas categorías (Figura 7.1). Para contrastar si esta tendencia creciente es significativa, se asignan las puntuaciones $s_1 = 4{,}65$, $s_2 = 5{,}72$ y $s_3 = 6{,}90$ correspondientes a la mediana del colesterol total de cada categoría. Aunque podrían asignarse las puntuaciones 1, 2 y 3, es preferible utilizar una medida de tendencia central de cada categoría (media o mediana) para preservar la distancia entre las mismas. Así, el numerador del estadístico del test de tendencia vendría dado por
>
> $$\begin{aligned} N = \{&2.591(0{,}052-0{,}073)(4{,}65-5{,}78) \\ &+ 2.408(0{,}072-0{,}073)(5{,}72-5{,}78) \\ &+ 2.713(0{,}094-0{,}073)(6{,}90-5{,}78)\}^2 = 15.364{,}56 \end{aligned}$$

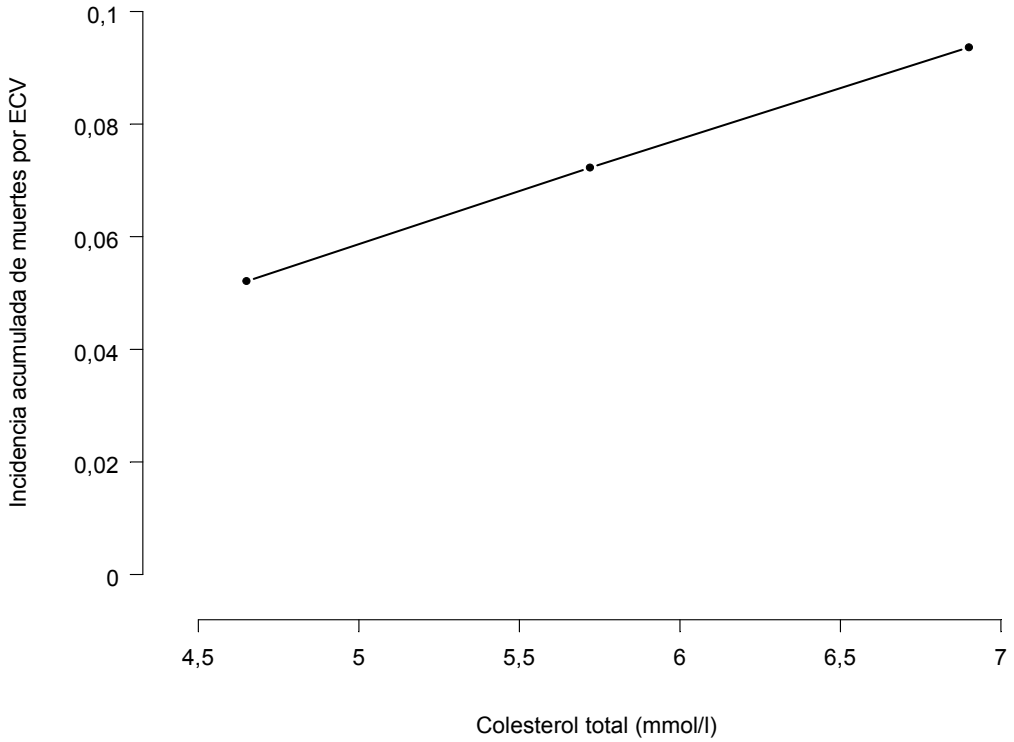


Figura 7.1 **Incidencia acumulada de muertes por enfermedad cardiovascular (ECV) en 15 años de seguimiento del estudio NHANES II según niveles de colesterol total < 5,20, 5,20-6,19 y ≥ 6,20 mmol/l.**

y el denominador por

$$\begin{aligned} D = 0{,}073(1-0{,}073)\{&2.591(4{,}65-5{,}78)^2 \\ &+ 2.408(5{,}72-5{,}78)^2 \\ &+ 2.713(6{,}90-5{,}78)^2\} = 454{,}78, \end{aligned}$$

donde $\bar{p}$ = 563/7.712 = 0,073 es la proporción global de muertes por enfermedad cardiovascular en todos los participantes del NHANES II y $\bar{s}$ = (2.591·4,65 + 2.408·5,72 + 2.713·6,90)/7.712 = 5,78 es la puntuación media. El estadístico resulta entonces $\chi^2 = N/D = 33{,}78$, que corresponde a un valor $P = P(\chi^2_1 \geq 33{,}78) < 0{,}005$ en la distribución chi-cuadrado con 1 grado de libertad (Tabla 6 del Apéndice). Este resultado confirma que el riesgo de mortalidad por enfermedad cardiovascular aumenta significativamente al aumentar el nivel de colesterol total.

7.6 MEDIDAS DE EFECTO EN UNA TABLA DE CONTINGENCIA

En epidemiología y en otras aplicaciones del análisis de datos en salud pública, no sólo interesa determinar el grado de significación estadística sino también obtener estimadores de efecto o medidas de la magnitud de la asociación. A partir de una tabla 2×2 pueden obtenerse distintas medidas de efecto, tales como la diferencia de riesgos, el riesgo relativo y el odds ratio. La **diferencia de riesgos** o proporciones, que ya se discutió en el Apartado 7.3, permite determinar la diferencia en la tasa de incidencia o prevalencia de la enfermedad entre los sujetos expuestos y no expuestos en un estudio prospectivo o transversal, respectivamente. En este apartado se revisan los métodos de inferencia sobre el riesgo relativo y el odds ratio, así como sus respectivos ámbitos de aplicación.

7.6.1 Riesgo relativo

El riesgo relativo o razón de riesgos es la medida de efecto más utilizada en estudios prospectivos para comparar la incidencia de la enfermedad entre expuestos y no expuestos, y se define como

$$\psi = \frac{\pi_1}{\pi_2} = \frac{P(D \mid E)}{P(D \mid E^c)},$$

donde $\pi_1 = P(D|E)$ y $\pi_2 = P(D|E^c)$ representan la probabilidad de desarrollar la enfermedad D entre los sujetos expuestos E y no expuestos E^c, respectivamente. Así, el riesgo relativo determina cuántas veces es más frecuente la enfermedad en expuestos que en no expuestos. Se trata, por tanto, de una medida de efecto multiplicativa que puede tomar cualquier valor no negativo, de tal forma que:

- $\psi = 1$ indica la misma probabilidad de enfermar en expuestos y no expuestos $P(D|E) = P(D|E^c)$; es decir, la exposición y la enfermedad son independientes. Cuanto más alejado esté ψ de 1 en cualquier sentido, mayor será la magnitud de la asociación entre exposición y enfermedad.
- $\psi > 1$ indica una mayor probabilidad de desarrollar la enfermedad en expuestos que en no expuestos. Por ejemplo, si $\psi = 1{,}25$, los sujetos expuestos tienen 1,25 veces más riesgo o son un 25% más propensos a desarrollar la enfermedad que los no expuestos ($100(\psi - 1) = 100(1{,}25 - 1) = 25\%$).
- $\psi < 1$ indica una menor probabilidad de contraer la enfermedad en expuestos que en no expuestos. Por ejemplo, si $\psi = 0{,}80$, los sujetos expuestos son un 20% menos propensos a desarrollar la enfermedad que los no expuestos ($100(0{,}80 - 1) = -20\%$).
- Un valor de ψ y su inverso $1/\psi$ representan el mismo nivel de asociación, pero en sentido opuesto. Por ejemplo, si $\psi = 4$, los sujetos expuestos son 4 veces más propensos a desarrollar la enfermedad que los no expuestos, o equivalentemente los no expuestos son un 75% menos propensos a contraer la enfermedad que los expuestos ($100(1/\psi - 1) = 100(0{,}25 - 1) = -75\%$).

Esta medida de efecto también puede aplicarse a estudios transversales en términos de la razón de prevalencias. Sin embargo, y al igual que ocurría con la diferencia de riesgos, el riesgo relativo no es directamente estimable a partir de estudios retrospectivos ya que la proporción de casos está predeterminada por el propio diseño del estudio.

A partir de los datos observados en una tabla 2×2 (Tabla 7.1), un estimador puntual del riesgo relativo viene determinado por

$$RR = \frac{p_1}{p_2} = \frac{a/n_1}{c/n_2},$$

que corresponde al cociente entre la proporción de enfermos en la muestra de sujetos expuestos $p_1 = a/n_1$ y no expuestos $p_2 = c/n_2$.

> ***Ejemplo 7.10*** De la Tabla 7.2 se desprende que la proporción de muertes por enfermedad cardiovascular es $p_1 = 254/2.713 = 0{,}094$ en los participantes del estudio NHANES II con niveles de colesterol total superiores a 6,20 mmol/l y $p_2 = 309/4.999 = 0{,}062$ en aquellos con niveles inferiores a 6,20 mmol/l. Así, la estimación puntual del riesgo relativo es
>
> $$RR = 0{,}094/0{,}062 = 1{,}51;$$

es decir, la incidencia acumulada de muertes por enfermedad cardiovascular en 15 años de seguimiento es un 51% superior en los sujetos con niveles altos de colesterol total que en quienes tienen niveles más bajos.

El cálculo de un intervalo de confianza y un test de hipótesis para ψ no resulta sencillo ya que la distribución muestral de su estimador RR es muy asimétrica, particularmente cuando el riesgo relativo subyacente dista mucho del valor nulo 1. Para solventar este problema de inferencia, es preferible trabajar con el logaritmo natural del riesgo relativo, cuya distribución presenta una mayor simetría. De hecho, puede probarse que si los tamaños de ambas muestras son suficientemente grandes $n_1\pi_1(1-\pi_1) \geq 5$ y $n_2\pi_2(1-\pi_2) \geq 5$, el $\log(RR)$ tiende a distribuirse de forma normal con media $\log(\psi)$ y varianza aproximada $1/a - 1/n_1 + 1/c - 1/n_2$,

$$\log(RR) \overset{\sim}{\to} N\left(\log(\psi), \frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_2}\right).$$

Ejemplo 7.11 En las Figuras 7.2(*a*) y (*b*) se presentan las distribuciones muestrales del RR y del $\log(RR)$ de mortalidad por enfermedad cardiovascular entre los sujetos con un colesterol total ≥ 6,20 y < 6,20 mmol/l obtenidos a partir de 1000 muestras aleatorias simples de tamaño 500 del estudio NHANES II. Como puede observarse, ambas distribuciones están centradas alrededor de los parámetros subyacentes 1,51 y $\log(1{,}51) = 0{,}42$ en todos los participantes del estudio. Sin embargo, la distribución muestral del RR presenta una clara asimetría, mientras que el $\log(RR)$ se distribuye de forma aproximadamente normal.

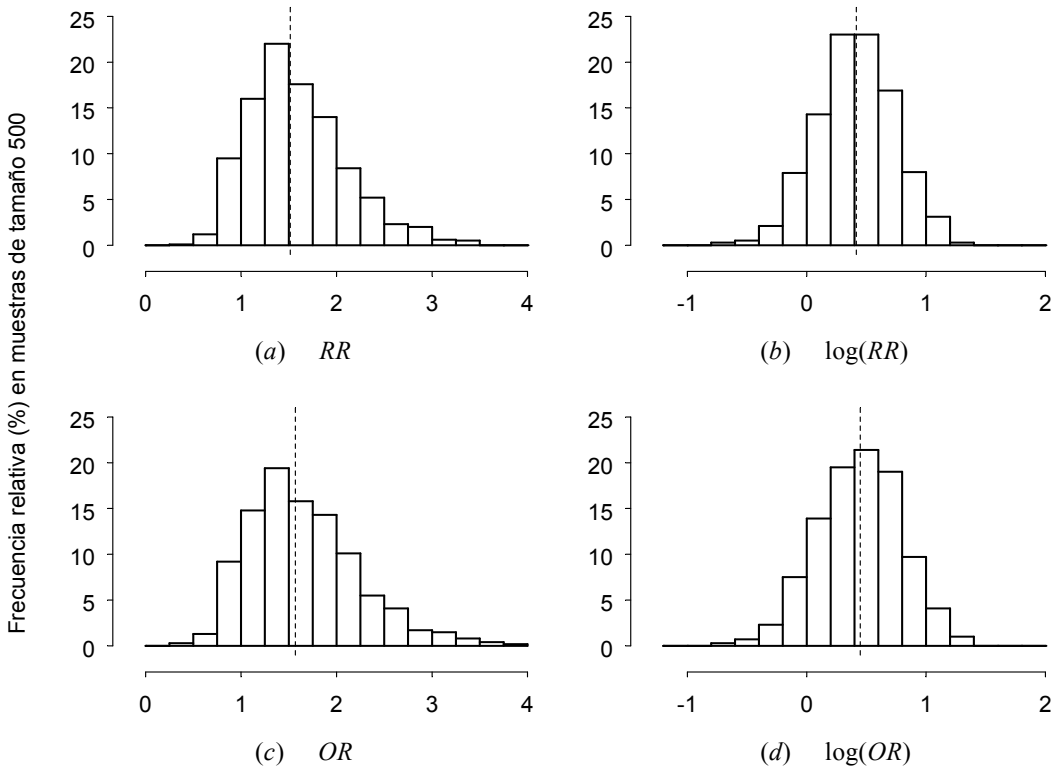


Figura 7.2 **Distribución muestral del *RR* (*a*), log(*RR*) (*b*), *OR* (*c*) y log(*OR*) (*d*) de mortalidad por enfermedad cardiovascular entre los sujetos con un colesterol total ≥ 6,20 y < 6,20 mmol/l en 1000 muestras aleatorias simples de tamaño $n = 500$ obtenidas a partir del estudio NHANES II. Las líneas verticales en trazo discontinuo corresponden a los parámetros subyacentes $\psi = 1{,}51$, $\log(\psi) = 0{,}42$, $\omega = 1{,}57$ y $\log(\omega) = 0{,}45$.**

En base a la distribución aproximadamente normal del log(RR), puede obtenerse un intervalo de confianza al 100(1 – α)% para el log(ψ) como

$$\log(RR) \pm z_{1-\alpha/2}\sqrt{\frac{1}{a}-\frac{1}{n_1}+\frac{1}{c}-\frac{1}{n_2}}\,.$$

Deshaciendo la transformación logarítmica en ambos límites de este intervalo, el IC al 100(1 – α)% para el riesgo relativo subyacente ψ queda entonces determinado por

$$\exp\left\{\log(RR) \pm z_{1-\alpha/2}\sqrt{\frac{1}{a}-\frac{1}{n_1}+\frac{1}{c}-\frac{1}{n_2}}\right\}.$$

Notar que por tratarse de una medida de efecto multiplicativa, el intervalo de confianza no es simétrico alrededor de la estimación puntual RR. Asimismo, la hipótesis nula de no efecto H_0: $\psi = 1$ puede contrastarse frente a la hipótesis alternativa bilateral H_1: $\psi \neq 1$ mediante el estadístico

$$z = \frac{\log(RR)}{\sqrt{\frac{1}{a}-\frac{1}{n_1}+\frac{1}{c}-\frac{1}{n_2}}},$$

que bajo H_0 sigue aproximadamente una distribución normal estandarizada. Conviene destacar que esta hipótesis nula H_0: $\psi = 1$ coincide con la hipótesis H_0: $\pi_1 = \pi_2$ de la comparación de proporciones en dos muestras independientes, así como con la hipótesis nula de independencia del test χ^2 de Pearson en una tabla 2×2. Este test es, por tanto, un procedimiento alternativo para contrastar la misma hipótesis nula, que arroja resultados muy similares cuando el tamaño muestral es grande. No obstante, si la muestra es moderada o pequeña, el valor P de este test puede resultar algo impreciso, en cuyo caso es preferible utilizar los contrastes basados en la diferencia de proporciones o el test χ^2 de Pearson.

Ejemplo 7.12 Retomando de nuevo los datos del NHANES II presentados en la Tabla 7.2, el IC al 95% para el log(ψ) resulta ser

$$\log(1{,}51) \pm z_{0{,}975}\sqrt{\frac{1}{254}-\frac{1}{2.713}+\frac{1}{309}-\frac{1}{4.999}}$$

$$= 0{,}415 \pm 1{,}96 \cdot 0{,}081 = (0{,}256;\ 0{,}574).$$

Aplicando la exponencial a ambos límites del intervalo, el IC al 95% para ψ vendría dado por

$$(\exp\{0{,}256\},\ \exp\{0{,}574\}) = (1{,}29;\ 1{,}78),$$

que es ligeramente asimétrico respecto a la estimación puntual RR = 1,51. El estadístico para el contraste de la hipótesis de no efecto H_0: $\psi = 1$ es

$$z = \frac{\log(1{,}51)}{\sqrt{\frac{1}{254}-\frac{1}{2.713}+\frac{1}{309}-\frac{1}{4.999}}} = 5{,}11,$$

que corresponde a un valor P bilateral $2P(Z \geq 5{,}11) = 2\{1 - \Phi(5{,}11)\} < 0{,}001$. Como cabía esperar, este test arroja un resultado significativo dado que el valor nulo $\psi = 1$ queda fuera de los límites del intervalo de confianza. Así, se concluye que los sujetos con niveles de colesterol total superiores a 6,20 mmol/l presentan un 51% (IC al 95% 29-78%; $P < 0{,}001$) más riesgo de morir por enfermedad cardiovascular que quienes tienen niveles inferiores a este umbral.

7.6.2 Odds ratio

La frecuencia de una enfermedad D en una población expuesta a un factor E suele medirse mediante la probabilidad $P(D|E)$ de que un sujeto de la población expuesta presente o desarrolle dicha enfermedad. Otra medida de frecuencia de la enfermedad vendría dada por

$$\frac{P(D \mid E)}{P(D^c \mid E)},$$

que se conoce como el **odds** de estar enfermo entre los expuestos y puede estimarse mediante

$$\frac{a/n_1}{b/n_1} = \frac{a}{b}.$$

Ejemplo 7.13 La proporción de muertes por enfermedad cardiovascular entre los participantes del NHANES II con niveles de colesterol total ≥ 6,20 mmol/l es

$$\frac{a}{n_1} = \frac{254}{2.713} = 0{,}094;$$

es decir, aproximadamente 1 de cada 11 sujetos con niveles altos de colesterol fallecerá por enfermedad cardiovascular a los 15 años de seguimiento. Por otra parte, el odds de morir por enfermedad cardiovascular entre estos sujetos es

$$\frac{a}{b} = \frac{254}{2.459} = 0{,}103;$$

esto es, por cada 10 sujetos con niveles altos de colesterol que no fallezcan por enfermedad cardiovascular, habrá aproximadamente 1 muerte por dicha causa a los 15 años de seguimiento. Aunque la interpretación difiere, ambas medidas de frecuencia facilitan la misma información.

De forma equivalente, el odds de estar enfermo entre los no expuestos se define como

$$\frac{P(D \mid E^c)}{P(D^c \mid E^c)},$$

y el odds ratio o razón de odds entre expuestos y no expuestos queda entonces determinado por

$$\omega = \frac{P(D \mid E)/P(D^c \mid E)}{P(D \mid E^c)/P(D^c \mid E^c)} = \frac{P(D \mid E)\,P(D^c \mid E^c)}{P(D^c \mid E)\,P(D \mid E^c)},$$

cuya estimación puntual

$$OR = \frac{(a/n_1)(d/n_2)}{(b/n_1)(c/n_2)} = \frac{ad}{bc}$$

coincide con la razón del producto cruzado de las celdas de una tabla 2×2.

Al igual que el riesgo relativo, el odds ratio es una medida de efecto multiplicativa que toma valores no negativos. Si $\omega = 1$, las probabilidades de enfermar en expuestos y no expuestos coinciden $P(D|E) = P(D|E^c)$, indicando independencia entre exposición y enfermedad. Si por el contrario $\omega > 1$, la probabilidad de contraer la enfermedad será mayor en expuestos que en no expuestos; mientras que si $\omega < 1$, la probabilidad de desarrollar la enfermedad será menor en expuestos que en no expuestos. Resulta sencillo probar que el odds ratio estará siempre más

alejado del valor nulo 1 que el riesgo relativo. Además, si la probabilidad de enfermar es baja en los sujetos expuestos y no expuestos, de tal forma que $P(D^c|E)$ y $P(D^c|E^c)$ estén próximas a 1, el odds ratio será entonces aproximadamente igual al riesgo relativo.

Ejemplo 7.14 A partir de los datos observados en el estudio NHANES II (Tabla 7.2), la estimación puntual del odds ratio es

$$OR = \frac{254 \cdot 4.690}{2.459 \cdot 309} = 1{,}57.$$

Por tanto, el odds de mortalidad por enfermedad cardiovascular es un 57% superior en los sujetos con niveles de colesterol total superiores a 6,20 mmol/l que en aquellos con niveles inferiores a 6,20 mmol/l. Este odds ratio es ligeramente mayor que el riesgo relativo $RR = 1{,}51$ estimado en el Ejemplo 7.10, aunque la diferencia no es muy grande porque la incidencia acumulada es relativamente baja tanto en expuestos 254/2.713 = 0,094 como en no expuestos 309/4.999 = 0,062.

De la propia definición de ω, resulta obvio que el odds ratio puede estimarse a partir de estudios prospectivos y transversales, ya que ambos diseños facilitan estimaciones de las probabilidades de enfermar $P(D|E)$ y $P(D|E^c)$. Aplicando la definición de probabilidad condicional (ver Tema 2), el odds ratio puede expresarse a su vez en términos de la probabilidad de estar expuesto en enfermos y no enfermos como

$$\omega = \frac{P(D \mid E)\, P(D^c \mid E^c)}{P(D^c \mid E)\, P(D \mid E^c)} = \frac{P(D \cap E)\, P(D^c \cap E^c)}{P(D^c \cap E)\, P(D \cap E^c)}$$

$$= \frac{P(E \mid D)\, P(E^c \mid D^c)}{P(E \mid D^c)\, P(E^c \mid D)},$$

de donde se desprende que el odds ratio es también estimable a partir de estudios retrospectivos, aun cuando estos diseños no facilitan información alguna sobre las probabilidades absolutas de enfermar en expuestos y no expuestos. Por supuesto, la estimación puntual del odds ratio en estudios retrospectivos coincide con la razón del producto cruzado

$$OR = \frac{(a/m_1)(d/m_2)}{(b/m_2)(c/m_1)} = \frac{ad}{bc}.$$

Los estudios retrospectivos suelen conducirse en enfermedades de baja incidencia, para las cuales la obtención de un número suficiente de casos requeriría de estudios prospectivos con gran tamaño muestral y amplio seguimiento. En tales circunstancias, si la incidencia de la enfermedad es baja y el diseño del estudio retrospectivo es adecuado (esto es, casos incidentes y controles representativos del nivel de exposición en la población libre de enfermedad), el odds ratio constituye una buena aproximación al riesgo relativo subyacente. En adelante, el odds ratio se utilizará e interpretará como estimación del riesgo relativo, asumiendo que se cumplen las condiciones citadas anteriormente.

Ejemplo 7.15 En el estudio EURAMIC se obtuvo una muestra de casos incidentes de infarto de miocardio procedentes de las unidades de cuidados intensivos y una muestra aleatoria de controles seleccionados a partir de la población de referencia. El número de casos y controles con valores de colesterol HDL superiores o inferiores a 0,90 mmol/l se presenta en la Tabla 7.3. Aunque el diseño retrospectivo del estudio no permite conocer la

incidencia de infartos entre los sujetos con valores altos y bajos de colesterol HDL, sí es posible obtener una medida relativa de la asociación entre el colesterol HDL y el riesgo de infarto de miocardio mediante el odds ratio

$$OR = \frac{269 \cdot 158}{381 \cdot 193} = 0{,}58.$$

Como la incidencia de infarto agudo de miocardio es relativamente baja en la población de hombres adultos, este odds ratio puede interpretarse como un riesgo relativo y concluir que los sujetos con un colesterol HDL superior a 0,90 mmol/l presentan un 42% menos riesgo de padecer un infarto de miocardio que aquellos con un colesterol HDL inferior a 0,90 mmol/l (100(0,58 – 1) = –42%).

El odds ratio es una medida de efecto multiplicativa cuya distribución muestral es notablemente asimétrica (Figura 7.2(*c*)), mientras que su transformación logarítmica log(*OR*) tiende a distribuirse normalmente (Figura 7.2(*d*)) con una varianza aproximadamente igual a la suma de los inversos de las frecuencias de una tabla 2×2

$$\text{var}\{\log(OR)\} \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}.$$

Utilizando esta aproximación normal a la distribución muestral del log(*OR*) y deshaciendo a continuación la transformación logarítmica, se obtiene el intervalo de confianza al 100(1 – α)% para el odds ratio subyacente ω

$$\exp\left\{\log(OR) \pm z_{1-\alpha/2}\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right\},$$

que no es simétrico alrededor de la estimación puntual *OR*. De forma análoga, la significación estadística del contraste bilateral de la hipótesis nula H_0: $\omega = 1$ se obtiene a partir del estadístico

$$z = \frac{\log(OR)}{\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}},$$

que bajo H_0 sigue aproximadamente una distribución normal estandarizada.

Ejemplo 7.16 Continuando con el ejemplo anterior, el IC al 95% para el odds ratio de infarto agudo de miocardio entre los sujetos con niveles altos y bajos de colesterol HDL es

$$\exp\left\{\log(0{,}58) \pm z_{0{,}975}\sqrt{\frac{1}{269} + \frac{1}{381} + \frac{1}{193} + \frac{1}{158}}\right\}$$
$$= \exp(-0{,}55 \pm 1{,}96 \cdot 0{,}134) = (0{,}44;\ 0{,}75).$$

Por tanto, puede afirmarse con una confianza del 95% que los sujetos con niveles altos de colesterol HDL tienen entre un 25 y un 56% menos riesgo de padecer un infarto de miocardio que quienes tienen niveles más bajos (100(0,75 – 1) = –25% y 100(0,44 – 1) = –56%). Asimismo, el contraste bilateral de la hipótesis de no efecto H_0: $\omega = 1$ mediante el estadístico

$$z = \frac{\log(0{,}58)}{\sqrt{\frac{1}{269} + \frac{1}{381} + \frac{1}{193} + \frac{1}{158}}} = -4{,}10$$

arroja un resultado muy significativo $P = 2P(Z \leq -4{,}10) = 2\{1 - \Phi(4{,}10)\} < 0{,}001$. Notar que este test es equivalente al contraste de hipótesis realizado en el Ejemplo 7.5 sobre la igualdad en la proporción de sujetos con niveles bajos de colesterol HDL entre los casos de infarto y los sujetos libres de la enfermedad, de tal forma que los valores P resultantes de ambos procedimientos son virtualmente idénticos.

7.7 COMPARACIÓN DE PROPORCIONES EN DOS MUESTRAS DEPENDIENTES

Hasta este punto se han presentado distintos métodos para la comparación de proporciones a partir de muestras independientes. Con cierta frecuencia, sin embargo, suelen emplearse muestras dependientes, que surgen tanto de observaciones tomadas en los mismos sujetos como en distintos sujetos emparejados de acuerdo a determinados factores pronósticos. En el Apartado 6.4 del tema anterior, se presentaron diversos diseños o mecanismos de generación de datos dependientes. En general, el propósito de los diseños emparejados es aumentar la precisión de las comparaciones y, en mayor medida, mejorar la validez de las inferencias al controlar por posibles factores de confusión. En este apartado se aborda el tratamiento estadístico de datos binarios o dicotómicos procedentes de parejas dependientes.

La muestra consiste en n parejas dependientes o correlacionadas, donde cada pareja está compuesta por dos observaciones de una variable dicotómica procedentes de distintas poblaciones. Así, por ejemplo, en comparaciones antes y después de un tratamiento, cada pareja de datos está constituida por la respuesta en un mismo sujeto antes y después de dicho tratamiento. Igualmente, en un estudio de casos y controles emparejados, cada pareja de observaciones está formada por la presencia o ausencia de exposición en cada caso y su correspondiente control. Para simplificar la presentación, nos centraremos en adelante en un estudio de casos y controles emparejados.

Para preservar el emparejamiento muestral, la unidad de análisis será cada pareja y no cada individuo. Así, la organización de los datos por individuo mediante la Tabla 7.1 no resulta adecuada ya que se pierde la información relativa al emparejamiento. La forma apropiada de presentar los datos se muestra en la Tabla 7.6. Cada unidad de esta tabla representa una pareja, de tal forma que hay a parejas donde ambos caso y control están expuestos al factor de riesgo, b parejas donde el caso está expuesto y el control no, c parejas donde el control está expuesto y el caso no, y d parejas donde ninguno está expuesto. Las $a + d$ parejas donde ambos o ninguno de los miembros están expuestos se denominan parejas concordantes, mientras las restantes $b + c$ parejas son discordantes.

Ejemplo 7.17 En el Ejemplo 6.12 se seleccionaron 50 casos de infarto de miocardio y 50 controles del estudio EURAMIC emparejados por grupos quinquenales de edad. A partir de sus valores del colesterol HDL (Tabla 6.1), se desprende que hay 23 parejas donde el caso de infarto y su correspondiente control presentan niveles altos de colesterol HDL (superior a 0,90 mmol/l), 6 parejas donde el caso tiene un nivel alto y el control bajo, 17 parejas donde el caso tiene un nivel bajo y el control alto, y 4 parejas donde ambos presentan niveles bajos de colesterol HDL. Los datos de este estudio de casos y controles emparejados se resumen en la Tabla 7.7.

Tabla 7.6 Tabla de contingencia en un estudio de casos y controles emparejados.

	Controles		
Casos	**Expuestos**	**No expuestos**	**Total**
Expuestos	a	b	$a + b$
No expuestos	c	d	$c + d$
Total	$a + c$	$b + d$	n

Tabla 7.7 Colesterol HDL en 50 casos de infarto de miocardio y 50 controles del estudio EURAMIC emparejados por grupos quinquenales de edad.

	Controles		
Casos	**HDL > 0,90 mmol/l**	**HDL ≤ 0,90 mmol/l**	**Total**
HDL > 0,90 mmol/l	23	6	29
HDL ≤ 0,90 mmol/l	17	4	21
Total	40	10	50

Con objeto de evaluar la asociación entre exposición y enfermedad controlando por aquellos factores de confusión utilizados en el emparejamiento, cada caso ha de ser comparado con su correspondiente control; es decir, las comparaciones deben estar condicionadas a cada pareja. Por ello, los pares concordantes, donde ambos miembros están o no expuestos, no aportan información sobre la asociación a estudio y, en consecuencia, el análisis estadístico se limita a las parejas discordantes. La probabilidad de observar una pareja con el caso expuesto y el control no expuesto viene dada por $P(E|D)P(E^c|D^c)$, mientras que la probabilidad de obtener una pareja con el control expuesto y el caso no expuesto es $P(E|D^c)P(E^c|D)$. Así, dado que una pareja es discordante, la probabilidad de que el caso esté expuesto es

$$\pi = \frac{P(E \mid D)\, P(E^c \mid D^c)}{P(E \mid D)\, P(E^c \mid D^c) + P(E \mid D^c)\, P(E^c \mid D)} = \frac{\omega}{\omega + 1},$$

donde la última igualdad refleja su relación con el odds ratio subyacente ω. Despejando ω de esta expresión, se tiene que

$$\omega = \frac{\pi}{1 - \pi}.$$

Como la probabilidad π puede estimarse mediante la proporción observada $b/(b + c)$ de parejas discordantes donde el caso está expuesto, la estimación puntual del odds ratio de enfermar entre expuestos y no expuestos es

$$OR = \frac{b/(b+c)}{1 - b/(b+c)} = \frac{b/(b+c)}{c/(b+c)} = \frac{b}{c},$$

que coincide con la razón entre ambos tipos de pares discordantes. Si el número de parejas discordantes b con el caso expuesto es superior al número de parejas discordantes c con el control expuesto, el odds ratio será mayor de 1 y la exposición estará directamente asociada con la enfermedad; mientras que si b es inferior a c, el odds ratio será menor de 1 y la exposición estará inversamente asociada con la enfermedad.

Al igual que en muestras independientes, el $\log(OR)$ también se distribuye de forma aproximadamente normal en muestras dependientes, con media $\log(\omega)$ y varianza aproximada $1/b + 1/c$. El intervalo de confianza al $100(1 - \alpha)\%$ para el odds ratio subyacente ω resulta entonces

$$\exp\left\{\log(OR) \pm z_{1-\alpha/2}\sqrt{\frac{1}{b} + \frac{1}{c}}\right\}.$$

Ejemplo 7.18 En la Tabla 7.7 se tienen 6 parejas discordantes donde sólo el caso de infarto tiene un nivel alto de colesterol HDL y 17 parejas discordantes donde sólo el

control presenta un nivel alto, de lo cual se deduce que la estimación puntual del odds ratio es

$$OR = \frac{6}{17} = 0{,}35,$$

y su IC al 95%

$$\exp\left\{\log(0{,}35) \pm z_{0{,}975}\sqrt{\frac{1}{6}+\frac{1}{17}}\right\}$$
$$= \exp(-1{,}04 \pm 1{,}96 \cdot 0{,}475) = (0{,}14;\ 0{,}90).$$

Por tanto, el riesgo de infarto agudo de miocardio es inferior en un 65% (IC al 95% 10-86%) en los sujetos con niveles de colesterol HDL > 0,90 mmol/l respecto a aquellos con niveles ≤ 0,90 mmol/l. La conclusión de este estudio emparejado es consistente con la obtenida en los Ejemplos 7.15 y 7.16 en la muestra completa e independiente de casos y controles del estudio EURAMIC. Aunque esta estimación de efecto es más imprecisa por disponer únicamente de 50 parejas, será menos propensa a posibles sesgos derivados de la diferencia de edad entre casos y controles.

El método más extendido para contrastar la hipótesis nula de independencia entre exposición y enfermedad en un estudio emparejado consiste en comparar la frecuencia observada b de pares discordantes donde el caso está expuesto con su frecuencia esperada bajo la hipótesis nula. Si no hubiera asociación entre exposición y enfermedad, esta frecuencia esperada sería simplemente la mitad del número total de parejas discordantes $(b+c)/2$, con lo cual el estadístico del contraste viene determinado por

$$\chi^2 = \frac{\{b - E(b)\}^2}{\text{var}(b)} = \frac{\left(b - \frac{b+c}{2}\right)^2}{\frac{b+c}{4}} = \frac{(b-c)^2}{b+c}.$$

Bajo la hipótesis nula de no efecto, este estadístico sigue aproximadamente una distribución chi-cuadrado con 1 grado de libertad, lo que permite obtener el valor P como la probabilidad a la derecha del estadístico χ^2 en la distribución χ^2_1. Este contraste se conoce como el **test de McNemar** y se aplica cuando la varianza de b bajo la hipótesis nula es $\text{var}(b) = (b + c)\,\pi(1 - \pi) = (b + c)/4 \geq 5$; es decir, cuando el número de parejas discordantes es superior o igual a 20.

Ejemplo 7.19 El estadístico del test de McNemar en la Tabla 7.7 toma el valor

$$\chi^2 = \frac{(6-17)^2}{6+17} = 5{,}26.$$

A partir de la distribución chi-cuadrado con 1 grado de libertad (Tabla 6 del Apéndice), puede comprobarse que este estadístico está comprendido entre los percentiles $\chi^2_{1;0{,}975} = 5{,}02$ y $\chi^2_{1;0{,}99} = 6{,}63$, de lo cual se tiene que $0{,}01 < P < 0{,}025$. Así, el riesgo de infarto agudo de miocardio difiere significativamente entre los sujetos con niveles de colesterol HDL superiores e inferiores a 0,90 mmol/l.

La inferencia sobre proporciones puede extenderse a estudios donde se empareja más de un sujeto por muestra (por ejemplo, un estudio de casos y controles donde cada caso está emparejado con múltiples controles, o un ensayo clínico donde cada paciente que recibe un nuevo tratamiento está emparejado con varios pacientes bajo tratamiento estándar), así como a estudios donde se comparan más de dos muestras dependientes (por ejemplo, un ensayo clínico donde se asignan aleatoriamente distintos tratamientos a cada paciente que conforma un grupo de emparejamiento). Estas generalizaciones siguen argumentos similares a los descritos en este apartado y pueden consultarse en los libros de análisis de datos categóricos referenciados en este tema.

7.8 APÉNDICE: CORRECCIÓN POR CONTINUIDAD

En este apéndice se derivan las versiones con corrección por continuidad del intervalo de confianza y del test de hipótesis para una proporción poblacional π. Si k es el número observado de eventos en una muestra aleatoria de tamaño n, el intervalo de confianza al $100(1-\alpha)\%$ para π vendrá determinado por aquellos valores (π_{inf}, π_{sup}) que verifiquen

$$P(X \geq k \mid \pi = \pi_{inf}) = \alpha/2,$$
$$P(X \leq k \mid \pi = \pi_{sup}) = \alpha/2,$$

donde X es una distribución binomial de parámetros n y π. Como se discutió en el Apartado 3.3.2, si $n\pi(1-\pi) \geq 5$, estas probabilidades binomiales pueden aproximarse mediante la distribución normal estandarizada Z como

$$P(X \geq k \mid \pi = \pi_{inf}) \approx P\left(Z > \frac{k - 1/2 - n\pi_{inf}}{\sqrt{n\pi_{inf}(1-\pi_{inf})}}\right) = \alpha/2,$$

$$P(X \leq k \mid \pi = \pi_{sup}) \approx P\left(Z < \frac{k + 1/2 - n\pi_{sup}}{\sqrt{n\pi_{sup}(1-\pi_{sup})}}\right) = \alpha/2.$$

Notar que el término 1/2 de la corrección por continuidad se añade a ambas expresiones con objeto de incluir la probabilidad de observar exactamente k eventos. Para simplificar los cálculos, las desviaciones típicas de estas distribuciones normales se sustituyen por la estimación $\sqrt{np(1-p)}$, de lo cual se deduce que

$$\frac{k - 1/2 - n\pi_{inf}}{\sqrt{np(1-p)}} = z_{1-\alpha/2},$$

$$\frac{k + 1/2 - n\pi_{sup}}{\sqrt{np(1-p)}} = -z_{1-\alpha/2}.$$

Finalmente, despejando π_{inf} y π_{sup}, se obtiene el intervalo de confianza al $100(1-\alpha)\%$ para π

$$p \pm \left(z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} + \frac{1}{2n}\right).$$

Este intervalo de confianza difiere de la versión sin corrección presentada en el Apartado 7.2 en que ambos límites del intervalo se amplían en una cantidad $1/(2n)$ inversamente proporcional al tamaño muestral. La utilización de esta corrección se fundamenta en el hecho de aproximar una distribución binomial discreta mediante una distribución normal continua. Cuanto menor sea el tamaño muestral, más imprecisa será la aproximación normal y, en consecuencia, la corrección por

continuidad $1/(2n)$ ha de ser mayor. Por el contrario, si el tamaño muestral es grande, la distribución binomial estará muy próxima a la normal, por lo que la corrección $1/(2n)$ será insignificante.

El valor P para el contraste bilateral de la hipótesis nula H_0: $\pi = \pi_0$ puede obtenerse a partir de la aproximación normal a la distribución binomial como

$$P = 2P(X \geq k \mid H_0) \approx 2P\left(Z > \frac{k - n\pi_0 - 1/2}{\sqrt{n\pi_0(1-\pi_0)}} \right),$$

si la proporción observada $p > \pi_0$, o alternativamente como

$$P = 2P(X \leq k \mid H_0) \approx 2P\left(Z < \frac{k - n\pi_0 + 1/2}{\sqrt{n\pi_0(1-\pi_0)}} \right)$$

$$= 2P\left(Z > \frac{n\pi_0 - k - 1/2}{\sqrt{n\pi_0(1-\pi_0)}} \right),$$

si $p \leq \pi_0$. Combinando ambos resultados, se tiene que el valor P corresponde al doble de la probabilidad normal estandarizada a la derecha del test estadístico

$$z = \frac{|k - n\pi_0| - 1/2}{\sqrt{n\pi_0(1-\pi_0)}} = \frac{|p - \pi_0| - \dfrac{1}{2n}}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}}.$$

El test con corrección por continuidad incorpora el término $-1/(2n)$ en el numerador del estadístico, de tal forma que el valor P será ligeramente mayor que el obtenido en el correspondiente contraste sin corrección por continuidad (Apartado 7.2). Esta corrección será tanto mayor cuanto más reducido sea el tamaño muestral.

Ejemplo 7.20 En el Ejemplo 7.1 se utilizaron los controles del estudio EURAMIC para realizar inferencias sobre la prevalencia poblacional π de hombres adultos con niveles bajos de colesterol HDL ($\leq 0{,}90$ mmol/l). A continuación se calculan los correspondientes intervalos de confianza y test de hipótesis utilizando la corrección por continuidad. El IC al 95% para π vendría dado por

$$0{,}293 \pm \left(z_{0{,}975} \sqrt{\frac{0{,}293(1-0{,}293)}{539}} + \frac{1}{2 \cdot 539} \right)$$

$$= 0{,}293 \pm (1{,}96 \cdot 0{,}020 + 0{,}001) = (0{,}254;\ 0{,}333),$$

y el estadístico corregido para el contraste bilateral de la hipótesis nula H_0: $\pi = 0{,}30$ sería

$$z = \frac{|p - \pi_0| - \dfrac{1}{2n}}{\sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}} = \frac{|0{,}293 - 0{,}30| - \dfrac{1}{2 \cdot 539}}{\sqrt{\dfrac{0{,}30(1-0{,}30)}{539}}} = 0{,}30,$$

con un valor P asociado en las tablas de la distribución normal estandarizada $P = 2P(Z \geq 0{,}30) = 2\{1 - \Phi(0{,}30)\} = 0{,}764$. Como cabría esperar, el intervalo de confianza corregido

Tabla 7.8 Intervalos de confianza (IC) y tests de hipótesis con corrección por continuidad.

	IC al 100(1 – α)%	Test estadístico
Una muestra	$p \pm \left(z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} + \frac{1}{2n} \right)$	$z = \frac{\lvert p - \pi_0 \rvert - \frac{1}{2n}}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$
Dos muestras independientes	$p_1 - p_2 \pm \left\{ z_{1-\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} + \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right\}$	$z = \frac{\lvert p_1 - p_2 \rvert - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2} \right)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$
Test χ^2 de Pearson*	—	$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(\lvert O_{ij} - E_{ij} \rvert - 1/2)^2}{E_{ij}}$
Test de McNemar	—	$\chi^2 = \frac{(\lvert b - c \rvert - 1)^2}{b + c}$

* La corrección por continuidad no se aplica al test χ^2 de Pearson en tablas de contingencia mayores de 2×2.

por continuidad (IC al 95% 25,4-33,3%) es ligeramente más amplio que su correspondiente intervalo sin corrección (25,5-33,2%, Ejemplo 7.1), y el valor *P* aumenta al aplicar dicha corrección (*P* = 0,764 versus 0,726, Ejemplo 7.1). No obstante, los resultados con y sin corrección son muy similares dado que el tamaño muestral utilizado en este ejemplo es moderadamente grande.

La corrección por continuidad también se aplica a la comparación de proporciones en muestras independientes o dependientes y al test chi-cuadrado de asociación en una tabla 2×2, ya que estos métodos de inferencia utilizan una distribución continua (normal o chi-cuadrado) para representar una distribución de frecuencias discreta. Las versiones corregidas de estos procedimientos, cuya derivación es similar al caso de una proporción, se presentan en la Tabla 7.8. En general, la utilización de la corrección por continuidad da lugar a resultados más conservadores; esto es, intervalos de confianza más amplios y mayores valores *P* de los contrastes. El principal objetivo de esta corrección es aumentar la cobertura de los intervalos de confianza y reducir la probabilidad de un error de tipo I en los contrastes, especialmente cuando el tamaño muestral es reducido.

7.9 REFERENCIAS

1. Agresti A. *Categorical Data Analysis, Second Edition.* New York: John Wiley & Sons, 2002.
2. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research, Fourth Edition.* Oxford: Blackwell Science, 2001.
3. Breslow NE, Day NE. *Statistical Methods in Cancer Research, Volume 1, The Analysis of Case-Control Studies.* Lyon: International Agency for Research on Cancer, 1980.
4. Breslow NE, Day NE. *Statistical Methods in Cancer Research, Volume 2, The Design and Analysis of Cohort Studies.* Lyon: International Agency for Research on Cancer, 1987.
5. Collett D. *Modelling Binary Data, Second Edition.* London: Chapman & Hall, 2002.
6. Colton T. *Estadística en Medicina.* Barcelona: Salvat, 1979.
7. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions, Third Edition.* New York: John Wiley & Sons, 2003.
8. Hennekens CH, Buring JE. *Epidemiology in Medicine.* Boston: Little, Brown and Company, 1987.
9. Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research: Principles and Quantitative Methods.* New York: John Wiley & Sons, 1982.
10. Rosner B. *Fundamentals of Biostatistics, Fifth Edition*. Belmont, CA: Duxbury Press, 1999.
11. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology, Third Edition*. Philadelphia: Lippincott Williams & Wilkins, 2008.

TEMA 8

MÉTODOS NO PARAMÉTRICOS

8.1 INTRODUCCIÓN

En los temas anteriores se han presentado distintos métodos de inferencia para datos de carácter continuo (Tema 6) y categórico (Tema 7). Estos procedimientos se conocen como **métodos paramétricos** y asumen que los datos proceden de una población cuya distribución de probabilidad es conocida (normal o binomial), o que al menos la distribución de los estadísticos empleados puede aproximarse mediante el teorema central del límite. Así, las inferencias se fundamentaban en la aproximación normal a la distribución de las medias y proporciones muestrales. Aunque en la mayoría de las ocasiones estas asunciones son razonables, pudiera ocurrir que no se cumplan las condiciones necesarias para la realización de análisis paramétricos, especialmente cuando los tamaños muestrales son muy reducidos. En tales circunstancias, es posible utilizar métodos alternativos que realizan asunciones mínimas acerca de la distribución de la variable a estudio, y que reciben colectivamente el nombre de **métodos no paramétricos** o de distribución libre.

Antes de proceder a la descripción de los métodos no paramétricos más utilizados, conviene apuntar sus principales ventajas e inconvenientes. Entre las ventajas fundamentales cabe destacar que:

- Los métodos no paramétricos son muy robustos y, en consecuencia, pueden aplicarse a situaciones donde la utilización de pruebas paramétricas es cuestionable. Así, por ejemplo, la comparación de medias en dos muestras independientes requiere de tamaños muestrales suficientemente grandes para aplicar el teorema central del límite y de una varianza homogénea en ambas poblaciones, mientras que su equivalente no paramétrico permite contrastar globalmente la igualdad de distribuciones bajo la única asunción de que ambas distribuciones sean continuas.
- Como se verá más adelante, la propia naturaleza de las pruebas no paramétricas las hace particularmente útiles para comparar variables cualitativas ordinales, cuyo tratamiento mediante métodos paramétricos clásicos entraña problemas conceptuales ya que estas variables carecen de interpretación numérica (ver definición de tipos de variables en el Tema 1).

Sin embargo, los métodos no paramétricos presentan una serie de limitaciones que impiden su uso generalizado:

- Los métodos no paramétricos se emplean casi exclusivamente para determinar la significación estadística de la comparación entre grupos. Aunque existen procedimientos no paramétricos para obtener estimadores de efecto e intervalos de confianza, éstos requieren de asunciones adicionales y su aplicación es más compleja.
- Si se cumplen las condiciones de aplicación de las pruebas paramétricas, el uso de métodos no paramétricos es un tanto ineficiente, lo que conlleva una leve pérdida de potencia en el análisis. Estudios de simulación bajo la asunción de normalidad han mostrado una perdida de potencia aproximada del 5% de las pruebas no paramétricas respecto a sus equivalentes paramétricos.
- Los métodos paramétricos pueden extenderse fácilmente al análisis multivariante de situaciones más complejas. Aunque en la actualidad los métodos no paramétricos han experimentado un fuerte desarrollo, su utilización es aún limitada por la mayor complejidad y menor disponibilidad en los programas de análisis estadístico de uso rutinario.

En general, los métodos no paramétricos se emplean como complemento o alternativa a las pruebas paramétricas cuando no se cumplen las condiciones mínimas para la aplicación de estas últimas. En este tema se revisan los métodos no paramétricos de uso más frecuente, tales como el test de la suma de rangos de Wilcoxon, el test de los rangos con signo de Wilcoxon y el test exacto de Fisher.

8.2 TEST DE LA SUMA DE RANGOS DE WILCOXON

En el Apartado 6.3 se trató el problema de la comparación de variables continuas en dos muestras independientes. Si ambos tamaños muestrales n_1 y n_2 son suficientemente grandes para aplicar el teorema central del límite, el test de la t de Student permite realizar inferencias acerca de la diferencia de medias entre ambas poblaciones. Sin embargo, si la distribución subyacente dista mucho de ser normal y las muestras son muy pequeñas, las medias muestrales no se distribuirán de forma normal y la anterior prueba paramétrica no será aplicable. Bajo estas circunstancias, ha de utilizarse el equivalente no paramétrico al test de la t de Student para muestras independientes, que se conoce como el test de la suma de rangos de Wilcoxon. Este procedimiento permite contrastar globalmente la igualdad de distribuciones bajo la única asunción de que la variable a estudio tenga una distribución subyacente continua.

Si no se asume nada sobre la forma de la distribución, parece razonable basar el contraste en el orden de las observaciones de ambas muestras y no en sus verdaderos valores. Para ello, se combinan las dos muestras ordenando los valores de menor a mayor. A continuación, se asigna el **rango** r_i o posición que ocupa cada observación dentro de la muestra combinada. Si existen varias observaciones con el mismo valor de la variable (empates), se asigna a cada una de ellas la media de los rangos correspondientes. Finalmente, se suman los rangos de una cualquiera de las dos muestras, seleccionemos por ejemplo la primera muestra,

$$U = \sum_{i=1}^{n_1} r_i.$$

El estadístico del test de Wilcoxon se basa en esta suma de rangos.

Ejemplo 8.1 Supongamos que la muestra consiste en $n_1 = 10$ casos de infarto de miocardio y $n_2 = 10$ controles seleccionados aleatoriamente del estudio EURAMIC. La Tabla 8.1 muestra los niveles de β-caroteno en tejido adiposo para estos 20 sujetos. Al menor valor de ambas muestras 0,04 µg/g se le asigna el rango 1, al siguiente valor 0,05 µg/g se le otorga el rango 2 y así sucesivamente hasta asignar el rango 20 al mayor valor 0,57 µg/g. A los dos sujetos con idéntico nivel 0,13 µg/g de β-caroteno les corresponden las posiciones 7 y 8 y, en consecuencia, se asigna el rango medio (7 + 8)/2 = 7,5 a ambas observaciones. Así, la suma de rangos en los casos de infarto es

$$\sum_{i=1}^{10} r_i = 1 + 9 + \ldots + 19 = 96{,}5$$

y en los controles

$$\sum_{j=1}^{10} r_j = 13 + 2 + \ldots + 6 = 113{,}5.$$

Notar que la elección entre una u otra suma de rangos es arbitraria. La suma total de rangos en ambas muestras es $(n_1 + n_2)(n_1 + n_2 + 1)/2 = 20 \cdot 21/2 = 210$, de tal forma que una vez calculada la suma de rangos 96,5 en la primera muestra, la otra queda determinada por 210 – 96,5 = 113,5.

Tabla 8.1 **β-caroteno en tejido adiposo en 10 casos de infarto de miocardio y 10 controles seleccionados aleatoriamente del estudio EURAMIC.**

Caso		Control	
β-caroteno (µg/g)	Rango (r_i)	β-caroteno (µg/g)	Rango (r_j)
0,04	1	0,25	13
0,14	9	0,05	2
0,20	11	0,36	17
0,08	3	0,09	4
0,21	12	0,33	16
0,10	5	0,37	18
0,28	14	0,13	7,5
0,29	15	0,17	10
0,13	7,5	0,57	20
0,48	19	0,12	6
	$\sum_{i=1}^{10} r_i = 96{,}5$		$\sum_{j=1}^{10} r_j = 113{,}5$

El objetivo es contrastar si las distribuciones F_1 y F_2 en ambas poblaciones son iguales H_0: $F_1 = F_2$ frente a la hipótesis alternativa bilateral H_1: $F_1 \neq F_2$. Bajo esta hipótesis nula, la suma de rangos esperada en la primera muestra sería igual a la suma total de rangos por la proporción de sujetos en dicha muestra,

$$E(U) = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} \frac{n_1}{n_1 + n_2} = \frac{n_1(n_1 + n_2 + 1)}{2}.$$

Por tanto, si u denota la suma de rangos observada en la primera muestra, el valor exacto de P vendría determinado por la probabilidad bajo H_0 de obtener una suma de rangos tanto o más distante de $E(U)$ que el valor observado u; es decir,

$$P = 2P(U \geq u \mid H_0),$$

si $u > E(U)$, o alternativamente

$$P = 2P(U \leq u \mid H_0),$$

si $u \leq E(U)$. Esta probabilidad puede calcularse teniendo en cuenta que bajo la hipótesis nula de igualdad de distribuciones, cualquier combinación de rangos en la primera muestra es igualmente probable. Así, como el número de combinaciones de los $n_1 + n_2$ posibles rangos tomados de n_1 en n_1 es $\binom{n_1 + n_2}{n_1}$, la probabilidad bajo H_0 para cualquier combinación $r_1, ..., r_{n_1}$ viene dada por

$$\frac{1}{\binom{n_1 + n_2}{n_1}}.$$

El cálculo del valor exacto de P se ilustra en el siguiente ejemplo.

Ejemplo 8.2 Si la distribución del β-caroteno fuera igual en los casos de infarto y en los controles libres de enfermedad, la suma de rangos esperada en los 10 casos de infarto del ejemplo anterior sería igual a

$$E(U) = \frac{10(10+10+1)}{2} = 105.$$

Como el valor observado de esta suma de rangos $u = 96{,}5$ es inferior al esperado, el valor P se obtiene mediante

$$P = 2P(U \leq 96{,}5 \mid H_0) = 2\sum_{k=55}^{96} P(U = k \mid H_0).$$

Notar que la suma arranca en el valor mínimo posible 1 + 2 + ... + 10 = 55 y sólo toma valores enteros (se excluyen posibles empates para facilitar los cálculos). La probabilidad bajo H_0 para cualquier combinación de rangos en la primera muestra es

$$\frac{1}{\binom{20}{10}} = \frac{10!(20-10)!}{20!} = \frac{1}{184.756},$$

de lo cual se sigue que

$$\begin{aligned}
P(U = 55 \mid H_0) &= P(1, 2, 3, 4, 5, 6, 7, 8, 9, 10 \mid H_0) = 1/184.756,\\
P(U = 56 \mid H_0) &= P(1, 2, 3, 4, 5, 6, 7, 8, 9, 11 \mid H_0) = 1/184.756,\\
P(U = 57 \mid H_0) &= P(1, 2, 3, 4, 5, 6, 7, 8, 9, 12 \mid H_0)\\
&\quad + P(1, 2, 3, 4, 5, 6, 7, 8, 10, 11 \mid H_0) = 2/184.756
\end{aligned}$$

y así sucesivamente. Como puede intuirse, el procedimiento resulta muy laborioso incluso para estas pequeñas muestras de tamaño 10, ya que requiere determinar el número de combinaciones con igual suma de rangos. Después de múltiples cálculos, se tiene que

$$\begin{aligned}
P &= 2\sum_{k=55}^{96} P(U = k \mid H_0) = 2(1 + 1 + 2 + ... + 4.397)/184.756\\
&= 97.708/184.756 = 0{,}529.
\end{aligned}$$

Aunque los casos de infarto muestran niveles inferiores de β-caroteno que los controles (la suma de rangos observada en los casos es menor que la esperada), no se alcanzan diferencias estadísticamente significativas. No obstante, dado el reducido tamaño muestral, cabe esperar que la potencia de este contraste sea muy pequeña para detectar cualquier posible diferencia en los niveles subyacentes de β-caroteno entre los casos de infarto y los sujetos libres de la enfermedad.

Para simplificar los cálculos de este test, la Tabla 8 del Apéndice facilita los percentiles de la distribución de la suma de rangos de Wilcoxon bajo la hipótesis nula de igualdad de distribuciones, cuando la menor de las dos muestras es de tamaño inferior o igual a 8. Para un nivel de significación α bilateral, la hipótesis nula se rechazará si la suma de rangos en la muestra de menor tamaño es inferior al percentil $\alpha/2$ o superior al percentil $1 - \alpha/2$ de dicha tabla.

Ejemplo 8.3 En un estudio hipotético a partir de dos muestras independientes de tamaños $n_1 = 5$ y $n_2 = 10$, la suma de rangos en la muestra más pequeña es 23. Como la distribución bajo H_0 de la suma de rangos es simétrica alrededor de $E(U) = n_1(n_1 + n_2 + 1)/2 = 5(5 + 10 + 1)/2 = 40$, se tiene que

$$P = 2P(U \leq 23 \mid H_0) = 2P(U \geq 57 \mid H_0).$$

Utilizando la Tabla 8 del Apéndice con $n_1 = 5$ y $n_2 = 10$, puede comprobarse que el valor $u = 57$ está comprendido entre los percentiles $u_{0,975} = 56$ y $u_{0,99} = 58$, de lo cual se deduce la desigualdad $0{,}01 < P(U \geq 57 \mid H_0) < 0{,}025$, que corresponde a $0{,}02 < P < 0{,}05$.

En el caso de que ambos tamaños muestrales sean superiores a 8, puede emplearse el siguiente método aproximado. Como el contraste para la igualdad de distribuciones se basa en el rango o posición de las observaciones, resulta lícito sustituir los valores observados x_i por sus correspondientes rangos r_i en el estadístico de la t de Student para muestras independientes con igual varianza (Apartado 6.3.1), obteniéndose

$$z = \frac{\bar{r}_1 - \bar{r}_2}{s_r \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}},$$

donde la diferencia de rangos medios es

$$\begin{aligned}
\bar{r}_1 - \bar{r}_2 &= \frac{1}{n_1}\sum_{i=1}^{n_1} r_i - \frac{1}{n_2}\sum_{j=1}^{n_2} r_j \\
&= \frac{1}{n_1}\sum_{i=1}^{n_1} r_i - \frac{1}{n_2}\left(\frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} - \sum_{i=1}^{n_1} r_i\right) \\
&= \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left(\sum_{i=1}^{n_1} r_i - \frac{n_1(n_1 + n_2 + 1)}{2}\right)
\end{aligned}$$

y, si no hay empates, la varianza de los rangos en la muestra combinada es

$$\begin{aligned}
s_r^2 &= \frac{1}{n_1 + n_2 - 1}\sum_{i=1}^{n_1+n_2} (r_i - \bar{r})^2 \\
&= \frac{1}{n_1 + n_2 - 1}\sum_{i=1}^{n_1+n_2} \left(i - \frac{n_1 + n_2 + 1}{2}\right)^2 \\
&= \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{12}.
\end{aligned}$$

Sustituyendo en la expresión anterior, se tiene

$$z = \frac{\displaystyle\sum_{i=1}^{n_1} r_i - \frac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\dfrac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = \frac{U - E(U)}{SE(U)},$$

que corresponde simplemente a la suma de rangos estandarizada; es decir, la diferencia entre la suma de rangos observada y esperada en la primera muestra dividida por su error estándar bajo la hipótesis nula de igualdad de distribuciones. Bajo H_0, este estadístico seguirá aproximadamente una distribución normal estandarizada si n_1, $n_2 > 8$. Notar que, en general, este tamaño muestral es muy inferior al que se requeriría para aplicar la prueba paramétrica de la *t* de Student en dos muestras independientes.

Ejemplo 8.4 A partir del estudio EURAMIC, se seleccionan 1000 muestras aleatorias simples de $n_1 = 10$ casos de infarto de miocardio y $n_2 = 10$ controles. En cada una de estas muestras, se calcula la diferencia de niveles medios de β-caroteno entre casos y controles, así como la suma de rangos para los casos de infarto. Las Figuras 8.1(*a*) y (*b*) presentan las distribuciones muestrales de la diferencia de medias $\bar{x}_1 - \bar{x}_2$ y de la suma de rangos *U*, respectivamente. Como la distribución poblacional del β-caroteno es marcadamente asimétrica (ver Figura 4.3) y las muestras son muy pequeñas, la diferencia de medias muestrales se distribuye de forma asimétrica alrededor de la diferencia subyacente $\mu_1 - \mu_2 = -0{,}09$ µg/g, de tal forma que no se cumple la condición de normalidad necesaria para aplicar el test de la *t* de Student. Por el contrario, la suma de rangos sí se distribuye de forma aproximadamente normal en torno a su valor esperado en esta población $E(U) = 96{,}9$. Así, aun cuando se disponga de muestras tan reducidas, se podría aplicar la aproximación normal al test de la suma de rangos de Wilcoxon.

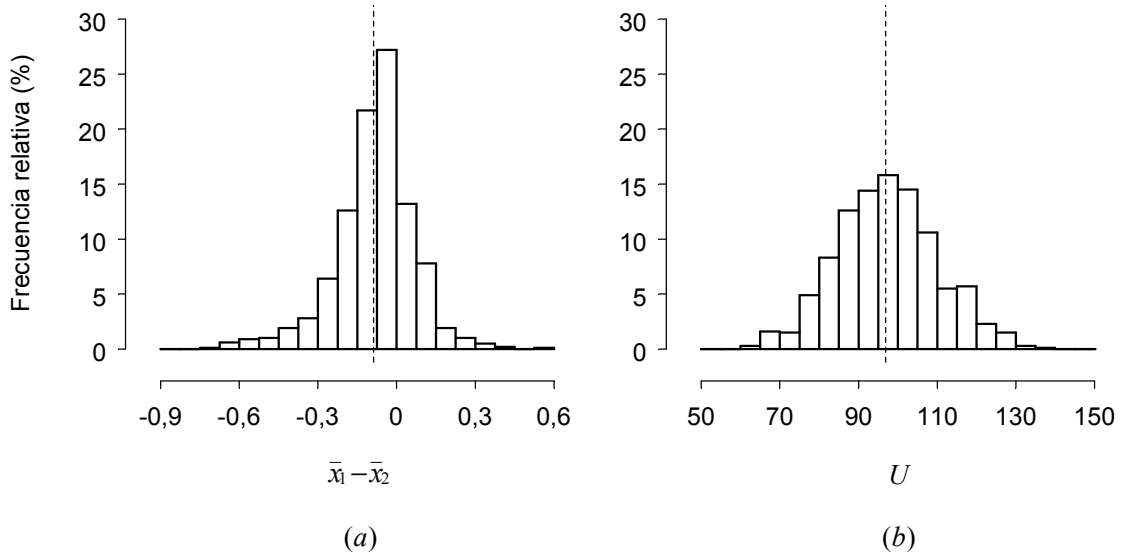


Figura 8.1 **Distribución muestral de la diferencia de niveles medios de β-caroteno $\bar{x}_1 - \bar{x}_2$ entre casos y controles (*a*) y de la suma de rangos *U* en los casos de infarto (*b*) en 1000 muestras aleatorias simples de $n_1 = 10$ casos de infarto de miocardio y $n_2 = 10$ controles obtenidos a partir del estudio EURAMIC. Las líneas verticales en trazo discontinuo corresponden a los parámetros subyacentes $\mu_1 - \mu_2 = -0{,}09$ µg/g y $E(U) = 96{,}9$.**

Si se producen empates en la asignación de rangos en la muestra combinada, la varianza de la suma de rangos es menor que la obtenida en ausencia de empates y el estadístico del test de la suma de los rangos de Wilcoxon resulta

$$z = \frac{\sum_{i=1}^{n_1} r_i - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)(1-f)}{12}}},$$

donde

$$f = \frac{\sum_{i=1}^{T} t_i (t_i+1)(t_i-1)}{(n_1+n_2)(n_1+n_2+1)(n_1+n_2-1)},$$

con t_i el número de empates para el valor i-ésimo de la variable. Notar que, si no hay empates, $f = 0$ y este estadístico se reduce al citado anteriormente. Finalmente, como la suma de rangos es una variable discreta que se aproxima mediante una distribución normal continua, es frecuente aplicar la corrección por continuidad a estos estadísticos. La versión con corrección por continuidad del test de la suma de rangos de Wilcoxon (con o sin empates) se presenta en la Tabla 8.2.

Ejemplo 8.5 Como la muestra de casos y controles de la Tabla 8.1 es $n_1 = n_2 = 10 > 8$, puede aplicarse la aproximación normal a la suma de rangos $U = 96{,}5$ en los casos de infarto. Bajo la hipótesis nula de una misma distribución del β-caroteno en casos y controles, el valor esperado de esta suma de rangos sería

$$E(U) = \frac{10(10+10+1)}{2} = 105$$

y su varianza

$$\text{var}(U) = \frac{10 \cdot 10(10+10+1)(1-0{,}00075)}{12} = 174{,}87,$$

donde

$$f = \frac{2(2+1)(2-1)}{(10+10)(10+10+1)(10+10-1)} = 0{,}00075$$

es el factor de corrección de la varianza debido a la presencia de $t_1 = 2$ observaciones empatadas para el valor 0,13 μg/g. Por tanto, el estadístico de la suma de rangos de Wilcoxon con corrección por continuidad es

$$z = \frac{|96{,}5 - 105| - 1/2}{\sqrt{174{,}87}} = 0{,}60,$$

que corresponde a un valor $P = 2P(Z \geq 0{,}60) = 2\{1 - \Phi(0{,}60)\} = 0{,}549$ a partir de la distribución normal estandarizada de la Tabla 3 del Apéndice. Este valor aproximado de P es muy similar al valor exacto calculado en el Ejemplo 8.2, no habiendo así suficiente evidencia para rechazar la hipótesis de igualdad de distribuciones del nivel de β-caroteno en los casos de infarto de miocardio y los sujetos libres de la enfermedad.

Tabla 8.2 **Estadísticos para el test de la suma de rangos y de los rangos con signo de Wilcoxon con corrección por continuidad.**

	Sin empates	Con empates
Test de la suma de rangos	$z = \dfrac{\left\lvert \sum_{i=1}^{n_1} r_i - \dfrac{n_1(n_1+n_2+1)}{2} \right\rvert - \dfrac{1}{2}}{\sqrt{\dfrac{n_1 n_2 (n_1+n_2+1)}{12}}}$	$z = \dfrac{\left\lvert \sum_{i=1}^{n_1} r_i - \dfrac{n_1(n_1+n_2+1)}{2} \right\rvert - \dfrac{1}{2}}{\sqrt{\dfrac{n_1 n_2 (n_1+n_2+1)(1-f)}{12}}}$ con $f = \dfrac{\sum_{i=1}^{T} t_i (t_i+1)(t_i-1)}{(n_1+n_2)(n_1+n_2+1)(n_1+n_2-1)}$
Test de los rangos con signo	$z = \dfrac{\left\lvert \sum_{i=1}^{m} r_i - \dfrac{n(n+1)}{4} \right\rvert - \dfrac{1}{2}}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}}$	$z = \dfrac{\left\lvert \sum_{i=1}^{m} r_i - \dfrac{n(n+1)}{4} \right\rvert - \dfrac{1}{2}}{\sqrt{\dfrac{n(n+1)(2n+1) - f}{24}}}$ con $f = \dfrac{\sum_{i=1}^{T} t_i (t_i+1)(t_i-1)}{2}$

El test de la suma de rangos de Wilcoxon es también conocido como el **test de Mann-Whitney**. Aunque este último se deriva siguiendo un procedimiento distinto, ambas pruebas de hipótesis son completamente equivalentes, obteniéndose el mismo valor P con cualquiera de ellas. La comparación no paramétrica de distribuciones continuas en más de dos muestras independientes se conoce como el **test de Kruskal-Wallis**. Este procedimiento es una generalización del test de la suma de rangos de Wilcoxon y puede consultarse en los textos sobre métodos no paramétricos referenciados en este tema.

8.3 TEST DE LOS RANGOS CON SIGNO DE WILCOXON

En este apartado se describe el procedimiento de contraste no paramétrico equivalente al test de la t de Student para muestras dependientes. Como se discutió en el Apartado 6.4, la prueba t para datos emparejados permite comparar dos medias poblacionales a partir de las diferencias observadas en cada pareja de datos dependientes. Esta prueba paramétrica requiere que el número de parejas sea suficientemente grande para asegurar que la media de las diferencias se distribuya de forma normal. En aquellas circunstancias donde se produzcan violaciones claras de este supuesto de normalidad (particularmente cuando el número de parejas sea muy reducido), resulta más apropiado utilizar el test no paramétrico de los rangos con signo de Wilcoxon. Bajo la asunción de que la variable a estudio sea continua, este procedimiento permite contrastar si las diferencias se distribuyen simétricamente alrededor de 0. La hipótesis nula establece, por tanto, que las diferencias de cualquier magnitud a favor de los sujetos de una población son igualmente probables que a favor de los sujetos de la otra población.

Con objeto de preservar el emparejamiento, se calculan las diferencias d_i en cada pareja de datos dependientes. La asignación de rangos a estas diferencias se realiza mediante el siguiente procedimiento. En primer lugar, se excluyen las parejas donde $d_i = 0$ y se asignan rangos r_i a las restantes n diferencias no nulas, comenzando en 1 para la diferencia con menor valor absoluto hasta n para aquella con mayor valor absoluto. Si existen diferencias con el mismo valor absoluto (empates), se asigna a cada una de ellas la media de los rangos correspondientes. Finalmente, a cada rango se le otorga el signo correspondiente a su diferencia. Estos **rangos con signo** constituyen así una representación estandarizada de las diferencias, que preserva tanto el orden de magnitud como el signo de las mismas. El test de los rangos con signo de Wilcoxon se basa en la suma de los rangos positivos (o, equivalentemente, de los rangos negativos)

$$W = \sum_{i=1}^{m} r_i,$$

donde m denota el número de rangos positivos.

Ejemplo 8.6 A partir del estudio EURAMIC, se seleccionan aleatoriamente 20 casos de infarto de miocardio y 20 controles emparejados por grupos quinquenales de edad. Los niveles de β-caroteno para estas 20 parejas de casos y controles se presentan en la Tabla 8.3. Una vez excluida la pareja con $d_i = 0$, el número efectivo de parejas es $n = 19$. A partir de estas parejas con diferencias no nulas, se asignan rangos del 1 al 19 comenzando en la menor diferencia absoluta 0,01 μg/g hasta la mayor diferencia absoluta 1,00 μg/g. A las dos parejas con diferencia absoluta 0,27 μg/g se les otorga el rango medio (9 + 10)/2 = 9,5, y a otras dos parejas con diferencia absoluta 0,38 μg/g se les asigna su rango medio (12 + 13)/2 = 12,5. Finalmente, se otorga un signo positivo a los rangos correspondientes

a diferencias positivas y un signo negativo a los rangos correspondientes a diferencias negativas. La suma de rangos positivos resulta

$$\sum_{i=1}^{9} r_i = 17 + 12{,}5 + \ldots + 3 = 91$$

y la suma de rangos negativos

$$\sum_{j=1}^{10} r_j = (-4) + (-14) + \ldots + (-9{,}5) = -99.$$

En este ejemplo la suma total de los rangos absolutos es $n(n + 1)/2 = 19 \cdot 20/2 = 190$. Así, una vez determinada la suma de rangos positivos 91, la suma de rangos negativos viene dada por 91 – 190 = –99.

Tabla 8.3 β-caroteno en tejido adiposo en 20 casos y controles del estudio EURAMIC emparejados según grupos quinquenales de edad.

	β-caroteno (µg/g)					
Pareja	Caso	Control	Diferencia (d_i)	Diferencia absoluta	Rango absoluto	Rango con signo (r_i)
1	0,47	0,55	–0,08	0,08	4	–4
2	0,75	0,09	0,66	0,66	17	17
3	0,78	0,40	0,38	0,38	12,5	12,5
4	0,66	0,13	0,53	0,53	15	15
5	0,09	0,49	–0,40	0,40	14	–14
6	0,20	0,31	–0,11	0,11	5	–5
7	0,08	0,28	–0,20	0,20	7	–7
8	0,08	0,46	–0,38	0,38	12,5	–12,5
9	0,31	0,16	0,15	0,15	6	6
10	0,30	0,87	–0,57	0,57	16	–16
11	0,16	1,16	–1,00	1,00	19	–19
12	0,13	0,13	0	0	—	—
13	0,06	0,37	–0,31	0,31	11	–11
14	0,25	0,04	0,21	0,21	8	8
15	0,39	0,37	0,02	0,02	2	2
16	0,95	0,14	0,81	0,81	18	18
17	0,33	0,06	0,27	0,27	9,5	9,5
18	0,53	0,50	0,03	0,03	3	3
19	0,16	0,17	–0,01	0,01	1	–1
20	0,23	0,50	–0,27	0,27	9,5	–9,5
				Suma de rangos positivos	$\sum_{i=1}^{9} r_i = 91$	
				Suma de rangos negativos	$\sum_{j=1}^{10} r_j = -99$	

Bajo la hipótesis nula de que las diferencias se distribuyen simétricamente alrededor de 0, se esperaría la misma suma de rangos positivos que negativos y, por consiguiente, la suma esperada de rangos positivos sería la mitad de la suma total de rangos absolutos

$$E(W) = \frac{1}{2}\frac{n(n+1)}{2} = \frac{n(n+1)}{4},$$

donde n indica el número de diferencias no nulas. Al igual que en el apartado anterior, el valor exacto de P para el contraste bilateral vendrá dado por la probabilidad bajo H_0 de obtener una suma de rangos positivos tanto o más distante de $E(W)$ que su valor observado w; esto es, si $w > E(W)$,

$$P = 2P(W \geq w \mid H_0)$$

y, si $w \leq E(W)$,

$$P = 2P(W \leq w \mid H_0).$$

Bajo dicha hipótesis nula, cualquier combinación de un número arbitrario de rangos positivos $r_1, ..., r_m$ es igualmente probable y su probabilidad viene determinada por

$$\frac{1}{2^n},$$

donde 2^n es el número de subconjuntos de cualquier tamaño que pueden obtenerse a partir de las n parejas con diferencias no nulas. Haciendo uso de este resultado, la Tabla 9 del Apéndice facilita los percentiles de la distribución de la suma de rangos positivos bajo la hipótesis nula de que las diferencias se distribuyen simétricamente alrededor de 0, cuando el número de diferencias no nulas es $n \leq 16$. Para un nivel de significación α preestablecido, la hipótesis nula se rechazará si la suma de rangos positivos es inferior al percentil $\alpha/2$ o superior al percentil $1 - \alpha/2$.

Ejemplo 8.7 Como ilustración, supongamos que la suma de rangos positivos es $w = 25$ a partir de $n = 12$ parejas de datos dependientes con diferencias no nulas. La distribución bajo H_0 de la suma de rangos positivos es simétrica alrededor de $E(W) = n(n + 1)/4 = 12(12 + 1)/4 = 39$, de lo cual se deduce que

$$w_{0,05} = n(n + 1)/2 - w_{0,95} = 78 - 60 = 18,$$

donde $w_{0,95} = 60$ se obtiene de la Tabla 9 del Apéndice para $n = 12$. Como la suma observada $w = 25 > w_{0,05} = 18$, se sigue que $P(W \leq 25 \mid H_0) > 0{,}05$. Así, el contraste bilateral arroja un valor $P > 0{,}10$.

En aquellas muestras donde el número de diferencias no nulas sea superior a 16, puede utilizarse la siguiente aproximación normal. Dado que los rangos con signo constituyen una representación estandarizada de las diferencias observadas en cada pareja de datos dependientes, podría construirse un estadístico sustituyendo las diferencias no nulas d_i por los rangos con signo r_i en el test de la t de Student para muestras dependientes (Apartado 6.4). Así, el estadístico resulta

$$z = \frac{\bar{r}}{s_r / \sqrt{n}},$$

donde la media de los m rangos positivos y $n - m$ rangos negativos es

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i = \frac{1}{n}\left(\sum_{i=1}^{m} r_i + \sum_{j=1}^{n-m} r_j\right)$$

$$= \frac{1}{n}\left\{\sum_{i=1}^{m} r_i + \left(\sum_{i=1}^{m} r_i - \frac{n(n+1)}{2}\right)\right\}$$

$$= \frac{2}{n}\left(\sum_{i=1}^{m} r_i - \frac{n(n+1)}{4}\right)$$

y, en el caso de que no haya empates, la varianza bajo H_0 de los rangos con signo se estima mediante

$$s_r^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \frac{1}{n}\sum_{i=1}^{n} i^2 = \frac{(n+1)(2n+1)}{6}.$$

Aplicando ambos resultados, se tiene el estadístico

$$z = \frac{\sum_{i=1}^{m} r_i - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{W - E(W)}{SE(W)},$$

que representa la diferencia entre el valor observado y esperado de la suma de rangos positivos, dividida por su error estándar bajo H_0. Si el número de parejas con diferencias no nulas es $n > 16$, este estadístico sigue aproximadamente una distribución normal estandarizada bajo la hipótesis nula de simetría de las diferencias alrededor de 0.

> ***Ejemplo 8.8*** A partir del estudio EURAMIC, se seleccionan 1000 muestras aleatorias de 20 parejas de casos y controles agrupados según quinquenios de edad. La Figura 8.2 presenta la distribución muestral de la diferencia media de β-caroteno $\bar{d}$ entre casos y controles, así como la distribución muestral de la suma de rangos positivos W (esto es, la suma de rangos en las parejas donde el caso presenta un nivel superior de β-caroteno que el control). Debido al reducido número de parejas, la media de las diferencias de β-caroteno presenta una distribución asimétrica y, en consecuencia, la utilización de la prueba de la t de Student para muestras dependientes resulta cuestionable. Sin embargo, a pesar de contar únicamente con 20 parejas, la distribución de la suma de rangos positivos presenta un aspecto mucho más normal, permitiendo así el uso de la aproximación normal al test de los rangos con signo de Wilcoxon.

En el caso de existir diferencias con el mismo valor absoluto, ha de utilizarse la siguiente versión corregida del estadístico del test de los rangos con signo

$$z = \frac{\sum_{i=1}^{m} r_i - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1) - f}{24}}},$$

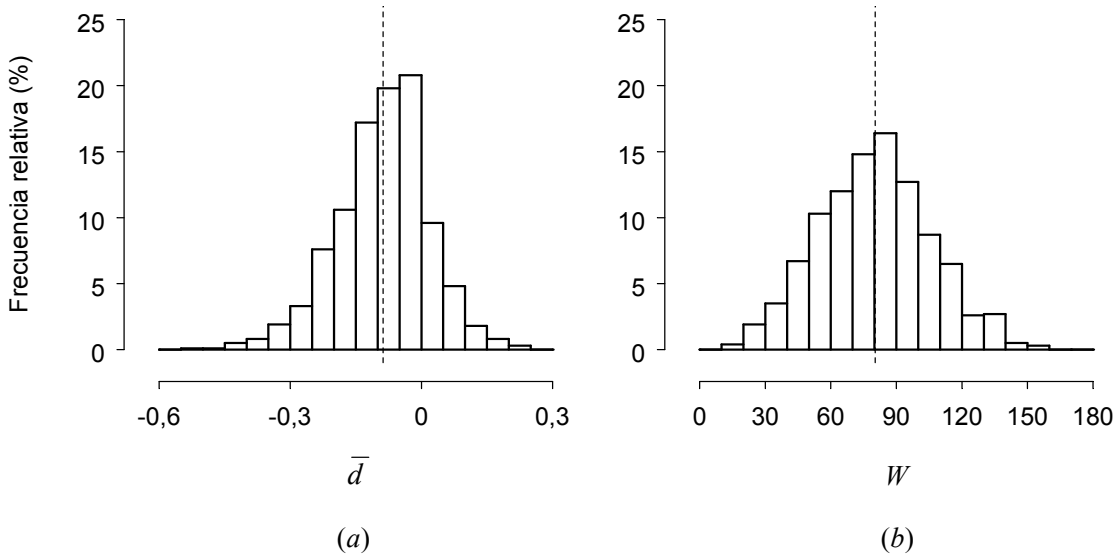


Figura 8.2 **Distribución muestral de la diferencia media de β-caroteno $\bar{d}$ entre casos y controles (*a*) y de la suma de rangos positivos W (*b*) en 1000 muestras aleatorias de $n = 20$ parejas de casos y controles agrupados según quinquenios de edad a partir del estudio EURAMIC. Las líneas verticales en trazo discontinuo corresponden a los parámetros subyacentes $\mu_1 - \mu_2 = -0{,}09$ µg/g y $E(W) = 80{,}3$.**

cuya varianza incluye el término de corrección por empates

$$f = \frac{\sum_{i=1}^{T} t_i (t_i + 1)(t_i - 1)}{2},$$

donde t_i es el número de empates para la i-ésima diferencia absoluta. Esta corrección conlleva una reducción de la varianza y su efecto sobre el estadístico será apreciable cuando el número de empates sea elevado (tal es el caso de las variables cualitativas ordinales). Dado el carácter discreto de la suma de rangos y el reducido tamaño muestral inherente a las pruebas no paramétricas, la aproximación normal a estos estadísticos suele incorporar además la corrección por continuidad de la Tabla 8.2 para reducir la probabilidad de incurrir en un error de tipo I.

Ejemplo 8.9 En la Tabla 8.3 se obtuvieron $n = 19 > 16$ parejas de casos y controles con diferencias no nulas de β-caroteno y, en consecuencia, puede utilizarse la aproximación normal a la suma de rangos positivos $W = 91$. Bajo la hipótesis nula de simetría de las diferencias alrededor de 0, el valor esperado de la suma de rangos positivos es

$$E(W) = \frac{19(19+1)}{4} = 95$$

y la varianza

$$\text{var}(W) = \frac{19(19+1)(2 \cdot 19+1) - 6}{24} = 617{,}25,$$

donde el término de corrección de la varianza por los $t_1 = 2$ empates con diferencia absoluta 0,27 µg/g y los $t_2 = 2$ empates con diferencia absoluta 0,38 µg/g es

$$f = \frac{2(2+1)(2-1) + 2(2+1)(2-1)}{2} = 6.$$

Aplicando la corrección por continuidad, el test estadístico de los rangos con signo de Wilcoxon resulta entonces

$$z = \frac{|91 - 95| - 1/2}{\sqrt{617{,}25}} = 0{,}14,$$

con un valor $P = 2P(Z \geq 0{,}14) = 2\{1 - \Phi(0{,}14)\} = 0{,}889$. Notar que el resultado del test sería idéntico de utilizar la suma de rangos negativos $W = -99$, ya que su valor esperado es $E(W) = -95$ y su varianza coincide con $\text{var}(W) = 617{,}25$. Así, una vez controladas las diferencias de edad, las diferencias de β-caroteno a favor de los casos de infarto no son significativamente distintas de las diferencias a favor de los sujetos libres de la enfermedad.

La comparación no paramétrica de una variable continua en más de dos muestras dependientes puede realizarse mediante el **test de Friedman**. Bajo la asunción de que la variable sigue la misma distribución continua excepto posibles diferencias de localización (traslaciones), esta prueba permite contrastar la hipótesis nula de una misma localización de la variable en cada una de las poblaciones. Este procedimiento también se fundamenta en la definición de rangos y puede consultarse en los libros específicos de métodos no paramétricos.

8.4 TEST EXACTO DE FISHER

En el Apartado 7.4 se presentó el test χ^2 de Pearson como un procedimiento general para evaluar la asociación estadística entre las variables de una tabla 2×2. Esta prueba se basa en la asunción de que el tamaño muestral es suficientemente grande para justificar la aproximación chi-cuadrado a la distribución nula del estadístico χ^2 de Pearson. En concreto, si los marginales de la tabla son pequeños, de tal forma que la frecuencia esperada en alguna de las celdas sea inferior a 5, esta aproximación puede resultar imprecisa. En tales circunstancias, es preferible utilizar métodos alternativos basados en la distribución exacta de las frecuencias de las celdas de una tabla 2×2. En este apartado se describe el más conocido de estos procedimientos, el test exacto de Fisher.

Ejemplo 8.10 La Tabla 8.4 presenta el número de sujetos con niveles de β-caroteno superiores e inferiores a 0,30 μg/g entre los 10 casos de infarto y los 10 controles del estudio EURAMIC seleccionados de forma independiente en el Ejemplo 8.1. Bajo la hipótesis de independencia entre el nivel de β-caroteno y el riesgo de infarto de miocardio, la frecuencia esperada en cada celda sería

$$E_{11} = E_{12} = \frac{5 \cdot 10}{20} = 2{,}5,$$

$$E_{21} = E_{22} = \frac{15 \cdot 10}{20} = 7{,}5.$$

Como los valores esperados en dos de las cuatro celdas son inferiores a 5, la prueba χ^2 de Pearson no será aplicable a esta tabla 2×2 y la asociación ha de contrastarse mediante otro procedimiento.

Tabla 8.4 **β-caroteno en tejido adiposo en 10 casos de infarto de miocardio y 10 controles seleccionados aleatoriamente del estudio EURAMIC.**

β-caroteno (µg/g)	Infarto de miocardio		
	Caso	Control	Total
> 0,30	1	4	5
≤ 0,30	9	6	15
Total	10	10	20

El test exacto de Fisher se basa en determinar la probabilidad exacta de observar una tabla cualquiera con frecuencias a, b, c y d, bajo la hipótesis nula de independencia y asumiendo que todos los marginales n_1, n_2, m_1 y m_2 son fijos (Tabla 7.1). La condición de marginales fijos se impone por conveniencia matemática, ya que los cálculos se simplifican notablemente y los marginales contienen poca información sobre la asociación a estudio. Bajo H_0, la probabilidad de enfermar π es común en los sujetos expuestos y los no expuestos. Así, el número de enfermos entre los expuestos sigue una distribución binomial de parámetros n_1 y π, mientras que entre los no expuestos sigue una distribución binomial de parámetros n_2 y π. Como las muestras de expuestos y no expuestos son independientes, la probabilidad de obtener una tabla con frecuencias a, b, c y d es el producto de las probabilidades binomiales de observar a sujetos enfermos entre los expuestos y c entre los no expuestos,

$$P(a, b, c, d \mid H_0) = \binom{n_1}{a} \pi^a (1-\pi)^{n_1-a} \binom{n_2}{c} \pi^c (1-\pi)^{n_2-c}$$

$$= \binom{n_1}{a} \binom{n_2}{m_1-a} \pi^{m_1} (1-\pi)^{m_2}.$$

Para marginales n_1, n_2, m_1 y m_2 fijos, el rango de valores posibles k para el número de casos expuestos varía entre $k_1 = \max(0, m_1 - n_2)$ y $k_2 = \min(n_1, m_1)$. Por tanto, la probabilidad de obtener una tabla con frecuencias a, b, c y d condicionada a unos marginales n_1, n_2, m_1 y m_2 fijos viene dada por

$$P(a, b, c, d \mid n_1, n_2, m_1, m_2; H_0) = \frac{\binom{n_1}{a}\binom{n_2}{m_1-a} \pi^{m_1} (1-\pi)^{m_2}}{\sum_{k=k_1}^{k_2} \binom{n_1}{k}\binom{n_2}{m_1-k} \pi^{m_1} (1-\pi)^{m_2}}$$

$$= \frac{\binom{n_1}{a}\binom{n_2}{m_1-a}}{\sum_{k=k_1}^{k_2} \binom{n_1}{k}\binom{n_2}{m_1-k}} = \frac{\binom{n_1}{a}\binom{n_2}{m_1-a}}{\binom{n_1+n_2}{m_1}},$$

donde el denominador de la última igualdad se obtiene de las propiedades de los coeficientes binomiales. Esta distribución de probabilidades entre todas las posibles tablas con los mismos marginales se conoce como **distribución hipergeométrica** y determina la distribución bajo H_0

del número de casos expuestos y no expuestos en una muestra de m_1 casos obtenidos a partir de un total de n_1 sujetos expuestos y n_2 sujetos no expuestos. Notar que esta probabilidad depende únicamente del número a de casos expuestos, dado que una vez conocido a las frecuencias de las restantes celdas quedan determinadas por los marginales de la tabla. Cabe destacar también que aunque los cálculos se han derivado de un estudio prospectivo, se obtendría el mismo resultado a partir de un estudio retrospectivo en términos del número de sujetos expuestos entre casos y controles,

$$P(a \mid n_1, n_2, m_1, m_2; H_0) = \frac{\binom{m_1}{a}\binom{m_2}{n_1 - a}}{\binom{m_1 + m_2}{n_1}} = \frac{\binom{n_1}{a}\binom{n_2}{m_1 - a}}{\binom{n_1 + n_2}{m_1}}$$

$$= \frac{n_1! n_2! m_1! m_2!}{n! a! b! c! d!},$$

lo cual confirma que la probabilidad condicional asociada a una determinada tabla no varía en función del diseño prospectivo o retrospectivo del estudio.

Ejemplo 8.11 Bajo la hipótesis nula de independencia entre el nivel de β-caroteno y el riesgo de infarto agudo de miocardio, la probabilidad exacta de obtener la Tabla 8.4 manteniendo los marginales fijos es

$$P(1 \mid 5, 15, 10, 10; H_0) = \frac{\binom{10}{1}\binom{10}{4}}{\binom{20}{5}} = \frac{5!15!10!10!}{20!1!4!9!6!} = 0{,}136,$$

que corresponde a la probabilidad de que, de los 5 sujetos observados con niveles de β-caroteno superiores a 0,30 μg/g, 1 sea caso y los restantes 4 sean controles. Notar que la tabla se refiere por la frecuencia $a = 1$ observada en la primera celda, dado que las demás frecuencias $b = 4$, $c = 9$ y $d = 6$ vienen entonces dadas por los marginales.

Para contrastar la independencia entre las variables de una tabla 2×2, el test exacto de Fisher consiste en enumerar todas las posibles tablas con los mismos marginales que la tabla observada, para a continuación calcular la probabilidad exacta asociada a cada una de estas tablas bajo la hipótesis nula de independencia. El valor P bilateral del test exacto de Fisher corresponde entonces a la suma de probabilidades para todas aquellas tablas con probabilidad inferior o igual a la de la tabla observada (esto es, la suma de probabilidades de las tablas tanto o menos compatibles con la hipótesis nula que la tabla observada).

Ejemplo 8.12 La Tabla 8.5 presenta todas las posibles tablas con los mismos marginales $n_1 = 5$, $n_2 = 15$, $m_1 = 10$ y $m_2 = 10$ observados en la Tabla 8.4 para la asociación entre el β-caroteno y el infarto de miocardio. Bajo la hipótesis nula de independencia entre ambas variables, la probabilidad exacta asociada a cada tabla viene dada por la distribución hipergeométrica

Tabla 8.5 Todas las posibles tablas con los mismos marginales que la Tabla 8.4, junto con sus probabilidades asociadas bajo la hipótesis nula de independencia.

Tabla		Probabilidad bajo H_0	Odds ratio
0	5		
10	5	0,016	0
1	4		
9	6	0,136	0,17
2	3		
8	7	0,348	0,58
3	2		
7	8	0,348	1,71
4	1		
6	9	0,136	6
5	0		
5	10	0,016	∞

$$P(0) = P(5) = \frac{5!15!10!10!}{20!0!5!10!5!} = 0{,}016,$$

$$P(1) = P(4) = \frac{5!15!10!10!}{20!1!4!9!6!} = 0{,}136,$$

$$P(2) = P(3) = \frac{5!15!10!10!}{20!2!3!8!7!} = 0{,}348,$$

cuya suma de probabilidades es igual a 1. Como las tablas con a = 0, 1, 4 y 5 tienen asociadas probabilidades menores o iguales que la probabilidad $P(1) = 0{,}136$ de la tabla observada, el valor P bilateral del test exacto de Fisher es

$$P = P(0) + P(1) + P(4) + P(5)$$
$$= 0{,}016 + 0{,}136 + 0{,}136 + 0{,}016 = 0{,}304.$$

Notar que se obtendría el mismo valor P si se sumaran las probabilidades asociadas a todas aquellas tablas con un odds ratio tanto o más alejado del valor nulo 1 que el $OR = 1{\cdot}6/(4{\cdot}9) = 0{,}17$ de la tabla observada; es decir, las probabilidades de las tablas con $OR \leq 0{,}17$ o $OR \geq 1/0{,}17 = 6$. Así, a partir de esta muestra tan reducida, no puede concluirse que exista una asociación significativa entre el nivel de β-caroteno y el riesgo de infarto de miocardio.

Cuando el tamaño muestral es muy pequeño, el número de posibles tablas con los mismos marginales será muy reducido, de tal forma que el valor P del test exacto de Fisher podrá tomar muy pocos valores, siendo así particularmente difícil obtener resultados significativos. Para un nivel de significación α preestablecido, el test exacto de Fisher tenderá a ser conservador con una verdadera probabilidad de error de Tipo I menor que el valor nominal α. Un contraste alternativo menos conservador consiste en calcular el **valor mid-*P*** bilateral, que se define como la probabilidad de la tabla observada más la probabilidad de las tablas menos verosímiles bajo H_0. Este valor mid-P será siempre inferior o igual al valor exacto de P, obteniéndose resultados muy similares si el tamaño muestral es grande.

Ejemplo 8.13 De todas las posibles tablas enumeradas en la Tabla 8.5, sólo las tablas con $a = 0$ y 5 tienen probabilidades bajo H_0 menores que la probabilidad $P(1) = 0{,}136$ de la tabla observada, así que el valor mid-P bilateral se calcula como

$$\text{mid-}P = P(0) + P(1) + P(5) = 0{,}016 + 0{,}136 + 0{,}016 = 0{,}168,$$

que es considerablemente menor que el valor exacto de $P = 0{,}304$ calculado en el ejemplo anterior. No obstante, ambos valores de P arrojan resultados no significativos para el nivel de significación estándar $\alpha = 0{,}05$.

El test exacto de Fisher puede generalizarse para evaluar la asociación estadística entre las variables categóricas de una tabla $r \times c$, cuando algunas frecuencias esperadas sean muy bajas y no pueda aplicarse el test χ^2 de Pearson. Aunque el valor P del test exacto de Fisher para tablas mayores de 2×2 se define igualmente como la suma de probabilidades para aquellas tablas tanto o menos probables que la tabla observada, su cálculo requiere de algoritmos de computación dado el elevado número de posibles tablas con los mismos marginales.

8.5 REFERENCIAS

1. Agresti A. *Categorical Data Analysis, Second Edition*. New York: John Wiley & Sons, 2002.
2. Bickel PJ, Doksum KA. *Mathematical Statistics: Basic Ideas and Selected Topics*. Englewood Cliffs, NJ: Prentice Hall, 1977.
3. Breslow NE, Day NE. *Statistical Methods in Cancer Research, Volume 1, The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer, 1980.
4. Colton T. *Estadística en Medicina*. Barcelona: Salvat, 1979.
5. Conover WJ. *Practical Nonparametric Statistics, Third Edition*. New York: John Wiley & Sons, 1998.
6. Fleiss JL. *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons, 1986.
7. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions, Third Edition*. New York: John Wiley & Sons, 2003.
8. Hollander M, Wolfe DA. *Nonparametric Statistical Methods, Second Edition*. New York: John Wiley & Sons, 1999.
9. Lehmann EL. *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden and Day, 1975.
10. Rosner B. *Fundamentals of Biostatistics, Fifth Edition*. Belmont, CA: Duxbury Press, 1999.
11. Snedecor GW, Cochran WG. *Statistical Methods, Eighth Edition*. Ames, IA: Iowa State University Press, 1989.

TEMA 9

DETERMINACIÓN DEL TAMAÑO MUESTRAL

9.1 INTRODUCCIÓN

Las inferencias poblacionales derivadas a partir de una muestra conllevan indefectiblemente un margen de error. Así, en el diseño de un estudio epidemiológico o clínico, es necesario plantearse de antemano el número de sujetos que deben ser estudiados para responder a la pregunta de investigación con un grado razonable de certidumbre. La determinación a priori del tamaño muestral es una parte importante del diseño de un estudio por distintos motivos:

- Permite concretar la hipótesis de trabajo. El investigador ha de precisar la hipótesis principal del estudio y, en función de su experiencia, investigaciones previas o estudios piloto, especificar la magnitud de efecto clínica o biológicamente relevante que se pretende detectar.
- Permite evaluar la factibilidad del estudio. Una de las limitaciones más frecuentes en los estudios epidemiológicos es la imposibilidad de reclutar un número suficiente de pacientes, bien sea por limitaciones en los recursos económicos, en el número de pacientes disponibles o en el tiempo de duración del estudio.
- Previene la obtención de resultados no concluyentes. Como se describió en el Tema 5, la precisión de una estimación y la potencia estadística de un contraste de hipótesis aumentan conforme aumenta el tamaño muestral, de tal forma que una muestra insuficiente dará lugar a estimaciones imprecisas y contrastes de baja potencia.

Desde un punto de vista puramente teórico, basta con aumentar el tamaño muestral para obtener estimaciones arbitrariamente precisas o para detectar como estadísticamente significativo cualquier efecto por pequeño que sea. Aun cuando esto sea posible en la práctica, la utilización de muestras excesivamente grandes es ineficiente, ya que la posible detección de efectos trivialmente pequeños y de escasa utilidad práctica no justificaría los recursos empleados. En último término, el objetivo de la determinación a priori del tamaño muestral consiste en estimar la muestra mínima necesaria para asegurar estimaciones razonablemente precisas o para tener una potencia suficiente en la detección de efectos clínicamente relevantes.

Con cierta frecuencia, el número de sujetos disponibles para un estudio viene dictado de antemano por las limitaciones económicas o temporales. En tales circunstancias, es importante determinar qué magnitudes de efecto tendrían una probabilidad razonable de ser detectadas con la muestra disponible, para contar así con una idea aproximada de las posibilidades que ofrecería la realización de dicho estudio.

Como se verá a continuación, el cálculo del tamaño muestral requiere de información previa a la realización del estudio. Estos datos suelen proceder de investigaciones previas relacionadas y, en la medida de lo posible, han de ajustarse a unas hipótesis de trabajo verosímiles. En cualquier caso, las asunciones realizadas en el cálculo del tamaño muestral pueden diferir de los resultados posteriores del estudio y, en consecuencia, estas determinaciones deben servir como guía orientativa más que como norma rígida para la estimación del tamaño muestral. Conviene apuntar también que la muestra resultante se refiere al número de sujetos necesarios para el

análisis y no a los inicialmente incluidos. Así, la muestra estimada ha de incrementarse en previsión de las posibles pérdidas de sujetos que pudieran ocurrir en el estudio.

En este tema se revisan las fórmulas del tamaño muestral más frecuentemente utilizadas en el diseño de estudios epidemiológicos y clínicos, tanto para la estimación de una media y una proporción en una única muestra, como para la comparación de medias y proporciones en muestras dependientes e independientes. En adelante, se asume que las muestras se obtienen mediante un muestreo aleatorio simple a partir de una población de tamaño esencialmente infinito. La corrección de las fórmulas del tamaño muestral para otros tipos de muestreo y para poblaciones finitas puede consultarse en los libros sobre muestreos complejos citados al final del tema.

9.2 TAMAÑO MUESTRAL PARA LA ESTIMACIÓN DE UN PARÁMETRO POBLACIONAL

En esta sección se presentan las fórmulas para determinar el tamaño muestral necesario para obtener estimaciones fiables de un parámetro poblacional (típicamente la media de una variable continua o la proporción de sujetos con una determinada característica) a partir de una única muestra. Esta situación concierne esencialmente a los estudios descriptivos o transversales. El objetivo se centra en calcular el tamaño muestral mínimo necesario para estimar el parámetro poblacional con un determinado grado de precisión, que suele cuantificarse mediante la amplitud del intervalo de confianza.

9.2.1 Tamaño muestral para la estimación de una media

A partir de la aproximación normal $N(\mu, \sigma^2/n)$ a la distribución de una media muestral $\bar{x}$, puede construirse un intervalo de confianza al $100(1 - \alpha)\%$ para la media poblacional μ como $\bar{x} \pm z_{1-\alpha/2}\sigma/\sqrt{n}$. Notar que este intervalo incluye la desviación típica poblacional σ en lugar de su estimación muestral, ya que la determinación del tamaño de una muestra precede a su selección y, en consecuencia, no se dispone de información muestral. La precisión de la estimación δ queda entonces determinada por la amplitud del intervalo de confianza o, más concretamente, por la distancia del centro a los límites del intervalo

$$\delta = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},$$

de donde puede despejarse el tamaño muestral n para obtener

$$n = \frac{z^2_{1-\alpha/2}\sigma^2}{\delta^2}.$$

De esta expresión se desprende que el tamaño muestral para la estimación de una media poblacional depende de tres elementos, que deben ser determinados de antemano para poder aplicar la fórmula:

- El **nivel de confianza** $100(1 - \alpha)\%$. Cuanto mayor sea este nivel de confianza, mayor será el tamaño muestral. En la práctica, suele utilizarse por convenio una confianza del 95% ($\alpha = 0{,}05$), de tal forma que el percentil de la distribución normal estandarizada es $z_{1-\alpha/2} = z_{0{,}975} = 1{,}96$.
- La **varianza poblacional** σ^2. Cuanto más dispersa sea una variable, mayor será la muestra necesaria para describirla aceptablemente. Se requiere, por tanto, de un valor aproximado

de la varianza de la variable a estudio, que suele obtenerse a partir de trabajos similares ya realizados o de un estudio piloto.

- La **precisión deseada** δ. El tamaño muestral será tanto mayor cuanto mayor sea la precisión exigida a la estimación (esto es, cuanto menor sea δ). El criterio para establecer la precisión de una estimación ha de fundamentarse en el conocimiento previo sobre la magnitud aproximada del parámetro. Así, por ejemplo, una precisión de un kilogramo puede ser aceptable para estimar el peso medio en personas adultas, pero resulta claramente insuficiente en recién nacidos.

Ejemplo 9.1 En un pequeño estudio piloto realizado en personas adultas de una determinada población, la media y la desviación típica de la presión arterial sistólica resultaron ser 130 y 20 mm Hg, respectivamente. Utilizando esta información preliminar, se planea obtener una muestra aleatoria simple de mayor tamaño para estimar el nivel medio de presión arterial sistólica con una precisión de ±2 mm Hg. Asumiendo un nivel de confianza del 95% y una desviación típica similar a la del estudio piloto, se tiene

$$n = \frac{1{,}96^2 20^2}{2^2} = 384{,}16;$$

es decir, se requerirían aproximadamente 385 sujetos para estimar la presión arterial sistólica media de esta población con una precisión de ±2 mm Hg. Obsérvese que el tamaño muestral aumenta de forma cuadrática con la precisión deseada, de tal forma que para el doble de precisión $\delta = 1$ mm Hg, el tamaño muestral mínimo necesario sería cuatro veces mayor

$$n = \frac{1{,}96^2 20^2}{1^2} = 1.536{,}64 \approx 1.537.$$

9.2.2 Tamaño muestral para la estimación de una proporción

Siguiendo un argumento similar al del apartado anterior, puede utilizarse la aproximación normal $N(\pi, \pi(1-\pi)/n)$ a la distribución de una proporción muestral p para obtener un intervalo de confianza al $100(1-\alpha)\%$ para la proporción poblacional π mediante $p \pm z_{1-\alpha/2}\sqrt{\pi(1-\pi)/n}$. Así, la precisión δ en la estimación de una proporción poblacional viene determinada por

$$\delta = z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}},$$

y el tamaño muestral mínimo necesario para alcanzar dicha precisión es

$$n = \frac{z^2_{1-\alpha/2}\,\pi(1-\pi)}{\delta^2}.$$

El cálculo del tamaño muestral para la estimación de una proporción precisa, por tanto, de los siguientes elementos:

- El **nivel de confianza** $100(1-\alpha)\%$, que se establece habitualmente en el 95%.
- La **proporción poblacional** π.
- La **precisión deseada** δ o el error absoluto que se considere aceptable.

El conocimiento previo del valor aproximado de la proporción objeto de estudio es necesario no sólo para sustituirlo explícitamente en la fórmula, sino también para establecer la precisión deseada en la estimación. Por ejemplo, un error absoluto del ±5% podría ser admisible en la estimación de una proporción próxima al 50%, mientras que este mismo error sería claramente inaceptable para una proporción pequeña, pongamos del 5% (o equivalentemente para una proporción muy grande, ya que cuando se estima una proporción también se está estimando su complementario). Así, para determinar de antemano qué error se considera admisible, ha de contarse con alguna información sobre la magnitud de π, bien sea a través de investigaciones previas o, en su defecto, de un estudio piloto.

Ejemplo 9.2 En el estudio piloto del ejemplo anterior, la proporción de hipertensos (presión arterial sistólica ≥ 140 mm Hg) fue del 30%. En base a esta información, se pretende realizar un estudio transversal para estimar la prevalencia de hipertensión en esta población con un error absoluto del ±3% (error relativo del ±10%). Asumiendo el nivel de confianza estándar del 95%, $\pi = 0{,}30$ y $\delta = 0{,}03$, se necesitaría una muestra mínima de

$$n = \frac{1{,}96^2 0{,}30(1-0{,}30)}{0{,}03^2} = 896{,}37 \approx 897.$$

Si, por el contrario, el estudio se diseñara para estimar la prevalencia de diabetes, que se asume próxima al 5%, con un error absoluto del ±1% (error relativo del ±20%), se requeriría un tamaño muestral considerablemente mayor

$$n = \frac{1{,}96^2 0{,}05(1-0{,}05)}{0{,}01^2} = 1.824{,}76 \approx 1.825.$$

Como se desprende de este ejemplo, para estimar fiablemente una proporción extrema (muy pequeña o muy grande) se necesitará una muestra mayor que para estimar una proporción cercana al 50%.

La fórmula del tamaño muestral presentada en este apartado se basa en la aproximación normal a la distribución muestral de una proporción. Aunque esta aproximación es razonable en la mayoría de las circunstancias, existen fórmulas alternativas, tales como las basadas en la aproximación normal con corrección por continuidad o en la aproximación de Poisson, que pueden ser útiles cuando se prevé trabajar con muestras de reducido tamaño o con proporciones muy extremas. Una descripción y comparación más detallada de los distintos métodos de cálculo del tamaño muestral puede encontrarse en la bibliografía de este tema.

9.3 TAMAÑO MUESTRAL PARA LA COMPARACIÓN DE MEDIAS

Muchos diseños epidemiológicos, bien sean observacionales (estudios de cohortes o de casos y controles) o experimentales (ensayos clínicos), se realizan con un afán comparativo, donde el objetivo no es tanto estimar la magnitud de un determinado parámetro poblacional, sino más bien comparar parámetros entre distintas poblaciones. En tales diseños, el problema radica en determinar el tamaño muestral mínimo necesario en cada grupo de comparación, de tal forma que el contraste de hipótesis que se pretende realizar tenga una potencia suficiente para detectar posibles diferencias clínica o epidemiológicamente relevantes. En este apartado se presentan

las fórmulas del tamaño muestral para contrastar diferencias en los niveles medios de una variable cuantitativa a partir de dos muestras dependientes o independientes.

9.3.1 Tamaño muestral para la comparación de medias en dos muestras independientes

Supongamos que se pretende contrastar la hipótesis nula H_0: $\mu_1 = \mu_2$ de igualdad de medias frente a la hipótesis alternativa bilateral H_1: $\mu_1 \neq \mu_2$ en dos distribuciones con igual varianza $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Según los resultados del Apartado 6.3, la distribución muestral de la diferencia de medias $\bar{x}_1 - \bar{x}_2$ en muestras independientes de tamaño n_1 y n_2 será aproximadamente normal con media $\mu_1 - \mu_2 = 0$ bajo H_0 y $\mu_1 - \mu_2 \neq 0$ bajo H_1, y varianza $\sigma_1^2/n_1 + \sigma_2^2/n_2 = \sigma^2(1/n_1 + 1/n_2)$ (Figura 9.1). Para asegurar una probabilidad α de cometer un error de tipo I, la hipótesis nula se rechazará sólo si el estadístico

$$\frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{1/n_1 + 1/n_2}} \leq -z_{1-\alpha/2} \text{ ó } \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{1/n_1 + 1/n_2}} \geq z_{1-\alpha/2}$$

o, equivalentemente, si la diferencia de medias

$$\bar{x}_1 - \bar{x}_2 \leq -z_{1-\alpha/2}\,\sigma\sqrt{1/n_1 + 1/n_2} \text{ ó } \bar{x}_1 - \bar{x}_2 \geq z_{1-\alpha/2}\,\sigma\sqrt{1/n_1 + 1/n_2}\,.$$

Así, bajo la hipótesis alternativa, la potencia del test para detectar una diferencia subyacente $\mu_1 - \mu_2$ vendrá dada por

$$1-\beta = P(\bar{x}_1 - \bar{x}_2 \leq -z_{1-\alpha/2}\,\sigma\sqrt{1/n_1 + 1/n_2} \mid H_1)$$
$$+ P(\bar{x}_1 - \bar{x}_2 \geq z_{1-\alpha/2}\,\sigma\sqrt{1/n_1 + 1/n_2} \mid H_1).$$

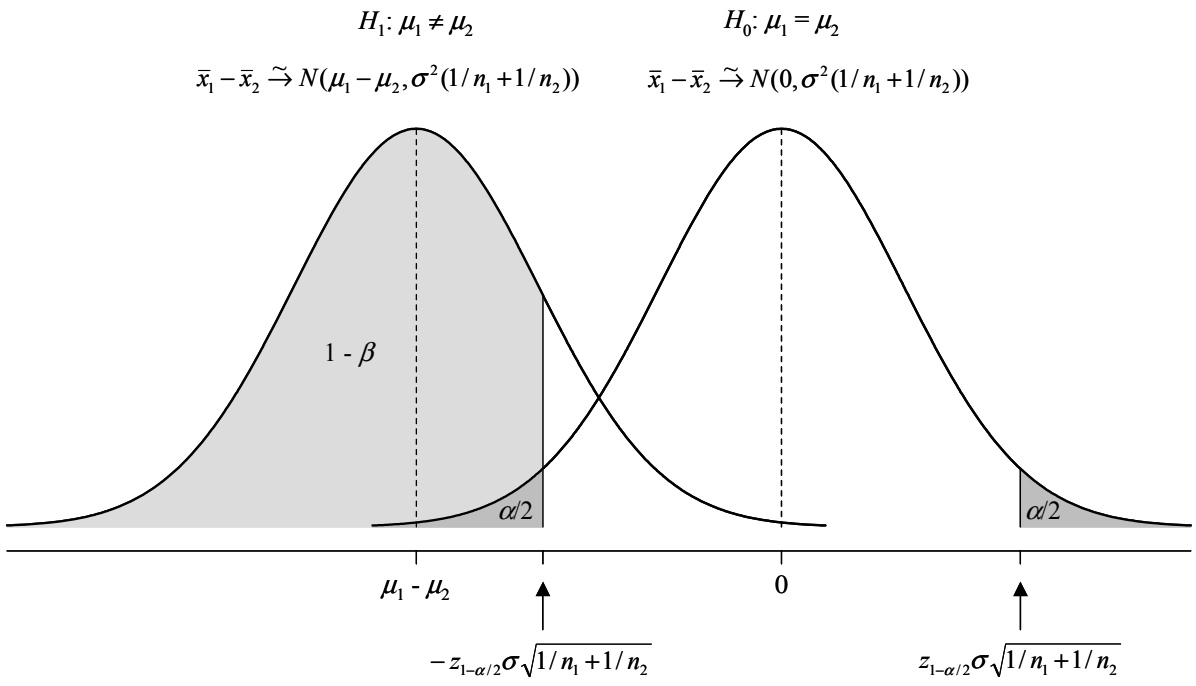


Figura 9.1 **Representación de la potencia del contraste bilateral de medias a partir de dos muestras independientes.**

Asumiendo sin pérdida de generalidad que $\mu_1 < \mu_2$ (Figura 9.1), la segunda probabilidad de la expresión anterior, que representa el evento de que $\bar{x}_1$ sea apreciablemente mayor que $\bar{x}_2$, será virtualmente cero. La potencia se reduce entonces a

$$\begin{aligned} 1-\beta &= P(\bar{x}_1 - \bar{x}_2 \leq -z_{1-\alpha/2}\,\sigma\sqrt{1/n_1 + 1/n_2} \mid H_1) \\ &= P\left(\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \leq \frac{-z_{1-\alpha/2}\,\sigma\sqrt{1/n_1 + 1/n_2} - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \middle| H_1\right) \\ &= \Phi\left(-z_{1-\alpha/2} + \frac{|\mu_1 - \mu_2|}{\sigma\sqrt{1/n_1 + 1/n_2}}\right), \end{aligned}$$

donde la última igualdad se deriva de la distribución normal de $\bar{x}_1 - \bar{x}_2$ bajo la hipótesis alternativa. Notar que se alcanzaría el mismo resultado si $\mu_1 > \mu_2$. Esta expresión permite determinar a posteriori la potencia de un contraste para detectar una diferencia de medias subyacente $\mu_1 - \mu_2$ a partir de dos muestras independientes de tamaños n_1 y n_2.

Ejemplo 9.3 En un ensayo clínico para evaluar la eficacia antihipertensiva de un nuevo fármaco en combinación con un tratamiento estándar, se asignaron aleatoriamente 50 pacientes hipertensos al grupo de monoterapia estándar y otros 50 pacientes de similares características al grupo de tratamiento combinado con el nuevo fármaco. Después de 4 semanas de tratamiento, la media y la desviación típica de la presión arterial sistólica fueron 155 y 22 mm Hg en el grupo de monoterapia, y 150 y 18 mm Hg en el grupo de tratamiento combinado. Como paso previo a la comparación de medias, se contrasta la igualdad de varianzas mediante el estadístico

$$F = \frac{s_1^2}{s_2^2} = \frac{22^2}{18^2} = 1{,}49,$$

que bajo la distribución F de Fisher con $n_1 - 1 = 49$ y $n_2 - 1 = 49$ grados de libertad, corresponde a un valor P bilateral $2P(F_{49,49} \geq 1{,}49) = 2 \cdot 0{,}082 = 0{,}164$. Así, la comparación del nivel medio de presión arterial sistólica entre ambos grupos puede realizarse mediante la prueba t de Student para muestras independientes asumiendo igualdad de varianzas, cuyo estadístico resulta

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{155 - 150}{20{,}1\sqrt{\frac{1}{50} + \frac{1}{50}}} = 1{,}24,$$

donde la varianza combinada es $s^2 = \{(50 - 1)22^2 + (50 - 1)18^2\}/(50 + 50 - 2) = 404$. Utilizando la distribución t de Student con $n_1 + n_2 - 2 = 98$ grados de libertad, el valor P bilateral es $2P(t_{98} \geq 1{,}24) = 2 \cdot 0{,}108 = 0{,}216$; es decir, los resultados del estudio no aportan suficiente evidencia para afirmar que el tratamiento combinado es más eficaz que la monoterapia.

A partir de estos resultados cabría preguntarse si en realidad ambos tratamientos son igualmente eficaces o si, por el contrario, el estudio carece de potencia suficiente para detectar una diferencia que, aun siendo moderada o pequeña, sea importante en términos clínicos. Si se considera clínicamente relevante una diferencia absoluta de $|\mu_1 - \mu_2| = 5$ mm Hg en la presión arterial sistólica media, y asumiendo un nivel de significación $\alpha =$

0,05 y una desviación típica $\sigma = 20$ mm Hg en ambos grupos, la potencia para detectar dicha diferencia en un estudio con $n_1 = n_2 = 50$ sería

$$1-\beta = \Phi\left(-1{,}96 + \frac{5}{20\sqrt{1/50+1/50}}\right) = \Phi(-0{,}71) = 0{,}239.$$

Es decir, únicamente un 23,9% de los estudios con este tamaño muestral detectarían como estadísticamente significativa una diferencia real de 5 mm Hg. Por tanto, no es sorprendente que el estudio anterior arrojara un resultado no significativo, aun cuando exista una diferencia subyacente de dicha magnitud entre ambos tratamientos.

Como ilustra el ejemplo anterior, en el diseño de un estudio es importante determinar a priori qué tamaño muestral será necesario en cada grupo de comparación para evitar la obtención de resultados no concluyentes por falta de potencia. Supongamos, en el caso más general, que se pretende asignar distinto tamaño a ambas muestras $n_2 = kn_1$, donde k es un número positivo prefijado. A partir de la fórmula de la potencia con $n_2 = kn_1$, y recordando que $\Phi(z_{1-\beta}) = 1 - \beta$, se sigue que

$$z_{1-\beta} = -z_{1-\alpha/2} + \frac{|\mu_1 - \mu_2|}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{kn_1}}},$$

de donde puede despejarse n_1 para obtener

$$n_1 = \frac{(k+1)(z_{1-\alpha/2} + z_{1-\beta})^2\sigma^2}{k(\mu_1 - \mu_2)^2},$$

que corresponde al tamaño necesario en la primera muestra y $n_2 = kn_1$ al de la segunda muestra. En el caso particular de que se desee un mismo tamaño muestral en ambos grupos $k = 1$, éste vendrá determinado por

$$n_1 = n_2 = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2\sigma^2}{(\mu_1 - \mu_2)^2}.$$

La asignación de igual tamaño a ambas muestras es, en general, más eficiente ya que da lugar a un menor tamaño total del estudio. No obstante, hay situaciones prácticas en las que es preferible seleccionar muestras de distinto tamaño, aun cuando ello conlleve un aumento de la muestra total para alcanzar la misma potencia; tal es el caso de los estudios donde la disponibilidad de sujetos o los costes difieren entre los grupos, o cuando se requieren estimaciones más precisas en uno de los grupos. Además de estas consideraciones, en el cálculo del tamaño muestral para la comparación de medias es necesario determinar previamente los siguientes elementos:

- El **nivel de significación** α del contraste bilateral, que representa la probabilidad de rechazar erróneamente la hipótesis nula y se establece usualmente en $\alpha = 0{,}05$.
- La **potencia** $1 - \beta$ del contraste, que determina la probabilidad de detectar hipótesis alternativas ciertas y se fija habitualmente en $1 - \beta = 0{,}80$ ó $0{,}90$.
- La **varianza poblacional** σ^2. En la determinación del tamaño muestral suele asumirse que la varianza es común para ambos grupos, ya que generalmente se carece de información previa suficiente para determinar una varianza específica en cada uno de los grupos.
- La **diferencia mínima detectable** $|\mu_1 - \mu_2|$. El tamaño muestral será tanto mayor cuanto menor sea la diferencia que se pretende detectar. La magnitud de esta diferencia debe ser

un valor plausible basado en conocimientos previos, o bien relevante desde el punto de vista clínico o epidemiológico.

Ejemplo 9.4 Dado que el estudio descrito en el ejemplo anterior carecía de potencia suficiente para detectar una diferencia subyacente de 5 mm Hg en la presión arterial sistólica media de los hipertensos bajo monoterapia y tratamiento combinado, se planea realizar un nuevo ensayo clínico que tenga una potencia $1-\beta = 0{,}80$ para detectar posibles diferencias de dicha magnitud. Asumiendo que se pretende asignar el mismo número de pacientes a ambos brazos del ensayo clínico, un nivel de significación $\alpha = 0{,}05$ y una desviación típica $\sigma = 20$ mm Hg similar a la del estudio anterior, el tamaño muestral necesario en cada uno de los grupos sería

$$n_1 = n_2 = \frac{2(z_{0{,}975} + z_{0{,}80})^2 \sigma^2}{(\mu_1 - \mu_2)^2} = \frac{2(1{,}96 + 0{,}84)^2 20^2}{5^2} = 250{,}88 \approx 251,$$

para una muestra total de 251 + 251 = 502 pacientes. Supongamos, por el contrario, que el tratamiento combinado con el nuevo fármaco es muy costoso y que se decide estudiar la mitad de sujetos bajo tratamiento combinado que bajo monoterapia estándar; esto es, $n_2 = 0{,}5 n_1$. En tal caso, el tamaño muestral necesario en el grupo de monoterapia sería

$$n_1 = \frac{(0{,}5+1)(1{,}96 + 0{,}84)^2 20^2}{0{,}5 \cdot 5^2} = 376{,}32 \approx 377$$

y en el grupo de tratamiento combinado $n_2 = 0{,}5 \cdot 376{,}32 = 188{,}16 \approx 189$. El número total de pacientes necesarios para el estudio sería entonces 377 + 189 = 566; es decir, 64 pacientes más de los requeridos en el caso de igual tamaño muestral para alcanzar una misma potencia.

9.3.2 Tamaño muestral para la comparación de medias en dos muestras dependientes

Supongamos que se planea seleccionar n parejas de datos dependientes procedentes de dos poblaciones para contrastar la hipótesis nula H_0: $\mu_1 = \mu_2$ frente a la hipótesis alternativa bilateral H_1: $\mu_1 \neq \mu_2$. Como se discutió en el Apartado 6.4, la media de las diferencias en cada pareja $\bar{d}$ se distribuirá de forma aproximadamente normal $N(0, \sigma_d^2/n)$ bajo H_0 y $N(\mu_1 - \mu_2, \sigma_d^2/n)$ bajo H_1, donde σ_d^2 es la varianza de las diferencias. Para un nivel de significación α preestablecido, el contraste arrojará un resultado significativo cuando la media de las diferencias

$$\bar{d} \leq -z_{1-\alpha/2}\sigma_d / \sqrt{n} \text{ ó } \bar{d} \geq z_{1-\alpha/2}\sigma_d / \sqrt{n}\,.$$

Por tanto, asumiendo como en el apartado anterior que $\mu_1 < \mu_2$, la potencia para detectar una diferencia de medias $\mu_1 - \mu_2$ será aproximadamente igual a

$$\begin{aligned} 1-\beta &= P(\bar{d} \leq -z_{1-\alpha/2}\sigma_d / \sqrt{n} \mid H_1) \\ &= P\left(\frac{\bar{d} - (\mu_1 - \mu_2)}{\sigma_d / \sqrt{n}} \leq \frac{-z_{1-\alpha/2}\sigma_d / \sqrt{n} - (\mu_1 - \mu_2)}{\sigma_d / \sqrt{n}} \middle| H_1 \right) \\ &= \Phi\left(-z_{1-\alpha/2} + \frac{|\mu_1 - \mu_2|}{\sigma_d / \sqrt{n}} \right). \end{aligned}$$

Como por definición $\Phi(z_{1-\beta}) = 1 - \beta$, se sigue que

$$z_{1-\beta} = -z_{1-\alpha/2} + \frac{|\mu_1 - \mu_2|}{\sigma_d / \sqrt{n}},$$

de donde puede despejarse n para obtener el número mínimo de parejas que serán necesarias para detectar una diferencia subyacente $\mu_1 - \mu_2$ con una potencia $1 - \beta$,

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma_d^2}{(\mu_1 - \mu_2)^2}.$$

En la práctica, resulta difícil determinar directamente la varianza de las diferencias σ_d^2 ya que los datos de una misma pareja están correlacionados. Asumiendo igual varianza σ^2 en ambas poblaciones y un coeficiente de correlación ρ entre los valores de una misma pareja, la varianza de las diferencias viene determinada según los resultados del Apartado 3.4 por

$$\sigma_d^2 = \sigma^2 + \sigma^2 - 2\sigma^2\rho = 2\sigma^2(1-\rho).$$

Así, el número de parejas necesarias también puede expresarse como

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma^2 (1-\rho)}{(\mu_1 - \mu_2)^2}$$

que, además de los parámetros descritos en el apartado anterior, depende de la **correlación** entre cada pareja de datos. Si el emparejamiento no es efectivo, de tal forma que ρ está próximo a 0, el número de parejas necesarias para un estudio emparejado será aproximadamente igual al número de sujetos por grupo para un estudio con muestras independientes (notar que si $\rho = 0$, la fórmula anterior se reduce a la obtenida en el caso de muestras independientes del mismo tamaño). Si, por el contrario, el emparejamiento es efectivo, los datos de cada pareja estarán correlacionados positivamente y, en consecuencia, el número de parejas será substancialmente inferior al número de sujetos requeridos en cada grupo de un estudio independiente bajo las mismas condiciones.

Ejemplo 9.5 Con objeto de asegurar la comparabilidad de los pacientes hipertensos bajo monoterapia y tratamiento combinado, se decide diseñar un ensayo clínico emparejado donde, en lugar de asignar distintos pacientes a ambos grupos, cada paciente es sometido a la monoterapia estándar durante un primer periodo de 4 semanas y al tratamiento combinado con el nuevo fármaco durante un segundo periodo de igual duración. Se asume que la desviación típica de la presión arterial sistólica bajo ambos tratamientos es 20 mm Hg, y que el coeficiente de correlación entre las determinaciones tomadas en un mismo sujeto con un intervalo de 4 semanas es aproximadamente 0,50. Para detectar una diferencia subyacente de 5 mm Hg en la presión arterial sistólica media al final de ambos tratamientos con una potencia de 0,80 y un nivel de significación de 0,05, el número de parejas necesarias sería

$$n = \frac{2(1{,}96 + 0{,}84)^2 \, 20^2 (1 - 0{,}50)}{5^2} = 125{,}44 \approx 126;$$

es decir, la mitad de los sujetos que serían necesarios en cada uno de los grupos de un diseño no emparejado (Ejemplo 9.4).

La determinación del tamaño muestral para la comparación de medias en más de dos muestras dependientes o independientes sigue argumentos similares a los descritos en este apartado. No

obstante, para preservar la incertidumbre global del proceso de inferencia, es necesario utilizar técnicas de corrección por las múltiples comparaciones que se pretendan realizar en el análisis (por ejemplo, un ensayo clínico en el que se comparan varios tratamientos frente a placebo). Estos métodos pueden consultarse en los libros de tamaño muestral referenciados al final del tema.

9.4 TAMAÑO MUESTRAL PARA LA COMPARACIÓN DE PROPORCIONES

En esta sección se aborda el problema de la determinación del tamaño muestral en estudios observacionales o ensayos clínicos donde se pretende contrastar diferencias entre proporciones a partir de dos muestras dependientes o independientes. Al igual que en el Apartado 9.2.2, las fórmulas descritas a continuación se fundamentan en la aproximación normal a la distribución muestral de una proporción y, en consecuencia, serán válidas siempre que $n\pi(1-\pi) \geq 5$ en ambos grupos de comparación. En las referencias de este tema pueden consultarse otros métodos alternativos de cálculo del tamaño muestral particularmente útiles para la comparación de proporciones muy extremas en muestras reducidas.

9.4.1 Tamaño muestral para la comparación de proporciones en dos muestras independientes

El propósito se centra en contrastar la hipótesis nula de igualdad de proporciones poblacionales H_0: $\pi_1 = \pi_2$ frente a la hipótesis alternativa bilateral H_1: $\pi_1 \neq \pi_2$ a partir de dos muestras independientes de tamaños n_1 y n_2. Del Apartado 7.3 se desprende que la diferencia de proporciones muestrales $p_1 - p_2$ seguirá aproximadamente una distribución normal $N(0, \pi(1-\pi)(1/n_1 + 1/n_2))$ bajo H_0 y $N(\pi_1 - \pi_2, \pi_1(1-\pi_1)/n_1 + \pi_2(1-\pi_2)/n_2)$ bajo H_1, donde $\pi = (n_1\pi_1 + n_2\pi_2)/(n_1 + n_2)$ es la proporción combinada que se asume común a ambos grupos bajo H_0. El contraste resultará significativo para un nivel α cuando la diferencia de proporciones muestrales

$$p_1 - p_2 \leq -z_{1-\alpha/2}\sqrt{\pi(1-\pi)(1/n_1 + 1/n_2)}$$

o

$$p_1 - p_2 \geq z_{1-\alpha/2}\sqrt{\pi(1-\pi)(1/n_1 + 1/n_2)}.$$

Así, asumiendo sin pérdida de generalidad que $\pi_1 < \pi_2$, la potencia para detectar una diferencia de proporciones subyacente $\pi_1 - \pi_2$ vendrá determinada por

$$\begin{aligned}
1-\beta &= P(p_1 - p_2 \leq -z_{1-\alpha/2}\sqrt{\pi(1-\pi)(1/n_1 + 1/n_2)} \mid H_1) \\
&= P\left(\frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\pi_1(1-\pi_1)/n_1 + \pi_2(1-\pi_2)/n_2}}\right. \\
&\qquad \left. \leq \frac{-z_{1-\alpha/2}\sqrt{\pi(1-\pi)(1/n_1 + 1/n_2)} - (\pi_1 - \pi_2)}{\sqrt{\pi_1(1-\pi_1)/n_1 + \pi_2(1-\pi_2)/n_2}} \middle| H_1\right) \\
&= \Phi\left(\frac{|\pi_1 - \pi_2| - z_{1-\alpha/2}\sqrt{\pi(1-\pi)(1/n_1 + 1/n_2)}}{\sqrt{\pi_1(1-\pi_1)/n_1 + \pi_2(1-\pi_2)/n_2}}\right).
\end{aligned}$$

Si las limitaciones prácticas determinan de antemano el tamaño muestral disponible para un estudio o si el estudio ya ha sido llevado a cabo, la fórmula anterior permitirá calcular la potencia estadística que tendría dicho estudio con la muestra disponible para detectar diferencias de una determinada magnitud.

Ejemplo 9.6 Se planea realizar un estudio de cohortes para evaluar la asociación entre el uso de anticonceptivos orales y el riesgo de cáncer de mama en mujeres entre 40 y 49 años. Para ello, se dispone de una cohorte de 6.000 mujeres en este rango de edad sin evidencia basal de cáncer de mama, que serán seguidas durante un periodo de 5 años para determinar casos incidentes de la enfermedad. Se estima que un 40% de estas mujeres han utilizado regularmente anticonceptivos orales y que la tasa de incidencia de cáncer de mama en este grupo de edad es de $I = 150$ casos por 100.000 personas-año. Para un nivel de significación $\alpha = 0{,}05$, ¿cuál sería la potencia de este estudio para detectar un hipotético aumento del riesgo de cáncer de mama del 50% entre las usuarias de anticonceptivos orales?

Asumiendo una tasa de incidencia constante en los 5 años de seguimiento, la incidencia acumulada o probabilidad de desarrollar un cáncer de mama en esta cohorte durante los próximos 5 años sería aproximadamente $\pi = IA_5 = 0{,}00150 \cdot 5 = 0{,}00750$. Aplicando la regla de la probabilidad total (véase Apartado 2.4), la relación entre esta probabilidad combinada de cáncer de mama en toda la cohorte y las probabilidades específicas por grupo de exposición vendrá dada por

$$\begin{aligned}\pi = P(D) &= P(E)P(D|E) + P(E^c)P(D|E^c) \\ &= 0{,}40\pi_1 + 0{,}60\pi_2 = 0{,}40 \cdot 1{,}50\pi_2 + 0{,}60\pi_2 = 1{,}20\pi_2,\end{aligned}$$

ya que se estima que un 40% de las mujeres son usuarias de anticonceptivos orales y que la probabilidad π_1 de padecer un cáncer de mama entre las usuarias es un 50% superior a la probabilidad π_2 entre las no usuarias. Así, la probabilidad de desarrollar un cáncer de mama en los 5 años de seguimiento sería $\pi_2 = \pi/1{,}20 = 0{,}00750/1{,}20 = 0{,}00625$ entre las no usuarias y $\pi_1 = 1{,}50\pi_2 = 1{,}50 \cdot 0{,}00625 = 0{,}00938$ entre las usuarias de anticonceptivos orales. Como se espera que $n_1 = 0{,}40 \cdot 6.000 = 2.400$ mujeres de la muestra sean usuarias de estos anticonceptivos y las restantes $n_2 = 0{,}60 \cdot 6.000 = 3.600$ no usuarias, la potencia de este estudio sería

$$\begin{aligned}1 - \beta &= \Phi\left(\frac{|\,0{,}00938 - 0{,}00625\,| - 1{,}96\sqrt{0{,}00750(1-0{,}00750)(1/2.400 + 1/3.600)}}{\sqrt{0{,}00938(1-0{,}00938)/2.400 + 0{,}00625(1-0{,}00625)/3.600}}\right) \\ &= \Phi\left(\frac{0{,}00313 - 1{,}96 \cdot 0{,}00227}{0{,}00237}\right) = \Phi(-0{,}56) = 0{,}287;\end{aligned}$$

es decir, la probabilidad de detectar un hipotético incremento del riesgo de cáncer de mama del 50% entre las usuarias y no usuarias de anticonceptivos orales sería únicamente del 28,7% a partir de una cohorte de 6.000 mujeres seguidas durante 5 años.

La expresión anterior de la potencia permite asimismo determinar a priori la muestra mínima que será necesaria en cada uno de los grupos para alcanzar una potencia preestablecida $1 - \beta$ en la detección de una diferencia subyacente de proporciones $\pi_1 - \pi_2$. En general, si se prevé asignar distinto tamaño a ambas muestras $n_2 = kn_1$, se sigue a partir de la fórmula de la potencia que

$$z_{1-\beta} = \frac{|\pi_1 - \pi_2| - z_{1-\alpha/2}\sqrt{\pi(1-\pi)\left(\frac{1}{n_1}+\frac{1}{kn_1}\right)}}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{kn_1}}}$$

$$= \frac{|\pi_1 - \pi_2| - z_{1-\alpha/2}\sqrt{\frac{(k+1)\pi(1-\pi)}{kn_1}}}{\sqrt{\frac{k\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}{kn_1}}},$$

de tal forma que el tamaño muestral requerido será

$$n_1 = \frac{(z_{1-\alpha/2}\sqrt{(k+1)\pi(1-\pi)} + z_{1-\beta}\sqrt{k\pi_1(1-\pi_1)+\pi_2(1-\pi_2)})^2}{k(\pi_1-\pi_2)^2}$$

en la primera muestra y $n_2 = kn_1$ en la segunda muestra, donde la proporción combinada en ambas muestras viene dada por $\pi = (n_1\pi_1 + n_2\pi_2)/(n_1 + n_2) = (\pi_1 + k\pi_2)/(1 + k)$. En el caso de asignar igual tamaño a ambos grupos de comparación $k = 1$, el tamaño muestral en cada una de las muestras se reduce a

$$n_1 = n_2 = \frac{(z_{1-\alpha/2}\sqrt{2\pi(1-\pi)} + z_{1-\beta}\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)})^2}{(\pi_1-\pi_2)^2},$$

donde la proporción combinada es $\pi = (\pi_1 + \pi_2)/2$. Como se comentó anteriormente, la asignación de igual tamaño a las dos muestras es más eficiente al requerir un menor tamaño total del estudio para alcanzar una misma potencia. Sin embargo, en el diseño de determinados estudios (ver ejemplos posteriores), la selección de muestras de distinto tamaño puede resultar más factible en términos de coste o disponibilidad de pacientes. En cualquier caso, la determinación del tamaño muestral para la comparación de proporciones en muestras independientes precisa de los siguientes elementos:

- El **nivel de significación** α del contraste bilateral, que suele establecerse por convenio en $\alpha = 0{,}05$.
- La **potencia** $1 - \beta$ para detectar hipótesis alternativas ciertas. La mayoría de los estudios se diseñan con una potencia $1 - \beta = 0{,}80$ ó $0{,}90$.
- Las **proporciones poblacionales** π_1 y π_2. A diferencia de la comparación de medias, no es suficiente con determinar la diferencia de proporciones que se pretende detectar, sino que es necesario especificar la magnitud aproximada de esta proporción en cada grupo de comparación, para contar así con un valor aproximado de las varianzas poblacionales $\pi_1(1 - \pi_1)$ y $\pi_2(1 - \pi_2)$.

Ejemplo 9.7 Como se vio en el ejemplo anterior, una cohorte de 6.000 mujeres carece de potencia suficiente para detectar un hipotético incremento del 50% en la incidencia acumulada de cáncer de mama en 5 años entre las mujeres usuarias y no usuarias de anticonceptivos orales. Según los cálculos del ejemplo anterior, la incidencia acumulada en este periodo en una cohorte de mujeres entre 40 y 49 años será aproximadamente $\pi = 0{,}00750$, siendo $\pi_1 = 0{,}00938$ y $\pi_2 = 0{,}00625$ las respectivas incidencias acumuladas

en usuarias y no usuarias. Como se prevé que la cohorte esté compuesta de un 40% de mujeres usuarias de anticonceptivos orales y un 60% de no usuarias, se tiene que $n_2 = 1{,}5n_1$. Asumiendo un nivel de significación $\alpha = 0{,}05$ y una potencia $1 - \beta = 0{,}80$, se necesitarían

$$n_1 = \frac{(1{,}96\sqrt{2{,}5 \cdot 0{,}00744} + 0{,}84\sqrt{1{,}5 \cdot 0{,}00929 + 0{,}00621})^2}{1{,}5(0{,}00938 - 0{,}00625)^2}$$

$$= 10.202{,}55 \approx 10.203$$

mujeres usuarias de estos anticonceptivos y $n_2 = 1{,}5 \cdot 10.202{,}55 = 15.303{,}82 \approx 15.304$ no usuarias. Así, para detectar un aumento subyacente del riesgo de cáncer de mama del 50% entre las usuarias de anticonceptivos orales con una potencia de 0,80, se precisaría de una cohorte inicial de 25.507 mujeres seguidas durante un periodo de 5 años.

El tamaño necesario de la cohorte se reduciría si el seguimiento del estudio se extendiera, por ejemplo, hasta los 10 años, ya que el número esperado de eventos aumentaría considerablemente. Siguiendo argumentos similares a los del ejemplo anterior, la incidencia acumulada en toda la cohorte durante 10 años sería $\pi = 0{,}01500$, y las incidencias acumuladas específicas entre las usuarias y no usuarias de anticonceptivos orales serían $\pi_1 = 0{,}01875$ y $\pi_2 = 0{,}01250$, respectivamente. La cohorte necesaria consistiría entonces en

$$n_1 = \frac{(1{,}96\sqrt{2{,}5 \cdot 0{,}01478} + 0{,}84\sqrt{1{,}5 \cdot 0{,}01840 + 0{,}01234})^2}{1{,}5(0{,}01875 - 0{,}01250)^2}$$

$$= 5.061{,}27 \approx 5.062$$

usuarias de anticonceptivos orales y $n_2 = 1{,}5 \cdot 5.061{,}27 = 7.591{,}90 \approx 7.592$ no usuarias; es decir, 12.654 mujeres seguidas a lo largo de 10 años.

Ejemplo 9.8 Dado que la realización de un estudio prospectivo requeriría de una gran cantidad de personas-año de seguimiento para obtener un número suficiente de casos de cáncer de mama, resultará más viable llevar a cabo un estudio de casos y controles. En tal caso, el propósito se centrará en seleccionar un número suficiente de casos y controles para detectar un odds ratio de cáncer de mama $\omega = 1{,}50$ entre las usuarias y no usuarias de anticonceptivos orales con una potencia $1 - \beta = 0{,}80$. Si los controles seleccionados constituyen una muestra representativa de la población de referencia, la proporción de utilización de anticonceptivos orales entre las mujeres del grupo control será aproximadamente $\pi_2 = 0{,}40$. A partir de la expresión del odds ratio en estudios de casos y controles (véase Apartado 7.6.2), se tiene que

$$\omega = \frac{P(E \mid D)P(E^c \mid D^c)}{P(E \mid D^c)P(E^c \mid D)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)},$$

de donde puede despejarse la proporción π_1 de mujeres que han usado anticonceptivos orales entre los casos de cáncer de mama como

$$\pi_1 = \frac{\omega\pi_2}{1 + (\omega - 1)\pi_2} = \frac{1{,}50 \cdot 0{,}40}{1 + 0{,}50 \cdot 0{,}40} = 0{,}50.$$

Para un nivel de significación estándar $\alpha = 0{,}05$ y asumiendo la selección del mismo número de casos que controles, de tal forma que la proporción combinada $\pi = (\pi_1 + \pi_2)/2 = (0{,}50 + 0{,}40)/2 = 0{,}45$, el número necesario de casos y controles sería

$$n_1 = n_2 = \frac{(1{,}96\sqrt{2 \cdot 0{,}45(1-0{,}45)} + 0{,}84\sqrt{0{,}50(1-0{,}50) + 0{,}40(1-0{,}40)}\ )^2}{(0{,}50-0{,}40)^2}$$

$$= 386{,}90 \approx 387,$$

para una muestra total de 774 mujeres.

Supongamos que, dada la baja incidencia de cáncer de mama, la disponibilidad de casos incidentes de esta enfermedad en la población es limitada y, por tanto, se decide reclutar el doble de controles que de casos. Así, $n_2 = 2n_1$ y la proporción combinada será $\pi = (\pi_1 + k\pi_2)/(1 + k) = (0{,}50 + 2 \cdot 0{,}40)/3 = 0{,}43$. La muestra necesaria estaría entonces compuesta por

$$n_1 = \frac{(1{,}96\sqrt{3 \cdot 0{,}43(1-0{,}43)} + 0{,}84\sqrt{2 \cdot 0{,}50(1-0{,}50) + 0{,}40(1-0{,}40)}\ )^2}{2(0{,}50-0{,}40)^2}$$

$$= 289{,}17 \approx 290$$

casos de cáncer de mama y $n_2 = 2 \cdot 289{,}17 = 578{,}33 \approx 579$ controles libres de la enfermedad. El tamaño total sería 290 + 579 = 869; es decir, 95 mujeres más de las requeridas en un estudio con el mismo número de casos que controles.

9.4.2 Tamaño muestral para la comparación de proporciones en dos muestras dependientes

Supongamos que se pretende contrastar la hipótesis nula H_0: $\pi_1 = \pi_2$ frente a la hipótesis alternativa bilateral H_1: $\pi_1 \neq \pi_2$ a partir de n parejas de datos dependientes. Para simplificar la exposición, supondremos además que se trata de un estudio de casos y controles emparejados uno a uno, donde π_1 y π_2 representan las respectivas proporciones poblacionales de expuestos a un determinado factor antecedente entre casos y controles. Como las parejas concordantes reflejan una misma exposición en caso y control, la hipótesis nula de igualdad de proporciones en un diseño emparejado es equivalente a H_0: $\pi_b = \pi_c$, donde π_b es la proporción de parejas discordantes con el caso expuesto y π_c es la proporción de parejas discordantes con el control expuesto. Según la notación de la Tabla 7.6, las proporciones muestrales de ambos tipos de pares discordantes serán $p_b = b/n$ y $p_c = c/n$. Estas proporciones estarán obviamente correlacionadas, de tal forma que el valor esperado de la diferencia será $E(p_b - p_c) = \pi_b - \pi_c$ y su varianza (véase Apartado 3.4)

$$\text{var}(p_b - p_c) = \text{var}(p_b) + \text{var}(p_c) - 2\text{cov}(p_b, p_c)$$

$$= \frac{\pi_b(1-\pi_b)}{n} + \frac{\pi_c(1-\pi_c)}{n} + \frac{2\pi_b\pi_c}{n}$$

$$= \frac{(\pi_b + \pi_c) - (\pi_b - \pi_c)^2}{n},$$

donde la covarianza negativa entre p_b y p_c viene dada por $\text{cov}(p_b, p_c) = -\pi_b\pi_c/n$. Así, la diferencia en la proporción muestral de parejas discordantes $p_b - p_c$ seguirá aproximadamente una distribución normal $N(0, (\pi_b + \pi_c)/n)$ bajo H_0 y $N(\pi_b - \pi_c, \{(\pi_b + \pi_c) - (\pi_b - \pi_c)^2\}/n)$ bajo H_1.

Para un nivel de significación α, el contraste arrojará un resultado significativo cuando

$$p_b - p_c \leq -z_{1-\alpha/2}\sqrt{(\pi_b + \pi_c)/n} \ \text{ ó } p_b - p_c \geq z_{1-\alpha/2}\sqrt{(\pi_b + \pi_c)/n}\,.$$

Asumiendo sin pérdida de generalidad que $\pi_b < \pi_c$, la probabilidad del segundo evento será despreciable bajo la hipótesis alternativa y la potencia podrá entonces aproximarse mediante

$$1-\beta = P(p_b - p_c \leq -z_{1-\alpha/2}\sqrt{(\pi_b+\pi_c)/n} \mid H_1)$$

$$= P\left(\frac{p_b - p_c - (\pi_b - \pi_c)}{\sqrt{\{(\pi_b+\pi_c)-(\pi_b-\pi_c)^2\}/n}} \leq \frac{-z_{1-\alpha/2}\sqrt{(\pi_b+\pi_c)/n} - (\pi_b-\pi_c)}{\sqrt{\{(\pi_b+\pi_c)-(\pi_b-\pi_c)^2\}/n}} \Big| H_1\right)$$

$$= \Phi\left(\frac{|\pi_b - \pi_c| - z_{1-\alpha/2}\sqrt{(\pi_b+\pi_c)/n}}{\sqrt{\{(\pi_b+\pi_c)-(\pi_b-\pi_c)^2\}/n}}\right).$$

A partir de esta expresión, se sigue que el número total de parejas necesarias para alcanzar una potencia $1-\beta$ es

$$n = \frac{(z_{1-\alpha/2}\sqrt{\pi_b+\pi_c} + z_{1-\beta}\sqrt{(\pi_b+\pi_c)-(\pi_b-\pi_c)^2})^2}{(\pi_b-\pi_c)^2},$$

para cuyo cálculo se precisa de una idea aproximada de las **probabilidades de obtener ambos tipos de parejas discordantes** π_b y π_c. Aunque son pocos los diseños emparejados donde se cuenta con información a priori de estas probabilidades, las siguientes consideraciones generales pueden resultar útiles en la práctica. Si el emparejamiento no fuera efectivo, pongamos por ejemplo un estudio de casos y controles donde las variables de emparejamiento no estuvieran asociadas con la exposición principal, el nivel de exposición sería entonces virtualmente independiente entre caso y control, de tal forma que la proporción esperada de parejas con el caso expuesto y el control no expuesto sería $\pi_b = \pi_1(1-\pi_2)$ y con el control expuesto y el caso no expuesto $\pi_c = \pi_2(1-\pi_1)$, para una proporción total de pares discordantes $\pi_b + \pi_c = \pi_1(1-\pi_2) + \pi_2(1-\pi_1)$. En tal caso, puede probarse que el número necesario de parejas coincidiría aproximadamente con el número de sujetos por grupo en un estudio de casos y controles independientes; resultado esperable siempre que se empareje por características irrelevantes. Por el contrario, si el emparejamiento fuera efectivo, esto es, si los factores pronósticos empleados en el emparejamiento estuvieran asociados con la exposición a estudio, los casos y controles se asemejarían en su nivel de exposición, induciendo así una correlación positiva en la exposición de cada pareja de caso y control. Las parejas discordantes serían entonces menos probables $\pi_b + \pi_c < \pi_1(1-\pi_2) + \pi_2(1-\pi_1)$ y, en consecuencia, para obtener un número suficiente de pares discordantes para el análisis, el número total de parejas habría de ser superior al número de sujetos por grupo en un estudio independiente. En general, la comparación de proporciones en muestras emparejadas tiene menor potencia que la comparación cruda de proporciones en muestras independientes, pero mayor validez interna al controlar los posibles sesgos derivados de los factores de confusión utilizados en el emparejamiento.

Ejemplo 9.9 En el estudio de casos y controles independientes del ejemplo anterior, cabría esperar que la edad media de los casos sea superior a la de los controles ya que la incidencia de cáncer de mama aumenta con la edad. Además, como la edad está inversamente relacionada con el uso de anticonceptivos orales, esta variable podría provocar una confusión negativa en la asociación a estudio, de tal forma que el odds ratio obtenido de la comparación cruda de casos y controles independientes tendería a infraestimar el potencial efecto nocivo del uso de anticonceptivos orales en el riesgo de cáncer de mama.

Para evitar esta posible confusión, se decide diseñar un estudio de casos y controles emparejados, donde cada caso de cáncer de mama se empareja aleatoriamente con un control de su misma edad. Como consecuencia de este emparejamiento por edad, se induciría un cierto grado de correlación positiva en la utilización de anticonceptivos de cada pareja. Así, la proporción esperada de pares discordantes sería inferior a $\pi_1(1-\pi_2) + \pi_2(1-\pi_1) = 0{,}50(1-0{,}40) + 0{,}40(1-0{,}50) = 0{,}50$, donde $\pi_1 = 0{,}50$ y $\pi_2 = 0{,}40$ son las proporciones poblacionales de usuarias de anticonceptivos orales entre casos y controles obtenidas del ejemplo anterior. Asumiendo una correlación moderada, podría establecerse a priori una proporción aproximada de parejas discordantes $\pi_b + \pi_c = 0{,}40$. Para un hipotético odds ratio de cáncer de mama $\omega = \pi_b/\pi_c = 1{,}50$, se esperaría entonces una proporción de parejas con el control usuario de anticonceptivos orales y el caso no usuario $\pi_c = (\pi_b + \pi_c)/(\omega + 1) = 0{,}40/2{,}50 = 0{,}16$, y con el caso usuario y el control no usuario $\pi_b = \omega\pi_c = 1{,}50 \cdot 0{,}16 = 0{,}24$. Así, el número total de parejas necesarias para detectar dicho efecto con una potencia $1-\beta = 0{,}80$ y un nivel de significación $\alpha = 0{,}05$ sería

$$n = \frac{(1{,}96\sqrt{0{,}24+0{,}16} + 0{,}84\sqrt{(0{,}24+0{,}16)-(0{,}24-0{,}16)^2}\,)^2}{(0{,}24-0{,}16)^2}$$

$$= 487{,}64 \approx 488,$$

con lo que se tendrían aproximadamente $0{,}40 \cdot 488 = 195$ pares discordantes para el análisis. Notar que el número de parejas requeridas para este estudio sería mayor que los 387 casos y controles necesarios en el correspondiente estudio independiente (Ejemplo 9.8). No obstante, el análisis emparejado de casos y controles de igual edad eliminaría la posibilidad de sesgos por diferencias de edad entre casos y controles.

El cálculo del tamaño muestral puede extenderse a la comparación de tres o más proporciones en muestras dependientes o independientes. Aunque las fórmulas se derivan siguiendo procedimientos similares a los aquí descritos, suelen emplearse métodos de corrección del nivel de significación α para preservar la probabilidad global de obtener un resultado significativo entre las múltiples comparaciones que se pretendan realizar (ver referencias bibliográficas).

9.5 REFERENCIAS

1. Breslow NE, Day NE. *Statistical Methods in Cancer Research, Volume 2, The Design and Analysis of Cohort Studies.* Lyon: International Agency for Research on Cancer, 1987.
2. Cochran WG. *Sampling Techniques, Third Edition.* New York: John Wiley & Sons, 1977.
3. Desu MM, Raghavarao D. *Sample Size Methodology.* Boston: Academic Press, 1990.
4. Fleiss JL. *The Design and Analysis of Clinical Experiments.* New York: John Wiley & Sons, 1986.
5. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions, Third Edition.* New York: John Wiley & Sons, 2003.
6. Lemeshow S, Hosmer DW, Klar J, Lwanga SK. *Adequacy of Sample Size in Health Studies.* New York: John Wiley & Sons, 1990.
7. Levy PS, Lemeshow S. *Sampling of Populations: Methods and Applications, Third Edition.* New York: John Wiley & Sons, 1999.
8. Rosner B. *Fundamentals of Biostatistics, Fifth Edition.* Belmont, CA: Duxbury Press, 1999.
9. Silva LC. *Diseño Razonado de Muestras y Captación de Datos para la Investigación Sanitaria.* Madrid: Díaz de Santos, 2000.

TEMA 10

CORRELACIÓN Y REGRESIÓN LINEAL SIMPLE

10.1 INTRODUCCIÓN

En el Tema 6 se discutieron las técnicas estadísticas adecuadas para comparar los niveles medios de una variable continua en dos grupos de sujetos definidos según la presencia o ausencia de una determinada característica dicotómica; esto es, la dependencia entre una variable continua y otra dicotómica. Asimismo, en el Tema 7 se presentaron distintos procedimientos para determinar la existencia o no de asociación entre dos variables dicotómicas. Queda pendiente, por tanto, describir los métodos necesarios para evaluar la relación entre dos variables continuas.

En este tema se presentan el coeficiente de correlación y la regresión lineal simple como las dos técnicas estadísticas más utilizadas para investigar la relación entre dos variables continuas X e Y. Como veremos más adelante, ambos procedimientos están estrechamente relacionados, aunque obedecen a estrategias de análisis un tanto diferentes. Por un lado, el coeficiente de correlación determina el grado de asociación lineal entre X e Y, sin establecer a priori ninguna direccionalidad en la relación entre ambas variables. Por el contrario, la regresión lineal simple permite cuantificar el cambio en el nivel medio de la variable Y conforme cambia la variable X, asumiendo implícitamente que X es la variable explicativa o independiente e Y es la variable respuesta o dependiente.

10.2 COEFICIENTE DE CORRELACIÓN

Como ya se anticipó en el Apartado 3.4, el parámetro más utilizado para medir la asociación lineal entre dos variables aleatorias X e Y es el **coeficiente de correlación poblacional** ρ_{xy}, que se define como

$$\rho_{xy} = \frac{\text{cov}(X,Y)}{\sigma_x\ \sigma_y} = \frac{E\{(X-\mu_x)(Y-\mu_y)\}}{\sigma_x\ \sigma_y},$$

donde μ_x y μ_y son las respectivas medias poblacionales de X e Y y σ_x y σ_y son sus correspondientes desviaciones típicas poblacionales. El numerador del coeficiente de correlación $\text{cov}(X, Y) = E\{(X-\mu_x)(Y-\mu_y)\}$ es la **covarianza poblacional** entre ambas variables y se define como la esperanza del producto de las desviaciones de cada variable respecto de su media. Así, si valores altos (o bajos) de X tienden a asociarse con valores altos (o bajos) de Y, el producto de las desviaciones $(x-\mu_x)(y-\mu_y)$ tenderá a ser positivo y la covarianza será positiva. Por el contrario, si valores altos de una variable se relacionan con valores bajos de la otra variable, el producto de las desviaciones tenderá a ser negativo y la covarianza será negativa. No obstante, resulta complicado determinar el grado de asociación lineal entre dos variables a partir de la magnitud de la covarianza, ya que ésta depende de las unidades de medida de las variables.

Al dividir la covarianza por el producto de las desviaciones típicas de X e Y, el coeficiente de correlación poblacional carece de unidades y permanece inalterable ante cambios de origen o escala en cualquiera de las dos variables. Puede comprobarse, además, que la covarianza entre X e Y es menor en valor absoluto que el producto de sus desviaciones típicas y, en consecuencia,

el coeficiente de correlación siempre está comprendido entre –1 y 1. En el caso extremo de que $\rho_{xy} = 1$, las variables estandarizadas $Z_x = (X - \mu_x)/\sigma_x$ y $Z_y = (Y - \mu_y)/\sigma_y$ verifican que (véase Apartado 3.4)

$$\text{var}(Z_x - Z_y) = \text{var}(Z_x) + \text{var}(Z_y) - 2\text{cov}(Z_x, Z_y) = 2(1 - \rho_{xy}) = 0;$$

es decir, $Z_x - Z_y$ es una variable aleatoria degenerada (constante) en su valor esperado, $Z_x - Z_y = E(Z_x - Z_y) = 0$, lo que implica que las variables X e Y presentan una relación lineal positiva perfecta, $Y = \mu_y + \sigma_y/\sigma_x(X - \mu_x)$. De igual forma, si $\rho_{xy} = -1$, se cumple que

$$\text{var}(Z_x + Z_y) = \text{var}(Z_x) + \text{var}(Z_y) + 2\text{cov}(Z_x, Z_y) = 2(1 + \rho_{xy}) = 0$$

y, por tanto, $Z_x + Z_y$ es una variable aleatoria constante igual a su valor esperado, $Z_x + Z_y = E(Z_x + Z_y) = 0$, de donde se deduce que las variables X e Y presentan una relación lineal negativa perfecta, $Y = \mu_y - \sigma_y/\sigma_x(X - \mu_x)$. Cuando $\rho_{xy} = 0$, se dice que las variables están linealmente **incorrelacionadas** ya que no existe relación lineal entre ambas variables. Notar que si dos variables son estadísticamente independientes, en el sentido de que el conocimiento del valor que toma una variable no aporta ninguna información sobre el valor de la otra variable, entonces están incorrelacionadas; pero que la incorrelación no implica necesariamente independencia, ya que las variables podrían presentar una dependencia no lineal aun cuando $\rho_{xy} = 0$.

El coeficiente de correlación permite, por tanto, cuantificar el grado de asociación lineal entre dos variables, de tal forma que cuanto más próximo esté el coeficiente de correlación a 1 ó –1, mayor será la dependencia lineal positiva o negativa entre las variables. Este hecho se ilustra en los **diagramas de dispersión** de la Figura 10.1, donde se representan los valores de la variable X en el eje horizontal y los correspondientes valores de Y en el eje vertical. A medida que los puntos del diagrama de dispersión se desvían de una línea recta perfecta con pendiente positiva o negativa, el coeficiente de correlación se aleja de 1 ó –1. Aunque la interpretación de la magnitud del coeficiente de correlación depende del contexto particular de aplicación, en términos generales se considera que una correlación es baja por debajo de 0,30 en valor absoluto, moderada entre 0,30 y 0,50, y alta por encima de 0,50.

Notar, por último, que en la interpretación del coeficiente de correlación hay dos errores frecuentes que deben ser evitados:

- El coeficiente de correlación entre X e Y no es una medida de la magnitud de la pendiente de la recta de regresión entre ambas variables. El coeficiente de correlación determina el grado de aproximación de los puntos del diagrama de dispersión a una línea recta, independientemente de cuál sea la magnitud de la pendiente de dicha recta. Como se ilustra en los paneles *a* y *b* de la Figura 10.2, el coeficiente de correlación es mayor en el panel *a*, a pesar de que la pendiente de la recta de regresión es mayor en el panel *b*. La pendiente de la recta de regresión no se determina mediante el coeficiente de correlación, sino mediante las técnicas de regresión lineal simple que se discutirán en la segunda parte de este tema.
- El coeficiente de correlación no es una medida de la idoneidad del modelo lineal. El coeficiente de correlación sólo determina la existencia de una componente lineal en la relación entre dos variables, independientemente de la forma subyacente de dicha relación. Así, por ejemplo, el coeficiente de correlación es mayor en el panel *d* que en el panel *c* de la Figura 10.2, aun cuando la relación subyacente entre las variables del panel *d* es claramente no lineal (en este caso, cuadrática). Por ello, antes de analizar el grado de asociación lineal entre dos variables, es aconsejable inspeccionar la naturaleza de la relación mediante un diagrama de dispersión.

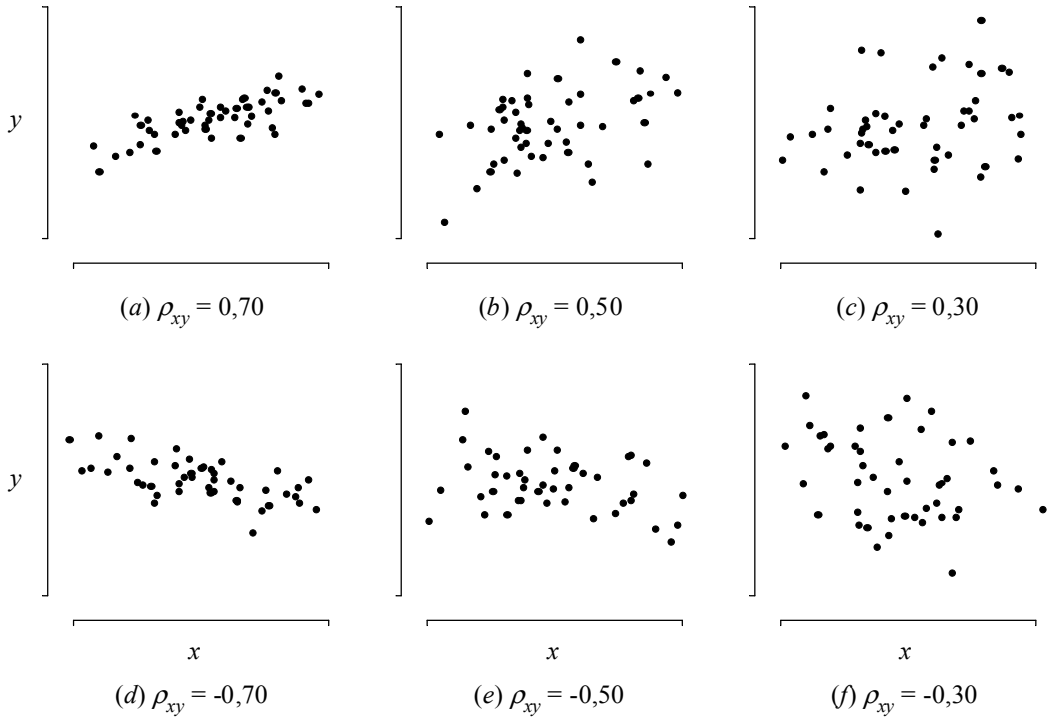


Figura 10.1 **Diagramas de dispersión entre dos variables aleatorias *X* e *Y* con coeficientes de correlación positivos ρ_{xy} = 0,70 (*a*), 0,50 (*b*) y 0,30 (*c*), así como con coeficientes de correlación negativos ρ_{xy} = −0,70 (*d*), −0,50 (*e*) y −0,30 (*f*).**

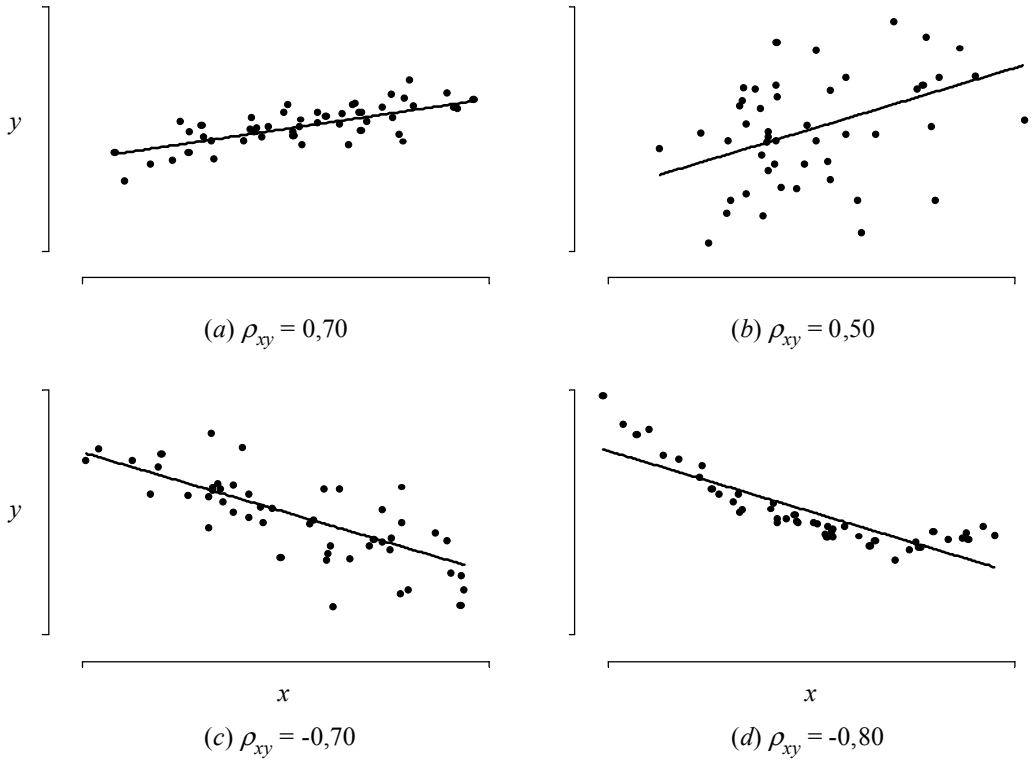


Figura 10.2 **Diagramas de dispersión, coeficientes de correlación y rectas de regresión entre dos variables aleatorias *X* e *Y* con distintas pendientes de la recta de regresión (paneles *a* y *b*) y distintas formas de la relación subyacente (paneles *c* y *d*).**

10.2.1 Coeficiente de correlación muestral de Pearson

Una vez descritas las propiedades e interpretación del coeficiente de correlación poblacional, en este apartado se presentan los métodos para estimar el coeficiente de correlación entre dos variables X e Y a partir de los valores observados de ambas variables (x_i, y_i) en una muestra de n sujetos mutuamente independientes, $i = 1, ..., n$.

El estimador muestral más utilizado para evaluar la dependencia lineal entre dos variables X e Y es el coeficiente de correlación muestral de Pearson, que se denota por r_{xy}, o simplemente por r, y se define como la covarianza muestral entre X e Y dividida por el producto de sus desviaciones típicas muestrales,

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

donde $\bar{x}$ y s_x son la media y la desviación típica muestral de X y $\bar{y}$ y s_y son la media y la desviación típica muestral de Y. Así, el coeficiente de correlación muestral de Pearson se define de forma análoga al coeficiente de correlación poblacional, reemplazando la covarianza y las desviaciones típicas poblacionales por sus correspondientes estimadores muestrales. Al igual que el coeficiente de correlación poblacional, el coeficiente de correlación muestral siempre toma valores entre -1 y 1, de tal forma que cuanto más se aproxime a 1 ó -1, mayor será la dependencia lineal positiva o negativa entre las variables.

Ejemplo 10.1 En la Figura 10.3 se presenta el diagrama de dispersión entre el índice de masa corporal, medida de obesidad que se obtiene de dividir el peso en kilogramos por la

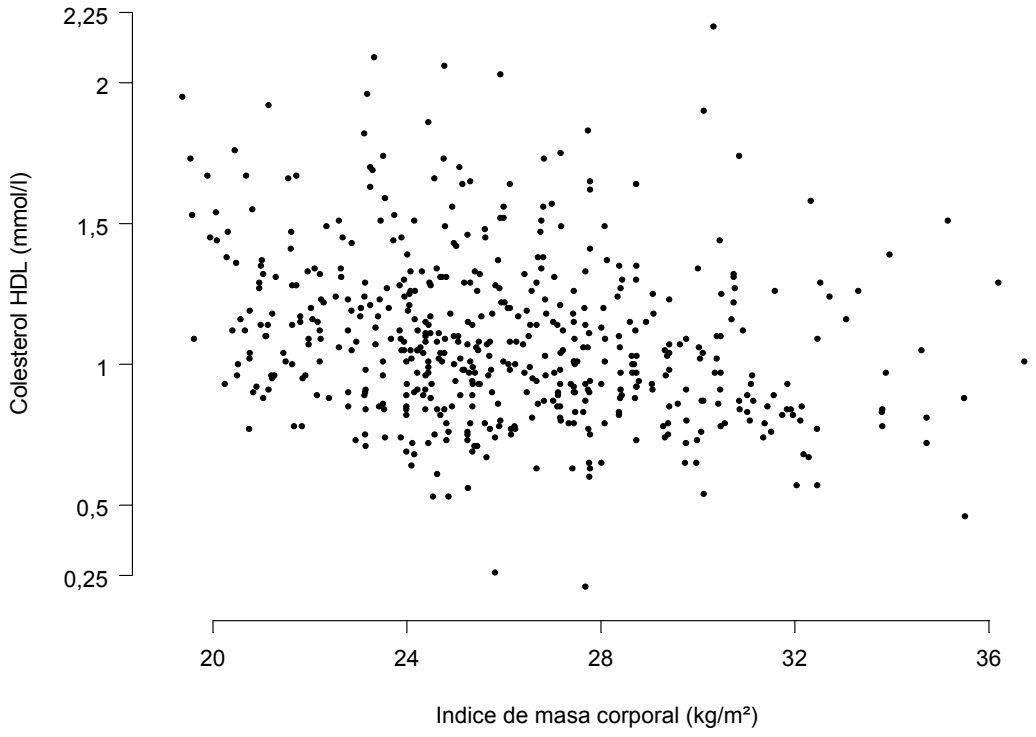


Figura 10.3 Diagrama de dispersión entre el índice de masa corporal y el colesterol HDL en el grupo control del estudio EURAMIC.

altura en metros al cuadrado, y el colesterol HDL en los 533 controles del estudio EURAMIC con valores para ambas variables. A simple vista, se aprecia un cierto grado de dependencia lineal negativa entre ambas variables; esto es, el colesterol HDL tiende a decrecer conforme aumenta el índice de masa corporal. Esta apreciación visual se confirma mediante el cálculo del coeficiente de correlación muestral de Pearson,

$$r = \frac{\frac{1}{532}\sum_{i=1}^{533}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{-0{,}285}{3{,}50 \cdot 0{,}295} = -0{,}276,$$

que indica una asociación lineal negativa moderada entre el índice de masa corporal y el colesterol HDL.

El coeficiente de correlación r de Pearson tiene una distribución muestral tanto más asimétrica cuanto más distante esté la correlación subyacente ρ del valor 0. Cuando ρ está relativamente próximo a 1 ó -1, las estimaciones muestrales del coeficiente de correlación tenderán por fuerza a desviarse más del parámetro ρ en la cola que no está limitada por el rango $[-1, 1]$ de valores posibles de r, resultando en una distribución con un marcado sesgo negativo o positivo. Por ello, el cálculo de un intervalo de confianza y un test de hipótesis para ρ no suele realizarse a partir de la distribución muestral de r, sino mediante la **transformación z de Fisher**

$$z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right),$$

cuya distribución muestral presenta una mayor simetría para cualquier valor de ρ. Puede probarse que si las distribuciones poblacionales de las variables X e Y no distan mucho del modelo normal y el tamaño muestral no es muy pequeño, típicamente $n > 50$, la transformación z de Fisher se distribuye de forma aproximadamente normal con media $\log\{(1+\rho)/(1-\rho)\}/2$ y varianza $1/(n-3)$,

$$z \overset{\sim}{\to} N\left(\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right).$$

Notar que la varianza de z es inversamente proporcional al tamaño muestral e independiente de la correlación subyacente ρ.

Ejemplo 10.2 Las Figuras 10.4(a) y (b) muestran las distribuciones del coeficiente de correlación r de Pearson y de la transformación z de Fisher entre el índice de masa corporal y el colesterol HDL en 1000 muestras aleatorias simples de tamaño 50 obtenidas a partir de los controles del estudio EURAMIC. La distribución muestral de r presenta un leve sesgo positivo ya que el percentil 75 ($-0{,}18$) está ligeramente más alejado de la mediana ($-0{,}28$) que el percentil 25 ($-0{,}36$). Para corregir esta leve asimetría, la transformación z de Fisher aumenta la dispersión de los valores de r más distantes de 0 (cola inferior de la distribución) y mantiene virtualmente constantes los valores próximos a 0 (cola superior), dando lugar así a una distribución sensiblemente más simétrica.

En este ejemplo, la distribución muestral del coeficiente de correlación r de Pearson presenta una leve asimetría ya que la correlación subyacente $-0{,}276$ en todos los controles del estudio EURAMIC es moderadamente baja. En otras situaciones donde la correlación subyacente ρ sea alta, la distribución muestral de r será notablemente asimétrica y, en consecuencia, el efecto normalizador de la transformación z de Fisher será mucho más marcado.

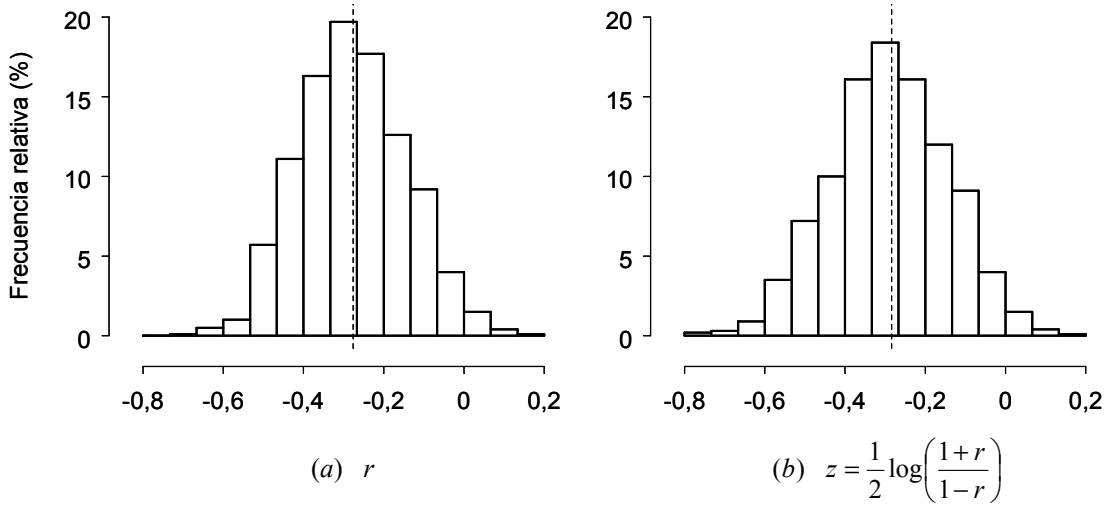


Figura 10.4 **Distribución muestral del coeficiente de correlación *r* de Pearson (*a*) y de la transformación *z* de Fisher (*b*) entre el índice de masa corporal y el colesterol HDL en 1000 muestras aleatorias simples de tamaño 50 obtenidas a partir de los controles del estudio EURAMIC. Las líneas verticales en trazo discontinuo representan los parámetros subyacentes $\rho = -0{,}276$ y $\log\{(1+\rho)/(1-\rho)\}/2 = -0{,}284$.**

En base a la distribución muestral de la transformación *z* de Fisher, el intervalo de confianza al $100(1-\alpha)\%$ para el parámetro $\log\{(1+\rho)/(1-\rho)\}/2$ viene dado por

$$(z_1, z_2) = z \pm z_{1-\alpha/2}\frac{1}{\sqrt{n-3}},$$

donde $z_{1-\alpha/2}$ es el percentil $1-\alpha/2$ de la distribución normal estandarizada. Así, el intervalo de confianza al $100(1-\alpha)\%$ para el coeficiente de correlación poblacional ρ se obtiene de aplicar el inverso de la transformación de Fisher a ambos límites del intervalo,

$$\left(\frac{\exp(2z_1)-1}{\exp(2z_1)+1}, \frac{\exp(2z_2)-1}{\exp(2z_2)+1}\right).$$

Este intervalo para ρ es tanto más asimétrico alrededor de la estimación puntual *r* cuanto mayor sea *r* en valor absoluto y menor sea el tamaño muestral. Asimismo, el contraste de la hipótesis nula $H_0: \rho = \rho_0$ frente a la hipótesis alternativa bilateral $H_1: \rho \neq \rho_0$ se realiza mediante el estadístico

$$\frac{z - \frac{1}{2}\log\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\frac{1}{\sqrt{n-3}}},$$

que bajo H_0 sigue aproximadamente una distribución normal estandarizada. El valor *P* del contraste se calcula, por tanto, como el área bajo la curva normal estandarizada para aquellos valores tanto o más distantes de 0 que el valor observado del estadístico.

Ejemplo 10.3 A partir de 533 controles del estudio EURAMIC, la estimación puntual del coeficiente de correlación entre el índice de masa corporal y el colesterol HDL fue $r = -0{,}276$. La transformación *z* de Fisher de esta correlación es $z = \log\{(1-0{,}276)/(1+0{,}276)\}/2 = -0{,}284$. Para obtener una estimación por intervalo de la correlación subyacente ρ entre ambas

variables en la población de referencia del estudio EURAMIC, se calcula en primer lugar el IC al 95% para el parámetro $\log\{(1+\rho)/(1-\rho)\}/2$ como

$$-0{,}284 \pm z_{0{,}975}\frac{1}{\sqrt{533-3}} = -0{,}284 \pm 1{,}96{\cdot}0{,}043 = (-0{,}369; -0{,}199)$$

y, a continuación, se aplica el inverso de la transformación de Fisher a ambos límites del intervalo

$$\left(\frac{\exp\{2(-0{,}369)\}-1}{\exp\{2(-0{,}369)\}+1}, \frac{\exp\{2(-0{,}199)\}-1}{\exp\{2(-0{,}199)\}+1}\right) = (-0{,}353; -0{,}196).$$

Notar que el intervalo resultante es ligeramente asimétrico respecto a la estimación puntual $r = -0{,}276$. Para contrastar la hipótesis de ausencia de asociación lineal entre ambas variables H_0: $\rho = 0$, se calcula el estadístico

$$-0{,}284\sqrt{533-3} = -6{,}53,$$

que corresponde a un valor P bilateral bajo la distribución normal estandarizada $2P(Z \leq -6{,}53) = 2\Phi(-6{,}53) < 0{,}001$. En conclusión, existe una asociación lineal moderada pero significativa entre el índice de masa corporal y el colesterol HDL con un coeficiente de correlación de −0,28 (IC al 95% −0,35 a −0,20; $P < 0{,}001$).

10.2.2 Coeficiente de correlación de los rangos de Spearman

Al igual que la media y la desviación típica muestral, el coeficiente de correlación de Pearson es sensible a la presencia de valores extremos en alguna de las variables, que podrían distorsionar la estimación resultante, no siendo entonces un buen reflejo de la asociación lineal subyacente entre ambas variables. Además, las inferencias basadas en la transformación de Fisher del coeficiente de correlación muestral asumen que las variables se distribuyen de forma aproximadamente normal y que el tamaño muestral es suficientemente grande. En aquellas situaciones donde exista una clara evidencia en contra de la normalidad, o bien cuando la muestra sea muy pequeña, estas inferencias pueden resultar engañosas y es preferible utilizar métodos no paramétricos. En este apartado se presenta el coeficiente de correlación de los rangos de Spearman como un procedimiento no paramétrico para detectar la existencia de una relación monótona (creciente o decreciente, aunque no necesariamente lineal) entre dos variables cualesquiera, que pueden ser variables continuas con distribuciones subyacentes no normales o incluso variables cualitativas ordinales.

Si se desea determinar el grado en que dos variables se relacionan de forma monótona sin realizar ninguna asunción sobre la distribución poblacional de ambas variables, basta con utilizar el orden de las observaciones de cada variable en lugar de sus verdaderos valores. Así, a cada sujeto se le asignan los **rangos** r_i y s_i en función de la posición que ocupan sus respectivos valores observados x_i e y_i dentro de la muestra ordenada ascendentemente por X e Y. En el caso de que existan varias observaciones con el mismo valor de una variable (empates), se asigna a cada una de ellas la media de los rangos correspondientes. El coeficiente de correlación r_s de Spearman se calcula simplemente como el coeficiente de correlación de Pearson reemplazando los valores observados (x_i, y_i) por sus correspondientes rangos (r_i, s_i),

$$r_s = \frac{\sum_{i=1}^{n}(r_i-\bar{r})(s_i-\bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i-\bar{r})^2}\sqrt{\sum_{i=1}^{n}(s_i-\bar{s})^2}},$$

donde los rangos medios son $\bar{r} = \bar{s} = (n + 1)/2$. El coeficiente de correlación de Spearman siempre toma valores entre -1 y 1. Si $r_s = 1$, los rangos son necesariamente idénticos $s_i = r_i$, de tal forma que si dos observaciones cualesquiera de la variable X verifican que $x_i < x_j$, sus correspondientes valores de la variable Y preservan dicho orden $y_i < y_j$; es decir, los valores observados de las variables X e Y presentan una relación monótona creciente perfecta. De igual forma, si $r_s = -1$, los rangos verifican que $s_i = n + 1 - r_i$, de donde se deduce que los valores de las variables X e Y presentan una relación monótona decreciente perfecta. Cuando $r_s = 0$, los rangos están incorrelacionados y no existe relación monótona alguna entre los valores de ambas variables.

En el caso de que no haya valores idénticos (empates) en ninguna de las variables, el cálculo del coeficiente de correlación de Spearman se simplifica notablemente ya que la varianza de los rangos es

$$\frac{1}{n-1}\sum_{i=1}^{n}(r_i-\bar{r})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(s_i-\bar{s})^2$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)^2 = \frac{n(n+1)}{12}$$

y su covarianza es

$$\frac{1}{n-1}\sum_{i=1}^{n}(r_i-\bar{r})(s_i-\bar{s}) = \frac{1}{2(n-1)}\sum_{i=1}^{n}\{(r_i-\bar{r})^2+(s_i-\bar{s})^2-(r_i-s_i)^2\}$$
$$= \frac{n(n+1)}{12} - \frac{1}{2(n-1)}\sum_{i=1}^{n}(r_i-s_i)^2.$$

Aplicando ambos resultados, el coeficiente de correlación de Spearman se reduce a

$$r_s = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(r_i-s_i)^2,$$

fórmula que sólo puede emplearse cuando no hay empates.

Ejemplo 10.4 En la Tabla 10.1 se presentan los niveles de α-tocoferol y β-caroteno en tejido adiposo en una muestra aleatoria de 10 controles del estudio EURAMIC, junto con los rangos correspondientes a los valores de ambas variables. A partir de estos rangos, el coeficiente de correlación de Spearman se calcula como

$$r_s = \frac{\frac{1}{9}\sum_{i=1}^{10}(r_i-\bar{r})(s_i-\bar{s})}{\sqrt{\frac{1}{9}\sum_{i=1}^{10}(r_i-\bar{r})^2}\sqrt{\frac{1}{9}\sum_{i=1}^{10}(s_i-\bar{s})^2}} = \frac{5{,}06}{3{,}03\cdot 3{,}03} = 0{,}552,$$

o de forma equivalente mediante la fórmula simplificada en ausencia de empates

$$r_s = 1 - \frac{6}{10(10^2-1)}\{(7-3)^2+\ldots+(6-6)^2\} = 1 - \frac{6\cdot 74}{10(10^2-1)} = 0{,}552,$$

que refleja una fuerte relación monótonamente creciente entre los niveles de α-tocoferol y β-caroteno. Cabe destacar que esta estimación no esta influenciada por el valor extremo 1,46 μg/g de β-caroteno ya que el rango de esta observación continuaría siendo 10 para cualquier valor arbitrariamente mayor que los demás.

Tabla 10.1 α-tocoferol y β-caroteno en tejido adiposo en una muestra aleatoria de 10 controles del estudio EURAMIC.

	α-tocoferol		β-caroteno	
Control	Valor (µg/g)	Rango (r_i)	Valor (µg/g)	Rango (s_i)
1	163,8	7	0,14	3
2	331,9	10	0,45	8
3	125,1	4	0,07	1
4	42,9	1	0,44	7
5	211,0	8	1,46	10
6	115,9	2	0,18	4
7	128,6	5	0,37	5
8	271,0	9	0,66	9
9	118,8	3	0,11	2
10	128,7	6	0,40	6

Al igual que otros procedimientos no paramétricos, el coeficiente de correlación de los rangos de Spearman permite contrastar la hipótesis nula de ausencia de asociación monótona entre dos variables. Bajo esta hipótesis nula, se ha comprobado que el coeficiente de correlación r_s de Spearman tiende a distribuirse de forma normal o, más concretamente, que el estadístico

$$t = \frac{r_s}{\sqrt{\frac{1-r_s^2}{n-2}}}$$

sigue aproximadamente una distribución t de Student con $n-2$ grados de libertad, siempre que el tamaño muestral sea $n > 10$. Así, el valor P bilateral del contraste puede aproximarse mediante el área bajo la distribución t_{n-2} para valores tanto o más alejados de 0 que el valor observado del estadístico t. Aparte del mínimo requerimiento muestral, este contraste tiene la ventaja adicional de poder aplicarse a cualquier distribución subyacente de las variables X e Y, a diferencia del contraste paramétrico basado en el coeficiente de correlación de Pearson que requiere de distribuciones poblacionales aproximadamente normales.

Ejemplo 10.5 Como las distribuciones subyacentes del α-tocoferol y el β-caroteno (Figura 4.3) son marcadamente asimétricas en los controles del estudio EURAMIC, el contraste bilateral de la hipótesis de no asociación entre ambas variables a partir de los 10 controles de la Tabla 10.1 ha de realizarse mediante el estadístico basado en la correlación de los rangos de Spearman

$$t = \frac{r_s}{\sqrt{\frac{1-r_s^2}{n-2}}} = \frac{0{,}552}{\sqrt{\frac{1-0{,}552^2}{8}}} = 1{,}87,$$

que bajo la distribución t de Student con 8 grados de libertad corresponde a un valor aproximado de $P = 2P(t_8 \geq 1{,}87) = 0{,}098$. Así, aunque el coeficiente de correlación de Spearman $r_s = 0{,}55$ estima una fuerte relación monótonamente creciente entre los valores observados de α-tocoferol y β-caroteno, esta asociación no llega a ser estadísticamente significativa, probablemente debido a la escasa potencia del test para detectar cualquier asociación subyacente con tan reducido tamaño muestral.

Cuando el tamaño muestral es inferior o igual a 10, la distribución t de Student no es una buena aproximación a la distribución muestral del estadístico t y, en consecuencia, el contraste

debe basarse en la distribución exacta del coeficiente de correlación de Spearman bajo la hipótesis nula. Si no existe ninguna relación monótona entre las variables, y los rangos r_i de la variable X se asumen constantes, cualquier permutación $s_1, ..., s_n$ de los rangos de la variable Y es igualmente probable y su probabilidad viene dada por $1/n!$. Haciendo uso de este resultado, es posible derivar la distribución bajo la hipótesis nula del coeficiente de correlación de Spearman, cuyos percentiles en muestras de tamaño $n \leq 10$ se presentan en la Tabla 10 del Apéndice. Para un contraste bilateral con un nivel de significación α preestablecido, la hipótesis de no asociación se rechazará si el coeficiente de correlación r_s de Spearman es inferior al percentil $\alpha/2$ o superior al percentil $1 - \alpha/2$ de dicha tabla.

Ejemplo 10.6 El valor exacto de P para el contraste bilateral de la hipótesis de no asociación entre el α-tocoferol y el β-caroteno viene dado por

$$P = P(r_s \geq 0{,}552|H_0) + P(r_s \leq -0{,}552|H_0) = 2P(r_s \geq 0{,}552|H_0),$$

ya que la distribución bajo H_0 del coeficiente de correlación de Spearman es simétrica alrededor de 0. Utilizando la Tabla 10 del Apéndice para $n = 10$, se tiene que el percentil $r_{s;0,95} = 0{,}552$, de lo cual se deduce que $P = 2P(r_s \geq 0{,}552|H_0) \geq 2 \cdot 0{,}05 = 0{,}10$. Este valor exacto de P es similar al valor aproximado mediante la distribución t de Student en el ejemplo anterior.

10.3 REGRESIÓN LINEAL SIMPLE

Las técnicas de regresión evalúan la relación entre dos variables siguiendo una estrategia de análisis distinta a la correlación. Mientras que el coeficiente de correlación determina el grado de asociación lineal entre X e Y tratando ambas variables de forma simétrica, la regresión lineal estudia la variación en el nivel medio de la variable respuesta Y a medida que cambia la variable explicativa X, estableciendo así una direccionalidad en la relación entre dichas variables. Aunque en ocasiones la elección entre la variable respuesta y explicativa es un tanto arbitraria (por ejemplo, en la asociación entre el α-tocoferol y el β-caroteno), la direccionalidad suele establecerse de forma natural por el propio diseño del estudio o la naturaleza de las variables (por ejemplo, los cambios medios en el colesterol HDL conforme aumenta el índice de masa corporal).

El modelo de regresión lineal asume que la media de la variable respuesta Y cambia linealmente con la variable explicativa X; esto es, para un valor fijo x de la variable explicativa, el valor esperado de la variable respuesta es

$$E(Y|x) = \beta_0 + \beta_1 x,$$

donde β_0 y β_1 son la constante y la pendiente de la **recta de regresión**, respectivamente. La constante β_0 determina la media de Y cuando $X = 0$, $E(Y|0) = \beta_0 + \beta_1 0 = \beta_0$, y la pendiente β_1 corresponde al cambio en el valor medio de Y por cada aumento de una unidad en X, $E(Y|x + 1) - E(Y|x) = \beta_0 + \beta_1(x + 1) - (\beta_0 + \beta_1 x) = \beta_1$. La especificación del modelo se completa asumiendo que los valores individuales de la variable respuesta se distribuyen de forma normal alrededor del valor esperado definido por la recta de regresión. Así, la estructura general del modelo de regresión lineal es

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

donde el término de error aleatorio ε, que representa la desviación de cada respuesta individual Y respecto de la recta de regresión $\beta_0 + \beta_1 x$, se distribuye de forma normal con media 0 y

varianza σ^2. Por tanto, la regresión lineal establece que para un valor fijo x de la variable explicativa, la variable respuesta Y sigue una distribución normal con media $E(Y|x) = \beta_0 + \beta_1 x + E(\varepsilon) = \beta_0 + \beta_1 x$ y varianza $\text{var}(Y|x) = \text{var}(\varepsilon) = \sigma^2$,

$$Y|x \sim N(\beta_0 + \beta_1 x, \sigma^2),$$

de donde se derivan las siguientes asunciones:

- **Linealidad**: El valor esperado de la variable respuesta Y es una función lineal de la variable explicativa X, de tal forma que cambios de magnitud constante a distintos niveles de X se asocian con un mismo cambio en el valor medio de Y.
- **Homogeneidad de la varianza**: La varianza de la variable respuesta Y es la misma para cualquier valor de la variable explicativa X; es decir, a diferencia de la media, la varianza de Y no está relacionada con X.
- **Normalidad**: Para un valor fijo de la variable explicativa X, la variable respuesta Y sigue una distribución normal.

Las asunciones subyacentes al modelo de regresión lineal se representan gráficamente en la Figura 10.5. Estas asunciones facilitan el proceso de inferencia sobre la recta de regresión y su idoneidad debe ser evaluada utilizando técnicas diagnósticas, algunas de las cuales se presentan al final de este tema.

En regresión lineal simple se estudia la distribución condicional de una variable respuesta continua en función de una única variable explicativa. Esta variable explicativa puede ser tanto continua como categórica ya que el modelo de regresión lineal no establece ninguna asunción respecto a su distribución. La extensión de estos modelos al análisis de regresión lineal múltiple, donde se consideran simultáneamente dos o más variables explicativas, se tratará en el Tema 11.

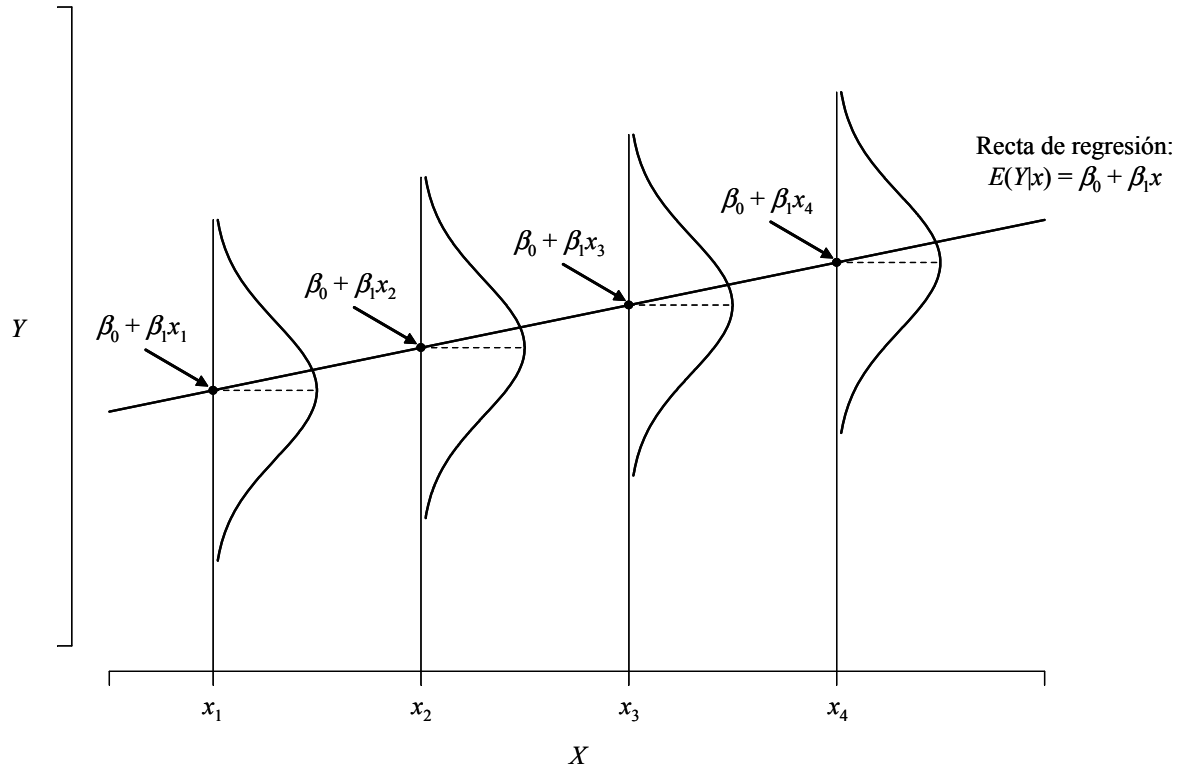


Figura 10.5 **Asunciones estadísticas subyacentes al modelo de regresión lineal simple.**

10.3.1 Estimación de la recta de regresión

El primer objetivo de la regresión lineal es obtener estimaciones puntuales b_0 y b_1 de la constante β_0 y la pendiente β_1 de la recta de regresión que mejor se ajuste a los valores observados (x_i, y_i) de las variables explicativa y respuesta en una muestra de n sujetos mutuamente independientes. Intuitivamente, se trataría de identificar la línea recta que más se aproxime al conjunto de todos los puntos del diagrama de dispersión entre ambas variables. Para formalizar esta idea, es preciso calcular la distancia de cada punto observado (x_i, y_i) respecto al punto correspondiente $(x_i, \hat{y}_i) = (x_i, b_0 + b_1 x_i)$ sobre la recta de regresión estimada en x_i. Esta distancia, que se representa en la Figura 10.6, viene dada por el error de estimación en la variable respuesta $e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$. Así, la recta de regresión vendrá determinada por aquellos valores b_0 y b_1 que hagan este error lo más pequeño posible para todas las observaciones o, equivalentemente, que minimicen la **suma de cuadrados del error**

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2,$$

también llamada suma de cuadrados residual. Notar que los errores se elevan al cuadrado para evitar que se compensen los errores positivos y negativos. Este procedimiento para estimar los parámetros de la recta de regresión se conoce como el **método de mínimos cuadrados**.

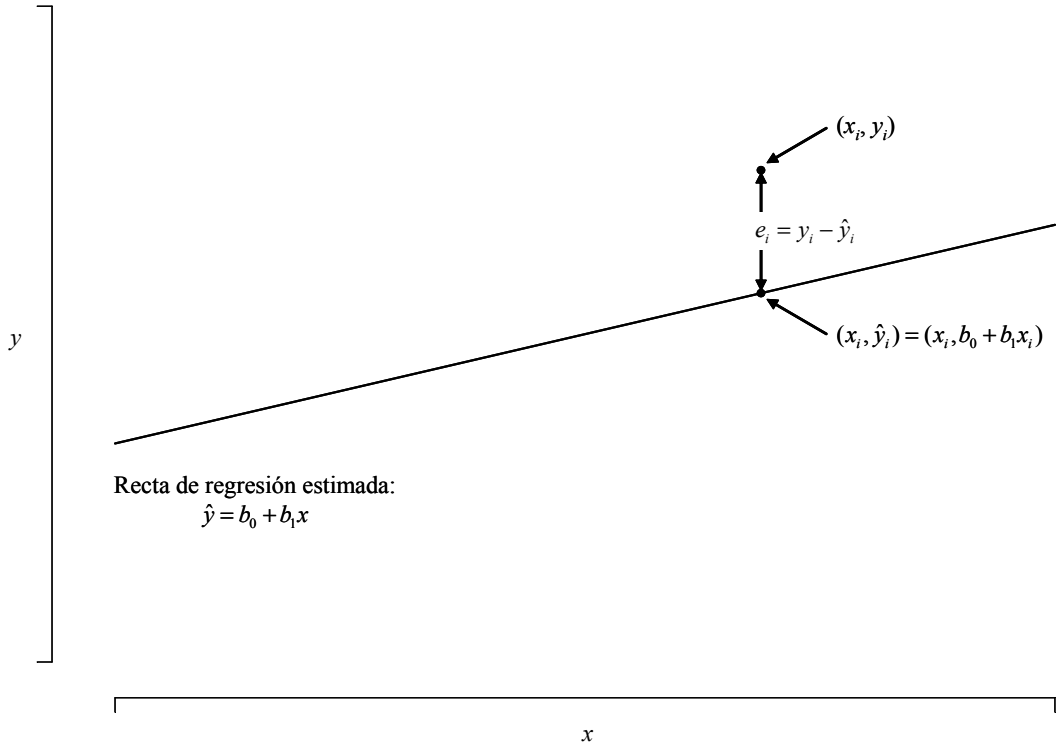


Figura 10.6 **Error o desviación del valor observado de la variable respuesta respecto a su valor estimado por la recta de regresión.**

Para obtener los valores b_0 y b_1 que minimizan la suma de cuadrados del error, se calculan las derivadas parciales de SSE respecto a b_0 y b_1 y se igualan a cero, resultando el sistema de ecuaciones lineales

$$\frac{\partial \text{SSE}}{\partial b_0} = -2\sum_{i=1}^{n} e_i = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0,$$

$$\frac{\partial \text{SSE}}{\partial b_1} = -2\sum_{i=1}^{n} x_i e_i = -2\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0,$$

cuya solución es

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = r\frac{s_y}{s_x},$$

$$b_0 = \bar{y} - b_1 \bar{x}.$$

La pendiente estimada b_1 de la recta de regresión es igual al producto del coeficiente de correlación r de Pearson por el cociente entre las desviaciones típicas muestrales de Y y X. Así, aunque los signos de b_1 y r coinciden, la magnitud de la pendiente b_1 no sólo depende del coeficiente de correlación r, sino también de las desviaciones típicas s_y y s_x de las variables. Una vez estimada la pendiente, la constante $b_0 = \bar{y} - b_1\bar{x}$ corresponde simplemente al valor que fuerza a la recta de regresión a atravesar el punto $(\bar{x}, \bar{y})$ correspondiente a la media muestral de ambas variables. Si la relación subyacente entre las variables es lineal (asunción de linealidad), b_0 y b_1 son estimadores insesgados de la constante β_0 y la pendiente β_1 de la recta de regresión.

La recta de regresión estimada viene entonces determinada por

$$\hat{y} = b_0 + b_1 x = \bar{y} + b_1(x - \bar{x}),$$

que facilita una estimación del valor esperado o predicho de la variable respuesta para cada valor fijo de la variable explicativa. Para completar la estimación de los parámetros del modelo lineal, ha de estimarse también la varianza σ^2 de la variable respuesta alrededor de dicha recta. A partir de la suma de cuadrados del error, esta **varianza residual** puede estimarse mediante

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

Cabe destacar que la suma de cuadrados del error se divide por $n - 2$ ya que, una vez estimadas la constante y la pendiente de la recta de regresión, los n errores o desviaciones de la variable respuesta respecto de la recta contienen $n - 2$ grados de libertad (conocidos b_0, b_1 y $n - 2$ errores, los 2 errores restantes se derivan automáticamente). Asumiendo que se cumplen las hipótesis de linealidad y homogeneidad de la varianza, la varianza residual s^2 es un estimador insesgado del parámetro poblacional σ^2.

Ejemplo 10.7 En el estudio de la relación entre el índice de masa corporal y el colesterol HDL, resulta natural considerar el índice de masa corporal como variable explicativa y el colesterol HDL como variable respuesta. El objetivo es, por tanto, estimar los cambios en

el nivel medio del colesterol HDL conforme aumenta el índice de masa corporal utilizando un modelo de regresión lineal simple. En este caso, tanto la variable respuesta como la variable explicativa son continuas.

En $n = 533$ controles del estudio EURAMIC, la media y la desviación típica del índice de masa corporal fueron $\bar{x} = 26{,}0$ y $s_x = 3{,}50$ kg/m^2, y los correspondientes valores del colesterol HDL fueron $\bar{y} = 1{,}09$ y $s_y = 0{,}295$ mmol/l. Además, en el Ejemplo 10.1 se obtuvo un coeficiente de correlación de Pearson entre ambas variables de $r = -0{,}276$. A partir de estos datos, las estimaciones de la pendiente y la constante de la recta de regresión por el método de mínimos cuadrados son

$$b_1 = r\frac{s_y}{s_x} = -0{,}276\frac{0{,}295}{3{,}50} = -0{,}023$$

y

$$b_0 = \bar{y} - b_1\bar{x} = 1{,}09 + 0{,}023 \cdot 26{,}0 = 1{,}69.$$

La constante $b_0 = 1{,}69$ mmol/l es una estimación del valor esperado de colesterol HDL para un sujeto con un índice de masa corporal igual a 0 kg/m^2, extrapolación que carece de sentido biológico. La pendiente $b_1 = -0{,}023$ estima que, por cada incremento de 1 kg/m^2 en el índice de masa corporal, el nivel medio de colesterol HDL disminuye en 0,023 mmol/l. En general, la pendiente puede utilizarse para calcular el efecto asociado a incrementos de cualquier magnitud c en la variable explicativa,

$$\hat{y}(x + c) - \hat{y}(x) = b_0 + b_1(x + c) - (b_0 + b_1 x) = cb_1.$$

Así, por ejemplo, incrementos de una desviación típica $c = 3{,}50$ kg/m^2 en el índice de masa corporal se asocian con una disminución media en el colesterol HDL de $cb_1 = 3{,}50(-0{,}023) = -0{,}081$ mmol/l. Notar que, como consecuencia de la hipótesis de linealidad, esta disminución se asume constante a lo largo de todo el rango observado del índice de masa corporal; esto es, el modelo de regresión lineal estima una misma reducción de 0,081 mmol/l en el colesterol HDL entre 25 y 28,5 kg/m^2 del índice de masa corporal que entre 28,5 y 32 kg/m^2.

La recta de regresión estimada del colesterol HDL sobre el índice de masa corporal es

$$\hat{y} = 1{,}69 - 0{,}023x,$$

que se muestra en la Figura 10.7. Esta recta de regresión puede utilizarse para estimar o predecir el valor esperado del colesterol HDL en función del índice de masa corporal. Por ejemplo, para un índice de masa corporal de 25 kg/m^2, el modelo estima un nivel medio de colesterol HDL de $\hat{y}(25) = 1{,}69 - 0{,}023 \cdot 25 = 1{,}11$ mmol/l. Por supuesto, los valores observados del colesterol HDL difieren de los valores medios predichos por la recta de regresión. La varianza residual del colesterol HDL respecto a la recta de regresión es

$$s^2 = \frac{\text{SSE}}{531} = \frac{1}{531}\sum_{i=1}^{533}\{y_i - (1{,}69 - 0{,}023x_i)\}^2 = \frac{42{,}63}{531} = 0{,}080.$$

Notar, por último, que debido a la hipótesis de homogeneidad de la varianza, la desviación típica residual del colesterol HDL $s = \sqrt{0{,}080} = 0{,}283$ mmol/l se asume constante alrededor de cualquier punto de la recta de regresión.

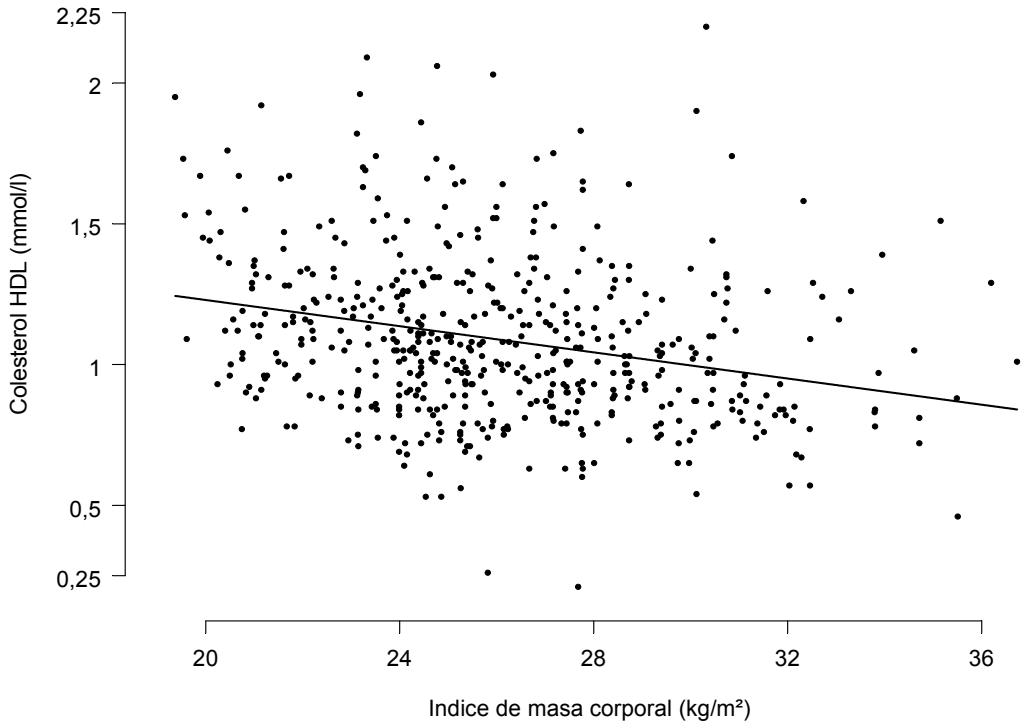


Figura 10.7 **Recta de regresión del colesterol HDL sobre el índice de masa corporal en el grupo control del estudio EURAMIC.**

10.3.2 Contraste del modelo de regresión lineal simple

En general, el contraste de regresión lineal permite evaluar si el modelo en su conjunto explica una parte significativa de la variabilidad de la variable respuesta. En el caso particular de la regresión lineal simple, la hipótesis nula del contraste es simplemente que la pendiente β_1 de la recta de regresión subyacente es 0, ya que en tal caso la variable respuesta no se relacionará linealmente con la única variable explicativa y, en consecuencia, el modelo lineal no aportará explicación alguna sobre la variabilidad de la variable respuesta. Es importante resaltar que este contraste de regresión asume linealidad y, por tanto, no debe interpretarse como un test de bondad del ajuste, en el sentido de que no facilita ninguna información sobre la idoneidad del modelo lineal para describir la relación subyacente entre las variables explicativa y respuesta.

La realización del contraste de regresión se basa en el análisis de la varianza de la variable respuesta. Una vez estimada la recta de regresión, la desviación de cada valor observado y_i respecto a la media muestral $\bar{y}$ puede separarse en dos componentes: el error o desviación del valor observado y_i respecto a su valor estimado por la recta de regresión $\hat{y}_i = b_0 + b_1 x_i$, y la distancia entre dicho valor estimado $\hat{y}_i$ y la media muestral $\bar{y}$; esto es,

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + y_i - \hat{y}_i .$$

Elevando al cuadrado estas desviaciones y sumando sobre todas las observaciones, se tiene que la **suma de cuadrados total** es

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n} (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

$$= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \text{SSR} + \text{SSE},$$

ya que ambas componentes están incorrelacionadas

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = b_1\sum_{i=1}^{n}(x_i - \bar{x})e_i = b_1\sum_{i=1}^{n}x_i e_i - b_1\bar{x}\sum_{i=1}^{n}e_i = 0$$

según las ecuaciones de regresión derivadas del método de mínimos cuadrados. Así, la suma de cuadrados total SST se descompone en dos términos independientes: la **suma de cuadrados de la regresión** SSR, que representa la variabilidad de la variable respuesta explicada por la única variable independiente del modelo de regresión, y la **suma de cuadrados del error** SSE, que corresponde a la variabilidad residual de la variable respuesta que queda sin explicar. Conviene recordar que la recta de regresión estimada por el procedimiento de mínimos cuadrados minimiza la suma de cuadrados del error, maximizando entonces la capacidad predictiva o explicativa del modelo de regresión. La Figura 10.8 ilustra gráficamente esta descomposición.

La descomposición de la variabilidad de la variable respuesta suele representarse mediante la denominada **tabla del análisis de la varianza** (Tabla 10.2). En primer lugar, esta tabla presenta las sumas de cuadrados junto con sus correspondientes grados de libertad. La suma de cuadrados de la regresión contiene únicamente 1 grado de libertad ya que, una vez conocida la media muestral $\bar{y}$, los valores estimados por la recta de regresión $\hat{y}_i = \bar{y} + b_1(x_i - \bar{x})$ quedan completamente determinados por su pendiente; mientras que, como se vio en el apartado anterior, la suma de cuadrados del error tiene $n - 2$ grados de libertad. A continuación, los términos de la varianza se obtienen de dividir las sumas de cuadrados por sus grados de libertad. Finalmente, la razón de varianzas se define como el cociente entre la varianza explicada por la regresión y la varianza residual, que constituye el estadístico del contraste de regresión.

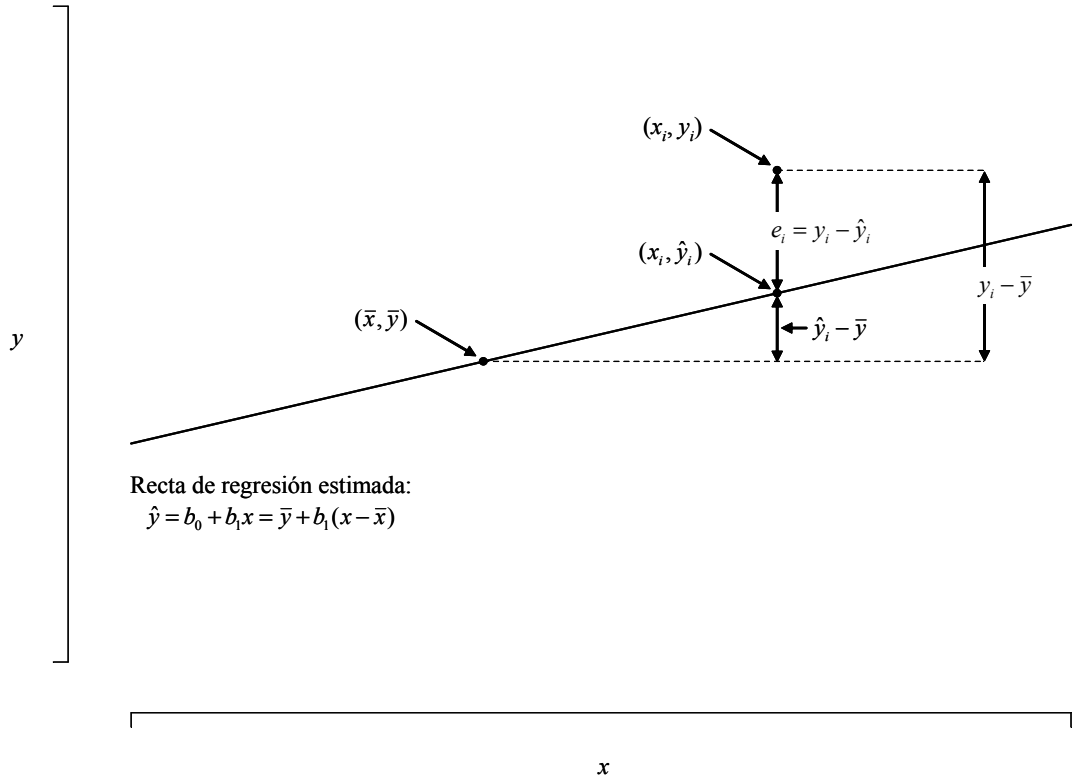


Figura 10.8 **Descomposición de la variabilidad de la variable respuesta en la parte explicada y no explicada por la regresión.**

Tabla 10.2 Tabla genérica del análisis de la varianza en regresión lineal simple.*

	Suma de cuadrados	Grados de libertad	Varianza	Razón de varianzas
Regresión	$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1	SSR	$F = \frac{\text{SSR}}{s^2}$
Error	$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$	$s^2 = \frac{\text{SSE}}{n-2}$	
Total	$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		

* Coeficiente de determinación $R^2 = \text{SSR/SST}$.

Para realizar el contraste de regresión, es preciso conocer la distribución de la razón de varianzas bajo la hipótesis nula H_0: $\beta_1 = 0$. Por un lado, se tiene que

$$\frac{\text{SSR}}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \frac{b_1^2}{\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{b_1^2(n-1)s_x^2}{\sigma^2} = \frac{b_1^2}{\text{var}(b_1)},$$

donde $\text{var}(b_1) = \sigma^2/\{(n-1)s_x^2\}$ es la varianza de la pendiente estimada. Como se comprobará en el siguiente apartado, si se cumplen las asunciones de la regresión lineal simple, la pendiente estimada b_1 seguirá una distribución normal con media β_1 y varianza $\text{var}(b_1)$. Así, bajo la hipótesis nula H_0: $\beta_1 = 0$, el cociente SSR/σ^2 es el cuadrado de una distribución normal estandarizada, que corresponde por definición a una distribución chi-cuadrado con 1 grado de libertad. Por otra parte, basta con que se cumplan las asunciones subyacentes al modelo lineal para que la varianza residual s^2 sea un estimador insesgado de σ^2 y el cociente

$$\frac{(n-2)s^2}{\sigma^2}$$

siga una distribución chi-cuadrado con $n-2$ grados de libertad. Combinando ambos resultados, se tiene que bajo la hipótesis nula H_0: $\beta_1 = 0$ la razón entre las varianzas explicada y residual

$$F = \frac{\text{SSR}}{s^2} = \frac{\text{SSR}/\sigma^2}{s^2/\sigma^2} \sim \frac{\chi_1^2}{\chi_{n-2}^2/(n-2)}$$

se distribuye como el cociente de dos chi-cuadrado independientes divididas por sus respectivos grados de libertad, que es una distribución F de Fisher con 1 grado de libertad en el numerador y $n-2$ grados de libertad en el denominador. El valor P del contraste de regresión de la hipótesis nula H_0: $\beta_1 = 0$ frente a la hipótesis alternativa bilateral H_1: $\beta_1 \neq 0$ se calcula entonces como la probabilidad a la derecha del estadístico F bajo la distribución $F_{1,n-2}$.

La tabla del análisis de la varianza suele ir acompañada del **coeficiente de determinación** R^2, que se define como la proporción de la variabilidad de la variable respuesta que se explica por el modelo de regresión,

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = b_1^2\frac{s_x^2}{s_y^2} = r^2.$$

En el caso de la regresión lineal simple, el coeficiente de determinación R^2 coincide con el cuadrado del coeficiente de correlación r de Pearson entre las variables explicativa y respuesta.

Ejemplo 10.8 La Tabla 10.3 presenta el análisis de la varianza de la regresión lineal del colesterol HDL sobre el índice de masa corporal en 533 controles del estudio EURAMIC. La suma de cuadrados de las desviaciones de los valores observados del colesterol HDL respecto a la media muestral $\bar{y} = 1{,}09$ mmol/l es

$$\text{SST} = \sum_{i=1}^{533} (y_i - 1{,}09)^2 = 46{,}15,$$

que se descompone en la suma de cuadrados de las desviaciones del colesterol HDL respecto a la recta de regresión $\hat{y}_i = 1{,}69 - 0{,}023x_i$

$$\text{SSE} = \sum_{i=1}^{533} \{y_i - (1{,}69 - 0{,}023x_i)\}^2 = 42{,}63$$

y la suma de cuadrados de las distancias entre los valores estimados por la recta de regresión y la media muestral

$$\text{SSR} = \sum_{i=1}^{533} (1{,}69 - 0{,}023x_i - 1{,}09)^2 = 3{,}53.$$

Así, la proporción de la variabilidad del colesterol HDL que se explica únicamente con el índice de masa corporal viene dada por el coeficiente de determinación

$$R^2 = 3{,}53/46{,}15 = 0{,}076,$$

que coincide con el cuadrado del coeficiente de correlación muestral entre el índice de masa corporal y el colesterol HDL $r^2 = (-0{,}276)^2 = 0{,}076$. Para determinar si esta variabilidad explicada por el índice de masa corporal es una parte significativa de la variabilidad total del colesterol HDL, se realiza el contraste de regresión de la hipótesis nula H_0: $\beta_1 = 0$ mediante la razón entre las varianzas explicada SSR = 3,53 y residual $s^2 = 42{,}63/531 = 0{,}080$,

$$F = 3{,}53/0{,}080 = 43{,}93.$$

Bajo la hipótesis nula, este estadístico sigue una distribución F de Fisher con 1 grado de libertad en el numerador y 531 grados de libertad en el denominador, luego el valor P bilateral del contraste es $P(F_{1,531} \geq 43{,}93) < 0{,}001$. En conclusión, las diferencias en el índice de masa corporal explican el 7,6% de la variabilidad del colesterol HDL en la población de referencia del estudio EURAMIC ($R^2 = 0{,}076$, $P < 0{,}001$).

Tabla 10.3 Tabla del análisis de la varianza de la regresión lineal del colesterol HDL sobre el índice de masa corporal en el grupo control del estudio EURAMIC.*

	Suma de cuadrados	Grados de libertad	Varianza	Razón de varianzas
Regresión	3,53	1	3,53	43,93
Error	42,63	531	0,080	
Total	46,15	532		

* Coeficiente de determinación $R^2 = 3{,}53/46{,}15 = 0{,}076$.

10.3.3 Inferencia sobre los parámetros de la recta de regresión

En el Apartado 10.3.1 se obtuvieron los estimadores b_0 y b_1 de la constante y la pendiente de la recta de regresión utilizando el método de mínimos cuadrados. A partir de las distribuciones muestrales de b_0 y b_1, se derivan a continuación los intervalos de confianza y tests de hipótesis para los parámetros subyacentes β_0 y β_1 del modelo de regresión lineal simple.

El estimador de mínimos cuadrados de la pendiente de la recta de regresión puede reescribirse como una combinación lineal de los valores de la variable respuesta

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} c_i y_i \,,$$

donde los coeficientes $c_i = (x_i - \bar{x})/\{(n-1)s_x^2\}$ dependen únicamente de los valores de la variable explicativa que se asumen constantes. Bajo las asunciones de linealidad y homogeneidad de la varianza, el valor esperado de b_1 es

$$E(b_1) = \sum_{i=1}^{n} c_i E(y_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i = \beta_1$$

y, como las observaciones y_i son independientes (véase Apartado 3.4), su varianza es

$$\text{var}(b_1) = \sum_{i=1}^{n} c_i^2 \ \text{var}(y_i) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{(n-1)s_x^2} \,.$$

Es decir, b_1 es un estimador insesgado de β_1 que será tanto más preciso cuanto menor sea la varianza de la variable respuesta alrededor de la recta de regresión y mayores sean el tamaño muestral y la dispersión de la variable explicativa. Además, si el tamaño muestral n es suficientemente grande, puede aplicarse una generalización del teorema central del límite (ver su versión más simple en el Apartado 4.3.3) para demostrar que b_1 se distribuye de forma aproximadamente normal con la media y varianza descritas anteriormente,

$$\frac{b_1 - \beta_1}{\dfrac{\sigma}{s_x\sqrt{n-1}}} \overset{\sim}{\to} N(0,\ 1) \,.$$

Para hacer uso de este resultado, el parámetro desconocido σ ha de sustituirse por la desviación típica residual s, que conlleva un error adicional de muestreo. La distribución resultante de b_1 será entonces más dispersa que la normal, siguiendo aproximadamente una distribución t de Student con los $n-2$ grados de libertad correspondientes a la estimación de la varianza residual,

$$\frac{b_1 - \beta_1}{\dfrac{s}{s_x\sqrt{n-1}}} \overset{\sim}{\to} t_{n-2} \,.$$

Cabe destacar que este resultado se ha derivado con independencia de la asunción de normalidad y, en consecuencia, es válido para cualquier distribución subyacente de la variable respuesta, siempre que el tamaño muestral sea suficientemente grande.

A partir de la distribución muestral de b_1, el intervalo de confianza al $100(1-\alpha)\%$ para la pendiente subyacente β_1 de la recta de regresión viene dado por

$$b_1 \pm t_{n-2,1-\alpha/2}\frac{s}{s_x\sqrt{n-1}}.$$

De igual forma, el contraste bilateral de la hipótesis de ausencia de asociación lineal entre las variables explicativa y respuesta H_0: $\beta_1 = 0$ se realiza mediante el estadístico

$$t = \frac{b_1}{\frac{s}{s_x\sqrt{n-1}}},$$

que se distribuye aproximadamente como una t de Student con $n-2$ grados de libertad si la hipótesis nula es cierta. Este test es equivalente al contraste de regresión lineal simple presentado en el apartado anterior. De hecho, el estadístico F del contraste de regresión es igual al cuadrado del estadístico t de este contraste,

$$F = \frac{\text{SSR}}{s^2} = \frac{b_1^2(n-1)s_x^2}{s^2} = t^2,$$

de tal forma que ambos procedimientos facilitan siempre los mismos valores P (la distribución F de Fisher con 1 grado de libertad en el numerador y $n-2$ grados de libertad en el denominador es, por definición, el cuadrado de la distribución t de Student con $n-2$ grados de libertad).

Para completar la exposición, se presentan el intervalo de confianza y el test de hipótesis para la constante de la recta de regresión, aunque estas inferencias suelen tener escasa importancia porque la relación en $x = 0$ carece de sentido en la mayoría de las aplicaciones. El estimador mínimo-cuadrático de la constante $b_0 = \bar{y} - b_1\bar{x}$ es una combinación lineal de dos estimadores independientes $\bar{y}$ y b_1 que tienden a distribuirse de forma normal conforme aumenta el tamaño muestral, de lo cual se deduce que la distribución muestral de b_0 también será aproximadamente normal con media

$$E(b_0) = E(\bar{y}) - E(b_1)\bar{x} = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$$

y varianza

$$\text{var}(b_0) = \text{var}(\bar{y}) + \text{var}(b_1)\bar{x}^2 = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right).$$

Reemplazando el parámetro σ^2 por su estimación s^2, el intervalo de confianza al $100(1-\alpha)\%$ para la constante poblacional β_0 es

$$b_0 \pm t_{n-2,1-\alpha/2}s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}$$

y el estadístico del contraste de la hipótesis nula H_0: $\beta_0 = 0$ es

$$t = \frac{b_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}},$$

que bajo H_0 seguirá aproximadamente una distribución t de Student con $n-2$ grados de libertad.

Ejemplo 10.9 Las estimaciones puntuales obtenidas en el Ejemplo 10.7 para los parámetros de la regresión del colesterol HDL sobre el índice de masa corporal fueron $b_0 = 1{,}69$, $b_1 = -0{,}023$ y $s = 0{,}283$. El error estándar de la estimación de la constante es

$$SE(b_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}} = 0{,}283\sqrt{\frac{1}{533} + \frac{26{,}0^2}{532 \cdot 3{,}50^2}} = 0{,}092$$

y de la pendiente

$$SE(b_1) = \frac{s}{s_x\sqrt{n-1}} = \frac{0{,}283}{3{,}50\sqrt{532}} = 0{,}0035.$$

Los ICs al 95% para la constante y la pendiente de la recta de regresión poblacional son entonces

$$b_0 \pm t_{531;0{,}975}\, SE(b_0) = 1{,}69 \pm 1{,}96 \cdot 0{,}092 = (1{,}51;\ 1{,}87)$$

y

$$b_1 \pm t_{531;0{,}975}\, SE(b_1) = -0{,}023 \pm 1{,}96 \cdot 0{,}0035 = (-0{,}030; -0{,}016).$$

Del intervalo para la pendiente puede concluirse con una confianza del 95% que el nivel medio de colesterol HDL en la población de referencia del estudio EURAMIC disminuye entre 0,016 y 0,030 mmol/l por cada incremento de 1 kg/m^2 en el índice de masa corporal. En general, el intervalo de confianza para el efecto subyacente $c\beta_1$ asociado a cualquier incremento c en la variable explicativa se obtiene multiplicando los límites del intervalo para β_1 por dicho incremento,

$$cb_1 \pm t_{n-2,1-\alpha/2} SE(cb_1) = c\{b_1 \pm t_{n-2,1-\alpha/2} SE(b_1)\}.$$

Así, por ejemplo, con un nivel de confianza del 95%, los incrementos de una desviación típica $c = 3{,}50$ kg/m^2 en el índice de masa corporal se asocian con una disminución media poblacional en el colesterol HDL de entre $3{,}50 \cdot 0{,}016 = 0{,}057$ y $3{,}50 \cdot 0{,}030 = 0{,}105$ mmol/l. Por supuesto, esta disminución es estadísticamente significativa ya que el contraste de la hipótesis nula H_0: $\beta_1 = 0$ mediante el estadístico

$$t = \frac{b_1}{SE(b_1)} = \frac{-0{,}023}{0{,}0035} = -6{,}63$$

resulta en un valor P bilateral $2P(t_{531} \leq -6{,}63) \approx 2\Phi(-6{,}63) < 0{,}001$. Notar que este test arroja el mismo valor P que el contraste de regresión del ejemplo anterior ya que $2P(t_{531} \leq -6{,}63) = P(t^2_{531} \geq 6{,}63^2) = P(F_{1,531} \geq 43{,}93)$.

10.3.4 Bandas de confianza y predicción para la recta de regresión

Además de realizar inferencias sobre los parámetros β_0 y β_1, es a menudo interesante calcular intervalos de confianza para la propia recta de regresión $\beta_0 + \beta_1 x$. Más concretamente, dado un determinado valor x_0 de la variable explicativa, se pretende obtener un intervalo de confianza para el valor esperado $\beta_0 + \beta_1 x_0$ de la variable respuesta. El estimador puntual de este valor esperado es $\hat{y}_0 = b_0 + b_1 x_0 = \bar{y} + b_1(x_0 - \bar{x})$ que, siguiendo un razonamiento análogo al del apartado anterior, presenta una distribución aproximadamente normal en muestras suficientemente grandes, con media

$$E(\hat{y}_0) = E(\bar{y}) + E(b_1)(x_0 - \bar{x}) = \beta_0 + \beta_1 \bar{x} + \beta_1(x_0 - \bar{x}) = \beta_0 + \beta_1 x_0$$

y varianza

$$\text{var}(\hat{y}_0) = \text{var}(\bar{y}) + \text{var}(b_1)(x_0 - \bar{x})^2 = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}\right).$$

Por tanto, utilizando la distribución t_{n-2} resultante de sustituir σ^2 por la estimación s^2, se tiene que el intervalo de confianza al $100(1-\alpha)\%$ para el valor esperado $\beta_0 + \beta_1 x_0$ es

$$b_0 + b_1 x_0 \pm t_{n-2,1-\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}\ .$$

La **banda de confianza** para la recta de regresión no es más que la representación gráfica de estos intervalos a lo largo de todo el rango observado de la variable explicativa. Esta banda de confianza está delimitada por las ramas de una hipérbola y su amplitud es mínima en $x_0 = \bar{x}$, aumentando a medida que x_0 se aleja de su media muestral $\bar{x}$, lo que confirma la intuición de que el valor esperado de la variable respuesta puede estimarse con mayor precisión en valores centrados que en valores extremos de la variable explicativa.

Ejemplo 10.10 Para cada valor fijo x_0 del índice de masa corporal, el modelo de regresión lineal estima un IC al 95% para el valor esperado del colesterol HDL de

$$1{,}69 - 0{,}023x_0 \pm 1{,}96 \cdot 0{,}283\sqrt{\frac{1}{533} + \frac{(x_0 - 26{,}0)^2}{532 \cdot 3{,}50^2}}\ .$$

El área en gris oscuro de la Figura 10.9 representa la banda de confianza al 95% para toda la recta de regresión del colesterol HDL sobre el índice de masa corporal, que se obtiene de calcular estos intervalos en sucesivos valores dentro del rango observado del índice de masa corporal. Los límites de esta banda de confianza tienen forma de hipérbola y su amplitud aumenta gradualmente conforme x_0 se aleja de la media $\bar{x} = 26{,}0$ kg/m^2 del índice de masa corporal. Así, por ejemplo, el IC al 95% para el valor medio del colesterol HDL entre los sujetos con un índice de masa corporal de 25 kg/m^2,

$$1{,}69 - 0{,}023 \cdot 25 \pm 1{,}96 \cdot 0{,}013 = (1{,}09;\ 1{,}14),$$

es sensiblemente más preciso que entre aquellos con un índice de masa corporal de 32 kg/m^2,

$$1{,}69 - 0{,}023 \cdot 32 \pm 1{,}96 \cdot 0{,}024 = (0{,}90;\ 1{,}00).$$

La recta de regresión puede utilizarse no sólo para estimar la media poblacional de la variable respuesta entre los sujetos con un determinado valor x_0 de la variable explicativa, sino también para predecir la respuesta individual y_0 de un nuevo sujeto dado su valor x_0. Según la estructura del modelo de regresión lineal, el valor subyacente de la variable respuesta para un determinado sujeto con $x = x_0$ viene dado por $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$, cuyo estimador insesgado es de nuevo $\hat{y}_0 = b_0 + b_1 x_0$ ya que

$$E(y_0 - \hat{y}_0) = \beta_0 + \beta_1 x_0 + E(\varepsilon_0) - \beta_0 - \beta_1 x_0 = E(\varepsilon_0) = 0.$$

Asimismo, como el valor estimado $\hat{y}_0$ por la recta de regresión en x_0 es independiente de la nueva observación y_0, se sigue que

$$\text{var}(y_0 - \hat{y}_0) = \text{var}(\varepsilon_0) + \text{var}(\hat{y}_0) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}\right);$$

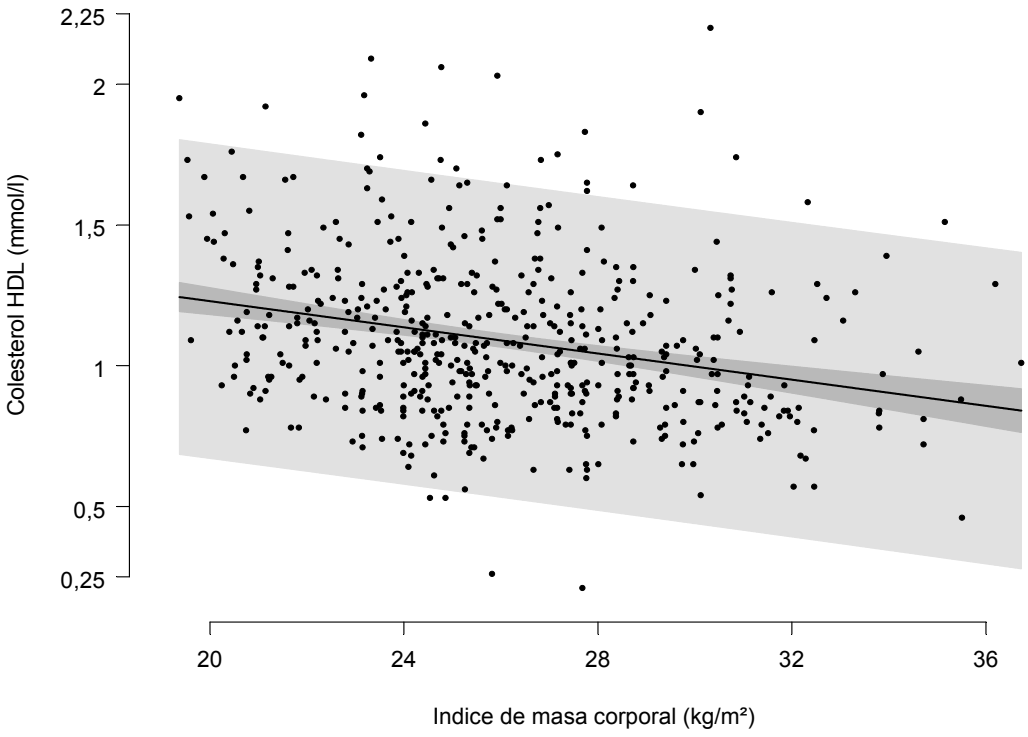


Figura 10.9 **Bandas de confianza (área en gris oscuro) y predicción (área en gris claro) al 95% para la recta de regresión del colesterol HDL sobre el índice de masa corporal en el grupo control del estudio EURAMIC.**

es decir, la predicción de una nueva observación a partir de la recta de regresión estimada está sujeta a dos fuentes de error: la varianza inherente de cada respuesta individual respecto a la recta de regresión subyacente y el error en la estimación de dicha recta. Además, si el término de error ε_0 se distribuye de forma normal (asunción de normalidad), la diferencia $y_0 - \hat{y}_0$ también seguirá una distribución normal, de tal forma que el intervalo de predicción al $100(1-\alpha)\%$ para una nueva observación individual y_0 es

$$b_0 + b_1 x_0 \pm t_{n-2,1-\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}}\,.$$

La **banda de predicción** viene entonces determinada por estos intervalos de predicción en los distintos valores observados x_0 de la variable explicativa. En general, la banda de predicción será substancialmente más amplia que la banda de confianza, particularmente cuando el tamaño muestral es grande, lo que refleja el hecho de que existe mucha más incertidumbre en la predicción de la respuesta individual de un único sujeto que en la estimación del valor medio de la variable respuesta para todos los sujetos con un mismo valor de la variable explicativa.

Cabe destacar, por último, que los intervalos de confianza para el valor esperado de la variable respuesta se basan únicamente en las asunciones de linealidad y homogeneidad de la varianza, mientras que los intervalos de predicción para una nueva observación requieren además de la hipótesis de normalidad, siendo estos últimos incorrectos si la distribución subyacente de la variable respuesta no es normal.

Ejemplo 10.11 A partir del modelo de regresión lineal del colesterol HDL sobre el índice de masa corporal se tiene que el intervalo de predicción al 95% para el nivel de colesterol HDL de un sujeto con un índice de masa corporal x_0 es

$$1{,}69 - 0{,}023x_0 \pm 1{,}96 \cdot 0{,}283\sqrt{1 + \frac{1}{533} + \frac{(x_0 - 26{,}0)^2}{532 \cdot 3{,}50^2}}\,.$$

El cálculo de estos intervalos en distintos valores x_0 del índice de masa corporal da lugar a la banda de predicción en gris claro de la Figura 10.9. Al igual que la banda de confianza, la banda de predicción está centrada alrededor de la recta de regresión estimada, pero su amplitud es notablemente mayor al incorporar la variabilidad de cada respuesta individual respecto a su valor esperado. Por ejemplo, el intervalo de predicción al 95% para el nivel de colesterol HDL de un sujeto con 25 kg/m^2 de índice de masa corporal viene dado por

$$1{,}69 - 0{,}023 \cdot 25 \ \pm 1{,}96 \cdot 0{,}284 = (0{,}56;\ 1{,}67),$$

que es mucho más impreciso que el intervalo de confianza calculado en el ejemplo anterior para el valor medio del colesterol HDL en todos los sujetos con dicho valor del índice de masa corporal (IC al 95% 1,09-1,14 mmol/l).

10.3.5 Evaluación de las asunciones del modelo de regresión lineal simple

Los procedimientos de estimación e inferencia derivados en los apartados anteriores se basan en las asunciones de linealidad, homogeneidad de la varianza y normalidad. La violación de estas asunciones puede dar lugar a conclusiones erróneas del modelo lineal, siendo así necesario evaluar su idoneidad en cada aplicación práctica. Aunque existen diversos tests para contrastar estadísticamente cada una de las hipótesis del modelo lineal (véase referencias al final del tema), en este apartado se presentan algunas técnicas diagnósticas basadas en el análisis gráfico de los residuos, proponiéndose asimismo extensiones básicas del modelo y transformaciones de los datos para acomodar posibles desviaciones de estas asunciones. En particular, se presta especial atención a las hipótesis de linealidad y homogeneidad de la varianza, ya que las principales inferencias relativas a la pendiente de la recta de regresión y al valor esperado de la variable respuesta son aproximadamente válidas en muestras moderadamente grandes aunque la distribución subyacente de la variable respuesta no sea normal.

El gráfico más simple para evaluar el grado de cumplimiento de las asunciones de la regresión lineal simple es el diagrama de dispersión entre las variables explicativa y respuesta, junto con la recta de regresión estimada. Si se cumplen las hipótesis de linealidad y homogeneidad de la varianza, los puntos del diagrama de dispersión han de distribuirse aleatoriamente alrededor de la recta de regresión sin evidencia de relaciones curvilíneas y con similar dispersión a lo largo de toda la recta. Tal parece ser el caso del diagrama de dispersión entre el índice de masa corporal y el colesterol HDL de la Figura 10.7, donde no se aprecian desviaciones obvias de estas asunciones. En la Figura 10.2(*d*), sin embargo, se muestra un claro ejemplo de violación de la asunción de linealidad, ya que la relación subyacente es visiblemente cuadrática. No obstante, el gráfico más utilizado para chequear las asunciones de la regresión lineal es el diagrama de dispersión de los residuos $e_i = y_i - \hat{y}_i$ frente a los valores predichos $\hat{y}_i = b_0 + b_1 x_i$ por la recta de regresión. Este gráfico es equivalente al diagrama de dispersión entre x_i e y_i en regresión lineal simple, pero tiene la ventaja de ser directamente generalizable a la presencia de más de una variable explicativa en regresión lineal múltiple.

Antes de proceder al análisis gráfico de los residuos, es importante describir algunas de sus propiedades. Bajo las hipótesis de linealidad y homogeneidad de la varianza, los residuos $e_i = y_i - \hat{y}_i$ tienen un valor esperado

$$E(e_i) = E(y_i) - E(\hat{y}_i) = 0$$

y una varianza

$$\text{var}(e_i) = \text{var}(y_i) + \text{var}(\hat{y}_i) - 2\text{cov}(y_i, \hat{y}_i) = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}\right).$$

Así, aun cuando se cumpla la asunción de homogeneidad de la varianza, los residuos e_i tendrán diferente varianza alrededor de los distintos puntos de la recta de regresión estimada. Más concretamente, los residuos tenderán a ser mayores en valores centrados que en valores extremos de la variable explicativa. Esto es debido a que los puntos (x_i, y_i) con x_i muy distante de $\bar{x}$ tienen mucha influencia en la estimación de la pendiente, de tal forma que la recta de regresión resultante tenderá a aproximarse a estos puntos que presentarán entonces pequeños residuos e_i. Por ello, y con objeto de que los residuos sean comparables a distintos niveles de la variable explicativa, es preferible realizar el diagnóstico del modelo mediante los **residuos estandarizados**

$$r_i = \frac{e_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}}} = \frac{e_i}{s\sqrt{1 - h_i}},$$

que se obtienen de dividir los residuos e_i por una estimación de su desviación típica. El término h_i se conoce como el **leverage** de una observación y es una medida estandarizada de la distancia entre cada valor x_i de la variable explicativa y su media $\bar{x}$ que se tratará en el apartado siguiente. No obstante, si el tamaño muestral es grande y no hay valores muy extremos de la variable explicativa (observaciones con alto leverage), ambos residuos e_i y r_i se comportan de forma análoga.

En determinados casos el gráfico de los residuos estandarizados r_i frente a los valores predichos $\hat{y}_i$ no permite apreciar claramente las posibles desviaciones de las asunciones de linealidad y homogeneidad de la varianza. Para obtener una representación más clara en tales circunstancias, es aconsejable dividir los n residuos r_i en K grupos de tamaño n_k ordenados por valores crecientes de $\hat{y}_i$ (por ejemplo, deciles) y calcular la media

$$\bar{r}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} r_i$$

y la varianza

$$s_k^2 = \frac{1}{n_k}\sum_{i=1}^{n_k} r_i^2$$

de los residuos en cada uno de los grupos. La presencia de curvatura en el gráfico de los residuos medios $\bar{r}_k$ frente a los valores predichos medios $\bar{\hat{y}}_k$ en los distintos grupos indicará falta de linealidad en la relación, mientras que la existencia de tendencia en el gráfico de las desviaciones típicas residuales s_k frente a los valores predichos medios $\bar{\hat{y}}_k$ de cada grupo aportará evidencia de heterogeneidad en la varianza.

Ejemplo 10.12 En la Figura 10.10(*a*) se representa el gráfico de los residuos estandarizados r_i frente a los valores predichos $\hat{y}_i$ de la regresión lineal del colesterol HDL sobre el índice de masa corporal. Este gráfico, al igual que el diagrama de dispersión entre el índice de masa corporal y el colesterol HDL de la Figura 10.7, parece compatible con las asunciones de linealidad y homogeneidad de la varianza. Para realizar una evaluación más detallada, en la Tabla 10.4 se presentan las medias $\bar{r}_k$ y desviaciones típicas s_k de los

residuos estandarizados por deciles de los valores predichos. La Figura 10.10(*b*) de los residuos medios $\bar{r}_k$ frente a los valores predichos medios $\bar{\hat{y}}_k$ de cada decil muestra indicios de una posible relación cuadrática entre el índice de masa corporal y el colesterol HDL, ya que los residuos del modelo lineal tienden a ser positivos para valores predichos altos y bajos del colesterol HDL y negativos para valores predichos intermedios. Por otra parte, en la Figura 10.10(*c*) no se aprecian desviaciones de la asunción de homogeneidad de la varianza, dado que las desviaciones típicas residuales s_k son similares en los distintos deciles de los valores predichos.

La alternativa más simple para acomodar una relación cuadrática entre el índice de masa corporal y el colesterol HDL es extender el modelo lineal a un modelo polinomial de segundo orden $E(Y|x) = \beta_0 + \beta_1 x + \beta_2 x^2$, que incluye el término cuadrático x^2 además del término lineal x del índice de masa corporal. La relación resultante entre ambas variables ya no será una línea recta sino una parábola, cuya curvatura vendrá determinada por el coeficiente β_2 asociado al término cuadrático. El ajuste de los modelos polinomiales se tratará en el Tema 11 ya que estos modelos pueden considerarse como casos particulares de la regresión lineal múltiple cuyas variables explicativas son distintas potencias de una misma variable básica.

Ejemplo 10.13 Los niveles de α-tocoferol y β-caroteno en tejido adiposo presentan distribuciones asimétricas en los 700 controles del estudio EURAMIC, con un marcado sesgo positivo en el caso del β-caroteno (Figura 4.3). La media y la desviación típica del α-tocoferol son $\bar{x} = 146{,}1$ y $s_x = 87{,}6$ µg/g y del β-caroteno $\bar{y} = 0{,}37$ y $s_y = 0{,}40$ µg/g, y el coeficiente de correlación de Pearson entre ambas variables es $r = 0{,}45$. A partir de estos datos se estima que la recta de regresión del β-caroteno sobre el α-tocoferol es

$$\hat{y} = 0{,}072 + 0{,}0021x,$$

con una desviación típica residual de los niveles de β-caroteno alrededor de dicha recta de $s = 0{,}36$ µg/g. El error estándar de la constante es $SE(b_0) = 0{,}026$ y de la pendiente $SE(b_1) = 0{,}00015$. Así, se tiene que incrementos de una desviación típica (87,6 µg/g) en el α-tocoferol se asocian con un aumento de $87{,}6 \cdot 0{,}0021 = 0{,}18$ µg/g en el nivel medio de β-caroteno, con un IC al 95% comprendido entre $87{,}6(0{,}0021 \pm 1{,}96 \cdot 0{,}00015) = (0{,}15;\ 0{,}21)$.

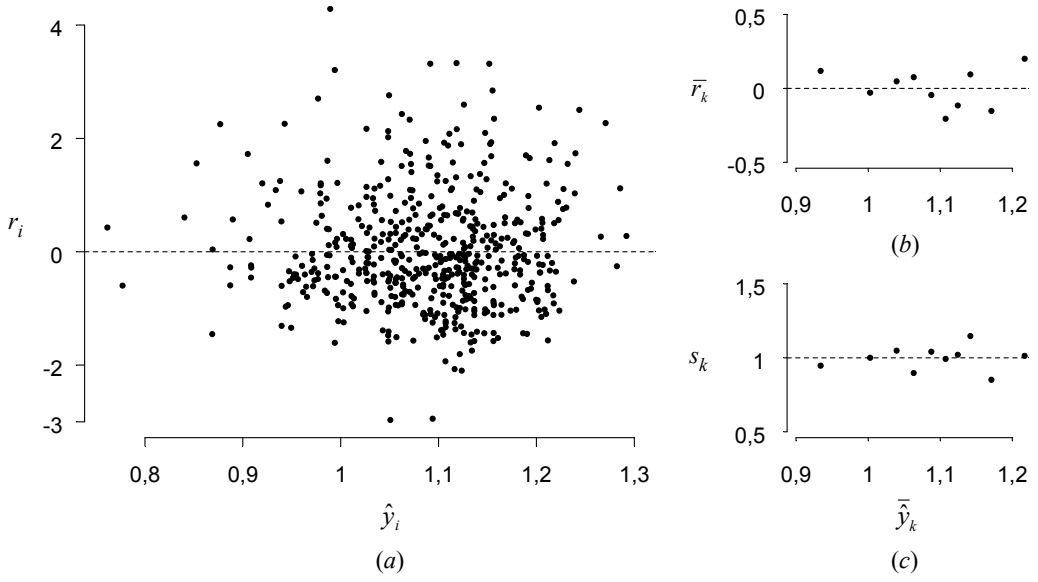


Figura 10.10 **Gráfico de los residuos estandarizados r_i frente a los valores predichos $\hat{y}_i$ (*a*), así como de las medias $\bar{r}_k$ (*b*) y desviaciones típicas s_k (*c*) de los residuos estandarizados por deciles de los valores predichos de la regresión lineal del colesterol HDL sobre el índice de masa corporal en el grupo control del estudio EURAMIC.**

Tabla 10.4 **Media y desviación típica de los residuos estandarizados r_i por deciles de los valores predichos $\hat{y}_i$ de la regresión lineal del colesterol HDL sobre el índice de masa corporal en el grupo control del estudio EURAMIC.**

Valores predichos (mmol/l)		Residuos estandarizados	
Decil (k)	Media ($\bar{\hat{y}}_k$)	Media ($\bar{r}_k$)	Desviación típica (s_k)
< 0,98	0,93	0,12	0,95
0,98-1,03	1,00	−0,03	1,00
1,03-1,05	1,04	0,05	1,05
1,05-1,07	1,06	0,08	0,90
1,07-1,10	1,09	−0,05	1,04
1,10-1,12	1,11	−0,21	0,99
1,12-1,13	1,12	−0,12	1,02
1,13-1,16	1,14	0,09	1,15
1,16-1,19	1,17	−0,15	0,85
≥ 1,19	1,22	0,20	1,01

Una simple inspección del diagrama de dispersión entre los niveles de α-tocoferol y β-caroteno de la Figura 10.11(*a*) evidencia una clara violación de la hipótesis de homogeneidad de la varianza, ya que hay mayor variabilidad de los puntos alrededor de la recta de regresión para valores altos del α-tocoferol que para valores bajos. Esta heterogeneidad se hace aún más evidente en la Figura 10.11(*c*), donde se observa cómo la desviación típica s_k de los residuos estandarizados aumenta linealmente con los deciles de los valores predichos.

Por otro lado, la Figura 10.11(*b*) no muestra una curvatura clara en la relación, pero sí se aprecia una cierta tendencia lineal negativa de los residuos medios $\bar{r}_k$ conforme aumenta el valor predicho. Esto podría deberse a que algunas observaciones con valores extremos de α-tocoferol y β-caroteno tienen excesiva influencia en la estimación de la pendiente, produciendo una sobreestimación de la misma que da lugar a residuos positivos para valores predichos bajos y residuos negativos para valores predichos altos. La identificación de observaciones influyentes se abordará en mayor detalle en el siguiente apartado.

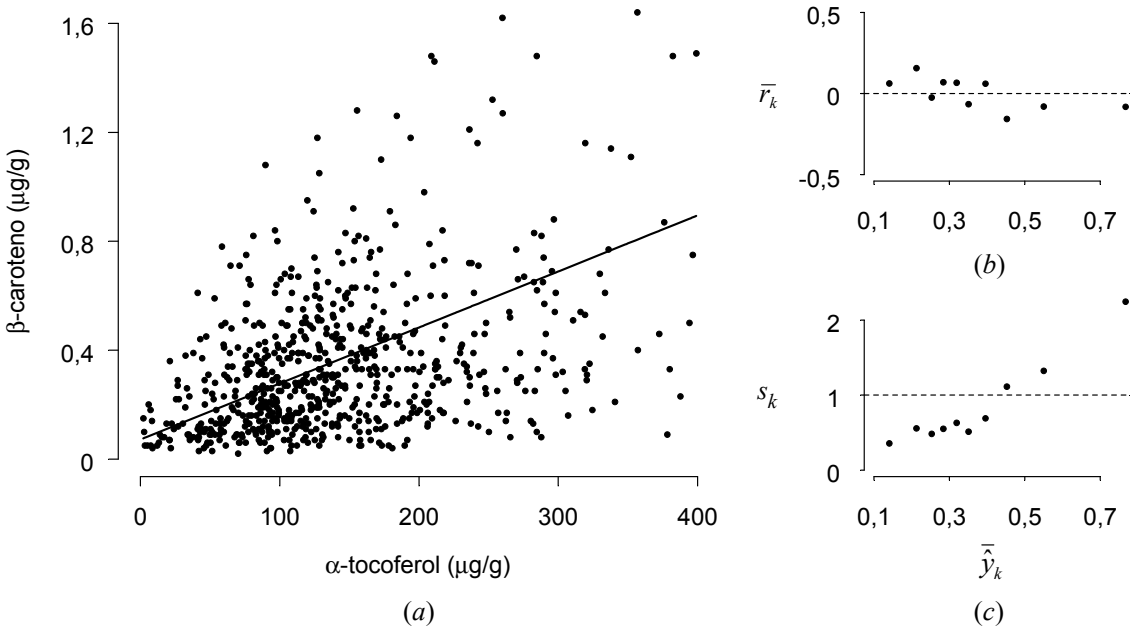


Figura 10.11 **Regresión lineal del β-caroteno sobre el α-tocoferol en el grupo control del estudio EURAMIC (*a*), junto con las medias $\bar{r}_k$ (*b*) y desviaciones típicas s_k (*c*) de los residuos estandarizados por deciles de los valores predichos.**

En presencia de heterogeneidad de la varianza, los estimadores puntuales b_0 y b_1, así como la propia recta de regresión estimada $\hat{y} = b_0 + b_1 x$, continúan siendo insesgados, pero la varianza residual s^2 está sesgada ya que infraestima la variabilidad de la variable respuesta alrededor de unos puntos de la recta de regresión y la sobreestima en otros. En consecuencia, los errores estándar de los estimadores no son correctos y sus correspondientes intervalos de confianza y tests de hipótesis dejan de ser válidos. En general, existen dos procedimientos alternativos para tratar con varianzas heterogéneas. El primer método consiste en realizar una **regresión lineal ponderada**, que es una extensión del modelo lineal ordinario donde cada observación de la variable respuesta recibe un peso inversamente proporcional a su varianza estimada alrededor de la recta de regresión. Así, cuanto más precisa sea una observación, mayor será su peso en la estimación de la recta de regresión. En el ejemplo anterior, la regresión lineal ponderada del β-caroteno sobre el α-tocoferol otorgaría más peso a los puntos con valores bajos del α-tocoferol que a aquellos con valores altos, ya que los primeros presentan menor variabilidad en el nivel de β-caroteno. Las técnicas de regresión lineal ponderada pueden consultarse en los textos específicos de regresión citados en este tema.

El segundo procedimiento para tratar con varianzas heterogéneas es encontrar una transformación de la variable respuesta que estabilice la varianza y ajustar el modelo lineal a esta variable transformada. La selección de la transformación adecuada suele basarse en la relación existente entre la varianza residual y el valor esperado de la variable respuesta. En el caso más frecuente de que la desviación típica residual tienda a aumentar linealmente con el valor predicho (tal como ocurre en la regresión del β-caroteno sobre el α-tocoferol), la heterogeneidad de la varianza se resuelve utilizando la **transformación logarítmica**, dado que el logaritmo de la respuesta tendrá entonces una varianza aproximadamente constante. Esta transformación logarítmica produce el mismo efecto en cualquier base y sólo puede aplicarse a variables respuestas positivas. Además de homogeneizar la varianza, la transformación logarítmica también suele emplearse para normalizar variables respuestas sesgadas positivamente, así como para linealizar relaciones con pendiente monótonamente creciente.

Aun cuando el uso de una respuesta logarítmica esté plenamente justificado en términos estadísticos, los resultados del modelo transformado han de interpretarse en la escala original de la variable respuesta. El modelo en escala logarítmica asume que el valor esperado del logaritmo de la variable respuesta Y cambia linealmente con la variable explicativa X,

$$E(\log Y|x) = \beta_0 + \beta_1 x.$$

Para volver a la escala original, se toma la exponencial en ambos lados de esta igualdad, resultando que la media geométrica de la variable respuesta (definida como la exponencial de la media de los logaritmos; véase Apartado 1.2.3) es una función exponencial de la variable explicativa,

$$E_G(Y|x) = \exp\{E(\log Y|x)\} = \exp(\beta_0 + \beta_1 x).$$

Así, el modelo en la escala original se interpreta en términos de la media geométrica de la variable respuesta, que varía exponencialmente con la variable explicativa. El coeficiente β_1 asociado a la variable explicativa tiene entonces una interpretación distinta de la habitual ya que su exponencial corresponde a la razón de medias geométricas de Y cuando X aumenta una unidad,

$$\frac{E_G(Y \mid x+1)}{E_G(Y \mid x)} = \exp\{\beta_0 + \beta_1(x+1) - (\beta_0 + \beta_1 x)\} = \exp(\beta_1);$$

es decir, $100\{\exp(\beta_1) - 1\}$ representa el cambio porcentual en la media geométrica de Y por cada incremento de una unidad en X. Este cambio relativo se asume constante a lo largo de todo el rango de la variable explicativa.

Ejemplo 10.14 En el análisis de regresión lineal del β-caroteno sobre el α-tocoferol del ejemplo anterior se observó un aumento lineal de la desviación típica residual conforme aumentaba el valor predicho, lo que sugiere la utilización de una transformación logarítmica de la variable respuesta. La Figura 10.12(*a*) muestra la recta de regresión estimada entre el logaritmo del β-caroteno y el α-tocoferol,

$$\log \bar{y}_G = -1{,}91 + 0{,}0040x,$$

donde el error estándar de la constante es $SE(b_0) = 0{,}055$ y de la pendiente $SE(b_1) = 0{,}00032$. Aunque el ajuste se ha realizado en escala logarítmica, el modelo tiene una interpretación directa en términos de la media geométrica de la variable respuesta. La razón de medias geométricas asociada a un aumento de c unidades en la variable explicativa viene dada por

$$\frac{\bar{y}_G(x+c)}{\bar{y}_G(x)} = \exp\{b_0 + b_1(x+c) - (b_0 + b_1x)\} = \exp(cb_1).$$

Así, por ejemplo, por cada incremento de una desviación típica $c = 87{,}6$ µg/g en el nivel de α-tocoferol, la media geométrica de β-caroteno aumenta un $100\{\exp(87{,}6 \cdot 0{,}0040) - 1\} = 100(1{,}42 - 1) = 42\%$. Este incremento porcentual en la media geométrica de β-caroteno permanece constante a través de todo el rango observado del α-tocoferol. Como consecuencia, la tendencia resultante en la escala original del β-caroteno es exponencial, tal como se muestra en la Figura 10.12(*b*).

El IC al 95% para la razón de medias geométricas asociada a un aumento de 87,6 µg/g en el α-tocoferol se calcula multiplicando primero los límites del intervalo para β_1 por dicho incremento y después exponenciando,

$$\begin{aligned}\exp[c\{b_1 \pm t_{698;0,975}\, SE(b_1)\}] &= \exp\{87{,}6(0{,}0040 \pm 1{,}96 \cdot 0{,}00032)\} \\ &= (1{,}34;\ 1{,}50),\end{aligned}$$

de donde se concluye con una confianza del 95% que la media geométrica de β-caroteno aumenta entre un 34 y un 50% por cada incremento de 87,6 µg/g en el nivel de α-tocoferol. Este cambio relativo es muy significativo dado que el contraste bilateral de la hipótesis nula H_0: $\beta_1 = 0$ mediante el estadístico

$$t = \frac{b_1}{SE(b_1)} = \frac{0{,}0040}{0{,}00032} = 12{,}44$$

arroja un valor $P = 2P(t_{698} \geq 12{,}44) \approx 2\{1 - \Phi(12{,}44)\} < 0{,}001$.

Como cabía esperar, la hipótesis de homogeneidad de la varianza se hace mucho más plausible utilizando la escala logarítmica (paneles *a* y *c* de la Figura 10.13). Sin embargo, la curvatura de los residuos de la Figura 10.13(*b*) sugiere que el efecto del α-tocoferol no es lineal en el logaritmo del β-caroteno o, dicho de forma equivalente, la relación subyacente entre el α-tocoferol y el β-caroteno no parece responder fielmente a un modelo exponencial. Así, la transformación logarítmica de la variable respuesta elimina la heterogeneidad de la varianza pero introduce una desviación de la asunción de linealidad. Como veremos más adelante, este problema podría paliarse transformando también la variable explicativa para restaurar la linealidad en la relación. Alternativamente, se podría haber ajustado un modelo de regresión lineal ponderado entre el α-tocoferol y el β-caroteno, que permite trabajar directamente con varianzas heterogéneas sin necesidad de transformar los datos ni modificar la estructura lineal del modelo.

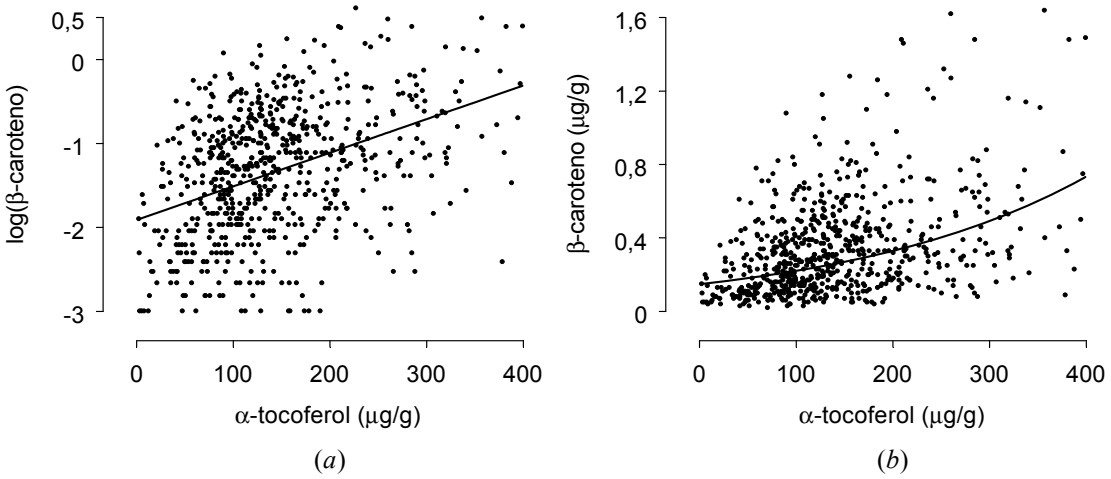


Figura 10.12 **Recta de regresión del logaritmo del β-caroteno sobre el α-tocoferol en el grupo control del estudio EURAMIC (*a*) y tendencia exponencial resultante en la escala original del β-caroteno (*b*).**

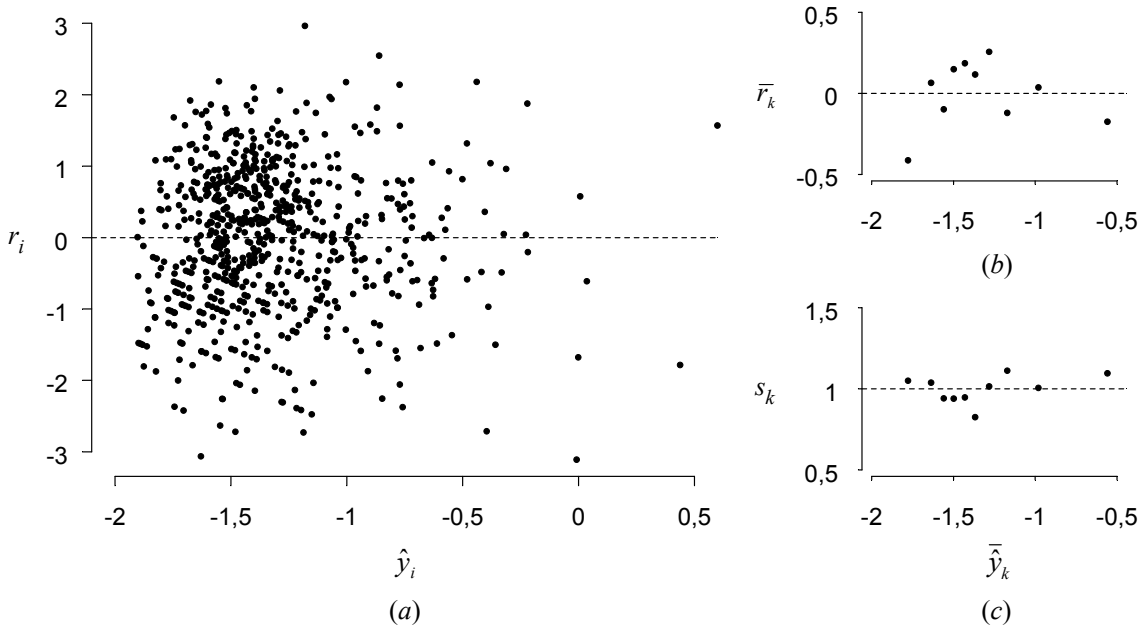


Figura 10.13 **Gráfico de los residuos estandarizados r_i frente a los valores predichos $\hat{y}_i$ de la regresión lineal del logaritmo del β-caroteno sobre el α-tocoferol en el grupo control del estudio EURAMIC (*a*), junto con las medias $\bar{r}_k$ (*b*) y desviaciones típicas s_k (*c*) de los residuos estandarizados por deciles de los valores predichos.**

10.3.6 Observaciones atípicas e influyentes

En el diagnóstico de un modelo de regresión lineal, tan importante como evaluar las asunciones de linealidad y homogeneidad de la varianza es examinar la contribución o influencia de cada observación en el modelo estimado. En general, es deseable que el modelo estimado responda al patrón global de los datos; esto es, las estimaciones de los parámetros del modelo deben basarse en el conjunto de todas las observaciones y no únicamente en un reducido número de observaciones muy influyentes. De esta forma, se tendrá un mayor grado de confianza a la hora de inferir los resultados del modelo a toda la población.

La forma más natural de medir la influencia de una observación en un modelo de regresión lineal simple es comparar las estimaciones de la constante y la pendiente obtenidas en la muestra

completa con sus correspondientes estimaciones tras excluir dicha observación. Una medida estandarizada del cambio global que se produce en las estimaciones b_0 y b_1 al eliminar la i-ésima observación es la **distancia de Cook** D_i, que en su forma más simple puede expresarse como

$$D_i = \frac{r_i^2 h_i}{2(1-h_i)}.$$

De esta fórmula se desprende que la influencia de una observación en las estimaciones b_0 y b_1 depende tanto de su residuo estandarizado r_i como de su leverage h_i. Los residuos estandarizados r_i determinan la desviación del valor observado de la variable respuesta respecto al valor predicho por la recta de regresión, de tal forma que valores altos de r_i en valor absoluto corresponden a observaciones pobremente ajustadas, que se conocen como **observaciones atípicas** o **outliers**. Estos outliers provocan una disminución de la calidad global del ajuste, lo que redunda en un aumento de la varianza residual s^2 y del error estándar de las estimaciones b_0 y b_1. Sin embargo, los outliers no son necesariamente influyentes en las estimaciones puntuales b_0 y b_1, ya que su influencia también depende del leverage. El **leverage** h_i de una observación es una medida estandarizada de la distancia entre el valor de la variable explicativa y su media, que se define como

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}$$

y toma valores entre $1/n$ y 1 con una media de $\bar{h} = 2/n$. A diferencia de los outliers que corresponden a observaciones con valores atípicos de la variable respuesta, las observaciones con alto leverage son aquellas con valores extremos de la variable explicativa. El leverage juega un papel determinante en la distinción entre outliers y observaciones influyentes. Así, por ejemplo, el punto A de la Figura 10.14(a) es un outlier extremo (residuo muy elevado) que tiene poca influencia en la recta de regresión estimada ya que ésta no varía sensiblemente tras excluir dicho punto. Esto se debe a que la observación A presenta un valor centrado de la variable explicativa (leverage muy bajo) que mitiga en gran medida su influencia sobre las estimaciones b_0 y b_1 (distancia de Cook moderada). Por el contrario, el punto B de la Figura 10.14(b) no es un outlier tan marcado pero tiene una influencia mucho mayor en la recta de regresión estimada, particularmente en la pendiente b_1, debido a que este punto presenta un valor muy extremo de la variable explicativa.

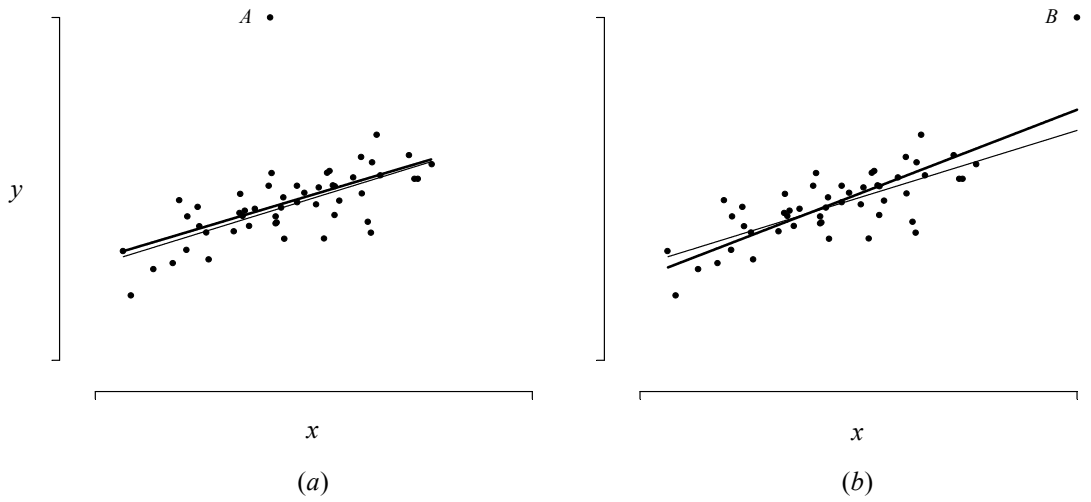


Figura 10.14 **Rectas de regresión resultantes de incluir (línea gruesa) y excluir (línea fina) los puntos *A* y *B* del ajuste del modelo lineal.**

Una **observación** será tanto más **influyente** en las estimaciones b_0 y b_1 de la recta de regresión cuanto mayor sea su distancia de Cook D_i. En general, se recomienda examinar detenidamente aquellas observaciones con una distancia de Cook superior a $4/(n-2)$, que corresponde, por ejemplo, a un punto con un leverage medio $h_i = 2/n$ y un residuo estandarizado alto $r_i = \pm 2$. No obstante, la selección de un valor crítico para D_i es un tanto arbitraria y es preferible evaluar la influencia relativa de cada observación en comparación con las restantes observaciones. Un gráfico útil es el diagrama de dispersión de los residuos estandarizados r_i frente a los leverages h_i, donde cada observación se representa mediante un círculo de área proporcional a su distancia de Cook D_i. En este gráfico, el tamaño de los círculos identificará claramente las observaciones más influyentes, mientras que la posición permitirá discernir la contribución de los residuos y leverages a la influencia de dichas observaciones.

Ejemplo 10.15 La Figura 10.15 muestra los residuos estandarizados r_i frente a los leverages h_i de la regresión lineal del colesterol HDL sobre el índice de masa corporal, donde se incluyen líneas de referencia horizontales en $r_i = -2$, 0 y 2 y verticales en el doble $h_i = 0{,}0075$ y el triple $h_i = 0{,}0113$ del leverage medio $\bar{h} = 2/533 = 0{,}0038$. El área de los círculos es proporcional a la distancia de Cook D_i e indica la influencia relativa de cada observación. Por supuesto, la influencia de las observaciones aumenta conforme aumentan sus residuos estandarizados en valor absoluto (dirección vertical del gráfico) y sus leverages (dirección horizontal). Sin embargo, no se aprecian observaciones marcadamente influyentes que pudieran conducir los resultados globales del modelo. La observación más influyente $D_i = 0{,}043$ se presenta en el cuadrante superior izquierda de la Figura 10.15, que corresponde a un outlier con un residuo muy alto $r_i = 4{,}28$ y un leverage moderado $h_i = 0{,}0047$. Las estimaciones de la constante y la pendiente de la recta de regresión excluyendo este outlier son $b_0^{(i)} = 1{,}71$ y $b_1^{(i)} = -0{,}024$ que, comparadas con las estimaciones (error estándar) $b_0 = 1{,}69$ (0,092) y $b_1 = -0{,}023$ (0,0035) obtenidas en la muestra completa (Ejemplo 10.9), suponen un cambio estandarizado de $(b_0^{(i)} - b_0)/SE(b_0) = (1{,}71 - 1{,}69)/0{,}092 = 0{,}20$ en la constante y $(b_1^{(i)} - b_1)/SE(b_1) = (-0{,}024 + 0{,}023)/0{,}0035 = -0{,}23$ en la pendiente. Así, a pesar de que este outlier está muy mal ajustado, no afecta substancialmente a la recta de regresión estimada.

Ejemplo 10.16 En la Figura 10.16 se representan los residuos estandarizados r_i frente a los leverages h_i de la regresión lineal del logaritmo del β-caroteno sobre el α-tocoferol. En una primera inspección visual se distinguen al menos 3 observaciones con una influencia sensiblemente mayor que las demás, que corresponden a los círculos de mayor tamaño situados a la derecha del gráfico. Los valores observados, predichos y las medidas diagnósticas asociadas a dichas observaciones se presentan en la Tabla 10.5. A diferencia del ejemplo anterior, donde la observación más influyente correspondía a un outlier, estas 3 observaciones presentan leverages muy altos $h_i = 0{,}044$, 0,038 y 0,022 debidos a valores muy elevados del α-tocoferol, y sólo una de ellas está pobremente ajustada con $r_i = -3{,}11$. Para evaluar la influencia conjunta de dichas observaciones en la recta de regresión estimada, se calcularon los coeficientes del modelo excluyendo simultáneamente las 3 observaciones, que resultaron ser $b_0^{(i)} = -1{,}93$ y $b_1^{(i)} = 0{,}0042$. En comparación con las estimaciones (error estándar) $b_0 = -1{,}91$ (0,055) y $b_1 = 0{,}0040$ (0,00032) obtenidas en la muestra completa (Ejemplo 10.14), la eliminación de estas 3 observaciones provoca un cambio estandarizado en la constante de $(-1{,}93 + 1{,}91)/0{,}055 = -0{,}36$ y en la pendiente de $(0{,}0042 - 0{,}0040)/0{,}00032 = 0{,}50$. Esto es, la exclusión de dichas observaciones conlleva una disminución en la constante de aproximadamente un tercio de su error estándar y un aumento en la pendiente de la mitad del error estándar. Así, aunque estas 3 observaciones no son extremadamente influyentes por sí mismas, el modelo sí parece ser sensible a la presencia de observaciones con alto leverage (Figura 10.16).

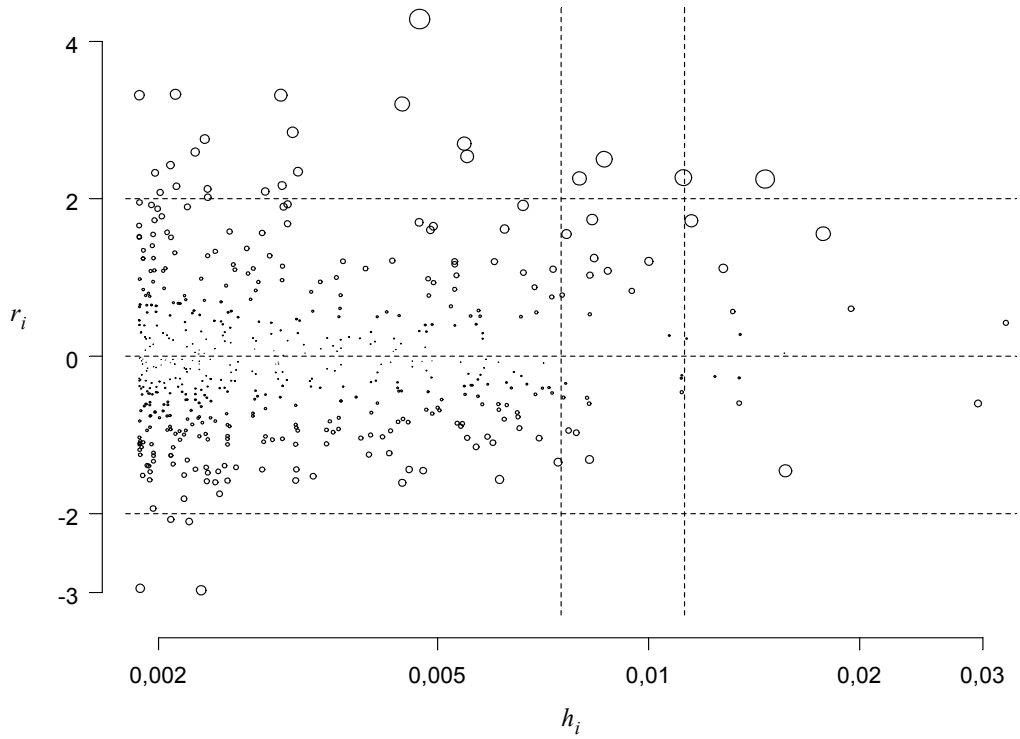


Figura 10.15 **Gráfico de los residuos estandarizados r_i frente a los leverages h_i de la regresión lineal del colesterol HDL sobre el índice de masa corporal en el grupo control del estudio EURAMIC. El área de los círculos es proporcional a la distancia de Cook D_i. Las líneas de referencia horizontales corresponden a $r_i = -2$, 0 y 2, y las verticales a $h_i = 2\bar{h} = 0{,}0075$ y $3\bar{h} = 0{,}0113$. El eje horizontal está en escala logarítmica para mejorar la representación gráfica.**

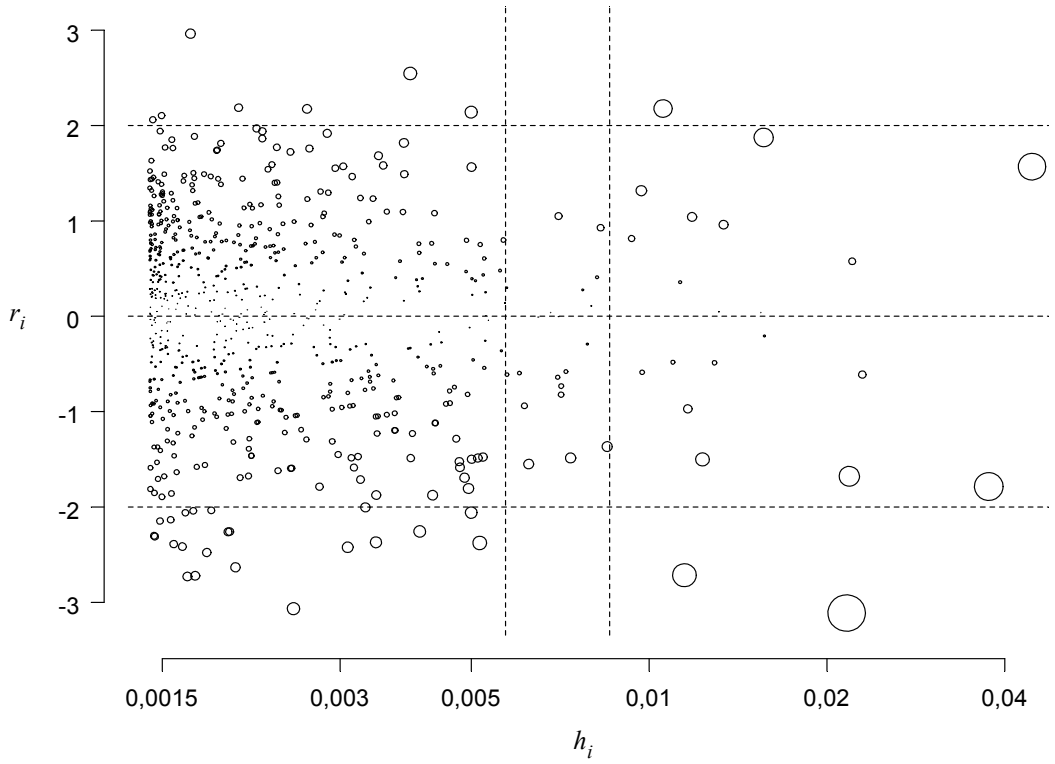


Figura 10.16 **Gráfico de los residuos estandarizados r_i frente a los leverages h_i de la regresión lineal del logaritmo del β-caroteno sobre el α-tocoferol en el grupo control del estudio EURAMIC. El área de los círculos es proporcional a la distancia de Cook D_i. Las líneas de referencia horizontales corresponden a $r_i = -2$, 0 y 2, y las verticales a $h_i = 2\bar{h} = 0{,}0057$ y $3\bar{h} = 0{,}0086$. El eje horizontal está en escala logarítmica.**

Tabla 10.5 Observaciones más influyentes en la regresión lineal del logaritmo del β-caroteno sobre el α-tocoferol en el grupo control del estudio EURAMIC.

Valores observados		Valor predicho	Medidas diagnósticas			Estimaciones*	
x_i	y_i	$\hat{y}_i$	r_i	h_i	D_i	$b_0^{(i)}$	$b_1^{(i)}$
626,8	1,74	0,60	1,57	0,044	0,057	−1,90	0,0039
586,6	−0,87	0,44	−1,79	0,038	0,062	−1,92	0,0041
475,1	−2,30	−0,01	−3,11	0,022	0,107	−1,93	0,0041

* Estimaciones de la constante y la pendiente de la recta de regresión tras excluir la observación correspondiente. Las estimaciones (y su error estándar) en la muestra completa de 700 controles fueron $b_0 = -1{,}91$ (0,055) y $b_1 = 0{,}0040$ (0,00032).

En ocasiones resulta lícito eliminar las observaciones marcadamente influyentes, bien por tratarse de valores atípicos de la variable respuesta o bien por presentar valores extremos de la variable explicativa. En tal caso, las inferencias derivadas del modelo deben limitarse exclusivamente al rango de valores observados en el resto de la muestra. No obstante, el tratamiento de observaciones influyentes no pasa necesariamente por su exclusión del ajuste del modelo. Un procedimiento alternativo de uso generalizado consiste en encontrar una transformación de la variable explicativa o respuesta que permita reducir la influencia de dichas observaciones. Por un lado, las transformaciones de la variable respuesta afectan al residuo estandarizado pero no al leverage de una observación, por lo que sólo son potencialmente útiles para atenuar la influencia de outliers. Por el contrario, las transformaciones de la variable explicativa influyen tanto en los residuos como en los leverages, de tal forma que estas transformaciones también pueden utilizarse para mitigar la influencia de observaciones extremas en la variable explicativa.

Ejemplo 10.17 Con objeto de reducir la influencia de las observaciones con valores muy elevados del α-tocoferol (alto leverage) en el modelo de regresión lineal del logaritmo del β-caroteno sobre el α-tocoferol, se podría aplicar a su vez una transformación logarítmica a la variable explicativa. En la Figura 10.17(*a*) se muestra la recta de regresión estimada entre el logaritmo del β-caroteno y el logaritmo del α-tocoferol,

$$\log \bar{y}_G = -3{,}76 + 0{,}51 \log x,$$

con errores estándar $SE(b_0) = 0{,}19$ y $SE(b_1) = 0{,}039$. Al exponenciar ambos lados de la igualdad, se tiene que la media geométrica de la variable respuesta es una función potencial de la variable explicativa (panel *b* de la Figura 10.17),

$$\bar{y}_G = \exp(-3{,}76 + 0{,}51 \log x) = 0{,}023x^{0{,}51}.$$

Este modelo tiene entonces una interpretación simple en la escala original de ambas variables ya que, al aumentar c veces la variable explicativa, la razón de medias geométricas es constante e igual a

$$\frac{\bar{y}_G(cx)}{\bar{y}_G(x)} = \frac{0{,}023(cx)^{0{,}51}}{0{,}023x^{0{,}51}} = c^{0{,}51};$$

es decir, a incrementos relativos en la variable explicativa les corresponde un mismo cambio relativo en la variable respuesta. Por ejemplo, incrementos del 50% ($c = 1{,}50$) en el nivel de α-tocoferol se asocian con un aumento del $100(1{,}50^{0{,}51} - 1) = 100(1{,}23 - 1) = 23\%$ en la media geométrica de β-caroteno. El IC al 95% para la razón de medias geométricas viene dado por

$$c^{b_1 \pm t_{698;0{,}975} SE(b_1)} = 1{,}50^{0{,}51 \pm 1{,}96 \cdot 0{,}039} = (1{,}19;\ 1{,}27),$$

de donde se concluye con una confianza del 95% que la media geométrica de β-caroteno aumenta entre un 19 y un 27% por cada incremento del 50% en el nivel de α-tocoferol.

La utilización de una transformación logarítmica para el α-tocoferol ha producido un doble efecto beneficioso en el ajuste del modelo. Por un lado, aunque persisten las observaciones con alto leverage (debidas, en este caso, a valores muy bajos del α-tocoferol), su influencia es ahora sensiblemente menor, como indica el tamaño de los círculos de la Figura 10.18(*a*). Por otro lado, la relación subyacente entre el α-tocoferol y el β-caroteno parece responder mejor al modelo potencial de la Figura 10.17(*b*), obtenido mediante transformaciones logarítmicas de ambas variables, que al modelo exponencial de la Figura 10.12(*b*), resultante de transformar únicamente el β-caroteno. Esta apreciación se fundamenta en que la curvatura de los residuos de la regresión lineal del logaritmo del β-caroteno sobre el α-tocoferol (panel *b* de la Figura 10.13) desaparece al transformar también el α-tocoferol (panel *b* de la Figura 10.18).

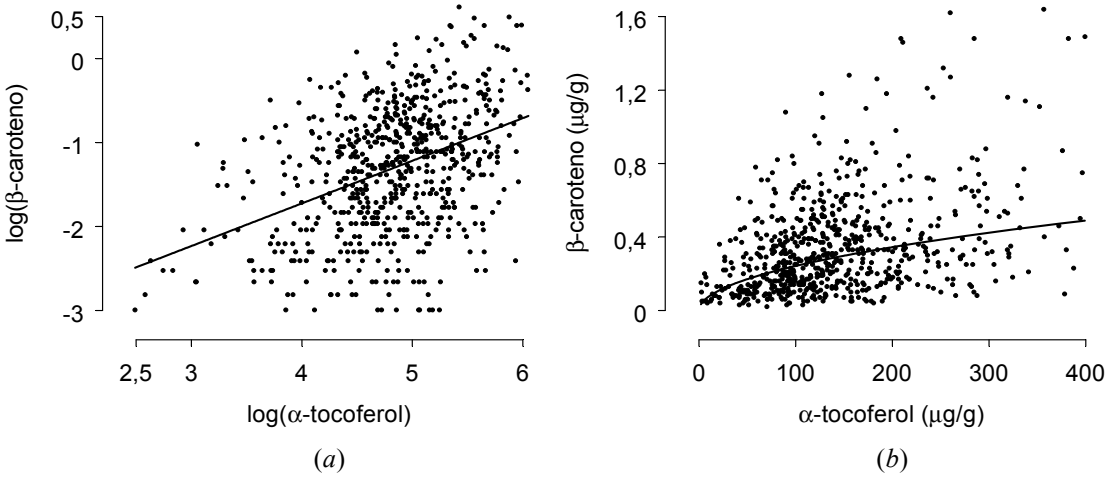


Figura 10.17 **Recta de regresión del logaritmo del β-caroteno sobre el logaritmo del α-tocoferol en el grupo control del estudio EURAMIC (*a*) y tendencia potencial resultante en la escala original de ambas variables (*b*).**

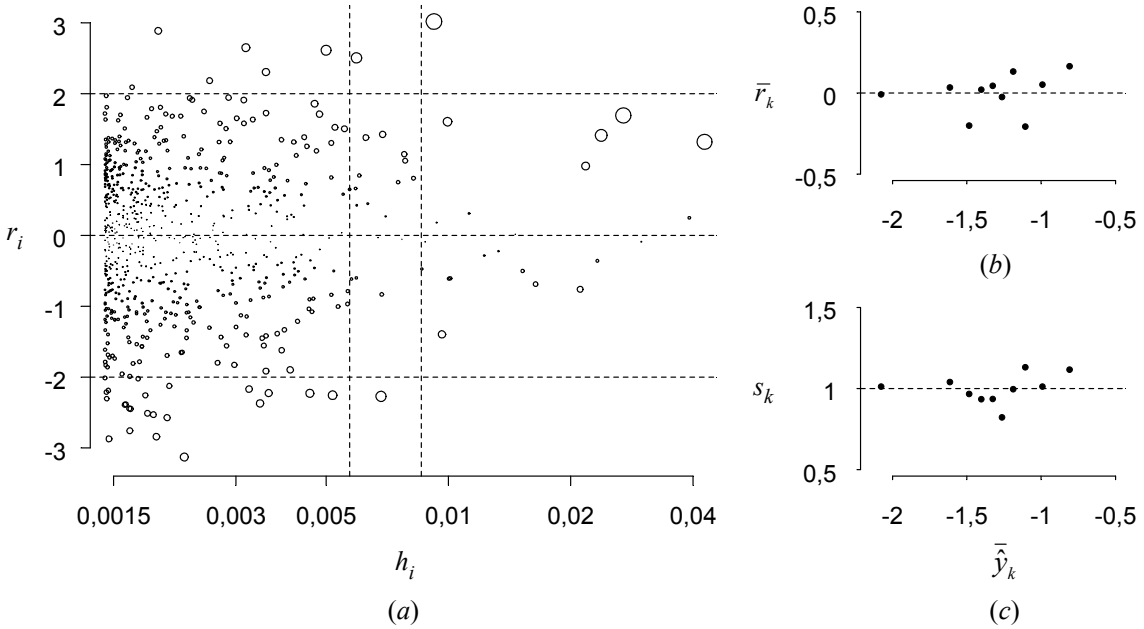


Figura 10.18 **Gráfico de los residuos estandarizados r_i frente a los leverages h_i de la regresión lineal del logaritmo del β-caroteno sobre el logaritmo del α-tocoferol en el grupo control del estudio EURAMIC (*a*), donde el área de los círculos es proporcional a la distancia de Cook D_i, y gráfico de las medias $\bar{r}_k$ (*b*) y desviaciones típicas s_k (*c*) de los residuos estandarizados por deciles de los valores predichos.**

10.3.7 Variable explicativa dicotómica

Hasta el momento se han considerado únicamente modelos de regresión lineal con variables explicativas continuas. No obstante, las variables explicativas pueden ser tanto continuas como categóricas ya que la regresión lineal no establece ninguna asunción respecto a su distribución. En este apartado se revisa el ajuste e interpretación de modelos de regresión lineal simple con una única variable explicativa dicotómica, que clasifica a los sujetos en dos grupos o categorías según la presencia o ausencia de una determinada característica. El tratamiento de variables explicativas politómicas con tres o más categorías se abordará en el Tema 11 ya que estas variables requieren de múltiples variables indicadoras para las distintas categorías.

Las variables explicativas dicotómicas se introducen en los modelos de regresión mediante una única **variable indicadora** X, que toma distintos valores x_i en cada una de las dos categorías de la variable. Aunque la elección de estos valores es arbitraria, la codificación más frecuente es $x_i = 1$ en los n_1 sujetos pertenecientes al primer grupo y 0 en los restantes $n_2 = n - n_1$ sujetos del segundo grupo. Bajo esta codificación, la interpretación del modelo de regresión lineal de la variable respuesta Y sobre la variable indicadora X es particularmente sencilla, dado que la estimación de la pendiente se reduce a

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n}{n_1 n_2}\sum_{i=1}^{n_1}(y_i - \bar{y}) = \frac{n}{n_2}(\bar{y}_1 - \bar{y}) = \bar{y}_1 - \bar{y}_2$$

y la constante a

$$b_0 = \bar{y} - b_1\bar{x} = \bar{y} - \frac{n_1}{n}(\bar{y}_1 - \bar{y}_2) = \bar{y}_2,$$

donde $\bar{y}_1$ y $\bar{y}_2$ son las medias muestrales de la variable respuesta en la primera y segunda categoría de la variable explicativa, respectivamente. Así, la constante corresponde simplemente a la media de la variable respuesta en el segundo grupo ($x_i = 0$) y la pendiente a la diferencia de medias entre el primer ($x_i = 1$) y el segundo grupo ($x_i = 0$). Asimismo, el error estándar de la constante viene dado por

$$SE(b_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}} = s\sqrt{\frac{1}{n} + \frac{n_1}{n_2 n}} = \frac{s}{\sqrt{n_2}}$$

y el error estándar de la pendiente por

$$SE(b_1) = \frac{s}{s_x\sqrt{n-1}} = s\sqrt{\frac{n}{n_1 n_2}} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

donde la varianza residual s^2 no es más que la combinación de las varianzas s_1^2 y s_2^2 de la variable respuesta en ambos grupos,

$$\begin{aligned} s^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 \\ &= \frac{\sum_{i=1}^{n_1}(y_i - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_j - \bar{y}_2)^2}{n-2} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n-2}. \end{aligned}$$

De estos resultados se desprende que la pendiente b_1 y su error estándar $SE(b_1)$ coinciden exactamente con la estimación puntual y el error estándar de la diferencia de medias en distribuciones con igual varianza (véase Apartado 6.3.1). Puede concluirse, por tanto, que las inferencias relativas a la pendiente de un modelo de regresión lineal con una única variable explicativa dicotómica son algebraicamente equivalentes a la comparación de medias mediante el **test de la *t* de Student para muestras independientes con igual varianza**.

Ejemplo 10.18 Para comparar los niveles medios de colesterol HDL entre los casos de infarto de miocardio y los controles libres de la enfermedad, se podría ajustar un modelo de regresión lineal simple del colesterol HDL sobre la variable indicadora del estatus caso/control ($x_i = 1$ en los casos y 0 en los controles) en la muestra completa de $n_1 = 462$ casos de infarto y $n_2 = 539$ controles del estudio EURAMIC con valores del colesterol HDL. La recta de regresión estimada entre el colesterol HDL y la variable indicadora del estatus caso/control es

$$\hat{y} = 1{,}09 - 0{,}11x,$$

con una desviación típica residual del colesterol HDL de $s = 0{,}27$ mmol/l que, debido a la hipótesis de homogeneidad de la varianza, se asume constante en casos y controles. El error estándar de la constante es $SE(b_0) = 0{,}012$ y de la pendiente $SE(b_1) = 0{,}017$. La constante $b_0 = 1{,}09$ mmol/l estima la media del colesterol HDL en los sujetos con valor 0 de la variable indicadora; esto es, el valor esperado del colesterol HDL en los controles libres de la enfermedad, cuyo IC al 95% es

$$b_0 \pm t_{999;0{,}975}\ SE(b_0) = 1{,}09 \pm 1{,}96 \cdot 0{,}012 = (1{,}06;\ 1{,}11).$$

Por otra parte, la pendiente $b_1 = -0{,}11$ mmol/l determina el cambio en el nivel medio de colesterol HDL por cada incremento de una unidad en la variable indicadora, lo que equivale a la diferencia de medias entre casos ($x_i = 1$) y controles ($x_i = 0$). El IC al 95% para la diferencia de medias subyacente viene dado por

$$b_1 \pm t_{999;0{,}975}\ SE(b_1) = -0{,}11 \pm 1{,}96 \cdot 0{,}017 = (-0{,}14;\ -0{,}08)$$

y el contraste bilateral de la hipótesis de igualdad de medias H_0: $\beta_1 = 0$ mediante el estadístico

$$t = \frac{b_1}{SE(b_1)} = \frac{-0{,}11}{0{,}017} = -6{,}35$$

resulta en un valor $P = 2P(t_{999} \leq -6{,}35) \approx 2\Phi(-6{,}35) < 0{,}001$. Así, los casos de infarto de miocardio presentan un nivel medio de colesterol HDL significativamente inferior que los sujetos libres de la enfermedad ($P < 0{,}001$), con una diferencia estimada en 0,11 mmol/l (IC al 95% 0,08-0,14 mmol/l). Notar, por último, que estos resultados son exactamente iguales a los obtenidos mediante el test de la *t* de Student para muestras independientes con igual varianza (Ejemplos 6.7 y 6.8).

10.4 REFERENCIAS

1. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research, Fourth Edition*. Oxford: Blackwell Science, 2002.
2. Bickel PJ, Doksum KA. *Mathematical Statistics: Basic Ideas and Selected Topics*. Englewood Cliffs, NJ: Prentice Hall, 1977.

3. Casella G, Berger RL. *Statistical Inference, Second Edition*. Belmont, CA: Duxbury Press, 2002.
4. Colton T. *Estadística en Medicina*. Barcelona: Salvat, 1979.
5. Conover WJ. *Practical Nonparametric Statistics, Third Edition*. New York: John Wiley & Sons, 1999.
6. Draper NR, Smith H. *Applied Regression Analysis, Third Edition*. New York: John Wiley & Sons, 1998.
7. Kleinbaum DG, Kupper LL, Nizam A, Muller KE. *Applied Regression Analysis and Other Multivariable Methods, Fourth Edition*. Belmont, CA: Duxbury Press, 2008.
8. Peña D. *Estadística: Modelos y Métodos, Volumen 2, Modelos Lineales y Series Temporales*. Madrid: Alianza Editorial, 1987.
9. Rosner B. *Fundamentals of Biostatistics, Sixth Edition*. Belmont, CA: Duxbury Press, 2006.
10. Seber GAF, Lee AJ. *Linear Regression Analysis, Second Edition*. New York: John Wiley & Sons, 2003.
11. Snedecor GW, Cochran WG. *Statistical Methods, Eighth Edition*. Ames, IA: Iowa State University Press, 1989.
12. Stuart A, Ord JK, Arnold S. *Kendall's Advanced Theory of Statistics, Volume 2A, Classical Inference and the Linear Model, Sixth Edition*. London: Edward Arnold, 1999.
13. Weisberg S. *Applied Linear Regression, Third Edition*. New York: John Wiley & Sons, 2005.

TEMA 11

REGRESIÓN LINEAL MÚLTIPLE

11.1 INTRODUCCIÓN

En el Tema 10 se presentó la regresión lineal simple como una herramienta para analizar la relación lineal entre una variable respuesta continua y una única variable explicativa. En la práctica, sin embargo, suele contarse con más de una variable explicativa y el interés se centra en estudiar la relación de cada una de las variables explicativas con la variable respuesta, teniendo en cuenta a su vez las restantes variables explicativas. De este tipo de problemas se ocupa la regresión lineal múltiple.

En presencia de múltiples variables explicativas asociadas con la variable respuesta, la utilización de distintos modelos de regresión lineal simple para cada variable explicativa da lugar a estimaciones imprecisas y a menudo sesgadas de las asociaciones subyacentes con la variable respuesta. Para ilustrar este hecho, la Figura 11.1 presenta los diagramas de dispersión entre una variable respuesta Y y una variable explicativa X_1, diferenciando mediante puntos y círculos los valores de otra variable explicativa dicotómica X_2. En la Figura 11.1(a), la variable explicativa X_2 está asociada con la variable respuesta Y (los valores de Y tienden a ser mayores en uno que en otro grupo de X_2), pero no con la variable explicativa X_1 (los valores de X_1 se distribuyen por igual en ambas categorías de X_2). Si se ignora la variable X_2 y se ajusta un modelo de regresión lineal simple entre X_1 e Y a toda la nube de puntos (línea gruesa), se obtiene la misma pendiente que al ajustar distintas rectas para cada valor de X_2 (líneas finas) y, en consecuencia, la asociación entre X_1 e Y no estará confundida por X_2. No obstante, la varianza residual alrededor de la recta de regresión es mayor al ignorar la variable explicativa X_2, lo que ocasionará un mayor error estándar en la estimación de la pendiente. Por el contrario, en la Figura 11.1(b), la variable explicativa X_2 está asociada de forma independiente con la variable respuesta Y y con la variable explicativa X_1 (para valores fijos de X_1 o Y, los valores de la otra variable difieren según categorías de X_2). La pendiente de la recta de regresión simple entre X_1 e Y (línea gruesa) sobreestima el efecto independiente de X_1 sobre Y cuando X_2 permanece constante (líneas finas). Esto es debido a que las variables explicativas X_1 y X_2 están correlacionadas y la regresión lineal simple estimará los efectos confundidos de ambas variables al no poder discernir entre el efecto independiente de X_1 y el efecto inducido por su asociación con X_2.

La principal conclusión del ejemplo anterior es que, si las variables explicativas están relacionadas entre sí, lo que sucede con cierta frecuencia, la regresión lineal simple puede proporcionar estimaciones sesgadas de las asociaciones subyacentes de cada variable explicativa con la variable respuesta. Por ello, los efectos de distintas variables explicativas deben estudiarse conjuntamente mediante modelos de regresión lineal múltiple. Estos modelos son una extensión de la regresión lineal simple a la presencia de dos o más variables explicativas, que pueden ser tanto continuas como categóricas. Como veremos a continuación, la regresión lineal múltiple permite estimar el efecto independiente de cada variable explicativa, manteniendo constantes las restantes variables incluidas en el modelo. Su utilidad en los análisis epidemiológicos es, por tanto, directa ya que facilita estimaciones ajustadas del efecto de cada variable explicativa.

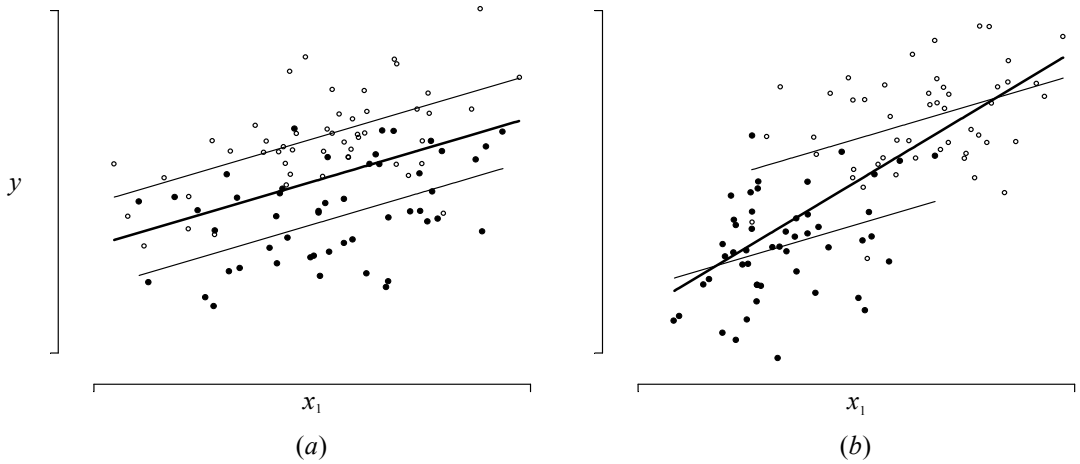


Figura 11.1 Diagramas de dispersión de la variable respuesta *Y* frente a la variable explicativa X_1 para distintos valores (puntos y círculos) de otra variable explicativa dicotómica X_2 asociada con *Y* pero no con X_1 (panel *a*) y asociada tanto con *Y* como con X_1 (panel *b*). Las líneas gruesas representan las rectas de regresión simple entre X_1 e *Y* ignorando la variable X_2 y las líneas finas corresponden a las rectas de regresión para cada valor de X_2.

11.2 ESTRUCTURA DE LA REGRESIÓN LINEAL MÚLTIPLE

El modelo de regresión lineal múltiple asume que la media de la variable respuesta Y puede expresarse como una combinación lineal de las variables explicativas $X_1, ..., X_p$; es decir, para valores fijos $x_1, ..., x_p$ de estas variables explicativas, el valor esperado de la variable respuesta es

$$E(Y|x_1, ..., x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \ .$$

La constante β_0 corresponde al valor esperado de Y cuando todas las variables explicativas son 0, $E(Y|0, \ldots, 0) = \beta_0 + \beta_1 0 + \ldots + \beta_p 0 = \beta_0$; mientras que cada **coeficiente de regresión** β_j determina el cambio esperado en Y por cada incremento de una unidad en X_j, manteniendo constantes el resto de variables explicativas,

$$\begin{aligned} &E(Y|x_1, ..., x_{j-1}, x_j + 1, x_{j+1}, ..., x_p) - E(Y|x_1, ..., x_p) \\ &\quad = \beta_0 + \beta_1 x_1 + \ldots + \beta_{j-1} x_{j-1} + \beta_j (x_j + 1) + \beta_{j+1} x_{j+1} + \ldots + \beta_p x_p \\ &\quad - (\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) = \beta_j. \end{aligned}$$

Así, los coeficientes de regresión asociados a cada variable explicativa no pueden estar confundidos por las demás variables explicativas, ya que éstas permanecen constantes. En este sentido, y a diferencia de la regresión simple, los coeficientes de regresión lineal múltiple facilitan el efecto independiente de cada variable explicativa sobre la variable respuesta ajustando o controlando por posibles diferencias en la distribución de las restantes variables explicativas incluidas en el modelo.

Para completar la estructura general de la regresión lineal múltiple, se asume que los valores individuales de la variable respuesta se distribuyen normalmente alrededor del valor esperado definido por la **ecuación de regresión**,

$$Y|x_1, ..., x_p \sim N(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \sigma^2),$$

o equivalentemente

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon,$$

donde el error aleatorio ε en la variable respuesta sigue una distribución normal con media 0 y varianza σ^2 para cualquier valor de las variables explicativas. De esta especificación del modelo de regresión lineal múltiple, se desprenden las siguientes asunciones:

- **Linealidad**: El valor esperado de la variable respuesta Y cambia linealmente con cada variable explicativa X_j, de tal forma que para valores fijos de las demás variables explicativas, cambios de magnitud constante a distintos niveles de X_j se asocian con un mismo cambio en la media de Y.
- **Aditividad**: El efecto conjunto de varias variables explicativas sobre la variable respuesta es la suma de sus efectos independientes.
- **Homogeneidad de la varianza**: La varianza de la variable respuesta permanece constante para cualquier valor de las variables explicativas.
- **Normalidad**: Dados unos valores fijos de las variables explicativas, la variable respuesta se distribuye de forma normal.

En el caso de dos variables explicativas, estas asunciones pueden representarse mediante el gráfico tridimensional de la Figura 11.2. Debido a las hipótesis de linealidad y aditividad, los valores esperados de Y para cualquier combinación de X_1 y X_2 se sitúan en el plano definido por la ecuación de regresión $\beta_0 + \beta_1 x_1 + \beta_2 x_2$. Asimismo, por las asunciones de homogeneidad de la varianza y normalidad, los valores individuales de Y para cualquier combinación de X_1 y X_2 se distribuyen de forma normal y con la misma varianza alrededor de dicho plano de regresión. Las hipótesis de linealidad y homogeneidad de la varianza se evaluarán utilizando procedimientos de diagnóstico gráfico similares a los empleados en regresión lineal simple. Las desviaciones de la asunción de aditividad se explorarán, por su parte, mediante la inclusión de términos de interacción entre las variables explicativas.

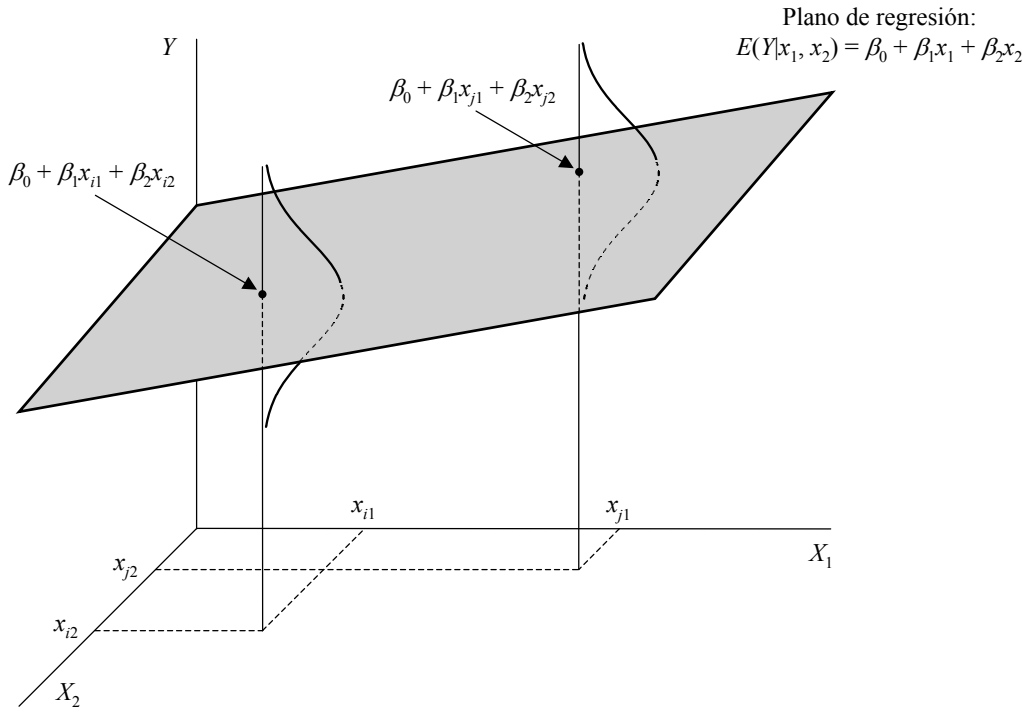


Figura 11.2 **Asunciones subyacentes al modelo de regresión lineal múltiple con dos variables explicativas.**

A estas asunciones, análogas a las utilizadas en regresión lineal simple, se añaden dos nuevas condiciones necesarias para poder estimar la ecuación de regresión:

- **Independencia lineal de las variables explicativas**: Ninguna variable explicativa es una combinación lineal exacta de las demás ya que, en tal caso, sus efectos individuales sobre la variable respuesta serían indiscernibles.

  Ejemplo 11.1 Supongamos que un modelo de regresión lineal múltiple incluye como variables explicativas la presión arterial sistólica X_1 y la presión arterial diastólica X_2,

  $$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

  Si se añade además la presión del pulso, definida como la diferencia entre la presión arterial sistólica y diastólica $X_3 = X_1 - X_2$, el modelo resultante puede reescribirse como

  $$\begin{aligned} Y &= \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \varepsilon \\ &= \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 (x_1 - x_2) + \varepsilon \\ &= \alpha_0 + (\alpha_1 + \alpha_3) x_1 + (\alpha_2 - \alpha_3) x_2 + \varepsilon, \end{aligned}$$

  que es algebraicamente equivalente al modelo anterior con $\beta_1 = \alpha_1 + \alpha_3$ y $\beta_2 = \alpha_2 - \alpha_3$. Existen, por tanto, infinitas combinaciones de los parámetros α_1, α_2 y α_3 que dan lugar a la misma ecuación de regresión (para cualquier valor de α_3, basta tomar $\alpha_1 = \beta_1 - \alpha_3$ y $\alpha_2 = \beta_2 + \alpha_3$ para obtener los mismos coeficientes de regresión β_1 y β_2). Así, como la presión del pulso es una combinación lineal exacta de la presión arterial sistólica y diastólica, no es posible determinar unívocamente los efectos independientes de cada una de estas tres variables explicativas.

- El número de observaciones n debe ser superior o igual al número de coeficientes $p + 1$ de la ecuación de regresión. Este requerimiento resulta obvio en el caso de $p = 2$ variables explicativas (véase Figura 11.2), ya que para determinar el plano de regresión se necesitan al menos $n = 3$ puntos u observaciones no alineadas.

Cabe destacar que estas dos condiciones son requerimientos teóricos mínimos para estimar la ecuación de regresión. En la práctica, sin embargo, el número de observaciones ha de ser muy superior al número de coeficientes de regresión para poder obtener estimaciones precisas de estos coeficientes y no incurrir en problemas de sobreajuste (esto es, modelar el error aleatorio en lugar de la relación subyacente). Un criterio habitual es no incluir más variables explicativas que el número de observaciones dividido por 10. Asimismo, aunque las variables explicativas no presenten una correlación lineal perfecta, es importante evaluar su grado de colinealidad. Si las variables explicativas son muy dependientes entre sí, resulta muy difícil separar sus efectos e identificar la contribución individual de cada una de ellas, lo que provocará estimaciones inestables de los coeficientes de regresión. Este problema se conoce como **multicolinealidad** y se tratará más adelante en el apartado de diagnóstico del modelo de regresión lineal múltiple.

11.3 ESTIMACIÓN E INFERENCIA DE LA ECUACIÓN DE REGRESIÓN

En este apartado se presenta, en primer lugar, el procedimiento de estimación de los coeficientes de regresión lineal múltiple. A continuación, se describen las propiedades de los estimadores y se derivan intervalos de confianza y tests de hipótesis para los coeficientes de regresión. Finalmente, se presentan intervalos de confianza para el valor esperado de la variable respuesta e intervalos de predicción para una nueva observación en función de los valores de las variables explicativas.

11.3.1 Estimación de los coeficientes de regresión

Al igual que en regresión lineal simple, las estimaciones puntuales $b_0, b_1, \ldots, b_p$ de los coeficientes de regresión $\beta_0, \beta_1, \ldots, \beta_p$ se obtienen mediante el **método de mínimos cuadrados** a partir de una muestra de n observaciones $(y_i, x_{i1}, \ldots, x_{ip})$ mutuamente independientes. En concreto, tal y como se muestra en la Figura 11.3 para dos variables explicativas, se trata de estimar los valores $b_0, b_1, \ldots, b_p$ que minimicen la **suma de cuadrados de los errores** o residuos $e_i = y_i - \hat{y}_i$, que corresponden a las distancias entre los valores observados y_i de la variable respuesta y los correspondientes valores estimados o predichos por la ecuación de regresión $\hat{y}_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip}$,

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \ldots - b_p x_{ip})^2.$$

Para estimar los coeficientes de regresión que minimizan esta suma de cuadrados del error, se calculan las derivadas parciales de SSE respecto a $b_0, b_1, \ldots, b_p$ y se igualan a cero, resultando el sistema de $p + 1$ ecuaciones lineales

$$\frac{\partial \text{SSE}}{\partial b_0} = -2\sum_{i=1}^{n} e_i = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \ldots - b_p x_{ip}) = 0,$$

$$\frac{\partial \text{SSE}}{\partial b_j} = -2\sum_{i=1}^{n} x_{ij} e_i = -2\sum_{i=1}^{n} x_{ij} (y_i - b_0 - b_1 x_{i1} - \ldots - b_p x_{ip}) = 0, \qquad j = 1, \ldots, p.$$

En general, este sistema lineal se resuelve utilizando álgebra de matrices. En el Apéndice al final del tema se derivan las fórmulas matriciales para calcular $b_0, b_1, \ldots, b_p$ que, bajo las asunciones de linealidad y aditividad, son estimadores insesgados de los coeficientes de

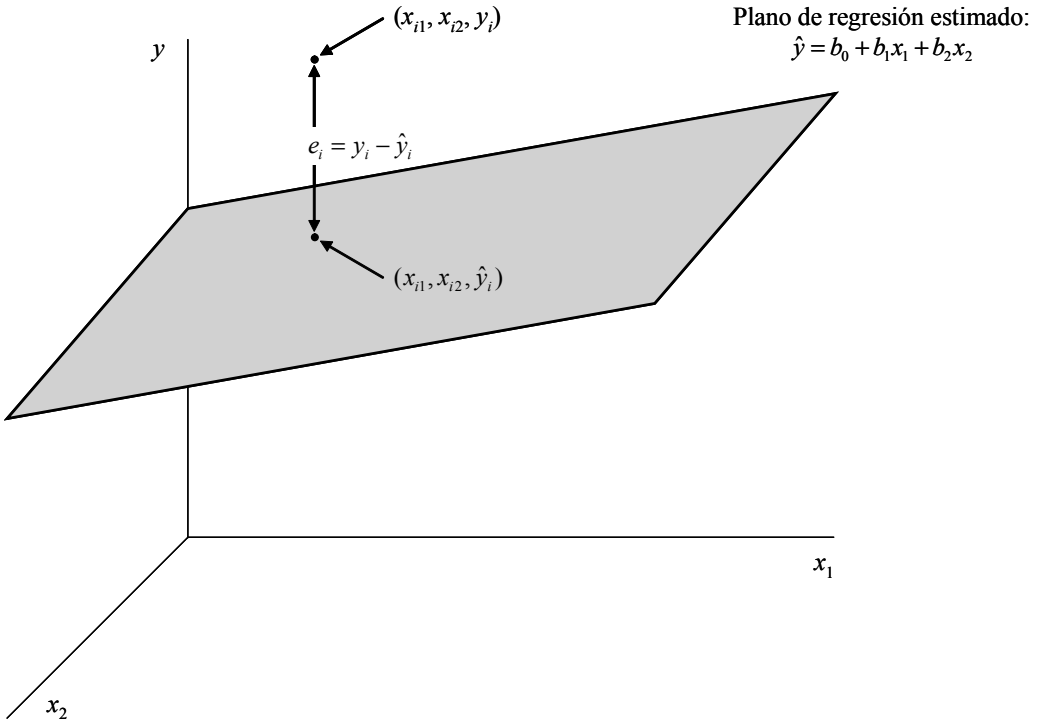


Figura 11.3 Error o desviación del valor observado de la variable respuesta respecto a su valor estimado por el plano de regresión.

regresión $\beta_0, \beta_1, \ldots, \beta_p$. En el caso particular de dos variables explicativas, puede comprobarse que estos estimadores vienen dados por

$$b_1 = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{1 - r^2_{x_1x_2}} \frac{s_y}{s_{x_1}},$$

$$b_2 = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{1 - r^2_{x_1x_2}} \frac{s_y}{s_{x_2}},$$

$$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 .$$

De estas expresiones se deduce que, si las variables explicativas X_1 y X_2 están incorrelacionadas $r_{x_1x_2} = 0$, las estimaciones de los coeficientes de regresión múltiple se reducen a $b_1 = r_{yx_1} s_y / s_{x_1}$ y $b_2 = r_{yx_2} s_y / s_{x_2}$, que son iguales a las obtenidas en regresión simple (véase Apartado 10.3.1). Por tanto, cuando las variables explicativas están incorrelacionadas, sus coeficientes estimados por regresión múltiple coinciden con los obtenidos de distintas regresiones simples para cada variable explicativa. Por el contrario, cuando las variables explicativas están correlacionadas, sus efectos ajustados mediante regresión múltiple pueden diferir notablemente de sus efectos crudos ignorando las restantes variables explicativas. Así, por ejemplo, la relación de la variable explicativa X_1 con la variable respuesta Y ajustando por la variable X_2 se estima mediante el coeficiente de regresión múltiple b_1, que depende no sólo de la correlación entre X_1 e Y r_{yx_1} (como ocurre en regresión lineal simple), sino también de sus respectivas correlaciones con la variable X_2 r_{yx_2} y $r_{x_1x_2}$.

Una vez estimada la ecuación de regresión, la varianza σ^2 de la variable respuesta alrededor de dicha ecuación se estima mediante la **varianza residual**

$$s^2 = \frac{\text{SSE}}{n - p - 1} = \frac{1}{n - p - 1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \ldots - b_p x_{ip})^2 ,$$

donde la suma de cuadrados del error SSE se divide por $n - p - 1$ ya que, una vez estimados los $p + 1$ coeficientes de regresión, los n errores o desviaciones de la variable respuesta respecto a la ecuación de regresión contienen $n - p - 1$ grados de libertad. Bajo las hipótesis de linealidad, aditividad y homogeneidad de la varianza, la varianza residual s^2 es un estimador insesgado del parámetro poblacional σ^2.

Ejemplo 11.2 En el Ejemplo 10.7 se estudió la relación del índice de masa corporal con el colesterol HDL utilizando un modelo de regresión lineal simple. No obstante, existen otros muchos determinantes de los niveles de colesterol HDL como, por ejemplo, el consumo de alcohol. Para obtener el efecto independiente de cada uno de estos determinantes, se podría ajustar un modelo de regresión lineal múltiple con el colesterol HDL como variable respuesta y el índice de masa corporal y el consumo de alcohol como variables explicativas.

En $n = 449$ controles del estudio EURAMIC con datos disponibles de estas variables, la media y la desviación típica fueron $\bar{x}_1 = 26{,}2$ y $s_{x_1} = 3{,}61$ kg/m^2 para el índice de masa corporal, $\bar{x}_2 = 16{,}5$ y $s_{x_2} = 21{,}8$ g/día para el consumo de alcohol y $\bar{y} = 1{,}08$ y $s_y = 0{,}295$ mmol/l para el colesterol HDL. El coeficiente de correlación de Pearson entre el índice de masa corporal y el consumo de alcohol fue $r_{x_1x_2} = -0{,}091$ y las correlaciones de estas variables explicativas con el colesterol HDL fueron $r_{yx_1} = -0{,}273$ y $r_{yx_2} = 0{,}232$, respectivamente. Las estimaciones de los coeficientes de regresión múltiple se obtienen entonces como

$$b_1 = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{1 - r^2_{x_1x_2}} \frac{s_y}{s_{x_1}} = \frac{-0{,}273 + 0{,}232 \cdot 0{,}091}{1 - 0{,}091^2} \frac{0{,}295}{3{,}61} = -0{,}0207,$$

$$b_2 = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{1 - r^2_{x_1x_2}} \frac{s_y}{s_{x_2}} = \frac{0{,}232 - 0{,}273 \cdot 0{,}091}{1 - 0{,}091^2} \frac{0{,}295}{21{,}8} = 0{,}0028,$$

$$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 = 1{,}08 + 0{,}0207 \cdot 26{,}2 - 0{,}0028 \cdot 16{,}5 = 1{,}58,$$

de donde resulta la ecuación de regresión

$$\hat{y} = 1{,}58 - 0{,}0207x_1 + 0{,}0028x_2,$$

con una varianza residual del colesterol HDL respecto a dicha ecuación

$$s^2 = \frac{\text{SSE}}{446} = \frac{1}{446} \sum_{i=1}^{449} \{y_i - (1{,}58 - 0{,}0207x_{i1} + 0{,}0028x_{i2})\}^2 = \frac{34{,}33}{446} = 0{,}077.$$

Estas estimaciones pueden obtenerse directamente de ajustar una regresión lineal múltiple del colesterol HDL sobre el índice de masa corporal y el consumo de alcohol en los programas estadísticos convencionales, cuyos resultados completos se muestran en la Tabla 11.1.

La ecuación de regresión puede utilizarse para estimar el valor esperado del colesterol HDL en función del índice de masa corporal y el consumo de alcohol. Así, por ejemplo, para un índice de masa corporal de 25 kg/m^2 y un consumo de alcohol de 20 g/día, el modelo estima un nivel medio de colesterol HDL de $\hat{y}(25, 20) = 1{,}58 - 0{,}0207 \cdot 25 + 0{,}0028 \cdot 20 = 1{,}12$ mmol/l.

Las estimaciones b_1 y b_2 determinan el efecto independiente de cada variable explicativa sobre la variable respuesta, una vez controladas las posibles diferencias en la otra variable explicativa.

Tabla 11.1 Resultados de la regresión lineal múltiple del colesterol HDL sobre el índice de masa corporal (IMC) y la ingesta de alcohol en los controles del estudio EURAMIC.

Análisis de la varianza*

	Suma de cuadrados	Grados de libertad	Varianza	Razón de varianzas
Regresión	4,58	2	2,29	29,72
Error	34,33	446	0,077	
Total	38,91	448		

* Coeficiente de determinación $R^2 = 4{,}58/38{,}91 = 0{,}118$.

Coeficientes de regresión

				Test H_0: $\beta_j = 0$	
	Estimación	Error estándar	IC al 95%	*t*	Valor *P*
Constante	1,58	0,098	(1,39; 1,77)	16,14	< 0,001
IMC	−0,0207	0,0036	(−0,0278; −0,0135)	−5,68	< 0,001
Alcohol	0,0028	0,0006	(0,0016; 0,0040)	4,68	< 0,001

Por un lado, manteniendo constante el consumo de alcohol, cada incremento de $c_1 = 3{,}50$ kg/m^2 en el índice de masa corporal se asocia con una disminución media en el colesterol HDL de

$$\begin{aligned}\hat{y}\,(x_1 + c_1, x_2) - \hat{y}\,(x_1, x_2) &= b_0 + b_1(x_1 + c_1) + b_2x_2 - (b_0 + b_1x_1 + b_2x_2)\\ &= c_1b_1 = 3{,}50(-0{,}0207) = -0{,}072.\end{aligned}$$

Por otro lado, para un mismo índice de masa corporal, incrementos de $c_2 = 20$ g/día (aproximadamente una desviación típica) en la ingesta de alcohol se asocian con un aumento medio en el colesterol HDL de

$$\begin{aligned}\hat{y}\,(x_1, x_2 + c_2) - \hat{y}\,(x_1, x_2) &= b_0 + b_1x_1 + b_2(x_2 + c_2) - (b_0 + b_1x_1 + b_2x_2)\\ &= c_2b_2 = 20 \cdot 0{,}0028 = 0{,}056.\end{aligned}$$

Para evaluar el grado de confusión inducido por el consumo de alcohol en la asociación entre el índice de masa corporal y el colesterol HDL, basta comparar el coeficiente ajustado mediante regresión múltiple $b_1 = -0{,}0207$ con el coeficiente crudo obtenido de una regresión simple en la misma muestra de 449 controles $b_1^* = r_{yx_1}s_y / s_{x_1} = -0{,}273 \cdot 0{,}295/3{,}61 = -0{,}0222$. La razón entre los coeficientes crudo y ajustado

$$\frac{b_1^*}{b_1} = \frac{-0{,}0222}{-0{,}0207} = 1{,}08$$

indica que, si no se ajusta por el consumo de alcohol, se sobreestima un $100(1{,}08 - 1) = 8\%$ la asociación inversa del índice de masa corporal con el colesterol HDL. Esto es debido a que el consumo de alcohol presenta una leve correlación negativa con el índice de masa corporal, lo que induce un pequeño sesgo en la estimación cruda (una pequeña parte de la reducción del colesterol HDL entre los sujetos con sobrepeso no se debe a su mayor índice de masa corporal sino a un consumo de alcohol ligeramente menor). No obstante, los efectos crudo y ajustado no difieren substancialmente, por lo que el consumo de alcohol no parece ser un factor de confusión importante para la asociación entre el índice de masa corporal y el colesterol HDL en los controles del estudio EURAMIC.

11.3.2 Inferencia sobre los coeficientes de regresión

En el Apéndice al final del tema se demuestra que, bajo las asunciones de linealidad, aditividad y homogeneidad de la varianza, los estimadores de mínimos cuadrados b_j siguen aproximadamente una distribución normal con media β_j y varianza $\sigma^2 v_{jj}$ en muestras suficientemente grandes,

$$\frac{b_j - \beta_j}{\sigma\sqrt{v_{jj}}} \overset{\sim}{\to} N(0,\ 1), \quad j = 0,\ 1,\ \ldots,\ p,$$

donde v_{jj} es un valor conocido que depende del tamaño muestral y de las varianzas y covarianzas entre las variables explicativas. Si se reemplaza el parámetro desconocido σ por la desviación típica residual s, puede probarse que los estadísticos resultantes siguen aproximadamente una distribución t de Student con los $n - p - 1$ grados de libertad correspondientes a la estimación de la desviación típica residual,

$$\frac{b_j - \beta_j}{s\sqrt{v_{jj}}} \overset{\sim}{\to} t_{n-p-1}, \qquad j = 0,\ 1,\ \ldots,\ p.$$

Notar que estas distribuciones de los estimadores b_j en muestras suficientemente grandes no requieren de la asunción de normalidad y, por tanto, son válidas para cualquier distribución subyacente de la variable respuesta.

Utilizando estos resultados, los intervalos de confianza al $100(1-\alpha)\%$ para los coeficientes de regresión β_j vienen dados por

$$b_j \pm t_{n-p-1,1-\alpha/2}\, s\sqrt{v_{jj}}$$

y los contrastes bilaterales de las hipótesis de ausencia de efecto independiente de cada variable explicativa H_0: $\beta_j = 0$ se realizan mediante los estadísticos

$$t = \frac{b_j}{s\sqrt{v_{jj}}},$$

que bajo dichas hipótesis nulas se distribuyen aproximadamente como una t de Student con $n-p-1$ grados de libertad.

Ejemplo 11.3 Los programas estadísticos convencionales facilitan directamente las estimaciones puntuales de los coeficientes de regresión lineal múltiple y sus errores estándar. Según la Tabla 11.1, los errores estándar de los coeficientes estimados para el índice de masa corporal y el consumo de alcohol son respectivamente $SE(b_1) = s\sqrt{v_{11}} = 0{,}0036$ y $SE(b_2) = s\sqrt{v_{22}} = 0{,}0006$. Por tanto, los ICs al 95% para estos coeficientes de regresión son

$$b_1 \pm t_{446;0,975}\, SE(b_1) = -0{,}0207 \pm 1{,}97 \cdot 0{,}0036 = (-0{,}0278;\ -0{,}0135),$$
$$b_2 \pm t_{446;0,975}\, SE(b_2) = 0{,}0028 \pm 1{,}97 \cdot 0{,}0006 = (0{,}0016;\ 0{,}0040),$$

que también se incluyen dentro de los resultados de la Tabla 11.1. En general, el intervalo de confianza para el efecto subyacente $c_j\beta_j$ asociado a un aumento de c_j unidades en la variable explicativa X_j se calcula como

$$c_j b_j \pm t_{n-p-1,1-\alpha/2}\, SE(c_j b_j) = c_j\{b_j \pm t_{n-p-1,1-\alpha/2}\, SE(b_j)\}.$$

Así, puede afirmarse con una confianza del 95% que el nivel medio de colesterol HDL en la población de referencia del estudio EURAMIC disminuye entre $3{,}50 \cdot 0{,}0135 = 0{,}047$ y $3{,}50 \cdot 0{,}0278 = 0{,}097$ mmol/l por cada incremento de $c_1 = 3{,}50$ kg/m^2 en el índice de masa corporal entre sujetos con la misma ingesta de alcohol, y que la media poblacional del colesterol HDL aumenta entre $20 \cdot 0{,}0016 = 0{,}032$ y $20 \cdot 0{,}0040 = 0{,}080$ mmol/l por cada incremento de $c_2 = 20$ g/día en el consumo de alcohol entre sujetos con el mismo índice de masa corporal. Estos efectos independientes del índice de masa corporal y de la ingesta de alcohol sobre el colesterol HDL son muy significativos, ya que sus correspondientes test estadísticos

$$t = \frac{b_1}{SE(b_1)} = \frac{-0{,}0207}{0{,}0036} = -5{,}68,$$

$$t = \frac{b_2}{SE(b_2)} = \frac{0{,}0028}{0{,}0006} = 4{,}68,$$

arrojan valores P bilaterales $2P(t_{446} \leq -5{,}68) \approx 2\Phi(-5{,}68) < 0{,}001$ y $2P(t_{446} \geq 4{,}68) \approx 2\{1 - \Phi(4{,}68)\} < 0{,}001$, tal como muestra la Tabla 11.1.

11.3.3 Inferencia sobre la ecuación de regresión

La ecuación de regresión puede utilizarse para estimar el valor esperado de la variable respuesta en función de los valores de las variables explicativas. Dados unos determinados valores $x_{01}, \ldots,$

x_{0p} de las variables explicativas, el estimador insesgado del valor esperado de la variable respuesta es

$$\hat{y}_0 = b_0 + b_1 x_{01} + \ldots + b_p x_{0p}$$

que, como se muestra en el Apéndice de este tema, se distribuye de forma aproximadamente normal con media $\beta_0 + \beta_1 x_{01} + \ldots + \beta_p x_{0p}$ y varianza $\sigma^2 h_0$ en muestras suficientemente grandes,

$$\hat{y}_0 \overset{\sim}{\rightarrow} N(\beta_0 + \beta_1 x_{01} + \ldots + \beta_p x_{0p}, \sigma^2 h_0),$$

donde h_0 es el **leverage** del punto $(x_{01}, \ldots, x_{0p})$ que puede interpretarse como una medida estandarizada de su distancia respecto al centro de las medias muestrales $(\bar{x}_1, \ldots, \bar{x}_p)$ de las variables explicativas. A partir de la distribución t_{n-p-1} resultante de sustituir σ^2 por su estimación s^2, se sigue que el **intervalo de confianza** al $100(1-\alpha)\%$ para el valor esperado $\beta_0 + \beta_1 x_{01} + \ldots + \beta_p x_{0p}$ es

$$\hat{y}_0 \pm t_{n-p-1,1-\alpha/2}\, s \sqrt{h_0}\ .$$

Como cabría esperar, la estimación del valor esperado de la variable respuesta en el punto $(x_{01}, \ldots, x_{0p})$ será tanto más imprecisa cuanto más extremo sea dicho punto o, más concretamente, cuanto mayor sea su distancia estandarizada h_0 respecto al centro de las medias muestrales $(\bar{x}_1, \ldots, \bar{x}_p)$.

> ***Ejemplo 11.4*** Para un índice de masa corporal de $x_{01} = 25$ kg/m^2 y un consumo de alcohol de $x_{02} = 20$ g/día, el modelo de regresión múltiple estima un nivel medio de colesterol HDL de $\hat{y}_0 = 1{,}58 - 0{,}0207 \cdot 25 + 0{,}0028 \cdot 20 = 1{,}12$ mmol/l. El punto de estimación $(x_{01}, x_{02}) = (25, 20)$ está próximo al centro de las medias muestrales $(\bar{x}_1, \bar{x}_2) = (26{,}2;\ 16{,}5)$ de ambas variables explicativas y, en consecuencia, su leverage $h_0 = 0{,}0025$ es bajo. Así, el IC al 95% para el valor esperado del colesterol HDL entre los sujetos con un índice de masa corporal de 25 kg/m^2 y un consumo de alcohol de 20 g/día es
>
> $$\hat{y}_0 \pm t_{446;0,975}\, s \sqrt{h_0} = 1{,}12 \pm 1{,}97 \sqrt{0{,}077 \cdot 0{,}0025} = (1{,}09;\ 1{,}15).$$
>
> Por el contrario, el valor esperado del colesterol HDL entre los sujetos con un índice de masa corporal de 32 kg/m^2 y un consumo de alcohol de 40 g/día se estima en $1{,}58 - 0{,}0207 \cdot 32 + 0{,}0028 \cdot 40 = 1{,}03$ mmol/l, cuyo IC al 95%
>
> $$1{,}03 \pm 1{,}97 \sqrt{0{,}077 \cdot 0{,}0113} = (0{,}97;\ 1{,}09)$$
>
> es sensiblemente más impreciso, ya que el punto de estimación (32, 40) está distante del centro de las medias muestrales (26,2; 16,5) y presenta un leverage alto de 0,0113.

El valor predicho $\hat{y}_0$ es un estimador insesgado no sólo de la esperanza o media poblacional de la variable respuesta entre aquellos sujetos con los mismos valores de las variables explicativas, sino también de la respuesta individual de un nuevo sujeto $y_0 = \beta_0 + \beta_1 x_{01} + \ldots + \beta_p x_{0p} + \varepsilon_0$. En el Apéndice de este tema se demuestra que, bajo las asunciones de la regresión lineal múltiple (linealidad, aditividad, homogeneidad de la varianza y normalidad), la diferencia $\hat{y}_0 - y_0$ sigue la distribución normal

$$\hat{y}_0 - y_0 \sim N(0,\ \sigma^2(1 + h_0)),$$

de tal forma que el **intervalo de predicción** al $100(1-\alpha)\%$ para una nueva observación individual y_0 viene dado por

$$\hat{y}_0 \pm t_{n-p-1,1-\alpha/2}\, s \sqrt{1 + h_0}\ .$$

Este intervalo de predicción para la respuesta individual de un único sujeto será substancialmente más amplio que el intervalo de confianza para la respuesta media de todos los sujetos con un mismo patrón de variables explicativas ya que, además del error en la estimación del valor predicho por la ecuación de regresión, el intervalo de predicción incorpora la varianza residual de cada respuesta individual alrededor de dicha ecuación de regresión. Notar, además, que los intervalos de predicción para una nueva observación requieren de la hipótesis de normalidad, mientras que los intervalos de confianza para el valor esperado tienden a ser correctos en muestras suficientemente grandes, independientemente de la distribución subyacente de la variable respuesta.

Ejemplo 11.5 El valor predicho del colesterol HDL para un nuevo sujeto con un índice de masa corporal de 25 kg/m^2 y un consumo de alcohol de 20 g/día es de nuevo $\hat{y}_0 = 1{,}58 - 0{,}0207 \cdot 25 + 0{,}0028 \cdot 20 = 1{,}12$ mmol/l. Sin embargo, el intervalo de predicción al 95% para esta nueva observación

$$\hat{y}_0 \pm t_{446;0,975}\, s\sqrt{1+h_0} = 1{,}12 \pm 1{,}97\sqrt{0{,}077(1+0{,}0025)} = (0{,}57;\ 1{,}67)$$

es notablemente más impreciso que el intervalo de confianza calculado en el ejemplo anterior para el valor medio del colesterol HDL en todos los sujetos con dichos valores del índice de masa corporal y del consumo de alcohol (IC al 95% 1,09-1,15 mmol/l).

11.4 CONTRASTES DE HIPÓTESIS EN REGRESIÓN LINEAL MÚLTIPLE

Como se vio en el Apartado 10.3.2 del tema anterior, el contraste de un modelo de regresión lineal simple se reduce a evaluar si el coeficiente β_1 asociado a la única variable explicativa es 0, en cuyo caso el modelo no aportará explicación alguna sobre la variabilidad de la variable respuesta. En regresión lineal múltiple, sin embargo, la presencia de múltiples variables explicativas permite realizar distintos contrastes de hipótesis, que dan respuesta a diferentes preguntas de investigación. En general, los contrastes de hipótesis en regresión lineal múltiple pueden clasificarse en tres grandes grupos, a saber:

- El contraste global determina si el modelo en su conjunto explica una parte significativa de la variabilidad de la variable respuesta.
- Los contrastes parciales individuales evalúan la contribución independiente de cada variable explicativa una vez controlados los efectos de las restantes variables explicativas.
- Los contrastes parciales múltiples valoran si un determinado subgrupo de dos o más variables explicativas contribuye significativamente a explicar la variabilidad residual de la variable respuesta que no se explica por las otras variables incluidas en el modelo.

En los siguientes apartados se describen los procedimientos estadísticos necesarios para realizar dichos contrastes. Conviene resaltar que estos contrastes de hipótesis asumen linealidad y aditividad en los efectos de las variables explicativas y, en consecuencia, no deben interpretarse como pruebas de bondad del ajuste, ya que no facilitan ninguna información sobre la idoneidad del modelo lineal aditivo para describir la relación subyacente de las variables explicativas con la variable respuesta.

11.4.1 Contraste global del modelo de regresión lineal múltiple

La hipótesis nula del contraste global de un modelo de regresión lineal múltiple establece que ninguna de las variables explicativas se asocia linealmente con la variable respuesta, que puede formularse

como H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$. Bajo esta hipótesis nula, la ecuación de regresión se reduce al término constante β_0 y el modelo no aportará entonces ninguna explicación sobre la variabilidad de la variable respuesta. El propósito es, por tanto, contrastar la hipótesis nula H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ frente a la hipótesis alternativa bilateral de que al menos una de las variables explicativas se relaciona linealmente con la respuesta, que corresponde a H_1: $\beta_j \neq 0$ para algún $j = 1, \ldots, p$.

Al igual que en regresión lineal simple, este contraste global se realiza descomponiendo la variabilidad de la variable respuesta. Una vez estimada la ecuación de regresión $\hat{y} = b_0 + b_1x_1 + \ldots + b_px_p$, la **suma de cuadrados total** SST de la variable respuesta puede descomponerse como

$$\begin{aligned}\text{SST} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y} + y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) \\ &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \text{SSR} + \text{SSE},\end{aligned}$$

ya que las desviaciones $\hat{y}_i - \bar{y}$ y $y_i - \hat{y}_i$ están incorrelacionadas

$$\begin{aligned}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) &= \sum_{i=1}^{n}\hat{y}_i e_i - \bar{y}\sum_{i=1}^{n}e_i \\ &= b_0\sum_{i=1}^{n}e_i + \sum_{j=1}^{p}b_j\sum_{i=1}^{n}x_{ij}e_i - \bar{y}\sum_{i=1}^{n}e_i = 0\end{aligned}$$

de acuerdo a las ecuaciones lineales derivadas del método de mínimos cuadrados (véase Apartado 11.3.1). En consecuencia, la suma de cuadrados total SST se descompone en dos términos independientes: la **suma de cuadrados de la regresión** SSR, que representa la variabilidad de la variable respuesta explicada por el modelo de regresión, y la **suma de cuadrados del error** SSE, que representa la variabilidad residual que permanece sin explicar. Por un lado, la suma de cuadrados de la regresión SSR contiene p grados de libertad ya que, conocida la media muestral $\bar{y}$, los valores estimados por la ecuación de regresión $\hat{y}_i = b_0 + b_1x_{i1} + \ldots + b_px_{ip} = \bar{y} + b_1(x_{i1} - \bar{x}_1) + \ldots + b_p(x_{ip} - \bar{x}_p)$ quedan completamente determinados por los p coeficientes asociados a las variables explicativas. De hecho, puede probarse que el cociente SSR/σ^2 sigue una distribución chi-cuadrado con p grados de libertad cuando la hipótesis nula H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ es cierta. Por otro lado, como se vio en el Apartado 11.3.1, la suma de cuadrados del error SSE contiene $n - p - 1$ grados de libertad. Además, bajo las asunciones del modelo de regresión lineal múltiple, se comprueba que el cociente SSE/σ^2 se distribuye conforme a una chi-cuadrado con $n - p - 1$ grados de libertad con independencia de la hipótesis nula. Combinando las distribuciones muestrales de ambas sumas de cuadrados, se tiene que bajo la hipótesis nula H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ la razón entre la varianza explicada por la regresión SSR/p y la varianza residual $s^2 = \text{SSE}/(n - p - 1)$

$$F = \frac{\text{SSR}}{ps^2} = \frac{\dfrac{\text{SSR}}{p\sigma^2}}{\dfrac{\text{SSE}}{(n-p-1)\sigma^2}} \sim \frac{\chi^2_p / p}{\chi^2_{n-p-1}/(n-p-1)} = F_{p,n-p-1}$$

se distribuye como el cociente de dos distribuciones chi-cuadrado independientes divididas por sus correspondientes grados de libertad, que equivale a una distribución F de Fisher con p grados de libertad en el numerador y $n - p - 1$ en el denominador. La razón entre las varianzas

Tabla 11.2 Tabla genérica del análisis de la varianza en regresión lineal múltiple.*

	Suma de cuadrados	Grados de libertad	Varianza	Razón de varianzas
Regresión	$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	p	$\frac{\text{SSR}}{p}$	$F = \frac{\text{SSR}}{ps^2}$
Error	$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-p-1$	$s^2 = \frac{\text{SSE}}{n-p-1}$	
Total	$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		

* Coeficiente de determinación $R^2 = \text{SSR/SST}$.

explicada y residual constituye, por tanto, el estadístico para el contraste global del modelo de regresión lineal múltiple. La descomposición de la variabilidad de la variable respuesta, junto con la razón de varianzas resultante, suele resumirse en la **tabla del análisis de la varianza** (Tabla 11.2).

Como complemento al contraste global del modelo, suele calcularse el **coeficiente de determinación** $R^2 = \text{SSR/SST}$, que es una medida cuantitativa de la proporción de la variabilidad de la variable respuesta explicada por el modelo de regresión múltiple. El coeficiente de determinación R^2 varía entre 0 y 1 y aumenta siempre que se incluyen nuevas variables explicativas en el modelo, aunque este incremento puede no ser significativo (ver apartado siguiente). Otra de sus principales propiedades es que equivale al cuadrado del coeficiente de correlación $r_{y\hat{y}}$ entre los valores observados y_i de la variable respuesta y los valores predichos $\hat{y}_i$ por la ecuación de regresión, que se conoce como **coeficiente de correlación múltiple**,

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\left(\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}$$

$$= \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y}) - \sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}$$

$$= \frac{\left(\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} = r_{y\hat{y}}^2.$$

Notar que las estimaciones de los coeficientes de regresión minimizan la suma de cuadrados del error SSE y, en consecuencia, maximizan el coeficiente de determinación R^2 del modelo. De la relación entre los coeficientes de determinación y correlación múltiple, se deriva entonces que las estimaciones b_0, b_1, …, b_p maximizan la correlación entre los valores observados y_i y los

valores predichos $\hat{y}_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip}$, de tal forma que cualquier otra combinación lineal de las variables explicativas tendrá menor correlación con la variable respuesta.

Ejemplo 11.6 En la primera parte de la Tabla 11.1 se presenta el análisis de la varianza de la regresión lineal múltiple del colesterol HDL sobre el índice de masa corporal y el consumo de alcohol. La suma de cuadrados total del colesterol HDL

$$\text{SST} = \sum_{i=1}^{449} (y_i - 1{,}08)^2 = 38{,}91$$

se descompone en la suma de cuadrados explicada por la ecuación de regresión estimada $\hat{y} = 1{,}58 - 0{,}0207 x_1 + 0{,}0028 x_2$

$$\text{SSR} = \sum_{i=1}^{449} (1{,}58 - 0{,}0207 x_{i1} + 0{,}0028 x_{i2} - 1{,}08)^2 = 4{,}58$$

y la suma de cuadrados residual

$$\text{SSE} = \sum_{i=1}^{449} \{y_i - (1{,}58 - 0{,}0207 x_{i1} + 0{,}0028 x_{i2})\}^2 = 34{,}33.$$

Por tanto, el coeficiente de determinación se estima en $R^2 = 4{,}58/38{,}91 = 0{,}118$ y el coeficiente de correlación múltiple en $r_{y\hat{y}} = \sqrt{0{,}118} = 0{,}343$. Es decir, la combinación lineal del índice de masa corporal y el consumo de alcohol presenta una correlación de 0,343 con el colesterol HDL, consiguiendo así explicar el 11,8% de la variabilidad del colesterol HDL en los controles del estudio EURAMIC. Esta variabilidad explicada por el modelo de regresión lineal múltiple representa una parte significativa de la variabilidad total del colesterol HDL, ya que el contraste global del modelo mediante la razón entre las varianzas explicada y residual

$$F = \frac{4{,}58/2}{34{,}33/446} = \frac{2{,}29}{0{,}077} = 29{,}72$$

resulta en un valor $P = P(F_{2,446} \geq 29{,}72) < 0{,}001$ bajo la distribución F de Fisher con 2 grados de libertad en el numerador y 446 en el denominador.

11.4.2 Contrastes parciales

Cuando el contraste global de regresión es significativo, el modelo en su conjunto resulta efectivo a la hora de explicar la variabilidad observada en la variable respuesta. No obstante, esto no implica necesariamente que todas las variables explicativas incluidas en el modelo contribuyan de forma significativa a explicar una parte de la variabilidad de la respuesta, pudiendo haber una o varias variables que tengan nula o escasa contribución. En este sentido, cabría preguntarse si es posible eliminar algunas variables explicativas del modelo sin afectar sensiblemente a la capacidad predictiva del mismo. Los contrastes parciales se ocupan de dar respuesta a este tipo de preguntas, valorando la contribución adicional de una o más variables explicativas a lo ya explicado por las otras variables presentes en el modelo.

La hipótesis nula del contraste parcial establece que, una vez incluidas las variables explicativas $X_1, \ldots, X_{p-r}$, $1 \leq r < p$, las restantes r variables $X_{p-r+1}, \ldots, X_p$ del modelo no se relacionan linealmente con la variable respuesta. Más concretamente, se pretende contrastar la hipótesis nula H_0: $\beta_{p-r+1} = \ldots = \beta_p = 0$ frente a la hipótesis alternativa bilateral H_1: $\beta_j \neq 0$, para algún $j = p - r + 1, \ldots, p$, en el modelo de regresión lineal múltiple

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-r} x_{p-r} + \beta_{p-r+1} x_{p-r+1} + \ldots + \beta_p x_p + \varepsilon.$$

Notar que este contraste parcial es equivalente a la comparación de dos modelos: el anterior **modelo completo** que incorpora las p variables explicativas y el **modelo reducido** que resulta de excluir las r variables X_{p-r+1}, ..., X_p objeto del contraste,

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-r} x_{p-r} + \varepsilon,$$

dado que los coeficientes asociados a dichas variables son 0 bajo la hipótesis nula. Así, los contrastes parciales son particularmente útiles para comparar el ajuste de dos modelos anidados, lo que permite decantarse entre el modelo más simple o el modelo extendido con variables adicionales en función del resultado del contraste.

El procedimiento más sencillo para realizar un contraste parcial es ajustar por separado el modelo completo y el modelo reducido excluyendo las r variables explicativas sometidas al contraste, asegurándose de utilizar las mismas observaciones en ambos modelos. Al incluir nuevas variables explicativas sobre la misma muestra de observaciones, la variabilidad de la variable respuesta explicada por el modelo completo SSR_1 será siempre mayor o igual que la variabilidad explicada por el modelo reducido SSR_0, de tal forma que la diferencia $\text{SSR}_1 - \text{SSR}_0$ representa el incremento en la variabilidad explicada al incluir las variables X_{p-r+1}, ..., X_p. Puede probarse que, si la hipótesis nula H_0: $\beta_{p-r+1} = \ldots = \beta_p = 0$ es cierta, el cociente $(\text{SSR}_1 - \text{SSR}_0)/\sigma^2$ sigue una distribución chi-cuadrado con los r grados de libertad correspondientes al número de variables explicativas a contrastar. Asimismo, la suma de cuadrados del error del modelo completo SSE_1 es independiente del incremento en la variabilidad explicada $\text{SSR}_1 - \text{SSR}_0$ y el cociente SSE_1/σ^2 se distribuye según una chi-cuadrado con $n-p-1$ grados de libertad. De estos resultados se deriva que, bajo H_0: $\beta_{p-r+1} = \ldots = \beta_p = 0$, la razón entre el incremento de la varianza explicada por ambos modelos $(\text{SSR}_1 - \text{SSR}_0)/r$ y la varianza residual del modelo completo $s_1^2 = \text{SSE}_1/(n-p-1)$

$$F = \frac{\text{SSR}_1 - \text{SSR}_0}{r s_1^2} = \frac{\dfrac{\text{SSR}_1 - \text{SSR}_0}{r\sigma^2}}{\dfrac{\text{SSE}_1}{(n-p-1)\sigma^2}} \sim \frac{\chi_r^2 / r}{\chi_{n-p-1}^2/(n-p-1)} = F_{r,n-p-1}$$

sigue una distribución F de Fisher con r y $n-p-1$ grados de libertad al ser el cociente de dos distribuciones chi-cuadrado independientes divididas por sus respectivos grados de libertad. Este análisis de la varianza para el contraste parcial de un modelo de regresión lineal múltiple se representa esquemáticamente en la Tabla 11.3.

Tabla 11.3 Análisis de la varianza para el contraste parcial en regresión lineal múltiple.

	Suma de cuadrados	Grados de libertad	Varianza	Razón de varianzas
Regresión	SSR_1	p		
$X_1,\ldots, X_{p-r}$	SSR_0	$p-r$		
$X_{p-r+1},\ldots, X_p \mid X_1,\ldots, X_{p-r}$	$\text{SSR}_1 - \text{SSR}_0$	r	$\dfrac{\text{SSR}_1 - \text{SSR}_0}{r}$	$F = \dfrac{\text{SSR}_1 - \text{SSR}_0}{r s_1^2}$
Error	SSE_1	$n-p-1$	$s_1^2 = \dfrac{\text{SSE}_1}{n-p-1}$	
Total	SST	$n-1$		

Ejemplo 11.7 La Tabla 11.4 muestra los resultados obtenidos en el grupo control del estudio EURAMIC al ajustar un modelo de regresión lineal múltiple con el colesterol HDL como variable respuesta, el índice de masa corporal, el consumo de alcohol y la edad en años como variables explicativas continuas y el estatus socioeconómico como variable explicativa dicotómica ($x_{i4} = 1$ en sujetos con bajo nivel socioeconómico y 0 en sujetos con alto nivel socioeconómico). De la tabla del análisis de la varianza se desprende que el modelo en su conjunto explica el 11,9% de la variabilidad del colesterol HDL, lo que representa una parte significativa de la variabilidad total de la respuesta ya que la razón de varianzas del contraste global del modelo $F = 14{,}85$ resulta en un valor $P = P(F_{4,440} \geq 14{,}85) < 0{,}001$ bajo la distribución F de Fisher con 4 y 440 grados de libertad. No obstante, una vez incluidos el índice de masa corporal y la ingesta de alcohol, ni la edad ($t = b_3/SE(b_3) = 0{,}0002/0{,}0014 = 0{,}12$, $P = 2P(t_{440} \geq 0{,}12) \approx 2\{1 - \Phi(0{,}12)\} = 0{,}90$) ni el estatus socioeconómico ($t = b_4/SE(b_4) = 0{,}021/0{,}027 = 0{,}80$, $P = 2P(t_{440} \geq 0{,}80) \approx 2\{1 - \Phi(0{,}80)\} = 0{,}43$) presentan efectos independientes significativos sobre los niveles de colesterol HDL. De hecho, cada incremento de 10 años en la edad se asocia con un aumento despreciable de $10 \cdot 0{,}0002 = 0{,}002$ mmol/l en la media del colesterol HDL entre sujetos con igual índice de masa corporal, consumo de alcohol y nivel socioeconómico. De igual forma, ajustando por diferencias en el índice de masa corporal, la ingesta de alcohol y la edad, la media del colesterol HDL difiere únicamente en 0,021 mmol/l entre los sujetos con nivel socioeconómico bajo y alto.

A partir de estos resultados, sería razonable preguntarse si la edad y el estatus socioeconómico contribuyen conjuntamente a explicar la variabilidad residual del colesterol HDL que permanece sin explicar por el índice de masa corporal y el consumo de alcohol, lo que equivale a contrastar este modelo frente al modelo reducido de la Tabla 11.1 que incluye únicamente el índice de masa corporal y la ingesta de alcohol como variables explicativas. No obstante, los resultados de ambos modelos no son

Tabla 11.4 Resultados de la regresión lineal múltiple del colesterol HDL sobre el índice de masa corporal (IMC), el consumo de alcohol, la edad y el estatus socioeconómico (ESE) en el grupo control del estudio EURAMIC.

Análisis de la varianza*

	Suma de cuadrados	Grados de libertad	Varianza	Razón de varianzas
Regresión	4,58	4	1,14	14,85
Error	33,93	440	0,077	
Total	38,51	444		

* Coeficiente de determinación $R^2 = 4{,}58/38{,}51 = 0{,}119$.

Coeficientes de regresión

				Test H_0: $\beta_j = 0$	
	Estimación	Error estándar	IC al 95%	t	Valor P
Constante	1,56	0,12	(1,33; 1,79)	13,24	< 0,001
IMC	−0,021	0,0037	(−0,028; −0,014)	−5,66	< 0,001
Alcohol	0,0028	0,0006	(0,0016; 0,0040)	4,64	< 0,001
Edad	0,0002	0,0014	(−0,0026; 0,0030)	0,12	0,90
ESE	0,021	0,027	(−0,031; 0,074)	0,80	0,43

Tabla 11.5 Análisis de la varianza para el contraste parcial múltiple de la edad y el estatus socioeconómico (ESE) en la regresión lineal del colesterol HDL sobre el índice de masa corporal (IMC), el consumo de alcohol, la edad y el ESE en el grupo control del estudio EURAMIC.

	Suma de cuadrados	Grados de libertad	Varianza	Razón de varianzas
Regresión	4,58	4		
IMC, alcohol	4,53	2		
Edad, ESE|IMC, alcohol	0,053	2	0,026	0,34
Error	33,93	440	0,077	
Total	38,51	444		

directamente comparables ya que el modelo reducido emplea 4 observaciones más que el modelo completo (449 versus 445). Esto es debido a que hay 4 sujetos con valores ausentes para el estatus socioeconómico, que pueden utilizarse en el ajuste del modelo reducido, pero no en el modelo completo que incluye dicha variable. Para comparar ambos modelos, es preciso ajustar el modelo reducido a la misma muestra de 445 controles del estudio EURAMIC, de donde se obtiene una suma de cuadrados explicada por el modelo reducido de $SSR_0 = 4{,}53$. Así, el incremento en la variabilidad explicada al incluir la edad y el estatus socioeconómico en el modelo completo es $SSR_1 - SSR_0 = 4{,}58 - 4{,}53 = 0{,}053$. La razón entre el incremento de la varianza explicada y la varianza residual del modelo completo es entonces

$$F = \frac{0{,}053/2}{33{,}93/440} = \frac{0{,}026}{0{,}077} = 0{,}34,$$

que corresponde a un valor $P = P(F_{2,440} \geq 0{,}34) = 0{,}71$ bajo la distribución F de Fisher con 2 y 440 grados de libertad. Este contraste parcial múltiple se representa en la Tabla 11.5. En conclusión, la edad y el estatus socioeconómico no contribuyen significativamente a explicar la variabilidad del colesterol HDL una vez tenidos en cuenta el índice de masa corporal y el consumo de alcohol, de tal forma que el modelo reducido a estas dos últimas variables explicativas resulta igualmente efectivo.

Los contrastes parciales pueden emplearse para evaluar la contribución adicional de una única variable explicativa o de múltiples variables explicativas. El contraste parcial individual de la variable explicativa X_j se reduce a evaluar la hipótesis nula H_0: $\beta_j = 0$ frente a la hipótesis alternativa H_1: $\beta_j \neq 0$ y, en consecuencia, es equivalente al test para los coeficientes de regresión presentado en el Apartado 11.3.2. De hecho, puede probarse que el estadístico F de la razón de varianzas del contraste parcial individual es igual al cuadrado del estadístico $t = b_j/SE(b_j)$ del correspondiente coeficiente, de tal forma que los valores P resultantes de ambos procedimientos son idénticos (la distribución F de Fisher con 1 grado de libertad en el numerador y $n - p - 1$ en el denominador es, por definición, el cuadrado de la distribución t de Student con $n - p - 1$ grados de libertad).

Ejemplo 11.8 Para evaluar si el estatus socioeconómico contribuye a explicar la variabilidad del colesterol HDL que no se explica por las diferencias de índice de masa corporal, consumo de alcohol y edad, se podría comparar la variabilidad explicada por el modelo completo con la variabilidad explicada por el modelo que excluye el estatus

socioeconómico en la misma muestra de 445 controles, obteniéndose una diferencia $SSR_1 - SSR_0 = 4{,}58 - 4{,}53 = 0{,}049$. Así, el estadístico F del contraste parcial individual es

$$F = \frac{0{,}049}{33{,}93/440} = \frac{0{,}049}{0{,}077} = 0{,}64,$$

que corresponde a un valor $P = P(F_{1,440} \geq 0{,}64) = 0{,}43$ bajo la distribución F de Fisher con 1 y 440 grados de libertad. Notar que este contraste es equivalente al test del coeficiente asociado al estatus socioeconómico en la Tabla 11.4 ya que $2P(t_{440} \geq 0{,}80) = P(t^2_{440} \geq 0{,}80^2) = P(F_{1,440} \geq 0{,}64)$.

11.5 VARIABLES EXPLICATIVAS POLITÓMICAS

La regresión lineal no establece ninguna asunción respecto a la distribución de las variables explicativas, que pueden ser tanto continuas como categóricas. En anteriores apartados, se ha tratado con modelos de regresión lineal que incorporan variables explicativas continuas y dicotómicas. Queda pendiente de estudiar, por tanto, el ajuste e interpretación de modelos de regresión lineal múltiple con variables explicativas politómicas, que clasifican a los sujetos en tres o más categorías en función de sus distintas características. Estas variables politómicas pueden ser nominales (nunca fumadores, ex fumadores o fumadores actuales), ordinales (nivel socioeconómico bajo, medio o alto) o incluso variables continuas categorizadas (normopeso, sobrepeso u obesidad para un índice de masa corporal < 25, 25-30 ó ≥ 30 kg/m^2, respectivamente).

En general, las variables explicativas politómicas no se introducen directamente en los modelos de regresión ya que los valores asignados a estas variables sólo sirven para discernir u ordenar las distintas categorías, pero no tienen interpretación numérica. La forma adecuada de incluir este tipo de variables explicativas en una regresión es mediante **variables indicadoras** que identifiquen cada una de las categorías de la variable. Existen diversos métodos para codificar adecuadamente variables indicadoras. La elección entre uno u otro procedimiento de codificación no afecta al ajuste del modelo (la tabla del análisis de la varianza permanece inalterable ante cualquier codificación que permita diferenciar todas las categorías de una variable politómica), pero sí a las estimaciones e interpretación de los coeficientes asociados a las variables indicadoras. En este apartado se presenta la **codificación de la categoría de referencia**, que es el método más extendido para definir variables indicadoras, de fácil interpretación y válido para cualquier tipo de variable politómica. Para cada una de las k categorías $j = 1, \ldots, k$ de la variable politómica, se define la variable indicadora $X_j = 1$ en los sujetos pertenecientes a la categoría j y 0 en los restantes sujetos, tal como se indica en la Tabla 11.6. Estas variables indicadoras $X_1, \ldots, X_k$ no pueden incluirse simultáneamente en un modelo de regresión que contenga el término constante, ya que su suma $X_1 + \ldots + X_k = 1$ para todos los sujetos y cualquier variable indicadora puede expresarse entonces como una combinación lineal exacta de la constante y de las demás variables indicadoras, con lo que el modelo incurriría en un problema de colinealidad perfecta (véase Ejemplo 11.1). Para solventar este problema, basta con excluir una cualquiera de las variables indicadoras, digamos X_k, manteniendo en el modelo las otras variables indicadoras $X_1, \ldots, X_{k-1}$,

$$E(Y|x_1, \ldots, x_{k-1}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1},$$

donde por simplicidad se omiten otras posibles variables explicativas. En este modelo, la constante β_0 corresponde al valor esperado de la respuesta en la categoría k de la variable politómica, que toma valores cero en todas las variables indicadoras incluidas en el modelo, $E(Y|x_1 = 0, \ldots, x_{k-1} = 0) = \beta_0$. Asimismo, cada coeficiente de regresión β_j determina el cambio en

Tabla 11.6 Variables indicadoras para las *k* categorías de una variable politómica.

	Variable indicadora			
Categoría	X_1	X_2	…	X_k
1	1	0	…	0
2	0	1	…	0
⋮	⋮	⋮		⋮
k	0	0	…	1

el valor esperado de la respuesta en la categoría $j = 1, \ldots, k-1$ respecto a la categoría k de la variable politómica,

$$E(Y|x_1 = 0, ..., x_{j-1} = 0, x_j = 1, x_{j+1} = 0, ..., x_{k-1} = 0) - E(Y|x_1 = 0, ..., x_{k-1} = 0) = \beta_0 + \beta_j - \beta_0 = \beta_j.$$

Como puede apreciarse, la categoría cuya variable indicadora se deja fuera del modelo actúa como grupo de referencia, de tal forma que los coeficientes asociados a las variables indicadoras presentes en el modelo determinan los cambios medios en la respuesta respecto a dicha categoría de referencia. Aunque en principio la elección del grupo de referencia es arbitraria, en la práctica suele utilizarse como categoría de referencia aquella que representa la ausencia o el menor nivel de exposición (nunca fumadores, nivel socioeconómico alto, normopeso), siempre y cuando su tamaño muestral sea lo suficientemente grande para obtener comparaciones precisas con el resto de categorías de la variable politómica.

En general, la contribución de las variables indicadoras a la capacidad predictiva del modelo debe evaluarse conjuntamente, dado que estas variables no representan más que las distintas categorías de una misma variable politómica. En este sentido, los contrastes parciales presentados en el apartado anterior pueden aplicarse al conjunto de todas las variables indicadoras para contrastar la hipótesis nula H_0: $\beta_1 = \ldots = \beta_{k-1} = 0$, lo que equivale a un **test de homogeneidad** del valor medio de la respuesta en las k categorías de la variable politómica. Notar que este test de homogeneidad permanece inalterable ante cualquier codificación de las variables indicadoras o selección del grupo de referencia, ya que éstas alteran los coeficientes de regresión, pero no cambian la contribución global de la variable politómica al ajuste del modelo.

Ejemplo 11.9 En la Tabla 11.7 se presentan los resultados de ajustar un modelo de regresión lineal múltiple del colesterol HDL sobre el índice de masa corporal, el consumo de alcohol y el hábito tabáquico en 448 controles del estudio EURAMIC con información completa de estas variables. El hábito tabáquico es una variable politómica con tres categorías, que diferencia a los sujetos según sean nunca fumadores (113 sujetos), ex fumadores (163) o fumadores actuales (172). Se designa como categoría de referencia a los nunca fumadores y, en consecuencia, el modelo incluye dos variables indicadoras para los ex fumadores ($x_{i3} = 1$ en ex fumadores y 0 en el resto) y los fumadores actuales ($x_{i4} = 1$ en fumadores actuales y 0 en el resto).

Para evaluar si el nivel medio de colesterol HDL difiere en las tres categorías del hábito tabáquico una vez tenidas en cuenta las diferencias de índice de masa corporal y consumo de alcohol, se realiza el contraste parcial múltiple de las dos variables indicadoras del hábito tabáquico H_0: $\beta_3 = \beta_4 = 0$. Para ello, se compara la variabilidad explicada $SSR_1 = 5{,}44$ por el

modelo completo de la Tabla 11.7 con la variabilidad explicada $SSR_0 = 4{,}58$ por el modelo que excluye ambas variables indicadoras en la misma muestra de 448 controles, obteniéndose un test estadístico

$$F = \frac{(5{,}44 - 4{,}58)/2}{33{,}42/443} = \frac{0{,}43}{0{,}075} = 5{,}69,$$

que corresponde a un valor $P = P(F_{2,443} \geq 5{,}69) = 0{,}004$ bajo la distribución F de Fisher con 2 y 443 grados de libertad. Así, se detectan diferencias significativas en las medias ajustadas del colesterol HDL entre los nunca fumadores, ex fumadores y fumadores actuales. Los coeficientes asociados a las variables indicadoras del hábito tabáquico permiten cuantificar estas diferencias de acuerdo a la codificación elegida. Por un lado, una vez controladas las diferencias en el índice de masa corporal y la ingesta de alcohol, la media del colesterol HDL presenta una diferencia insignificante de $b_3 = 0{,}009$ mmol/l entre los ex fumadores y los nunca fumadores. Sin embargo, los fumadores actuales presentan una disminución significativa en el nivel medio de colesterol HDL de $b_4 = -0{,}085$ mmol/l en comparación con los nunca fumadores, incluso después de ajustar por el índice de masa corporal y el consumo de alcohol.

En general, las variables indicadoras deben tratarse conjuntamente para preservar su interpretación. No obstante, en vista de que los niveles medios de colesterol HDL no difieren en nunca fumadores y ex fumadores, se podría eliminar del modelo la variable indicadora de los ex fumadores. En tal caso, el coeficiente asociado a la variable indicadora de los fumadores actuales cambiaría de interpretación, pasando a representar el cambio medio en el colesterol HDL entre fumadores actuales y no fumadores actuales (nueva categoría de referencia donde se englobarían tanto los nunca como los ex fumadores).

El test de homogeneidad permite contrastar si el nivel medio de la respuesta difiere significativamente en al menos 2 de las k categorías de una variable explicativa politómica. En el caso de que las categorías estén intrínsecamente ordenadas, como ocurre con las variables

Tabla 11.7 Resultados de la regresión lineal múltiple del colesterol HDL sobre el índice de masa corporal (IMC), el consumo de alcohol y las variables indicadoras de ex fumadores y fumadores actuales en el grupo control del estudio EURAMIC.

Análisis de la varianza*

	Suma de cuadrados	Grados de libertad	Varianza	Razón de varianzas
Regresión	5,44	4	1,36	18,03
Error	33,42	443	0,075	
Total	38,86	447		

* Coeficiente de determinación $R^2 = 5{,}44/38{,}86 = 0{,}140$.

Coeficientes de regresión

				Test H_0: $\beta_j = 0$	
	Estimación	Error estándar	IC al 95%	t	Valor P
Constante	1,61	0,099	(1,42; 1,81)	16,31	< 0,001
IMC	−0,021	0,0036	(−0,028; −0,014)	−5,79	< 0,001
Alcohol	0,0030	0,0006	(0,0018; 0,0042)	5,03	< 0,001
Ex fumador	0,009	0,034	(−0,058; 0,075)	0,26	0,80
Fumador actual	−0,085	0,034	(−0,151; −0,019)	−2,53	0,012

ordinales y las variables continuas categorizadas, cabría preguntarse además si los niveles medios de la respuesta siguen algún patrón específico a lo largo de las categorías. En particular, sería relevante contar con un **test de tendencia** que permitiera detectar la existencia de una componente lineal creciente o decreciente entre las respuestas medias de las sucesivas categorías. Para ello, la variable explicativa politómica X debe tomar valores que preserven el orden de las categorías. En el caso de variables ordinales, suelen asignarse los valores $x_i = 1, 2, ..., k$ según el sujeto pertenezca a la primera, segunda o sucesivas categorías. En el caso de variables continuas categorizadas, es preferible utilizar valores x_i que representen alguna medida de tendencia central de cada categoría (media o mediana) para preservar no sólo el orden de las categorías, sino también la distancia entre las mismas. La variable politómica así codificada se incluye directamente en el modelo de regresión, de tal forma que el contraste de su coeficiente determina la existencia de una tendencia lineal creciente o decreciente en el valor medio de la respuesta al aumentar la categoría de exposición. Conviene resaltar que este test de tendencia no permite evaluar la idoneidad de la relación lineal, sino únicamente la existencia de una componente lineal significativa a través de las categorías, independientemente de cuál sea la relación subyacente.

Ejemplo 11.10 Dado que en el ejemplo anterior los niveles medios de colesterol HDL no diferían significativamente en nunca fumadores y ex fumadores, ambas categorías se colapsaron en una única categoría de no fumadores actuales. Además, como se dispone de información sobre el número de cigarrillos al día en 154 de los 172 fumadores actuales, se construyó una nueva variable politómica que clasificaba a los sujetos en no fumadores actuales (276 sujetos), fumadores actuales de 1-10 (50 sujetos), 11-20 (67 sujetos) y > 20 cigarrillos/día (37 sujetos). La Tabla 11.8 muestra los resultados obtenidos en los controles del estudio EURAMIC al ajustar una regresión lineal múltiple del colesterol HDL sobre el índice de masa corporal, el consumo de alcohol y esta nueva variable explicativa politómica, donde los no fumadores actuales constituyen la categoría de referencia.

Tabla 11.8 Resultados de la regresión lineal múltiple del colesterol HDL sobre el índice de masa corporal (IMC), la ingesta de alcohol y las variables indicadoras de fumadores actuales de 1-10, 11-20 y > 20 cigarrillos/día en los controles del estudio EURAMIC.

Análisis de la varianza*

	Suma de cuadrados	Grados de libertad	Varianza	Razón de varianzas
Regresión	4,70	5	0,94	12,62
Error	31,59	424	0,075	
Total	36,29	429		

* Coeficiente de determinación $R^2 = 4{,}70/36{,}29 = 0{,}130$.

Coeficientes de regresión

				Test H_0: $\beta_j = 0$	
	Estimación	Error estándar	IC al 95%	t	Valor P
Constante	1,59	0,10	(1,40; 1,79)	15,90	< 0,001
IMC	−0,020	0,0037	(−0,027; −0,013)	−5,36	< 0,001
Alcohol	0,0028	0,0006	(0,0017; 0,0040)	4,70	< 0,001
Fumador 1-10	−0,086	0,042	(−0,168; −0,003)	−2,04	0,042
Fumador 11-20	−0,120	0,038	(−0,193; −0,046)	−3,19	0,002
Fumador > 20	−0,055	0,048	(−0,149; 0,040)	−1,14	0,26

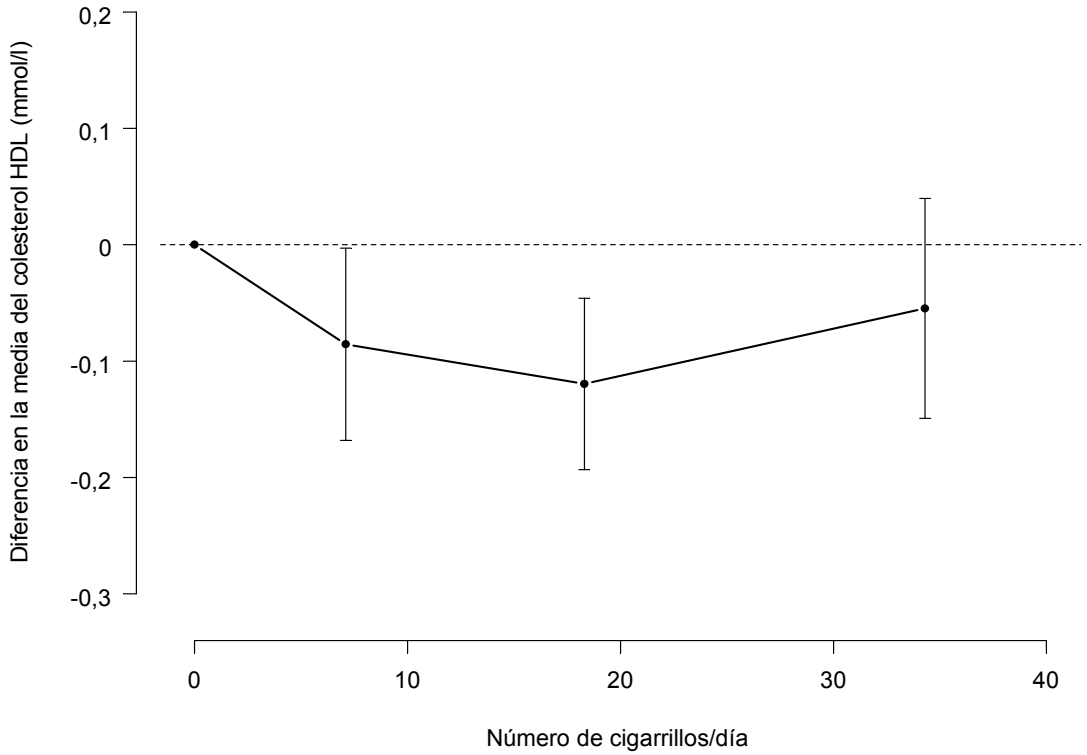


Figura 11.4 **Diferencia en la media ajustada del colesterol HDL de los fumadores actuales de 1-10, 11-20 y > 20 cigarrillos/día respecto a los no fumadores actuales del grupo control del estudio EURAMIC. Las barras verticales representan los intervalos de confianza al 95% para estas diferencias.**

El contraste parcial múltiple de las tres variables indicadoras H_0: $\beta_3 = \beta_4 = \beta_5 = 0$ revela que existen diferencias significativas en las medias ajustadas del colesterol HDL entre los no fumadores actuales y los fumadores de 1-10, 11-20 y > 20 cigarrillos/día, ya que la comparación de la variabilidad explicada $SSR_1 = 4{,}70$ por el modelo completo de la Tabla 11.8 y la variabilidad explicada $SSR_0 = 3{,}76$ por el modelo que excluye las tres variables indicadoras en la misma muestra de 430 controles resulta en un test estadístico

$$F = \frac{(4{,}70 - 3{,}76)/3}{31{,}59/424} = \frac{0{,}31}{0{,}075} = 4{,}22,$$

que corresponde a un valor $P = P(F_{3,424} \geq 4{,}22) = 0{,}006$. En comparación con los no fumadores actuales de igual índice de masa corporal y consumo de alcohol, los fumadores de 1-10, 11-20 y > 20 cigarrillos/día presentan una disminución en el nivel medio de colesterol HDL de $b_3 = -0{,}086$, $b_4 = -0{,}120$ y $b_5 = -0{,}055$ mmol/l, respectivamente. Esta tendencia decreciente en la media ajustada del colesterol HDL se representa en la Figura 11.4, donde el eje horizontal corresponde al número medio de cigarrillos diarios para cada categoría (0 en el caso de no fumadores actuales).

Para contrastar si esta tendencia decreciente es significativa, se crea una variable politómica con valores x_i = 0, 7,1, 18,3 y 34,3 correspondientes al número medio de cigarrillos diarios de los sujetos no fumadores y fumadores de 1-10, 11-20 y > 20 cigarrillos/día, respectivamente. Esta variable politómica se incluye directamente en un modelo de regresión múltiple junto con el índice de masa corporal y la ingesta de alcohol. El coeficiente asociado a la variable politómica y su error estándar se estiman en $b_3 = -0{,}0030$ y $SE(b_3) = 0{,}0012$, de donde se obtiene un estadístico $t = b_3/SE(b_3) = -0{,}0030/0{,}0012 = -2{,}46$ y un valor $P = 2P(t_{426} \leq -2{,}46) \approx 2\Phi(-2{,}46) = 0{,}014$ bajo la distribución t de

Student con $n - p - 1 = 430 - 3 - 1 = 426$ grados de libertad. Así, puede concluirse que la media ajustada del colesterol HDL no sólo difiere entre las categorías (P de homogeneidad = 0,006), sino que tiende a decrecer significativamente conforme aumenta la categoría de exposición (P de tendencia = 0,014). No obstante, la Figura 11.4 muestra que la relación subyacente podría no ser estrictamente lineal al presentar un leve repunte en la categoría de fumadores de más de 20 cigarrillos/día.

11.6 REGRESIÓN POLINOMIAL

La regresión lineal múltiple permite explorar relaciones no lineales entre las variables explicativas y la variable respuesta. El modelo más habitual para acomodar un efecto no lineal de una variable explicativa continua X es la regresión polinomial de orden k, que incorpora en el modelo los términos polinomiales X^2, …, X^k además del propio término lineal X,

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \varepsilon,$$

donde sin pérdida de generalidad se omiten otras posibles variables explicativas. Estos modelos polinomiales pueden considerarse como casos particulares de la regresión lineal múltiple cuyas variables explicativas son distintas potencias de una misma variable básica y, en consecuencia, los procedimientos de estimación e inferencia son idénticos a los descritos anteriormente para el modelo general de regresión.

En teoría, los modelos polinomiales de orden k elevado permiten aproximar cualquier tipo de relación curvilínea. No obstante, si el número requerido de términos polinomiales es muy elevado, la regresión polinomial puede ocasionar problemas de sobreajuste y dar lugar a estimaciones inestables de los coeficientes de regresión. Los polinomios de orden superior al cuadrático tienden además a producir curvas con puntos de inflexión y otras formas extrañas de difícil interpretación en términos epidemiológicos. Por ello, esta presentación se limita a los **modelos polinomiales de segundo orden o cuadráticos**, que incluyen un término lineal X y otro cuadrático X^2 de la variable explicativa. La tendencia resultante de estos modelos cuadráticos será una parábola que, aunque no se amolda a cualquier forma subyacente de la relación, sí permite capturar las desviaciones más frecuentes del modelo lineal, incluyendo tendencias monótonas cuya pendiente aumenta o disminuye progresivamente, así como curvas en forma de U o de U invertida con un cambio de dirección.

Aunque los modelos cuadráticos se ajustan mediante los métodos estándar de regresión múltiple, las variables X y X^2 están a menudo muy correlacionadas (típicamente, $r_{xx^2} > 0{,}95$), provocando estimaciones inestables de sus coeficientes de regresión. Para mitigar este problema de colinealidad, conviene centrar primero la variable original X e incluir después dicha variable centrada y su cuadrado en el modelo de regresión,

$$Y = \beta_0 + \beta_1 (x - \bar{x}) + \beta_2 (x - \bar{x})^2 + \varepsilon.$$

Las desviaciones respecto de la media $x - \bar{x}$ y sus cuadrados $(x - \bar{x})^2$ estarán menos correlacionadas que los valores x y x^2, ya que los cuadrados de las desviaciones $(x - \bar{x})^2$ serán elevados tanto para valores altos de X (desviaciones $x - \bar{x}$ positivas) como para valores bajos (desviaciones $x - \bar{x}$ negativas). El centrado de la variable explicativa X no afecta al ajuste global del modelo cuadrático ni a la tendencia parabólica resultante, se trata únicamente de una reparametrización del modelo que reduce la correlación entre el término lineal y cuadrático, produciendo así estimaciones más estables de sus coeficientes y contrastes más fácilmente interpretables.

Una vez ajustado el modelo cuadrático, el primer paso es contrastar si el coeficiente β_2 asociado al término cuadrático es 0. Si este coeficiente no difiere significativamente del valor

nulo, la inclusión del término cuadrático no mejorará significativamente la capacidad predictiva del modelo, de tal forma que podrá eliminarse dicho término cuadrático y volver al modelo lineal en la variable explicativa X. Por el contrario, si el coeficiente del término cuadrático resulta significativo, el modelo cuadrático presentará un mejor ajuste que el modelo lineal, debiendo mantener ambos términos lineal y cuadrático en el modelo. La interpretación del modelo cuadrático no es tan sencilla como la del modelo lineal, ya que la pendiente de la relación varía a lo largo del rango de la variable explicativa. En un modelo cuadrático con la variable X centrada, la pendiente de la relación viene dada por $\beta_1 + 2\beta_2(x - \bar{x})$; es decir, β_1 corresponde a la pendiente en la media $\bar{x}$ de la variable explicativa y $2\beta_2$ representa el cambio de pendiente por cada incremento de una unidad en X. No obstante, el interés no es tanto interpretar los coeficientes individuales, sino representar gráficamente la tendencia global resultante del modelo cuadrático.

Ejemplo 11.11 En la Figura 10.10(*b*) del tema anterior, el análisis de los residuos de la regresión lineal simple del colesterol HDL sobre el índice de masa corporal en los controles del estudio EURAMIC mostró indicios de una posible relación cuadrática entre ambas variables. Para contrastar formalmente esta tendencia, se ajustó un modelo de regresión múltiple para el colesterol HDL que incluía un término lineal y otro cuadrático del índice de masa corporal, además del consumo de alcohol y de la variable indicadora de los fumadores actuales (Tabla 11.9). Como el índice de masa corporal X_1 y su cuadrado X_1^2 presentaban una correlación lineal casi perfecta de 0,995, esta variable fue previamente centrada alrededor de su media muestral $\bar{x}_1 = 26{,}2$ kg/m^2 antes de incluir en el modelo los términos lineal $X_1 - 26{,}2$ y cuadrático $(X_1 - 26{,}2)^2$, cuya correlación era únicamente de 0,297.

El contraste para la nulidad del coeficiente asociado al término cuadrático del índice de masa corporal resulta en un valor $P = 0{,}021$, lo que indica que el modelo cuadrático mejora

Tabla 11.9 Resultados de la regresión múltiple del colesterol HDL sobre los términos lineal y cuadrático del índice de masa corporal (IMC), el consumo de alcohol y la variable indicadora de fumadores actuales en el grupo control del estudio EURAMIC.

Análisis de la varianza*

	Suma de cuadrados	Grados de libertad	Varianza	Razón de varianzas
Regresión	5,84	4	1,46	19,57
Error	33,02	443	0,075	
Total	38,86	447		

* Coeficiente de determinación $R^2 = 5{,}84/38{,}86 = 0{,}150$.

Coeficientes de regresión

				Test H_0: $\beta_j = 0$	
	Estimación	Error estándar	IC al 95%	*t*	Valor *P*
Constante	1,05	0,020	(1,01; 1,09)	52,62	< 0,001
IMC – 26,2	–0,024	0,0038	(–0,031; –0,016)	–6,25	< 0,001
(IMC – 26,2)^2	0,0016	0,0007	(0,0002; 0,0029)	2,32	0,021
Alcohol	0,0030	0,0006	(0,0018; 0,0042)	5,00	< 0,001
Fumador actual	–0,098	0,027	(–0,150; –0,045)	–3,63	< 0,001

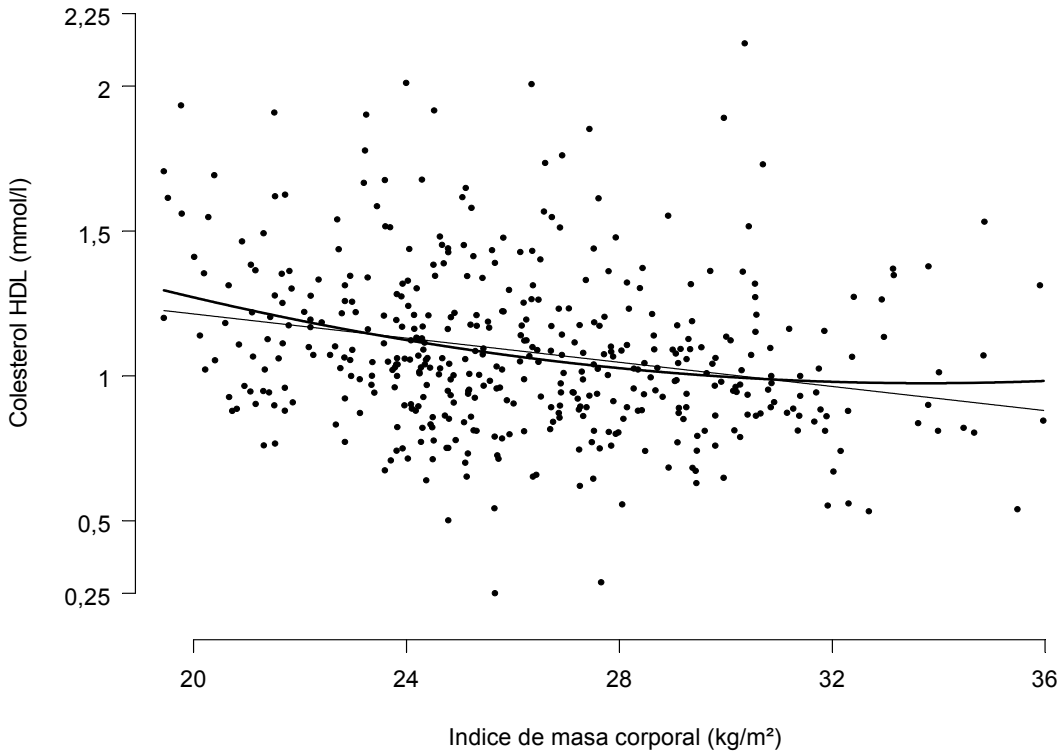


Figura 11.5 **Relación lineal (línea fina) y cuadrática (curva gruesa) entre el índice de masa corporal y el colesterol HDL obtenidas de modelos de regresión múltiple ajustados por consumo de alcohol y hábito tabáquico actual en el grupo control del estudio EURAMIC.**

significativamente el ajuste del modelo lineal. En consecuencia, la pendiente de la relación entre el colesterol HDL y el índice de masa corporal varía según el nivel de exposición, siendo $b_1 = -0{,}024$ la estimación de la pendiente en el nivel medio $\bar{x}_1 = 26{,}2$ kg/m^2 del índice de masa corporal y $2b_2 = 2 \cdot 0{,}0016 = 0{,}0032$ el cambio de pendiente por cada incremento de 1 kg/m^2 en el índice de masa corporal. No obstante, es más informativo representar la tendencia global estimada a partir del modelo cuadrático. Para ello, se calculan los valores medios del colesterol HDL predichos por el modelo cuadrático para los distintos valores observados x_1 del índice de masa corporal, manteniendo constantes el consumo de alcohol y la variable indicadora de fumadores actuales en sus respectivas medias $\bar{x}_2 = 16{,}5$ g/día y $\bar{x}_3 = 172/448 = 0{,}38$ (proporción de fumadores actuales),

$$\begin{aligned}\hat{y}\,(x_1;\,16{,}5;\,0{,}38) &= 1{,}05 - 0{,}024(x_1 - 26{,}2) + 0{,}0016(x_1 - 26{,}2)^2 \\ &\quad + 0{,}0030 \cdot 16{,}5 - 0{,}098 \cdot 0{,}38 \\ &= 1{,}06 - 0{,}024(x_1 - 26{,}2) + 0{,}0016(x_1 - 26{,}2)^2.\end{aligned}$$

Notar que la elección de los valores fijos de las otras variables explicativas es arbitraria, ya que sólo afectan a la constante de la relación cuadrática. En la práctica, es habitual fijar las restantes variables de ajuste en sus medias muestrales para obtener valores absolutos de la variable respuesta representativos de la población a estudio. La tendencia cuadrática estimada entre el índice de masa corporal y el colesterol HDL se representa en la Figura 11.5, junto con la relación lineal obtenida del mismo modelo de la Tabla 11.9 excluyendo el término cuadrático del índice de masa corporal. En comparación con la tendencia lineal, el modelo cuadrático estima una disminución más pronunciada de la

media del colesterol HDL dentro del rango de normopeso (< 25 kg/m^2), que se atenúa progresivamente al aumentar los niveles del índice de masa corporal.

Aunque los modelos cuadráticos permiten detectar efectos no lineales de las variables explicativas, la tendencia global resultante de estos modelos puede estar fuertemente influenciada por una o muy pocas observaciones con valores extremos de la variable explicativa. En este sentido, resulta especialmente importante evaluar los cambios que se producen en la tendencia cuadrática, o incluso la propia idoneidad del modelo cuadrático, al excluir del análisis las observaciones más influyentes (véase apartado de análisis diagnóstico).

11.7 CONFUSIÓN E INTERACCIÓN EN REGRESIÓN LINEAL

La regresión lineal múltiple puede utilizarse con dos propósitos claramente diferenciados. Por un lado, los modelos de regresión pueden emplearse para predecir el valor de la variable respuesta en función de los valores de las variables explicativas. En tal caso, el interés se centra en identificar e incluir todas aquellas variables explicativas que se asocien de forma significativa e independiente con la variable respuesta, de tal forma que el modelo resultante se ajuste bien a los datos observados (elevado coeficiente de determinación) y prediga con cierta precisión la respuesta en nuevos sujetos. Los contrastes parciales descritos en el Apartado 11.4.2 son particularmente útiles para este propósito, ya que permiten seleccionar las variables explicativas que mejoran significativamente la capacidad predictiva del modelo. Por otro lado, los modelos de regresión pueden utilizarse para estudiar la relación de una o varias variables explicativas de interés con la variable respuesta, controlando por otras variables explicativas o covariables que pudieran afectar a dicha relación. En este caso, no es necesario que el modelo incluya todos los determinantes de la variable respuesta, sino únicamente aquellos que influyan en la asociación objeto de estudio; es decir, aquellas covariables cuya inclusión afecte a las estimaciones de los coeficientes de regresión asociados a las variables explicativas de interés.

La confusión y la interacción son dos conceptos epidemiológicos estrechamente relacionados con este segundo propósito. A continuación se presenta una descripción general de ambos conceptos y su tratamiento dentro de los modelos de regresión lineal múltiple.

11.7.1 Control de la confusión en regresión lineal

La confusión se define como una distorsión en el efecto estimado de una variable explicativa sobre la variable respuesta debido a la interposición de otra covariable, denominada **factor de confusión** o simplemente confusor, cuyo efecto se confunde o se mezcla con el verdadero efecto de la variable explicativa de interés. La distorsión inducida por el factor de confusión puede ser grande y dar lugar tanto a una sobreestimación como a una infraestimación del efecto subyacente, dependiendo de la dirección de las asociaciones del factor de confusión con las variables explicativa y respuesta. El factor de confusión puede producir incluso un cambio en la dirección del efecto observado.

Para que una covariable X_2 pueda confundir la asociación entre la variable explicativa de interés X_1 y la variable respuesta Y en un modelo de regresión lineal debe cumplir tres condiciones necesarias:

- El factor de confusión X_2 debe estar linealmente relacionado con la variable explicativa X_1. Si las variables X_1 y X_2 están incorrelacionadas, sus efectos sobre la variable respuesta Y no podrán confundirse o mezclarse, de tal forma que la estimación del coeficiente asociado a la variable explicativa de interés X_1 no se verá afectada por la inclusión de la covariable

X_2 en el modelo. Este requisito ya se comprobó formalmente en el Apartado 11.3.1 y se ilustró gráficamente en la Figura 11.1(*a*).

- El factor de confusión X_2 debe estar asociado con la variable respuesta Y independientemente de su asociación con la variable explicativa X_1. Aunque las variables X_1 y X_2 estén correlacionadas, el efecto estimado de la variable explicativa X_1 sólo podrá estar confundido por la covariable X_2 cuando ésta tenga un efecto independiente sobre la variable respuesta Y. Si la covariable X_2 se relaciona con la respuesta Y únicamente a través de su asociación con la variable explicativa X_1, puede probarse que $r_{yx_2} = r_{yx_1} r_{x_1x_2}$, de donde se deriva que las estimaciones de los coeficientes de regresión múltiple asociados a las variables X_1 y X_2 se reducen a

$$b_1 = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{1 - r^2_{x_1x_2}} \frac{s_y}{s_{x_1}} = \frac{r_{yx_1}(1 - r^2_{x_1x_2})}{1 - r^2_{x_1x_2}} \frac{s_y}{s_{x_1}} = r_{yx_1} \frac{s_y}{s_{x_1}},$$

$$b_2 = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{1 - r^2_{x_1x_2}} \frac{s_y}{s_{x_2}} = \frac{r_{yx_1} r_{x_1x_2} - r_{yx_1} r_{x_1x_2}}{1 - r^2_{x_1x_2}} \frac{s_y}{s_{x_2}} = 0.$$

  Así, la covariable X_2 no se relacionará con la respuesta al controlar por X_1, mientras que el efecto estimado para la variable explicativa X_1 permanecerá inalterable al ajustar por X_2, con lo que la covariable X_2 no será un factor de confusión para la asociación entre X_1 e Y.

- El factor de confusión X_2 no debe ser un paso intermedio en la relación de la variable explicativa X_1 con la variable respuesta Y. A diferencia de las dos condiciones anteriores, este requisito epidemiológico no puede comprobarse con los datos disponibles y requiere de información externa o juicio experto sobre los mecanismos subyacentes que relacionan la variable explicativa con la respuesta. Por ejemplo, el índice de masa corporal podría considerarse a simple vista un potencial factor de confusión para la asociación entre la actividad física y el colesterol HDL, ya que se relaciona de forma independiente con ambas variables. Sin embargo, el índice de masa corporal no es un factor extraño que distorsiona dicha asociación, sino más bien un factor intermedio, ya que la actividad física reduce el índice de masa corporal, que a su vez provoca un aumento del colesterol HDL. En general, los modelos de regresión no deben incluir factores intermedios para la asociación objeto de estudio, a no ser que se pretenda estimar el efecto que no está mediado por dichos factores.

La selección de los potenciales factores de confusión debe limitarse, por tanto, a las covariables que satisfagan estas tres condiciones necesarias, a saber, aquellas covariables que se asocien de forma independiente con las variables explicativa y respuesta y que no constituyan un paso intermedio en la relación entre ambas variables. No obstante, es posible que una covariable cumpla los tres requisitos y no sea un factor de confusión, en el sentido de no introducir un sesgo en la asociación a estudio. Esto puede ocurrir, por ejemplo, cuando existen varios factores cuyos potenciales sesgos de confusión se compensan al actuar en direcciones opuestas.

En la práctica, para determinar si una o varias covariables son en realidad factores de confusión, se compara la **estimación cruda** de la asociación objeto de estudio con la **estimación ajustada** por los potenciales factores de confusión. Como se vio en el Apartado 11.2, estas estimaciones ajustadas pueden obtenerse directamente a partir de modelos de regresión múltiple que incorporen los potenciales factores de confusión además de la variable explicativa de interés. Así, los factores de confusión vendrán determinados por aquellas covariables cuya inclusión en el modelo produzca un cambio substancial en la estimación del coeficiente de regresión asociado a la variable explicativa de interés. La comparación entre los coeficientes

estimados con y sin ajuste por los potenciales factores de confusión no se realiza mediante pruebas estadísticas, ya que la significación estadística no depende únicamente de la magnitud del cambio, sino también del tamaño muestral (véase Apartado 5.4.2). Aunque el criterio varía según el ámbito de aplicación, en general se considera necesario controlar la confusión cuando la estimación cruda difiere de la ajustada en más del 10%.

Ejemplo 11.12 En los ejemplos anteriores se han considerado otros determinantes del colesterol HDL distintos del índice de masa corporal, pero no se ha prestado especial atención a la confusión que podrían inducir estos factores en la asociación entre el índice de masa corporal y el colesterol HDL. La edad y el estatus socioeconómico no mostraron un efecto independiente sobre los niveles de colesterol HDL (Tabla 11.4), por lo que no cumplen una de las condiciones necesarias para ser factores de confusión. Sin embargo, el consumo de alcohol y el hábito tabáquico actual sí se asociaron con el colesterol HDL independientemente del índice de masa corporal (Tablas 11.7, 11.8 y 11.9). Además, el alcohol y el tabaco son factores externos que no median en la relación del índice de masa corporal con el colesterol HDL. Si ambas covariables se asociaran también con el índice de masa corporal, verificarían los tres requisitos para ser potenciales factores de confusión.

La Tabla 11.10 muestra las estimaciones del coeficiente asociado al índice de masa corporal en distintos modelos de regresión lineal, a saber, un primer modelo sin covariables de ajuste, un segundo modelo ajustado por el consumo de alcohol, un tercer modelo ajustado por el hábito tabáquico actual y un último modelo ajustado por ambas covariables. Todos los modelos se obtuvieron a partir de la misma muestra de 448 controles del estudio EURAMIC con información completa de todas las variables. Tomando como referencia el modelo ajustado por ambas covariables, el cambio relativo que se produce en el coeficiente estimado del índice de masa corporal al excluir el consumo de alcohol es

$$\frac{b_{1|3}}{b_{1|2,3}} = \frac{-0{,}0225}{-0{,}0209} = 1{,}08;$$

es decir, una vez tenido en cuenta el hábito tabáquico actual, las diferencias en el consumo de alcohol provocan una sobreestimación del $100(1{,}08 - 1) = 8\%$ en la asociación inversa del índice de masa corporal con el colesterol HDL. Como se apuntó en el Ejemplo 11.2, esto es debido a que una pequeña parte de la reducción del colesterol HDL entre los sujetos con mayor índice de masa corporal se debe en realidad a su menor consumo de alcohol. Por otra parte, si se excluye la variable indicadora de los fumadores actuales, el cambio relativo es

$$\frac{b_{1|2}}{b_{1|2,3}} = \frac{-0{,}0206}{-0{,}0209} = 0{,}99;$$

esto es, una vez controladas las diferencias en la ingesta de alcohol, el hábito tabáquico actual no introduce virtualmente ningún sesgo en la asociación objeto de estudio (infraestimación del $100(0{,}99 - 1) = -1\%$). Esto es consecuencia de que el hábito tabáquico no se asocia con el índice de masa corporal en el grupo control del estudio EURAMIC (la media del índice de masa corporal es 26,3 kg/m^2 en los no fumadores y 26,1 kg/m^2 en los fumadores actuales). Por último, si se excluyen simultáneamente ambas covariables del modelo, el cambio relativo en el coeficiente estimado del índice de masa corporal es

$$\frac{b_{1}}{b_{1|2,3}} = \frac{-0{,}0222}{-0{,}0209} = 1{,}06.$$

Tabla 11.10 Estimación de la relación del índice de masa corporal (IMC) con el colesterol HDL a partir de diferentes modelos de regresión lineal múltiple ajustados por distintas combinaciones del consumo de alcohol y el hábito tabáquico actual en el grupo control del estudio EURAMIC.

	Coeficiente asociado al IMC		
Covariable de ajuste	**Estimación**	**Error estándar**	**IC al 95%**
Ninguna	−0,0222	0,0037	(−0,0295; −0,0149)
Alcohol	−0,0206	0,0036	(−0,0278; −0,0135)
Fumador actual	−0,0225	0,0037	(−0,0297; −0,0152)
Alcohol, fumador actual	−0,0209	0,0036	(−0,0279; −0,0138)

Notar que esta sobreestimación del 6% es resultado de la combinación de los sesgos inducidos de forma independiente por el consumo de alcohol y el hábito tabáquico. Si se adoptara el criterio estándar del 10%, se concluiría que el consumo de alcohol y el hábito tabáquico no son factores de confusión importantes para la asociación entre el índice de masa corporal y el colesterol HDL en los controles del estudio EURAMIC. No obstante, a pesar de no cumplir este criterio cuantitativo, se podría decidir ajustar por ambas covariables por razones de credibilidad, ya que el alcohol y el tabaco son determinantes conocidos del colesterol HDL y cualquier estudio sobre este tópico generaría desconfianza si no incluyera estas variables en el análisis.

La confusión es un sesgo introducido por un factor externo en la asociación objeto de estudio que debe prevenirse en el diseño o controlarse en el análisis de los datos. En este sentido, la regresión lineal múltiple es una herramienta útil para controlar la confusión en el análisis, ya que facilita estimaciones ajustadas por las restantes variables explicativas incluidas en el modelo. No obstante, la capacidad de ajuste de los modelos de regresión está condicionada por los siguientes factores:

- La disponibilidad de información sobre los potenciales factores de confusión. Obviamente, no se podrá controlar en el análisis ningún factor de confusión que no se haya medido previamente.
- El efecto conjunto de la variable explicativa de interés y de los factores de confusión. La regresión lineal múltiple asume que los efectos conjuntos son aditivos, de tal forma que si esta asunción no se cumple, la estimación del coeficiente de regresión asociado a la variable explicativa de interés puede estar sesgada.
- Los errores de medida y la especificación de los factores de confusión. Si los factores de confusión están medidos con un error considerable, o si su efecto sobre la variable respuesta se modela de forma inadecuada (por ejemplo, usando términos lineales para relaciones subyacentes curvilíneas), el ajuste no será completo, pudiendo quedar una apreciable confusión residual.

11.7.2 Evaluación de la interacción en regresión lineal

La interacción o modificación de efecto se refiere al cambio en la magnitud de la asociación entre la variable explicativa de interés y la variable respuesta a diferentes niveles de otra variable, que se denomina **modificador de efecto**. A diferencia de la confusión, que es un sesgo

a corregir en la estimación del efecto, la interacción es una característica inherente de la asociación a estudio, que debe describirse mediante **estimaciones específicas** del efecto de la variable explicativa de interés en los distintos niveles del modificador de efecto.

La confusión y la interacción son fenómenos diferentes que pueden o no ocurrir simultáneamente. No obstante, cuando existe evidencia de interacción con una determinada covariable, la valoración de la confusión inducida por dicha covariable es irrelevante. En presencia de interacción, la magnitud del efecto varía según el nivel de la covariable y, en consecuencia, deben obtenerse estimaciones específicas para cada nivel, que están libres de confusión al referirse a sujetos con idéntico valor de la covariable. Por el contrario, cuando no existe interacción, el efecto se asume igual en todos los niveles de la covariable y basta entonces con obtener una única estimación para todos los sujetos, que sí podría estar confundida por diferencias en la distribución de la covariable. Por ello, en la práctica sólo tiene sentido controlar la confusión cuando se ha descartado previamente la presencia de interacción.

Los efectos independientes de una variable explicativa de interés X_1 y otra covariable X_2 sobre la variable respuesta Y se obtienen a partir del modelo de regresión lineal múltiple

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

que incluye distintos términos para cada variable explicativa. Bajo este modelo, la relación entre X_1 e Y para un determinado valor fijo c_2 de la covariable X_2 viene dada por $E(Y|x_1, c_2) = (\beta_0 + \beta_2 c_2) + \beta_1 x_1$. Así, este modelo asume que no existe interacción entre X_1 y X_2 ya que el cambio esperado en Y por cada incremento de una unidad en X_1 es siempre igual a β_1, independientemente del nivel de X_2. De hecho, los cambios en el valor de la covariable X_2 sólo afectan a la constante de la recta de regresión de Y sobre X_1, pero no a su pendiente. Esta ausencia de interacción se representa gráficamente en la Figura 11.6(*a*), donde las rectas de regresión de Y sobre X_1 son líneas paralelas de igual pendiente para los distintos valores (puntos y círculos) de una covariable dicotómica X_2.

En regresión lineal múltiple, la forma más sencilla de modelar la interacción entre la variable explicativa de interés X_1 y la covariable X_2 consiste en añadir al modelo un nuevo término con el producto de ambas variables,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon.$$

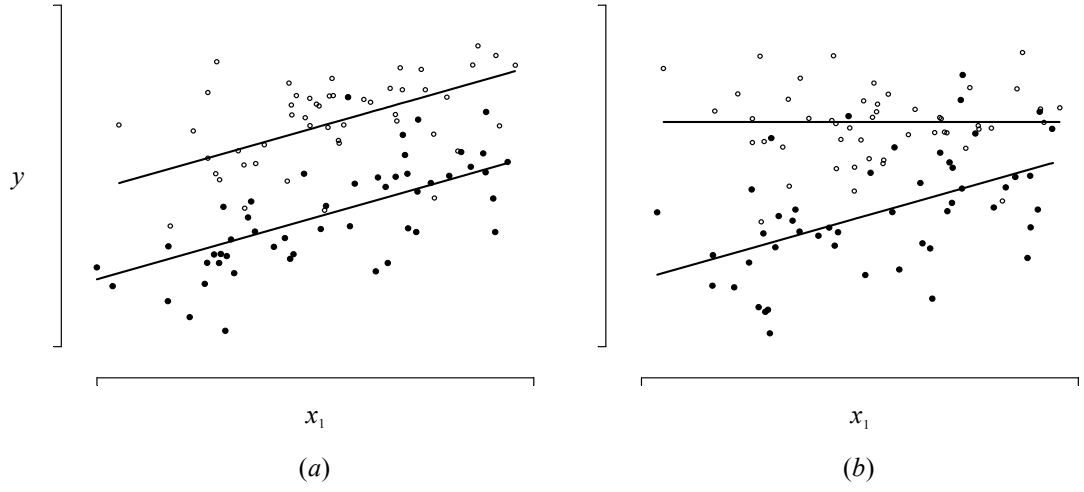


Figura 11.6 **Rectas de regresión de la variable respuesta Y sobre la variable explicativa X_1 para distintos valores (puntos y círculos) de una covariable dicotómica X_2 que no interacciona con X_1 (panel *a*) y que interacciona con X_1 (panel *b*).**

Notar que el modelo ha de incluir el **término producto** X_1X_2 además de los términos principales para las variables X_1 y X_2. Bajo este modelo extendido con el término producto, la recta de regresión de Y sobre X_1 para un determinado valor fijo c_2 de la covariable X_2 viene dada por $E(Y|x_1, c_2) = (\beta_0 + \beta_2 c_2) + (\beta_1 + \beta_3 c_2)x_1$. Así, el nuevo modelo contempla la posibilidad de interacción entre X_1 y X_2 ya que el cambio esperado en Y por cada incremento de una unidad en X_1 es igual a $\beta_1 + \beta_3 c_2$, que depende del nivel de X_2 siempre que el coeficiente β_3 del término producto sea distinto de 0. La presencia de interacción se ilustra en la Figura 11.6(*b*), donde las rectas de regresión de Y sobre X_1 presentan distintas pendientes para los dos valores (puntos y círculos) de una covariable dicotómica X_2.

A diferencia de la confusión, la interacción sí se evalúa estadísticamente mediante el contraste parcial del coeficiente β_3 asociado al término producto. Si este coeficiente no difiere significativamente del valor nulo, el efecto de X_1 sobre la variable respuesta Y no variará significativamente en los distintos niveles de X_2. En ausencia de interacción, ha de eliminarse el término producto y volver al modelo con los términos principales de ambas variables, que permite estimar el efecto global de X_1 ajustado por X_2. Por el contrario, si el coeficiente β_3 del término producto resulta significativo, el efecto de X_1 diferirá significativamente según el nivel de X_2 y, en consecuencia, se tendrá una interacción significativa entre ambas variables. Aunque las estimaciones de los coeficientes del modelo con el término producto no tienen en general una interpretación directa, pueden combinarse para obtener estimaciones especificas de la relación de X_1 con la variable respuesta Y en los distintos niveles de X_2. Para un determinado valor fijo c_2 de la covariable X_2, la ecuación de regresión estimada es $\hat{y}(x_1, c_2) = (b_0 + b_2 c_2) + (b_1 + b_3 c_2)x_1$, de tal forma que el cambio en el nivel medio de Y por cada incremento de una unidad en X_1 se estima mediante $b_1 + b_3 c_2$. Esta combinación constituye un estimador insesgado de la pendiente específica subyacente,

$$E(b_1 + b_3 c_2) = E(b_1) + E(b_3)c_2 = \beta_1 + \beta_3 c_2,$$

cuya varianza viene dada por (véase Apartado 3.4)

$$\begin{aligned} \text{var}(b_1 + b_3 c_2) &= \text{var}(b_1) + c_2^2\,\text{var}(b_3) + 2c_2\text{cov}(b_1, b_3) \\ &= \sigma^2(v_{11} + c_2^2 v_{33} + 2c_2 v_{13}), \end{aligned}$$

que depende de las varianzas de b_1 y b_3 y también de su covarianza ya que, como se muestra en el Apéndice de este tema, las estimaciones de los coeficientes de regresión múltiple están correlacionadas. Así, el intervalo de confianza al $100(1 - \alpha)\%$ para la pendiente subyacente $\beta_1 + \beta_3 c_2$ de la relación entre X_1 e Y en el valor c_2 de la covariable X_2 se calcula como

$$b_1 + b_3 c_2 \pm t_{n-p-1,1-\alpha/2}\; s\sqrt{v_{11} + c_2^2 v_{33} + 2c_2 v_{13}}\;.$$

Ejemplo 11.13 Para evaluar una posible modificación del efecto del índice de masa corporal sobre el colesterol HDL en los estratos de fumadores actuales y no fumadores actuales, se ajustó un modelo de regresión lineal múltiple en los controles del estudio EURAMIC que incluía los términos principales del índice de masa corporal X_1, el consumo de alcohol X_2 y la variable indicadora X_3 de los fumadores actuales, así como un término adicional con el producto entre el índice de masa corporal y la variable indicadora de los fumadores actuales,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \varepsilon.$$

La Tabla 11.11 muestra las estimaciones de los coeficientes de este modelo y las correlaciones entre los distintos pares de coeficientes, que forman parte de los resultados facilitados por los programas estadísticos convencionales. El contraste para la nulidad del coeficiente β_4 asociado al término producto arroja un valor $P = 0{,}16$, lo que indica que no existe una interacción estadísticamente significativa entre el índice de masa corporal y el hábito tabáquico actual en los controles del estudio EURAMIC. No obstante, este contraste podría tener escasa potencia estadística para detectar cambios relevantes en la magnitud de los efectos específicos del índice de masa corporal sobre el colesterol HDL dentro de cada estrato, ya que el estudio cuenta únicamente con 276 no fumadores actuales y 172 fumadores actuales. En este sentido, es aconsejable utilizar los resultados del modelo con el término producto para estimar los efectos específicos y valorar la relevancia del cambio.

Por un lado, en el estrato de los no fumadores actuales, la variable indicadora X_3 toma valor 0 y la ecuación de regresión estimada se reduce a

$$\hat{y}\,(x_1, x_2, 0) = b_0 + b_1x_1 + b_2x_2.$$

Así, una vez controladas las diferencias en el consumo de alcohol, cada incremento de 3,50 kg/m^2 en el índice de masa corporal de los no fumadores actuales se asocia con una disminución media en el colesterol HDL de $3{,}50b_1 = 3{,}50(-0{,}016) = -0{,}057$ mmol/l, con un IC al 95% comprendido entre

$$3{,}50\{b_1 \pm t_{443;0{,}975}SE(b_1)\} = 3{,}50(-0{,}016 \pm 1{,}97\cdot 0{,}0049) = (-0{,}090;\ -0{,}023).$$

La Figura 11.7 muestra en trazo fino la recta de regresión estimada del colesterol HDL sobre el índice de masa corporal entre los no fumadores actuales con un consumo medio de alcohol de $\bar{x}_2 = 16{,}5$ g/día,

$$\hat{y}\,(x_1;\ 16{,}5;\ 0) = 1{,}49 - 0{,}016x_1 + 0{,}0029\cdot 16{,}5 = 1{,}54 - 0{,}016x_1.$$

Por otro lado, en el estrato de los fumadores actuales, la variable indicadora X_3 toma valor 1 y la ecuación de regresión estimada viene dada por

$$\hat{y}\,(x_1, x_2, 1) = (b_0 + b_3) + (b_1 + b_4)x_1 + b_2x_2.$$

Así, después de ajustar por la ingesta de alcohol, los incrementos de 3,50 kg/m^2 en el índice de masa corporal de los fumadores actuales se asocian con una disminución media en el colesterol HDL de $3{,}50(b_1 + b_4) = 3{,}50(-0{,}016 - 0{,}010) = -0{,}092$ mmol/l. Para obtener una estimación por intervalo del efecto específico en este estrato, se calcula en primer lugar la varianza muestral de $b_1 + b_4$

$$\begin{aligned}\text{var}(b_1 + b_4) &= \text{var}(b_1) + \text{var}(b_4) + 2\text{cov}(b_1, b_4)\\ &= SE(b_1)^2 + SE(b_4)^2 + 2SE(b_1)SE(b_4)\,r_{b_1b_4}\\ &= 0{,}0049^2 + 0{,}0072^2 + 2\cdot 0{,}0049\cdot 0{,}0072(-0{,}679) = 0{,}000028,\end{aligned}$$

donde la correlación entre b_1 y b_4 se obtiene de la segunda parte de la Tabla 11.11. El IC al 95% para el efecto específico del índice de masa corporal en los fumadores actuales se calcula entonces como

$$\begin{aligned}&3{,}50\{b_1 + b_4 \pm t_{443;0{,}975}\ SE(b_1 + b_4)\}\\ &\quad = 3{,}50(-0{,}016 - 0{,}010 \pm 1{,}97\sqrt{0{,}000028}\,) = (-0{,}129;\ -0{,}056).\end{aligned}$$

En la Figura 11.7 se representa en trazo grueso la recta de regresión estimada del colesterol HDL sobre el índice de masa corporal entre los fumadores actuales con una ingesta media de alcohol de $\bar{x}_2 = 16{,}5$ g/día,

$$\begin{aligned}\hat{y}\,(x_1;\,16{,}5;\,1) &= (1{,}49 + 0{,}18) - (0{,}016 + 0{,}010)x_1 + 0{,}0029 \cdot 16{,}5 \\ &= 1{,}72 - 0{,}026x_1.\end{aligned}$$

En conclusión, a partir del modelo con la interacción se tiene que un mismo incremento de 3,50 kg/m^2 en el índice de masa corporal se asocia con distintas disminuciones en el nivel medio de colesterol HDL de $-0{,}057$ mmol/l en los no fumadores y $-0{,}092$ mmol/l en los fumadores actuales. El cambio en la magnitud del efecto es notable pero, debido al limitado tamaño muestral de ambos estratos, las estimaciones específicas son relativamente imprecisas y el test de interacción no alcanza la significación estadística. Por tanto, los resultados de este estudio no son concluyentes respecto a la posible acción sinérgica del índice de masa corporal y el tabaco en los niveles de colesterol HDL, y se requeriría de un estudio más potente para detectar un cambio subyacente de dicha magnitud en los efectos específicos del índice de masa corporal en fumadores y no fumadores actuales.

Tabla 11.11 Resultados de la regresión lineal múltiple del colesterol HDL sobre el índice de masa corporal (IMC), el consumo de alcohol, la variable indicadora de fumadores actuales y el producto entre IMC y fumador actual en el grupo control del estudio EURAMIC.

Coeficientes de regresión

				Test H_0: $\beta_j = 0$	
	Estimación	Error estándar	IC al 95%	*t*	Valor *P*
Constante	1,49	0,13	(1,24; 1,75)	11,47	< 0,001
IMC	−0,016	0,0049	(−0,026; −0,007)	−3,30	0,001
Alcohol	0,0029	0,0006	(0,0018; 0,0041)	4,88	< 0,001
Fumador	0,18	0,19	(−0,20; 0,55)	0,91	0,36
IMC·Fumador	−0,010	0,0072	(−0,024; 0,004)	−1,40	0,16

Matriz de correlaciones de las estimaciones

	IMC	Alcohol	Fumador	IMC·Fumador
Constante	−0,990	−0,052	−0,670	0,664
IMC		−0,016	0,674	−0,679
Alcohol			−0,134	0,120
Fumador				−0,990

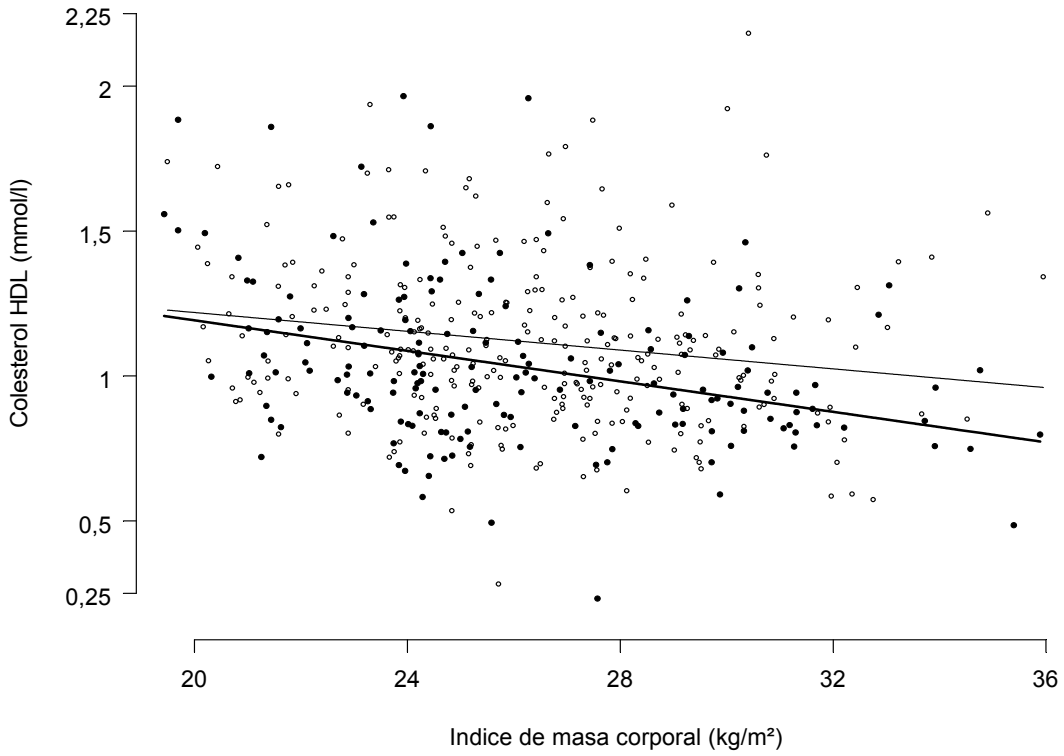


Figura 11.7 **Rectas de regresión del colesterol HDL sobre el índice de masa corporal en fumadores actuales (puntos y línea gruesa) y no fumadores actuales (círculos y línea fina) obtenidas de un modelo con interacción entre el índice de masa corporal y el hábito tabáquico actual en el grupo control del estudio EURAMIC.**

En regresión lineal, la ausencia de interacción entre dos variables explicativas X_1 y X_2 implica que sus efectos sobre la variable respuesta son aditivos; es decir, el efecto conjunto de ambas variables es la suma de sus efectos independientes. La presencia de interacción puede interpretarse, por tanto, como una **desviación de la aditividad**, que puede deberse tanto a efectos subaditivos como a efectos supraaditivos. Más concretamente, en un modelo de regresión lineal con el término producto entre X_1 y X_2, el cambio esperado en Y al aumentar simultáneamente una unidad ambas variables explicativas es

$$\begin{aligned} &E(Y|x_1+1, x_2+1) - E(Y|x_1, x_2) \\ &\quad = \beta_0 + \beta_1(x_1+1) + \beta_2(x_2+1) + \beta_3(x_1+1)(x_2+1) \\ &\quad - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2) = \beta_1 + \beta_2 + \beta_3(x_1 + x_2 + 1). \end{aligned}$$

En el mismo modelo, los cambios esperados en Y al aumentar por separado una unidad cada variable explicativa son

$$\begin{aligned} E(Y|x_1+1, x_2) - E(Y|x_1, x_2) &= \beta_0 + \beta_1(x_1+1) + \beta_2 x_2 + \beta_3(x_1+1)x_2 \\ &\quad - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2) = \beta_1 + \beta_3 x_2 \end{aligned}$$

y

$$\begin{aligned} E(Y|x_1, x_2+1) - E(Y|x_1, x_2) &= \beta_0 + \beta_1 x_1 + \beta_2(x_2+1) + \beta_3 x_1(x_2+1) \\ &\quad - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2) = \beta_2 + \beta_3 x_1. \end{aligned}$$

Así, si existe interacción entre X_1 y X_2, el coeficiente β_3 asociado al término producto será distinto de 0 y el efecto conjunto de ambas variables diferirá de la suma de sus efectos independientes,

$$\begin{aligned} &E(Y|x_1+1, x_2+1) - E(Y|x_1, x_2) \\ &\quad -\{E(Y|x_1+1, x_2) - E(Y|x_1, x_2) + E(Y|x_1, x_2+1) - E(Y|x_1, x_2)\} \\ &\quad = \beta_1 + \beta_2 + \beta_3(x_1 + x_2 + 1) - (\beta_1 + \beta_3 x_2 + \beta_2 + \beta_3 x_1) = \beta_3. \end{aligned}$$

Ejemplo 11.14 A partir de las estimaciones del modelo con el término producto de la Tabla 11.11, el nivel medio de colesterol HDL de los no fumadores con un índice de masa corporal de 25 kg/m^2 y un consumo de alcohol de 20 g/día es

$$\hat{y}\,(25, 20, 0) = 1{,}49 - 0{,}016 \cdot 25 + 0{,}0029 \cdot 20 = 1{,}148,$$

el de los no fumadores con un elevado índice de masa corporal de 28,5 kg/m^2 y un consumo de alcohol de 20 g/día es

$$\hat{y}\,(28{,}5;\ 20;\ 0) = 1{,}49 - 0{,}016 \cdot 28{,}5 + 0{,}0029 \cdot 20 = 1{,}091,$$

el de los fumadores actuales con un índice de masa corporal de 25 kg/m^2 y un consumo de alcohol de 20 g/día es

$$\hat{y}\,(25, 20, 1) = 1{,}49 - 0{,}016 \cdot 25 + 0{,}0029 \cdot 20 + 0{,}18 - 0{,}010 \cdot 25 = 1{,}070$$

y el de los fumadores actuales con un elevado índice de masa corporal de 28,5 kg/m^2 y un consumo de alcohol de 20 g/día es

$$\hat{y}\,(28{,}5;\ 20;\ 1) = 1{,}49 - 0{,}016 \cdot 28{,}5 + 0{,}0029 \cdot 20 + 0{,}18 - 0{,}010 \cdot 28{,}5 = 0{,}978.$$

Tomando como referencia a los sujetos no fumadores con un índice de masa corporal de 25 kg/m^2, los no fumadores con un elevado índice de masa corporal de 28,5 kg/m^2 presentan una disminución en la media del colesterol HDL de

$$\hat{y}\,(28{,}5;\ 20;\ 0) - \hat{y}\,(25, 20, 0) = 1{,}091 - 1{,}148 = -0{,}057,$$

los fumadores actuales con el mismo índice de masa corporal de 25 kg/m^2 de

$$\hat{y}\,(25, 20, 1) - \hat{y}\,(25, 20, 0) = 1{,}070 - 1{,}148 = -0{,}078$$

y los fumadores actuales con un elevado índice de masa corporal de 28,5 kg/m^2 de

$$\hat{y}\,(28{,}5;\ 20;\ 1) - \hat{y}\,(25, 20, 0) = 0{,}978 - 1{,}148 = -0{,}170.$$

Así, la disminución media del colesterol HDL de $-0{,}170$ mmol/l debida conjuntamente a fumar y aumentar el índice de masa corporal es mayor en valor absoluto que la suma de las disminuciones $-0{,}057 - 0{,}078 = -0{,}135$ mmol/l debidas a cada factor por separado. En otras palabras, los datos del estudio EURAMIC apuntan a un posible efecto supraaditivo o sinérgico del índice de masa corporal y el tabaco sobre los niveles de colesterol HDL.

11.8 APÉNDICE: FORMULACIÓN MATRICIAL DE LA REGRESIÓN LINEAL MÚLTIPLE

Según la estructura de la regresión lineal múltiple presentada en el Apartado 11.2, cada una de las n observaciones independientes $(y_i, x_{i1}, \ldots, x_{ip})$ presenta la relación lineal

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

donde los errores aleatorios ε_i son independientes y están distribuidos normalmente con media 0 y varianza constante σ^2. Estas n ecuaciones lineales pueden reescribirse en forma matricial como

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

o, abreviadamente,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

donde $\mathbf{y}$ es un vector $n \times 1$ con los valores de la variable respuesta, $\mathbf{X}$ es una matriz de dimensión $n \times (p+1)$ cuyas columnas son los valores de cada variable explicativa más una primera columna de unos, $\boldsymbol{\beta}$ es un vector $(p+1) \times 1$ con los coeficientes de regresión y $\boldsymbol{\varepsilon}$ es un vector $n \times 1$ con los errores aleatorios. El vector de errores aleatorios $\boldsymbol{\varepsilon}$ sigue entonces una distribución normal multivariante con media $\mathbf{0}$ y matriz diagonal de varianzas-covarianzas $\sigma^2\mathbf{I}$,

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}),$$

donde $\mathbf{0}$ denota el vector nulo $n \times 1$ con todos sus componentes iguales a cero e $\mathbf{I}$ denota la matriz identidad $n \times n$ con unos en la diagonal y ceros fuera de ella. Notar que, por la asunción de homogeneidad de la varianza, todas las varianzas de la diagonal de la matriz de varianzas-covarianzas son iguales a σ^2 y que, por tratarse de observaciones independientes, las covarianzas de fuera de la diagonal son iguales a cero.

A partir de esta formulación matricial del modelo de regresión lineal múltiple, resulta sencillo calcular las estimaciones de los coeficientes de regresión por el método de mínimos cuadrados. En el Apartado 11.3.1, se comprobó que estas estimaciones vienen dadas por la solución al sistema de $p+1$ ecuaciones lineales

$$\begin{aligned} \sum_{i=1}^{n} y_i &= nb_0 + b_1 \sum_{i=1}^{n} x_{i1} + \ldots + b_p \sum_{i=1}^{n} x_{ip}, \\ \sum_{i=1}^{n} x_{i1} y_i &= b_0 \sum_{i=1}^{n} x_{i1} + b_1 \sum_{i=1}^{n} x_{i1}^2 + \ldots + b_p \sum_{i=1}^{n} x_{i1} x_{ip}, \\ &\vdots \\ \sum_{i=1}^{n} x_{ip} y_i &= b_0 \sum_{i=1}^{n} x_{ip} + b_1 \sum_{i=1}^{n} x_{i1} x_{ip} + \ldots + b_p \sum_{i=1}^{n} x_{ip}^2, \end{aligned}$$

que puede representarse matricialmente como

$$\begin{bmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{n1} \\ \vdots & & \vdots \\ x_{1p} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} y_1 \\ \\ \vdots \\ \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{n1} \\ \vdots & & \vdots \\ x_{1p} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}$$

o, abreviadamente,

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b},$$

donde $\mathbf{X}'$ es la matriz traspuesta de $\mathbf{X}$ y $\mathbf{b}$ es el vector $(p + 1) \times 1$ con las estimaciones de los coeficientes. Como el modelo de regresión lineal múltiple asume que las variables explicativas son linealmente independientes y que el número de observaciones n es superior o igual al número de coeficientes $p + 1$, la matriz $\mathbf{X}$ tiene rango $p + 1$ y, en consecuencia, la matriz cuadrada $\mathbf{X}'\mathbf{X}$ es no singular. Multiplicando ambos lados de la ecuación anterior por la matriz inversa $(\mathbf{X}'\mathbf{X})^{-1}$, se obtienen las estimaciones de los coeficientes de regresión

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

De esta formula matricial se desprende que los estimadores de mínimos cuadrados $\mathbf{b}$ son combinaciones lineales de los valores de la variable respuesta $\mathbf{y}$, cuyos coeficientes dependen de los valores de las variables explicativas $\mathbf{X}$ que se asumen constantes. En consecuencia, si el tamaño muestral n es suficientemente grande, puede aplicarse una generalización del teorema central del límite para demostrar que los estimadores $\mathbf{b}$ siguen aproximadamente una distribución normal multivariante con media

$$\begin{aligned} E(\mathbf{b}) &= E\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\} = E\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}) = \boldsymbol{\beta} \end{aligned}$$

y matriz de varianzas-covarianzas

$$\begin{aligned} \text{var}(\mathbf{b}) &= E\{(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'\} = E\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}, \end{aligned}$$

ya que $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ y $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \text{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ por las asunciones de linealidad, aditividad, homogeneidad de la varianza e independencia. Cada estimador de mínimos cuadrados b_j es entonces un estimador insesgado de su correspondiente coeficiente de regresión β_j y sigue aproximadamente la distribución normal

$$b_j \overset{\sim}{\to} N(\beta_j, \sigma^2 v_{jj}), \qquad j = 0, 1, \ldots, p,$$

donde v_{jj} es el elemento (j, j)-ésimo de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$. Además, los estimadores b_j y b_k de distintos coeficientes de regresión están correlacionados con una covarianza $\text{cov}(b_j, b_k) = \sigma^2 v_{jk}$. Cabe destacar que estas distribuciones muestrales no requieren de la asunción de normalidad y son válidas para cualquier distribución subyacente de la variable respuesta, siempre que el tamaño muestral sea suficientemente grande.

Una vez estimados los coeficientes de regresión, el valor esperado de la variable respuesta Y dados unos valores fijos $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0p})'$ de las variables explicativas puede estimarse como

$$\hat{y}_0 = b_0 + b_1 x_{01} + \ldots + b_p x_{0p} = \mathbf{x}_0' \mathbf{b}$$

que, al ser una combinación lineal de $\mathbf{b}$, también se distribuye de forma aproximadamente normal en muestras suficientemente grandes, con media

$$E(\hat{y}_0) = \mathbf{x}_0' E(\mathbf{b}) = \mathbf{x}_0' \boldsymbol{\beta}$$

y varianza

$$\text{var}(\hat{y}_0) = \mathbf{x}_0' E\{(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'\}\mathbf{x}_0 = \sigma^2 \mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = \sigma^2 h_0;$$

es decir,

$$\hat{y}_0 \overset{\sim}{\rightarrow} N(\mathbf{x}_0' \boldsymbol{\beta}, \sigma^2 h_0),$$

donde el leverage $h_0 = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$ es una medida estandarizada de la desviación de $\mathbf{x}_0$ respecto de las medias muestrales de las variables explicativas. El valor predicho $\hat{y}_0$ es un estimador insesgado no sólo de la esperanza o media poblacional de la variable respuesta $\mathbf{x}_0'\boldsymbol{\beta}$, sino también de la respuesta individual de un nuevo sujeto $y_0 = \mathbf{x}_0'\boldsymbol{\beta} + \varepsilon_0$ ya que

$$E(\hat{y}_0 - y_0) = E\{\mathbf{x}_0'(\mathbf{b} - \boldsymbol{\beta}) - \varepsilon_0\} = \mathbf{x}_0' E(\mathbf{b} - \boldsymbol{\beta}) - E(\varepsilon_0) = 0.$$

Como el valor predicho $\hat{y}_0$ no depende de la nueva observación y_0, la varianza de esta diferencia es

$$\begin{aligned} \text{var}(\hat{y}_0 - y_0) &= \mathbf{x}_0' E\{(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'\}\mathbf{x}_0 + \text{var}(\varepsilon_0) \\ &= \sigma^2\{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\} = \sigma^2(1 + h_0). \end{aligned}$$

Si además el error ε_0 de la nueva observación se distribuye de forma normal (asunción de normalidad), la diferencia $\hat{y}_0 - y_0$ también seguirá la distribución normal

$$\hat{y}_0 - y_0 \sim N(0, \sigma^2(1 + h_0)).$$

En el caso particular de una única variable explicativa, todos los resultados anteriores se reducen a los obtenidos en regresión lineal simple (véase Apartados 10.3.1, 10.3.3 y 10.3.4). Así, se tiene que

$$\begin{aligned} \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} &= \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \\ &= \frac{1}{n\sum (x_i - \bar{x})^2} \begin{bmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ n\sum x_i y_i - \sum x_i \sum y_i \end{bmatrix}, \end{aligned}$$

donde todos los sumatorios son sobre $i = 1, \ldots, n$. Por tanto, la estimación de la pendiente es

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = r \frac{s_y}{s_x}$$

y la estimación de la constante es

$$b_0 = \frac{\bar{y}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \bar{x}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \bar{y} - b_1\bar{x}\,.$$

Además, la matriz de varianzas-covarianzas de estos estimadores es

$$\text{var}(\mathbf{b}) = \begin{bmatrix} \text{var}(b_0) & \text{cov}(b_0, b_1) \\ \text{cov}(b_0, b_1) & \text{var}(b_1) \end{bmatrix} = \sigma^2 \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1}$$
$$= \frac{\sigma^2}{n\sum(x_i - \bar{x})^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix},$$

de donde se sigue que

$$\text{var}(b_0) = \sigma^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + n\bar{x}^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right),$$

$$\text{var}(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)s_x^2},$$

$$\text{cov}(b_0, b_1) = \frac{-\sigma^2\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{-\sigma^2\bar{x}}{(n-1)s_x^2}\,.$$

Por último, para un valor fijo x_0 de la variable explicativa, la varianza del valor predicho $\hat{y}_0 = b_0 + b_1 x_0$ es

$$\text{var}(\hat{y}_0) = \sigma^2 [1 \quad x_0] \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ x_0 \end{bmatrix}$$
$$= \frac{\sigma^2}{n\sum(x_i - \bar{x})^2} [1 \quad x_0] \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} 1 \\ x_0 \end{bmatrix}$$
$$= \sigma^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(x_0 - \bar{x})^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}\right),$$

donde se observa que el leverage del valor x_0

$$h_0 = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}$$

es una medida estandarizada de su desviación respecto de la media muestral $\bar{x}$ de la variable explicativa.

11.9 REFERENCIAS

1. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research, Fourth Edition*. Oxford: Blackwell Science, 2002.
2. Bickel PJ, Doksum KA. *Mathematical Statistics: Basic Ideas and Selected Topics*. Englewood Cliffs, NJ: Prentice Hall, 1977.
3. Casella G, Berger RL. *Statistical Inference, Second Edition*. Belmont, CA: Duxbury Press, 2002.
4. Draper NR, Smith H. *Applied Regression Analysis, Third Edition*. New York: John Wiley & Sons, 1998.
5. Kleinbaum DG, Kupper LL, Nizam A, Muller KE. *Applied Regression Analysis and Other Multivariable Methods, Fourth Edition*. Belmont, CA: Duxbury Press, 2008.
6. McCullagh P, Nelder JA. *Generalized Linear Models, Second Edition*. London: Chapman & Hall, 1989.
7. Peña D. *Estadística: Modelos y Métodos, Volumen 2, Modelos Lineales y Series Temporales*. Madrid: Alianza Editorial, 1987.
8. Rosner B. *Fundamentals of Biostatistics, Sixth Edition*. Belmont, CA: Duxbury Press, 2006.
9. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology, Third Edition*. Philadelphia: Lippincott Williams & Wilkins, 2008.
10. Seber GAF, Lee AJ. *Linear Regression Analysis, Second Edition*. New York: John Wiley & Sons, 2003.
11. Snedecor GW, Cochran WG. *Statistical Methods, Eighth Edition*. Ames, IA: Iowa State University Press, 1989.
12. Stuart A, Ord JK, Arnold S. *Kendall's Advanced Theory of Statistics, Volume 2A, Classical Inference and the Linear Model, Sixth Edition*. London: Edward Arnold, 1999.
13. Weisberg S. *Applied Linear Regression, Third Edition*. New York: John Wiley & Sons, 2005.

APÉNDICE

TABLAS ESTADÍSTICAS

Tabla 1 **Probabilidades** $P(X=k)=\binom{n}{k}\pi^k(1-\pi)^{n-k}$ **para la distribución binomial** X **con parámetros** $n = 2, 3, ..., 20$ **y** $\pi = 0{,}05, 0{,}10, ..., 0{,}50$.*

		π									
n	***k***	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
2	0	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
	1	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000
	2	0,0025	0,0100	0,0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500
3	0	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
	1	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750
	2	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750
	3	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4	0	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
	1	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500
	2	0,0135	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750
	3	0,0005	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500
	4	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5	0	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0313
	1	0,2036	0,3281	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1563
	2	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125
	3	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125
	4	0,0000	0,0005	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1563
	5	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0313
6	0	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
	1	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938
	2	0,0305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3110	0,2780	0,2344
	3	0,0021	0,0146	0,0415	0,0819	0,1318	0,1852	0,2355	0,2765	0,3032	0,3125
	4	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344
	5	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938
	6	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7	0	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
	1	0,2573	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547
	2	0,0406	0,1240	0,2097	0,2753	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641
	3	0,0036	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734
	4	0,0002	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734
	5	0,0000	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641
	6	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547
	7	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078
8	0	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
	1	0,2793	0,3826	0,3847	0,3355	0,2670	0,1977	0,1373	0,0896	0,0548	0,0313
	2	0,0515	0,1488	0,2376	0,2936	0,3115	0,2965	0,2587	0,2090	0,1569	0,1094
	3	0,0054	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188
	4	0,0004	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734
	5	0,0000	0,0004	0,0026	0,0092	0,0231	0,0467	0,0808	0,1239	0,1719	0,2188
	6	0,0000	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094
	7	0,0000	0,0000	0,0000	0,0001	0,0004	0,0012	0,0033	0,0079	0,0164	0,0313
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0017	0,0039

Tabla 1 (Continuación)

		π									
n	***k***	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
9	0	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0207	0,0101	0,0046	0,0020
	1	0,2985	0,3874	0,3679	0,3020	0,2253	0,1556	0,1004	0,0605	0,0339	0,0176
	2	0,0629	0,1722	0,2597	0,3020	0,3003	0,2668	0,2162	0,1612	0,1110	0,0703
	3	0,0077	0,0446	0,1069	0,1762	0,2336	0,2668	0,2716	0,2508	0,2119	0,1641
	4	0,0006	0,0074	0,0283	0,0661	0,1168	0,1715	0,2194	0,2508	0,2600	0,2461
	5	0,0000	0,0008	0,0050	0,0165	0,0389	0,0735	0,1181	0,1672	0,2128	0,2461
	6	0,0000	0,0001	0,0006	0,0028	0,0087	0,0210	0,0424	0,0743	0,1160	0,1641
	7	0,0000	0,0000	0,0000	0,0003	0,0012	0,0039	0,0098	0,0212	0,0407	0,0703
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0035	0,0083	0,0176
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0020
10	0	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
	1	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0725	0,0403	0,0207	0,0098
	2	0,0746	0,1937	0,2759	0,3020	0,2816	0,2335	0,1757	0,1209	0,0763	0,0439
	3	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2522	0,2150	0,1665	0,1172
	4	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2377	0,2508	0,2384	0,2051
	5	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1536	0,2007	0,2340	0,2461
	6	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0689	0,1115	0,1596	0,2051
	7	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0212	0,0425	0,0746	0,1172
	8	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0043	0,0106	0,0229	0,0439
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016	0,0042	0,0098
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
11	0	0,5688	0,3138	0,1673	0,0859	0,0422	0,0198	0,0088	0,0036	0,0014	0,0005
	1	0,3293	0,3835	0,3248	0,2362	0,1549	0,0932	0,0518	0,0266	0,0125	0,0054
	2	0,0867	0,2131	0,2866	0,2953	0,2581	0,1998	0,1395	0,0887	0,0513	0,0269
	3	0,0137	0,0710	0,1517	0,2215	0,2581	0,2568	0,2254	0,1774	0,1259	0,0806
	4	0,0014	0,0158	0,0536	0,1107	0,1721	0,2201	0,2428	0,2365	0,2060	0,1611
	5	0,0001	0,0025	0,0132	0,0388	0,0803	0,1321	0,1830	0,2207	0,2360	0,2256
	6	0,0000	0,0003	0,0023	0,0097	0,0268	0,0566	0,0985	0,1471	0,1931	0,2256
	7	0,0000	0,0000	0,0003	0,0017	0,0064	0,0173	0,0379	0,0701	0,1128	0,1611
	8	0,0000	0,0000	0,0000	0,0002	0,0011	0,0037	0,0102	0,0234	0,0462	0,0806
	9	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018	0,0052	0,0126	0,0269
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0007	0,0021	0,0054
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0005
12	0	0,5404	0,2824	0,1422	0,0687	0,0317	0,0138	0,0057	0,0022	0,0008	0,0002
	1	0,3413	0,3766	0,3012	0,2062	0,1267	0,0712	0,0368	0,0174	0,0075	0,0029
	2	0,0988	0,2301	0,2924	0,2835	0,2323	0,1678	0,1088	0,0639	0,0339	0,0161
	3	0,0173	0,0852	0,1720	0,2362	0,2581	0,2397	0,1954	0,1419	0,0923	0,0537
	4	0,0021	0,0213	0,0683	0,1329	0,1936	0,2311	0,2367	0,2128	0,1700	0,1208
	5	0,0002	0,0038	0,0193	0,0532	0,1032	0,1585	0,2039	0,2270	0,2225	0,1934
	6	0,0000	0,0005	0,0040	0,0155	0,0401	0,0792	0,1281	0,1766	0,2124	0,2256
	7	0,0000	0,0000	0,0006	0,0033	0,0115	0,0291	0,0591	0,1009	0,1489	0,1934
	8	0,0000	0,0000	0,0001	0,0005	0,0024	0,0078	0,0199	0,0420	0,0762	0,1208
	9	0,0000	0,0000	0,0000	0,0001	0,0004	0,0015	0,0048	0,0125	0,0277	0,0537
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0025	0,0068	0,0161
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0029
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002

Tabla 1 (Continuación)

		π									
n	***k***	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
13	0	0,5133	0,2542	0,1209	0,0550	0,0238	0,0097	0,0037	0,0013	0,0004	0,0001
	1	0,3512	0,3672	0,2774	0,1787	0,1029	0,0540	0,0259	0,0113	0,0045	0,0016
	2	0,1109	0,2448	0,2937	0,2680	0,2059	0,1388	0,0836	0,0453	0,0220	0,0095
	3	0,0214	0,0997	0,1900	0,2457	0,2517	0,2181	0,1651	0,1107	0,0660	0,0349
	4	0,0028	0,0277	0,0838	0,1535	0,2097	0,2337	0,2222	0,1845	0,1350	0,0873
	5	0,0003	0,0055	0,0266	0,0691	0,1258	0,1803	0,2154	0,2214	0,1989	0,1571
	6	0,0000	0,0008	0,0063	0,0230	0,0559	0,1030	0,1546	0,1968	0,2169	0,2095
	7	0,0000	0,0001	0,0011	0,0058	0,0186	0,0442	0,0833	0,1312	0,1775	0,2095
	8	0,0000	0,0000	0,0001	0,0011	0,0047	0,0142	0,0336	0,0656	0,1089	0,1571
	9	0,0000	0,0000	0,0000	0,0001	0,0009	0,0034	0,0101	0,0243	0,0495	0,0873
	10	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0022	0,0065	0,0162	0,0349
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0012	0,0036	0,0095
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
14	0	0,4877	0,2288	0,1028	0,0440	0,0178	0,0068	0,0024	0,0008	0,0002	0,0001
	1	0,3593	0,3559	0,2539	0,1539	0,0832	0,0407	0,0181	0,0073	0,0027	0,0009
	2	0,1229	0,2570	0,2912	0,2501	0,1802	0,1134	0,0634	0,0317	0,0141	0,0056
	3	0,0259	0,1142	0,2056	0,2501	0,2402	0,1943	0,1366	0,0845	0,0462	0,0222
	4	0,0037	0,0349	0,0998	0,1720	0,2202	0,2290	0,2022	0,1549	0,1040	0,0611
	5	0,0004	0,0078	0,0352	0,0860	0,1468	0,1963	0,2178	0,2066	0,1701	0,1222
	6	0,0000	0,0013	0,0093	0,0322	0,0734	0,1262	0,1759	0,2066	0,2088	0,1833
	7	0,0000	0,0002	0,0019	0,0092	0,0280	0,0618	0,1082	0,1574	0,1952	0,2095
	8	0,0000	0,0000	0,0003	0,0020	0,0082	0,0232	0,0510	0,0918	0,1398	0,1833
	9	0,0000	0,0000	0,0000	0,0003	0,0018	0,0066	0,0183	0,0408	0,0762	0,1222
	10	0,0000	0,0000	0,0000	0,0000	0,0003	0,0014	0,0049	0,0136	0,0312	0,0611
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0033	0,0093	0,0222
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0019	0,0056
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0009
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
15	0	0,4633	0,2059	0,0874	0,0352	0,0134	0,0047	0,0016	0,0005	0,0001	0,0000
	1	0,3658	0,3432	0,2312	0,1319	0,0668	0,0305	0,0126	0,0047	0,0016	0,0005
	2	0,1348	0,2669	0,2856	0,2309	0,1559	0,0916	0,0476	0,0219	0,0090	0,0032
	3	0,0307	0,1285	0,2184	0,2501	0,2252	0,1700	0,1110	0,0634	0,0318	0,0139
	4	0,0049	0,0428	0,1156	0,1876	0,2252	0,2186	0,1792	0,1268	0,0780	0,0417
	5	0,0006	0,0105	0,0449	0,1032	0,1651	0,2061	0,2123	0,1859	0,1404	0,0916
	6	0,0000	0,0019	0,0132	0,0430	0,0917	0,1472	0,1906	0,2066	0,1914	0,1527
	7	0,0000	0,0003	0,0030	0,0138	0,0393	0,0811	0,1319	0,1771	0,2013	0,1964
	8	0,0000	0,0000	0,0005	0,0035	0,0131	0,0348	0,0710	0,1181	0,1647	0,1964
	9	0,0000	0,0000	0,0001	0,0007	0,0034	0,0116	0,0298	0,0612	0,1048	0,1527
	10	0,0000	0,0000	0,0000	0,0001	0,0007	0,0030	0,0096	0,0245	0,0515	0,0916
	11	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0074	0,0191	0,0417
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052	0,0139
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0032
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
16	0	0,4401	0,1853	0,0743	0,0281	0,0100	0,0033	0,0010	0,0003	0,0001	0,0000
	1	0,3706	0,3294	0,2097	0,1126	0,0535	0,0228	0,0087	0,0030	0,0009	0,0002
	2	0,1463	0,2745	0,2775	0,2111	0,1336	0,0732	0,0353	0,0150	0,0056	0,0018
	3	0,0359	0,1423	0,2285	0,2463	0,2079	0,1465	0,0888	0,0468	0,0215	0,0085

Tabla 1 (Continuación)

		π									
n	***k***	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
	4	0,0061	0,0514	0,1311	0,2001	0,2252	0,2040	0,1553	0,1014	0,0572	0,0278
	5	0,0008	0,0137	0,0555	0,1201	0,1802	0,2099	0,2008	0,1623	0,1123	0,0667
	6	0,0001	0,0028	0,0180	0,0550	0,1101	0,1649	0,1982	0,1983	0,1684	0,1222
	7	0,0000	0,0004	0,0045	0,0197	0,0524	0,1010	0,1524	0,1889	0,1969	0,1746
	8	0,0000	0,0001	0,0009	0,0055	0,0197	0,0487	0,0923	0,1417	0,1812	0,1964
	9	0,0000	0,0000	0,0001	0,0012	0,0058	0,0185	0,0442	0,0840	0,1318	0,1746
	10	0,0000	0,0000	0,0000	0,0002	0,0014	0,0056	0,0167	0,0392	0,0755	0,1222
	11	0,0000	0,0000	0,0000	0,0000	0,0002	0,0013	0,0049	0,0142	0,0337	0,0667
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0040	0,0115	0,0278
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0029	0,0085
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
17	0	0,4181	0,1668	0,0631	0,0225	0,0075	0,0023	0,0007	0,0002	0,0000	0,0000
	1	0,3741	0,3150	0,1893	0,0957	0,0426	0,0169	0,0060	0,0019	0,0005	0,0001
	2	0,1575	0,2800	0,2673	0,1914	0,1136	0,0581	0,0260	0,0102	0,0035	0,0010
	3	0,0415	0,1556	0,2359	0,2393	0,1893	0,1245	0,0701	0,0341	0,0144	0,0052
	4	0,0076	0,0605	0,1457	0,2093	0,2209	0,1868	0,1320	0,0796	0,0411	0,0182
	5	0,0010	0,0175	0,0668	0,1361	0,1914	0,2081	0,1849	0,1379	0,0875	0,0472
	6	0,0001	0,0039	0,0236	0,0680	0,1276	0,1784	0,1991	0,1839	0,1432	0,0944
	7	0,0000	0,0007	0,0065	0,0267	0,0668	0,1201	0,1685	0,1927	0,1841	0,1484
	8	0,0000	0,0001	0,0014	0,0084	0,0279	0,0644	0,1134	0,1606	0,1883	0,1855
	9	0,0000	0,0000	0,0003	0,0021	0,0093	0,0276	0,0611	0,1070	0,1540	0,1855
	10	0,0000	0,0000	0,0000	0,0004	0,0025	0,0095	0,0263	0,0571	0,1008	0,1484
	11	0,0000	0,0000	0,0000	0,0001	0,0005	0,0026	0,0090	0,0242	0,0525	0,0944
	12	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0081	0,0215	0,0472
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0021	0,0068	0,0182
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
18	0	0,3972	0,1501	0,0536	0,0180	0,0056	0,0016	0,0004	0,0001	0,0000	0,0000
	1	0,3763	0,3002	0,1704	0,0811	0,0338	0,0126	0,0042	0,0012	0,0003	0,0001
	2	0,1683	0,2835	0,2556	0,1723	0,0958	0,0458	0,0190	0,0069	0,0022	0,0006
	3	0,0473	0,1680	0,2406	0,2297	0,1704	0,1046	0,0547	0,0246	0,0095	0,0031
	4	0,0093	0,0700	0,1592	0,2153	0,2130	0,1681	0,1104	0,0614	0,0291	0,0117
	5	0,0014	0,0218	0,0787	0,1507	0,1988	0,2017	0,1664	0,1146	0,0666	0,0327
	6	0,0002	0,0052	0,0301	0,0816	0,1436	0,1873	0,1941	0,1655	0,1181	0,0708
	7	0,0000	0,0010	0,0091	0,0350	0,0820	0,1376	0,1792	0,1892	0,1657	0,1214
	8	0,0000	0,0002	0,0022	0,0120	0,0376	0,0811	0,1327	0,1734	0,1864	0,1669
	9	0,0000	0,0000	0,0004	0,0033	0,0139	0,0386	0,0794	0,1284	0,1694	0,1855
	10	0,0000	0,0000	0,0001	0,0008	0,0042	0,0149	0,0385	0,0771	0,1248	0,1669
	11	0,0000	0,0000	0,0000	0,0001	0,0010	0,0046	0,0151	0,0374	0,0742	0,1214
	12	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0047	0,0145	0,0354	0,0708
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0045	0,0134	0,0327
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0039	0,0117
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0009	0,0031
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006

Tabla 1 (Continuación)

		π									
n	***k***	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
19	0	0,3774	0,1351	0,0456	0,0144	0,0042	0,0011	0,0003	0,0001	0,0000	0,0000
	1	0,3774	0,2852	0,1529	0,0685	0,0268	0,0093	0,0029	0,0008	0,0002	0,0000
	2	0,1787	0,2852	0,2428	0,1540	0,0803	0,0358	0,0138	0,0046	0,0013	0,0003
	3	0,0533	0,1796	0,2428	0,2182	0,1517	0,0869	0,0422	0,0175	0,0062	0,0018
	4	0,0112	0,0798	0,1714	0,2182	0,2023	0,1491	0,0909	0,0467	0,0203	0,0074
	5	0,0018	0,0266	0,0907	0,1636	0,2023	0,1916	0,1468	0,0933	0,0497	0,0222
	6	0,0002	0,0069	0,0374	0,0955	0,1574	0,1916	0,1844	0,1451	0,0949	0,0518
	7	0,0000	0,0014	0,0122	0,0443	0,0974	0,1525	0,1844	0,1797	0,1443	0,0961
	8	0,0000	0,0002	0,0032	0,0166	0,0487	0,0981	0,1489	0,1797	0,1771	0,1442
	9	0,0000	0,0000	0,0007	0,0051	0,0198	0,0514	0,0980	0,1464	0,1771	0,1762
	10	0,0000	0,0000	0,0001	0,0013	0,0066	0,0220	0,0528	0,0976	0,1449	0,1762
	11	0,0000	0,0000	0,0000	0,0003	0,0018	0,0077	0,0233	0,0532	0,0970	0,1442
	12	0,0000	0,0000	0,0000	0,0000	0,0004	0,0022	0,0083	0,0237	0,0529	0,0961
	13	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0024	0,0085	0,0233	0,0518
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0082	0,0222
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0022	0,0074
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
20	0	0,3585	0,1216	0,0388	0,0115	0,0032	0,0008	0,0002	0,0000	0,0000	0,0000
	1	0,3774	0,2702	0,1368	0,0576	0,0211	0,0068	0,0020	0,0005	0,0001	0,0000
	2	0,1887	0,2852	0,2293	0,1369	0,0669	0,0278	0,0100	0,0031	0,0008	0,0002
	3	0,0596	0,1901	0,2428	0,2054	0,1339	0,0716	0,0323	0,0123	0,0040	0,0011
	4	0,0133	0,0898	0,1821	0,2182	0,1897	0,1304	0,0738	0,0350	0,0139	0,0046
	5	0,0022	0,0319	0,1028	0,1746	0,2023	0,1789	0,1272	0,0746	0,0365	0,0148
	6	0,0003	0,0089	0,0454	0,1091	0,1686	0,1916	0,1712	0,1244	0,0746	0,0370
	7	0,0000	0,0020	0,0160	0,0545	0,1124	0,1643	0,1844	0,1659	0,1221	0,0739
	8	0,0000	0,0004	0,0046	0,0222	0,0609	0,1144	0,1614	0,1797	0,1623	0,1201
	9	0,0000	0,0001	0,0011	0,0074	0,0271	0,0654	0,1158	0,1597	0,1771	0,1602
	10	0,0000	0,0000	0,0002	0,0020	0,0099	0,0308	0,0686	0,1171	0,1593	0,1762
	11	0,0000	0,0000	0,0000	0,0005	0,0030	0,0120	0,0336	0,0710	0,1185	0,1602
	12	0,0000	0,0000	0,0000	0,0001	0,0008	0,0039	0,0136	0,0355	0,0727	0,1201
	13	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0045	0,0146	0,0366	0,0739
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0049	0,0150	0,0370
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0049	0,0148
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0046
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002
	19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	20	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

* Para $\pi = 0{,}55, 0{,}60, \ldots, 0{,}95$, $P(X = k) = P(Y = n - k)$ donde Y es la distribución binomial con parámetros n y $1 - \pi$.

Tabla 2 **Probabilidades $P(X=k)=\dfrac{e^{-\mu}\mu^{k}}{k!}$ para la distribución de Poisson X con parámetro μ de 0,5 a 20 en intervalos de 0,5.**

	μ									
k	0,5	1,0	1,5	2,0	2,5	3,0	3,5	4,0	4,5	5,0
0	0,6065	0,3679	0,2231	0,1353	0,0821	0,0498	0,0302	0,0183	0,0111	0,0067
1	0,3033	0,3679	0,3347	0,2707	0,2052	0,1494	0,1057	0,0733	0,0500	0,0337
2	0,0758	0,1839	0,2510	0,2707	0,2565	0,2240	0,1850	0,1465	0,1125	0,0842
3	0,0126	0,0613	0,1255	0,1804	0,2138	0,2240	0,2158	0,1954	0,1687	0,1404
4	0,0016	0,0153	0,0471	0,0902	0,1336	0,1680	0,1888	0,1954	0,1898	0,1755
5	0,0002	0,0031	0,0141	0,0361	0,0668	0,1008	0,1322	0,1563	0,1708	0,1755
6	0,0000	0,0005	0,0035	0,0120	0,0278	0,0504	0,0771	0,1042	0,1281	0,1462
7	0,0000	0,0001	0,0008	0,0034	0,0099	0,0216	0,0385	0,0595	0,0824	0,1044
8	0,0000	0,0000	0,0001	0,0009	0,0031	0,0081	0,0169	0,0298	0,0463	0,0653
9	0,0000	0,0000	0,0000	0,0002	0,0009	0,0027	0,0066	0,0132	0,0232	0,0363
10	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0023	0,0053	0,0104	0,0181
11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0007	0,0019	0,0043	0,0082
12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0034
13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0013
14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0005
15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	5,5	6,0	6,5	7,0	7,5	8,0	8,5	9,0	9,5	10,0
0	0,0041	0,0025	0,0015	0,0009	0,0006	0,0003	0,0002	0,0001	0,0001	0,0000
1	0,0225	0,0149	0,0098	0,0064	0,0041	0,0027	0,0017	0,0011	0,0007	0,0005
2	0,0618	0,0446	0,0318	0,0223	0,0156	0,0107	0,0074	0,0050	0,0034	0,0023
3	0,1133	0,0892	0,0688	0,0521	0,0389	0,0286	0,0208	0,0150	0,0107	0,0076
4	0,1558	0,1339	0,1118	0,0912	0,0729	0,0573	0,0443	0,0337	0,0254	0,0189
5	0,1714	0,1606	0,1454	0,1277	0,1094	0,0916	0,0752	0,0607	0,0483	0,0378
6	0,1571	0,1606	0,1575	0,1490	0,1367	0,1221	0,1066	0,0911	0,0764	0,0631
7	0,1234	0,1377	0,1462	0,1490	0,1465	0,1396	0,1294	0,1171	0,1037	0,0901
8	0,0849	0,1033	0,1188	0,1304	0,1373	0,1396	0,1375	0,1318	0,1232	0,1126
9	0,0519	0,0688	0,0858	0,1014	0,1144	0,1241	0,1299	0,1318	0,1300	0,1251
10	0,0285	0,0413	0,0558	0,0710	0,0858	0,0993	0,1104	0,1186	0,1235	0,1251
11	0,0143	0,0225	0,0330	0,0452	0,0585	0,0722	0,0853	0,0970	0,1067	0,1137
12	0,0065	0,0113	0,0179	0,0263	0,0366	0,0481	0,0604	0,0728	0,0844	0,0948
13	0,0028	0,0052	0,0089	0,0142	0,0211	0,0296	0,0395	0,0504	0,0617	0,0729
14	0,0011	0,0022	0,0041	0,0071	0,0113	0,0169	0,0240	0,0324	0,0419	0,0521
15	0,0004	0,0009	0,0018	0,0033	0,0057	0,0090	0,0136	0,0194	0,0265	0,0347
16	0,0001	0,0003	0,0007	0,0014	0,0026	0,0045	0,0072	0,0109	0,0157	0,0217
17	0,0000	0,0001	0,0003	0,0006	0,0012	0,0021	0,0036	0,0058	0,0088	0,0128
18	0,0000	0,0000	0,0001	0,0002	0,0005	0,0009	0,0017	0,0029	0,0046	0,0071
19	0,0000	0,0000	0,0000	0,0001	0,0002	0,0004	0,0008	0,0014	0,0023	0,0037
20	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0006	0,0011	0,0019
21	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0003	0,0005	0,0009
22	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0002	0,0004
23	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
24	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
25	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Tabla 2 (Continuación)

	μ									
k	10,5	11,0	11,5	12,0	12,5	13,0	13,5	14,0	14,5	15,0
0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
1	0,0003	0,0002	0,0001	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
2	0,0015	0,0010	0,0007	0,0004	0,0003	0,0002	0,0001	0,0001	0,0001	0,0000
3	0,0053	0,0037	0,0026	0,0018	0,0012	0,0008	0,0006	0,0004	0,0003	0,0002
4	0,0139	0,0102	0,0074	0,0053	0,0038	0,0027	0,0019	0,0013	0,0009	0,0006
5	0,0293	0,0224	0,0170	0,0127	0,0095	0,0070	0,0051	0,0037	0,0027	0,0019
6	0,0513	0,0411	0,0325	0,0255	0,0197	0,0152	0,0115	0,0087	0,0065	0,0048
7	0,0769	0,0646	0,0535	0,0437	0,0353	0,0281	0,0222	0,0174	0,0135	0,0104
8	0,1009	0,0888	0,0769	0,0655	0,0551	0,0457	0,0375	0,0304	0,0244	0,0194
9	0,1177	0,1085	0,0982	0,0874	0,0765	0,0661	0,0563	0,0473	0,0394	0,0324
10	0,1236	0,1194	0,1129	0,1048	0,0956	0,0859	0,0760	0,0663	0,0571	0,0486
11	0,1180	0,1194	0,1181	0,1144	0,1087	0,1015	0,0932	0,0844	0,0753	0,0663
12	0,1032	0,1094	0,1131	0,1144	0,1132	0,1099	0,1049	0,0984	0,0910	0,0829
13	0,0834	0,0926	0,1001	0,1056	0,1089	0,1099	0,1089	0,1060	0,1014	0,0956
14	0,0625	0,0728	0,0822	0,0905	0,0972	0,1021	0,1050	0,1060	0,1051	0,1024
15	0,0438	0,0534	0,0630	0,0724	0,0810	0,0885	0,0945	0,0989	0,1016	0,1024
16	0,0287	0,0367	0,0453	0,0543	0,0633	0,0719	0,0798	0,0866	0,0920	0,0960
17	0,0177	0,0237	0,0306	0,0383	0,0465	0,0550	0,0633	0,0713	0,0785	0,0847
18	0,0104	0,0145	0,0196	0,0255	0,0323	0,0397	0,0475	0,0554	0,0632	0,0706
19	0,0057	0,0084	0,0119	0,0161	0,0213	0,0272	0,0337	0,0409	0,0483	0,0557
20	0,0030	0,0046	0,0068	0,0097	0,0133	0,0177	0,0228	0,0286	0,0350	0,0418
21	0,0015	0,0024	0,0037	0,0055	0,0079	0,0109	0,0146	0,0191	0,0242	0,0299
22	0,0007	0,0012	0,0020	0,0030	0,0045	0,0065	0,0090	0,0121	0,0159	0,0204
23	0,0003	0,0006	0,0010	0,0016	0,0024	0,0037	0,0053	0,0074	0,0100	0,0133
24	0,0001	0,0003	0,0005	0,0008	0,0013	0,0020	0,0030	0,0043	0,0061	0,0083
25	0,0001	0,0001	0,0002	0,0004	0,0006	0,0010	0,0016	0,0024	0,0035	0,0050
26	0,0000	0,0000	0,0001	0,0002	0,0003	0,0005	0,0008	0,0013	0,0020	0,0029
27	0,0000	0,0000	0,0000	0,0001	0,0001	0,0002	0,0004	0,0007	0,0011	0,0016
28	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0002	0,0003	0,0005	0,0009
29	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0002	0,0003	0,0004
30	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0002
31	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001
32	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
33	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	15,5	16,0	16,5	17,0	17,5	18,0	18,5	19,0	19,5	20,0
0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
1	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
2	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
3	0,0001	0,0001	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
4	0,0004	0,0003	0,0002	0,0001	0,0001	0,0001	0,0000	0,0000	0,0000	0,0000
5	0,0014	0,0010	0,0007	0,0005	0,0003	0,0002	0,0002	0,0001	0,0001	0,0001
6	0,0036	0,0026	0,0019	0,0014	0,0010	0,0007	0,0005	0,0004	0,0003	0,0002
7	0,0079	0,0060	0,0045	0,0034	0,0025	0,0019	0,0014	0,0010	0,0007	0,0005
8	0,0153	0,0120	0,0093	0,0072	0,0055	0,0042	0,0031	0,0024	0,0018	0,0013
9	0,0264	0,0213	0,0171	0,0135	0,0107	0,0083	0,0065	0,0050	0,0038	0,0029
10	0,0409	0,0341	0,0281	0,0230	0,0186	0,0150	0,0120	0,0095	0,0074	0,0058
11	0,0577	0,0496	0,0422	0,0355	0,0297	0,0245	0,0201	0,0164	0,0132	0,0106
12	0,0745	0,0661	0,0580	0,0504	0,0432	0,0368	0,0310	0,0259	0,0214	0,0176

Tabla 2 (Continuación)

	μ									
k	15,5	16,0	16,5	17,0	17,5	18,0	18,5	19,0	19,5	20,0
13	0,0888	0,0814	0,0736	0,0658	0,0582	0,0509	0,0441	0,0378	0,0322	0,0271
14	0,0983	0,0930	0,0868	0,0800	0,0728	0,0655	0,0583	0,0514	0,0448	0,0387
15	0,1016	0,0992	0,0955	0,0906	0,0849	0,0786	0,0719	0,0650	0,0582	0,0516
16	0,0984	0,0992	0,0985	0,0963	0,0929	0,0884	0,0831	0,0772	0,0710	0,0646
17	0,0897	0,0934	0,0956	0,0963	0,0956	0,0936	0,0904	0,0863	0,0814	0,0760
18	0,0773	0,0830	0,0876	0,0909	0,0929	0,0936	0,0930	0,0911	0,0882	0,0844
19	0,0630	0,0699	0,0761	0,0814	0,0856	0,0887	0,0905	0,0911	0,0905	0,0888
20	0,0489	0,0559	0,0628	0,0692	0,0749	0,0798	0,0837	0,0866	0,0883	0,0888
21	0,0361	0,0426	0,0493	0,0560	0,0624	0,0684	0,0738	0,0783	0,0820	0,0846
22	0,0254	0,0310	0,0370	0,0433	0,0496	0,0560	0,0620	0,0676	0,0727	0,0769
23	0,0171	0,0216	0,0265	0,0320	0,0378	0,0438	0,0499	0,0559	0,0616	0,0669
24	0,0111	0,0144	0,0182	0,0226	0,0275	0,0328	0,0385	0,0442	0,0500	0,0557
25	0,0069	0,0092	0,0120	0,0154	0,0193	0,0237	0,0285	0,0336	0,0390	0,0446
26	0,0041	0,0057	0,0076	0,0101	0,0130	0,0164	0,0202	0,0246	0,0293	0,0343
27	0,0023	0,0034	0,0047	0,0063	0,0084	0,0109	0,0139	0,0173	0,0211	0,0254
28	0,0013	0,0019	0,0028	0,0038	0,0053	0,0070	0,0092	0,0117	0,0147	0,0181
29	0,0007	0,0011	0,0016	0,0023	0,0032	0,0044	0,0058	0,0077	0,0099	0,0125
30	0,0004	0,0006	0,0009	0,0013	0,0019	0,0026	0,0036	0,0049	0,0064	0,0083
31	0,0002	0,0003	0,0005	0,0007	0,0010	0,0015	0,0022	0,0030	0,0040	0,0054
32	0,0001	0,0001	0,0002	0,0004	0,0006	0,0009	0,0012	0,0018	0,0025	0,0034
33	0,0000	0,0001	0,0001	0,0002	0,0003	0,0005	0,0007	0,0010	0,0015	0,0020
34	0,0000	0,0000	0,0001	0,0001	0,0002	0,0002	0,0004	0,0006	0,0008	0,0012
35	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0002	0,0003	0,0005	0,0007
36	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0002	0,0003	0,0004
37	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0001	0,0002
38	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001
39	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
40	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Tabla 3 Función de distribución normal estandarizada $\Phi(z) = P(Z \leq z)$ para valores z de 0 a 3,99 en intervalos de 0,01.*

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,10	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,20	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,30	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,40	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,50	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,60	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,70	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,80	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,90	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,00	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,10	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,20	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,30	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,40	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,50	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,60	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,70	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,80	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,90	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,00	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,10	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,20	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,30	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,40	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,50	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,60	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,70	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,80	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,90	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,00	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,10	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,20	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,30	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,40	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,50	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,60	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,70	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,80	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,90	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

* Para valores z negativos, $\Phi(z) = P(Z \leq z) = P(Z \geq -z) = 1 - P(Z \leq -z) = 1 - \Phi(-z)$.

Tabla 4 Tabla de 1000 dígitos aleatorios.

	1–5	6–10	11–15	16–20	21–25	26–30	31–35	36–40	41–45	46–50
1	28068	97497	24717	94945	71584	46975	80676	37564	85194	26562
2	77798	61589	36980	18859	78471	07605	41910	98737	97310	76984
3	33911	76198	97068	89844	07886	96716	18354	66921	85958	59963
4	45302	20953	65158	70637	42792	85207	32911	93401	90088	88104
5	31759	68429	61028	00200	02062	92555	82037	69832	74185	76010
6	81262	04831	92203	25447	65875	71086	12676	42753	79223	63135
7	27510	88900	41437	07409	87437	79309	83499	50721	40752	82801
8	84888	90443	23200	86340	07731	64171	76935	02931	66982	30842
9	92551	42420	29984	87522	19370	30357	33530	58101	59423	91700
10	48644	97274	33475	71381	27387	50740	03176	96910	94049	65052
11	71226	14223	27559	00943	46943	40680	96829	09265	94401	98461
12	59902	65129	28077	80487	79160	56426	47978	08556	20753	10206
13	24973	51863	86605	16991	58423	33341	70147	06005	81833	00868
14	27005	74018	05569	70982	80438	76901	80061	11144	91733	07228
15	25651	65765	98249	24231	32819	26680	17613	29917	47814	92539
16	34255	68331	66861	37285	34606	68167	55636	70101	51328	57528
17	74791	18769	92325	19959	90031	27008	25857	68520	41469	45100
18	63485	89564	62107	80055	08094	85412	33589	71900	05892	63260
19	99762	44503	91645	15352	25957	73662	71146	26161	98418	10195
20	85157	99008	25927	31118	65466	48706	20302	26133	04751	34701

Tabla 5 Percentiles de la distribución *t* de Student para distintos grados de libertad.

Grados de libertad	Percentil								
	0,75	0,80	0,85	0,90	0,95	0,975	0,99	0,995	0,9995
1	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,599
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,768
24	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660	3,460
120	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Tabla 6 Percentiles de la distribución chi-cuadrado para distintos grados de libertad *d*.

	Percentil												
d	**0,005**	**0,01**	**0,025**	**0,05**	**0,10**	**0,25**	**0,50**	**0,75**	**0,90**	**0,95**	**0,975**	**0,99**	**0,995**
1	0,000	0,0002	0,001	0,004	0,02	0,10	0,45	1,32	2,71	3,84	5,02	6,63	7,88
2	0,01	0,02	0,05	0,10	0,21	0,58	1,39	2,77	4,61	5,99	7,38	9,21	10,60
3	0,07	0,11	0,22	0,35	0,58	1,21	2,37	4,11	6,25	7,81	9,35	11,34	12,84
4	0,21	0,30	0,48	0,71	1,06	1,92	3,36	5,39	7,78	9,49	11,14	13,28	14,86
5	0,41	0,55	0,83	1,15	1,61	2,67	4,35	6,63	9,24	11,07	12,83	15,09	16,75
6	0,68	0,87	1,24	1,64	2,20	3,45	5,35	7,84	10,64	12,59	14,45	16,81	18,55
7	0,99	1,24	1,69	2,17	2,83	4,25	6,35	9,04	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	5,07	7,34	10,22	13,36	15,51	17,53	20,09	21,95
9	1,73	2,09	2,70	3,33	4,17	5,90	8,34	11,39	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	6,74	9,34	12,55	15,99	18,31	20,48	23,21	25,19
11	2,60	3,05	3,82	4,57	5,58	7,58	10,34	13,70	17,28	19,68	21,92	24,72	26,76
12	3,07	3,57	4,40	5,23	6,30	8,44	11,34	14,85	18,55	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	7,04	9,30	12,34	15,98	19,81	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	7,79	10,17	13,34	17,12	21,06	23,68	26,12	29,14	31,32
15	4,60	5,23	6,26	7,26	8,55	11,04	14,34	18,25	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	9,31	11,91	15,34	19,37	23,54	26,30	28,85	32,00	34,27
17	5,70	6,41	7,56	8,67	10,09	12,79	16,34	20,49	24,77	27,59	30,19	33,41	35,72
18	6,26	7,01	8,23	9,39	10,86	13,68	17,34	21,60	25,99	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	11,65	14,56	18,34	22,72	27,20	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	12,44	15,45	19,34	23,83	28,41	31,41	34,17	37,57	40,00
21	8,03	8,90	10,28	11,59	13,24	16,34	20,34	24,93	29,62	32,67	35,48	38,93	41,40
22	8,64	9,54	10,98	12,34	14,04	17,24	21,34	26,04	30,81	33,92	36,78	40,29	42,80
23	9,26	10,20	11,69	13,09	14,85	18,14	22,34	27,14	32,01	35,17	38,08	41,64	44,18
24	9,89	10,86	12,40	13,85	15,66	19,04	23,34	28,24	33,20	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	16,47	19,94	24,34	29,34	34,38	37,65	40,65	44,31	46,93
26	11,16	12,20	13,84	15,38	17,29	20,84	25,34	30,43	35,56	38,89	41,92	45,64	48,29
27	11,81	12,88	14,57	16,15	18,11	21,75	26,34	31,53	36,74	40,11	43,19	46,96	49,64
28	12,46	13,56	15,31	16,93	18,94	22,66	27,34	32,62	37,92	41,34	44,46	48,28	50,99
29	13,12	14,26	16,05	17,71	19,77	23,57	28,34	33,71	39,09	42,56	45,72	49,59	52,34
30	13,79	14,95	16,79	18,49	20,60	24,48	29,34	34,80	40,26	43,77	46,98	50,89	53,67
35	17,19	18,51	20,57	22,47	24,80	29,05	34,34	40,22	46,06	49,80	53,20	57,34	60,27
40	20,71	22,16	24,43	26,51	29,05	33,66	39,34	45,62	51,81	55,76	59,34	63,69	66,77
45	24,31	25,90	28,37	30,61	33,35	38,29	44,34	50,98	57,51	61,66	65,41	69,96	73,17
50	27,99	29,71	32,36	34,76	37,69	42,94	49,33	56,33	63,17	67,50	71,42	76,15	79,49
60	35,53	37,48	40,48	43,19	46,46	52,29	59,33	66,98	74,40	79,08	83,30	88,38	91,95
70	43,28	45,44	48,76	51,74	55,33	61,70	69,33	77,58	85,53	90,53	95,02	100,43	104,21
80	51,17	53,54	57,15	60,39	64,28	71,14	79,33	88,13	96,58	101,88	106,63	112,33	116,32
90	59,20	61,75	65,65	69,13	73,29	80,62	89,33	98,65	107,57	113,15	118,14	124,12	128,30
100	67,33	70,06	74,22	77,93	82,36	90,13	99,33	109,14	118,50	124,34	129,56	135,81	140,17

Tabla 7 Percentiles de la distribución *F* de Fisher para distintos grados de libertad del numerador d_1 y del denominador d_2.*

		d_1											
d_2	Percentil	1	2	3	4	5	6	8	10	15	20	30	∞
1	0,90	39,86	49,50	53,59	55,83	57,24	58,20	59,44	60,19	61,22	61,74	62,26	63,33
	0,95	161,45	199,50	215,71	224,58	230,16	233,99	238,88	241,88	245,95	248,01	250,10	254,31
	0,975	647,79	799,50	864,16	899,58	921,85	937,11	956,66	968,63	984,87	993,10	1001,4	1018,3
	0,99	4052,2	4999,5	5403,4	5624,6	5763,7	5859,0	5981,1	6055,9	6157,3	6208,7	6260,7	6365,9
	0,995	16211	20000	21615	22500	23056	23437	23925	24224	24630	24836	25044	25464
2	0,90	8,53	9,00	9,16	9,24	9,29	9,33	9,37	9,39	9,42	9,44	9,46	9,49
	0,95	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,40	19,43	19,45	19,46	19,50
	0,975	38,51	39,00	39,17	39,25	39,30	39,33	39,37	39,40	39,43	39,45	39,46	39,50
	0,99	98,50	99,00	99,17	99,25	99,30	99,33	99,37	99,40	99,43	99,45	99,47	99,50
	0,995	198,50	199,00	199,17	199,25	199,30	199,33	199,37	199,40	199,43	199,45	199,47	199,50
3	0,90	5,54	5,46	5,39	5,34	5,31	5,28	5,25	5,23	5,20	5,18	5,17	5,13
	0,95	10,13	9,55	9,28	9,12	9,01	8,94	8,85	8,79	8,70	8,66	8,62	8,53
	0,975	17,44	16,04	15,44	15,10	14,88	14,73	14,54	14,42	14,25	14,17	14,08	13,90
	0,99	34,12	30,82	29,46	28,71	28,24	27,91	27,49	27,23	26,87	26,69	26,50	26,13
	0,995	55,55	49,80	47,47	46,19	45,39	44,84	44,13	43,69	43,08	42,78	42,47	41,83
4	0,90	4,54	4,32	4,19	4,11	4,05	4,01	3,95	3,92	3,87	3,84	3,82	3,76
	0,95	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,96	5,86	5,80	5,75	5,63
	0,975	12,22	10,65	9,98	9,60	9,36	9,20	8,98	8,84	8,66	8,56	8,46	8,26
	0,99	21,20	18,00	16,69	15,98	15,52	15,21	14,80	14,55	14,20	14,02	13,84	13,46
	0,995	31,33	26,28	24,26	23,15	22,46	21,97	21,35	20,97	20,44	20,17	19,89	19,32
5	0,90	4,06	3,78	3,62	3,52	3,45	3,40	3,34	3,30	3,24	3,21	3,17	3,10
	0,95	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,74	4,62	4,56	4,50	4,36
	0,975	10,01	8,43	7,76	7,39	7,15	6,98	6,76	6,62	6,43	6,33	6,23	6,02
	0,99	16,26	13,27	12,06	11,39	10,97	10,67	10,29	10,05	9,72	9,55	9,38	9,02
	0,995	22,78	18,31	16,53	15,56	14,94	14,51	13,96	13,62	13,15	12,90	12,66	12,14
6	0,90	3,78	3,46	3,29	3,18	3,11	3,05	2,98	2,94	2,87	2,84	2,80	2,72
	0,95	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,06	3,94	3,87	3,81	3,67
	0,975	8,81	7,26	6,60	6,23	5,99	5,82	5,60	5,46	5,27	5,17	5,07	4,85
	0,99	13,75	10,92	9,78	9,15	8,75	8,47	8,10	7,87	7,56	7,40	7,23	6,88
	0,995	18,63	14,54	12,92	12,03	11,46	11,07	10,57	10,25	9,81	9,59	9,36	8,88
7	0,90	3,59	3,26	3,07	2,96	2,88	2,83	2,75	2,70	2,63	2,59	2,56	2,47
	0,95	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,64	3,51	3,44	3,38	3,23
	0,975	8,07	6,54	5,89	5,52	5,29	5,12	4,90	4,76	4,57	4,47	4,36	4,14
	0,99	12,25	9,55	8,45	7,85	7,46	7,19	6,84	6,62	6,31	6,16	5,99	5,65
	0,995	16,24	12,40	10,88	10,05	9,52	9,16	8,68	8,38	7,97	7,75	7,53	7,08
8	0,90	3,46	3,11	2,92	2,81	2,73	2,67	2,59	2,54	2,46	2,42	2,38	2,29
	0,95	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,35	3,22	3,15	3,08	2,93
	0,975	7,57	6,06	5,42	5,05	4,82	4,65	4,43	4,30	4,10	4,00	3,89	3,67
	0,99	11,26	8,65	7,59	7,01	6,63	6,37	6,03	5,81	5,52	5,36	5,20	4,86
	0,995	14,69	11,04	9,60	8,81	8,30	7,95	7,50	7,21	6,81	6,61	6,40	5,95
9	0,90	3,36	3,01	2,81	2,69	2,61	2,55	2,47	2,42	2,34	2,30	2,25	2,16
	0,95	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,14	3,01	2,94	2,86	2,71
	0,975	7,21	5,71	5,08	4,72	4,48	4,32	4,10	3,96	3,77	3,67	3,56	3,33
	0,99	10,56	8,02	6,99	6,42	6,06	5,80	5,47	5,26	4,96	4,81	4,65	4,31
	0,995	13,61	10,11	8,72	7,96	7,47	7,13	6,69	6,42	6,03	5,83	5,62	5,19
10	0,90	3,29	2,92	2,73	2,61	2,52	2,46	2,38	2,32	2,24	2,20	2,16	2,06
	0,95	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,98	2,85	2,77	2,70	2,54
	0,975	6,94	5,46	4,83	4,47	4,24	4,07	3,85	3,72	3,52	3,42	3,31	3,08
	0,99	10,04	7,56	6,55	5,99	5,64	5,39	5,06	4,85	4,56	4,41	4,25	3,91
	0,995	12,83	9,43	8,08	7,34	6,87	6,54	6,12	5,85	5,47	5,27	5,07	4,64
12	0,90	3,18	2,81	2,61	2,48	2,39	2,33	2,24	2,19	2,10	2,06	2,01	1,90
	0,95	4,75	3,89	3,49	3,26	3,11	3,00	2,85	2,75	2,62	2,54	2,47	2,30
	0,975	6,55	5,10	4,47	4,12	3,89	3,73	3,51	3,37	3,18	3,07	2,96	2,72

Tabla 7 (Continuación)

		d_1											
d_2	Percentil	1	2	3	4	5	6	8	10	15	20	30	∞
	0,99	9,33	6,93	5,95	5,41	5,06	4,82	4,50	4,30	4,01	3,86	3,70	3,36
	0,995	11,75	8,51	7,23	6,52	6,07	5,76	5,35	5,09	4,72	4,53	4,33	3,90
14	0,90	3,10	2,73	2,52	2,39	2,31	2,24	2,15	2,10	2,01	1,96	1,91	1,80
	0,95	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,60	2,46	2,39	2,31	2,13
	0,975	6,30	4,86	4,24	3,89	3,66	3,50	3,29	3,15	2,95	2,84	2,73	2,49
	0,99	8,86	6,51	5,56	5,04	4,69	4,46	4,14	3,94	3,66	3,51	3,35	3,00
	0,995	11,06	7,92	6,68	6,00	5,56	5,26	4,86	4,60	4,25	4,06	3,86	3,44
16	0,90	3,05	2,67	2,46	2,33	2,24	2,18	2,09	2,03	1,94	1,89	1,84	1,72
	0,95	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,49	2,35	2,28	2,19	2,01
	0,975	6,12	4,69	4,08	3,73	3,50	3,34	3,12	2,99	2,79	2,68	2,57	2,32
	0,99	8,53	6,23	5,29	4,77	4,44	4,20	3,89	3,69	3,41	3,26	3,10	2,75
	0,995	10,58	7,51	6,30	5,64	5,21	4,91	4,52	4,27	3,92	3,73	3,54	3,11
18	0,90	3,01	2,62	2,42	2,29	2,20	2,13	2,04	1,98	1,89	1,84	1,78	1,66
	0,95	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,41	2,27	2,19	2,11	1,92
	0,975	5,98	4,56	3,95	3,61	3,38	3,22	3,01	2,87	2,67	2,56	2,44	2,19
	0,99	8,29	6,01	5,09	4,58	4,25	4,01	3,71	3,51	3,23	3,08	2,92	2,57
	0,995	10,22	7,21	6,03	5,37	4,96	4,66	4,28	4,03	3,68	3,50	3,30	2,87
20	0,90	2,97	2,59	2,38	2,25	2,16	2,09	2,00	1,94	1,84	1,79	1,74	1,61
	0,95	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,35	2,20	2,12	2,04	1,84
	0,975	5,87	4,46	3,86	3,51	3,29	3,13	2,91	2,77	2,57	2,46	2,35	2,09
	0,99	8,10	5,85	4,94	4,43	4,10	3,87	3,56	3,37	3,09	2,94	2,78	2,42
	0,995	9,94	6,99	5,82	5,17	4,76	4,47	4,09	3,85	3,50	3,32	3,12	2,69
25	0,90	2,92	2,53	2,32	2,18	2,09	2,02	1,93	1,87	1,77	1,72	1,66	1,52
	0,95	4,24	3,39	2,99	2,76	2,60	2,49	2,34	2,24	2,09	2,01	1,92	1,71
	0,975	5,69	4,29	3,69	3,35	3,13	2,97	2,75	2,61	2,41	2,30	2,18	1,91
	0,99	7,77	5,57	4,68	4,18	3,85	3,63	3,32	3,13	2,85	2,70	2,54	2,17
	0,995	9,48	6,60	5,46	4,84	4,43	4,15	3,78	3,54	3,20	3,01	2,82	2,38
30	0,90	2,88	2,49	2,28	2,14	2,05	1,98	1,88	1,82	1,72	1,67	1,61	1,46
	0,95	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,16	2,01	1,93	1,84	1,62
	0,975	5,57	4,18	3,59	3,25	3,03	2,87	2,65	2,51	2,31	2,20	2,07	1,79
	0,99	7,56	5,39	4,51	4,02	3,70	3,47	3,17	2,98	2,70	2,55	2,39	2,01
	0,995	9,18	6,35	5,24	4,62	4,23	3,95	3,58	3,34	3,01	2,82	2,63	2,18
35	0,90	2,85	2,46	2,25	2,11	2,02	1,95	1,85	1,79	1,69	1,63	1,57	1,41
	0,95	4,12	3,27	2,87	2,64	2,49	2,37	2,22	2,11	1,96	1,88	1,79	1,56
	0,975	5,48	4,11	3,52	3,18	2,96	2,80	2,58	2,44	2,23	2,12	2,00	1,70
	0,99	7,42	5,27	4,40	3,91	3,59	3,37	3,07	2,88	2,60	2,44	2,28	1,89
	0,995	8,98	6,19	5,09	4,48	4,09	3,81	3,45	3,21	2,88	2,69	2,50	2,04
40	0,90	2,84	2,44	2,23	2,09	2,00	1,93	1,83	1,76	1,66	1,61	1,54	1,38
	0,95	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,08	1,92	1,84	1,74	1,51
	0,975	5,42	4,05	3,46	3,13	2,90	2,74	2,53	2,39	2,18	2,07	1,94	1,64
	0,99	7,31	5,18	4,31	3,83	3,51	3,29	2,99	2,80	2,52	2,37	2,20	1,80
	0,995	8,83	6,07	4,98	4,37	3,99	3,71	3,35	3,12	2,78	2,60	2,40	1,93
60	0,90	2,79	2,39	2,18	2,04	1,95	1,87	1,77	1,71	1,60	1,54	1,48	1,29
	0,95	4,00	3,15	2,76	2,53	2,37	2,25	2,10	1,99	1,84	1,75	1,65	1,39
	0,975	5,29	3,93	3,34	3,01	2,79	2,63	2,41	2,27	2,06	1,94	1,82	1,48
	0,99	7,08	4,98	4,13	3,65	3,34	3,12	2,82	2,63	2,35	2,20	2,03	1,60
	0,995	8,49	5,79	4,73	4,14	3,76	3,49	3,13	2,90	2,57	2,39	2,19	1,69
120	0,90	2,75	2,35	2,13	1,99	1,90	1,82	1,72	1,65	1,55	1,48	1,41	1,19
	0,95	3,92	3,07	2,68	2,45	2,29	2,18	2,02	1,91	1,75	1,66	1,55	1,25
	0,975	5,15	3,80	3,23	2,89	2,67	2,52	2,30	2,16	1,94	1,82	1,69	1,31
	0,99	6,85	4,79	3,95	3,48	3,17	2,96	2,66	2,47	2,19	2,03	1,86	1,38
	0,995	8,18	5,54	4,50	3,92	3,55	3,28	2,93	2,71	2,37	2,19	1,98	1,43

Tabla 7 (Continuación)

		d_1											
d_2	Percentil	1	2	3	4	5	6	8	10	15	20	30	∞
∞	0,90	2,71	2,30	2,08	1,94	1,85	1,77	1,67	1,60	1,49	1,42	1,34	1,00
	0,95	3,84	3,00	2,60	2,37	2,21	2,10	1,94	1,83	1,67	1,57	1,46	1,00
	0,975	5,02	3,69	3,12	2,79	2,57	2,41	2,19	2,05	1,83	1,71	1,57	1,00
	0,99	6,63	4,61	3,78	3,32	3,02	2,80	2,51	2,32	2,04	1,88	1,70	1,00
	0,995	7,88	5,30	4,28	3,72	3,35	3,09	2,74	2,52	2,19	2,00	1,79	1,00

* Para percentiles inferiores $\alpha = 0{,}005$, 0,01, 0,025, 0,05 y 0,10, $F_{d_1,d_2,\alpha} = 1/F_{d_2,d_1,1-\alpha}$.

Tabla 8 Percentiles de la distribución bajo H_0 de la suma de rangos de Wilcoxon $U = \sum_{i=1}^{n_1} r_i$ en la muestra de menor tamaño $n_1 \leq n_2$ para $n_1 = 3, 4, ..., 8$.*

	Percentil 0,95 n_1						Percentil 0,975 n_1					
n_2	3	4	5	6	7	8	3	4	5	6	7	8
3	14						15					
4	17	24					18	25				
5	19	27	35				20	28	37			
6	21	30	39	49			22	31	41	51		
7	24	33	43	54	65		25	34	44	56	68	
8	26	36	46	58	70	84	27	37	48	60	73	86
9	28	39	50	62	75	89	30	41	52	64	78	92
10	31	42	53	66	80	95	32	44	56	69	83	98
11	33	45	57	70	85	100	35	47	60	73	88	104
12	36	48	61	75	90	105	37	50	63	78	93	109
13	38	51	64	79	94	111	40	53	67	82	98	115
14	40	54	68	83	99	116	42	56	71	87	103	121
15	43	57	71	87	104	122	45	59	75	91	108	126
16	45	59	75	91	109	127	47	62	79	95	113	132
17	47	62	79	96	113	132	50	66	82	100	118	137
18	50	65	82	100	118	138	52	69	86	104	123	143
19	52	68	86	104	123	143	55	72	90	109	128	149
20	54	71	89	108	128	148	57	75	94	113	133	154
21	57	74	93	112	133	154	60	78	97	117	138	160
22	59	77	96	116	137	159	62	81	101	122	143	166
23	61	80	100	121	142	165	65	84	105	126	148	171
24	64	83	104	125	147	170	67	88	109	131	153	177
25	66	86	107	129	152	175	70	91	112	135	158	182
26	68	89	111	133	156	181	72	94	116	139	163	188
27	71	92	114	137	161	186	75	97	120	144	168	194
28	73	95	118	142	166	191	77	100	124	148	173	199
29	75	98	121	146	171	197	79	103	127	152	178	205
30	78	101	125	150	176	202	82	106	131	157	183	210
31	80	104	129	154	180	207	84	109	135	161	188	216
32	82	107	132	158	185	213	87	113	139	166	193	222
33	85	110	136	162	190	218	89	116	142	170	198	227
34	87	113	139	167	195	223	92	119	146	174	203	233
35	89	116	143	171	199	229	94	122	150	179	208	238
36	92	119	146	175	204	234	97	125	154	183	213	244
37	94	122	150	179	209	240	99	128	158	187	218	250
38	96	125	154	183	214	245	102	131	161	192	223	255
39	99	127	157	187	218	250	104	134	165	196	228	261
40	101	130	161	192	223	256	107	138	169	201	233	266
41	103	133	164	196	228	261	109	141	173	205	238	272
42	106	136	168	200	233	266	112	144	176	209	243	278
43	108	139	171	204	237	272	114	147	180	214	248	283
44	110	142	175	208	242	277	117	150	184	218	253	289
45	113	145	179	212	247	282	119	153	188	223	258	294
46	115	148	182	217	252	288	121	156	191	227	263	300
47	117	151	186	221	257	293	124	159	195	231	268	306
48	119	154	189	225	261	298	126	162	199	236	273	311
49	122	157	193	229	266	304	129	166	203	240	278	317

Tabla 8 (Continuación)

	Percentil 0,99 n_1						Percentil 0,995 n_1					
n_2	3	4	5	6	7	8	3	4	5	6	7	8
3	15						15					
4	18	26					18	26				
5	21	29	38				21	30	39			
6	24	32	42	53			24	33	43	54		
7	26	36	46	58	70		27	37	48	59	72	
8	29	39	50	62	76	90	30	40	52	64	77	92
9	31	42	54	67	81	96	32	44	56	69	83	98
10	34	46	58	72	86	102	35	47	60	74	88	104
11	37	49	62	77	92	108	38	51	64	79	94	110
12	39	52	66	81	97	114	40	54	68	83	99	116
13	42	56	70	86	102	119	43	58	72	88	105	122
14	45	59	74	91	108	125	46	61	77	93	110	129
15	47	62	78	95	113	131	48	64	81	98	116	135
16	50	66	82	100	118	137	51	68	85	103	121	141
17	52	69	86	104	123	143	54	71	89	107	127	147
18	55	72	90	109	129	149	57	75	93	112	132	153
19	58	76	94	114	134	155	59	78	97	117	138	159
20	60	79	98	118	139	161	62	81	101	122	143	165
21	63	82	102	123	144	167	65	85	105	127	149	171
22	66	86	106	128	150	173	67	88	110	131	154	177
23	68	89	110	132	155	179	70	92	114	136	159	184
24	71	92	114	137	160	185	73	95	118	141	165	190
25	73	96	118	141	166	190	75	99	122	146	170	196
26	76	99	122	146	171	196	78	102	126	151	176	202
27	79	102	126	151	176	202	81	105	130	155	181	208
28	81	105	130	155	181	208	84	109	134	160	187	214
29	84	109	134	160	187	214	86	112	138	165	192	220
30	86	112	138	165	192	220	89	116	142	170	197	226
31	89	115	142	169	197	226	92	119	147	174	203	232
32	92	119	146	174	202	232	94	123	151	179	208	238
33	94	122	150	178	208	238	97	126	155	184	214	244
34	97	125	154	183	213	243	100	129	159	189	219	250
35	99	129	158	188	218	249	102	133	163	193	225	256
36	102	132	162	192	223	255	105	136	167	198	230	263
37	105	135	166	197	229	261	108	140	171	203	235	269
38	107	139	170	202	234	267	110	143	175	208	241	275
39	110	142	174	206	239	273	113	146	179	213	246	281
40	112	145	178	211	244	279	116	150	183	217	252	287
41	115	148	182	215	250	285	119	153	188	222	257	293
42	118	152	186	220	255	290	121	157	192	227	263	299
43	120	155	190	225	260	296	124	160	196	232	268	305
44	123	158	194	229	265	302	127	164	200	236	273	311
45	126	162	198	234	271	308	129	167	204	241	279	317
46	128	165	202	238	276	314	132	170	208	246	284	323
47	131	168	205	243	281	320	135	174	212	251	290	329
48	133	172	209	248	286	326	137	177	216	255	295	335
49	136	175	213	252	292	332	140	181	220	260	301	341

* Para percentiles inferiores $\alpha = 0{,}005$, 0,01, 0,025 y 0,05, $u_\alpha = n_1(n_1 + n_2 + 1) - u_{1-\alpha}$.

Tabla 9 **Percentiles de la distribución bajo H_0 de la suma de rangos positivos de Wilcoxon $W = \sum_{i=1}^{m} r_i$ para un número de parejas con diferencias no nulas $n \leq 16$.***

	Percentil			
n	**0,95**	**0,975**	**0,99**	**0,995**
5	14	15	15	15
6	18	20	21	21
7	24	25	27	28
8	30	32	34	35
9	36	39	41	43
10	44	46	49	51
11	52	55	58	60
12	60	64	68	70
13	69	73	78	81
14	79	83	89	92
15	89	94	100	104
16	100	106	112	116

* Para percentiles inferiores $\alpha = 0{,}005$, 0,01, 0,025 y 0,05, $w_\alpha = n(n+1)/2 - w_{1-\alpha}$.

Tabla 10 **Percentiles de la distribución bajo H_0 del coeficiente de correlación r_s de Spearman en muestras de tamaño $n \leq 10$.***

	Percentil			
n	**0,95**	**0,975**	**0,99**	**0,995**
4	0,800	1,000	1,000	1,000
5	0,800	0,900	0,900	1,000
6	0,771	0,829	0,886	0,943
7	0,679	0,750	0,857	0,893
8	0,619	0,714	0,810	0,857
9	0,583	0,683	0,767	0,817
10	0,552	0,636	0,733	0,782

* Para percentiles inferiores $\alpha = 0{,}005$, 0,01, 0,025 y 0,05, $r_{s,\alpha} = -r_{s,1-\alpha}$.