COLECCIÓN CUIDADOS DE SALUD AVANZADOS

Directora *Loreto Maciá Soler*

ESTADÍSTICA APLICADA A LAS CIENCIAS DE LA SALUD

Joaquín Moncho Vasallo

ELSEVIER

Ámsterdam Barcelona Beijing Boston Filadelfia Londres Madrid
México Milán Múnich Orlando París Roma Sídney Tokio Toronto

ELSEVIER

Travessera de Gràcia, 17-21
08021 Barcelona, España

ISBN (versión impresa): 978-84-9022-446-5
ISBN (versión electrónica): 978-84-9022-641-4

Depósito legal (versión impresa): B. 16.512-2014
Depósito legal (versión electrónica): B. 16.513-2014
Servicios editoriales: ***GEA*** **Consultoría Editorial, s. l.**

Advertencia

La enfermería es un área en constante evolución. Aunque deben seguirse unas precauciones de seguridad estándar, a medida que aumenten nuestros conocimientos gracias a la investigación básica y clínica habrá que introducir cambios en los tratamientos y en los fármacos. En consecuencia, se recomienda a los lectores que analicen los últimos datos aportados por los fabricantes sobre cada fármaco para comprobar las dosis recomendadas, la vía y duración de la administración y las contraindicaciones. Es responsabilidad ineludible del médico determinar las dosis y el tratamiento más indicados para cada paciente, en función de su experiencia y del conocimiento de cada caso concreto. Ni los editores ni los directores asumen responsabilidad alguna por los daños que pudieran generarse a personas o propiedades como consecuencia del contenido de esta obra.

El Editor

Presentación de la colección Cuidados de Salud Avanzados

Cuidados de Salud Avanzados es una colección de monografías dirigidas a profesionales de la salud y estudiantes de posgrado, máster y doctorado dentro del ámbito de las ciencias de la salud. Su orientación recoge las cuatro funciones que la Organización Mundial de la Salud otorga a las profesiones sanitarias: asistencial, docente, investigadora y gestora.

Actualmente, la formación sanitaria tiene tres niveles para todas las titulaciones (grado, máster y doctorado), además de las especialidades propias de cada disciplina. El nivel de grado otorga competencias para el ejercicio de una profesión, por lo que existen grandes diferencias formativas entre titulaciones. Sin embargo, en niveles de formación superior, la orientación de los estudios máster hacia una especialización o formación superior, ya sea con perfil profesional o investigador, a la que tienen acceso en condiciones de igualdad todos los titulados universitarios de grado, permite que la literatura de consulta resulte más homogénea. Lo mismo sucede en los programas de doctorado. Los requisitos y las exigencias formativas e investigadoras no distinguen entre titulaciones de origen, de manera que la bibliografía de consulta debe cumplir unos mínimos acordes con la formación superior requerida, útil para todos los ámbitos de la salud y que considere la formación de grado previa con el fin de que no se repitan competencias.

Todas las monografías han sido escritas por autores de reconocido prestigio en su ámbito, que han constituido equipos de trabajo con expertos en el área, de tal forma que el objetivo principal de la colección queda garantizado: ser una referencia de consulta y apoyo docente dirigida a posgraduados en el ámbito de las ciencias de la salud.

Loreto Maciá Soler

Introducción

Durante las últimas décadas se ha producido una creciente aplicación de los métodos estadísticos en todas las disciplinas del ámbito de las ciencias de la salud, dando lugar, por su amplia utilización, a la implantación de la estadística en los planes de estudios de numerosas titulaciones de este campo, como Medicina, Enfermería, Fisioterapia, Veterinaria, Biología, etc. Y es que muchos de los fenómenos objeto de estudio en este ámbito varían de individuo a individuo, resultando imposible predecir con certeza su resultado de antemano. Dos seres vivos nunca son iguales; es más, ni siquiera un individuo es igual a sí mismo en diferentes etapas de la vida.

En el área clínica, los profesionales se enfrentan con frecuencia a preguntas como: ¿qué patología presenta el paciente?, ¿qué posibilidades de éxito tendrá el tratamiento?, ¿sobrevivirá más de cinco años tras el tratamiento?, ¿cuál es el rango de normalidad de este parámetro clínico?, ¿es lo suficientemente fiable esta prueba diagnóstica? En el área comunitaria se intenta, entre otras cosas, establecer el estado de salud de la comunidad, detectando aquellos grupos de población que requieren una mayor atención sanitaria, o evaluar la efectividad de un programa dedicado a incrementar el nivel de salud de la población. La respuesta a estas cuestiones precisa de las herramientas que proporciona la estadística como parte fundamental del método científico.

El veloz desarrollo informático experimentado en los últimos años, por una parte, ha propiciado la aparición de modelos matemáticos y estadísticos cuya sofisticación y complejidad aumentan día a día, y por otra, ha extendido la utilización de estos procedimientos a través de programas informáticos estadísticos, que facilitan enormemente el análisis de datos. Sin embargo, se requiere de un conocimiento de los conceptos y los procedimientos de análisis estadístico que permita una utilización adecuada de estos recursos.

Esta monografía ha sido estructurada en cinco capítulos: «Conceptos básicos de estadística descriptiva y probabilidad», «Inferencia estadística», «Pruebas no paramétricas», «Análisis de la varianza. ANOVA» y «Análisis de regresión lineal simple y múltiple». La experiencia docente de los autores en multitud de estudios de pregrado, másteres oficiales, cursos de posgrado, cursos de doctorado, etc., permitió reflexionar y discutir cuál debía ser el punto de partida, llegándose a la conclusión de que era necesario comenzar por los conceptos y las técnicas estadísticas básicas, intentando generar una forma de entender y abordar los problemas que permitiera la comprensión y la utilización de técnicas algo más complejas. En este sentido, se ha intentado no abusar en exceso del lenguaje matemático, haciendo especial hincapié en las ideas intuitivas sobre los procedimientos

y la interpretación de los resultados. A partir del capítulo 3, «Pruebas no paramétricas», se han incorporado al texto resultados obtenidos mediante un programa de análisis estadístico, el SPSS v18. No se ha pretendido en ningún momento explicar el funcionamiento del programa, sino, por un lado, proporcionar resultados que de otra manera requerirían numerosos cálculos y, por otro, familiarizar al lector con las salidas que nos ofrecen habitualmente este tipo de programas y su correspondencia con los conceptos y los elementos discutidos.

En ocasiones se hace referencia en el texto a tablas de probabilidad correspondientes a diferentes modelos de probabilidad (normal, *t* de Student, F de Snedecor, etc.) o tablas de valoración de estadísticos como el de Durbin-Watson, U de Mann-Whitney, etc. Estas tablas no han sido incluidas en el texto por su amplitud y porque pueden encontrarse con facilidad en la propia red. Además, conocidas hojas de cálculo incorporan entre las funciones disponibles las correspondientes a estos modelos, permitiendo el cálculo de las probabilidades que se precisen.

En el capítulo 5 se ha realizado un abordaje relativamente completo del modelo de regresión lineal simple tratando que el modelo de regresión lineal múltiple se perciba como una extensión natural. Si bien la parte correspondiente al análisis de regresión lineal múltiple incorpora métodos de diagnóstico de las hipótesis del modelo y de detección de problemas derivados de la inclusión de más de una variable explicativa, como la colinealidad o multicolinealidad, debe hacerse notar que no deja de ser una introducción al estudio de esta técnica que, por motivos de extensión, no ha permitido incluir aspectos como: inclusión de variables cualitativas en el modelo de regresión lineal múltiple, confusión e interacción, métodos de construcción automática, coeficiente de correlación parcial, etc. Sin embargo, pensamos que el lector cuenta, en este texto, con los elementos y la base necesarios para comprender estas nuevas cuestiones.

Al finalizar el estudio de la monografía, los lectores habrán adquirido las siguientes competencias:

- Aplicar los métodos estadísticos como herramienta fundamental en investigación en ciencias de la salud.
- Analizar e interpretar los datos estadísticos referidos a estudios poblacionales.
- Redactar trabajos científicos en ciencias de la salud.
- Desarrollar razonamientos críticos y la capacidad para definir y dar respuesta a problemas utilizando la evidencia científica disponible.

Colaboradores

Joaquín Moncho Vasallo
Departamento de Enfermería Comunitaria, Medicina Preventiva y Salud Pública e Historia de la Ciencia, Facultad de Ciencias de la Salud, Universidad de Alicante, España.

Andreu Nolasco Bonmatí
Departamento de Enfermería Comunitaria, Medicina Preventiva y Salud Pública e Historia de la Ciencia, Facultad de Ciencias de la Salud, Universidad de Alicante, España.

CAPÍTULO

1

Conceptos básicos de estadística descriptiva y probabilidad

Joaquín Moncho Vasallo y Andreu Nolasco Bonmatí

INTRODUCCIÓN

La estadística proporciona al investigador un conjunto de herramientas de análisis que le permiten resumir y describir la información sobre determinadas características de interés de los individuos o elementos objeto de estudio, así como inferir o extraer conclusiones sobre una población a partir de los resultados obtenidos en una muestra. En el presente capítulo se realizará un breve recorrido por las diferentes técnicas de resumen de la información y los conceptos básicos de probabilidad necesarios para la comprensión de los conceptos y técnicas de análisis que se desarrollarán con posterioridad y que constituyen el objetivo principal de la presente obra.

ESTADÍSTICA DESCRIPTIVA

CONCEPTOS PREVIOS

VARIABLES

Las características de interés sobre los individuos o elementos de una población se denominan *variables*. Así, por ejemplo, la *edad*, el *nivel de colesterol* (en mg/100 ml), el *sexo* (hombre/mujer), el *nivel de estudios* (sin estudios, primaria, secundaria, universitarios), etc. serían variables siempre y cuando varíen de individuo a individuo o elemento a elemento de una población. En caso contrario se hablaría de una *constante* y no de una *variable*.

Tipos de variables

Las variables pueden clasificarse en diferentes tipos dependiendo de los valores a los que dan lugar. Esta clasificación es importante porque determinará el tipo de técnicas de análisis que pueden utilizarse para su estudio.

Las variables se clasifican en dos grandes grupos: las variables *cualitativas* o *categóricas* y las variables *cuantitativas.* Las variables cualitativas no toman valores numéricos y pueden clasificarse en un determinado número de categorías o estados (sexo, nivel de estudios, etc.). Las variables cuantitativas toman valores numéricos (edad, número de hijos, nivel de colesterol, etc.). Las variables cualitativas se clasifican a su vez en variables cualitativas *ordinales* y *no ordinales*, dependiendo de si sus categorías o estados pueden ordenarse o no. Así, *sexo* sería una variable cualitativa no ordinal, mientras que *nivel de estudios* sería una variable cualitativa ordinal. Las variables cuantitativas se subdividen a su vez en variables cuantitativas *discretas* y cuantitativas *continuas*, dependiendo de si toman un número finito o infinito numerable de valores (discretas) o infinito no numerable (continuas). A las variables cuantitativas continuas también se las llama variables de razón o intervalo. Así, por ejemplo, las variables *número de hijos*, *número de ingresos en un hospital*, etc., serían variables cuantitativas discretas (obsérvese que entre 0 y 1 hijo no hay valores posibles), mientras *que nivel de colesterol*, *edad* o *nivel de ácido úrico* serían variables continuas, puesto que cualquier valor entre dos dados es posible (toman valores en un intervalo).

RESUMEN DE DATOS. TABLAS DE DISTRIBUCIÓN DE FRECUENCIAS

La primera técnica de resumen de datos consiste en agrupar las diferentes observaciones en cada una de las categorías correspondientes y plasmarlas en una tabla. Cuando las variables son cualitativas las categorías de agrupación vienen determinadas por la propia variable, aunque podrían realizarse reagrupaciones de varias categorías para obtener los resultados deseados. Este tipo de tablas se conoce como tablas de distribución de frecuencias.

Ejemplo 1-1

Se cuenta con información sobre el nivel de estudios de un grupo de $n = 120$ individuos. La variable nivel de estudios ha sido recogida de la siguiente forma: sin estudios, primaria, secundaria, universitarios. Los resultados de las observaciones se resumen en la tabla 1-1.

TABLA 1-1 Tabla de distribución de frecuencias para la variable «nivel de estudios»

Nivel de estudios	f_i	fr_i	F_i	Fr_i
Sin estudios	5	0,042	5	0,042
Primaria	30	0,25	35	0,292
Secundaria	45	0,375	80	0,667
Universitarios	40	0,333	120	1
Total	120	1		

Donde f_i es el número de individuos en la categoría *i*. En este caso, por ejemplo, $f_3 = 45$ (individuos con estudios de secundaria). A cada uno de los valores de f_i se les denomina *frecuencia absoluta*. Suele ser útil construir, a partir de las frecuencias absolutas, algunas columnas adicionales. Se denomina fr_i a la frecuencia relativa que se obtiene dividiendo la frecuencia absoluta en cada categoría entre el número total de observaciones. Si se multiplica por 100, el resultado p_i será el porcentaje de individuos en la categoría correspondiente.

$$\mathrm{fr_i} = \frac{\mathrm{f_i}}{\mathrm{n}}; \mathrm{p_i} = \mathrm{fr_i} \cdot 100$$

La columna F_i se conoce como la *frecuencia acumulada* y expresa el número de individuos hasta la categoría *i*. Así, por ejemplo:

$$F_3 = f_1 + f_2 + f_3 = 5 + 30 + 45 = 80$$

Al igual que en el caso de la frecuencia relativa, la frecuencia acumulada puede hacerse relativa al total de individuos estudiados obteniendo una medida en tanto por uno Fr_i conocida como *frecuencia relativa acumulada*, que, a su vez, puede expresarse en forma de porcentaje multiplicando por 100:

$$\mathrm{Fr_i} = \frac{\mathrm{F_i}}{\mathrm{n}}; \mathrm{P_i} = \mathrm{Fr_i} \cdot 100$$

En el caso de las variables cuantitativas continuas, en las que el rango de valores distintos puede ser elevado, sería poco operativo construir una tabla como la anterior basándose en las frecuencias de repetición de los valores correspondientes, ya que las frecuencias absolutas serían bajas y con elevadas posibilidades de que hubiera valores de la variable con frecuencia nula entre otros valores observados. En este caso, suele ser oportuno agrupar la variable en intervalos.

Ejemplo 1-2

Se dispone de las observaciones del nivel de colesterol de un grupo de 120 individuos en mg/100 ml.

$$220, 165, 159, 170, 246, 234, 200, \ldots, 176$$

Los datos podrían resumirse construyendo una tabla de distribución de frecuencias en las que se agrupan las observaciones en intervalos, por ejemplo, de amplitud 30 (tabla 1-2).

Procediendo de esta forma puede establecerse que el intervalo con mayor frecuencia de observaciones es el [210, 240], que cuenta con 42 individuos.

TABLA 1-2 Tabla de distribución de frecuencias para la variable «nivel de colesterol»

Nivel de colesterol	f_i	fr_i	F_i	Fr_i
150-180	10	0,083	10	0,083
180-210	36	0,3	46	0,383
210-240	42	0,35	88	0,733
240-270	25	0,208	113	0,941
270-300	7	0,058	120	1

También podría concluirse que un 94,1% de los individuos observados tiene un nivel de colesterol inferior a 270 mg/100 ml ($P_i = Fr_i \cdot 100 = 0{,}941 \cdot 100 = 94{,}1$).

RESUMEN DE DATOS. REPRESENTACIONES GRÁFICAS

Si las tablas de distribución de frecuencias proporcionan información sobre el comportamiento de la variable a través del estudio de la distribución de las observaciones entre las diferentes categorías de clasificación, las representaciones gráficas la complementan de forma eficaz, proporcionando una imagen que permite extraer conclusiones de forma rápida acerca de la misma. Dependiendo del tipo de variable será más oportuno utilizar un tipo de representación u otra. A continuación se muestran algunos ejemplos de gráficos para diferentes variables según su tipo.

El diagrama de sectores (fig. 1-1) y el diagrama de barras (fig. 1-2) suelen representar frecuencias absolutas o relativas y están especialmente indicados para variables cualitativas. Sin embargo, en el caso de variables cuantitativas discretas de pocos valores, podrían ser utilizados con la misma eficacia. Incluso en el caso de variables cuantitativas continuas podría utilizarse el diagrama de sectores si previamente la variable ha sido agrupada en un número relativamente reducido de intervalos.

El histograma (fig. 1-3) y el diagrama de cajas *(box-plot)* (fig. 1-4) son gráficos propios de la representación de variables cuantitativas

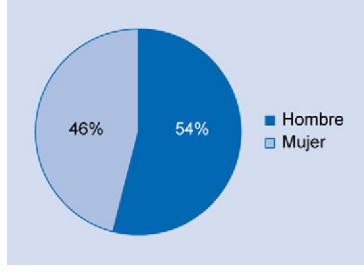

FIGURA 1-1 Diagrama de sectores para la variable *sexo*.

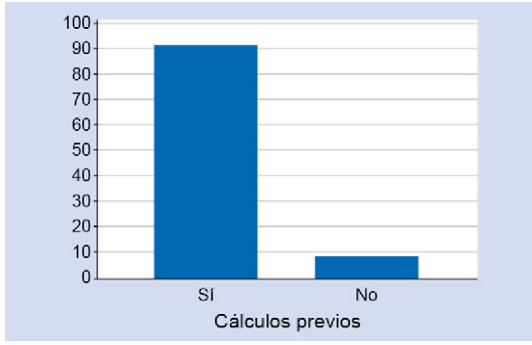

FIGURA 1-2 Diagrama de barras para la variable *cálculos previos*.

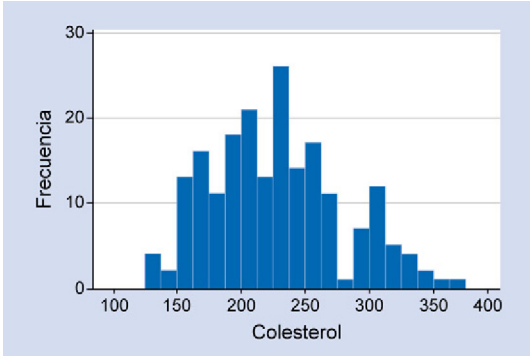

FIGURA 1-3 Histograma para la variable *nivel de colesterol*.

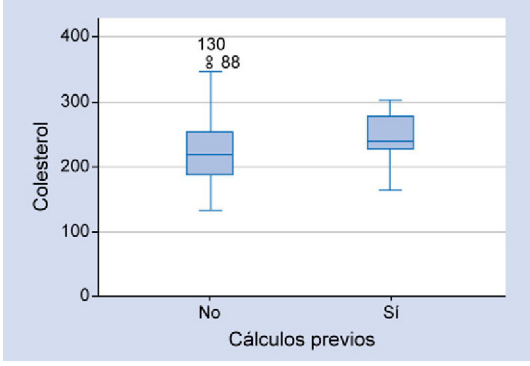

FIGURA 1-4 Diagrama de cajas *(box-plot)* para la variable *colesterol* según *cálculos previos*.

continuas, si bien en el caso de variables cuantitativas discretas con un amplio recorrido de valores posibles, también podrían ser utilizados con éxito.

El histograma muestra, en este caso, un comportamiento de la variable *nivel de colesterol*, en el que la mayoría de las observaciones se concentran en la zona central de la distribución, disminuyendo de forma progresiva la frecuencia de observaciones a uno y otro extremo.

El diagrama de cajas, por su parte, muestra en este caso el valor de la mediana de colesterol (percentil 50) en cada uno de los dos grupos (con antecedentes de cálculos biliares previos y sin antecedentes) en negrita. La caja está determinada por los percentiles 25 y 75 (valores de colesterol no superados por el 25 y el 75% de las observaciones, respectivamente), de forma que en el interior se encuentra el 50% de las observaciones.

En este caso se observa que la mediana de colesterol en el segundo grupo (con antecedentes de cálculos) es superior a la del primer grupo. Por otra parte, la amplitud de la caja del primer grupo (distancia entre el percentil 25 y 75) es superior a la del segundo, sugiriendo una mayor variabilidad en las observaciones de nivel de colesterol. Los conceptos de percentil, mediana o variabilidad serán abordados con posterioridad.

MEDIDAS DESCRIPTIVAS

Las tablas de distribución de frecuencias y las representaciones gráficas constituyen una primera aproximación al resumen de la información proporcionada por los datos disponibles correspondientes a una variable. En el caso de las variables cuantitativas puede profundizarse todavía más en el resumen de los datos, de forma que puedan construirse medidas que informen al investigador sobre la *localización* y *dispersión* de los datos, así como de la *forma* en la que se distribuyen.

MEDIDAS DE TENDENCIA CENTRAL

Las medidas de tendencia central proporcionan información sobre la posición o localización de los datos observados. Entre las medidas de este tipo se encuentran la *media,* la *mediana* o la *moda.* Para ilustrar el cálculo de estas medidas se propone el siguiente ejemplo:

Ejemplo 1-3

En un estudio se obtuvo información sobre el peso (en kg) de un grupo de 15 individuos que se relacionan a continuación:

64, 54, 82, 76, 75, 90, 64, 55, 71, 69, 73, 78, 74, 80, 75

Media

La media (aritmética) es una de las medidas de tendencia central más utilizadas. Se interpreta como el promedio de los datos y se construye de forma que intervienen todos los datos observados en su cálculo de la siguiente forma:

$$\overline{x} = \frac{\sum x_i}{n} = \frac{64 + 54 + 82 + 76 + 75 + 90 + \cdots + 80 + 75}{15} = 72$$

Donde x_1, x_2, x_3…, x_n son las observaciones de la variable y n el número total de observaciones.

La media $\overline{x}$, única para un conjunto de datos, se sitúa en el centro de gravedad de la distribución de los mismos reforzando su papel de medida de tendencia central. Sin embargo, debe tenerse en cuenta que la media es una medida sensible a observaciones atípicas o extremas. Un valor alejado del resto tendría un efecto importante sobre el valor de la media. Por ejemplo, si se trata de un valor considerablemente mayor que el conjunto de las observaciones, la media se desplazará hacia la derecha (aumentará su valor), pudiendo situarse en un lugar poco representativo del conjunto de datos. Existen alternativas ajustadas del cálculo de la media *(medias robustas)* que tratan de corregir este problema otorgando un menor peso a las observaciones alejadas.

Mediana

Una alternativa al cálculo de la media, no sensible a observaciones atípicas o extremas, la constituye la mediana. El valor de la mediana, para un conjunto de datos, se obtiene de forma que deja el mismo número de observaciones a su izquierda que a su derecha.

Aunque podría haber infinitos valores que cumplieran este requisito para un conjunto de observaciones, la forma habitual de cálculo garantiza que la mediana será única para un conjunto de datos. En primer lugar, será necesario ordenar los datos de menor a mayor:

$$54, 55, 64, 64, 69, 71, 73, 74, 75, 75, 76, 78, 80, 82, 90$$

Dado que, en este caso, el número de datos es impar, solo hay un valor que se sitúa en el centro, dejando el mismo número de datos a izquierda y a derecha, que es el que ocupa la posición 8 (deja siete datos a su izquierda y siete a su derecha). Si el número total de datos fuera par, se calcularía la semisuma entre los dos datos centrales. En general, se calculará en primer lugar el rango de la mediana, que informará sobre la posición que debe ocupar esta, una vez ordenados los datos de menor a mayor, de la siguiente forma:

$$r_{md} = \frac{n+1}{2} = \frac{15+1}{2} = 8$$

En este caso, la mediana es el dato que ocupa la posición 8 y sería Md = 74 kg. Si se considerara un conjunto de $n = 16$ observaciones (se añade la observación 93 kg al grupo anterior), se tendrá que:

$$54,55,64,64,69,71,73,74,75,75,76,78,80,82,90,93$$

Donde:

$$r_{Md} = \frac{n+1}{2} = \frac{16+1}{2} = 8,5$$

La mediana sería un valor entre el dato que ocupa la posición 8 y el dato que ocupa la posición 9, que en este caso corresponde a los valores 74 y 75. La mediana se obtendrá entonces:

$$Md = \frac{74+75}{2} = 74,5$$

Adviértase que si el valor observado más elevado fuera 120 kg, el valor de la mediana no cambiaría (algo que sí ocurriría con la media). Por otra parte, en el cálculo de la mediana intervienen solo uno o dos datos directamente con su valor y todos indirectamente a través de su orden, por lo que se deduce que, con la utilización de la mediana, se pierde parte de la información que proporcionan los datos en comparación con la media.

Moda

La moda se define, para un conjunto de datos, como el valor más frecuente, es decir, el valor que más veces se repite. Si se trabaja con los datos del ejemplo, se observa que dos datos se repiten exactamente el mismo número de veces y representan la mayor frecuencia observada y son: 64 y 75 kg. Por tanto, se dispondría de dos valores para la moda:

$$Mo = \{64,75\}$$

Este resultado evidencia que la moda no tiene por qué ser única para un conjunto de datos. Por otra parte, basta que un dato se repita más veces que el resto para considerarse moda, aunque no sea una buena medida resumen de los datos, siendo, por tanto, la medida más débil de las estudiadas hasta el momento.

Una alternativa para el cálculo de la moda en el caso de variables cuantitativas continuas, donde es habitual observar frecuencias bajas en la mayoría de los valores observados, es agrupar en intervalos y detectar el intervalo o intervalos con mayor frecuencia absoluta lo que podría definirse como *intervalo modal.*

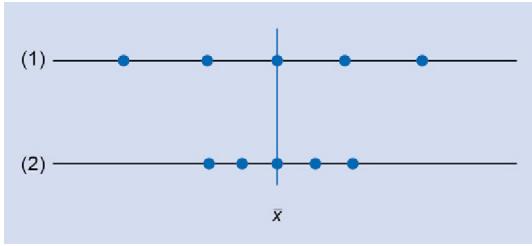

FIGURA 1-5 Comparación de la dispersión de los datos en dos grupos de observaciones.

MEDIDAS DE DISPERSIÓN

Las medidas de tendencia central proporcionaban información sobre la localización de los datos pero no sobre la dispersión o variabilidad con la que se sitúan en torno a dichas medidas.

En la figura 1-5 se presentan dos conjuntos de datos en los que la media coincide. Además, como la distribución de los datos es simétrica, la mediana coincide con la media. Sin embargo, la distribución de los dos conjuntos de datos difiere, ya que puede observarse que en el caso (2) los datos se encuentran mucho más concentrados en torno a la media (o mediana) que en el caso (1), donde la dispersión es mayor. Es necesario, por tanto, disponer de medidas que informen sobre la dispersión de los datos y que permitan distinguir situaciones como la planteada.

Rango o recorrido

La medida más sencilla y visualmente intuitiva para cuantificar la dispersión de los datos es el *rango* y se obtendrá calculando la distancia entre el mayor y el menor valor observado. Si se trabaja con los datos del ejemplo 1-1 se tendrá que:

$$R = X_{máx} - X_{mín} = 90 - 54 = 36$$

Luego el rango de valores observados muestra una distancia de 36 kg entre el menor y el mayor valor observado. La obtención del rango es sencilla, sin embargo, en su construcción solo intervienen dos de los datos observados, que, además, son los más extremos. Esto tiene como consecuencia que el rango será una medida extremadamente sensible a observaciones extremas y que no tiene en cuenta gran parte de la información disponible.

Varianza y desviación típica o estándar

La desviación típica o estándar es la medida de dispersión más utilizada por sus propiedades y porque involucra a todos los datos en su construcción.

La idea es obtener una medida resumen de la distancia de cada dato a la media (desviación a la media). Cuanto mayor sea la medida resumen de las distancias, más alejados estarán los datos de la media y, por tanto, existirá una mayor dispersión o variabilidad. Se utilizan las distancias al cuadrado para obviar el signo de la distancia y valorar únicamente su magnitud. Así, en primer lugar, se define la varianza como el promedio de las distancias (al cuadrado) de cada dato a la media, que con los datos del ejemplo quedará:

$$S^2 = \frac{\sum (x_i - \overline{x})^2}{n} = \frac{(64-72)^2 + (54-72)^2 + \cdots + (75-72)^2}{15} = 87,6\,kg^2$$

La varianza está expresada, por tanto, en unidades al cuadrado de la variable. Para conseguir una medida en las mismas unidades que la variable original se extrae la raíz cuadrada, obteniéndose la denominada desviación típica o estándar. En el ejemplo se tendrá que:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}} = \sqrt{87,6} = 9,36\,kg$$

Puede afirmarse, por la desigualdad de Tchebychev, que entre la media $\overline{x}$ y k desviaciones típicas se encuentran, al menos, el $100(1 - 1/k^2)\%$ de los datos. Trabajando con los datos del ejemplo se tendrá que el intervalo:

$$\left[\overline{x} - 2S; \overline{x} + 2S\right] = \left[72 - 2 \cdot 9,36; 72 + 2 \cdot 9,36\right] = \left[53,28; 90,72\right]$$

Contendrá, al menos, el siguiente porcentaje de datos:

$$100\left(1 - \frac{1}{2^2}\right) 100 \cdot 0,75 = 75\%$$

Coeficiente de variación

La desviación típica proporcionaba una medida resumen de las distancias de cada dato a la media (desviaciones) en las mismas unidades que la variable original y, por tanto, depende de dichas unidades de medida. ¿Son comparables las desviaciones típicas de dos conjuntos de datos? ¿Puede afirmarse, en general, que a mayor desviación típica mayor dispersión? La respuesta es que esto solo es posible si los conjuntos de datos que se pretenden comparar tienen la misma media. Para ilustrar esta cuestión se propone el siguiente ejemplo:

Ejemplo 1-4

Supóngase que se dispone de información sobre el número de hijos y la edad de un grupo de mujeres. La media y la desviación típica del número de hijos fueron de 1,3 y 1,2, respectivamente, mientras que para la edad la media fue de 34,2, con una desviación típica de 6 años. La cuestión es ¿qué variable presenta una mayor dispersión o variabilidad?

Si se atiende únicamente al valor de la desviación típica se decidiría que la *edad* presenta mayor dispersión que la variable *número de hijos* (S = 6 años frente a S = 1,2 hijos). Sin embargo, no es lo mismo desviarse 6 unidades en magnitudes en torno a 34,2 que desviarse 1,2 unidades en magnitudes alrededor de 1,3, y parece lógico pensar que la dispersión es mayor en este segundo caso. Será necesario construir una medida de dispersión relativa que no dependa de las unidades de medida (adimensional).

El coeficiente de variación es una medida adimensional de la dispersión relativa de los datos que se obtiene dividiendo la desviación típica por la media. Si se multiplica por 100, podrá interpretarse como el porcentaje de variabilidad de los datos para los que se calcula. Trabajando con los datos del ejemplo se tendrá que:

$$CV_{edad} = \frac{6}{34,2} = 0,175(17,5\%)$$

$$CV_{n.^{o}hijos} = \frac{1,2}{1,3} = 0,923(92,3\%)$$

Donde se pone de manifiesto que el porcentaje de variabilidad observada en la variable número de hijos (92,3%) es mucho mayor que la correspondiente a la edad (17,5%). Si se trabaja con los datos del ejemplo 1-3, se obtendrá:

$$CV = \frac{S}{\bar{x}} = \frac{9,36}{72} = 0,13(13\%)$$

PERCENTILES O CUANTILES

Los percentiles o cuantiles son valores de la variable no superados por un determinado porcentaje de observaciones o datos (equivalentemente, también puede definirse como el valor superado por el resto). Así, el percentil de orden k para un conjunto de datos será el valor de la variable no superado por el k% de las observaciones.

Si se trabaja con los datos del ejemplo 1-3, los percentiles de orden 30 y 70 serán los valores de peso no superados por el 30 y 70% de las observaciones. Téngase en cuenta que, bajo esta perspectiva, la mediana es un caso particular en el ámbito de los percentiles, puesto que coincide con el percentil

de orden 50. Para calcular los percentiles será necesario, en primer lugar, ordenar los datos de menor a mayor:

$$54, 55, 64, 64, 69, 71, 73, 74, 75, 75, 76, 78, 80, 82, 90$$

A continuación, de forma similar al caso de la mediana, se calculará el rango del percentil correspondiente. Para los percentiles de orden 30 y 70 quedará:

$$r_{p_k} = \frac{k}{100}(n+1) = r_{p_{30}} = \frac{30}{100}(15+1) = 4,8$$

$$r_{p_k} = \frac{k}{100}(n+1) = r_{p_{70}} = \frac{70}{100}(15+1) = 11,2$$

Luego el percentil p_{30} será el dato que ocupa la posición 4,8 (será, por tanto, un valor entre el dato que ocupa la posición 4 y el dato que ocupa la posición 5), y el percentil $p_{70,}$ el dato que ocupa la posición 11,2 (será un valor entre el dato que ocupa la posición 11 y el que ocupa la posición 12). Será necesario calcular una media ponderada para obtener el valor final de la siguiente forma:

$$p_{30} = (1-f)x_{(i)} + fx_{(i+1)} = (1-0,8)\cdot 64 + 0,8\cdot 69 = 68$$

$$p_{70} = (1-f)x_{(i)} + fx_{(i+1)} = (1-0,2)\cdot 76 + 0,2\cdot 78 = 76,4$$

Donde *f* es la parte fraccionaria del rango del percentil correspondiente.

INTERVALO INTERPERCENTÍLICO O INTERCUANTÍLICO DE ORDEN K

El intervalo interpercentílico de orden *k* es el intervalo centrado que contiene al (100 − 2 k)% de los datos. Se entiende por intervalo centrado aquel que deja el mismo porcentaje de observaciones a su izquierda que a su derecha (fig. 1-6).

Para garantizar esta distribución de los datos, el intervalo deberá construirse a partir de los percentiles p_k y p_{100-k}.

$$I_k = [p_k, p_{100-k}]$$

Si se trabaja con los datos del ejemplo, el intervalo interpercentílico de orden 30 quedará de la siguiente forma:

$$I_k = [p_k, p_{100-k}] = [p_{30}; p_{70}] = [68; 76,4]$$

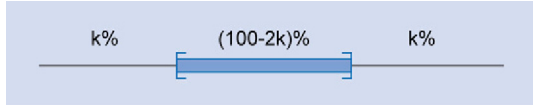

FIGURA 1-6 Intervalo interpercentílico de orden *k*.

Este intervalo contendrá al (100 − 2k) = (100 − 2 · 30) = 40% de los datos observados y dejará, a cada lado, el 30% de los mismos.

MEDIDAS DE FORMA

Las medidas de forma proporcionan información sobre el comportamiento de los datos correspondientes a una variable atendiendo a la simetría o el apuntamiento de la distribución de los mismos.

Coeficiente de asimetría

Una primera aproximación sencilla al estudio de la simetría de la distribución de los datos consiste en comparar la media y la mediana. Si están muy próximas, la distribución será aproximadamente simétrica. Si, por el contrario, la media es significativamente mayor que la mediana o significativamente menor, la distribución será asimétrica por la derecha o por la izquierda. El coeficiente de asimetría es una medida más sofisticada para el estudio de la simetría y se calcula de la siguiente forma:

$$As = \frac{\frac{\sum (x_i - \bar{x})^3}{n}}{S^3} = \frac{(64-72)^3 + (54-72)^3 + \cdots + (75-72)^3}{15 \cdot 9,36^3} = -0,33$$

Si el valor del coeficiente *As* está cerca de cero, la distribución de los datos será aproximadamente simétrica. Si el valor del coeficiente *As* es superior a cero, la distribución será asimétrica por la derecha, mientras que si es inferior a cero será asimétrica por la izquierda.

En el ejemplo puede observarse que el valor del coeficiente de asimetría es −0,33, valor muy próximo a 0, por lo que la distribución será aproximadamente simétrica o muy ligeramente asimétrica por la izquierda.

Coeficiente de apuntamiento o curtosis

La curtosis es una medida del apuntamiento de la distribución de los datos en comparación con la distribución *normal* que más se le aproxima a los mismos (la distribución normal será abordada en próximos apartados). Su cálculo viene determinado por la siguiente expresión, que en el caso de los datos del ejemplo 1-3 quedará:

$$Cu = \frac{\frac{\sum (x_i - \bar{x})^4}{n}}{S^4} = \frac{(64-72)^4 + (54-72)^4 + \cdots + (75-72)^4}{15 \cdot 9,36^4} = 2,76$$

Si se tiene en cuenta que el valor de la curtosis en cualquier modelo de distribución normal es siempre 3, podrá procederse del siguiente modo:

Si Cu >> 3 → La distribución es más apuntada que la normal (leptocúrtica).
Si Cu ≃ 3 → La distribución es igual de apuntada que la normal (mesocúrtica).
Si Cu << 3 → La distribución es menos apuntada que la normal que más se le aproxima (platicúrtica).

En este caso, el valor de la curtosis es ligeramente inferior a 3, por lo que el apuntamiento de la distribución es similar al de la normal o ligeramente menos apuntada. En la mayoría de los programas de análisis estadístico el coeficiente de apuntamiento o curtosis se calcula:

$$Cu = \frac{\frac{\sum (x_i - \bar{x})^4}{n}}{S^4} - 3$$

Consiguiéndose de esta forma que la interpretación del coeficiente se base en el valor 0, al igual que ocurría con el coeficiente de asimetría.

PROBABILIDAD

La teoría de la probabilidad, que experimentó un gran desarrollo a partir del estudio de los juegos de azar, tratará de cuantificar la incertidumbre que rodea a un suceso dada la imposibilidad de predecir con exactitud el resultado del fenómeno aleatorio del que proviene y constituirá la herramienta necesaria para extraer conclusiones sobre determinadas características de interés de una población, a partir de los datos de una muestra (inferencia estadística). En este apartado se abordarán los conceptos básicos de probabilidad que permitan la compresión de posteriores desarrollos.

CONCEPTOS PREVIOS

FENÓMENOS ALEATORIOS

Los fenómenos aleatorios son aquellos cuyos resultados son impredecibles de antemano. Gran parte de las características de interés sobre los individuos o elementos de una población se consideran *fenómenos aleatorios.* Así, el grupo sanguíneo, nivel de colesterol, la edad, el nivel de ácido úrico, el sexo, el estado civil o la respuesta a un tratamiento de un individuo seleccionado al azar de una población constituyen fenómenos aleatorios, ya que se desconoce su resultado hasta que se produce la observación.

SUCESOS SIMPLES O ELEMENTALES

A cada uno de los resultados posibles de un fenómeno aleatorio se le denomina *suceso simple o elemental.* Por ejemplo, si se trabaja con el fenómeno aleatorio *sexo,* los sucesos simples o elementales serán dos:

$$H = \{\text{ser hombre}\}; M = \{\text{ser mujer}\}$$

En el caso de trabajar con el fenómeno aleatorio *grupo sanguíneo,* se tendrían cuatro sucesos simples o elementales:

$$A = \{\text{ser del grupo A}\}; B = \{\text{ser del grupo B}\}$$

$$AB = \{\text{ser del grupo AB}\}; O = \{\text{ser del grupo O}\}$$

ESPACIO MUESTRAL

El conjunto de sucesos simples o elementales recibe el nombre de *espacio muestral.* Así, si se trabaja con el fenómeno aleatorio *grupo sanguíneo,* el espacio muestral quedaría de la siguiente forma:

$$\text{Espacio muestral} = \{A, B, AB, O\}$$

OPERACIONES CON SUCESOS: UNIÓN E INTERSECCIÓN DE SUCESOS

Los sucesos elementales pueden combinarse de forma que pueden obtenerse otros sucesos denominados *sucesos aleatorios.* La forma habitual de combinar sucesos consiste en considerar si ocurre un suceso y otro o si ocurre un suceso u otro. La notación utilizada para expresar estas combinaciones de sucesos tiene sus raíces en la teoría de conjuntos. Así, si se combina el suceso {*ser hombre*} y el suceso {*ser del grupo A*} pueden obtenerse los siguientes sucesos aleatorios:

$$H \cap A = \{\text{ser hombre y del grupo A}\}$$

$$H \cup A = \{\text{ser hombre o del grupo A}\}$$

El primer suceso aleatorio ($H \cap A$) requiere que el individuo posea las dos características a la vez (ser hombre y del grupo sanguíneo A). El segundo suceso aleatorio requiere que el individuo posea, al menos, alguna de las dos.

Ejemplo 1-5

En la tabla 1-3 se muestra información sobre el sexo y el grupo sanguíneo de un conjunto de 150 individuos.

Se observan 30 casos favorables al suceso $H \cap A$ y 113 (30 + 10 + 6 + 34 + 33) favorables al suceso $H \cup A$. Del mismo modo, puede establecerse que se observan 80 casos favorables al suceso H y 81 (63 + 18) favorables al suceso $A \cup B$.

Suceso seguro, imposible y complementario

Algunos sucesos aleatorios reciben nombres especiales que recuerdan la naturaleza de los mismos. Así, el suceso imposible será aquel que no puede ocurrir nunca y se suele expresar mediante el símbolo ϕ. El suceso seguro es aquel que ocurre siempre y se expresa como Ω, mientras que el suceso contrario o complementario al suceso A requiere la no ocurrencia del suceso correspondiente y se denotará $\overline{A}$.

Si se trabaja con los datos del ejemplo 1-5 se aprecian varios sucesos imposibles. Así, por ejemplo:

$$A \cap O = \phi ; H \cap M = \phi$$

Obsérvese, además, que estos sucesos imposibles no cuentan con ningún caso favorable. Por su parte, entre los sucesos seguros se encuentran:

$$A \cup B \cup AB \cup O = \Omega ; H \cup M = \Omega$$

El suceso seguro cuenta con todos los casos como favorables al suceso. Es obvio en este caso que un individuo cualquiera de los estudiados en el ejemplo pertenecerá al grupo sanguíneo A o B o AB o O.

Por su parte, pueden proponerse multitud de sucesos complementarios a partir de los sucesos simples e, incluso, a partir de combinaciones de estos sucesos o sucesos aleatorios. En el ejemplo 1-5 se tendrá, por ejemplo:

$$\overline{A} = B \cup AB \cup O ; \overline{AB} = A \cup B \cup O ; \overline{A \cup AB} = B \cup O$$
$$\overline{H} = M$$

TABLA 1-3 Distribución de frecuencias de grupo sanguíneo por sexos

	Sexo		
Grupo sanguíneo	**Hombre**	**Mujer**	**Total**
A	30	33	**63**
B	10	8	**18**
AB	6	2	**8**
O	34	27	**61**
Total	**80**	**70**	**150**

SUCESOS MUTUAMENTE EXCLUYENTES, INCOMPATIBLES O DISJUNTOS

Se dice que los sucesos $A_1, A_2, A_3 \ldots, A_k$ constituyen una familia de sucesos mutuamente excluyentes, incompatibles o disjuntos si para cualesquiera dos sucesos pertenecientes a la familia se verifica:

$$A_i \cap A_j = \phi$$

En el ejemplo 1-5, los sucesos A, B y AB constituyen una familia de sucesos mutuamente excluyentes, ya que:

$$A \cap B = \phi; A \cap AB = \phi; B \cap AB = \phi$$

PARTICIÓN DEL ESPACIO MUESTRAL

Se dice que los sucesos $A_1, A_2, A_3 \ldots, A_k$, constituyen una partición del espacio muestral si son una familia de sucesos mutuamente excluyentes que, además, verifica que:

$$A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_k = \Omega$$

Así, por ejemplo, si se trabaja con la familia de sucesos del apartado anterior (A, B y AB) se tendrá que, a pesar de que son mutuamente excluyentes, no constituyen una partición del espacio muestral, ya que:

$$A \cup B \cup AB \neq \Omega$$

Para conseguir una partición del espacio muestral sería necesario añadir el suceso O a la familia, de forma que quede conformada por los sucesos A, B, AB y O. Esta familia es partición del espacio muestral ya que:

$$A \cup B \cup AB \cup O = \Omega$$

LEYES DE DE MORGAN

La unión e intersección de sucesos complementarios pueden expresarse de forma equivalente que, en ocasiones, facilita el conteo (y posteriormente el cálculo de probabilidades) de los casos favorables correspondientes. Las siguientes equivalencias se conocen como las *leyes de De Morgan,* y establecen que, para dos sucesos A y B cualesquiera se verifica que:

$$\overline{A} \cap \overline{B} = \overline{A \cup B}$$

$$\overline{A} \cup \overline{B} = \overline{A \cap B}$$

MEDIDA DE PROBABILIDAD

REGLA DE LAPLACE

Si bien es cierto que un fenómeno aleatorio se caracteriza por la imposibilidad de predecir su resultado de antemano, es posible cuantificar el

grado de incertidumbre que rodea un suceso determinado de forma que puede establecerse si existen resultados del fenómeno aleatorio o sucesos más probables que otros.

Por ejemplo, si se consideran los datos del ejemplo 1-5 y se extrae un individuo al azar de los 150 observados, no se puede asegurar con certeza el grupo sanguíneo al que pertenecerá pero sí se puede afirmar que existen más posibilidades (probabilidades) de que pertenezca al grupo O que al grupo AB, puesto que son muchos más individuos los pertenecientes al primer grupo. De hecho podría calcularse una medida de probabilidad de pertenencia a cada uno de los dos grupos considerados de la siguiente forma:

$$\text{Probabilidad}(\text{O}) = \frac{61}{150} = 0,407\,;\text{Probabilidad}(\text{AB}) = \frac{8}{150} = 0,053$$

De hecho la posibilidad de que, en este caso, el individuo pertenezca al grupo O es del 40,7%, mientras que la posibilidad de que pertenezca al grupo AB es del 5,3%. Para la construcción de esta medida se han considerado los casos favorables al suceso y los casos posibles. En general, para un suceso A cualquiera, se calculará la probabilidad del suceso A de la siguiente forma:

$$\text{Probabilidad}(\text{A}) = P(\text{A}) = \frac{\text{número de casos favorables al suceso A}}{\text{número de casos posibles}}$$

Esta regla de cálculo de probabilidades se conoce como la *regla de Laplace*, si bien su enunciado se atribuye a Cardano. De este resultado, basado en el postulado de indiferencia o de sucesos equiprobables, se desprende, por ejemplo, que la probabilidad de un suceso cualquiera tomará siempre un valor entre 0 y 1. Basta con tener en cuenta que si no existe ningún caso favorable el numerador será 0 y, por tanto, el cociente. Si el suceso es seguro, contará con todos los casos como favorables, de forma que el cociente será 1.

Por otra parte, un individuo no puede pertenecer al grupo sanguíneo A y O a la vez ($\text{A} \cap \text{O} = \varnothing$). En este caso, podrá calcularse la probabilidad de que un individuo seleccionado al azar pertenezca al grupo A o O ($\text{A} \cup \text{O}$) de la forma:

$$\text{P}(\text{A} \cup \text{O}) = \frac{63+61}{150} = \frac{124}{150} = \frac{63}{150} + \frac{61}{150} = \text{P}(\text{A}) + \text{P}(\text{O})$$

DEFINICIÓN FRECUENTISTA DE LA PROBABILIDAD

La aproximación frecuentista al cálculo de la medida de probabilidad de un suceso aleatorio *A* se basa en utilizar la frecuencia relativa del suceso

A para estimar su probabilidad cuando el proceso se repite un número considerable de veces. Así se tendrá que:

$$P(A) = \lim_{n \to +\infty} fr(A)$$

Si se pretendiera cuantificar la probabilidad de que un individuo de una población perteneciera al grupo sanguíneo AB podría extraerse una muestra de dicha población y estimar la probabilidad mediante la frecuencia relativa correspondiente al grupo AB.

DEFINICIÓN AXIOMÁTICA DE LA PROBABILIDAD

La forma en que debe calcularse la probabilidad asociada a un suceso ha sido objeto de controversia y debate a lo largo de la historia de la evolución de la teoría de la probabilidad. No obstante, pueden enunciarse los siguientes axiomas o premisas que no se derivan de ningún resultado anterior, y a partir de los cuales pueden obtenerse el resto de propiedades o teoremas relativos al cálculo de probabilidades:

- $P(A) \geq 0$
- $P(A \cup B) = P(A) + P(B)$ si $A \cap B = \varnothing$
- $P(\Omega) = 1$

Obsérvese que estos axiomas responden a lo esperado en cuanto al cálculo de probabilidades se refiere que ilustraban los resultados obtenidos en el ejemplo 1-5 en aplicación de la regla de Laplace para el cálculo de probabilidades.

PROPIEDADES DE LA PROBABILIDAD

De los axiomas anteriores se derivan una serie de propiedades entre las que se encuentran las siguientes:

- $0 \leq P(A) \leq 1$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ para A, B sucesos aleatorios cualesquiera
- $P(\bar{A}) = 1 - P(A)$

Si se trabaja con los datos del ejemplo 1-5 y se aplican los diferentes axiomas y propiedades se obtendrá que:

$$P(\bar{H}) = 1 - P(H) = 1 - \frac{80}{150} = \frac{70}{150} = P(M)$$

$$P(\bar{A}) = 1 - P(A) = 1 - \frac{63}{150} = \frac{18+8+61}{150} = P(B \cup AB \cup O)$$

$$P(A \cup H) = P(A) + P(H) - P(A \cap H) = \frac{63}{150} + \frac{80}{150} - \frac{30}{150} = \frac{63+80-30}{150}$$

Para el cálculo de la probabilidad de la unión de cualesquiera dos sucesos, en el ejemplo A ∪ H, téngase en cuenta que, a la suma de las probabilidades de cada uno de los sucesos, debe restarse la probabilidad de la intersección. Para comprender este hecho basta con observar que los casos favorables al suceso A ∩ H han sido contabilizados, tanto en el cálculo de la probabilidad de A como en el de la probabilidad de H, debiendo ser descontados una vez:

$$P(A \cup H) = \frac{\text{casos favorables}}{\text{casos posibles}} = \frac{30+33+10+6+34}{150} = \frac{63+80-30}{150}$$

PROBABILIDAD CONDICIONAL Y LEY MULTIPLICATIVA

En ocasiones resulta necesario cuantificar la probabilidad de ocurrencia de un determinado suceso condicionado por el hecho de que previamente ha sucedido otro. Por ejemplo, ¿cuál es la probabilidad de que un individuo padezca un cáncer de pulmón si es fumador? En este caso, será necesario distinguir entre la probabilidad general de padecer un cáncer de pulmón de la probabilidad de padecerlo entre los fumadores.

Si se trabaja con los datos del ejemplo 1-5, ¿cuál sería la probabilidad de que un individuo de los estudiados seleccionado al azar perteneciera al grupo A si se sabe que es un hombre? Parece razonable proceder del siguiente modo:

$$P(A \text{ si } H) = P\left(A/H\right) = \frac{30}{80}$$

Como se cuenta con 80 hombres, se trataría de construir la probabilidad como el cociente entre los hombres pertenecientes al grupo A y el número total de hombres. (Los datos de las mujeres no intervendrían en el resultado.) Este cálculo ha sido posible gracias a que se dispone de las frecuencias en cada uno de los casos como muestra la tabla 1-3. Sin embargo, es necesario contar con una regla general de cálculo en función de las probabilidades de los sucesos involucrados en ausencia de frecuencias. En este sentido puede observarse que:

$$P\left(A/H\right) = \frac{30}{80} = \frac{30/150}{80/150} = \frac{P(A \cap H)}{P(H)}$$

Por tanto, puede establecerse que si A y B son sucesos aleatorios cualesquiera, se tendrá que la probabilidad de A condicionado por la ocurrencia del suceso B podrá expresarse de la forma:

$$P\left(A/B\right) = \frac{P(A \cap B)}{P(B)}$$

Obsérvese que para que pueda calcularse esta probabilidad condicional se requiere que $P(B) \neq 0$. De esta expresión para el cálculo de la probabilidad condicional se deriva otra que permite desarrollar la probabilidad de la intersección de dos sucesos cualesquiera de la forma:

$$P(A \cap B) = P\left(A/B\right) \cdot P(B)$$

Esta regla de cálculo para la probabilidad de la intersección se denomina *ley multiplicativa.* Obsérvese que para obtener este resultado basta con despejar el valor de la probabilidad de la intersección en la expresión de la probabilidad condicional. De forma análoga, la probabilidad de la intersección de dos sucesos A y B podría expresarse también:

$$P(A \cap B) = P\left(B/A\right) \cdot P(A)$$

INDEPENDENCIA DE SUCESOS

La consideración de probabilidades condicionadas o condicionales facilita el estudio de la independencia o dependencia entre sucesos. Así, si la probabilidad general de padecer un cáncer de pulmón aumenta o disminuye si se conoce que el individuo es fumador, podrá concluirse que existe algún tipo de relación o dependencia entre los sucesos estudiados. Por el contrario, si la probabilidad general de padecer cáncer de pulmón no se ve modificada por el hecho de conocer que el individuo fuma, los sucesos serían independientes.

En general, para dos sucesos aleatorios cualesquiera A y B, se dirá que son independientes si se verifica:

$$P\left(A/B\right) = P(A)$$
$$P\left(B/A\right) = P(B)$$

En el caso del ejemplo 1-5 se observa, por ejemplo, que los sucesos O y H no serían independientes para los sujetos estudiados, ya que:

$$p\left(O/H\right) = \frac{34}{80} = 0,425 \neq 0,407 = \frac{61}{150} = P(O)$$

La probabilidad general de ser del grupo O pasa de 0,407 a 0,425 si se sabe que el individuo seleccionado es un hombre.

CARACTERIZACIÓN DE SUCESOS INDEPENDIENTES

De la definición de sucesos independientes y de la ley multiplicativa puede obtenerse una regla de cálculo simplificada para la probabilidad de

la intersección de dos sucesos independientes. Según la ley multiplicativa la probabilidad de la intersección de dos sucesos A y B se expresaría de la forma:

$$P(A \cap B) = P\left(A/B\right) \cdot P(B)$$

Si los sucesos son independientes se tiene que: $P\left(A/B\right) = P(A)$, que aplicado a la expresión anterior proporcionará el siguiente resultado:

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

Luego si dos sucesos *A* y *B* son independientes, entonces la probabilidad de la intersección puede expresarse como el producto de probabilidades. Este resultado puede extenderse a una familia de sucesos independientes, de forma que, si A_1, A_2, A_3…, A_k son independientes entre sí, entonces se verificará que:

$$P(A_1 \cap A_2 \cap A_3 \cap \ldots \cap A_k) = P(A_1) \cdot P(A_2) \cdot P(A_3) \ldots \cdot P(A_k) = \prod P(A_i)$$

Si se trabaja con los datos del ejemplo 1-5 puede comprobarse si dos sucesos son independientes verificando si la probabilidad de la intersección coincide con el producto de probabilidades. Así, al igual que se verificara anteriormente, los sucesos O y H no serían independientes, ya que:

$$P(O \cap H) = \frac{34}{150} = 0,227 \neq 0,217 = \frac{61}{150}\frac{80}{150} = P(O) \cdot P(H)$$

Por otra parte, también se verifica que si A y B son independientes, entonces también lo serán las parejas de sucesos: *A* y $\overline{B}$; $\overline{A}$ y B; $\overline{A}$ y $\overline{B}$.

SUCESO COMPLEMENTARIO CONDICIONADO

El interés por el estudio de sucesos condicionados hace necesario reflexionar sobre el cálculo de probabilidades correspondientes a los complementarios de dichos sucesos. En general, para cualesquiera dos sucesos A y B se tendrá que:

$$P\left(\overline{A}/B\right) = 1 - P\left(A/B\right)$$

Es decir, manteniendo el suceso previamente conocido o condicionante B invariable, puede ocurrir A o $\overline{A}$. Si se trabaja con los datos del ejemplo 1-5, se tendrá, por ejemplo:

$$P\left(\overline{O}/H\right) = 1 - P\left(O/H\right) = 1 - \frac{34}{80} = 0,575$$

Ejemplo 1-6

Se sabe que en una población el 15% de los individuos padece una determinada patología (E), que el 25% son obesos (Ob) y que el 10% padecen la patología y son obesos. Si se selecciona a un individuo al azar de dicha población, ¿cuál será la probabilidad de que:

1. no padezca la patología?
2. padezca la patología o sea obeso?
3. padezca la patología si es obeso?
4. sea obeso si padece la patología?
5. no sea obeso ni padezca la patología?
6. no padezca la patología si es obeso?
7. ¿Son los sucesos *padecer la patología* y *ser obeso* independientes?

1. $P\left(\overline{E}\right) = 1 - P(E) = 1 - 0,15 = 0,85$

2. $P(E \cup Ob) = P(E) + P(Ob) - P(E \cap Ob) = 0,15 + 0,25 - 0,1 = 0,3$

3. $P\left(E/_{Ob}\right) = \dfrac{P\left(E \cap Ob\right)}{P(Ob)} = \dfrac{0,1}{0,25} = 0,4$

4. $P\left(Ob/_{E}\right) = \dfrac{P\left(Ob \cap E\right)}{P(E)} = \dfrac{0,1}{0,15} = 0,67$

5. $P\left(\overline{Ob} \cap \overline{E}\right) = P\left(\overline{Ob \cup E}\right) = 1 - P\left(Ob \cup E\right) = 1 - 0,3 = 0,7$

6. $P\left(\overline{E}/_{Ob}\right) = 1 - P\left(E/_{Ob}\right) = 1 - 0,4 = 0,6$

7. E y Ob no son independientes, ya que: $P(E) = 0,15 \neq 0,4 = P\left(E/_{Ob}\right)$

TEOREMAS BÁSICOS DE LA PROBABILIDAD: TEOREMA DE LA PROBABILIDAD TOTAL Y TEOREMA DE BAYES

Los axiomas y propiedades de la probabilidad que de ellos se derivan, así como los resultados obtenidos para la probabilidad condicional permiten el abordaje de problemas, a priori, más complejos. Para ilustrar este tipo de situaciones se propone el siguiente ejemplo.

Ejemplo 1-7

La gravedad de una determinada patología ha sido clasificada en cuatro niveles de menor a mayor (estadio I, II, III y IV). Se sabe que el éxito del tratamiento es del 95% si la patología se encuentra en estadio I, del 80% en el II, del 60% en el III y del 30% en el IV. Por otra parte, se conoce que el 35% de los individuos diagnosticados son clasificados en estadio I, el 30% en estadio II, el 25% en estadio III y el 10% en estadio IV. Si se determina que un individuo de la población padece la patología pero se desconoce todavía su estadio, ¿cuál es la probabilidad de que el tratamiento tenga éxito?

En primer lugar será necesario traducir al lenguaje de sucesos y probabilidades los datos proporcionados.

$$\begin{aligned}
&e = \{\text{éxito del tratamiento}\},\\
&I = \{\text{patología en estadio I}\};\ II = \{\text{patología en estadio II}\}\\
&III = \{\text{patología en estadio III}\};\ IV = \{\text{patología en estadio IV}\}\\
&P\left(e/_{I}\right) = 0,95;\ P\left(e/_{II}\right) = 0,8;\ P\left(e/_{III}\right) = 0,6;\ P\left(e/_{IV}\right) = 0,3\\
&P(I) = 0,35;\ P(II) = 0,3;\ P(III) = 0,25;\ P(IV) = 0,1
\end{aligned}$$

Si la probabilidad de que la patología se clasificara en cada uno de los estadios fuera la misma, para obtener la probabilidad total de éxito bastaría con calcular la media de las cuatro probabilidades de éxito de la forma:

$$\begin{aligned}
P(e) &= \frac{P\left(e/_{I}\right) + P\left(e/_{III}\right) + P\left(e/_{III}\right) + P\left(e/_{IV}\right)}{4}\\
&= \frac{1}{4}P\left(e/_{I}\right) + \frac{1}{4}P\left(e/_{II}\right) + \frac{1}{4}P\left(e/_{III}\right) + \frac{1}{4}\left(e/_{IV}\right)\\
&= 0,25P\left(e/_{I}\right) + 0,25P\left(e/_{II}\right) + 0,25P\left(e/_{III}\right) + 0,25P\left(e/_{IV}\right)
\end{aligned}$$

Sin embargo, según los datos proporcionados, el 35% de los casos son clasificados en el estadio I, luego sería lógico pensar que debería pesar más en el resultado final de éxito del tratamiento la probabilidad asociada al estadio I que, por ejemplo, la correspondiente al estadio IV, que solo cuenta con un 10% de casos diagnosticados. Sería lógico, por tanto, calcular una media ponderada de la forma:

$$\begin{aligned}
P(e) &= P\left(e/_{I}\right)0,35 + P\left(e/_{II}\right)0,3 + P\left(e/_{III}\right)0,25 + P\left(e/_{IV}\right)0,1\\
&= P\left(e/_{I}\right)P(I) + P\left(e/_{II}\right)P(II) + P\left(e/_{III}\right)P(III) + P\left(e/_{IV}\right)P(IV)\\
&= (0,95 \cdot 0,35) + (0,8 \cdot 0,3) + (0,6 \cdot 0,25) + (0,3 \cdot 0,1) = 0,752
\end{aligned}$$

Obsérvese que los sucesos I, II, III y IV son una partición del espacio muestral. En general podrá procederse de esta forma en las condiciones que establece el teorema de la probabilidad total.

TEOREMA DE LA PROBABILIDAD TOTAL

Sean A_1, A_2, A_3…, A_k una partición del espacio muestral y sea B otro suceso aleatorio cualesquiera. Entonces se verificará que:

$$P(B) = P\left(B/A_1\right)P(A_1) + P\left(B/A_2\right)P(A_2) + \cdots + P\left(B/A_k\right)P(A_k) = \sum P\left(B/A_i\right)P(A_i)$$

Continuando con el ejemplo 1-7, si un individuo diagnosticado de dicha patología ha sido tratado con éxito, ¿cuál es la probabilidad de que hubiera sido clasificado en el estadio III?

Para resolver esta cuestión, será necesario utilizar la definición de la probabilidad condicional combinada con el teorema de la probabilidad total. Así se tiene que, por la definición de la probabilidad condicional:

$$P\left(III/e\right) = \frac{P(III \cap e)}{P(e)}$$

Teniendo en cuenta que:

$$P\left(e/III\right) = \frac{P(e \cap III)}{P(III)}; P(e \cap III) = P\left(e/III\right) \cdot P(III)$$

Se tendrá que:

$$P\left(III/e\right) = \frac{P\left(e/III\right) \cdot P(III)}{P(e)}$$

Desarrollando el denominador aplicando el teorema de la probabilidad total, se obtendrá finalmente que:

$$P\left(III/e\right) = \frac{P\left(e/III\right) \cdot P(III)}{P(e)}$$

$$= \frac{P\left(e/III\right) \cdot P(III)}{P\left(e/I\right)P(I) + P\left(e/II\right)P(II) + P\left(e/III\right)P(III) + P\left(e/IV\right)P(IV)}$$

$$= \frac{0,60 \cdot 0,25}{0,752} = 0,199$$

Puede observarse que para la obtención del resultado final ha sido necesario considerar nuevamente que I, II, III y IV constituyen una partición del espacio muestral. Este resultado puede generalizarse en las condiciones que establece el teorema de Bayes.

TEOREMA DE BAYES

Sean A_1, A_2, A_3…, A_k una partición del espacio muestral y sea B otro suceso aleatorio cualesquiera. Entonces se verificará que:

$$P\left(A_i/B\right) = \frac{P\left(B/A_i\right)P(A_i)}{P(B)} = \frac{P\left(B/A_i\right)P(A_i)}{\sum P\left(B/A_i\right)P(A_i)}$$

APLICACIONES DE LOS TEOREMAS BÁSICOS DE LA PROBABILIDAD AL DIAGNÓSTICO/DETECCIÓN DE UNA ENFERMEDAD

Una prueba diagnóstica es un conjunto de intervenciones sobre un individuo que pretenden establecer la presencia o ausencia de una determinada patología. Una gran cantidad de pruebas diagnósticas utilizadas en la actualidad, sobre todo en primera instancia, no son capaces de determinar con exactitud el estado de salud del individuo pero sí contribuyen a reducir la incertidumbre que lo rodea. ¿Qué quiere decir esto? Simplemente que, a pesar de que se cometen errores captando como positivos personas sanas y como negativos personas enfermas, se puede determinar la probabilidad, por ejemplo, de que un individuo que ha dado positivo en la prueba padezca realmente la enfermedad. Por otra parte, los conceptos y herramientas descritos en este apartado son igualmente aplicables a otros ámbitos, pudiéndose estar interesado en determinar la presencia o ausencia de una determinada característica en un individuo o elemento (p. ej., presencia o ausencia de agentes contaminantes en los alimentos). Los sucesos según el estado de salud del individuo y el resultado de la prueba diagnóstica se resumen en la tabla 1-4.

Las probabilidades que intervienen en el estudio y aplicación de una prueba diagnóstica reciben nombres especiales y se recogen en la tabla 1-5.

La *prueba diagnóstica perfecta* no produciría falsos positivos ni negativos y se tendría que:

$$P\left(+/\bar{E}\right) = P\left(-/E\right) = 0$$

TABLA 1-4 Sucesos según el estado de salud del individuo y el resultado de la prueba diagnóstica

Estado de salud del individuo	Resultado de la prueba
Enfermo (E)	Positivo: +
No enfermo ($\bar{E}$)	Negativo: −

TABLA 1-5 Probabilidades según el estado de salud del individuo y el resultado de la prueba diagnóstica

Probabilidades sobre el estado de salud	Probabilidades sobre el resultado de la prueba
$P\left(\frac{E}{+}\right)$ = Valor predictivo positivo	$P\left(\frac{+}{E}\right)$ = Sensibilidad
$P\left(\frac{\bar{E}}{-}\right)$ = Valor predictivo negativo	$P\left(\frac{-}{\bar{E}}\right)$ = Especificidad
$P\left(\frac{E}{-}\right)$ = Valor predictivo falso negativo	$P\left(\frac{+}{\bar{E}}\right)$ = Probabilidad de falso positivo
$P\left(\frac{\bar{E}}{+}\right)$ = Valor predictivo falso positivo	$P\left(\frac{-}{E}\right)$ = Probabilidad de falso negativo
P(E) = Probabilidad de estar enfermo (prevalencia de la enfermedad)	P(+) = Probabilidad de resultado positivo
$P(\bar{E})$ = Probabilidad de no estar enfermo	P(−) = Probabilidad de resultado negativo

En este caso, la sensibilidad y la especificidad de la prueba serían iguales a 1, ya que son sus sucesos complementarios condicionales:

$$\text{Sensibilidad} = P\left(\frac{+}{E}\right) = 1 - P\left(\frac{-}{E}\right);\ \text{Especificidad} = P\left(\frac{-}{\bar{E}}\right) = 1 - P\left(\frac{+}{\bar{E}}\right)$$

De donde:

$$\text{Sensibilidad} = P\left(\frac{+}{E}\right) = 1 = P\left(\frac{-}{\bar{E}}\right) = \text{Especificidad}$$

Lo habitual es que la *sensibilidad* y *especificidad* de una prueba sean conocidas (la prueba fue evaluada sobre un grupo de individuos enfermos y sanos seleccionados previamente determinándose los valores de sensibilidad, especificidad, falsos positivos y negativos) y se pretenda utilizar en una población determinada. En este caso, será necesario conocer la prevalencia de la enfermedad en dicha población (probabilidad de estar enfermo) para calcular probabilidades relacionadas con el estado de salud del individuo (habitualmente los valores predictivos positivo y negativo). En este proceso, será determinante la consideración de los teoremas de la probabilidad total y del teorema de Bayes.

Ejemplo 1-8

Una prueba diagnóstica para la detección de una determinada patología fue ensayada sobre un grupo de individuos, determinándose una sensibilidad de 0,97 y una especificidad de 0,92. Si la prevalencia de la enfermedad en la población es del 6%, ¿cuál será la probabilidad de que una persona de dicha

población sobre la que la prueba ha resultado positiva padezca realmente la enfermedad?

$$\text{Sensibilidad} = P\left(+ \middle/ E\right) = 0,97 \,; \text{Especificidad} = P\left(- \middle/ \overline{E}\right) = 0,92$$

$$\text{Probabilidad de falso negativo} = P\left(- \middle/ E\right) = 1 - P\left(+ \middle/ E\right) = 1 - 0,97 = 0,03$$

$$\text{Probabilidad de falso positivo} = P\left(+ \middle/ \overline{E}\right) = 1 - P\left(- \middle/ \overline{E}\right) = 1 - 0,92 = 0,08$$

Además, dado que la prevalencia de la enfermedad en la población es del 6%, se tendrá que:

$$P(E) = 0,06$$

Lo que se pretende determinar es el valor predictivo positivo: $P\left(E \middle/ +\right)$. Aplicando el teorema de Bayes se tendrá que:

$$P\left(E \middle/ +\right) = \frac{P\left(+ \middle/ E\right)P(E)}{P(+)} = \frac{0,97 \cdot 0,06}{P(+)} = \frac{0,97 \cdot 0,06}{P\left(+ \middle/ E\right)P(E) + P\left(+ \middle/ \overline{E}\right)P\left(\overline{E}\right)}$$

Teniendo en cuenta que $P\left(\overline{E}\right) = 1 - P(E) = 1 - 0,06 - 0,94$, se tendrá que:

$$P\left(E \middle/ +\right) = \frac{0,97 \cdot 0,06}{\left(0,97 \cdot 0,06\right) + \left(0,08 \cdot 0,94\right)} = 0,436$$

Luego, un individuo sobre el que la prueba ha resultado positiva tiene un 43,6% de posibilidades de padecer realmente la enfermedad y, en consecuencia, un 56,4% de no padecerla. Obsérvese que, en este caso, se tienen más probabilidades de no padecer la enfermedad que de padecerla habiendo dado positivo en la prueba diagnóstica. Por su parte, si se calcula el valor predictivo negativo se tendrá que:

$$P\left(\overline{E} \middle/ -\right) = \frac{P\left(- \middle/ \overline{E}\right)P\left(\overline{E}\right)}{P(-)} = \frac{0,92 \cdot 0,94}{P\left(- \middle/ E\right)P(E) + P\left(- \middle/ \overline{E}\right)P\left(\overline{E}\right)} = \frac{0,92 \cdot 0,94}{\left(0,03 \cdot 0,06\right) + \left(0,92 \cdot 0,94\right)}$$

$$= 0,998$$

Como puede observarse, el valor predictivo negativo es prácticamente 1, por lo que si sobre un individuo la prueba resulta negativa es casi seguro que no padecerá la enfermedad.

Tal y como fue expresado con anterioridad, ha sido posible obtener los valores predictivos de una prueba diagnóstica con una determinada sensibilidad y especificidad, gracias a que la prevalencia de la enfermedad en la población sobre la que se aplica era conocida. ¿Qué hubiera ocurrido con los valores predictivos (positivo y negativo) si esta misma prueba hubiera sido aplicada sobre una población con una prevalencia de la enfermedad distinta, por ejemplo, del 1%?

Los valores de la sensibilidad y especificidad no varían, pero la probabilidad de padecer la enfermedad sería ahora:

$$P(E) = 0,01$$

Los valores predictivos positivo y negativo serán en este caso:

$$P\left(\frac{E}{+}\right) = \frac{P\left(\frac{+}{E}\right)P(E)}{P\left(\frac{+}{E}\right)P(E) + P\left(\frac{+}{\bar{E}}\right)P\left(\bar{E}\right)} = \frac{0,97 \cdot 0,01}{\left(0,97 \cdot 0,01\right) + \left(0,08 \cdot 0,99\right)} = 0,109$$

$$P\left(\frac{E}{-}\right) = \frac{P\left(\frac{-}{E}\right)P\left(\bar{E}\right)}{P\left(\frac{-}{E}\right)P(E) + P\left(\frac{-}{\bar{E}}\right)P\left(\bar{E}\right)} = \frac{0,92 \cdot 0,99}{\left(0,03 \cdot 0,01\right) + \left(0,92 \cdot 0,99\right)} = 0,999$$

Obsérvese que al disminuir la prevalencia de la enfermedad también disminuye el valor predictivo positivo (pasa de 0,436 a 0,109), mientras que el valor predictivo negativo aumenta (pasa de 0,998 a 0,999). Siguiendo este mismo razonamiento a la inversa, un aumento en la prevalencia de la enfermedad en la población conllevaría un aumento del valor predictivo positivo y una disminución del valor predictivo negativo.

VARIABLES ALEATORIAS Y MODELOS DE PROBABILIDAD

CONCEPTOS PREVIOS

Desde un punto de vista formal una variable aleatoria se define como una función que asigna a cada uno de los posibles resultados de un fenómeno aleatorio un valor numérico. Por ejemplo, en el caso de variables cuantitativas como el *nivel de colesterol*, *nivel de ácido úrico* o *número de ingresos en un servicio de urgencias*, la variable aleatoria vendría definida por cada una de las posibilidades de las variables consideradas, puesto que en este caso ya son valores numéricos. Cuando las variables son de tipo cualitativo como el *sexo* o el *nivel de estudios*, será necesario asignar valores numéricos a cada una de las posibilidades (p. ej., 1 Hombre, 2 Mujer, para el sexo;

1 Sin estudios, 2 Primaria, 3 Secundaria y 4 Universitarios, para el nivel de estudios).

Una consecuencia inmediata derivada de este proceso de asignación de valores numéricos es que las variables aleatorias se clasifican únicamente en: variables aleatorias *discretas* y variables aleatorias *continuas.* Las variables aleatorias discretas pueden tomar un número finito o infinito numerable de valores, mientras que las continuas pueden alcanzar un número infinito no numerable de posibles valores, es decir, pueden tomar cualquier valor en un intervalo.

Desde un punto de vista más práctico, las variables aleatorias podrían considerarse como variables cuyos resultados se rigen por el azar. En este sentido es importante tener en cuenta que, como se ha mencionado con anterioridad, no todos los posibles valores de una variable aleatoria tienen la misma probabilidad de ser observados. Por tanto, sería útil contar con herramientas que proporcionen información sobre la probabilidad asociada a cada uno de los valores de una variable aleatoria.

FUNCIÓN DE PROBABILIDAD PARA UNA VARIABLE ALEATORIA DISCRETA

Se dice que $p(x)$ es una función de probabilidad para la variable aleatoria X si $p(x)$ es una función no negativa (no puede dar lugar a valores negativos) tal que, para cualquier valor de la variable, por ejemplo x_0, la función devuelve su probabilidad, es decir:

$$p(x_0) = P(X = x_0)$$

Si se trabaja con los datos del ejemplo 1-5, relativos al grupo sanguíneo, puede construirse una variable aleatoria asignando un número a cada una de las categorías, obteniéndose el resultado contenido en la tabla 1-6.

La probabilidad de que un individuo seleccionado al azar de entre los 150 pertenezca al grupo sanguíneo AB podría expresarse de la siguiente forma:

$$P(AB) = P(X = 3) = p(3) = \frac{8}{150}$$

TABLA 1-6 Función de probabilidad para la variable aleatoria «grupo sanguíneo»

Grupo sanguíneo	X = grupo sanguíneo	Individuos	$p(x_0) = P(X = x_0)$
A	1	63	63/150
B	2	18	18/150
AB	3	8	8/150
O	4	61	61/150
Total		**150**	

De forma análoga, la probabilidad de que pertenezca al grupo AB o al grupo O podría escribirse:

$$P(AB \cup O) = P(X = 3) + P(X = 4) = p(3) + p(4) = \frac{8}{150} + \frac{61}{150} = \frac{69}{150}$$

Puede establecerse que, en general, para cualquier variable aleatoria discreta X su función de probabilidad $p(x)$ verificará que:

$$\sum p(x) = 1$$

FUNCIÓN DE DISTRIBUCIÓN PARA UNA VARIABLE ALEATORIA DISCRETA

Se dice que $F(x)$ es una función de distribución para la variable aleatoria X si $F(x)$ es una función no negativa tal que, para cada valor de la variable, por ejemplo x_0, la función devuelve la probabilidad de que la variable tome un valor menor o igual a x_0, es decir:

$$F(x_0) = P(x \leq x_0)$$

En el caso del ejemplo 1-5, puede construirse una nueva columna para la función de distribución que quedará como se recoge en la tabla 1-7.

La probabilidad de que un individuo seleccionado al azar de entre los 150 pertenezca al grupo A, B o AB podría expresarse de la siguiente forma:

$$P(A \cup B \cup AB) = P(X \leq 3) = F(3)$$

Además, se tendrá que:

$$F(3) = P(X = 1) + P(X = 2) + P(X = 3) = p(1) + p(2) + p(3) = \frac{63 + 18 + 8}{150} = \frac{89}{150}$$

TABLA 1-7 Función de probabilidad y función de distribución para la variable aleatoria «grupo sanguíneo»

X = grupo sanguíneo	Individuos	$p(x_0) = P(X = x_0)$	$F(x_0) = P(X \leq x_0)$
1	63	63/150	63/150
2	18	18/150	(63 + 18)/150
3	8	8/150	(63 + 18 + 8)/150
4	61	61/150	(63 + 18 + 8 + 61)/150 = 1
	150		

Puede establecerse que, en general, para cualquier variable aleatoria discreta su función de distribución verificará que:

$$F(x_0) = \sum_{x_i \leq x_0} x_i$$

ESPERANZA O MEDIA Y DESVIACIÓN TÍPICA DE UNA VARIABLE ALEATORIA DISCRETA

Las expresiones para el cálculo de la media y la desviación típica para variables aleatorias discretas constituyen una extensión del cálculo de estas medidas descriptivas basadas en valores y frecuencias de dichos valores al caso en que se cuenta con los valores y las probabilidades de dichos valores. Así, mientras que para el caso basado en frecuencias la media se obtenía:

$$\bar{x} = \frac{\sum f_i x_i}{n} = \sum \frac{f_i x_i}{n} = \sum x_i \frac{f_i}{n}$$

En el caso de contar con probabilidades se calculará de la forma:

$$E(X) = \sum x_i \, p(x_i)$$

Puede observarse la similitud entre las expresiones anteriores, ya que f_i/n es utilizado como la estimación de la probabilidad $p(x_i)$. Si se trabaja con los datos del ejemplo 1-5 se tendrá que:

$$\begin{aligned} E(X) &= \sum x_i p(x_i) = 1 \cdot p(1) + 2 \cdot p(2) + 3 \cdot p(3) + 4 \cdot p(4) \\ &= 1 \cdot \frac{63}{150} + 2 \cdot \frac{18}{150} + 3 \cdot \frac{8}{150} + 4 \cdot \frac{61}{150} = \frac{367}{150} = 2,447 \end{aligned}$$

Por su parte, la varianza y la desviación típica que para valores y frecuencias se expresaba:

$$S^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n} = \sum \frac{f_i (x_i - \bar{x})^2}{n} = \sum (x_i - \bar{x})^2 \frac{f_i}{n}; S = \sqrt{S^2}$$

Se expresarán ahora:

$$V(X) = \sum (x_i - E(X))^2 \, p(x_i) \quad ; \quad D(X) = \sqrt{V(X)}$$

Debe tenerse en cuenta que, como la variable *grupo sanguíneo* no toma originalmente valores numéricos, los resultados de la media o esperanza y la desviación típica no aportan información relevante e, incluso, carecen de sentido en este caso. Sin embargo, ha sido útil para ilustrar el procedimiento.

INTRODUCCIÓN A LOS MODELOS DE PROBABILIDAD PARA VARIABLES ALEATORIAS DISCRETAS

El estudio de determinados fenómenos aleatorios ha permitido generalizar y reproducir el cálculo de probabilidades sobre sus resultados en situaciones equivalentes. Debe hacerse notar que un modelo de probabilidad para una variable aleatoria discreta debe proporcionar una expresión concreta para la *función de probabilidad* correspondiente *p(x)*. En este apartado se pretender abordar el estudio de algunos modelos de probabilidad para variables aleatorias discretas a modo de introducción y con el objetivo de familiarizar al lector con los elementos que intervienen en el proceso de cálculo de probabilidades. Se propone el siguiente ejemplo:

Ejemplo 1-9

Suponga que se conoce que el 10% de los individuos de una población posee una determinada característica (alergia, cáncer de colon, alteraciones visuales, etc.). Si se seleccionan al azar tres individuos de dicha población, ¿cuál será la probabilidad de que exactamente dos de ellos posean la característica?

En la tabla 1-8 se muestran todas las situaciones favorables al suceso en el que solo dos de los tres individuos seleccionados posean la característica.

Como puede apreciarse, se observan tres casos distintos en los que se verificaría que solo dos de los tres individuos poseerían la característica. La probabilidad de que exactamente dos de los tres la posean podría expresarse de la forma:

$$P(2 \text{ de los } 3 \text{ posean la característica}) = P(C_1 \cup C_2 \cup C_3)$$

Como los sucesos C_1, C_2 y C_3 son mutuamente excluyentes o incompatibles (no pueden darse a la vez, ya que eso implicaría que, por ejemplo, el individuo 2 poseyera la característica según C_1 y no la poseyera al mismo tiempo según establece C_2), se tendrá que esta probabilidad puede expresarse como la suma de probabilidades:

$$P(C_1 \cup C_2 \cup C_3) = P(C_1) + P(C_2) + P(C_3)$$

TABLA 1-8 Posibilidades favorables al caso en que 2 de los 3 individuos analizados posean la característica

Suceso	Individuo 1	Individuo 2	Individuo 3
C_1	Sí	Sí	No
C_2	Sí	No	Sí
C_3	No	Sí	Sí

Por otra parte, la probabilidad de cada uno de estos sucesos que exige que solo dos de los tres posean la característica se expresará de la siguiente forma:

$$P(C_1) = P(Sí_1 \cap Sí_2 \cap No_3); P(C_2) = P(Sí_1 \cap No_2 \cap Sí_3)$$

$$P(C_3) = P(No_1 \cap Sí_2 \cap Sí_3)$$

Si se supone que las ocurrencias del suceso son independientes entre sí (que un individuo de la población posea la característica es independiente de que otro la posea o no), la probabilidad de la intersección podrá expresarse como el producto de probabilidades:

$$\begin{aligned} P(C_1 \cup C_2 \cup C_3) &= P(Sí_1)P(Sí_2)P(No_3)+P(Sí_1)P(No_2)P(Sí_3)+P(No_1)P(Sí_2)P(Sí_3) \\ &= 0,1 \cdot 0,1 \cdot (1-0,1)+0,1 \cdot (1-0,1) \cdot 0,1+(1-0,1) \cdot 0,1 \cdot 0,1 \\ &= 3 \cdot 0,1^2 (1-0,1)^1 = 0,027 \end{aligned}$$

Como puede observarse, al final del proceso, por cada individuo que posea la característica se multiplicará por 0,1. Por cada individuo que no posea la característica se multiplicará por (1 − 0,1). Esta cantidad se multiplicará, además, por el número de combinaciones posibles en las que se puede extraer dos elementos entre tres que, en este caso, son tres combinaciones. Por tanto, el resultado final, podrá expresarse de la forma:

$$P(\text{2 de los 3 posean la característica}) = \binom{3}{2} 0,1^2 (1-0,1)^{3-2} = 0,027$$

EL MODELO DE PROBABILIDAD BINOMIAL

Sea un fenómeno aleatorio que puede dar lugar a dos posibles resultados (enfermo/no enfermo; defectuoso/no defectuoso; etc.) y sea *p* la probabilidad de ocurrencia de uno de los dos sucesos en una determinada población. Si se selecciona un grupo de *N* individuos o elementos de la población y se considera que las ocurrencias del suceso son independientes entre sí, entonces, la probabilidad de que se produzcan *k* ocurrencias del suceso entre los *N* individuos o elementos seleccionados puede calcularse de la siguiente forma:

$$P(X = k) = \binom{N}{k} p^k (1-p)^{N-k}; k = 0, 1, 2, 3 \ldots, N$$

Donde *X = número de ocurrencias del suceso entre los N individuos o elementos*. Por su parte, la media o esperanza y la desviación típica, en aplicación de las expresiones obtenidas en el apartado anterior, quedarán:

$$E(X) = Np$$

$$V(X) = Np(1-p) \quad ; \quad D(X) = \sqrt{Np(1-p)}$$

Ejemplo 1-10

Un estudio permitió establecer que el 15% de los individuos de una población padece una determinada patología. Se selecciona un grupo de cinco individuos de dicha población y se desea determinar la probabilidad de que:

- Dos de ellos padezcan la patología.
- No más de dos padezcan la patología.
- Entre dos y cuatro padezcan la patología.
- Ninguno padezca la patología.

Además, se desea conocer el número esperado de afectados por la patología y su desviación típica.

Si se supone que el hecho de que un individuo padezca la patología es independiente de que otro individuo la padezca o no, entonces la variable aleatoria X = *número de afectados por la patología entre los cinco* se distribuirá según un modelo de probabilidad binomial con $N = 5$ y $p = 0{,}15$. Se tendrá, por tanto, que:

$$P(\text{k de los 5 estén afectados}) = P(X=k) = \binom{5}{k} = 0{,}15^k\,(1-0{,}15)^{5-k}$$

En consecuencia, se tendrá que:

- $P(X=2) = \binom{5}{2} 0{,}15^2\,(1-0{,}15)^3 = 0{,}138$
- $P(X \le 2) = P(X=0) + P(X=1) + P(X=2) = \binom{5}{0} 0{,}15^0\,(1-0{,}15)^5 + \binom{5}{1} 0{,}15^1\,(1-0{,}15)^4 + \binom{5}{2} 0{,}15^2\,(1-0{,}15)^3 = 0{,}973$
- $P(2 \le X \le 4) = P(X=2) + P(X=3) + P(X=4) = \binom{5}{2} 0{,}15^2\,(1-0{,}15)^3 + \binom{5}{3} 0{,}15^3\,(1-0{,}15)^2 + \binom{5}{4} 0{,}15^4\,(1-0{,}15)^1 = 0{,}165$
- $P(X=0) = \binom{5}{0} 0{,}15^0\,(1-0{,}15)^5 = 0{,}444$

En cuanto a la media o esperanza y la desviación típica:

$$E(X) = Np = 5 \cdot 0{,}15 = 0{,}75 \quad ; \quad D(X) = \sqrt{Np(1-p)} = \sqrt{5 \,.\, 0{,}15 \,.\, 0{,}85} = 0{,}798$$

EL MODELO DE PROBABILIDAD POISSON

Sea un fenómeno aleatorio que puede dar lugar a dos posibles resultados y sea λ el promedio de ocurrencias de uno de los dos sucesos en un intervalo, generalmente de tiempo o espacio. Si se considera que el número de ocurrencias en un intervalo de tiempo o espacio es proporcional a la amplitud del intervalo y que las ocurrencias del suceso son independientes entre sí, entonces, la probabilidad de que se produzcan *k* ocurrencias del suceso en ese mismo intervalo de tiempo o espacio puede calcularse de la siguiente forma:

$$P(X=k)=\frac{e^{-\lambda}\lambda^{k}}{k!};\ k=0,1,2,3\ldots+\infty$$

Donde X = *número de ocurrencias del suceso en el intervalo considerado.* La media o esperanza y la desviación típica en este caso se calcularán:

$$E(X)=\lambda$$

$$V(X)=\lambda;\ D(X)=\sqrt{\lambda}$$

Ejemplo 1-11

El número de llegadas al servicio de urgencias de un hospital es de siete cada 15 min. Si se supone que el número de llegadas se distribuye según un modelo de Poisson, se pretende calcular la probabilidad de que en los próximos 15 min se produzcan:

- Ocho llegadas.
- Entre seis y ocho llegadas.
- No más de dos llegadas.
- Más de una llegada.

Además, se desea conocer el número esperado de llegadas en los próximos 30 min y su desviación típica.

La variable X = *número de ocurrencias del suceso en los próximos 15 min* se distribuirá según un modelo de Poisson con media $\lambda = 7$. Se tendrá que:

$$P(X=k)=\frac{e^{-7}7^{k}}{k!}$$

Las preguntas podrán expresarse de la forma:

- $P(X=8)=\dfrac{e^{-7}7^{8}}{8!}=0,13$
- $P(6\le X\le 8)=P(X=6)+P(X=7)+P(X=8)=\dfrac{e^{-7}7^{6}}{6!}+\dfrac{e^{-7}7^{7}}{7!}+\dfrac{e^{-7}7^{8}}{8!}=0,428$

- $P(X \leq 2) = P(X=0) + P(X=1) + P(X=2) = \frac{e^{-7}7^0}{0!} + \frac{e^{-7}7^1}{1!} + \frac{e^{-7}7^2}{2!} = 0,03$
- $P(X>1) = 1 - P(X \leq 1) = 1 - \left[P(X=0) + P(X=1)\right] =$

$$= 1 - \left[\frac{e^{-7}7^0}{0!} + \frac{e^{-7}7^1}{1!}\right] = 1 - 0,007 = 0,993$$

Por otra parte, se pretende averiguar el número esperado de llegadas en los próximos 30 min y su desviación típica. Deberá tenerse en cuenta que, en este caso, la variable y el modelo de probabilidad de Poisson deberá ajustarse al período de tiempo para el que se pretende calcular. Si el promedio de llegadas en 15 min es de siete, entonces el promedio de llegadas en 30 min será $7 \cdot 2 = 14$. Así, la variable será ahora X = *número de ocurrencias del suceso en los próximos 30 min* y el modelo de Poisson adecuado quedará:

$$P(X=k) = \frac{e^{-14}14^k}{k!}$$

Y la esperanza o media y la desviación típica se obtendrán de la siguiente forma:

$$E(X) = 14\,;\, D(X) = \sqrt{14} = 3,74$$

INTRODUCCIÓN A LOS MODELOS DE PROBABILIDAD PARA VARIABLES ALEATORIAS CONTINUAS

Cuando se trabaja con variables aleatorias continuas debe tenerse en cuenta que, al tomar cualquier valor en un intervalo, no es posible proceder como en el caso de las variables discretas a la hora de proponer un modelo de probabilidad que rija su comportamiento ya que toman un número infinito no numerable de valores.

Para comprender la definición y el funcionamiento de los modelos de probabilidad para este tipo de variables se partirá de uno de los tipos de representación gráfica más adecuados como el histograma. En las figuras 1-7 a 1-9 se muestra el histograma para una variable continua (p. ej., la *talla* en centímetros) basado en un determinado grupo de observaciones cada vez más elevado.

Como puede verse, a medida que aumenta el número de observaciones, los intervalos que determinan la anchura de las barras del histograma pueden hacerse cada vez más estrechos. En el límite (llevando la anchura de los intervalos a 0) podrá construirse una curva suave $f(x)$ que represente la distribución de los datos, tal y como se muestra en la figura 1-9.

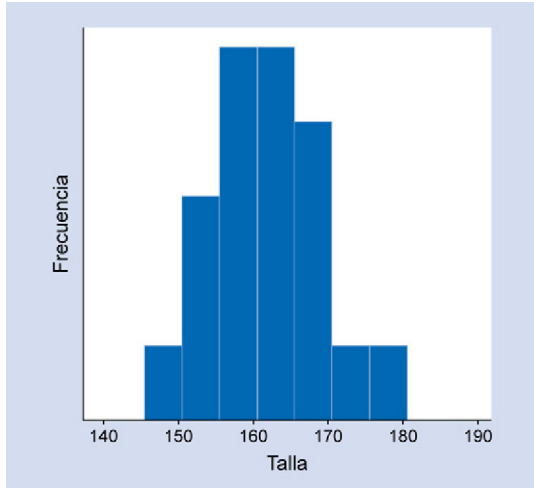

FIGURA 1-7 Histograma de talla para $n = 80$.

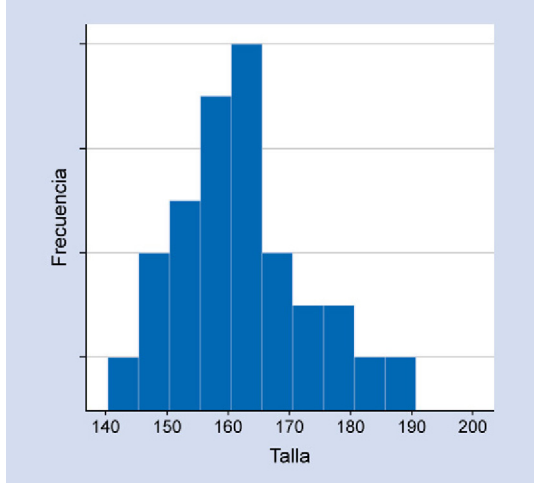

FIGURA 1-8 Histograma de talla para $n = 200$.

Esta función se conoce como la *función de densidad de probabilidad*. Si se tiene la precaución de construir los histogramas de forma que el área de cada barra coincida con la frecuencia relativa en cada uno de los intervalos, se conseguirá que la suma de las áreas de las barras sea 1 y que la probabilidad de que la variable tome un valor entre dos dados coincida con la suma de las áreas de las barras correspondientes. Como consecuencia, se

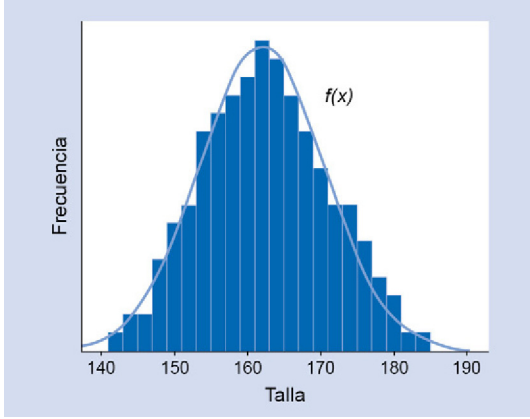

FIGURA 1-9 Histograma de talla para $n = 2.500$.

tendrá que, tomando como base la función *f(x)*, la probabilidad de que la variable tome un valor entre dos dados coincidirá con el área bajo la curva entre los dos valores y que el área total bajo la curva será 1.

FUNCIÓN DE DENSIDAD DE PROBABILIDAD PARA UNA VARIABLE ALEATORIA CONTINUA

Se dice que *f(x)* es una función de densidad de probabilidad para la variable aleatoria *X* si *f(x)* es una función no negativa (no puede dar lugar a valores negativos) tal que, para cualesquiera dos valores x_1 y x_2, se verifica:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)dx$$

Es decir, la probabilidad de que la variable tome un valor entre dos dados se obtiene como el área bajo la curva entre dichos valores, tal y como puede observarse en la figura 1-10. En consecuencia, el área total bajo la curva debe ser igual a 1.

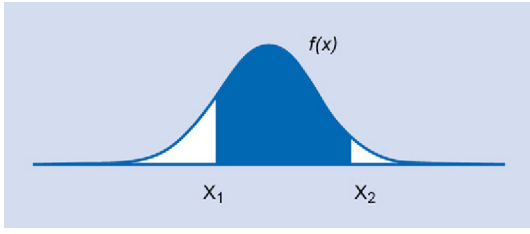

FIGURA 1-10 Función de densidad de probabilidad.

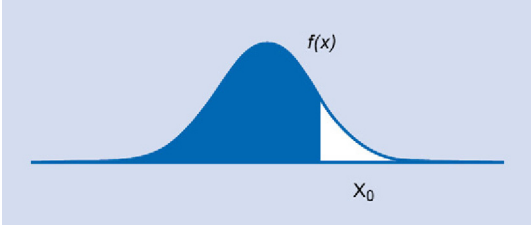

FIGURA 1-11 Área correspondiente a la función de distribución en x_0.

$$P(-\infty \leq X \leq +\infty) = \int_{-\infty}^{+\infty} f(x)dx = 1$$

FUNCIÓN DE DISTRIBUCIÓN PARA UNA VARIABLE ALEATORIA CONTINUA

Se dice que *F(x)* es una función de distribución para la variable aleatoria *X* si *F(x)* es una función no negativa tal que, para cada valor de la variable, por ejemplo x_0, la función devuelve la probabilidad de que la variable tome un valor menor o igual a x_0, es decir:

$$F(x_0) = P(X \leq x_0) = \int_{-\infty}^{x_0} f(x)dx$$

Gráficamente, esta probabilidad vendría determinada por el área bajo la curva desde $-\infty$ hasta x_0 tal y como se refleja en la figura 1-11.

ESPERANZA O MEDIA Y DESVIACIÓN TÍPICA DE UNA VARIABLE ALEATORIA CONTINUA

Las expresiones para el cálculo de la media y la desviación típica para variables aleatorias continuas, constituyen una extensión del cálculo de estas medidas descriptivas en variables discretas llevadas al ámbito del cálculo infinitesimal. Así, mientras que para el caso de las variables aleatorias discretas la media o esperanza se obtenía:

$$E(X) = \sum x_i \, p(x_i)$$

En el caso de las variables aleatorias continuas se sumará infinitesimalmente *xf(x)*, obteniéndose:

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

Por su parte la varianza y la desviación típica en el caso de variables aleatorias discretas se expresaba de la forma:

$$V(X) = \sum (x_i - E(X))^2 \, p(x_i) \, ; D(X) = \sqrt{V(X)}$$

Procediendo de manera análoga y sumando infinitesimalmente se obtendrá que, en el caso de variables aleatorias continuas, la varianza y la desviación típica podrán calcularse del siguiente modo:

$$V(X)=\int_{-\infty}^{+\infty}\left(x_i-E(X)\right)^2 f(x)dx\ ;\ D(X)=\sqrt{V(X)}$$

MODELO DE DISTRIBUCIÓN DE PROBABILIDAD NORMAL

El modelo normal constituye la distribución de probabilidad para variables aleatorias continuas más importante de toda la estadística, debido a que en la naturaleza muchas de las variables, y en particular las relacionadas con errores en los procesos de medición, se comportarían de forma aproximada según este modelo de probabilidad y, sobre todo, porque un resultado como el *teorema central del límite* asignará al modelo normal un papel destacado en el ámbito de la estadística inferencial.

La función de densidad de probabilidad que describe el modelo normal para una variable X con media μ y desviación típica σ es:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Debe tenerse en cuenta que, para cada valor de la media μ y de la desviación típica σ, se tratará de un modelo de probabilidad normal distinto, por lo que se puede afirmar que el modelo normal es una familia de distribuciones normales. En la figura 1-12 se pueden observar tres modelos normales. El primero y el segundo modelo normal tienen la misma media pero distinta desviación típica. El segundo y el tercer modelo normal tienen la misma desviación típica pero distinta media. Obsérvese además que, en este modelo, es evidente que:

$$E(X)=\mu;\ V(X)=\sigma^2\ ;\ D(X)=\sigma$$

En el modelo de probabilidad normal, la *media* coincide con la *mediana* y con la *moda*. (Téngase en cuenta que la distribución es simétrica y que

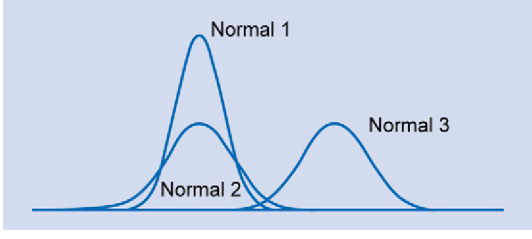

FIGURA 1-12 Familia de distribuciones normales.

el valor máximo de la función *f(x)* se alcanza en la media.) Por otra parte, la curva normal nunca toca el eje de abscisas, si bien, a medida que nos alejamos a la derecha o a la izquierda, cada vez se situará más cerca de él. Esto puede expresarse matemáticamente de la forma:

$$\lim_{x\to-\infty} f(x) = 0 = \lim_{x\to+\infty} f(x)$$

Para calcular manualmente probabilidades sobre un modelo normal suele utilizarse una tabla correspondiente a la normal con media 0 y desviación típica 1, conocida como la distribución *normal estándar*. Para ilustrar el método de cálculo se propone el siguiente ejemplo.

Ejemplo 1-12

Se sabe que el nivel de colesterol (en mg/100 ml) en una población se distribuye, aproximadamente, según un modelo de normal con media 205 mg/100 ml y desviación típica 36 mg/100 ml. Se selecciona al azar un individuo de dicha población y se pretende calcular la probabilidad de que:

- Su nivel de colesterol sea inferior a 190 mg/100 ml.
- Su nivel de colesterol se encuentre entre 180 y 220 mg/100 ml.
- Su nivel de colesterol sea superior a 230 mg/100 ml.

Sea X = nivel de colesterol.

- $P(X < 190)$.
- $P(180 < X < 220)$.
- $P(X > 230)$.

Dado que únicamente se dispone de la tabla correspondiente a la normal estándar, será necesario transformar la variable y la probabilidad en una probabilidad equivalente pero planteada sobre una normal de media 0 y desviación típica 1. Para conseguir que una variable tenga media 0 y desviación típica 1 se le resta la media y se divide por la desviación típica. Este proceso se conoce como *tipificación* o *estandarización* de la variable. Así, se tendrá que la variable estandarizada:

$$Z = \frac{X-\mu}{\sigma} = \frac{X-205}{36}$$

Se distribuirá según un modelo normal de media 0 y desviación típica 1 o normal estándar. Las probabilidades anteriores podrán expresarse ahora de la siguiente forma:

- $P(X<190)=P\left(\frac{X-205}{36}<\frac{190-205}{36}\right)=P\left(Z<\frac{190-205}{36}\right)$
  $=P(Z<-0,42)=0,337$
- $P(180<X<220)=P\left(\frac{180-205}{36}<\frac{X-205}{36}<\frac{220-205}{36}\right)$
  $=P(-0,69<X<0,42)=P(X<0,42)-P(X<-0,69)$
  $=0,663-0,245=0,418$
- $P(X>230)=P\left(\frac{X-205}{36}>\frac{230-205}{36}\right)=P(Z>0,69)=1-P(X<0,69)$
  $=1-0,755=0,245$

AUTOEVALUACIÓN

1. El número de llegadas al servicio de urgencias de un hospital es una variable:
- **a.** Cuantitativa continua.
- **b.** Cualitativa ordinal.
- **c.** Cualitativa no ordinal.
- **d.** Cuantitativa discreta.
- **e.** No procede porque no es una variable.

2. La mediana:
- **a.** Es una medida de tendencia central sensible a observaciones atípicas.
- **b.** Es una medida de dispersión de los datos observados.
- **c.** En el caso de existencia de observaciones atípicas es preferible a la media.
- **d.** Deja el mismo número de observaciones a su izquierda que a su derecha.
- **e.** c y d son ciertas.

3. Si A y B son sucesos aleatorios cualesquiera:
- **a.** $P(\overline{A})=1-P(A)$
- **b.** $P(A\cup B)=P(A)+P(B)$
- **c.** $P(A/B)=P(A)/P(B)$
- **d.** $P(A)\leq 0$
- **e.** $P(A\cap B)=P(A)\cdot P(B)$

4. El valor predictivo positivo de una prueba para el diagnóstico de una determinada enfermedad:
- **a.** Es la probabilidad de que un individuo sobre el que la prueba ha resultado positiva esté realmente enfermo.
- **b.** Es la probabilidad de que sobre un individuo enfermo la prueba dé positivo.

c. Es la probabilidad de que un individuo sano dé positivo en la prueba.
d. Es la probabilidad de que un individuo que padece la enfermedad dé negativo en la prueba.
e. Es la probabilidad de que un individuo sobre el que la prueba ha resultado positiva esté sano.

5. En el modelo normal:
a. La media coincide con la mediana, pero no con la moda.
b. El percentil 0,975 es 1,65 si se trata del modelo normal estándar.
c. La media es 0.
d. El coeficiente de asimetría es 0.
e. El 60% de los valores serán superiores a la media.

CAPÍTULO

2

Inferencia estadística

Joaquín Moncho Vasallo y Andreu Nolasco Bonmatí

INTRODUCCIÓN

Uno de los objetivos básicos en la investigación de un determinado fenómeno aleatorio consiste en extraer conclusiones acerca de una característica de interés sobre la *población* objeto de estudio (nivel promedio de colesterol, proporción de fumadores, diferencia en el nivel promedio de ácido úrico según sexo…) cuando únicamente se dispone de la información contenida en una *muestra* de dicha población. En este proceso de generalización de resultados de la muestra a la población, que recibe el nombre de *inferencia estadística,* jugará un papel decisivo la forma en que hayan sido obtenidos los datos de la muestra observada.

En este capítulo se abordarán los conceptos básicos de inferencia estadística, las técnicas inferenciales (intervalos de confianza y contrastes de hipótesis) y el muestreo aleatorio como base para la realización de las mismas.

La primera cuestión que cabe plantearse es por qué, en algunas ocasiones, únicamente se puede disponer de información sobre algunos de los individuos o elementos de la población y no sobre todos ellos. Por ejemplo:

- Un técnico especialista desea estimar el tiempo medio de duración de un lote correspondiente a un determinado tipo de prótesis sometiéndolas a diferentes pruebas de desgaste.
- En un estudio se pretende estimar el tiempo medio de espera en la sala de urgencias de un centro hospitalario.
- Se desea contrastar si la proporción de fumadores en España es superior al 45%.

En cada uno de los casos anteriores, ¿cuál sería la población objeto de estudio? ¿Es posible observar a todos los individuos de dicha población?

En el primer caso, la población la componen todas las prótesis del lote. Dado que la observación del tiempo de duración de una prótesis implica en este caso su destrucción, no tendría ningún sentido observar el tiempo de duración de todas las prótesis del lote, puesto que esto supondría la desaparición de todas ellas.

En el segundo caso, la población la formarían todos los posibles usuarios del servicio de urgencias. Esta población, en contra de lo que pudiera parecer, es una población infinita, puesto que un mismo usuario puede acudir más de una vez y existen infinitos instantes de tiempo en los que podrían producirse llegadas al servicio. Existe, por tanto, una imposibilidad real de observar a todos los individuos de esta población.

En el tercer caso, la población estaría compuesta por todos los habitantes del estado español. Aunque esta población podría llegar a ser teóricamente observada en su totalidad, el coste económico que supondría entrevistar a todos los habitantes podría hacerlo inviable.

En consecuencia, entre las razones por las que se cuenta con los datos relativos a una muestra y no a toda la población, se encontrarían las siguientes:

- En algunos estudios la recogida de la información puede suponer la destrucción del elemento.
- Los elementos que componen la población pueden existir conceptualmente pero no en la realidad.
- Puede ser económicamente inviable observar a toda la población objeto de estudio.

En cualquiera de estos casos, será necesario seleccionar un subconjunto de elementos de la población objeto de estudio sobre los que será observada la variable relacionada con la característica de interés.

POBLACIÓN Y MUESTRA

Se define como población a cualquier conjunto de individuos o elementos sobre el que se pretende estudiar una determinada característica. Cualquier subconjunto de individuos o elementos de dicha población constituirá una muestra. Adicionalmente, esta muestra será aleatoria si los individuos o elementos han sido seleccionados al azar mediante una técnica de muestreo aleatorio determinada.

PARÁMETRO

Se define como parámetro a cualquier característica cuantitativa de una o más variables de la población, generalmente desconocida, sobre la que se pretende realizar algún tipo de inferencia (estimar o contrastar).

TÉCNICAS INFERENCIALES. CONSIDERACIONES PREVIAS

Tal y como se ha discutido anteriormente, continuamente se pretende realizar afirmaciones sobre determinadas características *(parámetros)* de la población objeto de estudio cuando solamente se dispone de la información

contenida en una muestra de dicha población. El tipo de afirmaciones que se realicen sobre estos parámetros desconocidos de la población determinará la técnica que deberá emplearse en cada caso. Para ilustrar esta situación se proponen los siguientes ejemplos:

Ejemplo 2-1

Se pretende estimar el tiempo medio de estancia en el hospital de los pacientes afectados por una determinada patología. Se cuenta con información correspondiente a 450 pacientes, procedentes de varios hospitales de la geografía española, a partir de los cuales se obtuvo un tiempo medio de estancia de 8 días con una desviación típica de 1,5 días.

Ejemplo 2-2

En un estudio sobre litiasis biliar realizado en un municipio se obtuvo información sobre 500 individuos. El nivel promedio de colesterol observado fue de 226,58 mg/100 ml, con una desviación típica de 49,39 mg/100 ml. A partir de estos datos se desea comprobar si el nivel promedio de colesterol de los individuos del municipio es de 210 mg/100 ml.

En el primero de los ejemplos, el objetivo del estudio es *estimar* o *cuantificar* el valor de un determinado parámetro de la población (en este caso, el de la media de tiempo de estancia en el hospital), mientras que en el segundo, lo que se pretende es *contrastar* si el nivel promedio de colesterol en la población es igual a 210 mg/100 ml y, en ambos casos, a partir de la información contenida en una muestra de la población. Estas dos situaciones constituyen el punto de partida para la aplicación de dos técnicas inferenciales (fig. 2-1): la *estimación* y los *contrastes de hipótesis.*

ESTIMACIÓN POR INTERVALOS

Cuando lo que se pretende es estimar o cuantificar el valor de un parámetro desconocido de la población podría procederse: 1) proporcionando un único valor para el parámetro, o 2) proporcionando un intervalo que contendrá al verdadero valor del parámetro de la población con una determinada seguridad o confianza. La diferencia entre estas dos aproximaciones radicaría en el hecho de que, mientras que en el primero de los casos (estimación *puntual*) no se proporciona ningún tipo de información sobre la magnitud probable del parámetro objeto de estudio ni del error que pudiera cometerse (es evidente que si se utiliza la información de una pequeña parte

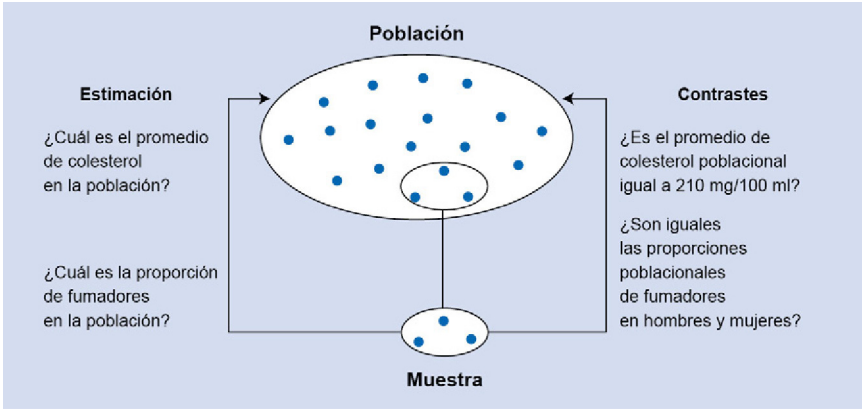

FIGURA 2-1 Inferencia estadística. Herramientas inferenciales.

de la población para estimar el valor del parámetro correspondiente a toda la población, el valor de la estimación será aproximado y, por tanto, sujeto a error), mediante la aproximación por intervalos (estimación *por intervalos* o *intervalos de confianza*) sí se responde a estas cuestiones.

A partir de los datos del ejemplo 2-1, podría utilizarse la media del tiempo de estancia en el hospital de los 450 pacientes estudiados $(\bar{x})$ para estimar la media de estancia de todos los individuos afectados por esa patología (μ). El estimador puntual de la media poblacional sería 8 días de estancia.

Por otra parte podría construirse un intervalo al 95% de confianza para la media del tiempo de estancia de la forma:

$$I_{0,95}(\mu)=\left[\bar{x}-1,96\frac{S^*}{\sqrt{n}};\bar{x}+1,96\frac{S^*}{\sqrt{n}}\right]=[7,86\ ;8,14]$$

Donde:

$$S^*=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$$

En este caso se concluiría que el tiempo medio de estancia en un hospital de los pacientes afectados por esa patología se situaría entre los 7,86 y los 8,14 días con una confianza o seguridad del 95%, es decir, con una posibilidad de error del 5%. Si bien la información proporcionada por el intervalo de confianza es evidentemente mayor, es importante destacar que para su construcción ha sido necesario contar con la media muestral como estimador puntual del parámetro (media de la población).

En general, en el proceso de construcción de intervalos de confianza para un parámetro poblacional será necesario disponer del estimador puntual de dicho parámetro. La pregunta es: ¿cuáles son los estimadores puntuales

de los diversos parámetros poblacionales de interés (medias, proporciones, varianzas, diferencias de medias…) y qué características deben reunir? Este problema, así como el método de construcción de los intervalos de confianza, serán discutidos en el apartado «Estimación».

CONTRASTES DE HIPÓTESIS

Cuando el objetivo del estudio es contrastar alguna hipótesis sobre uno o varios parámetros desconocidos de la población (¿existirán diferencias en los niveles medios de ácido úrico entre hombres y mujeres? ¿Fuman más los hombres que las mujeres? ¿La proporción de mejoría es igual para los dos tratamientos?) a partir de la información contenida en una muestra, se recurrirá a los contrastes de hipótesis. El funcionamiento de esta segunda técnica inferencial se basa en la realización de una afirmación sobre el parámetro o parámetros poblacionales (hipótesis) y en el estudio de la compatibilidad entre dicha afirmación y la información que proporciona la muestra observada.

En principio, parece lógico considerar que cuanta mayor discrepancia exista entre la hipótesis realizada y lo observado en la muestra mayor será la evidencia en contra de la hipótesis mencionada, pero ¿cómo medir esa discrepancia?

En el ejemplo 2-2 se pretendía contrastar si el nivel promedio de colesterol en el municipio (μ) era igual a 210 mg/100 ml (μ_0) a partir de la información sobre 500 individuos de dicha población. Si el nivel promedio de colesterol en los 500 individuos observados fue de $\bar{x} = 226{,}58$ mg/100 ml, parece razonable utilizar la diferencia entre el promedio de nivel de colesterol observado $(\bar{x})$ y el nivel promedio de colesterol hipotético (μ_0) como medida de discrepancia. Así, se tendrá que:

$$\text{discrepancia} = \bar{x} - \mu_0 = 226{,}58 - 210 = 16{,}58$$

Luego la discrepancia entre el nivel promedio de colesterol propuesto y el obtenido a partir de los datos de la muestra es de 16,58 mg/100 ml. Debe hacerse notar que, al igual que en el caso de los intervalos de confianza, ha sido necesario contar con la media muestral como estimador puntual de la media poblacional. La cuestión es ¿esta discrepancia es lo suficientemente grande como para rechazar la hipótesis planteada? Este problema y el concerniente a la realización de contrastes serán discutidos en el apartado «Contrastes de hipótesis».

MUESTREO E INFERENCIA ESTADÍSTICA

Los intervalos de confianza y los contrastes de hipótesis son técnicas inferenciales que permiten realizar afirmaciones sobre parámetros desconocidos de la población a partir de la información contenida en una

muestra. Si solo se dispone de esta información, que constituye una parte de toda la población objeto de estudio, cualquier afirmación que se realice sobre los parámetros poblacionales estará sujeta a un error inherente al propio proceso de muestreo que recibe el nombre de *error muestral,* también denominado *error aleatorio,* cuando la muestra ha sido obtenida mediante algún procedimiento de muestreo aleatorio.

Por otra parte, como ha podido comprobarse en ejemplos anteriores, en el desarrollo de ambas técnicas ha sido indispensable contar con estimadores puntuales de los parámetros poblacionales. Si estos estimadores se construyen a partir de los datos de la muestra, la forma en que hayan sido seleccionados los individuos o elementos de la misma influirá enormemente sobre los resultados que puedan obtenerse y, en este caso, sobre el valor del estimador puntual.

Lo deseable sería que la muestra seleccionada sea *representativa* de la población sobre la que se pretende realizar algún tipo de inferencia *(población objetivo)* pues, de lo contrario, las estimaciones se alejarán de los verdaderos valores de los parámetros poblacionales, produciéndose un error denominado error *sistemático* o *sesgo* (fig. 2-2).

Será necesario disponer de métodos o técnicas de muestreo en el proceso de selección de los individuos o elementos que contribuyan (aunque no lo garanticen) a conseguir muestras representativas de la población objetivo (reduciendo o eliminando el error de tipo sistemático) y, por otra parte, a cuantificar el error muestral o aleatorio asociado a las estimaciones que se obtengan a partir de ellas.

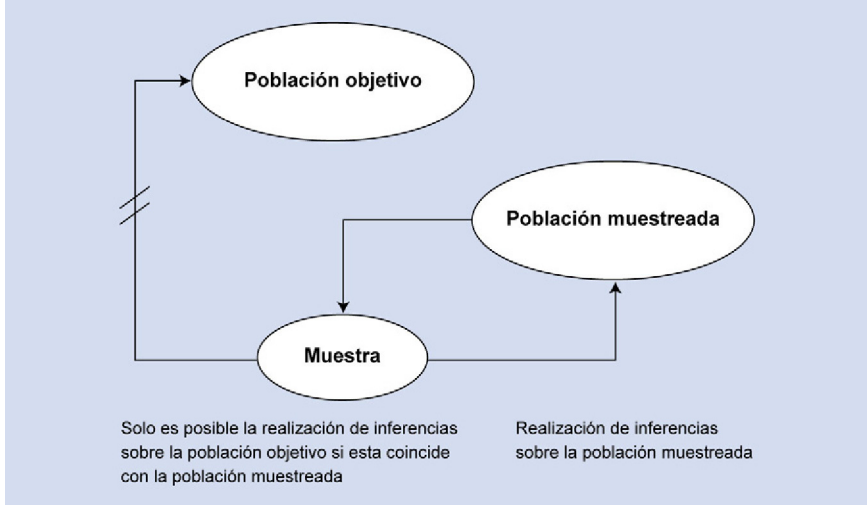

FIGURA 2-2 Población objetivo. Error sistemático o sesgo.

MUESTREO ALEATORIO O PROBABILÍSTICO

Las distintas técnicas de muestreo se clasifican en *probabilísticas* o *aleatorias* y *no probabilísticas.* La diferencia estriba fundamentalmente en que, en las primeras, cada uno de los individuos o elementos de la población tiene una probabilidad conocida y distinta de 0 de ser incluido en la muestra, mientras que en las últimas esta cuestión se desconoce. La principal consecuencia del conocimiento o desconocimiento de la probabilidad de inclusión de un individuo o elemento en la muestra es que las técnicas probabilísticas serán las únicas que harán posible tanto la cuantificación del error muestral como la aplicación de las técnicas inferenciales (construcción de intervalos de confianza y contrastes de hipótesis).

MUESTREO ALEATORIO SIMPLE

Supóngase que una población está formada por *N* individuos o elementos, el muestreo aleatorio simple selecciona una muestra de tamaño *n,* de forma que todas las unidades poblacionales tienen la misma probabilidad de ser incluidas en la muestra:

$$n/N$$

Dependiendo de si un mismo elemento puede ser seleccionado en más de una ocasión o no, el muestreo se denomina *con reemplazamiento o sin reemplazamiento* (la elección del muestreo aleatorio con o sin reemplazamiento cobrará especial importancia en el caso de poblaciones finitas, especialmente cuando $n/N > 0{,}05$, aspecto que será discutido en apartados posteriores). Como consecuencia, cada una de las muestras distintas que podrían obtenerse tendrá una probabilidad de ser seleccionada de:

$$1/N^n \quad \text{o} \quad 1/\binom{N}{n}$$

En este proceso suele ser habitual utilizar una tabla de *números aleatorios* o algunos programas de análisis estadístico, si bien, en este último caso, debe tenerse en cuenta que los números aleatorios obtenidos no son aleatorios puros sino pseudoaleatorios, ya que los programas parten de una «semilla» inicial a partir de la cual se genera el resto de números.

Las tablas de números aleatorios (tabla 2-1) están compuestas por un conjunto de números enteros (de 0 hasta 9) generados de forma aleatoria, de modo que la tabla contendrá a todos estos enteros en proporciones aproximadamente iguales y sin tendencias en el patrón con que fueron generados los datos. Ello implica que el lugar de la tabla a partir del cual comience la búsqueda (desde arriba, abajo, izquierda, derecha, centro, etc.) es indiferente, aunque sí es necesario adoptar algún tipo de criterio

TABLA 2-1 Tabla de números aleatorios (reducida)

018095	451846	894272	042572	000895	512267	899462	416203
856719	175549	664998	708414	592502	903636	567050	204870
117559	678983	966177	743175	264227	516232	611864	225475
147361	751548	405976	209540	917949	245822	437322	973331
426815	577815	559584	155811	406849	797925	424843	373144
914382	895487	045268	760994	337886	680083	104835	040504
289462	673288	577309	053879	743535	340608	390705	553281

(de izquierda a derecha y de arriba hacia abajo, de derecha a izquierda y de abajo hasta arriba, etc.).

El procedimiento de selección de una muestra por muestreo aleatorio simple sería el siguiente:

- Disponer de un listado enumerado de todos los individuos o elementos de la población $\{e_1, e_2, e_3 \ldots, e_N\}$.
- Seleccionar n números aleatorios entre 1 y N utilizando una tabla de números aleatorios.
- Seleccionar los individuos o elementos de la población correspondientes a los números aleatorios seleccionados.

Este procedimiento sería equivalente a introducir en un sombrero N papeletas numeradas, cada una de ellas correspondiente a cada uno de los individuos o elementos de la población, y extraer n de ellas una vez hayan sido mezcladas convenientemente.

Ejemplo 2-3

A partir de una población formada por $N = 250$ individuos se desea obtener por muestreo aleatorio simple sin reemplazamiento una muestra de tamaño 10.

- En primer lugar se asignaría un número a cada uno de los 250 individuos o elementos de la población:

$$e_1, e_2, e_3 \ldots, e_{250}$$

- En segundo lugar se seleccionarían en la tabla de números aleatorios, por ejemplo, de izquierda a derecha y de arriba abajo, números de tres dígitos (tener en cuenta que hay 250 individuos en la población y se necesitarán tres dígitos para enumerarlos) hasta completar un total de 10 números distintos entre 1 y 250. En la tabla 2-1, los números de tres dígitos que se irían encontrando dentro de los límites establecidos serían los siguientes (los 10 números aleatorios seleccionados serían los que aparecen en negrita, puesto que son los primeros 10 números de tres dígitos distintos que se encuentran entre 1 y 250):

018, 095, 451, 846, 894, 272, **042,** 572, 000, 895, 512, 267, 899, 462, 416, **203,** 856, 719, **175,** 549, 664, 998, 708, 414, 592, 502, 903, 636, 567, **050, 204,** 870, **117,** 559, 678, 983, 966, **177,** 743, **175,** 264, **227**

Obsérvese que el número 175, señalado con un subrayado, no se tendría en cuenta en este caso al haber sido seleccionado previamente.

- Por último se seleccionarían los siguientes individuos de la población:

$$e_{18}, e_{95}, e_{42}, e_{203}, e_{175}, e_{50}, e_{204}, e_{117}, e_{177}, e_{227}$$

En el caso en que el muestreo se hubiera realizado con reemplazamiento, el número 175 habría sido seleccionado en dos ocasiones en detrimento del último valor 227.

Existen multitud de técnicas de muestreo aleatorio alternativas al muestreo aleatorio simple como el *muestreo aleatorio sistemático, estratificado, por conglomerados, polietápico*, etc., que no son objeto de estudio en este libro.

ESTIMACIÓN

Con frecuencia el interés del investigador se centra en estimar o cuantificar el valor de un parámetro desconocido de la población a partir de la información contenida en una muestra.

Ejemplo 2-4

Una nueva técnica quirúrgica fue practicada con éxito en 40 de los 50 pacientes intervenidos y seleccionados al azar entre los afectados por una determinada patología. ¿Cuál sería la proporción de éxito en la intervención en la población de afectados por dicha patología?

Ejemplo 2-5

En un estudio sobre la práctica de ejercicio físico se obtuvo información sobre 350 personas, de las cuales 161 manifestaron realizar ejercicio físico de forma regular. El promedio de edad entre las que declararon realizar ejercicio físico fue de 24,5 años, con una desviación típica de 3,6 años. ¿Cuál es el promedio de edad de las personas que realizan ejercicio de forma regular en la población de la que partió la muestra?

En el proceso de estimación de la proporción de éxito en la intervención y el promedio de edad de los individuos que realizan ejercicio de forma regular será muy importante, como fue discutido en apartados anteriores, la forma en que hayan sido seleccionados los individuos de la muestra observada. Pero, suponiendo que los individuos hayan sido seleccionados de forma aleatoria, ¿cómo puede estimarse el valor del parámetro poblacional?

ESTIMACIÓN PUNTUAL

La primera etapa en el proceso de estimación de un parámetro desconocido de la población consiste en obtener, a partir de los datos de la muestra, un valor que será utilizado como estimación de dicho parámetro. Este valor, denominado *estimador puntual*, ayudaría a situar el valor del parámetro aunque sin especificar la magnitud probable del mismo. En general, es habitual utilizar letras del alfabeto griego para referirse a los parámetros poblacionales, y el acento circunflejo para referirse al estimador puntual del parámetro correspondiente que se calcula a partir de los datos de la muestra. Así se tendría que, por ejemplo, para referirse a una media poblacional se utilizaría la letra μ y para su estimador puntual $\hat{\mu}$. En general, para un parámetro cualquiera θ el estimador puntual se denominará $\hat{\theta}$.

En ejemplo 2-4 se pretende estimar el valor de la proporción poblacional de éxito de una determinada intervención quirúrgica a partir de la información contenida en una muestra de 50 individuos. La proporción de éxito tras la intervención observada se calcularía de la siguiente forma:

$$\hat{p} = \frac{r}{n} = \frac{40}{50} = 0,8$$

Donde *r* es el número de individuos de la muestra en los que la intervención ha sido un éxito y *n* el tamaño de la muestra. Esto significa que en el 80% de los pacientes la intervención ha sido un éxito. ¿Sería factible utilizar la proporción muestral como estimador puntual de la proporción poblacional de éxito en la intervención?

En el ejemplo 2-5 se pretendía estimar el valor de la media de edad de los individuos que practican ejercicio de forma regular. Se cuenta con información sobre 161 individuos que practican algún tipo de ejercicio en los que la media de edad es:

$$\hat{\mu} = \overline{x} = 24,5$$

Al igual que en el caso de la proporción, la cuestión es: ¿podría utilizarse la media muestral como estimador puntual de la media de edad poblacional? Como puede observarse, tanto la proporción como la media muestral son características cuantitativas calculables a partir de los datos de la muestra, pero ¿puede asegurarse que se aproximarán a los verdaderos valores de los parámetros poblacionales? En primer lugar será necesario definir los conceptos de *estadístico* y *estimador.*

Estadístico: un estadístico es cualquier función de los datos de la muestra o, equivalentemente, cualquier característica cuantitativa calculada a partir de los datos de la muestra.
Estimador: un estimador es un estadístico (luego es calculable a partir de los datos de la muestra) que, por su construcción, intenta acercarse al verdadero valor de un parámetro desconocido de la población.

Tanto la proporción como la media muestral son *estadísticos*, ya que son calculables a partir de los datos de la muestra. Faltaría comprobar si son *estimadores*, es decir, si se acercan a los valores de los parámetros poblacionales.

ESTADÍSTICOS EN EL MUESTREO

La muestra aleatoria observada constituye la base para la realización de inferencias sobre un determinado parámetro desconocido de la población θ. El estimador puntual $\hat{\theta}$, tal y como se ha definido, debería calcularse a partir de ella y de forma que, por su construcción, se acercara al verdadero valor del parámetro que pretende estimar (fig. 2-3).

En un muestreo aleatorio simple, por ejemplo con reemplazamiento, el número de muestras de tamaño n que podrían obtenerse a partir de una población de tamaño N es N^n. Sin embargo, en el proceso de muestreo previo a la estimación del parámetro desconocido de la población, se obtendría tan solo una de estas muestras como base para la realización de inferencias.

La pregunta que surge de forma inmediata es: ¿se mantendrá constante el valor del estimador independientemente de la muestra seleccionada o, por el contrario, se producirán variaciones en función de la muestra escogida? En el caso de que se produzcan variaciones, ¿tendrán todos los valores del estimador calculables a partir de las diferentes muestras la misma posibilidad de ser observados una vez finalizado el proceso de muestreo, o existirán unos valores más probables que otros?

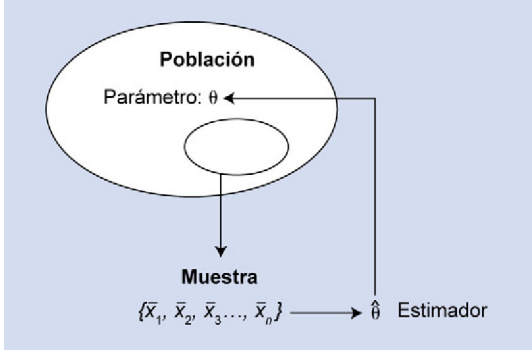

FIGURA 2-3 Población, muestra y estimador.

Para profundizar en esta cuestión se plantea un ejemplo en el que se parte de una situación ideal relativa a una población muy pequeña donde es posible conocer de antemano el valor de un determinado parámetro poblacional (en este caso la media relativa a toda la población). De este modo será posible analizar todas las muestras posibles y los valores de las estimaciones puntuales del parámetro poblacional en cada una de ellas.

Ejemplo 2-6

Supóngase, sin pérdida de generalidad y a efectos de simplicidad en los cálculos, una población formada por seis individuos sobre la que se pretende estimar el consumo medio de cigarrillos diarios. Supóngase, además, que ha sido posible obtener las seis observaciones correspondientes a los seis individuos de la población y que son:

$$\{5, 10, 15, 20, 25, 30\}$$

El conocimiento de las observaciones de la variable (número de cigarrillos consumidos diariamente) correspondientes a la totalidad de los individuos de la población permitiría calcular la media y la desviación típica poblacionales de la mencionada variable de la siguiente forma:

$$\mu = \frac{\sum x_i}{N} = \frac{5+10+15+20+25+30}{6} = 17,5$$

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{(5-17,5)^2 + (10-17,5)^2 + \ldots + (30-17,5)^2}{6} = 72,92$$

Estos valores de media y varianza serían los parámetros de la población, habitualmente desconocidos, ya que no se dispondría de las observaciones de la variable en todos los individuos o elementos que la componen. Si se hubiera efectuado un muestreo aleatorio simple con reemplazamiento para obtener una muestra de tamaño $n = 2$, el número de muestras distintas que podrían obtenerse es $N^n = 6^2 = 36$, tal y como se refleja en la tabla 2-2.

TABLA 2-2 Valores de la media muestral que podrían obtenerse por muestreo con reemplazamiento a partir de todas las muestras posibles de tamaño $n = 2$

	5	**10**	**15**	**20**	**25**	**30**
5	5 (5;5)	7,5 (5;1)	10 (5;15)	12,5 (5;20)	15 (5;25)	17,5 (5;30)
10	7,5 (10;5)	10 (10;10)	12,5 (10;15)	15 (10;20)	17,5 (10;25)	20 (10;30)
15	10 (15;5)	12,5 (15;10)	15 (15;15)	17,5 (15;20)	20 (15;25)	22,5 (15;30)
20	12,5 (20;5)	15 (20;10)	17,5 (20;15)	20 (20;20)	22,5 (20;25)	25 (20;30)
25	15 (25;5)	17,5 (25;10)	20 (25;15)	22,5 (25;20)	25 (25;25)	27,5 (25;30)
30	17,5 (30;5)	20 (30;10)	22,5 (30;15)	25 (30;20)	27,5 (30;25)	30 (30;30)

Como puede observarse, el valor de la media muestral calculado a partir de los datos de la muestra depende de la que haya sido seleccionada en el proceso de muestreo. Además, no todos los valores obtenidos para la media muestral se repiten con la misma frecuencia. Así, se tiene que en solo en una de las muestras, la correspondiente a los valores {5,5}, se obtendría un promedio de cinco cigarrillos diarios. En distinta situación se encuentra el valor 10, que podría obtenerse a partir de tres muestras distintas {5, 15}, {10, 10}, {15, 5}. Se observa sin dificultad que el valor de la media muestral que tendría más posibilidades de ser observado (el que se repite con mayor frecuencia) es 17,5, que aparece en 6 de las 36 muestras posibles.

Puede concluirse, por tanto, que la media muestral es una variable aleatoria, ya que varía en función de la muestra aleatoria seleccionada. Dado que la media muestral es una variable y que se dispone de los 36 valores de la media muestral correspondientes a cada una de las muestras, puede calcularse la media o esperanza de todos estos valores (media de todas las medias muestrales posibles) y la varianza. Así, se tendrá que:

$$E\left(\overline{X}\right)=\frac{\sum \overline{x}_i}{36}=\frac{5+7,5+10+12,5+\ldots+30}{36}=17,5$$

$$Var\left(\overline{X}\right)=\frac{\sum\left(\overline{x}_i-E\left(\overline{X}\right)\right)^2}{36}=\frac{(5-17,5)^2+(7,5-17,5)^2+\ldots+(30-17,5)^2}{36}=36,46$$

Esto quiere decir que la media de todos los valores de la media muestral que podrían ser calculados a partir de las diferentes muestras aleatorias del mismo tamaño de la población coincide con el promedio poblacional de consumo de cigarrillos (17,5). Además, la varianza asociada a la media muestral es significativamente inferior a la de la variable original.

$$E\left(\overline{X}\right)=E(X)=17,5$$

$$Var\left(\overline{X}\right)=\frac{Var(X)}{n}=\frac{72,92}{2}=36,46$$

En conclusión, puede afirmarse que los valores del estimador $\hat{\theta}$ (en este caso $\overline{x}$) varían en función de la muestra aleatoria seleccionada y, por tanto, el estimador es una variable aleatoria. En segundo lugar, y no por ello menos importante, no todos los valores posibles del estimador tienen la misma probabilidad de ser observados. De hecho, puede demostrarse que en el caso concreto del estimador media muestral $\overline{x}$ y a partir de un muestreo aleatorio con reemplazamiento, la media y la varianza del estimador verificarán:

$$E\left(\overline{X}\right)=E(X)=\mu$$

$$Var\left(\overline{X}\right)=\frac{Var(X)}{n}=\frac{\sigma^2}{n}$$

PROPIEDADES DESEABLES PARA UN ESTIMADOR PUNTUAL

En general, parece deseable que los estimadores de los parámetros poblacionales verifiquen que la esperanza o media del estimador coincida con el parámetro que se pretende estimar *(insesgadez)* y que la varianza del estimador sea lo más pequeña posible *(eficiencia).* El interés en que la varianza del estimador sea lo más pequeña posible reside en que cuanto menor sea la variabilidad del estimador, menor será el error que se cometerá en la estimación del parámetro poblacional. Para comprender esto basta tener en cuenta que, cuanto menor es la varianza de una variable, más concentrados están los datos en torno a su media. Por otra parte, si la esperanza del estimador coincide con el parámetro poblacional, más cerca del parámetro se encuentran todos los posibles valores del estimador independientemente de la muestra que haya sido seleccionada en el proceso de muestreo.

En la tabla 2-3 se presentan algunos de los parámetros poblacionales más habituales y sus correspondientes estimadores puntuales.

Como puede observarse, los estimadores puntuales propuestos para estos parámetros parecen responder a un criterio de selección puramente analógico (la media muestral como estimador de la media poblacional, la proporción muestral como estimador de la proporción poblacional, la diferencia de medias muestrales como estimador de la diferencia de medias poblacionales, etc.) cuando, en realidad, el proceso es mucho más complejo y tiene en cuenta las propiedades mencionadas con anterioridad. Obsérvese que en el caso de la estimación de una varianza poblacional se efectúa una ligera corrección sobre la varianza muestral. Esto se debe a que la varianza muestral S^2 no es un estimador insesgado para la varianza poblacional. Por este motivo, muchos autores prefieren definir directamente la varianza muestral de la siguiente forma:

$$S^{*2} = \frac{n}{n-1} S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

TABLA 2-3 Estimadores puntuales para parámetros habituales

Parámetro		Estimador puntual	
Media	μ	Media muestral	$\hat{\mu} = \bar{x} = \frac{\sum x_i}{n}$
Proporción	p	Proporción muestral	$\hat{p} = \frac{r}{n}$
Diferencia de medias	$\mu_1 - \mu_2$	Diferencia de medias muestrales	$\widehat{\mu_1 - \mu_2} = \hat{\mu}_1 - \hat{\mu}_1 = \bar{x}_1 - \bar{x}_2$
Diferencia de proporciones	$p_1 - p_2$	Diferencia de proporciones muestrales	$\widehat{p_1 - p_2} = \hat{p}_1 - \hat{p}_2$
Varianza	σ^2	Varianza muestral corregida	$\hat{\sigma}^2 = S^{*2} = \frac{n}{n-1} S^2$

DISTRIBUCIONES MUESTRALES

Una de las conclusiones más importantes a las que se llegaba en el apartado «Estadísticos en el muestreo» hacía referencia al hecho de que un estimador es una variable aleatoria, ya que su valor depende de la muestra aleatoria que haya sido seleccionada previamente en el proceso de muestreo. Si el estimador es una variable aleatoria tiene sentido plantearse, en primer lugar, cuál es su media o esperanza, su varianza y, lo que es todavía más importante, cuál es la distribución de probabilidad que gobierna su comportamiento.

DISTRIBUCIÓN DE LA MEDIA MUESTRAL EN POBLACIONES NORMALES

En el caso de que la variable objeto de estudio X siga un modelo de distribución normal con media μ y varianza σ^2, la variable media muestral $\bar{X}$ se distribuirá según un modelo de distribución normal con la misma media μ y varianza σ^2/n. Este resultado constituye la base de las denominadas *pruebas paramétricas* que permiten la realización de inferencias sobre los parámetros correspondientes. Sin embargo, cabe preguntarse qué ocurre cuando se desconoce la distribución poblacional de la variable X o si los datos obtenidos en la muestra no permiten establecer si la distribución normal es adecuada o evidencian una distribución distinta a la normal.

DISTRIBUCIÓN ASINTÓTICA DE LA MEDIA MUESTRAL

Anteriormente ha sido comprobado que, en el caso de que los datos de la muestra se hubieran obtenido por muestreo aleatorio simple con reemplazamiento, la media y la varianza de la media muestral verificaban:

$$E\left(\bar{X}\right) = E(X) = \mu$$

$$\mathrm{Var}\left(\bar{X}\right) = \frac{\mathrm{Var}(X)}{n} = \frac{\sigma^2}{n}$$

Puede comprobarse que, en el caso de que los datos hayan sido seleccionados por muestreo aleatorio simple sin reemplazamiento, la esperanza de la media muestral sigue coincidiendo con el parámetro desconocido de la población μ mientras que la varianza sufre una ligera variación:

$$E\left(\bar{X}\right) = E(X) = \mu$$

$$\mathrm{Var}\left(\bar{X}\right) = \frac{\mathrm{Var}(X)}{n} = \frac{\sigma^2}{n}\frac{N-n}{N-1}$$

Sin embargo, cuando el tamaño de la población N es grande, la cantidad $(N - n)/(N - 1)$ se acerca a 1 y, por tanto, la distinción entre el muestreo

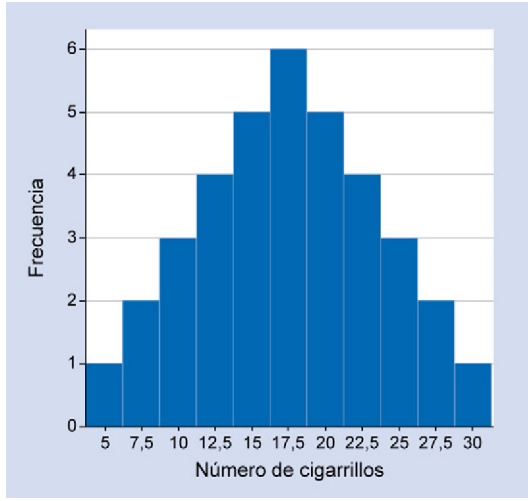

FIGURA 2-4 Gráfico de distribución de frecuencias de la media muestral.

con o sin reemplazamiento es absolutamente innecesaria. En la práctica suele despreciarse esta corrección cuando $n/N \leq 0{,}05$.

Se denomina *error estándar* a la desviación típica de un estadístico en el muestreo. Por tanto, el error estándar de un estimador es una medida de la variabilidad del mismo en el proceso de muestreo. Conviene recordar que una de las propiedades deseables para los estimadores puntuales, la eficiencia, requería que la varianza del estimador (y, en consecuencia, su error estándar asociado) fuera la mínima posible.

En la figura 2-4 se representa el histograma correspondiente a la media muestral construido a partir de la tabla de distribución de frecuencias que se obtendría a partir de los datos de la tabla 2-2.

Se observa que la distribución de la media muestral es simétrica respecto del valor 17,5 que coincide además con la esperanza o media de la variable E(X). La mayor frecuencia se concentra alrededor de este valor central descendiendo paulatinamente hacia los extremos de la distribución y bien podría recordar el modelo de distribución de probabilidad normal, aunque evidentemente de forma aproximada. De hecho, puede comprobarse que, a medida que se aumenta el tamaño de la muestra, la distribución de la media muestral $\overline{X}$ se aproxima de forma asintótica (cada vez más cuanto mayor sea el tamaño de la muestra n) a la distribución normal. Este resultado se conoce como el teorema central del límite, que se expone a continuación.

Teorema central del límite

Sea X una variable aleatoria cualquiera con media μ y varianza σ^2; entonces, si el tamaño muestral es lo suficientemente grande (habitualmente

$n \geq 30$), la media muestral $\bar{X}$ se distribuirá asintóticamente según un modelo de probabilidad normal con la misma media que la variable original μ y con varianza σ^2/n. Este resultado puede expresarse de las siguientes formas equivalentes:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad ; \quad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

INTERVALO DE CONFIANZA PARA UNA MEDIA POBLACIONAL CON VARIANZA POBLACIONAL CONOCIDA

El conocimiento de la distribución de probabilidad que gobierna el comportamiento de una variable aleatoria permite calcular cualquier tipo de probabilidad sobre la misma. Así, si X es una variable que sigue un modelo de distribución normal estándar (de media 0 y desviación típica 1), puede concluirse sin dificultad que:

$$P(-1,96 < X < 1,96) = 0,95$$

Puede afirmarse entonces que el intervalo [−1,96; 1,96] contiene al 95% de todos los posibles valores de la variable aleatoria X. Si la media muestral se comporta según un modelo de distribución de probabilidad normal con media μ y varianza σ^2/n (resultado discutido anteriormente), entonces:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1); \; P\left(-1,96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1,96\right) = 0,95$$

Despejando de la desigualdad el valor de la media muestral $\bar{X}$ se tendría que el siguiente intervalo contendría al 95% de los posibles valores de la media muestral $\bar{X}$.

$$\left[\mu - 1,96\frac{\sigma}{\sqrt{n}}; \mu + 1,96\frac{\sigma}{\sqrt{n}}\right]$$

Por tanto se verificará que, con una probabilidad de 0,95:

$$\mu - 1,96\frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 1,96\frac{\sigma}{\sqrt{n}}$$

Si se le da la vuelta a este resultado y se despeja μ en esta doble desigualdad se obtendrá la siguiente expresión:

$$\bar{X} - 1,96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1,96\frac{\sigma}{\sqrt{n}}$$

Luego, si se construye un intervalo para la media poblacional μ tomando como límite inferior y límite superior:

$$\lim_{inf} = \bar{X} - 1,96\frac{\sigma}{\sqrt{n}}; \lim_{sup} = \bar{X} + 1,96\frac{\sigma}{\sqrt{n}}$$

este contendrá a μ en el 95% de todos los posibles intervalos que podrían construirse a partir de cada una de las muestras aleatorias que podrían obtenerse en el proceso de muestreo. Este proceso se refleja en la figura 2-5.

Como puede observarse, dependiendo de la muestra seleccionada en el proceso de muestreo se obtendrá un intervalo de confianza distinto para la media poblacional μ. De los seis intervalos de confianza que a modo de ejemplo han sido representados en la figura 2-5, solo uno no contiene al valor del parámetro, hecho que se debe exclusivamente a que la media muestral calculada a partir de los datos de la muestra se sitúa en la zona sombreada y que corresponde a los valores más extremos de la distribución de $\bar{X}$.

Se sabe, como fue demostrado con anterioridad, que la probabilidad de que, en este caso, la media muestral tome un valor comprendido en esa zona es 0,05. Por tanto, puede afirmarse que el 95% de los intervalos que pudieran construirse a partir de las diferentes muestras contendrán al verdadero valor del parámetro. Este valor 0,05 que cuantifica el error se suele denominar α. Es importante señalar que se habla de confianza de que el intervalo contenga al parámetro y no de probabilidad de que el parámetro esté en el intervalo debido a que el valor del parámetro μ es fijo y no una variable. Si no es una variable no tendría sentido hablar de probabilidades sobre algo que es constante, si bien la interpretación es similar.

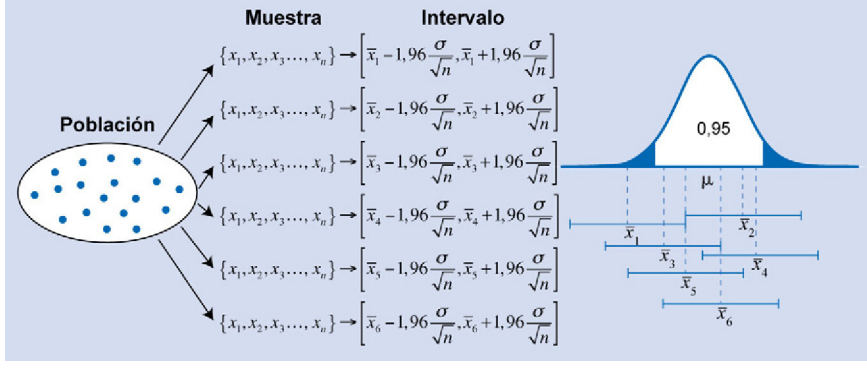

FIGURA 2-5 Proceso de muestreo y construcción de intervalos de confianza para la media poblacional, supuesta la desviación típica poblacional conocida.

Ejemplo 2-7

En un estudio se obtuvo información sobre la presión sistólica de un grupo de 150 pacientes a los que se les administró un determinado fármaco. Si la presión sistólica media fue de 120 mmHg y suponiendo que la desviación típica poblacional de la presión sistólica en este tipo de pacientes es $\sigma = 12$, ¿cuál será el promedio de presión sistólica media en la población de pacientes que utiliza este tipo de fármaco?

Dado que el tamaño muestral $n = 150$ es superior a 30, puede afirmarse, por el teorema central del límite, que la distribución de probabilidad de la media muestral será aproximadamente normal con media μ y varianza σ^2/n. Esto permitirá construir el intervalo de confianza para la media poblacional de la siguiente forma:

$$\left[\bar{X}-1,96\frac{\sigma}{\sqrt{n}};\bar{X}+1,96\frac{\sigma}{\sqrt{n}}\right]=\left[120-1,96\frac{12}{\sqrt{150}};120+1,96\frac{12}{\sqrt{150}}\right]=[118,08;121,92]$$

Luego podría concluirse que la presión sistólica media en este tipo de pacientes estará comprendida entre 118,08 y 121,92 mmHg, con una confianza del 95%.

INTERVALO DE CONFIANZA PARA UNA MEDIA POBLACIONAL CON VARIANZA DESCONOCIDA

La construcción del intervalo de confianza, para una media poblacional descrita anteriormente, se apoyaba en la suposición de que la varianza poblacional σ^2 era conocida. Sin embargo, esa situación constituye habitualmente un puro ejercicio teórico, ya que la varianza poblacional, al igual que la media, suele ser desconocida. ¿Cómo pueden aprovecharse los resultados obtenidos hasta el momento para construir el intervalo de confianza en este caso?

El conocimiento de la distribución de probabilidad del estadístico:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

fue determinante en este proceso de construcción. Si la variable X se distribuye según un modelo normal o, en la práctica, se cuenta con un número de datos superior o igual a 30 (requisito de aplicación del teorema central del límite) y se sustituye el valor de σ por el de su estimador puntual S^*, puede demostrarse que el estadístico:

$$\frac{\bar{X}-\mu}{S^*/\sqrt{n}}$$

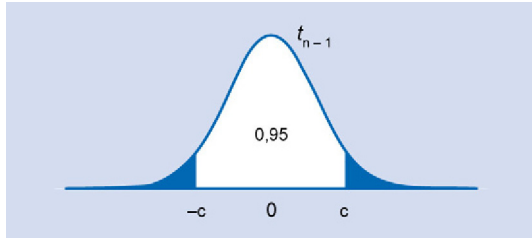

FIGURA 2-6 Percentiles 0,025 y 0,975 sobre la *t* de Student.

se distribuirá de muestra a muestra según un modelo de probabilidad *t* de Student con $n - 1$ grados de libertad. Del mismo modo en el que se procedió en el caso anterior, se obtiene que el intervalo de confianza al 95% para la media poblacional cuando la varianza poblacional es desconocida puede construirse de la siguiente forma:

$$\left[\bar{X} - c\frac{S^*}{\sqrt{n}}; \bar{X} + c\frac{S^*}{\sqrt{n}}\right]$$

donde el coeficiente *c* es el valor que deberá obtenerse en una *t* de Student con $n - 1$ grados de libertad (fig. 2-6) de forma que se verifique:

$$P\left(-c < \frac{\bar{X} - \mu}{S^*/\sqrt{n}} < c\right) = 0,95$$

Lo que significa que el valor *c* es el percentil 0,975 y, en consecuencia, deja 0,975 de probabilidad a la izquierda. Si se tiene en cuenta que el error es, en este caso, $\alpha = 0,05$, tiene sentido que el valor del coeficiente *c* suela expresarse como $c_{1-\frac{\alpha}{2}}$.

Obsérvese que en el ejemplo se tendrá que:

$$c_{1-\frac{\alpha}{2}} = c_{1-\frac{0,05}{2}} = c_{1-0,025} = c_{0,975}$$

Si se trabaja con los datos del ejemplo 2-5, en las 161 personas que manifestaron realizar ejercicio físico de forma regular, el promedio de edad fue de 24,5 años con una desviación típica de 3,6 años. Se pretendía cuantificar el promedio poblacional de edad en la población de la que partió la muestra.

Dado que el tamaño muestral es superior a 30 se verifican las condiciones del teorema central del límite por el que la media muestral se distribuirá, aproximadamente, según un modelo de distribución de probabilidad normal. Como la desviación típica poblacional es desconocida, el estadístico:

$$\frac{\bar{X}-\mu}{S^*/\sqrt{n}}$$

se comportará según un modelo de distribución de probabilidad t de Student con $n-1=161-1=160$ grados de libertad. El intervalo quedará:

$$\left[\bar{X}-c\frac{S^*}{\sqrt{n}};\bar{X}+c\frac{S^*}{\sqrt{n}}\right]=\left[24,5-c\frac{\sqrt{\frac{n}{n-1}}S}{\sqrt{n}};24,5+c\frac{\sqrt{\frac{n}{n-1}}S}{\sqrt{n}}\right]$$

$$=\left[24-c\frac{S}{\sqrt{n-1}};24+c\frac{S}{\sqrt{n-1}}\right]$$

$$=\left[24-1,9749\frac{3,6}{\sqrt{161-1}};120+1,9749\frac{3,6}{\sqrt{161-1}}\right]=[23,44;\ 24,56]$$

Luego el promedio poblacional de edad estará comprendido entre 23,44 y 24,56 años con una confianza del 95%.

CONSTRUCCIÓN DE INTERVALOS DE CONFIANZA. GENERALIZACIÓN

Si se analizan con detenimiento los intervalos de confianza para la media poblacional obtenidos en el apartado anterior podrá observarse que, tanto en el caso en que la varianza poblacional sea conocida como desconocida, los elementos que intervienen en la construcción de los intervalos de confianza son los siguientes:

- El estimador puntual (EP) de la media poblacional μ: $\bar{x}$.
- Un coeficiente que depende de la distribución muestral del estadístico y del nivel de confianza exigido: $c_{1-\frac{\alpha}{2}}$.
- El error estándar (EE): $\sigma/\sqrt{n}$ o $S^*/\sqrt{n}$ según el caso.

La forma en que se combinan estos elementos en el proceso de construcción de los intervalos, según los resultados obtenidos con anterioridad, sería la siguiente:

$$I_{1-\alpha}(\mu)=\left[EP-c_{1-\frac{\alpha}{2}}\cdot EE;EP+c_{1-\frac{\alpha}{2}}\cdot EE\right]$$

Este resultado es consecuencia de que la distribución de probabilidad asociada, la normal y la t de Student, según el caso, es simétrica. Esta forma de construir los intervalos de confianza para diferentes parámetros poblacionales es generalizable siempre que la distribución muestral asociada sea simétrica. En la tabla 2-4 se presentan algunos de los parámetros poblacionales más habituales, sus estimadores puntuales, la distribución muestral asociada y el error estándar correspondiente.

TABLA 2-4 Elementos necesarios para la construcción de intervalos de confianza

Parámetro poblacional (θ)	Estimador puntual (EP)	Distribución muestral	Error estándar (EE)
Media μ:			
σ^2 conocida	$\hat{\mu} = \bar{x}$	Normal	$\frac{\sigma}{\sqrt{n}}$
σ^2 desconocida		t_{n-1}	$\frac{S^*}{\sqrt{n}}$
Proporción p	$\hat{p} = \frac{r}{n}$	Aprox. normal	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Diferencia de medias ($\mu_1 - \mu_2$):			
$\sigma_1^2 = \sigma_2^2$	$\widehat{\mu_1 - \mu_2} = \bar{x}_1 - \bar{x}_2$	$t_{n1+n2-2}$	$\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$
$\sigma_1^2 \neq \sigma_2^2$		t_f	$\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$
Diferencia de proporciones ($p_1 - p_2$)	$\widehat{p_1 - p_2} = \hat{p}_1 - \hat{p}_2$	Aprox. normal	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Donde:

$$I_{1-\alpha}(\theta) = \left[EP - c_{1-\frac{\alpha}{2}} \cdot EE;\ EP + c_{1-\frac{\alpha}{2}} \cdot EE \right]$$

$$S_p^2 = \frac{(n_1-1)S_1^{*2} + (n_2-1)S_2^{*2}}{n_1+n_2-2};\quad f = \frac{\left(\frac{S_1^{*2}}{n_1} + \frac{S_2^{*2}}{n_2}\right)^2}{\frac{\left(\frac{S_1^{*2}}{n_1}\right)^2}{n_1} + \frac{\left(\frac{S_2^{*2}}{n_2}\right)^2}{n_2}}$$

INTERVALO DE CONFIANZA PARA UNA PROPORCIÓN

Considérese el ejemplo 2-4 de este capítulo, en el que se contaba con información sobre la proporción de éxito de una intervención quirúrgica en un grupo de pacientes afectados por una determinada patología. De los 50 pacientes intervenidos, 40 respondieron de forma satisfactoria. Para estimar la proporción de éxito de este tipo de intervención en la población de todos los afectados por esta patología será necesario construir un intervalo de confianza. Supóngase que se utiliza el nivel de confianza 0,95 o del 95%.

En primer lugar se comprueba que se verifican los requisitos necesarios para utilizar la aproximación normal (en caso contrario sería necesario utilizar la distribución binomial exacta). Así, se tendrá que:

$$n\hat{p} = 50 \cdot \frac{40}{50} = 40 \geq 5$$

$$n(1-\hat{p}) = 50 \cdot \left(1 - \frac{40}{50}\right) = 10 \geq 5$$

Dado que la distribución muestral asociada es simétrica (en este caso la distribución normal), el intervalo de confianza se construirá:

$$I_{0,95}(p) = \left[EP - c\ EE;\ EP - c\ EE\right] = \left[\hat{p} - c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}};\ \hat{p} - c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Sustituyendo los valores correspondientes se tendrá que:

$$I_{0,95}(p) = \left[\frac{40}{50} - c\sqrt{\frac{\frac{40}{50}\left(1 - \frac{40}{50}\right)}{50}};\ \frac{40}{50} + c\sqrt{\frac{\frac{40}{50}\left(1 - \frac{40}{50}\right)}{50}}\right]$$

El valor del coeficiente *c* corresponde al percentil 0,975 en la distribución normal estándar que es 1,96. El intervalo quedará:

$$\begin{aligned} I_{0,95}(p) &= \left[0,8 - 1,96\sqrt{\frac{0,8(1-0,8)}{50}};\ 0,8 + 1,96\sqrt{\frac{0,8(1-0,8)}{50}};\right] \\ &= \left[0,8 - 1,96\sqrt{\frac{0,8(1-0,8)}{50}};\ 0,8 + 1,96\sqrt{\frac{0,8(1-0,8)}{50}};\right] \\ &= \left[0,69; 0,91\right] \end{aligned}$$

Luego, la proporción de éxito en la intervención en la población de afectados por la mencionada patología estará comprendida entre 0,69 y 0,91 con una confianza de 0,95 o del 95%. El error que se puede cometer en esta estimación es $\alpha = 0,05$ o del 5%.

PRECISIÓN DE UN INTERVALO DE CONFIANZA

Se define como precisión de un intervalo de confianza y se denominará τ a la mayor de las distancias entre el estimador puntual y cada uno de los dos límites del intervalo (fig. 2-7). Así, se tendrá que:

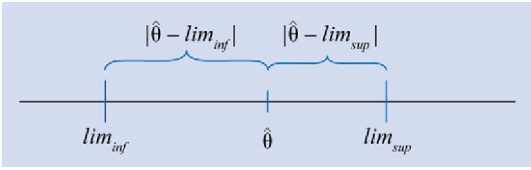

FIGURA 2-7 Precisión de un intervalo de confianza.

$$\tau = \text{máx}\left(\left|\hat{\theta} - \lim_{\text{inf}}\right|, \left|\hat{\theta} - \lim_{\text{sup}}\right|\right)$$

Esto quiere decir que cuanto mayor sea el valor de τ menos preciso será el intervalo de confianza. En el caso de que la distribución de probabilidad asociada sea simétrica, las dos distancias anteriores serán coincidentes, ya que el estimador puntual se sitúa en el centro del intervalo.

NIVEL DE CONFIANZA Y PRECISIÓN

Para un tamaño muestral fijo, el nivel de confianza se relaciona de forma inversamente proporcional a la precisión del intervalo. Esto quiere decir que, al aumentar el nivel de confianza disminuye la precisión del intervalo (aumenta, en consecuencia, el valor de τ). La situación puede ilustrarse fácilmente si se tiene en cuenta que, por ejemplo, en el caso de la media de una población la precisión puede expresarse:

$$\tau = \text{máx}\left(\left|\bar{x} - \left(\bar{x} - c_{1-\alpha/2}\frac{S^*}{\sqrt{n}}\right)\right|, \left|\bar{x} - \left(\bar{x} + c_{1-\alpha/2}\frac{S^*}{\sqrt{n}}\right)\right|\right) = c_{1-\alpha/2}\frac{S^*}{\sqrt{n}}$$

donde el coeficiente $c_{1-\alpha/2}$ crece a medida que lo hace la confianza exigida. Esta expresión pone de manifiesto, de forma adicional que, para un nivel de confianza fijo, el valor de τ y el tamaño muestral se relacionan de forma inversamente proporcional, con lo que, a mayor tamaño muestral más preciso será el intervalo de confianza (el valor de τ será pequeño a medida que aumenta n).

Considérense los datos del ejemplo 2-4, en el que ahora se pretende trabajar al nivel de confianza del 0,99 o del 99%. Esto supondrá que el error que puede cometerse con esta estimación no será superior al 1%.

Todos los resultados obtenidos con anterioridad relativos al estimador puntual, la distribución de probabilidad asociada, la estructura del intervalo de confianza y el error estándar no sufren variaciones. El cambio en el nivel de confianza exigido únicamente afectará al valor del coeficiente, que en este caso será el percentil $1 - \alpha/2 = 1 - 0{,}01/2 = 0{,}995$ en una distribución normal estándar.

$$P(Z \leq 2{,}576) = 0{,}995$$

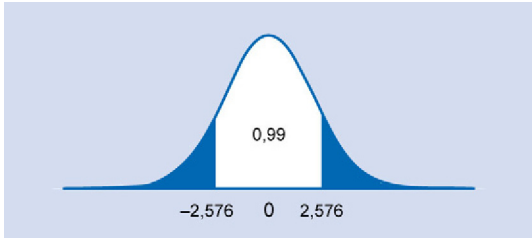

FIGURA 2-8 Percentiles 0,995 y 0,005 en la normal estándar.

Luego el intervalo de confianza para la proporción poblacional de éxito de la intervención en este tipo de pacientes quedará:

$$I_{1-\alpha}(p)=\left[0,8-2,576\sqrt{\frac{0,8(1-0,8)}{50}}\;;0,8+2,576\sqrt{\frac{0,8(1-0,8)}{50}}\right]=[0,65;0,95]$$

Esto quiere decir que la proporción de éxito de la intervención estará comprendida entre 0,65 y 0,95 con una confianza del 0,99 o del 99%. Si se compara este resultado con el obtenido anteriormente, en el que se trabajaba a un nivel de confianza del 95%, se observa que la amplitud del intervalo es mayor.

DETERMINACIÓN DE TAMAÑOS MUESTRALES

Las relaciones entre nivel de confianza, precisión y tamaño muestral sugieren que es posible determinar el tamaño muestral necesario para garantizar un determinado nivel de confianza y precisión en el proceso de construcción de un intervalo de confianza para un parámetro poblacional.

DETERMINACIÓN DEL TAMAÑO MUESTRAL PARA ESTIMAR UNA MEDIA POBLACIONAL

En el proceso de diseño muestral es importante determinar el tamaño muestral necesario para construir estimaciones de los parámetros de interés con un determinado nivel de confianza y precisión prefijados.

Ejemplo 2-8

Supóngase que se pretende calcular el tamaño muestral necesario para estimar la edad media de los individuos que practican deporte de forma regular en una determinada población con una confianza del 0,95 o del 95% y una precisión de 0,5 años. De la expresión:

$$\tau = c_{1-\alpha/2}\frac{S^*}{\sqrt{n}}$$

puede despejarse el valor de n de la siguiente forma:

$$\sqrt{n} = c_{1-\alpha/2}\frac{S^*}{\tau} \; ; \; n = \frac{c^2_{1-\alpha/2}\, S^{*2}}{\tau^2}$$

Como puede observarse, se necesitaría disponer del valor de S^{*2} para determinar el tamaño muestral, algo que es imposible sin haber observado previamente la muestra. La solución es seleccionar en una primera fase una muestra relativamente pequeña de la población de tamaño n_1 para obtener una primera estimación de S^{*2}. Supóngase que $n_1 = 30$ y que la varianza de la edad en esos 30 datos es $S^{*2} = 11$. Entonces, sustituyendo en la expresión se tendrá que:

$$n = \frac{c^2_{1-\alpha/2}\, S^{*2}}{\tau^2} = \frac{1{,}96^2 \cdot 11}{0{,}5^2} = 169{,}03$$

donde 1,96 es el percentil 0,975 en una normal estándar (téngase en cuenta que la t de Student tiende a la normal cuando n tiende a infinito). Como ya se cuenta con una muestra de tamaño n_1, faltarán $n - n_1$ datos para completar la muestra necesaria, es decir, $169{,}03 - 30 = 139{,}03 \simeq 140$ datos más.

DETERMINACIÓN DEL TAMAÑO MUESTRAL PARA ESTIMAR UNA PROPORCIÓN POBLACIONAL

Siguiendo con el ejemplo 2-8, supóngase que se pretende estimar, al 95% de confianza y precisión del 5%, la proporción de individuos de la población que practica deporte de forma regular. La precisión del intervalo vendrá dada por la expresión:

$$\tau = c_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

De donde, despejando el valor de n se obtendrá que:

$$n = \frac{c_{1-\alpha/2}{}^2 \cdot \hat{p}(1-\hat{p})}{\tau^2}$$

Al igual que en el caso de la media, no puede obtenerse un valor de $\hat{p}$ hasta que se haya obtenido una muestra. Podría procederse de diferentes formas:

- Seleccionar una primera muestra piloto de tamaño n_1 para obtener una estimación de $\hat{p}$ y sustituir en la expresión anterior. El tamaño final necesario vendrá determinado por $n = n - n_1$. Por ejemplo, si $n_1 = 30$ y $\hat{p} = 0{,}28$, entonces:

$$n = \frac{c_{1-\alpha/2}^2 \cdot \hat{p}(1-\hat{p})}{\tau^2} = \frac{1{,}96^2 \cdot 0{,}28(1-0{,}28)}{0{,}05^2} = 309{,}79$$

- Si se sabe que la proporción no puede ser superior o inferior a un determinado valor p_0 se utilizará este valor en la expresión. Por ejemplo, si el valor de p no puede ser superior a 0,1, entonces:

$$n = \frac{c_{1-\alpha/2}^2 \cdot \hat{p}(1-\hat{p})}{\tau^2} = \frac{1{,}96^2 \cdot 0{,}1(1-0{,}1)}{0{,}05^2} = 138{,}3$$

- Si se desconoce el rango de valores de la proporción que se pretende estimar puede utilizarse el valor $\hat{p} = 0{,}5$ como estimación de la proporción y sustituir en la expresión anterior. Debe tenerse en cuenta que, con independencia del valor de la proporción poblacional real, este último método garantizará que el tamaño muestral necesario será inferior o igual al obtenido. En este caso se tendrá que:

$$n = \frac{c_{1-\alpha/2}^2 \cdot \hat{p}(1-\hat{p})}{\tau^2} = \frac{1{,}96^2 \cdot 0{,}5(1-0{,}5)}{0{,}05^2} = 384{,}16$$

CONTRASTES DE HIPÓTESIS

Al igual que ocurre con la estimación de parámetros desconocidos de la población mediante intervalos de confianza, los contrastes de hipótesis también constituyen una herramienta para la realización de inferencias sobre determinados parámetros de la población objeto de estudio, a partir de una muestra observada de dicha población. El funcionamiento de esta segunda técnica inferencial se basa en la realización de una afirmación acerca de un parámetro de una o más poblaciones *(hipótesis)* y en el estudio de la compatibilidad entre esta afirmación y lo observado en la muestra. En principio, cuanto mayor sea la discrepancia entre la hipótesis realizada y la información proporcionada por la muestra observada, mayor será la evidencia en contra de dicha hipótesis.

Ejemplo 2-9

En un estudio se recabó información sobre el nivel de colesterol de un grupo de 46 pacientes seleccionados al azar de entre los afectados por una determinada patología. El promedio de nivel de colesterol fue de 235 mg/100 ml con una desviación típica de 28 mg/100 ml. ¿Proporcionan estos datos evidencia suficiente que indique que el nivel promedio de colesterol en este tipo de pacientes es superior a 220 mg/100 ml?

Ejemplo 2-10

De los 250 individuos a los que se les administró un determinado tratamiento, 180 respondieron de forma positiva. ¿Puede considerarse que la proporción de éxito del tratamiento es del 80%?

En el ejemplo 2-9, la población objeto de estudio la constituiría toda la población de pacientes afectados por dicha patología. El parámetro sobre el que se realiza la afirmación que se pretende contrastar sería la media de nivel de colesterol en mg/100 ml. La afirmación que se realiza sobre el parámetro (hipótesis) sería que la media es inferior o igual a 220 mg/100 ml.

En el ejemplo 2-10, la población objeto de estudio la conformarían todos los individuos afectados o que pudieran estar afectados por la patología que el tratamiento pretende tratar. El parámetro poblacional sobre el que se realiza la afirmación sería la proporción de éxito del tratamiento. La afirmación que se realiza sobre el parámetro (hipótesis) es que la proporción poblacional es de 0,8 (o del 80%). En ambos casos el resultado del contraste debe conducir a la confirmación o a la negación de la afirmación realizada.

CONTRASTE DE HIPÓTESIS SOBRE LA MEDIA DE UNA POBLACIÓN

Para ilustrar mejor el procedimiento de realización de un contraste de hipótesis, así como los elementos que intervienen en el mismo, se utilizarán los datos del ejemplo 2-9. Se cuenta con información sobre el nivel de colesterol de 46 pacientes en los que el promedio es de 235 mg/100 ml y la desviación típica de $S^* = 28$ mg/100 ml. A efectos de simplicidad se supondrá, en un primer momento, que lo que el investigador pretende demostrar es si el promedio de colesterol poblacional es igual o distinto de 220 mg/100 ml. Los pasos para la realización del contraste serían los siguientes:

1. DEFINICIÓN DE LAS HIPÓTESIS DEL CONTRASTE. HIPÓTESIS NULA E HIPÓTESIS ALTERNATIVA

El parámetro sobre el que se pretende realizar un contraste de hipótesis es, en este caso, una media poblacional. Se trata de decidir si puede considerarse que la media poblacional es 220 o, por el contrario, la media poblacional es distinta de 220. Estas dos posibles decisiones se expresan mediante las siguientes hipótesis:

$$H_0 : \mu = 220$$

$$H_1 : \mu \neq 220$$

Donde H_0 se denomina *hipótesis nula* y H_1 *hipótesis alternativa.* La mecánica del contraste establece que la hipótesis nula se mantendrá a no ser que los datos muestren una fuerte evidencia en contra de la misma, en cuyo caso, se optará por la hipótesis alternativa.

2. DEFINICIÓN DE UNA MEDIDA DE DISCREPANCIA O ESTADÍSTICO DE CONTRASTE ENTRE LO QUE SE AFIRMA EN LA HIPÓTESIS NULA Y LA INFORMACIÓN QUE PROPORCIONAN LOS DATOS DE LA MUESTRA OBSERVADA

En el caso del contraste de una media, cuando se desconoce la desviación típica poblacional, el estadístico de contraste utilizado es:

$$EC = \frac{\overline{x} - \mu_0}{S^* / \sqrt{n}}$$

donde μ_0 es el valor de la media que se especifica en la hipótesis nula, $\overline{x}$ es la media muestral, S^* el estimador puntual de la desviación típica poblacional y n el tamaño de la muestra. En este caso se tendrá:

$$EC = \frac{\overline{x} - \mu_0}{S^* / \sqrt{n}} = \frac{235 - 220}{28 / \sqrt{46}} = 3,63$$

Puede observarse que, dado que la desviación típica es siempre una cantidad positiva, la medida de discrepancia o estadístico de contraste tomará el valor 0 únicamente cuando la media observada en la muestra coincida exactamente con la que se propone en la hipótesis nula. Además, cuanto mayor sea la discrepancia entre la media observada y la media a la que se refiere la hipótesis nula, mayor será el numerador (en términos absolutos) y, por tanto, el valor del estadístico de contraste. De este modo puede concluirse que valores del estadístico de contraste próximos a 0 favorecerían a la hipótesis nula, mientras que valores muy distantes de 0 indicarían la existencia de evidencia en contra de dicha hipótesis. Será necesario decidir cuándo la discrepancia es lo suficientemente grande como para rechazar la hipótesis nula. En este sentido será muy útil la consideración del siguiente paso.

3. CONOCER LA DISTRIBUCIÓN DE PROBABILIDAD ASOCIADA A LA MEDIDA DE DISCREPANCIA O ESTADÍSTICO DE CONTRASTE

El conocimiento de la distribución de probabilidad que gobierna el comportamiento del estadístico de contraste será vital para el desarrollo final del contraste de hipótesis. En el ejemplo se cuenta con 46 datos. Si la variable nivel de colesterol sigue un modelo de distribución normal o por aplicación del teorema central del límite, puede establecerse que la distribución de probabilidad asociada al estadístico de contraste es,

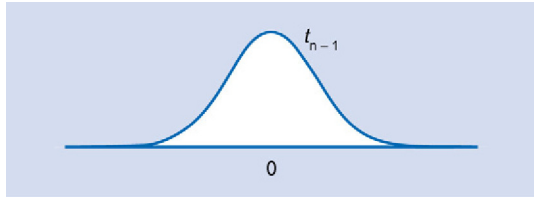

FIGURA 2-9 Distribución muestral del estadístico de contraste.

como fue discutido con anterioridad, una t de Student con $n - 1$ grados de libertad, tal y como se refleja en la figura 2-9.

En este caso se trataría de una distribución t de Student con $n - 1 = 46 - 1 = 45$ grados de libertad. La distribución t_{45} es simétrica y está centrada en 0. Si la hipótesis nula fuera cierta, el estadístico de contraste debería tomar valores en la zona central de la distribución, siendo muy improbable observar valores en cualquiera de sus dos extremos.

4. ESTABLECIMIENTO DEL NIVEL DE SIGNIFICACIÓN DEL CONTRASTE

Para decidir exactamente qué valores del estadístico de contraste tendrían una probabilidad de observarse prácticamente despreciable si la hipótesis nula fuera cierta, se establece el denominado *nivel de significación* del contraste o nivel de significación *a priori* α. Habitualmente se considera el valor $\alpha = 0{,}05$, aunque dependiendo del caso puede utilizarse el nivel 0,01 o 0,001. En este ejemplo se decide utilizar un nivel $\alpha = 0{,}05$.

5. CONSTRUCCIÓN DE LA REGLA DE DECISIÓN

Si el nivel de significación elegido es $\alpha = 0{,}05$, deberían despreciarse valores del estadístico de contraste con una probabilidad inferior a 0,05. Como fue discutido con anterioridad, los valores menos probables del estadístico de contraste si la hipótesis nula fuera cierta se concentran en ambos extremos de la distribución, por tanto, se define como *región crítica de contraste* o *región de rechazo* de la hipótesis nula a la región $]-\infty, t_{0,025}[\cup]t_{0,975}+\infty[$ que aparece sombreada en la figura 2-10.

De forma complementaria, se define como *región de aceptación* de la hipótesis nula a la región comprendida entre $t_{0,025}$ y $t_{0,975}$: $[t_{0,025}, t_{0,975}]$.

Los valores $t_{0,025}$ y $t_{0,975}$ sobre una distribución t_{45} que determinan la región de aceptación y de rechazo de la hipótesis nula en este caso son: $t_{0,025} = -2{,}014$ y $t_{0,975} = 2{,}014$. La regla de decisión quedará, por tanto:

- Si EC $> 2{,}014$ o EC $< -2{,}014 \rightarrow$ Se rechaza la hipótesis nula H_0.
- Si $-2{,}014 \leq$ EC $\leq 2{,}014 \rightarrow$ Se acepta la hipótesis nula H_0.

6. APLICACIÓN DE LA REGLA DE DECISIÓN

Por último, será necesario aplicar la regla de decisión del contraste de hipótesis. Para ello habrá que establecer la región en la que se sitúa el estadístico de contraste calculado con anterioridad en el paso 2.

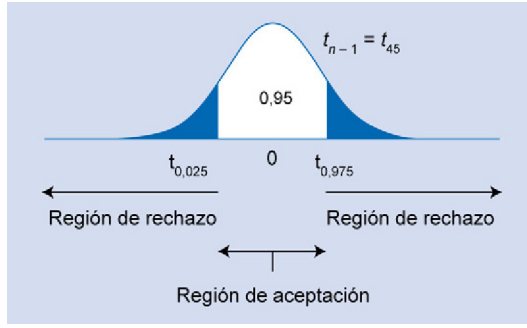

FIGURA 2-10 Región crítica de contraste y de aceptación de H_0.

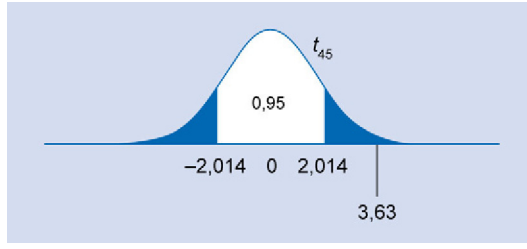

FIGURA 2-11 Situación del estadístico de contraste.

Como puede apreciarse en la figura 2-11, el valor del estadístico de contraste EC = 3,63 se encuentra situado en la región de rechazo de la hipótesis nula (3,63 > 2,014) y la regla de decisión conducirá a rechazarla, considerándose que existe evidencia suficiente que indica que el promedio de nivel de colesterol es *significativamente* distinto de 220 mg/100 ml.

ERRORES EN UN CONTRASTE DE HIPÓTESIS

Dado que en un contraste de hipótesis son dos los posibles resultados (rechazar la hipótesis nula o no rechazarla), también serán dos los posibles errores que puedan cometerse, consecuencia de cada una de las dos decisiones posibles. En la tabla 2-5 se refleja esta situación.

El error de tipo I se define como el error que se comete al rechazar la hipótesis nula cuando esta es cierta. El contraste conducirá al rechazo de

TABLA 2-5 Tipos de error en un contraste de hipótesis

Decisión del contraste	**Realidad**	
	H_0 cierta	**H_1 cierta (H_0 falsa)**
Se acepta H_0	Ausencia de error	Error tipo II
Se rechaza H_0	Error tipo I	Ausencia de error

la hipótesis nula únicamente cuando el estadístico de contraste se sitúe en la región crítica y esto solo puede ocurrir con una probabilidad α.

$$\alpha = P(\text{Rechazar } H_0 | H_0 \text{ cierta})$$

Luego, al establecer un nivel de significación para el contraste se está controlando la probabilidad de cometer un error de tipo I. Por otra parte, el error de tipo II es el que se comete cuando no se rechaza la hipótesis nula a pesar de que es falsa y suele denominarse β.

$$\beta = P(\text{No rechazar } H_0 | H_0 \text{ falsa})$$

HIPÓTESIS NULA E HIPÓTESIS ALTERNATIVA

Los contrastes de hipótesis se basan en la definición de dos hipótesis enfrentadas, la *hipótesis nula* H_0 y la *hipótesis alternativa* H_1. La hipótesis nula es la hipótesis que se pretende contrastar y será mantenida a menos que los datos observados en la muestra indiquen una fuerte evidencia de que no es cierta. Si en el procedimiento de realización del contraste de hipótesis únicamente se ha controlado el error de tipo I a través del establecimiento del nivel de significación *a priori* α y se desconoce la probabilidad de cometer un error de tipo II, la aceptación de la hipótesis nula no implica evidencia de su certeza sino, simplemente, que no se ha encontrado evidencia de que no lo sea. Por otra parte, si el contraste de hipótesis se decide por la hipótesis alternativa y, por tanto, rechaza la hipótesis nula, será porque la evidencia en contra de dicha hipótesis es manifiesta. Por este motivo, algunos autores prefieren afirmar que una hipótesis nula, contrastada en dichas condiciones, nunca puede ser aceptada, sino simplemente rechazada o no rechazada. Debe tenerse en cuenta que puede determinarse el tamaño muestral necesario para contrastar una hipótesis controlando simultáneamente el error de tipo I y de tipo II, de forma que las decisiones por la hipótesis nula tengan asociado un error tan pequeño como se requiera.

CONTRASTE Y NIVEL DE SIGNIFICACIÓN

La posibilidad de rechazar una hipótesis nula depende en gran parte de la magnitud de la región crítica de contraste. El tamaño de esta región crítica está determinado por el valor del nivel de significación del contraste α. Por tanto, el resultado del contraste de hipótesis será dependiente del nivel de significación elegido, pudiéndose dar el caso, en que se rechace la hipótesis nula trabajando con un nivel de significación, por ejemplo, $\alpha = 0{,}05$ y no rechazarse al nivel $\alpha = 0{,}01$.

En la figura 2-12 puede observarse que para todos los valores de la región (a) y (b) se rechazaría la hipótesis nula al nivel $\alpha = 0{,}05$, pero no al nivel $\alpha = 0{,}01$. Por otra parte, la aplicación de la regla de decisión del contraste tan solo permite concluir si se consigue o no se consigue rechazar la

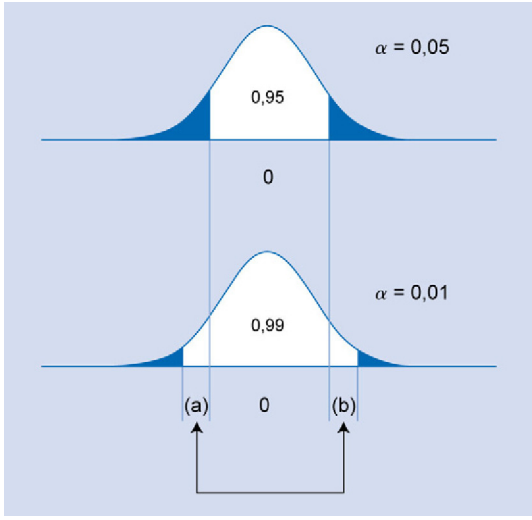

FIGURA 2-12 Diferencia en la región crítica de contraste según el nivel de significación.

hipótesis nula pero no informa sobre la magnitud de la evidencia en contra de dicha hipótesis. Ambos problemas pueden solucionarse definiendo el que se denomina *nivel crítico p, p-valor* o nivel de significación *a posteriori.*

NIVEL DE SIGNIFICACIÓN *A POSTERIORI* O P-VALOR

El *p-valor* se define como la probabilidad de observar, bajo la suposición de que la hipótesis nula es cierta, un valor del estadístico de contraste o medida de discrepancia igual o más extremo que el observado en la muestra. Por tanto, el valor de *p* no se fija a priori sino que es calculado a partir de los datos de la muestra *(a posteriori).* Valores muy pequeños de *p* estarían indicando que el estadístico de contraste se encuentra situado en cualquiera de los dos extremos de la distribución y la evidencia en contra de la hipótesis nula sería patente. Además, cuanto más pequeño sea el valor de *p* mayor será la evidencia en contra de la hipótesis nula.

En el ejemplo 2-9 el valor del estadístico de contraste era EC = 3,63. El valor de *p* será la probabilidad de observar un valor del estadístico o medida de discrepancia igual o más extremo al observado, tal y como puede observarse en la figura 2-13.

Dado que la región crítica de contraste tiene dos colas (se rechaza tanto con valores en el extremo derecho de la distribución como en el extremo izquierdo), el valor de la *p* del contraste debe calcularse teniendo en cuenta las dos regiones. En el ejemplo, el valor de *p* viene determinado por:

$$p = P(EC \geq 3,63) + P(EC \leq -3,63) = 0,00036 + 0,00036 = 0,00072$$

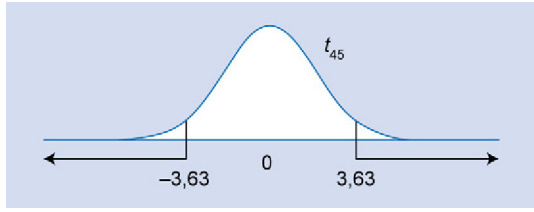

FIGURA 2-13 Región que determina el valor de p.

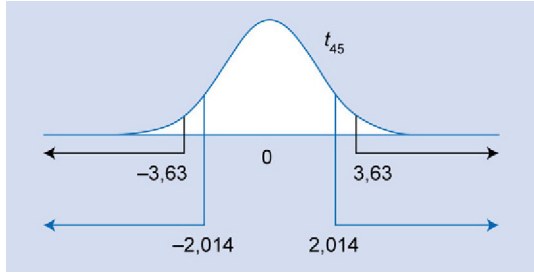

FIGURA 2-14 Valor p del contraste y nivel de significación α.

Un valor de $p = 0{,}00072$ indica que el EC se ha situado en el extremo de la distribución. Recuérdese que cuanto más pequeño sea este valor más evidencia en contra de la hipótesis nula. ¿Qué hubiera ocurrido si se hubiera trabajado al nivel de significación $\alpha = 0{,}05$?

Como puede observarse, en la figura 2-14, al nivel de significación $\alpha = 0{,}05$ se rechazaría la hipótesis nula, ya que el EC se situaría dentro de la región de rechazo. En general puede establecerse que, para un nivel de significación del contraste cualquiera α, se verificará:

- Si $p < \alpha \rightarrow$ Se rechaza H_0 al nivel α.
- Si $p \geq \alpha \rightarrow$ Se acepta H_0 al nivel α.

Obsérvese que en el ejemplo se rechazaría también la hipótesis nula trabajando al nivel de significación $\alpha = 0{,}01$ e incluso al nivel $\alpha = 0{,}001$, ya que el valor de la *p* del contraste es inferior en todos los casos.

CONTRASTES BILATERALES Y UNILATERALES

Los contrastes de hipótesis pueden ser *unilaterales* o *de una cola* y *bilaterales* o *de dos colas* dependiendo de la forma en que se planteen las hipótesis. Así, podría considerarse que cuando el interés se centra en contrastar la hipótesis de que un determinado parámetro de la población tome exacta-

mente un valor dado frente a la hipótesis de que el parámetro tome un valor distinto al propuesto, el contraste será bilateral o de dos colas.

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

Obsérvese que la región de rechazo definida por un contraste de este tipo quedaría situada a ambos extremos de la distribución, puesto que debería rechazarse la hipótesis nula tanto cuando $\theta > \theta_0$ como cuando $\theta < \theta_0$. Por otra parte, si la hipótesis se plantea de forma que se atiende únicamente al hecho de que ese mismo parámetro de la población tome un valor superior (análogamente inferior) a un valor dado, el contraste será unilateral o de una cola.

$$H_0 : \theta \leq \theta_0 \qquad H_0 : \theta \geq \theta_0$$
$$H_1 : \theta > \theta_0 \qquad H_1 : \theta < \theta_0$$

En este caso la región de rechazo definida por el contraste se situaría bien a la derecha de la distribución (se rechaza H_0 cuando $\theta > \theta_0$), o bien a la izquierda de la distribución (se rechaza H_0 cuando $\theta < \theta_0$).

Si se trabaja con los datos del ejemplo 2-9, dado que la hipótesis nula nunca se considera probada a menos que sea controlado el error de tipo II, la única posibilidad para demostrar que el nivel promedio de colesterol poblacional es superior a 220 mg/100 ml es colocar este supuesto en la hipótesis alternativa. Las hipótesis podrían formularse del siguiente modo para el contraste unilateral:

$$H_0 : \mu \leq 220$$
$$H_1 : \mu > 220$$

El estadístico de contraste y la distribución de probabilidad asociada coinciden exactamente con los expuestos en el contraste bilateral realizado con anterioridad. Se tendrá que:

$$EC = \frac{\overline{x} - \mu_0}{S^* / \sqrt{n}} = \frac{235 - 220}{28 / \sqrt{46}} = 3,63$$

que se distribuirá según una *t* de Student con $n - 1 = 46 - 1 = 45$ grados de libertad. Sin embargo, si se trabaja al mismo nivel de significación $\alpha = 0,05$ se producirán cambios en la región crítica de contraste (fig. 2-15) y la regla de decisión, que quedará de la siguiente forma:

- Si EC > 1,679 → Se rechaza la hipótesis nula H_0.
- Si EC ≤ 1,679 → Se acepta la hipótesis nula H_0.

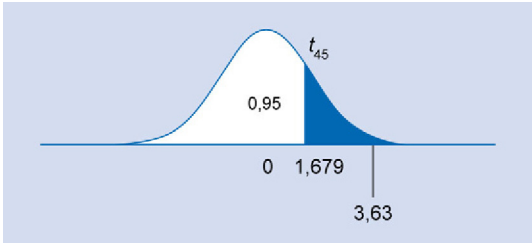

FIGURA 2-15 Región crítica para el contraste unilateral.

Como puede observarse, en el contraste unilateral se agranda la región crítica de contraste en el sentido de la discrepancia observada (obsérvese que en el contraste bilateral se rechazaba a la derecha de la distribución a partir del valor 2,014, mientras que en el contraste unilateral se rechaza a partir del valor 1,679), incrementándose la probabilidad de rechazar la hipótesis nula para los mismos datos observados. Por este motivo, muchos autores desaconsejan realizar contrastes unilaterales a menos que el conocimiento sobre el fenómeno objeto de estudio indique que los resultados solo puedan orientarse en una dirección (p. ej., valoración de un nuevo tratamiento que solo puede dejar igual o mejorar el estado de salud del paciente, pero no empeorarlo).

Por su parte, los contrastes bilaterales permiten la extracción de conclusiones en todos los casos y no tienen en cuenta el sentido de la discrepancia tras la observación de los datos. Así, en el ejemplo 2-9, si se parte de un contraste bilateral como el realizado con anterioridad, se llegaba a la conclusión de que debía rechazarse la hipótesis nula, luego, $\mu \neq 220$. Si el valor del promedio poblacional de colesterol es significativamente distinto de 220 será porque es mayor a 220 o porque es menor que 220. Como el valor del estadístico de contraste fue de EC = 3,63 se rechazaba a la derecha de la distribución (EC > 2,014). Esto es así porque el numerador del estadístico de contraste resulta un valor positivo:

$$EC = \frac{\bar{x} - \mu_0}{S^*/\sqrt{n}} = \frac{235 - 220}{28/\sqrt{46}} = 3,63$$

$$\bar{x} - \mu_0 = 235 - 220 = 15 > 0$$

Lo que ocurrirá siempre que la media observada sea superior al valor que se propone en la hipótesis nula. En este caso, podría concluirse que el promedio poblacional de colesterol es significativamente superior a 220 mg/100 ml a partir del contraste bilateral.

En conclusión, existen mayores probabilidades de rechazar la hipótesis nula bajo el planteamiento de contrastes unilaterales. Probablemente el investigador del ejemplo anterior planteó el contraste unilateral una vez calculada la media de la muestra y después de haber observado que esta era superior a 220, pretendiendo entonces demostrar si esa diferencia a favor de una media superior era significativa. Dado que previamente a la observación de los datos de la muestra se desconoce si la media será o no superior a la que se propone en la hipótesis nula, no tendría sentido plantearse si la media de la población será o no significativamente superior a determinado valor sino, en todo caso, si sería significativamente distinta de una valor dado (mayor o menor). Por tanto, el planteamiento bilateral del contraste sería el más indicado.

POTENCIA DE UN CONTRASTE

La potencia de un contraste de hipótesis se define como la probabilidad de rechazar H_0 cuando es falsa o, equivalentemente, la probabilidad de decidirse por la hipótesis alternativa cuando esta es cierta.

$$\text{Potencia} = P(\text{rechazar } H_0 | H_0 \text{falsa})$$

En algunos estudios, las características de la muestra seleccionada pueden impedir la detección de evidencia significativa en contra de la hipótesis nula que se plantea aunque esta sea falsa. En este sentido, la potencia del contraste podría interpretarse como la probabilidad de encontrar en el estudio evidencia significativa en contra de la hipótesis nula, en el caso de que efectivamente la hipótesis nula fuera falsa.

Cuando la hipótesis alternativa es una hipótesis simple (especifica un único valor para el parámetro) el valor de la potencia es único. Sin embargo, cuando la hipótesis alternativa es compuesta existirá un valor de *potencia* asociado a cada una de las posibilidades. Esto sugiere la definición de la que se denomina *función de potencia* de un contraste de la forma:

$$\text{Pot}(\theta) = P(\text{Rechazar } H_0 | \theta)$$

donde θ representa todas las posibilidades del parámetro sobre el que se pretende contrastar alguna hipótesis.

Cuando sustituimos θ por el valor que se especifica en la hipótesis nula θ_0, la función de potencia toma el valor α.

$$\text{Pot}(\theta_0) = P(\text{Rechazar } H_0 | \theta_0) = \alpha = P(\text{Error tipo I})$$

Por otra parte, cuando θ toma cualquier otro valor, la función de potencia quedará:

$$\begin{aligned}\text{Pot}(\theta) &= P(\text{Rechazar } H_0 | \theta) = 1 - P(\text{Aceptar } H_0 | \theta) = 1 - \beta(\theta) \\ &= 1 - P(\text{Error tipo II})\end{aligned}$$

Lo deseable es que tanto el error de tipo I como el error de tipo II sean lo más bajos posible, siendo habitual trabajar al nivel de significación $\alpha = 0,05$ y potencia no superior a 0,2.

CONTRASTE DE HIPÓTESIS SOBRE UNA PROPORCIÓN

El ejemplo 2-10 planteaba que de los 250 individuos a los que se les administró un determinado tipo de tratamiento, 180 respondieron de forma positiva. ¿Puede considerarse que la proporción de éxito del tratamiento es del 0,8 o del 80%? Para responder a esta cuestión será necesario realizar un contraste de hipótesis sobre la proporción poblacional de respuesta positiva al tratamiento. Si se siguen los pasos para la realización del contraste, establecidos con anterioridad se tendrá que:

$$H_0 : p = 0,8$$

$$H_0 : p \neq 0,8$$

La medida de discrepancia o *estadístico de contraste* en el caso de una proporción poblacional será:

$$EC = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

donde p_0 es el valor de la proporción poblacional que se especifica en la hipótesis nula, $\hat{p}$ la proporción calculada a partir de los datos de la muestra y n el tamaño de la muestra. En este caso se tendrá que:

$$EC = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = \frac{\frac{180}{250} - 0,8}{\sqrt{\frac{\frac{180}{250}\left(1-\frac{180}{250}\right)}{250}}} = \frac{0,72 - 0,8}{\sqrt{\frac{0,72(1-0,72)}{250}}} = 2,81$$

La distribución de probabilidad asociada sería aproximadamente la normal estándar si se verifica que:

$$n\hat{p} = 250\frac{180}{250} = 180 \geq 5$$

$$n(1-\hat{p}) = 150\left(1-\frac{180}{250}\right) = 70 \geq 5$$

Como puede observarse, en este caso se verifican las condiciones para la aproximación normal. En caso contrario sería necesario utilizar la distribución binomial exacta. Si se trabaja al nivel habitual $\alpha = 0,05$, como el valor del estadístico de contraste (EC = 2,81) es mayor que 1,96, se situará en

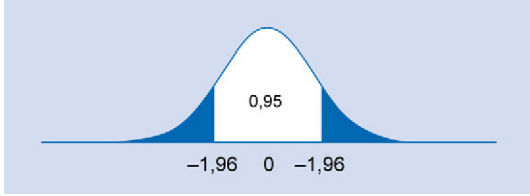

FIGURA 2-16 Región crítica de contraste. Contraste de una proporción.

la región crítica de contraste (fig. 2-16) y, por tanto, se procederá a rechazar la hipótesis nula. Podrá concluirse, por tanto, que la proporción poblacional de individuos con respuesta positiva al tratamiento es significativamente distinta de 0,8 (o del 80%). Dado que el rechazo se ha producido a la derecha de la distribución, puede concluirse que la proporción poblacional es significativamente superior a 0,8.

COMPARACIÓN DE DOS PROPORCIONES POBLACIONALES. MUESTRAS INDEPENDIENTES

Supóngase que de los 250 individuos del ejemplo anterior 150 eran hombres y el resto mujeres. Además, de los 180 que respondieron positivamente al tratamiento, 110 eran hombres y 70 mujeres. A partir de estos datos, ¿puede afirmarse que existen diferencias significativas en las proporciones de respuesta positiva al tratamiento entre hombres y mujeres?

En primer lugar, se procede al establecimiento de las hipótesis del contraste que, en este caso, quedarán de la siguiente forma:

$$H_0 : p_1 = p_2 \qquad H_0 : p_1 - p_2 = 0$$
$$H_0 : p_1 \neq p_2 \qquad H_0 : p_1 - p_2 \neq 0$$

Donde, por ejemplo, p_1 es la proporción poblacional de respuesta positiva en hombres y p_2 en mujeres. La medida de discrepancia o *estadístico de contraste* en el caso de la diferencia de dos proporciones poblacionales será:

$$EC = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

donde $\hat{p}_1$ y $\hat{p}_2$ son las proporciones calculadas a partir de los datos de la muestra en hombres y mujeres respectivamente, y $\hat{p}$ es la proporción global (sin distinguir entre grupos) de respuesta positiva. En este caso se tendrá que:

$$\hat{p} = \frac{r_1 + r_2}{n_1 + n_2} = \frac{110 + 70}{150 + 100} = 0,72$$

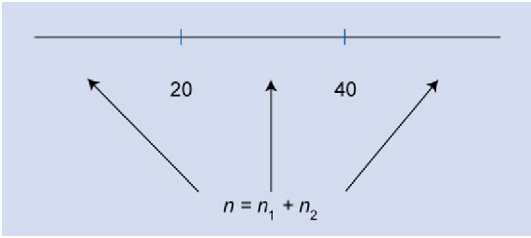

FIGURA 2-17 Zonas de decisión para la aproximación normal.

$$EC = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} = \frac{\frac{110}{150} - \frac{70}{100}}{\sqrt{\frac{0{,}72(1-0{,}72)}{150} + \frac{0{,}72(1-0{,}72)}{100}}} = 0{,}575$$

La distribución de probabilidad asociada sería aproximadamente la normal estándar si se verifican determinados requisitos sobre los tamaños de las muestras. En este sentido, si se calcula $n = n_1 + n_2$, entonces puede ocurrir que se sitúe en una de las tres zonas contempladas en la figura 2-17.

Si $n = n_1 + n_2 = 150 + 100 = 250$ es menor que 20 entonces será necesario recurrir a la distribución binomial exacta. Si *n* toma un valor mayor que 40 la distribución será aproximadamente la normal estándar. Si *n* toma un valor entre 20 y 40 entonces, para garantizar la aproximación de la normal estándar será necesario comprobar que:

$$n_1\hat{p}_1 = 150\frac{110}{150} = 110 \geq 5 \; ; \; n_2\hat{p}_2 = 100\frac{70}{100} = 70 \geq 5$$

$$n_1(1-\hat{p}_1) = 150\left(1-\frac{110}{150}\right) = 40 \geq 5 \; ; \; n_2(1-\hat{p}_2) = 100\left(1-\frac{70}{100}\right) = 30 \geq 5$$

Como puede observarse en este caso, $n = 150 + 100 = 250 > 40$ y se verifican las condiciones para la aproximación normal. En caso contrario sería necesario utilizar la distribución binomial exacta. En la figura 2-18 se representa la región crítica de contraste para el nivel de significación habitual $\alpha = 0{,}05$.

Como el valor del estadístico de contraste EC = 0,575 se encuentra entre −1,96 y 1,96, se procederá a aceptar la hipótesis nula. Podrá concluirse, por tanto, que no existe evidencia de que las proporciones poblacionales de respuesta positiva al tratamiento sean significativamente distintas entre hombres y mujeres.

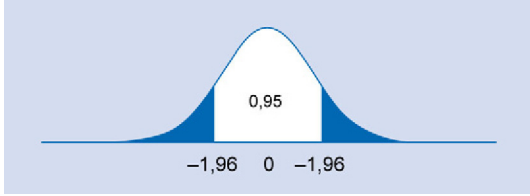

FIGURA 2-18 Región crítica de contraste. Comparación de dos proporciones.

PRUEBA JI-CUADRADO

Una alternativa a la prueba de comparación de dos proporciones poblacionales más general que la anterior y que permite incluso la comparación de tres o más proporciones la constituye la prueba ji-cuadrado. Si se trabaja con los datos del ejemplo 2-10, en primer lugar será necesario construir una tabla 2×2 como la descrita en la tabla 2-6.

La hipótesis del contraste puede ser, por ejemplo, que las proporciones de respuesta positiva al tratamiento sean iguales en hombres y mujeres frente a la alternativa de que sean distintas. Por tanto, se tendrá:

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

Para construir el estadístico de contraste deberá tenerse en cuenta cuáles serían las frecuencias en cada una de las casillas o celdas de la tabla si la hipótesis nula fuera cierta. Si las dos proporciones fueran iguales en hombres y mujeres existiría una única proporción global de respuesta positiva al tratamiento, que sería:

$$\hat{p} = \frac{180}{250} = 0,72$$

Luego se esperaría que, tanto en hombres como en mujeres, el 72% mostrara una respuesta positiva, mientras que el 28% tendría una respuesta negativa en ambos grupos. Así se tendrá que:

$$\textit{frecuencia esperada respuesta positiva}_{\text{Hombres}} = 0,72 \cdot 150 = \frac{180 \cdot 150}{250} = 108$$

TABLA 2-6 Distribución de frecuencias de respuesta al tratamiento según sexo

Respuesta al tratamiento	Hombres	Mujeres	Total
Positiva	110	70	180
Negativa	40	30	70

TABLA 2-7 Frecuencias esperadas bajo la hipótesis nula

Respuesta al tratamiento	Hombres	Mujeres	Total
Positiva	108	72	180
Negativa	42	28	70

$$\textit{frecuencia esperada respuesta positiva}_{\text{Mujeres}} = 0,72 \cdot 100 = \frac{180 \cdot 100}{250} = 72$$

De esta forma, podría completarse una tabla con las frecuencias esperadas E_{ij} en cada una de las casillas. Una forma rápida para el cálculo de los valores esperados en cada celda se basa en los totales por fila, columna y total. Así, por ejemplo, para obtener la frecuencia esperada E_{ij} se calculará:

$$E_{ij} = \frac{\text{Total por fila} \cdot \text{Total por columna}}{\text{Total}}$$

La tabla 2-7 muestra todas las frecuencias esperadas obtenidas.

Finalmente el estadístico de contraste se basará en la diferencia entre las frecuencias observadas en cada una de las celdas y los valores esperados bajo la hipótesis de igualdad de proporciones y quedará de la siguiente forma:

$$EC = \sum \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

que se distribuirá según un modelo de distribución ji-cuadrado con $(n°_{\text{filas}} - 1)(n°_{\text{columnas}} - 1)$ grados de libertad que, en este caso, quedará: $(2 - 1)(2 - 1) = 1$.

Así, se tendrá que:

$$EC = \frac{(110-108)^2}{108} + \frac{(70-72)^2}{72} + \frac{(40-42)^2}{42} + \frac{(30-28)^2}{28} = 0,331$$

Que se distribuirá según un modelo de distribución ji-cuadrado con 1 grado de libertad. Se requiere para garantizar la distribución muestral que la frecuencia esperada en cada una de las casillas sea igual o superior a 5. En el caso de que alguna o algunas de las casillas contengan un valor esperado inferior a 5 podrá utilizarse la distribución ji-cuadrado siempre que el porcentaje de casillas en los que se dé esta circunstancia no sea superior al 20%. En el ejemplo, todas las casillas muestran unos valores esperados superiores a 5.

Si las frecuencias esperadas bajo la hipótesis nula coincidieran con las observadas, el valor del estadístico de contraste sería cero. Por el contrario, cuanto mayores sean las diferencias entre las frecuencias esperadas y las

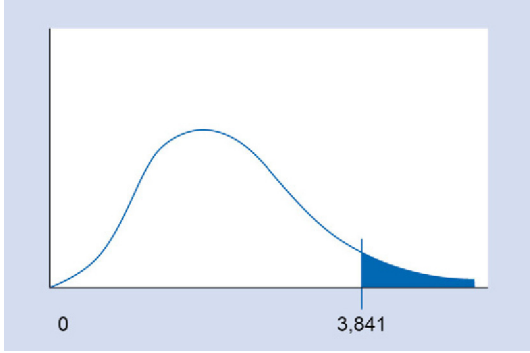

FIGURA 2-19 Región crítica de contraste. Prueba ji-cuadrado.

observadas mayor evidencia en contra de la hipótesis nula. Así, la región crítica de contraste se situará únicamente a la derecha de la distribución. Si se trabaja al nivel $\alpha = 0{,}05$, el percentil 0,95 es 3,841 y la región crítica quedará como se describe en la figura 2-19.

Como el valor del estadístico de contraste EC = 0,331 es inferior a 3,841, se procederá a aceptar la hipótesis nula y concluir que no existe evidencia que indique diferencias en las proporciones de respuesta positiva al tratamiento entre hombres y mujeres.

COMPARACIÓN DE DOS VARIANZAS POBLACIONALES

En determinadas pruebas de hipótesis como, por ejemplo, los contrastes de hipótesis sobre la diferencia de dos o más medias poblacionales para muestras independientes que se abordarán posteriormente, se requiere comprobar si las varianzas poblacionales correspondientes son iguales o distintas. La decisión sobre la igualdad o diferencia entre las varianzas contempladas en el análisis tendrá una influencia decisiva en el estadístico de contraste y la distribución muestral asociada, necesarios para la comparación de las medias poblacionales.

Ejemplo 2-11

Supóngase que se cuenta con información sobre la edad en dos grupos de pacientes seleccionados al azar. El promedio de edad en el primer grupo formado por 40 individuos fue de 56 años, con una desviación típica de $S_1^* = 12$ años, mientras que en el segundo de 35 individuos el promedio de edad fue de 62 años y una desviación típica de $S_2^* = 14$ años.

Para comprobar si existe evidencia o no de que las varianzas poblacionales son significativamente distintas se plantea el siguiente contraste de hipótesis:

$$H_0: \sigma_1^2 = \sigma_2^2 \quad ; \quad H_0: \sigma_1^2 / \sigma_2^2 = 1$$

$$H_0: \sigma_1^2 \neq \sigma_2^2 \quad ; \quad H_0: \sigma_1^2 / \sigma_2^2 \neq 1$$

El estadístico de contraste o medida de discrepancia en el caso de la comparación de dos varianzas poblacionales será:

$$EC = \frac{S^{*2}_{mayor}}{S^{*2}_{menor}} = \frac{14^2}{12^2} = \frac{196}{144} = 1,361$$

Téngase en cuenta que, para facilitar el cálculo, se ha situado la varianza mayor en el numerador y la varianza menor en el denominador. De esta forma, el estadístico de contraste se situará siempre a la derecha de la distribución aunque se trabaje en un contraste bilateral. Si la variable nivel de colesterol se comporta según un modelo de distribución normal, este estadístico de contraste se distribuirá según una distribución de probabilidad F de Snedecor con $n_{numerador} - 1$ y $n_{denominador} - 1$ grados de libertad. En este caso será una distribución $F_{35-1,\ 40-1}$. En la figura 2-20 se representa la distribución F de Snedecor con 34 y 39 grados de libertad. Si se trabaja al nivel habitual $\alpha = 0,05$ será necesario obtener el valor del percentil 0,975 que, en este caso, es 1,921 (consulte las tablas de la F de Snedecor). Como el estadístico de contraste EC = 1,361 es inferior a 1,921, puede concluirse que no existe evidencia que indique que las varianzas poblacionales sean significativamente distintas.

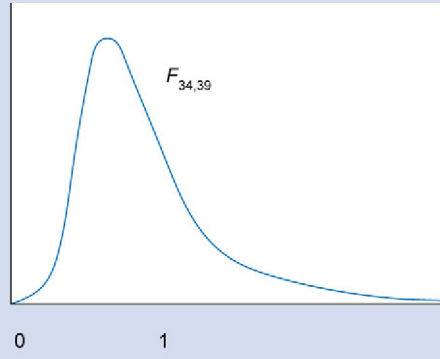

FIGURA 2-20 Distribución F de Snedecor. Prueba de igualdad de varianzas.

MUESTRAS INDEPENDIENTES Y RELACIONADAS O APAREADAS

Las muestras independientes se caracterizan porque los individuos o elementos que forman parte de cada una de las muestras han sido obtenidos de forma independiente a los de la otra. Por ejemplo, se obtiene información sobre un grupo de 230 individuos seleccionados al azar, de los cuales 150 fueron hombres y 80 mujeres. Se pretende comparar las medias de cualquier característica de los individuos entre estos dos grupos. Por otra parte, las muestras apareadas o relacionadas requieren que por cada individuo o elemento que forma parte de la primera de las muestras se seleccione otro de las mismas características (p. ej., sexo, edad o ambos) que constituirá su pareja. De esta forma, los elementos de la segunda muestra vienen determinados por los elementos de la primera. Una situación habitual la constituye el caso de un grupo de individuos observados en dos momentos del tiempo.

COMPARACIÓN DE DOS MEDIAS POBLACIONALES

Para la realización de un contraste sobre la diferencia de dos medias poblacionales es necesario tener en cuenta:

- Si se trata de muestras independientes o apareadas (relacionadas).
- Si las varianzas son iguales o distintas.

Adicionalmente, será necesario contar con un número mínimo de datos para utilizar las pruebas paramétricas correspondientes. En otro caso, deberá recurrirse a pruebas no paramétricas que se abordarán en el capítulo 3.

MUESTRAS INDEPENDIENTES. VARIANZAS POBLACIONALES IGUALES (PRUEBA T *DE COMPARACIÓN DE MEDIAS PARA MUESTRAS INDEPENDIENTES. VARIANZAS IGUALES)*

Para la utilización de la prueba *t* de comparación de medias se requiere que las muestras sean independientes y que la variable siga una distribución normal. En la práctica suele utilizarse cuando el tamaño de cada una de las muestras es superior o igual a 30. Si se trabaja con los datos del ejemplo 2-11, se tendrá que:

$$\overline{x}_1 = 56 \; ; \overline{x}_2 = 62$$

$$S_1^* = 12 \; ; S_2^* = 14$$

$$n_1 = 40 \; ; n_2 = 35$$

Puede observarse que el tamaño de las muestras es superior a 30 en los dos casos. Por otra parte, anteriormente pudo comprobarse que, en este

caso, el contraste de hipótesis de comparación de varianzas poblacionales llevó a la conclusión de que no existía evidencia en contra de la hipótesis de igualdad de varianzas. Por tanto, estará indicada la utilización de la prueba *t* para muestras independientes y varianzas poblacionales iguales. El estadístico de contraste será:

$$EC = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Donde:

$$S_p^2 = \frac{(n_1 - 1)S_1^{*2} + (n_2 - 1)S_2^{*2}}{n_1 + n_2 - 2} = \frac{(40 - 1)12^2 + (35 - 1)14^2}{40 + 35 - 2} = 168,22$$

Téngase en cuenta que, dado que se considera que las varianzas poblacionales son iguales y se cuenta con dos varianzas (una para cada muestra), se construye una varianza común ponderándolas adecuadamente y obteniendo la cantidad S_p^2. El valor del estadístico quedará:

$$EC = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{56 - 62}{\sqrt{168,22\left(\frac{1}{40} + \frac{1}{35}\right)}} = -1,999$$

Que se distribuirá según una *t* de Student con $n_1 + n_2 - 2 = 40 + 35 - 2 = 73$ grados de libertad. Si se trabaja al nivel de significación $\alpha = 0,05$ se tendrá que el percentil 0,975 será 1,993. Así, la región crítica de contraste quedará definida como se describe en la figura 2-21.

Como el estadístico de contraste EC = −1,999 es menor que −1,993 se situará en la región crítica de contraste (cola de la izquierda), pudiéndose concluir que existe evidencia de que las medias de edad son significativamente distintas entre los dos grupos de pacientes.

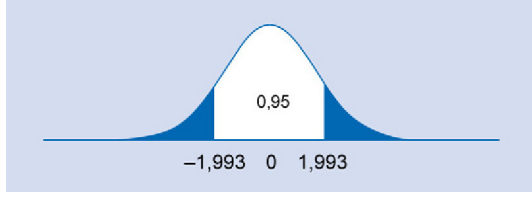

FIGURA 2-21 Región crítica de contraste. Comparación de medias. Varianzas iguales.

MUESTRAS INDEPENDIENTES. VARIANZAS POBLACIONALES DISTINTAS (PRUEBA T DE COMPARACIÓN DE MEDIAS PARA MUESTRAS INDEPENDIENTES. VARIANZAS DISTINTAS)

Para ilustrar los cambios que provoca la consideración de varianzas poblacionales distintas se trabajará con los mismos datos que en el ejemplo 2-11, a excepción de las desviaciones típicas muestrales que se considerarán, en este caso, de 10 años en el primer grupo y de 18 años en el segundo. Así, se tendrá que:

$$\overline{x}_1 = 56 \; ; \overline{x}_2 = 62$$

$$S_1^* = 10 \; ; S_2^* = 18$$

$$n_1 = 40 \; ; n_2 = 35$$

En primer lugar, habría que realizar un contraste de igualdad de varianzas. Con los datos actuales quedará:

$$EC = \frac{S^{*2}_{mayor}}{S^{*2}_{menor}} = \frac{18^2}{10^2} = \frac{324}{100} = 3,24$$

La distribución de probabilidad asociada seguirá siendo la $F_{34,39}$ con lo que el percentil 0,975 coincidirá con el obtenido anteriormente que, en este caso es 1,921. Como 3,24 es superior a 1,921, se situará en la región crítica de contraste, concluyéndose que existe evidencia de que las varianzas poblacionales son significativamente distintas. El estadístico de contraste deberá tener en cuenta cada una de las varianzas muestrales observadas de forma separada, quedando de la siguiente forma:

$$EC = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{S^{*2}_1}{n_1} + \frac{S^{*2}_2}{n_2}}} = \frac{56 - 62}{\sqrt{\frac{10^2}{40} + \frac{18^2}{35}}} = -1,75$$

Que se distribuirá según una distribución de probabilidad *t* de Student con *f* grados de libertad, donde:

$$f = \frac{\left(\frac{S^{*2}_1}{n_1} + \frac{S^{*2}_2}{n_2}\right)^2}{\frac{\left(\frac{S^{*2}_1}{n_1}\right)^2}{n_1} + \frac{\left(\frac{S^{*2}_2}{n_2}\right)^2}{n_2}} = \frac{\left(\frac{10^2}{40} + \frac{18^2}{35}\right)^2}{\frac{\left(\frac{10^2}{40}\right)^2}{40} + \frac{\left(\frac{18^2}{35}\right)^2}{35}} \simeq 53$$

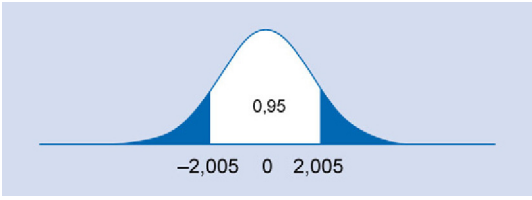

FIGURA 2-22 Región crítica de contraste. Comparación de medias. Varianzas distintas.

En la figura 2-22 puede observarse la región crítica de contraste para el nivel de significación $\alpha = 0{,}05$.

Como el estadístico de contraste EC = −1,75 se encuentra situado entre −2,005 y 2,005, puede concluirse que no existe evidencia de que los promedios de edad sean significativamente distintos entre los dos grupos.

MUESTRAS RELACIONADAS O APAREADAS (PRUEBA T *DE COMPARACIÓN DE MEDIAS PARA MUESTRAS RELACIONADAS)*

Para ilustrar el proceso de comparación de dos medias poblacionales a partir de muestras relacionadas se propone un ejemplo con un número reducido de observaciones para facilitar los cálculos. (Debe tenerse en cuenta que, en la práctica, el tamaño muestral mínimo para su utilización es de 30.)

Ejemplo 2-12

En un estudio sobre la eficacia de un determinado tratamiento en el que participaron 10 individuos se recogió información sobre el nivel de triglicéridos antes y después de ser sometidos al tratamiento. Los resultados se muestran en la tabla 2-8.

La idea principal es construir una nueva variable x_D basada en la diferencia entre el nivel de triglicéridos, después y antes del tratamiento, en cada uno de los individuos. Así, las hipótesis podrán expresarse de la siguiente forma:

$$H_0 : \mu_D = 0$$

$$H_1 : \mu_D \neq 0$$

Donde μ_D es el promedio poblacional de las diferencias. Si este promedio es cero significará que, en conjunto, no se observan cambios en el nivel de

TABLA 2-8 Nivel de triglicéridos antes y después del tratamiento

	Nivel de triglicéridos (mg/dl)		
Individuo	**Antes**	**Después**	**x_D = diferencia**
1	350	215	−135
2	175	180	5
3	254	221	−33
4	402	340	−62
5	324	260	−64
6	180	142	−38
7	285	251	−34
8	410	320	−90
9	260	270	10
10	372	201	−171

triglicéridos antes y después del tratamiento. Como puede observarse, el problema de la comparación de dos medias poblacionales para muestras relacionadas ha sido transformado en un simple contraste para una media poblacional, solo que trabajando con la variable x_D. El estadístico de contraste quedará, por tanto:

$$EC = \frac{\overline{x}_D - 0}{S_D/\sqrt{n}} = \frac{\overline{x}_D}{S_D/\sqrt{n}}$$

Si se calculan la media y la desviación típica de la variable x_D se tendrá que:

$$\overline{x}_D = \frac{(-135)+5+(-33)+(-62)+(-64)+\ldots+(-171)}{10} = -61,2$$

$$S_D^* = \sqrt{\frac{\left(-135-(-61,2)\right)^2+\left(5-(-61,2)\right)^2+\ldots+\left(-171-(-61,2)\right)^2}{10-1}} = 57,61$$

Luego el estadístico de contraste quedará:

$$EC = \frac{\overline{x}_D}{S_D^*/\sqrt{n}} = \frac{-61,2}{57,61/\sqrt{10}} = -3,359$$

La distribución muestral del estadístico de contraste sería la *t* de Student con $n - 1 = 10 - 1 = 9$ grados de libertad (recuérdese que, en la práctica, se requiere que el tamaño muestral sea 30 o superior). Los percentiles 0,975 y 0,025 son, en este caso, 2,262 y −2,262, respectivamente.

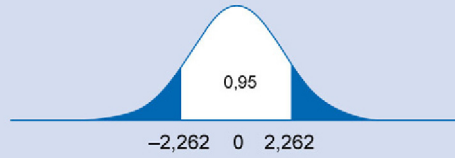

FIGURA 2-23 Región crítica de contraste. Prueba *t* apareada.

Como el valor del estadístico de contraste EC = −3,359 es inferior a −2,262, se situará en la región crítica de contraste (fig. 2-23), concluyéndose que existe evidencia que indica la existencia de diferencias significativas en los niveles promedio de triglicéridos antes y después del tratamiento.

AUTOEVALUACIÓN

1. El muestreo aleatorio:
- **a.** Garantiza que la muestra sea representativa de la población.
- **b.** Requiere que todos los individuos o elementos de la población tengan una probabilidad conocida y distinta de cero de ser incluidos en la muestra.
- **c.** Solo es aplicable en poblaciones finitas.
- **d.** Permite la aplicación de las técnicas inferenciales: intervalos de confianza y contrastes de hipótesis.
- **e.** b y d son ciertas.

2. Para una muestra aleatoria determinada se verifica que:
- **a.** A mayor nivel de confianza mayor precisión del intervalo de confianza.
- **b.** A mayor nivel de confianza menor precisión del intervalo de confianza.
- **c.** A menor nivel de confianza menor precisión del intervalo de confianza.
- **d.** A menor nivel de confianza mayor precisión del intervalo de confianza.
- **e.** b y d son ciertas.

3. Se pretende estimar la proporción de individuos de una población afectados por una patología:
- **a.** Será necesario realizar un contraste de hipótesis sobre la proporción poblacional de afectados.
- **b.** Será necesario seleccionar una muestra de la población aunque no sea aleatoria.
- **c.** Deberá construirse un intervalo de confianza para la proporción poblacional de afectados.
- **d.** Deberá compararse la proporción de afectados con la de no afectados.
- **e.** b y c son ciertas.

4. En un contraste de hipótesis:
   - **a.** El error de tipo I es la probabilidad de cometer un error cuando se rechaza la hipótesis nula.
   - **b.** El nivel de significación α representa la probabilidad de cometer un error de tipo II.
   - **c.** La hipótesis alternativa se mantendrá como cierta a menos que los datos muestren evidencia de lo contrario.
   - **d.** La región crítica de contraste representa la zona de aceptación de la hipótesis nula.
   - **e.** Si el p-valor es de 0,025, se procederá a rechazar la hipótesis nula al nivel $\alpha = 0{,}01$.

5. Se dispone de una muestra aleatoria de 45 individuos observados en dos instantes de tiempo. Para comparar el promedio de ácido úrico antes y después:
   - **a.** Podrá utilizarse la prueba *t* de comparación de medias para muestras independientes con varianzas poblacionales conocidas.
   - **b.** Podrá utilizarse la prueba *t* de comparación de medias para muestras independientes con varianzas poblacionales desconocidas pero iguales.
   - **c.** Podrá utilizarse la prueba *t* de comparación de medias para muestras independientes con varianzas poblacionales desconocidas y distintas.
   - **d.** Podrá utilizarse la prueba *t* para muestras relacionadas o apareadas.
   - **e.** Se requiere un tamaño muestral mayor para poder realizar la prueba correspondiente.

CAPÍTULO

3

Pruebas no paramétricas

Joaquín Moncho Vasallo

INTRODUCCIÓN

Las pruebas paramétricas requieren la estimación de uno o varios parámetros poblacionales, con el propósito de establecer la distribución de la variable objeto de estudio necesaria para la realización de inferencias. Así, una prueba como la prueba *t* para el contraste de una media de una población requiere que los datos provengan de una población normal que, como es sabido, depende de dos parámetros: la media poblacional μ y su desviación típica σ. En situaciones en que el tamaño de la muestra es lo suficientemente grande (habitualmente $n \geq 30$) el teorema central del límite garantiza que la distribución de la media muestral $\bar{x}$ será asintóticamente normal con la misma media que la variable original μ y con desviación típica $\sigma/\sqrt{n}$ con independencia de la distribución real de la variable analizada. Se dice, por tanto, que una prueba como la prueba *t* para el contraste de una media es robusta ante la falta de normalidad de la variable objeto de estudio, pero siempre que el tamaño muestral sea grande. La cuestión es: ¿qué ocurre cuando se desconoce la distribución poblacional de la variable objeto de estudio o el número de observaciones es pequeño y no permite la utilización del teorema central del límite para garantizar su distribución asintótica?

A continuación se muestran algunas situaciones a modo de ejemplo:

- Se dispone de un total de 10 observaciones del nivel de ácido úrico de un conjunto de individuos seleccionados al azar de una población. ¿Existe evidencia de que el nivel promedio de ácido úrico poblacional sea inferior a 5,2 mg/dl?
- ¿Existen diferencias en los niveles promedio de ácido úrico poblacionales entre dos grupos de individuos de dicha población? El tamaño de las muestras disponibles es inferior a 30 en uno o ambos grupos.

Aunque se dispone de estrategias para la comprobación de la hipótesis de normalidad de la variable objeto de estudio, basadas en el análisis de

la simetría y curtosis, gráficos comparativos de distribución (Q-Q normal o P-P normal) o pruebas de hipótesis como la prueba de Kolmogorov-Smirnov, en la práctica existen serias dudas sobre la conveniencia de su utilización cuando el número de datos en la muestra es pequeño, ya que sería difícil rechazar la normalidad de la distribución a menos que los escasos datos disponibles muestren discrepancias muy acentuadas.

En otros casos en los que se pretende comparar el comportamiento de una variable en dos o más grupos (p. ej., el ANOVA) se requiere adicionalmente que la varianza de la variable sea homogénea en las poblaciones a las que corresponden los distintos grupos considerados (igualdad de varianzas). Aunque, al igual que en el caso anterior, se dispone de pruebas para la comparación de varianzas entre dos o más grupos, su utilización cuando el número de datos es escaso está igualmente en entredicho. Por otro lado, aunque el número de datos disponible hiciera posible su utilización, podría concluirse, como resultado de la prueba de igualdad de varianzas, una desigualdad entre las mismas, situando al investigador en una posición complicada si no se dispone de alternativas de análisis.

Por otra parte, la mayoría de las pruebas paramétricas parten de la base de que la variable objeto de estudio es una variable cuantitativa continua (en escala de razón o intervalo). Sin embargo, en multitud de ocasiones las variables con las que se trabaja toman un número muy concreto y reducido de valores numéricos (variables discretas con estrecho rango de valores posibles) o son variables cualitativas ordinales (no toman valores numéricos pero pueden ser clasificadas en un número determinado de categorías ordenadas).

En estos casos ¿es posible plantear pruebas de hipótesis que permitan concluir diferencias significativas con respecto a un valor poblacional o entre diferentes grupos considerados? Es más, ¿tienen las pruebas paramétricas habituales una alternativa que no precise de los requisitos relativos a sus parámetros o distribución de la variable de estudio?

Las pruebas no paramétricas proporcionan un conjunto de recursos para el análisis de datos que responden a las situaciones planteadas con anterioridad. Aunque desde un punto de vista formal habría que distinguir entre las *pruebas no paramétricas,* que serían aquellas que no plantean hipótesis sobre determinados parámetros poblacionales (como puede ser el caso de las pruebas de bondad de ajuste) y las *pruebas de libre distribución,* que serían aquellas que no requieren conocer la distribución de la variable en la población objeto de estudio o no realizan ningún tipo de suposición sobre la misma, tradicionalmente todo este tipo de pruebas se engloba bajo la denominación de *pruebas no paramétricas.* Si las pruebas no paramétricas no requieren suposiciones acerca de la distribución de la variable, ¿en qué basan su funcionamiento?

Hay que reconocer que las pruebas no paramétricas se basan en ideas muy originales, por no calificarlas de brillantes, que involucran, en multitud

de ocasiones, a la distribución binomial y el «rango» o posición que ocupa cada uno de los datos una vez ordenados de menor a mayor. Adicionalmente, y como consecuencia de este tipo de estrategias, las hipótesis se basan, habitualmente, en la mediana de los datos en lugar de la media. Para ilustrar el funcionamiento de este tipo de pruebas se proponen a continuación dos ejemplos (ejemplos 3-1 y 3-2) basados en las situaciones introducidas al principio del presente capítulo.

Ejemplo 3-1

Se dispone de un total de 10 observaciones del nivel de ácido úrico (en mg/dl) de un conjunto de individuos seleccionados al azar de una población. Los niveles observados son los siguientes:

$$3,5; 4,6; 4,2; 6,1; 5,4; 5,1; 3,8; 5,8; 4,7; 5,3$$

¿Existe evidencia de que el nivel promedio de ácido úrico poblacional sea superior a 5,2 mg/dl?

PRUEBA DE LA MEDIANA

A partir de los datos del ejemplo 3-1, en primer lugar se procede a ordenar los datos observados de menor a mayor valor, con lo que quedarán de la siguiente forma:

$$3,5;\ 3,8;\ 4,2;\ 4,6;\ 4,7;\ 5,1;\ 5,3;\ 5,4;\ 5,8;\ 6,1$$

Se plantea el siguiente contraste de hipótesis:

$$H_0 : Md = 5,2$$

$$H_1 : Md \neq 5,2$$

Si la mediana poblacional fuera efectivamente 5,2 la probabilidad de que un individuo de dicha población tuviera un nivel de ácido úrico inferior a 5,2 sería igual a la probabilidad de que tuviera un nivel superior a 5,2. (Téngase en cuenta que la mediana se sitúa en el centro de la distribución ordenada de datos, garantizando que el número de observaciones a su izquierda será el mismo que el número de observaciones a su derecha.)

En el caso del ejemplo donde se dispone de una muestra de 10 individuos seleccionados al azar de una población, si se construye la variable aleatoria:

X = número de observaciones con un nivel de ácido úrico por encima de 5,2

esta se distribuirá, bajo la hipótesis de que la mediana es 5,2, según un modelo binomial con N = 10 y p = 0,5. Es decir:

$$P(X=k)=\binom{10}{k}0,5^{k}(1-0,5)^{10-k}$$

Para calcular el valor de la *p* del contraste de hipótesis planteado habrá que averiguar el número de observaciones en la muestra (k_0) con un nivel de ácido úrico por encima de 5,2 y calcular la siguiente probabilidad que representa la situación igual o más extrema que la observada:

$$P(X\leq k_0)$$

A continuación será necesario multiplicar esta cantidad por 2 para obtener el valor de la *p* para el contraste bilateral. En el ejemplo se detectan cuatro individuos con un nivel de ácido úrico superior a 5,2 (sus niveles son 5,3; 5,4; 5,8; 6,1). Por tanto, $k_0 = 4$ y se tendrá que:

$$P(X\leq 4)=P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4)$$

donde:

$$P(X=0)=\binom{10}{0}0,5^{0}(1-0,5)^{10-0}=0,000977$$

$$P(X=1)=\binom{10}{1}0,5^{1}(1-0,5)^{10-1}=0,009766$$

$$P(X=2)=\binom{10}{2}0,5^{2}(1-0,5)^{10-2}=0,043945$$

$$P(X=3)=\binom{10}{3}0,5^{3}(1-0,5)^{10-3}=0,117187$$

$$P(X=4)=\binom{10}{4}0,5^{4}(1-0,5)^{10-4}=0,205078$$

$$P(X\leq 4)=0,000977+0,009766+0,043945+0,117187+0,205078=0,37695$$

Por tanto, el valor de la *p* del contraste bilateral quedará:

$$p=2\cdot P(X\leq 4)=2\cdot 0,37695=0,7539$$

Con este valor de la *p* del contraste no es posible rechazar la hipótesis nula al nivel habitual $\alpha = 0,05$.

¿Qué hubiera ocurrido si la pregunta del investigador planteara la existencia o no de evidencia de que el promedio poblacional de ácido úrico fuera inferior a 5,9?

En este caso, el contraste de hipótesis sería:

$$H_0 : Md = 5,9$$

$$H_1 : Md \neq 5,9$$

Procediendo del mismo modo que en el caso anterior se tendría que $k_0 = 1$ (solo un individuo presenta un nivel de ácido úrico por encima de 5,9). El valor de la *p* del contraste bilateral vendría dado por:

$$p = 2 \cdot P(X \leq 1) = 2 \cdot [P(X = 0) + P(X = 1)] = 2(0,000977 + 0,009766) = 0,02148$$

Como 0,02148 es inferior a $\alpha = 0,05$, en este caso podría concluirse que la mediana poblacional de nivel de ácido úrico es significativamente distinta de 5,9. Como la mediana calculada a partir de los datos de la muestra es inferior a 5,9 (toma un valor entre 4,7 y 5,1) podría concluirse que la mediana poblacional es significativamente inferior a 5,9.

Ejemplo 3-2

¿Existen diferencias en los niveles promedio de ácido úrico poblacionales entre dos grupos de individuos? Para responder a esta cuestión se distinguirá entre el caso en que las dos muestras disponibles son independientes y el caso en que las muestras seleccionadas son apareadas o relacionadas. Por otra parte, las soluciones que se proporcionarán constituirán una primera aproximación al problema, disponiendo de soluciones más elaboradas, que serán estudiadas con posterioridad.

COMPARACIÓN DE DOS MEDIAS (MEDIANAS) EN DOS MUESTRAS RELACIONADAS (PRUEBA DEL SIGNO)

A partir del enunciado del ejemplo 3-2 supóngase que se dispone de información sobre 10 individuos seleccionados aleatoriamente de una población a los que se les administró un determinado tratamiento. Por cada uno de estos individuos se seleccionó, al azar, otro individuo de la misma edad y sexo que se asignó al grupo control. Los datos se reflejan en la tabla 3-1.

TABLA 3-1 Nivel de ácido úrico en dos grupos de tratamiento. Muestras relacionadas

Grupo control	Grupo tratamiento
5,2	3,5
5,8	4,6
4,4	4,2
6	6,1
6,8	5,4
5,3	5,1
5,1	3,8
3,9	5,8
4,8	4,7
5,5	5,3

Para el estudio de las posibles diferencias en los promedios de nivel de ácido úrico entre los dos grupos, en primer lugar se plantean las hipótesis del contraste, que quedarán de la siguiente forma:

$$H_0 : Md_{trat} = Md_{control}$$

$$H_1 : Md_{trat} \neq Md_{control}$$

A continuación, para cada una de las parejas de datos se obtendrá el signo de la diferencia entre el nivel de ácido úrico perteneciente al individuo del grupo de tratamiento y el nivel de ácido úrico de su pareja en el grupo control, de la siguiente forma:

$$\text{signo(diferencia)} = \text{signo}(x_{control} - x_{trat})$$

Los resultados se muestran en la tabla 3-2.

TABLA 3-2 Diferencias por parejas del nivel de ácido úrico

Grupo control	Grupo tratamiento	Diferencia = $x_{control} - x_{trat}$	Signo
5,2	3,5	5,2 − 3,5 = 1,7	+
5,8	4,6	5,8 − 4,6 = 1,2	+
4,4	4,2	4,4 − 4,2 = 0,2	+
6	6,1	6 − 6,1 = −0,1	−
6,8	5,4	6,8 − 5,4 = 1,4	+
5,3	5,1	5,3 − 5,1 = 0,2	+
5,1	3,8	5,1 − 3,8 = 1,3	+
3,9	5,8	3,9 − 5,8 = −1,9	−
4,8	4,7	4,8 − 4,7 = 0,1	+
5,5	5,3	5,5 − 5,3 = 0,2	+

Si las medianas en los dos grupos son iguales debería observarse el mismo número de signos positivos que negativos para las diferencias por parejas. Si se construye la variable aleatoria:

$$X = \text{número de signos positivos para la diferencia}$$

esta se distribuirá, bajo la hipótesis nula, según un modelo binomial con n = 10 y p = 0,5, es decir:

$$P(X=k)=\binom{10}{k}0,5^{k}(1-0,5)^{10-k}$$

Para calcular el valor de la *p* del contraste de hipótesis planteado habrá que averiguar el número de parejas de observaciones en la muestra (k_0) con un signo positivo para la diferencia en los niveles de ácido úrico correspondiente y calcular la siguiente probabilidad que representa la situación igual o más extrema que la observada:

$$P(X \geq k_0)$$

A continuación será necesario multiplicar esta cantidad por 2 para obtener el valor de la *p* para el contraste bilateral. En el ejemplo se detectan ocho parejas de individuos con un signo positivo para la diferencia en los niveles de ácido úrico (sus diferencias son: 1,7; 1,2; 0,2; 1,4; 0,2; 1,3; 0,1; 0,2). Por tanto, $k_0 = 8$ y se tendrá que:

$$P(X \geq 8) = P(X=8) + P(X=9) + P(X=10)$$

donde:

$$P(X=8)=\binom{10}{8}0,5^{8}(1-0,5)^{10-8}=0,043945$$

$$P(X=9)=\binom{10}{9}0,5^{9}(1-0,5)^{10-9}=0,009766$$

$$P(X=10)=\binom{10}{10}0,5^{10}(1-0,5)^{10-10}=0,000977$$

$$P(X \geq 8) = 0,043945 + 0,009766 + 0,000977 = 0,05469$$

Por tanto, el valor de la p del contraste bilateral quedará:

$$p = 2 \cdot P(X \geq 8) = 2 \cdot 0,05469 = 0,10938$$

Con este valor de la *p* del contraste no es posible rechazar la hipótesis nula al nivel habitual $\alpha = 0,05$. ¿Qué hubiera tenido que ocurrir para detectar significación estadística en este caso?

Si el número de diferencias entre los niveles de ácido úrico con signo positivo hubiera sido igual 9, entonces:

$$P(X \geq 9) = P(X = 9) + P(X = 10) = 0,009766 + 0,000977 = 0,010743$$

Y el valor de la *p* del contraste sería:

$$p = 2 \cdot P(X \geq 9) = 2 \cdot 0,010743 = 0,021486$$

Como el valor de la *p* sería de 0,021486, inferior al nivel de significación habitual $\alpha = 0,05$, se rechazaría la hipótesis de igualdad de medianas concluyéndose una diferencia significativa entre ellas.

COMPARACIÓN DE DOS MEDIAS (MEDIANAS) EN DOS MUESTRAS INDEPENDIENTES (PRUEBA DE LA MEDIANA)

A partir del enunciado del ejemplo 3-2, supóngase que se dispone de dos muestras aleatorias independientes de individuos de la población de tamaño 14 y 10 respectivamente. Los niveles de ácido úrico observados se muestran en la tabla 3-3.

Para el estudio de las posibles diferencias en los promedios de nivel de ácido úrico entre los dos grupos, en primer lugar, se plantean las hipótesis del contraste, que quedarán de la siguiente forma:

$$H_0 : Md_{\text{grupo 1}} = Md_{\text{grupo 2}}$$

$$H_1 : Md_{\text{grupo 1}} \neq Md_{\text{grupo 2}}$$

Si los dos grupos provienen de poblaciones con la misma mediana de ácido úrico existirá una única mediana global. A continuación se calcula la mediana global de los datos observados (incluyendo los datos de los dos grupos conjuntamente). Para ello se ordenan todos los datos de menor a mayor. Así se tendrá que:

3,5; 3,8; 3,9; 4,2; 4,4; 4,6; 4,7; 4,8; 4,9; 5,1; 5,1; 5,2; 5,3; 5,3; 5,4; 5,4; 5,5; 5,5; 5,8; 5,8; 6; 6,1; 6,2; 6,8

TABLA 3-3 Nivel de ácido úrico en dos grupos de tratamiento. Muestras independientes

Grupo	Nivel de ácido úrico
1	5,2; 5,8; 4,4; 6; 6,8; 5,3; 5,1; 3,9; 4,8; 5,5; 5,4; 4,9; 6,2; 5,5
2	3,5; 4,6; 4,2; 6,1; 5,4; 5,1; 3,8; 5,8; 4,7; 5,3

La posición que ocupará la mediana vendrá dada por el rango de la mediana que se calcula de la forma siguiente:

$$r_{Md} = \frac{n+1}{2} = \frac{24+1}{2} = 12,5$$

La mediana global se situará, por tanto, entre el dato que ocupa la posición 12 y el dato que ocupa la posición 13, que son 5,2 y 5,3. Para obtener un valor concreto para la mediana se calcula la semisuma entre los dos valores:

$$Md = \frac{5,2+5,3}{2} = 5,25$$

Si los datos provienen de dos poblaciones con la misma mediana, el número de datos que quedan, por ejemplo, por encima de la mediana global debería ser el mismo en los dos grupos. A continuación se muestran en negrita los datos que superan el valor Md = 5,25 en cada uno de los grupos.

Grupo 1: 3,9; 4,4; 4,8; 4,9; 5,1; 5,2; **5,3; 5,4; 5,5; 5,5; 5,8; 6; 6,2; 6,8**
Grupo 2: 3,5; 3,8; 4,2; 4,6; 4,7; 5,1; **5,3; 5,4; 5,8; 6,1**

En la tabla 3-4 se muestra el recuento de datos que quedan por encima y por debajo de la mediana global en cada uno de los grupos.

En la tabla 3-5 se recogen los valores esperados bajo la hipótesis de igualdad de proporciones (que equivale en este caso a la hipótesis de igualdad de medianas) en los dos grupos.

Como puede observarse los valores esperados son mayores o iguales a 5 en todas las casillas, por lo que puede utilizarse la prueba ji-cuadrado para valorar si las diferencias observadas respecto a la hipótesis de igualdad de medianas son significativas. El estadístico ji-cuadrado se obtendrá de la siguiente forma:

$$EC = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(8-7)^2}{7} + \frac{(6-7)^2}{7} + \frac{(4-5)^2}{5} + \frac{(6-5)^2}{5} = 0,68571$$

TABLA 3-4 Valores observados por encima y por debajo de la Md = 5,25

	Por encima de la Md	Por debajo de la Md	Total
Grupo 1	8	6	14
Grupo 2	4	6	10

TABLA 3-5 Valores esperados por encima y por debajo de la Md = 5,25

	Por encima de la Md	Por debajo de la Md	Total
Grupo 1	7	7	14
Grupo 2	5	5	10

Que se distribuirá según un modelo de probabilidad ji-cuadrado con 1 grado de libertad y que, para el estadístico calculado proporcionará un valor de la *p* del contraste de 0,40763, mayor que el nivel habitual $\alpha = 0,05$. En consecuencia no existiría evidencia de que las medianas de ácido úrico entre los dos grupos fueran significativamente distintas.

REFLEXIONES SOBRE LAS PRUEBAS NO PARAMÉTRICAS

Las pruebas no paramétricas del signo y de la mediana analizadas en los ejemplos anteriores permiten la extracción de conclusiones basadas en la mediana de uno o dos grupos de población y no en sus medias, siendo esta una de las principales consecuencias de la aplicación de este tipo de técnicas. Aunque debe tenerse en cuenta que bajo la suposición de simetría en la distribución de los datos en la población correspondiente la media coincidirá con la mediana, es igualmente útil, a los propósitos del investigador, utilizar una medida de tendencia central como la mediana en lugar de la media para llegar a las conclusiones deseadas sobre el comportamiento de los datos. Es importante señalar que en los ejemplos analizados no ha sido necesario suponer una determinada distribución de los datos en la población correspondiente y que la técnica hubiera podido ser empleada incluso con datos cualitativos ordinales.

Sin embargo, no todo son ventajas en la aplicación de este tipo de procedimientos. Las pruebas no paramétricas requerirán, habitualmente, que los datos observados muestren mayores diferencias con respecto a la hipótesis nula para que esta pueda ser rechazada, es decir, será más difícil detectar significación estadística que si hubiera podido utilizarse la prueba paramétrica correspondiente.

Por otra parte, si bien gran parte de los procedimientos basados en pruebas no paramétricas desaprovecharán información sobre los datos observados, en este caso ambas pruebas adolecen de una pérdida excesiva de información ya que el procedimiento únicamente ha tenido en cuenta si cada una de las observaciones se sitúa por encima o por debajo de la mediana hipotética.

La cuestión es: ¿es posible construir procedimientos de análisis no paramétricos en los que la pérdida de información sobre los datos disponibles sea menor?

Evidentemente, la respuesta es afirmativa, y la clave en la construcción de este tipo de pruebas reside en la consideración de la posición que ocupa cada dato (llamada tradicionalmente «rango») y su inclusión en los cálculos correspondientes de forma individualizada.

Para ilustrar este tipo de técnicas se considerarán a continuación los dos casos analizados con anterioridad para la comparación del nivel promedio

de ácido úrico en dos grupos de población, tanto para muestras relacionadas como independientes.

COMPARACIÓN DE DOS MEDIAS (MEDIANAS) EN DOS MUESTRAS RELACIONADAS (PRUEBA DE LOS RANGOS CON SIGNO DE WILCOXON)

Utilizando los datos correspondientes al ejemplo 3-2, en su versión de muestras relacionadas (datos de la tabla 3-1), se propone ahora un método que incorpore información sobre la posición de los datos o rango.

En primer lugar se calculan las diferencias entre los valores de ácido úrico en cada una de las parejas de datos. Como lo que interesa es la magnitud de la diferencia, se extrae el valor absoluto de dicha cantidad (para expresarla en positivo) y se asigna el *rango* o posición que ocuparía dicha diferencia si se ordenaran todas de menor a mayor. En el caso de que se produzca algún empate se asignará la media de los rangos. Los resultados de dicho procedimiento se muestran en la tabla 3-6.

Para el estudio de las posibles diferencias en los promedios (en este caso medianas) de nivel de ácido úrico entre los dos grupos, en primer lugar, se plantean las hipótesis del contraste, que quedarán de la siguiente forma:

$$H_0 : Md_{diferencia} = 0$$

$$H_1 : Md_{diferencia} \neq 0$$

Los rangos asignables a las diferencias son 1, 2, 3, 4…, 10, ya que se dispone de 10 diferencias calculadas (en valor absoluto). Obsérvese que existen dos diferencias de la misma magnitud (0,1) que es la más baja y, por tanto,

TABLA 3-6 Rangos de las diferencias de nivel de ácido úrico por parejas

Grupo control	Grupo tratamiento	$d_i = x_{control} - x_{trat}$	$\lvert d_i \rvert$	Rango = $r(\lvert d_i \rvert)$
5,2	3,5	5,2 − 3,5 = 1,7	1,7	9
5,8	4,6	5,8 − 4,6 = 1,2	1,2	6
4,4	4,2	4,4 − 4,2 = 0,2	0,2	4
6	6,1	6 − 6,1 = −0,1	0,1	1,5
6,8	5,4	6,8 − 5,4 = 1,4	1,4	8
5,3	5,1	5,3 − 5,1 = 0,2	0,2	4
5,1	3,8	5,1 − 3,8 = 1,3	1,3	7
3,9	5,8	3,9 − 5,8 = −1,9	1,9	10
4,8	4,7	4,8 − 4,7 = 0,1	0,1	1,5
5,5	5,3	5,5 − 5,3 = 0,2	0,2	4

les correspondería los dos primeros rangos 1 y 2. Como hay un empate, se asigna a cada diferencia el rango 1,5 (media de los rangos 1 y 2). De igual modo, en el caso del valor 0,2 se produce un triple empate. Dado que les correspondería los rangos 3, 4 y 5, se les asigna a los tres la media de los mismos, es decir, 4.

Para que la mediana de las diferencias sea 0, debería verificarse que la suma de los rangos correspondientes a las diferencias originalmente positivas sea igual a la suma de los rangos correspondientes a las diferencias negativas. Se calcula el estadístico basado exclusivamente en la suma de los rangos procedentes de las diferencias d_i originalmente positivas, que puede expresarse de forma general de la siguiente forma:

$$T^+ = \sum sig^+(d_i) \cdot r(|d_i|)$$

donde:

$$sig^+(d_i) = 1 \text{ si } d_i > 0$$

$$sig^+(d_i) = 0 \text{ si } d_i < 0$$

Con los datos del ejemplo se tendrá que:

$$W^+ = \sum sig^+(d_i) \cdot r(|d_i|) = 9+6+4+8+4+7+1,5+4 = 43,5$$

Si se calcula ahora el estadístico basado exclusivamente en la suma de los rangos procedentes de las diferencias d_i originalmente negativas, se tendrá que:

$$W^- = \sum sig^-(d_i) \cdot r(|d_i|) = 1,5+10 = 11,5$$

Como puede observarse, existen diferencias importantes entre la suma de los rangos positivos y la suma de los rangos negativos. Cuanto más grande sea la diferencia, mayor evidencia existirá en contra de la hipótesis nula que presupone que la mediana de las diferencias es 0. Téngase en cuenta que la suma de una progresión aritmética desde 1 hasta n, como es el caso de los rangos, puede calcularse de la forma:

$$1+2+3+4+\ldots+n = \frac{n(n+1)}{2}$$

Para que $W^+ = W^-$ tendría que ocurrir que:

$$W^+ = W^- = \frac{\left(\frac{n(n+1)}{2}\right)}{2} = \frac{n(n+1)}{4} = \frac{10(10+1)}{4} = 27,5$$

Parece, a la vista de los resultados, que ambos grupos (los positivos y los negativos) se alejan de ese valor común para la suma de rangos que debería ser 27,5.

Para el cálculo de la *p* exacta del contraste será necesario contemplar todas las disposiciones posibles de los datos y la suma de los rangos asociados, por ejemplo, a las diferencias positivas. Se obtendrá, por tanto, un valor de la suma de los rangos asociados a las cantidades positivas para cada posible disposición de los datos. Esto incluye todas las posibles permutaciones de las disposiciones de los mismos. En el caso del ejemplo, aunque parece que el número de parejas es pequeño (n = 10), el número de disposiciones posibles de los datos es muy elevado. Para su cálculo es necesario recurrir a tablas específicas o a los resultados proporcionados por diferentes programas de análisis estadístico. En este caso el valor de la *p* del contraste bilateral es de 0,11328, superior al nivel habitual $\alpha = 0{,}05$ y, por tanto, no podría rechazarse la hipótesis nula de igualdad de medianas de ácido úrico entre los dos grupos.

Para ilustrar el cálculo de la *p* exacta, basado en las múltiples disposiciones de los datos, se propone el siguiente ejemplo ilustrativo (tabla 3-7) con un número muy reducido de parejas, en concreto, n = 3.

Obsérvese que se parte de una situación extrema en la que, en todas las parejas, el nivel de ácido úrico es superior en el grupo control al del grupo tratamiento. Si se calcula W^+ se tendrá que:

$$W^+ = \sum \text{sig}^+(d_i) \cdot r(|d_i|) = 3+2+1 = 6$$

Por su parte, el valor de W^- será 0 porque no hay ningún rango que proceda de una diferencia originalmente negativa. La cuestión es: ¿cuántas disposiciones posibles de los datos podrían producirse a partir de estos resultados?

En la tabla 3-8 se muestran las posibles disposiciones de los rangos (repartidas en rangos de origen positivo o negativo) y la suma de los rangos originalmente positivos. Debe tenerse en cuenta que en cada una de estas disposiciones deben contemplarse las posibles permutaciones de los datos que mantendrían la misma disposición. Las posibles permutaciones entre tres elementos en cada una de las disposiciones se expresan como 3! = 3.2.1 = 6.

TABLA 3-7 Ejemplo ilustrativo con tamaño muestral reducido. Cálculo de la *p* exacta. Prueba de Wilcoxon

Grupo control	Grupo tratamiento	$d_i = x_{control} - x_{trat}$	$\lvert d_i \rvert$	Rango = $r(\lvert d_i \rvert)$
5,2	3,5	5,2 − 3,5 = 1,7	1,7	3
5,8	4,6	5,8 − 4,6 = 1,2	1,2	2
4,4	4,2	4,4 − 4,2 = 0,2	0,2	1

TABLA 3-8 Disposiciones posibles de los datos, valor de W^+ y número de permutaciones asociadas

Signo	+	−	+	−	+	−	+	−	+	−	+	−	+	−	+	−
Disposiciones y rangos	1		1		1		1			1		1		1		1
	2			2	2			2		2	2			2	2	
	3			3		3	3			3		3	3		3	
W^+	6		1		3		4		0		2		3		5	
Permutaciones	6		6		6		6		6		6		6		6	

TABLA 3-9 Distribución de frecuencias de los valores posibles de W^+

W^+	**Frecuencia**
0	6
1	6
2	6
3	12
4	6
5	6
6	6

En la tabla 3-9 se resume la distribución de frecuencias de la cantidad W^+. Puede observarse que existen 48 posibles disposiciones de los datos (incluidas las permutaciones asociadas). Para obtener el valor de la *p* exacta será necesario calcular la probabilidad de obtener un valor de W^+ igual o más extremo que el observado y, posteriormente, multiplicarlo por 2 para el contraste bilateral. En este caso, el valor observado de W^+ es 6, por lo que quedará de la siguiente forma:

$$p = 2 \cdot P\left(W^+ \geq 6\right) = 2 \cdot P\left(W^+ = 6\right) = 2 \cdot \frac{6}{48} = \frac{12}{48} = 0,25$$

APROXIMACIÓN ASINTÓTICA EN EL CASO DE MUESTRAS GRANDES

En el caso de que el número de datos analizados sea lo suficientemente elevado, puede utilizarse la aproximación normal basada en la aproximación asintótica del estadístico correspondiente. El número mínimo de datos que se requiere para su utilización varía según diferentes autores y oscila, en este caso, entre 20 y 30 parejas de datos.

$$Z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Donde el valor de W es el mínimo de los valores W^+ y W^-, que se distribuirá asintóticamente según un modelo normal de media 0 y desviación típica 1.

Aunque el número de datos disponibles en el ejemplo 3-2 (situación de muestras relacionadas) y en el ejemplo ilustrativo es inferior al recomendado se procede al cálculo de estadístico Z a efectos didácticos. Dado que el mínimo de los valores W^+ y W^- para los datos de la tabla 3-1 es 11,5 ($W^+ = 43{,}5$ y $W^- = 11{,}5$), se tendrá que:

$$Z = \frac{W^- - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{11{,}5 - \frac{10(10+1)}{4}}{\sqrt{\frac{10(10+1)(2\cdot 10+1)}{24}}} = \frac{-16}{9{,}81} = -1{,}63$$

En el caso del ejemplo ilustrativo (datos de la tabla 3-7) el valor será:

$$Z = \frac{W^- - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{0 - \frac{3(3+1)}{4}}{\sqrt{\frac{3(3+1)(2\cdot 3+1)}{24}}} = \frac{-3}{1{,}87} = -1{,}6$$

En las tablas 3-10 y 3-11 se muestran los resultados del análisis de los datos del ejemplo 3-2 (muestras relacionadas incluyendo las 10 parejas de datos), obtenido mediante el programa SPSS, mientras que en las tablas 3-12 y 3-13 se presentan los resultados correspondientes a los datos del ejemplo ilustrativo (incluyendo únicamente 3 parejas de datos).

Obsérvese cómo los resultados coinciden con los obtenidos manualmente y cómo se reproducen en la salida los elementos básicos utilizados en el análisis. Adviértase adicionalmente la diferencia del valor de la *p* exacta y el calculado mediante la aproximación asintótica que es mucho más acusada en el caso del ejemplo ilustrativo dado el escaso número de datos y que refuerza la no aplicación de la aproximación para muestras muy pequeñas.

TABLA 3-10 Valores de W^+ y W^- y rango promedio para los datos de la tabla 3-1

		N	Rango promedio	Suma de rangos
Control − tratamiento	Rangos negativos	8[a]	5,44	43,5
	Rangos positivos	2[b]	5,75	11,5
	Empates	0[c]		
	Total	10		

[a] *Control < tratamiento.*
[b] *Control > tratamiento.*
[c] *Control = tratamiento.*

TABLA 3-11 Prueba de Wilcoxon para los datos de la tabla 3-1. Estadísticos de contraste[a]

	Control − tratamiento
Z	−1,636[b]
Sig. asintótica (bilateral)	,102
Sig. exacta (bilateral)	,113
Sig. exacta (unilateral)	,057
Probabilidad en el punto	,008

[a] *Prueba de los rangos con signo de Wilcoxon.*
[b] *Basado en los rangos positivos.*

TABLA 3-12 Valores de W^+ y W^- y rango promedio para los datos del ejemplo ilustrativo de la tabla 3-7

		N	Rango promedio	Suma de rangos
Control − Tratamiento	Rangos negativos	0[a]	,00	,00
	Rangos positivos	3[b]	2	6
	Empates	0[c]		
	Total	3		

[a] *Control < tratamiento.*
[b] *Control > tratamiento.*
[c] *Control = tratamiento.*

TABLA 3-13 Prueba de Wilcoxon para los datos de la tabla 3-7. Estadísticos de contraste[a]

	Grupo 2 − Grupo 1
Z	–1,604[b]
Sig. asintótica (bilateral)	,109
Sig. exacta (bilateral)	,25
Sig. exacta (unilateral)	,125
Probabilidad en el punto	,125

[a] *Prueba de los rangos con signo de Wilcoxon.*
[b] *Basado en los rangos positivos.*

COMPARACIÓN DE DOS MEDIAS (MEDIANAS) EN DOS MUESTRAS INDEPENDIENTES (PRUEBA U DE MANN-WHITNEY-WILCOXON)

Utilizando los datos correspondientes al ejemplo 3-2 en su versión de muestras independientes (datos de la tabla 3-3) estudiado con anterioridad, se pretende construir una prueba que tenga en cuenta la posición de los datos.

En primer lugar se ordenan todos los datos (incluyendo los correspondientes a los dos grupos) de menor a mayor y se asigna a cada uno de ellos

TABLA 3-14 Observaciones y rangos ordenados según el grupo de tratamiento. Muestras independientes

Observaciones grupo 1	Observaciones grupo 2	Rangos grupo 1	Rangos grupo 2
	3,5		1
	3,8		2
3,9		3	
	4,2		4
4,4		5	
	4,6		6
	4,7		7
4,8		8	
4,9		9	
5,1	5,1	10,5	10,5
5,2		12	
5,3	5,3	13,5	13,5
5,4	5,4	15,5	15,5
5,5		17,5	
5,5		17,5	
5,8	5,8	19,5	19,5
6		21	
	6,1		22
6,2		23	
6,8		24	
Suma de rangos		199	101

el rango o posición que ocupa. En el caso de existir empates se procede de forma análoga al caso anterior, asignando la media de los rangos a cada uno de ellos. Una vez obtenidos los rangos se procede a la suma de los mismos en cada uno de los dos grupos. Los resultados de dicho procedimiento se muestran la tabla 3-14.

Para el estudio de las posibles diferencias en los promedios de nivel de ácido úrico entre los dos grupos, en primer lugar, se plantean las hipótesis del contraste que quedarán de la siguiente forma:

$$H_0 : Md_{\text{grupo 1}} = Md_{\text{grupo 2}}$$

$$H_1 : Md_{\text{grupo 1}} \neq Md_{\text{grupo 2}}$$

En principio, si las medianas de los dos grupos fueran iguales, entonces la suma de los rangos correspondientes al primer grupo coincidiría con la suma de los rangos de los datos del segundo grupo. Sin embargo, esto solo sería cierto si los dos grupos tuvieran el mismo número de observaciones. Apréciese que si un grupo tiene un mayor número

de observaciones es lógico que la suma de los rangos tienda a ser mayor al sumar un número mayor de rangos asociados. Sin embargo, puede corregirse este efecto, calculando la media de los rangos en cada uno de los grupos (dividiendo la suma de rangos por el número de observaciones en cada grupo).

En este caso las medias de los rangos, o «rangos medios», quedarían de la siguiente forma:

$$\text{Rango medio grupo } 1 = {}^{199}\!/_{14} = 14,21$$

$$\text{Rango medio grupo } 2 = {}^{101}\!/_{10} = 10,1$$

Como puede observarse, el rango medio del grupo 1 es sensiblemente mayor al rango medio del grupo 2, indicando que los datos correspondientes al grupo 1 tienden a ocupar posiciones más elevadas en el conjunto de datos ordenados y, por tanto, corresponden a valores de ácido úrico que, en global, son superiores a los del grupo 2. Esta información a nivel descriptivo muestra hacia dónde se dirigen las diferencias observadas en el caso de que fueran significativas.

Para poder contrastar la hipótesis de igualdad de medianas de ácido úrico entre los dos grupos será necesario reflexionar sobre las situaciones de las que se desprendería evidencia en contra de dicha hipótesis. En este sentido, ¿cuál sería la disposición de los datos más extrema en contra de la hipótesis de igualdad de medianas?

La situación más extrema, en el mismo sentido que la observada, implicaría que todos los individuos correspondientes al grupo 2 tuvieran un nivel de ácido úrico inferior a cualquiera de los pertenecientes al grupo 1. Así, una vez ordenados todos los datos conjuntamente, las primeras 10 posiciones serían ocupadas por los datos del grupo 2 y las siguientes 14, por las del grupo 1. (Téngase en cuenta que se procedería de forma equivalente en la situación contraria en la que todos los individuos del grupo 1 tuvieran un nivel de ácido úrico inferior a los del grupo 2, si los datos hubieran mostrado esta tendencia.)

La hipotética tabla de asignación de rangos tendría un aspecto similar a la disposición recogida en la tabla 3-15. En consecuencia, cuanto más cerca esté la disposición real de los datos observados de la descrita en esta tabla, existirá una mayor evidencia en contra de la hipótesis nula de igualdad de medianas. Por tanto, puede construirse un estadístico de contraste de la siguiente forma:

$$U = S_2 - (1+2+3+\ldots+10) = S_2 - \frac{n_2(n_2+1)}{2}$$

TABLA 3-15 Disposición más extrema de los datos ordenados indicativa de diferencias entre los dos grupos

Observaciones grupo 1	Observaciones grupo 2	Rangos grupo 1	Rangos grupo 2
	X		1
	X		2
	X		3
	X		4
	.		.
	.		.
	X		10
X		11	
X		12	
X		13	
.		.	
.		.	
X		24	
Suma de rangos		$S_1 = 245$	$S_2 = 55$

Donde S_2 es la suma de los rangos correspondientes al grupo 2, y n_2 el tamaño de la muestra del grupo 2. Con los datos disponibles se tendrá que:

$$U = 101 - 55 = 46$$

Cuanto más pequeño sea el valor de la cantidad U más cerca se encontrarán los datos de la situación más extrema de diferencia entre los dos grupos. Al igual que en la prueba de Wilcoxon para muestras relacionadas, sería necesario contabilizar todas las posibles disposiciones de los datos (incluyendo las permutaciones correspondientes) para averiguar la probabilidad de haber obtenido una disposición igual o más extrema que la observada. Multiplicando por 2 dicha probabilidad, se obtendría el valor de la p del contraste bilateral.

$$p = 2 \cdot P(U \leq 46)$$

Debido a que la cantidad de disposiciones en este caso es elevadísima y que ha sido ilustrado el procedimiento en el caso anterior, se proporciona directamente el valor de la p, que en este caso es de 0,16704.

Dado que el valor 0,16704 es superior al nivel habitual $\alpha = 0,05$ no existiría evidencia de que las medianas de ácido úrico fueran significativamente distintas entre los dos grupos.

APROXIMACIÓN ASINTÓTICA PARA MUESTRAS GRANDES

En el caso de que el número de datos analizados sea lo suficientemente elevado, puede utilizarse la aproximación normal basada en la aproximación asintótica del estadístico correspondiente. El número mínimo de

datos que se requiere para su utilización varía según diferentes autores y oscila, en este caso, en un mínimo de entre 10 y 20 datos por grupo. El valor del estadístico utilizado mediante esta aproximación se expresará de la siguiente forma:

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

Que se distribuirá según un modelo normal de media 0 y desviación típica 1. En el ejemplo 3-2 (muestras independientes) el valor del estadístico de contraste basado en la aproximación asintótica quedará:

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = \frac{46 - \frac{14 \cdot 10}{2}}{\sqrt{\frac{14 \cdot 10(14+10+1)}{12}}} = \frac{-24}{17,078} = -1,4$$

Al igual que en el caso anterior, en las tablas 3-16 y 3-17 se muestran los resultados del análisis efectuado con el programa SPSS. Obsérvese que, en este caso, el valor de la *p* exacta (0,167) y el valor de la *p* proporcionado por la aproximación asintótica (0,159) son relativamente próximos.

TABLA 3-16 Valores de S_1 y S_2 y rango promedio para los datos de la tabla 3-3

	Grupo	N	Rango promedio	Suma de rangos
Ácido úrico	1	14	14,21	199
	2	10	10,1	101
	Total	24		

TABLA 3-17 Prueba U de Mann-Whitney. Estadísticos de contraste[a]

	Grupo 1
U de Mann-Whitney	46,000
W de Wilcoxon	101,000
Z	−1,407
Sig. asintótica (bilateral)	,159
Sig. exacta [2*(Sig. unilateral)]	,172[b]
Sig. exacta (bilateral)	,167
Sig. exacta (unilateral)	,083
Probabilidad en el punto	,004

[a] *Variable de agrupación: grupo.*
[b] *No corregidos para los empates.*

COMPARACIÓN DE TRES O MÁS MEDIAS (MEDIANAS) PARA MUESTRAS INDEPENDIENTES (PRUEBA DE KRUSKAL-WALLIS)

La prueba de Kruskal-Wallis está indicada para la comparación de una variable cuantitativa en tres o más grupos cuando las muestras son independientes.

Ejemplo 3-3

Se pretende comparar el promedio de edad de tres grupos de pacientes. Para ello, se cuenta con tres muestras independientes seleccionadas al azar, correspondientes a cada uno de los grupos, cuyos resultados se muestran en la tabla 3-18.

TABLA 3-18 Edad (en años) según grupo de pacientes

Grupo	Observaciones
1	25, 27, 18, 20, 23
2	42, 38, 35, 29, 32, 34, 31
3	28, 32, 25, 21, 26, 30

Para el estudio de las posibles diferencias en los promedios de edad (en este caso medianas) entre los tres grupos, en primer lugar se plantean las hipótesis del contraste, que quedarán de la siguiente forma:

$$H_0 : Md_1 = Md_2 = Md_3$$

$$H_1 : \text{al menos una Md es distinta}$$

La construcción del estadístico de contraste que permita contrastar la hipótesis planteada de igualdad de medianas constituye una extensión de la prueba de suma de rangos de Mann-Whitney-Wilcoxon para el caso de tres o más grupos y se basa, por tanto, en la suma de rangos correspondientes a cada uno de los grupos una vez ordenados de menor a mayor, tal y como se refleja en la tabla 3-19.

El rango medio de cada uno de los grupos analizados podrá calcularse de la siguiente forma:

TABLA 3-19 Observaciones y rangos ordenados según grupo de pacientes. Muestras independientes

Datos			Rangos		
Grupo 1	**Grupo 2**	**Grupo 3**	**Grupo 1**	**Grupo 2**	**Grupo 3**
18			1		
20			2		
		21			3
23			4		
25		25	5,5		5,5
		26			7
27			8		
		28			9
	29			10	
		30			11
	31			12	
	32	32		13,5	13,5
	34			15	
	35			16	
	38			17	
	42			18	
Suma de rangos			$S_1 = 20,5$	$S_1 = 101,5$	$S_1 = 49$

$$\text{Rango medio grupo } 1 = \bar{R}_1 = \frac{S_1}{n_1} = 20,5/5 = 4,1$$

$$\text{Rango medio grupo } 2 = \bar{R}_2 = \frac{S_2}{n_2} = 101,5/7 = 14,5$$

$$\text{Rango medio grupo } 3 = \bar{R}_3 = \frac{S_3}{n_3} = 49/6 = 8,17$$

Puede observarse que los rangos promedio muestran diferencias importantes entre los tres grupos considerados (mayores en el grupo 2, seguidos por el grupo 3 y, por último, el grupo 1). El rango medio global de los datos sin distinguir entre grupos se calcularía:

$$\text{Rango medio global} = \frac{\frac{n(n+1)}{2}}{n} = \frac{n+1}{2} = \frac{18+1}{2} = 9,5$$

El estadístico de Kruskal-Wallis se construirá a partir de las diferencias entre el rango medio observado de cada grupo y la media global. Así, se tendrá que:

$$KW = \sum n_i \left(\bar{R}_i - \frac{n+1}{2} \right)^2$$

Si los datos provienen de una población con idéntica mediana, entonces se espera que los rangos promedio de cada grupo coincidan con el rango promedio global y, entonces, el valor de KW se aproximaría a 0.

Con los datos del ejemplo se tendría que:

$$KW = \sum n_i \left(\bar{R}_i - \frac{n+1}{2} \right)^2 = 5(4,1-9,5)^2 + 7(14,5-9,5)^2 + 6(8,17-9,5)^2$$
$$= 145,8 + 175 + 14,11 = 306,68$$

Al igual que en pruebas anteriores, para obtener el valor de la *p* exacta habría que realizar todas las disposiciones y permutaciones posibles de los *n* datos y calcular la probabilidad de obtener un valor igual o más extremo que el observado. En este caso:

$$p = P(KW \geq 306,68) = 0,000316$$

Como el valor de la *p* exacta es inferior al nivel habitual $\alpha = 0,05$, se rechazaría la hipótesis de igualdad de medianas, concluyéndose diferencias significativas entre las mismas.

Una expresión alternativa, y más ampliamente conocida, para el estadístico de Kruskal-Wallis se obtendría procediendo con los rangos promedio de forma similar a las medias en el caso del ANOVA, para construir un estadístico basado en la comparación de la variabilidad entre grupos dividido por la variabilidad total. Así, se tendrá que:

$$H = (n-1) \frac{\sum n_i \left(\bar{R}_i - \frac{n+1}{2} \right)^2}{\sum \sum \left(\bar{R}_{ij} - \frac{n+1}{2} \right)^2}$$
$$= \frac{12}{n(n+1)} \sum n \left(\bar{R}_i - \frac{n+1}{2} \right)^2 = \frac{12}{n(n+1)} \sum \frac{S_i^2}{n_i} - 3(n+1)$$

APROXIMACIÓN ASINTÓTICA

El estadístico *H* de Kruskal Wallis se distribuirá asintóticamente según una distribución ji-cuadrado con $k - 1$ grados de libertad, donde k = número de grupos. El requisito necesario para esta aproximación difiere según diferentes autores aunque suele exigirse al menos cinco datos en cada uno de los grupos que se pretende comparar.

Este estadístico de contraste sufre una ligera corrección en el caso de que existan empates en los rangos asignados, apenas apreciable cuando

TABLA 3-20 Rangos promedio según grupo de pacientes

	Grupo	N	Rango promedio
Edad	1	5	4,1
	2	7	14,5
	3	6	8,17
	Total	18	

TABLA 3-21 Prueba de Kruskal-Wallis. Estadísticos de contraste

	Edad
Ji-cuadrado	11,654
gl	2
Sig. asintótica	,003
Sig. exacta	,000
Probabilidad en el punto	,000

gl, *grados de libertad*.

el número de empates es pequeño. A partir de los datos del ejemplo (si se ignoran los empates) se tendrá que:

$$H = \frac{12}{18(18+1)}\left(\frac{20,5^2}{5}+\frac{101,5^2}{7}+\frac{49^2}{6}\right)-3(18+1)=11,63$$

Este valor del estadístico de contraste sobre una distribución ji-cuadrado con 2 grados de libertad proporcionará un valor de la p del contraste de 0,00298.

En las tablas 3-20 y 3-21 se muestran los resultados obtenidos mediante el programa SPSS.

Obsérvese que los datos coinciden con los obtenidos manualmente, si bien existe una pequeña diferencia en el valor del estadístico ji-cuadrado (11,654) debida a la existencia de empates en los rangos, aunque sin apenas trascendencia en la significación estadística del contraste. Adicionalmente puede observarse una ligera diferencia en el valor de la p del contraste obtenido mediante la prueba exacta ($p < 0,001$) y la asintótica basada en la distribución ji-cuadrado ($p = 0,003$).

COMPARACIÓN DE TRES O MÁS MEDIAS (MEDIANAS) PARA MUESTRAS RELACIONADAS (PRUEBA DE LOS RANGOS DE FRIEDMAN)

La prueba de Friedman está indicada para la comparación de una variable cuantitativa en tres o más grupos cuando las muestras están relacionadas.

Ejemplo 3-4

Se pretende valorar la eficacia de un tratamiento sobre el nivel de colesterol. Se dispone de un grupo de seis individuos y de los niveles de colesterol (en mg/100 ml) medidos antes del tratamiento, a los 30 y a los 90 días del inicio del mismo. Los resultados obtenidos se muestran en la tabla 3-22.

Para el estudio de las posibles diferencias en los promedios de colesterol entre los tres grupos o instantes de tiempo, en primer lugar se plantean las hipótesis del contraste, que quedarán de la siguiente forma:

$$H_0: \text{No existen diferencias entre los grupos}$$

$$H_1: \text{Al menos un grupo muestra diferencias}$$

Se requiere que la variable sea cuantitativa continua, discreta u ordinal. Las muestras deben ser relacionadas.

La construcción del estadístico de contraste que permita contrastar la hipótesis planteada requiere asignar rangos en cada uno de los tríos de observaciones tal y como se describe en la tabla 3-23.

Si los datos provienen de una población en la que no existen diferencias entre los tres grupos de observaciones, entonces se espera que la suma de los

TABLA 3-22 Niveles de colesterol obtenidos en tres instantes de tiempo

Antes del tratamiento	A los 30 días	A los 90 días
283	268	242
245	250	210
320	270	234
275	265	220
345	315	285
290	240	221

TABLA 3-23 Asignación de rangos en cada uno de los tríos de observaciones

G_1 = antes del tratamiento	G_2 = a los 30 días	G_3 = a los 90 días	Rango G_1	Rango G_2	Rango G_3
283	268	242	3	2	1
245	250	210	2	3	1
320	270	234	3	2	1
275	265	270	3	1	2
345	315	285	3	2	1
290	240	221	3	2	1
Suma de rangos			$S_1 = 17$	$S_2 = 12$	$S_3 = 7$

rangos en cada uno de los grupos sea la misma. En este caso la suma total de los rangos de las observaciones sería:

$$\text{Suma total de rangos} = 6\cdot(1+2+3) = n\frac{k(k+1)}{2} = \frac{nk(k+1)}{2} = 36$$

Donde k = número de grupos y n el número total de tríos. Si no hubiera diferencia entre los grupos de observaciones a cada uno de los grupos debería corresponderle, en este caso, la tercera parte de la suma total de rangos. En general, a cada grupo le correspondería la suma total de rangos dividida entre el número de grupos, es decir:

$$\frac{\frac{nk(k+1)}{2}}{k} = \frac{n(k+1)}{2} = \frac{6(3+1)}{2} = 12$$

La base fundamental del estadístico de contraste tratará de cuantificar la diferencia entre el valor observado y el esperado bajo la hipótesis de no existencia de diferencias entre los grupos y, por tanto, se apoyará en la siguiente cantidad:

$$\sum\left(S_i - \frac{n(k+1)}{2}\right)^2$$

Al igual que en el caso de la prueba de Kruskal-Wallis, puede construirse una expresión basada en esta cantidad que se distribuirá asintóticamente según un modelo de distribución ji-cuadrado con $k - 1$ grados de libertad.

$$F = \frac{12}{nk(k+1)}\sum S_i^2 - 3n(k+1)$$

Si se trabaja con los datos del ejemplo, se tendrá que:

$$F = \frac{12}{6\cdot 3(3+1)}\left(17^2+12^2+7^2\right) - 3\cdot 6(3+1) = \frac{12}{72}\cdot 482 - 72 = 8,33$$

El valor de la *p* exacta para este estadístico de contraste es de 0,01196, con lo que se rechazaría la hipótesis nula al nivel habitual $\alpha = 0,05$ y concluyéndose una diferencia significativa en los niveles de colesterol entre los tres grupos (instantes de tiempo de tratamiento).

APROXIMACIÓN ASINTÓTICA

El valor de la p del contraste calculado a partir de la distribución asintótica (ji-cuadrado con $k - 1 = 3 - 1 = 2$ grados de libertad es de 0,016

TABLA 3-24 Rangos promedio

Grupo	Rango promedio
Antes del tratamiento	2,83
A los 30 días	2
A los 90 días	1,17

TABLA 3-25 Prueba de Friedman. Estadísticos de contraste

N	6
Ji-cuadrado	8,333
gl	2
Sig. asintótica	,016
Sig. exacta	,012
Probabilidad en el punto	,004

gl, *grados de libertad.*

llegándose a la misma conclusión. El número mínimo de datos para la utilización de esta aproximación se sitúa en torno a los 8. En las tablas 3-24 y 3-25 se proporcionan los resultados obtenidos mediante el programa SPSS.

Obsérvese que los resultados coinciden con los obtenidos manualmente, y la ligera diferencia en el valor de la *p* del contraste obtenido mediante la prueba exacta y la asintótica que afectan a la tercera cifra decimal. Por otra parte, el programa proporciona los rangos promedio en cada uno de los grupos obtenidos como la suma de los rangos en cada uno de los grupos divida entre los seis individuos (tríos) estudiados.

COMPARACIÓN DE DOS PROPORCIONES PARA MUESTRAS RELACIONADAS (PRUEBA DE McNEMAR)

La prueba de McNemar está indicada para la comparación de una variable cualitativa dicotómica en dos grupos cuando las muestras están relacionadas, es decir, compara dos proporciones en muestras relacionadas.

Ejemplo 3-5

Se pretende estudiar la efectividad de un programa de educación para la salud para concienciar a la población sobre la necesidad de realizar ejercicio físico con frecuencia. Para ello se seleccionó aleatoriamente un grupo de 80 individuos que cumplimentaron una encuesta antes y después del programa. De los 80 individuos incluidos en el estudio, 50 declararon no realizar ningún

tipo de actividad física antes del programa. Tras el programa de concienciación, de los 50 que no realizaban ejercicio, 28 declararon haber comenzado a realizar ejercicio físico de forma regular, mientras que de los 30 restantes, 5 habían dejado de realizar ejercicio. ¿Existe evidencia de que el programa haya mejorado los hábitos de realización de ejercicio físico?

Para responder a esta cuestión será necesario comparar las proporciones de realización de ejercicio físico antes y después del programa de educación para la salud pero teniendo en cuenta el cambio, si es que se produce, en cada uno de los individuos. Esta situación de muestras relacionadas requiere que los datos se expresen en una tabla que recoja las concordancias y discordancias en la respuesta del individuo antes y después del programa. Por ejemplo, hay un total de 50 individuos que no realizaban ejercicio antes del programa, de los cuales 22 (50-28) continuaron sin realizar ejercicio físico después (par concordante). Por otra parte, de los 30 que declararon realizar ejercicio físico antes del programa, 5 dejaron de realizarlo después del mismo (par discordante). En la tabla 3-26 se proporciona el recuento de las frecuencias en cada una de las cuatro posibilidades antes-después.

Se cuenta, por tanto, con 80 pares de datos, de los cuales 47 (25 + 22) son concordantes (coincide la respuesta antes-después) y 33 = (5 + 28) son discordantes (difiere la respuesta del individuo antes y después del programa).

El test de McNemar trataría de estudiar la diferencia en la práctica de ejercicio físico antes y después del programa a través, por ejemplo, de las proporciones de respuesta afirmativa a la práctica de ejercicio antes y después del programa. Las hipótesis del contraste podrían establecerse del siguiente modo:

$$H_0 : p_{antes} = p_{después}$$

$$H_1 : p_{antes} = p_{después}$$

Si se observan los datos de la tabla de pares concordantes y discordantes puede apreciarse que la diferencia en las proporciones vendrá determinada por los pares discordantes *b* y *c*, ya que:

$$\hat{p}_{antes} = (a+b)/(a+b+c+d) = (a+b)/n$$

$$\hat{p}_{después} = (a+c)/(a+d+c+d) = (a+c)/n$$

TABLA 3-26 Distribución de frecuencias de la variable «ejercicio físico» antes y después del programa. Pares concordantes y discordantes

Ejercicio antes del programa	Ejercicio después del programa		Total
	Sí	**No**	
Sí	25 *(a)*	5 *(b)*	30 *(a + b)*
No	28 *(c)*	22 *(d)*	50 *(c + d)*
			$n = a + b + c + d = 80$

Bajo la hipótesis de igualdad en las proporciones de respuesta afirmativa antes-después el total de pares discordantes debería distribuirse de forma equitativa entre el antes y el después del programa (para que b = c) y, en consecuencia, $\hat{p}_{antes} = \hat{p}_{después}$. Por tanto, la variable aleatoria:

X = número de pares discordantes (respuesta afirmativa-antes)

se distribuirá, bajo la hipótesis de igualdad, según un modelo binomial con N = b + c y p = 0,5.

En el ejemplo se tiene que N = b + c = 33 (número total de pares discordantes) y que b = 5. Bajo la hipótesis nula el valor de *b* debería haberse situado cerca de 33/2 = 16,5, con lo que el valor de la *p* exacta del contraste se calculará de la siguiente forma:

$$p = 2 \cdot P(X \leq 5) = 2 \cdot \left[P(X=0) + P(X=1) + \ldots + P(X=5) \right]$$

$$P(X=0) = \binom{33}{0} 0,5^0 (1-0,5)^{33-0} = 1,1641 \cdot 10^{-10}$$

$$P(X=1) = \binom{33}{1} 0,5^1 (1-0,5)^{33-1} = 3,8417 \cdot 10^{-9}$$

$$P(X=2) = \binom{33}{2} 0,5^2 (1-0,5)^{33-2} = 6,1467 \cdot 10^{-8}$$

$$P(X=3) = \binom{33}{3} 0,5^3 (1-0,5)^{33-3} = 6,3516 \cdot 10^{-7}$$

$$P(X=4) = \binom{33}{4} 0,5^4 (1-0,5)^{33-4} = 4,7637 \cdot 10^{-6}$$

$$P(X=5) = \binom{33}{5} 0,5^5 (1-0,5)^{33-5} = 2,7629 \cdot 10^{-5}$$

$$p = 2 \cdot \left[1,1641 \,.\, 10^{-10} + 3,8417 \cdot 10^{-9} + \ldots + 2,7629 \cdot 10^{-5} \right] = 2 \cdot 3,3093 \cdot 10^{-5}$$
$$= 6,61877 \cdot 10^{-5}$$

Como el valor de la *p* exacta es inferior al nivel habitual $\alpha = 0,05$ se rechazaría la hipótesis de igualdad de proporciones. Para averiguar el sentido de la diferencia bastará con observar hacia dónde se ha desplazado el mayor número de pares discordantes que, como puede observarse en el ejemplo, ha sido a la casilla *(c)* que corresponde a los que antes no practicaban ejercicio físico y que, después del programa, sí lo han hecho mostrando, por tanto, evidencia de la efectividad del programa de concienciación.

APROXIMACIÓN ASINTÓTICA

Cuando tanto el número de casos discordantes tipo *(b)* como los de tipo *(c)* son mayores o iguales a 10, puede utilizarse la distribución asintótica basada en la aproximación normal al modelo binomial. Así se tendrá que:

$$Z = \frac{X - E(X)}{\sqrt{V(x)}} = \frac{X - \frac{b+c}{2}}{\sqrt{\frac{b+c}{4}}}$$

Se distribuirá según un modelo de distribución normal de media 0 y desviación típica 1. Si se eleva al cuadrado, la variable resultante se distribuirá según un modelo de distribución ji-cuadrado con 1 grado de libertad. Además, dado que el número de pares discordantes con respuesta afirmativa-antes ha sido designado como de tipo *(b)* se tendrá que el estadístico de contraste se expresará de la siguiente forma:

$$J = \left(\frac{X - \frac{b+c}{2}}{\sqrt{\frac{b+c}{4}}} \right)^2 = \frac{\left(b - \left(\frac{b+c}{2} \right) \right)^2}{\frac{b+c}{4}} = \frac{\frac{(2b-(b+c))^2}{4}}{\frac{b+c}{4}} = \frac{(b-c)^2}{b+c}$$

Si se trabaja con los datos del ejemplo se tendrá que:

$$J = \frac{(b-c)^2}{b+c} = \frac{(5-28)^2}{5+28} = \frac{529}{33} = 16,03$$

En las tablas 3-27 y 3-28 se muestran los resultados del análisis obtenidos a partir del programa de análisis estadístico SPSS.

Como puede observarse, los resultados de la prueba exacta coinciden con los obtenidos manualmente (p = 0,000066) y es similar al obtenido mediante la aproximación asintótica (p = 0,000128), ya que difieren a partir de la cuarta cifra decimal. Sin embargo, debe hacerse notar que tanto el

TABLA 3-27 Tabla de contingencia. Práctica de ejercicio físico

Antes	Después	
	No	**Sí**
No	22	28
Sí	5	25

TABLA 3-28 Prueba de McNemar

Estadísticos de contraste	Antes y después
N	80
Ji-cuadrado[a]	14,667
Sig. asintótica	,000128
Sig. exacta (bilateral)	,000066
Sig. exacta (unilateral)	,000
Probabilidad en el punto	,000

[a] *Corregido por continuidad.*

estadístico ji-cuadrado proporcionado por el SPSS como, en consecuencia, la *p* asociada difieren del obtenido manualmente, ya que el programa utiliza una corrección por continuidad que se obtendría de la siguiente forma.

$$J_{\text{corregido}} = \left(\frac{|X - 0,5| - \frac{b+c}{2}}{\sqrt{\frac{b+c}{4}}} \right)^2 = \frac{\left(|b - 0,5| - \left(\frac{b+c}{2} \right) \right)^2}{\frac{b+c}{4}} = \frac{\left(|b-c| - 1 \right)^2}{b+c}$$

En este caso se tendrá que:

$$J_{\text{corregido}} = \frac{\left(|b-c| - 1 \right)^2}{b+c} = \frac{\left(|5-28| - 1 \right)^2}{5+28} = 14,667$$

Que, como puede observarse, coincide con el obtenido con el SPSS.

COMPARACIÓN DE TRES O MÁS PROPORCIONES PARA MUESTRAS RELACIONADAS (PRUEBA Q DE COCHRAN)

La prueba Q de Cochran está indicada para la comparación de una variable cualitativa dicotómica en tres o más grupos cuando las muestras están relacionadas y constituye una extensión de la prueba de McNemar.

Ejemplo 3-6

Se pretende estudiar la efectividad de tres tratamientos para una determinada patología. Se dispone de tres muestras de 15 pacientes cada una, de forma que, para cada paciente al que se le suministró el primer tratamiento (T_1), se seleccionaron otros dos de la misma edad, sexo y características clínicas a los que se les suministró el segundo (T_2) y tercer tratamiento (T_3), respectivamente.

TABLA 3-29 Éxito/fracaso según el tipo de tratamiento. Muestras relacionadas

Bloque	T_1	T_2	T_3	F_i **= suma fila**
1	1	0	1	2
2	1	1	1	3
3	1	0	0	1
4	1	0	0	1
5	0	1	1	2
6	1	0	1	2
7	0	0	1	1
8	0	0	0	0
9	1	1	0	2
10	1	0	1	2
11	0	1	0	1
12	0	0	1	1
13	1	0	1	2
14	1	1	1	3
15	1	1	0	2
C_i **= suma columna**	10	6	9	$S_T = 25$

Los resultados de éxito o fracaso del tratamiento se codificaron de la siguiente forma: 1 (éxito) y 0 (fracaso). Los resultados se muestran en la tabla 3-29.

Cada uno de los bloques representa a un trío (en general *k* grupos) de datos relacionados (pacientes con las mismas características de cada una de las tres muestras seleccionadas). En total se dispone de $N = 15$ tríos o bloques. Se ha añadido una última fila y una columna donde se hace constar la suma por fila y por columna de los códigos correspondientes al éxito/fracaso. La suma total de códigos se ha representado como S_T, que en este caso es de $10 + 6 + 9 = 2 + 3 + 1 + 1 + \ldots + 3 + 2 = 25$. Obsérvese que la suma de los totales por fila y por columna debe coincidir y ser igual a S_T.

¿Existe evidencia de diferencias en la efectividad de los tratamientos? Si se trabaja con la proporción de éxito de cada uno de los tratamientos, las hipótesis del contraste quedarían establecidas de la siguiente forma:

$$H_0 : p_1 = p_2 = p_3$$

$$H_1 : \text{Al menos una proporción distinta}$$

Las proporciones de éxito observadas en los datos correspondientes a cada uno de los tratamientos pueden calcularse de la siguiente forma:

$$\hat{p}_1 = \frac{C_1}{N} = \frac{10}{15} = 0,667; \hat{p}_2 = \frac{C_2}{N} = \frac{6}{15} = 0,4; \hat{p}_3 = \frac{C_3}{N} = \frac{9}{15} = 0,6$$

A priori, parece observarse una diferencia de éxito del segundo tratamiento con respecto a los otros dos. Bajo la hipótesis de que no existen diferencias en las proporciones de éxito de los tratamientos, todas deberían ser iguales a la proporción global de éxito p_g, que viene dada por:

$$\hat{p}_g = \frac{S_T / N}{k} = \frac{25/15}{3} = 0,55$$

donde k es el número de grupos. En consecuencia, podría construirse un estadístico de contraste basado en las diferencias de cada una de las proporciones de éxito observadas con respecto a la proporción global de éxito esperada:

$$\sum \left(\hat{p}_i - \hat{p}_g\right)^2 = \sum \left(\frac{C_i}{N} - \frac{\frac{S_T}{N}}{k} \right)^2 = \sum \frac{1}{N^2} \left(C_i - \frac{S_T}{k} \right)^2 = \frac{1}{N^2} \sum \left(C_i - \frac{S_T}{k} \right)^2$$

El estadístico Q de Cochran se expresará finalmente de la siguiente forma y se distribuirá asintóticamente según un modelo de distribución ji-cuadrado con k − 1 grados de libertad (en este caso k = 3 y, por tanto, k − 1 = 2 grados de libertad):

$$Q = k(k-1) \frac{\sum \left(C_i - \frac{S_T}{k} \right)^2}{\sum F_i \left(k - F_i \right)} = (k-1) \frac{k \sum C_i^2 - S_T^2}{k S_T - \sum F_i^2}$$

Obsérvese que la parte fundamental del estadístico se encuentra en el numerador y corresponde a la diferencia entre la proporción observada en cada uno de los grupos y la esperada o global, bajo la hipótesis de no existencia de diferencias descrita con anterioridad.

A partir de los datos del ejemplo, se tendrá que:

$$Q = (k-1) \frac{k \sum C_i^2 - S_T^2}{k S_T - \sum F_i^2} = (3-1) \frac{3\left(10^2 + 6^2 + 9^2\right) - 25^2}{3 \cdot 25 - \left(2^2 + 3^2 + 1^2 + \ldots + 2^2\right)} = 2,167$$

En una distribución ji-cuadrado con 2 grados de libertad, proporciona un valor de la p del contraste de 0,33847. De forma análoga a las pruebas estudiadas con anterioridad se puede calcular la p exacta cuando el número de datos disponibles no es suficiente para garantizar la distribución asintótica y que, en este caso, proporciona un valor de 0,4407. En ambos casos, el valor de la p es superior al nivel habitual $\alpha = 0,05$, por lo que no podrá rechazarse la hipótesis nula, concluyendo la no existencia de significación estadística para la diferencia en las proporciones de éxito de los tratamientos.

TABLA 3-30 Frecuencias por grupo

	Resultado (éxito/fracaso)	
	(fracaso)	(éxito)
Tratamiento 1	5	10
Tratamiento 2	9	6
Tratamiento 3	6	9

TABLA 3-31 Prueba de Cochran. Estadísticos de contraste

N	15
Q de Cochran	2,167[a]
gl	2
Sig. asintótica	,338
Sig. exacta	,441
Probabilidad en el punto	,145

gl, *grados de libertad.*
[a] *1 se trata como un éxito.*

En las tablas 3-30 y 3-31 se muestran los resultados del análisis efectuado con el programa de análisis estadístico SPSS.

Puede observarse que los resultados obtenidos coinciden con los calculados manualmente. La diferencia entre la p obtenida mediante la aproximación asintótica y la exacta da una idea del número de datos (tríos de datos) necesarios para una buena aproximación entre las dos cantidades.

COMPARACIÓN DE DOS O MÁS PROPORCIONES PARA MUESTRAS INDEPENDIENTES (PRUEBA EXACTA DE FISHER)

La prueba exacta de Fisher está indicada para el estudio de la relación entre dos variables cualitativas y habitualmente se emplea cuando no se verifican las condiciones necesarias para la utilización de la prueba ji-cuadrado. Recuérdese que esta última prueba requiere que, al menos en el 80% de las casillas, la frecuencia esperada sea igual o superior a 5. Aunque esta prueba es aplicable a variables cualitativas politómicas, se analizará en este apartado, con objeto de ilustrar el procedimiento, el caso en que las dos variables cualitativas sean dicotómicas, dando lugar, en consecuencia, a una tabla 2×2.

Ejemplo 3-7

En un estudio sobre hábitos nutricionales se obtuvo información sobre el consumo diario de grasas de un grupo de 18 individuos. El interés se centra en establecer algún tipo de relación entre el consumo excesivo de grasas (se adoptó el criterio de consumo excesivo si superaba el 35% del total de calorías diarias) y la obesidad. Los resultados obtenidos se muestran en la tabla 3-32.

Si se calculan las proporciones de obesidad observadas en cada uno de los grupos se tiene que:

$$\hat{p}_{exc} = \frac{7}{10} = 0,7; \hat{p}_{no\,exc} = \frac{2}{8} = 0,25$$

A nivel descriptivo parece apreciarse que la proporción de individuos con obesidad es superior en el grupo de los que presentan un consumo excesivo de grasas. Sin embargo, habría que comprobar si esta diferencia es significativa. La prueba que, en principio, podría utilizarse es la ji-cuadrado, siempre que se verificaran las condiciones necesarias. En la tabla 3-33 se proporcionan los valores de las frecuencias esperadas.

Dado que el 75% de las casillas muestra un valor esperado inferior a 5, no sería aplicable la prueba ji-cuadrado para la comparación de proporciones, siendo necesaria alguna otra alternativa.

En primer lugar, las hipótesis del contraste quedarían establecidas de la siguiente forma:

$$H_0 : p_{exc} = p_{no\,exc}$$

$$H_1 : p_{exc} = p_{no\,exc}$$

TABLA 3-32 Distribución de frecuencias de obesidad y consumo excesivo de grasas de 18 individuos

Consumo excesivo	Obesidad		Total
	Sí	No	
Sí	7 *(a)*	3 *(b)*	10 *(a + b)*
No	2 *(c)*	5 *(d)*	7 *(c + d)*
Total	9 *(a + c)*	8 *(b + d)*	17 *(a + b + c + d)*

TABLA 3-33 Frecuencias esperadas bajo la hipótesis de igualdad de proporciones

Consumo excesivo	Obesidad	
	Sí	No
Sí	5,29	4,7
No	3,7	3,29

La probabilidad de obtener una disposición de datos concreta (frecuencias en cada una de las casillas) manteniendo constantes los totales por fila y por columna puede calcularse de la siguiente forma:

$$\text{Probabilidad} = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}}$$

En el caso del ejemplo, la probabilidad de haber obtenido la tabla será:

$$\text{Probabilidad} = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}} = \frac{\binom{7+2}{7}\binom{3+5}{3}}{\binom{17}{7+3}} = \frac{\binom{9}{7}\binom{8}{3}}{\binom{17}{10}} = 0,10366$$

La estrategia de análisis se basará en la consideración de todas las disposiciones posibles de los datos (frecuencias en cada casilla) con la condición de mantener constantes los totales por fila y por columna. En este caso las disposiciones posibles serían las mostradas en la figura 3-1.

A continuación se calcula la probabilidad de cada una de estas disposiciones de la forma descrita con anterioridad, obteniéndose:

$$0,10366; 0,30234; 0,36281; 0,18141; 0,03455; 0,00185; 0,01296; 0,00041$$

El valor de la *p* exacta del contraste vendrá dado por la probabilidad de obtener una disposición igual o más extrema que la observada con independencia de hacia dónde se dirige la diferencia en las proporciones (proporción de obesidad mayor en uno u otro grupo con independencia de que se haya observado en la disposición real que la proporción de obesidad es mayor en el primer grupo de consumo excesivo de grasas). Dado que la probabilidad de la disposición real observada es 0,10366, habría que sumarle las probabilidades de inferior valor (ya que corresponden a disposiciones más extremas). Así se tendrá que:

$$p = 0,10366 + 0,03455 + 0,00185 + 0,01296 + 0,00041 = 0,15343$$

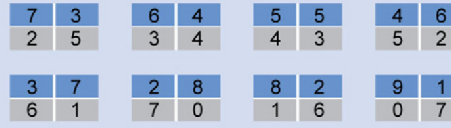

FIGURA 3-1 Disposiciones posibles de los datos en el ejemplo 3-7.

TABLA 3-34 Prueba exacta de Fisher. Pruebas de ji-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)	Probabilidad en el punto
Ji-cuadrado de Pearson	2,837	1	,092	,153	,117	
Corrección por continuidad	1,418	1	,234			
Razón de verosimilitudes	2,915	1	,088	,153	,117	
Estadístico exacto de Fisher				,153	,117	
Asociación lineal por lineal	2,670	1	,102	,153	,117	,104
N de casos válidos	17					

Como el valor de la p del contraste 0,15343 es superior al nivel habitual $\alpha = 0{,}05$, no puede rechazarse la hipótesis nula y, por tanto, no puede concluirse la existencia de diferencias significativas en las proporciones de obesidad entre los dos grupos considerados.

En la tabla 3-34 se muestran los resultados del análisis con el programa SPSS. Como puede observarse, los resultados coinciden con los obtenidos manualmente. Recuérdese que en este caso no procede comparar con el resultado obtenido mediante el estadístico ji-cuadrado (aunque en este caso coincida el valor de la p que proporciona) porque no es aplicable.

AUTOEVALUACIÓN

1. Las pruebas no paramétricas:
- **a.** Tienen mayor capacidad para detectar significación estadística.
- **b.** Pueden utilizarse en muestras pequeñas.
- **c.** No requieren que el muestreo sea aleatorio.
- **d.** Habitualmente son preferibles a las pruebas paramétricas.
- **e.** Requieren la suposición de una distribución de probabilidad para la variable objeto de estudio.

2. Se desea contrastar si el nivel promedio de colesterol en un determinado tipo de pacientes es distinto de 240 mg/100 ml. Se dispone de 10 observaciones:
- **a.** Podría utilizarse la prueba t para una media.
- **b.** Deberá formularse la hipótesis sobre la mediana de colesterol poblacional.

c. Se requiere suponer una distribución de probabilidad para la variable *nivel de colesterol*.
d. Podría utilizarse la prueba de la mediana.
e. b y d son ciertas.

3. Se desea comparar los niveles promedio de triglicéridos en dos grupos de pacientes de 12 individuos cada uno:
a. Si las muestras son independientes podría utilizarse la prueba U de Mann-Whitney.
b. Si las muestras son apareadas podría utilizarse la prueba de McNemar.
c. En cualquier caso, la prueba más indicada sería la de Kruskal-Wallis.
d. La prueba más indicada sería la de los rangos de Friedman.
e. Ninguna de las anteriores.

4. Se desea comparar las proporciones poblacionales de consumo de tabaco en hombres y mujeres. Se dispone de dos muestras de 15 y 17 individuos respectivamente:
a. En cualquier caso, podría utilizarse la prueba ji-cuadrado de comparación de proporciones.
b. La prueba más indicada sería la prueba de McNemar.
c. Se requieren hipótesis adicionales sobre la distribución asociada para poder utilizar la prueba ji-cuadrado.
d. Podría utilizarse la prueba exacta de Fisher si más del 20% de las casillas tiene un valor esperado inferior a 5.
e. No es posible utilizar la prueba exacta de Fisher.

CAPÍTULO

4

Análisis de la varianza. ANOVA

Joaquín Moncho Vasallo

INTRODUCCIÓN

En multitud de ocasiones el investigador trata de analizar la posible relación entre una variable cuantitativa y una variable cualitativa o factor. La aproximación habitual consiste en comparar las medias de la variable cuantitativa en cada uno de los grupos o niveles de la variable cualitativa. En caso de ser significativamente distintas, podría concluirse una relación entre dichas variables, ya que habría dado lugar a un comportamiento diferencial de la variable cuantitativa en cada uno de los grupos analizados.

- ¿Existen diferencias en los niveles promedio de colesterol (en mg/100 ml) según el nivel de obesidad? El nivel de obesidad se ha clasificado en tres grupos: delgado-normal (IMC < 25), sobrepeso ($25 \leq$ IMC < 30) y obesidad (IMC ≥ 30)
- ¿Existe relación entre el nivel de colesterol (en mg/100 ml) y el consumo de alcohol? El consumo de alcohol se ha clasificado en tres niveles: nunca, moderado y alto.

Anteriormente se abordó el caso de la comparación de dos medias para muestras independientes mediante la prueba *t*. El ANOVA es una prueba de comparación de medias resultado de la extensión natural de la prueba *t* para la comparación de dos muestras independientes al caso de la comparación de tres o más medias. En el caso de tres o más muestras apareadas o relacionadas podrá utilizarse el ANOVA *de medidas repetidas*, aunque no será objeto de estudio en el presente capítulo. En adelante se supondrá que las observaciones de los diferentes grupos o niveles del factor constituyen muestras independientes.

ANOVA DE EFECTOS FIJOS Y ALEATORIOS

El ANOVA puede ser de efectos fijos o de efectos aleatorios (también podría ser una combinación de efectos fijos y aleatorios) en función de las características de la variable que define los grupos o factor. Cuando

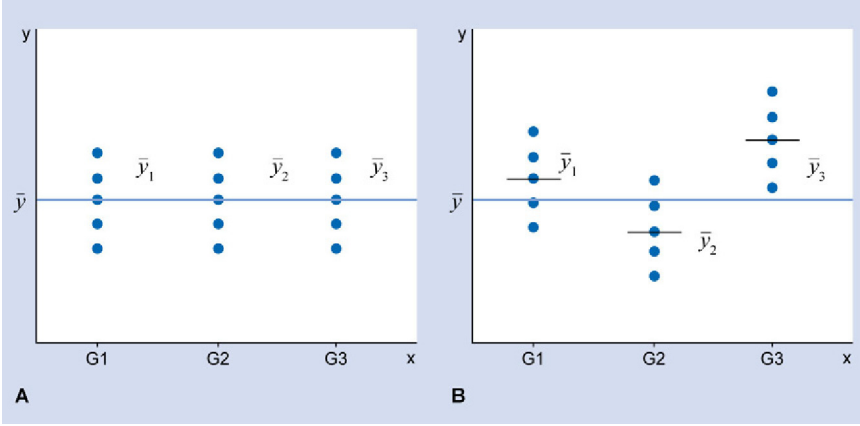

FIGURA 4-1 Diagrama de dispersión de una variable cuantitativa *(y)* según el grupo de pertenencia (G_1, G_2 o G_3).

los niveles o grupos del factor son los únicos de interés o los únicos que pueden tener un efecto sobre la variable dependiente, se denominará factor fijo (p. ej., sexo, estado civil, etc.). Si, por el contrario, los niveles del factor constituyen una muestra de una población de niveles más amplia, se denominará efecto aleatorio (p. ej., se dispone de observaciones de determinadas características para una muestra de ciudades, países, hospitales, etc.). En el presente capítulo se abordará el análisis correspondiente al ANOVA de efectos fijos.

En la figura 4-1 pueden observarse dos situaciones generadas a partir de la comparación de la distribución de una variable cuantitativa en tres grupos (definidos por la variable cualitativa). Por ejemplo, la variable *y* podría ser el *nivel de colesterol* y la variable *x* el *nivel de obesidad*.

En la figura 4-1A las medias de nivel de colesterol de cada uno de los grupos o niveles de obesidad son iguales y, por tanto, coincidirán con la media global. En la figura 4-1B, por el contrario, se observa que las medias de nivel de colesterol de los grupos difieren entre sí y, en consecuencia, de la media global. Por otra parte, se ha de ser consciente de que, al igual que en situaciones anteriores, los resultados obtenidos corresponden a una muestra de una población mucho mayor y que el interés se centra en establecer la existencia o no de diferencias significativas entre las medias poblacionales a través del contraste de hipótesis, cuya hipótesis nula sería:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

La hipótesis alternativa contemplaría el caso en el que al menos una de las medias poblacionales es diferente del resto. ¿Cómo se puede valorar si las medias de los grupos son distintas entre sí?

Debe tenerse en cuenta, al proponer un método de análisis de comparación de tres o más medias, que se precisa un procedimiento que compare todas las medias conjuntamente. Una comparación sucesiva de pares de medias (prueba *t* para muestras independientes) podría llevar a resultados distintos ya que el nivel de significación real al que se estaría trabajando sería mayor, incrementando el error de tipo I y, en consecuencia, las posibilidades de detección de significación estadística.

DESCOMPOSICIÓN DE LA VARIABILIDAD

En el proceso de construcción de un método para la comparación de más de dos medias en muestras independientes será muy útil utilizar el gráfico de dispersión de los datos. En la figura 4-2 se observa la distribución de la variable cuantitativa *y* en cada uno de los tres grupos que define la variable cualitativa *x*.

El gráfico sugiere un comportamiento diferencial de la variable cuantitativa *y* en cada uno de los tres grupos que define la variable cualitativa *x*. La distancia de una observación cualquiera y_{ij} a la media de las observaciones $\overline{y}$ puede descomponerse de la forma:

$$y_{ij} - \overline{y} = \left(y_{ij} - \overline{y}_i\right) + \left(\overline{y}_i - \overline{y}\right)$$

Es decir, la distancia de una observación cualquiera y_{ij} a la media de las observaciones $\overline{y}$ puede descomponerse como la suma de la distancia de cada observación a la media de su grupo y la distancia de la media de su grupo a la media global. Una forma de obtener un resumen de las distancias

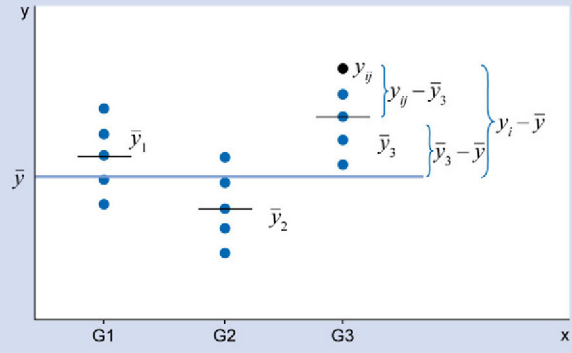

FIGURA 4-2 Descomposición de la variabilidad.

sería calcular la suma de todas las distancias (una para cada observación) de la siguiente forma:

$$\sum (y_{ij} - \overline{y}) = \sum \left(y_{ij} - \overline{y}_i\right) + \sum \left(\overline{y}_i - \overline{y}\right)$$

Sin embargo, dado que la media de las observaciones se sitúa en el centro de gravedad de los datos, las distancias positivas cancelarían las distancias negativas y esta cantidad sería siempre 0. Para eliminar el efecto del signo, pero conservar la magnitud de la distancia, se eleva al cuadrado, de forma que se obtiene la siguiente expresión:

$$(y_{ij} - \overline{y})^2 = \left(y_{ij} - \overline{y}_i\right)^2 + \left(\overline{y}_i - \overline{y}\right)^2 + 2\left(y_{ij} - \overline{y}_i\right)\left(\overline{y}_i - \overline{y}\right)$$

Si se calcula la suma de todas estas distancias, ahora al cuadrado y, por tanto, positivas, se tendrá que:

$$\sum (y_{ij} - \overline{y})^2 = \sum \left(y_{ij} - \overline{y}_i\right)^2 + \sum \left(\overline{y}_i - \overline{y}\right)^2 + 2\sum \left(y_{ij} - \overline{y}_i\right)\left(\overline{y}_i - \overline{y}\right)$$

donde:

$$\sum \left(y_{ij} - \overline{y}_i\right)\left(\overline{y}_i - \overline{y}\right) = 0$$

Y en consecuencia:

$$\sum (y_{ij} - \overline{y})^2 = \sum \left(y_{ij} - \overline{y}_i\right)^2 + \sum \left(\overline{y}_i - \overline{y}\right)^2$$

Esta expresión se conoce como la descomposición de la variabilidad, ya que, como puede observarse, a la izquierda de la igualdad se tiene el numerador de la varianza total de la variable cuantitativa *y* expresada como suma de dos variabilidades, donde:

$\Sigma(y_{ij} - \overline{y}_i)^2$ = Variabilidad no explicada por la variable explicativa o factor. Apréciese que resume las distancias al cuadrado entre el verdadero valor observado de la variable cuantitativa y la media del grupo al que pertenece cada observación (varianza no explicada por el factor). Se denota VNE.

$\Sigma(\overline{y}_i - \overline{y})^2$ = Variabilidad explicada por la variable explicativa o factor. Obsérvese que resume las distancias entre la media de cada uno de los grupos o niveles del factor y la media global (valor que se proporcionaría para estimar el valor de *y* si no se tuviera en cuenta el factor). Se denota VE.

$\Sigma(y_{ij} - \overline{y})^2$ = Variabilidad total observada en la variable cuantitativa. Obsérvese que coincide con el numerador de la varianza de la variable *y*. Se denota VT.

Con la notación adoptada se tiene que la varianza total de la variable cuantitativa *y* puede expresarse:

$$VT = VNE + VE$$

Adviértase que cuanto más grande sea la cantidad VE mayor es la diferencia entre las medias de cada uno de los grupos y la media global y, en consecuencia, mayor es la diferencia entre ellas. Además, cuanto mayor sea VE menor será VNE ya que VT es fijo para un conjunto de datos.

COEFICIENTE DE DETERMINACIÓN

La descomposición de la variabilidad ofrece la oportunidad de definir una medida interpretada como la proporción de variabilidad de la variable cuantitativa o dependiente explicada por la variable que define los grupos o factor, que recibe el nombre de *coeficiente de determinación*. Así, se tendrá que:

$$\text{Coeficiente de determinación} = R^2 = \frac{VE}{VNE} = \frac{\sum \left(\bar{y}_i - \bar{y}\right)^2}{\sum \left(y_{ij} - \bar{y}_i\right)^2}$$

Si se multiplica por 100, se interpretará como el porcentaje de variabilidad de la variable dependiente explicada por el factor.

Para ilustrar el procedimiento se propone un ejemplo con un número reducido de datos, si bien serían necesarios, al menos, 30 datos por grupo.

Ejemplo 4-1

En un estudio se obtuvo información sobre el nivel de colesterol (en mg/100 ml) y el nivel de obesidad (delgado-normal, sobrepeso y obesidad, construidos a partir del índice de masa corporal) de un grupo de 25 pacientes. Los resultados se muestran en la tabla 4-1.

TABLA 4-1 Datos de nivel de obesidad y nivel de colesterol

Nivel de obesidad	Nivel de colesterol	n
Delgado-normal	145, 160, 166, 173, 202, 218, 224, 251, 172	9
Sobrepeso	156, 232, 237, 239, 245, 247, 302, 241, 235, 240	10
Obesidad	154, 228, 236, 242, 245, 266, 269, 305, 280, 275, 304	11

Si se calculan las medias del nivel de colesterol en cada uno de los grupos o niveles del factor *obesidad* y la media global se obtendrá que:

$$\bar{y} = \frac{145+160+166+173+\ldots+304}{30} = 229,63$$

$$\bar{y}_1 = \frac{145+160+166+173+202+218+224+251+172}{9} = 190,11$$

$$\bar{y}_2 = \frac{156+232+237+239+245+247+302+241+235+240}{10} = 237,4$$

$$\bar{y}_3 = \frac{154+228+236+242+245+266+269+305+280+275+304}{11} = 254,91$$

A continuación se calculan las sumas de cuadrados correspondientes a las variabilidades totales, explicada y no explicada, de la siguiente forma:

$$VT = \sum (y_{ij} - \bar{y})^2 = (145-229,63)^2 + (160-229,63)^2 + \ldots + (304-229,63)^2$$

$$\begin{aligned} VNE = \sum \left(y_{ij} - \bar{y}_i\right)^2 &= (145-190,11)^2 + (160-190,11)^2 + \ldots + (172-190,11)^2 + \\ &+ (156-237,4)^2 + (232-237,4)^2 + \ldots + (240-237,4)^2 + \\ &+ (154-254,91)^2 + (228-254,91)^2 + \ldots + (304-254,91)^2 \end{aligned}$$

$$\begin{aligned} VE &= \sum \left(\bar{y}_i - \bar{y}\right)^2 \\ &= (190,11-229,63)^2 + (190,11-229,63)^2 + \ldots + (190,11-229,63)^2 \\ &+ (237,11-229,63)^2 + (237,4-229,63)^2 + \ldots + (237,4-229,63)^2 \\ &+ (254,91-229,63)^2 + (254,91-229,63)^2 + \ldots + (254,91-229,63)^2 \\ &= 9(190,11-229,63)^2 + 10(237,11-229,63)^2 + 11(254,91-229,63)^2 \end{aligned}$$

Se tendrá que:

$$VT = 60.416,967; VNE = 38.728,198; VE = 21.688,769$$

donde la suma de la variabilidad explicada por el factor y la no explicada coincidirá con la variabilidad total:

$$VT = 60.416,967 = 21.688,769 + 38.728,198 = VE + VNE$$

Por su parte, el coeficiente de determinación quedará de la siguiente forma:

$$R^2 = \frac{VE}{VNE} = \frac{21.688,769}{60.416,967} = 0,359$$

Este resultado implicaría que el factor nivel de obesidad lograría explicar el 35,9% de la variabilidad observada en la variable dependiente *nivel de colesterol.*

INFERENCIA Y TABLA DE ANOVA

Cuando se comparan tres o más medias a partir de los datos contenidos en una muestra aleatoria de la población debe tenerse en cuenta que las medias muestrales calculadas en cada uno de los grupos o niveles del factor no son más que unas de todas las posibles medias que podrían calcularse, a partir de cada una de las muestras posibles de la población que hubieran podido ser seleccionadas en el proceso de muestreo. Esto es, para cada muestra se obtendrían unas medias que podrían ser similares a las de otra muestra, pero que variarán al variar los datos de partida.

En la figura 4-3 puede observarse que las medias correspondientes a cada uno de los grupos difieren en función de la muestra aleatoria seleccionada. En consecuencia, las medias muestrales de los grupos son variables aleatorias que varían de muestra a muestra de la población.

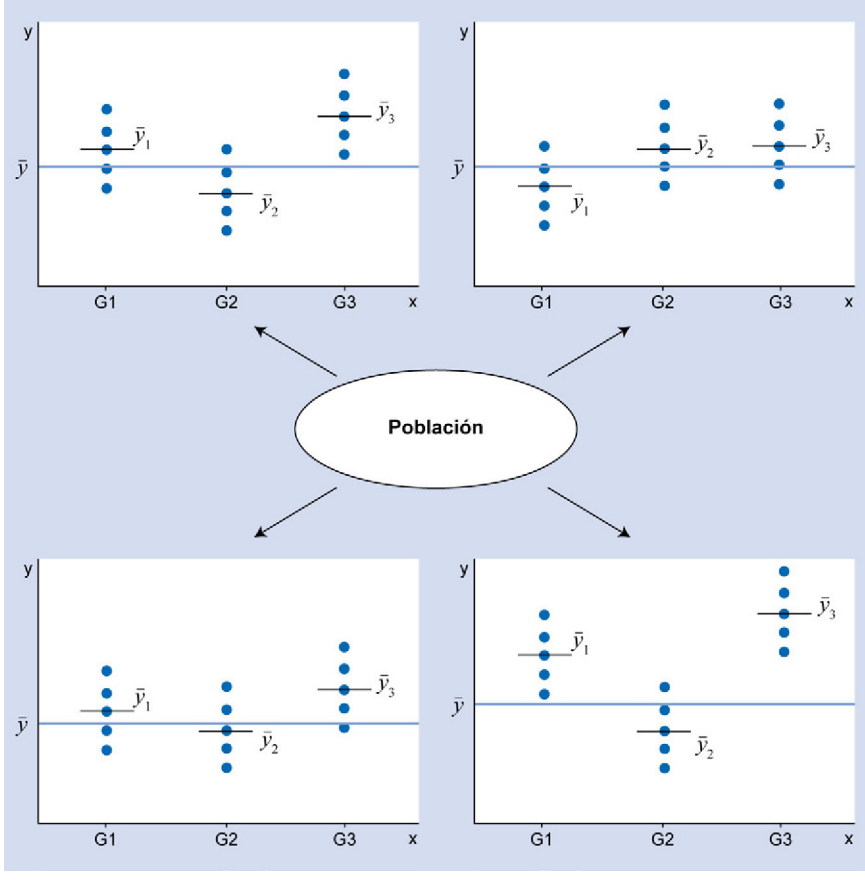

FIGURA 4-3 Diferencias en las medias en función de la muestra seleccionada.

Un contraste de interés trataría de establecer si todas las medias poblacionales de los diferentes grupos o niveles del factor son iguales o, por el contrario, alguna o algunas son distintas de las demás.

$$H_0: \mu_1 = \mu_2 = \mu_3$$
$$H_1: \text{Alguna } \mu_1 \text{ es distinta}$$

En caso de aceptación de la hipótesis nula, no habría evidencia de que las medias poblacionales sean distintas entre sí. En caso de rechazar la hipótesis nula, alguna o algunas de las medias poblacionales serían distintas.

El estadístico de contraste es, en este caso:

$$EC = \frac{VE/(k-1)}{VNE/(n-k)} = \frac{VE/(3-1)}{VNE/(n-3)} = \frac{\sum(\bar{y}_i - \bar{y})^2/2}{\sum(y_i - \bar{y}_i)^2/(n-3)}$$

donde k = número de grupos y n = tamaño de la muestra global ($n = n_1 + n_2 + n_3$). Este estadístico de contraste se distribuirá según una F de Snedecor con $k - 1$ y $n - k$ grados de libertad en el caso de verificarse las hipótesis necesarias que se abordarán más adelante. Como puede observarse, si la variabilidad explicada por el modelo es 0 el valor del estadístico de contraste será 0 (situación compatible con la hipótesis nula de no existencia de diferencias significativas entre las medias de los grupos). Por otro lado, si la variabilidad explicada comienza a aumentar (crece el numerador), la variabilidad no explicada tendrá que disminuir (disminuye el denominador) con lo que el valor del estadístico de contraste crecerá cada vez más hasta el punto de poder rechazar, en su caso, la hipótesis nula (las medias de los grupos son significativamente distintas entre sí).

Suele ser habitual, presentar los resultados anteriores en forma de tabla conocida como la tabla de ANOVA (tabla 4-2). Esta reproduce la secuencia de razonamiento que culmina con la construcción de un estadístico de

TABLA 4-2 Tabla de ANOVA

Fuente	Suma de cuadrados	Media de cuadrados	Cociente de varianzas
Variabilidad entre grupos (VE)	$\sum(\bar{y}_i - \bar{y})^2$	$\frac{\sum(\bar{y}_i - \bar{y})^2}{k-1}$	$\frac{\sum(\bar{y}_i - \bar{y})^2/(k-1)}{\sum(y_{ij} - \bar{y}_i)^2/(n-k)}$
Variabilidad intragrupos (VNE)	$\sum(y_{ij} - \bar{y}_i)^2$	$\frac{\sum(y_{ij} - \bar{y}_i)^2}{n-k}$	
Total (VT)	$\sum(y_{ij} - \bar{y})^2$	$\frac{\sum(y_{ij} - \bar{y})^2}{n-1}$	

TABLA 4-3 Tabla de ANOVA para los datos del ejemplo 4-1

Fuente	Suma de cuadrados	Media de cuadrados	Cociente de varianzas
Variabilidad entre grupos (VE)	21.688,769	10.844,38	$\frac{10.844,38}{1.434,38} = 7,56$
Variabilidad intragrupos (VNE)	38.728,198	1.434,38	
Total (VT)	60.416,96	2.083,34	

contraste como el propuesto con anterioridad, y que tiene como actores principales a la varianza explicada (entre grupos) y la no explicada (intragrupos). En la tabla 4-3 se muestra la tabla de ANOVA para los datos del ejemplo 4-1.

El estadístico de contraste sería EC = 7,56 que, comprobado en las tablas de la F de Snedecor con 2 y 27 grados de libertad, proporcionaría un valor de la *p* de 0,002. Dado que el valor de la *p* del contraste es inferior al nivel habitual 0,05, se rechazaría la hipótesis nula y se concluiría que al menos una media poblacional es distinta a las demás.

HIPÓTESIS BÁSICAS SOBRE EL ANOVA

La aplicación del método de ANOVA para la comparación de tres o más medias para muestras independientes requiere la asunción de una serie de requisitos sobre las variables y observaciones, que pueden resumirse en las siguientes:

- La distribución de las observaciones es normal en cada uno de los grupos o niveles del factor. En la práctica se requiere un número mínimo de 30 datos por grupo.
- La varianza es homogénea en todos los grupos o niveles del factor.
- Las observaciones de cada uno de los grupos o niveles del factor constituyen una muestra aleatoria en cada uno de ellos y son independientes entre sí.

Para la comprobación de las hipótesis se utilizan recursos gráficos y analíticos como los que se describen a continuación:

- Criterios gráficos:
  - Histogramas y gráficos P-P normal para la variable cuantitativa en conjunto y para cada uno de los grupos o niveles del factor considerados.
  - Gráficos comparativos de cajas *(box-plots)* y diagramas de dispersión de la variable cuantitativa según cada uno de los grupos o niveles del factor.

  - Gráficos de secuencia, aunque la posibilidad de su construcción depende de la forma en que se hayan obtenido los datos.
- Criterios analíticos:
  - Test de Kolmogorov-Smirnov o contraste ji-cuadrado de bondad de ajuste para comprobar la hipótesis de normalidad.
  - Test de Levene para comprobar la hipótesis de homogeneidad de varianzas (menos exigente con la hipótesis de normalidad que la prueba de Barlett).

REFLEXIONES SOBRE LAS HIPÓTESIS

El ANOVA es una técnica que ha demostrado un comportamiento robusto ante la violación de las hipótesis necesarias descritas con anterioridad. De hecho, algunos autores señalan que lo inusual es que se verifiquen todas y cada una de las hipótesis, siendo su aplicación extendida para comparar tres o más medias, sobre todo en el caso del ANOVA de efectos fijos. Sin embargo, será necesario tener en cuenta las consecuencias que su incumplimiento tiene sobre los resultados obtenidos.

- La hipótesis de normalidad no tiene por qué verificarse exactamente cuando se trabaja con muestras grandes, sobre todo en el ANOVA de efectos fijos. En el caso de ANOVA de efectos aleatorios su violación tendrá consecuencias más importantes.
- La violación de la hipótesis de homogeneidad de varianzas tampoco tiene una gran trascendencia cuando el número de observaciones por grupo es similar, aunque en el caso de ANOVA de efectos aleatorios precisaría de algún tipo de análisis complementario.
- Por su parte, la violación de la hipótesis de independencia de las observaciones puede conducir a errores graves en las inferencias, tanto en el ANOVA de efectos fijos como en el ANOVA de efectos aleatorios.

IDENTIFICACIÓN DE LAS MEDIAS SIGNIFICATIVAMENTE DISTINTAS. CONTRASTES *POST HOC*

El contraste F del ANOVA permite contrastar la hipótesis de igualdad de medias de la variable cuantitativa entre los diferentes grupos. En el caso de que el contraste condujera al rechazo de dicha hipótesis podía afirmarse que existe evidencia de que al menos la media correspondiente a uno de los grupos era distinta del resto. La cuestión es si existe algún método para determinar qué media o medias son significativamente distintas, ya que no tienen por qué ser todas ellas las responsables de las diferencias halladas. Existen varias propuestas para responder a esta cuestión y cuya

utilización se aconseja en función del tamaño de los grupos, número de grupos a comparar, estrategia de comparación entre los grupos, equilibrio en el tamaño de los grupos o similitud de varianzas. A continuación se destacan algunas de estas medidas:

- En el caso de varianzas similares y equilibrio en el tamaño de los grupos que se pretende comparar:
  - HSD de Tuckey: realiza todas las pruebas de comparación de medias por pares corrigiendo el nivel de significación en función del número de pruebas realizadas.
  - Prueba de Dunnett: útil cuando lo que interesa al investigador es comparar todos los grupos con respecto a un grupo control.
- En el caso de varianzas similares y desequilibrio en el tamaño de los grupos que se pretende comparar:
  - Scheffé: no solo compara parejas, sino combinaciones lineales de las medias de los grupos, permitiendo contrastar la diferencia entre, por ejemplo, una media y el resto. Constituye una buena alternativa además cuando no existe equilibrio en el tamaño de los grupos. Para el caso en que el interés se centra únicamente en las comparaciones por pares, es un método más conservador, que tiende a detectar menos diferencias significativas de las que debería. Tiene un buen comportamiento incluso cuando el tamaño de los grupos es similar.

Ejemplo 4-2

Se cuenta con información sobre el nivel de colesterol e índice de masa corporal de un grupo de 199 pacientes. El nivel de obesidad ha sido clasificado en tres niveles dependiendo del índice de masa corporal. Los resultados descriptivos se muestran en la tabla 4-4.

A nivel descriptivo, se observa una diferencia en los niveles promedio de colesterol muestrales entre los tres grupos considerados. ¿Son significativas las diferencias?

TABLA 4-4 Media y desviación típica del nivel de colesterol según el nivel de obesidad

Nivel de obesidad	Media	N	Desviación típica
Delgado o normal (IMC < 25)	205,83	69	44,326
Sobrepeso (25 ≤ IMC < 30)	237,57	93	50,819
Obesidad (IMC ≥ 30)	241,08	37	49,743
Total	227,22	199	50,699

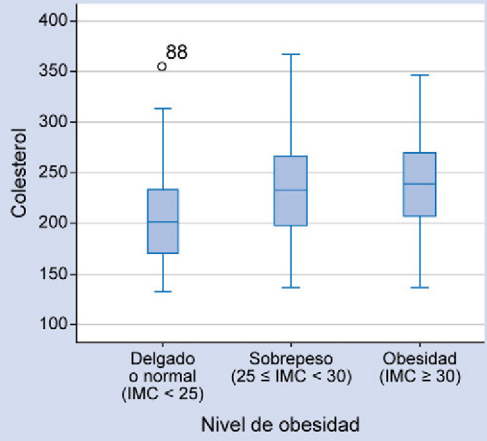

FIGURA 4-4 Diagrama de cajas. Nivel de colesterol según nivel de obesidad.

La técnica que permitiría establecer la existencia de diferencias significativas entre los promedios de colesterol entre los tres grupos considerados sería el ANOVA (de efectos fijos) para muestras independientes.

En primer lugar, se propone comprobar los requisitos para la utilización del ANOVA. El número de observaciones por grupo es lo suficientemente elevado (superior a 30 en todos los grupos considerados) lo que minimiza el efecto de una posible desviación de la normalidad de la variable dependiente en cada uno de los grupos sobre los resultados que se obtengan en el ANOVA.

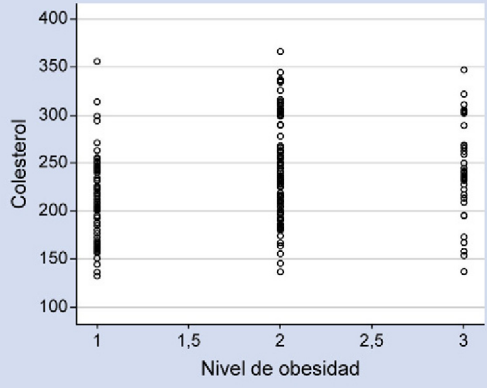

FIGURA 4-5 Diagrama de dispersión. Nivel de colesterol según nivel de obesidad.

TABLA 4-5 Prueba de Kolmogorov-Smirnov para una muestra

		Colesterol
N		199
Parámetros normales[a,b]	Media	227,22
	Desviación típica	50,699
Diferencias más extremas	Absoluta	,059
	Positiva	,047
	Negativa	−,059
Z de Kolmogorov-Smirnov		,838
Sig. asintótica (bilateral)		,484

[a] *La distribución de contraste es la normal.*
[b] *Se han calculado a partir de los datos.*

Los diagramas de cajas (fig. 4-4) y de dispersión (fig. 4-5) muestran, no obstante, situaciones compatibles con la normalidad en cada uno de los grupos y la homogeneidad de varianzas. En el diagrama de cajas se observa que en todos los casos la mediana se sitúa en el centro de la caja correspondiente al intervalo intercuartílico evidenciando una simetría aparente en la distribución de los datos.

En el diagrama de dispersión, se observa una mayor concentración o densidad de observaciones en torno a los valores centrales de cada uno de los grupos, compatible con lo esperado si los datos se distribuyeran según un modelo de distribución normal. Por otra parte, la altura de las cajas en cada uno de los grupos (distancia entre el cuartil 1 y el cuartil 3) es similar, lo que indica una cierta homogeneidad en la dispersión de los datos y, en consecuencia, de sus varianzas. En el mismo sentido, en el diagrama de dispersión se observa que el rango de valores en los tres grupos es muy similar.

En las tablas 4-5 a 4-8 se proporcionan resultados analíticos para la comprobación de las hipótesis de normalidad a través del contraste de Kolmogorov-Smirnov para el conjunto de datos y para cada uno de los grupos.

El contraste de normalidad de la distribución de la variable dependiente *nivel de colesterol*, tanto en conjunto como para cada uno de los grupos, resulta no significativo, por lo que no se evidencian diferencias significativas con respecto a la distribución normal.

TABLA 4-6 Prueba de Kolmogorov-Smirnov para una muestra[a]

	Colesterol
N	69
Z de Kolmogorov-Smirnov	,756
Sig. asintótica (bilateral)	,617

[a] *Nivel de obesidad = delgado o normal (IMC < 25).*

TABLA 4-7 Prueba de Kolmogorov-Smirnov para una muestra[a]

	Colesterol
N	93
Z de Kolmogorov-Smirnov	,736
Sig. asintótica (bilateral)	,65

[a] *Nivel de obesidad = sobrepeso (25 ≤ IMC < 30).*

TABLA 4-8 Prueba de Kolmogorov-Smirnov para una muestra[a]

	Colesterol
N	37
Z de Kolmogorov-Smirnov	,48
Sig. asintótica (bilateral)	,976

[a] *Nivel de obesidad = obesidad (IMC ≥ 30).*

En cuanto al requerimiento de homogeneidad de varianzas, en la tabla 4-9 se muestra el resultado de la prueba de Levene.

En este caso, dado que el valor de la *p* del contraste de igualdad de varianzas es superior al nivel habitual 0,05, puede concluirse que no existen diferencias significativas entre las varianzas de los grupos. Por último, dado que los datos han sido obtenidos en un estudio observacional en el que la secuencia temporal no parece tener ningún tipo de efecto sobre los datos, se supone que se verifica la hipótesis de independencia de las observaciones.

En la tabla 4-10 se proporciona la tabla de ANOVA basada en la descomposición de la variabilidad.

Puede observarse que el valor de la p del contraste de igualdad de medias es inferior a la milésima ($p < 0{,}001$) por lo que puede concluirse que al menos una de las medias de los tres grupos, es significativamente distinta al resto.

Para profundizar en el análisis e intentar detectar la media o medias que presentan diferencias significativamente distintas, se propone utilizar el contraste *post hoc* de Scheffé, dado que los grupos están desequilibrados

TABLA 4-9 Prueba de homogeneidad de varianzas

Colesterol			
Estadístico de Levene	**gl1**	**gl2**	**Sig.**
,743	2	196	,477

TABLA 4-10 Tabla de ANOVA

Colesterol

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Intergrupos	48.652,243	2	24.326,122	10,359	,000
Intragrupos	460.283,465	196	2.348,385		
Total	508.935,709	198			

(el tamaño de los grupos es bastante desigual), cuyos resultados se muestran en la tabla 4-11.

De los resultados obtenidos se desprende que existen diferencias significativas entre los promedios de nivel de colesterol de los grupos delgado/normal y los otros dos grupos (sobrepeso y obesidad), pero no entre en los correspondientes a sobrepeso y obesidad. Este resultado puede resumirse, además, en los datos de la tabla 4-12, que presenta los grupos homogéneos que se desprenderían de los resultados del análisis.

Puede observarse que el análisis proporciona dos subconjuntos homogéneos. En un primer subconjunto se situaría el primer grupo (delgado/normal) y en el segundo subconjunto los otros dos grupos (sobrepeso y obesidad).

TABLA 4-11 Comparaciones múltiples

Variable dependiente: colesterol

	(I) Nivel de obesidad	(J) Nivel de obesidad	Diferencia de medias (I − J)	Error típico	Sig.	IC 95%	
						Límite inferior	Límite superior
Scheffé	Delgado/normal (IMC < 25)	Sobrepeso (25 ≤ IMC < 30)	−31,744*	7,7	,000	−50,74	−12,75
		Obesidad (IMC ≥ 30)	−35,255*	9,874	,002	−59,61	−10,9
	Sobrepeso (25 ≤ IMC < 30)	Delgado/normal (IMC < 25)	31,744*	7,7	,000	12,75	50,74
		Obesidad (IMC ≥ 30)	−3,511	9,419	,933	−26,74	19,72
	Obesidad (IMC ≥ 30)	Delgado/normal (IMC < 25)	35,255*	9,874	,002	10,9	59,61
		Sobrepeso (25 ≤ IMC < 30)	3,511	9,419	,933	−19,72	26,74

* *La diferencia de medias es significativa al nivel 0,05.*

TABLA 4-12 Grupos homogéneos

Colesterol

	Nivel de obesidad	N	Subconjunto para α = 0,05	
			1	2
Scheffé	Delgado o normal (IMC < 25)	69	205,83	
	Sobrepeso (25 ≤ IMC < 30)	93		237,57
	Obesidad (IMC ≥ 30)	37		241,08
	Sig.		1	,927

Se muestran las medias para los grupos en los subconjuntos homogéneos.

MODELIZACIÓN DEL ANOVA. MODELO DE ANOVA DE EFECTOS FIJOS

El ANOVA de efectos fijos puede ser expresado en forma de modelo para un conjunto de datos. Así, si se pretende valorar el efecto de un factor (variable cualitativa que define los grupos) sobre una variable respuesta o dependiente (variable cuantitativa) que define *k* grupos, la relación podrá expresarse del siguiente modo:

$$y_{ij} = \mu + \alpha_i \quad y_{ij} = \mu + \alpha_i + e_{ij}; i = 1, 2..., k$$

donde:

y_{ij} es cada una de las observaciones de la variable dependiente.
μ es la media global (sin distinguir entre grupos).
α_i es el efecto diferencial del grupo i respecto a la media global ($\alpha_i = \mu_i - \mu$).
e_{ij} es el error [$e_{ij} = y_{ij} - (\mu + \alpha_i) = y_{ij} - \mu - \alpha_i$].

De la modelización propuesta se desprende que cada una de las observaciones de la variable dependiente puede expresarse como un efecto global (común) más un efecto grupo (si es que existe) más el error (fruto de la variabilidad observada en cada uno de los grupos considerados).

La estimación de cada uno de los parámetros del modelo se realizaría de la siguiente forma:

$$\mu = \frac{\overline{y}_1 + \overline{y}_2 + \ldots + \overline{y}_k}{k} = \overline{y}^*$$

$$\alpha_i = \overline{y}_i - \overline{y}^*$$

Debe hacerse notar que la media $\overline{y}^*$ no coincide con la media calculada a partir de todos los datos a no ser que los grupos sean equilibrados

(el tamaño de las muestras de los grupos sea el mismo). De la expresión anterior se desprende que la suma de todos los efectos diferenciales de los grupos o niveles del factor debe ser cero:

$$\alpha_1 + \alpha_2 + \ldots + \alpha_k = \sum \alpha_i = \sum \left(\overline{y}_i - \overline{y}^*\right) = \sum \overline{y}_i - \sum \overline{y}^* = \sum \overline{y}_i - 3\overline{y}^* = 0$$

En el caso de que cada uno de los efectos α_i sea igual a 0 el modelo estará mostrando que los datos pueden expresarse como una media global más el error, no existiendo un efecto diferencial por grupo y, en consecuencia, no existiendo diferencias en los promedios de la variable dependiente entre los grupos considerados.

Si $0 = \alpha_i = \mu_i - \mu$, entonces $\mu_i - \mu = 0$ y $\mu_i = \mu$.

Puede comprobarse que cada una de las medias de los grupos coincidiría con la media global y, por tanto, no existirían diferencias entre las mismas.

RELACIÓN ENTRE EL MODELO DE ANOVA Y EL MODELO DE REGRESIÓN LINEAL

El análisis de regresión lineal y la modelización del ANOVA de efectos fijos guardan una estrecha relación. Tanto es así que podría construirse un modelo de regresión lineal múltiple que proporcionara los mismos resultados que el ANOVA correspondiente, si bien lo habitual es utilizar el ANOVA cuando no se introducen variables cuantitativas entre las explicativas.

Para comprender mejor este resultado se recomienda al lector revisar previamente el capítulo correspondiente a la regresión lineal.

Supóngase que se trabaja con los datos del ejemplo 4-1 (v. tabla 4-1) para el establecimiento de una posible relación entre el nivel de obesidad categorizado en tres grupos y el nivel de colesterol.

El modelo de ANOVA se expresará:

$$y_{ij} = \mu + \alpha_i$$

Si se estiman el efecto general y el efecto grupo para los datos del ejemplo se tendrá que:

$$\mu = \frac{\overline{y}_1 + \overline{y}_2 + \overline{y}_3}{3} = \overline{y}^* = \frac{190,11 + 237,4 + 254,91}{3} = 227,47$$

$$\alpha_1 = \overline{y}_1 - \overline{y}^* = 190,11 - 227,47 = -37,36$$

$$\alpha_2 = \overline{y}_2 - \overline{y}^* = 237,4 - 227,47 = 9,93$$

$$\alpha_3 = \overline{y}_3 - \overline{y}^* = 254,91 - 227,47 = 27,44$$

Puede observarse que la suma de los efectos α_i es 0:

$$\sum \alpha_i = \alpha_1 + \alpha_2 + \alpha_3 = -37,36 + 9,93 + 27,44 = 0$$

El modelo de regresión lineal equivalente precisaría de la construcción de dos variables ficticias o *dummies* (tabla 4-13) para la separación de los efectos de los tres grupos de la variable *nivel de obesidad*.

TABLA 4-13 Variables ficticias o *dummies*

Variable independiente (factor)	Valor en variable ficticia 1 (X_1 = Ob1)	Valor en variable ficticia 2 (X_2 = Ob2)
Delgado/normal	1	0
Sobrepeso	0	1
Obesidad	−1	−1

El modelo de regresión lineal quedaría de la siguiente forma:

$$\text{Colesterol} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Al ajustar este modelo de regresión lineal múltiple se obtiene el resultado descrito en la tabla 4-14.

Puede comprobarse que el parámetro β_0 del modelo de regresión coincide con la media ajustada $\overline{y}^*$ = 227,47. Por otra parte, los efectos de los grupos también coinciden con los parámetros ajustados que acompañan las variables ficticias ob1 y ob2 ($\beta_1 = -37,36 = \alpha_1$ y $\beta_2 = 9,93 = \alpha_2$). El valor de α_3 no aparece en el modelo de regresión al ser redundante, ya que, si se sabe que $\Sigma\alpha_i = 0$, entonces:

$$\alpha_3 = -(\alpha_1 + \alpha_2) = -(-37,362 + 9,927) = -(-27,44) = 27,44$$

TABLA 4-14 Coeficientes del modelo de regresión lineal múltiple[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	227,473	6,938		32,787	,000
	ob1	−37,362	10,063	−,678	−3,713	,001
	ob2	9,927	9,795	,185	1,013	,32

[a] *Variable dependiente: colesterol.*

ESTRATEGIAS DE ANÁLISIS ANTE LA FALTA DE CUMPLIMIENTO DE LOS REQUISITOS DEL ANOVA

Cuando los requerimientos necesarios para la aplicación del ANOVA no se verifican pueden adoptarse diversas estrategias:

- Estrategias encaminadas a corregir las desviaciones de las hipótesis sobre el ANOVA para que pueda utilizarse.
- Utilizar otras técnicas de análisis.

En el primer caso suele ser habitual utilizar transformaciones de los datos para conseguir normalidad u homocedasticidad. Debe tenerse en cuenta la robustez del ANOVA en lo relativo a la hipótesis de normalidad y homocedasticidad comentadas anteriormente en este capítulo, sobre todo en el caso del ANOVA de efectos fijos con grupos equilibrados. En consecuencia, cuando los grupos muestren tamaños y varianzas significativamente desiguales sería recomendable adoptar alguna solución alternativa.

Una segunda posibilidad es recurrir a técnicas no paramétricas que no precisan estos requerimientos como es el caso de la prueba de Kruskall-Wallis, si bien debe tenerse en cuenta que son menos potentes a la hora de detectar significación estadística.

Existen sin embargo, otras posibilidades de análisis cuando las varianzas no son homogéneas, basadas en pruebas paramétricas que surgen como una extensión de la prueba *t* de comparación de medias para el caso de varianzas desiguales. La idea fundamental es corregir el estadístico de contraste F del ANOVA, de forma que tenga en cuenta la disparidad en las varianzas de los grupos considerados. Los programas de análisis estadístico suelen incorporar dos pruebas construidas de esta forma como son:

- La prueba de Welch.
- La prueba de Brown-Forsythe.

En la tabla 4-15 se proporcionan los resultados de las pruebas de Welch y Brown-Forsythe para los datos del ejemplo 4-2.

Ambas pruebas proporcionarían, en este caso, un valor de *p* significativo. Obsérvese que el estadístico de contraste no solo difiere en ambos casos del

TABLA 4-15 Pruebas de Welch y Brown-Forsythe. Pruebas robustas de igualdad de las medias del nivel de colesterol

	Estadístico[a]	gl1	gl2	Sig.
Welch	11,154	2	95,931	,000
Brown-Forsythe	10,41	2	138,672	,000

[a] *Distribuidos en F asintóticamente.*

obtenido en el ANOVA, sino que también difieren los grados de libertad asociados a la distribución F correspondiente. Habitualmente la prueba de Welch suele ser la preferida por los investigadores como alternativa al ANOVA cuando las varianzas no son homogéneas.

En caso de haber optado por la utilización de alguna de estas dos pruebas (Welch o Brown-Forsythe) la interpretación de los resultados obtenidos será similar a la del ANOVA. Así, si se obtiene un valor de *p* significativo se concluirá que al menos una de las medias de los grupos es distinta a las demás. Al igual que en el caso del ANOVA, será necesario identificar, posteriormente, qué media o medias son las responsables de las diferencias encontradas.

En este caso, cuando se ha optado por una prueba basada en la disparidad o heterogeneidad de varianzas será necesario utilizar una prueba *post hoc* que tenga en cuenta esta situación. Entre las pruebas más habituales en este caso se proponen las siguientes:

- Games-Howell: se basa en la corrección de Welch y tiene un buen comportamiento incluso en el caso de que los tamaños de los grupos no sean equilibrados, o de que haya pocos casos.
- Dunnett T3: indicada cuando lo que interesa es la comparación por pares de medias y los tamaños de las muestras por grupos son pequeños.

TABLA 4-16 Comparaciones múltiples

Variable dependiente: nivel de colesterol

	(I) Nivel de obesidad	(J) Nivel de obesidad	Diferencia de medias (I − J)	Error típico	Sig.	IC al 95%	
						Límite inferior	Límite superior
Games-Howell	Delgado/ normal	Sobrepeso	−31,744*	7,5	,000	−49,49	−14
		Obesidad	−35,255*	9,765	,002	−58,66	−11,85
	Sobrepeso	Delgado o normal	31,744*	7,5	,000	14	49,49
		Obesidad	−3,511	9,729	,931	−26,83	19,8
	Obesidad	Delgado o normal	35,255*	9,765	,002	11,85	58,66
		Sobrepeso	3,511	9,729	,931	−19,8	26,83
C de Dunnett	Delgado/ normal	Sobrepeso	−31,744*	7,5		−49,66	−13,83
		Obesidad	−35,255*	9,765		−58,98	−11,53
	Sobrepeso	Delgado o normal	31,744*	7,5		13,83	49,66
		Obesidad	−3,511	9,729		−27,11	20,09
	Obesidad	Delgado o normal	35,255*	9,765		11,53	58,98
		Sobrepeso	3,511	9,729		−20,09	27,11

** La diferencia de medias es significativa al nivel 0,05.*

- Dunnett C: indicada también cuando el interés se centra en las comparaciones de pares de medias pero los tamaños de las muestras por grupos son grandes.

En la tabla 4-16 se muestran los resultados de las pruebas *post hoc* de Games-Howell y Dunnett C (las muestras por grupo son relativamente grandes) para los datos del ejemplo 4-2.

Puede observarse que la media correspondiente al grupo de delgado/normal es significativamente distinta de las otras dos. Por su parte las medias de los grupos *sobrepeso* y *obesidad* no muestran diferencias significativas entre ellas.

AUTOEVALUACIÓN

1. Se desea comparar el promedio poblacional de peso en tres grupos de pacientes con tamaños muestrales 35, 45 y 40, respectivamente:
   a. Deben compararse todos los pares de medias posibles para detectar cuáles presentan diferencias significativas.
   b. Es suficiente con comparar las medias muestrales.
   c. No puede utilizarse el ANOVA, ya que el tamaño muestral es insuficiente.
   d. La prueba más indicada sería, en principio, el ANOVA, si se cumplen los requisitos necesarios.
   e. Las muestras son relacionadas y no es posible utilizar el ANOVA para muestras independientes.

2. En principio, la aplicación de la técnica de ANOVA para la comparación de medias requiere que:
   a. La variable dependiente siga una distribución normal.
   b. La población sobre la que se quiere inferir sea finita.
   c. La varianza sea homogénea en al menos dos de los grupos considerados.
   d. Las observaciones estén relacionadas entre sí.
   e. Se requiere únicamente que el tamaño muestral sea igual o superior a 30 en cada uno de los grupos.

3. El método de comparación por pares de Scheffé:
   a. Compara únicamente pares de medias.
   b. Compara no solo parejas de medias, sino combinaciones lineales de las mismas.
   c. Tiende a detectar más diferencias de las que debería.
   d. Es una buena alternativa cuando las varianzas de los grupos son muy distintas entre sí.
   e. Debe realizarse antes del contraste del ANOVA.

4. En caso de que las varianzas de los grupos (tres o más) muestren una excesiva heterogeneidad:
   a. No existen soluciones estadísticas que permitan la comparación de medias.
   b. Deberá utilizarse necesariamente la prueba de Kruskal-Wallis.
   c. Existen pruebas paramétricas alternativas de análisis.
   d. Solo pueden compararse parejas de medias.
   e. b y d son ciertas.

CAPÍTULO

5

Análisis de regresión lineal simple y múltiple

Joaquín Moncho Vasallo

INTRODUCCIÓN

En anteriores capítulos han sido abordadas situaciones en las que se pretendía estudiar la posible relación entre una variable cuantitativa y otra cualitativa (comparando medias de la variable cuantitativa en cada uno de los grupos de la variable cualitativa) o entre dos variables cualitativas (comparando proporciones de ocurrencia de categorías de una variable en cada uno de los grupos de la otra variable). Sin embargo, con frecuencia se pretende estudiar la posible relación entre dos variables cuantitativas:

- ¿Existe relación entre el nivel de colesterol y la edad? ¿Y entre el índice de masa corporal y el nivel de colesterol? Se dispone de información sobre 150 sujetos seleccionados al azar.
- ¿Existe relación entre el peso de la madre y el peso del niño al nacimiento? Se dispone de una muestra aleatoria de 85 madres.

La primera reflexión consistiría en preguntarse cuándo se decidirá que existe una relación entre las dos variables consideradas. Parece lógico pensar que existirá una relación, por ejemplo, entre la edad y el nivel de colesterol si, en conjunto, a mayor edad mayor nivel de colesterol. A este tipo de relación se la llamaría «directa». Pero también podría darse el caso en que a mayor valor de una variable menor valor de la otra, en cuyo caso se trataría de una relación «inversa». En el caso de que al aumentar el valor de una variable la otra no mostrara ninguna variación significativa en su comportamiento, se diría que no se observa relación entre las mismas.

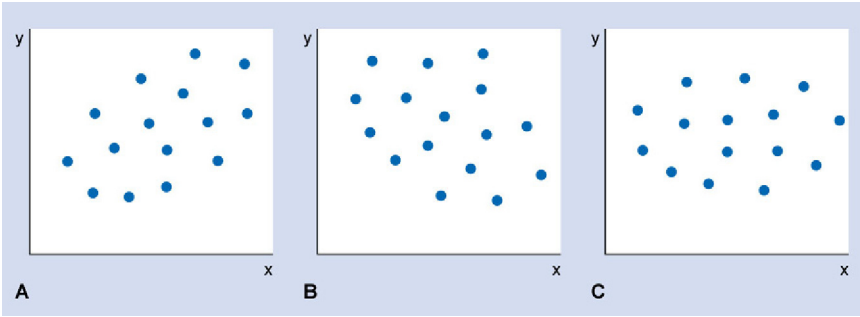

FIGURA 5-1 Diagramas de dispersión para diferentes situaciones.

CONCEPTOS PREVIOS. COVARIANZA Y COEFICIENTE DE CORRELACIÓN LINEAL

En el proceso de estudio de la posible relación entre dos variables cuantitativas será muy útil utilizar instrumentos de representación gráfica que permitan sugerir diferentes tipos de relación detectables. Para representar gráficamente la distribución conjunta de dos variables cuantitativas se utiliza habitualmente el *diagrama de dispersión* (fig. 5-1), en el que se sitúa una variable en el eje de abscisas y la otra en el eje de ordenadas. La situación A sería compatible con una relación directa (a mayor valor de x, mayor valor de y), la situación B con una relación inversa, y la situación C con una ausencia de relación.

COVARIANZA ENTRE DOS VARIABLES CUANTITATIVAS

La *covarianza* es una medida que ofrece información sobre la posible relación entre dos variables cuantitativas. Para ilustrar su proceso de cálculo será muy útil apoyarse en el diagrama de dispersión y tener en cuenta dónde se sitúa el punto $(\overline{x}, \overline{y})$. Si se considera la situación A de la figura 5-1 , se tendrá que el punto $(\overline{x}, \overline{y})$ se situará en el centro de gravedad de la distribución de los datos (nube de puntos), determinando cuatro cuadrantes, tal y como se describe en la figura 5-2. Se propone calcular la cantidad $(x_i - \overline{x})(y_i - \overline{y})$ para cada uno de los pares de observaciones (puntos). Para todos los puntos situados en el primer cuadrante (1) y en el cuadrante (3) estas cantidades tendrán signo positivo. [Obsérvese, por ejemplo, que para todos los puntos situados en el primer cuadrante se verifica $x_i > \overline{x}$ e $y_i > \overline{y}$ por lo que $(x_i - \overline{x})(y_i - \overline{y}) > 0$.] Sin embargo, en el caso de los puntos situados en los cuadrantes (2) y (4), estas cantidades serán negativas. ¿Qué sucedería si se calculara, en este caso, la media de todas las cantidades?

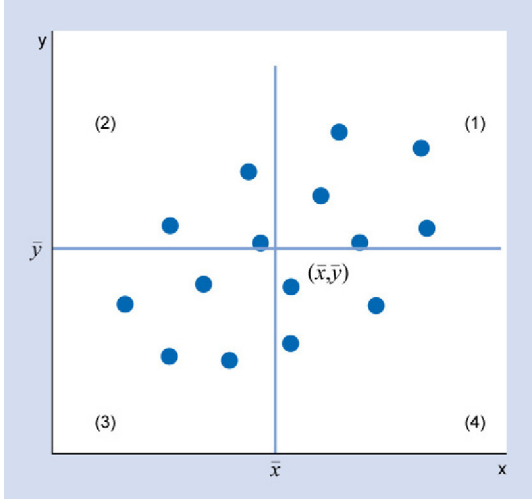

FIGURA 5-2 Situación de cada observación respecto al punto ($\bar{x}, \bar{y}$).

$$\mathrm{Cov}(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Esta cantidad recibe el nombre de *covarianza.* Si se observa la figura 5-2, la mayoría de las observaciones se sitúan en el primer y tercer cuadrante, presentando además unas distancias al punto $(\bar{x}, \bar{y})$ mayores que en el caso de los puntos situados en el segundo y cuarto cuadrante. En consecuencia, las cantidades positivas serán superiores a las cantidades negativas y la media de todas las cantidades tendrá un signo positivo.

¿Qué hubiera ocurrido si se hubiera trabajado con los datos de la figura 5-1B? En ese caso, la mayoría de las observaciones se hubieran situado en los cuadrantes (2) y (4), presentando además unas distancias al punto $(\bar{x}, \bar{y})$ superiores a los puntos situados en los cuadrantes (1) y (2). En este caso, la cantidad *Cov(x,y)* tendría un valor negativo.

En conclusión, parece que, si los datos muestran una relación directa entre las variables cuantitativas, el valor de *Cov(x,y)* es positivo y, si la relación es inversa, el valor de *Cov(x,y)* es negativo. Por último, si los datos describen una situación de ausencia de relación como la de la figura 5-1C, las cantidades positivas se cancelarían con la negativas y el valor de *Cov(x,y)* sería 0.

A pesar del buen comportamiento de la covarianza como medida de la relación entre dos variables cuantitativas, sobre todo en lo concerniente

al sentido de la relación (directa o inversa) si es que existe, presenta dos problemas importantes:

- *Depende de las unidades de medida.* Esto quiere decir que dos representaciones gráficas idénticas y, por tanto, que describen una misma relación entre las variables, podrían ofrecer dos valores de la covarianza distintos.
- *No tiene cota superior ni inferior.* Esto supone que no es posible conocer, a partir de la covarianza, lo cerca o lejos que se encuentran los datos de describir la relación perfecta (todos los puntos se situarían sobre una recta), sea directa o inversa.

Se precisa, por tanto, una medida adimensional (que no dependa de las unidades de medida) que además tenga una cota superior e inferior. El valor máximo y mínimo se alcanzaría cuando la relación entre las variables fuera perfecta (directa e inversa, respectivamente). Un valor próximo a 0 indicaría, en principio, ausencia de relación, aunque esta cuestión precisará algunas matizaciones.

COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON

Para obtener una medida adimensional y acotada superior e inferiormente se divide la covarianza por el producto de las desviaciones típicas de cada una de las variables. Así, se tendrá:

$$r = \frac{\mathrm{Cov}(x,y)}{S_x S_y}$$

Esta medida toma valores entre –1 y 1. Como puede observarse, el actor principal de la cantidad así obtenida es la covarianza, que es quien atesora el sentido de la relación entre las dos variables cuantitativas. El valor de *r* será, por ejemplo, positivo, solo si el valor de la *Cov(x,y)* lo es y, por tanto, la relación entre las variables es directa. Además cuanto más se acerque el valor de *r* a 1 o –1, mayor es la magnitud de la relación entre las dos variables y más cerca se encuentran de describir la relación lineal perfecta. Por otro lado, si el valor de *r* se acerca a 0 será porque la *Cov(x,y)* se acerca a 0 y se supone una ausencia de relación entre las mismas, pero ¿es realmente así? ¿Qué tipo de relaciones es capaz de captar el coeficiente de correlación lineal de Pearson?

En la figura 5-3 se muestra el diagrama de dispersión correspondiente a dos variables cuantitativas. Si se calcula el coeficiente de correlación lineal de Pearson su valor se aproximará a 0. Obsérvese que las cantidades $(x_i - \overline{x})\,(y_i - \overline{y})$ correspondientes al primer cuadrante se cancelarían con las del segundo cuadrante y, por su parte, las del tercero con las del cuarto cuadrante. Este resultado haría pensar en una ausencia de relación entre las variables.

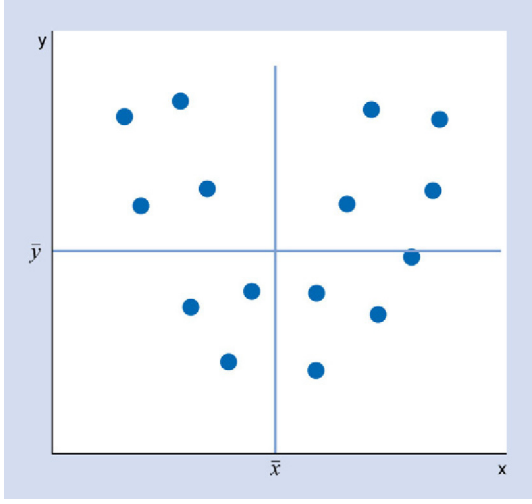

FIGURA 5-3 Relación no lineal con coeficiente de correlación lineal $r \approx 0$.

Sin embargo, el gráfico muestra que podría existir una relación entre las dos variables que no sería lineal sino cuadrática (parábola). Por tanto, debe tenerse muy en cuenta, que el coeficiente de correlación lineal de Pearson es capaz de captar, en principio, relaciones lineales y que un valor próximo a 0 no implica necesariamente una ausencia de relación entre las variables estudiadas, sino más bien una ausencia de relación lineal.

INFERENCIA SOBRE EL COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON

Al igual que en capítulos anteriores, es importante tener en cuenta que el coeficiente de correlación lineal de Pearson calculado a partir de un conjunto de datos es un estadístico (calculado a partir de los datos de la muestra) y que, por tanto, varía en función de la muestra aleatoria seleccionada. En la mayoría de ocasiones, el objetivo se centra en averiguar si existe relación entre dos variables cuantitativas en la población de la que partió la muestra.

Un contraste de hipótesis de interés, en este caso, trataría de establecer si el coeficiente de correlación lineal de Pearson poblacional es o no significativamente distinto de 0. Así, se tendrá que:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

El estadístico de contraste utilizado en ese caso se basa en una transformación del coeficiente de correlación lineal de Pearson:

$$EC = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Esta cantidad se distribuirá, bajo la hipótesis nula, según un modelo de distribución *t* de Student con n – 2 grados de libertad. Para garantizar la distribución muestral del estadístico de contraste se precisaría que cada una de las variables cuantitativas se distribuyera según un modelo de probabilidad normal.

Si se rechaza la hipótesis nula, el coeficiente de correlación lineal poblacional sería significativamente distinto de 0 y se concluiría que existe relación significativa entre las dos variables.

MODELO DE REGRESIÓN LINEAL SIMPLE

El coeficiente de correlación lineal de Pearson informa sobre la magnitud o fuerza de la asociación lineal entre dos variables cuantitativas (cuánto se asocian), sin embargo, se precisa de otras técnicas, como el análisis de regresión lineal, para poder disponer de información sobre la naturaleza de la relación existente (cómo se asocian). Además, disponer de un modelo lineal que expresara la relación entre las dos variables permitiría la realización de predicciones.

- ¿Qué cambio se produce en el nivel de colesterol si se aumenta en un año la edad del individuo?
- ¿Cuál sería el nivel de colesterol estimado para un individuo de 35 años? ¿Y el nivel promedio de colesterol de los individuos de 35 años?

ESTRUCTURA DEL MODELO DE REGRESIÓN LINEAL SIMPLE

Para ilustrar la construcción e interpretación del modelo de regresión lineal simple se utilizará de nuevo el diagrama de dispersión. Así, para los datos del estudio de la relación, por ejemplo, entre el nivel de colesterol y la edad, el modelo de regresión lineal simple sería la recta que mejor resume los datos (fig. 5-4). La expresión funcional del modelo sería, por tanto:

$$y = \beta_0 + \beta_1 x$$

Como puede observarse, esta expresión funcional establece que el valor de *y* depende del valor de *x*. Esto es: una vez que se dispone del modelo, si se hace variar el valor de *x* variará el valor de *y*. En el caso de que el modelo propuesto fuera $x = \beta_0 + \beta_1 y$, entonces sería *x* la variable que dependería del valor de *y*. Esta cuestión es importante porque, a diferencia del análisis

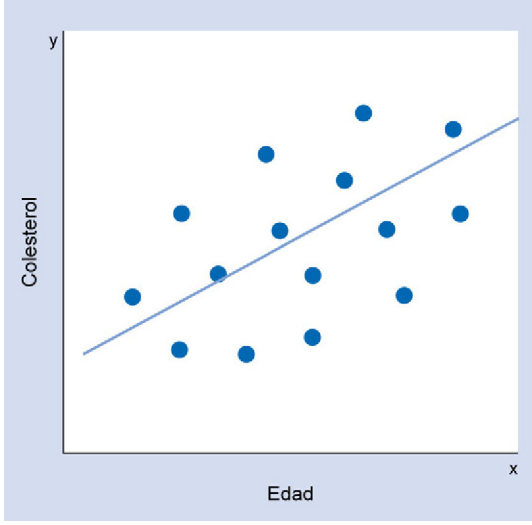

FIGURA 5-4 Recta de regresión lineal: nivel de colesterol en función de la edad.

de correlación lineal, el análisis de regresión requiere que se establezca qué variable representará el papel de *variable dependiente, explicada* o *respuesta* y qué variable jugará el papel de *variable independiente, explicativa* o *predictora.* En el ejemplo, parece lógico considerar que la variable dependiente sea el nivel de colesterol y la variable independiente la edad.

OBTENCIÓN DE LA RECTA DE REGRESIÓN LINEAL SIMPLE

Para que la recta de regresión lineal simple esté perfectamente definida será necesario obtener los valores de β_0 y β_1. Uno de los criterios más comunes para la obtención de la recta que mejor ajusta los datos se basa en la minimización del error cometido al proporcionar el valor estimado por la recta $\hat{y}_i$ en lugar del verdadero valor observado y_i, tal y como se refleja en la figura 5-5.

La mejor recta sería la que hiciera mínimas todas las distancias entre el verdadero valor observado y_i y el valor estimado por la recta $\hat{y}_i$, llamado *error* o *residuo* y expresado como e_i. Así, para cada individuo tendríamos:

$$e_i = y_i - \hat{y}_i$$

La cantidad Σe_i expresaría la suma total de los residuos del modelo y, *a priori,* la mejor recta sería la que proporcionara valores de β_0 y β_1 que minimizaran esta cantidad global. Sin embargo, debe tenerse en cuenta que, una recta que se situara de la forma descrita en la figura 5-5, verificaría que los e_i correspondientes a valores por encima de la recta serían positivos y se cancelarían con los e_i de los valores por debajo de la recta. Para resolver

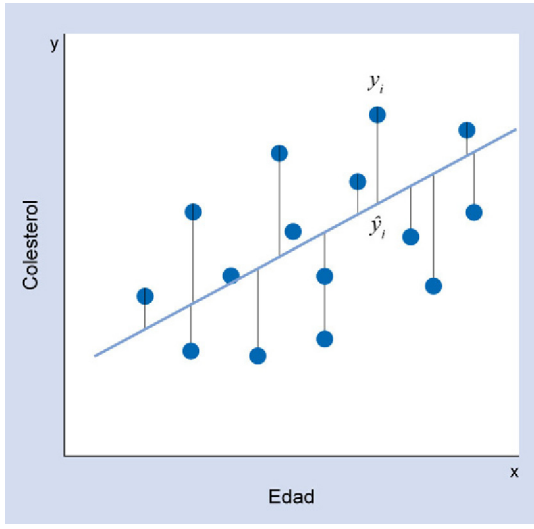

FIGURA 5-5 Ajuste de la recta de regresión lineal simple. Minimización de $y_i - \hat{y}_i$.

este problema se adopta el criterio de elevar al cuadrado cada uno de los errores y conseguir que todos tomen valores positivos, con lo que se tendría en cuenta la magnitud del error en cada observación (error al cuadrado en este caso) pero no el signo. El objetivo sería hallar los valores de β_0 y β_1 que minimicen la cantidad:

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

Este método para la estimación de los coeficientes del modelo β_0 y β_1 se conoce como el *método de mínimos cuadrados* (se minimiza una suma de cantidades/errores al cuadrado). Obsérvese que en esta expresión todo son números conocidos excepto β_0 y β_1. Para obtener la solución final debe derivarse parcialmente $\Sigma[y_i - (\beta_0 + \beta_1 x)]^2$ con respecto a β_0 y a β_1 e igualarse a 0, que es la forma en la que se determina la existencia de un máximo o un mínimo para una función cualquiera. De estas dos expresiones se obtendrá un sistema de dos ecuaciones con dos incógnitas. Los valores finales para β_0 y β_1 vendrán determinados por las siguientes expresiones:

$$\beta_1 = \frac{\mathrm{Cov}(x,y)}{S_x^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Ejemplo 5-1

En un estudio se obtuvo información sobre la edad y el nivel de colesterol de 12 pacientes, resultados presentados en la tabla 5-1.

Se pretende estudiar la posible relación entre el nivel de colesterol y la edad del individuo. El diagrama de dispersión (fig. 5-6) muestra una ligera inclinación hacia arriba de la nube de puntos que sugiere la posibilidad de una relación lineal directa. Se calculará el coeficiente de correlación lineal de Pearson para valorar la magnitud de la asociación lineal y la recta de regresión lineal para estudiar la naturaleza de la misma.

Se calculan en primer lugar las medias y desviaciones típicas de cada una de las variables:

$$\overline{x} = \frac{\sum x_i}{n} = \frac{(34+57+27+34+\ldots+47)}{12} = 45$$

$$\overline{y} = \frac{\sum y_i}{n} = \frac{(235+322+185+299+\ldots+193)}{12} = 228,9$$

$$S_x = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}} = \sqrt{\frac{(34-45)^2+(57-45)^2+\ldots+(47-45)^2}{12}} = 16,4$$

$$S_y = \sqrt{\frac{\sum (y_i - \overline{y})^2}{n}} = \sqrt{\frac{(235-228,9)^2+(322-228,9)^2+\ldots+(193-228,9)^2}{12}} = 47,1$$

TABLA 5-1 Edad y nivel de colesterol de un grupo de 12 individuos

Edad	Colesterol
34	235
57	322
27	185
34	299
26	198
51	266
70	188
30	203
75	235
31	164
58	259
47	193

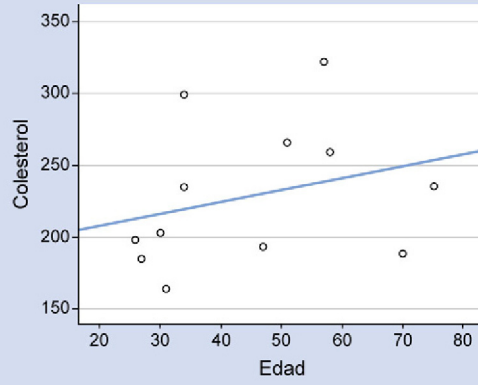

FIGURA 5-6 Diagrama de dispersión y recta de regresión lineal ajustada.

A partir de estos resultados se calcula el valor de la covarianza que quedará como sigue a continuación:

$$\text{Cov}(x,y)=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n}$$
$$=\frac{(34-45)(235-228,9)+\ldots+(47-45)(193-228,9)}{12}=221,3$$

Como puede observarse, el valor de la covarianza es positivo, resultado coherente con la previsión de posible relación lineal directa entre las dos variables a la que se llegaba a partir del diagrama de dispersión. El coeficiente de correlación lineal quedará:

$$r=\frac{\text{Cov}(x,y)}{S_xS_y}=\frac{221,3}{16,4\cdot 47,1}=0,29$$

Aprovechando los cálculos realizados anteriormente, para obtener los valores de los coeficientes del modelo de regresión lineal simple β_0 y β_1 se tendrá que:

$$\beta_1=\frac{\text{Cov}(x,y)}{S_x^2}=\frac{221,3}{16,4^2}=0,823$$

$$\beta_0=\bar{y}-\beta_1\bar{x}=228,9-0,823\cdot 45=191,6$$

El modelo de regresión lineal simple que expresa el valor del nivel de colesterol en función del valor de la edad del individuo quedará:

$$\text{Colesterol}=191,6+0,823\cdot\text{Edad}$$

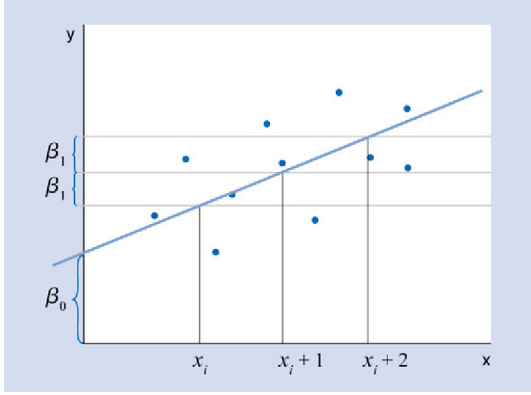

FIGURA 5-7 Interpretación del coeficiente β_1 de la recta de regresión.

INTERPRETACIÓN DE LOS COEFICIENTES DEL MODELO DE REGRESIÓN LINEAL SIMPLE

El coeficiente β_0 indica, simplemente, el punto de corte de la recta de regresión con el eje de ordenadas (obsérvese que si la edad fuera 0, el nivel de colesterol sería 191,6), mientras que el coeficiente β_1 es quien recoge el efecto de la variable independiente sobre la variable dependiente (fig. 5-7). ¿Qué quiere decir esto?

Utilizando los datos del ejemplo 5-1 se tiene que $\beta_0 = 191{,}6$ y $\beta_1 = 0{,}823$.

Si se aumenta progresivamente en una unidad (en este caso 1 año) el valor de la variable independiente (en este caso la edad) se tiene que:

$$\left.\begin{array}{l}\text{Si } Edad = 35 \rightarrow Colesterol = 191{,}6 + 0{,}823 \cdot 35 = 220{,}405\\ \text{Si } Edad = 36 \rightarrow Colesterol = 191{,}6 + 0{,}823 \cdot 36 = 221{,}228\end{array}\right\} 0{,}823$$

$$\left.\begin{array}{l}\text{Si } Edad = 36 \rightarrow Colesterol = 191{,}6 + 0{,}823 \cdot 36 = 221{,}228\\ \text{Si } Edad = 37 \rightarrow Colesterol = 191{,}6 + 0{,}823 \cdot 37 = 222{,}051\end{array}\right\} 0{,}823$$

Como puede observarse, por cada año más del individuo el nivel de colesterol que le pronostica la recta aumenta en 0,823 unidades que coincide con el valor de β_1. Luego el coeficiente β_1 se interpreta como el *cambio en la variable dependiente por unidad de cambio en la variable independiente* que, a su vez, coincide con la pendiente de la recta de regresión lineal simple. Esta es, por tanto, la naturaleza de la relación entre las dos variables cuantitativas estudiadas.

Por otra parte, obsérvese que el coeficiente de correlación lineal de Pearson y el coeficiente β_1 de la recta de regresión están íntimamente relacionados, ya que la Cov(x,y) interviene en el cálculo de ambas cantidades y con un papel protagonista. Dado que los denominadores de ambas cantidades

(r y β_1) son siempre positivos, su signo vendrá determinado por el valor de la Cov(x,y). De este modo, si la covarianza es positiva (lo que indicaba una posible relación lineal directa) también lo será el valor del coeficiente de correlación lineal de Pearson y el valor β_1 que es la pendiente de la recta. Si por el contrario el valor de la *Cov(x,y)* es negativo, también lo serán los valores de *r* y β_1.

BONDAD DEL AJUSTE DEL MODELO DE REGRESIÓN LINEAL SIMPLE

Hasta el momento se ha incidido en el ajuste e interpretación del modelo de regresión lineal simple. Sin embargo, se ha de tener presente que, para un conjunto de datos cualesquiera, este procedimiento proporcionará siempre una recta, con independencia de que ajuste bien o no a los datos. Es necesario, por tanto, profundizar en la construcción de una medida que ofrezca información sobre la bondad del ajuste de la recta a los datos observados.

Como punto de partida, se tendrá en cuenta que la distancia de cada observación de la variable dependiente y_i a la media de las observaciones de la variable $\overline{y}$ para un determinado valor de *x* puede expresarse de la forma:

$$y_i - \overline{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})$$

que en el diagrama de dispersión se representaría de la forma descrita en la figura 5-8.

Una forma de obtener un resumen de las distancias consistiría en calcular la suma de todas las distancias (una para cada observación):

$$\sum (y_i - \overline{y}) = \sum (y_i - \hat{y}_i) + \sum (\hat{y}_i - \overline{y})$$

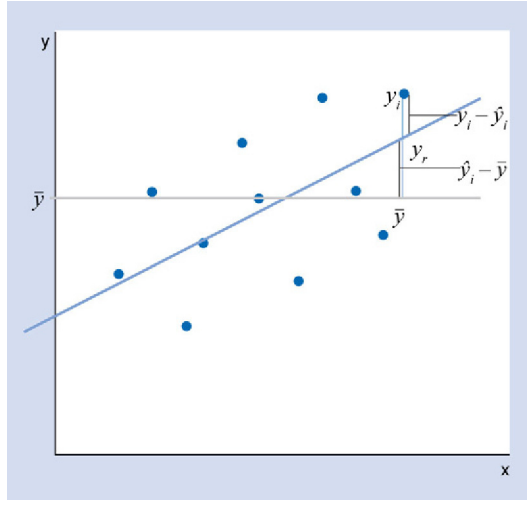

FIGURA 5-8 Descomposición de la distancia de cada observación a la media.

Sin embargo, dado que la media de las observaciones se sitúa en el centro de gravedad de los datos, las distancias positivas cancelarían las distancias negativas y esta cantidad sería siempre 0. Para eliminar el efecto del signo pero conservar la magnitud de la distancia se eleva al cuadrado, de forma que se obtiene la siguiente expresión:

$$(y_i - \overline{y})^2 = (y_i - \hat{y}_i)^2 + (\hat{y}_i - \overline{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$$

Si se calcula la suma de todas estas distancias, ahora al cuadrado y por tanto positivas, se tendrá que:

$$\sum(y_i - \overline{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \overline{y})^2 + 2\sum(y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$$

Dado que se verifica:

$$\sum(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = 0$$

la expresión final quedará:

$$\sum(y_i - \overline{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \overline{y})^2$$

Esta expresión se conoce como la *descomposición de la variabilidad*, ya que, como puede observarse, a la izquierda de la igualdad se tiene el numerador de la varianza total de la variable dependiente *y* expresada como suma de dos variabilidades, donde:

$\sum(y_i - \hat{y}_i)^2$ = Variabilidad no explicada por el modelo de regresión. Apréciese que resume las distancias al cuadrado entre el verdadero valor observado de la variable dependiente y el valor que pronostica la recta. Se interpreta como la varianza no explicada por el modelo o varianza residual y se denota VNE.

$\sum(\hat{y}_i - \overline{y})^2$ = Variabilidad explicada por el modelo de regresión. Obsérvese que resume las distancias entre el valor promedio de la variable dependiente (valor que se proporcionaría para estimar el valor de *y* si no se tuviera en cuenta el modelo de regresión) y el valor que pronostica la recta. Se interpreta como la varianza explicada por el modelo o varianza de la regresión y se denota VE.

$\sum(y_i - \overline{y})^2$ = Variabilidad total observada en la variable dependiente. Adviértase que coincide con el numerador de la varianza de la variable *y*. Se denota VT.

En consecuencia, se tiene que:

$$VT = VNE + VE$$

COEFICIENTE DE DETERMINACIÓN DE LA RECTA

Dado que la variabilidad total observada en la variable dependiente ha podido ser expresada (descompuesta) como la suma de dos variabilidades (la variabilidad no explicada por el modelo de regresión y la variabilidad explicada por el modelo de regresión), puede obtenerse una medida de la bondad del ajuste del modelo calculando la proporción de variabilidad explicada. Para ello, será suficiente con dividir la variabilidad explicada entre la variabilidad total. Se tendrá entonces que:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{VE}{VT}$$

Esta cantidad se conoce como el *coeficiente de determinación* de la recta de regresión lineal simple. Puede ser interpretada como la proporción de variabilidad explicada por el modelo (recta de regresión) o, si se multiplica el resultado por 100, como porcentaje de variabilidad explicada por el modelo. Debe tenerse en cuenta que, como toda proporción, toma valores entre 0 y 1 (entre 0 y 100% si se opta por la expresión en forma de porcentaje).

En la tabla 5-2 pueden observarse los pasos necesarios para la obtención de las sumas de cuadrados, utilizando los datos del ejemplo 5-1.

De donde:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{2.185,9}{26.664,9} = 0,082$$

TABLA 5-2 Descomposición de la variabilidad. Sumas de cuadrados. Datos del ejemplo 5-1

x = edad	y = colesterol	$\hat{y}$ = 191,6 + 0,823 x	$(\hat{y}_i - \bar{y})^2$	$(y_i - \bar{y})^2$
34	235	219,58	86,83	37,2
57	322	238,51	92,37	8667,6
27	185	213,82	227,38	1927,2
34	299	219,58	86,83	4.914
26	198	213	252,87	954,8
51	266	233,57	21,84	1.376,4
70	188	249,21	412,5	1.672,8
30	203	216,29	159,01	670,8
75	235	253,33	596,58	37,2
31	164	217,11	138,93	4.212
58	259	239,33	108,87	906
47	193	230,28	1,91	1.288,8
			Σ = 2.185,9	Σ = 26.664,9

Luego podría concluirse que el modelo consigue explicar en torno al 8,2% de la variabilidad observada en la variable *nivel de colesterol*. Dado que en el modelo solo se dispone de una variable explicativa, toda la explicación estimada es atribuible a ella, por lo que puede decirse que la edad ha logrado explicar el 8,2% de la variabilidad del nivel de colesterol de los individuos observados.

INFERENCIA SOBRE EL MODELO DE REGRESIÓN LINEAL SIMPLE

Cuando se construye un modelo de regresión lineal simple a partir de los datos contenidos en una muestra aleatoria de la población, debe tenerse en cuenta que el modelo ajustado no es más que uno de todos los posibles modelos que podrían ajustarse a partir de cada una de las muestras posibles de la población que hubieran podido ser seleccionadas en el proceso de muestreo. Esto es, para cada muestra se ajustaría un modelo que podría ser similar, pero que mostrará diferencias en el valor de sus coeficientes β_0 y β_1 al variar los datos de partida.

En la figura 5-9 puede observarse que tanto el punto de corte con el eje de ordenadas β_0 como la pendiente de la recta β_1 varían en función de los datos de la muestra correspondiente. En consecuencia, y al igual que sucedía con el coeficiente de correlación lineal de Pearson, los valores de β_0 y β_1, así como el del coeficiente de determinación de la recta R^2, serán variables aleatorias que varían de muestra a muestra de la población.

Un contraste de interés trataría de establecer si el modelo de regresión lineal simple explica de forma significativa parte de la variabilidad observada en la variable dependiente. En caso afirmativo, toda la explicación sería atribuible a la variable explicativa o independiente ya que es la única variable introducida en el modelo.

Se plantea, por tanto, un contraste sobre el coeficiente de determinación poblacional de la recta ρ^2 como el siguiente:

$$H_0 : \rho^2 = 0$$
$$H_1 : \rho^2 \neq 0$$

En caso de aceptación de la hipótesis nula no habría evidencia de que el coeficiente de determinación poblacional sea significativamente distinto de 0. Por tanto, la variabilidad explicada podría ser 0 y el modelo no explicaría nada. En caso de rechazar la hipótesis nula el coeficiente de determinación sería significativamente distinto de 0 y el modelo explicaría, de forma significativa, parte de la variabilidad observada en la variable dependiente.

Es importante tener en cuenta que no se plantea que el modelo explique una cantidad importante de la variabilidad, sino si explica una parte, aunque sea pequeña, de la misma, pero de forma significativa. Esto querría decir que la variable independiente muestra un efecto significativo sobre la variable dependiente que podría cuantificarse mediante un intervalo de confianza para el coeficiente de determinación.

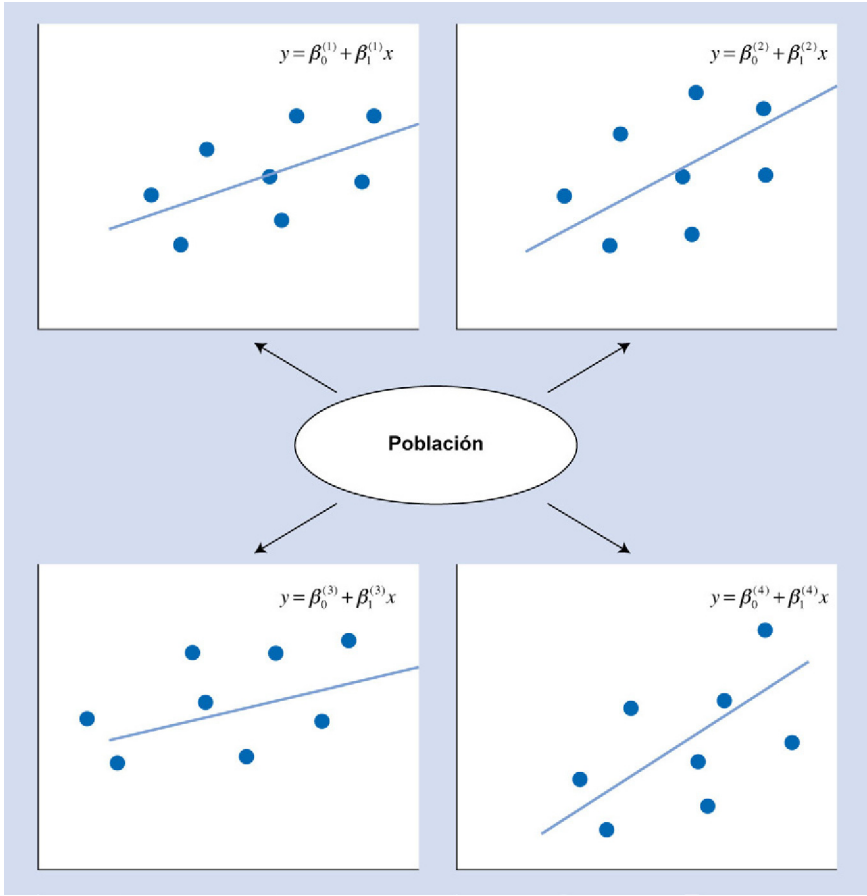

FIGURA 5-9 Diferencias en el modelo de regresión lineal ajustado en función de la muestra seleccionada.

El estadístico de contraste es, en este caso:

$$EC = \frac{VE/1}{VNE/(n-2)} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \hat{y}_i)^2/(n-2)}$$

La distribución muestral asociada sería un F de Snedecor con 1 y n − 2 grados de libertad, en el caso de verificarse las hipótesis necesarias que se abordarán posteriormente. Como puede observarse, si la variabilidad explicada por el modelo es 0 el valor del estadístico de contraste será 0 (situación compatible con la hipótesis nula de no explicación significativa). Por otro lado, si la variabilidad explicada comienza a aumentar (crece el numerador), la variabilidad no explicada tendrá que disminuir (disminuye

TABLA 5-3 Tabla de ANOVA de la regresión

Fuente	Suma de cuadrados	Media de cuadrados	Cociente de varianzas
Regresión (VE)	$\sum(\hat{y}_i - \bar{y})^2$	$\frac{\sum(\hat{y}_i - \bar{y})^2}{1}$	$\frac{\sum(\hat{y}_i - \bar{y})^2 / 1}{\sum(y_i - \hat{y}_i)^2 / (n-2)}$
Residual (VNE)	$\sum(y_i - \hat{y}_i)^2$	$\frac{\sum(y_i - \hat{y}_i)^2}{n-2}$	
Total (VT)	$\sum(y_i - \bar{y})^2$	$\frac{\sum(y_i - \bar{y})^2}{n-1}$	

el denominador) con lo que el valor del estadístico de contraste crecerá cada vez más hasta el punto de poder rechazar, en su caso, la hipótesis nula.

TABLA DE ANOVA DE LA REGRESIÓN

Para representar los datos anteriores, suele ser habitual, proporcionar una tabla conocida como la tabla de ANOVA de la regresión (tabla 5-3). Esta tabla reproduce la secuencia de razonamiento, que culmina con la construcción de un contraste sobre el coeficiente de determinación poblacional ρ^2 de la recta de regresión que inició su andadura en la descomposición de la variabilidad (obsérvese la segunda columna de la tabla). Al dividir cada una de las sumas de cuadrados por sus grados de libertad se obtienen las varianzas correspondientes (obsérvese la columna *media de cuadrados*). Finalmente se construye el estadístico de contraste como el cociente entre la varianza explicada y la no explicada por el modelo de regresión corregidas por sus grados de libertad (medias de cuadrados) que seguirá una distribución F de Snedecor.

En la tabla 5-4 se muestra la tabla de ANOVA para los datos del ejemplo 5-1. El estadístico de contraste sería 0,89 que, comprobado en las tablas de la F de Snedecor con 1 y n − 2 grados de libertad, proporcionaría un valor de la p de 0,367. Dado que el valor de la *p* del contraste es superior al nivel habitual 0,05, se aceptaría la hipótesis nula y se concluiría que el modelo no explica y, por tanto, que la variable edad no tiene un efecto significativo sobre el nivel de colesterol.

TABLA 5-4 Tabla de ANOVA de la regresión para los datos del ejemplo 5-1

Fuente	Suma de cuadrados	Media de cuadrados	Cociente de varianzas
Regresión (VE)	2.185,9	2.185,9	$\frac{2.185,9}{2.447,9} = 0,89$
Residual (VNE)	24.479,2	$\frac{24.479,2}{12-2} = 2.447,92$	
Total (VT)	26.664,9		

REQUERIMIENTOS SOBRE EL MODELO DE REGRESIÓN LINEAL SIMPLE

La propuesta de un modelo de regresión lineal para el estudio de la relación entre dos variables cuantitativas normalmente va más allá de la posible relación detectada en los datos observados (muestra), ya que trata de extraer conclusiones sobre la población general de la que los datos observados no son más que una pequeña parte. Por tanto, los requisitos necesarios para la realización de inferencias a partir del modelo de regresión lineal simple contendrán hipótesis sobre la pertinencia de la relación lineal entre las dos variables estudiadas e hipótesis sobre los datos y su distribución que permitan la realización de las inferencias deseadas. Estas hipótesis podrían resumirse en las siguientes:

- Pertinencia de la linealidad.
- Homocedasticidad.
- Normalidad.
- Independencia de las observaciones.

PERTINENCIA DE LA LINEALIDAD

Esta hipótesis requiere que el modelo de regresión lineal sea pertinente para el estudio de la relación lineal entre las dos variables cuantitativas consideradas.

En la figura 5-10A puede observarse que la media de la variable dependiente *y* para cada uno de los valores de la variable independiente $x\left(\bar{y} \mid x_i\right)$ se sitúa sobre la recta de regresión lineal. Matemáticamente esta hipótesis de pertinencia de la linealidad podría expresarse:

$$E\left(y \mid x\right) = \beta_0 + \beta_1 x$$

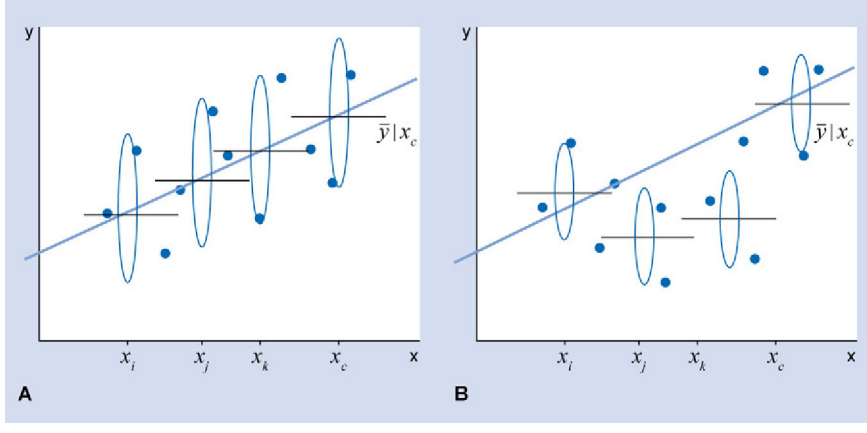

FIGURA 5-10 Hipótesis de pertinencia de la linealidad.

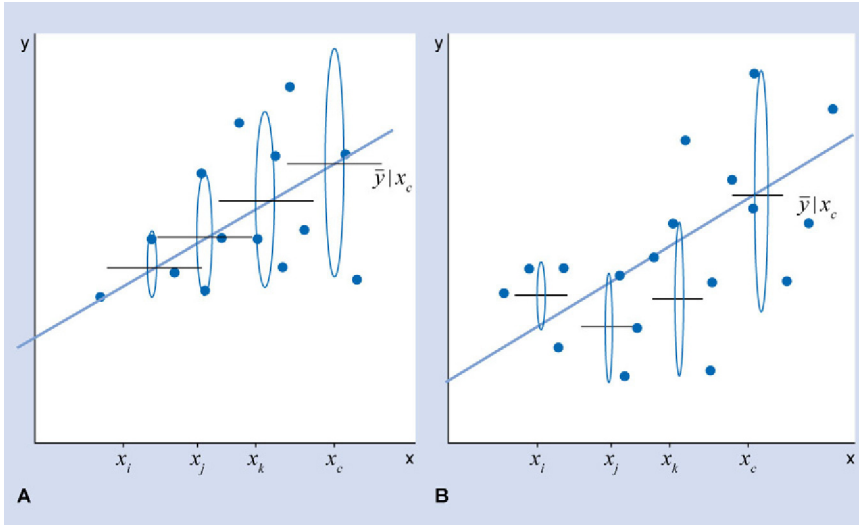

FIGURA 5-11 Hipótesis de homocedasticidad.

Sin embargo, en la figura 5-10B puede observarse que las medias de y para cada uno de los valores de la variable independiente $x\left(\bar{y}\,|\,x_i\right)$ no se encuentran sobre la recta de regresión lineal en todos los casos, violándose la hipótesis de linealidad. Nótese que la imagen del diagrama de dispersión sugiere en este caso una relación no lineal (p. ej., cuadrática) entre las variables consideradas.

HOMOCEDASTICIDAD

La hipótesis de homocedasticidad implica que la varianza de la variable dependiente y debe ser constante, es decir, para cualquier valor de la variable explicativa x la varianza de y será la misma.

En la figura 5-10 la varianza permanece constante para todos los valores de x. Sin embargo, en la figura 5-11 la varianza es distinta según el valor de la variable independiente. Obsérvese que en ambos casos la varianza aumenta con el valor de x, por lo que se violaría la hipótesis de homocedasticidad. Además, nótese que en la figura 5-11A se verifica la hipótesis de linealidad, pero en la figura 5-11B se violarían ambas hipótesis (relación posiblemente cuadrática y con varianza creciente). Matemáticamente la hipótesis de homocedasticidad se expresaría de la siguiente forma:

$$\mathrm{Var}\left(y\,|\,x\right)=\sigma^2$$

Puede observarse que, según la expresión, aunque varíe el valor de la variable independiente x el valor de la varianza de la variable dependiente y permanecerá constante. La falta de homocedasticidad influye en la

varianza de los estimadores, invalidando las expresiones del contraste F del ANOVA de la regresión.

NORMALIDAD

La hipótesis de normalidad requiere que la distribución de las observaciones de la variable dependiente *y* sea normal para cada uno de los valores de la variable independiente *x*.

En la figura 5-12A puede observarse que la distribución de la variable dependiente *y* es la normal para todo valor de la variable independiente *x*. Sin embargo, en la figura 5-12B la distribución de las observaciones de la variable *y* varía en función del valor de la variable *x*. Matemáticamente la hipótesis de normalidad se expresaría:

$$y \mid x \sim \text{Normal}$$

Las hipótesis de pertinencia de la linealidad, homocedasticidad y normalidad pueden expresarse matemáticamente y de forma unificada de la siguiente forma:

$$y \mid x \sim \text{Normal}\left(\beta_0 + \beta_1 x; \sigma^2\right)$$

La principal consecuencia del incumplimiento de la hipótesis de normalidad es la falta de eficiencia de los estimadores (no son de mínima varianza) y, por tanto, su efecto sobre los intervalos de confianza y contrastes de hipótesis sobre los parámetros del modelo.

INDEPENDENCIA

Se requiere que las observaciones sean independientes, requisito exigido en muchas de las pruebas estadísticas abordadas con anterioridad.

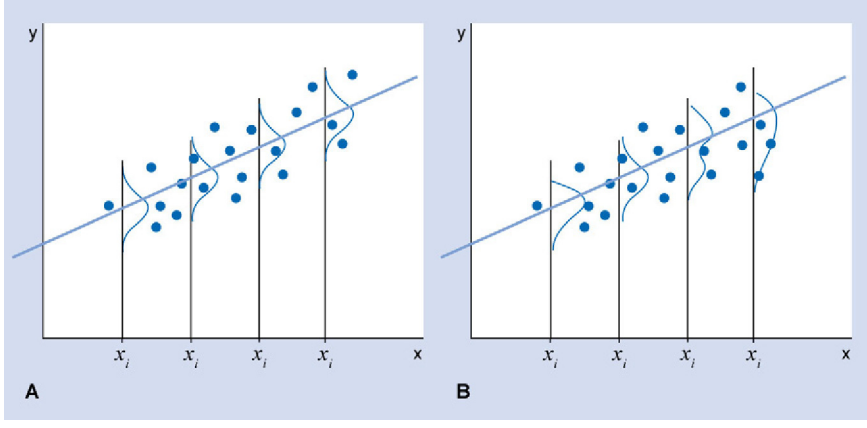

FIGURA 5-12 Hipótesis de normalidad.

DIAGNOSIS DEL MODELO. ANÁLISIS DE LOS RESIDUOS

Tradicionalmente las hipótesis sobre el modelo de regresión lineal han sido comprobadas mediante el análisis de los residuos del modelo. Las hipótesis se reformularían de la siguiente forma:

- Linealidad: $E(e \mid x) = 0$.
- Homocedasticidad: $Var(e \mid x) = cte$.
- Normalidad: $e \mid x \sim$ normal.
- Independencia: e_i, e_j independientes para cualquier i, j.

HIPÓTESIS DE LINEALIDAD Y HOMOCEDASTICIDAD

Para el estudio de las dos primeras hipótesis (linealidad y homocedasticidad) resulta de utilidad construir un gráfico de dispersión con las variables $\hat{y}$ (valor que predice la recta) en el eje de abscisas y la variable e (error o residuo del modelo) en el eje de ordenadas. En la figura 5-13A puede observarse una situación en la que se verificarían las dos hipótesis (los datos se disponen alrededor de la horizontal en 0 que actúa de forma similar a un eje de simetría y, además, se sitúan en una banda que se mantiene constante a lo largo de toda la recta).

En la figura 5-13B puede observarse que se viola la hipótesis de linealidad (la media de los residuos no es 0 a lo largo de la recta) aunque se mantiene la hipótesis de homocedasticidad (los datos mantienen una variabilidad constante a lo largo de la recta).

En la figura 5-13C se muestra una situación en la que se verifica la hipótesis de linealidad pero no la de homocedasticidad. Mientras que en la figura 5-13D se violan ambas hipótesis.

Analíticamente puede completarse el análisis gráfico de la linealidad comprobando que la media de los residuos del modelo es 0, $\Sigma e_i = 0$ (aunque esto no sería suficiente porque se precisaría que fuera 0 para cada valor de la variable independiente x).

Para la evaluación analítica de la homocedasticidad existen varias posibilidades, como los contrastes de Gresjer, Brensh-Pagan o White. Una aproximación sencilla al estudio de la homocedasticidad, si se sospecha de un crecimiento más o menos lineal de la varianza a medida que aumenta el valor de la variable independiente x, consiste en construir el siguiente modelo de regresión lineal:

$$|e| = \beta_0 + \beta_1 x$$

En el caso descrito en la figura 5-13C se observa un crecimiento lineal de la varianza a medida que se avanza en el valor predicho por la recta y, en consecuencia, de la variable independiente x. Al considerar el valor absoluto de los residuos del modelo, el gráfico de dispersión quedaría como se describe en la figura 5-14.

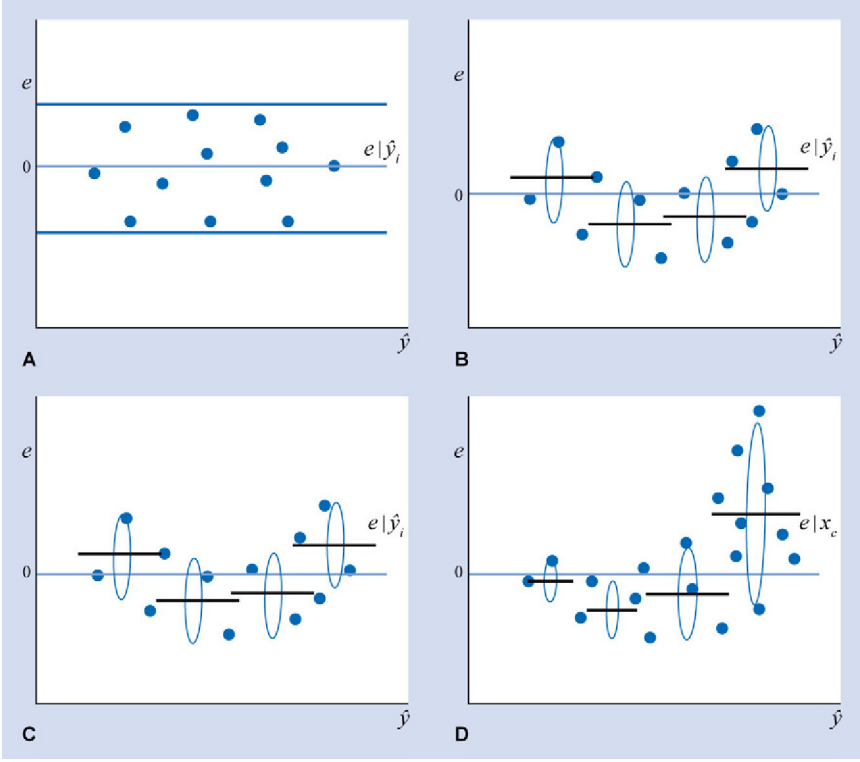

FIGURA 5-13 Diagnosis del modelo. Análisis de los residuos.

Como puede observarse en la figura 5-13C, la pendiente de la recta que ajustaba los residuos era 0. Al trabajar con el valor absoluto de los residuos se consigue que todos los datos queden por encima de la horizontal en 0 (v. fig. 5-14A). En el caso de no existencia de un aumento de la varianza, la recta que ajustaría estos nuevos datos sería otra recta en la horizontal (en un valor mayor que 0). Sin embargo, en este caso la recta que ajusta los datos tiene una marcada pendiente hacia arriba que sugiere un aumento del valor de los residuos con el valor de x y, en consecuencia, de la varianza (v. fig. 5-14B). Si el contraste sobre el coeficiente de determinación de esta recta resulta significativo se concluiría una violación de la hipótesis de homocedasticidad.

HIPÓTESIS DE NORMALIDAD

Dado que en multitud de situaciones es poco probable obtener suficientes observaciones de la variable dependiente para un mismo valor de la variable independiente, la hipótesis de normalidad suele comprobarse sobre el total de residuos del modelo y no sobre los residuos para cada valor de

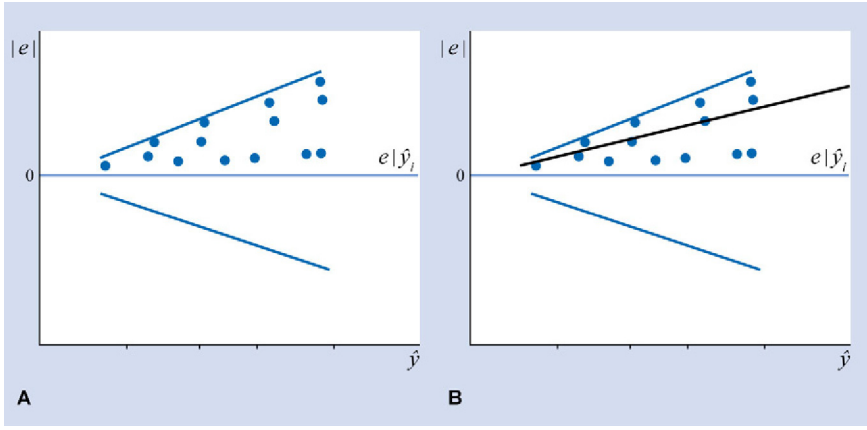

FIGURA 5-14 Estudio de la hipótesis de homocedasticidad. Posible crecimiento lineal de la varianza.

x. Para ello se construye el histograma de los residuos con superposición de la curva normal que más se le aproxime (fig. 5-15) y el gráfico P-P de probabilidad normal (fig. 5-16) que compara la función de distribución de los datos (observada) con la función de distribución del modelo normal (esperada).

El estudio de la normalidad puede completarse con el contraste de Kolmogorov-Smirnov que compara la función de distribución de los datos y la función de distribución normal. Debe tenerse en cuenta que, si se dispone de un número considerable de datos, este contraste puede proporcionar

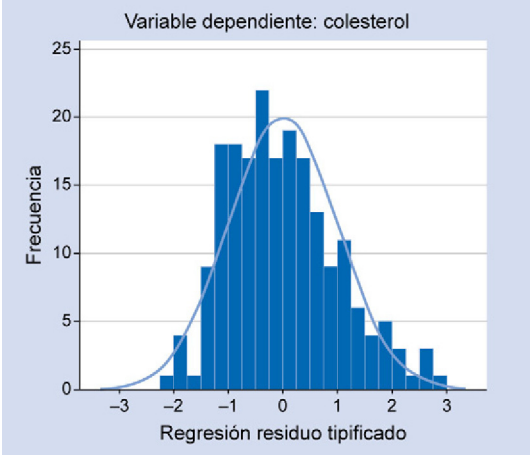

FIGURA 5-15 Histograma de los residuos.

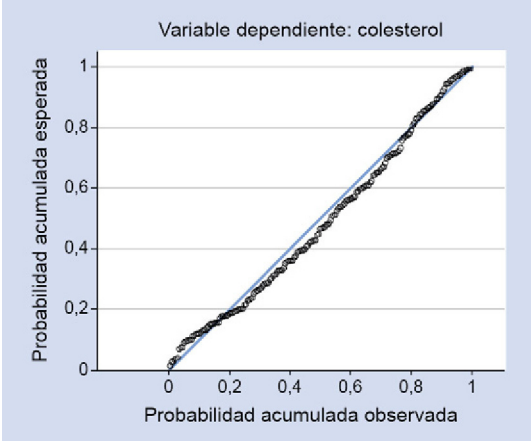

FIGURA 5-16 Gráfico P-P normal de los residuos.

una diferencia significativa con la distribución normal pero cuya magnitud sea despreciable y, en consecuencia, irrelevante para el cumplimiento de la hipótesis de normalidad, por lo que habrá que prestar atención a los gráficos de distribución anteriores.

De forma adicional puede utilizarse el diagrama de dispersión e (en el eje de ordenadas) comparado con $\hat{y}$ (en el eje de abcisas) para, de forma muy intuitiva, verificar si los datos muestran una distribución compatible con la distribución normal. Para ello debería observarse una mayor densidad de datos alrededor de la horizontal en 0 y una menor densidad de datos a medida que nos alejamos (arriba o abajo) de dicha horizontal.

HIPÓTESIS DE INDEPENDENCIA

Para verificar la hipótesis de independencia de forma gráfica debería representarse un diagrama de dispersión con la variable $\hat{y}$ en el eje de abscisas y la variable e en el eje de ordenadas, pero en el que las observaciones $(\hat{y}, e_i)$ se van incluyendo en el orden en que fueron obtenidas. Si se observa que valores de dichas observaciones por encima de la horizontal en 0 son sucedidas sistemáticamente por valores por encima de dicha horizontal habrá una autocorrelación positiva (fig. 5-17A), mientras que si valores por encima son sucedidos por valores por debajo y al contrario, la autocorrelación será negativa (v. fig. 5-17B).

Sin embargo, en muchas situaciones, la forma en la que se obtienen los datos no permite establecer un orden de las observaciones que tenga sentido y este tipo de gráficos no puede construirse. Habitualmente se recurre a un contraste de hipótesis de independencia de las observaciones basado en el estadístico de Durbin-Watson.

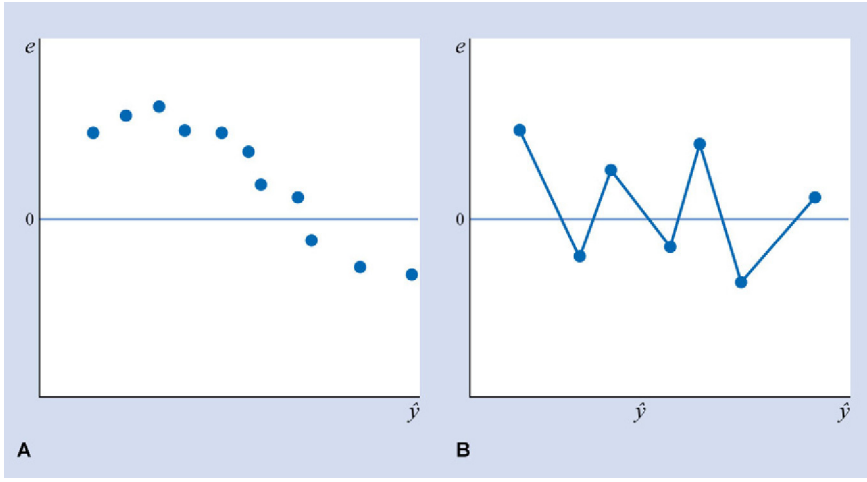

FIGURA 5-17 Estudio de la hipótesis de independencia. Autocorrelación de los residuos.

$$H_0 : \text{Los residuos son independientes}$$
$$H_1 : \text{Los residuos están autocorrelados}$$

El estadístico de contraste de Durbin-Watson se calcula de la siguiente forma:

$$EC = \frac{\sum (e_t - e_{t-1})^2}{\sum e_t^2} \simeq 2(1-r)$$

donde *r* es el coeficiente de autocorrelación de los residuos del modelo. Este estadístico toma valores entre 0 y 4.

$$\text{Si } r = 1 \rightarrow 2(1-r) = 2(1-1) = 0$$
$$\text{Si } r = 0 \rightarrow 2(1-r) = 2(1-0) = 2$$
$$\text{Si } r = -1 \rightarrow 2(1-r) = 2\left[1-(-1)\right] = 4$$

Valores próximos a 2 indicarán una ausencia de autocorrelación entre los residuos del modelo. Valores próximos a 4 indicarán una autocorrelación negativa, mientras que valores próximos a 0 indicarán una autocorrelación positiva. En la figura 5-18 se reflejan las zonas de decisión del contraste de independencia de los residuos para el estadístico de Durbin-Watson. Los valores de EC_U y EC_L pueden consultarse en tablas para el estadístico de Durbin-Watson.

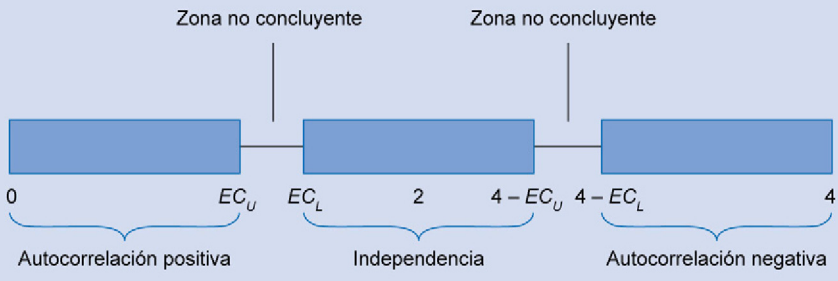

FIGURA 5-18 Zonas de decisión. Estadístico de Durbin-Watson.

Ejemplo 5-2

En un estudio se obtuvo información sobre la edad y el índice de masa corporal de un grupo de 200 pacientes. El objetivo es estudiar el posible efecto de la edad sobre el índice de masa corporal (IMC).

$$r = 0,408$$
$$R^2 = 0,167$$

La recta ajustada (figura 5-19) parece indicar una relación lineal directa entre las dos variables. Se propone trabajar según el esquema propuesto en la figura 5-20.

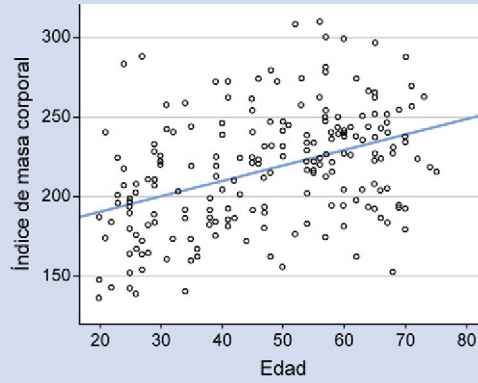

FIGURA 5-19 Recta de regresión lineal ajustada a los datos.

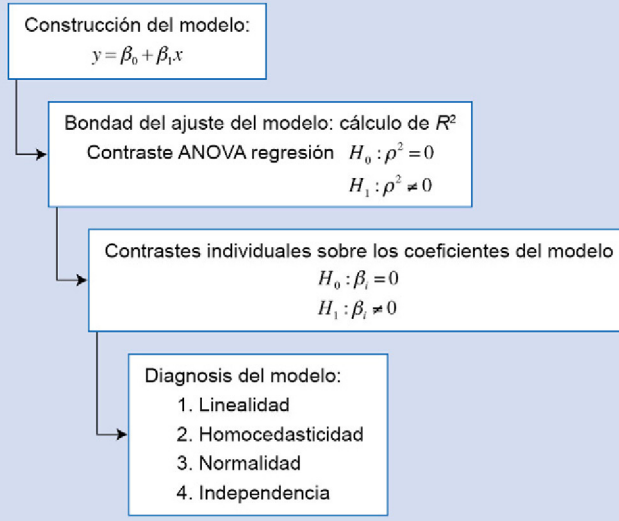

FIGURA 5-20 Esquema de construcción y análisis del modelo de regresión lineal.

Construcción del modelo

$$\text{IMC} = 22,086 + 0,098 \cdot \text{edad}$$

Se observa que en los datos observados el modelo de regresión lineal simple ajustado cuantifica el efecto de la edad sobre el índice de masa corporal de la siguiente forma: por cada año más del paciente el índice de masa corporal aumenta en 0,098 unidades.

Bondad del ajuste del modelo

Para valorar la bondad del ajuste del modelo se calcula el coeficiente de determinación de la recta. En este caso, se tiene que:

$$R^2 = 0,167$$

Esto quiere decir que el modelo logra explicar un 16,7% de la variabilidad observada en los índices de masa corporal de los pacientes estudiados. La magnitud de la asociación lineal entre las dos variables vendrá dada por el coeficiente de correlación lineal de Pearson:

$$r = \sqrt{R^2} = \sqrt{0,167} = 0,408$$

Sin embargo, el interés se centra en valorar si el modelo logra explicar de forma significativa una parte de la variabilidad de la variable *índice de masa*

TABLA 5-5 Tabla del ANOVA de la regresión

Modelo	Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	453,022	1	453,022	39,655 ,000
	Residual	2.261,986	198	11,424	
	Total	2.715,008	199		

corporal en la población de la que partió la muestra. Se plantea por tanto el contraste:

$$H_0 : \rho^2 = 0$$
$$H_1 : \rho^2 \neq 0$$

La tabla de ANOVA de la regresión (tabla 5-5) proporciona un valor del estadístico de contraste:

$$EC = \frac{VE/1}{VNE/(n-2)} = \frac{\sum(y_r - \bar{y})^2}{\sum(y_i - y_r)^2/(n-2)} = \frac{453,022}{11,424} = 39,655$$

Con un valor de la significación $p < 0,001$.

Por tanto, se rechazaría la hipótesis nula al nivel habitual $\alpha = 0,05$ y se concluirá que el modelo explica significativamente una parte de la variabilidad de la variable *índice de masa corporal.*

Contrastes individuales sobre los coeficientes del modelo

Este tipo de contrastes no serían necesarios en el caso de la regresión lineal simple (salvo que fuera de interés contrastar si la recta pasa o no por el origen de coordenadas o se pretendiera obtener los intervalos de confianza para los coeficientes) pero es conveniente introducirlos en este esquema de análisis porque alcanzarán un marcado protagonismo en los modelos de regresión lineal múltiple. Para entender esta cuestión, es necesario plantearse qué ocurriría si en el modelo de regresión lineal hubiera más de una variable explicativa. En este escenario podría darse el caso de que el modelo en conjunto explicara de forma significativa una parte de la variabilidad de la variable dependiente pero que no lo hicieran todas y cada una de las variables explicativas incluidas en el mismo, sino solo algunas de ellas. El contraste F del ANOVA de la regresión es, por tanto, un contraste de conjunto que requiere, posteriormente, identificar las variables que tienen un aporte significativo.

En el caso de la regresión lineal simple, al contar únicamente con una variable explicativa, toda la explicación del modelo es atribuible a dicha variable, con lo que el contraste de conjunto del modelo coincidirá con el contraste sobre el coeficiente correspondiente.

TABLA 5-6 Coeficientes[a] del modelo y contrastes correspondientes para los datos del ejemplo 5-2

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	22,086	,775		28,481	,000
	Edad	,098	,016	,408	6,297	,000

[a] *Variable dependiente: índice de masa corporal.*

En la tabla 5-6 se muestran los resultados obtenidos para los coeficientes del modelo.

El estadístico de contraste del coeficiente β_1 viene dado por la expresión:

$$EC = \frac{\beta_1}{\sqrt{Var(\beta_1)}} = \frac{0,098}{0,016} = 6,297$$

Que se distribuirá según un modelo de probabilidad $t_{n-2} = t_{198}$ que arroja un valor de $p < 0,001$, que coincide con la obtenida en el contraste de conjunto del ANOVA de la regresión.

Diagnosis del modelo

El diagrama de dispersión de la figura 5-21, muestra una dispersión de los residuos en torno a la horizontal en 0 que actúa a modo de eje de simetría entre la parte superior e inferior a medida que se avanza en valor pronosticado y, por tanto, en valor de la variable independiente, describiendo una situación

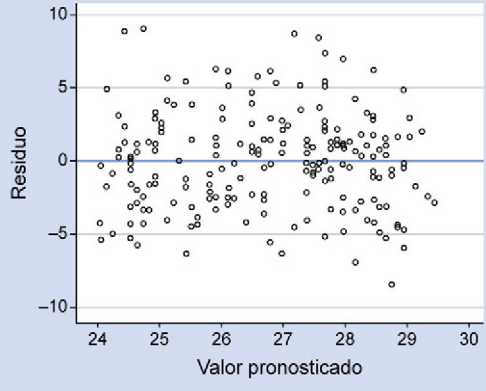

FIGURA 5-21 Diagrama de dispersión de los residuos frente al valor pronosticado.

compatible con la hipótesis de linealidad. Además, se verifica que la media global de los residuos es 0. Por otra parte, y analizando este mismo gráfico, se observa que la dispersión de los residuos permanece aproximadamente constante a lo largo de toda la recta.

$$\sum e_i = 2,52 \cdot 10^{-15} \simeq 0$$

Se completa esta imagen gráfica con el análisis del modelo de regresión que utiliza la variable $|e|$ como dependiente y la variable *edad* como independiente con objeto de detectar algún aumento o disminución progresivo de la varianza de los residuos. Así, se tiene que:

$$|e| = \text{Abs(Residuo)} = 2,937 - 0,005 \cdot \text{Edad}$$

$$R^2 = 0,002; p = 0,565$$

Como puede observarse, el efecto de la variable *edad* sobre el valor absoluto de los residuos es prácticamente 0 y no significativo (p = 0,565), por lo que se descarta que la varianza de los residuos aumente o disminuya linealmente en función de la edad, resultado en consonancia con el obtenido en el análisis gráfico anterior.

A continuación se proporcionan el histograma con superposición de la curva normal (fig. 5-22) para los residuos del modelo y el gráfico P-P normal (fig. 5-23), con objeto de comprobar la hipótesis de normalidad.

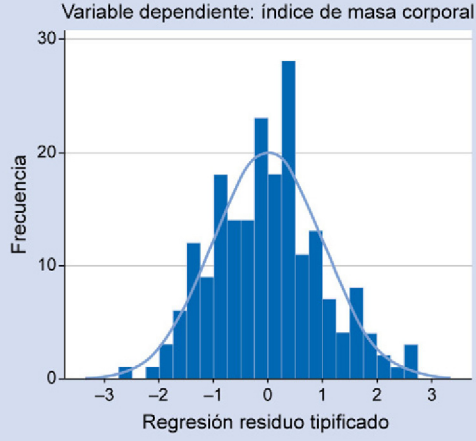

FIGURA 5-22 Histograma de los residuos del ejemplo 5-2.

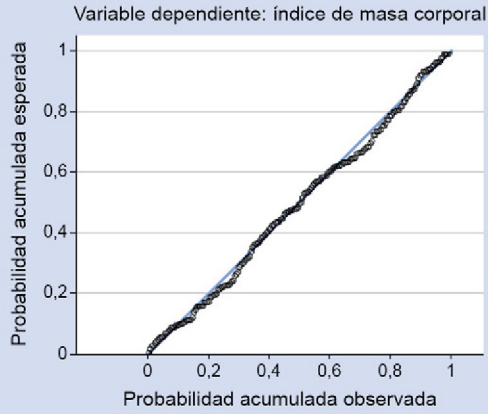

FIGURA 5-23 Gráfico P-P normal para los datos del ejemplo 5-2.

El histograma de los residuos muestra una distribución aproximadamente normal (aunque con algunos defectos o excesos de observaciones en la zona central y derecha de la distribución). Por su parte, la función de distribución observada se asemeja a la distribución normal teórica (obsérvese que en el gráfico P-P normal prácticamente se describe la diagonal). Se completa este resultado con la prueba de Kolmogorov-Smirnov para una muestra. La hipótesis nula será que la distribución es normal y la alternativa que no lo es. En este caso se obtiene el siguiente resultado:

$$EC = 0,656; p = 0,782$$

Por lo que no se puede rechazar la hipótesis nula de normalidad de los residuos. Debe recordarse que estos resultados se centran en la comprobación de la normalidad de los residuos en conjunto, y suponen una relajación de la hipótesis de normalidad de los residuos para cada uno de los valores de la variable independiente o, equivalentemente, del valor pronosticado por la recta. Una aproximación gráfica que permitiría, aunque de forma subjetiva, completar el estudio anterior valorando una distribución de los residuos compatible con la distribución normal a lo largo de toda la recta consistiría en observar en el diagrama de dispersión (e_i frente a $\hat{y}$) una mayor «densidad» de observaciones en torno a la horizontal en 0 y una menor densidad a medida que nos alejamos arriba o abajo. La figura 5-21 parece ilustrar esta situación.

Por último, el estadístico de Durbin-Watson tiene un valor de 2,023, muy próximo al valor 2 en el que se concluye ausencia de autocorrelación entre los residuos del modelo, por lo que se verificaría la hipótesis de independencia.

VALORES DE INFLUENCIA

El estudio de posibles valores de influencia en el modelo de regresión lineal ajustado es de gran importancia ya que, tanto la magnitud del efecto de la variable independiente, como la bondad del ajuste y su significación estadística, pueden verse seriamente afectadas.

En la figura 5-24A la recta de regresión ajustada a los datos observados es horizontal y, por tanto, la variable independiente no tendría ningún efecto sobre la variable dependiente al tener pendiente nula. La figura 5-24B representa el modelo ajustado en el caso de que se hubiera observado el dato A. Como puede observarse, la pendiente de la recta se modifica enormemente como consecuencia del «esfuerzo» del modelo por contemplar el dato A. En consecuencia, el efecto de la variable independiente podría pasar de nulo a significativo únicamente por una observación.

En general, el efecto de una variable sobre otra (no tiene por qué ser nulo) podría verse enormemente afectado por la presencia de algún o algunos valores de influencia sobre el modelo. Por otra parte, si estas observaciones influyentes corresponden, por ejemplo, a errores de observación, pueden resolverse de forma clara. En general la forma de proceder ante la presencia de valores de influencia (no siempre habría que excluirlos del análisis) dependerá de cada caso.

En primer lugar será útil analizar la existencia de observaciones atípicas en los datos analizados. Estas observaciones atípicas son susceptibles de jugar el papel de valores de influencia (son observaciones que, tal y como ocurre con el dato A, se alejan del resto de forma considerable), aunque no tendrían por qué ser necesariamente influyentes.

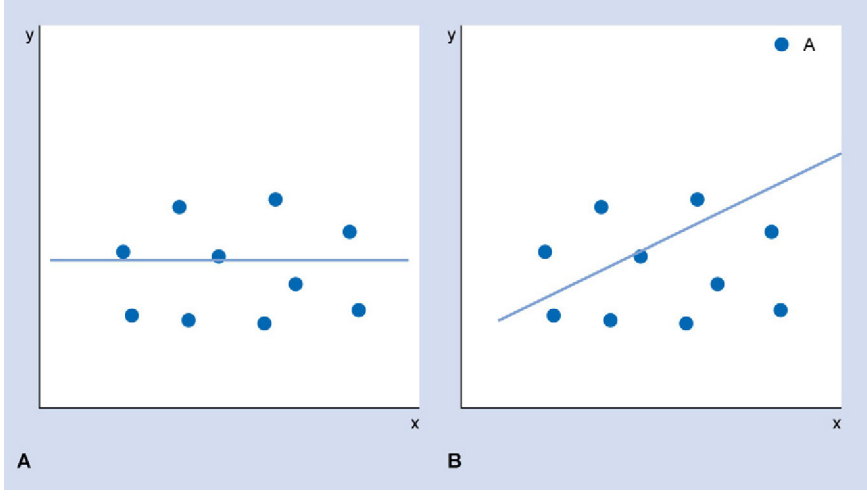

FIGURA 5-24 Efecto de un valor de influencia sobre el modelo ajustado.

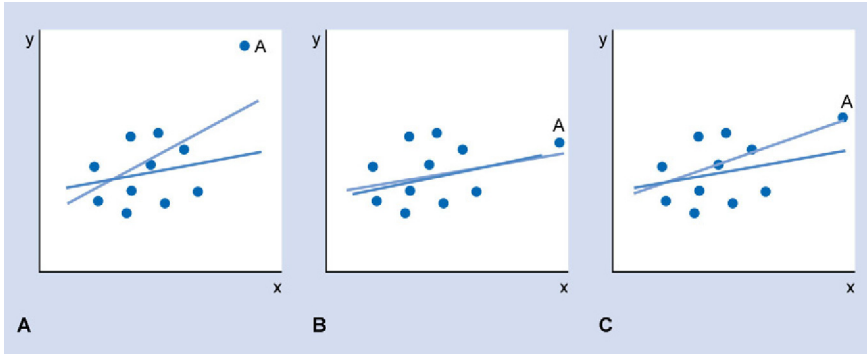

FIGURA 5-25 Observaciones atípicas influyentes y no influyentes.

En las figuras 5-25A y 5-25C el dato A es una observación atípica e influyente en el modelo (si se omitiera se modificaría la pendiente de la recta). Sin embargo, en la figura 5-24B el dato A es una observación atípica pero no influyente (si se omitiera no se modificaría apenas la pendiente de la recta de regresión). Es importante señalar que en todos los casos el punto A se aleja de forma considerable del conjunto de la nube de puntos (observación atípica) pero no de la misma forma. Mientras que en la figura 5-25A el dato A es un valor atípico con respecto a la variable y en las figuras 5-25B y 5-25C lo es respecto a la variable x.

Las observaciones atípicas con respecto a la variable y son candidatas a valores de influencia en el modelo, al presentar un valor del residuo sensiblemente superior al resto de observaciones, si bien puede darse el caso de que una observación sea influyente sin necesidad de ser atípica. Por otra parte, las observaciones atípicas con respecto a la variable x pueden ser influyentes si se alejan en exceso del conjunto de datos observados y su coordenada y, sin ser atípica, toma un valor relativamente alejado del que se obtendría en el ajuste de la recta sin dicha observación (v. fig. 5-25C).

Para la detección de observaciones atípicas con posible influencia en el modelo pueden adoptarse criterios sencillos, como los siguientes, que incluyen la mayoría de los programas de análisis estadístico:

- Identificar como casos atípicos aquellos cuyo valor del residuo correspondiente sea mayor de tres desviaciones típicas. Esto es: $e_i < -3 \cdot S_e$ o $e_i > 3 \cdot S_e$. En el caso de trabajar con los residuos estandarizados se identificarían como atípicos con posible influencia sobre el modelo aquellos que quedaran fuera del intervalo [–3,3].
- Sea d la distancia entre el percentil 25 (p_{25}) y el percentil 75 (p_{75}) de los residuos del modelo ($d = p_{75} - p_{25}$). Se identificará como valor atípico con posible influencia sobre el modelo aquel que verifique:

$$e_i < p_{25} - 1,5d \text{ o } e_i > p_{75} + 1,5d$$

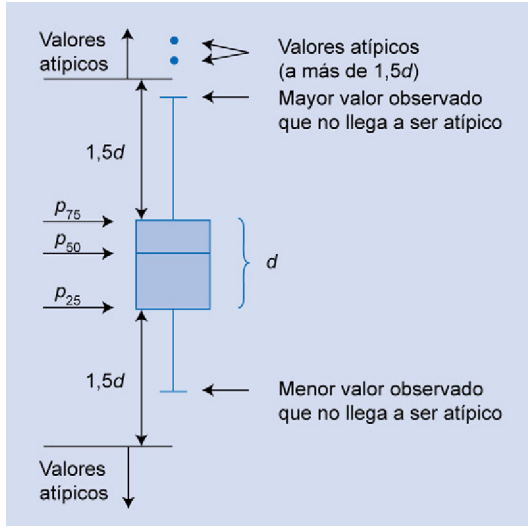

FIGURA 5-26 Diagrama de cajas *(box-plot)* y valores atípicos.

A los valores atípicos que superan 3*d* se les suele denominar «extremos». Para la identificación de valores atípicos con posible influencia sobre el modelo basados en el criterio anterior suele utilizarse el diagrama de cajas o *box-plot*, que se interpreta tal y como se describe en la figura 5-26.

Estas estrategias de detección de valores de influencia pueden completarse con medidas específicas. Los programas de análisis estadístico incorporan habitualmente medidas más sofisticadas para detectar valores atípicos, potencialmente influyentes e influyentes en el modelo como la distancia de Mahalanobis, apalancamiento *(leverage)*, distancia de Cook, DfBetas, DfAjuste y razón de covarianzas.

DISTANCIA DE MAHALANOBIS

La distancia de Mahalanobis es una medida estandarizada de la distancia de un punto (teniendo en cuenta únicamente las coordenadas de las variables predictoras o independientes) al centro de la nube de puntos que conforman las variables predictoras y que tiene en cuenta la correlación entre las mismas. Su expresión de cálculo en modo matricial es el siguiente:

$$D_i = (x_i - \bar{x})^T \Sigma^{-1} (x_i - \bar{x})$$

Donde Σ^{-1} es la inversa de la matriz de varianzas-covarianzas entre las variables predictoras o independientes. La distancia d Mahalanobis coincide con la distancia euclídea de un punto al centro cuando las coordenadas se calculan en base a las componentes principales. Valores atípicos de la

distancia de Mahalanobis indicarán que se encuentran más alejados del conjunto de las observaciones teniendo en cuenta la correlación.

Según la distancia de Mahalanobis para el caso de dos variables explicativas o predictoras, el punto A de la figura 5-27 estaría más alejado del punto $(\bar{x}_1, \bar{x}_2)$ que el punto B a pesar de que están exactamente a la misma distancia euclídea. Esto es debido a que el punto B se encuentra en el sentido de la correlación entre las dos variables.

Obsérvese que las rectas C_1 y C_2 representarían las dos componentes principales para los datos observados. La componente C_1 se dispone en el sentido de la correlación de forma que la distancia de Mahalanobis irá aumentando ligeramente a medida que nos alejamos del punto $(\bar{x}_1, \bar{x}_2)$ en esa dirección. Por el contrario, si nos alejamos en la dirección de la componente principal C_2, la distancia de Mahalanobis aumentará de forma muy importante a medida que nos alejamos del punto $(\bar{x}_1, \bar{x}_2)$. Se estarían penalizando, por tanto, las observaciones que no se disponen en el sentido de la correlación existente entre las dos variables.

APALANCAMIENTO (LEVERAGE)

El apalancamiento es una medida de la influencia de cada una de las observaciones de la variable dependiente sobre los valores ajustados por el modelo y toma valores en el intervalo [1/n;1]. Se calcula, para el caso de una variable explicativa, de la siguiente forma:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$

Como puede observarse, de la expresión puede deducirse que si el valor de la variable independiente *x* para un caso cualquiera está muy próximo a

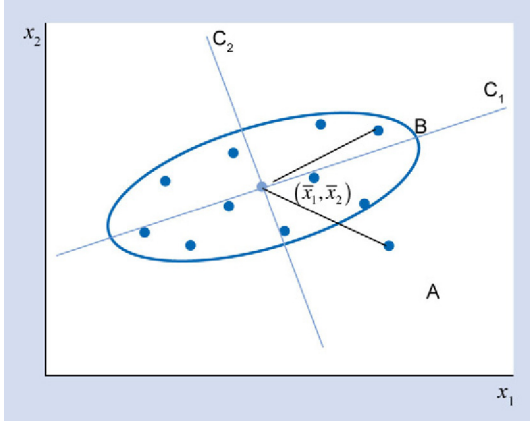

FIGURA 5-27 Distancia de Mahalanobis.

la media $\overline{x}$ su influencia en el modelo será nula (valor mínimo 1/n), mientras que a medida que se vaya alejando de la media su influencia será mayor y su valor se acercará a 1. Suele ser habitual trabajar con una medida centrada del apalancamiento que se consigue restando a la expresión anterior 1/n. De este modo quedará:

$$\textit{Apalancamiento centrado} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum (x_i - \overline{x})^2} - \frac{1}{n} = \frac{(x_i - \overline{x})^2}{\sum (x_i - \overline{x})^2}$$

Que tomará valores en el intervalo $\left[0, \frac{n-1}{n}\right]$.

Cuando el número de observaciones del que se dispone es elevado tomará valores entre [0,1].

En la figura 5-28 se representan dos situaciones en las que existe un valor que se aleja del conjunto de la nube de puntos. Como puede observarse, en la figura 5-28A el punto A no modifica prácticamente la pendiente de la recta ajustada, aunque tiene un efecto sobre el término de interceptación (en color claro se representa el ajuste sin el punto A y en color más oscuro con el punto A), mientras que en el caso de la figura 5-28B la pendiente (y, por tanto, el efecto de la variable independiente sobre la dependiente) se modifica de forma considerable. Obsérvese que en este segundo caso el valor de la coordenada *x* está muy alejado de la media $\overline{x}$. En consecuencia, valores alejados de la media de la variable independiente tendrán mayor influencia sobre la pendiente de la recta que los que están cerca.

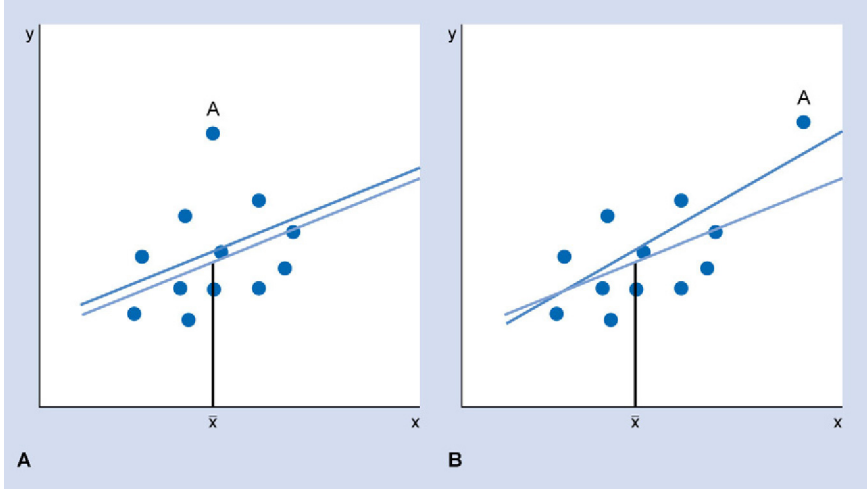

FIGURA 5-28 Apalancamiento *(leverage)*.

DISTANCIA DE COOK

La distancia de Cook es una medida de la influencia de una determinada observación sobre el modelo, medida en términos de distancia entre los parámetros ajustados del modelo con y sin esa observación. Su cálculo en forma matricial quedaría:

$$D_i = \frac{\left(\hat{\beta} - \hat{\beta}_{(i)}\right)' X' X \left(\hat{\beta} - \hat{\beta}_{(i)}\right)}{(k+1) S_r^2} = \frac{r_i^2}{k+1}\left(\frac{h_{ii}}{1-h_{ii}}\right)$$

donde $\hat{\beta}$ es el vector de parámetros del modelo, para el caso de una única variable explicativa (β_0, β_1,) ajustados con todas las observaciones, y $\hat{\beta}_{(i)}$ es el vector de parámetros del modelo ajustados sin la observación (i). Por otra parte, h_{ii} es el elemento correspondiente de la diagonal de la matriz $X(X'X)^{-1}X$, donde X representa la matriz de datos. Un valor próximo a 0 de esta distancia indicará que los parámetros del modelo prácticamente no varían cuando se elimina del ajuste a la observación (i) y, por tanto, la observación (i) no será influyente en el modelo. Un valor cada vez mayor indicará que existe mayor diferencia entre los parámetros ajustados con todas las observaciones y sin la observación (i) y, por tanto, la observación (i) tendrá una influencia considerable en el modelo. Es importante tener en cuenta que la distancia de Cook es una medida de conjunto que involucra a la vez a todos los parámetros del modelo, en este caso, β_0 y β_1.

DFBETAS

A diferencia del estadístico de Cook, que proporcionaba una medida del cambio conjunto de los coeficientes del modelo al excluir una determinada observación, los DfBetas constituyen una medida de la influencia de una observación en el modelo de regresión lineal medida en términos de cambio en cada uno de los coeficientes del modelo, al excluir una determinada observación. Por tanto, para cada una de las observaciones se obtendrán, en el caso de la regresión lineal simple, dos valores de DfBetas: uno para el cambio en el coeficiente β_0 y otro para el cambio en β_1. La expresión para el cálculo de estas medidas quedará:

$$\mathrm{DfBeta}_i = \frac{\beta_k - \beta_{k(i)}}{S_{(i)}\sqrt{a_{kk}}}$$

donde a_{kk} es el elemento correspondiente de la diagonal de la matriz $(X'X)^{-1}$. Se considerará que una observación es influyente sobre el coeficiente del modelo correspondiente si su DfBeta es superior o igual a $2/\sqrt{n}$, donde n es el número total de datos.

DFAJUSTE

Los DfAjuste proporcionan una medida del cambio en el valor pronosticado como consecuencia de la eliminación de una observación. La expresión para el cálculo de estas medidas quedará:

$$\text{DfAjuste}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{S_{(i)}\sqrt{h_{ii}}}$$

donde h_{ii} es el elemento correspondiente de la diagonal de la matriz $X(X'X)^{-1}X$. Se considera que una observación es influyente sobre las predicciones del modelo si su DfAjuste es superior a $2\sqrt{p/n}$, donde n es el número total de datos y p el número de coeficientes del modelo incluyendo el término de interceptación.

RAZÓN DE COVARIANZAS

Es una medida del cambio en la matriz de varianzas-covarianzas de los coeficientes del modelo al eliminar una observación. Su valor vendrá determinado por la expresión:

$$\text{CovRatio} = \frac{\left|S^2_{(i)}\left(X'_{(i)}X_{(i)}\right)^{-1}\right|}{\left|S^2\left(X'X\right)^{-1}\right|}$$

Un valor próximo a 1 asociado a la observación *(i)* indicaría que apenas se producen cambios en la matriz de varianzas-covarianzas y, en consecuencia, no sería influyente. Valores de esta razón alejados de 1 implicarían cambios importantes en la matriz de varianzas-covarianzas y, en consecuencia, se decidiría que tiene gran influencia sobre el modelo.

Si se trabaja con los datos del ejemplo 5-2, resulta de utilidad la construcción de los gráficos *box-plot* para cada una de estas medidas de influencia.

En la figura 5-29 puede observarse que la distancia de Mahalanobis y el valor del apalancamiento centrado tienen un comportamiento muy similar. Esto es debido a que solo hay una variable explicativa en el modelo y, por tanto, están captando el mismo efecto.

Los valores de la distancia de Cook, DfBetas, DfAjuste y CovRatio parecen identificar algunos valores como influyentes. Obsérvese, por ejemplo, que al calcular estas medidas las observaciones 171, 184 y 95 se muestran como valores atípicos en prácticamente todas ellas. Sería útil al investigador construir el modelo de regresión con y sin estos datos al fin de observar la magnitud de los cambios producidos en los coeficientes, en las estimaciones y en la significación estadística alcanzada.

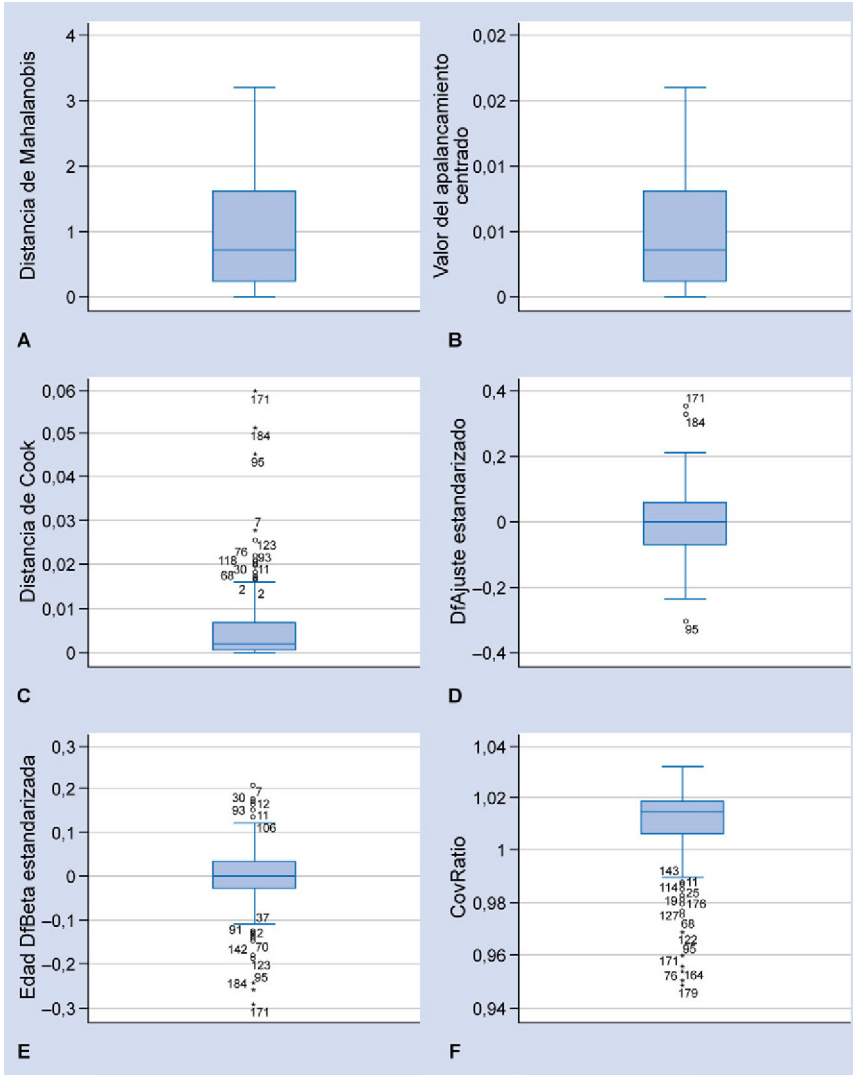

FIGURA 5-29 Estudio de los valores de influencia en el modelo de regresión.

PREDICCIONES

Una de las utilidades del modelo de regresión lineal simple, al margen de la valoración del efecto de la variable independiente o explicativa sobre la variable dependiente o explicada, es la posibilidad de realizar predicciones. Es obvio pensar que cuanto mejor ajuste el modelo, mejores serán las predicciones obtenidas, puesto que se reducirá notablemente la varianza

asociada a la estimación de la predicción. Las predicciones pueden ser de dos tipos:

1. Sobre una nueva observación.
2. Sobre la media de un conjunto de observaciones con el mismo valor de la variable independiente.

PREDICCIÓN DE UNA NUEVA OBSERVACIÓN

$$\hat{y}_i = \beta_i + \beta_i x_i$$

$$I_{1-\alpha}(y) = \left[\hat{y}_i - t_{n-2} S_r \sqrt{1 + \frac{1}{n}\left(1 + \frac{(x-\overline{x})^2}{S_x^2}\right)}, \hat{y}_i + t_{n-2} S_r \sqrt{1 + \frac{1}{n}\left(1 + \frac{(x-\overline{x})^2}{S_x^2}\right)} \right]$$

PREDICCIÓN DE UNA MEDIA DE OBSERVACIONES

$$\hat{y}_i = \beta_i + \beta_i x_i$$

$$I_{1-\alpha}(\overline{y}) = \left[\hat{y}_i - t_{n-2} S_r \sqrt{\frac{1}{n}\left(1 + \frac{(x-\overline{x})^2}{S_x^2}\right)}, \hat{y}_i + t_{n-2} S_r \sqrt{\frac{1}{n}\left(1 + \frac{(x-\overline{x})^2}{S_x^2}\right)} \right]$$

De esta forma podría obtenerse una estimación para una nueva observación o para la media para cada uno de los valores de la variable independiente.

Como puede observarse en la expresión analítica, la amplitud de los intervalos de confianza es mayor al estimar una nueva observación que al estimar una media. (Obsérvese que el contenido de la raíz cuadrada es una cantidad superior en el caso de una nueva observación que en el caso de la media, por lo que la amplitud del intervalo aumentará.) Para comprender este efecto debe tenerse en cuenta que, en caso de predicción de una nueva observación, al error asociado a la estimación de la media que proporciona la recta ajustada debe añadirse el error o distancia de cada observación a la media.

Por otra parte, se observa que la amplitud de los intervalos de confianza aumenta a medida que la variable independiente x se aleja de la media $\overline{x}$. [Obsérvese que la cantidad $(x-\overline{x})^2$ se encuentra en el numerador y aumentará la varianza a medida que aumenta la distancia entre x y $\overline{x}$.]

Si se unen los extremos de los intervalos de confianza construidos para cada valor de x para el mismo nivel de confianza se obtendrán las *bandas de confianza* para la predicción representadas en la figura 5-30.

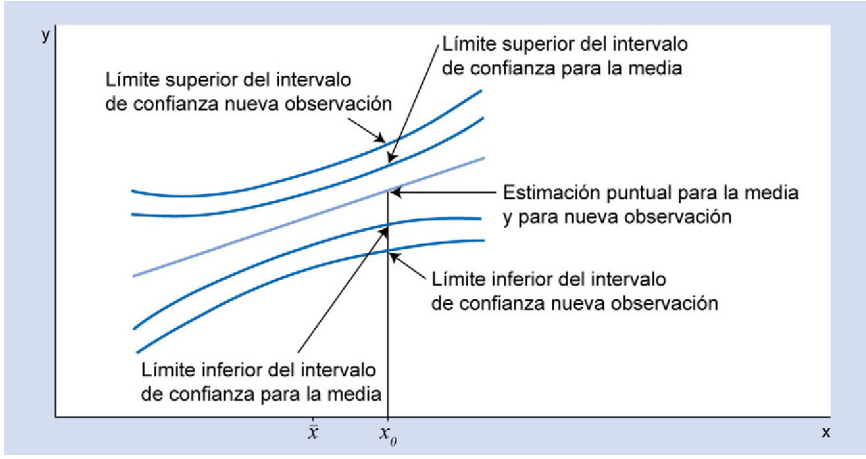

FIGURA 5-30 Bandas de confianza para la predicción.

Debe tenerse especial cuidado en la realización de predicciones más allá del rango de valores a partir del cual se ha ajustado el modelo, ya que se desconoce el comportamiento de los datos y la pertinencia del modelo lineal. Evidentemente, el riesgo aumenta cuanto más se aleja el valor de la variable x utilizado para la predicción del rango de valores utilizado en el ajuste.

MODELO DE REGRESIÓN LINEAL MÚLTIPLE

En la primera parte del presente capítulo se abordó con relativa profundidad el estudio de la relación entre dos variables cuantitativas. El modelo de regresión lineal simple permitía expresar la posible relación entre las dos variables en términos de «cambio en la variable dependiente por unidad de cambio en la variable independiente», mediante el ajuste de una recta. La variable independiente jugaba el papel de variable explicativa y la variable dependiente el de explicada. La pendiente de la recta, en este caso, recogía el efecto de la variable explicativa sobre la explicada. Sin embargo, en multitud de ocasiones, el comportamiento de una variable suele relacionarse, no con una única variable, sino con un conjunto de ellas.

- ¿Qué parte de la variabilidad del nivel de colesterol puede explicarse mediante la edad y el índice de masa corporal de un individuo? ¿Tanto la edad como el índice de masa corporal tienen un efecto significativo sobre el colesterol?
- ¿Qué ocurre si en el caso anterior se incluye además información sobre la existencia o no de antecedentes de cálculos biliares? ¿Se modifican los efectos de las otras dos variables explicativas?

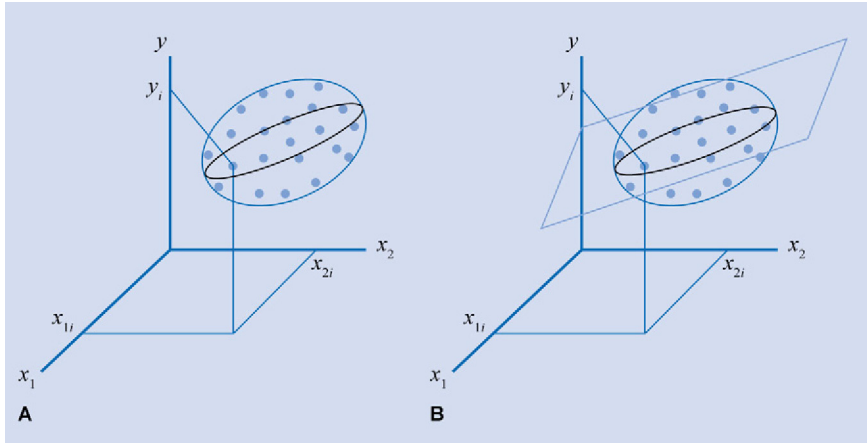

FIGURA 5-31 Diagrama de dispersión y modelo de regresión lineal múltiple ajustado.

Para responder a estas preguntas será necesario expresar la variable nivel de colesterol en función de las variables *edad* e *índice de masa corporal* en el primero de los casos y en función de la *edad*, *índice de masa corporal* y *antecedentes de cálculos biliares* en el segundo.

En este apartado se estudiará el modelo de regresión lineal múltiple como una extensión natural del modelo de regresión lineal simple que permitirá expresar la relación entre una variable dependiente y un conjunto de variables explicativas, profundizando en los aspectos fundamentales del análisis e incorporando a la discusión nuevos conceptos y situaciones derivadas de la inclusión de más de una variable explicativa en el modelo.

ESTRUCTURA

Para ilustrar la construcción e interpretación del modelo de regresión lineal múltiple se propone iniciar el estudio del caso en el que se incluyen únicamente dos variables explicativas en el modelo, situación que permite la representación gráfica en tres dimensiones mediante un diagrama de dispersión.

En la figura 5-31A se observa el diagrama de dispersión en tres dimensiones necesario para representar las observaciones correspondientes a cada uno de los individuos en el estudio de la relación lineal entre el *nivel de colesterol* como variable dependiente y la *edad* y el *índice de masa corporal* (IMC) como variables explicativas o independientes. En concreto, el punto en el espacio (x_{1i}, x_{2i}, y_i) representa edad, IMC y colesterol del individuo *i*.

El modelo de regresión lineal múltiple sería, en este caso, el plano que mejor resumiera el conjunto de la nube de puntos representados (v. fig. 5-31B). La expresión funcional del modelo quedaría:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Como puede observarse, esta expresión funcional establece que el valor de y depende de los valores que tomen las variables x_1 y x_2. Esto es, una vez que se dispone del modelo, si se hace variar el valor de x_1 y x_2 variará el valor de y.

En general, el modelo de regresión lineal múltiple para un conjunto de k variables explicativas podrá expresarse de la siguiente forma, aunque no pueda ser representado gráficamente:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$$

OBTENCIÓN DE LA RECTA DE REGRESIÓN LINEAL MÚLTIPLE

El modelo de regresión lineal múltiple estará perfectamente definido cuando se obtengan los valores de β_0, β_1, $\beta_2 \ldots$, β_k. Al igual que en el caso de la regresión lineal simple, uno de los criterios más comunes para la obtención del modelo final es el método de mínimos cuadrados, que se basa en la minimización del cuadrado del error cometido al proporcionar el valor estimado por el modelo $\hat{y}_i$ en lugar del verdadero valor observado y_i. Gráficamente, para el caso de dos variables explicativas, podría expresarse como se describe en la figura 5-32.

El mejor plano sería el que hiciera mínimos todos los cuadrados de las distancias entre el verdadero valor observado y_i y el valor estimado por el plano $\hat{y}_i$, llamado *error* o *residuo* e_i. Así, para cada individuo se tendrá que:

$$e_i = y_i - \hat{y}_i$$

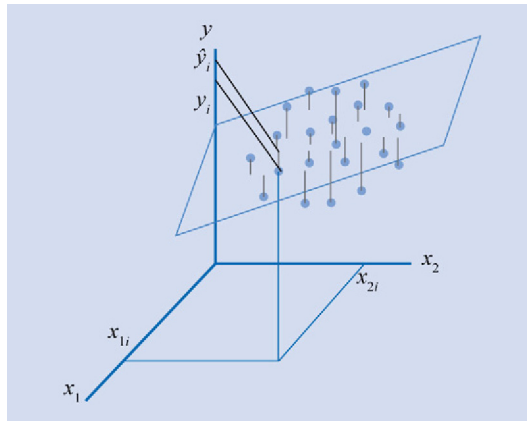

FIGURA 5-32 Método de mínimos cuadrados. Minimización de los errores.

Para obtener el mejor plano habría que minimizar la siguiente cantidad:

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum \left(y_i - \left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \right) \right)^2$$

Que, en general, quedará:

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum \left(y_i - \left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} \right) \right)^2$$

Obsérvese que en esta expresión todo son números conocidos excepto β_0, β_1 y β_2 o β_0, β_1, $\beta_2 \ldots$, β_k, respectivamente. Para obtener la solución final debe derivarse parcialmente $\Sigma[y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki})]^2$ con respecto a cada uno de los coeficientes del modelo e igualar a 0 para determinar la existencia de un mínimo. De estas k + 1 expresiones se obtendrá un sistema de k + 1 ecuaciones con k + 1 incógnitas. Los valores finales para β_0, β_1, $\beta_2 \ldots$, β_k vendrán determinados, en notación matricial, por la siguiente expresión:

$$\hat{\beta} = (X'X)^{-1} X'Y$$

Debe tenerse en cuenta que serán necesarias hipótesis adicionales sobre las variables independientes incluidas en el modelo de regresión para que el sistema de ecuaciones tenga solución única:

- El número de variables explicativas debe ser inferior al número de observaciones.
- Las variables explicativas son distintas entre sí y no existen entre ellas relaciones lineales exactas. (De no ser así, la matriz X'X no sería invertible.)

En el caso de que se incluyan únicamente dos variables explicativas en el modelo de regresión lineal múltiple, las expresiones para los coeficientes del modelo, obviando el término de interceptación β_0, quedarán:

$$\beta_1 = \frac{\mathrm{Cov}(x_1, y) S_{x_2}^2 - \mathrm{Cov}(x_2, y) \mathrm{Cov}(x_1, x_2)}{S_{x_1}^2 S_{x_2}^2 - \mathrm{Cov}(x_1, x_2)^2}$$

$$\beta_2 = \frac{\mathrm{Cov}(x_2, y) S_{x_1}^2 - \mathrm{Cov}(x_1, y) \mathrm{Cov}(x_1, x_2)}{S_{x_1}^2 S_{x_2}^2 - \mathrm{Cov}(x_1, x_2)^2}$$

Como puede observarse, en las expresiones de ambos coeficientes interviene de forma determinante la covarianza entre las variables explicativas x_1 y x_2. ¿Qué ocurriría si entre las dos variables explicativas existiera una relación lineal perfecta? En este caso el coeficiente de correlación lineal de Pearson sería 1 o –1. Supóngase una relación lineal perfecta y directa entre

las dos variables explicativas. Entonces el coeficiente de correlación de Pearson sería 1 y se tendría que:

$$r_{x_1x_2} = \frac{Cov(x_1, x_2)}{S_{x_1} S_{x_2}} = 1$$

Despejando el valor de la covarianza se obtendría la expresión:

$$Cov(x_1, x_2) = S_{x_1} S_{x_2}$$

Sustituyendo en las expresiones de los coeficientes del modelo anteriormente obtenidas se tendría que:

$$\beta_1 = \frac{Cov(x_1, y) S^2_{x_2} - Cov(x_2, y) Cov(x_1, x_2)}{S^2_{x_1} S^2_{x_2} - Cov(x_1, x_2)^2} = \frac{Cov(x_1, y) S^2_{x_2} - Cov(x_2, y) S_{x_1} S_{x_2}}{S^2_{x_1} S^2_{x_2} - S^2_{x_1} S^2_{x_2}}$$

$$\beta_2 = \frac{Cov(x_2, y) S^2_{x_1} - Cov(x_1, y) Cov(x_1, x_2)}{S^2_{x_1} S^2_{x_2} - Cov(x_1, x_2)^2} = \frac{Cov(x_2, y) S^2_{x_1} - Cov(x_1, y) S_{x_1} S_{x_2}}{S^2_{x_1} S^2_{x_2} - S^2_{x_1} S^2_{x_2}}$$

En ambos casos el denominador sería 0 y no podría calcularse un valor único para los coeficientes del modelo. En consecuencia, no pueden existir relaciones lineales exactas entre las variables explicativas del modelo.

Por otra parte, ¿qué ocurriría si la relación lineal entre las dos variables explicativas fuera nula? En este caso el valor del coeficiente de correlación lineal de Pearson sería 0 y, por tanto, la covarianza entre las dos variables explicativas sería 0. Sustituyendo el valor 0 de la covarianza en las expresiones para el cálculo de los coeficientes del modelo se obtendría:

$$\beta_1 = \frac{Cov(x_1, y) S^2_{x_2}}{S^2_{x_1} S^2_{x_2}} = \frac{Cov(x_1, y)}{S^2_{x_1}}$$

$$\beta_2 = \frac{Cov(x_2, y) S^2_{x_1}}{S^2_{x_1} S^2_{x_2}} = \frac{Cov(x_2, y)}{S^2_{x_2}}$$

Puede observarse que, en este caso, los coeficientes del modelo de regresión lineal múltiple que recogen los efectos de cada una de las variables explicativas consideradas, coinciden con los coeficientes de los correspondientes modelos de regresión lineal simple entre la variable dependiente y cada una de las variables explicativas por separado. Además, puede concluirse de la misma forma, y a partir de las expresiones de cálculo de los coeficientes y de la intervención de la covarianza entre las variables explicativas en el cálculo de los mismos, que cuanta mayor relación

exista entre las variables explicativas mayormente se verán afectados los coeficientes del modelo en comparación con los que se obtendrían en sus respectivas regresiones lineales simples.

INTERPRETACIÓN DE LOS COEFICIENTES DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

De los resultados obtenidos en el apartado anterior puede deducirse que la interpretación de los coeficientes del modelo de regresión lineal múltiple requiere realizar alguna reflexión adicional, ya que el valor del coeficiente correspondiente a una variable explicativa x_1 puede verse afectado por la presencia o no de otra variable explicativa x_2, dependiendo de si esta última guarda algún tipo de relación lineal con la primera. Para ilustrar la interpretación de los coeficientes del modelo de regresión lineal múltiple se propone trabajar con dos variables explicativas que han permitido ajustar el siguiente modelo en el que $\beta_0 = 6$, $\beta_1 = 2$ y $\beta_2 = 4$:

$$y = 6 + 2x_1 + 4x_2$$

Si se fija el valor de la segunda variable explicativa en $x_2 = 2$, ¿qué ocurrirá si se aumenta progresivamente en una unidad la variable x_1? ¿Y si se fija el valor de la segunda variable explicativa en $x_2 = 6$?

En la tabla 5-7 puede observarse que al aumentar en una unidad la variable x_1, manteniendo constante la variable x_2, el valor de la variable dependiente y aumenta en 2 unidades, que coincide con el valor del coeficiente que acompaña a la variable explicativa x_1.

Si se fija el valor de la primera variable explicativa en $x_1 = 2$, ¿qué ocurrirá si se aumenta progresivamente en una unidad la variable x_2? ¿Y si se fija el valor de la primera variable explicativa en $x_1 = 6$?

En la tabla 5-8 se observa que al aumentar en una unidad la variable x_2 manteniendo constante la variable x_1, el valor de la variable dependiente

TABLA 5-7 Interpretación del coeficiente β_1

Si $x_2 = 2$	**Si $x_2 = 6$**
$y = 6 + 2 \cdot 0 + 4 \cdot 2 = 14$ si $x_1 = 0$	$y = 6 + 2 \cdot 0 + 4 \cdot 6 = 30$ si $x_1 = 0$
$y = 6 + 2 \cdot 1 + 4 \cdot 2 = 16$ si $x_1 = 1$	$y = 6 + 2 \cdot 1 + 4 \cdot 6 = 32$ si $x_1 = 1$
$y = 6 + 2 \cdot 2 + 4 \cdot 2 = 18$ si $x_1 = 2$	$y = 6 + 2 \cdot 2 + 4 \cdot 6 = 34$ si $x_1 = 2$

TABLA 5-8 Interpretación del coeficiente β_2

Si $x_1 = 2$	**Si $x_1 = 6$**
$y = 6 + 2 \cdot 2 + 4 \cdot 0 = 10$ si $x_2 = 0$	$y = 6 + 2 \cdot 6 + 4 \cdot 0 = 18$ si $x_2 = 0$
$y = 6 + 2 \cdot 2 + 4 \cdot 1 = 14$ si $x_2 = 1$	$y = 6 + 2 \cdot 6 + 4 \cdot 1 = 22$ si $x_2 = 1$
$y = 6 + 2 \cdot 2 + 4 \cdot 2 = 18$ si $x_2 = 2$	$y = 6 + 2 \cdot 6 + 4 \cdot 2 = 26$ si $x_2 = 2$

y aumenta en 4 unidades, que coincide con el valor del coeficiente que acompaña a la variable explicativa x_2.

En consecuencia, cada uno de los coeficientes del modelo de regresión lineal múltiple que acompañan a las correspondientes variables explicativas incluidas en el mismo, por ejemplo x_k, puede interpretarse como el *cambio en la variable dependiente por unidad de cambio en la variable independiente* x_k *manteniendo constantes el resto de las variables explicativas*.

Si se tiene en cuenta este resultado y el discutido anteriormente sobre la influencia de la covarianza y, en consecuencia, de la correlación lineal entre las variables explicativas sobre los valores de los coeficientes del modelo de regresión lineal múltiple, puede afirmarse que, para interpretar cada uno de los coeficientes del modelo de regresión lineal múltiple, será necesario mencionar el resto de variables explicativas incluidas. Cada uno de los coeficientes estará «ajustado» por el resto de variables explicativas del modelo.

BONDAD DEL AJUSTE DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

Al igual que en el caso de la regresión lineal simple, este procedimiento de construcción del modelo siempre proporcionará un resultado (a excepción del caso de relaciones lineales exactas entre las variables explicativas) con independencia de que el modelo propuesto ajuste suficientemente a los datos observados.

DESCOMPOSICIÓN DE LA VARIABILIDAD

Como punto de partida se tendrá en cuenta que la distancia de cada observación de la variable dependiente *y* a la media de las observaciones de la variable puede expresarse como:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

donde:

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki}$$

Repitiendo el mismo procedimiento que en el caso de la regresión lineal simple la descomposición de la variabilidad quedará de la siguiente forma:

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

donde:

$\sum (y_i - \hat{y}_i)^2$ = Variabilidad no explicada por el modelo de regresión. Adviértase que resume las distancias al cuadrado entre el verdadero valor observado de la variable dependiente y el valor que pronostica el modelo (varianza no explicada por el modelo o varianza residual). Se denota VNE.

$\sum(\hat{y}_i - \overline{y})^2$ = Variabilidad explicada por el modelo de regresión. Obsérvese que resume las distancias entre el valor promedio de la variable dependiente (valor que se proporcionaría para estimar el valor de *y* si no se tuviera en cuenta el modelo de regresión) y el valor que pronostica el modelo (varianza explicada por el modelo o varianza de la regresión). Se denota VE.

$\sum(y_i - \overline{y})^2$ = Variabilidad total observada en la variable dependiente. Apréciese que coincide con el numerador de la varianza de la variable *y*. Se denota VT.

En consecuencia, se tiene que:

$$VT = VNE + VE$$

COEFICIENTE DE DETERMINACIÓN Y CORRELACIÓN MÚLTIPLE

Dado que la variabilidad total observada en la variable dependiente ha podido ser expresada (descompuesta) como la suma de dos variabilidades (la variabilidad no explicada por el modelo de regresión y la variabilidad explicada por el modelo de regresión), puede obtenerse una medida de la bondad del ajuste del modelo calculando la proporción de variabilidad explicada por el modelo. El resultado es una extensión natural del coeficiente de determinación de la recta de regresión lineal simple al caso de más de una variable explicativa. Se tendrá entonces que:

$$R^2 = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = \frac{VE}{VT}$$

Este coeficiente se interpreta como la proporción de variabilidad explicada por el modelo o, si se multiplica el resultado por 100, como porcentaje de variabilidad explicada por el modelo.

Dado que la inclusión de nuevas variables explicativas en el modelo conducirá a una mejora del valor de R^2, aunque no aporten explicación suficiente a la variabilidad observada en la variable dependiente, puede calcularse un coeficiente de determinación múltiple corregido por el número de variables explicativas incluidas en el modelo de la siguiente forma:

$$R^2_{\text{correg}} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Por otra parte, el coeficiente de correlación múltiple representará la magnitud de la relación lineal entre la variable dependiente y un conjunto de variables explicativas. Este coeficiente podrá obtenerse de la forma:

$$r_{\text{múltiple}} = \sqrt{R^2}$$

Ejemplo 5-3

En un estudio se obtuvo información sobre el nivel de colesterol, la edad y el índice de masa corporal de un grupo de 200 pacientes. Se pretende estudiar el posible efecto de la edad y el índice de masa corporal sobre el nivel de colesterol. El modelo de regresión lineal múltiple para el estudio propuesto quedará:

$$\text{Colesterol} = \beta_0 + \beta_1 \text{edad} + \beta_2 \text{IMC}$$

En las tablas 5-9 y 5-10 se muestran los resultados del ajuste del modelo obtenidos utilizando el programa SPSS. Como puede observarse, la magnitud de la asociación lineal entre el colesterol y las variables edad e IMC viene dada por su coeficiente de correlación múltiple que, en este caso, es 0,422. Además el coeficiente de determinación múltiple del modelo es $R^2 = 0{,}178$, con lo que se estaría explicando el 17,8% de la variabilidad observada en el nivel de colesterol de los individuos estudiados. El coeficiente de determinación múltiple corregido por el número de variables incluidas en el modelo se sitúa en 0,17 y el porcentaje de variabilidad explicada en el 17%.

Según los resultados obtenidos, el modelo ajustado vendrá dado por la siguiente expresión:

$$\text{Colesterol} = 108{,}041 + 0{,}989 \cdot \text{edad} + 2{,}704 \cdot \text{IMC}$$

Esto significaría que, para individuos de la misma edad, por aumentar en una unidad el IMC se produciría un aumento del nivel de colesterol de 2,704 unidades (en mg/100 ml). Por otra parte, para individuos del mismo IMC, por cada año más de edad se produciría un aumento de 0,989 unidades en el nivel de colesterol (en mg/100 ml).

A partir de este resultado podría concluirse erróneamente que la variable IMC tiene un mayor efecto sobre el nivel de colesterol que la variable edad al provocar un mayor cambio en la variable dependiente (2,704 unidades para el

TABLA 5-9 Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,422	,178	,17	46,19

TABLA 5-10 Coeficientes

Modelo	Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
	B	Error típ.	Beta		
(Constante)	108,041	23,926		4,516	,000
Edad	,989	,233	,301	4,247	,000
IMC	2,704	,971	,197	2,784	,006

IMC frente a las 0,989 unidades en el caso de la *edad*). Sin embargo debe tenerse en cuenta que cada uno de los coeficientes β_i depende no solo de las unidades de medida de la variable dependiente, sino también de las unidades de medida de la variable x_i a la que acompañan. En este sentido, es esperable que variables explicativas que se mueven en un rango de valores superior al resto, presenten un coeficiente ajustado inferior como consecuencia de las unidades de medida y no porque tengan un menor efecto que las demás.

Para intentar valorar el «*peso*» de cada una de las variable explicativas en el modelo de regresión lineal múltiple, entendido como la variable que produce un mayor cambio en la variable dependiente, eliminado el posible efecto de las unidades de medida, suele ser habitual trabajar con las variables (dependiente e independientes) estandarizadas. De esta forma, todas las variables incluidas en el ajuste serán adimensionales y tendrán media 0 y desviación típica 1.

En la columna de coeficientes tipificados en la tabla 5-10 pueden observarse los valores de los coeficientes del modelo que se habrían obtenido en el caso de trabajar con las variables estandarizadas. Una de las consecuencias es que el plano pasará por el origen de coordenadas y no habrá término de interceptación β_0. Además, el coeficiente que acompaña a la edad es, en este caso, de 0,301 frente al 0,197 de la variable IMC, indicando que es la variable *edad* la que parece mostrar un mayor efecto sobre el nivel de colesterol.

INFERENCIA SOBRE EL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

Cuando se construye un modelo de regresión lineal múltiple a partir de los datos contenidos en una muestra aleatoria de la población debe tenerse en cuenta que el modelo ajustado no es más que uno de todos los posibles modelos que podrían ajustarse a partir de cada una de las muestras posibles de la población que hubieran podido ser seleccionadas en el proceso de muestreo. Esto es, para cada muestra se ajustaría un modelo que podría ser similar, pero que mostrará diferencias en el valor de sus coeficientes β_0, β_1, β_2…, β_k, ajustados al variar los datos de partida.

En consecuencia, los valores de β_0, β_1, β_2…, β_k, así como el del coeficiente de determinación múltiple R^2, serán variables aleatorias que varían de muestra a muestra de la población. Un contraste de interés trataría de establecer si el modelo de regresión lineal múltiple explica de forma significativa parte de la variabilidad observada en la variable dependiente. Se plantea, por tanto, un contraste sobre el coeficiente de determinación múltiple del modelo como el siguiente:

$$H_0 : \rho^2 = 0$$
$$H_1 : \rho^2 \neq 0$$

Este contraste es idéntico al construido para la regresión lineal simple, salvo por el hecho de que, en este caso, en el modelo se cuenta con más de una variable explicativa. En caso de aceptación de la hipótesis nula no habría evidencia de que el coeficiente de determinación múltiple poblacional fuera significativamente distinto de 0. Por tanto, la variabilidad explicada podría ser 0 y el modelo no explicaría nada (se espera, por tanto, que ninguna de las variables explicativas tenga un efecto significativo sobre la variable dependiente). En caso de rechazar la hipótesis nula, el coeficiente de determinación múltiple poblacional sería significativamente distinto de 0 y el modelo explicaría, de forma significativa, parte de la variabilidad observada en la variable dependiente. En este caso, sería necesario establecer si todas y cada una de las variables explicativas tienen un efecto significativo sobre la variable dependiente o solo algunas de ellas. El estadístico de contraste sobre el coeficiente de determinación múltiple poblacional es:

$$EC = \frac{VE/k}{VNE/(n-k-1)} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \hat{y}_i)^2/(n-k-1)}$$

La distribución muestral asociada sería un F de Snedecor con k y n-k-1 grados de libertad en el caso de verificarse las hipótesis necesarias que se abordarán más adelante. Como puede observarse, si la variabilidad explicada por el modelo es 0, el valor del estadístico de contraste será 0 (situación compatible con la hipótesis nula de no explicación significativa). Por otro lado, si la variabilidad explicada comienza a aumentar (crece el numerador), la variabilidad no explicada tendrá que disminuir (disminuye el denominador) con lo que el valor del estadístico de contraste crecerá cada vez más hasta el punto de poder rechazar, en su caso, la hipótesis nula (el modelo explicaría una parte significativa de la variabilidad de la variable dependiente).

TABLA DE ANOVA DE LA REGRESIÓN

Al igual que en la regresión lineal simple, suele ser habitual, presentar los resultados anteriores en forma de tabla conocida como la tabla de ANOVA de la regresión (tabla 5-11).

En la tabla 5-12 se muestran los resultados obtenidos a partir de los datos del ejemplo 5-3.

El estadístico de contraste sería 21,27 que, comprobado en las tablas de la F de Snedecor con 2 y n – 2 − 1 = 196 grados de libertad, proporcionaría un valor de $p < 0{,}001$.

$$EC = \frac{VE/k}{VNE/(n-k-1)} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \hat{y}_i)^2/(n-k-1)} = \frac{\frac{90.761,468}{2}}{\frac{418.174,24}{196}}$$

$$= \frac{45.380,734}{2.133,542} = 21,27$$

TABLA 5-11 Tabla de ANOVA de la regresión lineal múltiple

Fuente	Suma de cuadrados	Media de cuadrados	Cociente de varianzas
Regresión (VE)	$\sum(\hat{y}_i - \bar{y})^2$	$\frac{\sum(\hat{y}_i - \bar{y})^2}{1}$	$\frac{\sum(\hat{y}_i - \bar{y})^2 / 1}{\sum(y_i - \hat{y}_i)^2 / (n-2)}$
Residual (VNE)	$\sum(y_i - \hat{y}_i)^2$	$\frac{\sum(y_i - \hat{y}_i)^2}{n-2}$	
Total (VT)	$\sum(y_i - \bar{y})^2$	$\frac{\sum(y_i - \bar{y})^2}{n-1}$	

TABLA 5-12 Tabla de ANOVA de la regresión lineal múltiple para los datos del ejemplo 5-3

Modelo	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Regresión (VE)	90.761,468	2	45.380,734	21,270	,000
Residual (VNE)	418.174,24	196	2.133,542		
Total (VT)	508.935,709	198			

Dado que el valor de la *p* del contraste es inferior al nivel habitual 0,05, se rechazaría la hipótesis nula y se concluiría que el modelo explica significativamente una parte de la variabilidad observada en la variable *nivel de colesterol*. Sin embargo, sería necesario profundizar en el efecto de cada una de las dos variables incluidas en el modelo con objeto de averiguar si las dos tienen un efecto significativo sobre la variable dependiente o solo una de ellas. Se plantean por tanto, los contrastes individuales sobre los coeficientes:

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

Si se recupera la tabla 5-10 de coeficientes del modelo, puede observarse que los contrastes individuales sobre los coeficientes son significativos, tanto en el caso de la *edad* como del *IMC*, con unos valores de *p* inferiores a la milésima y de 0,006 respectivamente, e inferiores al nivel de significación habitual 0,05. Por tanto, las dos variables tienen un efecto significativo sobre el nivel colesterol.

REQUERIMIENTOS SOBRE EL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

Los requisitos necesarios para la realización de inferencias a partir del modelo de regresión lineal múltiple recogen hipótesis sobre la pertinencia de la relación lineal entre la variable dependiente y el resto de variables explicativas estudiadas e hipótesis sobre los datos y su distribución muestral

que permitan la realización de las inferencias deseadas y suponen una extensión natural de las hipótesis sobre el modelo de regresión lineal simple. Estas hipótesis podrían resumirse en las siguientes:

- Pertinencia de la linealidad: $E(y \mid x_1, x_2, x_3 \ldots, x_k) = 0$.
- Homocedasticidad: $Var(y \mid x_1, x_2, x_3 \ldots, x_k) = cte$.
- Distribución de probabilidad normal: $y \mid x_1, x_2, x_3 \ldots, x_k \sim$ normal.
- Independencia de las observaciones: y_i, y_j independientes para cualquier i, j.
- No existen relaciones lineales exactas entre las variables explicativas. Sería también problemático que, aunque no fueran exactas, presentaran buenas relaciones lineales entre ellas (multicolinealidad).

Tradicionalmente las hipótesis sobre el modelo de regresión lineal han sido comprobadas mediante el análisis de los residuos del modelo. Las hipótesis se reformularían de la siguiente forma:

- Linealidad: $E(e \mid x_1, x_2, x_3 \ldots, x_k) = 0$.
- Homocedasticidad: $Var(e \mid x_1, x_2, x_3 \ldots, x_k) = cte$.
- Normalidad: $e \mid x_1, x_2, x_3 \ldots, x_k \sim$ normal.
- Independencia: e_i, e_j independientes para cualquier i, j.
- Estudio de la posible colinealidad o multicolinealidad.

Para la comprobación de las hipótesis del modelo se utilizarían representaciones gráficas como las estudiadas en el modelo de regresión lineal simple (v. el apartado «Requerimientos sobre el modelo de regresión lineal simple»). La forma de condicionar la distribución de los residuos a los valores de las variables explicativas $e \mid x_1, x_2, x_3 \ldots, x_k$ consiste en utilizar el valor estimado por el modelo y trabajar con $e \mid \hat{y}$, donde se sabe que: $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ \ldots \ + \beta_k x_k$. Si se trabaja con los datos del ejemplo 5-3 se obtendrían los siguientes resultados para la comprobación de las hipótesis sobre el modelo de regresión lineal múltiple:

$$\sum e_i = 3{,}18 \cdot 10^{-15} \simeq 0$$

En la figura 5-33 se observa que la horizontal en 0 actúa como un eje de simetría con respecto a los residuos del modelo, reforzando la hipótesis de linealidad. Aunque no se observan cambios relevantes de la dispersión de los residuos a lo largo del valor pronosticado por el modelo, sí parece observarse un pequeño aumento de la dispersión con el valor pronosticado y requerirá una especial atención para la valoración de la hipótesis de homocedasticidad.

Para identificar qué variable o variables explicativas podrían ser las causantes de una posible homocedasticidad resulta de utilidad la construcción de gráficos parciales de residuos, es decir, diagramas de dispersión de los residuos en función de cada una de las variables explicativas.

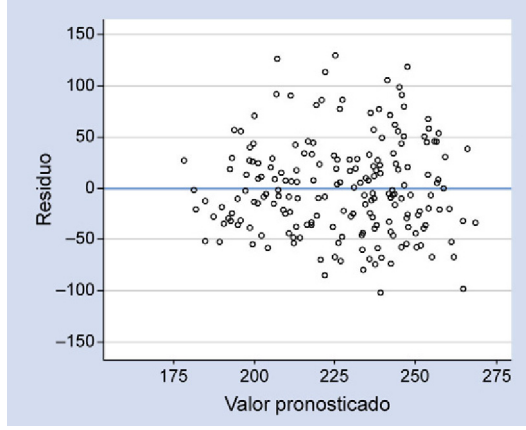

FIGURA 5-33 Diagrama de dispersión de los residuos frente al valor pronosticado.

La figura 5-34 muestra los gráficos parciales de los residuos con respecto a cada una de las variables explicativas.

En ambos gráficos la horizontal en 0 actúa como eje de simetría con respecto a los residuos. En el gráfico parcial con respecto a la *edad* no se observan cambios en la dispersión de los mismos a lo largo del valor pronosticado por el modelo, resultado coherente con el representado en la figura 5-32 correspondiente al modelo de regresión múltiple. En el gráfico parcial con respecto al IMC parece observarse un ligero aumento de la dispersión de los datos con el IMC aunque parece no significativo a simple vista. Se completa el estudio de la homocedasticidad construyendo los modelos de regresión siguientes utilizando el valor absoluto de los residuos como variable dependiente.

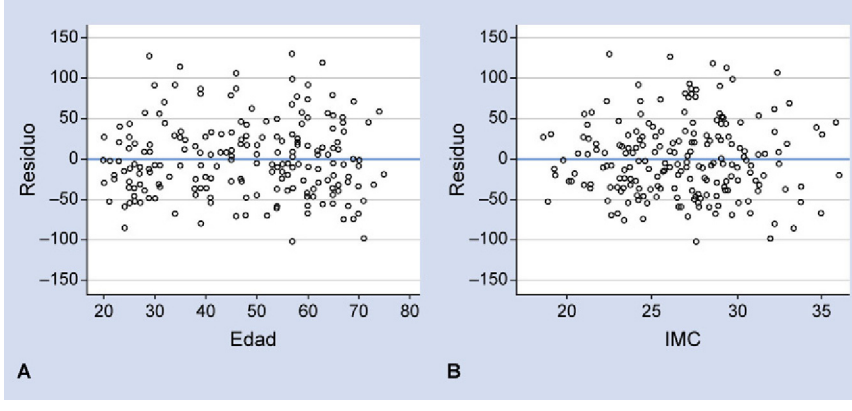

FIGURA 5-34 Gráficos parciales de residuos.

TABLA 5-13 Estudio de la homocedasticidad. Modelos de regresión para los datos del ejemplo 5-3

Modelo	R^2	p
$\|e\| = \beta_0 + \beta_1 \hat{y}$	0,016	0,076
$\|e\| = \beta_0 + \beta_1$edad	0,008	0,21
$\|e\| = \beta_0 + \beta_1$IMC	0,018	0,061

En la tabla 5-13 puede observarse que en todos los casos el contraste sobre el coeficiente de determinación del modelo no es significativo al nivel 0,05, aunque en el caso del IMC se obtiene un valor muy próximo de 0,061 y que refleja el efecto visual observado en los gráficos anteriores.

A continuación se proporcionan los gráficos correspondientes al histograma (fig. 5-35) y gráfico de probabilidad normal (fig. 5-36) de los residuos del modelo que muestran un comportamiento razonablemente similar a la distribución normal.

El contraste de Kolmogorov-Smirnov en este caso proporciona un estadístico de contraste de 0,849 con un valor de $p = 0{,}467$ no significativo al nivel $\alpha = 0{,}05$.

Por último, y a falta del estudio de la posible multicolinealidad, se aborda la comprobación de la hipótesis de independencia. El estadístico de contraste de Durbin-Watson en este caso es:

$$\text{Durbin} - \text{Watson} = 2{,}007$$

Que se encuentra claramente situado en la región de aceptación de la hipótesis de no autocorrelación entre los residuos.

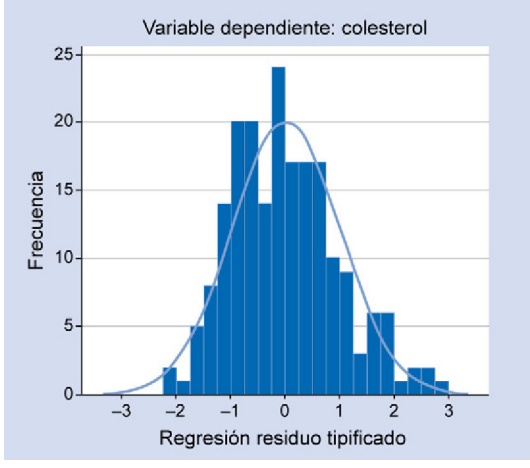

FIGURA 5-35 Histograma de los residuos.

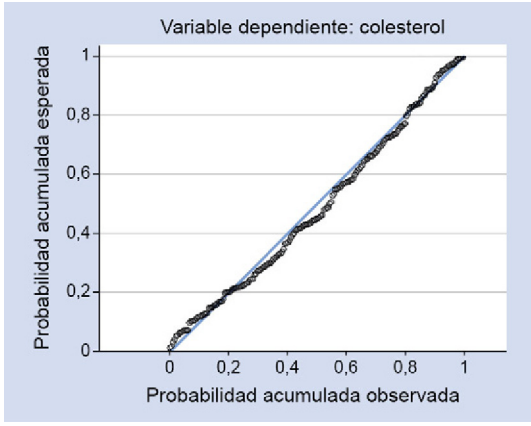

FIGURA 5-36 Gráfico P-P normal.

COLINEALIDAD O MULTICOLINEALIDAD

Anteriormente fue discutido que la inclusión de más de una variable explicativa en el modelo de regresión lineal implicaba la consideración de requisitos adicionales sobre el mismo, y, concretamente, la comprobación de la inexistencia de relaciones lineales exactas entre las variables explicativas incluidas que garantizaran que el sistema de ecuaciones para el cálculo de los coeficientes tuviera una solución única. Adicionalmente se estableció que cuanta mayor relación exista entre las variables explicativas, mayormente se verán afectados los coeficientes del modelo en comparación con los que se obtendrían en sus respectivas regresiones lineales simples.

De hecho, supóngase que se construye un modelo de regresión lineal múltiple con dos variables explicativas x_1 y x_2 de forma que se sabe que entre ellas existe una relación lineal exacta del tipo $x_2 = a + bx_1$. El modelo quedará:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

de donde:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = \beta_0 + \beta_1 x_1 + \beta_2 (a + bx_1) = \beta_0 + \beta_1 x_1 + \beta_2 a + \beta_2 b x_1 =$$
$$= (\beta_0 + \beta_2 a) + (\beta_1 + \beta_2 b) x_1$$
$$= \beta_0^* + \beta_1^* x_1$$

Como puede observarse, la relación entre las dos variables explicativas y la variable dependiente puede expresarse, finalmente, mediante un modelo de regresión que solo incluye a una de las variables. Por tanto, al proponer un modelo que incluya las dos variables explicativas se estaría repartiendo un único efecto β_1^* en dos efectos (β_1 y β_2) de inferior magnitud. Este hecho

será igualmente preocupante cuando existan buenas relaciones lineales, aunque no sean exactas, entre las variables explicativas.

Analíticamente, ello implica que el valor de los coeficientes del modelo correspondientes a un conjunto de variables explicativas disminuirá enormemente en presencia de variables que tengan una buena relación lineal con ellas y, por tanto, será más difícil detectar significación estadística en los respectivos contrastes individuales. Téngase en cuenta que los contrastes individuales de los coeficientes se construyen a partir del estadístico de contraste:

$$EC = \frac{\beta_i}{\sqrt{Var(\beta_i)}}$$

Cuanto menor sea el valor del coeficiente β_i menor será el valor del estadístico de contraste y será más difícil rechazar la hipótesis nula que presupone la inexistencia de un efecto significativo de la variable explicativa correspondiente.

Por otra parte, la varianza asociada a los coeficientes de un modelo de regresión lineal múltiple con dos variables explicativas cualesquiera puede expresarse de la siguiente forma:

$$Var(\beta_1) \simeq \frac{S_r^2}{S_1^2(1-r^2)n} = \frac{1}{1-r^2}\left(\frac{S_r^2}{S_1^2 n}\right)$$

$$Var(\beta_2) \simeq \frac{S_r^2}{S_2^2(1-r^2)n} = \frac{1}{1-r^2}\left(\frac{S_r^2}{S_2^2 n}\right)$$

donde r es el coeficiente de correlación lineal de Pearson entre las variables explicativas x_1 y x_2. En consecuencia, r^2 sería el coeficiente de determinación de la recta ajustada utilizando una de las variables explicativas como variable dependiente y la otra como independiente.

En general puede demostrarse que la varianza de cada uno de los coeficientes de un modelo de regresión lineal múltiple es proporcional a $1/(1-R^2)$, donde R^2 es el coeficiente de determinación múltiple del modelo construido con una variable explicativa en función de las restantes. Obsérvese que, cuanto mayor sea la relación lineal entre una variable explicativa y el resto, mayor será el valor del coeficiente de determinación múltiple correspondiente (tiende a 1), con lo que el denominador de $1/(1-R^2)$ tenderá a 0 y la varianza tenderá a infinito.

En consecuencia, cuanto mayor sea la relación lineal entre una variable explicativa y el resto de variables explicativas del modelo mayor será la varianza asociada a los coeficientes β_i y menor será el valor del estadístico de contraste sobre los coeficientes, dificultando la detección de significación estadística.

Una situación que se da con relativa frecuencia en un análisis de regresión lineal múltiple es obtener un valor del coeficiente de determinación múltiple significativo (el modelo explica) y, sin embargo, todos los contrastes individuales sobre los coeficientes resultan no significativos. La explicación podría encontrarse en la existencia de una fuerte multicolinealidad entre las variables explicativas que podría resolverse, por ejemplo, identificando la variable o variables responsables y excluirlas del análisis. Es necesario, por tanto, contar con estrategias de análisis para el estudio de la posible colinealidad o multicolinealidad entre las variables explicativas en un modelo de regresión lineal múltiple.

CORRELACIONES BIVARIADAS

Una primera aproximación al estudio de la multicolinealidad consiste en el cálculo de todas las correlaciones bivariadas (correlaciones lineales de Pearson) entre todas las variables explicativas incluidas en el modelo. Se obtendría una matriz de correlaciones expresada de la siguiente forma:

$$\begin{pmatrix} r_{x_1x_1} & r_{x_1x_2} \cdots & r_{x_1x_k} \\ \vdots & \ddots & \vdots \\ r_{x_kx_1} & r_{x_kx_2} \cdots & r_{x_kx_k} \end{pmatrix}$$

Este resultado permitiría identificar pares de variables que tienen buena relación lineal entre ellas. Obsérvese que el coeficiente de determinación de la recta de regresión lineal simple correspondiente podría obtenerse elevando al cuadrado la correlación.

TOLERANCIA Y FACTOR DE INCREMENTO DE LA VARIANZA

La inexistencia de buenas relaciones lineales por pares no implica que no exista un problema de multicolinealidad, ya que es igualmente problemático que una variable explicativa tenga una buena relación lineal con el resto de variables explicativas incluidas en el modelo. Sería de utilidad construir un modelo de regresión lineal múltiple para cada una de las variables explicativas que expresara cada una de ellas en función de las demás. Así, si se considera el modelo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$$

se construirían los siguientes modelos de regresión lineal múltiple y se calcularían sus respectivos coeficientes de determinación múltiples:

$$x_1 = \beta_0 + \beta_1 x_2 + \beta_2 x_3 + \beta_3 x_4 + \ldots + \beta_k x_k \quad R^2_{x_1 \cdot x_2 x_3 \ldots x_k}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad \ldots\ldots$$

$$x_k = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_4 + \ldots + \beta_k x_{k-1} \quad R^2_{x_k \cdot x_1 x_2 \ldots x_{k-1}}$$

Un coeficiente de determinación múltiple próximo a 1 para uno o varios de estos modelos indicaría que la variable o variables correspondientes serían una buena combinación lineal del resto y, por tanto, indicativo de multicolinealidad.

Una variable sería bien tolerada en el modelo de regresión lineal múltiple, desde el punto de vista de la multicolinealidad, si el coeficiente de determinación múltiple correspondiente a su regresión con respecto al resto de variables explicativas del modelo toma un valor próximo a 0 (ausencia de explicación).

Se define la *tolerancia* de una variable de la siguiente forma:

$$Tolerancia_i = \mathrm{Tol}_i = 1 - R^2_{x_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_k}$$

Esta cantidad proporcionará un valor próximo a 0 cuando el coeficiente de determinación múltiple tome un valor próximo a 1, indicativo de multicolinealidad. Por otra parte, tomará un valor próximo 1 cuando el coeficiente de determinación lineal múltiple tome un valor próximo a 0, indicando ausencia de multicolinealidad.

Adicionalmente, tal y como pudo comprobarse con anterioridad, la varianza asociada a los estimadores de los coeficientes puede expresarse de la siguiente forma:

$$\mathrm{Var}(\beta_i) = \mathrm{cantidad}_i \cdot \frac{1}{1 - R^2_{x_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_k}}$$

Que como puede observarse sería proporcional a $1/\mathrm{Tol}_i$. Por tanto, cuanto menor sea la tolerancia para una variable explicativa cualquiera, mayor será el valor de $1/\mathrm{Tol}_i$ y la varianza asociada a su coeficiente en el modelo crecerá.

Se define como factor de incremento de la varianza (FIV) a la cantidad:

$$\mathrm{FIV}_i = \frac{1}{\mathrm{Tol}_i} = \frac{1}{1 - R^2_{x_i \cdot x_1 \cdot x_2 \ldots x_{i-1} x_{i+1} \ldots x_k}}$$

NÚMERO DE CONDICIÓN E ÍNDICE DE CONDICIÓN

Otra aproximación al estudio de la multicolinealidad lo constituye el análisis de componentes principales para las variables explicativas del modelo de regresión lineal múltiple. Recuérdese que el análisis de componentes principales es utilizado habitualmente para reducir el número de variables con las que trabajar por medio de la construcción de componentes, que son combinaciones lineales de las variables utilizadas que verifican:

- Las componentes principales son linealmente independientes entre sí.
- Maximizan, en orden descendente, la varianza explicada, es decir: la primera componente principal es la que más varianza explica, seguida de la siguiente y así sucesivamente.

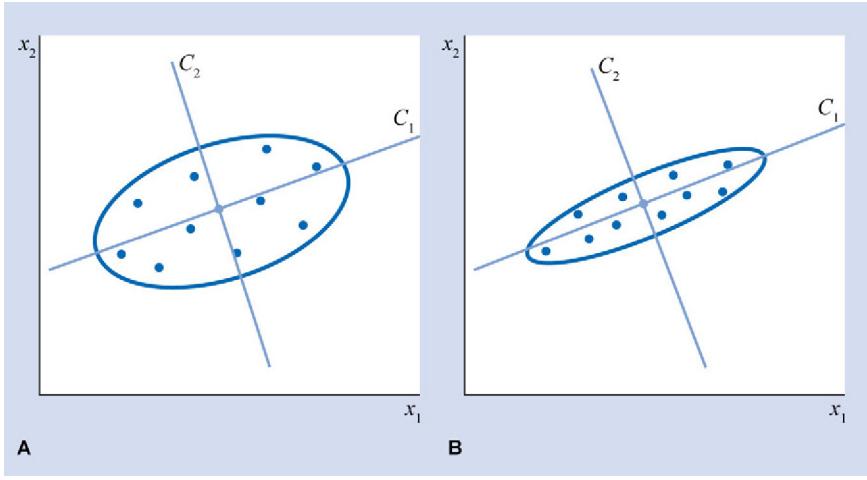

FIGURA 5-37 Representación de las componentes principales basadas en dos variables explicativas.

Si se cuenta con *k*-variables explicativas, serán necesarias *k* componentes principales para explicar el 100% de la varianza observada. En la figura 5-37 pueden observarse las componentes principales para el caso de dos variables explicativas.

A pesar de que en los dos casos son necesarias dos componentes principales para explicar el 100% de la variabilidad observada, en la figura 5-37B la primera componente C_1 logra explicar un porcentaje de variabilidad muy elevado. Esto es debido a que entre las variables x_1 y x_2 existe una muy buena relación lineal. La idea es, por tanto, que si se realiza un análisis de componentes principales para las *k* variables explicativas en el modelo de regresión lineal múltiple y el número de componentes necesarias para explicar un alto porcentaje de la variabilidad total es inferior al número de variables explicativas, será indicativo de que alguna o algunas de ellas son una buena combinación lineal de las restantes.

Si se tiene en cuenta que la varianza explicada por cada una de las componentes principales queda recogida en los valores propios de la matriz de datos X'X, para *k* variables explicativas se obtendrán *k* componentes principales y los correspondientes valores propios λ_1, λ_2, λ_3…, λ_k tal que:

$$\lambda_1 > \lambda_2 > \lambda_3 \ldots > \lambda_k$$

Cuanto más se aproxime un valor propio λ_i a 0, menos variabilidad explicará la correspondiente componente principal y las siguientes (ya que su valor propio será todavía inferior a este).

El *número de condición* se construye de la siguiente forma:

$$\textit{Número de condición}_i = \mathrm{NC}_i = \sqrt{\frac{\lambda_1}{\lambda_i}}$$

Cuanto menor sea el valor de λ_i en comparación con el primer valor propio (varianza explicada por la primera componente principal que es la que más explica) mayor será el valor de NC_i. Valores elevados de este número de condición serán indicativos, por tanto, de multicolinealidad. La idea de construcción de este índice de condición asociado a cada una de las componentes principales da lugar a la obtención de una medida global de la posible multicolinealidad basada en la comparación del primer y último valor propio λ_1 y λ_k. Así, se tendrá que:

$$\textit{Índice de condición} = \mathrm{IC} = \sqrt{\frac{\lambda_1}{\lambda_k}}$$

Análogamente al caso anterior, valores elevados del índice de condición indicarán un problema de multicolinealidad entre las variables explicativas incluidas en el modelo de regresión lineal múltiple. La mayoría de autores considera que un valor inferior a 10 implicaría que no existe un problema de multicolinealidad. Un valor entre 10 y 30 supondría una multicolinealidad moderada, mientras que un valor por encima de 30 indicaría una multicolinealidad severa.

PROPORCIONES DE LA VARIANZA

El análisis de componentes principales ofrece, además, otras posibilidades de profundización en el estudio de la multicolinealidad, analizando las «cargas» (proporción de la varianza de la variable explicativa) de cada una de las variables explicativas sobre cada una de las componentes principales. Debe tenerse en cuenta que la suma de las cargas de una variable explicativa sobre cada una de las componentes principales será 1 (entre todas las componentes principales se logra explicar el 100% de la variabilidad total). Una carga elevada de una variable explicativa sobre una componente principal cualquiera indicará que está bien representada en dicha componente. La situación ideal que sugeriría una ausencia de multicolinealidad implicaría que cada una de las variables explicativas estuviera bien representada en una componente principal distinta a las demás y, como son linealmente independientes entre sí, también lo serían las variables explicativas. Por el contrario, si dos o más variables explicativas están bien representadas en la misma componente principal indicaría un problema de multicolinealidad.

TABLA 5-14 Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,494	,244	,232	44,416

TABLA 5-15 Tabla de ANOVA

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	124.241,013	3	41.413,671	20,992	,000
	Residual	384.694,696	195	1.972,793		
	Total	508.935,709	198			

Si se trabaja con los datos del ejemplo y se incluye, además, información sobre el nivel de triglicéridos entre las variables explicativas se tendrá que el modelo quedará:

$$\text{Colesterol} = \beta_0 + \beta_1 \text{edad} + \beta_2 \text{IMC} + \beta_3 \text{triglicéridos}$$

En la tabla 5-14 puede observarse que el modelo propuesto logra explicar el 24,4% de la variabilidad de la variable nivel de colesterol (el 23,2% corregido por el número de variables incluidas en el modelo).

El contraste F del ANOVA de la regresión (tabla 5-15) indica que el modelo logra explicar significativamente, una parte de la variabilidad del nivel de colesterol que, según el resultado anterior, se sitúa en torno al 23-24%.

Los contrastes individuales de los coeficientes (tabla 5-16) indican que solo las variables edad y triglicéridos tendrían un efecto significativo sobre el nivel de colesterol. Si se compara con el modelo con dos variables explicativas ajustado con anterioridad (colesterol frente a edad e IMC) se observa que al introducir el nivel de triglicéridos el efecto del IMC pasa a ser no significativo.

No obstante, la posible existencia de buenas relaciones lineales entre las variables explicativas tendría consecuencias sobre los resultados obtenidos. En primer lugar, se proporciona información sobre las correlaciones bivariadas de las variables explicativas.

En la tabla 5-17 puede observarse que existen buenas relaciones lineales entre *edad* e *IMC* y *triglicéridos* e *IMC* que, además, son significativas al nivel $\alpha = 0{,}05$. Por tanto, el IMC tiene una buena relación lineal con las otras dos variables explicativas y puede ser una de las razones por la que no ha podido mantener su significación en el modelo.

En la tabla 5-18 se muestran los valores de las tolerancias y factores de incremento de la varianza (FIV). Como puede observarse, el valor de la tolerancia asociado al IMC (Tol = 0,769) es el menor de los tres que han sido calculados y el que proporciona un mayor incremento de la varianza asociada a su coeficiente del modelo de regresión lineal múltiple (FIV = 1,301). No obstante los valores de las tolerancias son superiores en todos los casos a 0,75.

TABLA 5-16 Coeficientes

Modelo	Coeficientes no estandarizados		Coeficientes tipificados		t	Sig.
	B	Error típ.	Beta			
1	(Constante)	121,43	23,235		5,226	,000
	Edad	,955	,224	,291	4,261	,000
	IMC	1,592	,972	,116	1,637	,103
	Triglicéridos	,149	,036	,271	4,12	,000

TABLA 5-17 Correlaciones entre las variables explicativas

	Edad	IMC	Triglicéridos
Edad	1	,408	,161
IMC	,408	1	,316
Triglicéridos	,161	,316	1

Por último, se realiza un análisis de componentes principales. Habitualmente, los programas de análisis estadístico incorporan al análisis de componentes principales, en el caso de la regresión lineal múltiple, una dimensión por cada coeficiente del modelo, incluida la constante β_0. Por tanto, en el modelo de regresión del ejemplo se obtendrán un total de cuatro componentes principales (en el caso de la componente asociada a la constante se trabaja como si tuviera asociada una variable que tomara el valor constante 1).

En la tabla 5-19 se muestran los valores propios (autovalores) que cuantifican la varianza explicada por cada una de las componentes principales:

$$\lambda_1 = 3,656; \lambda_2 = 0,279; \lambda 3 = 0,057; \lambda 4 = 0,008$$

Puede observarse que la primera componente principal es la que más varianza explica, seguida por la segunda y, así, sucesivamente. El número de condición asociado a cada una de las componentes principales y el índice de condición sería:

TABLA 5-18 Tolerancias y factores de incremento de la varianza

Modelo		Estadísticos de colinealidad	
		Tolerancia	FIV
1	(Constante)		
	Edad	,832	1,202
	IMC	,769	1,301
	Triglicéridos	,899	1,113

TABLA 5-19 Análisis de componentes principales

Dimensión	Autovalores	Índice de condición	Proporciones de la varianza			
			(Constante)	Edad	IMC	Triglicéridos
1	3,656	1	,00	,01	,00	,02
2	,279	3,619	,01	,02	,00	,93
3	,057	8,02	,07	,92	,03	,00
4	,008	20,773	,92	,05	,97	,05

$$NC_1 = 1; NC_2 = 3,619; NC_3 = 8,02; NC_4 = 20,773$$

$$\text{Índice de condición} = \text{IC} = \sqrt{\lambda_1 / \lambda_4} = 3,656 / 0,008 = 20,773$$

Los tres primeros números de condición son inferiores a 10, sugiriendo ausencia de multicolinelidad. Sin embargo, el último número de condición asociado a la última componente principal (equivalente al índice de condición) toma un valor entre 10 y 30 sugiriendo una multicolinealidad moderada. Obsérvese que es en esta última componente principal donde se encuentra muy bien representada la variable IMC (su carga es 0,97 en esta componente). Por otra parte, cada una de las variables explicativas está bien representada en una componente principal distinta: la variable *edad* está bien representada en la tercera componente principal con una carga de 0,92, y el *nivel de triglicéridos* en la segunda componente principal con una carga de 0,93.

Lo ideal sería que, efectivamente, cada una de las variables explicativas estuviera bien representada en una componente principal distinta como es el caso, aunque algo ocurre con la variable *IMC*. Si se observa la variable que acompaña a la constante, puede apreciarse que está bien representada en la última componente principal, al igual que la variable *IMC*, y con una carga de 0,92. Esto indicaría, en principio, una buena relación lineal entre ambas. ¿Qué quiere decir esto?

Si la varianza asociada a una variable explicativa es pequeña en relación al resto de variables presentes en el modelo de regresión lineal múltiple, podría asociarse linealmente, de forma aceptable, con una variable que tomara el valor constante 1 (variable que acompaña al término de interceptación) provocando un problema de colinealidad. Esto sucede, habitualmente, cuando las unidades o escalas de medida de las variables explicativas o predictoras son muy diferentes entre sí. Algunos autores proponen no tener en cuenta la colinealidad respecto al término de interceptación, puesto que no se corresponde con un problema generado entre las propias variables explicativas y que podría corregirse: 1) centrando (restando a cada variable en

TABLA 5-20 Análisis de componentes principales con las variables centradas

Dimensión	Autovalores	Índice de condición	Proporciones de la varianza		
			Edad	IMC	Triglicéridos
1	1,602	1	,2	,17	,14
2	,845	1,377	,01	,34	,69
3	,553	1,701	,79	,49	,17

el modelo, incluida la variable dependiente, su media), o 2) estandarizando las variables (restando a cada variable su media y dividiendo por su desviación típica). A saber:

$$y - \overline{y} = \beta_1 (x_1 - \overline{x}_1) + \beta_2 (x_2 - \overline{x}_2) + \ldots + \beta_k (x_k - \overline{x}_k)$$

$$\frac{y - \overline{y}}{S_y} = \beta_1 \frac{(x_1 - \overline{x}_1)}{S_{x1}} + \beta_2 \frac{(x_2 - \overline{x}_2)}{S_{x2}} + \ldots + \beta_k \frac{(x_k - \overline{x}_k)}{S_{xk}}$$

Aplicando estas estrategias de análisis a los datos del ejemplo se obtendrían los resultados descritos en la tabla 5-20.

Puede observarse que tanto el número de condición asociado a cada componente principal como el índice de condición toman, en todos los casos, un valor inferior a 10. Por otra parte, las variables *IMC* y *edad* muestran cargas relativamente elevadas en la misma componente principal (0,79 y 0,49 respectivamente en la tercera componente principal), aunque parecen no ser problemáticas porque están asociadas a un número de condición muy pequeño.

En general, si en una misma componente principal se observan dos o más proporciones de la varianza superiores o iguales a 0,5 (dos o más variables bien representadas en la misma componente principal) se sospechará un problema de multicolinealidad y tendrá que ser estudiado el efecto de la convivencia de dichas variables en el modelo sobre su significación estadística. Especialmente graves pueden ser las situaciones en las que dos o más proporciones de la varianza sean superiores a 0,9.

Entre las soluciones para resolver el problema de la multicolinealidad cabe mencionar las siguientes:

- Si las variables implicadas no alcanzan significación estadística en los contrastes individuales sobres los coeficientes del modelo, podría ensayarse la eliminación de una de ellas con objeto de valorar si las restantes pudieran pasar a ser significativas.
- Aumentar, si es posible, el tamaño de la muestra, de forma que pueda reducirse la varianza asociada a los coeficientes del modelo e incrementar las posibilidades de detección de significación estadística.

- Transformar las variables. Entre las posibilidades se encuentra el centrado y la estandarización.
- Utilizar las componentes principales construidas a partir de las variables explicativas como las variables predictoras del modelo.

AUTOEVALUACIÓN

1. El coeficiente de correlación lineal de Pearson para dos variables es de 0,38; entonces:
   a. Por aumentar en 1 unidad una de las variables se produce un aumento de 0,38 unidades en la otra variable.
   b. Es posible que exista una relación lineal directa entre las dos variables.
   c. Es posible que exista una relación lineal inversa entre las dos variables.
   d. No existe relación lineal entre las variables puesto que el valor es muy bajo.
   e. La relación no es lineal.

2. El coeficiente de determinación de la recta:
   a. Se interpreta como el cambio en la variable dependiente por unidad de cambio en la variable independiente.
   b. Es una medida de la influencia de una observación sobre los coeficientes del modelo.
   c. Es una medida de la pendiente de la recta.
   d. Toma valores entre –1 y 1.
   e. Se interpreta como la proporción de la varianza explicada por el modelo.

3. El modelo de regresión lineal múltiple ajustado es $y = 2 + 3x_1 - 5x_2$.
   a. Por aumento en una unidad de x_1 se produce una disminución de 5 unidades en y.
   b. Por aumento en una unidad de x_2 se produce una aumento de 3 unidades en y.
   c. Se sospecha una relación lineal directa entre y y cada una de las dos variables explicativas.
   d. Se verifica:

$$y = \beta_0 + 3x_1$$
$$y = \beta_0^* - 5x_2$$

   e. Cada uno de los coeficientes (3 y –5) está ajustado por la presencia de las variables explicativas incluidas en el modelo.

4. La distancia de Cook:
   a. Es una medida de la variabilidad de las predicciones.
   b. Es una medida del efecto de una variable explicativa sobre otra.

c. Es una medida de la influencia de una observación sobre el conjunto de coeficientes del modelo.
d. Es una medida de la influencia de una observación sobre las predicciones del modelo.
e. Es una medida de la influencia de una variable en el modelo de regresión.

5. El índice de condición:
a. Es una medida de la influencia de una observación sobre el modelo de regresión.
b. Es una medida útil para valorar la independencia de las observaciones.
c. Es útil para valorar la posible multicolinealidad entre las variables explicativas.
d. Se utiliza para comprobar la hipótesis de homogeneidad de varianzas.
e. Cuanto mayor sea su valor menor es la posible multicolinealidad entre las variables explicativas.

APÉNDICE

Respuestas de las autoevaluaciones

CAPÍTULO 1

1. Respuesta correcta: *d.*
Toma un número finito o infinito numerable de valores (obedece a la idea de contar cuántas veces ocurre un suceso).

2. Respuesta correcta: *e.*
La mediana no se ve afectada por observaciones atípicas, mientras que la media sí. La mediana, por construcción, deja el mismo número de observaciones por encima que por debajo de ella.

3. Respuesta correcta: *a.*
Probabilidad del suceso complementario.

4. Respuesta correcta: *a.*
El valor predictivo positivo se define como $P(E/+)$, la probabilidad de que un individuo que ha dado positivo en la prueba padezca la enfermedad.

5. Respuesta correcta: *d.*
La media coincide con la mediana y la moda. El percentil 0,975 en la normal estándar es 1,96. La media depende del modelo normal del que se trate. El coeficiente de asimetría es 0 por ser simétrica. El 50% de los valores serán superiores a la media, ya que coincide con la mediana.

CAPÍTULO 2

1. Respuesta correcta: *e.*
El muestreo aleatorio no garantiza muestras representativas, aunque intenta conseguirlas. Es aplicable en poblaciones finitas e infinitas. Cuando se muestrea siempre se cometerá un error muestral.

2. Respuesta correcta: *e.*
Véase el apartado «Nivel de confianza y precisión», en el que se discute que a mayor nivel de confianza menor precisión y, en consecuencia, a menor nivel de confianza mayor precisión para los mismos datos.

3. Respuesta correcta: *c.*
Cuando se pretende cuantificar el valor de un parámetro desconocido de la población debe construirse un intervalo de confianza. La

muestra siempre debe ser aleatoria. No tiene ningún sentido realizar un contraste ni comparar las proporciones de afectados y no afectados (una es complementaria de la otra).

4. Respuesta correcta: *a.*
   Véase el apartado «Errores en un contraste de hipótesis». El nivel α representa la probabilidad de cometer un error de tipo I. Es la hipótesis nula la que se mantendrá como cierta a menos que los datos muestren evidencia de lo contrario. La región crítica de contraste es la región de rechazo de la hipótesis nula. Si el p-valor es mayor que el nivel α, se procederá a aceptar la hipótesis nula.
5. Respuesta correcta: *d.*
   Las muestras son apareadas o relacionadas, ya que para cada individuo se cuenta con una pareja de observaciones (antes-después). El tamaño muestral es suficiente, ya que permite la aplicación del teorema central del límite ($n \geq 30$).

CAPÍTULO 3

1. Respuesta correcta: *b.*
   Las pruebas no paramétricas o de libre distribución no requieren la suposición de una distribución de probabilidad para la variable objeto de estudio. Requieren muestreo aleatorio. Tienen menor capacidad para detectar significación estadística y, por esta razón, suelen utilizarse cuando no se cumplen los criterios de aplicación de las pruebas paramétricas.
2. Respuesta correcta: *e.*
   El tamaño muestral (inferior a 30) impide la utilización de la prueba t para una media. Se precisa la utilización de una prueba no paramétrica que no necesita suposiciones sobre la distribución de probabilidad de la variable. La prueba de la mediana permitiría contrastar la hipótesis de que la media poblacional fuera distinta de 220 mg/100 ml.
3. Respuesta correcta: *a.*
   La prueba de McNemar está indicada para comparar proporciones para muestras apareadas. Se precisa conocer si las muestras son independientes o apareadas para poder seleccionar la prueba más adecuada.
4. Respuesta correcta: *d.*
   La prueba más indicada sería la prueba ji-cuadrado o la prueba exacta de Fisher, dependiendo de si el número de casillas con un valor esperado inferior a 5 es inferior o superior al 20%. No se requieren hipótesis adicionales. La prueba de McNemar está indicada para muestras relacionadas, algo imposible en este caso porque el número de sujetos en cada uno de los grupos es distinto.

CAPÍTULO 4

1. Respuesta correcta: *d.*
El tamaño es suficiente (más de 30 casos por grupo). Diferencias en las medias muestrales no implican diferencias en las medias poblacionales. Las muestras no son relacionadas (distintos tamaños por grupo).

2. Respuesta correcta: *a.*
Deben verificarse: hipótesis de normalidad de la variable dependiente, homogeneidad de varianzas (en todos los grupos) e independencia de las observaciones. En la práctica se requieren al menos 30 datos por grupo.

3. Respuesta correcta: *b.*
Compara parejas y combinaciones lineales de las mismas. Es una técnica conservadora que tiende a detectar menos diferencias de las que debiera. Debe realizarse después del ANOVA. Las varianzas deben ser homogéneas en los grupos considerados.

4. Respuesta correcta: *c.*
Existen pruebas alternativas como las de Welch o Brown-Forsythe.

CAPÍTULO 5

1. Respuesta correcta: *b.*
El coeficiente de correlación lineal de Pearson es una medida de la magnitud de la asociación lineal entre dos variables. Dado que es positivo, se sospecha relación directa, pero será necesario comprobar su significación estadística. Se desconoce si la relación es no lineal.

2. Respuesta correcta: *e.*
Véase la definición del coeficiente de determinación. Toma valores entre 0 y 1 (se interpreta como una proporción). Es una medida del ajuste del modelo y no de la influencia de una observación.

3. Respuesta correcta: *e.*
Por aumento en una unidad de la variable x_1 se produciría un aumento de 3 unidades en la variable dependiente y manteniendo constante el valor de la otra variable explicativa. Se sospecha una relación directa entre x_1 e y; inversa entre x_2 e y. Los coeficientes 3 y −5 solo coincidirán con los de las respectivas regresiones lineales simples si el coeficiente de correlación lineal entre x_1 y x_2 es 0.

4. Respuesta correcta: *c.*
Véase la distancia de Cook en el apartado «Valores de influencia». Es una medida de la influencia de cada observación sobre el conjunto de los coeficientes del modelo, incluido el término *interceptación.*

5. Respuesta correcta: *c.*
El número de condición se usa para valorar la multicolinealidad de las variables predictoras. Cuanto mayor sea su valor, mayor posibilidad de existencia de multicolinealidad entre las variables explicativas.