Schaum

ESTADÍSTICA

SEGUNDA EDICIÓN

Murray R. Spiegel

- 975 problemas resueltos con soluciones completamente detalladas
- Más de 700 problemas suplementarios con solución
- Especial énfasis en la comprensión de los métodos de resolución de problemas prácticos
- Abarca los aspectos teóricos esenciales de la estadística

Utilizado por millones de estudiantes y recomendado por profesores de todo el mundo

Mc Graw Hill

ESTADISTICA

Segunda edición

ESTADISTICA

Segunda edición

MURRAY R. SPIEGEL
Hartford Graduate Center

Traducción
RAFAEL HERNANDEZ HEREDERO
Dpto. de Métodos Matemáticos de la Física
Universidad Complutense de Madrid

Revisión Técnica
LORENZO ABELLANAS RAPUN
Catedrático de Métodos Matemáticos de la Física
Universidad Complutense de Madrid



MADRID • BUENOS AIRES • CARACAS • GUATEMALA • LISBOA
MEXICO • NUEVA YORK • PANAMA • SAN JUAN • SANTA FE DE BOGOTA • SANTIAGO • SAO PAULO
AUCKLAND • HAMBURGO • LONDRES • MILAN • MONTREAL • NUEVA DELHI
PARIS • SAN FRANCISCO • SIDNEY • SINGAPUR • ST. LOUIS • TOKIO • TORONTO

ESTADISTICA (Segunda edición)

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

DERECHOS RESERVADOS © 1991, respecto a la primera edición en español por McGRAW-HILL/INTERAMERICANA DE ESPAÑA, S. A.

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

Traducido de la segunda edición en inglés de **STATISTICS**

Copyright © MCMLXXXVIII, por McGraw-Hill, Inc.

ISBN: 0-07-060234-4

ISBN: 84-7615-562-X
Depósito legal: M. 522-1997

Fotocompuesto en MonoComp, S. A.

IMPRESO POR INDUSTRIAS GRAFICAS S.A.

IMPRESO EN CHILE - PRINTED IN CHILE

519.5
Sp43
1991
(BC)

Contenido

UNIVERSIDAD DE CONCEPCION
DIRECCION
BIBLIOTECA
SECCION
PROCESOS
TECNICOS

PROLOGO xi

Capítulo *1* **VARIABLES Y GRAFICOS** 1
Estadística. Población y muestreo; estadística inductiva y descriptiva. Variables: discretas y continuas. Redondeo de datos. Notación científica. Dígitos significativos. Cálculos. Funciones. Coordenadas rectangulares. Gráficos. Ecuaciones. Desigualdades. Logaritmos. Antilogaritmos. Cálculos usando logaritmos.

Capítulo *2* **DISTRIBUCIONES DE FRECUENCIAS** 37
Filas de datos. Ordenaciones. Distribuciones de frecuencias. Intervalos de clase y límites de clase. Fronteras de clase. Tamaño o anchura de un intervalo de clase. Marca de clase. Reglas generales para formar distribuciones de frecuencias. Histogramas y polígonos de frecuencias. Distribuciones de frecuencias relativas. Distribuciones de frecuencias acumuladas y ojivas. Distribuciones de frecuencias relativas y ojivas de porcentajes. Curvas de frecuencia y ojivas suavizadas. Tipos de curvas de frecuencia.

Capítulo *3* **MEDIA, MEDIANA, MODA Y OTRAS MEDIDAS DE TENDENCIA CENTRAL** 60
Notación de índices. Notación de suma. Promedios o medidas de tendencia central. La media aritmética. La media aritmética ponderada. Propiedades de la media aritmética. Cálculo de la media aritmética para datos agrupados. La mediana. La moda. Relación empírica entre media, mediana y moda. La media geométrica *G*. La media armónica *H*. Relación entre las medias aritmética, geométrica y armónica. La media cuadrática (MQ). Cuartiles, deciles y percentiles.

Capítulo *4* **LA DESVIACION TIPICA Y OTRAS MEDIDAS DE DISPERSION** 91
Dispersión o variación. El rango. La desviación media. El rango semi-intercuartil. El rango percentil 10-90. La desviación típica. La varianza. Métodos cortos para calcular la desviación típica. Propiedades de la desviación típica. Comprobación de Charlier. Corrección

252874

de Sheppard para la varianza. Relaciones empíricas entre medidas de dispersión. Dispersión absoluta y relativa; coeficiente de variación. Variables tipificadas: unidades estándar.

Capítulo 5 **MOMENTOS, SESGO Y CURTOSIS** 116
Momentos. Momentos para datos agrupados. Relaciones entre momentos. Cálculo de momentos para datos agrupados. Comprobación de Charlier y correcciones de Sheppard. Momentos adimensionales. Sesgo. Curtosis. Momentos, sesgo y curtosis de una población.

Capítulo 6 **TEORIA ELEMENTAL DE PROBABILIDADES** 129
Definiciones de probabilidad. Probabilidad condicional; sucesos independientes y sucesos dependientes. Sucesos mutuamente excluyentes. Distribuciones de probabilidad. Esperanza matemática. Relación entre población, media muestral y varianza. Análisis combinatorio. Combinaciones. Aproximación de Stirling a $n!$. Relación de la probabilidad con la teoría de conjuntos.

Capítulo 7 **LAS DISTRIBUCIONES BINOMIAL, NORMAL Y DE POISSON** 159
La distribución binomial. La distribución normal. Relación entre la distribución binomial y la distribución normal. La distribución de Poisson. Relación entre la distribución binomial y la distribución de Poisson. La distribución multinomial. Ajuste de distribuciones de frecuencias muestrales mediante distribuciones teóricas.

Capítulo 8 **TEORIA ELEMENTAL DEL MUESTREO** 186
Teoría del muestreo. Muestras aleatorias y números aleatorios. Muestreo con y sin reposición. Distribuciones de muestreo. Distribución de muestreo de medias. Distribución de muestreo de proporciones. Distribución de muestreo de diferencias y sumas. Errores típicos.

Capítulo 9 **TEORIA DE LA ESTIMACION ESTADISTICA** 208
Estimación de parámetros. Estimaciones sin sesgo. Estimación eficiente. Estimaciones de punto y estimaciones de intervalo; su fiabilidad. Estimaciones de intervalo de confianza para parámetros de población. Error probable.

Capítulo 10 **TEORIA ESTADISTICA DE LAS DECISIONES** 223
Decisiones estadísticas. Hipótesis estadísticas. Contrastes de hipótesis y significación, o reglas de decisión. Errores de Tipo I y de Tipo II. Nivel

de significación. Contrastes mediante la distribución normal. Contrastes de una y de dos colas. Contrastes especiales. Curvas de operación características; potencia de un contraste. Gráficos de control. Contrastes mediante diferencias muestrales. Contrastes mediante la distribución binomial.

Capítulo *11* **TEORIA DE PEQUEÑAS MUESTRAS** 251
Pequeñas muestras. Distribución *t* de Student. Intervalos de confianza. Contrastes de hipótesis y significación. Distribución ji-cuadrado. Intervalos de confianza para la distribución ji-cuadrado. Grados de libertad. La distribución *F*.

Capítulo *12* **TEST JI-CUADRADO** 268
Frecuencias observadas y teóricas. Definición de χ^2. Contrastes de significación. El test ji-cuadrado para la bondad de ajuste. Tablas de contingencia. Corrección de Yates a la continuidad. Fórmulas simples para calcular. Coeficiente de contingencia. Correlación de atributos. Propiedad aditiva de χ^2.

Capítulo *13* **AJUSTE DE CURVAS Y EL METODO DE MINIMOS CUADRADOS** 289
Relaciones entre variables. Ajuste de curvas. Ecuaciones de curvas aproximantes. Ajuste de curvas a mano. La recta. El método de mínimos cuadrados. La recta de mínimos cuadrados. Relaciones no lineales. La parábola de mínimos cuadrados. Regresión. Aplicaciones a series en el tiempo. Problemas en más de dos variables.

Capítulo *14* **TEORIA DE LA CORRELACION** 322
Correlación y regresión. Correlación lineal. Medidas de correlación. La recta de regresión de mínimos cuadrados. Error típico de estimación. Variación explicada y variación inexplicada. Coeficiente de correlación. Observaciones sobre el coeficiente de correlación. Fórmulas momento-producto para el coeficiente de correlación lineal. Fórmulas cortas de cálculo. Rectas de regresión y el coeficiente de correlación lineal. Correlación de series en el tiempo. Correlación de atributos. Teoría muestral de la correlación. Teoría muestral de la regresión.

Capítulo *15* **CORRELACION MULTIPLE Y PARCIAL** 357
Correlación múltiple. Notación de subíndices. Ecuaciones de regresión y planos de regresión. Ecuaciones normales para el plano de regresión de mínimos cuadrados. Planos de regresión y coeficientes de correlación. Error típico de estimación. Coeficiente de correlación múltiple. Cambio

de variable dependiente. Generalización a más de tres variables. Correlación parcial. Relaciones entre coeficientes de correlación parcial y múltiple. Regresión múltiple no lineal.

Capítulo *16* **ANALISIS DE VARIANZA** 375
Objetivo del análisis de varianza. Experimentos de factor único. Variación total, variación dentro de los tratamientos y variación entre tratamientos. Métodos abreviados para calcular variaciones. Modelos matemáticos para el análisis de varianza. Valores esperados de las variaciones. Distribuciones de las variaciones. El contraste F para la hipótesis nula de igualdad de medias. Tablas de análisis de varianza. Modificaciones para números distintos de observaciones. Experimentos de dos factores. Notación para experimentos de dos factores. Variaciones para experimentos de dos factores. Análisis de varianza para experimentos de dos factores. Experimentos de dos factores con repetición. Diseño experimental.

Capítulo *17* **CONTRASTES NO PARAMETRICOS** 411
Introducción. El test de los signos. El U-test de Mann-Whitney. El H-test de Kruskal-Wallis. El H-test corregido por coincidencias. El test de las rachas para el carácter aleatorio. Otras aplicaciones del test de las rachas. Correlación de rango de Spearman.

Capítulo *18* **ANALISIS DE SERIES EN EL TIEMPO** 440
Series en el tiempo. Gráficos de series en el tiempo. Movimientos característicos de series en el tiempo. Clasificación de movimientos de series en el tiempo. Análisis de series en el tiempo. Promedios móviles; suavización de series en el tiempo. Estimación de la tendencia. Estimación de las variaciones estacionales; el índice estacional. Datos ajustados a la variación estacional. Estimación de las variaciones cíclicas. Estimación de las variaciones irregulares. Comparación de datos. Predicción. Resumen de los pasos fundamentales en el análisis de series en el tiempo.

Capítulo *19* **NUMEROS INDICE** 478
Número índice. Aplicaciones de los números índice. Relaciones de precios. Propiedades de las relaciones de precios. Relaciones de cantidad o de volumen. Relaciones de valor. Relaciones de enlace y en cadena. Problemas implícitos en el cálculo de números índice. El uso de promedios. Criterios teóricos para números índice. Notación. El método de agregación simple. El método del promedio simple de relaciones. El método de agregación ponderada. Indice ideal de Fisher. El índice de Marshall-Edgeworth. El método del promedio ponderado de relaciones. Números índice de cantidad o volumen. Números índice de valor. Cambio del período base en los números índice. Deflación de series en el tiempo.

SOLUCIONES A LOS PROBLEMAS SUPLEMENTARIOS **511**

APENDICES **533**

I Ordenadas (Y) de la curva normal canónica en z 535

II Areas bajo la curva normal canónica entre 0 y z 536

III Valores percentiles (t_p) para la distribución t de Student con v grados de libertad 537

IV Valores percentiles (χ^2_p) para la distribución ji-cuadrado con v grados de libertad 538

V Valores de los 95-ésimos percentiles para la distribución F 539

VI Valores de los 99-ésimos percentiles para la distribución F 540

VII Logaritmos decimales con cuatro cifras 541

VIII Valores de $e^{-\lambda}$ 544

IX Números aleatorios 545

INDICE **546**

Prólogo

La Estadística o los métodos estadísticos, como se denomina a veces, está jugando un papel más y más importante en casi todas las facetas del comportamiento humano. Ocupada inicialmente en asuntos de Estado, y de ahí su nombre, la influencia de la Estadística se ha extendido ahora a la agricultura, biología, negocios, química, comunicaciones, economía, educación, electrónica, medicina, física, ciencias políticas, psicología, sociología y otros muchos campos de la ciencia y la ingeniería.

El propósito de este libro es presentar una introducción a los principios básicos de la Estadística que serán de utilidad con independencia del campo de interés específico del lector. Se ha diseñado para ser usado como suplemento a un texto estándar o como libro de texto para un curso formal de Estadística. Será de considerable interés, asimismo, como libro de consulta, para todos aquellos que estén implicados en aplicar la Estadística a sus propios problemas de investigación.

Cada capítulo comienza con enunciados claros de las definiciones pertinentes, teoremas y principios, junto con otro material ilustrativo y descriptivo. Ello viene seguido de problemas resueltos y suplementarios que en muchos casos utilizan datos obtenidos en situaciones estadísticas reales. Los problemas resueltos sirven para ilustrar y ampliar la teoría, arrojan luz sobre los puntos sutiles, sin lo cual el estudiante se sentiría siempre sobre arenas movedizas, y proporcionan la oportunidad de repetir los principios básicos, vital para un aprendizaje eficaz. Numerosas demostraciones de fórmulas han quedado incluidas entre los problemas resueltos. El elevado número de problemas suplementarios con solución, completa la revisión del material expuesto en cada capítulo.

La única base matemática requerida para la comprensión del libro consiste en aritmética y rudimentos de álgebra. En el primer capítulo se presenta un repaso de los conceptos matemáticos usados posteriormente. Puede leerse al comienzo o guardarlo como referencia para cuando sea preciso.

La primera parte del libro trata el análisis de las distribuciones de frecuencia y las medidas asociadas de tendencia central, dispersión, sesgo (asimetría) y curtosis (aplastamiento). Lo cual conduce naturalmente a una discusión de teoría elemental de probabilidades y sus aplicaciones, que allana el camino para la teoría del muestreo. Se consideran en primer lugar las técnicas de grandes muestras, que involucran a la distribución normal, y aplicaciones a la estimación estadística y al contraste de hipótesis y significación. La teoría de pequeñas muestras, que emplea la distribución t de Student, la ji-cuadrado y la distribución F, aparece en un capítulo posterior, junto con sus aplicaciones. Otro capítulo sobre ajuste de curvas y el método de mínimos cuadrados lleva lógicamente a los temas de correlación y regresión en dos variables. La correlación parcial y múltiple, en más de dos variables, se estudia en un capítulo aparte. Luego siguen capítulos sobre el análisis de varianza y los métodos no paramétricos, nuevos en esta segunda edición. Dos capítulos finales tratan el análisis de series en el tiempo y los números índice, respectivamente.

Hemos incluido más material del que puede cubrirse en un curso habitual, con el fin de hacer el libro más flexible, ampliarlo y mejorarlo como libro de consulta y estimular el interés por otros temas. Al usar el libro es posible alterar el orden de muchos capítulos e incluso omitir algunos. Así,

los Capítulos 13-15 y 18-19 en su casi totalidad pueden introducirse tras el Capítulo 5, si se desea estudiar correlación, regresión, series en el tiempo y números índice antes que la teoría de muestreo. Análogamente, el Capítulo 6 puede omitirse casi completo si no se quiere perder mucho tiempo en las probabilidades. En un primer curso, todo el Capítulo 15 pude ser omitido. Hemos elegido el orden que aparece porque existe la tendencia creciente, en los cursos modernos, de introducir la teoría del muestreo y la inferencia estadística lo antes posible.

Deseo agradecer a las diversas instituciones, tanto gubernamentales como privadas, por su cooperación al proporcionarme datos para las tablas. En el texto figuran las referencias oportunas a las fuentes consultadas. En particular, estoy agradecido al profesor Sir Ronald A. Fisher, F. R. S., Cambridge; doctor Frank Yates, F. R. S., Rothamsted, y Messrs. Oliver y Boyd Ltd., Edinburgh, por conceder autorización para utilizar los datos de la Tabla III de su libro *Statistical Tables for Biological, Agricultural, and Medical Research*. Quiero dar las gracias, asimismo, a Esther y Meyer Scher por su apoyo y al personal de McGraw-Hill por su colaboración.

MURRAY R. SPIEGEL

CAPITULO 1

Variables y gráficos

ESTADISTICA

La Estadística estudia los métodos científicos para recoger, organizar, resumir y analizar datos, así como para sacar conclusiones válidas y tomar decisiones razonables basadas en tal análisis.

En un sentido menos amplio, el término *estadística* se usa para denotar los propios datos, o números derivados de ellos, tales como los promedios. Así se habla de estadística de empleo, estadística de accidentes, etc.

POBLACION Y MUESTREO; ESTADISTICA INDUCTIVA Y DESCRIPTIVA

Al recoger datos relativos a las características de un grupo de individuos u objetos, sean alturas y pesos de estudiantes de una universidad o tuercas defectuosas producidas en una fábrica, suele ser imposible o nada práctico observar todo el grupo, en especial si es muy grande. En vez de examinar el grupo entero, llamado *población* o *universo*, se examina una pequeña parte del grupo, llamada *muestra*.

Una población puede ser *finita* o *infinita*. Por ejemplo, la población consistente en todas las tuercas producidas por una fábrica un cierto día es finita, mientras que la determinada por todos los posibles resultados (caras, cruces) de sucesivas tiradas de una moneda, es infinita.

Si una muestra es representativa de una población, es posible inferir importantes conclusiones sobre la población a partir del análisis de la muestra. La fase de la Estadística que trata con las condiciones bajo las cuales tal diferencia es válida se llama *estadística inductiva* o *inferencia estadística*. Ya que dicha inferencia no es del todo exacta, el lenguaje de las *probabilidades* aparecerá al establecer nuestras conclusiones.

La parte de la Estadística que sólo se ocupa de describir y analizar un grupo dado, sin sacar conclusiones sobre un grupo mayor, se llama *estadística descriptiva* o *deductiva*.

Antes de entrar en el estudio de la Estadística, recordemos algunas nociones matemáticas relevantes.

VARIABLES: DISCRETAS Y CONTINUAS

Una *variable* es un símbolo, tal como X, Y, H, x o B, que puede tomar un conjunto prefijado de valores, llamado *dominio* de esa variable. Si la variable puede tomar un solo valor, se llama *constante*.

Una variable que puede tomar cualquier valor entre dos valores dados se dice que es una *variable continua*; en caso contrario diremos que la *variable es discreta*.

EJEMPLO 1. El número N de hijos en una familia puede ser 0, 1, 2, 3, ... pero no 2.5 ó 3.842. Es una variable discreta.

EJEMPLO 2. La altura H de una persona, que puede ser 62 pulgadas (abreviatura «in»), 63.8 in o 65.8341 in, dependiendo de la precisión de la medida, es una variable continua.

Los datos que admiten descripción mediante una variable discreta o continua se denominan respectivamente *datos discretos* y *continuos*. El número de hijos en cada una de 1000 familias es un ejemplo de datos discretos, mientras que las alturas de 100 universitarios lo es de datos continuos. En general, las *mediciones* dan lugar a datos continuos, y las *enumeraciones o recuentos*, a datos discretos.

A veces conviene extender la noción de variable a entidades no numéricas; por ejemplo, el color C en un arco iris es una variable que puede tomar los «valores» rojo, anaranjado, amarillo, verde, azul, añil y violeta. Suele ser posible sustituir tales variables por entidades numéricas; por ejemplo, denotando el rojo como 1, el anaranjado como 2, etc.

REDONDEO DE DATOS

El resultado de redondear un número como 72.8 en unidades es 73, pues 72.8 está más próximo de 73 que de 72. Análogamente, 72.8146 se redondea en centésimas (o sea con dos decimales) a 72.81, porque 72.8146 está más cerca de 72.81 que de 72.82.

Al redondear 72.465 en centésimas nos hallamos ante un dilema, ya que está *equidistante* de 72.46 y de 72.47. Se adopta en tales casos la costumbre de redondear al *entero par* que preceda al 5. Así pues, 72.465 se redondea a 72.46, 183.575 se redondea a 183.58 y 116,500,000 se redondea en millones a 116,000,000. Esta estrategia es particularmente útil para minimizar los *errores de redondeo acumulados* cuando se efectúa un gran número de operaciones (véase Prob. 1.4).

NOTACION CIENTIFICA

Al escribir números, especialmente los que tienen muchos ceros antes o después del punto decimal, interesa emplear la notación científica mediante potencias de 10.

EJEMPLO 3. $10^1 = 10$, $10^2 = 10 \times 10 = 100$, $10^5 = 10 \times 10 \times 10 \times 10 \times 10 = 100{,}000$ y $10^8 = 100{,}000{,}000$.

EJEMPLO 4. $10^0 = 1$; $10^{-1} = .1$, o sea 0.1; $10^{-2} = .01$, o sea 0.01, y $10^{-5} = .00001$, o sea 0.00001.

EJEMPLO 5. $864{,}000{,}000 = 8.64 \times 10^8$, y $0.00003416 = 3.416 \times 10^{-5}$.

Nótese que al multiplicar un número por 10^8, por ejemplo, el punto decimal se mueve ocho posiciones *a la derecha*, y al multiplicar por 10^{-6} se mueve seis posiciones *a la izquierda*.

A menudo escribiremos 0.1253 en vez de .1253 para recalcar el hecho de que no se ha omitido accidentalmente un entero no nulo delante del punto decimal. Sin embargo, ese cero puede omitirse cuando no exista riesgo de confusión, por ejemplo, en tablas.

Con frecuencia usamos paréntesis o puntos para denotar el producto de dos o más números. Así pues, $(5)(3) = 5 \cdot 3 = 5 \times 3 = 15$, y $(10)(10)(10) = 10 \cdot 10 \cdot 10 = 10 \times 10 \times 10 = 1000$. Si se usan letras para representar los números, se suelen omitir los paréntesis y los puntos; por ejemplo, $ab = (a)(b) = a \cdot b = a \times b$.

La notación científica resulta útil en el cálculo, sobre todo para localizar puntos decimales. Se utilizan entonces las reglas

$$(10^p)(10^q) = 10^{p+q} \qquad \frac{10^p}{10^q} = 10^{p-q}$$

donde p y q son números arbitrarios.

En 10^p, p se llama *exponente* y 10 *base*.

EJEMPLO 6. $(10^3)(10^2) = 1000 \times 100 = 100{,}000 = 10^5$ es decir, 10^{3+2}

$$\frac{10^6}{10^4} = \frac{1{,}000{,}000}{10{,}000} = 100 = 10^2 \quad \text{es decir, } 10^{6-4}$$

EJEMPLO 7. $(4{,}000{,}000)(0.0000000002) = (4 \times 10^6)(2 \times 10^{-10}) = (4)(2)(10^6)(10^{-10}) = 8 \times 10^{6-10}$
$= 8 \times 10^{-4} = 0.0008$

EJEMPLO 8. $\dfrac{(0.006)(80{,}000)}{0.04} = \dfrac{(6 \times 10^{-3})(8 \times 10^4)}{4 \times 10^{-2}} = \dfrac{48 \times 10^1}{4 \times 10^{-2}} = \left(\dfrac{48}{4}\right) \times 10^{1-(-2)}$

$= 12 \times 10^3 = 12{,}000$

DIGITOS SIGNIFICATIVOS

Si una altura se anota con la mejor precisión posible como 65.4 in, eso significa que está entre 65.35 y 65.45. Los dígitos empleados, aparte de los ceros necesarios para localizar el punto decimal, se llaman *dígitos significativos* o *cifras significativas*, del número.

EJEMPLO 9. 65.4 tiene tres cifras significativas.

EJEMPLO 10. 4.5300 tiene cinco cifras significativas.

EJEMPLO 11. $.0018 = 0.0018 = 1.8 \times 10^{-3}$ tiene dos cifras significativas.

EJEMPLO 12. $.001800 = 0.001800 = 1.800 \times 10^{-3}$ tiene cuatro cifras significativas.

Los números asociados a enumeraciones, por contraposición a los obtenidos por mediciones, son exactos y tienen una cantidad ilimitada de cifras significativas. No obstante, en algunos de estos casos puede resultar difícil decidir qué cifras son significativas sin información adicional. Así, el número 186,000,000 puede tener 3, 4, ..., 9 cifras significativas. Si se sabe que tiene cinco, es mejor escribirlo como 186.00 millones o bien 1.8600×10^8.

CALCULOS

Al efectuar cálculos que impliquen productos, divisiones y raíces de números, el resultado final no puede tener más dígitos significativos que el ingrediente con menor cantidad de ellos (véase Problema 1.9).

EJEMPLO 13. $73.24 \times 4.52 = (73.24)(4.52) = 331$.

EJEMPLO 14. $1.648/0.023 = 72$.

EJEMPLO 15. $\sqrt{38.7} = 6.22$.

EJEMPLO 16. $(8.416)(50) = 420.8$ (si 50 es exacto).

Al hacer sumas y restas, el resultado final no puede tener más cifras significativas tras el punto decimal que el ingrediente con menor cantidad de ellas (véase Prob. 1.10).

EJEMPLO 17. $3.16 + 2.7 = 5.9$.

EJEMPLO 18. $83.42 - 72 = 11$.

EJEMPLO 19. $47.816 - 25 = 22.816$ (si 25 es exacto).

La regla precedente admite generalización (véase Prob. 1.11).

FUNCIONES

Si a cada valor posible de una variable X le corresponden uno o más valores de otra variable Y, decimos que Y es *función* de X y escribimos $Y = F(X)$ (léase «Y igual a F de X») para indicar esa dependencia funcional. Cabe utilizar en vez de F otras letras (G, ϕ, etc.).

La variable X se llama la *variable independiente* e Y la *variable dependiente*.

Si a cada valor de X le corresponde un solo valor de Y, se dice que Y es *función univaluada* de X; en caso contrario, se dice *multivaluada*.

EJEMPLO 20. La población total P de EE.UU. es función del tiempo t, y escribimos $P = F(t)$.

EJEMPLO 21. La longitud L de un muelle vertical es función del peso P que soporta. En símbolos, $L = G(P)$.

La dependencia funcional (o correspondencia) entre variables se anota a veces en una tabla. Sin embargo, puede también indicarse con una ecuación que conecta ambas variables, tal como $Y = 2X - 3$, de la que Y se determina a partir de X.

Si $Y = F(X)$, se suele denotar por $F(3)$ el «valor de Y cuando $X = 3$», por $F(10)$ el «valor de Y cuando $X = 10$», etc. Así que si $Y = F(X) = X^2$, entonces $F(3) = 3^2 = 9$ es el valor de Y para $X = 3$.

El concepto de función admite extensión a varias variables (véase Prob. 1.17).

COORDENADAS RECTANGULARES

Consideremos dos rectas perpendiculares $X'OX$ e $Y'OY$, llamadas *ejes* X e Y, respectivamente (véase Fig. 1.1), sobre las que se indican escalas apropiadas. Estas rectas dividen el plano que determinan, llamado *plano* XY, en cuatro regiones denotadas por I, II, III y IV, que llamaremos primero, segundo, tercero y cuarto *cuadrantes*, respectivamente.

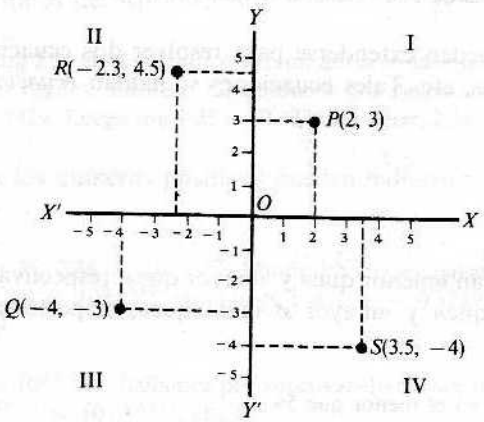


Figura 1.1.

El punto O se llama *origen* o *punto cero*. Dado un punto P, tracemos perpendiculares a los ejes X e Y desde P. Los valores de X, Y en los puntos donde tales perpendiculares cortan a los ejes se conocen como las *coordenadas rectangulares*, o simplemente *coordenadas* de P y se denotan (X, Y). La coordenada X se llama *abscisa*, y la Y *ordenada*, del punto. En la Figura 1.1 la abscisa del punto P es 2 y la ordenada es 3, de modo que las coordenadas de P son (2, 3).

Recíprocamente, dadas las coordenadas de un punto, podemos localizar (marcar) el punto. Así, los puntos con coordenadas $(-4, -3)$, $(-2.3, 4.5)$ y $(3.5, -4)$ están representados en la Figura 1.1 por Q, R y S, respectivamente.

Construyendo un eje Z que pase por O y sea perpendicular al plano XY, podemos extender fácilmente las ideas anteriores. En tal caso, las coordenadas de un punto P se denotan (X, Y, Z).

GRAFICOS

Un *gráfico* es una representación de la relación entre variables. Muchos tipos de gráficos aparecen en Estadística, según la naturaleza de los datos involucrados y el propósito del gráfico. Entre ellos citemos los *gráficos de barras*, *circulares*, etc. Estos gráficos se refieren a veces como diagramas. Hablaremos, por tanto, de *diagramas* de barras, circulares, etc. (véanse Probs. 1.23, 1.24, 1.26 y 1.27).

ECUACIONES

Las ecuaciones son enunciados del tipo $A = B$, donde A se llama *miembro* (o *lado*) *izquierdo*, y B *miembro derecho*, de la ecuación. Siempre que se efectúe sobre ambos miembros de una ecuación

una misma operación, se obtendrán ecuaciones equivalentes. Por tanto, se puede sumar, restar, multiplicar o dividir ambos lados de una ecuación por el mismo número y se llegará a una ecuación equivalente, con la única excepción de *la división por cero*, que *no está permitida.*

EJEMPLO 22. Dada la ecuación $2X + 3 = 9$, restemos 3 de ambos lados: $2X + 3 - 3 = 9 - 3$, o sea $2X = 6$. Dividimos ambos miembros por 2: $2X/2 = 6/2$, es decir $X = 3$. Este valor de X es una *solución* de la ecuación dada, como se ve sustituyendo X por 3, obteniéndose $2(3) + 3 = 9$ ó $9 = 9$, que es una *identidad.* Este proceso de hallar soluciones de una ecuación se llama *resolver* la ecuación.

Las ideas precedentes pueden extenderse para resolver dos ecuaciones en dos incógnitas, tres ecuaciones en tres incógnitas, etc. Tales ecuaciones se llaman *ecuaciones simultáneas* (véase Problema 1.30).

DESIGUALDADES

Los símbolos $<$ y $>$ significan «menor que» y «mayor que», respectivamente. Los símbolos $\leqslant$ y $\geqslant$ significan «menor o igual que» y «mayor o igual que», respectivamente. Son los *símbolos de desigualdad.*

EJEMPLO 23. $3 < 5$ se lee «3 es menor que 5».

EJEMPLO 24. $5 > 3$ se lee «5 es mayor que 3».

EJEMPLO 25. $X < 8$ se lee «X es menor que 8».

EJEMPLO 26. $X \geqslant 10$ se lee «X es mayor o igual que 10».

EJEMPLO 27. $4 < Y \leqslant 6$ se lee «4 es menor que Y, que es menor o igual que 6», o bien «Y está entre 4 y 6, excluyendo el 4, pero incluyendo el 6», o sea, «Y es mayor que 4, y menor o igual que 6».

Las relaciones que usan símbolos de desigualdad se llaman *desigualdades.* Igual que hablamos de miembros de una ecuación, hablaremos de *miembros* (o *lados*) *de una desigualdad.* De modo que en la desigualdad $4 < Y \leqslant 6$, los miembros son 4, Y y 6.

Una desigualdad válida permanece válida si:

1. Se suma o resta el mismo número de ambos lados

EJEMPLO 28. Como $15 > 12$, $15 + 3 > 12 + 3$ (es decir, $18 > 15$) y $15 - 3 > 12 - 3$ (es decir, $12 > 9$).

2. Se multiplica o divide cada lado por un mismo número *positivo.*

EJEMPLO 29. Como $15 > 12$, $(15)(3) > (12)(3)$ (es decir, $45 > 36$) y $15/3 > 12/3$ (es decir, $5 > 4$).

3. Se multiplica o divide cada lado por un mismo número *negativo* y se invierte el símbolo de desigualdad.

EJEMPLO 30. Como $15 > 12$, $(15)(-3) < (12)(-3)$ (es decir, $-45 < -36$) y $15/(-3) < 12/(-3)$ (es decir, $-5 < -4$).

LOGARITMOS

Todo número positivo N puede expresarse como potencia de 10; es decir, podemos encontrar p tal que $N = 10^p$. Se dice que p es el *logaritmo de N en base 10*, o el *logaritmo común o decimal* de N, y se escribe en breve $p = \log N$, o bien $p = \log_{10} N$. Por ejemplo, como $1000 = 10^3$, $\log 1000 = 3$. Del mismo modo, como $0.01 = 10^{-2}$, $\log 0.01 = -2$.

Cuando N está entre 1 y 10 (o sea, 10^0 y 10^1), $p = \log N$ es un número entre 0 y 1, y se puede hallar con la tabla de logaritmos del Apéndice VII.

EJEMPLO 31. Para hallar log 2.36 en el Apéndice VII, miramos en la columna de la izquierda, encabezada por N, hasta encontrar los dos dígitos iniciales, 23. Entonces nos desplazamos a la derecha a la columna encabezada por 6. Allí leemos 3729. Luego $\log 2.36 = 0.3729$ (es decir, $2.36 = 10^{0.3729}$).

Los logaritmos de todos los números positivos pueden hallarse a partir de los de los números comprendidos entre 1 y 10.

EJEMPLO 32. Del Ejemplo 31, $2.36 = 10^{0.3729}$. Multiplicando sucesivamente por 10, tenemos $23.6 = 10^{1.3729}$, $236 = 10^{2.3729}$, $2360 = 10^{3.3729}$, etc. Luego $\log 2.36 = 0.3729$, $\log 23.6 = 1.3729$, $\log 236 = 2.3729$, y $\log 2360 = 3.3729$.

EJEMPLO 33. Como $2.36 = 10^{0.03729}$, hallamos por sucesivas divisiones por 10 que $0.236 = 10^{0.3729-1} = 10^{-0.6271}$, $0.0236 = 10^{0.3729-2} = 10^{-1,6271}$, etc.

A menudo escribimos $0.3729 - 1$ como $9.3729 - 10$, o $\bar{1}.3729$; y $0.3729 - 2$ como $8.3729 - 10$, o $\bar{2}.3729$; etcétera. Con esa notación se tiene

$$\log 0.236 = 9.3729 - 10 = \bar{1}.3729 = -0.6271$$

$$\log 0.0236 = 8.3729 - 10 = \bar{2}.3729 = -1.6271$$

etcétera.

La parte decimal .3729 en todos esos logaritmos se llama *mantisa*. El resto, que antecede al punto decimal [o sea, 1, 2, 3, y $\bar{1}$ y $\bar{2}$ (o sea $9 - 10$, $8 - 10$, respectivamente)] se llama la *característica*.

Es sencillo demostrar las siguientes reglas:

1. Para un número mayor que 1 la característica es positiva y vale una unidad *menos* que el número de dígitos que preceden al punto decimal.

EJEMPLO 34. Las características de los logaritmos de 2360, 236, 23.6 y 2.36 son 3, 2, 1 y 0, y los logaritmos son 3.3729, 2.3729, 1.3729 y 0.3729.

2. Para un número menor que 1, la característica es negativa y vale uno *más* que el número de ceros que siguen al punto decimal.

EJEMPLO 35. Las características de los logaritmos de 0.236, 0.0236 y 0.00236 son -1, -2 y -3, y los logaritmos son $\bar{1}.3729$, $\bar{2}.3729$ y $\bar{3}.3729$, o sea $9.3729 - 10$, $8.3729 - 10$ y $7.3729 - 10$, respectivamente.

Si se precisan logaritmos de números de cuatro cifras (como 2.364 y 758.2) debe usarse *interpolación* (véase Prob. 1.36).

ANTILOGARITMOS

En la forma exponencial $2.36 = 10^{0.3729}$, el número 2.36 se llama el *antilogaritmo* de 0.3729, o sea antilog 0.3729. Es el número cuyo logaritmo es 0.3729. Se sigue que antilog 1.3729 = 23.6, antilog 2.3729 = 236, antilog 3.3729 = 2360, antilog 9.3729 − 10 = antilog $\bar{1}$.3729 = 0.236 y antilog 8.3729 − 10 = antilog $\bar{2}$.3729 = 0.0236. El antilogaritmo de cualquier número se puede hallar con el Apéndice VII.

EJEMPLO 36. Para hallar antilog 8.6284 − 10, miramos la mantisa .6284 dentro de la tabla. Como aparece en la fila del 42 y en la columna encabezada con 5, los dígitos requeridos son 425. Y ya que la característica es 8 − 10, el número es 0.0425.

Análogamente, antilog 3.6284 = 4250 y antilog 5.6284 = 425,000.

Si no se encuentra la mantisa en el Apéndice VII, úsese interpolación (véase Prob. 1.37).

CALCULOS USANDO LOGARITMOS

Estos cálculos recurren a las siguientes propiedades:

$$\log MN = \log M + \log N$$

$$\log \frac{M}{N} = \log M - \log N$$

$$\log M^p = p \log M$$

Combinando esos resultados obtenemos, por ejemplo,

$$\log \frac{A^p B^q C^r}{D^s E^t} = p \log A + q \log B + r \log C - s \log D - t \log E$$

Véanse Problemas 1.38 al 1.45.

PROBLEMAS RESUELTOS

VARIABLES

1.1. Decir cuáles de estos datos son discretos y cuáles continuos:

(*a*) Número de acciones vendidas un día en la Bolsa de Valores.
(*b*) Temperaturas medidas en un observatorio cada media hora.
(*c*) Vida media de los tubos de televisión producidos por una fábrica.
(*d*) Ingresos anuales de los profesores de Enseñanza Media.
(*e*) Longitudes de 1000 tornillos producidos en una empresa.

Solución

(*a*) Discretos; (*b*) continuos; (*c*) continuos; (*d*) discretos; (*e*) continuos.

1.2. Dar el dominio de las siguientes variables y decir cuáles son continuas:

(*a*) Número G de galones (gal) de agua en una lavadora.
(*b*) Número B de libros en una estantería.
(*c*) Suma S de los puntos obtenidos al lanzar un par de dados.
(*d*) Diámetro D de una esfera.
(*e*) País P de Europa.

Solución

(*a*) *Dominio:* Cualquier valor entre 0 gal y la capacidad de la lavadora. *Variable:* Continua.
(*b*) *Dominio:* 0, 1, 2, 3, ... hasta el número total de libros que caben en la estantería. *Variable:* Discreta.
(*c*) *Dominio:* Los puntos de un dado pueden ser 1, 2, 3, 4, 5 ó 6. Luego la suma de dos dados puede ser 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 ó 12, que es el dominio de S. *Variable:* Discreta.
(*d*) *Dominio:* Todos los valores positivos. *Variable:* Continua.
(*e*) *Dominio:* Francia, Italia, ..., etc., que pueden representarse numéricamente como 1, 2, ... *Variable:* Discreta.

REDONDEO DE DATOS

1.3. Redondear cada número con la precisión establecida:

(*a*)	48.6	unidades	(*f*)	143.95	decenas
(*b*)	136.5	unidades	(*g*)	368	centenas
(*c*)	2.484	centésimas	(*h*)	24,448	millares
(*d*)	0.0435	milésimas	(*i*)	5.56500	centésimas
(*e*)	4.50001	unidades	(*j*)	5.56501	centésimas

Solución

(*a*) 49; (*b*) 136; (*c*) 2.48; (*d*) 0.044; (*e*) 5; (*f*) 144.0; (*g*) 400; (*h*) 24,000; (*i*) 5.56; (*j*) 5.57.

1.4. Sumar los números 4.35, 8.65, 2.95, 12.45, 6.65, 7.55 y 9.75 (*a*) directamente, (*b*) redondeando en décimas con la regla del «entero par», y (*c*) idem eligiendo el entero dudoso más alto anterior al 5.

Solución

(*a*)		(*b*)		(*c*)	
	4.35		4.4		4.4
	8.65		8.6		8.7
	2.95		3.0		3.0
	12.45		12.4		12.5
	6.65		6.6		6.7
	7.55		7.6		7.6
	9.75		9.8		9.8
Total	52.35	Total	52.4	Total	52.7

Nótese que el método (*b*) es mejor que el (*c*) por cuanto minimiza la *acumulación de errores de redondeo.*

NOTACION CIENTIFICA Y DIGITOS SIGNIFICATIVOS

1.5. Expresar los siguientes números sin usar potencias de 10:

(*a*) 4.823×10^{7} (*c*) 3.80×10^{-4} (*e*) 300×10^{8}

(*b*) 8.4×10^{-6} (*d*) 1.86×10^{5} (*f*) $70{,}000 \times 10^{-10}$

Solución

(*a*) Movemos el punto decimal siete lugares a la derecha y obtenemos 48,230,000; (*b*) moviendo ahora seis posiciones a la izquierda queda 0.0000084; (*c*) 0.000380; (*d*) 186,000; (*e*) 30,000,000,000; (*f*) 0.0000070000.

1.6. ¿Cuántas cifras significativas hay en cada uno de estos números, supuesto que han sido redondeados correctamente?

(*a*) 149.8 in (*d*) 0.00280 m (*g*) 9 casas

(*b*) 149.80 in (*e*) 1.00280 m (*h*) 4.0×10^{3} libras (lb)

(*c*) 0.0028 metros (m) (*f*) 9 gramos (g) (*i*) 7.58400×10^{-5} dinas

Solución

(*a*) cuatro; (*b*) cinco; (*c*) dos; (*d*) tres; (*e*) seis; (*f*) uno; (*g*) sin límite; (*h*) dos; (*i*) seis.

1.7. ¿Cuál es el máximo error en cada una de estas medidas, supuesto que se han anotado del modo más preciso posible?

(*a*) 73.854 in (*b*) 0.09800 pies cúbicos (ft^3) (*c*) 3.867×10^{8} kilómetros (km)

Solución

(*a*) La medida debe estar entre 73.8535 a 73.8545 in; luego el máximo error es 0.0005 in. Hay 5 cifras significativas.

(*b*) El número de pies cúbicos está entre 0.097995 a 0.098005 pies cúbicos; luego el error máximo es 0.000005 pies cúbicos. Cuatro cifras significativas.

(*c*) El número real de kilómetros es mayor que 3.8665×10^{8} pero menor que 3.8675×10^{8}; por tanto, el máximo error posible es 0.0005×10^{8}, o sea 50,000 km. Cuatro cifras significativas.

1.8. Escribir cada número en notación científica. Salvo mención expresa en contra, se suponen todas las cifras significativas.

(*a*) 24,380,000 (cuatro cifras significativas) (*c*) 7,300,000,000 (cinco cifras significativas)

(*b*) 0.000009851 (*d*) 0.00018400

Solución

(*a*) 2.438×10^{7}; (*b*) 9.851×10^{-6}; (*c*) 7.3000×10^{9}; (*d*) 1.8400×10^{-4}.

CALCULOS

1.9. Probar que el producto de 5.74 y 3.8, supuesto que tienen tres y dos cifras significativas, no puede lograrse con más de dos cifras significativas.

Solución

Primer método

$5.74 \times 3.8 = 21.812$, pero no todas las cifras de este producto son significativas. Para ver cuántas lo son, nótese que 5.74 puede ser cualquier número entre 5.735 y 5.745, mientras que 3.8 es cualquiera entre 3.75 y 3.85. Luego el menor valor posible del producto es $5.735 \times 3.75 = 21.50625$, y el mayor $5.745 \times 3.85 = 21.11825$.

Como el posible rango de valores es 21.50625 a 22.11825, es claro que sólo las dos primeras cifras del producto son cifras significativas, pudiendo escribir el resultado como 22. Observemos que 22 debe interpretarse como cualquier número entre 21.5 y 22.5.

Segundo método

Con las cifras dudosas en cursiva, el producto es:

$$\begin{array}{r} 5.7\,\mathit{4} \\ 3.\mathit{8} \\ \hline \mathit{4\,5\,9\,2} \\ 1\,7\,\mathit{2\,2} \\ \hline \mathit{2\,1.8\,1\,2} \end{array}$$

No debemos conservar más de una cifra dudosa en el producto, que es en consecuencia 22 con dos cifras significativas. Es, por tanto, innecesario arrastrar más cifras significativas de las que figuren en el factor menos preciso; así, si 5.74 se redondea a 5.7, el producto es $5.7 \times 3.8 = 21.66 = 22$ con dos cifras significativas, de acuerdo con el resultado ya sabido.

Al calcular a mano, se ahorra trabajo no guardando más que una o dos cifras más allá de las que tenga el factor menos preciso, y redondeando al número adecuado de cifras significativas el resultado final. Con calculadoras que manejan muchos dígitos, debe tenerse cuidado en no creer que todas las obtenidas son cifras significativas.

1.10. Sumar 4.19355, 15.28, 5.9561, 12.3 y 8.472, suponiendo que todas son cifras significativas.

Solución

Pondremos en el cálculo (*a*) las cifras dudosas en cursiva. La respuesta final con sólo una cifra dudosa se presenta como 46.2.

$$\begin{array}{rr} (a) \quad 4.19\mathit{355} & (b) \quad 4.19 \\ \quad 15.\mathit{28} & \quad 15.28 \\ \quad 5.956\mathit{1} & \quad 5.96 \\ \quad 12.\mathit{3} & \quad 12.3 \\ \quad 8.\mathit{472} & \quad 8.47 \\ \hline \quad 46.\mathit{20165} & \quad 46.20 \end{array}$$

Se ahorra esfuerzo guardando, como en (*b*), un decimal significativo más que en el número preciso. La respuesta final, redondeada a 46.2, coincide con el cálculo (*a*).

1.11. Calcular 475,000,000 + 12,684,000 − 1,372,410 si esos números tienen 3,5 y 7 cifras significativas, respectivamente.

Solución

En (*a*) conservaremos todas las cifras y redondearemos el resultado final. En (*b*), usamos un método análogo al del Problema 1.10(*b*). En ambos casos, las cifras dudosas están en cursiva.

$$(a)\quad \begin{array}{r} 475{,}000{,}000 \\ +\ 12{,}684{,}000 \\ \hline 487{,}684{,}000 \end{array} \qquad \begin{array}{r} 487{,}684{,}000 \\ -\ 1{,}372{,}410 \\ \hline 486{,}311{,}590 \end{array} \qquad (b)\quad \begin{array}{r} 475{,}000{,}000 \\ +\ 12{,}700{,}000 \\ \hline 487{,}700{,}000 \end{array} \qquad \begin{array}{r} 487{,}700{,}00 \\ -\ 1{,}400{,}000 \\ \hline 486{,}300{,}000 \end{array}$$

El resultado final se redondea a 486,000,000; o mejor, para mostrar que hay 3 cifras significativas, escribirlo como 486 millones o 4.86×10^8.

1.12. Efectuar cada operación indicada.

(*a*) 48.0×943

(*b*) $8.35/98$

(*c*) $(28)(4193)(182)$

(*d*) $\dfrac{(526.7)(0.001280)}{0.000034921}$

(*e*) $\dfrac{(1.47562 - 1.47322)(4895.36)}{0.000159180}$

(*f*) Si los denominadores 5 y 6 son exactos, $\dfrac{(4.38)^2}{5} + \dfrac{(5.482)^2}{6}$

(*g*) $3.1416\sqrt{71.35}$

(*h*) $\sqrt{128.5 - 89.24}$

Solución

(*a*) $48.0 \times 943 = (48.0)(943) = 45{,}300$

(*b*) $8.35/98 = 0.085$

(*c*) $(28)(4193)(182) = (2.8 \times 10^1)(4.193 \times 10^3)(1.82 \times 10^2)$

$= (2.8)(4.193)(1.82) \times 10^{1+3+2} = 21 \times 10^6 = 2.1 \times 10^7$

Esto puede escribirse también como 21 millones para mostrar las dos cifras significativas.

(*d*) $$\frac{(526.7)(0.001280)}{0.000034921} = \frac{(5.267 \times 10^2)(1.280 \times 10^{-3})}{3.4921 \times 10^{-5}} = \frac{(5.267)(1.280)}{3.4921} \times \frac{(10^2)(10^{-3})}{10^{-5}}$$

$$= 1.931 \times \frac{10^{2-3}}{10^{-5}} = 1.931 \times \frac{10^{-1}}{10^{-5}}$$

$$= 1.931 \times 10^{-1+5} = 1.931 \times 10^4$$

Que cabe presentar como 19.31 miles mostrando las cuatro cifras significativas.

(*e*) $$\frac{(1.47562 - 1.47322)(4895.36)}{0.000159180} = \frac{(0.00240)(4895.36)}{0.000159180} = \frac{(2.40 \times 10^{-3})(4.89536 \times 10^3)}{1.59180 \times 10^{-4}}$$

$$= \frac{(2.40)(4.89536)}{1.59180} \times \frac{(10^{-3})(10^3)}{10^{-4}} = 7.38 \times \frac{10^0}{10^{-4}} = 7.38 \times 10^4$$

Esto puede expresarse como 73.8 miles, mostrando sus tres cifras significativas. Nótese que aunque había seis cifras significativas en cada número inicial, algunas se han perdido al restar 1.47322 de 1.47562.

(f) Si los denominadores 5 y 6 son exactos, $\dfrac{(4.38)^2}{5} + \dfrac{(5.482)^2}{6} = 3.84 + 5.009 = 8.85$

(g) $3.1416\sqrt{71.35} = (3.1416)(8.447) = 26.54$

(h) $\sqrt{128.5 - 89.24} = \sqrt{39.3} = 6.27$

1.13. Evaluar lo que sigue, dado que $X = 3$, $Y = -5$, $A = 4$ y $B = -7$, donde todos los números son exactos:

(a) $2X - 3Y$

(b) $4Y - 8X + 28$

(c) $\dfrac{AX + BY}{BX - AY}$

(d) $X^2 - 3XY - 2Y^2$

(e) $2(X + 3Y) - 4(3X - 2Y)$

(f) $\dfrac{X^2 - Y^2}{A^2 - B^2 + 1}$

(g) $\sqrt{2X^2 - Y^2 - 3A^2 + 4B^2 + 3}$

(h) $\sqrt{\dfrac{6A^2}{X} + \dfrac{2B^2}{Y}}$

Solución

(a) $2X - 3Y = 2(3) - 3(-5) = 6 + 15 = 21$

(b) $4Y - 8X + 28 = 4(-5) - 8(3) + 28 = -20 - 24 + 28 = -16$

(c) $\dfrac{AX + BY}{BX - AY} = \dfrac{(4)(3) + (-7)(-5)}{(-7)(3) - (4)(-5)} = \dfrac{12 + 35}{-21 + 20} = \dfrac{47}{-1} = -47$

(d) $X^2 - 3XY - 2Y^2 = (3)^2 - 3(3)(-5) - 2(-5)^2 = 9 + 45 - 50 = 4$

(e) $2(X + 3Y) - 4(3X - 2Y) = 2[(3) + 3(-5)] - 4[3(3) - 2(-5)]$
$= 2(3 - 15) - 4(9 + 10) = 2(-12) - 4(19) = -24 - 76 = -100$

Otro método

$2(X + 3Y) - 4(3X - 2Y) = 2X + 6Y - 12X + 8Y = -10X + 14Y = -10(3) + 14(-5)$
$= -30 - 70 = -100$

(f) $\dfrac{X^2 - Y^2}{A^2 - B^2 + 1} = \dfrac{(3)^2 - (-5)^2}{(4)^2 - (-7)^2 + 1} = \dfrac{9 - 25}{16 - 49 + 1} = \dfrac{-16}{-32} = \dfrac{1}{2} = 0.5$

(g) $\sqrt{2X^2 - Y^2 - 3A^2 + 4B^2 + 3} = \sqrt{2(3)^2 - (-5)^2 - 3(4)^2 + 4(-7)^2 + 3}$
$= \sqrt{18 - 25 - 48 + 196 + 3} = \sqrt{144} = 12$

(h) $\sqrt{\dfrac{6A^2}{X} + \dfrac{2B^2}{Y}} = \sqrt{\dfrac{6(4)^2}{3} + \dfrac{2(-7)^2}{-5}} = \sqrt{\dfrac{96}{3} + \dfrac{98}{-5}} = \sqrt{12.4} = 3.52$ aproximadamente

FUNCIONES

1.14. La Tabla 1.1 muestra el número de bushels (bu) de trigo y maíz producidos en la cooperativa PQR durante los años 1975-1985. Con referencia a esa tabla, determinar el año o años durante los cuales: (a) la producción de trigo fue mínima, (b) la de maíz fue máxima, (c) se dio el mayor descenso en la producción de trigo, (d) decreció la producción de maíz respecto del año anterior y creció la de trigo, (e) se produjo idéntica cantidad de trigo y (f) la producción conjunta de trigo y maíz fue máxima.

Tabla 1.1

Año	Número de bushels de trigo	Número de bushels de maíz
1975	200	75
1976	185	90
1977	225	100
1978	250	85
1979	240	80
1980	195	100
1981	210	110
1982	225	105
1983	250	95
1984	230	110
1985	235	100

Solución

(*a*) 1976; (*b*) 1981 y 1984; (*c*) 1980; (*d*) 1978, 1982, 1983 y 1985; (*e*) 1977 y 1982, y 1978 y 1983; (*f*) 1983.

1.15. Sean W y C, respectivamente el número de bushels de trigo y maíz producidos en el año t en la cooperativa PQR del Problema 1.14. Es claro que W y C son ambas funciones de t, lo que podemos indicar como $W = F(t)$ y $C = G(t)$.

(*a*) Hallar W cuando $t = 1981$.
(*b*) Hallar C cuando $t = 1978$ y 1984.
(*c*) Hallar t cuando $W = 225$.
(*d*) Hallar F (1979).
(*e*) Hallar G (1983).
(*f*) Hallar C cuando $W = 210$.
(*g*) ¿Cuál es el dominio de la variable t?
(*h*) ¿Es W función univaluada de t?
(*i*) ¿Es t función de W? Si lo es, ¿es univaluada?
(*j*) ¿Es C función de W?
(*k*) ¿Qué variable es independiente, t o W?

Solución

(*a*) 210; (*b*) 85 y 110, respectivamente; (*c*) 1977 y 1982; (*d*) 240; (*e*) 95; (*f*) 110; (*g*) los años 1975, 1976, ..., 1985.

(*h*) Sí, pues a cada valor de t en su dominio le corresponde uno y sólo un valor de W.

(*i*) Sí, porque a cada valor de W podemos suponer que le corresponden uno o más valores de t, que pueden hallarse con la Tabla 1.1. Como puede haber más de un valor de t para cada valor de W (así ocurre con $W = 225$ y $t = 1977$ ó 1982), la función es multivaluada. Esta dependencia funcional de t en W se puede expresar como $t = H(W)$.

(*j*) Sí, pues a cada valor que puede tomar W le corresponden uno o más valores de C, como enseña la Tabla 1.1. Análogamente, W es función de C.

(*k*) Físicamente, suele pensarse en W como determinado por t, y no al revés. Así pues, físicamente t es la variable independiente y W la dependiente. Matemáticamente, sin embargo, cualquiera de las variables puede verse como independiente y la otra como dependiente. A la que se asignan diversos valores es la independiente; la que viene determinada como resultado es la dependiente.

1.16. Una variable Y queda determinada por la variable X mediante la ecuación $Y = 2X - 3$, donde 2 y 3 son exactos.

(*a*) Hallar Y cuando $X = 3, -2$ y 1.5.
(*b*) Poner en una tabla los valores de Y para $X = -2, -1, 0, 1, 2, 3$ y 4.
(*c*) Si denotamos la dependencia de Y en X por $Y = F(X)$, determinar $F(2.4)$ y $F(0.8)$.
(*d*) ¿Qué valor de X corresponde a $Y = 15$?
(*e*) ¿Puede expresarse X como función de Y?
(*f*) ¿Es Y función univaluada de X?
(*g*) ¿Es X función univaluada de Y?

Solución

(*a*) Cuando $X = 3$, $Y = 2X - 3 = 2(3) - 3 = 6 - 3 = 3$. Cuando $X = -2$, $Y = 2X - 3 = 2(-2) - 3 = -4 - 3 = -7$. Cuando $X = 1.5$, $Y = 2X - 3 = 2(1.5) - 3 = 3 - 3 = 0$.
(*b*) Los valores de Y, calculados como en (*a*), se indican en la Tabla 1.2. Nótese que pueden construirse otras tablas escogiendo otros valores de X. La relación $Y = 2X - 3$ es equivalente a la colección de todas las posibles tablas.

Tabla 1.2

X	-2	-1	0	1	2	3	4
Y	-7	-5	-3	-1	1	3	5

(*c*) $F(2.4) = 2(2.4) - 3 = 4.8 - 3 = 1.8$ y $F(0.8) = 2(0.8) - 3 = 1.6 - 3 = -1.4$.
(*d*) Sustituir $Y = 15$ en $Y = 2X - 3$. Se obtiene $15 = 2X - 3$, $2X = 18$ y $X = 9$.
(*e*) Sí. Como $Y = 2X - 3$, $Y + 3 = 2X$ y $X = \frac{1}{2}(Y + 3)$. Esto expresa X *explícitamente* como función de Y.
(*f*) Sí, porque para cada valor posible de X (hay infinitos) le corresponde un solo de Y.
(*g*) Sí, porque de la parte (*e*), $X = \frac{1}{2}(Y + 3)$, de modo que correspondiente a cada valor de Y hay uno y uno sólo de X.

1.17. Si $Z = 16 + 4X - 3Y$, hallar el valor de Z correspondiente a: (*a*) $X = 2$, $Y = 5$; (*b*) $X = -3$, $Y = -7$; (*c*) $X = -4$, $Y = 2$.

Solución

(*a*) $Z = 16 + 4(2) - 3(5) = 16 + 8 - 15 = 9$.
(*b*) $Z = 16 + 4(-3) - 3(-7) = 16 - 12 + 21 = 25$.
(*c*) $Z = 16 + 4(-4) - 3(2) = 16 - 16 - 6 = -6$.

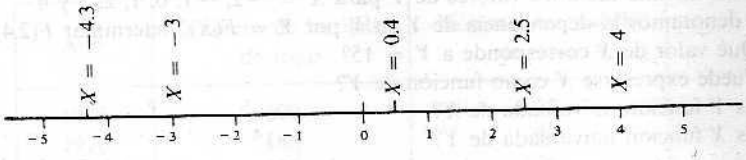
Dados valores de X e Y, les corresponde uno de Z. Podemos denotar esta dependencia de Z en X e Y como $Z = F(X, Y)$ (se lee «Z es función de X e Y»). $F(2.5)$ denota el valor de Z cuando $X = 2$ e $Y = 5$, que es 9; véase (*a*). De la misma manera, $F(-3, -7) = 25$ y $F(-4, 2) = -6$ por las partes (*b*) y (*c*), respectivamente.

Las variables X, Y se llaman *variables independientes*, y Z la *variable dependiente*.

GRAFICOS

1.18. Localizar en el eje X de un sistema coordenado los puntos correspondientes a: (*a*) $X = 4$, (*b*) $X = -3$, (*c*) $X = 2.5$, (*d*) $X = -4.3$ y (*e*) $X = 0.4$, suponiendo que esos valores son exactos.

Solución



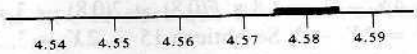
Cada valor exacto de X corresponde a un punto y sólo uno sobre el eje X. Recíprocamente, se demuestra en matemáticas más avanzadas que a cada punto del eje le corresponde un valor de X y sólo uno.

Así pues, teóricamente existe un punto asociado a $X = 22/7 = 3.142857142857...$, o al $X = \pi = 3.14159265358...$ En la práctica, naturalmente, no es factible su localización exacta, porque el lápiz hace una marca de cierta anchura y cubre una infinidad de puntos. El propio eje X tiene grosor. De modo que el diagrama adjunto es una representación física de la situación matemática.

1.19. Sea X el diámetro en centímetros (cm) de una bola. Si $X = 4.58$ con tres cifras significativas, ¿cómo debe representarse en el eje X?

Solución

La verdadera medida está entre 4.575 y 4.585 cm, luego hay que representarla por el segmento grueso de la figura adjunta.



1.20. Localizar en un sistema de coordenadas rectangulares los puntos de coordenadas: (*a*) (5, 2), (*b*) (2, 5), (*c*) (−3, 1), (*d*) (1, −3), (*e*) (3, −4), (*f*) (−2.5, −4.8), (*g*) (0, −2.5) y (*h*) (4, 0). Suponemos exactos todos esos números.

Solución

Véase Figura 1.2.

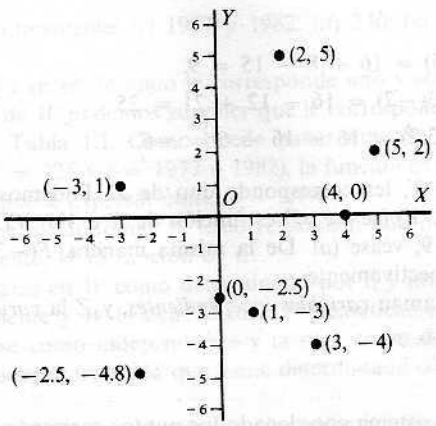


Figura 1.2.

1.21. Representar la ecuación $Y = 2X - 3$.

Solución

Tomando $X = -2, -1, 0, 1, 2, 3$ y 4, obtenemos que $Y = -7, -5, -3, -1, 1, 3$ y 5, respectivamente [véase Prob. 1.16(*b*)]. Luego los puntos vienen dados en el gráfico por $(-2, -7)$, $(-1, -5)$, $(0, -3)$, $(1, -1)$, $(2, 1)$, $(3, 3)$ y $(4, 5)$, que pueden verse representados en coordenadas rectangulares en la Figura 1.3. Todos ellos, así como los obtenidos a partir de otros valores de X, yacen en una recta que es la gráfica pedida.

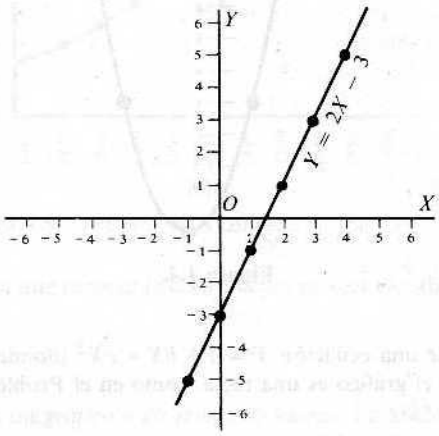


Figura 1.3.

Como la gráfica de $Y = 2X - 3$ es una línea recta, se dice que $F(X) = 2X - 3$ es una *función lineal*. En general, $F(X) = aX + b$ (con a, b constantes) es una función lineal cuya gráfica es una recta.

Nótese que sólo se necesitan dos puntos para hallar la gráfica de una función lineal, pues dos puntos determinan una recta.

1.22. Representar la ecuación $Y = X^2 - 2X - 8$.

Solución

La Tabla 1.3 muestra los valores de Y correspondientes a algunos valores de X; por ejemplo, cuando $X = -2$, $Y = (-2)^2 - 2(-2) - 8 = 4 + 4 - 8 = 0$. De esa tabla vemos que están sobre la gráfica los puntos $(-3, 7)$, $(-2, 0)$, $(-1, -5)$, $(0, -8)$, $(1, -9)$, $(2, -8)$, $(3, -5)$, $(4, 0)$ y $(5, 7)$. Estos puntos, y otros calculados mediante otros valores de X, están sobre la curva de la Figura 1.4, llamada parábola. La función $F(X) = X^2 - 2X - 8$ se llama una *función cuadrática*.

Tabla 1.3

X	-3	-2	-1	0	1	2	3	4	5
Y	7	0	-5	-8	-9	-8	-5	0	7

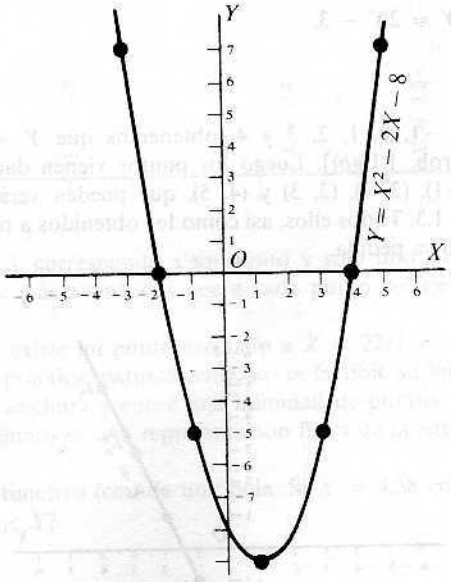


Figura 1.4.

En general, el gráfico de una ecuación $Y = a + bX + cX^2$ (donde a, b y c son constantes y $c \neq 0$) es una parábola. Si $c = 0$, el gráfico es una recta, como en el Problema 1.21.

1.23. La Tabla 1.4 muestra la población de EE.UU. (en millones) en los años 1860-1980. Representar esos datos.

Solución

Primer método

En la Figura 1.5, la población P es la variable dependiente y el tiempo t la variable independiente. Los puntos se localizan del modo habitual por las coordenadas leídas en la tabla, como (1880, 50.2). Se conectan los puntos sucesivos con trazos rectos, ya que no disponemos de información sobre P en los tiempos intermedios; de ahí que el gráfico se llame un *gráfico de trazos*.

Obsérvese que las unidades en los ejes son distintas, como al dibujar el gráfico de $Y = 2X - 3$. Ello es correcto, pues de hecho las dos variables son magnitudes completamente diferentes.

Asimismo, el cero se ha indicado en el eje vertical, pero (por razones obvias) no en el horizontal. Debe indicarse el cero siempre que sea posible, sobre todo en el eje vertical. Si no fuese posible por alguna razón, y si tal omisión pudiera provocar alguna conclusión errónea, es aconsejable advertirlo de algún modo, por ejemplo como en el Problema 1.26.

Tabla 1.4. Población de EE.UU., 1860-1980

Año	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980
Población (millones)	31.4	39.8	50.2	62.9	76.0	92.0	105.7	122.8	131.7	151.1	179.3	203.3	226.5

Fuente: U.S. Bureau of the Census.

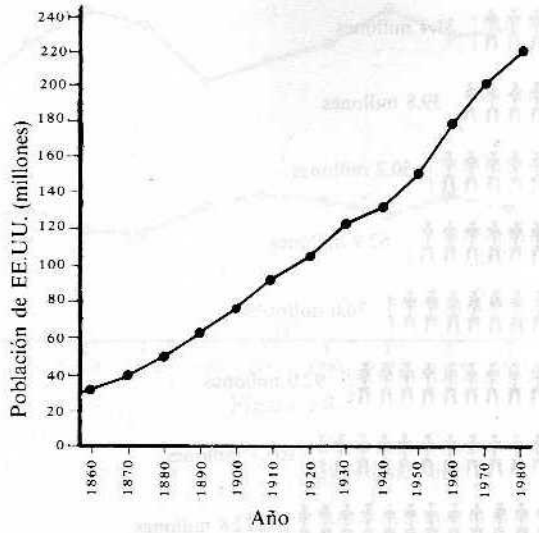


Figura 1.5. (*Fuente:* U.S. Bureau of the Census.)

Una tabla o una gráfica que recojan la distribución de una variable en función del tiempo, se llaman *series en el tiempo.*

Segundo método

La Figura 1.6 se llama un *gráfico o diagrama de barras.* La anchura de cada barra, todas idénticas, no tienen importancia en este caso y se escoge a capricho (siempre que las barras no se solapen).

Los números sobre las barras pueden omitirse. Si se mantienen, la escala vertical de la izquierda es innecesaria.

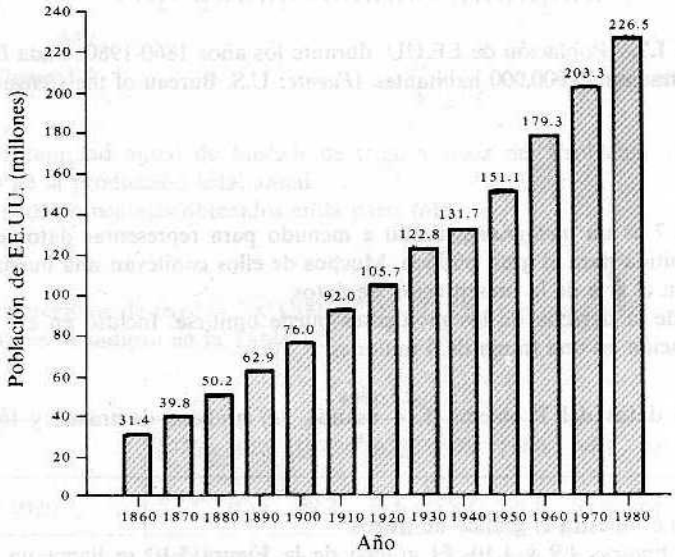


Figura 1.6. (*Fuente:* U.S. Bureau of the Census.)

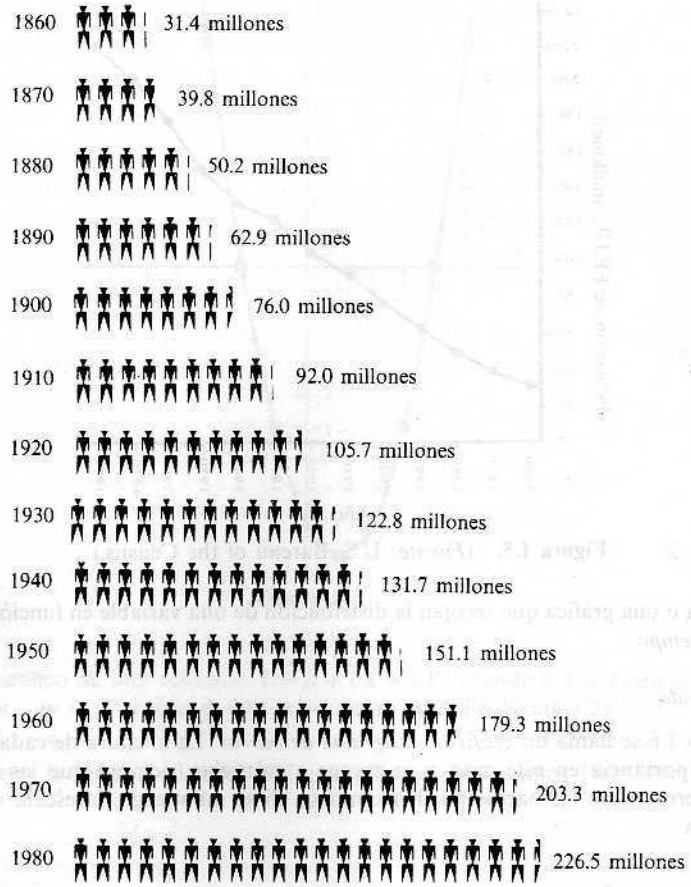


Figura 1.7. Población de EE.UU. durante los años 1860-1980. Cada figura representa 10,000,000 habitantes. (*Fuente:* U.S. Bureau of the Census.)

Tercer método

La Figura 1.7 es un *pictograma*, usado a menudo para representar datos en Estadística de una forma que sea nítida para el gran público. Muchos de ellos conllevan una buena dosis de ingenuidad y originalidad en el arte de la presentación de datos.

El número de la derecha de los monigotes puede omitirse. Incluso en ese caso, el lector podrá estimar la población en una franja de 5 millones.

1.24. Representar los datos del Problema 1.14 usando: (*a*) gráficos de trazos y (*b*) gráficos de barras.

Solución

(*a*) La Figura 1.8 muestra el gráfico de trazos.

(*b*) Véanse las Figuras 1.9 y 1.10. El gráfico de la Figura 1.10 se llama un *gráfico de barras en componentes*.

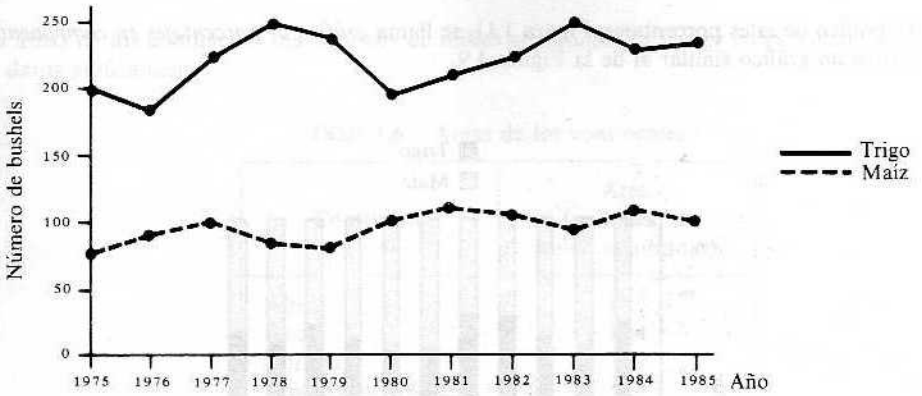


Figura 1.8.

Primer método

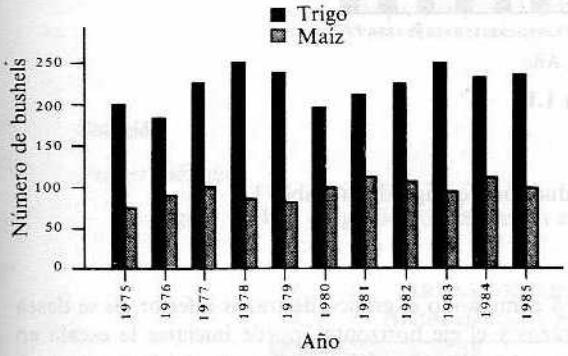


Figura 1.9.

Segundo método

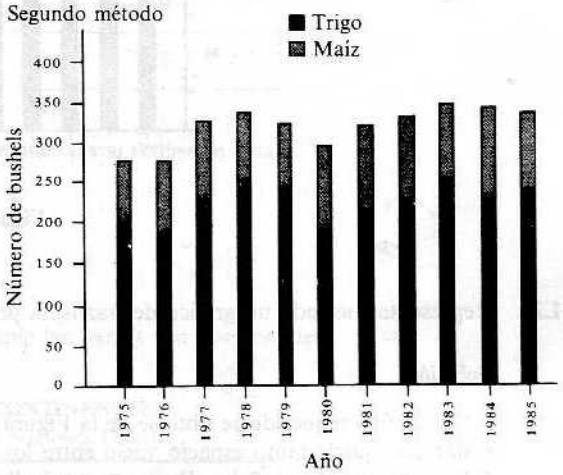


Figura 1.10.

1.25. (*a*) Expresar la cantidad anual de bushels de trigo y maíz del Problema 1.14 (Tabla 1.1) como porcentajes de la producción total anual.

(*b*) Representar los porcentajes obtenidos en la parte (*a*).

Solución

(*a*) En 1975 el porcentaje de trigo = 200/(200 + 75) = 72.7%, y el maiz 100% − 72.7% = 27.3%; etc. Los porcentajes se indican en la Tabla 1.5.

Tabla 1.5

Año	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
Porcentaje de trigo	72.7	67.3	69.2	74.6	75.0	66.1	65.6	68.2	72.5	67.6	70.1
Porcentaje de maiz	27.3	32.7	30.8	25.4	25.0	33.9	34.4	31.8	27.5	32.4	29.9

(*b*) El gráfico de tales porcentajes, Figura 1.11, se llama *gráfico de porcentajes en componentes*. Puede usarse un gráfico similar al de la Figura 1.9.

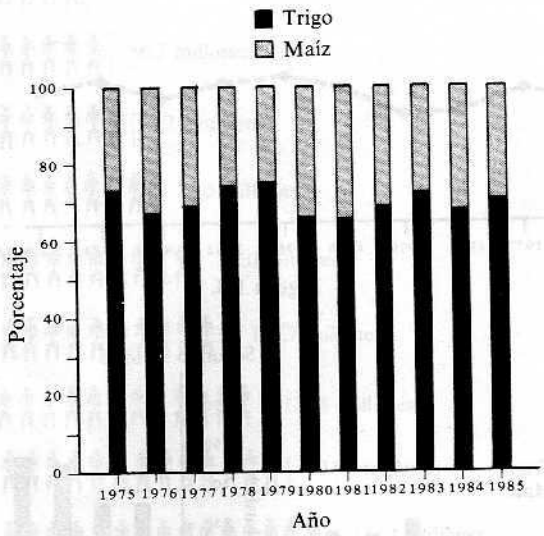


Figura 1.11.

1.26. Representar, usando un gráfico de trazos, la producción de trigo de la Tabla 1.1.

Solución

El gráfico requerido se obtiene de la Figura 1.8 eliminando el gráfico de trazos inferior. Si se desea evitar que quede tanto espacio vacío entre los trazos y el eje horizontal, puede iniciarse la escala en 150 bu en vez de en 0 bu. Pero eso puede llevar a conclusiones falsas por parte del lector que no advierta la omisión del cero. Para advertirle de ello, cabe construir el gráfico de la Figura 1.12.

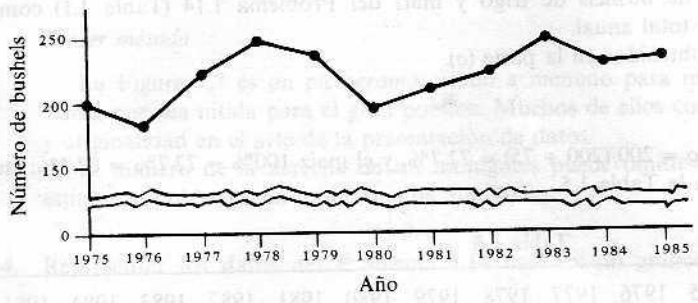


Figura 1.12.

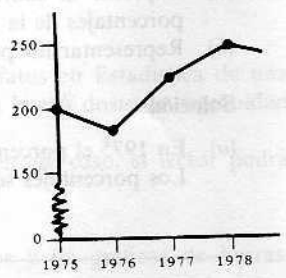


Figura 1.13.

Otro truco frecuente para llamar la atención sobre la supresión del cero es el uso de una línea en zigzag en uno de los ejes (Fig. 1.13).

1.27. Las áreas de los continentes (en millones de millas cuadradas) se recoge en la Tabla 1.6. Representar los datos gráficamente.

Tabla 1.6. Areas de los continentes

Continente	Area (millones de millas cuadradas)
Africa	11.7
Asia	10.4
Europa	1.9
América del Norte	9.4
Oceania	3.3
América del Sur	6.9
Unión Soviética	7.9
	Total 51.5

Fuente: Naciones Unidas.
Nota: Europa excluye Turquía, que se incluye en Asia.

Solución

Primer método

La Figura 1.14 es un gráfico de barras en el que las barras son horizontales.

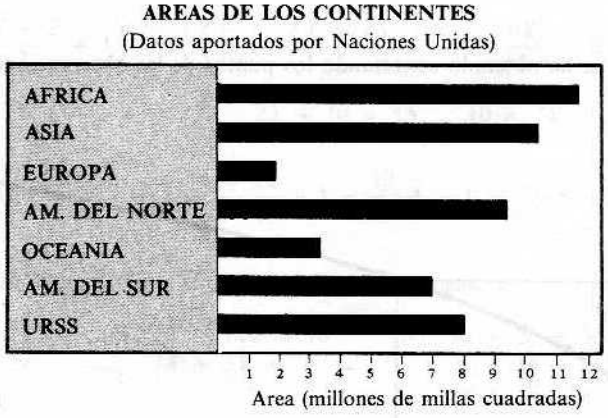


Figura 1.14.

La Figura 1.15 se llama un *diagrama circular*. Para construirlo, hacemos que el área total, 51.5 millones de millas cuadradas, corresponda a los 360° del círculo. Así, un millón corresponde a 360°/51.5. Se deduce que Africa, con 11.7 millones, ocupa un arco de 11.7(360°/51.5) = 82°, mientras Asia, Europa, Norteamérica, Oceanía, América del Sur y la URSS ocupan 73°, 13°, 66°, 23°, 48° y 55°, respectivamente.

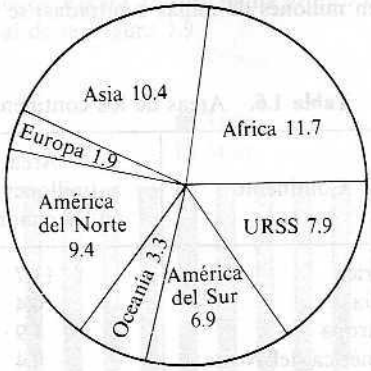


Figura 1.15. Areas de los continentes (en millones de millas cuadradas).

1.28. El tiempo T (en segundos) requerido para una oscilación completa de un péndulo simple de longitud L cm, se ve en la Tabla 1.7, que da las observaciones obtenidas en un laboratorio de Física.

(*a*) Representar gráficamente T como función de L.
(*b*) De la gráfica en (*a*), estimar T para un péndulo de 40 cm.

Tabla 1.7

L	10.1	16.2	22.2	33.8	42.0	53.4	66.7	74.5	86.6	100.0
T	0.64	0.81	0.95	1.17	1.30	1.47	1.65	1.74	1.87	2.01

Solución

(*a*) La Figura 1.16 se ha obtenido conectando los puntos de las observaciones con una curva suave.

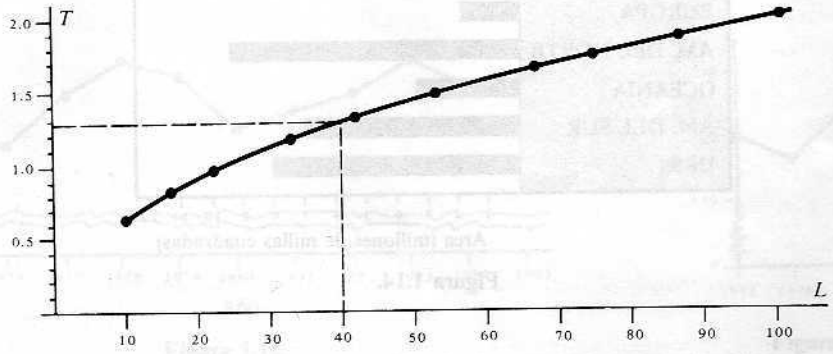


Figura 1.16.

(*b*) El valor estimado de T es 1.27 segundos.

ECUACIONES

1.29. Resolver las ecuaciones:

(*a*) $4a - 20 = 8$ (*c*) $18 - 5b = 3(b + 8) + 10$

(*b*) $3X + 4 = 24 - 2X$ (*d*) $\dfrac{Y + 2}{3} + 1 = \dfrac{Y}{2}$

Solución

(*a*) Sumar 20 a cada lado: $4a - 20 + 20 = 8 + 20$, o sea $4a = 28$.
Dividir ambos lados por 4: $4a/4 = 28/4$ y $a = 7$.
Comprobación: $4(7) - 20 = 8$, $28 - 20 = 8$ y $8 = 8$.

(*b*) Restar 4 de ambos miembros: $3X + 4 - 4 = 24 - 2X - 4$, o sea $3X = 20 - 2X$.
Sumar $2X$ a ambos lados: $3X + 2X = 20 - 2X + 2X$, o sea $5X = 20$.
Dividir por 5: $5X/5 = 20/5$ y $X = 4$.
Comprobación: $3(4) + 4 = 24 - 2(4)$, $12 + 4 = 24 - 8$ y $16 = 16$.

Puede obtenerse el resultado mucho más fácilmente dándose cuenta de que cada término puede ser trasladado de un miembro de la ecuación al otro sin más que cambiarle el signo. Así, podemos hacer

$$3X + 4 = 24 - 2X \qquad 3X + 2X = 24 - 4 \qquad 5X = 20 \qquad X = 4$$

(*c*) $18 - 5b = 3b + 24 + 10$ y $18 - 5b = 3b + 34$.
Trasponiendo, $-5b - 3b = 34 - 18$, o sea $-8b = 16$.
Dividiendo por -8, $-8b/(-8) = 16/(-8)$ y $b = -2$.
Comprobación: $18 - 5(-2) = 3(-2 + 8) + 10$, $18 + 10 = 3(6) + 10$ y $28 = 28$.

(*d*) Multiplicamos primero ambos lados por 6, el común denominador.

$$6\left(\frac{Y + 2}{3} + 1\right) = 6\left(\frac{Y}{2}\right) \qquad 6\left(\frac{Y + 2}{3}\right) + 6(1) = \frac{6Y}{2} \qquad 2(Y + 2) + 6 = 3Y$$

$$2Y + 4 + 6 = 3Y \qquad 2Y + 10 = 3Y \qquad 10 = 3Y - 2Y \qquad Y = 10$$

Comprobación: $\dfrac{10 + 2}{3} + 1 = \dfrac{10}{2}$, $\dfrac{12}{3} + 1 = \dfrac{10}{2}$, $4 + 1 = 5$ y $5 = 5$.

1.30. Resolver cada uno de los conjuntos de ecuaciones simultáneas:

(*a*) $3a - 2b = 11$, $5a + 7b = 39$ (*b*) $5X + 14Y = 78$, $7X + 3Y = -7$ (*c*) $3a + 2b + 5c = 15$, $7a - 3b + 2c = 52$, $5a + b - 4c = 2$

Solución

(*a*) Multiplicar la primera ecuación por 7: $21a - 14b = 77$ (1)
Multiplicar la segunda ecuación por 2: $10a + 14b = 78$ (2)
Sumar: $31a = 155$
Dividir por 31: $a = 5$

Nótese que al multiplicar cada ecuación por un número apropiado, somos capaces de escribir dos ecuaciones equivalentes, (1) y (2), en las que los coeficientes de la incógnita b son iguales, de modo que al sumar se elimina b y hallamos a.

Sustituimos $a = 5$ en la primera ecuación: $3(5) - 2b = 11$, $-2b = -4$ y $b = 2$. Así pues, $a = 5$ y $b = 2$.

Comprobación: $3(5) - 2(2) = 11$, $15 - 4 = 11$ y $11 = 11$; $5(5) + 7(2) = 39$, $25 + 14 = 39$ y $39 = 39$.

(*b*) Multiplicar la primera ecuación por 3: $15X + 42Y = 234$ (3)
Multiplicar la segunda ecuación por -14: $-98X - 42Y = 98$ (4)
Sumar: $-83X = 332$
Dividir por -83: $X = -4$

Sustituimos $X = -4$ en la primera ecuación: $5(-4) + 14Y = 78$, $14Y = 98$ e $Y = 7$.
Luego $X = -4$ e $Y = 7$.
Comprobación: $5(-4) + 14(7) = 78$, $-20 + 98 = 78$ y $78 = 78$; $7(-4) + 3(7) = -7$, $-28 + 21 = -7$ y $-7 = -7$.

(*c*) Multiplicar la segunda por 2: $6a + 4b + 10c = 77$
Repetir la tercera ecuación por -5: $-35a + 15b - 10c = -260$
Sumar: $-29a + 19b = -230$ (5)

Multiplicar la segunda por 2: $14a - 6b + 4c = 104$
Repetir la tercera ecuación: $5a + b - 4c = 2$
Sumar: $19a - 5b = 106$ (6)

Así hemos eliminado c y nos quedan dos ecuaciones, (5) y (6), para deducir a y b.

Multiplicar la ecuación (5) por 5: $-145a + 95b = -1150$
Multiplicar la ecuación (6) por 19: $361a - 95b = 2014$
Sumar: $216a = 864$
Dividir por 216: $a = 4$

Sustituyendo $a = 4$ en (5) o (6) vemos que $b = -6$.
Sustituyendo $a = 4$ y $b = -6$ en alguna de las ecuaciones dadas, se obtiene $c = 3$.
Así pues, $a = 4$, $b = -6$ y $c = 3$.
Comprobación: $3(4) + 2(-6) + 5(3) = 15$ y $15 = 15$; $7(4) - 3(-6) + 2(3) = 52$ y $52 = 52$; $5(4) + (-6) - 4(3) = 2$ y $2 = 2$.

DESIGUALDADES

1.31. Expresar en palabras el significado de:

(*a*) $N > 30$ (*b*) $X \leqslant 12$ (*c*) $0 < p \leqslant 1$ (*d*) $\mu - 2t < X < \mu + 2t$

Solución

(*a*) N es mayor que 30.
(*b*) X es menor o igual que 12.

(*c*) p es mayor que 0, pero menor o igual que 1.
(*d*) X es mayor que $\mu - 2t$, pero menor que $\mu + 2t$.

1.32. Traducir lo que sigue en símbolos:

(*a*) La variable X tiene valores entre 2 y 5 inclusive.
(*b*) La media aritmética X es mayor que 28.42, pero menor que 31.56.
(*c*) m es un número positivo menor o igual que 10.
(*d*) P es un número no negativo.

Solución

(*a*) $2 \leqslant X \leqslant 5$; (*b*) $28.42 < \bar{X} < 31.56$; (*c*) $0 < m \leqslant 10$; (*d*) $P \geqslant 0$.

1.33. Usando símbolos de desigualdad, poner 3.42, -0.6, -2.1, 1.45 y -3 en: (*a*) orden creciente y (*b*) orden decreciente.

Solución

(*a*) $-3 < -2.1 < -0.6 < 1.45 < 3.42$
(*b*) $3.42 > 1.45 > -0.6 > -2.1 > -3$

Nótese que al marcar los puntos en una recta, crecen de izquierda a derecha.

1.34. Escribir como desigualdades en X (o sea, despejar X):

(*a*) $2X < 6$
(*b*) $3X - 8 \geqslant 4$
(*c*) $6 - 4X < -2$
(*d*) $-3 < \dfrac{X - 5}{2} < 3$
(*e*) $-1 \leqslant \dfrac{3 - 2X}{5} \leqslant 7$

Solución

(*a*) Dividiendo ambos lados por 2 resulta $X < 3$.
(*b*) Sumando 8 a ambos lados, $3X \geqslant 12$; dividiendo ambos lados por 3, $X \geqslant 4$.
(*c*) Sumando -6 queda $-4X < -8$; dividiendo por -4, $X > 2$. Hagamos constar que, como en las ecuaciones, podemos pasar un término al otro lado sin más que cambiarle el signo. Por la parte (*b*), por ejemplo, $3X \geqslant 8 + 4$.
(*d*) Multiplicar por 2, $-6 < X - 5 < 6$; sumando 5, $-1 < X < 11$.
(*e*) Multiplicando por 5, $-5 \leqslant 3 - 2X \leqslant 35$; sumando -3, $-8 \leqslant -2X \leqslant 32$; dividiendo por -2, $4 \geqslant X \geqslant -16$, es decir $-16 \leqslant X \leqslant 4$.

LOGARITMOS Y ANTILOGARITMOS

1.35. Determinar la característica de los logaritmos comunes (base 10) de los números:

(*a*)	57	(*d*)	35.63	(*g*)	186,000	(*j*)	0.0325
(*b*)	57.4	(*e*)	982.5	(*h*)	0.71	(*k*)	0.0071
(*c*)	5.63	(*f*)	7824	(*i*)	0.7314	(*l*)	0.0003

Solución

(*a*) 1; (*b*) 1; (*c*) 0; (*d*) 1; (*e*) 2; (*f*) 3; (*g*) 5; (*h*) $9-10$; (*i*) $9-10$; (*j*) $8-10$; (*k*) $7-10$; (*l*) $6-10$.

1.36. Calcular los siguientes logaritmos:

(*a*)	log 87.2	(*f*)	log 0.382	(*k*)	log 4.638	(*p*)	log 0.2548
(*b*)	log 37,300	(*g*)	log 0.00159	(*l*)	log 6.753	(*q*)	log 0.04372
(*c*)	log 753	(*h*)	log 0.0753	(*m*)	log 183.2	(*r*)	log 0.009848
(*d*)	log 9.21	(*i*)	log 0.000827	(*n*)	log 43.15	(*s*)	log 0.0001788
(*e*)	log 54.50	(*j*)	log 0.0503	(*o*)	log 876,400		

Solución

(*a*) Mantisa = .9405, y característica = 1; de modo que log 87.2 = 1.9405; (*b*) 4.5717; (*c*) 2.8768; (*d*) 0.9643; (*e*) 1.7364; (*f*) Mantisa = .5821, y característica = 9 − 10; por tanto log 0.382 = 9.5821 − 10; (*g*) 7.2014 − 10; (*h*) 8.8768 − 10; (*i*) 6.9175 − 10; (*j*) 8.7016 − 10; (*k*) La mantisa de log 4638 está a 0.8 de camino entre la de log 4630 y la de log 4640.

$$\begin{aligned} \text{Mantisa de log } 4640 &= .6665 \\ \text{Mantisa de log } 4630 &= .6656 \\ \text{Diferencia tabular} &= .0009 \end{aligned}$$

La mantisa de log 4.638 = .6656 + (0.8)(.0009) = .6663 con cuatro dígitos; luego log 4.638 = .6663. Este proceso se llama interpolación lineal. Si se desea, la tabla de partes proporcionales del Apéndice VII permite deducir la mantisa directamente (6656 + 7).

(*l*) 0.8295 (8293 + 2); (*m*) 2.2630 (2625 + 5); (*n*) 1.6350 (6345 + 5); (*o*) 5.9427 (9425 + 2); (*p*) 9.4062 − 10 (4048 + 14); (*q*) 8.6407 − 10 (6405 + 2); (*r*) 7.9933 − 10 (9930 + 3); (*s*) 6.2524 − − 10 (2504 + 20).

1.37. Calcular los siguientes antilogaritmos:

(*a*)	antilog 1.9058	(*c*)	antilog 7.8657 − 10	(*f*)	antilog 2.6715
(*b*)	antilog 3.8531	(*d*)	antilog 9.8267 − 10		antilog 4.1853
	antilog 2.1875		antilog $\bar{2}$.3927		antilog 0.9245
	antilog 0.4997		antilog 7.7443 − 10	(*g*)	antilog $\bar{1}$.6089
	antilog 4.9360	(*e*)	antilog 9.3842 − 10		antilog 8.8907 − 10
					antilog 1.2000

Solución

(*a*) En el Apéndice VII la mantisa .9058 corresponde al número 805. Como la característica es 1, el número debe tener dos cifras delante del punto decimal; por tanto, es 80.5 (es decir, antilog 1.9058 = 80.5).

(*b*) antilog 3.8531 = 7130, antilog 2.1875 = 154, antilog 0.4997 = 3.16 y antilog 4.9360 = 86,300.

(*c*) En el Apéndice VII la mantisa .8657 corresponde al número 734. Como la característica es 7 − 10, el número tiene dos ceros tras el punto decimal. En consecuencia, el número es 0.00734 (o sea, antilog 7.8657 − 10 = 0.00734). La tabla de partes proporcionales del Apéndice VII la daría también.

(*d*) antilog 9.8267 − 10 = 0.671, antilog $\bar{2}$.3927 = 0.0247 y antilog 7.7443 − 10 = 0.00555.

(*e*) Como la mantisa no aparece en la tabla, hay que usar interpolación:

$$\begin{aligned} \text{Mantisa de log } 2430 &= .3856 \\ \text{Mantisa de log } 2420 &= .3838 \\ \text{Diferencia tabular} &= .0018 \end{aligned} \qquad \begin{aligned} \text{Mantisa dada} &= .3842 \\ \text{Mantisa inferior más próxima} &= .3838 \\ \text{Diferencia} &= .0004 \end{aligned}$$

Luego 2420 + (4/18)(2430 − 2420) = 2422 con cuatro dígitos, y el número pedido es 0.2422.

(f) antilog 2.6715 = 469.3 ($3/9 \times 10 = 3$ aproximadamente), antilog 4.1853 = 15,320 ($6/28 \times 10 = 2$ aproximadamente), y antilog 0.9245 = 8.404 ($2/5 \times 10 = 4$).

(g) antilog $\bar{1}.6089$ = 0.4064 ($4/11 \times 10 = 4$ aproximadamente), antilog 8.8907 − 10 = 0.07775 ($3/6 \times 10 = 5$) y antilog 1.2000 = 15.85 ($13/27 \times 10 = 5$ aproximadamente).

CALCULOS USANDO LOGARITMOS

Calcular cada una de las cantidades que siguen, usando logaritmos.

1.38. $P = (3.81)(43.4)$.

Solución

$\log P = \log 3.81 + \log 43.4$:

$$\begin{aligned} \log 3.81 &= 0.5809 \\ (+)\ \log 43.4 &= 1.6375 \\ \log P &= 2.2184 \end{aligned}$$

Por tanto, P = antilog 2.2184 = 165.3, o sea, 165 con tres dígitos significativos. Nótese el significado del cálculo en exponenciales:

$$(3.81)(43.4) = (10^{0.5809})(10^{1.6375}) = 10^{0.5809+1.6375} = 10^{2.2184} = 165.3$$

1.39. $P = (73.42)(0.004620)(0.5143)$.

Solución

$\log P = \log 73.42 + \log 0.004620 + \log 0.5143$:

$$\begin{aligned} \log 73.42 &= 1.8685 \\ (+)\ \log 0.004620 &= 7.6646 - 10 \\ (+)\ \log 0.5143 &= 9.7112 - 10 \\ \log P &= 19.2416 - 20 = 9.2416 - 10 \end{aligned}$$

Luego $P = 0.1744$.

1.40. $P = \dfrac{(784.6)(0.0431)}{28.23}$.

Solución

$\log P = \log 784.6 + \log 0.0431 - \log 28.23$:

$$\begin{aligned} \log 784.6 &= 2.8947 \\ (+)\ \log 0.0431 &= 8.6345 - 10 \\ &\quad 11.5292 - 10 \\ (-)\ \log 28.23 &= 1.4507 \\ \log P &= 10.0785 - 10 = 0.0785 \end{aligned}$$

Así pues, P = 1.198, o sea 1.20 con tres dígitos significativos. En términos de exponenciales:

$$\frac{(784.6)(0.0431)}{28.23} = \frac{(10^{2.8947})(10^{8.6345-10})}{10^{1.4507}} = 10^{2.8947+8.6345-10-1.4507} = 10^{0.0785} = 1.198$$

1.41. $P = (5.395)^8$

Solución

$\log P = 8 \log 5.395 = 8(0.7320) = 5.8560$ y $P = 717{,}800$, o sea 7.178×10^5.

1.42. $P = \sqrt{387.2} = (387.2)^{1/2}$.

Solución

$\log P = \frac{1}{2} \log 387.2 = \frac{1}{2}(2.5879) = 1.2940$ y $P = 19.68$.

1.43. $P = (0.08317)^{1/5}$.

Solución

$\log P = \frac{1}{5} \log 0.08317 = \frac{1}{5}(8.9200 - 10) = \frac{1}{5}(48.9200 - 50) = 9.7840 - 10$ y $P = 0.6081$.

1.44. $P = \dfrac{\sqrt{0.003654}(18.37)^3}{(8.724)^4 \sqrt[4]{743.8}}$.

Solución

$\log P = \frac{1}{2} \log 0.003654 + 3 \log 18.37 - (4 \log 8.724 + \frac{1}{4} \log 743.8)$:

Numerador N			*Denominador D*		
$\frac{1}{2} \log 0.003654$	$= \frac{1}{2}(7.5628 - 10)$		$4 \log 8.724$	$= 4(0.9407)$	$= 3.7628$
	$= \frac{1}{2}(17.5628 - 20)$	$= 8.7814 - 10$	$\frac{1}{4} \log 743.6$	$= \frac{1}{4}(2.8714)$	$= 0.7178$
$3 \log 18.37$	$= 3(1.2641)$	$= 3.7923$	Sumar:	$\log D$	$= 4.4806$
Sumar:	$\log N$	$= 12.5737 - 10$			
	$(-) \log D$	$= 4.4806$			
	$\log P$	$= 8.0931 - 10$			
	P	$= 0.01239$			

1.45. $P = \sqrt{\dfrac{(874.3)(0.03816)(28.53)^3}{(1.754)^4(0.007352)}}$.

Solución

$\log P = \frac{1}{2}[\log 874.3 + \log 0.03816 + 3 \log 28.53 - (4 \log 1.754 + \log 0.007352)]$:

$\log 874.3$	$= 2.9417$	$= 2.9417$	
$\log 0.03816$	$= 8.5816 - 10$	$= 8.5816 - 10$	
$3 \log 28.53$	$= 3(1.4553)$	$= 4.3659$	
Sumar:		$15.8892 - 10$	(1)
$4 \log 1.754$	$= 4(0.2440)$	$= 0.9760$	
$\log 0.007352$		$= 7.8664 - 10$	
Sumar:		$8.8424 - 10$	(2)

De (1) y (2) tenemos que

$$\log P = \tfrac{1}{2}[(15.8892 - 10) - (8.8424 - 10)] = \tfrac{1}{2}(7.0468) = 3.5234 \text{ y } P = 3338$$

PROBLEMAS SUPLEMENTARIOS

VARIABLES

1.46. Decir cuáles de los que siguen representan datos discretos y cuáles continuos:

(*a*) Centímetros de lluvia en una ciudad durante varios meses.
(*b*) Velocidad de un coche (km/h).
(*c*) Número de billetes de $20 en circulación en EE.UU. en cada momento.
(*d*) Volumen de negocio diario en la Bolsa de Tokio.
(*e*) Número de estudiantes matriculados en una Universidad en varios años.

1.47. Dar el dominio de cada variable y decir si son discretas o continuas:

(*a*) Número W de bushels de trigo producidos por acre en un campo en varios años.
(*b*) Número N de miembros en una familia.
(*c*) Estado civil de una persona.
(*d*) Tiempo de vuelo T de un misil.
(*e*) Número P de pétalos de una flor.

REDONDEO DE DATOS, NOTACION CIENTIFICA Y DIGITOS SIGNIFICATIVOS

1.48. Redondear cada número con la precisión indicada:

(*a*) 3256 centenas.
(*b*) 5.781 decenas.
(*c*) 0.0045 milésimas.
(*d*) 46.7385 centésimas.
(*e*) 125.9995 dos cifras decimales.
(*f*) 3,502,378 millones.
(*g*) 148.475 unidades.
(*h*) 0.000098501 millonésimas.
(*i*) 2184.73 decenas.
(*j*) 43.87500 centésimas.

1.49. Expresar cada número sin usar potencias de 10:

(*a*) 132.5×10^{4}
(*b*) 418.72×10^{-5}
(*c*) 280×10^{-7}
(*d*) 7300×10^{6}
(*e*) 3.487×10^{-4}
(*f*) 0.0001850×10^{5}

1.50. ¿Cuántos dígitos significativos hay en estos números, supuesto que se dan con la mayor precisión posible?

(*a*) 2.54 cm
(*b*) 0.004500 yd
(*c*) 3,510,000 bu
(*d*) 3.51 millones bu
(*e*) 10.000100 pies
(*f*) 378 personas
(*g*) 378 oz
(*h*) 4.50×10^{-3} km
(*i*) 500.8×10^{5} kg
(*j*) 100.00 mi

1.51. ¿Cuál es el error máximo en cada una de las medidas siguientes, supuesto que se dan con la mayor precisión posible? Decir en cada caso el número de dígitos significativos.

(*a*) 7.20 millones bu
(*b*) 0.00004835 cm
(*c*) 5280 pies
(*d*) 3.0×10^{8} m
(*e*) 186,000 mi/seg
(*f*) 186 miles mi/seg

1.52. Escribir estos números en notación científica, supuesto que todos son dígitos significativos salvo mención expresa en contra.

(*a*) 0.000317
(*b*) 428,000,000 (cuatro cifras significativas)
(*c*) 21,600.00
(*d*) 0.000009810
(*e*) 732 miles
(*f*) 18.0 diezmilésimas

CALCULOS

1.53. Probar que: (*a*) el producto y (*b*) el cociente de 72.48 y 5.16, supuesto que tienen cuatro y tres dígitos significativos, respectivamente, no admiten más de tres dígitos significativos. Escribir los resultados con la mejor precisión posible.

1.54. Efectuar cada operación, suponiendo que los números se dan en la mayor precisión posible.

(*a*) 0.36×781.4

(*b*) $\dfrac{873.00}{4.881}$

(*c*) $5.78 \times 2700 \times 16.00$

(*d*) $\dfrac{0.00480 \times 2300}{0.2084}$

(*e*) $\sqrt{120 \times 0.5386 \times 0.4614}$ (120 exacto)

(*f*) $\dfrac{(416{,}000)(0.000187)}{\sqrt{73.84}}$

(*g*) $14.8641 + 4.48 - 8.168 + 0.36125$

(*h*) $4{,}173{,}00 - 170{,}264 + 1{,}820{,}470 - 78{,}320$ (los números son exactos en, respectivamente, 4, 6, 6 y 5 cifras significativas)

(*i*) $\sqrt{\dfrac{7(4.386)^2 - 3(6.47)^2}{6}}$ (3, 6 y 7 son exactos)

(*j*) $4.120\sqrt{\dfrac{3.1416[(9.483)^2 - (5.075)^2]}{0.0001980}}$

1.55. Evaluar lo que sigue, sabiendo que $U = -2$, $V = \frac{1}{2}$, $W = 3$, $X = -4$, $Y = 9$ y $Z = \frac{1}{6}$, donde todos los números son exactos.

(*a*) $4U + 6V - 2W$

(*b*) $\dfrac{XYZ}{UVW}$

(*c*) $\dfrac{2X - 3Y}{UW + XV}$

(*d*) $3(U - X)^2 + Y$

(*e*) $\sqrt{U^2 - 2UV + W}$

(*f*) $3X(4Y + 3Z) - 2Y(6X - 5Z) - 25$

(*g*) $\sqrt{\dfrac{(W - 2)^2}{V} + \dfrac{(Y - 5)^2}{Z}}$

(*h*) $\dfrac{X - 3}{\sqrt{(Y - 4)^2 + (U + 5)^2}}$

(*i*) $X^3 + 5X^2 - 6X - 8$

(*j*) $\dfrac{U - V}{\sqrt{U^2 + V^2}}[U^2V(W + X)]$

FUNCIONES, TABLAS Y GRAFICOS

1.56. Una variable Y queda determinada por otra X mediante $Y = 10 - 4X$.

(*a*) Hallar Y tal que $X = -3, -2, -1, 0, 1, 2, 3, 4$ y 5, y poner los resultados en una tabla.

(*b*) Hallar Y tal que $X = -2.4, -1.6, -0.8, 1.8, 2.7, 3.5$ y 4.6.

(*c*) Si denotamos la dependencia entre X e Y por $Y = F(X)$, calcular $F(2.8)$, $F(-5)$, $F(\sqrt{2})$ y $F(-\pi)$.

(*d*) ¿Qué valor de X corresponde a $Y = -2, 6, -10, 1.6, 16, 0$ y 10?

(*e*) Expresar X explícitamente como función de Y.

1.57. Si $Z = X^2 - Y^2$, calcular Z cuando: (*a*) $X = -2$, $Y = 3$, y (*b*) $X = 1$, $Y = 5$. (*c*) En la notación funcional $Z = F(X, Y)$, cuando $F(-3, -1)$.

1.58. Si $W = 3XZ - 4Y^2 + 2XY$, calcular W cuando: (*a*) $X = 1$, $Y = -2$, $Z = 4$, y (*b*) $X = -5$, $Y = -2$, $Z = 0$. (*c*) Con la notación funcional $W = F(X, Y, Z)$, calcular $F(3, 1, -2)$.

1.59. Localizar en un sistema de coordenadas rectangulares los puntos de coordenadas: (*a*) (3, 2), (*b*) (2, 3), (*c*) (−4, 4), (*d*) (4, −4), (*e*) (−3, −2), (*f*) (−2, −3), (*g*) (−4.5, 3), (*h*) (−1.2, −2.4), (*i*) (0, −3) y (*j*) (1.8, 0).

1.60. Representar las ecuaciones: (*a*) $Y = 10 - 4X$ (véase Prob. 1.56), (*b*) $Y = 2X + 5$, (*c*) $Y = \frac{1}{3}(X - 6)$, (*d*) $2X + 3Y = 12$ y (*e*) $3X - 2Y = 6$.

1.61. Representar las ecuaciones: (*a*) $Y = 2X^2 + X - 10$ y (*b*) $Y = 6 - 3X - X^2$.

1.62. Representar $Y = X^3 - 4X^2 + 12X - 6$.

1.63. La Tabla 1.8 muestra el número de trabajadores, agrícolas o no, en EE.UU. durante 1840-1980. Representar los datos usando: (*a*) gráfico de trazos, (*b*) gráfico de barras y (*c*) gráfico de barras en componentes.

Tabla 1.8

Año	Trabajadores agrícolas (millones)	Trabajadores no agrícolas (millones)
1840	3.72	1.70
1860	6.20	4.33
1880	8.59	8.80
1900	10.90	18.17
1920	11.46	30.97
1940	9.22	43.75
1960	4.19	65.70
1980	2.33	103.76

Fuente: U.S. Bureau of the Census.

1.64. Con los datos de la Tabla 1.8, diseñar un pictograma que muestre la variación en el número de trabajadores: (*a*) agrícolas y (*b*) no agrícolas. ¿Puede diseñar otro que las muestre a la vez?

1.65. Con los datos de la Tabla 1.8, construir un gráfico que muestre el porcentaje de trabajadores: (*a*) agrícolas y (*b*) no agrícolas. ¿Puede diseñar otro que las muestre a la vez?

1.66. La Tabla 1.9 da la expectativa de vida de un niño nacido en EE.UU. durante 1920-1980. Llevar los datos a un gráfico.

Tabla 1.9

Año	Varones	Hembras
1920	53.6	54.6
1930	58.1	61.6
1940	60.8	65.2
1950	65.6	71.1
1960	66.6	73.1
1970	67.1	74.7
1980	70.0	77.4

Fuente: National Center for Health Statistics.

1.67. La Tabla 1.10 recoge las velocidades orbitales de los planetas del sistema solar. Representar esos datos.

Tabla 1.10

Planeta	Velocidad (m/seg)
Mercurio	29.7
Venus	21.8
Tierra	18.5
Marte	15.0
Júpiter	8.1
Saturno	6.0
Urano	4.2
Neptuno	3.4
Plutón	3.0

1.68. En la Tabla 1.11 se ven los números (en millones) de estudiantes en enseñanza elemental, media y superior («colleges») en EE.UU. Representar los datos, usando: (*a*) gráficos de trazos, (*b*) gráficos de barras y (*c*) gráficos de barras en componentes.

Tabla 1.11

Año	1960	1965	1970	1975	1980
Elemental	32.4	35.5	37.1	33.8	30.6
Media	10.2	13.0	14.7	15.7	14.6
Superior	3.6	5.7	7.4	9.7	10.2

Fuente: U.S. Bureau of the Census.

1.69. Representar los datos de la Tabla 1.11 en un gráfico de porcentajes en componentes.

1.70. La Tabla 1.12 muestra el estado civil de hombres y mujeres (de más de 18 años) en EE.UU. en 1983. Representar los datos mediante: (*a*) dos gráficos circulares de igual diámetro y (*b*) un gráfico de diseño propio.

Tabla 1.12

Estado civil	Varones (% total)	Hembras (% total)
Soltero	25.1	18.4
Casado	66.7	61.3
Viudo	2.4	12.4
Divorciado	5.8	7.9

Fuente: U.S. Bureau of the Census.

1.71. En la Tabla 1.13 figuran las declaraciones de quiebra habidas en EE.UU. en 1975-1986. Representar los datos usando gráficos adecuados.

Tabla 1.13

Año	Total de declaraciones de quiebra
1975	11,432
1976	9,628
1977	7,919
1978	6,619
1979	7,564
1980	11,742
1981	16,794
1982	24,908
1983	31,334
1984	52,078
1985	57,252
1986	61,183

Fuente: Survey of Current Business.

1.72. La Tabla 1.14 recoge la relación entre divorcios y bodas en EE.UU. durante 1900-1980. Representar los datos en dos tipos de gráficos.

Tabla 1.14

Año	Relación entre divorcios y bodas
1900	0.079
1910	0.088
1920	0.134
1930	0.174
1940	0.165
1950	0.231
1960	0.258
1970	0.328
1980	0.491

Fuente: U.S. Department of Health and Human Services.

1.73. La Tabla 1.15 da, redondeados al millón, los países más poblados en 1986. Representar los datos por dos métodos diferentes.

Tabla 1.15

País	Población (millones)
China	1038
India	768
U.R.S.S.	278
EE.UU.	239
Indonesia	173
Brasil	135
Japón	121

Fuente: Naciones Unidas.

1.74. Representar los datos de la Tabla 1.15 teniendo en cuenta que la población mundial era en 1986 de 4850 millones.

1.75. En la Tabla 1.16 se ven las áreas de los océanos en millones de millas cuadradas. Representar los datos usando: (*a*) un gráfico de barras y (*b*) un gráfico circular.

Tabla 1.16

Océano	Area (millones de millas cuadradas)
Pacífico	63.8
Atlántico	31.5
Indico	28.4
Antártico	7.6
Artico	4.8

Fuente: Naciones Unidas.

ECUACIONES

1.76. Resolver las ecuaciones:

(*a*) $16 - 5c = 36$

(*b*) $2Y - 6 = 4 - 3Y$

(*c*) $4(X - 3) - 11 = 15 - 2(X + 4)$

(*d*) $3(2U + 1) = 5(3 - U) + 3(U - 2)$

(*e*) $3[2(X + 1) - 4] = 10 - 5(4 - 2X)$

(*f*) $\frac{2}{5}(12 + Y) = 6 - \frac{1}{4}(9 - Y)$

1.77. Resolver las ecuaciones simultáneas:

(*a*) $2a + b = 10$
$7a - 3b = 9$

(*b*) $3a + 5b = 24$
$2a + 3b = 14$

(*c*) $8X - 3Y = 2$
$3X + 7Y = -9$

(*d*) $5A - 9B = -10$
$3A - 4B = 16$

(*e*) $2a + b - c = 2$
$3a - 4b + 2c = 4$
$4a + 3b - 5c = -8$

(*f*) $5X + 2Y + 3Z = -5$
$2X - 3Y - 6Z = 1$
$X + 5Y - 4Z = 22$

(*g*) $3U - 5V + 6W = 7$
$5U + 3V - 2W = -1$
$4U - 8V + 10W = 11$

1.78. (*a*) Representar las ecuaciones $5X+2Y=4$ y $7X - 3Y = 23$, usando el mismo sistema coordenado.

(*b*) Determinar, con tales gráficos, la solución simultánea de ambas ecuaciones.

(*c*) Repetir las partes (*a*) y (*b*) para las ecuaciones simultáneas (*a*)-(*d*) del Problema 1.77.

1.79. (*a*) Usar el gráfico del Problema 1.61(*a*) para resolver la ecuación $2X^2 + X - - 10 = 0$. (*Ayuda:* Hallar los valores de X en que la parábola corta al eje X, es decir, donde $Y = 0$.)

(*b*) Por el método de la parte (*a*), resuélvase $3X^2 - 4X - 5 = 0$.

1.80. Las soluciones de la ecuación cuadrática $aX^2 + bX + c = 0$ vienen dadas por la fórmula cuadrática:

$$X = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Usarla para hallar las soluciones de: (*a*) $3X^2 - -4X - 5 = 0$, (*b*) $2X^2 + X - 10 = 0$, (*c*) $5X^2 + 10X = 7$ y (*d*) $X^2 + 8X + + 25 = 0$.

DESIGUALDADES

1.81. Usando símbolos de desigualdad, poner los números -4.3, -6.15, 2.37, 1.52 y -1.5 en orden: (*a*) creciente y (*b*) decreciente.

1.82. Expresar con símbolos de desigualdad las afirmaciones siguientes:

(*a*) N está entre 30 y 50 inclusive.
(*b*) S no es menor que 7.
(*c*) X es mayor o igual que -4, pero menor que 3.
(*d*) P es a lo sumo 5.
(*e*) X sobrepasa a Y en al menos 2.

1.83. Resolver las desigualdades:

(*a*) $3X \geqslant 12$
(*b*) $4X < 5X - 3$
(*c*) $2N + 15 > 10 + 3N$
(*d*) $3 + 5(Y - 2) \leqslant 7 - 3(4 - Y)$
(*e*) $-3 \leqslant \frac{1}{5}(2X + 1) \leqslant 3$
(*f*) $0 < \frac{1}{2}(15 - 5N) \leqslant 12$
(*g*) $-2 \leqslant 3 + \frac{1}{2}(a - 12) < 8$

LOGARITMOS Y ANTILOGARITMOS

1.84. Hallar los logaritmos comunes de:

(*a*) 387
(*b*) 0.387
(*c*) 0.0792
(*d*) 14,630
(*e*) 0.6042
(*f*) 0.002795
(*g*) 476.3
(*h*) 1.007
(*i*) 7.146
(*j*) 71.46
(*k*) 0.00098
(*l*) 84,620,000

1.85. Hallar los antilogaritmos de:

(*a*) 3.5611
(*b*) 9.8293 − 10
(*c*) 1.7045
(*d*) 8.9266 − 10
(*e*) 2.4700
(*f*) 6.4700 − 10
(*g*) $\bar{2}$.8003
(*h*) 3.7072
(*i*) 0.0800
(*j*) 6.3841

1.86. Evaluar mediante logaritmos:

(*a*) (783.6)(1654)

(*b*) $\dfrac{21.7}{378.2}$

(*c*) $\dfrac{(0.04556)(624.1)}{(14.32)(0.003572)}$

(*d*) $(1.562)^{15}$

(*e*) $\dfrac{(0.3854)^4(12.48)^2}{(0.04382)^3}$

(*f*) $0.04182\sqrt{0.6758}$

(*g*) $\sqrt[3]{3728}$

(*h*) $\sqrt[5]{(21.63)(33.81)(47.53)(65.28)(87.47)}$

(*i*) $\sqrt{\dfrac{(48.79)(0.00574)^3}{(2.143)^5}}$

(*j*) $\dfrac{3.781}{0.01873}\sqrt{\dfrac{(43.25)(0.08743)}{(0.002356)(6.824)}}$

1.87. Representar: (*a*) $Y = \log X$ y (*b*) $Y = 10^X$ y discutir las analogías entre ambos gráficos.

1.88. Escribir sin usar logaritmos las ecuaciones: (*a*) $2 \log X - 3 \log Y = 2$ y (*b*) $\log Y + 2X = \log 3$.

1.89. Si $a^p = N$, donde a y p son positivos y $a \neq 1$, llamamos a p el *logaritmo de N en base a*, y escribimos $p = \log_a N$. Evaluar: (*a*) $\log_2 8$, (*b*) $\log_{25} 125$, (*c*) $\log_4 1/16$, (*d*) $\log_{1/2} 32$ y (*e*) $\log_5 1$.

1.90. Probar que $\log_e N = 2.303 \log_{10} N$, aproximadamente, donde $e = 2.71828...$ se llama *base natural* de logaritmos y donde $N > 0$.

1.91. Probar que $(\log_b a)(\log_a b) = 1$, donde $a > 0$, $b > 0$, $a \neq 1$ y $b \neq 1$.

CAPITULO 2

Distribuciones de frecuencias

FILAS DE DATOS

Una *fila de datos* consiste en datos recogidos que no han sido organizados numéricamente, por ejemplo, las alturas de 100 estudiantes por letra alfabética.

ORDENACIONES

Una *ordenación* es un conjunto de datos numéricos en orden creciente o decreciente. La diferencia entre el mayor y el menor se llama *rango* de ese conjunto de datos. Así, si la mayor altura de entre los 100 estudiantes era de 74 in y la menor de 60 in, el rango es $74 - 60 = 14$ in.

DISTRIBUCIONES DE FRECUENCIAS

Al resumir grandes colecciones de datos, es útil distribuirlos en *clases* o *categorías*, y determinar el número de individuos que pertenecen a cada clase, llamado *frecuencia de clase*. Una disposición tabular de los datos por clases junto con las correspondientes frecuencias de clase, se llama *distribución de frecuencias* (o *tabla de frecuencias*). La Tabla 2.1 es una distribución de frecuencias de alturas (con precisión de 1 pulgada) de 100 estudiantes varones de la Universidad XYZ.

Tabla 2.1. Alturas de 100 estudiantes varones de la Universidad XYZ

Altura (in)	Número de estudiantes
60-62	5
63-65	18
66-68	42
69-71	27
72-74	8
	Total 100

La primera clase (o categoría), por ejemplo, consta de las alturas entre 60 y 62 in, y se indica por el rango 60-62. Como hay 5 estudiantes en esta clase, la correspondiente frecuencia de clase es 5.

Los datos así organizados en clases como en la anterior distribución de frecuencias se llaman *datos agrupados*. Aunque el proceso de agrupamiento destruye en general detalles de los datos iniciales, es muy ventajosa la visión nítida obtenida y las relaciones evidentes que saca a la luz.

INTERVALOS DE CLASE Y LIMITES DE CLASE

El símbolo que define una clase, como el 60-62 en la Tabla 2.1, se llama un *intervalo de clase*. Los números extremos, 60 y 62, se llaman *límite inferior de clase* (60) y *límite superior de clase* (62). Con frecuencia se intercambian los términos clase e intervalo de clase, aunque el intervalo de clase es un símbolo para la clase.

Un intervalo de clase que, al menos en teoría, carece de límite superior o inferior indicado, se llama *intervalo de clase abierto*. Por ejemplo, refiriéndonos a edades de personas, la clase «65 años o más» es un intervalo de clase abierto.

FRONTERAS DE CLASE

Si se dan alturas con precisión de 1 pulgada, el intervalo de clase 60-62 incluye teóricamente todas las medidas desde 59.5000 a 62.5000. in. Estos números, indicados más brevemente por los números exactos 59.5 y 62.5, se llaman *fronteras de clase* o verdaderos límites de clase; el menor (59.5) es la *frontera inferior* y el mayor (62.5) la *frontera superior*.

En la práctica, las fronteras de clase se obtienen promediando el límite superior de una clase con el inferior de la siguiente.

A veces se usan las fronteras de clase como símbolos para la clase. Así, las clases de la primera columna de la Tabla 2.1 se pueden indicar por 59.5-62.5, 62.5-65.5, etc. Para evitar ambigüedad en tal notación, las fronteras no deben coincidir con valores realmente medidos. De modo que si una observación diera 62.5, no sería posible decidir si pertenece al intervalo de clase 59.5-62.5 o al 62.5-65.5.

TAMAÑO O ANCHURA DE UN INTERVALO DE CLASE

El *tamaño* o *anchura de un intervalo de clase* es la diferencia entre las fronteras de clase superior e inferior. Si todos los intervalos de clase de una distribución de frecuencias tienen la misma anchura, la denotaremos por c. En tal caso, c es igual a la diferencia entre dos límites inferiores (o superiores) de clases sucesivas. Para los datos de la Tabla 2.1, por ejemplo, la anchura del intervalo de clase es $c = 62.5 - 59.5 = 65.5 - 62.5 = 3$.

MARCA DE CLASE

La *marca de clase* es el punto medio del intervalo de clase y se obtiene promediando los límites inferior y superior de clase. Así que las marcas de clase del intervalo 60-62 es $(60 + 62)/2 = 61$. La marca de clase se denomina también *punto medio de la clase*.

A efectos de análisis subsiguientes, todas las observaciones pertenecientes a un mismo intervalo de clase se supone que coinciden con la marca de clase. De manera que todas las alturas en el intervalo de clase 60-62 in se considerarán de 61 in.

REGLAS GENERALES PARA FORMAR DISTRIBUCIONES DE FRECUENCIAS

1. Determinar el mayor y el menor de todos los datos, hallando así el rango (diferencia entre ambos).
2. Dividir el rango en un número adecuado de intervalos de clase del mismo tamaño. Si ello no es factible, usar intervalos de clase de distintos tamaños o intervalos de clase abiertos (véase Problema 2.12). Se suelen tomar entre 5 y 20 intervalos de clase, según los datos. Los intervalos de clase se eligen también de modo tal que las marcas de clase (o puntos medios) coincidan con datos realmente observados. Ello tiende a disminuir el llamado *error de agrupamiento* que se produce en análisis ulteriores. No obstante, las fronteras de clase no debieran coincidir con datos realmente observados.
3. Determinar el número de observaciones que caen dentro de cada intervalo de clase; esto es, hallar las frecuencias de clase. Esto se logra mejor con una *hoja de recuentos* (véase Prob. 2.8).

HISTOGRAMAS Y POLIGONOS DE FRECUENCIAS

Los histogramas y los polígonos de frecuencias son dos representaciones gráficas de las distribuciones de frecuencias.

1. Un *histograma* o *histograma de frecuencias*, consiste en un conjunto de rectángulos con: (*a*) bases en el eje X horizontal, centros en las marcas de clase y longitudes iguales a los tamaños de los intervalos de clase y (*b*) áreas proporcionales a las frecuencias de clase.

   Si los intervalos de clase tienen todos la misma anchura, las alturas de los rectángulos son proporcionales a las frecuencias de clase, y entonces es costumbre tomar las alturas iguales a las frecuencias de clase. En caso contrario, deben ajustarse las alturas (véase Problema 2.13).

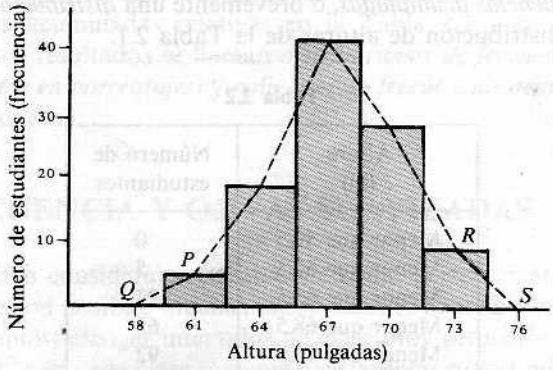


Figura 2.1.

2. Un *polígono de frecuencias* es un gráfico de trozos de la frecuencia de clase con relación a la marca de clase. Puede obtenerse conectando los puntos medios de las partes superiores de los rectángulos del histograma.

Histograma y polígono de frecuencias correspondientes a la distribución de frecuencias de alturas en la Tabla 2.1 se indican sobre los mismos ejes en la Figura 2.1. Suelen añadirse las longitudes PQ y RS a las marcas de clase extremas como asociadas a una frecuencia de clase cero. En tal caso, la suma de las áreas de los rectángulos del histograma es igual al área total limitada por el polígono de frecuencias y el eje X (véase Prob. 2.11).

DISTRIBUCIONES DE FRECUENCIAS RELATIVAS

La *frecuencia relativa de una clase* es su frecuencia dividida por la frecuencia total de todas las clases y se expresa generalmente como un porcentaje. Por ejemplo, la frecuencia relativa de la clase 66-68 en la Tabla 2.1 es 42/100 = 42%. La suma de las frecuencias relativas de todas las clases da obviamente 1, o sea 100 por 100.

Si se sustituyen las frecuencias de la Tabla 2.1 por las correspondientes frecuencias relativas, la tabla resultante se llama una *distribución de frecuencias relativas, distribución de porcentajes o tablas de frecuencias relativas.*

La representación gráfica de distribuciones de frecuencias relativas se puede obtener del histograma o del polígono de frecuencias sin más que cambiar la escala vertical de frecuencias a frecuencias relativas, manteniendo exactamente el mismo diagrama. Los gráficos resultantes se llaman *histogramas de frecuencias relativas* (o *histogramas de porcentajes*) y *polígonos de frecuencias relativas* (o *polígonos de porcentajes*), respectivamente.

DISTRIBUCIONES DE FRECUENCIAS ACUMULADAS Y OJIVAS

La frecuencia total de todos los valores menores que la frontera de clase superior de un intervalo de clase dado se llama *frecuencia acumulada* hasta ese intervalo de clase inclusive. Por ejemplo, la frecuencia acumulada hasta el intervalo de clase 66-68 en la Tabla 2.1 es 5 + 18 + 42 = 65, lo que significa que 65 estudiantes tienen alturas por debajo de 68.5 in.

Una tabla que presente tales frecuencias acumuladas se llama una *distribución de frecuencias acumuladas, tabla de frecuencias acumuladas*, o brevemente una *distribución acumulada*, y se muestra en la Tabla 2.2 para la distribución de alturas de la Tabla 2.1.

Tabla 2.2

Altura (in)	Número de estudiantes
Menor que 59.5	0
Menor que 62.5	5
Menor que 65.5	23
Menor que 68.5	65
Menor que 71.5	92
Menor que 74.5	100

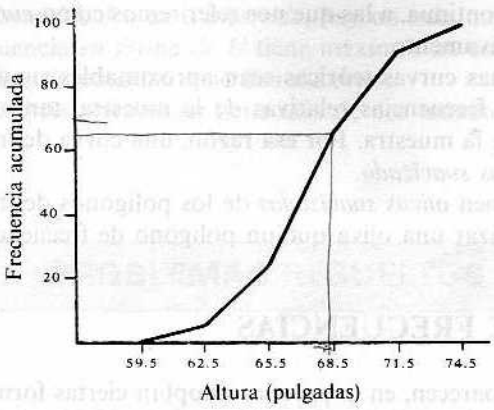


Figura 2.2.

Un gráfico que recoja las frecuencias acumuladas por debajo de cualquiera de las fronteras de clase superiores respecto de dicha frontera se llama un *polígono de frecuencias acumuladas* u *ojiva*, y se ilustra en la Figura 2.2 para las alturas de estudiantes de la Tabla 2.1.

A ciertos efectos, es deseable considerar una distribución de frecuencias acumuladas de todos los valores mayores o iguales que la frontera de clase inferior de cada intervalo de clase. Como eso hace considerar alturas de 59.5 in o más, de 62.5 in o más, etc., se le suele llamar una *distribución acumulada «o más»*, mientras que la antes considerada es una *distribución acumulada «menor que»*. Es fácil deducir una de otra (véase Prob. 2.15). Las correspondientes ojivas se conocen con los mismos apodos. Siempre que nos refiramos a distribuciones acumuladas u ojivas sin más, estaremos hablando del caso «menor que».

DISTRIBUCIONES DE FRECUENCIAS RELATIVAS Y OJIVAS DE PORCENTAJES

La *frecuencia acumulada relativa* o *frecuencia acumulada en porcentajes*, es la frecuencia acumulada dividida por la frecuencia total. Así, la frecuencia acumulada relativa de alturas menores que 68.5 in es 65/100 = 65%, lo que significa que el 65% de los estudiantes mide menos de 68.5 in.

Si se usan frecuencias acumuladas relativas en la Tabla 2.2 y en la Figura 2.2 en vez de frecuencias acumuladas, los resultados se llaman *distribuciones de frecuencias acumuladas relativas* (o *distribuciones acumuladas en porcentajes*) y *polígonos de frecuencias acumuladas relativas* (u *ojivas de porcentajes*), respectivamente.

CURVAS DE FRECUENCIA Y OJIVAS SUAVIZADAS

Los datos recogidos pueden considerarse usualmente como pertenecientes a una muestra de una población grande. Ya que son posibles muchas observaciones sobre esa población, es teóricamente posible (para datos continuos) escoger intervalos de clase muy pequeños y tener todavía números razonables de observaciones en cada clase. Así que cabe esperar que el polígono de frecuencias o el polígono de frecuencias relativas para una gran población tenga tantos pequeños segmentos que

aparezca como casi una curva continua, a las que nos referiremos como *curva de frecuencias* o *curva de frecuencias relativas*, respectivamente.

Es sensato esperar que dichas curvas teóricas sean aproximables suavizando los polígonos de frecuencias o los polígonos de frecuencias relativas de la muestra, tanto mejor la aproximación cuanto mayor sea el tamaño de la muestra. Por esa razón, una curva de frecuencias se cita a veces como un *polígono de frecuencias suavizado*.

De forma análoga, se obtienen *ojivas suavizadas* de los polígonos de frecuencias acumuladas u ojivas. Suele ser más fácil suavizar una ojiva que un polígono de frecuencias (véase Prob. 2.18).

TIPOS DE CURVAS DE FRECUENCIAS

Las curvas de frecuencia que aparecen, en la práctica adoptan ciertas formas características, como ilustra la Figura 2.3.

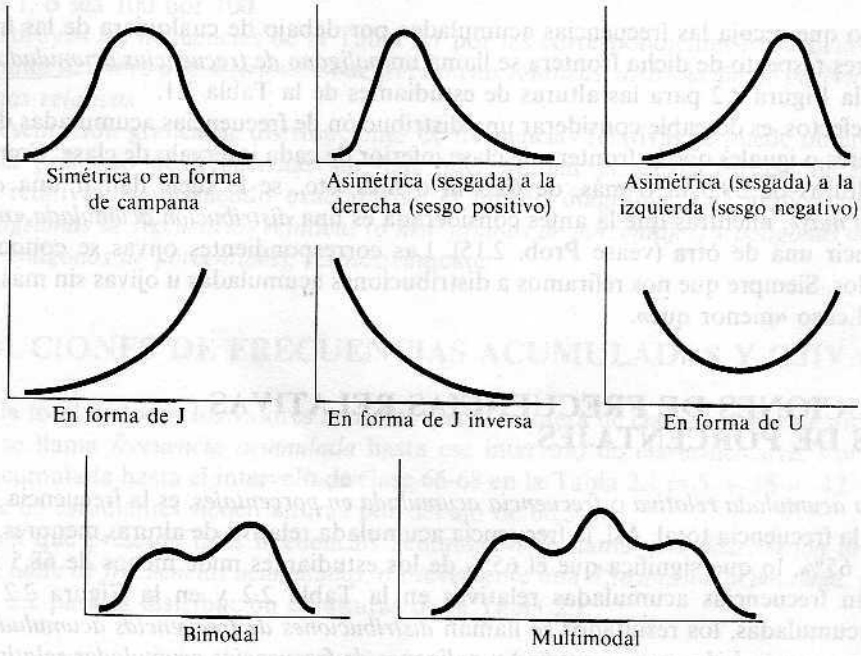


Figura 2.3.

1. Las curvas de frecuencias *simétricas* o *en forma de campana*, se caracterizan porque las observaciones equidistantes del máximo central tienen la misma frecuencia. Ejemplo importante es la curva normal.
2. En las curvas de frecuencia *poco asimétricas*, o *sesgadas*, la cola de la curva a un lado del máximo central es más larga que al otro lado. Si la cola mayor está a la derecha, la curva se dice *asimétrica a la derecha* o de *asimetría positiva*. En caso contrario, se dice *asimétrica a la izquierda* o de *asimetría negativa*.

3. En una curva *en forma de J* o de *J invertida*, hay un máximo en un extremo.
4. Una curva de frecuencia *en forma de U* tiene máximos en ambos extremos.
5. Una curva de frecuencia *bimodal* tiene dos máximos.
6. Una curva de frecuencia *multimodal* tiene más de dos máximos.

PROBLEMAS RESUELTOS

ORDENACIONES

2.1. (*a*) Disponer los números 17, 45, 38, 27, 6, 48, 11, 57, 34 y 22 en lista ordenada.
(*b*) Determinar el rango de esos números.

Solución

(*a*) En orden creciente: 6, 11, 17, 22, 27, 34, 38, 45, 48, 57. En orden decreciente: 57, 48, 45, 38, 34, 27, 22, 17, 11, 6.
(*b*) El menor es 6 y el mayor 57, luego el rango es $57 - 6 = 51$.

2.2. Las calificaciones finales en Matemáticas de 80 estudiantes figuran en la tabla adjunta.

68	84	75	82	68	90	62	88	76	93
73	79	88	73	60	93	71	59	85	75
61	65	75	87	74	62	95	78	63	72
66	78	82	75	94	77	69	74	68	60
96	78	89	61	75	95	60	79	83	71
79	62	67	97	78	85	76	65	71	75
65	80	73	57	88	78	62	76	53	74
86	67	73	81	72	63	76	75	85	77

Hallar en esa tabla:

(*a*) La calificación más alta.
(*b*) La más baja.
(*c*) El rango.
(*d*) Las cinco más altas.
(*e*) Las cinco más bajas.
(*f*) La décima de mayor a menor.
(*g*) El número de estudiantes con calificaciones de 75 o más.
(*h*) Idem por debajo de 85.
(*i*) El porcentaje de estudiantes con calificaciones mayores que 65 pero no superiores a 85.
(*j*) Las calificaciones que no aparecen.

Solución

Algunas de estas cuestiones son tan de detalle que se contestan mejor en una ordenación, lo cual se hace subdividiendo los datos en clases y colocando cada número de la tabla en su clase, como

en la Tabla 2.3, llamada *tabla de entrada única*. Ordenando entonces los de cada clase, como en la Tabla 2.4 es fácil deducir las respuestas a las cuestiones planteadas.

(*a*) 97.
(*b*) 53.
(*c*) Rango = 97 − 53 = 44.
(*d*) Las cinco más altas son 97, 96, 95, 95 y 94.
(*e*) Las cinco más bajas son 53, 57, 59, 60 y 60.
(*f*) 88.

Tabla 2.3

50-54	53
55-59	59, 57
60-64	62, 60, 61, 62, 63, 60, 61, 60, 62, 62, 63
65-69	68, 68, 65, 66, 69, 68, 67, 65, 65, 67
70-74	73, 73, 71, 74, 72, 74, 71, 71, 73, 74, 73, 72
75-79	75, 76, 79, 75, 75, 78, 78, 75, 77, 78, 75, 79, 79, 78, 76, 75, 78, 76, 76, 75, 77
80-84	84, 82, 82, 83, 80, 81
85-89	88, 88, 85, 87, 89, 85, 88, 86, 85
90-94	90, 93, 93, 94
95-99	95, 96, 95, 97

Tabla 2.4

50-54	53
55-59	57, 59
60-64	60, 60, 60, 61, 61, 62, 62, 62, 62, 63, 63
65-69	65, 65, 65, 66, 67, 67, 68, 68, 68, 69
70-74	71, 71, 71, 72, 72, 73, 73, 73, 73, 74, 74, 74
75-79	75, 75, 75, 75, 75, 75, 75, 76, 76, 76, 76, 77, 77, 78, 78, 78, 78, 78, 79, 79, 79
80-84	80, 81, 82, 82, 83, 84
85-89	85, 85, 85, 86, 87, 88, 88, 88, 89
90-94	90, 93, 93, 94
95-99	95, 95, 96, 97

(*g*) 44 estudiantes.
(*h*) 63 estudiantes.
(*i*) El porcentaje 85 es 49/80 = 61.2%.
(*j*) No aparecen 0, 1, 2, 3, ..., 52, 54, 55, 56, 58, 64, 70, 91, 92, 98, 99 y 100.

DISTRIBUCIONES DE FRECUENCIA, HISTOGRAMAS Y POLIGONOS DE FRECUENCIAS

2.3. La Tabla 2.5 muestra una distribución de frecuencia de los salarios semanales de 65 empleados de la empresa P&R. Determinar de esa tabla:

(*a*) El límite inferior de la sexta clase.
(*b*) El límite superior de la cuarta clase.
(*c*) La marca de clase (o punto medio) de la tercera clase.
(*d*) Las fronteras de clase del quinto intervalo.

(*e*) La anchura del quinto intervalo de clase.
(*f*) La frecuencia de la tercera clase.
(*g*) La frecuencia relativa de la tercera clase.
(*h*) El intervalo de clase con máxima frecuencia, que se llama *intervalo de clase modal*. Su frecuencia es la *frecuencia de clase modal*.
(*i*) El porcentaje de empleados que cobran menos de \$280.00 a la semana.
(*j*) El porcentaje de empleados que cobran menos de \$300.00 pero al menos \$260.00 por semana.

Tabla 2.5

Salarios	Número de empleados
\$250.00-\$259.99	8
260.00-269.99	10
270.00-279.99	16
280.00-289.99	14
290.00-299.99	10
300.00-309.99	5
310.00-319.99	2
Total	65

Solución

(*a*) \$300.00.
(*b*) \$289.99.
(*c*) La marca de clase de la tercera clase = $\frac{1}{2}$(\$270.00 + \$279.99) = \$274.995. A efectos prácticos se redondeará a \$275.00.
(*d*) La frontera de clase inferior de la quinta clase = $\frac{1}{2}$(\$290.00 + \$289.99) = \$289.995. La superior = $\frac{1}{2}$(\$299.99 + \$300.00) = \$299.995.
(*e*) Anchura del quinto intervalo de clase = frontera superior de la quinta clase − frontera inferior de la quinta clase = \$299.995 − \$289.985 = \$10.00. En este caso, todos los intervalos de clase son de la misma anchura: \$10.00.
(*f*) 16.
(*g*) 16/65 = 0.246 = 24.6%.
(*h*) \$270.00-\$279.99.
(*i*) Número de empleados que ganan menos de \$280 por semana = 16 + 10 + 8 = 34. Porcentaje de empleados que ganan menos de \$280 por semana = 34/65 = 52.3%.
(*j*) Número de empleados que cobran menos de \$300.00 pero al menos \$260 por semana = 10 + + 14 + 16 + 10 = 50. Porcentaje de empleados que cobran menos de \$300.00 pero al menos \$260 por semana = 50/65 = 76.9%.

2.4. Si las marcas de clase en una distribución de frecuencias de pesos de estudiantes son 128, 137, 146, 155, 164, 173 y 182 libras (lb), hallar: (*a*) la anchura del intervalo de clase, (*b*) las fronteras de clase y (*c*) los límites de clase, suponiendo que los pesos se midieron con 1 libra de precisión.

Solución

(*a*) Anchura del intervalo de clase = diferencia común entre marcas de clase sucesivas = 137 − 128 = = 146 − 137 = etc. = 9 lb.

(*b*) Como los intervalos de clase son de igual anchura, las fronteras de clase están a mitad de camino entre las marcas de clase, luego son

$$\tfrac{1}{2}(128 + 137),\ \tfrac{1}{2}(137 + 146),\ \ldots,\ \tfrac{1}{2}(173 + 182) \quad \text{o sea} \quad 132.5,\ 141.5,\ 150.5,\ \ldots,\ 177.5 \text{ lb}$$

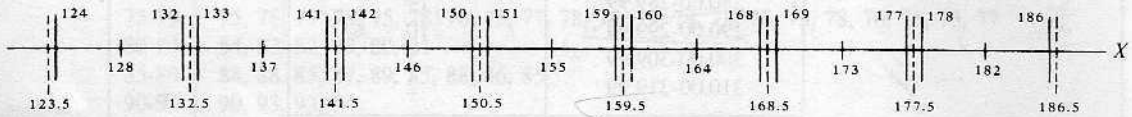
La primera frontera de clase es $132.5 - 9 = 123.5$ y la última $177.5 + 9 = 186.5$, ya que la anchura común de los intervalos de clase es 9 lb. Así pues, las fronteras de clase son

$$123.5,\ 132,\ 141.5,\ 150.5,\ 159.5,\ 168.5,\ 177.5,\ 186.5 \text{ lb}$$

(*c*) Como los límites de clase son enteros, los elegimos como los enteros más cercanos a las fronteras de clase, a saber, 123, 124, 132, 133, 141, 142, ... Luego la primera clase tiene límites 124-132, la siguiente 133-141, etc.

2.5. Representar gráficamente los resultados del Problema 2.4.

Solución



El gráfico se ve en el diagrama adjunto. Las marcas de clase 128, 137, 146, ..., 182 están localizadas en el eje *X*. Las fronteras de clase se indican por los segmentos verticales discontinuos, y los límites de clase por segmentos verticales sólidos.

2.6. La menor de 150 medidas es 5.18 in y la mayor 7.44 in. Determinar un conjunto apropiados de: (*a*) intervalos de clase, (*b*) fronteras de clase y (*c*) marcas de clase que puedan usarse para formar la distribución de frecuencias de esas medidas.

Solución

El rango es $7.44 - 5.18 = 2.26$ in. Para un mínimo de cinco intervalos de clase, la anchura de estos es $2.26/5 = 0.45$ aproximadamente; y para un máximo de 20 intervalos de clase la anchura es $2.26/20 = 0.11$ aproximadamente. Elecciones convenientes de la anchura de los intervalos de clase están entre 0.11 y 0.45, es decir, podrían ser 0.20, 0.30 ó 0.40.

(*a*) Las columnas I, II y III de la tabla adjunta muestran intervalos de clase de anchuras 0.20, 0.30 y 0.40, respectivamente.

I	II	III
5.10-5.29	5.10-5.39	5.10-5.49
5.30-5.49	5.40-5.69	5.50-5.89
5.50-5.69	5.70-5.99	5.90-6.29
5.70-5.89	6.00-6.29	6.30-6.69
5.90-6.09	6.30-6.59	6.70-7.09
6.10-6.29	6.60-6.89	7.10-7.49
6.30-6.49	6.90-7.19	
6.50-6.69	7.20-7.49	
6.70-6.89		
6.90-7.09		
7.10-7.29		
7.30-7.49		

Nótese que el límite inferior de cada primera clase podría haber sido distinto de 5.10; por ejemplo, si en la columna I hubiéramos partido de 5.15 como límite inferior, el primer intervalo de clase hubiera sido 5.15-5.34.

(*b*) Las fronteras de clase correspondientes a las columnas I, II y III de la parte (*a*) vienen dadas, respectivamente, por

I 5.095-5.295, 5.295-5.495, 5.495-5.695, ..., 7.295-7.495
II 5.095-5.395, 5.395-5.695, 5.695-5.995, ..., 7.195-7.495
III 5.095-5.495, 5.495-5.895, 5.895-6.295, ..., 7.095-7.495

Obsérvese que tales fronteras de clase son correctas, pues no coinciden con medidas obtenidas.

(*c*) Las marcas de clase correspondientes a las columnas I, II y III de (*a*) son

I 5.195, 5.395, ..., 7.395 II 5.245, 5.545, ..., 7.345 III 5.295, 5.695, ..., 7.295

Estas marcas de clase tienen la desventaja de no coincidir con medidas observadas.

2.7. Al contestar el Problema 2.6(*a*), un estudiante escogió los intervalos de clase 5.10-5.40, 5.40-5.70, ..., 6.90-7.20 y 7.20-7.50. ¿Hay algo incorrecto en su elección?

Solución

Esos intervalos de clase se solapan en 5.40, 5.70, ..., 7.20. Luego una medida anotada como 5.40, por ejemplo, podría ser colocada en cualquiera de los dos primeros intervalos de clase. Algunos estadísticos justifican esta elección decidiendo asignar la mitad de los casos dudosos a una clase y la otra mitad a la otra.

La ambigüedad desaparece escribiendo los intervalos de clase como 5.10 hasta 5.40, 5.40 hasta 5.70, etc. En este caso, los límites de clase coinciden con las fronteras de clase, y las marcas de clase pueden coincidir con datos observados.

En general, es deseable evitar solapamientos de intervalos de clase si es posible y escogerlos de modo que las fronteras de clase no coincidan con los datos observados. Por ejemplo, los intervalos de clase del Problema 2.6 podían haberse escogido como 5.095-5.395, 5.395-5.695, etc., sin ambigüedad. Una desventaja de esta elección particular es que las marcas de clase no coinciden con los datos observados.

2.8. En la tabla que sigue se recogen los pesos de 40 estudiantes varones de una universidad, con precisión de 1 libra. Construir una distribución de frecuencias.

138	164	150	132	144	125	149	157
146	158	140	147	136	148	152	144
168	126	138	176	163	199	154	165
146	173	142	147	135	153	140	135
161	145	135	142	150	156	145	128

Solución

Los pesos extremos son 176 y 119 lb, luego el rango es $176-119=57$ lb. Si se usan 5 intervalos de clase, su anchura será $57/5=11$ aproximadamente; si se usan 20 intervalos de clase, será de $57/20=3$, aproximadamente.

Una colección razonable es 5 lb. Es conveniente, asimismo, elegir las marcas de clase como 120, 125, 130, 135, ..., lb. De modo que los intervalos de clase pueden tomarse como 118-122, 123-127, 128-132, ... Con tal elección, las fronteras de clase son 117.5, 122.5, 127.5, ..., que no coinciden con los datos observados.

$\frac{57}{x}=5 \rightarrow x=11,4 \simeq 12$

Tabla 2.6

Peso (lb)	Recuento	Frecuencia
118-122	/	1
123-127	//	2
128-132	//	2
133-137	////	4
138-142	/// /	6
143-147	/// ///	8
148-152	///	5
153-157	///	4
158-162	//	2
163-167	///	3
168-172	/	1
173-177	//	2
		Total 40

Tabla 2.7

Peso (lb)	Recuento	Frecuencia
118-126	///	3
127-135	///	5
136-144	/// ////	9
145-153	/// /// //	12
154-162	///	5
163-171	////	4
172-180	//	2
		Total 40

La distribución de frecuencias requerida se ve en la Tabla 2.6. La columna central, llamada *hoja de recuentos*, se usa para tabular las frecuencias de clase y suele omitirse en la presentación final de la distribución de frecuencias. No es necesario hacer ordenación, aunque si se dispone de ella puede utilizarse para tabular las frecuencias.

Otro método

Naturalmente, existen otras distribuciones de frecuencias. La Tabla 2.7, por ejemplo, muestra una distribución de frecuencias con sólo 7 clases, en la que la anchura del intervalo de clase es 9 lb.

2.9. Construir: (*a*) un histograma y (*b*) un polígono de frecuencias para la distribución de pesos del Problema 2.8.

Solución

El histograma y el polígono de frecuencias para cada caso del Problema 2.8 vienen dados en las Figuras 2.4(*a*) y 2.4(*b*). Nótese que los centros de las bases de los rectángulos están localizados en las marcas de clase.

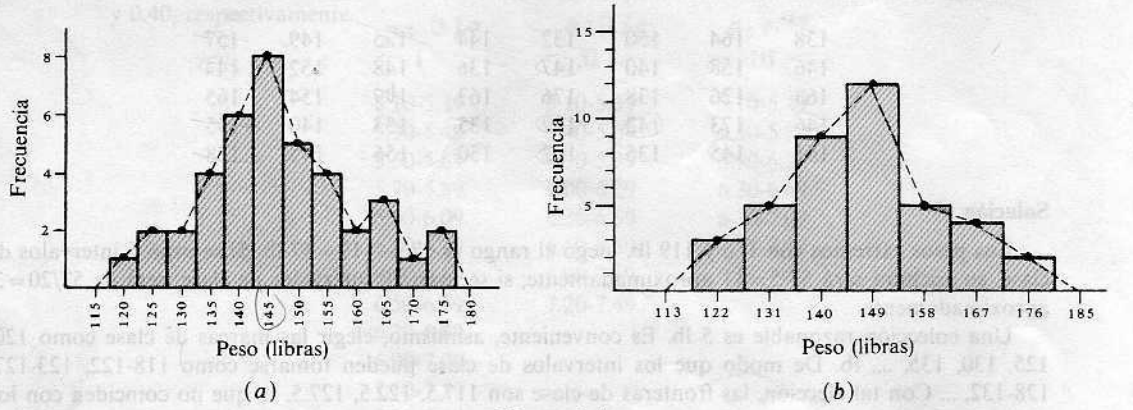


Figura 2.4.

2.10. Con los datos de la Tabla 2.5 del Problema 2.3, construir: (*a*) una distribución de frecuencias relativas, (*b*) un histograma, (*c*) un histograma de frecuencias relativas, (*d*) un polígono de frecuencias y (*e*) un polígono de frecuencias relativas.

Solución

(*a*) La distribución de frecuencias relativas de la Tabla 2.8 se obtiene de la distribución de frecuencias de la Tabla 2.5 dividiendo cada frecuencia de clase por la frecuencia total (65) y expresando el resultado como porcentaje.

(*b*) y (*c*) El histograma y el histograma de frecuencias relativas se muestran en la Figura 2.5. Nótese que para pasar de uno a otro sólo es necesario añadir al histograma una escala vertical con las frecuencias relativas, como se ve a la derecha en la Figura 2.5.

(*d*) y (*e*) El polígono de frecuencias y el polígono de frecuencias relativas se indican por la gráfica de trazos en la Figura 2.5. Así pues, para convertir un polígono de frecuencias en un polígono de frecuencias relativas, basta añadir una escala vertical con las frecuencias relativas.

Si sólo se desea un polígono de frecuencias relativas, la figura adjunta no contendría el histograma y el eje de las frecuencias relativas aparecería en la izquierda en lugar del eje de frecuencias.

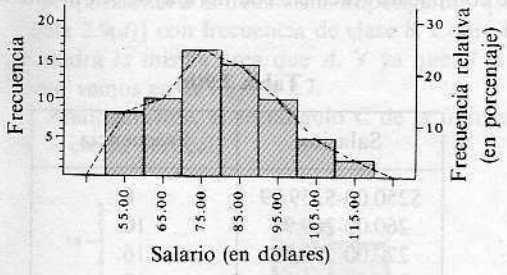


Figura 2.5.

Tabla 2.8

Salarios	Frecuencia relativa (como porcentaje)
\$250.00-\$259.99	12.3
260.00-269.99	15.4
270.00-279.99	24.6
280.00-289.99	21.5
290.00-299.99	15.4
300.00-309.99	7.7
310.00-319.99	3.1
	Total 100.0

2.11. Probar que en un histograma el área total de los rectángulos es igual al área total limitada por el correspondiente polígono de frecuencias y el eje X.

Solución

Lo probaremos para el caso de un histograma con tres rectángulos (Fig. 2.6) y el polígono de frecuencias asociado, que se indica con trazo discontinuo.

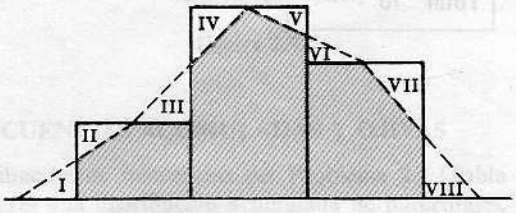


Figura 2.6.

Area total de los rectángulos = área sombreada + área II + área IV + área V + área VII
= área sombreada + área I + área III + área VI + área VIII
= área total acotada por el polígono de frecuencias y el eje X

Como área I = área II, entonces área III = área IV, área V = área VI y área VII = área VIII.

2.12. En la empresa P&R (Prob. 2.3), se ha contratado a cinco nuevos trabajadores con salarios semanales de \$285.34, \$316.83, \$335.78, \$356.21 y \$374.50. Construir una distribución de frecuencia de los salarios de los 70 trabajadores.

Solución

La Tabla 2.9 muestra posibles distribuciones de frecuencia.

En la Tabla 2.9(*a*) se ha usado un mismo tamaño de intervalos de clase \$10.00. Como consecuencia, hay demasiadas clases vacías y la información es más detallada en el extremo superior de la escala de salarios.

En la Tabla 2.9(*b*) las clases vacías y los detalles finos han sido evitados usando el intervalo de clase abierto «\$320.00 o más». Una desventaja sería que la tabla se haría menos cómoda al efectuar ciertos cálculos. Así, es imposible determinar la cantidad total pagada a la semana porque «\$320.00 o más» podría significar que hay individuos que cobran incluso \$1400.00 a la semana.

En la Tabla 2.9(*c*) se usa una anchura de intervalo de clase de \$20.00, con la desventaja de que se que ciertas operaciones matemáticas posteriores se complican. Además, cuanto mayor sea la anchura, mayor el error de agrupamiento.

Tabla 2.9(a)

Salarios	Frecuencia
\$250.00-\$259.99	8
260.00-269.99	10
270.00-279.99	16
280.00-289.99	15
290.00-299.99	10
300.00-309.99	5
310.00-319.99	3
320.00-329.99	0
330.00-339.99	1
340.00-349.99	0
350.00-359.99	1
360.00-369.99	0
370.00-379.99	1
	Total 70

Tabla 2.9(b)

Salarios	Frecuencia
\$250.00-\$259.99	8
260.00-269.99	10
270.00-279.99	16
280.00-289.99	15
290.00-299.99	10
300.00-309.99	5
310.00-319.99	3
320.00 en adelante	3
	Total 70

Tabla 2.9(c)

Salarios	Frecuencia
\$250.00-\$269.99	18
270.00-289.99	31
290.00-309.99	15
310.00-329.99	3
330.00-349.99	1
350.00-369.99	1
370.00-389.99	1
	Total 70

Tabla 2.9(d)

Salarios	Frecuencia
\$250.00-\$259.99	8
260.00-269.99	10
270.00-279.99	16
280.00-289.99	15
290.00-299.99	10
300.00-319.99	8
320.00-379.99	3
	Total 70

2.13. Construir un histograma para la distribución de frecuencias de la Tabla 2.9(*d*).

Solución

La Figura 2.7 muestra el histograma solicitado. Para construirlo usamos el hecho de que el área es proporcional a la frecuencia. Supongamos que el rectángulo *A* corresponde a la primera clase [véase Tabla 2.9(*d*)] con frecuencia de clase 8. Como la sexta clase tiene también frecuencia 8, su rectángulo *B* tendrá la misma área que *A*. Y ya que *B* es doble ancho que *A*, tendrá la mitad de su altura, tal como vemos en la Figura 2.7.

Análogamente, el rectángulo *C* de la última clase en la Tabla 2.9(*d*) tiene media unidad de altura en la escala vertical.

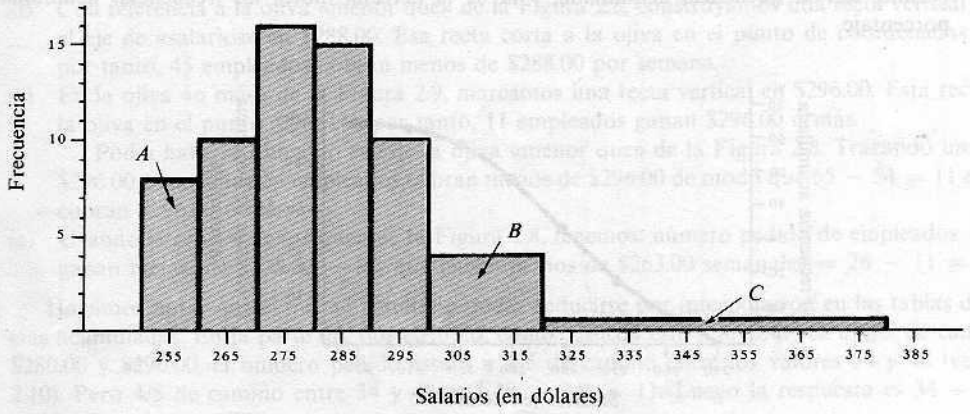


Figura 2.7.

DISTRIBUCIONES DE FRECUENCIAS ACUMULADAS Y OJIVAS

2.14. Construir para la distribución de frecuencias del Problema 2.3 (Tabla 2.5): (*a*) una distribución de frecuencias acumuladas, (*b*) una distribución acumulada de porcentajes, (*c*) una ojiva y (*d*) una ojiva de porcentajes.

Solución

(*a*) y (*b*) La distribución de frecuencias acumuladas y la distribución acumulada en porcentajes (o distribución de frecuencias relativas acumuladas) se combinan en la Tabla 2.10.

Tabla 2.10

Salarios	Frecuencia acumulada	Porcentaje acumulativo de distribución
Menor que 250.00	0	0.0
Menor que 260.00	8	12.3
Menor que 270.00	18	27.7
Menor que 280.00	34	52.3
Menor que 290.00	48	73.8
Menor que 300.00	58	89.2
Menor que 310.00	63	96.9
Menor que 320.00	65	100.0

Nótese que cada entrada de la columna 2 se obtiene sumando entradas sucesivas de la columna 2 de la Tabla 2.5. Luego 18 = 8 + 10, 34 = 8 + 10 + 16, etc.

Cada entrada en la columna 3 se obtiene de la anterior dividiendo por 65, la frecuencia total, y expresando el resultado como porcentaje. Así, 34/65 = 52.3%. También podían haberse obtenido sumando entradas sucesivas de la columna 2 de la Tabla 2.8. Así, 27.7 = 12.3 + 15.4, 52.3 = = 12.3 + 15.4 + 24.6, etc.

(*c*) y (*d*) La ojiva (o polígono de frecuencias acumuladas) y la ojiva de porcentajes (o polígono de frecuencias acumuladas relativas) se ven en la Figura 2.8. La escala vertical de la izquierda nos permite leer la frecuencia acumulada, y la de la derecha indica las frecuencias acumuladas en porcentaje.

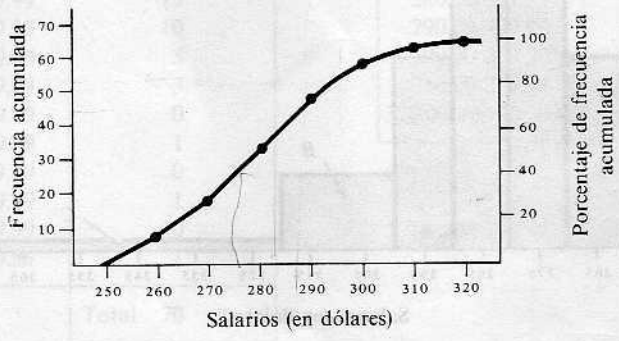


Figura 2.8.

Las anteriores suelen llamarse ojiva o distribución de frecuencias acumuladas «menor que», por la manera de acumular las frecuencias.

2.15. A partir de la distribución de frecuencias de la Tabla 2.5 del Problema 2.3, construir: (*a*) una distribución de frecuencias acumuladas «o más» y (*b*) una ojiva «o más».

Solución

(*a*) Cada entrada de la columna 2 en la Tabla 2.11 se obtiene sumando entradas sucesivas de la columna 2 de la Tabla 2.55, *comenzando por abajo;* así pues, 7 = 2 + 5, 17 = 2 + 5 + 10, etc. Estas entradas pueden obtenerse también restando cada entrada en la columna 2 de la Tabla 2.10 de la frecuencia total, 65, es decir, 57 = 65 − 8, 47 = 65 − 18, etc.

(*b*) La Figura 2.9 muestra una ojiva «o más».

Tabla 2.11

Salarios	Frecuencia acumulada «o más»
$250.00 o más	65
260.00 o más	57
270.00 o más	47
280.00 o más	31
290.00 o más	17
300.00 o más	7
310.00 o más	2
320.00 o más	0

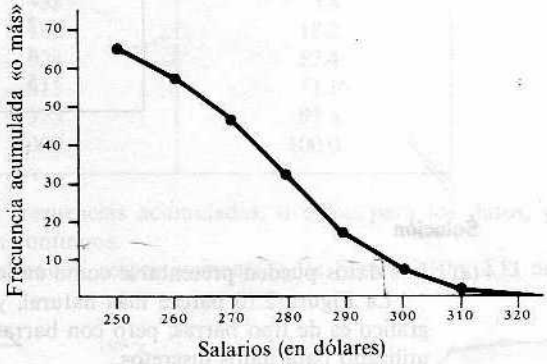


Figura 2.9.

2.16. De las ojivas en las Figuras 2.8 y 2.9 (de los Probs. 2.14 y 2.15, respectivamente), estimar el número de empleados que cobran por semana: (*a*) menos de $288.00, (*b*) $296.00 o más y (*c*) al menos $263.00, pero menos de $275.00.

Solución

(*a*) Con referencia a la ojiva «menor que» de la Figura 2.8, construyamos una recta vertical que corte al eje de «salarios» en $288.00. Esa recta corta a la ojiva en el punto de coordenadas (288, 45); por tanto, 45 empleados cobran menos de $288.00 por semana.

(*b*) En la ojiva «o más» de la Figura 2.9, marcamos una recta vertical en $296.00. Esta recta corta a la ojiva en el punto (296, 11); por tanto, 11 empleados ganan $296.00 o más.

Podía haberse obtenido eso de la ojiva «menor que» de la Figura 2.8. Trazando una recta en $296.00, vemos que 54 empleados cobran menos de $296.00 de modo que 65 − 54 = 11 empleados cobran $296.00 o más.

(*c*) Usando la ojiva «menor que» de la Figura 2.8, tenemos: número pedido de empleados = los que ganan menos de $275.00 − los que ganan menos de $263.00 semanales = 26 − 11 = 15.

Hagamos notar que el mismo resultado podía deducirse por interpolación en las tablas de frecuencias acumuladas. En la parte (*a*), por ejemplo, como $288.00 está a 8/10, o sea a 4/5, de camino entre $280.00 y $290.00, el número pedido estará a 4/5 de camino entre los valores 34 y 48 (véase Tabla 2.10). Pero 4/5 de camino entre 34 y 48 es $\frac{4}{5}(48 - 34) = 11$. Luego la respuesta es 34 + 11 = 45 empleados.

2.17. Se lanzan cinco monedas 1000 veces. El número de lanzamientos en los que han salido 0, 1, 2, 3, 4 y 5 caras se indican en la Tabla 2.12.

(*a*) Representar los datos de esa tabla.

(*b*) Construir una tabla que muestre los porcentajes de tiradas que han dado un número de caras menor que 0, 1, 2, 3, 4, 5 ó 6.

(*c*) Representar los datos de la tabla de la parte (*b*).

Tabla 2.12

Número de caras	Número de tiradas (frecuencia)
0	38
1	144
2	342
3	287
4	164
5	25
	Total 1000

Solución

(*a*) Los datos pueden presentarse como en las Figuras 2.10 ó 2.11.

La Figura 2.10 parece más natural, ya que el número de caras no puede ser 1.5 ó 3.2. Este gráfico es de tipo barras, pero con barras de anchura cero. Se llama *gráfico de varillas* y es muy utilizado para datos discretos.

La Figura 2.11 es un histograma de los datos. El área total del histograma es la frecuencia total, 1000, como debe ser. Al usar la representación en histograma o el correspondiente polígono de frecuencias, estamos tratando los datos como si fueran continuos. Luego veremos que tal perspectiva es útil. Recuérdese que ya hemos utilizado el histograma y los polígonos de frecuencias para datos discretos en el Problema 2.10.

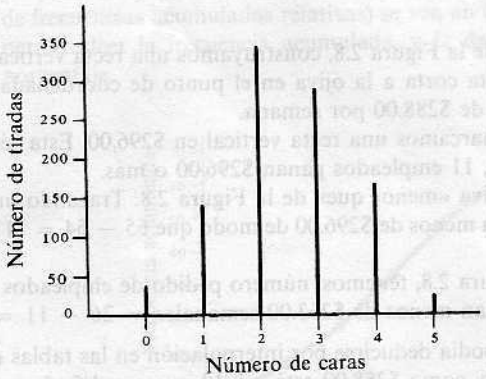


Figura 2.10.

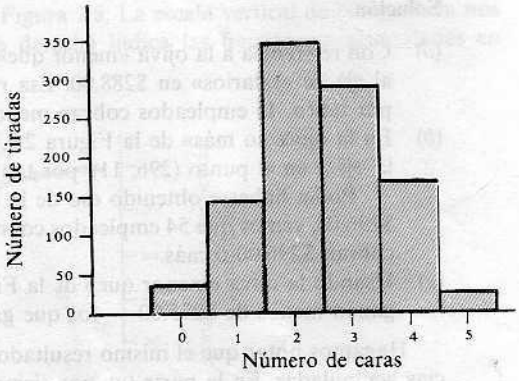


Figura 2.11.

(*b*) La Tabla 2.13 muestra simplemente una distribución de frecuencias acumuladas y una distribución de porcentajes acumulados del número de caras. Debe observarse que las entradas «menor que 1», «menor que 2», etc., podrían haberse sustituido por entradas «menor o igual que».

(*c*) El gráfico pedido puede presentarse como en la Figura 2.12 o como en la Figura 2.13.

La Figura 2.12 parece más natural para presentar datos discretos, pues el porcentaje de tiradas con menos de 2 caras ha de ser igual que para menos de 1.75, 1.56 ó 1.23 caras, de manera que debe verse el mismo porcentaje (18.2%) para esos valores (indicado por un segmento horizontal).

Tabla 2.13

Número de caras	Número de tiradas (frecuencia acumulada)	Porcentaje de número de tiradas (porcentaje de frecuencia acumulada)
Menor que 0	0	0.0
Menor que 1	38	3.8
Menor que 2	182	18.2
Menor que 3	524	52.4
Menor que 4	811	81.1
Menor que 5	975	97.5
Menor que 6	1000	100.0

La Figura 2.13 muestra el polígono de frecuencias acumuladas, u ojiva, para los datos, y esencialmente trata los datos como si fueran continuos.

Nótese que las Figuras 2.12 y 2.13 corresponden, respectivamente, a las Figura 2.10 y 2.11 de la parte (*a*).

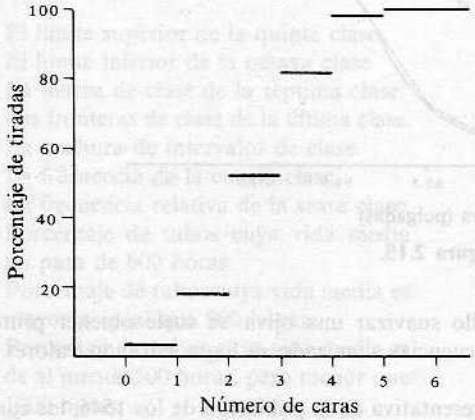


Figura 2.12.

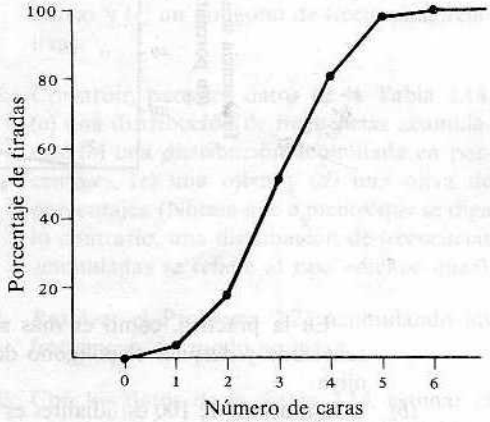


Figura 2.13.

CURVAS DE FRECUENCIA Y OJIVAS SUAVIZADAS

2.18. Los 100 estudiantes de la Universidad XYZ (Tabla 2.1) constituían en realidad una muestra de los 1546 estudiantes varones de esa universidad.

(*a*) De los datos de esa muestra, construir un polígono de frecuencias en porcentajes suavizado (curva de frecuencias) y una ojiva suavizada en porcentajes «menor que».

(*b*) De los resultados de una de las construcciones de la parte (*a*), estimar el número de estudiantes con alturas entre 65 y 70 in. ¿Qué hipótesis hay que hacer?

(*c*) ¿Cabe utilizar los resultados para estimar la proporción de varones en EE.UU. con alturas entre 65 y 70 in?

Solución

(*a*) En las Figuras 2.14 y 2.15 los gráficos discontinuos representan los polígonos de frecuencias y las ojivas, y se han obtenido de las Figuras 2.1 y 2.2, respectivamente. Las gráficas suavizadas (en trazo continuo) se obtienen aproximando los anteriores mediante curvas continuas.

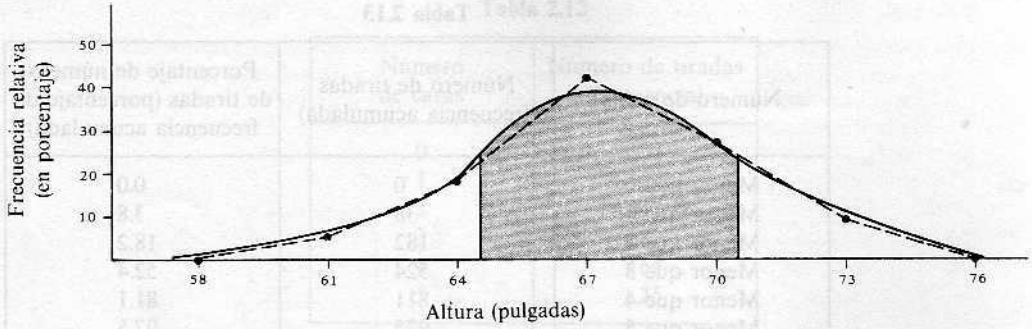


Figura 2.14.

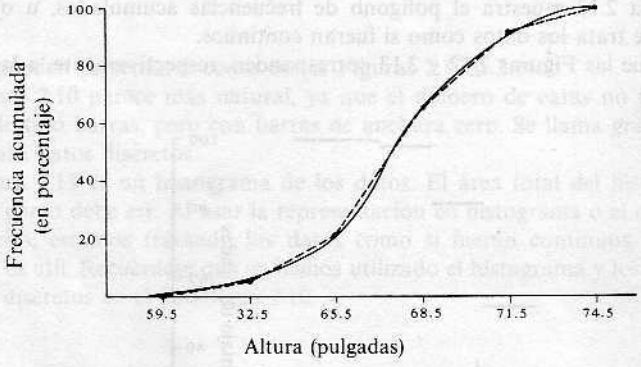


Figura 2.15.

En la práctica, como es más sencillo suavizar una ojiva, se suele obtener primero la ojiva suavizada y después el polígono de frecuencias suavizado se logra mirando valores en la citada ojiva.

(*b*) Si la muestra de 100 estudiantes es representativa de la población de los 1546, las curvas suavizadas de las Figuras 2.14 y 2.15 pueden considerarse como la curva de frecuencias en porcentajes y la ojiva de porcentajes de esa población. Esta hipótesis es correcta sólo si la muestra es aleatoria (o sea, si cada estudiante tiene la misma probabilidad de salir elegido en la muestra).

Como las alturas anotadas entre 65 y 70 in, con precisión de pulgada, en realidad representan alturas entre 64.5 y 70.5 in, el porcentaje de estudiantes en la población que tiene esas alturas se encuentra dividiendo el área sombreada de la Figura 2.14 por el área total acotada por la curva suavizada y el eje X.

Es más sencillo, no obstante, usar la Figura 2.15, de la que vemos que

$$\text{Porcentaje de estudiantes por debajo de } 70.5 \text{ in} = 82\%$$
$$\text{Porcentaje de estudiantes por debajo de } 64.5 \text{ in} = 18\%$$

luego el porcentaje con alturas entre 64.5 y 70.5 in = 82% − 18% = 64%. Así pues, el número de estudiantes de esa universidad que miden entre 65 y 70 in es el 64% de 1546 = 989.

Otra forma de decir eso es afirmar que la probabilidad de que una persona, elegida al azar de entre esas 1546, tenga altura comprendida entre 65 y 70 in, es 64%, 0,64 ó 64 de cada 100. A causa de la relación con las probabilidades (tratadas en el Capítulo 6), las curvas de frecuencia relativa se llaman *curvas de probabilidad* o *distribuciones de probabilidad.*

(*c*) Podríamos estimar la requerida proporción en un 64% (ahora con mucho más margen de error) sólo si estuviéramos convencidos de que la muestra de 100 estudiantes fuera realmente aleatoria vista desde la población masculina de EE.UU. Lo cual es improbable, porque algunos estudiantes no habrán alcanzado aún su altura tope y las generaciones jóvenes tienden a ser más altas que las anteriores, aparte de otros factores.

PROBLEMAS SUPLEMENTARIOS

2.19. (*a*) Ordenar los números 12, 56, 42, 21, 5, 18, 10, 3, 61, 34, 65 y 24 y (*b*) hallar su rango.

2.20. La Tabla 2.14 muestra una distribución de frecuencias de las vidas medias de 400 válvulas de radio probadas en la empresa L&M. Determinar de esa tabla:

(*a*) El límite superior de la quinta clase.
(*b*) El límite inferior de la octava clase.
(*c*) La marca de clase de la séptima clase.
(*d*) Las fronteras de clase de la última clase.
(*e*) La anchura de intervalos de clase.
(*f*) La frecuencia de la cuarta clase.
(*g*) La frecuencia relativa de la sexta clase.
(*h*) Porcentaje de tubos cuya vida media no pasa de 600 horas.
(*i*) Porcentaje de tubos cuya vida media es mayor o igual que 900 horas.
(*j*) Porcentaje de tubos cuya vida media es de al menos 500 horas, pero menor que 1000 horas.

Tabla 2.14

Vida media (horas)	Número de tubos
300-399	14
400-499	46
500-599	58
600-699	76
700-799	68
800-899	62
900-999	48
1000-1099	22
1100-1199	6
	Total 400

2.21. Construir: (*a*) un histograma y (*b*) un polígono de frecuencias correspondientes a la distribución de frecuencias de la Tabla 2.14.

2.22. Para los datos de la Tabla 2.14 (Prob. 2.20), construir: (*a*) una distribución de frecuencias relativas, (*b*) un histograma de frecuencias relativas y (*c*) un polígono de frecuencias relativas.

2.23. Construir, para los datos de la Tabla 2.14, (*a*) una distribución de frecuencias acumuladas, (*b*) una distribución acumulada en porcentajes, (*c*) una ojiva y (*d*) una ojiva de porcentajes. (Nótese que a menos que se diga lo contrario, una distribución de frecuencias acumuladas se refiere al tipo «menor que»).

2.24. Resolver el Problema 2.23 acumulando las frecuencias del modo «o más».

2.25. Con los datos de la Tabla 2.14, estimar el porcentaje de tubos con vida media: (*a*) menor que 560 horas, (*b*) 970 horas o más y (*c*) entre 620 y 890 horas.

2.26. Los diámetros internos de los tubos fabricados por una empresa se miden con precisión de milésima de pulgada. Si las marcas de clase de una distribución de frecuencias de esos diámetros vienen dadas por 0.321, 0.324, 0.327, 0.330, 0.333 y 0.336, hallar: (*a*) la anchura del intervalo de clase, (*b*) las fronteras de clase y (*c*) los límites de clase.

2.27. La tabla adjunta muestra los diámetros en centímetros de una muestra de 60 bolas de cojinete manufacturadas por una fábrica. Construir una distribución de frecuencias con intervalos de clase apropiados.

1.738	1.729	1.743	1.740	1.736	1.741
1.735	1.731	1.726	1.737	1.728	1.737
1.736	1.735	1.724	1.733	1.742	1.736
1.739	1.735	1.745	1.736	1.742	1.740
1.728	1.738	1.725	1.733	1.734	1.732
1.733	1.730	1.732	1.730	1.739	1.734
1.738	1.739	1.727	1.735	1.735	1.732
1.735	1.727	1.734	1.732	1.736	1.741
1.736	1.744	1.732	1.737	1.731	1.746
1.735	1.735	1.729	1.734	1.730	1.740

2.28. Para los datos del Problema 2.27, construir: (*a*) un histograma, (*b*) un polígono de frecuencias, (*c*) una distribución de frecuencias relativas, (*d*) un histograma de frecuencias relativas, (*e*) un polígono de frecuencias relativas, (*f*) una distribución de frecuencias acumuladas, (*g*) una distribución acumulada en porcentajes, (*h*) una ojiva e (*i*) una ojiva de porcentajes.

2.29. Determinar, a partir de los resultados del Problema 2.28, el porcentaje de bolas con diámetros: (*a*) mayores que 1.732 cm, (*b*) no mayor que 1.736 cm y (*c*) entre 1.730 y 1.738 cm. Comparar los resultados con los obtenidos directamente de los datos del Problema 2.27.

2.30. Repetir el Problema 2.28 para los datos del Problema 2.20.

2.31. La Tabla 2.15 muestra la distribución de porcentajes de ventas totales para plantaciones de tipo familiar en EE.UU. en 1982. Usando esa tabla, responder las siguientes cuestiones:

(*a*) ¿Cuál es la anchura del segundo intervalo de clase? ¿Y del séptimo?

(*b*) Cuántos tamaños diferentes de intervalos de clase hay?

(*c*) ¿Cuántos intervalos de clase abiertos hay?

(*d*) ¿Cómo habría que escribir el primer intervalo de clase para que su anchura sea igual a la del segundo?

(*e*) ¿Cuál es la marca de clase del segundo intervalo de clase? ¿Y del séptimo?

(*f*) ¿Cuáles son las fronteras de clase del cuarto intervalo de clase?

(*g*) ¿Qué porcentaje de las plantaciones tuvo ventas de $20,000 o más? ¿Y por debajo de $10,000?

(*h*) ¿Qué porcentaje logró ventas de al menos $10,000, pero no mayores que $40,000?

(*i*) ¿Qué porcentaje tuvo ventas entre $15,000 y $25,000? ¿Qué hipótesis se han hecho en ese cálculo?

(*j*) ¿Por qué los porcentajes de la Tabla 2.15 no suman 100%?

Tabla 2.15

Ventas (dólares)	Explotaciones (%)
Menos de 2,500	25.9
2,500-4,999	13.2
5,000-9,999	13.0
10,000-19,999	11.7
20,000-39,999	11.0
40,000-99,999	14.4
100,000-249,999	8.5
250,000-499,999	1.8
500,000 o más	0.6

2.32. (*a*) ¿Por qué es imposible construir un histograma de porcentajes o un polígono de frecuencias para la distribución de la Tabla 2.15?

(*b*) ¿Cómo modificaría la distribución para que pudieran construirse ambos?

(*c*) Llevar a cabo la modificación y la construcción.

2.33. El número total de plantaciones en la distribución de la Tabla 2.15 es 1,945,000. A partir de ese dato, determinar el número de plantaciones con ventas: (*a*) superiores a $40,000, (*b*) menores que $40,000 y (*c*) entre $30,000 y $50,000.

2.34. (*a*) Construir un polígono de frecuencias en porcentajes suavizado y una ojiva en porcentajes suavizada para los datos de la Tabla 2.14.

(*b*) Estimar con ellos la probabilidad de que un tubo se deteriore antes de 600 horas.

(*c*) Discutir el riesgo del fabricante al garantizar los tubos por 425 horas. Idem con 875 horas.

(*d*) Si el fabricante ofrece una garantía de 90 días para la devolución del importe de un tubo, ¿cuál es la probabilidad de que devuelva el importe, supuesto que el tubo esté en uso 4 horas diarias? ¿Y con 8 horas diarias?

2.35. (*a*) Lanzar 4 monedas 50 veces y anotar el número de caras en cada ocasión.

(*b*) Construir una distribución de frecuencias que indique el número de veces que se han obtenido 0, 1, 2, 3 y 4 caras.

(*c*) Construir una distribución de porcentajes correspondiente a la parte (*b*).

(*d*) Comparar el porcentaje obtenido en (*c*) con los teóricos 6,25%, 25%, 37.5%, 25% y 6.25% (proporcional a 1, 4, 6, 4 y 1) deducidos por las leyes de las probabilidades.

(*e*) Representar las distribuciones de las partes (*b*) y (*c*).

(*f*) Construir una ojiva de porcentajes para los datos.

2.36. Repetir el problema anterior con otros 50 lanzamientos y véase si el experimento está más de acuerdo con lo esperado teóricamente. Si no, dar posibles razones para tales discrepancias.

CAPITULO 3

Media, mediana, moda y otras medidas de tendencia central

NOTACION DE INDICES

Denotemos por X_j (léase «X sub j») cualquiera de los N *valores* $X_1, X_2, X_3, \ldots, X_N$ que toma una variable X. La letra j en X_j, que puede valer 1, 2, 3, ..., N se llama *subíndice*. Es claro que podíamos haber empleado cualquier otra letra en vez de j, por ejemplo, i, k, p, q o s.

NOTACION DE SUMA

El símbolo $\sum_{j=1}^{N} X_j$ denotará la suma de todos los X_j desde $j = 1$ a $j = N$; por definición,

$$\sum_{j=1}^{N} X_j = X_1 + X_2 + X_3 + \cdots + X_N$$

Cuando no ocasione confusión, denotaremos esa suma simplemente por $\sum X$, $\sum X_j$ o $\sum_j X_j$. El símbolo $\sum$ es la letra griega *sigma* mayúscula, que denota suma.

EJEMPLO 1. $\sum_{j=1}^{N} X_j Y_j = X_1 Y_1 + X_2 Y_2 + X_3 Y_3 + \cdots + X_N Y_N$.

EJEMPLO 2. $\sum_{j=1}^{N} aX_j = aX_1 + aX_2 + \cdots + zX_N = a(X_1 + X_2 + \cdots + X_N) = a \sum_{j=1}^{N} X_j$, donde a es una constante. Más sencillamente, $\sum aX = a \sum X$.

EJEMPLO 3. Si a, b, c son constantes, entonces $\sum (aX + bY - cZ) = a \sum X + b \sum Y - c \sum Z$. Véase Problema 3.3.

PROMEDIOS O MEDIDAS DE TENDENCIA CENTRAL

Un *promedio* es un valor típico o representativo de un conjunto de datos. Como tales valores suelen situarse hacia el centro del conjunto de datos ordenados por magnitud, los promedios se conocen como *medidas de tendencia central*.

Se definen varios tipos, siendo los más comunes la *media aritmética*, la *mediana*, la *moda*, la *media geométrica* y la *media armónica*. Cada una tiene ventajas y desventajas, según los datos y el objetivo perseguido.

LA MEDIA ARITMETICA

La *media aritmética*, o simplemente *media*, de un conjunto de N números $X_1, X_2, X_3, \ldots, X_N$ se denota por $\bar{X}$ (léase «X barra») y se define por

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\sum_{j=1}^{N} X_j}{N} = \frac{\sum X}{N} \tag{1}$$

EJEMPLO 4. La media aritmética de los números 8, 3, 5, 12 y 10 es

$$\bar{X} = \frac{8 + 3 + 5 + 12 + 10}{5} = \frac{38}{5} = 7.6$$

Si los números $X_1, X_2, \ldots, X_K$ ocurren $f_1, f_2, \ldots, f_k$ veces, respectivamente (o sea, con frecuencias $f_1, f_2, \ldots, f_k$), la media aritmética es

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + \cdots + f_KX_K}{f_1 + f_2 + \cdots + f_K} = \frac{\sum_{j=1}^{K} f_jX_j}{\sum_{j=1}^{K} f_j} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{N} \tag{2}$$

donde $N = \sum f$ es la *frecuencia total* (o sea, el número total de casos).

EJEMPLO 5. Si 5, 8, 6 y 2 ocurren con frecuencias 3, 2, 4 y 1, respectivamente, su media aritmética es

$$\bar{X} = \frac{(3)(5) + (2)(8) + (4)(6) + (1)(2)}{3 + 2 + 4 + 1} = \frac{15 + 16 + 24 + 2}{10} = 5.7$$

LA MEDIA ARITMETICA PONDERADA

A veces asociamos con los números $X_1, X_2, \ldots, X_K$ ciertos *factores peso* (o *pesos*) $w_1, w_2, \ldots, w_K$, dependientes de la relevancia asignada a cada número. En tal caso,

$$\bar{X} = \frac{w_1X_1 + w_2X_2 + \cdots + w_KX_k}{w_1 + w_2 + \cdots + w_K} = \frac{\sum wX}{\sum w} \tag{3}$$

se llama la *media aritmética ponderada* con pesos $f_1, f_2, \ldots, f_K$.

EJEMPLO 6. Si el examen final de un curso cuenta tres veces más que una evaluación parcial, y un estudiante tiene calificación 85 en el examen final y 70 y 90 en los dos parciales, la calificación media es

$$\bar{X} = \frac{(1)(70) + (1)(90) + (3)(85)}{1 + 1 + 3} = \frac{415}{5} = 83$$

PROPIEDADES DE LA MEDIA ARITMETICA

1. La suma algebraica de las desviaciones de un conjunto de números respecto de su media aritmética es cero.

EJEMPLO 7. Las desviaciones de los números 8, 3, 5, 12 y 10 respecto de su media aritmética 7.6 son $8 - 7.6$, $3 - 7.6$, $5 - 7.6$, $12 - 7.6$ y $10 - 7.6$, o sea 0.4, -4.6, -2.6, 4.4 y 2.4, con suma algebraica $0.4 - 4.6 - 2.6 + 4.4 + 2.4 = 0$.

2. La suma de los cuadrados de las desviaciones de un conjunto de números X_j respecto de un cierto número a es mínima si y sólo si $a = \bar{X}$ (véase Prob. 4.27).
3. Si f_1 números tienen media m_1, f_2 números tiene media m_2, ..., f_K números tienen media m_K, entonces la media de todos los números es

$$\bar{X} = \frac{f_1 m_1 + f_2 m_2 + \cdots + f_K m_K}{f_1 + f_2 + \cdots + f_K} \tag{4}$$

es decir, una media aritmética ponderada de todas las medias (véase Prob. 3.12).

4. Si A es una *media aritmética supuesta o conjeturada* (que puede ser cualquier número) y si $d_j = X_j - A$ son las desviaciones de X_j respecto de A, las ecuaciones (1) y (2) se convierten, respectivamente, en

$$\bar{X} = A + \frac{\sum_{j=1}^{N} d_j}{N} = A + \frac{\sum d}{N} \tag{5}$$

$$\bar{X} = A + \frac{\sum_{j=1}^{K} f_j d_j}{\sum_{j=1}^{K} f_j} = A + \frac{\sum fd}{N} \tag{6}$$

donde $N = \sum_{j=1}^{K} f_j = \sum f$. Nótese que las fórmulas (5) y (6) se resumen en $\bar{X} = A + \bar{d}$ (véase Prob. 3.18).

CALCULO DE LA MEDIA ARITMETICA PARA DATOS AGRUPADOS

Cuando los datos se presentan en una distribución de frecuencias, todos los valores que caen dentro de un intervalo de clase dado se consideran iguales a la marca de clase, o punto medio, del

intervalo. Las fórmulas (2) y (6) son válidas para tales datos agrupados si interpretamos X_j como la marca de clase, f_j como su correspondiente frecuencia de clase, A como cualquier marca de clase conjeturada y $d_j = X_j - A$ como las desviaciones de X_j respecto de A.

Los cálculos con (2) y (6) se llaman *métodos largos y cortos*, respectivamente (véanse Probs. 3.15 y 3.20).

Si todos los intervalos de clase tienen idéntica anchura c, las desviaciones $d_j = X_j - A$ pueden expresarse como cu_j, donde u_j pueden ser $0, \pm 1, \pm 2, \pm 3, \ldots$, y la fórmula (6) se convierte en

$$\bar{X} = A + \left(\frac{\sum_{j=1}^{K} f_j u_j}{N}\right) = A + \left(\frac{\sum fu}{N}\right)c \tag{7}$$

que es equivalente a la ecuación $\bar{X} = A + c\bar{u}$ (véase Prob. 3.21). Esto se conoce como *método de compilación* para calcular la media. Es un método muy breve y debe usarse siempre para datos agrupados con intervalos de clase de anchuras iguales (véanse Probs. 3.22 y 3.23). Nótese que en el método de compilación los valores de la variable X *se transforman* en los valores de la variable u de acuerdo con $X = A + cu$.

LA MEDIANA

La *mediana* de un conjunto de números ordenados en magnitud es o el valor central o la media de los dos valores centrales.

EJEMPLO 8. El conjunto de números 3, 4, 4, 5, 6, 8, 8, 8 y 10 tiene mediana 6.

EJEMPLO 9. El conjunto de números 5, 5, 7, 9, 11, 12, 15 y 18 tiene mediana $\frac{1}{2}(9 + 11) = 10$.

Para datos agrupados, la mediana obtenida por interpolación viene dada por

$$\text{Mediana} = L_1 + \left(\frac{\frac{N}{2} - (\sum f)_1}{f_{\text{mediana}}}\right)c \tag{8}$$

donde:

L_1 = frontera inferior de la clase de la mediana.
N = número de datos (frecuencia total).
$(\sum f)_1$ = suma de frecuencias de las clases inferiores a la de la mediana.
f_{mediana} = frecuencia de la clase de la mediana.
c = anchura del intervalo de clase de la mediana.

Geométricamente la mediana es el valor de X (abscisa) que corresponde a la recta vertical que divide un histograma en dos partes de igual área. Ese valor de X se suele denotar por $\tilde{X}$.

LA MODA

La *moda* de un conjunto de números es el valor que ocurre con mayor frecuencia; es decir, el valor más frecuente. La moda puede no existir, e incluso no ser única en caso de existir.

EJEMPLO 10. El conjunto 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12 y 18 tiene moda 9.

EJEMPLO 11. El conjunto 3, 5, 8, 10, 12, 15 y 16 no tiene moda.

EJEMPLO 12. El conjunto 2, 3, 4, 4, 4, 5, 5, 7, 7, 7 y 9 tiene dos modas, 4 y 7, y se llama *bimodal*.

Una distribución con moda única se dice *unimodal*.

En el caso de datos agrupados donde se haya construido una curva de frecuencias para ajustar los datos, la moda será el valor (o valores) de X correspondiente al máximo (o máximos) de la curva. Ese valor de X se denota por $\hat{X}$.

La moda puede deducirse de una distribución de frecuencias o de un histograma a partir de la fórmula

$$\text{Moda} = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)c \tag{9}$$

donde:

L_1 = frontera inferior de la clase modal (clase que contiene a la moda).
Δ_1 = exceso de la frecuencia modal sobre la de la clase inferior inmediata.
Δ_2 = exceso de la frecuencia modal sobre la clase superior inmediata.
c = anchura del intervalo de clase modal.

RELACION EMPIRICA ENTRE MEDIA, MEDIANA Y MODA

Para curvas de frecuencia unimodales que sean poco asimétricas tenemos la siguiente relación empírica

$$\text{Media} - \text{moda} = 3(\text{media} - \text{mediana}) \tag{10}$$

Las Figuras 3.1 y 3.2 muestran las posiciones relativas de la media, la mediana y la moda para curvas de frecuencia asimétricas a derecha e izquierda, respectivamente. Para curvas simétricas, los tres valores coinciden.

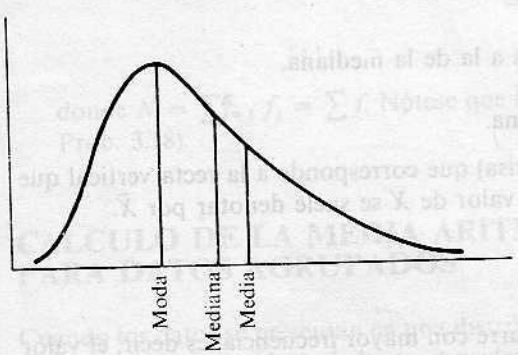


Figura 3.1.

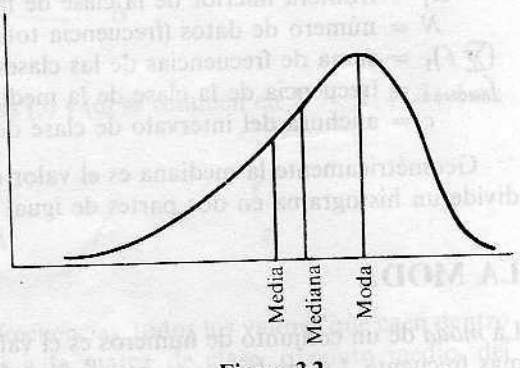


Figura 3.2.

LA MEDIA GEOMETRICA G

La *media geométrica* G de un conjunto de N números positivos $X_1, X_2, X_3, \ldots, X_N$ es la raíz N-ésima del producto de esos números:

$$G = \sqrt[N]{X_1 X_2 X_3 \cdots X_N} \tag{11}$$

EJEMPLO 13. La media geométrica de 2, 4 y 8 es $G = \sqrt[3]{(2)(4)(8)} = \sqrt[3]{64} = 4$.

Podemos calcular G por logaritmos (véase Prob. 3.35) o con una calculadora. Para la media geométrica de datos agrupados, véanse Problemas 3.36 y 3.91.

LA MEDIA ARMONICA H

La *media armónica* H de un conjunto de números $X_1, X_2, X_3, \ldots, X_N$ es el recíproco de la media aritmética de los recíprocos de esos números:

$$H = \frac{1}{\frac{1}{N}\sum_{j=1}^{N}\frac{1}{X_j}} = \frac{N}{\sum\frac{1}{X}} \tag{12}$$

En la práctica es más fácil recordar que

$$\frac{1}{H} = \frac{\sum\frac{1}{X}}{N} = \frac{1}{N}\sum\frac{1}{X} \tag{13}$$

EJEMPLO 14. La media armónica de los números 2, 4 y 8 es

$$H = \frac{3}{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}} = \frac{3}{\frac{7}{8}} = 3.43$$

Para la media armónica de datos agrupados, véanse Problemas 3.99 y 3.100.

RELACION ENTRE LAS MEDIAS ARITMETICA, GEOMETRICA Y ARMONICA

La media geométrica de una colección de números positivos $X_1, X_2, \ldots, X_N$ es menor o igual que su media aritmética, pero mayor o igual que su media armónica. En símbolos,

$$H \leqslant G \leqslant \bar{X} \tag{14}$$

La igualdad ocurre si y sólo si todos los números $X_1, X_2, \ldots, X_N$ son idénticos.

EJEMPLO 15. El conjunto 2, 4, 8 tiene media aritmética 4.67, media geométrica 4 y media armónica 3.43.

LA MEDIA CUADRATICA (MQ)

La *media cuadrática* (MQ) de un conjunto de números $X_1, X_2, ..., X_N$ se suele denotar por $\sqrt{\bar{X}^2}$ y se define como

$$\text{MQ} = \sqrt{\bar{X}^2} = \sqrt{\frac{\sum_{j=1}^{N} X_j^2}{N}} = \sqrt{\frac{\sum X^2}{N}} \tag{15}$$

Este tipo de promedio se utiliza con frecuencia en las aplicaciones físicas.

EJEMPLO 16. La media cuadrática del conjunto 1, 3, 4, 5 y 7 es

$$\sqrt{\frac{1^2 + 3^2 + 4^2 + 5^2 + 7^2}{5}} = \sqrt{20} = 4.47$$

CUARTILES, DECILES Y PERCENTILES

Si un conjunto de datos está ordenado por magnitud, el valor central (o la media de los dos centrales) que divide al conjunto en dos mitades iguales, es la mediana. Extendiendo esa idea, podemos pensar en aquellos valores que dividen al conjunto en cuatro partes iguales. Esos valores, denotados Q_1, Q_2 y Q_3, se llaman primer, segundo y tercer *cuartiles*, respectivamente. El Q_2 coincide con la mediana.

Análogamente, los valores que dividen a los datos en 10 partes iguales se llaman *deciles*, y se denotan $D_1, D_2, ..., D_9$, mientras que los valores que los dividen en 100 partes iguales se llaman *percentiles*, denotados por $P_1, P_2, ..., P_{99}$. El 5.° decil y el 50.° percentil coinciden con la mediana. Los 25.° y 75.° percentiles coinciden con el primer y tercer cuartiles.

Colectivamente, cuartiles, deciles y percentiles se denominan *cuantiles*. Para su cálculo con datos agrupados, véanse Problemas 3.44 al 3.46.

PROBLEMAS RESUELTOS

NOTACION DE SUMA

3.1. Escribir explícitos los términos en cada suma:

(a) $\sum_{j=1}^{6} X_j$ (c) $\sum_{j=1}^{N} a$ (e) $\sum_{j=1}^{3} (X_j - a)$

(b) $\sum_{j=1}^{4} (Y_j - 3)^2$ (d) $\sum_{k=1}^{5} f_k X_k$

Solución

(a) $X_1 + X_2 + X_3 + X_4 + X_5 + X_6$

(b) $(Y_1 - 3)^2 + (Y_2 - 3)^2 + (Y_3 - 3)^2 + (Y_4 - 3)^2$

(*c*) $a + a + a + \cdots + a = Na$

(*d*) $f_1X_1 + f_2X_2 + f_3X_3 + f_4X_4 + f_5X_5$

(*e*) $(X_1 - a) + (X_2 - a) + (X_3 - a) = X_1 + X_2 + X_3 - 3a$

3.2. Expresar cada suma en notación abreviada de suma:

(*a*) $X_1^2 + X_2^2 + X_3^2 + \cdots + X_{10}^2$

(*b*) $(X_1 + Y_1) + (X_2 + Y_2) + \cdots + (X_8 + Y_8)$

(*c*) $f_1X_1^3 + f_2X_2^3 + \cdots + f_{20}X_{20}^3$

(*d*) $a_1b_1 + a_2b_2 + a_3b_3 + \cdots + a_Nb_N$

(*e*) $f_1X_1Y_1 + f_2X_2Y_2 + f_3X_3Y_3 + f_4X_4Y_4$

Solución

(*a*) $\sum_{j=1}^{10} X_j^2$ (*c*) $\sum_{j=1}^{20} f_jX_j^3$ (*e*) $\sum_{j=1}^{4} f_jX_jY_j$

(*b*) $\sum_{j=1}^{8} (X_j + Y_j)$ (*d*) $\sum_{j=1}^{N} a_jb_j$

3.3. Probar que $\sum_{j=1}^{N} (aX_j + bY_j - cZ_j) = a\sum_{j=1}^{N} X_j + b\sum_{j=1}^{N} Y_j - c\sum_{j=1}^{N} Z_j$, donde a, b y c son constantes.

Solución

$$\begin{aligned}\sum_{j=1}^{N}(aX_j+bY_j-cZ_j) &= (aX_1+bY_1-cZ_1)+(aX_2+bY_2-cZ_2)+\cdots+(aX_N+bY_N-cZ_N)\\ &= (aX_1+aX_2+\cdots+aX_N)+(bY_1+bY_2+\cdots+bY_N)-(cZ_1+cZ_2+\cdots+cZ_N)\\ &= a(X_1+X_2+\cdots+X_N)+b(Y_1+Y_2+\cdots+Y_N)-c(Z_1+Z_2+\cdots+Z_N)\\ &= a\sum_{j=1}^{N}X_j+b\sum_{j=1}^{N}Y_j-c\sum_{j=1}^{N}Z_j\end{aligned}$$

o más abreviado, $\sum (aX + bY - cZ) = a\sum X + b\sum Y - c\sum Z$.

3.4. Dos variables X e Y toman los valores $X_1 = 2$, $X_2 = -5$, $X_3 = 4$, $X_4 = -8$ e $Y_1 = -3$, $Y_2 = -8$, $Y_3 = 10$, $Y_4 = 6$, respectivamente. Calcular: (*a*) $\sum X$, (*b*) $\sum Y$, (*c*) $\sum XY$, (*d*) $\sum X^2$, (*e*) $\sum Y^2$, (*f*) $(\sum X)(\sum Y)$, (*g*) $\sum XY^2$ y (*h*) $\sum (X + Y)(X - Y)$.

Solución

Nótese que en cada caso el subíndice j de X e Y ha sido omitido, y la $\sum$ se entiende como $\sum_{j=1}^{4}$. Así pues, $\sum X$, por ejemplo, es una abreviatura para $\sum_{j=1}^{4} X_j$.

(*a*) $\sum X = (2) + (-5) + (4) + (-8) = 2 - 5 + 4 - 8 = -7$

(*b*) $\sum Y = (-3) + (-8) + (10) + (6) = -3 - 8 + 10 + 6 = 5$

(*c*) $\sum XY = (2)(-3) + (-5)(-8) + (4)(10) + (-8)(6) = -6 + 40 + 40 - 48 = 26$

(*d*) $\sum X^2 = (2)^2 + (-5)^2 + (4)^2 + (-8)^2 = 4 + 25 + 16 + 64 = 109$

(*e*) $\sum Y^2 = (-3)^2 + (-8)^2 + (10)^2 + (6)^2 = 9 + 64 + 100 + 36 = 209$

(*f*) $(\sum X)(\sum Y) = (-7)(5) = -35$, usando las partes (*a*) y (*b*). Nótese que $(\sum X)(\sum Y) \neq \sum XY$.

(*g*) $\sum XY^2 = (2)(-3)^2 + (-5)(-8)^2 + (4)(10)^2 + (-8)(6)^2 = -190$

(*h*) $\sum (X + Y)(X - Y) = \sum (X^2 - Y^2) = \sum X^2 - \sum Y^2 = 109 - 209 = -100$, usando las partes (*d*) y (*e*).

3.5. Si $\sum_{j=1}^{6} X_j = -4$ y $\sum_{j=1}^{6} X_j^2 = 10$, calcular: (*a*) $\sum_{j=1}^{6} (2X_j + 3)$, (*b*) $\sum_{j=1}^{6} X_j(X_j - 1)$ y (*c*) $\sum_{j=1}^{6} (X_j - 5)^2$.

Solución

(*a*) $$\sum_{j=1}^{6} (2X_j + 3) = \sum_{j=1}^{6} 2X_j + \sum_{j=1}^{6} 3 = 2\sum_{j=1}^{6} X_j + (6)(3) = 2(-4) + 18 = 10$$

(*b*) $$\sum_{j=1}^{6} X_j(X_j - 1) = \sum_{j=1}^{6} (X_j^2 - X_j) = \sum_{j=1}^{6} X_j^2 - \sum_{j=1}^{6} X_j = 10 - (-4) = 14$$

(*c*) $$\sum_{j=1}^{6} (X_j - 5)^2 = \sum_{j=1}^{6} (X_j^2 - 10X_j + 25) = \sum_{j=1}^{6} X_j^2 - 10\sum_{j=1}^{6} X_j + 25(6) = 10 - 10(-4) + 25(6) = 200$$

Si se desea, puede omitirse el subíndice j y usar $\sum$ en lugar de $\sum_{j=1}^{6}$ siempre que se manejen con soltura estas abreviaturas.

LA MEDIA ARITMETICA

3.6. Las notas de un estudiante en seis exámenes fueron 84, 91, 72, 68, 87 y 78. Hallar la media aritmética.

Solución

$$\bar{X} = \frac{\sum X}{N} = \frac{84 + 91 + 72 + 68 + 87 + 78}{6} = \frac{480}{6} = 80$$

A menudo se usa el término *promedio* como sinónimo de *media aritmética*. Estrictamente hablando, sin embargo, esto es incorrecto, porque hay otros promedios además de la media aritmética.

3.7. Diez medidas del diámetro de un cilindro fueron anotadas por un científico como 3.88, 4.09, 3.92, 3.97, 4.02, 3.95, 4.03, 3.92, 3.98 y 4.06 centímetros (cm). Hallar la media aritmética de tales medidas.

Solución

$$\bar{X} = \frac{\sum X}{N} = \frac{3.88 + 4.09 + 3.92 + 3.97 + 4.02 + 3.95 + 4.03 + 3.92 + 3.98 + 4.06}{10} = \frac{39.82}{10} = 3.98 \text{ cm}$$

3.8. Los salarios anuales de 4 individuos son \$15,000, \$16,000, \$16,500 y \$40,000.
(*a*) Hallar su media aritmética.
(*b*) ¿Puede decirse que ese promedio es *típico* de dichos salarios?

Solución

(*a*) Supuesto que todas las cifras eran significativas en los salarios anotados,

$$\bar{X} = \frac{\$15{,}000 + \$16{,}000 + \$16{,}500 + \$40{,}000}{4} = \frac{\$87{,}500}{4} = \$21{,}875$$

(*b*) La media \$21,875 no es ciertamente típica de esos salarios, y presentarla como un promedio sin más comentarios sería muy engañoso. Una gran desventaja de la media es que se ve muy afectada por valores extremos.

3.9. Hallar la media aritmética de los números 5, 3, 6, 5, 4, 5, 2, 8, 6, 5, 4, 8, 3, 4, 5, 4, 8, 2, 5 y 4.

Solución

Primer método

$$\bar{X} = \frac{\sum X}{N} = \frac{5+3+6+5+4+5+2+8+6+5+4+8+3+4+5+4+8+2+5+4}{20} = \frac{96}{20} = 4.8$$

Segundo método

Hay 6 cincos, 2 treses, 5 cuatros, 2 doses y 3 ochos. Luego

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{N} = \frac{(6)(5) + (2)(3) + (2)(6) + (5)(4) + (2)(2) + (3)(8)}{6 + 2 + 2 + 5 + 2 + 3} = \frac{96}{20} = 4.8$$

3.10. De entre 100 números, 20 son cuatros, 40 son cincos, 30 son seises y los restantes sietes. Hallar su media aritmética.

Solución

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{N} = \frac{(20)(4) + (40)(5) + (30)(6) + (10)(7)}{100} = \frac{530}{100} = 5.30$$

3.11. Las calificaciones finales de un estudiante en cuatro asignaturas fueron 82, 86, 90 y 70. Si los respectivos créditos otorgados a esos cursos son 3, 5, 3 y 1, determinar una calificación media apropiada.

Solución

Usamos una media aritmética ponderada, con pesos dados por los créditos otorgados. Así pues,

$$\bar{X} = \frac{\sum wX}{\sum w} = \frac{(3)(82) + (5)(86) + (3)(90) + (1)(70)}{3 + 5 + 3 + 1} = 85$$

3.12. De los 80 empleados de una empresa, 60 cobran \$7,00 a la hora y el resto \$4,00 a la hora.

(*a*) Hallar cuánto cobran de media por hora.
(*b*) ¿Sería idéntica la respuesta si los 60 cobraran de media \$4,00 a la hora? Demuestre su respuesta.
(*c*) ¿Cree que la media es representativa?

Solución

(*a*)
$$\bar{X} = \frac{\sum fX}{N} = \frac{(60)(\$7{,}00) + (20)(\$4{,}00)}{60 + 20} = \frac{\$500.00}{80} = \$6.25$$

(*b*) Sí, el resultado es el mismo. Para verlo, supongamos que f_1 números tienen media m_1 y que f_2 números tienen media m_2. Debemos probar que la media de todos esos números es

$$\bar{X} = \frac{f_1 m_1 + f_2 m_2}{f_1 + f_2}$$

Sea M_1 la suma de los f_1 números y M_2 la de los otros f_2. Entonces, por definición de media aritmética,

$$m_1 = \frac{M_1}{f_1} \qquad m_2 = \frac{M_2}{f_2}$$

o sea $M_1 = f_1 m_1$ y $M_2 = f_2 m_2$. Cuando los $(f_1 + f_2)$ números suman $(M_1 + M_2)$, la media aritmética de todos los números es

$$\bar{X} = \frac{M_1 + M_2}{f_1 + f_2} = \frac{f_1 m_1 + f_2 m_2}{f_1 + f_2}$$

como habíamos anunciado. El resultado se generaliza con facilidad.

(*c*) Podemos decir que \$6.25 es representativo en el sentido de que la mayoría de los trabajadores cobra \$7.00 a la hora, que no difiere mucho de \$6.25. Hay que recordar que siempre que resumimos datos numéricos en un solo número (un promedio, por ejemplo), estamos abocados a cometer algun error. No obstante, el resultado no es tan engañoso como el del Problema 3.8.

Realmente, para pisar suelo firme, es preciso dar alguna estimación de la «dispersión» o «variación» de los datos respecto de la media (u otro promedio). Eso se llama *dispersión* de los datos. En el Capítulo 4 veremos diversas medidas de la dispersión.

3.13. Cuatro grupos de estudiantes, consistentes en 15, 20, 10 y 18 individuos, dieron pesos medios de 162, 148, 153 y 140 lb, respectivamente. Hallar el peso medio de todos esos estudiantes.

Solución

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{(15)(162) + (20)(148) + (10)(153) + (18)(140)}{15 + 20 + 10 + 18} = 150 \text{ lb}$$

3.14. Si los ingresos medios anuales de los trabajadores agrícolas y no agrícolas en EE.UU. son \$9000 y \$15,000, respectivamente, ¿la media anual de todos ellos sería $\frac{1}{2}(\$9000 + \$15{,}000) = \$12{,}000$?

Solución

Sería \$12,000 sólo si hubiera tantos trabajadores de un tipo como de otro. Para hallar la verdadera media sería necesario conocer los números relativos de trabajadores de cada tipo. Si, por ejemplo, hay uno agrícola por cada diez no agrícolas, la media será

$$\bar{X} = \frac{(1)(\$9000) + (11)(\$15{,}000)}{1 + 11} = \$14{,}500$$

Es una media aritmética ponderada.

3.15. Usar la distribución de frecuencias de alturas en la Tabla 2.1 para hallar la altura media de esos 100 estudiantes.

Solución

La Tabla 3.1 indica cómo se hace. Nótese que todos los estudiantes que tienen entre 60 y 62 in, o entre 63 y 65, etc., se consideran como de 61 in, 64 in, etc. El problema se reduce entonces a hallar la altura media de 100 estudiantes, de los cuales 5 miden 61 in, 18 miden 64 in, etc.

Los cálculos exigidos pueden ser tediosos, sobre todo para casos de números grandes y con muchas clases. Hay técnicas que acortan el trabajo; véanse, por ejemplo, los Problemas 3.20 y 3.22.

Tabla 3.1

Altura (in)	Marca de clase (X)	Frecuencia (f)	fX
60-62	61	5	305
63-65	64	18	1152
66-68	67	42	2814
69-71	70	27	1890
72-74	73	8	584
		$N=\sum f=100$	$\sum fX=6745$

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{N} = \frac{6745}{100} = 67.45 \text{ in}$$

PROPIEDADES DE LA MEDIA ARITMETICA

3.16. Probar que la suma de desviaciones de $X_1, X_2, \ldots, X_N$ respecto de su media $\bar{X}$ es cero.

Solución

Sean $d_1 = X_1 - \bar{X}$, $d_2 = X_2 - \bar{X}$, ..., $d_N = X_N - \bar{X}$ las desviaciones de $X_1, X_2, \ldots, X_N$ respecto de su media $\bar{X}$. Entonces

$$\text{Suma de desviaciones} = \sum d_j = \sum (X_j - \bar{X}) = \sum X_j - N\bar{X}$$

$$= \sum X_j - N\left(\frac{\sum X_j}{N}\right) = \sum X_j - \sum X_j = 0$$

donde hemos usado $\sum$ en vez de $\sum_{j=1}^{N}$. Hubiéramos podido omitir el subíndice j en X_j, supuesto que queda *sobreentendido*.

3.17. Si $Z_1 = X_1 + Y_1$, $Z_2 = X_2 + Y_2$, ..., $Z_N = X_N + Y_N$, probar que $\bar{Z} = \bar{X} + \bar{Y}$.

Solución

Por definición,

$$\bar{X} = \frac{\sum X}{N} \qquad \bar{Y} = \frac{\sum Y}{N} \qquad \bar{Z} = \frac{\sum Z}{N}$$

Luego

$$\bar{Z} = \frac{\sum Z}{N} = \frac{\sum (X + Y)}{N} = \frac{\sum X + \sum Y}{N} = \frac{\sum X}{N} + \frac{\sum Y}{N} = \bar{X} + \bar{Y}$$

donde los subíndices j en X, Y y Z han sido suprimidos, y donde $\sum$ significa $\sum_{j=1}^{N}$.

3.18. (*a*) Si N números $X_1, X_2, \ldots, X_N$ tienen desviaciones respecto de un número A dadas por $d_1 = X_1 - A$, $d_2 = X_2 - A$, ..., $d_N = X_N - A$, respectivamente, probar que

$$\bar{X} = A + \frac{\sum_{j=1}^{N} d_j}{N} = A + \frac{\sum d}{N}$$

(*b*) En el caso de que $X_1, X_2, \ldots, X_K$ tengan, respectivamente, frecuencias $f_1, f_2, \ldots, f_K$ y $d_1 = X_1 - A, \ldots, d_K = X_K - A$, probar que el resultado de la parte (*a*) queda sustituido por

$$\bar{X} = A + \frac{\sum_{j=1}^{K} f_j d_j}{\sum_{j=1}^{K} f_j} = A + \frac{\sum fd}{N} \quad \text{donde} \sum_{j=1}^{K} f_j = \sum f = N$$

Solución

(*a*) *Primer método:*

Como $d_j = X_j - A$ y $X_j = A + d_j$, se tiene

$$\bar{X} = \frac{\sum X_j}{N} = \frac{\sum (A + d_j)}{N} = \frac{\sum A + \sum d_j}{N} = \frac{NA + \sum d_j}{N} = A + \frac{\sum d_j}{N}$$

donde hemos usado $\sum$ en vez de $\sum_{j=1}^{N}$ por brevedad.

Segundo método:

Tenemos $d = X - A$, o sea $X = A + d$, omitiendo los subíndices en d y X. Luego por el Problema 3.17,

$$\bar{X} = \bar{A} + \bar{d} = A + \frac{\sum d}{N}$$

ya que la media de un conjunto de constantes iguales todas a A es A.

(*b*)

$$\bar{X} = \frac{\sum_{j=1}^{K} f_j X_j}{\sum_{j=1}^{K} f_j} = \frac{\sum f_j X_j}{N} = \frac{\sum f_j (A + d_j)}{N} = \frac{\sum A f_j + \sum f_j d_j}{N} = \frac{A \sum f_j + \sum f_j d_j}{N}$$

$$= \frac{AN + \sum f_j d_j}{N} = A + \frac{\sum f_j d_j}{N} = A + \frac{\sum fd}{N}$$

Hagamos notar que *formalmente* el resultado se obtiene de (*a*) sustituyendo d_j por $f_j d_j$ y sumando desde $j = 1$ hasta K en vez de hacerlo desde $j = 1$ hasta N. El resultado es equivalente a $\bar{X} = A + \bar{d}$, donde $\bar{d} = (\sum fd)/N$.

CALCULO DE LA MEDIA ARITMETICA PARA DATOS AGRUPADOS

3.19. Usar el método del Problema 3.18(*a*) para hallar la media aritmética de los números 5, 8, 11, 9, 12, 6, 14 y 10, escogiendo como «media conjeturada» A los valores: (*a*) 9 y (*b*) 20.

Solución

(*a*) Las desviaciones de los números dados respecto de A son $-4, -1, 2, 0, 3, -3, 5$ y 1, y la suma de estas desviaciones es $\sum d = -4 - 1 + 2 + 0 + 3 - 3 + 5 + 1 = 3$. Por tanto

$$\bar{X} = A + \frac{\sum d}{N} = 9 + \frac{3}{8} = 9.375$$

(*b*) Las desviaciones respecto de 20 son $-15, -12, -9, -11, -8, -14, -6$ y -10, y $\sum d = -85$. Así pues,

$$\bar{X} = A + \frac{\sum d}{N} = 20 + \frac{(-85)}{8} = 9.375$$

3.20. Usar el método del Problema 3.18(*b*) para hallar la media aritmética de las alturas del Problema 3.15.

Solución

El método queda indicado en la Tabla 3.2. Tomamos como media conjeturada la marca de clase 67 (que tiene la máxima frecuencia), aunque podría usarse cualquier marca de clase. Observemos que los cálculos son más sencillos que los del Problema 3.15. Para abreviarlos aún más, podemos proceder como en el Problema 3.22, haciendo uso de que las desviaciones (columna 2 de la Tabla 3.2) son todas múltiplos enteros de la anchura del intervalo de clase.

Tabla 3.2

Marca de clase (X)	Desviación $d = X - A$	Frecuencia (f)	fd
61	-6	5	-30
64	-3	18	-54
$A \rightarrow 67$	0	42	0
70	3	27	81
73	6	8	48
		$N = \sum f = 100$	$\sum fd = 45$

$$\bar{X} = A + \frac{\sum fd}{N} = 67 + \frac{45}{100} = 67.45 \text{ in}$$

3.21. Sea $d_j = X_j - A$ las desviaciones de cada marca de clase en una distribución de frecuencias respecto de una marca de clase dada A. Probar que si todos los intervalos de clase tienen la misma anchura c, entonces: (*a*) las desviaciones son todas múltiplos de c (es decir, $d_j = cu_j$, donde $u_j = 0, \pm 1, \pm 2, \ldots$) y (*b*) la media aritmética es calculable mediante la fórmula

$$\bar{X} = A + \left(\frac{\sum fu}{N}\right)c$$

Solución

(*a*) El resultado se ilustra en la Tabla 3.2 del Problema 3.20, donde se ve que las desviaciones en la columna 2 son todas múltiplos de la anchura $c = 3$ in.

Para ver que el resultado es cierto en general, notemos que si $X_1, X_2, X_3, \ldots$ son sucesivas marcas de clase, su diferencia común será igual a c, de modo que $X_2 = X_1 + c$, $X_3 = X_1 + 2c$, y en general $X_j = X_1 + (j - 1)c$. Entonces, cualquier par de marcas de clase, digamos X_p y X_q, difieren en

$$X_p - X_q = [X_1 + (p - 1)c] - [X_1 + (q - 1)c] = (p - q)c$$

que es múltiplo de c.

(*b*) Por la parte (*a*), las desviaciones de todas las marcas de clase respecto de cualquiera de ellas son múltiplos de c (o sea, $d_j = cu_j$). Usando el Problema 3.18(*b*), tendremos

$$\bar{X} = A + \frac{\sum f_j d_j}{N} = A + \frac{\sum f_j(cu_j)}{N} = A + c\frac{\sum f_j u_j}{N} = A + \left(\frac{\sum fu}{N}\right)c$$

Nótese que esto es equivalente al resultado $\bar{X} = A + c\bar{u}$, que puede obtenerse de $\bar{X} = A + \bar{d}$ haciendo $d = cu$ y observando que $\bar{d} = c\bar{u}$ (véase Prob. 3.18).

3.22. Hacer uso del resultado del Problema 3.21(*b*) para hallar la altura media de los 100 estudiantes del Problema 3.20.

Solución

El método, resumido en la Tabla 3.3, se llama *método de compilación*, y debe utilizarse siempre que sea posible.

Tabla 3.3

	X	u	f	fu
	61	−2	5	−10
	64	−1	18	−18
$A \longrightarrow$	67	0	42	0
	70	1	27	27
	73	2	8	16
			$N = 100$	$\sum fu = 15$

$$\bar{X} = A + \left(\frac{\sum fu}{N}\right)c = 67 + \left(\frac{15}{100}\right)(3) = 67.45 \text{ in}$$

3.23. Calcular el salario semanal medio de los 65 empleados de la empresa P&R a partir de la distribución de frecuencias de la Tabla 2.5, usando: (*a*) el método largo y (*b*) el método de compilación.

Solución

Las Tablas 3.4 y 3.5 muestran las respectivas soluciones a (*a*) y (*b*).

Cabe suponer que se ha introducido error en esas tablas porque las marcas de clase verdaderas son \$254.995, \$264.995, etc., en lugar de \$255.00, \$265.00, etc. Si se usan en la Tabla 3.4 esas marcas de clase verdaderas en vez de las otras, $\bar{X}$ resulta ser \$279.76 en vez de \$279.77, y la diferencia es despreciable.

$$\bar{X} = \frac{\sum fX}{N} = \frac{\$18{,}185.00}{65} = \$279.77 \qquad \bar{X} = A + \left(\frac{\sum fu}{N}\right)c = \$275.000 + \frac{31}{65}(\$10.00) = \$279.77$$

Tabla 3.4

X	f	fX
\$255.00	8	\$2040.00
265.00	10	2650.00
275.00	16	4400.00
285.00	14	3990.00
295.00	10	2950.00
305.00	5	1525.00
315.00	2	630.00
	$N=65$	$\sum fX=\$18{,}185.00$

Tabla 3.5

X	u	f	fu
\$255.00	−2	8	−16
265.00	−1	10	−10
$A\rightarrow$ 275.00	0	16	0
285.00	1	14	14
295.00	2	10	20
305.00	3	5	15
315.00	4	2	8
		$N=65$	$\sum fu=31$

3.24. Usando la Tabla 2.9(*d*), hallar el salario medio de los 70 trabajadores de la empresa P&R.

Solución

En este caso, los intervalos de clase no son de la misma anchura y hemos de recurrir al método largo, como muestra la Tabla 3.6.

Tabla 3.6

X	f	fX
\$255.00	8	\$2040.00
265.00	10	2650.00
275.00	16	4400.00
285.00	15	4275.00
295.00	10	2950.00
310.00	8	2480.00
350.00	3	1050.00
	$N=70$	$\sum fX=\$19{,}845.00$

$$\bar{X} = \frac{\sum fX}{N} = \frac{\$19{,}845.00}{70} = \$283.50$$

LA MEDIANA

3.25. Las notas de un estudiante en seis exámenes han sido 84, 91, 72, 68, 87 y 78. Hallar la mediana de esas notas.

Solución

Las notas ordenadas son 68, 72, 78, 84, 87 y 91. Como hay un número par de ellas, hay dos valores centrales, 78 y 84, cuya media aritmética $\frac{1}{2}(78 + 84) = 81$ es la nota pedida. Comparar con el Problema 3.6, donde la media aritmética era 80.

3.26. Cinco oficinistas cobran \$4.52, \$5.96, \$5.28, \$11.20 y \$5.75 a la hora. Hallar: (*a*) la mediana y (*b*) la media de esas cantidades.

Solución

(*a*) Los salarios, en ordenación, son \$4.52, \$5.28, \$5.75, \$5.96 y \$11.20. Como hay un número impar de ellos, sólo hay un valor central, \$5.75, que es la mediana.

(*b*) La media aritmética es

$$\frac{\$4.52 + \$5.96 + \$5.28 + \$11.20 + \$5.75}{5} = \$6.54$$

Nótese que la mediana no se ve afectada por el valor extremo \$11.20, mientras que la media sí. En este caso, la mediana da mejor indicación del salario medio que la media.

3.27. Si (*a*) 85 y (*b*) 150 números se ordenan, ¿cómo calcularía la mediana de esos números?

Solución

(*a*) Como hay 85 números, y 85 es impar, el único valor central es el 43.°, y ese es la mediana. Deja 42 números a cada lado.

(*b*) Ahora 150 es par, y hay dos valores centrales, el 75.° y el 76.°. Dejan 74 números a cada lado. Su promedio es la mediana.

3.28. Del Problema 2.8, hallar la mediana de los pesos de esos 40 estudiantes, usando: (*a*) la distribución de frecuencias de la Tabla 2.7 (reproducida aquí como Tabla 3.7) y (*b*) los datos originales.

Solución

(*a*) *Primer método* (por interpolación)

Los pesos en la distribución de frecuencias de la Tabla 3.7 se suponen distribuidos continuamente. En tal caso, la mediana es aquel peso para el que la mitad de la frecuencia total (40/2 = 20) quede por encima y la mitad por debajo.

Tabla 3.7

Peso (lb)	Frecuencia
118-126	3
127-135	5
136-144	9
⇒ 145-153	12
154-162	5
163-171	4
172-180	2
	Total 40

Ahora bien, la suma de las tres primeras frecuencias de clase es 3 + 5 + 9 = 17. Luego para llegar al 20 deseado tomamos 3 más de entre los 12 casos de la cuarta clase. Puesto que el cuarto intervalo de clase, 145-153, realmente corresponde a pesos desde 144.5 a 153.5, la mediana debe estar a 3/12 de camino entre 144.5 y 153.5; es decir, la mediana es

$$144.5 + \frac{3}{12}(153.5 - 144.5) = 144.5 + \frac{3}{12}(9) = 146.8 \text{ lb}$$

Segundo método (usando la fórmula)

Como la suma de las frecuencias de las tres y cuatro primeras clases son $3 + 5 + 9 = 17$ y $3 + 5 + 9 + 12 = 29$, respectivamente, es claro que la mediana cae en la cuarta clase, que es, por tanto, la clase de la mediana. Entonces

L_1 = frontera de la clase inferior a la de la mediana = 144.5

N = Número de datos = 40

$(\sum f)_1$ = suma de las clases inferiores a la de la mediana = $3 + 5 + 9 = 17$

f_{mediana} = frecuencia de la clase de la mediana = 12

c = tamaño del intervalo de la clase de la mediana = 9

luego

$$\text{Mediana} = L_1 + \left(\frac{N/2 - (\sum f)_1}{f_{\text{mediana}}}\right)c = 144.5 + \left(\frac{40/2 - 17}{12}\right)(9) = 146.8 \text{ lb}$$

(*b*) Ordenados, los pesos originales eran

119, 125, 126, 128, 132, 135, 135, 135, 136, 138, 138, 140, 140, 142, 142, 144, 144, 145, 145, 146, 146, 147, 147, 148, 149, 150, 150, 152, 153, 154, 156, 157, 158, 161, 163, 164, 165, 168, 173, 176

La mediana es la media aritmética de los pesos 20.° y 21.° en esa ordenación, a saber, 146 lb.

3.29. Mostrar cómo se puede obtener el peso mediana en el Problema 3.28 de: (*a*) un histograma y (*b*) una ojiva de porcentajes.

Solución

(*a*) La Figura 3.3(*a*) muestra el histograma de los pesos del Problema 3.28. La mediana es la abscisa correspondiente a la recta LM, que divide el histograma en dos áreas iguales. Como en un histograma el área corresponde a la frecuencia, LM es tal que el área total a izquierda y a derecha es la mitad de la frecuencia total, o sea, 20. Así pues, las áreas $AMLD$ y $MBEL$ corresponden a frecuencias de 3 y 9. Entonces, $AM = \frac{3}{12}AB = \frac{3}{12}(9) = 2.25$, y la mediana es $144.5 + 2.25 = 146.75$, o sea 146.8 lb redondeada a la décima de libra. El valor aproximado puede adivinarse del gráfico.

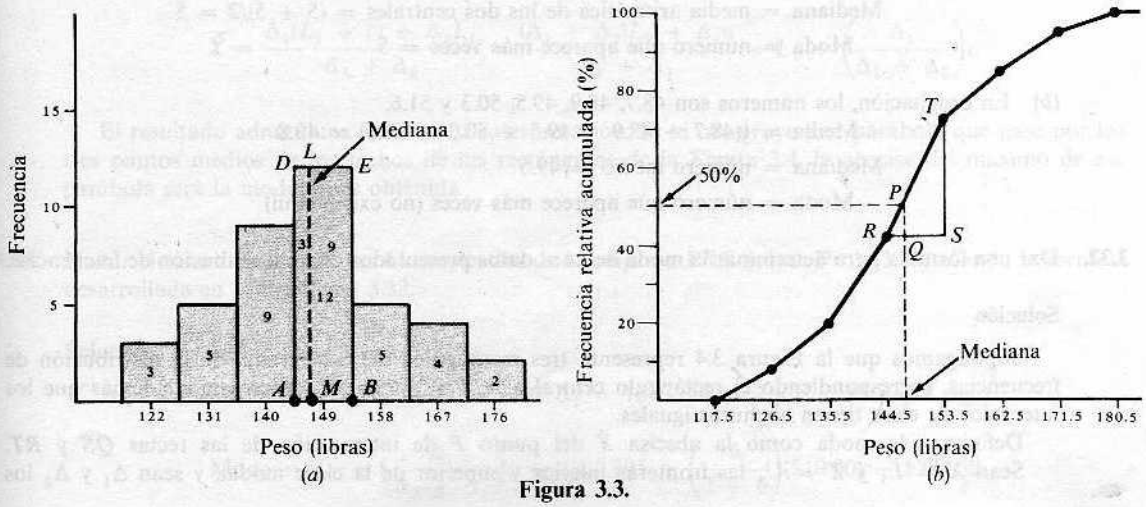


Figura 3.3.

(*b*) La Figura 3.3(*b*) muestra el polígono de frecuencia relativa acumulada (u ojiva de porcentajes) para los pesos del Problema 3.28. La mediana es la abscisa del punto P en esa ojiva, cuya ordenada es 50%. Para calcular ese valor, vemos de los triángulos semejantes PQR y RST que

$$\frac{RQ}{RS} = \frac{PQ}{ST} \quad \text{o sea} \quad \frac{RQ}{9} = \frac{50\% - 42\%}{72.5\% - 42.5\%} = \frac{1}{4} \quad \text{así que} \quad RQ = \frac{9}{4} = 2.25$$

Por tanto

$$\text{Mediana} = 144.5 + RQ = 144.5 + 2.25 = 146.75 \text{ lb}$$

o sea 146.8 lb, con precisión de décima de libra.. Este valor puede verse también aproximadamente en el gráfico.

3.30. Hallar la paga media de los 65 empleados de la empresa P&R (véase Prob. 2.3).

Solución

Aquí $N = 65$ y $N/2 = 32.5$. Como las sumas de las primeras dos y tres frecuencias de clase son $8 + 10 = 18$ y $8 + 10 + 16 = 34$, respectivamente, la clase de la mediana es la tercera. Usando la fórmula,

$$\text{Mediana} = L_1 + \left(\frac{N/2 - (\sum f)_1}{f_{\text{mediana}}}\right)c = \$269.995 + \left(\frac{32.5 - 18}{16}\right)(\$10.00) = \$279.06$$

LA MODA

3.31. Hallar la media, la mediana y la moda para los conjuntos: (*a*) 3, 5, 2, 6, 5, 9, 5, 2, 8, 6 y (*b*) 51.6, 48.7, 50.3, 49.5, 48.9.

Solución

(*a*) Ordenados, los números son 2, 2, 3, 5, 5, 5, 6, 6, 8 y 9.

$$\text{Media} = \tfrac{1}{10}(2 + 2 + 3 + 5 + 5 + 5 + 6 + 6 + 8 + 9) = 5.1$$
$$\text{Mediana} = \text{media aritmética de los dos centrales} = (5 + 5)/2 = 5$$
$$\text{Moda} = \text{número que aparece más veces} = 5$$

(*b*) En ordenación, los números son 48.7, 48.9, 49.5, 50.3 y 51.6.

$$\text{Media} = \tfrac{1}{5}(48.7 + 48.9 + 49.5 + 50.3 + 51.6) = 49.8$$
$$\text{Mediana} = \text{número medio} = 49.5$$
$$\text{Moda} = \text{número que aparece más veces (no existe aquí)}$$

3.32. Dar una fórmula para determinar la moda de unos datos presentados como distribución de frecuencias.

Solución

Supongamos que la Figura 3.4 representa tres rectángulos del histograma de la distribución de frecuencias, correspondiendo el rectángulo central a la clase modal. Y supongamos además que los intervalos de clase tienen anchuras iguales.

Definimos la moda como la abscisa $\hat{X}$ del punto P de intersección de las rectas QS y RT. Sean $X = L_1$ y $X = U_1$ las fronteras inferior y superior de la clase modal, y sean Δ_1 y Δ_2 los

excesos de frecuencia de la clase modal sobre las de las clases adyacentes a izquierda y derecha, respectivamente.

De los triángulos semejantes PQR y PST, tenemos

$$\frac{EP}{RQ} = \frac{PF}{ST} \quad \text{o sea} \quad \frac{\hat{X} - L_1}{\Delta_1} = \frac{U_1 - \hat{X}}{\Delta_2}$$

Entonces

$$\Delta_2(\hat{X} - L_1) = \Delta_1(U_1 - \hat{X}) \qquad \Delta_2\hat{X} - \Delta_2 L_1 = \Delta_1 U_1 - \Delta_1\hat{X} \qquad (\Delta_1 + \Delta_2)\hat{X} = \Delta_1 U_1 + \Delta_2 L_1$$

o

$$\hat{X} = \frac{\Delta_1 U_1 + \Delta_2 L_1}{\Delta_1 + \Delta_2}$$

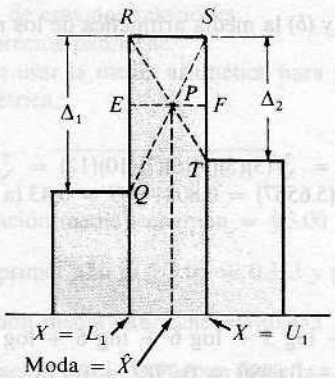


Figura 3.4.

Como $U_1 = L_1 + c$, donde c es la anchura de los intervalos de clase, eso se convierte en

$$\hat{X} = \frac{\Delta_1(L_1 + c) + \Delta_2 L_1}{\Delta_1 + \Delta_2} = \frac{(\Delta_1 + \Delta_2)L_1 + \Delta_1 c}{\Delta_1 + \Delta_2} = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)c$$

El resultado admite una interesante interpretación: Si se construye una parábola que pase por los tres puntos medios de los techos de los rectángulos de la Figura 3.4, la abscisa del máximo de esa parábola será la moda antes obtenida.

3.33. Hallar el salario modal de los 65 empleados de la empresa P&R (véase Prob. 3.23) usando la fórmula desarrollada en el Problema 3.32.

Solución

Aquí $L_1 = \$269.995$, $\Delta_1 = 16 - 10 = 6$, $\Delta_2 = 16 - 14 = 2$ y $c = \$10.00$. Luego

$$\text{Moda} = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)c = \$269.995 + \left(\frac{6}{2 + 6}\right)(\$10.00) = \$277.50$$

RELACION EMPIRICA ENTRE MEDIA, MEDIANA Y MODA

3.34. (*a*) Usar la fórmula empírica media − moda = 3(media − mediana) para hallar el salario modal de los 65 empleados de la empresa P&R.
(*b*) Comparar el resultado con el del Problema 3.33.

Solución

(*a*) De los Problemas 3.23 y 3.30 tenemos media = \$279.77 y mediana = \$279.06. Entonces

$$\text{Moda} = \text{media} - 3(\text{media} - \text{mediana}) = \$279.77 - 3(\$279.77 - \$279.06) = \$277.64$$

(*b*) Del Problema 3.33 vemos que el salario modal es \$277.50, así que está en buen acuerdo con el resultado empírico.

LA MEDIA GEOMETRICA

3.35. Hallar: (*a*) la media geométrica y (*b*) la media aritmética de los números 3, 5, 6, 6, 7, 10 y 12, supuestos exactos.

Solución

(*a*) La media geométrica $= G = \sqrt[7]{(3)(5)(6)(6)(7)(10)(12)} = \sqrt[7]{453{,}600}$. Usando logaritmos comunes, $\log G = \frac{1}{7}\log 453{,}600 = \frac{1}{7}(5.6567) = 0.8081$ y $G = 6.43$ (a la centésima). Alternativamente, puede usarse una calculadora.

Otro método

$$\begin{aligned}\log G &= \tfrac{1}{7}(\log 3 + \log 5 + \log 6 + \log 6 + \log 7 + \log 10 + \log 12)\\ &= \tfrac{1}{7}(0.4771 + 0.6990 + 0.7782 + 0.7782 + 0.8451 + 1.0000 + 1.0792)\\ &= 0.8081\end{aligned}$$

y $G = 6.43$

(*b*) Media aritmética $= \hat{X} = \frac{1}{7}(3 + 5 + 6 + 6 + 7 + 10 + 12) = 7$. Esto ilustra el hecho de que la media geométrica de un conjunto de números distintos positivos es menor que la media aritmética.

3.36. Los números $X_1, X_2, \ldots, X_K$ ocurren con frecuencias $f_1, f_2, \ldots, f_K$, donde $f_1 + f_2 + \cdots + f_K = N$ es la frecuencia total.

(*a*) Hallar su media geométrica G.
(*b*) Deducir una expresión para $\log G$.
(*c*) ¿Cómo pueden usarse esos resultados para hallar la media geométrica de datos agrupados en una distribución de frecuencias?

Solución

(*a*)

$$G = \sqrt[N]{\underbrace{X_1X_1\cdots X_1}_{f_1 \text{ veces}}\ \underbrace{X_2X_2\cdots X_2}_{f_2 \text{ veces}}\ \cdots\ \underbrace{X_KX_K\cdots X_K}_{f_K \text{ veces}}} = \sqrt[N]{X_1^{f_1}X_2^{f_2}\cdots X_K^{f_K}}$$

donde $N = \sum f$. Esto se llama a veces la *media geométrica ponderada.*

(*b*)

$$\log G = \frac{1}{N}\log\left(X_1^{f_1}X_2^{f_2}\cdots X_K^{f_K}\right) = \frac{1}{N}\left(f_1\log X_1 + f_2\log X_2 + \cdots + f_K\log X_K\right)$$

$$= \frac{1}{N}\sum_{j=1}^{K} f_j \log X_j = \frac{\sum f\log X}{N}$$

donde suponemos que todos los números son positivos; de lo contrario, los logaritmos no estarían definidos.

Nótese que el logaritmo de la media geométrica de un conjunto de números es la media aritmética de los logaritmos de tales números.

(*c*) El resultado puede aplicarse para calcular la media geométrica de datos agrupados tomando X_1 X_2, ..., X_K como marca de clase y $f_1, f_2, \ldots, f_K$ como las correspondientes frecuencias de clase.

3.37. Mientras durante un año la relación entre el precio de la leche (un cuarto de galón) y el de la hogaza de pan era 3.00, al año siguiente pasó a ser 2.00.

(*a*) Hallar la media aritmética de esas dos relaciones.
(*b*) Idem para la relación de precios pan/leche.
(*c*) Discutir la conveniencia de usar la media aritmética para promediar relaciones.
(*d*) Idem para la media geométrica.

Solución

(*a*)

$$\text{Relación media leche/pan} = \tfrac{1}{2}(3.00 + 2.00) = 2.50$$

(*b*) La relación pan/leche del primer año es $1/3.00 = 0.333$ y para el segundo $1/2.00 = 0.500$. Luego

$$\text{Relación media pan/leche} = \tfrac{1}{2}(0.333 + 0.500) = 0.417$$

(*c*) Sería de esperar que la relación media leche/pan fuese la recíproca de la pan/leche, si la media es un promedio adecuado. Sin embargo, $1/0.417 = 2.40 \neq 2.50$. Eso demuestra que la media aritmética es un pobre promedio para manejar cocientes entre magnitudes.

(*d*)

$$\text{Media geométrica de las relaciones leche/pan} = \sqrt{(3.00)(2.00)} = \sqrt{6.00}$$

$$\text{Media geométrica de las relaciones pan/leche} = \sqrt{(0.333)(0.500)} = \sqrt{0.0167} = 1/\sqrt{6.00}$$

Como estos promedios son recíprocos, la conclusión es que la media geométrica es más adecuada que la media aritmética para promediar relaciones del tipo propuesto en este problema.

3.38. La población de bacterias en un cultivo creció de 1000 a 4000 en 3 días. ¿Cuál fue el crecimiento medio diario?

Solución

Ya que de 1000 a 4000 es un 300% de crecimiento, uno podría sospechar que el crecimiento medio diario es 300%/3 = 100%. Sin embargo, eso implicaría que el primer día subiría ya de 1000 a 2000, el segundo a 4000 y el tercero a 8000, contra lo dicho.

Denotemos el crecimiento medio diario por r. Entonces

$$\text{Población de bacterias tras 1 día} = 1000 + 1000r = 1000(1 + r)$$
$$\text{Población de bacterias tras 2 días} = 1000(1 + r) + 1000(1 + r)r = 1000(1 + r)^2$$
$$\text{Población de bacterias tras 3 días} = 1000(1 + r)^2 + 1000(1 + r)^2 r = 1000(1 + r)^3$$

Esta última expresión debe dar 4000. Por tanto, $1000(1 + r)^3 = 4000$, $(1 + r)^3 = 4$, $1 + r = \sqrt[3]{4}$ y $r = \sqrt[3]{4} - 1 = 1.587 - 1 = 0.587$, así que $r = 58.7\%$.

En general, si arrancamos con una cantidad P y crece a razón constante r por unidad de tiempo, tendremos, tras n unidades de tiempo, la cantidad

$$A = P(1 + r)^n$$

Esta es la *fórmula del interés compuesto* (véanse Probs. 3.94 y 3.95).

LA MEDIA ARMONICA

3.39. Hallar la media armónica de los números 3, 5, 6, 6, 7, 10 y 12.

Solución

$$\frac{1}{H} = \frac{1}{N}\sum\frac{1}{X} = \frac{1}{7}\left(\frac{1}{3} + \frac{1}{5} + \frac{1}{6} + \frac{1}{6} + \frac{1}{7} + \frac{1}{10} + \frac{1}{12}\right) = \frac{1}{7}\left(\frac{140 + 84 + 70 + 70 + 60 + 42 + 35}{420}\right)$$

$$= \frac{501}{2940}$$

y $$H = \frac{2940}{501} = 5.87$$

A menudo conviene expresar antes las fracciones en forma decimal. Así

$$\frac{1}{H} = \tfrac{1}{7}(0.3333 + 0.2000 + 0.1667 + 0.1667 + 0.1429 + 0.1000 + 0.0833)$$

$$= \frac{1.1929}{7}$$

y $$H = \frac{7}{1.1929} = 5.87$$

La comparación con el Problema 3.35 ilustra el hecho de que la media es menor que la media geométrica, la cual a su vez es menor que la media aritmética.

3.40. Durante cuatro años sucesivos, una familia compró el fuel para su calefacción a \$0.80, \$0.90, \$1.05 y \$1.25 por galón (gal), respectivamente. Hallar el coste medio del fuel en ese período.

Solución

Caso 1

Supongamos que consumieron todos los años la misma cantidad, digamos 1000 gal. Entonces

$$\text{Coste medio} = \frac{\text{coste total}}{\text{cantidad total adquirida}} = \frac{\$800 + \$900 + \$1050 + \$1250}{400 \text{ gal}} = \$1.00/\text{gal}$$

Eso es lo mismo que la media aritmética del coste por galón; es decir, $\frac{1}{4}(\$0.80 + \$0.90 + \$1.05 + \$1.25) =$ $= 1.00$/gal. El resultado sería el mismo si consumieran x galones al año.

Caso 2

Supongamos que la familia gasta cada año la misma cantidad de dinero en fuel, digamos \$1000. Entonces

$$\text{Coste medio} = \frac{\text{coste total}}{\text{cantidad total adquirida}} = \frac{\$4000}{(1250 + 1111 + 952 + 800)\ \text{gal}} = \$0.975/\text{gal}$$

Esto es lo mismo que la media armónica de los costes por galón:

$$\frac{4}{\frac{1}{0.80} + \frac{1}{0.90} + \frac{1}{1.05} + \frac{1}{1.25}} = 0.975$$

El resultado sería el mismo si gastasen y dólares al año.

Ambos procedimientos de promediar son correctos, cada uno en ciertas circunstancias.

Debe observarse que en caso de que el consumo en galones cambiase de año en año, la media aritmética del primer caso vendría sustituida por la media aritmética ponderada. Análogamente, ante un gasto variable en dólares de año en año, la media armónica del segundo caso sería reemplazada por una media armónica ponderada.

3.41. Una persona viaja de A a B con una velocidad media de 30 millas por hora(mi/h) y regresa de B a A a una velocidad media de 60 mi/h. Hallar su velocidad media en el viaje completo.

Solución

Supongamos que A y B distan 60 millas (aunque cualquier distancia valdría). Entonces

$$\text{Tiempo para ir de } A \text{ a } B = \frac{60\ \text{mi}}{30\ \text{mi/h}} = 2\ \text{h} \qquad \text{Tiempo para ir de } B \text{ a } A = \frac{60\ \text{mi}}{60\ \text{mi/h}} = 1\ \text{h}$$

y

$$\text{Velocidad media del viaje total} = \frac{\text{distancia total}}{\text{tiempo total}} = \frac{120\ \text{mi}}{3\ \text{h}} = 40\ \text{mi/h}$$

El promedio anterior es la media armónica de 30 y 60; esto es,

$$\frac{2}{1/30 + 1/60} = 40\ \text{mi/h}$$

Si las distancias recorridas no son iguales, se llega a una media armónica ponderada, donde los pesos son las distancias (véase Prob. 3.102).

Nótese que uno hubiera estado tentado de tomar la media aritmética de 30 y 60 mi/h obteniendo 45 mi/h, lo cual es incorrecto.

LA MEDIA CUADRATICA

3.42. Hallar la media cuadrática de los números 3, 5, 6, 6, 7, 10 y 12.

Solución

$$\text{Media cuadrática} = \text{MQ} = \sqrt{\frac{3^2 + 5^2 + 6^2 + 6^2 + 7^2 + 10^2 + 12^2}{7}} = \sqrt{57} = 7.55$$

3.43. Probar que la media cuadrática de dos números positivos distintos, a y b, es mayor que su media geométrica.

Solución

Tenemos que probar que $\sqrt{\frac{1}{2}(a^2 + b^2)} > \sqrt{ab}$. Si eso es verdad, entonces completando el cuadrado de ambos lados, $\frac{1}{2}(a^2 + b^2) > ab$, de manera que $a^2 + b^2 > 2ab$, $a^2 - 2ab + b^2 > 0$, o sea $(a - b)^2 > 0$. Pero esta última desigualdad es cierta, pues el cuadrado de todo número real no nulo es positivo.

La demostración consiste en volver hacia atrás esos pasos. Así, partiendo de $(a - b)^2 > 0$, que sabemos es cierta, podemos probar que $a^2 + b^2 > 2ab$, $\frac{1}{2}(a^2 + b^2) > ab$, y finalmente $\sqrt{\frac{1}{2}(a^2 + b^2)} > \sqrt{ab}$, como se quería.

Nótese que $\sqrt{\frac{1}{2}(a^2 + b^2)} = \sqrt{ab}$, si y sólo si, $a = b$.

CUARTILES, DECILES Y PERCENTILES

3.44. Hallar: (*a*) los cuartiles Q_1, Q_2 y Q_3, y (*b*) los deciles D_1, D_2, ..., D_9 para los salarios de los 65 empleados de la empresa P&R (véase Prob. 2.3).

Solución

(*a*) El primer cuartil Q_1 es el salario obtenido contando $N/4 = 65/4 = 16.25$ de los casos, comenzando con la primera clase (la más baja). Como la primera clase contiene 8 casos, debemos tomar 8.25 (16.25 − 8) de los 10 casos de la segunda clase. Por interpolación lineal se tiene

$$Q_1 = \$259.995 + \frac{8.25}{10}(\$10.00) = \$268.25$$

El segundo cuartil Q_2 se obtiene contando los primeros $2N/4 = N/2 = 65/2 = 32.5$ casos. Como las dos primeras clases contienen 18 casos, hay que tomar $32.5 - 18 = 14.5$ de los 16 casos de la tercera clase, es decir

$$Q_2 = \$269.995 + \frac{14.5}{16}(\$10.00) = \$279.06$$

Notemos que Q_2 es la mediana.

El tercer cuartil Q_3 se obtiene contando los primeros $3N/4 = \frac{3}{4}(65) = 48.75$ casos. Ya que las cuatro primeras clases contienen 48 casos, hemos de tomar $48.75 - 48 = 0.75$ de los 10 casos de la quinta; luego

$$Q_3 = \$289.995 + \frac{0.75}{10}(\$10.00) = \$290.75$$

Por tanto, el 25% de los empleados ganan \$268.25 o menos, el 50% \$279.06 o menos, y el 75% \$290.75 o menos.

(*b*) Los deciles primero, segundo y noveno se obtienen contando $N/10$, $2N/10$, ..., $9N/10$ casos a partir de la primera clase. Así pues,

$$D_1 = \$249.995 + \frac{6.5}{8}(\$10.00) = \$258.12 \qquad D_6 = \$279.995 + \frac{5}{14}(\$10.00) = \$283.57$$

$$D_2 = \$259.995 + \frac{5}{10}(\$10.00) = \$265.0 \qquad D_7 = \$279.995 + \frac{11.5}{14}(\$10.00) = \$288.21$$

$$D_3 = \$269.995 + \frac{1.5}{16}(\$10.00) = \$270.94 \qquad D_8 = \$289.995 + \frac{4}{10}(\$10.00) = \$294.00$$

$$D_4 = \$269.995 + \frac{8}{16}(\$10.00) = \$275.00 \qquad D_9 = \$299.995 + \frac{0.5}{5}(\$10.00) = \$301.00$$

$$D_5 = \$269.995 + \frac{14.5}{16}(\$10.00) = \$279.06$$

Por tanto, el 10% de los empleados ganan \$258.12 o menos, el 20% ganan \$265.00 o menos, ..., el 90% ganan \$301.00 o menos.

Nótese que el quinto decil es la mediana. El segundo, cuarto, sexto y octavo deciles, que dividen la distribución en cinco partes iguales, se llaman *quintiles* y a veces son utilizados en la práctica.

3.45. Determinar: (*a*) el 35.° percentil y (*b*) el 60.° percentil para la distribución del Problema 3.44.

Solución

(*a*) El 35.° percentil P_{35} se obtiene contando los primeros $35N/100 = 35(65)/100 = 22.75$ casos, comenzando por la primera clase (la más baja). Entonces, como en el Problema 3.44,

$$P_{35} = \$269.995 + \frac{4.75}{16}(\$10.00) = \$272.97$$

Eso significa que el 35% de los empleados cobran \$272.97 o menos.

(*b*) El 60.° percentil es $P_{60} = \$279.995 + \frac{5}{14}(\$10.00) = \$283.57$. Coincide con el 6.° decil y el tercer quintil.

3.46. Probar que los resultados de los Problemas 3.44 y 3.45 se pueden deducir de una ojiva de porcentajes.

Solución

La ojiva de porcentajes correspondiente a los datos de los Problemas 3.44 y 3.45 se muestra en la Figura 3.5.

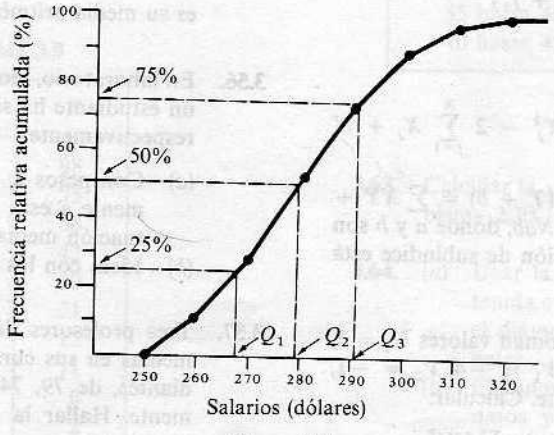


Figura 3.5.

El primer cuartil es la abscisa del punto de la ojiva cuya ordenada es 25%, y análogamente, los cuartiles segundo y tercero son las abscisas de aquellos puntos de la ojiva con ordenadas respectivas 50% y 75%.

De modo parecido se obtienen los deciles y percentiles. Por ejemplo, el 7.° decil y el 35.° percentil son las abscisas de aquellos puntos de la ojiva que tienen ordenadas respectivas 70% y 35%.

PROBLEMAS SUPLEMENTARIOS

NOTACION DE SUMA

3.47. Escribir los términos de cada suma indicada:

(*a*) $\sum_{j=1}^{4} (X_j + 2)$ (*d*) $\sum_{k=1}^{N} (Y_k^2 - 4)$

(*b*) $\sum_{j=1}^{5} f_j X_j^2$ (*e*) $\sum_{j=1}^{4} 4X_j Y_j$

(*c*) $\sum_{j=1}^{3} U_j(U_j + 6)$

3.48. Expresar en notación abreviada de suma:

(*a*) $(X_1 + 3)^3 + (X_2 + 3)^3 + (X_3 + 3)^3$

(*b*) $f_1(Y_1 - a)^2 + f_2(Y_2 - a)^2 + \cdots + f_{15}(Y_{15} - a)^2$

(*c*) $(2X_1 - 3Y_1) + (2X_2 - 3Y_2) + \cdots + (2X_N - 3Y_N)$

(*d*) $(X_1/Y_1 - 1)^2 + (X_2/Y_2 - 1)^2 + \cdots + (X_8/Y_8 - 1)^2$

(*e*) $\dfrac{f_1 a_1^2 + f_2 a_2^2 + \cdots + f_{12} a_{12}^2}{f_1 + f_2 + \cdots + f_{12}}$

3.49. Demostrar que

$$\sum_{j=1}^{N} (X_j - 1)^2 = \sum_{j=1}^{N} X_j^2 - 2 \sum_{j=1}^{N} X_j + N$$

3.50. Probar que $\sum (X + a)(Y + b) = \sum XY + a \sum Y + b \sum X + Nab$, donde a y b son constantes. ¿Qué notación de subíndice está implícita?

3.51. Dos variables, U y V, toman valores $U_1 = 3$, $U_2 = -2$, $U_3 = 5$ y $V_1 = -4$, $V_2 = -1$, $V_3 = 6$, respectivamente. Calcular:

(*a*) $\sum UV$ (*e*) $\sum UV^2$

(*b*) $\sum (U+3)(V-4)$ (*f*) $\sum (U^2 - 2V^2 + 2)$

(*c*) $\sum V^2$ (*g*) $\sum (U/V)$

(*d*) $(\sum U)(\sum V)^2$

3.52. Dado $\sum_{j=1}^{4} X_j = 7$, $\sum_{j=1}^{4} Y_j = -3$ y $\sum_{j=1}^{4} X_j Y_j = 5$, calcular: (*a*) $\sum_{j=1}^{4} (2X_j + 5Y_j)$ y (*b*) $\sum_{j=1}^{4} (X_j - 3)(2Y_j + 1)$.

LA MEDIA ARITMETICA

3.53. Las notas de un estudiante han sido 85, 76, 93, 82 y 96. Hallar su media aritmética.

3.54. Los tipos de reacción de un individuo ante diversos estímulos, medidos por un psicólogo, fueron 0.53, 0.46, 0.50, 0.49, 0.52, 0.53, 0.44 y 0.55 segundos, respectivamente. Determinar su tiempo medio de reacción.

3.55. Un conjunto de números contiene 6 seises, 7 sietes, 8 ochos, 9 nueves y 10 dieces. ¿Cuál es su media aritmética?

3.56. En laboratorio, teoría y problemas de Física, un estudiante ha sacado 71, 78 y 89 puntos, respectivamente.

(*a*) Con pesos 2, 4, 5 asignados respectivamente a esas pruebas, ¿cuál es su puntuación media?

(*b*) Idem con los tres pesos iguales.

3.57. Tres profesores de Economía dieron notas medias en sus cursos, con 32, 25 y 17 estudiantes, de 79, 74 y 82 puntos, respectivamente. Hallar la puntuación media de los tres cursos.

3.58. El salario medio anual en una empresa es de \$15,000. Los de hombres y mujeres fueron, respectivamente, de \$15,600 y \$12,600 en media. Hallar el porcentaje de mujeres empleadas en esa empresa.

3.59. La Tabla 3.8 muestra la distribución de cargas máximas en toneladas cortas (1 tonelada corta = 2000 lb) que soportan los cables producidos en cierta fábrica. Determinar la carga máxima media, usando: (*a*) el «método largo» y (*b*) el método de compilación.

Tabla 3.8

Carga máxima (toneladas cortas)	Número de cables
9.3-9.7	2
9.8-10.2	5
10.3-10.7	12
10.8-11.2	17
11.3-11.7	14
11.8-12.2	6
12.3-12.7	3
12.8-13.2	1
	Total 60

3.60. Hallar $\bar{X}$ para los datos de la Tabla 3.9, usando: (*a*) el «método largo» y (*b*) el método de compilación.

Tabla 3.9

X	f
462	98
480	75
498	56
516	42
534	30
552	21
570	15
588	11
606	6
624	2

3.61. La Tabla 3.10 muestra la distribución de los diámetros de los remaches salidos de una fábrica. Calcular el diámetro medio.

Tabla 3.10

Diámetro (cm)	Frecuencia
0.7247-0.7249	2
0.7250-0.7252	6
0.7253-0.7255	8
0.7256-0.7258	15
0.7259-0.7261	42
0.7262-0.7264	68
0.7265-0.7267	49
0.7268-0.7270	25
0.7271-0.7273	18
0.7274-0.7276	12
0.7277-0.7279	4
0.7280-0.7282	1
	Total 250

3.62. Calcular la media para los datos de la Tabla 3.11.

Tabla 3.11

Clase	Frecuencia
10 hasta 15	3
15 hasta 20	7
20 hasta 25	16
25 hasta 30	12
30 hasta 35	9
35 hasta 40	5
40 hasta 45	2
	Total 54

3.63. Calcular la vida media de los tubos del Problema 2.20.

3.64. (*a*) Usar la distribución de frecuencias obtenida en el Problema 2.27 para calcular el diámetro medio de las bolas de cojinetes.

(*b*) Calcular la media directamente de los datos y comparar con (*a*), explicando cualquier discrepancia.

LA MEDIANA

3.65. Hallar la media y la mediana de estos conjuntos de números: (*a*) 5, 4, 8, 3, 7, 2, 9 y (*b*) 18.3, 20.6, 19.3, 22.4, 20.2, 18.8, 19.7, 20.0.

3.66. Hallar la puntuación media del Problema 3.53.

3.67. Hallar el tiempo de reacción medio en el Problema 3.54.

3.68. Hallar la mediana del conjunto de números del Problema 3.55.

3.69. Hallar la mediana de las cargas máximas del Problema 3.59 (Tabla 3.8).

3.70. Hallar la mediana $\tilde{X}$ para la distribución del Problema 3.60 (Tabla 3.9).

3.71. Hallar el diámetro medio de los remaches de la Tabla 3.10, Problema 3.61.

3.72. Hallar la mediana de la distribución de la Tabla 3.11 del Problema 3.62.

3.73. La Tabla 3.12 muestra el número de bodas (incluidas posibles repeticiones) en EE.UU. para hombres y mujeres de distintos grupos de edad durante 1984.

(*a*) Hallar la mediana de edad de hombres y mujeres en esas bodas.

(*b*) ¿Por qué la mediana es una medida de tendencia central más adecuada que la media en este caso?

Tabla 3.12

Edad (años)	Varones (miles)	Hembras (miles)
18-19	121	481
20-24	2,441	4,184
25-29	5,930	6,952
30-34	6,587	7,193
35-44	11,788	11,893
45-54	9,049	9,022
55-64	8,749	8,171
65-74	5,786	4,654
75 y más	2,581	1,524

Fuente: U.S. Bureau of Census

3.74. Hallar la mediana de las ventas del Problema 2.31.

3.75. Hallar la mediana de las vidas medias de los tubos del Problema 2.20.

LA MODA

3.76. Hallar la media, la mediana y la moda de cada uno de estos conjuntos: (*a*) 7, 4, 10, 9, 15, 12, 7, 9, 7 y (*b*) 8, 11, 4, 3, 2, 5, 10, 6, 4, 1, 10, 8, 12, 6, 5, 7.

3.77. Hallar la puntuación modal del Problema 3.53.

3.78. Hallar el tiempo de reacción modal en el Problema 3.54.

3.79. Hallar la moda del conjunto de números del Problema 3.55.

3.80. Hallar la moda de las cargas máximas de los cables del Problema 3.59.

3.81. Hallar la moda $\hat{X}$ para la distribución de la Tabla 3.9 del Problema 3.60.

3.82. Hallar el diámetro modal de los remaches de la Tabla 3.10, del Problema 3.61.

3.83. Hallar la moda de la distribución del Problema 3.62.

3.84. Hallar la vida media modal de los tubos del Problema 2.20.

3.85. ¿Es posible determinar la moda para las distribuciones de los Problemas 3.73 y 2.31? Razonar la respuesta.

3.86. Usar la fórmula empírica media − moda = = 3(media − mediana) para calcular la moda de las distribuciones de los Problemas 3.59, 3.60, 3.61, 3.62 y 2.20. Comparar los resultados con los que da la fórmula (9) de este capítulo, explicando los acuerdos y las discrepancias.

3.87. Probar la afirmación del final del Problema 3.32.

LA MEDIA GEOMETRICA

3.88. Hallar la media geométrica de los números: (*a*) 4.2 y 16.8 y (*b*) 3.00 y 6.00.

3.89. Hallar (*a*) la media geométrica G y (*b*) la media aritmética $\bar{X}$ del conjunto 2, 4, 8, 16, 32.

3.90. Hallar la media geométrica de los conjuntos: (*a*) 3, 5, 8, 3, 7, 2 y (*b*) 28.5, 73.6, 47.2, 31.5, 64.8.

3.91. Hallar la media geométrica de las distribuciones en: (*a*) Problema 3.59 y (*b*) Problema 3.60. Verificar que la media geométrica es menor o igual que la media aritmética en estos casos.

3.92. Si el precio de un artículo se duplica en un período de 4 años, ¿cuál es el porcentaje medio de crecimiento anual?

3.93. En 1970 y 1980 la población de EE.UU. era de 203.3 y 226.5 millones, respectivamente.

(*a*) Hallar el porcentaje medio de crecimiento anual.
(*b*) Estimar la población en 1974.
(*c*) Si el porcentaje medio de crecimiento entre 1980 y 1990 es el de la parte (*a*), ¿cuál será la población en 1990?

3.94. ¿Qué capital final se tendrá al cabo de 6 años, si se invierten \$1000 al 8% de interés anual?

3.95. Si en el problema anterior se compone el interés trimestralmente (o sea, el capital aumenta un 2% cada trimestre), ¿cuál sería el capital final?

3.96. Hallar dos números cuya media aritmética es 9.0 y cuya media geométrica es 7.2.

LA MEDIA ARMONICA

3.97. Hallar la media armónica de los números: (*a*) 2, 3 y 6 y (*b*) 3.2, 5.2, 4.8, 6.1 y 4.2.

3.98. Hallar (*a*) la media aritmética, (*b*) la media geométrica y (*c*) la media armónica de los números 0, 2, 4 y 6.

3.99. Si X_1, X_2, X_3, ... representan las marcas de clase de una distribución de frecuencias con correspondientes frecuencias de clase f_1, f_2, f_3, ..., probar que la media armónica H de esa distribución viene dada por

$$\frac{1}{H} = \frac{1}{N}\left(\frac{f_1}{X_1} + \frac{f_2}{X_2} + \frac{f_3}{X_3} + \cdots\right) = \frac{1}{N}\sum\frac{f}{X}$$

donde $N = f_1 + f_2 + \cdots = \sum f$.

3.100. Usar el Problema 3.99 para hallar la media armónica de las distribuciones de: (*a*) Problema 3.59 y (*b*) Problema 3.60. Comparar con el Problema 3.91.

3.101. Las ciudades A, B y C están equidistantes entre sí. Un motorista viaja desde A hasta B a 30 mi/h, desde B hasta C a 40 mi/h, y desde C hasta A a 50 mi/h. Determinar su velocidad media en el viaje completo.

3.102. (*a*) Un avión vuela d_1, d_2 y d_3 millas a velocidades v_1, v_2 y v_3 mi/h, respectivamente. Probar que su velocidad media es V, dada por

$$\frac{d_1 + d_2 + d_3}{V} = \frac{d_1}{v_1} + \frac{d_2}{v_2} + \frac{d_3}{v_3}$$

Es una media armónica ponderada.

(*b*) Calcular V si $d_1 = 2500$, $d_2 = 1200$, $d_3 = 500$, $v_1 = 500$, $v_2 = 400$ y $v_3 = 250$.

3.103. Demostrar que la media geométrica de dos números positivos a y b es: (*a*) menor o igual que la media aritmética y (*b*) mayor o igual que la media armónica de esos números. ¿Puede extender la demostración a más de dos números?

LA MEDIA CUADRATICA

3.104. Hallar la media cuadrática de los números: (*a*) 11, 23 y 35 y (*b*) 2.7, 3.8, 3.2 y 4.3.

3.105. Probar que la media cuadrática de dos números positivos a y b es: (*a*) mayor o igual que la media aritmética y (*b*) mayor o igual que la media armónica. Extienda, si le es posible, la demostración a más de dos números.

3.106. Deducir una fórmula que sirva para hallar la media cuadrática de datos agrupados y aplíquese a alguna distribución de frecuencias ya considerada.

CUARTILES, DECILES Y PERCENTILES

3.107. La Tabla 3.13 muestra una distribución de frecuencias de puntuaciones de un examen final de álgebra. (*a*) Hallar los cuartiles de la distribución y (*b*) interpretar su significado.

Tabla 3.13

Grado	Número de estudiantes
90-100	9
80-89	32
70-79	43
60-69	21
50-59	11
40-49	3
30-39	1
	Total 120

3.108. Hallar los cuartiles Q_1, Q_2 y Q_3 para la distribución del: (*a*) Problema 3.59 y (*b*) Problema 3.60. Interpretar su significado.

3.109. Hallar: (*a*) el segundo decil, (*b*) el cuarto decil, (*c*) el 90.° percentil y (*d*) el 68.° percentil, para los datos del Problema 3.73, interpretando cada uno de ellos.

3.110. Hallar: (*a*) P_{10}, (*b*) P_{90}, (*c*) P_{25} y (*d*) P_{75} para los datos del Problema 3.59, interpretando cada uno de ellos.

3.111. (*a*) ¿Pueden todos los cuartiles ser expresados como percentiles? Explíquese.
(*b*) Idem con los quintiles.

3.112. Para los datos del Problema 3.107, determinar: (*a*) la puntuación más baja alcanzada por el 25% más alto del curso y (*b*) la más alta alcanzada por el 20% más bajo del curso. Interpretar la respuesta en términos de percentiles.

3.113. Interpretar los resultados del Problema 3.107 gráficamente usando: (*a*) un histograma de porcentajes, (*b*) un polígono de frecuencias en porcentajes y (*c*) una ojiva de porcentajes.

3.114. Resolver el Problema 3.113 con los datos del Problema 3.108.

3.115. (*a*) Desarrollar una fórmula, similar a la (8) de este capítulo, para calcular percentiles de una distribución de frecuencias.
(*b*) Ilustrar su uso obteniendo los resultados del Problema 3.110.

CAPITULO 4

La desviación típica y otras medidas de dispersión

DISPERSION O VARIACION

La *dispersión* o *variación* de los datos intenta dar una idea de cuán esparcidos se encuentran éstos. Hay varias medidas de tal dispersión, siendo las más comunes el rango, la desviación media, el rango semi-intercuartil, el rango percentil 10-90 y la desviación típica.

EL RANGO

El *rango* de un conjunto de números es la diferencia entre el mayor y el menor de todos ellos.

EJEMPLO 1. El rango del conjunto 2, 3, 3, 5, 5, 5, 8, 10, 12 es $12 - 2 = 10$. A veces el rango se indica dando el par de valores extremos; así, en este ejemplo, sería 2-12.

LA DESVIACION MEDIA

La *desviación media* o *desviación promedio*, de un conjunto de N números $X_1, X_2, \ldots, X_N$ es abreviada por MD y se define como

$$\text{Desviación media (MD)} = \frac{\sum_{j=1}^{N} |X_j - \bar{X}|}{N} = \frac{\sum |X - \bar{X}|}{N} = \overline{|X - \bar{X}|} \qquad (1)$$

donde $\bar{X}$ es la media aritmética de los números y $|X_j - \bar{X}|$ es el valor absoluto de la desviación de X_j respecto de $\bar{X}$. (El *valor absoluto* de un número es el número sin signo y se denota con dos barras verticales; así $|-4| = 4$, $|+3| = 3$, $|6| = 6$ y $|-0.84| = 0.84$.)

EJEMPLO 2. Hallar la desviación media del conjunto 2, 3, 6, 8, 11.

$$\text{Media aritmética } (\bar{X}) = \frac{2 + 3 + 6 + 8 + 11}{5} = 6$$

$$\text{MD} = \frac{|2-6| + |3-6| + |6-6| + |8-6| + |11-6|}{5} = \frac{|-4| + |-3| + |0| + |2| + |5|}{5} = \frac{4+3+0+2+5}{5} = 2.8$$

Si $X_1, X_2, \ldots, X_K$ ocurren con frecuencias $f_1, f_2, \ldots, f_K$, respectivamente, la desviación media se puede escribir como

$$\text{MD} = \frac{\sum_{j=1}^{K} f_j|X_j - \bar{X}|}{N} = \frac{\sum f|X - \bar{X}|}{N} = \overline{|X - \bar{X}|} \tag{2}$$

donde $N = \sum_{j=1}^{K} f_j = \sum f$. Esta forma es útil para datos agrupados, donde los X_j representan las marcas de clase y los f_j son las correspondientes frecuencias de clase.

Ocasionalmente se define la desviación media en términos de desviaciones absolutas respecto de la mediana u otro promedio, en vez de la media. Una propiedad interesante de la suma $\sum_{j=1}^{N} |X_j - a|$ es que es mínima cuando a es la mediana (o sea, la desviación media respecto de la mediana es mínima).

Nótese que sería más apropiado usar la terminología *desviación media absoluta* que *desviación media*.

EL RANGO SEMI-INTERCUARTIL

El *rango semi-intercuartil*, o *desviación cuartil*, de un conjunto de datos se denota por Q y se define como

$$Q = \frac{Q_3 - Q_1}{2} \tag{3}$$

donde Q_1 y Q_3 son el primer y tercer cuartil de esos datos (véanse Probs. 4.6 y 4.7). El rango intercuartil $Q_3 - Q_1$ también se usa a veces, pero menos que el rango semi-intercuartil, como medida de dispersión.

EL RANGO PERCENTIL 10-90

El *rango percentil 10-90* de un conjunto de datos se define por

$$\text{rango percentil 10-90} = P_{90} - P_{10} \tag{4}$$

donde P_{10} y P_{90} son los décimo y nonagésimo percentiles de esos datos (véase Prob. 4.8). Puede usarse también el rango percentil semi 10-90 $\frac{1}{2}(P_{90} - P_{10})$, pero no es frecuente.

LA DESVIACION TIPICA o DESVIACION STANDAR.

La *desviación típica* de un conjunto de N números $X_1, X_2, \ldots, X_N$ se denota por s y se define como

$$s = \sqrt{\frac{\sum_{j=1}^{N} (X_j - \bar{X})^2}{N}} = \sqrt{\frac{\sum (X - \bar{X})^2}{N}} = \sqrt{\frac{\sum x^2}{N}} = \sqrt{(X - \bar{X})^2} \tag{5}$$

donde x representa las desviaciones de cada uno de los números X_j respecto de la media $\bar{X}$. Así que s es la raíz cuadrada de la media de las desviaciones cuadráticas, o como se le llama en ocasiones, la *desviación raíz-media-cuadrado*.

Si $X_1, X_2, \ldots, X_K$ ocurren con frecuencias $f_1, f_2, \ldots, f_K$, respectivamente, la desviación típica puede expresarse

$$s = \sqrt{\frac{\sum_{j=1}^{K} f_j(X_j - \bar{X})^2}{N}} = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}} = \sqrt{\frac{\sum fx^2}{N}} = \sqrt{(X - \bar{X})^2} \tag{6}$$

donde $N = \sum_{j=1}^{K} f_j = \sum f$. En esta forma resulta útil para datos agrupados.

A veces se define la desviación típica de los datos de una muestra con $(N - 1)$ reemplazando a N en los denominadores de (5) y (6), porque el valor resultante da una mejor estimación de la desviación típica de la población total. Para grandes valores de N (ciertamente para $N > 30$), no hay prácticamente diferencia entre ambas definiciones. Además, cuando se necesita esa mejor estimación, siempre podemos obtenerla multiplicando la aquí definida por $\sqrt{N/(N-1)}$. Por tanto, nos quedaremos con la elección (5) y (6).

LA VARIANZA

La *varianza* de un conjunto de datos se define como el cuadrado de la desviación típica y viene dada en consecuencia por s^2 en las ecuaciones (5) y (6).

Cuando sea necesario distinguir la desviación típica de una población de la de una muestra de dicha población, usaremos el símbolo s para esta última y σ (*sigma* griega minúscula) para la primera. De modo que s^2 y σ^2 representarían la *varianza de la muestra* y la *varianza de la población*, respectivamente.

METODOS CORTOS PARA CALCULAR LA DESVIACION TIPICA

Las ecuaciones (5) y (6) se pueden escribir, respectivamente, en las formas equivalentes

$$s = \sqrt{\frac{\sum_{j=1}^{N} X_j^2}{N} - \left(\frac{\sum_{j=1}^{N} X_j}{N}\right)^2} = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2} \tag{7}$$

$$s = \sqrt{\frac{\sum_{j=1}^{K} f_j X_j^2}{N} - \left(\frac{\sum_{j=1}^{K} f_j X_j}{N}\right)^2} = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2} \tag{8}$$

donde $\overline{X^2}$ denota la media de los cuadros de los diversos valores de X, mientras $\bar{X}^2$ denota el cuadrado de la media de los valores de X (véanse Probs. 4.12 a 4.14).

Si $d_j = X_j - A$ son las desviaciones de X_j respecto de alguna constante arbitraria A, los resultados (7) y (8) se convierten, respectivamente, en

$$s = \sqrt{\frac{\sum_{j=1}^{N} d_j^2}{N} - \left(\frac{\sum_{j=1}^{N} d_j}{N}\right)^2} = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} = \sqrt{\overline{d^2} - \overline{d}^2} \tag{9}$$

$$s = \sqrt{\frac{\sum_{j=1}^{K} f_j d_j^2}{N} - \left(\frac{\sum_{j=1}^{K} f_j d_j}{N}\right)^2} = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\overline{d^2} - \overline{d}^2} \tag{10}$$

(Véanse Probs. 4.15 y 4.17.)

Cuando se tienen los datos agrupados en una distribución de frecuencias cuyos intervalos de clase tienen la misma anchura c, tenemos $d_j = cu_j$ o sea $X_j = A + cu_j$ y (10) pasa a ser

$$s = c\sqrt{\frac{\sum_{j=1}^{K} f_j u_j^2}{N} - \left(\frac{\sum_{j=1}^{K} f_j u_j}{N}\right)^2} = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = c\sqrt{\overline{u^2} - \overline{u}^2} \tag{11}$$

Esta última fórmula proporciona un método muy breve para calcular la desviación típica y debe usarse para datos agrupados con igual anchura en sus intervalos de clase. Se llama *método de compilación* y es similar al utilizado en el Capítulo 3 para el cálculo de la media aritmética de datos agrupados. (Véanse Probs 4.16 a 4.19.)

PROPIEDADES DE LA DESVIACION TIPICA

1. La desviación típica puede definirse como

$$s = \sqrt{\frac{\sum_{j=1}^{N} (X_j - a)^2}{N}}$$

donde a es un promedio distinto de la media aritmética. De tales desviaciones típicas, la mínima es aquella para la cual $a = \bar{X}$, debido a la Propiedad 2 del Capítulo 3. Esta propiedad da una buena razón para adoptar la definición del comienzo. Su demostración se verá en el Problema 4.27.

2. Para distribuciones normales (véase Cap. 7), resulta (como sugiere la Fig. 4.1):

   (*a*) 68.27% de los casos están entre $\bar{X} - s$ y $\bar{X} + s$ (o sea, una desviación típica a cada lado de la media).

   (*b*) 95.45% de los casos están entre $\bar{X} - 2s$ y $\bar{X} + 2s$ (o sea, dos desviaciones típicas a cada lado de la media).

   (*d*) 99.73% de los casos entre $\bar{X} - 3s$ y $\bar{X} + 3s$ (o sea, tres desviaciones típicas a cada lado de la media).

Para distribuciones poco asimétricas, los anteriores porcentajes son aproximadamente válidos (véase Prob. 4.24).

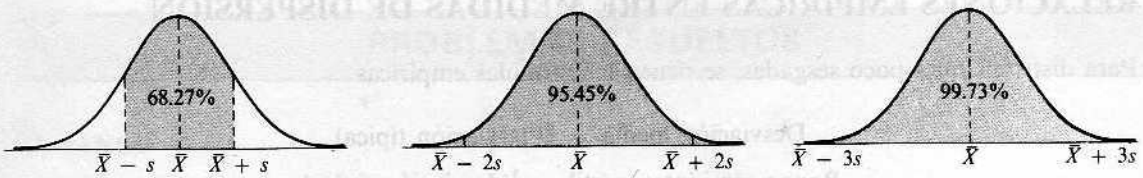


Figura 4.1.

3. Supongamos que dos conjuntos de N_1 y N_2 números (o dos distribuciones de frecuencias con frecuencias totales N_1 y N_2 tienen varianza dadas por s_1^2 y s_2^2, respectivamente, y tienen la *misma* media $\bar{X}$. Entonces la *varianza combinada* de ambos conjuntos (o de ambas distribuciones de frecuencias) vendrá dada por

$$s^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2} \tag{12}$$

Nótese que esto es una medida aritmética ponderada de las varianzas. El resultado admite generalización a más conjuntos.

COMPROBACION DE CHARLIER

La comprobación de Charlier en cálculos de la media y de la desviación típica por el método de compilación hace uso de las identidades

$$\sum f(u + 1) = \sum fu + \sum f = \Sigma fu + N$$

$$\sum f(u + 1)^2 = \sum f(u^2 + 2u + 1) = \sum fu^2 + 2\sum fu + \sum f = \sum fu^2 + 2\sum fu + N$$

(Véase Prob. 4.20.)

CORRECCION DE SHEPPARD PARA LA VARIANZA

El cálculo de la desviación típica es algo erróneo como resultado del agrupamiento de datos en clases (error de agrupamiento). Para corregirlo, se usa la fórmula

$$\text{Varianza corregida} = \text{varianza de los datos agrupados} - \frac{c^2}{12} \tag{13}$$

donde c es la anchura del intervalo de clase. La corrección $c^2/12$ (que se resta) se llama *corrección de Sheppard*. Se usa para distribuciones de variables continuas donde las «colas» van hacia cero en ambas direcciones.

Los estadísticos discrepan en cuanto a *si debe* aplicarse antes de examinar con corrección y *cuándo*. Ciertamente no debe aplicarse antes de examinar con cuidado la situación, pues a menudo tiende a *corregir en demasía*, con lo que sustituye un error por otro. En este libro, salvo indicación expresa, no la usaremos.

RELACIONES EMPIRICAS ENTRE MEDIDAS DE DISPERSION

Para distribuciones poco sesgadas, se tienen las fórmulas empíricas

$$\text{Desviación media} = \tfrac{4}{5}(\text{desviación típica})$$

$$\text{Rango semi-intercuartil} = \tfrac{2}{3}(\text{desviación típica})$$

Son consecuencia de que para la distribución normal la desviación media y el rango semi-intercuartil son iguales, respectivamente, a 0.7979 y 0.6745 veces la desviación típica.

DISPERSION ABSOLUTA Y RELATIVA: COEFICIENTE DE VARIACION

La variación o dispersión real, tal como se determina de la desviación típica u otra medida de dispersión, se llama la *dispersión absoluta*. Sin embargo, una dispersión (o variación) de 10 pulgadas (in) en la medida de 1000 pies es muy diferente de esa misma dispersión al medir una distancia de 20 pies. Una medida de este efecto la da la *dispersión relativa*, a saber

$$\text{Dispersión relativa} = \frac{\text{dispersión absoluta}}{\text{promedio}} \tag{14}$$

Si la dispersión absoluta es la desviación típica s y el promedio es la media $\bar{X}$, entonces la dispersión relativa se llama el *coeficiente de variación*, o *coeficiente de dispersión;* se denotará por V y se define como

$$\text{Coeficiente de variación } (V) = \frac{s}{\bar{X}} \tag{15}$$

y se expresa en general en forma de porcentaje. Hay otras posibilidades (véase Prob. 4.30).

Nótese que el coeficiente de variación es independiente de las unidades usadas. Por esa razón es útil al comparar distribuciones con unidades diferentes. Una desventaja del coeficiente de variación es que pierde su utilidad cuando $\bar{X}$ es próxima a cero.

VARIABLES TIPIFICADAS: UNIDADES ESTANDAR

La variable que mide la desviación de la medida en unidades de la desviación típica se llama una *variable tipificada*, es adimensional (independiente de las unidades usadas) y viene dada por

$$z = \frac{X - \bar{X}}{s} \tag{16}$$

Si las desviaciones de la media se dan en unidades de la desviación típica, se dicen expresadas en *unidades estándar*, o *recuentos estándar*. Son de gran valor al comparar distribuciones (véase Problema 4.31).

PROBLEMAS RESUELTOS

EL RANGO

4.1. Hallar el rango de los conjuntos (*a*) 12, 6, 7, 3, 15, 10, 18, 5 y (*b*) 9, 3, 8, 8, 9, 8, 9, 18.

Solución

En ambos casos, rango = número mayor − número menor = 18 − 3 = 15. Sin embargo, como se ve de sus ordenaciones (*a*) y (*b*),

(*a*) 3, 5, 6, 7, 10, 12, 15, 18 (*b*) 3, 8, 8, 8, 9, 9, 9, 18

hay mucha más dispersión en (*a*) que en (*b*). De hecho, (*b*) consiste esencialmente de ochos y nueves.

Como el rango no indica diferencia entre esos conjuntos, no es buena medida de la dispersión en este caso. Cuando hay valores muy extremos, el rango es una pobre medida de la dispersión.

Se mejora eliminando los valores extremos, 3 y 18. Entonces para (*a*) el rango es (15 − 5) = 10, y para (*b*) es (9 − 8) = 1, que muestran claramente que el (*a*) tiene más dispersión que el (*b*). No obstante, no es así como se define el rango. El rango semi-intercuartil y el rango percentil 10-90 están pensados para mejorar el rango suprimiendo los valores extremos.

4.2. Hallar el rango de las alturas de los estudiantes de la Tabla 2.1.

Solución

Hay dos formas de definir el rango para datos agrupados.

Primer método

Rango = marca de clase de la clase más alta − marca clase más baja = 73 − 61 = 12 in

Segundo método

Rango = frontera superior de la clase más alta − frontera inferior de la clase más baja = = 74.5 − 59.5 = 15 in

El primer método tiende a eliminar los casos extremos en cierto grado.

LA DESVIACION MEDIA

4.3. Hallar la desviación media de los conjuntos de números del Problema 4.1.

Solución

(*a*) La media aritmética es

$$\bar{X} = \frac{12 + 6 + 7 + 3 + 15 + 10 + 18 + 5}{8} = \frac{76}{8} = 9.5$$

La desviación media es

$$\text{MD} = \frac{\sum |X - \bar{X}|}{N} =$$

$$= \frac{|12-9.5|+|6-9.5|+|7-9.5|+|3-9.5|+|15-9.5|+|10-9.5|+|18-9.5|+|5-9.5|}{8} =$$

$$= \frac{2.5+3.5+2.5+6.5+5.5+0.5+8.5+4.5}{8} = \frac{34}{8} = 4.25$$

(*b*)
$$\bar{X} = \frac{9+3+8+8+9+8+9+18}{8} = \frac{72}{8} = 9$$

$$\text{MD} = \frac{\sum |X - \bar{X}|}{N} =$$

$$= \frac{|9-9|+|3-9|+|8-9|+|8-9|+|9-9|+|8-9|+|9-9|+|18-9|}{8} =$$

$$= \frac{0+6+1+1+0+1+0+9}{8} = 2.25$$

La desviación media indica que el conjunto (*b*) tiene menos dispersión que el (*a*), como debía ocurrir.

4.4. Hallar la desviación media de las alturas de los 100 estudiantes de la Universidad XYZ (Tabla 3.2 del Problema 3.20).

Solución

Del Problema 3.20, $\bar{X} = 67.45$ in. El trabajo se realiza como en la Tabla 4.1. Es posible diseñar un método de compilación para calcular la desviación media (véase Prob. 4.47).

Tabla 4.1

Altura (in)	Marca de clase (X)	$\lvert X - \bar{X}\rvert = \lvert X - 67.45\rvert$	Frecuencia (f)	$f\lvert X - \bar{X}\rvert$
50-62	61	6.45	5	32.25
63-65	64	3.45	18	62.10
66-68	67	0.45	42	18.90
69-71	70	2.55	27	68.85
72-74	73	5.55	8	44.40
			$N = \sum f = 100$	$\sum f\lvert X - \bar{X}\rvert = 226.50$

$$\text{MD} = \frac{\sum f|X - \bar{X}|}{N} = \frac{226.50}{100} = 2.26 \text{ in}$$

4.5. Determinar el porcentaje de estudiantes del Problema 4.4 que miden entre (*a*) $\bar{X} \pm \text{MD}$, (*b*) $\bar{X} \pm 2$ MD, (*c*) $\bar{X} \pm 3$ MD.

Solución

(*a*) El rango entre 65.19 y 69.71 in es $\bar{X} \pm \text{MD} = 67.45 \pm 2.26$. Este rango incluye a todos los individuos de la tercera clase; $+\frac{1}{3}(65.5 - 65.19)$, de los de la segunda; $+\frac{1}{3}(69.71 - 68.5)$, de los de la cuarta (como la anchura del intervalo de clase es 3 in, la frontera superior de la segunda clase es 65.5 in, y la inferior de la cuarta 68.5 in). El número de estudiantes en el rango $\bar{X} \pm 2$ MD es

$$42 + \frac{0.31}{3}(18) + \frac{1.21}{3}(27) = 42 + 1.86 + 10.89 = 54.75 \quad \text{o sea} \quad 55$$

que es el 55% del total.

(*b*) El rango desde 62.93 a 71.97 in es $\bar{X} \pm 2 \text{ MD} = 67.45 \pm 2(2.26) = 67.45 \pm 4.52$. El número de estudiantes en el rango $\bar{X} \pm 2$ MD es

$$18 - \left(\frac{62.93 - 62.5}{3}\right)(18) + 42 + 27 + \left(\frac{71.97 - 71.5}{3}\right)(8) = 85.67 \quad \text{o sea} \quad 86$$

que es el 86% del total.

(*c*) El rango desde 60.67 a 74.23 in es $\bar{X} \pm 3 \text{ MD} = 67.45 \pm 3(2.26) = 67.45 \pm 6.78$. El número de estudiantes en el rango $\bar{X} \pm 3$ MD es

$$5 - \left(\frac{60.67 - 59.5}{3}\right)(5) + 18 + 42 + 27 + \left(\frac{74.5 - 74.23}{3}\right)(8) = 97.33 \quad \text{o sea} \quad 97$$

que es el 97% del total.

EL RANGO SEMI-INTERCUARTIL

4.6. Hallar el rango semi-intercuartil para la distribución de alturas de la Universidad XYZ (Tabla 4.1 del Problema 4.4).

Solución

Los cuartiles inferior y superior son $Q_1 = 65.5 + \frac{2}{42}(3) = 65.64$ in y $Q_3 = 68.5 + \frac{10}{27}(3) = 69.61$ in, respectivamente, y el rango semi-intercuartil (o desviación cuartil) es $Q = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(69.61 - 65.64) = 1.98$ in. Nótese que el 50% de los casos cae entre Q_1 y Q_3 (o sea, 50 estudiantes miden entre 65.64 y 69.61 in).

Podemos considerar $\frac{1}{2}(Q_1 + Q_3) = 67.63$ in como una medida de tendencia central (o sea, un promedio de alturas). Se sigue que el 50% de las alturas caen en el rango 67.63 ± 1.98 in.

4.7. Hallar el rango semi-intercuartil para los salarios de los 65 empleados de la empresa P&R (Tabla 2.5 del Próblema 2.3).

Solución

Del Problema 3.44, $Q_1 = \$268.25$ y $Q_3 = \$290.75$. Así pues, el rango semi-intercuartil $Q = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(\$290.75 - \$268.25) = \$11.25$. Como $\frac{1}{2}(Q_1 + Q_3) = \$279.50$, podemos concluir que el 50% de los empleados cobra en el rango $\$279.50 \pm \$11.25$.

EL RANGO PERCENTIL 10-90

4.8. Hallar el rango percentil 10-90 de las alturas de la Tabla 2.1.

Solución

Aqui $P_{10} = 62.5 + \frac{5}{18}(3) = 63.33$ in, y $P_{90} = 68.5 + \frac{25}{27}(3) = 71.27$ in. Luego el rango percentil 10-90 es $P_{90} - P_{10} = 71.27 - 63.33 = 7.94$ in. Como $\frac{1}{2}(P_{10} + P_{90}) = 67.30$ in y $\frac{1}{2}(P_{90} - P_{10}) = 3.97$ in, podemos concluir que el 80% de los estudiantes tiene alturas en el rango 67.30 ± 3.97 in.

LA DESVIACION TIPICA

4.9. Hallar la desviación típica s de los conjuntos de números del Problema 4.1.

Solución

(*a*) $$\bar{X} = \frac{\sum X}{N} = \frac{12 + 6 + 7 + 3 + 15 + 10 + 18 + 5}{8} = \frac{76}{8} = 9.5$$

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N}} =$$

$$= \sqrt{\frac{(12-9.5)^2 + (6-9.5)^2 + (7-9.5)^2 + (3-9.5)^2 + (15-9.5)^2 + (10-9.5)^2 + (18-9.5)^2 + (5-9.5)^2}{8}} =$$

$$= \sqrt{23.75} = 4.87$$

(*b*) $$\bar{X} = \frac{9 + 3 + 8 + 8 + 9 + 8 + 9 + 18}{8} = \frac{72}{8} = 9$$

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N}} =$$

$$= \sqrt{\frac{(9-9)^2 + (3-9)^2 + (8-9)^2 + (8-9)^2 + (9-9)^2 + (8-9)^2 + (9-9)^2 + (18-9)^2}{8}} =$$

$$= \sqrt{15} = 3.87$$

Los resultados anteriores deben compararse con los del Problema 4.3. Se apreciará que la desviación típica indica que (*b*) es menos disperso que (*a*). Sin embargo, el efecto está enmascarado por el hecho de que los valores extremos afectan a la desviación típica mucho más que a la desviación media. Era de esperar, desde luego, porque las desviaciones se elevan al cuadrado al calcular la desviación típica.

4.10. Hallar la varianza de los conjuntos de números del Problema 4.1.

Solución

Varianza $= s^2$. Luego del Problema 4.9 deducimos (*a*) $s^2 = 23.75$ y (*b*) $s^2 = 15$.

4.11. Hallar la desviación típica de las alturas de estudiantes de la Tabla 2.1.

Solución

De los Problemas 3.15, 3.20 ó 3.22, $\bar{X} = 67.45$ in. El método de trabajo se recoge en la Tabla 4.2.

Tabla 4.2

Altura (in)	Marca de clase (X)	$X - \bar{X} = X - 67.45$	$(X - \bar{X})^2$	Frecuencia (f)	$f(X - \bar{X})^2$
60-62	61	−6.45	41.6025	5	208.0125
63-65	64	−3.45	11.9025	18	214.2450
66-68	67	−0.45	0.2025	42	8.5050
69-71	70	2.55	6.5025	27	175.5675
72-74	73	5.55	30.8025	8	246.4200
				$N = \sum f = 100$	$\sum f(X - \bar{X})^2 = 852.7500$

$$s = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}} = \sqrt{\frac{852.7500}{100}} = \sqrt{8.5275} = 2.92 \text{ in}$$

CALCULO DE LA DESVIACION TIPICA PARA DATOS AGRUPADOS

4.12. (*a*) Probar que

$$s = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2}$$

(*b*) Usar la fórmula en (*a*) para hallar la desviación típica del conjunto de números 12, 6, 7, 3, 15, 10, 18, 5.

Solución

(*a*) Por definición:

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

Entonces
$$s^2 = \frac{\sum (X - \bar{X})^2}{N} = \frac{\sum (X^2 - 2\bar{X}X + \bar{X}^2)}{N} = \frac{\sum X^2 - 2\bar{X}\sum X + N\bar{X}^2}{N} =$$

$$= \frac{\sum X^2}{N} - 2\bar{X}\frac{\sum X}{N} + \bar{X}^2 = \frac{\sum X^2}{N} - 2\bar{X}^2 + \bar{X}^2 = \frac{\sum X^2}{N} - \bar{X}^2 =$$

$$= \overline{X^2} = \bar{X}^2 = \frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2$$

o sea
$$s = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2}$$

Obsérvese que en las sumas precedentes hemos usado la forma abreviada, con X sustituyendo a X_j y $\sum$ a $\sum_{j=1}^{N}$.

Otro método

$$s^2 = \overline{(X - \bar{X})^2} = \overline{X^2 - 2X\bar{X} + \bar{X}^2} = \overline{X^2} - \overline{2X\bar{X}} + \overline{\bar{X}^2} = \overline{X^2} - 2\bar{X}\overline{X} + \bar{X}^2 = \overline{X^2} - \bar{X}^2$$

(*b*)

$$\overline{X^2} = \frac{\sum X^2}{N} = \frac{(12)^2 + (6)^2 + (7)^2 + (3)^2 + (15)^2 + (10)^2 + (18)^2 + (5)^2}{8} = \frac{912}{8} = 114$$

$$\bar{X} = \frac{\sum X}{N} = \frac{12 + 6 + 7 + 3 + 15 + 10 + 18 + 5}{8} = \frac{76}{8} = 9.5$$

Así pues, $$s = \sqrt{\overline{X^2} - \bar{X}^2} = \sqrt{114 - 90.25} = \sqrt{23.75} = 4.87$$

Compárese este método con el del Problema 4.9(*a*).

4.13. Modificar la fórmula del Problema 4.12(*a*) para permitir frecuencias asignadas a los diferentes valores de X.

Solución

La modificación adecuada es

$$s = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2}$$

Como en el Problema 4.12(*a*), ésta puede probarse partiendo de

$$s = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}}$$

Entonces $$s^2 = \frac{\sum f(X - \bar{X})^2}{N} = \frac{\sum f(X^2 - 2\bar{X}X + \bar{X}^2)}{N} = \frac{\sum fX^2 - 2\bar{X}\sum fX + \bar{X}^2 \sum f}{N} =$$

$$= \frac{\sum fX^2}{N} - 2\bar{X}\frac{\sum fX}{N} + \bar{X}^2 = \frac{\sum fX^2}{N} - 2\bar{X}^2 + \bar{X}^2 = \frac{\sum fX^2}{N} - \bar{X}^2 =$$

$$= \frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2$$

es decir $$s = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2}$$

Nótese que en las anteriores sumas se ha empleado la forma abreviada, con X y sustituyendo a X_j y f_j, $\sum$ a $\sum_{j=1}^{K}$ y $\sum_{j=1}^{K} f_j = N$.

4.14. Mediante la fórmula del Problema 4.13, hallar la desviación típica de los datos de la Tabla 4.2 del Problema 4.11.

Solución

Hágase como sugiere la Tabla 4.3, donde $\bar{X} = (\sum fX)/N = 67.45$ in, como se sigue del Problema 3.15. Nótese que este método, al igual que el del Problema 4.11, exige cálculos tediosos. El Problema 4.17 enseña que el método de compilación los simplifica en gran medida.

Tabla 4.3

Altura (in)	Marca de clase (X)	X^2	Frecuencia (f)	fX^2
60-62	61	3271	5	18,605
63-65	64	4096	18	73,728
66-68	67	4489	42	188,538
69-71	70	4900	27	132,300
72-74	73	5329	8	42,632
			$N = \sum f = 100$	$\sum fX^2 = 455{,}803$

$$s = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2} = \sqrt{\frac{455{,}803}{100} - (67.45)^2} = \sqrt{8.5275} = 2.92 \text{ in}$$

4.15. Si $d = X - A$ son las desviaciones de X respecto de una constante arbitraria A, probar que

$$s = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

Solución

Como $d = X - A$, $X = A + d$ y $\bar{X} = A + \bar{d}$ (véase Prob. 3.18), entonces

$$X - \bar{X} = (A + d) - (A + \bar{d}) = d - \bar{d}$$

así que

$$s = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}} = \sqrt{\frac{\sum f(d - \bar{d})^2}{N}} = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

usando el Problema 4.13 y sustituyendo X y $\bar{X}$ por d y $\bar{d}$, respectivamente.

Otro método

$$s^2 = \overline{(X - \bar{X})^2} = \overline{(d - \bar{d})^2} = \overline{d^2 - 2\bar{d}d + \bar{d}^2} =$$

$$= \overline{d^2} - 2\bar{d}^2 + \bar{d}^2 = \overline{d^2} - \bar{d}^2 = \frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2$$

y el resultado se sigue tomando la raíz cuadrada positiva.

4.16. Probar que si cada marca de clase X en una distribución de frecuencias con intervalos de clase de igual anchura c se compila en un valor asociado u según la relación $X = A + cu$, donde A es una marca de clase dada, entonces la desviación típica se escribe

$$s = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = c\sqrt{\overline{u^2} - \bar{u}^2}$$

Solución

Se deduce del Problema 4.15, ya que $d = X - A = cu$. Luego, al ser c constante,

$$s = \sqrt{\frac{\sum f(cu)^2}{N} - \left(\frac{\sum f(cu)}{N}\right)^2} = \sqrt{c^2\frac{\sum fu^2}{N} - c^2\left(\frac{\sum fu}{N}\right)^2} = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2}$$

Otro método

También se puede demostrar directamente sin apelar al Problema 4.15. Como $X = A + cu$, $\bar{X} = A + c\bar{u}$, y $X - \bar{X} = c(u - \bar{u})$, entonces

$$s^2 = \overline{(X - \bar{X})^2} = \overline{c^2(u - \bar{u})^2} = c^2\overline{(u^2 - 2\bar{u}u + \bar{u}^2)} = c^2(\overline{u^2} - 2\bar{u}^2 + \bar{u}^2) = c^2(\overline{u^2} - \bar{u}^2)$$

y

$$s = c\sqrt{\overline{u^2} - \bar{u}^2} = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2}$$

4.17. Hallar la desviación típica de las alturas de estudiantes de la Universidad XYZ (Tabla 2.1) mediante (*a*) la fórmula del Problema 4.15 y (*b*) el método del Problema 4.16.

Solución

En las Tablas 4.4 y 4.5, A se ha tomado arbitrariamente como la marca de clase 67. Nótese que en la Tabla 4.4 las desviaciones son todas múltiplos de la anchura del intervalo de clase $c = 3$. Ese factor se ha suprimido en la Tabla 4.5. En consecuencia, se simplifican muchos los cálculos de la Tabla 4.5 (a comparar con los de los Problemas 4.11 y 4.14). Por tal razón, el método de compilación es muy recomendable.

(*a*) Véase Tabla 4.4.

Tabla 4.4

Marca de clase (X)	$d = X - A$	Frecuencia (f)	fd	fd^2
61	−6	5	−30	180
64	−3	18	−54	162
$A \rightarrow 67$	0	42	0	0
70	3	27	81	243
73	6	8	48	288
		$N = \sum f = 100$	$\sum fd = 45$	$\sum fd^2 = 873$

$$s = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{873}{100} - \left(\frac{45}{100}\right)^2} = \sqrt{8.5275} = 2.92 \text{ in}$$

(*b*) Véase Tabla 4.5.

Tabla 4.5

Marca de clase (X)	$u = \dfrac{X - A}{c}$	Frecuencia (f)	fu	fu^2
61	−2	5	−10	20
64	−1	18	−18	18
$A \rightarrow 67$	0	42	0	0
70	1	27	27	27
73	2	8	16	32
		$N = \sum f = 100$	$\sum fu = 15$	$\sum fu^2 = 97$

$$s = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = 3\sqrt{\frac{97}{100} - \left(\frac{15}{100}\right)^2} = 3\sqrt{0.9475} = 2.92 \text{ in}$$

4.18. Por métodos de compilación, hallar (*a*) la media y (*b*) la desviación típica para la distribución de salarios del Problema 2.3.

Solución

La tarea es sencilla, como ilustra la Tabla 4.6.

Tabla 4.6

	X	u	f	fu	fu^2
	\$255.00	−2	8	−16	32
	265.00	−1	10	−10	10
$A \longrightarrow$	275.00	0	16	0	0
	285.00	1	14	14	14
	295.00	2	10	20	40
	305.00	3	5	15	45
	315.00	4	2	8	32
			$N = \sum f = 65$	$\sum fu = 31$	$\sum fu^2 = 173$

(*a*) $$\bar{X} = A + c\bar{u} = A + c\frac{\sum fu}{N} = \$275.00 + (\$10.00)\left(\frac{31}{65}\right) = \$279.77$$

(*b*) $$s = \sqrt{\bar{u^2} - \bar{u}^2} = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = (\$10.00)\sqrt{\frac{173}{65} - \left(\frac{31}{65}\right)^2} = (\$10.00)\sqrt{2.4341} = \$15.60$$

4.19. La Tabla 4.7 muestra los IQ (cocientes de inteligencia) de 480 niños de una escuela elemental. Mediante el método de compilación, hallar (*a*) la media y (*b*) la desviación típica.

Tabla 4.7

Marca de clase (X)	70	74	78	82	86	90	94	98	102	106	110	114	118	122	126
Frecuencia (f)	4	9	16	28	45	66	85	72	54	38	27	18	11	5	2

Solución

El cociente de inteligencia es

$$\text{IQ} = \frac{\text{edad mental}}{\text{edad cronológica}}$$

expresado como porcentaje. Por ejemplo, un niño de 8 años que (de acuerdo con ciertos procedimientos pedagógicos) tiene una mentalidad equivalente a uno de 10 años, tendría un IQ de $10/8 = 12.5 = 125\%$, o sencillamente 125, quedando sobreentendido el símbolo %.

Para hallar la media y la desviación típica de los IQ de la Tabla 4.7, podemos hacer lo que indica la Tabla 4.8.

(*a*) $$\bar{X} = A + c\bar{u} = A + c\frac{\sum fu}{N} = 94 + 4\left(\frac{236}{480}\right) = 95.97$$

(*b*) $$s = c\sqrt{\bar{u^2} - \bar{u}^2} = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = 4\sqrt{\frac{3404}{480} - \left(\frac{236}{480}\right)^2} = 4\sqrt{6.8499} = 10.47$$

Tabla 4.8

X	u	f	fu	fu^2
70	−6	4	−24	144
74	−5	9	−45	225
78	−4	16	−64	256
82	−3	28	−84	252
86	−2	45	−90	180
90	−1	66	−66	66
$A \longrightarrow$ 94	0	85	0	0
98	1	72	72	72
102	2	54	108	216
106	3	38	114	342
110	4	27	108	432
114	5	18	90	450
118	6	11	66	396
122	7	5	35	245
126	8	2	16	128
		$N = \sum f = 480$	$\sum fu = 236$	$\sum fu^2 = 3404$

COMPROBACION DE CHARLIER

4.20. Usar la comprobación de Charlier para verificar los cálculos de (*a*) la media y (*b*) la desviación típica, efectuados en el Problema 4.19.

Solución

Para aplicar esa comprobación hay que sumar las columnas de la Tabla 4.9 a las de la 4.8 (excepto la columna 2, que se repite en la Tabla 4.9 por conveniencia).

(*a*) De la Tabla 4.9, $\sum f(u + 1) = 716$; de la Tabla 4.8, $\sum fu + N = 236 + 480 = 716$. Eso da la requerida comprobación sobre la media.

(*b*) De la Tabla 4.9, $\sum f(u + 1)^2 = 4356$; de la Tabla 4.8, $\sum fu^2 + 2\sum fu + N = 3404 + 2(236) + 480 = 4356$. Lo cual proporciona la comprobación pedida sobre la desviación típica.

CORRECCIONES DE SHEPPARD PARA LA VARIANZA

4.21. Aplicar la corrección de Sheppard para determinar la desviación típica de los datos del (*a*) Problema 4.17, (*b*) Problema 4.18 y (*c*) Problema 4.19.

Solución

(*a*) $s^2 = 8.5275$ y $c = 3$. Varianza corregida $= s^2 - c^2/12 = 8.5275 - 3^2/12 = 7.7775$. Desviación típica corregida $= \sqrt{\text{varianza correcta}} = \sqrt{7.7775} = 2.79$ in.

(*b*) $s^2 = 243.41$ y $c = 10$. Varianza corregida $= s^2 - c^2/12 = 243.41 - 10^2/12 = 235.08$. Desviación típica corregida $= \sqrt{235.08} = \$15.33$.

(*c*) $s^2 = 109.60$ y $c = 4$. Varianza corregida $= s^2 - c^2/12 = 109.60 - 4^2/12 = 108.27$. Desviación típica corregida $= \sqrt{108.27} = 10.41$.

Tabla 4.9

$u + 1$	f	$f(u + 1)$	$f(u + 1)^2$
−5	4	−20	100
−4	9	−36	144
−3	16	−48	144
−2	28	−56	112
−1	45	−45	45
0	66	0	0
1	85	85	85
2	72	144	288
3	54	162	486
4	38	152	608
5	27	135	675
6	18	108	648
7	11	77	539
8	5	40	320
9	2	18	162
	$N = \sum f = 480$	$\sum f(u + 1) = 716$	$\sum f(u + 1)^2 = 4356$

4.22. Hallar, para la segunda distribución de frecuencias del Problema 2.8, (*a*) la media, (*b*) la desviación típica, (*c*) la desviación típica usando la corrección de Sheppard y (*d*) la verdadera desviación típica para los datos sin agrupar.

Solución

El trabajo lo resume la Tabla 4.10.

Tabla 4.10

X	u	f	fu	fu^2
122	−3	3	−9	27
131	−2	5	−10	20
140	−1	9	−9	9
$A \longrightarrow$ 149	0	12	0	0
158	1	5	5	5
167	2	4	8	16
176	3	2	6	18
		$N = \sum f = 40$	$\sum fu = -9$	$\sum fu^2 = 95$

(*a*) $$\bar{X} = A + c\bar{u} = A + c\frac{\sum fu}{N} = 149 + 9\left(\frac{-9}{40}\right) = 147.0 \text{ lb}$$

(*b*) $$s = c\sqrt{\bar{u^2} - \bar{u}^2} = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = 9\sqrt{\frac{95}{40} - \left(\frac{-9}{40}\right)^2} = 9\sqrt{2.324375} = 13.7 \text{ lb}$$

(*c*) Varianza corregida $= s^2 - c^2/12 = 188.27 - 9^2/12 = 181.52$. Desviación corregida típica = 13.5 lb.

(*d*) Para calcular la desviación típica de los propios datos originales, conviene restar primero un número adecuado, digamos $A = 150$ lb, de cada peso y usar entonces el método del Problema 4.15. Las desviaciones $d = X - A = X - 150$ son las que figuran en la siguiente tabla:

−12	14	0	−18	−6	−25	−1	7
−4	8	−10	−3	−14	−2	2	−6
18	−24	−12	26	13	−31	4	15
−4	23	−8	−3	−15	3	−10	−15
11	−5	−15	−8	0	6	−5	−22

de donde deducimos que $\sum d = -128$ y $\sum d^2 = 7052$. Entonces

$$s = \sqrt{\bar{d^2} - \bar{d}^2} = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} = \sqrt{\frac{7052}{40} - \left(\frac{-128}{40}\right)^2} = \sqrt{166.06} = 12.9 \text{ lb}$$

De modo que la corrección de Sheppard produce una cierta mejora en este caso.

RELACIONES EMPIRICAS ENTRE MEDIDAS DE DISPERSION

4.23. Para la distribución de alturas de la Universidad XYZ, discutir la validez de las fórmulas empiricas (*a*) desviación media = $\frac{4}{5}$(desviación típica) y (*b*) rango semi-intercuartil = $\frac{2}{3}$(desviación típica).

Solución

(*a*) De los Problemas 4.4 y 4.11, desviación media ÷ desviación típica = 2.26/2.92 = 0.77, que está cerca de 4/5.

(*b*) De los Problemas 4.6 y 4.11, rango semi-intercuartil ÷ desviación típica = 1.98/2.92 = 0.68, que es próximo a 2/3.

Luego las fórmulas empíricas son válidas en este caso.

Notemos que en lo anterior no hemos usado la desviación típica con corrección Sheppard para el agrupamiento, pues no se ha hecho corrección correspondiente para la desviación media o el rango semi-intercuartil.

PROPIEDADES DE LA DESVIACION TIPICA

4.24. Determinar el porcentaje de los IQ del Problema 4.19 que caen en los rangos (*a*) $\bar{X} \pm s$, (*b*) $\bar{X} \pm 2s$ y (*c*) $\bar{X} \pm 3s$.

Solución

(*a*) El rango de IQ desde 85.5 a 106.4 es $\bar{X} \pm s = 95.97 \pm 10.47$. El número de IQ en el rango $\bar{X} \pm s$ es

$$\left(\frac{88 - 85.5}{4}\right)(45) + 66 + 85 + 72 + 54 + \left(\frac{106.4 - 104}{4}\right)(38) = 339$$

El porcentaje de IQ en el rango $\bar{X} \pm s$ es 339/480 = 70.6%.

(*b*) El rango de IQ desde 75.0 a 116.9 es $\bar{X} \pm 2s = 95.97 \pm 2(10.47)$. El número de IQ en el rango $\bar{X} \pm 2s$ es

$$\left(\frac{76 - 75.0}{4}\right)(9) + 16 + 28 + 45 + 66 + 85 + 72 + 54 + 38 + 27 + 18 + \left(\frac{116.9 - 116}{4}\right)(11) = 451$$

El porcentaje de IQ en el rango $\bar{X} \pm 2s$ es 451/480 = 94.0%.

(*c*) El rango de IQ desde 64.6 a 127.4 es $\bar{X} \pm 3s = 95.97 \pm 3(10.47)$. El número de IQ en el rango $\bar{X} \pm 3s$ es

$$480 - \left(\frac{128 - 127.4}{4}\right)(2) = 479.7 \qquad \text{o sea} \qquad 480$$

El porcentaje de IQ en el rango $\bar{X} \pm 3s$ es 479.7/480 = 99.9%, es decir, prácticamente el 100 por 100.

Los porcentajes en las partes (*a*), (*b*) y (*c*) están en buen acuerdo con los esperados para una distribución normal: 68.27%, 95.45% y 99.73%, respectivamente.

Nótese que no hemos usado la corrección de Sheppard para la desviación típica. Si se usa, los resultados en este caso coinciden casi con lo obtenido aquí. Por cierto, que éstos pueden también obtenerse de la Tabla 4.11 del Problema 4.32.

4.25. Dados los conjuntos de números 2, 5, 8, 11, 14 y 2, 8, 14, hallar (*a*) la media de cada uno, (*b*) la varianza de cada uno, (*c*) la media combinada y (*d*) la varianza combinada

Solución

(*a*) Media del primer conjunto = $\frac{1}{5}(2 + 5 + 8 + 11 + 14) = 8$. Media del segundo conjunto = $\frac{1}{3}(2 + 8 + 14) = 8$.

(*b*) Varianza del primer conjunto = $s_1^2 = \frac{1}{5}[(2 - 8)^2 + (5 - 8)^2 + (8 - 8)^2 + (11 - 8)^2 + (14 - 8)^2] = 18$. Varianza del segundo conjunto = $s_2^2 = \frac{1}{3}[(2 - 8)^2 + (8 - 8)^2 + (14 - 8)^2] = 24$.

(*c*) La media de ambos conjuntos es

$$\frac{2 + 5 + 8 + 11 + 14 + 2 + 8 + 14}{5 + 3} = 8$$

(*d*) La varianza del conjunto total es

$$s^2 = \frac{(2 - 8)^2 + (5 - 8)^2 + (8 - 8)^2 + (11 - 8)^2 + (14 - 8)^2 + (2 - 8)^2 + (8 - 8)^2 + (14 - 8)^2}{5 + 3} = 20.25$$

Otro método (por fórmula)

$$s^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2} = \frac{(5)(18) + (3)(24)}{5 + 3} = 20.25$$

4.26. Resolver el Problema 4.25 para los conjuntos 2, 5, 8, 11, 14 y 10, 16, 22.

Solución

Aquí las medias de los dos conjuntos son 8 y 16, mientras que las varianzas son las mismas que las de los conjuntos del problema anterior, es decir, $s_1^2 = 18$ y $s_2^2 = 24$.

$$\text{Media de ambos conjuntos} = \frac{2 + 5 + 8 + 11 + 14 + 10 + 16 + 22}{5 + 3} = 11$$

$$s^2 = \frac{(2-11)^2 + (5-11)^2 + (8-11)^2 + (11-11)^2 + (14-11)^2 + (10-11)^2 + (16-11)^2 + (22-11)^2}{5+3} = 35.25$$

Nótese que la fórmula

$$s^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2}$$

que da el valor 20.25, *no* es aplicable en este caso porque las medias de los dos conjuntos *no* son iguales.

4.27. (*a*) Probar que $w^2 + pw + q$, donde p y q son constantes dadas, es un mínimo si y sólo si $w = -\frac{1}{2}p$.

(*b*) Usando la parte (*a*), probar que

$$\frac{\sum_{j=1}^{N} (X_j - a)^2}{N} \quad \text{o brevemente} \quad \frac{\sum (X - a)^2}{N}$$

es un mínimo si y sólo si $a = \bar{X}$.

Solución

(*a*) Tenemos $w^2 + pw + q = (w + \frac{1}{2}p)^2 + q - \frac{1}{4}p^2$. Como $(q - \frac{1}{4}p^2)$ es una constante, la expresión tiene el valor mínimo si y sólo si $w + \frac{1}{2}p = 0$ (i.e., $w = -\frac{1}{2}p$).

(*b*) $$\frac{\sum (X - a)^2}{N} = \frac{\sum (X^2 - 2aX + a^2)}{N} = \frac{\sum X^2 - 2a \sum X + Na^2}{N} = a^2 - 2a \frac{\sum X}{N} + \frac{\sum X^2}{N}$$

Comparando esta última expresión con $(w^2 + pw + q)$, se obtiene

$$w = a \qquad p = -2 \frac{\sum X}{N} \qquad q = \frac{\sum X^2}{N}$$

Así pues, la expresión es mínima cuando $a = -\frac{1}{2}p = (\sum X)/N = \bar{X}$, usando el resultado en (*a*).

DISPERSION ABSOLUTA Y RELATIVA: COEFICIENTE DE VARIACION

4.28. Un fabricante de tubos de televisión produce dos tipos de tubos, A y B, que tienen vidas medias respectivas $\bar{X}_A = 1495$ horas y $\bar{X}_B = 1875$ horas, y desviación típica de $s_A = 280$ horas y $s_B = 310$ horas. ¿Qué tubo tiene (*a*) mayor dispersión absoluta y (*b*) mayor dispersión relativa?

Solución

(*a*) La dispersión absoluta de A es $s_A = 280$ horas y la de B es $s_B = 310$ horas. Luego el tubo B tiene mayor dispersión absoluta.

(*b*) Los coeficientes de variación son

$$A = \frac{s_A}{\bar{X}_A} = \frac{280}{1495} = 18.7\% \qquad B = \frac{s_B}{\bar{X}_B} = \frac{310}{1875} = 16.5\%$$

Luego tiene más dispersión relativa el A.

4.29. Hallar los coeficientes de variación V para los datos del (*a*) Problema 4.14 y (*b*) Problema 4.18, usando tanto desviaciones típicas corregidas como no corregidas.

Solución

(*a*)
$$V(\text{sin corregir}) = \frac{s(\text{sin corregir})}{\bar{X}} = \frac{2.92}{67.45} = 0.0433 = 4.3\%$$

$$V(\text{corregido}) = \frac{s(\text{corregido})}{\bar{X}} = \frac{2.79}{67.45} = 0.0413 = 4.1\% \qquad \text{por el Problema 4.21}(a)$$

(*b*)
$$V(\text{sin corregir}) = \frac{s(\text{sin corregir})}{\bar{X}} = \frac{15.60}{79.77} = 0.196 = 19.6\%$$

$$V(\text{corregido}) = \frac{s(\text{corregido})}{\bar{X}} = \frac{15.33}{79.77} = 0.192 = 19.2\% \qquad \text{por el Problema 4.21}(b)$$

4.30. (*a*) Definir una medida de la dispersión relativa que pueda utilizarse para un conjunto de datos cuyos cuartiles son conocidos.

(*b*) Ilustrar el cálculo de la medida definida en (*a*) mediante los datos del Problema 4.6.

Solución

(*a*) Si Q_1 y Q_3 son conocidos para un conjunto de números, entonces $\frac{1}{2}(Q_1 + Q_3)$ es una medida de tendencia central de esos datos, o promedio, mientras que $Q = \frac{1}{2}(Q_3 - Q_1)$, el rango semi-intercuartil, es una medida de su dispersión. Podemos, pues, definir una medida de dispersión relativa como

$$V_Q = \frac{\frac{1}{2}(Q_3 - Q_1)}{\frac{1}{2}(Q_1 + Q_3)} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

que llamaremos el *coeficiente de variación cuartil*, o *coeficiente cuartil de dispersión relativa*.

(*b*)
$$V_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{69.61 - 65.64}{69.61 + 65.64} = \frac{3.97}{135.25} = 0.0293 = 2.9\%$$

VARIABLES TIPIFICADAS: UNIDADES ESTANDAR

4.31. Un estudiante obtuvo 84 puntos en el examen final de Matemáticas, en el que la nota media fue 76, y la desviación típica 10. En el examen final de Física obtuvo 90 puntos, siendo la media 82 y la desviación típica 16. ¿En qué examen sobresalió más?

Solución

La variable tipificada $z = (X - \bar{X})/s$ mide la desviación de X respecto de la media $\bar{X}$ en términos de la desviación típica s. En Matemáticas, $z = (84 - 76)/10 = 0.8$; para física, $z = (90 - 82)/16 = 0.5$. Luego su puntuación estaba 0.8 desviaciones típicas sobre la media en matemáticas y sólo 0.5 desviaciones típicas en física. Sobresalió más en matemáticas.

La variable $z = (X - \bar{X})/s$ se usa a menudo en niveles de enseñanza, donde se conoce como una *puntuación o recuento estándar*.

4.32. (*a*) Convertir los IQ del Problema 4.19 en un recuento estándar y (*b*) construir una gráfica de frecuencias relativas versus recuento estándar.

Solución

(*a*) La Tabla 4.11 resume el proceso de conversión. Añadidas a la tabla para su uso en la parte (*b*) están las marcas de clase de IQ 66 y 130, que tienen frecuencia cero. Asimismo, la corrección de Sheppard para la desviación típica no ha sido utilizada; las correcciones en este caso serían casi despreciables.

(*b*) El polígono de frecuencias relativas se muestra en la Figura 4.2. El eje horizontal se mide en términos de la desviación típica s como la unidad. Nótese que la distribución es poco asimétrica y algo sesgada a la derecha.

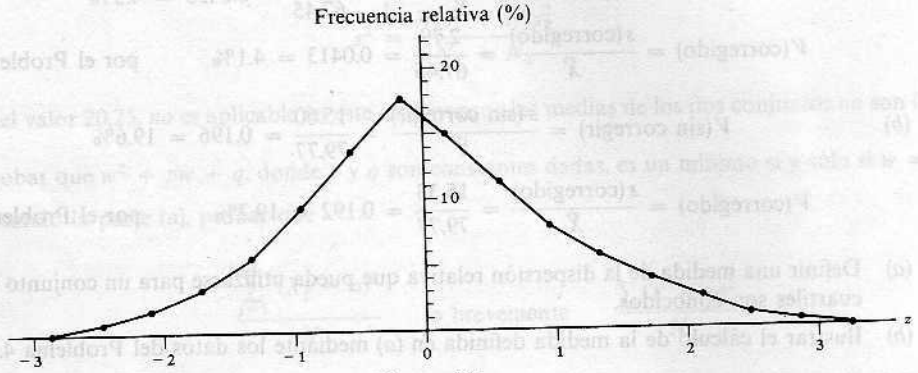


Figura 4.2.

Tabla 4.11. $\bar{X} = 96.0$, $s = 10.5$

IQ (X)	$X - \bar{X}$	$z = \dfrac{X - \bar{X}}{s}$	Frecuencia (f)	Frecuencia relativa (f)/N (%)
66	−30.0	−2.86	0	0.0
70	−26.0	−2.48	4	0.8
74	−22.0	−2.10	9	1.9
78	−18.0	−1.71	16	3.3
82	−14.0	−1.33	28	5.8
86	−10.0	−0.95	45	9.4
90	−6.0	−0.57	66	13.8
94	−2.0	−0.19	85	17.7
98	2.0	0.19	72	15.0
102	6.0	0.57	54	11.2
106	10.0	0.95	38	7.9
110	14.0	1.33	27	5.6
114	18.0	1.71	18	3.8
118	22.0	2.10	11	2.3
122	26.0	2.48	5	1.0
126	30.0	2.86	2	0.4
130	34.0	3.24	0	0.0
			480	100

PROBLEMAS SUPLEMENTARIOS

EL RANGO

4.33. Hallar el rango de los conjuntos de números (*a*) 5, 3, 8, 4, 7, 6, 12, 4, 3 y (*b*) 8.772, 6.453, 10.624, 8.628, 9.434, 6.351.

4.34. Hallar el rango de las cargas máximas del Problema 3.59, Tabla 3.8.

4.35. Hallar el rango de los diámetros de remaches del Problema 3.61, Tabla 3.10.

4.36. La mayor de 50 medidas es 8.34 kilogramos (kg). Si el rango es 0.46 kg, hallar la menor de esas medidas.

4.37. Determinar el rango de los datos en (*a*) Problema 3.62, (*b*) Problema 3.73 y (*c*) Problema 2.20.

LA DESVIACION MEDIA

4.38. Hallar los valores absolutos de (*a*) -18.2, (*b*) $+3.58$, (*c*) 6.21, (*d*) 0, (*e*) $-\sqrt{2}$ y (*f*) $4.00 - 2.36 - 3.52$.

4.39. Hallar la desviación media del conjunto (*a*) 3, 7, 9, 5 y (*b*) 2.4, 1.6, 3.8, 4.1, 3.4.

4.40. Hallar la desviación media de los conjuntos de números del Problema 4.33.

4.41. Hallar la desviación media de las cargas máximas del Problema 3.59, Tabla 3.8.

4.42. (*a*) Hallar la desviación media de los diámetros del Problema 3.61, Tabla 3.10.
(*b*) ¿Qué porcentaje de ellos está entre ($\bar{X} \pm$ $\pm$MD), ($\bar{X} \pm 2$ MD) y ($\bar{X} \pm 3$ MD)?

4.43. Para el conjunto de números 8, 10, 9, 12, 4, 8, 2, hallar la desviación media respecto de (*a*) la media y (*b*) la mediana. Verificar que la desviación media de la mediana no es mayor que la de la media.

4.44. Para la distribución de la Tabla 3.9, Problema 3.60, hallar la desviación media respecto de (*a*) la media y (*b*) la mediana. Usar los resultados de los Problemas 3.60 y 3.70.

4.45. Para la distribución de la Tabla 3.11, Problema 3.62, hallar la desviación media respecto de (*a*) la media y (*b*) la mediana. Usar los resultados de los Problemas 3.62 y 3.72.

4.46. Explicar por qué la desviación media es o no una buena medida de dispersión para la distribución de la Tabla 3.12 del Problema 3.73.

4.47. Deducir fórmulas de compilación para calcular la desviación media respecto de (*a*) la media y (*b*) la mediana, de una distribución de frecuencias. Aplicar estas fórmulas a la verificación de los resultados de los Problemas 4.44 y 4.45.

EL RANGO SEMI-INTERCUARTIL

4.48. Hallar el rango semi-intercuartil para la distribución del (*a*) Problema 3.59, (*b*) Problema 3.60 y (*c*) Problema 3.107. Interpretar los resultados claramente en cada caso.

4.49. Hallar el rango semi-intercuartil para la distribución de (*a*) Problema 2.31 y (*b*) Problema 3.73, interpretando los resultados en cada caso. Comparando con otras medidas de dispersión, explicar las ventajas del rango semi-intercuartil para este tipo de distribuciones.

4.50. Probar que para cualquier distribución de frecuencias el porcentaje total de casos que caen en el intervalo $\frac{1}{2}(Q_1 + Q_3) \pm \frac{1}{2}(Q_3 - Q_1)$ es 50%. ¿Es eso cierto para el intervalo $Q_2 \pm \frac{1}{2}(Q_3 - Q_1)$? Explicar la respuesta.

4.51. (*a*) ¿Cómo representaría el rango semi-intercuartil de una distribución de frecuencias dada?
(*b*) ¿Cuál es la relación del rango semi-intercuartil con la ojiva de la distribución?

EL RANGO PERCENTIL 10-90

4.52. Hallar el rango percentil 10-90 para las distribuciones de (*a*) Problema 3.59 y (*b*) Problema 3.107. Interpretar cada resultado.

4.53. Hallar el rango percentil 10-90 para las distribuciones de (*a*) Problema 2.31 y (*b*) Problema 3.73. Interpretar los resultados. ¿Qué ventajas y desventajas ofrece el rango percentil 10-90 frente a otras medidas de dispersión?

4.54. ¿Qué ventajas y desventajas tendría un rango percentil 20-80 comparado con el rango percentil 10-90?

4.55. Resolver el Problema 4.51 con referencia al (*a*) rango percentil 10-90, (*b*) rango percentil 20-80 y (*c*) rango percentil 25-75. ¿Cuál es la relación entre (*c*) y el rango semi-intercuartil?

LA DESVIACION TIPICA

4.56. Hallar la desviación típica de los conjuntos de números (*a*) 3, 6, 2, 1, 7, 5; (*b*) 3.2, 4.6, 2.8, 5.2, 4.4 y (*c*) 0, 0, 0, 0, 0, 1, 1, 1.

4.57. (*a*) Sumando 5 a cada número del conjunto 3, 6, 2, 1, 7, 5, obtenemos 8, 11, 7, 6, 12, 10. Probar que ambos conjuntos de números tienen la misma desviación típica pero diferentes medias. ¿Cómo están relacionadas las medias?

(*b*) Multiplicando cada número en 3, 6, 2, 1, 7, y 5 por 2 y sumando entonces 5, obtenemos el conjunto 11, 17, 9, 7, 19, 15. ¿Cuál es la relación entre la desviación típica y las medias de ambos conjuntos?

(*c*) ¿Qué propiedades de la media y de la desviación típica quedan ilustradas por los conjuntos particulares elegidos en las partes (*a*) y (*b*)?

4.58. Hallar la desviación típica del conjunto de números de la progresión aritmética 4, 10, 16, 22, ..., 154.

4.59. Hallar la desviación típica para las distribuciones de (*a*) Problema 3.59, (*b*) Problema 3.60 y (*c*) Problema 3.107.

4.60. Ilustrar el uso de la comprobación de Charlier en cada parte del Problema 4.59.

4.61. Hallar (*a*) la media y (*b*) la desviación típica para la distribución del Problema 2.17, y explicar la relevancia de los resultados obtenidos.

4.62. (*a*) Explicar por qué la desviación típica no es una medida apropiada de dispersión para la distribución del Problema 2.31.

(*b*) ¿Qué medida de dispersión debe utilizarse en su lugar? Ilustrar la respuesta.

4.63. (*a*) Hallar la desviación típica s de los diámetros de remaches de la Tabla 3.10.

(*b*) ¿Qué porcentajes de ellos cae entre $\bar{X} \pm s$, $\bar{X} \pm 2s$ y $\bar{X} \pm 3s$?

(*c*) Comparar los porcentajes de la parte (*b*) con los esperados teóricamente si la distribución fuese normal, y razonar la discrepancia.

4.64. Aplicar la corrección de Sheppard a cada desviación típica del Problema 4.59, y discutir en cada caso si tal aplicación está o no justificada.

4.65. ¿Qué modificaciones se producen en el Problema 4.63 al aplicar la corrección de Sheppard?

4.66. (*a*) Hallar la media y la desviación típica para los datos del Problema 2.8.

(*b*) Construir una distribución de frecuencias para los datos y hallar su desviación típica.

(*c*) Comparar los resultados de (*a*) y (*b*). Determinar si la aplicación de la corrección de Sheppard mejora los resultados.

4.67. Repetir el Problema 4.66 con los datos del Problema 2.27.

4.68. (*a*) De un total de N números, la fracción p son unos, y la fracción $q = 1 - p$ son ceros. Probar que la desviación típica de ese conjunto de números es $\sqrt{pq}$.

(*b*) Aplicar el resultado de (*a*) al Problema 4.56(*c*).

4.69. (*a*) Probar que la varianza de un conjunto de números a, $a + d$, $a + 2d$, ...

$a + (n - 1)d$ (o sea, una progresión aritmética con primer término a y razón d) viene dada por $\frac{1}{12}(n^2 - 1)d^2$.

(*b*) Usar (*a*) del Problema 4.58. [*Ayuda:* Use $1 + 2 + 3 + \cdots + (n - 1) = \frac{1}{2}n(n - 1)$, $1^2 + 2^2 + 3^2 + \cdots + (n - 1)^2 = \frac{1}{6}n(n - 1)(2n - 1)$.]

4.70. Generalizar y probar la Propiedad 3 de este capítulo (pág. 95).

RELACIONES EMPIRICAS ENTRE MEDIDAS DE DISPERSION

4.71. Comparando las desviaciones típicas obtenidas en el Problema 4.59 con las correspondientes desviaciones medias de los Problemas 4.41, 4.42 y 4.44, determinar si es válida la siguiente relación empírica: Desviación media $= \frac{4}{5}$(desviación típica). Explicar las posibles discrepancias.

4.72. Comparando las desviaciones típicas obtenidas en el Problema 4.59 con los correspondientes rangos semi-intercuartiles del Problema 4.48, determinar si es válida la relación empírica: rango semi-intercuartil $= \frac{2}{3}$(desviación típica). Explicar las posibles discrepancias.

4.73. ¿Qué relación empírica esperaría entre el rango semi-intercuartil y la desviación media de una distribución de frecuencias en forma de campana algo sesgada?

4.74. Una distribución de frecuencias que es casi normal tiene un rango semi-intercuartil igual a 10. ¿Qué valores esperaría para (*a*) la desviación típica y (*b*) la desviación media?

DISPERSION ABSOLUTA Y RELATIVA: COEFICIENTE DE VARIACION

4.75. En un examen final de Estadística, la puntuación media de 150 estudiantes fue de 78, y la desviación típica 8.0. En Algebra, la media fue 73 y la desviación típica 7.6. ¿En qué materia fue mayor (*a*) la dispersión absoluta y (*b*) la dispersión relativa?

4.76. Hallar el coeficiente de variación para los datos de (*a*) Problema 3.59 y (*b*) Problema 3.107.

4.77. (*a*) ¿Por qué no es posible calcular el coeficiente de variación para la distribución del Problema 2.31?

(*b*) Calcular el coeficiente cuartil de dispersión relativa para esta distribución. [Véanse Probs. 3.10(*c*) y 4.30.]

4.78. (*b*) Ilustrar el cálculo de tal medida con los datos del Problema 3.73.

VARIABLES TIPIFICADAS: UNIDADES ESTANDAR

4.79. En los exámenes a que se refiere el Problema 4.75, un alumno tuvo 75 en Estadística y 71 en Algebra. ¿En qué examen sobresalió más?

4.80. Convertir el conjunto 6, 2, 8, 7, 5 en un recuento estándar.

4.81. Probar que la media y la desviación típica de un recuento estándar son 0 y 1, respectivamente. Ilustrar esto mediante el Problema 4.80.

4.82. (*a*) Convertir las puntuaciones del Problema 3.107 en un recuento estándar y (*b*) construir un gráfico de frecuencias relativas versus ese recuento estándar.

CAPITULO 5

Momentos, sesgo y curtosis

MOMENTOS

Si $X_1, X_2, \ldots, X_N$ son los N valores de la variable X, definimos la cantidad

$$\overline{X^r} = \frac{X_1^r + X_2^r + \cdots + X_N^r}{N} = \frac{\sum_{j=1}^{N} X_j^r}{N} = \frac{\sum X^r}{N} \tag{1}$$

llamada r-ésimo *momento*. El primer momento, con $r = 1$, es la media aritmética $\bar{X}$.

El r-ésimo *momento respecto de la media* $\bar{X}$ se define como

$$m_r = \frac{\sum_{j=1}^{N} (X_j - \bar{X})^r}{N} = \frac{\sum (X - \bar{X})^r}{N} = \overline{(X - \bar{X})^r} \tag{2}$$

Si $r = 1$, entonces $m_1 = 0$ (véase Prob. 3.16). Si $r = 2$, entonces $m_2 = s^2$, la varianza.

El r-ésimo *momento respecto de cualquier origen* A se define como

$$m'_r = \frac{\sum_{j=1}^{N} (X_j - A)^r}{N} = \frac{\sum (X - A)^r}{N} = \frac{\sum d^r}{N} = \overline{(X - A)^r} \tag{3}$$

donde $d = X - A$ son las desviaciones de X respecto de A. Si $A = 0$, la ecuación (3) se reduce a la (1). Por esa razón, se suele llamar a (1) el r-ésimo *momento respecto de cero*.

MOMENTOS PARA DATOS AGRUPADOS

Si $X_1, X_2, \ldots, X_K$ ocurren con frecuencias $f_1, f_2, \ldots, f_K$, respectivamente, los momentos anteriores vienen dados por

$$\overline{X^r} = \frac{f_1 X_1^r + f_2 X_2^r + \cdots + f_K X_K^r}{N} = \frac{\sum_{j=1}^{K} f_j X_j^r}{N} = \frac{\sum f X^r}{N} \tag{4}$$

$$m_r = \frac{\sum_{j=1}^{K} f_j(X_j - \bar{X})^r}{N} = \frac{\sum f(X - \bar{X})^r}{N} = \overline{(X - \bar{X})^r} \tag{5}$$

$$m'_r = \frac{\sum_{j=1}^{K} f_j(X_j - A)^r}{N} = \frac{\sum f(X - A)^r}{N} = \overline{(X - A)^r} \tag{6}$$

donde $N = \sum_{j=1}^{K} f_j = \sum f$. Las fórmulas son adecuadas para calcular momentos en datos agrupados.

RELACIONES ENTRE MOMENTOS

Existen las siguientes relaciones entre momentos respecto de la media m_r y momentos respecto de un origen arbitrario m'_r:

$$\begin{aligned} m_2 &= m'_2 - m'^2_1 \\ m_3 &= m'_3 - 3m'_1m'_2 + 2m'^3_1 \\ m_4 &= m'_4 - 4m'_1m'_3 + 6m'^2_1m'_2 - 3m'^4_1 \end{aligned} \tag{7}$$

etcétera (véase Problema 5.5). Nótese que $m'_1 = \bar{X} - A$.

CALCULO DE MOMENTOS PARA DATOS AGRUPADOS

El método de compilación visto en capítulos precedentes para el cálculo de la media y de la desviación típica, puede usarse también como método breve para calcular momentos. Este método se apoya en que $X_j = A + cu_j$ (o más brevemente, $X = A + cu$), así que de la ecuación (6) tenemos

$$m'_r = c^r \frac{\sum fu^r}{N} = \overline{c^r u^r} \tag{8}$$

que puede utilizarse para hallar m_r aplicando las ecuaciones (7).

COMPROBACION DE CHARLIER Y CORRECCIONES DE SHEPPARD

La comprobación de Charlier para calcular momentos por compilación usa las identidades

$$\begin{aligned} \sum f(u+1) &= \sum fu + N \\ \sum f(u+1)^2 &= \sum fu^2 + 2\sum fu + N \\ \sum f(u+1)^3 &= \sum fu^3 + 3\sum fu^2 + 3\sum fu + N \\ \sum f(u+1)^4 &= \sum fu^4 + 4\sum fu^3 + 6\sum fu^2 + 4\sum fu + N \end{aligned} \tag{9}$$

Las correcciones de Sheppard para los momentos son como siguen:

$$m_2 \text{ corregido} = m_2 - \tfrac{1}{12}c^2 \qquad m_4 \text{ corregido} = m_4 - \tfrac{1}{2}c^2 m_2 + \tfrac{7}{240}c^4$$

Los momentos m_1 y m_3 no requieren corrección.

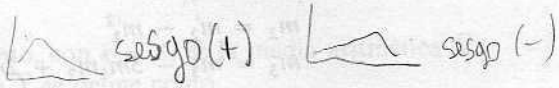
MOMENTOS ADIMENSIONALES

Para evitar unidades particulares, podemos definir los *momentos adimensionales* respecto de la media como

$$a_r = \frac{m_r}{s^r} = \frac{m_r}{(\sqrt{m_2})^r} = \frac{m_r}{\sqrt{m_2^r}} \tag{10}$$

donde $s = \sqrt{m_2}$ es la desviación típica. Ya que $m_1 = 0$ y $m_2 = s^2$, se tiene $a_1 = 0$ y $a_2 = 1$.

SESGO



Se conoce como *sesgo* el grado de asimetría de una distribución, es decir, cuánto se aparta de la simetría. Si la curva de frecuencias (polígono de frecuencias suavizado) de una distribución tiene a la derecha una cola más larga que a la izquierda, se dice *sesgada a la derecha*, o de *sesgo positivo*. En caso contrario, *sesgada a la izquierda*, o de *sesgo negativo*.

Para distribuciones sesgadas, la media tiende a estar del mismo lado de la moda que la cola larga (véanse Figs. 3.1 y 3.2). Luego una medida de la asimetría viene dada por la diferencia: media − moda, que puede hacerse adimensional dividiéndola por una medida de dispersión, tal como la desviación típica, lo que lleva a la definición

$$\text{Sesgo} = \frac{\text{media} - \text{moda}}{\text{desviación típica}} = \frac{\bar{X} - \text{moda}}{s} \tag{11}$$

Para evitar el uso de la moda, podemos recurrir a la fórmula empírica (10) del Capítulo 3 y definir

$$\text{Sesgo} = \frac{3(\text{media} - \text{mediana})}{\text{desviación típica}} = \frac{3(\bar{X} - \text{mediana})}{s} \tag{12}$$

Las ecuaciones (11) y (12) se llaman, respectivamente, *primer y segundo coeficientes de sesgo de Pearson*.

Otras medidas del sesgo, en términos de cuartiles y percentiles, son

$$\text{Coeficiente cuartil de sesgo} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} \tag{13}$$

$$\text{Coeficiente percentil 10-90 de sesgo} = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{P_{90} - P_{10}} = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} \tag{14}$$

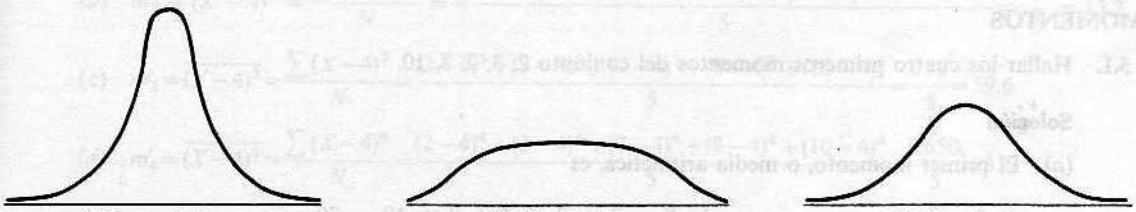
Una importante medida del sesgo usa el tercer momento respecto de la media expresado en forma adimensional y viene dado por

$$\text{Coeficiente momento de sesgo} = a_3 = \frac{m_3}{s^3} = \frac{m_3}{(\sqrt{m_2})^3} = \frac{m_3}{\sqrt{m_2^3}} \tag{15}$$

Otra usada a veces es $b_1 = a_3^2$. Para curvas perfectamente simétricas, como la curva normal, a_3 y b_1 son cero.

CURTOSIS

La *curtosis* mide cuán puntiaguda es una distribución, en general por referencia a la normal. Si tiene un pico alto, como la de la Figura 5.1(*a*), se dice *leptocúrtica*, mientras si es aplastada, como la de la Figura 5.1(*b*), se dice *platicúrtica*. La distribución normal, mostrada en la Figura 5.1(*c*), que no es ni muy puntiaguda ni muy aplastada, se llama *mesocúrtica*.



(*a*) Leptocúrtica (*b*) Platicúrtica (*c*) Mesocúrtica.

Figura 5.1.

Una medida de la curtosis utiliza el cuarto momento respecto de la media en forma adimensional y viene dada por

$$\text{Coeficiente momento de curtosis} = a_4 = \frac{m_4}{s^4} = \frac{m_4}{m_2^2} \tag{16}$$

que se suele denotar por b_2. Para la distribución normal, $b_2 = a_4 = 3$. De ahí que se defina a veces la curtosis como $(b_2 - 3)$, que es positivo para una distribución leptocúrtica, negativo para una platicúrtica y cero para la normal.

Otra medida de curtosis se basa en cuartiles y percentiles, y viene dada por

$$\kappa = \frac{Q}{P_{90} - P_{10}} \tag{17}$$

donde $Q = \frac{1}{2}(Q_3 - Q_1)$ es el rango semi-intercuartil. Nos referiremos a κ (letra griega minúscula *kappa*) como el *coeficiente percentil de curtosis;* para la distribución normal, κ vale 0.263 (véase Problema 5.14).

MOMENTOS, SESGO Y CURTOSIS DE UNA POBLACION

Cuando es necesario distinguir entre los momentos, medidas de sesgo y medidas de curtosis de una población y los de una muestra suya, se suelen usar símbolos latinos para los primeros y griegos para los segundos. Así, si los momentos de la muestra se denotan por m_r y m'_r, los correspondientes símbolos griegos serán μ_r y μ'_r (μ es la letra griega *mu*). Los subíndices serán siempre símbolos latinos.

Análogamente, si las medidas de sesgo y curtosis de la muestra se denotan a_3 y a_4, respectivamente, las de la población serán α_3 y α_4 (α es la letra griega *alfa*).

Ya sabemos, Capítulo 4, que la desviación típica de una muestra y de una población se denotan, respectivamente, por s y σ.

PROBLEMAS RESUELTOS

MOMENTOS

5.1. Hallar los cuatro primeros momentos del conjunto 2, 3, 7, 8, 10.

Solución

(*a*) El primer momento, o media aritmética, es

$$\bar{X} = \frac{\sum X}{N} = \frac{2 + 3 + 7 + 8 + 10}{5} = \frac{30}{5} = 6$$

(*b*) El segundo momento es

$$\overline{X^2} = \frac{\sum X^2}{N} = \frac{2^2 + 3^2 + 7^2 + 8^2 + 10^2}{5} = \frac{226}{5} = 45.2$$

(*c*) El tercer momento es

$$\overline{X^3} = \frac{\sum X^3}{N} = \frac{2^3 + 3^3 + 7^3 + 8^3 + 10^3}{5} = \frac{1890}{5} = 378$$

(*d*) El cuarto momento es

$$\overline{X^4} = \frac{\sum X^4}{N} = \frac{2^4 + 3^4 + 7^4 + 8^4 + 10^4}{5} = \frac{16{,}594}{5} = 3318.8$$

5.2. Hallar los cuatro primeros momentos respecto de la media para el conjunto de números del Problema 5.1.

Solución

(*a*) $m_1 = \overline{(X - \bar{X})} = \dfrac{\sum (X - \bar{X})}{N} = \dfrac{(2 - 6) + (3 - 6) + (7 - 6) + (8 - 6) + (10 - 6)}{5} = \dfrac{0}{5} = 0$

m_1 es siempre cero ya que $\overline{X - \bar{X}} = \bar{X} - \bar{X} = 0$ (véase Probl. 3.16).

(*b*) $m_2 = \overline{(X-\bar{X})^2} = \dfrac{\sum (X-\bar{X})^2}{N} = \dfrac{(2-6)^2+(3-6)^2+(7-6)^2+(8-6)^2+(10-6)^2}{6} = \dfrac{46}{5} = 9.2$

Nótese que m_2 es la varianza s^2.

(*c*) $m_3 = \overline{(X-\bar{X})^3} = \dfrac{\sum (X-\bar{X})^3}{N} = \dfrac{(2-6)^3+(3-6)^3+(7-6)^3+(8-6)^3+(10-6)^3}{5} = \dfrac{-18}{5} = -3.6$

(*d*) $m_4 = \overline{(X-\bar{X})^4} = \dfrac{\sum (X-\bar{X})^4}{N} = \dfrac{(2-6)^4+(3-6)^4+(7-6)^4+(8-6)^4+(10-6)^4}{5} = \dfrac{610}{5} = 122$

5.3. Hallar los cuatro primeros momentos respecto del origen para el conjunto de números del Problema 5.1.

Solución

(*a*) $m'_1 = \overline{(X-4)} = \dfrac{\sum (X-4)}{N} = \dfrac{(2-4)+(3-4)+(7-4)+(8-4)+(10-4)}{5} = 2$

(*b*) $m'_2 = \overline{(X-4)^2} = \dfrac{\sum (X-4)^2}{N} = \dfrac{(2-4)^2+(3-4)^2+(7-4)^2+(8-4)^2+(10-4)^2}{5} = \dfrac{66}{5} = 13.2$

(*c*) $m'_3 = \overline{(X-4)^3} = \dfrac{\sum (X-4)^3}{N} = \dfrac{(2-4)^3+(3-4)^3+(7-4)^3+(8-4)^3+(10-4)^3}{5} = \dfrac{298}{5} = 59.6$

(*d*) $m'_4 = \overline{(X-4)^4} = \dfrac{\sum (X-4)^4}{N} = \dfrac{(2-4)^4+(3-4)^4+(7-4)^4+(8-4)^4+(10-4)^4}{5} = \dfrac{1650}{5} = 330$

5.4. Usando los resultados de los Problemas 5.2 y 5.3, verificar las relaciones entre momentos: (*a*) $m_2 = m'_2 - m'^2_1$, (*b*) $m_3 = m'_3 - 3m'_1m'_2 + 2m'^3_1$ y (*c*) $m_4 = m'_4 - 4m'_1m'_3 + 6m'^2_1m'_2 - 3m'^4_1$.

Solución

Por el Problema 5.3 tenemos $m'_1 = 2$, $m'_2 = 13.2$, $m'_3 = 59.6$ y $m'_4 = 330$. Por tanto:

(*a*) $m_2 = m'_2 - m'^2_1 = 13.2 - (2)^2 = 13.2 - 4 = 9.2$.

(*b*) $m_3 = m'_3 - 3m'_1m'_2 + 2m'^3_1 = 59.6 - (3)(2)(13.2) + 2(2)^3 = 59.6 - 79.2 + 16 = -3.6$

(*c*) $m_4 = m'_4 - 4m'_1m'_3 + 6m'^2_1m'_2 - 3m'^4_1 = 330 - 4(2)(59.6) + 6(2)^2(13.2) - 3(2)^4 = 122$

de acuerdo con el Problema 5.2.

5.5. Probar que: (*a*) $m_2 = m'_2 - m'^2_1$, (*b*) $m_3 = m'_3 - 3m'_1m'_2 + 2m'^3_1$ y (*c*) $m_4 = m'_4 - 4m'_1m'_3 + 6m'^2_1m'_2 - 3m'^4_1$.

Solución

Si $d = X - A$, entonces $X = A + d$, $\bar{X} = A + \bar{d}$ y $X - \bar{X} = d - \bar{d}$. Luego:

(*a*)

$$m_2 = \overline{(X-\bar{X})^2} = \overline{(d-\bar{d})^2} = \overline{d^2 - 2\bar{d}d + \bar{d}^2}$$
$$= \overline{d^2} - 2\bar{d}^2 + \bar{d}^2 = \overline{d^2} - \bar{d}^2 = m'_2 - m'^2_1$$

(*b*)

$$m_3 = \overline{(X-\bar{X})^3} = \overline{(d-\bar{d})^3} = \overline{(d^3 - 3d^2\bar{d} + 3d\bar{d}^2 - \bar{d}^3)}$$
$$= \overline{d^3} - 3\bar{d}\overline{d^2} + 3\bar{d}^3 - \bar{d}^3 = \overline{d^3} - 3\bar{d}\overline{d^2} + 2\bar{d}^3 = m'_3 - 3m'_1m'_2 + 2m'^3_1$$

(*c*)

$$m_4 = \overline{(X - \bar{X})^4} = \overline{(d - \bar{d})^4} = \overline{(d^4 - 4d^3\bar{d} + 6d^2\bar{d}^2 - 4d\bar{d}^3 + \bar{d}^4)}$$
$$= \overline{d^4} - 4\bar{d}\overline{d^3} + 6\bar{d}^2\overline{d^2} - 4\bar{d}^4 + \bar{d}^4 = \overline{d^4} - 4\bar{d}\overline{d^3} + 6\bar{d}^2\overline{d^2} - 3\bar{d}^4$$
$$= m'_4 - 4m'_1m'_3 + 6m'^2_1m'_2 - 3m'^4_1$$

Por extensión de este método, se pueden deducir resultados similares para m_5, m_6, etc.

CALCULO DE MOMENTOS PARA DATOS AGRUPADOS

5.6. Hallar los cuatro primeros momentos respecto de la media para la distribución de alturas del Problema 3.22.

Solución

Tabla 5.1

X	u	f	fu	fu^2	fu^3	fu^4
61	−2	5	−10	20	−40	80
64	−1	18	−18	18	−18	18
67	0	42	0	0	0	0
70	1	27	27	27	27	27
73	2	8	16	32	64	128
		$N = \sum f = 100$	$\sum fu = 15$	$\sum fu^2 = 97$	$\sum fu^3 = 33$	$\sum fu^4 = 253$

El trabajo lo resume la Tabla 5.1, de la que vemos que

$$m'_1 = c\frac{\sum fu}{N} = (3)\left(\frac{15}{100}\right) = 0.45 \qquad m'_3 = c^3\frac{\sum fu^3}{N} = (3)^3\left(\frac{33}{100}\right) = 8.91$$

$$m'_2 = c^2\frac{\sum fu^2}{N} = (3)^2\left(\frac{97}{100}\right) = 8.73 \qquad m'_4 = c^4\frac{\sum fu^4}{N} = (3)^4\left(\frac{253}{100}\right) = 204.93$$

Así pues,

$$m_1 = 0$$
$$m_2 = m'_2 - m'^2_1 = 8.73 - (0.45)^2 = 8.5275$$
$$m_3 = m'_3 - 3m'_1m'_2 + m'^3_1 = 8.91 - 3(0.45)(8.73) + 2(0.45)^3 = -2.6932$$
$$m_4 = m'_4 - 4m'_1m'_3 + 6m'^2_1m'_2 - 3m'^4_1$$
$$= 204.93 - 4(0.45)(8.91) + 6(0.45)^2(8.73) - 3(0.45)^4 = 199.3759$$

5.7. Calcular: (*a*) m'_1, (*b*) m'_2, (*c*) m'_3, (*d*) m'_4, (*e*) m_1, (*f*) m_2, (*g*) m_3, (*h*) m_4, (*i*) $\bar{X}$, (*j*) s, (*k*) $\overline{X^2}$ y (*l*) $\overline{X^3}$ para la distribución de la Tabla 4.7 del Problema 4.19.

Solución

Procédase como indica la Tabla 5.2.

Tabla 5.2

X	u	f	fu	fu^2	fu^3	fu^4
70	−6	4	−24	144	−864	5184
74	−5	9	−45	225	−1125	5625
78	−4	16	−64	256	−1024	4096
82	−3	28	−84	252	−756	2268
86	−2	45	−90	180	−360	720
90	−1	66	−66	66	−66	66
$A \longrightarrow$ 94	0	85	0	0	0	0
98	1	72	72	72	72	72
102	2	54	108	216	432	864
106	3	38	114	342	1026	3078
110	4	27	108	432	1728	6912
114	5	18	90	450	2250	11250
118	6	11	66	396	2376	14256
122	7	5	35	245	1715	12005
126	8	2	16	128	1024	8192
		$N=\sum f=480$	$\sum fu=236$	$\sum fu^2=3404$	$\sum fu^3=6428$	$\sum fu^4=74{,}588$

$$(a)\quad m'_1 = c\frac{\sum fu}{N} = (4)\left(\frac{236}{480}\right) = 1.9667$$

$$(b)\quad m'_2 = c^2\frac{\sum fu^2}{N} = (4)^2\left(\frac{3404}{480}\right) = 113.4667$$

$$(c)\quad m'_3 = c^3\frac{\sum fu^3}{N} = (4)^3\left(\frac{6428}{480}\right) = 857.0667$$

$$(d)\quad m'_4 = c^4\frac{\sum fu^4}{N} = (4)^4\left(\frac{74{,}588}{480}\right) = 39{,}780.2667$$

$$(e)\quad m_1 = 0$$

$$(f)\quad m_2 = m'_2 - m'^2_1 = 113.4667 - (1.9667)^2 = 109.5988$$

$$(g)\quad m_3 = m'_3 - 3m'_1m'_2 + 2m'^3_1 = 857.0667 - 3(1.9667)(113.4667) + 2(1.9667)^3 = 202.8158$$

$$(h)\quad m_4 = m'_4 - 4m'_1m'_3 + 6m'^2_1m'_2 - 3m'^4_1 = 35{,}627.2853$$

$$(i)\quad \bar{X} = \overline{(A+d)} = A + m'_1 = A + c\frac{\sum fu}{N} = 94 + 1.9667 = 95.97$$

$$(j)\quad s = \sqrt{m_2} = \sqrt{109.5988} = 10.47$$

$$(k)\quad \overline{X^2} = \overline{(A+d)^2} = \overline{(A^2 + 2Ad + d^2)} = A^2 + 2A\bar{d} + \overline{d^2} = A^2 + 2Am'_1 + m'_2$$
$$= (94)^2 + 2(94)(1.9667) + 113.4667 = 9319.2063, \text{ o sea } 9319 \text{ con cuatro cifras significativas}$$

$$(l)\quad \overline{X^3} = \overline{(A+d)^3} = \overline{(A^3 + 3A^2d + 3Ad^2 + d^3)} = A^3 + 3A^2\bar{d} + 3A\overline{d^2} + \overline{d^3}$$
$$= A^3 + 3A^2m'_1 + 3Am'_2 + m'_3 = 915{,}571.9597, \text{ o sea } 915{,}600 \text{ con cuatro cifras significativas}$$

COMPROBACION DE CHARLIER

5.8. Ilustrar el uso de la comprobación de Charlier en los cálculos del Problema 5.7.

Solución

Para ello, sumamos a la Tabla 5.2 las columnas de la Tabla 5.3 (excepto la columna 2 que se repite en la Tabla 5.3 por conveniencia).

En cada uno de los siguientes agrupamientos, el primero está sacado de la Tabla 5.3 y el segundo de la Tabla 5.2. La igualdad de los resultados en cada grupo proporciona la deseada comprobación.

$$\sum f(u+1) = 716$$

$$\sum fu + N = 236 + 480 = 716$$

$$\sum f(u+1)^2 = 4356$$

$$\sum fu^2 + 2\sum fu + N = 3404 + 2(236) + 480 = 4356$$

$$\sum f(u+1)^3 = 17{,}828$$

$$\sum fu^3 + 3\sum fu^2 + 3\sum fu + N = 6428 + 3(3404) + 3(236) + 480 = 17{,}828$$

$$\sum f(u+1)^4 = 122{,}148$$

$$\sum fu^4 + 4\sum fu^3 + 6\sum fu^2 + 4\sum fu + N = 74{,}588 + 4(6428) + 6(3404) + 4(236) + 480 = 122{,}148$$

Tabla 5.3

$u+1$	f	$f(u+1)$	$f(u+1)^2$	$f(u+1)^3$	$f(u+1)^4$
−5	4	−20	100	−500	2500
−4	9	−36	144	−576	2304
−3	16	−48	144	−432	1296
−2	28	−56	112	−224	448
−1	45	−45	45	−45	45
0	66	0	0	0	0
1	85	85	85	85	85
2	72	144	288	576	1152
3	54	162	486	1458	4374
4	38	152	608	2432	9728
5	27	135	675	3375	16875
6	18	108	648	3888	23328
7	11	77	539	3773	26411
8	5	40	320	2560	20480
9	2	18	162	1458	13122
	$N = \sum f =$ $= 480$	$\sum f(u+1) =$ $= 716$	$\sum f(u+1)^2 =$ $= 4356$	$\sum f(u+1)^3 =$ $= 17828$	$\sum f(u+1)^4 =$ $= 122148$

CORRECCIONES DE SHEPPARD PARA LOS MOMENTOS

5.9. Aplicar las correcciones de Sheppard para determinar los momentos respecto de la media para los datos en: (*a*) Problema 5.6 y (*b*) Problema 5.7.

Solución

(*a*) m_2 corregido $= m_2 - c^2/12 = 8.5275 - 3^2/12 = 7.7775$

m_4 corregido $= m_4 - \frac{1}{2}c^2 m_2 + \frac{7}{240}c^4$

$= 199.3759 - \frac{1}{2}(3)^2(8.5275) + \frac{7}{240}(3)^4$

$= 163.3646$

m_1 y m_3 no necesitan corrección

(*b*) m_2 corregido $= m_2 - c^2/12 = 109.5988 - 4^2/12 = 108.2655$

m_4 corregido $= m_4 - \frac{1}{2}c^2 m_2 + \frac{7}{240}c^4$

$= 35{,}627.2853 - \frac{1}{2}(4)^2(109.5988) + \frac{7}{240}(4)^4$

$= 34{,}757.9616$

SESGO

5.10. Hallar el (*a*) primero y (*b*) segundo coeficientes de Pearson de sesgo para la distribución salarial de los 65 empleados de la empresa P&R (véanse Probs. 3.44 y 4.18).

Solución

Media = \$279.76, mediana = \$279.06, moda = \$277.50 y la desviación típica s = \$15.60. Así pues:

(*a*) Primer coeficiente de sesgo $= \dfrac{\text{media} - \text{moda}}{s} = \dfrac{\$279.76 - \$277.50}{\$15.60} = 0.1448$, o sea 0.14

(*b*) Segundo coeficiente de sesgo $= \dfrac{3(\text{media} - \text{mediana})}{s} = \dfrac{3(\$279.76 - \$279.06)}{\$15.60} = 0.1346$, o sea 0.13

Si se usa la desviación típica corregida [véase Prob. 4.21(*b*)], estos coeficientes pasan a ser, respectivamente:

(*a*) $\dfrac{\text{media} - \text{moda}}{s \text{ corregida}} = \dfrac{\$279.76 - \$277.50}{\$15.33} = 0.1474$, o sea 0,15

(*b*) $\dfrac{3(\text{media} - \text{mediana})}{s \text{ corregida}} = \dfrac{3(\$279.76 - \$279.06)}{\$15.33} = 0.1370$, o sea 0.14

Como los coeficientes son positivos, la distribución tiene sesgo positivo (o sea, a la derecha).

5.11. Hallar el coeficiente: (*a*) cuartil y (*b*) percentil de sesgo para la distribución del Problema 5.10 (véase Problema 3.44).

Solución

Q_1 = \$268.25, $Q_2 = P_{50}$ = \$279.06, Q_3 = \$290.75, $P_{10} = D_1$ = \$258.12 y $P_{90} = D_9$ = \$301.00. Luego:

(*a*) Coeficiente cuartil de sesgo $= \dfrac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} = \dfrac{\$290.75 - 2(\$279.06) + \$268.25}{\$290.75 - \$268.25} = 0.0391$

(*b*) Coeficiente percentil de sesgo $= \dfrac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} = \dfrac{\$301.00 - 2(\$279.06) + \$258.12}{\$301.00 - \$258.12} = 0.0233$

5.12. Hallar el coeficiente momento de sesgo a_3 para: (*a*) la distribución de alturas de estudiantes universitarios del Problema 5.6 y (*b*) los IQ de alumnos de escuela elemental del Problema 5.7.

Solución

(*a*) $m_2 = s^2 = 8.5275$ y $m_3 = -2.6932$. Luego

$$a_3 = \frac{m_3}{s^3} = \frac{m_3}{(\sqrt{m_2})^3} = \frac{-2.6932}{(\sqrt{8.5275})^3} = -0.1081 \quad \text{o sea} \quad -0.11$$

Si se usan correcciones de Sheppard para agrupar [véase Prob. 5.9(*a*)], entonces

$$a_3 \text{ corregido} = \frac{m_3}{(\sqrt{m_2 \text{ corregido}})^3} = \frac{-2.6932}{(\sqrt{7.7775})^3} = -0.1242 \quad \text{o sea} \quad -0.12$$

(*b*)

$$a_3 = \frac{m_3}{s^3} = \frac{m_3}{(\sqrt{m_2})^3} = \frac{202.8158}{(\sqrt{109.5988})^3} = 0.1768 \quad \text{o sea} \quad 0.18$$

Si se usan correcciones de Sheppard para agrupar [véase Prob. 5.9(*a*)], entonces

$$a_3 \text{ corregido} = \frac{m_3}{(\sqrt{m_2 \text{ corregido}})^3} = \frac{202.8158}{(\sqrt{108.2655})^3} = 0.1800 \quad \text{o sea} \quad 0.18$$

Nótese que ambas distribuciones son poco sesgadas, la (*a*) a la izquierda (negativamente) y la (*b*) a la derecha (positivamente). La (*b*) es más sesgada que la (*a*); esto es, (*a*) es más simétrica que (*b*), como queda patente por el hecho de que el valor numérico (o valor absoluto) del coeficiente de sesgo es mayor para (*b*) que para (*a*).

CURTOSIS

5.13. Hallar el coeficiente momento de curtosis a_4 para los datos de: (*a*) Problema 5.6 y (*b*) Problema 5.7.

Solución

(*a*)

$$a_4 = \frac{m_4}{s^4} = \frac{m_4}{m_2^2} = \frac{199.3759}{(8.5275)^2} = 2.7418 \quad \text{o sea} \quad 2.74$$

Si se usan correcciones de Sheppard [véase Prob. 5.9(*a*)], entonces

$$a_4 \text{ corregido} = \frac{m_4 \text{ corregido}}{(m_2 \text{ corregido})^2} = \frac{163.3646}{(7.7775)^2} = 2.7007 \quad \text{o sea} \quad 2.70$$

(*b*)

$$a_4 = \frac{m_4}{s^4} = \frac{m_4}{m_2^2} = \frac{35{,}627.2853}{(109.5988)^2} = 2.9660 \quad \text{o sea} \quad 2.97$$

Si se usan correcciones de Sheppard [véase Prob. 5.9(*b*)], entonces

$$a_4 \text{ corregido} = \frac{m_4 \text{ corregido}}{(m_2 \text{ corregido})} = \frac{34{,}757.9616}{(108.2655)^2} = 2.9653 \quad \text{o sea} \quad 2.97$$

Como para una distribución normal $a_4 = 3$, se sigue que ambas distribuciones, (*a*) y (*b*), son platicúrticas con respecto a la normal (o sea, más aplastadas que la distribución normal).

En lo referente a aplastamiento, la distribución (*b*) se aproxima a la normal mucho más que la (*a*). Sin embargo, sabemos del Problema 5.12 que en lo concerniente a la simetría, la (*a*) se aproxima más a la normal.

5.14. (*a*) Calcular el coeficiente percentil de curtosis $\kappa = Q/(P_{90} - P_{10})$, para la distribución del Problema 5.11.

(*b*) ¿Se aproximaría bien por una distribución normal?

Solución

(*a*) $Q = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(\$290.75 - \$268.25) = \$11.25$, $P_{90} - P_{10} = \$301.00 - \$258.12 = \$42.88$. Por tanto $\kappa = Q/(P_{90} - P_{10}) = 0.262$.

(*b*) Como para la distribución normal κ vale 0.263, se sigue que la distribución dada es mesocúrtica (o sea de aplastamiento más o menos normal). Así pues, la curtosis es la misma que para una distribución normal y nos lleva a creer que sería bien aproximada por ella, al menos en lo referente a curtosis.

PROBLEMAS SUPLEMENTARIOS

MOMENTOS

5.15. Hallar los cuatro primeros momentos del conjunto 4, 7, 5, 9, 8, 3, 6.

5.16. Hallar los cuatro primeros momentos respecto de la media para el conjunto de números del Problema 5.15.

5.17. Hallar los cuatro primeros momentos respecto del número 7 para el conjunto de números del Problema 5.15.

5.18. Usando los resultados de los Problemas 5.16 y 5.17, verificar las relaciones entre momentos: (*a*) $m_2 = m'_2 - m'^2_1$, (*b*) $m_3 = m'_3 - 3m'_1m'_2 + 2m'^3_1$ y (*c*) $m_4 = m'_4 - 4m'_1m'_3 + 6m'^2_1m'_2 - 3m'^4_1$.

5.19. Hallar los cuatro primeros momentos respecto de la media para el conjunto de números de la progresión aritmética 2, 5, 8, 11, 14, 17.

5.20. Probar que: (*a*) $m'_2 = m_2 + h^2$, (*b*) $m'_3 = m_3 + 3hm_2 + h^3$ y (*c*) $m'_4 = m_4 + 4hm_3 + 6h^2m_2 + h^4$, donde $h = m'_1$.

5.21. Si el primer momento respecto del número 2 es 5, ¿cuál es la media?

5.22. Si los primeros cuatro momentos de un conjunto de números respecto del número 3 son -2, 10, -25 y 50, determinar los correspondientes momentos respecto de: (*a*) la media, (*b*) el número 5 y (*c*) el cero.

5.23. Hallar los cuatro primeros momentos respecto de la media para el conjunto de números 0, 0, 0, 1, 1, 1, 1 y 1.

5.24. (*a*) Probar que $m_5 = m'_5 - 5m'_1m'_4 + 10m'^2_1m'_3 - 10m'^3_1m'_2 + 4m'^5_1$.

(*b*) Deducir una fórmula similar para m_6.

5.25. De un total de N números, la fracción p son unos y la fracción $q = 1 - p$ son ceros. Hallar: (*a*) m_1, (*b*) m_2, (*c*) m_3 y (*d*) m_4.

5.26. Probar que los primeros cuatro momentos respecto de la media de la progresión aritmética a, $a + d$, $a + 2d$, ..., $a + (n - 1)d$ son $m_1 = 0$, $m_2 = \frac{1}{12}(n^2 - 1)d^2$, $m_3 = 0$ y $m_4 = \frac{1}{240}(n^2 - 1)(3n^2 - 7)d^4$. Comparar

con el Problema 5.19 (véase también el Problema 4.69). [*Ayuda:* $1^4 + 2^4 + 3^4 + \cdots + (n-1)^4 = \frac{1}{30}n(n-1)(2n-1)(3n^2-3n-1)$.]

MOMENTOS PARA DATOS AGRUPADOS

5.27. Calcular los primeros cuatro momentos respecto de la media para la distribución de la Tabla 5.4.

Tabla 5.4

X	f
12	1
14	4
16	6
18	10
20	7
22	2
	Total 30

5.28. Ilustrar el uso de la comprobación de Charlier para los cálculos del Problema 5.27.

5.29. Aplicar las correcciones de Sheppard a los momentos obtenidos en el Problema 5.27.

5.30. Hallar los cuatro primeros momentos respecto de la media para la distribución del Problema 3.59: (*a*) sin correcciones de Sheppard y (*b*) con correcciones de Sheppard.

5.31. Hallar: (*a*) m_1, (*b*) m_2, (*c*) m_3, (*d*) m_4, (*e*) $\bar{X}$, (*f*) s, (*g*) $\overline{X^2}$, (*h*) $\overline{X^3}$, (*i*) $\overline{X^4}$ y (*j*) $\overline{(X+1)^3}$ para la distribución del Problema 3.62.

SESGO

5.32. Hallar el coeficiente momento de sesgo a_3 para la distribución del Problema 5.27: (*a*) sin y (*b*) con correcciones de Sheppard.

5.33. Hallar el coeficiente momento de sesgo a_3 para la distribución del Problema 3.59 (véase Problema 5.30).

5.34. Los segundos momentos respecto de la media de dos distribuciones son 9 y 16, mientras que los terceros momentos respecto de la media son -8.1 y -12.8, respectivamente. ¿Qué distribución es más sesgada a la izquierda?

5.35. Hallar los coeficientes de Pearson: (*a*) primero y (*b*) segundo, para la distribución del Problema 3.59, y explicar la diferencia.

5.36. Hallar el coeficiente de sesgo: (*a*) cuartil y (*b*) percentil, para la distribución del Problema 3.59. Comparar los resultados con los del Problema 5.35 y explicar lo que se aprecie.

5.37. (*a*) Explicar por qué los coeficientes de sesgo de Pearson no son apropiados para la distribución del Problema 2.31.
(*b*) Hallar el coeficiente cuartil de sesgo para ella e interpretar el resultado.

CURTOSIS

5.38. Hallar el coeficiente momento de curtosis a_4 para la distribución del Problema 5.27: (*a*) sin y (*b*) con correcciones de Sheppard.

5.39. Hallar el coeficiente momento de curtosis para la distribución del Problema 3.59: (*a*) sin y (*b*) con correcciones de Sheppard (véase Problema 5.30).

5.40. Los cuartos momentos respecto de la media de las distribuciones del Problema 5.34 son 230 y 780, respectivamente. ¿Qué distribución se aproxima más a la normal desde el punto de vista de: (*a*) aplastamiento y (*b*) sesgo?

5.41. ¿Cuál de las distribuciones del Problema 5.40 es: (*a*) leptocúrtica, (*b*) mesocúrtica y (*c*) platicúrtica?

5.42. La desviación típica de una distribución simétrica es 5. ¿Cuál debe ser el valor del cuarto momento respecto de la media para que la distribución sea: (*a*) leptocúrtica, (*b*) mesocúrtica y (*c*) platicúrtica?

5.43. (*a*) Calcular el coeficiente percentil de curtosis para la distribución del Problema 3.59.
(*b*) Comparar el resultado con el valor teórico 0.263 para la normal e interpretar.
(*c*) ¿Cómo se puede reconciliar este resultado con el del Problema 5.39?

CAPITULO 6

Teoría elemental de probabilidades

DEFINICIONES DE PROBABILIDAD

Definición clásica

Supongamos que un suceso E tiene h posibilidades de ocurrir entre un total de n posibilidades, cada una de las cuales tiene la misma oportunidad de ocurrir que las demás. Entonces, la probabilidad de que ocurra E (o sea un *éxito*) se denota por

$$p = \Pr\{E\} = \frac{h}{n}$$

La probabilidad de que no ocurra E (o sea, un *fracaso*) se denota por

$$q = \Pr\{\text{no } E\} = \frac{n-h}{n} = 1 - \frac{h}{n} = 1 - p = 1 - \Pr\{E\}$$

Así pues, $p + q = 1$, es decir, $\Pr\{E\} + \Pr\{\text{no } E\} = 1$. El suceso «no E» se denotará por $\bar{E}$, $\tilde{E}$ o $\sim E$.

EJEMPLO 1. Sea E el suceso de que al tirar un dado una vez salga un 3 o un 4. Hay seis formas de caer el dado, dando 1, 2, 3, 4, 5 ó 6; y si el dado es bueno (no trucado), como se supondrá en todo lo que sigue salvo mención explícita, podemos suponer que las seis tienen la misma oportunidad de salir. Como E puede ocurrir de dos formas, tenemos $p = \Pr\{E\} = \frac{2}{6} = \frac{1}{3}$.

La probabilidad de que no salga ni 3 ni 4 (o sea, de que salga 1, 2, 5 ó 6) es $q = \Pr\{\bar{E}\} = 1 - \frac{1}{3} = \frac{2}{3}$.

Nótese que la probabilidad de un suceso es un número entre 0 y 1. Si un suceso es imposible, su probabilidad es 0. Si un suceso debe ocurrir necesariamente (suceso seguro) su probabilidad es 1.

Si p es la probabilidad de que ocurra un suceso, las *apuestas* a su favor están $p:q$ (léase «p a q»). Luego las apuestas en su contra están $q:p$. Así, las apuestas contra la aparición de un 3 o un 4 al lanzar un dado bueno son $q:p = \frac{2}{3}:\frac{1}{3} = 2:1$ (o sea, 2 a 1).

Definición como frecuencia relativa

La definición clásica de probabilidad tiene la pega de que las palabras «misma oportunidad» aparecen como sinónimas de «equiprobables», lo cual produce un círculo vicioso. Por ello, algunos

defienden una definición estadística de la probabilidad. Para ellos, la probabilidad estimada, o *probabilidad empírica*, de un suceso se toma como la *frecuencia relativa* de ocurrencia del suceso cuando el número de observaciones es muy grande. La probabilidad misma es el *límite* de esa frecuencia relativa cuando el número de observaciones crece indefinidamente.

EJEMPLO 2. Si en 1000 tiradas de una moneda salen 529 caras, la frecuencia relativa de caras es 529/1000 = = 0.529. Si en otros 1000 lanzamientos salen 493 caras, la frecuencia relativa en el total de 2000 tiradas es (529 + + 493)/2000 = 0.511. De acuerdo con la definición estadística, continuando de este modo nos iremos acercando más y más a un número que representa la probabilidad de que salga cara en una sola tirada. De los resultados presentados, éste sería 0.5, con un dígito significativo. Para obtener más dígitos habría que hacer más tiradas.

La definición estadística, si bien útil en la práctica, tiene una desventaja matemática en el hecho de que un límite puede no existir. Por esa razón, la moderna teoría de la probabilidad es *axiomática* y deja el concepto de probabilidad sin definir, al igual que sucede en geometría con el punto y la recta.

PROBABILIDAD CONDICIONAL; SUCESOS INDEPENDIENTES Y SUCESOS DEPENDIENTES

Si E_1 y E_2 son dos sucesos, la probabilidad de que E_2 ocurra dado que haya ocurrido E_1 se denota por $\Pr\{E_2 \mid E_1\}$, o $\Pr\{E_2$ dado $E_1\}$, y se llama la *probabilidad condicional* de E_2 dado E_1.

Si la ocurrencia o no de E_1 no afecta para nada la probabilidad de ocurrencia de E_2, entonces $\Pr\{E_2 \mid E_1\} = \Pr\{E_2\}$, y diremos que E_1 y E_2 son *sucesos independientes;* en caso contrario, se dirá que son *sucesos dependientes*.

Si denotamos por E_1E_2 el suceso de que «ambos E_1 y E_2 ocurran», llamado un suceso compuesto, entonces

$$\Pr\{E_1E_2\} = \Pr\{E_1\} \Pr\{E_2 \mid E_1\} \tag{1}$$

En particular,

$$\Pr\{E_1E_2\} = \Pr\{E_1\} \Pr\{E_2\} \quad \text{para sucesos independientes} \tag{2}$$

Para tres sucesos E_1, E_2 y E_3, tenemos

$$\Pr\{E_1E_2E_3\} = \Pr\{E_1\} \Pr\{E_2 \mid E_1\} \Pr\{E_3 \mid E_1E_2\} \tag{3}$$

Esto es, la probabilidad de que ocurran E_1, E_2 y E_3 es igual a (la probabilidad de E_1) $\times$ (la probabilidad de E_2 dado E_1) $\times$ (la probabilidad de E_3 dados E_1 y E_2). En particular,

$$\Pr\{E_1E_2E_3\} = \Pr\{E_1\} \Pr\{E_2\} \Pr\{E_3\} \quad \text{para sucesos independientes} \tag{4}$$

En general, si E_1, E_2, E_3, ..., E_n son n sucesos independientes con probabilidades respectivas p_1, p_2, p_3, ..., p_n, entonces la probabilidad de que ocurran E_1 y E_2 y E_3 y $\cdots$ E_n es $p_1p_2p_3 \cdots p_n$.

EJEMPLO 3. Sean E_1 y E_2 los sucesos «cara en el quinto lanzamiento» y «cara en el sexto lanzamiento» de una moneda, respectivamente. Entonces, E_1 y E_2 son sucesos independientes y, por tanto, la probabilidad de que salga cara en ambos intentos (supuesta la moneda no trucada, aquí y en lo que sigue) es

$$\Pr\{E_1E_2\} = \Pr\{E_1\}\Pr\{E_2\} = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4}$$

EJEMPLO 4. Si las probabilidades de A y B de estar vivos dentro de 20 años son 0.7 y 0.5, respectivamente, entonces la probabilidad de que ambos lo estén es $(0.7)(0.5) = 0.35$.

EJEMPLO 5. Una caja contiene 3 bolas blancas y 2 bolas negras. Sea E_1 el suceso «la primera bola extraída es negra» y E_2 el suceso «la segunda bola extraída es negra». Las bolas extraídas no se devuelven a la caja. E_1 y E_2 son sucesos dependientes.

La probabilidad de que la primera bola sea negra es $\Pr\{E_1\} = 2/(3 + 2) = \frac{2}{5}$. La probabilidad de que la segunda sea negra, dado que ya lo haya sido la primera, es $\Pr\{E_2 \mid E_1\} = 1/(3 + 1) = \frac{1}{4}$. Luego la probabilidad de que ambas sean negras es

$$\Pr\{E_1E_2\} = \Pr\{E_1\}\Pr\{E_2 \mid E_1\} = \frac{2}{5}\cdot\frac{1}{4} = \frac{1}{10}$$

SUCESOS MUTUAMENTE EXCLUYENTES

Dos o más sucesos se llaman sucesos *mutuamente excluyentes* si la ocurrencia de cualquiera de ellos excluye la de los otros. De modo que si E_1 y E_2 son sucesos mutuamente excluyentes, entonces $\Pr\{E_1E_2\} = 0$.

Si $E_1 + E_2$ denota el suceso de que «ocurra E_1 o bien E_2 o ambos a la vez», entonces

$$\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1E_2\} \tag{5}$$

En particular,

$$\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\} \quad \text{para sucesos mutuamente excluyentes} \tag{6}$$

Como extensión de esto, si $E_1, E_2, \ldots, E_n$ son n sucesos mutuamente excluyentes con probabilidades respectivas E_1 o E_2 o $\cdots$ E_n es $p_1 + p_2 + \cdots + p_n$.

El resultado (5) se puede generalizar a tres o más sucesos mutuamente excluyentes (véase Problema 6.38).

EJEMPLO 6. Sean E_1 el suceso «sacar un as de una baraja» y E_2 «sacar un rey». Entonces $\Pr\{E_1\} = \frac{4}{52} = \frac{1}{13}$ y $\Pr\{E_2\} = \frac{4}{52} = \frac{1}{13}$. La probabilidad de sacar o un as o un rey en un solo ensayo es

$$\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\} = \frac{1}{13} + \frac{1}{13} = \frac{2}{13}$$

pues no es posible sacar ambos a la vez, y son, por tanto, sucesos mutuamente excluyentes.

EJEMPLO 7. Sean E_1 el suceso «sacar un as» de una baraja y E_2 «sacar una espada». Entonces E_1 y E_2 no son sucesos mutuamente excluyentes, porque puede sacarse el as de espadas. Luego la probabilidad de sacar un as o una espada o ambos es

$$\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1E_2\} = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$

DISTRIBUCIONES DE PROBABILIDAD

Discretas

Si una variable X puede tomar un conjunto discreto de valores $X_1, X_2, \ldots, X_K$, con probabilidades respectivas $p_1, p_2, \ldots, p_K$, donde $p_1 + p_2 + \cdots + p_K = 1$, decimos que tenemos definida una *distribución de probabilidad discreta* para X. La función $p(X)$, que tiene valores $p_1, p_2, \ldots, p_K$ para $X = X_1, X_2, \ldots, X_K$, se llama *función de probabilidad* o una *función de frecuencia* de X. Como X puede tomar ciertos valores con ciertas probabilidades, se le llama una *variable aleatoria discreta*. Una variable aleatoria se conoce también como *variable estocástica*.

EJEMPLO 8. Sea X la suma de puntos obtenida al lanzar dos dados. La distribución de probabilidad se muestra en la Tabla 6.1. Por ejemplo, la probabilidad de obtener suma 5 es $\frac{4}{36} = \frac{1}{9}$; así que en 900 tiradas se esperan 100 tiradas con suma 5.

Tabla 6.1

X	2	3	4	5	6	7	8	9	10	11	12
$p(X)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Nótese que esto es análogo a una distribución de frecuencias relativa, con probabilidad en lugar de frecuencia relativa. De manera que podemos pensar en las distribuciones de probabilidad como formas teóricas o ideales en el límite, de distribuciones de frecuencia relativa cuando el número de observaciones es muy grande. Por eso podemos pensar en las distribuciones de probabilidad como si fueran distribuciones de *poblaciones*, mientras que las distribuciones de frecuencia relativa son distribuciones de *muestras* de esa población.

La distribución de probabilidad se puede representar gráficamente dibujando $p(X)$ versus X, igual que para las distribuciones de frecuencia relativa (véase Prob. 6.11).

Acumulando probabilidades, obtenemos *distribuciones de probabilidad acumulada*, análogas a las distribuciones de frecuencia relativa acumulada. La función asociada con esa distribución se llama una *función de distribución*.

Continuas

Las ideas anteriores se extienden a variables X que pueden tomar un conjunto continuo de valores. El polígono de frecuencias relativas de una muestra se convierte, en el caso teórico o límite de una población, en una curva continua (como la de la Fig. 6.1) de ecuación $Y = p(X)$. El área total bajo esa curva y sobre el eje X es 1, y el área entre $X = a$ y $X = b$ (sombreada en la figura) da la probabilidad de que X esté entre a y b, que se denota por $\Pr\{a < X < b\}$.

Llamamos a $p(x)$ una *función densidad de probabilidad*, o brevemente una *función densidad*, y cuando tal función es dada decimos que se ha definido una *distribución de probabilidad continua* para X. La variable X se llama entonces una *variable aleatoria continua*.

Como en el caso discreto, podemos definir distribuciones de probabilidad acumulada y las asociadas funciones de distribución.

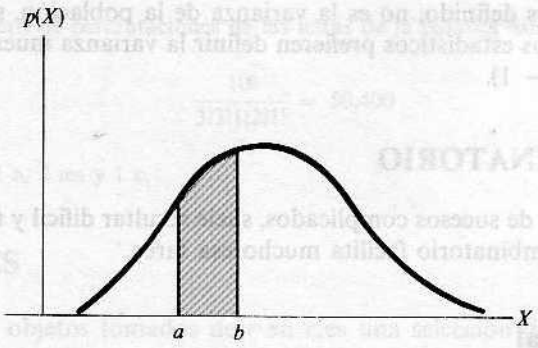


Figura 6.1.

ESPERANZA MATEMATICA

Si p es la probabilidad de que una persona reciba una cantidad S de dinero, la *esperanza matemática* (o simplemente *esperanza*) se define como pS.

EJEMPLO 9. Si la probabilidad de que un hombre gane un premio de \$10 es 1/5, su esperanza matemática es $\frac{1}{5}(\$10) = \$2$.

El concepto de esperanza matemática se extiende fácilmente. Si X denota una variable aleatoria discreta que puede tomar los valores $X_1, X_2, \ldots, X_K$ con probabilidades $p_1, p_2, \ldots, p_K$, donde $p_1 + p_2 + \cdots + p_K = 1$, la *esperanza matemática* de X (o simplemente *esperanza* de X), denotada $E(X)$, y se define como

$$E(X) = p_1X_1 + p_2X_2 + \cdots + p_KX_K = \sum_{j=1}^{K} p_jX_j = \sum pX \tag{7}$$

Si las probabilidades p_j en esa expresión se sustituyen por las frecuencias relativas f_j/N, donde $N = \sum f_j$, la esperanza matemática se reduce a $(\sum fX)/N$, que es la media aritmética $\bar{X}$ de una muestra de tamaño N en la que $X_1, X_2, \ldots, X_K$ aparecen con estas frecuencias relativas. Al crecer N más y más, las frecuencias relativas se acercan a las probabilidades p_j. Así que nos vemos abocados a interpretar $E(X)$ como la media de la población cuyo muestreo se consideraba. Si llamamos m a la media muestral, podemos denotar la media poblacional por la correspondiente letra griega μ (mu).

Puede definirse, asimismo, la esperanza matemática para variables aleatorias continuas, pero requiere el cálculo.

RELACION ENTRE POBLACION, MEDIA MUESTRAL Y VARIANZA

Si seleccionamos una muestra de tamaño N al azar de una población (o sea, suponemos que todas las posibles muestras son igualmente probables), entonces es posible mostrar que *el valor esperado de la media muestral m es la media poblacional* μ.

No se deduce, sin embargo, que el valor esperado de cualquier cantidad calculada sobre una muestra sea la cantidad correspondiente de la población. Así, el valor esperado de la varianza

muestral, como la hemos definido, no es la varianza de la población, sino $(N - 1)/N$ veces dicha varianza. Por eso algunos estadísticos prefieren definir la varianza muestral como nuestra varianza multiplicada por $N/(N - 1)$.

ANALISIS COMBINATORIO

Al hallar probabilidades de sucesos complicados, suele resultar difícil y tediosa una enumeración de los casos. El análisis combinatorio facilita mucho esa tarea.

Principio fundamental

Si un suceso puede ocurrir de n_1 maneras, y si cuando éste ha ocurrido otro suceso puede ocurrir de n_2 maneras, entonces el número de maneras en que ambos pueden ocurrir en el orden especificado es $n_1 n_2$.

EJEMPLO 10. Si hay 3 candidatos para gobernador y 5 para alcalde, los dos cargos pueden ocuparse de $3 \cdot 5 = 15$ formas.

Factorial de *n*

La factorial de n, denotada por $n!$, se define como

$$n! = n(n - 1)(n - 2) \cdots 1 \tag{8}$$

Así, $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$, y $4!3! = (4 \cdot 3 \cdot 2 \cdot 1)(3 \cdot 2 \cdot 1) = 144$. Conviene definir $0! = 1$.

Permutaciones

Una permutación de n objetos tomados de r en r es una *elección ordenada* de r objetos de entre n. El número de permutaciones de n objetos tomados de r en r se denota por ${}_nP_r$, $P(n, r)$, o $P_{n,r}$ y viene dado por

$${}_nP_r = n(n - 1)(n - 2) \cdots (n - r + 1) = \frac{n!}{(n - r)!} \tag{9}$$

En particular, el número de permutaciones de n objetos tomados de n en n es

$${}_n^nP = n(n - 1)(n - 2) \cdots 1 = n!$$

EJEMPLO 11. El números de permutaciones que se pueden dar de las letras a, b y c tomadas de dos en dos es ${}_3P_2 = 3 \cdot 2 = 6$. Son ab, ba, ac, ca, bc y cb.

El número de permutaciones de n objetos, de los que n_1 son iguales, n_2 son iguales, ... es

$$\frac{n!}{n_1!n_2!\cdots} \quad \text{donde } n = n_1 + n_2 + \cdots \tag{10}$$

EJEMPLO 12. El número de permutaciones de las letras de la palabra «statistics» es

$$\frac{10!}{3!3!1!2!1!} = 50{,}400$$

porque hay 3 eses, 3 tes, 1 a, 2 ies y 1 c.

COMBINACIONES

Una combinación de n objetos tomados de r en r es una selección de r de ellos, sin importar el orden de los r escogidos. El número de combinaciones de n objetos, tomados de r en r se denota por $\binom{n}{r}$ y viene dado por

$$\binom{n}{r} = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!} \tag{11}$$

EJEMPLO 13. El número de combinaciones de las letras a, b y c tomadas de dos en dos es

$$\binom{3}{2} = \frac{3 \cdot 2}{2!} = 3$$

Son ab, ac y bc. Nótese que ab es la misma combinación que ba, pero no la misma permutación.

APROXIMACION DE STIRLING A $n!$

Cuando n es grande, la evaluación directa de $n!$ es horrible. En tal caso, se usa una fórmula aproximada debida a James Stirling:

$$n! \approx \sqrt{2\pi n}\, n^n e^{-n} \tag{12}$$

donde $e = 2.71828\cdots$ es la base natural de logaritmos (véase Prob. 6.31).

RELACION DE LA PROBABILIDAD CON LA TEORIA DE CONJUNTOS

En la moderna teoría de probabilidad, se piensa en los posibles resultados de un ensayo, experimento, etc., como puntos de un espacio (que puede ser de 1, 2, 3, ..., dimensiones), llamado *espacio muestral* S. Si S contiene sólo un número finito de puntos, a cada punto está asociado un número no negativo, llamado *probabilidad*, tal que la suma de todos ellos es 1. Un suceso es un *conjunto* (o *colección*) de puntos de S, tal como E_1 o E_2 en la Figura 6.2; esa figura se llama un *diagrama de Euler* o de *Venn*.

El suceso $E_1 + E_2$ es el conjunto de puntos que *están en* E_1 *o en* E_2 *o en ambos*, y el suceso E_1E_2 es el conjunto de puntos *comunes a* E_1 *y a* E_2. Así que la probabilidad de un suceso tal como E_1 es la suma de las probabilidades asociadas a todos sus puntos. Análogamente, la probabilidad de $E_1 + E_2$, denotada $\Pr\{E_1 + E_2\}$, es la suma de las probabilidades asociadas a todos los puntos

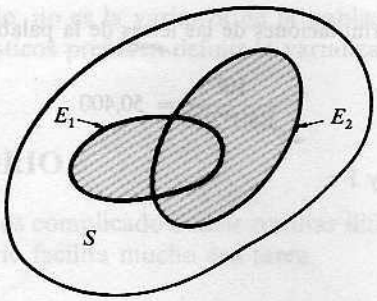


Figura 6.2.

contenidos en el conjunto $E_1 + E_2$. Si E_1 y E_2 no tienen puntos en común (o sea, si son sucesos mutuamente excluyentes), entonces $\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\}$. Si tienen puntos en común, entonces $\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1E_2\}$.

El conjunto $E_1 + E_2$ se denota a veces por $E_1 \cup E_2$ y se llama conjunto *unión* de los dos conjuntos. El conjunto E_1E_2 se suele denotar $E_1 \cap E_2$ y se llama *intersección* de los dos conjuntos. Cabe extender eso a más de dos conjuntos; así, en vez de $E_1 + E_2 + E_3$ y $E_1E_2E_3$, podríamos usar las notaciones $E_1 \cup E_2 \cup E_3$ y $E_1 \cap E_2 \cap E_3$, respectivamente.

El símbolo ϕ (letra griega *phi*) se usa para denotar el *conjunto vacío*, que no contiene punto alguno. La probabilidad asociada con un suceso correspondiente a este conjunto es cero (o sea, $\Pr\{\phi\} = 0$). Si E_1 y E_2 no tienen puntos en común, podemos escribir $E_1E_2 = \phi$, que significa que los correspondientes sucesos son sucesos mutuamente excluyentes, de donde $\Pr\{E_1E_2\} = 0$.

Con este enfoque moderno, una variable aleatoria es una función definida en cada punto del espacio muestral. Por ejemplo, en el Problema 6.36 la variable aleatoria es la suma de las coordenadas de cada punto.

En el caso de que S tenga infinitos puntos, lo anterior se extiende usando nociones del Cálculo.

PROBLEMAS RESUELTOS

REGLAS FUNDAMENTALES DE LA PROBABILIDAD

6.1. Determinar, o estimar, la probabilidad p de los siguientes sucesos:

(*a*) Una tirada de un dado resulte impar.

(*b*) Al menos una cara en dos tiradas de una moneda.

(*c*) Un as, el 10 de diamantes o el 2 de picas aparezca al sacar una sola carta de una baraja francesa de 52 naipes.

(*d*) La suma de dos dados sea 7.

(*e*) Que aparezca una cruz en la próxima tirada de una moneda si han salido 56 caras de 100 tiradas previas.

Solución

(*a*) De los 6 casos equiprobables, tres (si salen 1, 3 ó 5) son favorables al suceso. Luego $p = \frac{3}{6} = \frac{1}{2}$.

(*b*) Si H denota cara y T cruz, pueden salir HH, HT, TH y TT, con igual probabilidad. Sólo los tres primeros son favorables, luego $p = \frac{3}{4}$.

(c) El suceso puede ocurrir de 6 maneras (los 4 ases, el 10 de diamantes y el 2 de picas) de los 52 casos posibles. Luego $p = \frac{6}{52} = \frac{3}{26}$.

(d) Emparejando de todos los modos posibles las puntuaciones de los dos dados, hay $6 \cdot 6 = 36$ posibles casos. Pueden denotarse (1, 1), (2, 1), (3, 1), ..., (6, 6).

Las seis formas de que sumen 7 son (1, 6), (2, 5), (3, 4), (4, 3), (5, 2) y (6, 1) [véase Prob. 6.37(a)]. Luego $p = \frac{6}{36} = \frac{1}{6}$.

(e) Como salieron $100 - 56 = 44$ cruces en 100 tiradas, la probabilidad estimada (o empírica) de una cruz es la frecuencia relativa $44/100 = 0.44$.

6.2. Un experimento consiste en tirar un dado y una moneda. Si E_1 es el suceso «cara» al tirar la moneda, y E_2 es el suceso «3 ó 6» al tirar el dado, enunciar en palabras el significado de:

(a) $\bar{E}_1$ (c) E_1E_2 (e) $\Pr\{E_1 | E_2\}$

(b) $\bar{E}_2$ (d) $\Pr\{E_1\bar{E}_2\}$ (f) $\Pr\{\bar{E}_1 + \bar{E}_2\}$

Solución

(a) Cruz en la moneda y cualquier cosa en el dado.

(b) 1, 2, 4 ó 5 en el dado y cualquier cosa en la moneda.

(c) Cara en la moneda y 3 ó 6 en el dado.

(d) La probabilidad de cara en la moneda y 1, 2, 4 ó 5 en el dado.

(e) La probabilidad de cara en la moneda, dado que en el dado sale 3 ó 6.

(f) La probabilidad de cruz en la moneda o 1, 2, 4 ó 5 en el dado, o ambos.

6.3. Se saca al azar una bola de una caja que contiene 6 bolas rojas, 4 blancas y 5 azules. Hallar la probabilidad de que la bola extraída sea: (a) roja, (b) blanca, (c) azul, (d) no roja y (e) roja o blanca.

Solución

Denotemos R, W y B los sucesos de sacar una bola roja, blanca y azul, respectivamente. Entonces:

(a)
$$\Pr\{R\} = \frac{\text{formas de coger una bola roja}}{\text{formas totales de coger una bola}} = \frac{6}{6 + 4 + 5} = \frac{6}{15} = \frac{2}{5}$$

(b)
$$\Pr\{W\} = \frac{4}{6 + 4 + 5} = \frac{4}{15}$$

(c)
$$\Pr\{B\} = \frac{5}{6 + 4 + 5} = \frac{5}{15} = \frac{1}{3}$$

(d)
$$\Pr\{\bar{R}\} = 1 - \Pr\{R\} = 1 - \frac{2}{5} = \frac{3}{5} \quad \text{por la parte } (a)$$

(e)
$$\Pr\{R + W\} = \frac{\text{formas de coger una bola roja o una blanca}}{\text{formas totales de coger una bola}} = \frac{6 + 4}{6 + 4 + 5} = \frac{10}{15} = \frac{2}{3}$$

Otro método

$$\Pr\{R + W\} = \Pr\{\bar{B}\} = 1 - \Pr\{B\} = 1 - \frac{1}{3} = \frac{2}{3} \quad \text{por la parte } (c)$$

Nótese que $\Pr\{R + W\} = \Pr\{R\} + \Pr\{W\}$ (es decir, $\frac{2}{3} = \frac{2}{5} + \frac{4}{15}$). Esto ilustra la regla general $\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\}$ válida para sucesos mutuamente excluyentes E_1 y E_2.

6.4. Un dado se lanza dos veces. Hallar la probabilidad de obtener 4, 5 ó 6 en la primera tirada y 1, 2, 3 ó 4 en la segunda.

Solución

Sea E_1 = suceso «4, 5 ó 6» en la primera tirada, y E_2 = suceso «1, 2, 3 ó 4» en la segunda. Los diversos resultados de las dos tiradas se emparejan de $6 \times 6 = 36$ formas posibles, todas equiprobables. Las tres formas de salir el resultado apetecido en la primera y las cuatro de la segunda se emparejan de $3 \times 4 = 12$ formas, los casos favorables en que E_1 y E_2 ocurren ambos, es decir E_1E_2. Luego $\Pr\{E_1E_2\} = 12/36 = 1/3$.

Notemos que $\Pr\{E_1E_2\} = \Pr\{E_1\}\ \Pr\{E_2\}$ (es decir, $\frac{1}{3} = \frac{3}{6} \cdot \frac{4}{6}$) es válida para los *sucesos independientes* E_1 y E_2.

6.5. De una baraja de 52 naipes, mezclados al azar, se sacan dos naipes. Hallar la probabilidad de que ambos sean ases si la primera extraída: (*a*) se devuelve a la baraja y (*b*) si no se devuelve.

Solución

Sea E_1 = suceso «as» en la primera extracción, y E_2 = suceso «as» en la segunda.

(*a*) Si se repone, E_1 y E_2 son sucesos independientes. Así pues, $\Pr\{\text{ambas sean ases}\} = \Pr\{E_1E_2\} = \Pr\{E_1\}\ \Pr\{E_2\} = (\frac{4}{52})(\frac{4}{52}) = \frac{1}{169}$.

(*b*) Si no se repone, la primera carta se saca de entre 52 y la segunda de entre 51, luego ambas pueden sacarse de 52×51 formas, todas equiprobables.

Hay 4 casos favorables a E_1 y 3 a E_2, de modo que ambos, E_1 y E_2, o sea E_1E_2, pueden ocurrir de 4×3 formas. Luego $\Pr\{E_1E_2\} = (4 \cdot 3)/(52 \cdot 51) = \frac{1}{221}$.

Nótese que $\Pr\{E_2 \mid E_1\} = \Pr\{\text{la segunda es un as dado que la primera era un as}\} = \frac{3}{51}$. Por tanto, nuestro resultado ilustra la regla general de que $\Pr\{E_1E_2\} = \Pr\{E_1\}\ \Pr\{E_2 \mid E_1\}$ cuando E_1 y E_2 son sucesos dependientes.

6.6. Se sacan sucesivamente 3 bolas de la caja del Problema 6.3. Hallar la probabilidad de que salgan en el orden roja, blanca, azul si cada bola: (*a*) se repone y (*b*) no se repone.

Solución

Sea R = suceso «roja» en la primera extracción, W = suceso «blanca» en la segunda y B = suceso «azul» en la tercera. Se pide $\Pr\{RWB\}$.

(*a*) Con reposición, R, W y B son sucesos independientes, luego

$$\Pr\{RWB\} = \Pr\{R\}\ \Pr\{W\}\ \Pr\{B\} = \left(\frac{6}{6+4+5}\right)\left(\frac{4}{6+4+5}\right)\left(\frac{5}{6+4+5}\right) = \left(\frac{6}{15}\right)\left(\frac{4}{15}\right)\left(\frac{5}{15}\right) = \frac{8}{225}$$

(*b*) Sin reposición, R, W y B son sucesos dependientes y

$$\Pr\{RWB\} = \Pr\{R\}\ \Pr\{W \mid R\}\ \Pr\{B \mid WR\} = \left(\frac{6}{6+4+5}\right)\left(\frac{4}{5+4+5}\right)\left(\frac{5}{5+3+5}\right)$$

$$= \left(\frac{6}{15}\right)\left(\frac{4}{14}\right)\left(\frac{5}{13}\right) = \frac{4}{91}$$

donde $\Pr\{B \mid WR\}$ es la probabilidad condicional de sacar una azul si ya han salido una blanca y una roja.

6.7. Hallar la probabilidad de que salga al menos un 4 en dos tiradas de un dado.

Solución

Sea E_1 = suceso «4» en la primera tirada, E_2 = suceso «4» en la segunda y $E_1 + E_2$ = suceso «4» en la primera o «4» en la segunda o en ambas = suceso de que salga al menos un 4. Se pide $\Pr\{E_1 + E_2\}$.

Primer método

El número de formas en que pueden salir los dos dados es $6 \times 6 = 36$. Además,

Número de formas de que salga E_1 pero no $E_2 = 5$
Número de formas de que salga E_2 pero no $E_1 = 5$
Número de formas de que salgan ambos E_1 y $E_2 = 1$

Luego el número de formas en que al menos uno de ellos sale es $5 + 5 + 1 = 11$ y, por tanto, $\Pr\{E_1 + E_2\} = \frac{11}{36}$.

Segundo método

Como E_1 y E_2 no son sucesos mutuamente excluyentes, $\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1E_2\}$. Además, como E_1 y E_2 son sucesos independientes, $\Pr\{E_1E_2\} = \Pr\{E_1\}\Pr\{E_2\}$. Entonces $\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1\}\Pr\{E_2\} = \frac{1}{6} + \frac{1}{6} - (\frac{1}{6})(\frac{1}{6}) = \frac{11}{36}$.

Tercer método

$$\Pr\{\text{salir al menos un 4}\} + \Pr\{\text{no salga ningún 4}\} = 1$$

Por tanto

$$\begin{aligned}\Pr\{\text{al menos un 4}\} &= 1 - \Pr\{\text{ningún 4}\} \\ &= 1 - \Pr\{\text{ni 4 en la primera ni 4 en la segunda}\} \\ &= 1 - \Pr\{\bar{E}_1\bar{E}_2\} = 1 - \Pr\{\bar{E}_1\}\Pr\{\bar{E}_2\} \\ &= 1 - \left(\frac{5}{6}\right)\left(\frac{5}{6}\right) = \frac{11}{36}\end{aligned}$$

6.8. Una bolsa contiene 4 bolas blancas y 2 bolas negras; otra contiene 3 bolas blancas y 5 bolas negras. Si se saca una bola de cada bolsa, hallar la probabilidad de que: (*a*) ambas sean blancas, (*b*) ambas sean negras y (*c*) una sea blanca y la otra negra.

Solución

Sea W_1 = suceso «bola blanca» de la primera bolsa y W_2 = suceso «bola blanca» de la segunda.

(*a*)

$$\Pr\{W_1W_2\} = \Pr\{W_1\}\Pr\{W_2\} = \left(\frac{4}{4+2}\right)\left(\frac{3}{3+5}\right) = \frac{1}{4}$$

(*b*)

$$\Pr\{\bar{W}_1\bar{W}_2\} = \Pr\{\bar{W}_1\}\Pr\{\bar{W}_2\} = \left(\frac{2}{4+2}\right)\left(\frac{5}{3+5}\right) = \frac{5}{24}$$

(*c*) El suceso «una es blanca y la otra negra» es el mismo que «o la primera es blanca, o la segunda es negra o la primera negra y la segunda blanca»; esto es, $W_1\bar{W}_2 + \bar{W}_1W_2$. Como $W_1\bar{W}_2$ y $\bar{W}_1W_2$ son sucesos mutuamente excluyentes, tenemos

$$\begin{aligned}\Pr\{W_1\bar{W}_2 + \bar{W}_1W_2\} &= \Pr\{W_1\bar{W}_2\} + \Pr\{\bar{W}_1W_2\} \\ &= \Pr\{W_1\}\Pr\{\bar{W}_2\} + \Pr\{\bar{W}_1\}\Pr\{W_2\} \\ &= \left(\frac{4}{4+2}\right)\left(\frac{5}{3+5}\right) + \left(\frac{2}{4+2}\right)\left(\frac{3}{3+5}\right) = \frac{13}{24}\end{aligned}$$

Otro método

La probabilidad pedida es $1 - \Pr\{W_1W_2\} - \Pr\{\bar{W}_1\bar{W}_2\} = 1 - \frac{1}{4} - \frac{5}{24} = \frac{13}{24}$.

6.9. *A* y *B* juegan 12 partidas de ajedrez. *A* gana 6, *B* gana 4 y en 2 hacen tablas. Acuerdan jugar un torneo de 3 partidas. Hallar la probabilidad de que: (*a*) *A* gane las 3, (*b*) hagan tablas en 2, (*c*) *A* y *B* ganen alternadamente y (*d*) *B* gane al menos 1 partida.

Solución

Denotemos por A_1, A_2 y A_3 los sucesos «*A* gana» en la primera, segunda y tercera partidas, respectivamente; y por B_1, B_2 y B_3 lo análogo para *B*. Sean T_1, T_2 y T_3 los sucesos «tablas» en las tres partidas sucesivas.

Sobre la base de su experiencia pasada (probabilidad empírica), supondremos que $\Pr\{A \text{ gana cualquier partida}\} = \frac{6}{12} = \frac{1}{2}$, que $\Pr\{B \text{ gana cualquier partida}\} = \frac{4}{12} = \frac{1}{3}$, y que $\Pr\{\text{tablas en cualquier partida}\} = \frac{2}{12} = \frac{1}{6}$.

(*a*) $\Pr\{A \text{ gane los 3 juegos}\} = \Pr\{A_1A_2A_3\} = \Pr\{A_1\}\Pr\{A_2\}\Pr\{A_3\} = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{8}$

suponiendo que los resultados de cada partida sean independientes, lo cual parece justificable (a menos que los jugadores se dejen *influir psicológicamente* por las derrotas).

(*b*) $\Pr\{\text{tablas en 2 partidas}\} = \Pr\{\text{1.ª y 2.ª en tablas, o 1.ª y 3.ª en tablas, o 2.ª y 3.ª en tablas}\}$

$$\begin{aligned}&= \Pr\{T_1T_2\bar{T}_3\} + \Pr\{T_1\bar{T}_2T_3\} + \Pr\{\bar{T}_1T_2T_3\} \\ &= \Pr\{T_1\}\Pr\{T_2\}\Pr\{\bar{T}_3\} + \Pr\{T_1\}\Pr\{\bar{T}_2\}\Pr\{T_3\} + \Pr\{\bar{T}_1\}\Pr\{T_2\}\Pr\{T_3\} \\ &= \left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\left(\frac{5}{6}\right) + \left(\frac{1}{6}\right)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right) + \left(\frac{5}{6}\right)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{15}{216} = \frac{5}{72}\end{aligned}$$

(*c*) $\Pr\{A \text{ y } B \text{ ganan alternadamente}\} = \Pr\{\text{ganan } ABA \text{ o ganan } BAB\}$

$$\begin{aligned}&= \Pr\{A_1B_2A_3 + B_1A_2B_3\} = \Pr\{A_1B_2A_3\} + \Pr\{B_1A_2B_3\} \\ &= \Pr\{A_1\}\Pr\{B_2\}\Pr\{A_3\} + \Pr\{B_1\}\Pr\{A_2\}\Pr\{B_2\} \\ &= \left(\frac{1}{2}\right)\left(\frac{1}{3}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{2}\right)\left(\frac{1}{3}\right) = \frac{5}{36}\end{aligned}$$

(*d*) $\Pr\{B \text{ gana al menos 1 partida}\} = 1 - \Pr\{B \text{ pierde las tres}\}$

$$\begin{aligned}&= 1 - \Pr\{\bar{B}_1\bar{B}_2\bar{B}_3\} = 1 - \Pr\{\bar{B}_1\}\Pr\{\bar{B}_2\}\Pr\{\bar{B}_3\} \\ &= 1 - \left(\frac{2}{3}\right)\left(\frac{2}{3}\right)\left(\frac{2}{3}\right) = \frac{19}{27}\end{aligned}$$

DISTRIBUCIONES DE PROBABILIDAD

6.10. Hallar la probabilidad de cada reparto en chicos y chicas en familias con 3 hijos, supuesta igual probabilidad para ambos.

Solución

Sea B = suceso «chico» y G = suceso «chica». De acuerdo con la hipótesis de igual probabilidad, $\Pr\{B\} = \Pr\{G\} = \frac{1}{2}$. En familias de 3 hijos, pueden ocurrir los siguientes sucesos mutuamente excluyentes con las probabilidades indicadas:

(*a*) Tres chicos (BBB):

$$\Pr\{BBB\} = \Pr\{B\}\ \Pr\{B\}\ \Pr\{B\} = \frac{1}{8}$$

Aquí suponemos que el nacimiento de cada hijo es independiente de los demás nacimientos.

(*b*) Tres chicas (GGG): Como en la parte (*a*) por simetría,

$$\Pr\{GGG\} = \frac{1}{8}$$

(*c*) Dos chicos y una chica ($BBG + BGB + GBB$):

$$\begin{aligned}\Pr\{BBG + BGB + GBB\} &= \Pr\{BBG\} + \Pr\{BGB\} + \Pr\{GBB\} \\ &= \Pr\{B\}\ \Pr\{B\}\ \Pr\{G\} + \Pr\{B\}\ \Pr\{G\}\ \Pr\{B\} + \Pr\{G\}\ \Pr\{B\}\ \Pr\{B\} \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}\end{aligned}$$

(*d*) Dos chicas y un chico ($GGB + GBG + BGG$): Como en la parte (*c*) o por simetría, la probabilidad es $\frac{3}{8}$.

Si llamamos X a la *variable aleatoria* que indica el número de chicos en cada familia de 3 hijos, su distribución de probabilidad se muestra en la Tabla 6.2.

Tabla 6.2

Número de chicos X	0	1	2	3
Probabilidad $p(X)$	1/8	3/8	3/8	1/8

6.11. Representar la distribución del Problema 6.10.

Solución

El gráfico puede representarse como en la Figura 6.3 o como en la 6.4. La suma de las áreas de los rectángulos de la Figura 6.4 es 1; en ella, llamada un *histograma de probabilidad*, estamos considerando a X como una variable continua aunque es discreta en verdad, un procedimiento que resulta útil a menudo. La Figura 6.3, por su lado, se usa cuando uno no quiere tratar la variable como continua.

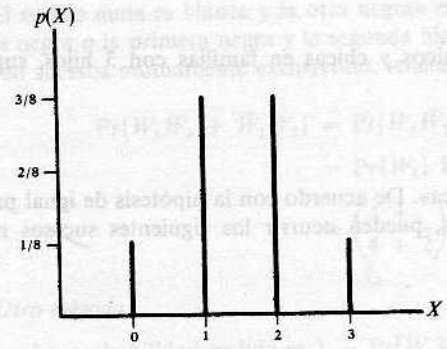


Figura 6.3.

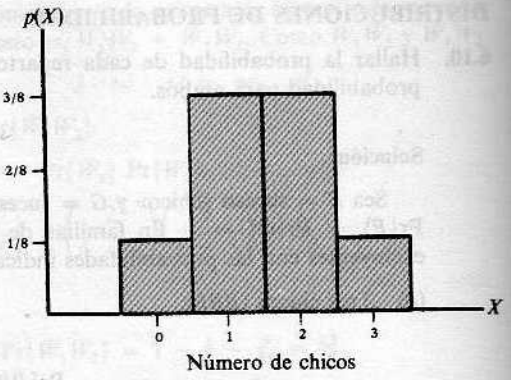


Figura 6.4.

6.12. Una variable aleatoria continua X con valores entre 0 y 4 tiene una función densidad dada por $p(X) = \frac{1}{2} - aX$, donde a es una constante.

(*a*) Calcular a.
(*b*) Hallar $\Pr\{1 < X < 2\}$.

Solución

(*a*) El gráfico de $p(X) = \frac{1}{2} - aX$ es una recta, como muestra la Figura 6.5. Para hallar a, debemos constatar primero que el área total bajo la recta entre $X = 0$ y $X = 4$, y sobre el eje X, ha de ser 1: en $X = 0$, $p(X) = \frac{1}{2}$, y en $X = 4$, $p(X) = \frac{1}{2} - 4a$. Entonces debemos elegir a de modo que el área del trapecio = 1. Area del trapecio = (altura) × (suma de las bases/2 = $\frac{1}{2}(4)(\frac{1}{2} + \frac{1}{2} - 4a) = 2(1 - 4a) = 1$, de donde $(1 - 4a) = \frac{1}{2}$, $4a = \frac{1}{2}$ y $a = \frac{1}{8}$. Luego $(\frac{1}{2} - 4a)$ es realmente igual a cero y, por tanto, la gráfica correcta se muestra en la Figura 6.6.

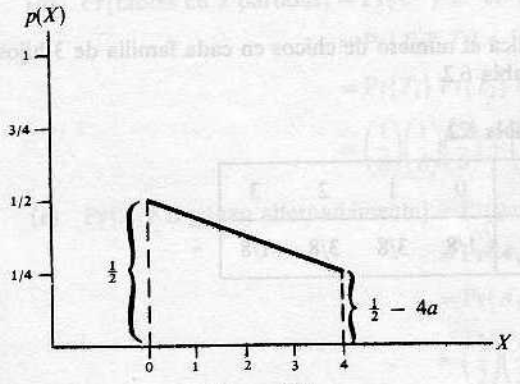


Figura 6.5.

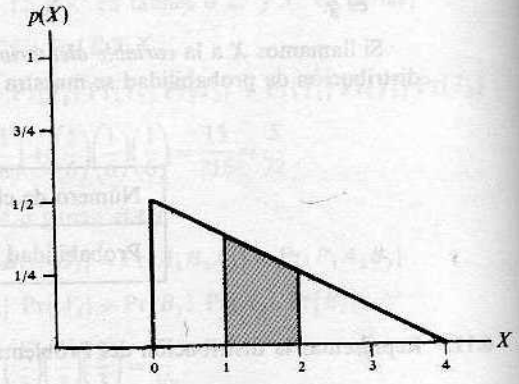


Figura 6.6.

(*b*) La requerida probabilidad es el área entre $X = 1$ y $X = 2$, sombreada en la Figura 6.6. De la parte (*a*), $p(X) = \frac{1}{2} - \frac{1}{8}X$; así que $p(1) = \frac{3}{8}$ y $p(2) = \frac{1}{4}$ son las ordenadas en $X = 1$ y $X = 2$, respectivamente. El área del trapecio pedida es $\frac{1}{2}(1)(\frac{3}{8} + \frac{1}{4}) = \frac{5}{16}$, que es la probabilidad deseada.

$P(1<X<2) = F(2) - F(1) - P(X=2)$
$\frac{1}{4} - \frac{3}{20} - F(2) + F(2)$

ESPERANZA MATEMATICA

6.13. Un boleto de una rifa ofrece dos premios, uno de \$5000 y otro de \$2000, con probabilidades 0.001 y 0.003. ¿Cuál sería el precio justo a pagar por él?

Solución

Su esperanza matemática es (\$5000)(0.001) + (\$2000)(0.003) = \$5 + \$6 = \$11, que es el precio justo.

6.14. En un negocio aventurado, una señora puede ganar \$300 con probabilidad 0.6 o perder \$100 con probabilidad 0.4. Hallar su esperanza matemática.

Solución

Su esperanza matemática es (\$300)(0.6) + (−\$100)(0.4) = \$180 − \$40 = \$140.

6.15. Hallar: (*a*) $E(X)$, (*b*) $E(X^2)$ y (*c*) $E[(X - \bar{X})^2]$ para la distribución de probabilidad que muestra la Tabla 6.3.

Tabla 6.3

X	8	12	16	20	24
$p(X)$	1/8	1/6	3/8	1/4	1/12

Solución

(*a*) $E(X) = \sum Xp(X) = (8)(\frac{1}{8}) + (12)(\frac{1}{6}) + (16)(\frac{3}{8}) + (20)(\frac{1}{4}) + (24)(\frac{1}{12}) = 16$; esto representa la *media* de la distribución.

(*b*) $E(X^2) = \sum X^2p(X) = (8)^2(\frac{1}{8}) + (12)^2(\frac{1}{6}) + (16)^2(\frac{3}{8}) + (20)^2(\frac{1}{4}) + (24)^2(\frac{1}{12}) = 276$; esto representa el segundo momento respecto del origen cero.

(*c*) $E[(X - \bar{X})^2] = \sum (X - \bar{X})^2p(X) = (8 - 16)^2(\frac{1}{8}) + (12 - 16)^2(\frac{1}{6}) + (16 - 16)^2(\frac{3}{8}) + (20 - 16)^2(\frac{1}{4}) + + (24 - 16)^2(\frac{1}{12}) = 20$; esto representa la *varianza* de la distribución.

6.16. Una bolsa contiene 2 bolas blancas y 3 bolas negras. Cada una de cuatro personas, *A*, *B*, *C* y *D*, en ese orden, saca una bola y no la repone. El primero que la saque blanca recibe \$10. Determinar las esperanzas matemáticas de *A*, *B*, *C* y *D*.

Solución

Como sólo hay 3 bolas blancas, alguien ganará en su primer intento. Sean *A*, *B*, *C* y *D* los sucesos «*A* gana», «*B* gana», «*C* gana» y «*D* gana», respectivamente.

$$\Pr\{A \text{ gana}\} = \Pr\{A\} = \frac{2}{3 + 2} = \frac{2}{5}$$

La esperanza matemática de $A = \frac{2}{5}(\$10) = \$4$.

$$\Pr\{A \text{ pierde y } B \text{ gana}\} = \Pr\{\bar{A}B\} = \Pr\{\bar{A}\}\Pr\{B \mid \bar{A}\} = \left(\frac{3}{5}\right)\left(\frac{2}{4}\right) = \frac{3}{10}$$

Así que la esperanza matemática de $B = \$3$.

$$\Pr\{A \text{ y } B \text{ pierden y } C \text{ gana}\} = \Pr\{\bar{A}\bar{B}C\} = \Pr\{\bar{A}\}\Pr\{\bar{B} \mid \bar{A}\}\Pr\{C \mid \bar{A}\bar{B}\} = \left(\frac{3}{5}\right)\left(\frac{2}{4}\right)\left(\frac{2}{3}\right) = \frac{1}{5}$$

Luego la esperanza matemática de $C = \$2$.

$$\Pr\{A, B \text{ y } C \text{ pierden y } D \text{ gana}\} = \Pr\{\bar{A}\bar{B}\bar{C}D\}$$

$$= \Pr\{\bar{A}\}\Pr\{\bar{B} \mid \bar{A}\}\Pr\{\bar{C} \mid \bar{A}\bar{B}\}\Pr\{D \mid \bar{A}\bar{B}\bar{C}\}$$

$$= \left(\frac{3}{5}\right)\left(\frac{2}{4}\right)\left(\frac{1}{3}\right)\left(\frac{1}{1}\right) = \frac{1}{10}$$

Y la de $D = \$1$.

Comprobación: $\$4 + \$3 + \$2 + \$1 = \$10$ y $\frac{2}{5} + \frac{3}{10} + \frac{1}{5} + \frac{1}{10} = 1$

PERMUTACIONES

6.17. ¿De cuántas maneras se pueden poner en fila 5 fichas de colores distintos?

Solución

Debemos colocarlas en cinco posiciones: $_____$. La primera posición puede ser ocupada por cualquier ficha (o sea, hay 5 formas de ocupar esa posición). Una vez ocupada ella, hay 4 maneras de ocupar la siguiente, y entonces 3 de ocupar la tercera, 2 de ocupar la cuarta y sólo una de ocupar la quinta y última. En consecuencia:

$$\text{Número de ordenaciones de 5 fichas en fila} = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5! = 120$$

En general,

$$\text{Número de ordenaciones de } n \text{ objetos distintos en fila} = n(n-1)(n-2)\cdots 1 = n!$$

Eso se llama el número de *permutaciones* de n objetos distintos tomados de n en n, y se denota por ${}_nP_n$

6.18. ¿De cuántas maneras se pueden sentar 10 personas en un banco si hay 4 sitios disponibles?

Solución

El primer sitio se puede ocupar de 10 formas, y una vez ocupado, el segundo se puede ocupar de 9 maneras, el tercero de 8 y el cuarto de 7. Por tanto,

$$\text{Número de colocaciones de 10 personas tomadas de 4 en 4} = 10 \cdot 9 \cdot 8 \cdot 7 = 5040$$

En general,

$$\text{Número de colocaciones de } n \text{ objetos distintos de } r \text{ en } r = n(n-1)\cdots(n-r+1)$$

Esto se llama el número de *permutaciones* de n objetos distintos tomados de n en n y se denota por ${}_nP_r$, $P(n, r)$ o $P_{n,r}$. Nótese que cuando $r = n$, ${}_nP_n = n!$, como en el Problema 6.17.

6.19. Evaluar: (*a*) ${}_8P_3$, (*b*) ${}_6P_4$, (*c*) ${}_{15}P_1$ y ${}_3P_3$.

Solución

(*a*) ${}_8P_3 = 8 \cdot 7 \cdot 6 = 336$, (*b*) ${}_6P_4 = 6 \cdot 5 \cdot 4 \cdot 3 = 360$, (*c*) ${}_{15}P_1 = 15$ y (*d*) ${}_3P_3 = 3 \cdot 2 \cdot 1 = 6$.

6.20. Hay que colocar a 5 hombres y 4 mujeres en una fila de modo que las mujeres ocupen los lugares pares. ¿De cuántas maneras puede hacerse?

Solución

Los hombres se pueden colocar de ${}_5P_5$ maneras y las mujeres de ${}_4P_4$ maneras; cada colocación de ellos se puede asociar con una de ellas, luego el número pedido es ${}_5P_5 \cdot {}_4P_4 = 5!4! = (120)(24) = 2880$.

6.21. ¿Cuántos números de 4 dígitos se pueden formar con las cifras 0, 1, 2, 3, ..., 9: (*a*) permitiendo repeticiones, (*b*) sin repeticiones y (*c*) si el último dígito ha de ser cero y no se permiten repeticiones?

Solución

(*a*) El primero de los dígitos puede ser cualquiera de los 9 no nulos (el cero no se permite en esta posición, pues daría lugar a un número de 3 cifras). El segundo, tercero y cuarto dígitos pueden ya ser cualquiera de los 10. Luego se pueden formar $9 \cdot 10 \cdot 10 \cdot 10 = 9000$ números.

(*b*) El primer dígito puede ser cualquiera salvo el 0. El segundo cualquiera de los 9 que quedan al suprimir el ya empleado. El tercero uno de los 8 que aún no se han colocado y el cuarto cualquiera de los 7 no utilizados todavía. Así que se pueden formar $9 \cdot 9 \cdot 8 \cdot 7 = 4536$ números.

Otro método

El primero de los dígitos puede ser elegido entre 9, y los tres restantes de ${}_9P_3$ maneras. Por tanto, hay $9 \cdot {}_9P_3 = 9 \cdot 9 \cdot 8 \cdot 7 = 4536$ números.

(*c*) El primer dígito se puede elegir de 9 formas, el segundo de 8 y el tercero de 7. Luego se podrán formar $9 \cdot 8 \cdot 7 = 504$ números.

Otro método

El primero de los dígitos se puede tomar de 9 maneras y los otros dos de ${}_8P_2$ maneras. Luego $9 \cdot {}_8P_2 = 9 \cdot 8 \cdot 7 = 504$ números se pueden formar.

6.22. Cuatro libros diferentes de matemáticas, 6 de física y 2 de química han de ser colocados en una estantería. ¿Cuántas colocaciones distintas admiten si: (*a*) los libros de cada materia han de estar juntos y (*b*) sólo los de matemáticas tienen que estar juntos?

Solución

(*a*) Los de matemáticas se pueden colocar entre sí de ${}_4P_4 = 4!$ formas, los de física en ${}_6P_6 = 6!$, los de química de ${}_2P_2 = 2!$ y los tres grupos de ${}_3P_3 = 3!$ maneras entre sí. Luego el número requerido es $= 4!6!2!3! = 207{,}360$.

(*b*) Consideremos los 4 de matemáticas como una sola obra. Entonces tenemos 9 libros, que se pueden colocar de ${}_9P_9 = 6!$ maneras. En cada una de ellas, los 4 de matemáticas están juntos. Pero estos 4 se pueden colocar entre sí de ${}_4P_4 = 4!$ maneras. Luego la solución es $9!4! = 8{,}709{,}120$.

6.23. Cinco fichas rojas, 2 blancas y 3 azules se colocan en fila. Las de un color no son distinguibles entre sí. ¿Cuántas colocaciones distintas son posibles?

Solución

Sea P el número de colocaciones. Multiplicando P por el número de colocaciones de: (a) las 5 rojas entre sí, (b) las 2 blancas entre sí y (c) las 3 azules entre sí (o sea, multiplicando por P es 5!2!3!), obtendremos el número de colocaciones de 10 fichas distinguibles (o sea 10!). Luego

$$(5!2!3!)P = 10! \quad \text{y} \quad P = \frac{10!}{5!2!3!}$$

En general, el número de colocaciones diferentes de n objetos, de los que n_1 son iguales, n_2 son iguales, ..., n_k son iguales, es

$$\frac{n!}{n_1!n_2! \cdots n_k!}$$

donde $n_1 + n_2 + \cdots + n_k = n$.

6.24. ¿De cuántas formas se pueden sentar 7 personas en torno a una mesa redonda si: (a) son libres de elegir el asiento que deseen y (b) 2 personas particulares no pueden sentarse juntas?

Solución

(a) Sentemos a una en una silla. Entonces, los 6 restantes se pueden sentar de 6! = 720 formas, que es el número total pedido.

(b) Consideremos a esas dos especiales como una sola persona. Entonces habría 6 personas, que se pueden sentar de 5! formas. Pero las 2 especiales se pueden colocar entre sí de 2! maneras, luego el número de formas en que se pueden situar 6 personas en una mesa redonda estando dos prefijadas juntas es = 5!2! = 240.

Usando la parte (a), la solución a (b) no es otra que = 720 − 240 = 480 maneras de sentarse con las condiciones impuestas.

COMBINACIONES

6.25. ¿De cuántas formas se pueden repartir 10 objetos en dos grupos de 4 y 6 objetos, respectivamente?

Solución

Es el mismo que el número de colocaciones de 10 objetos de los que 4 son iguales y los otros 6 son iguales. Por el Problema 6.23, es

$$\frac{10!}{4!6!} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{4!} = 210$$

El problema equivale a hallar el número de selecciones de 4 entre 10 objetos (o 6 entre 10), siendo irrelevante el orden de selección.

En general, el número de selecciones de r entre n objetos, llamado el número de *combinaciones* de n objetos tomados de r en r, se denota por $\binom{n}{r}$ y viene dado por

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{{}_nP_r}{r!}$$

6.26. Calcular: (*a*) $\binom{7}{4}$, (*b*) $\binom{6}{5}$ y (*c*) $\binom{4}{4}$.

Solución

(*a*)
$$\binom{7}{4} = \frac{7!}{4!3!} = \frac{7 \cdot 6 \cdot 5 \cdot 4}{4!} = \frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} = 35$$

(*b*)
$$\binom{6}{5} = \frac{6!}{5!1!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{5!} = 6 \quad \text{o} \quad \binom{6}{5} = \binom{6}{1} = 6$$

(*c*) $\binom{4}{4}$ es el número de selecciones de 4 objetos tomados todos de golpe, y hay una sola selección, así que $\binom{4}{4} = 1$. Nótese que formalmente

$$\binom{4}{4} = \frac{4!}{4!0!} = 1$$

si definimos $0! = 1$.

6.27. ¿De cuántas maneras se puede formar con 9 personas una comisión de 5 miembros?

Solución

$$\binom{9}{5} = \frac{9!}{5!4!} = \frac{9 \cdot 8 \cdot 7 \cdot 6 \cdot 5}{5!} = 126$$

6.28. De entre 5 matemáticos y 7 físicos hay que constituir una comisión de 2 matemáticos y 3 físicos. ¿De cuántas formas podrá hacerse si: (*a*) todos son elegibles, (*b*) un físico particular ha de estar en esa comisión y (*c*) dos matemáticos concretos tienen prohibido pertenecer a la comisión?

Solución

(*a*) Dos matemáticos entre 5 se pueden escoger de $\binom{5}{2}$ maneras, y 3 físicos de entre 7, de $\binom{7}{3}$ maneras. El número total de posibles selecciones es

$$\binom{5}{2} \cdot \binom{7}{3} = 10 \cdot 35 = 350$$

(*b*) Dos matemáticos entre 5 se pueden escoger de $\binom{5}{2}$ maneras, y los 2 físicos adicionales de entre 6 de $\binom{6}{2}$ formas. El número total de selecciones posibles es

$$\binom{5}{2} \cdot \binom{6}{2} = 10 \cdot 15 = 150$$

(*c*) Dos matemáticos entre 3 son elegibles de $\binom{3}{2}$ maneras, y 3 físicos de entre 7, de $\binom{7}{3}$ maneras. Luego el número total de selecciones posibles es

$$\binom{3}{2} \cdot \binom{7}{3} = 3 \cdot 35 = 105$$

6.29. ¿Cuántos ramilletes distintos se pueden formar con 5 flores de variedades diferentes?

Solución

Cada flor puede elegirse o no. Esas dos posibilidades ocurren para cada flor, luego en total 2^5. Pero de estas 2^5 opciones hay que excluir la consistente en no escoger ninguna. Luego el número de ramilletes es $= 2^5 - 1 = 31$.

Otro método

Podemos elegir 1 de las 5, o 2 de las 5, ..., o las 5 flores. De modo que el número pedido es

$$\binom{5}{1} + \binom{5}{2} + \binom{5}{3} + \binom{5}{4} + \binom{5}{5} = 5 + 10 + 10 + 5 + 1 = 31$$

En general, para todo entero n positivo,

$$\binom{n}{1} + \binom{n}{2} + \binom{n}{3} + \cdots + \binom{n}{n} = 2^n - 1$$

6.30. Con 7 consonantes y 5 vocales, ¿cuántas palabras se pueden formar que tengan 4 consonantes distintas y 3 vocales distintas? Se admiten palabras sin significado.

Solución

Las 4 consonantes se pueden escoger de $\binom{7}{4}$ maneras, las 3 vocales de $\binom{5}{3}$ maneras y las 7 letras ya elegidas se pueden colocar entre sí de ${}_7P_7 = 7!$ maneras. Así que el número requerido es

$$\binom{7}{4} \cdot \binom{5}{3} \cdot 7! = 35 \cdot 10 \cdot 5040 = 1{,}764{,}000$$

APROXIMACION DE STIRLING A $n!$

6.31. Calcular aproximadamente 50!

Solución

Para n grande, tenemos $n! \approx \sqrt{2\pi n}\, n^n e^{-n}$; así que

$$50! \approx \sqrt{2\pi(50)}\, 50^{50} e^{-50} = S$$

Para evaluar S usemos logaritmos en base 10. Tendremos

$$\begin{aligned} \log S = \log (\sqrt{100\pi}\, 50^{50} e^{-50}) &= \tfrac{1}{2} \log 100 + \tfrac{1}{2} \log \pi + 50 \log 50 - 50 \log e \\ &= \tfrac{1}{2} \log 100 + \tfrac{1}{2} \log 3.142 + 50 \log 50 - 50 \log 2.718 \\ &= \tfrac{1}{2}(2) + \tfrac{1}{2}(0.4972) + 50(1.6990) - 50(0.4343) = 64.4846 \end{aligned}$$

de donde $S = 3.05 \times 10^{64}$,

PROBABILIDAD Y ANALISIS COMBINATORIO

6.32. Una caja contiene 8 bolas rojas, 3 blancas y 9 azules. Si se sacan 3 bolas al azar, determinar la probabilidad de que: (*a*) las 3 sean rojas, (*b*) las 3 sean blancas, (*c*) 2 sean rojas y 1 blanca, (*d*) al menos 1 sea blanca, (*e*) sean una de cada color y (*f*) salgan en el orden roja, blanca, azul.

Solución

(*a*) *Primer método*

Denotemos por R_1, R_2 y R_3 los sucesos «la primera bola es roja», «la segunda bola es roja» y «la tercera bola es roja», respectivamente. Entonces $R_1R_2R_3$ denota el suceso de que las 3 sean rojas.

$$\Pr\{R_1R_2R_3\} = \Pr\{R_1\}\Pr\{R_2 \mid R_1\}\Pr\{R_3 \mid R_1R_2\} = \left(\frac{8}{20}\right)\left(\frac{7}{19}\right)\left(\frac{6}{18}\right) = \frac{14}{285}$$

Segundo método

$$\text{Probabilidad requerida} = \frac{\text{número selecciones de 3 entre 8}}{\text{número selecciones de 3 entre 20}} = \frac{\binom{8}{3}}{\binom{20}{3}} = \frac{14}{285}$$

(*b*) Usando el segundo método de la parte (*a*),

$$\Pr\{\text{las 3 son blancas}\} = \frac{\binom{3}{3}}{\binom{20}{3}} = \frac{1}{1140}$$

Podía usarse también el primer método de (*a*).

(*c*)

$$\Pr\{\text{2 son rojas y 1 blanca}\} = \frac{\binom{\text{selecciones de 2 entre}}{\text{8 bolas rojas}}\binom{\text{selecciones de 1 entre}}{\text{3 bolas blancas}}}{\text{número de selecciones de 3 entre 20 bolas}} = \frac{\binom{8}{2}\binom{3}{1}}{\binom{20}{3}} = \frac{7}{95}$$

(*d*)

$$\Pr\{\text{ninguna es blanca}\} = \frac{\binom{17}{3}}{\binom{20}{3}} = \frac{34}{57} \quad \text{o} \quad \Pr\{\text{al menos 1 es blanca}\} = 1 - \frac{34}{57} = \frac{23}{57}$$

(*e*)

$$\Pr\{\text{sacar 1 de cada color}\} = \frac{\binom{8}{1}\binom{3}{1}\binom{9}{1}}{\binom{20}{3}} = \frac{18}{95}$$

(*f*) Usando la parte (*e*),

$$\Pr\{\text{bolas en orden roja, blanca, azul}\} = \frac{1}{3!}\Pr\{\text{1 de cada color}\} = \frac{1}{6}\left(\frac{18}{95}\right) = \frac{3}{95}$$

Otro método

$$\Pr\{R_1 W_2 B_2\} = \Pr\{R_1\}\ \Pr\{W_2 \mid R_1\}\ \Pr\{B_3 \mid R_1 W_2\} = \left(\frac{8}{20}\right)\left(\frac{3}{19}\right)\left(\frac{9}{18}\right) = \frac{3}{95}$$

6.33. De una baraja de 52 naipes bien mezclada se sacan 5 naipes. Hallar la probabilidad de que: (*a*) 4 sean ases, (*b*) 4 sean ases y 1 rey, (*c*) 3 sean dieces y 2 sotas, (*d*) salgan nueve, diez, sota, caballo y rey en cualquier orden, (*e*) 3 son de un palo y 2 de otro y (*f*) al menos uno sea un as.

Solución

(*a*)

$$\Pr\{4 \text{ ases}\} = \binom{4}{4} \cdot \frac{\binom{48}{1}}{\binom{52}{5}} = \frac{1}{54{,}145}$$

(*b*)

$$\Pr\{4 \text{ ases y } 1 \text{ rey}\} = \binom{4}{4} \cdot \frac{\binom{4}{1}}{\binom{52}{5}} = \frac{1}{649{,}740}$$

(*c*)

$$\Pr\{3 \text{ son dieces y } 2 \text{ son sotas}\} = \binom{4}{3} \cdot \frac{\binom{4}{2}}{\binom{52}{5}} = \frac{1}{108{,}290}$$

(*d*)

$$\Pr\{\text{nueve, diez, sota, caballo, rey en cualquier orden}\} = \frac{\binom{4}{1} \cdot \binom{4}{1} \cdot \binom{4}{1} \cdot \binom{4}{1} \cdot \binom{4}{1}}{\binom{52}{5}} = \frac{64}{162{,}435}$$

(*e*) Como hay cuatro formas de escoger el primer palo y tres de elegir el segundo,

$$\Pr\{3 \text{ de cualquier figura, } 2 \text{ de otra}\} = \frac{4\binom{13}{3} \cdot 3\binom{13}{2}}{\binom{52}{5}} = \frac{429}{4165}$$

(*f*)

$$\Pr\{\text{ningún as}\} = \frac{\binom{48}{5}}{\binom{52}{5}} = \frac{35{,}673}{54{,}145} \qquad \text{y} \qquad \Pr\{\text{al menos 1 as}\} = 1 - \frac{35{,}673}{54{,}145} = \frac{18{,}482}{54{,}145}$$

6.34. Determinar la probabilidad de sacar 3 seises en 5 tiradas de un dado.

Solución

Representemos las 5 tiradas por 5 espacios $- - - - -$. En cada espacio tendremos los sucesos 6 o no 6 ($\bar{6}$); por ejemplo, tres 6 y dos no 6 pueden ocurrir como $6\ 6\ \bar{6}\ 6\ \bar{6}$ o como $6\ \bar{6}\ 6\ \bar{6}\ 6$, etc. Ahora bien, la probabilidad de un suceso tal como $6\ 6\ \bar{6}\ 6\ \bar{6}$ es

$$\Pr\{6\ 6\ \bar{6}\ 6\ \bar{6}\} = \Pr\{6\}\Pr\{6\}\Pr\{\bar{6}\}\Pr\{6\}\Pr\{\bar{6}\} = \frac{1}{6}\cdot\frac{1}{6}\cdot\frac{5}{6}\cdot\frac{1}{6}\cdot\frac{5}{6} = \left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2$$

Similar $\Pr\{6\ \bar{6}\ 6\ \bar{6}\ 6\} = (\frac{1}{6})^3(\frac{5}{6})^2$, etc., para todos los sucesos en los que salen tres 6 y dos no 6. Pero hay $\binom{5}{3} = 10$ de tales sucesos, y esos sucesos son sucesos mutuamente excluyentes; por tanto, la probabilidad requerida es

$$\Pr\{6\ 6\ \bar{6}\ 6\ \bar{6}\ \text{ó}\ 6\ \bar{6}\ 6\ \bar{6}\ 6\ \text{o etc.}\} = \binom{5}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 = \frac{125}{3888}$$

En general, si $p = \Pr\{E\}$ y $q = \Pr\{\bar{E}\}$, entonces usando el mismo argumento que antes, la probabilidad de obtener exactamente X veces E en N intentos es $\binom{N}{x}p^X q^{N-X}$.

6.35. Una factoría observa que, en promedio, el 20% de las tuercas producidas por una máquina son defectuosas. Si se toman 10 tuercas al azar, hallar la probabilidad de que: (*a*) exactamente 2 sean defectuosas, (*b*) 2 o más sean defectuosas y (*c*) más de 5 sean defectuosas.

Solución

(*a*) Por un razonamiento similar al del Problema 6.34,

$$\Pr\{2\ \text{tuercas defectuosas}\} = \binom{10}{2}(0.2)^2(0.8)^8 = 45(0.04)(0.1678) = 0.3020$$

(*b*)

$$\begin{aligned}\Pr\{2\ \text{o más tuercas defectuosas}\} &= 1 - \Pr\{0\ \text{tuercas defectuosas}\} - \Pr\{1\ \text{tuercas defectuosas}\}\\ &= 1 - \binom{10}{0}(0.2)^0(0.8)^{10} - \binom{10}{1}(0.2)^1(0.8)^9\\ &= 1 - (0.8)^{10} - 10(0.2)(0.8)^9\\ &= 1 - 0.1074 - 0.2684 = 0.6242\end{aligned}$$

(*c*)

$$\begin{aligned}\Pr\{\text{más de 5 tuercas defectuosas}\} &= \Pr\{6\ \text{tuercas defectuosas}\} + \Pr\{7\ \text{tuercas defectuosas}\}\\ &\quad + \Pr\{8\ \text{tuercas defectuosas}\} + \Pr\{9\ \text{tuercas defectuosas}\}\\ &\quad + \Pr\{10\ \text{tuercas defectuosas}\}\\ &= \binom{10}{6}(0.2)^6(0.8)^4 + \binom{10}{7}(0.2)^7(0.8)^3 + \binom{10}{8}(0.2)^8(0.8)^2\\ &\quad + \binom{10}{9}(0.2)^9(0.8) + \binom{10}{10}(0.2)^{10}\\ &= 0.00637\end{aligned}$$

6.36. Si se tomaran 1000 muestras de 10 tuercas cada una en el Problema 6.35, ¿de cuántas de ellas cabría esperar que tuvieran: (*a*) exactamente 2 defectuosas, (*b*) 2 o más defectuosas y (*c*) más de 5 defectuosas?

Solución

(*a*) Número esperado = (1000)(0.3020) = 302, por el Problema 6.35(*a*).
(*b*) Número esperado = (1000)(0.6242) = 624, por el Problema 6.35(*b*).
(*c*) Número esperado = (1000)(0.00637) = 6, por el Problema 6.35(*c*).

ESPACIO MUESTRAL Y DIAGRAMAS DE EULER

6.37. (*a*) Describir un espacio muestral para una tirada de un par de dados.
(*b*) Determinar a partir de él la probabilidad de que la suma de los dados sea 7 u 11.

Solución

(*a*) El espacio muestral consta de los puntos de la Figura 6.7, cuyas primeras coordenadas son las puntuaciones del primer dado y las segundas coordenadas son las puntuaciones del segundo dado. Hay 36 puntos, y a cada uno le asignamos una probabilidad de $\frac{1}{36}$. La suma de todas esas probabilidades es 1.

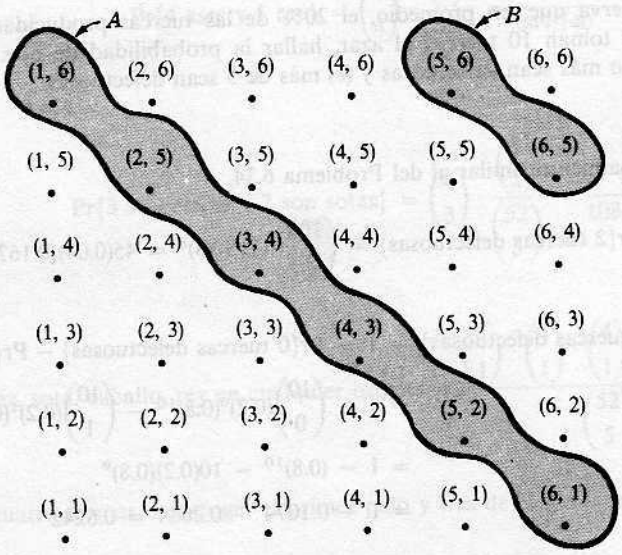


Figura 6.7.

(*b*) Los conjuntos de puntos correspondientes a los sucesos «suma 7» y «suma 11» se indican por A y B, respectivamente.

$\Pr\{A\}$ = suma de probabilidades asociadas con cada punto de $A = \frac{6}{36}$
$\Pr\{B\}$ = suma de probabilidades asociadas con cada punto de $B = \frac{2}{36}$
$\Pr\{A + B\}$ = suma de probabilidades de los puntos en A, en B o en ambos

Nótese que en este caso $\Pr\{A + B\} = \Pr\{A\} + \Pr\{B\}$. Ello ocurre porque A y B no tienen puntos en común (es decir, son sucesos mutuamente excluyentes).

6.38. Usando un espacio muestral, probar que:

(*a*) $\Pr\{A + B\} = \Pr\{A\} + \Pr\{B\} - \Pr\{AB\}$

(*b*) $\Pr\{A + B + C\} = \Pr\{A\} + \Pr\{B\} + \Pr\{C\} - \Pr\{AB\} - \Pr\{BC\} - \Pr\{AC\} + \Pr\{ABC\}$

Solución

(*a*) Sean A y B dos conjuntos de puntos con puntos comunes denotados por AB, como en la Figura 6.8. A consta de $A\bar{B}$ y de AB, mientras B está compuesto por $B\bar{A}$ y AB. La totalidad de puntos en $A + B$ (o bien A, o B o ambos) = totalidad de puntos en A + totalidad de puntos en B − totalidad de puntos en AB. Como la probabilidad de un suceso conjunto es la suma de las probabilidades asociadas a sus puntos, tenemos

$$\Pr\{A + B\} = \Pr\{A\} + \Pr\{B\} - \Pr\{AB\}$$

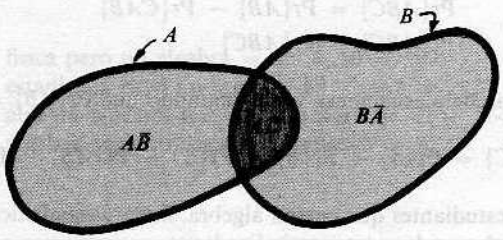


Figura 6.8.

Otro método

Denotemos por $A - AB$ el conjunto de puntos que están en A, pero no en B (es lo mismo que $A\bar{B}$); entonces $A - AB$ y B son mutuamente excluyentes (o sea, sin puntos en común). Además, $\Pr\{A - AB\} = \Pr\{A\} - \Pr\{AB\}$. Luego

$$\Pr\{A + B\} = \Pr\{A - AB\} + \Pr\{B\} = \Pr\{A\} - \Pr\{AB\} + \Pr\{B\} = \Pr\{A\} + \Pr\{B\} - \Pr\{AB\}$$

(*b*) Sean A, B y C tres conjuntos de puntos, como indica la Figura 6.9. El símbolo $AB\bar{C}$ significa el conjunto de puntos en A y B que no están en C, y los otros símbolos son análogos.

Podemos considerar puntos que están en A o B o C como incluidos en los 7 conjuntos mutuamente excluyentes de la Figura 6.9, cuatro de los cuales están sombreados y tres sin sombrear. La probabilidad pedida viene dada por

$$\Pr\{A+B+C\} = \Pr\{A\bar{B}\bar{C}\} + \Pr\{B\bar{C}\bar{A}\} + \Pr\{C\bar{A}\bar{B}\} + \Pr\{AB\bar{C}\} + \Pr\{BC\bar{A}\} + \Pr\{CA\bar{B}\} + \Pr\{ABC\}$$

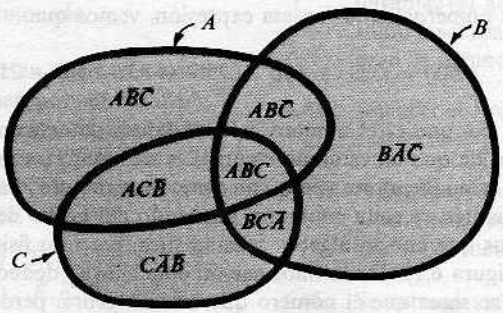


Figura 6.9.

Para obtener ahora $A\bar{B}\bar{C}$, por ejemplo, eliminamos los puntos comunes a A, B y a A, C; pero al hacerlo, hemos quitado los puntos comunes a A, B, C dos veces. Por tanto, $A\bar{B}\bar{C} = A - AB - AC + ABC$, y

$$\Pr\{A\bar{B}\bar{C}\} = \Pr\{A\} - \Pr\{AB\} - \Pr\{AC\} + \Pr\{ABC\}$$

Análogamente, se encuentra

$$\begin{aligned}
\Pr\{B\bar{C}\bar{A}\} &= \Pr\{B\} - \Pr\{BC\} - \Pr\{BA\} + \Pr\{BCA\} \\
\Pr\{C\bar{A}\bar{B}\} &= \Pr\{C\} - \Pr\{CA\} - \Pr\{CB\} + \Pr\{CAB\} \\
\Pr\{BC\bar{A}\} &= \Pr\{BC\} - \Pr\{ABC\} \\
\Pr\{CA\bar{B}\} &= \Pr\{CA\} - \Pr\{BCA\} \\
\Pr\{AB\bar{C}\} &= \Pr\{AB\} - \Pr\{CAB\} \\
\Pr\{ABC\} &= \Pr\{ABC\}
\end{aligned}$$

Sumando esas siete ecuaciones y considerando que $\Pr\{AB\} = \Pr\{BA\}$, etc., obtenemos

$$\Pr\{A + B + C\} = \Pr\{A\} + \Pr\{B\} + \Pr\{C\} - \Pr\{AB\} - \Pr\{BC\} - \Pr\{AC\} + \Pr\{ABC\}$$

6.39. Un recuento de 500 estudiantes que cursan álgebra, física y estadística reveló los siguintes números de estudiantes matriculados en las materias indicadas:

Algebra	329	Algebra y física	83
Física	186	Algebra y estadística	217
Estadística	295	Física y estadística	63

¿Cuántos estudiantes están matriculados en: (*a*) las tres, (*b*) álgebra pero no estadística, (*c*) física pero no álgebra, *d*) estadística pero no física, (*e*) álgebra o estadística pero no física y (*f*) álgebra pero no física ni estadística?

Solución

Sea A el conjunto de estudiantes matriculados en álgebra y (A) el número de ellos. Lo mismo con B, (B) para la física, y con C, (C) para la estadística. Entonces $(A + B + C)$ denota el número de estudiantes matriculados bien en álgebra o en física o en estadística o combinaciones de ellas, (AB) el de los matriculados en ambas, álgebra y física, etc. Como en el Problema 6.38, se sigue que

$$(A + B + C) = (A) + (B) + (C) - (AB) - (BC) - (AC) + (ABC)$$

(*a*) Sustituyendo los números dados en esa expresión, vemos que

$$500 = 329 + 186 + 295 - 83 - 63 - 217 + (ABC)$$

o sea $(ABC) = 53$, que es el número de estudiantes que cursan las tres. Nótese que la probabilidad (empírica) de que un estudiante curse las tres materias es $\frac{53}{500}$.

(*b*) Para obtener la deseada información, conviene construir un diagrama de Euler que muestre el número de estudiantes en cada conjunto. Partiendo del hecho de que 53 de ellos cursan las tres, deducimos que los que cursan álgebra y estadística, pero no física, son $217 - 53 = 164$, como se indica en la Figura 6.10. De la información conocida se deducen los otros números.

De los datos se sigue que el número que cursa álgebra, pero no estadística $= 329 - 217$; y por la Figura 6.10, $82 + 30 = 112$.

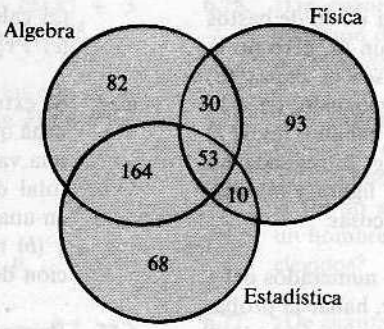


Figura 6.10.

(*c*) Número que cursa física pero no álgebra = 93 + 10 = 103
(*d*) Número que cursa estadística pero no física = 68 + 164 = 232
(*e*) Número que cursa álgebra o estadística pero no física = 82 + 164 + 68 = 314
(*f*) Número que cursa álgebra pero no física ni estadística = 82

PROBLEMAS SUPLEMENTARIOS

REGLAS FUNDAMENTALES DE LA PROBABILIDAD

6.40. Determinar la probabilidad p, o estimarla, para los sucesos:

(*a*) Al extraer una carta de una baraja bien mezclada se saca as, rey o la sota de bastos o el caballo de oros.
(*b*) Al lanzar un par de dados salga suma 8.
(*c*) Encontrar una tuerca defectuosa si entre 600 ya examinadas había 12 defectuosas.
(*d*) Sumar 7 u 11 en una tirada de un par de dados.
(*e*) Sacar al menos una cara en tres lanzamientos de una moneda.

6.41. Un experimento consiste en sacar tres cartas sucesivamente de una baraja bien mezclada. Sea E_1 el suceso «rey» en la primera extracción, E_2 el suceso «rey» en la segunda y E_3 el suceso «rey» en la tercera. Expresar en palabras el significado de:

(*a*) $\Pr\{E_1\bar{E}_2\}$ (*d*) $\Pr\{E_3 \mid E_1\bar{E}_2\}$
(*b*) $\Pr\{E_1+E_2\}$ (*e*) $\bar{E}_1\bar{E}_2\bar{E}_3$
(*c*) $\bar{E}_1+\bar{E}_2$ (*f*) $\Pr\{E_1E_2+\bar{E}_2E_3\}$

6.42. Se saca al azar una bola de una caja que contiene 10 rojas, 30 blancas, 20 azules y 15 naranja. Hallar la probabilidad de que la bola extraída sea: (*a*) roja o naranja, (*b*) ni roja ni azul, (*c*) no azul, (*d*) blanca y (*e*) roja, blanca o azul.

6.43. De la caja del Problema 6.42 se saca una bola, se repone y se hace una nueva extracción. Hallar la probabilidad de que: (*a*) ambas sean blancas, (*b*) la primera sea roja y la segunda blanca, (*c*) ninguna sea naranja, (*d*) ambas son rojas, o blancas o una de cada, (*e*) la segunda no sea azul, (*f*) la primera sea naranja, (*g*) al menos una sea azul, (*h*) a lo sumo una sea roja, (*i*) la primera sea azul, pero la segunda no y (*j*) sólo una sea roja.

6.44. Rehacer el Problema 6.43 sin reponer tras la extracción.

6.45. Hallar la probabilidad de obtener un total de 7 puntos en dos tiradas de un dado: (*a*) una vez, (*b*) al menos una vez y (*c*) dos veces.

6.46. Se extraen sucesivamente dos cartas de una baraja bien mezclada. Hallar la probabilidad

de que: (*a*) la primera no sea un 10 de bastos o un as, (*b*) la primera sea un as, pero no la segunda, (*c*) al menos una sea de copas, (*d*) las cartas no sean del mismo palo, (*e*) a lo sumo una sea figura (sota, caballo, rey), (*f*) la segunda no sea figura, (*g*) la segunda no sea figura si la primera era figura y (*h*) sean figuras o espadas o ambas cosas.

6.47. Una caja contiene 9 tickets numerados del 1 al 9. Si se extraen 3 a la vez, hallar la probabilidad de que sean: (*a*) impar, par, impar, o (*b*) par, impar, par.

6.48. Las apuestas a favor de que A gane una partida de ajedrez contra B están 3 : 2. Si se disputan 3 partidas, ¿cuáles son las apuestas: (*a*) a favor de que A gane al menos dos y (*b*) en contra de que A pierda las dos primeras?

6.49. Un bolso contiene 2 monedas de plata y 4 de cobre, y otro contiene 4 de plata y 3 de cobre. Si se coge al azar de uno de los bolsos una moneda, ¿cuál es la probabilidad de que sea de plata?

6.50. La probabilidad de que un hombre siga vivo dentro de 25 años es $\frac{3}{5}$, y la de que su esposa lo esté es de $\frac{2}{3}$. Hallar la probabilidad de que en ese momento: (*a*) ambos estén vivos, (*b*) sólo el hombre viva, (*c*) sólo viva la esposa y (*d*) al menos uno esté vivo.

6.51. De entre 800 familias con 4 hijos cada una, ¿qué porcentaje es de esperar que tenga: (*a*) 2 chicos y dos chicas, (*b*) al menos un chico, (*c*) ninguna chica y (*d*) a lo sumo 2 chicas? Se supone igual probabilidad para chicos y chicas.

DISTRIBUCIONES DE PROBABILIDAD

6.52. Si X es la variable aleatoria que da el número de chicos en familias de 4 hijos (véase Prob. 6.51): (*a*) construir una tabla que muestre su distribución de probabilidad y (*b*) representar la distribución de probabilidad de la parte (*a*) gráficamente.

6.53. Una variable aleatoria continua X que toma valores entre 2 y 8 inclusive, tiene una función densidad dada por $a(X + 3)$, con a constante: (*a*) calcular a; hallar: (*b*) $\Pr\{3 < X < 5\}$, (*c*) $\Pr\{X \geq 4\}$ y (*d*) $\Pr\{|X - 5| < 0.5\}$.

6.54. Se extraen, sin reposición, tres fichas de una urna que contiene 4 rojas y 6 blancas. Si X es una variable aleatoria que denota el número total de fichas rojas extraídas: (*a*) construir en una tabla su distribución de probabilidad y (*b*) representar gráficamente esa distribución de probabilidad.

6.55. Para el Problema 6.54, hallar: (*a*) $\Pr\{X = 2\}$, y (*b*) $(\Pr\{1 \leq X \leq 3\}$, e interpretar los resultados.

ESPERANZA MATEMATICA

6.56. ¿Cuál es el precio justo para participar en un juego en el que se ganan \$25 con probabilidad 0.2 y \$10 con probabilidad 0.4?

6.57. Si llueve, un vendedor de paraguas gana \$30 al día, y si no llueve pierde \$6 al día. ¿Cuál es su esperanza matemática si la probabilidad de lluvia es 0.3?

6.58. A y B juegan a tirar una moneda tres veces. Gana el primero que saque cara. Si A lanza primero y el montante de la apuesta es \$20, ¿cuánto debe poner cada uno para que el juego sea justo?

6.59. Hallar: (*a*) $E(X)$, (*b*) $E(X^2)$, (*c*) $E[(X - \bar{X})^2]$ y (*d*) $E(X^3)$ para la distribución de probabilidad de la Tabla 6.4.

Tabla 6.4

X	-10	-20	30
$p(X)$	1/5	3/10	1/2

6.60. Refiriéndonos al Problema 6.54, hallar: (*a*) la media, (*b*) la varianza y (*c*) la desviación típica de la distribución de X, e interpretar los resultados.

6.61. Una variable aleatoria toma el valor 1 con probabilidad p y el 0 con probabilidad $q = 1 - p$. Probar que: (*a*) $E(X) = p$ y (*b*) $E[(X - \bar{X})^2] = pq$.

6.62. Probar que: (*a*) $E(2X + 3) = 2E(X) + 3$ y (*b*) $E[(X - \bar{X})^2] = E(X^2) - [E(X)]^2$.

6.63. Sea X e Y dos variables aleatorias con idéntica distribución. Demostrar que $E(X+Y) = E(X) + E(Y)$.

PERMUTACIONES

6.64. Evaluar: (*a*) ${}_4P_2$, (*b*) ${}_7P_5$ y (*c*) ${}_{10}P_3$.

6.65. ¿Para qué valor de n es ${}_{n+1}P_3 = {}_nP_4$?

6.66. ¿De cuántas maneras pueden sentarse 5 personas en un sofá de 3 plazas?

6.67. ¿De cuántas maneras pueden colocarse 7 libros en una estantería si: (*a*) cualquier colocación es admitida, (*b*) 3 libros particulares han de estar juntos y (*c*) 2 libros particulares deben ocupar los extremos?

6.68. ¿Cuántos números de 5 cifras diferentes se pueden formar con los dígitos 1, 2, 3, ..., 9 si: (*a*) cada número ha de ser impar y (*b*) los dos primeros dígitos han de ser pares?

6.69. Resolver el Problema 6.68 permitiendo repeticiones de dígitos.

6.70. ¿Cuántos números de tres dígitos se pueden formar con 3 cuatros, 4 doses y 2 treses?

6.71. ¿De cuántas maneras pueden sentarse 3 hombres y 3 mujeres en una mesa redonda si: (*a*) no se imponen restricciones, (*b*) 2 mujeres particulares no pueden sentarse juntas y (*c*) cada mujer ha de estar entre dos hombres?

COMBINACIONES

6.72. Evaluar: (*a*) $\binom{7}{3}$, (*b*) $\binom{8}{4}$ y (*c*) $\binom{10}{8}$.

6.73. ¿Para qué valor de n es $3\binom{n+1}{3} = 7\binom{n}{2}$?

6.74. ¿De cuántas maneras pueden seleccionarse 6 cuestiones de entre un total de 10?

6.75. ¿De cuántas maneras puede formarse una comisión de 3 hombres y 4 mujeres de entre un total de 8 hombres y 6 mujeres?

6.76. ¿De cuántas maneras pueden escogerse 2 hombres, 4 mujeres, 3 niños y 3 niñas de entre 6 hombres, 8 mujeres, 4 niños y 5 niñas si: (*a*) no se impone restricción alguna y (*b*) un hombre y una mujer concretos deben ser elegidos?

6.77. ¿De cuántas maneras puede dividirse un grupo de 10 personas en dos grupos de 7 y 3 personas?

6.78. ¿De cuántas maneras puede elegirse una comisión de 3 estadísticos y 2 economistas de entre 5 estadísticos y 6 economistas si: (*a*) no se imponen restricciones, (*b*) 2 estadísticos particulares han de figurar en ella y (*c*) un economista concreto tiene vetado el figurar en ella?

6.79. Hallar el número de: (*a*) combinaciones y (*b*) permutaciones de 4 letras que pueden formarse con las letras de la palabra *Tennessee*.

6.80. Demostrar que $1 - \binom{n}{1} + \binom{n}{2} - \binom{n}{3} + \cdots + (-1)^n\binom{n}{n} = 0$.

APROXIMACION DE STIRLING A $n!$

6.81. ¿De cuántas maneras pueden seleccionarse 30 individuos de entre 100?

6.82. Probar que $\binom{2n}{n} = 2^{2n}/\sqrt{\pi n}$, aproximadamente, para grandes valores de n.

PROBLEMAS DIVERSOS

6.83. Se sacan 3 cartas de una baraja de 52 cartas. Hallar la probabilidad de que: (*a*) dos sean sotas y una rey, (*b*) todas sean del mismo palo, (*c*) sean de palos diferentes y (*d*) al menos dos sean ases.

6.84. Hallar la probabilidad de al menos dos sietes en 4 tiradas de un par de dados.

6.85. Si el 10% de los remaches producidos por una máquina son defectuosos, ¿cuál es la probabilidad de que entre 5 elegidos al azar: (*a*) ninguno sea defectuoso, (*b*) haya uno defectuoso y (*c*) al menos dos lo sean?

6.86. (*a*) Describir un espacio muestral para los resultados de dos lanzamientos de una moneda, usando 1 para representar «cara» y 0 para «cruz».
(*b*) Con tal espacio muestral, determinar la probabilidad de al menos una cara.
(*c*) ¿Puede dar un espacio muestral para los resultados de lanzar 3 veces una moneda? En caso afirmativo, determine con su ayuda la probabilidad de al menos 2 caras.

6.87. Un muestreo de 200 votantes revela la siguiente información referente a tres candidatos A, B y C de un cierto partido que se disputaban tres cargos diferentes:

28 a favor de ambos A y B
98 a favor de A o B pero no C
42 a favor de B pero no A o C
122 a favor de B o C pero no A
64 a favor de C pero no A o B
14 a favor de A y C pero no B

¿Cuántos de los votantes están a favor de: (*a*) los tres candidatos, (*b*) de A e indiferentes a B y C, (*c*) de B e indiferentes a A y C, (*d*) de C e indiferentes a A y B, (*e*) de A y B, pero no de C y (*f*) sólo de uno de los candidatos?

6.88. (*a*) Probar que para cualesquiera sucesos E_1 y E_2, $\Pr\{E_1 + E_2\} \leqslant \Pr\{E_1\} + \Pr\{E_2\}$.
(*b*) Generalizar el resultado de la parte (*a*).

6.89. Sean E_1, E_2 y E_3 tres sucesos diferentes, al menos uno de los cuales se sabe que ha ocurrido. Si todas las probabilidades $\Pr\{E_1\}$, $\Pr\{E_2\}$, $\Pr\{E_3\}$ y $\Pr\{A \mid E_1\}$, $\Pr\{A \mid E_2\}$, $\Pr\{A \mid E_3\}$ se suponen conocidas, probar que

$$\Pr\{E_1 \mid A\} = \frac{\Pr\{E_1\}\Pr\{A \mid E_1\}}{\sum_1^3 j \Pr\{E_j\}\Pr\{A \mid E_j\}}$$

con resultados similares para $\Pr\{E_2 \mid A\}$ y $\Pr\{E_3 \mid A\}$. Esto se conoce como *regla* o *teorema* de Bayes. Es útil al calcular probabilidades de varias hipótesis que han resultado en el sucesc A. El resultado es generalizable.

6.90. Tres joyeros idénticos tienen cada uno dos cajones. Cada cajón del primero contiene un reloj de oro, y cada uno del segundo un reloj de plata. En un cajón del tercero hay uno de oro y en el otro uno de plata. Si seleccionamos un joyero al azar, abrimos uno de sus cajones y en él hay un reloj de plata, ¿cuál es la probabilidad de que en el otro cajón haya un reloj de oro? [*Ayuda:* Aplicar el Problema 6.89.]

6.91. Hallar la probabilidad de acertar una loto en la que se deben marcar 6 números de entre 1, 2, 3, ..., 40 en cualquier orden.

6.92. Rehacer el Problema 6.91 si se marcan: (*a*) 5, (*b*) 4 y (*c*) 3 de los números.

6.93. En el póquer se dan a cada jugador 5 cartas de una baraja de 52 cartas. Determinar las apuestas en contra de que un jugador reciba:
(*a*) Escalera de color máxima (10, J, Q, K y as del mismo palo).
(*b*) Escalera de color (cinco cartas sucesivas del mismo palo, por ejemplo, 3, 4, 5, 6 y 7 de tréboles).
(*c*) Un póquer (cuatro cartas iguales, por ejemplo, cuatro sietes).
(*d*) Un «full» (un trío y una pareja, por ejemplo, tres reyes y dos cincos).

6.94. A y B deciden encontrarse entre las 3 y las 4 de la tarde, pero acuerdan que cada uno no espera más de 10 minutos al otro. Hallar la probabilidad de que se encuentren.

6.95. Se escogen al azar dos puntos en un segmento recto de longitud $a > 0$. Hallar la probabilidad de que los tres segmentos así formados puedan ser los lados de un triángulo.

CAPITULO 7

Las distribuciones binomial, normal y de Poisson

LA DISTRIBUCION BINOMIAL

Si p es la probabilidad de que ocurra un suceso en un solo intento (llamada probabilidad de *éxito*) y $q = 1 - p$ es la probabilidad de que no ocurra en un solo intento (llamada probabilidad de *fracaso*), entonces la probabilidad de que el suceso ocurra exactamente X veces en N intentos (o sea, X éxitos y $N - X$ fracasos) viene dada por

$$p(X) = \binom{N}{X} p^X q^{N-X} = \frac{N!}{X!(N-X)!} p^X q^{N-X} \tag{1}$$

donde $X = 0, 1, 2, \ldots, N$; $N! = N(N-1)(N-2)\cdots 1$; y $0! = 1$ por definición (véase Prob. 6.34).

EJEMPLO 1. La probabilidad de obtener exactamente 2 caras en 6 tiradas de una moneda es

$$\binom{6}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{6-2} = \frac{6}{2!4!}\left(\frac{1}{2}\right)^6 = \frac{15}{64}$$

usando la fórmula (1) con $N = 6$, $X = 2$ y $p = q = \frac{1}{2}$.

EJEMPLO 2. La probabilidad de obtener al menos 4 caras en 6 tiradas de una moneda es

$$\binom{6}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^{6-4} + \binom{6}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^{6-5} + \binom{6}{6}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^{6-6} = \frac{15}{64} + \frac{6}{64} + \frac{1}{64} = \frac{11}{32}$$

La distribución de probabilidad discreta (1) se llama *distribución binomial* porque para $X = 0, 1, 2, \ldots, N$ corresponde a términos sucesivos de la *fórmula binomial*, o *desarrollo del binomio*,

$$(q + p)^N = q^N + \binom{N}{1} q^{N-1}p + \binom{N}{2} q^{N-2}p^2 + \cdots + p^N \tag{2}$$

donde $1, \binom{N}{1}, \binom{N}{2}, \ldots$ se llaman *coeficientes binomiales*.

EJEMPLO 3.

$$(q + p)^4 = q^4 + \binom{4}{1}q^3p + \binom{4}{2}q^2p^2 + \binom{4}{3}qp^3 + p^4$$
$$= q^4 + 4q^3p + 6q^2p^2 + 4qp^3 + p^4$$

La distribución (1) se llama también *distribución de Bernoulli*, en honor de James Bernoulli, quien la descubrió a finales del siglo XVII. Algunas propiedades de la distribución binomial se recogen en la Tabla 7.1.

EJEMPLO 4. En 100 tiradas de una moneda el número medio de caras es $\mu = Np = (100)(\frac{1}{2}) = 50$; este es el número esperado de caras en 100 lanzamientos. La desviación típica es $\sigma = \sqrt{Npq} = \sqrt{(100)(\frac{1}{2})(\frac{1}{2})} = 5$.

LA DISTRIBUCION NORMAL

Uno de los más importantes ejemplos de una distribución de probabilidad continua es la *distribución normal, curva normal* o *distribución gaussiana*, definida por la ecuación

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(X-\mu)^2/\sigma^2} \tag{3}$$

Tabla 7.1. Distribución binomial

Media	$\mu = Np$
Varianza	$\sigma^2 = Npq$
Desviación típica	$\sigma = \sqrt{Npq}$
Coeficiente de sesgo	$\alpha_3 = \dfrac{q - p}{\sqrt{Npq}}$
Coeficiente de curtosis	$\alpha_4 = 3 + \dfrac{1 - 6pq}{Npq}$

donde μ = media, σ = desviación típica, $\pi = 3.14159\cdots$ y $e = 2.71828\cdots$. El área total limitada por la curva (3) y el eje X es 1; por tanto, el área bajo la curva entre $X = a$ y $X = b$, con $a < b$, representa la probabilidad de que X esté entre a y b. Esta probabilidad se denota por $\Pr\{a < X < b\}$.

Cuando se expresa la variable X en unidades estándar $[z = (X - \mu)/\sigma]$, la ecuación (3) es reemplazada por la llamada *forma canónica*

$$Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \tag{4}$$

En tal caso, decimos que z está *normalmente distribuida con media 0 y varianza* 1. La Figura 7.1 es un gráfico de esta forma canónica. Muestra que las áreas comprendidas entre $z = -1$ y $+1$, $z = -2$ y $+2$, y $z = -3$ y $+3$ son iguales, respectivamente, a 68.27%, 95.45% y 99.73% del área total, que es 1. La tabla del Apéndice II muestra las áreas bajo esta curva acotadas por las ordenadas $z = 0$ y cualquier valor positivo de z. De esa tabla se puede deducir el área entre todo par de coordenadas usando la simetría de la curva respecto de $z = 0$.

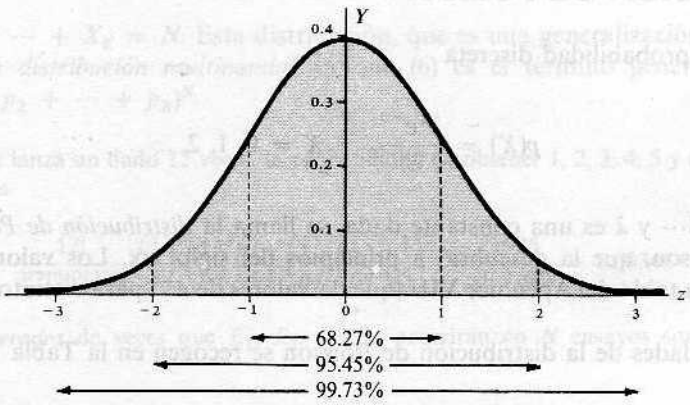


Figura 7.1.

Algunas propiedades de la distribución normal (3) se listan en la Tabla 7.2.

Tabla 7.2. Distribución normal

Media	μ
Varianza	σ^2
Desviación típica	σ
Coeficiente de sesgo	$\alpha_3 = 0$
Coeficiente de curtosis	$\alpha_4 = 3$
Desviación media	$\sigma\sqrt{2/\pi} = 0.7979\sigma$

RELACION ENTRE LA DISTRIBUCION BINOMIAL Y LA DISTRIBUCION NORMAL

Si N es grande y si ni p ni q son muy próximos a cero, la distribución binomial puede aproximarse estrechamente por una distribución normal con variable canónica dada por

$$z = \frac{X - Np}{\sqrt{Npq}}$$

La aproximación mejora al crecer N, y en el límite es exacta; esto se muestra en las Tablas 7.1 y 7.2, donde es claro que al crecer N, el sesgo y la curtosis de la distribución binomial se aproximan a los de la distribución normal. En la práctica, la aproximación es muy buena si tanto Np como Nq son mayores que 5.

LA DISTRIBUCION DE POISSON

La distribución de probabilidad discreta

$$p(X) = \frac{\lambda^X e^{-\lambda}}{X!} \qquad X = 0, 1, 2, \ldots \tag{5}$$

donde $e = 2.71828\cdots$ y λ es una constante dada, se llama la *distribución de Poisson* en honor de Siméon-Denis Poisson, que la descubrió a principios del siglo XIX. Los valores de $p(X)$ pueden calcularse usando la tabla del Apéndice VIII (que da valores de $e^{-\lambda}$ para distintos λ) o por medio de logaritmos.

Algunas propiedades de la distribución de Poisson se recogen en la Tabla 7.3.

Tabla 7.3. Distribución de Poisson

Media	$\mu = \lambda$
Varianza	$\sigma^2 = \lambda$
Desviación típica	$\sigma = \sqrt{\lambda}$
Coeficiente de sesgo	$\alpha_3 = 1/\sqrt{\lambda}$
Coeficiente de curtosis	$\alpha_4 = 3 + 1/\lambda$

RELACION ENTRE LA DISTRIBUCION BINOMIAL Y LA DISTRIBUCION DE POISSON

En la distribución binomial (1), si N es grande y la probabilidad p de ocurrencia de un suceso es muy pequeña, de modo que $q = 1 - p$ es casi 1, el suceso se llama un *suceso raro*. En la práctica, un suceso se considera raro si el número de ensayos es al menos 50 ($N \geq 50$) mientras Np es menor que 5. En tal caso, la distribución binomial queda aproximada muy estrechamente por la distribución de Poisson (5) con $\lambda = Np$. Esto se comprueba comparando las Tablas 7.1 y 7.3, pues al poner $\lambda = Np$, $q \approx 1$ y $p \approx 0$ en la Tabla 7.1 obtenemos los resultados de la Tabla 7.3.

Como hay una relación entre la distribución binomial y la distribución normal, se sigue que también están relacionadas la distribución de Poisson y la distribución normal. De hecho, puede probarse que la distribución de Poisson tiende a una distribución normal con variable canónica $(X - \lambda)/\sqrt{\lambda}$ cuando λ crece indefinidamente.

LA DISTRIBUCION MULTINOMIAL

Si los sucesos $E_1, E_2, \ldots, E_K$ pueden ocurrir con frecuencias $p_1, p_2, \ldots, p_K$, respectivamente, entonces la probabilidad de que $E_1, E_2, \ldots, E_K$ ocurran $X_1, X_2, \ldots, X_K$ veces, respectivamente, es

$$\frac{N!}{X_1!X_2!\cdots X_K!}p_1^{X_1}p_2^{X_2}\cdots p_K^{X_K} \tag{6}$$

donde $X_1 + X_2 + \cdots + X_K = N$. Esta distribución, que es una generalización de la distribución binomial, se llama *distribución multinomial* ya que (6) es el término general en el *desarrollo multinomial* $(p_1 + p_2 + \cdots + p_K)^N$.

EJEMPLO 5. Si se lanza un dado 12 veces, la probabilidad de obtener 1, 2, 3, 4, 5 y 6 puntos exactamente dos veces cada uno es

$$\frac{12!}{2!2!2!2!2!2!}\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2 = \frac{1925}{559{,}872} = 0.00344$$

Los números *esperados* de veces que $E_1, E_2, \ldots, E_K$ ocurrirán en N ensayos son $Np_1, Np_2, \ldots, Np_K$, respectivamente.

AJUSTE DE DISTRIBUCIONES DE FRECUENCIAS MUESTRALES MEDIANTE DISTRIBUCIONES TEORICAS

Cuando se tiene una cierta indicación sobre la distribución de una población por argumentos probabilísticos o de otra índole, suele ser posible ajustar esa distribución teórica (llamada también distribución *esperada* o *modelo*) a distribuciones de frecuencias obtenidas de una muestra de esa población. El método usado consiste en emplear la media y la desviación típica de la muestra para estimar las de la población (véanse Probs. 7.31, 7.33 y 7.34).

Para comprobar la *bondad* del ajuste de las distribuciones teóricas, usamos el *test ji-cuadrado* (Cap. 12). Al intentar determinar si una distribución normal representa un buen ajuste para datos dados, es conveniente usar *papel gráfico de curva normal*, o *papel gráfico de probabilidad* como se le llama a veces (véase Prob. 7.32).

PROBLEMAS RESUELTOS

DISTRIBUCION BINOMIAL

7.1. Calcular:

(a) $5!$ (c) $\binom{8}{3}$ (e) $\binom{4}{4}$

(b) $\frac{6!}{2!4!}$ (d) $\binom{7}{5}$ (f) $\binom{4}{0}$

Solución

(a) $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$

(b) $\dfrac{6!}{2!4!} = \dfrac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(2 \cdot 1)(4 \cdot 3 \cdot 2 \cdot 1)} = \dfrac{6 \cdot 5}{2 \cdot 1} = 15$

(c) $\binom{8}{3} = \dfrac{8!}{3!(8-3)!} = \dfrac{8!}{3!5!} = \dfrac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(3 \cdot 2 \cdot 1)(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)} = \dfrac{8 \cdot 7 \cdot 6}{3 \cdot 2 \cdot 1} = 56$

(d) $\binom{7}{5} = \dfrac{7!}{5!2!} = \dfrac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)(2 \cdot 1)} = \dfrac{7 \cdot 6}{2 \cdot 1} = 21$

(e) $\binom{4}{4} = \dfrac{4!}{4!0!} = 1$ porque $0! = 1$ por definición

(f) $\binom{4}{0} = \dfrac{4!}{0!4!} = 1$

7.2. Hallar la probabilidad de que al lanzar una moneda tres veces, aparezcan: (a) 3 caras, (b) 2 caras y una cruz, (c) 2 caras y una cara y (d) 3 cruces.

Solución

Primer método

Denotemos «cara» por H y «cruz» por T, y supongamos que designamos por HTH el que ocurra cara en el primer lanzamiento, cruz en el segundo y cara en el tercero. Como las posibilidades cara y cruz pueden aparecer en cada tirada, hay $(2)(2)(2) = 8$ posibles resultados. Son

$$HHH \quad HHT \quad HTH \quad HTT \quad TTH \quad THH \quad THT \quad TTT$$

Cada una de esas posibilidades es igualmente probable, con probabilidad $\frac{1}{8}$.

(a) 3 caras (HHH) sólo ocurren una vez; luego su probabilidad es $\frac{1}{8}$.
(b) 2 caras y 1 cruz ocurren tres veces (HHT, HTH y THH); luego $\Pr\{\text{2 caras y una cruz}\} = \frac{3}{8}$.
(c) 1 cara y dos cruces ocurren tres veces (HTT, TTH y THT); luego $\Pr\{\text{1 cara y 2 cruces}\} = \frac{3}{8}$.
(d) 3 cruces (TTT) ocurren sólo una vez; luego $\Pr\{TTT\} = \Pr\{\text{3 cruces}\} = \frac{1}{8}$.

Segundo método [usando la fórmula (1)]

(a) $\Pr\{\text{3 caras}\} = \binom{3}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^0 = (1)\left(\frac{1}{8}\right)(1) = \frac{1}{8}$

(b) $\Pr\{\text{2 caras y 1 cruz}\} = \binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right) = (3)\left(\frac{1}{4}\right)\left(\frac{1}{2}\right) = \frac{3}{8}$

(c) $\Pr\{\text{1 cara y 2 cruces}\} = \binom{3}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^2 = (3)\left(\frac{1}{2}\right)\left(\frac{1}{4}\right) = \frac{3}{8}$

(d) $\Pr\{\text{3 cruces}\} = \binom{3}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^3 = (1)(1)\left(\frac{1}{2}\right) = \frac{1}{8}$

Podría procederse, asimismo, como en el Problema 6.10.

7.3. Hallar la probabilidad de que en 5 tiradas de un dado aparezca el 3: (*a*) ninguna vez, (*b*) 1 vez, (*c*) 2 veces, (*d*) 3 veces, (*e*) 4 veces y (*f*) 5 veces.

Solución

La probabilidad del 3 en una sola tirada $= p = \frac{1}{6}$, y la de no sacar 3 $= q = 1 - p = \frac{5}{6}$; luego:

(*a*) $\Pr\{3 \text{ ocurra cero veces}\} = \binom{5}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^5 = (1)(1)\left(\frac{5}{6}\right)^5 = \frac{3125}{7776}$

(*b*) $\Pr\{3 \text{ ocurra una vez}\} = \binom{5}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^4 = (5)\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^4 = \frac{3125}{7776}$

(*c*) $\Pr\{3 \text{ ocurra dos veces}\} = \binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 = (10)\left(\frac{1}{36}\right)\left(\frac{125}{216}\right) = \frac{625}{3888}$

(*d*) $\Pr\{3 \text{ ocurra tres veces}\} = \binom{5}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 = (10)\left(\frac{1}{216}\right)\left(\frac{25}{36}\right) = \frac{125}{3888}$

(*e*) $\Pr\{3 \text{ ocurra cuatro veces}\} = \binom{5}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^1 = (5)\left(\frac{1}{1296}\right)\left(\frac{5}{6}\right) = \frac{25}{7776}$

(*f*) $\Pr\{3 \text{ ocurra cinco veces}\} = \binom{5}{5}\left(\frac{1}{6}\right)^5\left(\frac{5}{6}\right)^0 = (1)\left(\frac{1}{7776}\right)(1) = \frac{1}{7776}$

Nótese que estas probabilidades representan los términos del desarrollo binomial

$$\left(\frac{5}{6} + \frac{1}{6}\right)^5 = \left(\frac{5}{6}\right)^5 + \binom{5}{1}\left(\frac{5}{6}\right)^4\left(\frac{1}{6}\right) + \binom{5}{2}\left(\frac{5}{6}\right)^3\left(\frac{1}{6}\right)^2 + \binom{5}{3}\left(\frac{5}{6}\right)^2\left(\frac{1}{6}\right)^3 + \binom{5}{4}\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)^4 + \left(\frac{1}{6}\right)^5 = 1$$

7.4. Escribir el desarrollo binomial para: (*a*) $(q + p)^4$ y (*b*) $(q + p)^6$.

Solución

(*a*)

$$(q + p)^4 = q^4 + \binom{4}{1}q^3p + \binom{4}{2}q^2p^2 + \binom{4}{3}qp^3 + p^4$$
$$= q^4 + 4q^3p + 6q^2p^2 + 4qp^3 + p^4$$

(*b*)

$$(q + p)^6 = q^6 + \binom{6}{1}q^5p + \binom{6}{2}q^4p^2 + \binom{6}{3}q^3p^3 + \binom{6}{4}q^2p^4 + \binom{6}{5}qp^5 + p^6$$
$$= q^6 + 6q^5p + 15q^4p^2 + 20q^3p^3 + 15q^2p^4 + 6qp^5 + p^6$$

Los coeficientes 1, 4, 6, 4, 1 y 1, 6, 15, 20, 15, 6, 1 se llaman *coeficientes binomiales* correspondientes a $N = 4$ y $N = 6$, respectivamente. Escribiendo estos coeficientes para $N = 0, 1, 2, 3, \ldots$, como muestra la disposición triangular adjunta, obtenemos el llamado *triángulo de Pascal*. Notemos que el primero y el último de los números de cada fila son 1 y que todo otro número se obtiene sumando sus dos vecinos de la fila de encima.

$$
\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1
\end{array}
$$

7.5. Hallar la probabilidad de que en una familia con 5 hijos haya: (*a*) al menos un chico y (*b*) al menos un chico y una chica. Suponemos que la probabilidad de que nazca chico es $\frac{1}{2}$.

Solución

(*a*)

$$\Pr\{1 \text{ chico}\} = \binom{4}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^3 = \frac{1}{4} \qquad \Pr\{3 \text{ chicos}\} = \binom{4}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right) = \frac{1}{4}$$

$$\Pr\{2 \text{ chicos}\} = \binom{4}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 = \frac{3}{8} \qquad \Pr\{4 \text{ chicos}\} = \binom{4}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^0 = \frac{1}{16}$$

Por tanto

$$\Pr\{\text{al menos 1 chico}\} = \Pr\{1 \text{ chico}\} + \Pr\{2 \text{ chicos}\} + \Pr\{3 \text{ chicos}\} + \Pr\{4 \text{ chicos}\}$$
$$= \frac{1}{4} + \frac{3}{8} + \frac{1}{4} + \frac{1}{16} = \frac{15}{16}$$

Otro método

$$\Pr\{\text{al menos 1 chico}\} = 1 - \Pr\{\text{ningún chico}\} = 1 - \left(\frac{1}{2}\right)^4 = 1 - \frac{1}{16} = \frac{15}{16}$$

(*b*)

$$\Pr\{\text{al menos 1 chico y 1 chica}\} = 1 - \Pr\{\text{ningún chico}\} - \Pr\{\text{ninguna chica}\} = 1 - \frac{1}{16} - \frac{1}{16} = \frac{7}{8}$$

7.6. De entre 2000 familias con 4 hijos, ¿cuántas cabe esperar que tengan: (*a*) al menos 1 chico, (*b*) 2 chicos, (*c*) 1 ó 2 chicas y (*d*) ninguna chica? Véase el Problema 7.5(*a*).

Solución

(*a*) Número esperado de familias con al menos 1 chico $= 2000(\frac{15}{16}) = 1875$

(*b*) Número esperado de familias con 2 chicos $= 2000 \cdot \Pr\{2 \text{ chicos}\} = 2000(\frac{3}{8}) = 750$

(*c*) $\Pr\{1 \text{ ó } 2 \text{ chicas}\} = \Pr\{1 \text{ chica}\} + \Pr\{2 \text{ chicas}\} = \Pr\{1 \text{ chico}\} + \Pr\{2 \text{ chicos}\} = \frac{1}{4} + \frac{3}{8} = \frac{5}{8}$.
Número esperado de familias con 1 ó 2 chicas $= 2000(\frac{5}{8}) = 1250$

(*d*) Número esperado de familias con ninguna chica $= 2000(\frac{1}{16}) = 125$

7.7. Si el 20% de los pernos producidos por una máquina son defectuosos, determinar la probabilidad de que, entre 4 pernos elegidos al azar: (*a*) 1, (*b*) 0 y (*c*) a lo sumo 2 sean defectuosos.

Solución

La probabilidad de un perno defectuoso es $p = 0.2$ y la de uno no defectuoso es $q = 1 - p = 0.8$.

(*a*)

$$\Pr\{1 \text{ defectuoso entre } 4\} = \binom{4}{1}(0.2)^1(0.8)^3 = 0.4096$$

(*b*)

$$\Pr\{0 \text{ defectuosos}\} = \binom{4}{0}(0.2)^0(0.8)^4 = 0.4096$$

(*c*)

$$\Pr\{2 \text{ defectuosos}\} = \binom{4}{2}(0.2)^2(0.8)^2 = 0.1536$$

Entonces

$$\begin{aligned}\Pr\{\text{de más de 2 pernos defectuosos}\} &= \Pr\{0 \text{ defectuosos}\} + \Pr\{1 \text{ defectuoso}\} + \{2 \text{ defectuosos}\} \\ &= 0.4096 + 0.4096 + 0.1536 = 0.9728\end{aligned}$$

28. La probabilidad de que un estudiante que ingresa en la Universidad se licencie es 0.4. Hallar la probabilidad de que entre 5 estudiantes elegidos al azar: (*a*) ninguno, (*b*) 1, (*c*) al menos 1 y (*d*) todos, se licencien.

Solución

(*a*) $\Pr\{\text{ninguno se licencie}\} = \binom{5}{0}(0.4)^0(0.6)^5 = 0.07776$ o sea, aproximadamente 0.08

(*b*) $\Pr\{1 \text{ se licencie}\} = \binom{5}{1}(0.4)^1(0.6)^4 = 0.2592$ o sea, aproximadamente 0.26

(*c*) $\Pr\{\text{al menos 1 se licencie}\} = 1 - \Pr\{\text{ninguno se licencie}\} = 0.92224$ o sea, aprox. 0.92

(*d*) $\Pr\{\text{todos se licenciarán}\} = \binom{5}{5}(0.4)^5(0.6)^0 = 0.01024$ o sea, aproximadamente 0.01

29. ¿Cuál es la probabilidad de obtener un total de 9: (*a*) dos veces y (*b*) al menos dos veces, en 6 tiradas de un par de dados?

Solución

Asociando los 6 posibles resultados del primer dado con los 6 del segundo, resulta un total de $6 \cdot 6 = 36$ posibles formas de caer los dados. Son: 1 en el primero y 1 en el segundo, 1 en el primero y 2 en el segundo, etc., denotadas por (1, 1), (1, 2), etc.

De esas 36 posibilidades equiprobables, la suma 9 ocurre en cuatro de ellas: (3, 6), (4, 5), (5, 4) y (6, 3). Luego la probabilidad de sacar 9 en una tirada es $p = \frac{4}{36} = \frac{1}{9}$, y la de no sacar 9 es $q = 1 - p = \frac{8}{9}$.

(*a*) $\Pr\{2 \text{ nueves en 6 tiradas}\} = \binom{6}{2}\left(\frac{1}{9}\right)^2\left(\frac{8}{9}\right)^{6-2} = \frac{61{,}440}{531{,}441}$

(*b*) $\Pr\{\text{al menos 2 nueves}\} = \Pr\{2 \text{ nueves}\} + \Pr\{3 \text{ nueves}\} + \Pr\{4 \text{ nueves}\} + \Pr\{5 \text{ nueves}\} + \Pr\{6 \text{ nueves}\}$

$$= \binom{6}{2}\left(\frac{1}{9}\right)^2\left(\frac{8}{9}\right)^4 + \binom{6}{3}\left(\frac{1}{9}\right)^3\left(\frac{8}{9}\right)^3 + \binom{6}{4}\left(\frac{1}{9}\right)^4\left(\frac{8}{9}\right)^2 + \binom{6}{5}\left(\frac{1}{9}\right)^5\left(\frac{8}{9}\right)^1 + \binom{6}{6}\left(\frac{1}{9}\right)^6\left(\frac{8}{9}\right)^0$$

$$= \frac{61{,}440}{531{,}441} + \frac{10{,}240}{531{,}441} + \frac{960}{531{,}441} + \frac{48}{531{,}441} + \frac{1}{531{,}441} = \frac{72{,}689}{531{,}441}$$

Otro método

$$\Pr\{\text{al menos 2 nueves}\} = 1 - \Pr\{0 \text{ nueves}\} - \Pr\{1 \text{ nueve}\}$$

$$= 1 - \binom{6}{0}\left(\frac{1}{9}\right)^0\left(\frac{8}{9}\right)^6 - \binom{6}{1}\left(\frac{1}{9}\right)^1\left(\frac{8}{9}\right)^5 = \frac{72{,}689}{531{,}441}$$

7.10. Evaluar: (*a*) $\sum_{X=0}^{N} Xp(X)$ y (*b*) $\sum_{X=0}^{N} X^2p(X)$, donde $p(X) = \binom{N}{X}p^Xq^{N-X}$.

Solución

(*a*) Como $q + p = 1$,

$$\sum_{X=0}^{N} Xp(X) = \sum_{X=1}^{N} X\frac{N!}{X!(N-X)!}p^Xq^{N-X} = Np\sum_{X=1}^{N}\frac{(N-1)!}{(X-1)!(N-X)!}p^{X-1}q^{N-X}$$

$$= Np(q+p)^{N-1} = Np$$

(*b*) $$\sum_{X=0}^{N} X^2p(X) = \sum_{X=1}^{N}\frac{N!}{X!(N-X)!}p^Xq^{N-X} = \sum_{X=1}^{N}[X(X-1)+X]\frac{N!}{X!(N-X)!}p^Xq^{N-X}$$

$$= \sum_{X=2}^{N} X(X-1)\frac{N!}{X!(N-X)!}p^Xq^{N-X} + \sum_{X=1}^{N} X\frac{N!}{X!(N-X)!}p^Xq^{N-X}$$

$$= N(N-1)p^2\sum_{X=2}^{N}\frac{(N-2)!}{(X-2)!(N-X)!}p^{X-2}q^{N-X} + Np = N(N-1)p^2(q+p)^{N-2} + Np$$

$$= N(N-1)p^2 + Np$$

Nota: Los resultados en las partes (*a*) y (*b*) son las esperanzas de X y X^2, denotadas por $E(X)$ y $E(X^2)$, respectivamente (véase Cap. 6).

7.11. Si una variable está normalmente distribuida, determinar: (*a*) su media μ y (*b*) su varianza σ^2.

Solución

(*a*) Por el Problema 7.10(*a*),

$$\mu = \text{valor esperado de la variable} = \sum_{X=0}^{N} Xp(X) = Np$$

(*b*) Usando $\mu = Np$ y los resultados del Problema 7.10,

$$\sigma^2 = \sum_{X=0}^{N}(X-\mu)^2 p(X) = \sum_{X=0}^{N}(X^2 - 2\mu X + \mu^2)p(X) = \sum_{X=0}^{N} X^2 p(X) - 2\mu \sum_{X=0}^{N} Xp(X) + \mu^2 \sum_{X=0}^{N} p(X)$$
$$= N(N-1)p^2 + Np - 2(Np)(Np) + (Np)^2(1) = Np - Np^2 = Np(1-p) = Npq$$

Se desprende que la desviación típica de una variable normalmente distribuida es $\sigma = \sqrt{Npq}$.

Otro método

Por el Problema 6.62(*b*),

$$E[(X - \bar{X})]^2 = E(X^2) - [E(X)]^2 = N(N-1)p^2 + Np - N^2p^2 = Np - Np^2 = Npq$$

7.12. Si la probabilidad de un perno defectuoso es 0.1, hallar: (*a*) la media y (*b*) la desviación típica, para la distribución de pernos defectuosos en un total de 400.

Solución

(*a*) La media es $Np = 400(0.1) = 40$; esto es, esperamos 40 pernos defectuosos.

(*b*) La varianza es $Npq = 400(0.1)(0.9) = 36$. Por tanto, la desviación típica es $\sqrt{36} = 6$.

7.13. Hallar los coeficientes momento de: (*a*) sesgo y (*b*) curtosis de la distribución del Problema 7.12.

Solución

(*a*)

$$\text{Coeficiente momento de sesgo} = \frac{q-p}{\sqrt{Npq}} = \frac{0.9 - 0.1}{6} = 0.133$$

Como es positivo, la distribución es sesgada a la derecha.

(*b*)

$$\text{Coeficiente momento de curtosis} = 3 + \frac{1 - 6pq}{Npq} = 3 + \frac{1 - 6(0.1)(0.9)}{36} = 3.01$$

La distribución es un poco *leptocúrtica* con respecto a la distribución normal (o sea, algo más puntiaguda; véase Cap. 5).

LA DISTRIBUCION NORMAL

7.14. En un examen de matemáticas, la calificación media fue 72 y la desviación típica 15. Determinar en unidades estándar las puntuaciones de los alumnos que obtuvieron: (*a*) 60, (*b*) 93 y (*c*) 72.

Solución

(*a*) $z = \dfrac{X - \bar{X}}{s} = \dfrac{60 - 72}{15} = -0.8$

(*b*) $z = \dfrac{X - \bar{X}}{s} = \dfrac{93 - 72}{15} = 1.4$

(*c*) $z = \dfrac{X - \bar{X}}{s} = \dfrac{72 - 72}{15} = 0$

7.15. Con referencia al Problema 7.14, hallar las puntuaciones correspondientes a las puntuaciones estándar: (*a*) -1 y (*b*) 1.6.

Solución

(*a*) $X = \bar{X} + zs = 72 + (-1)(15) = 57$ (*b*) $X = \bar{X} + zs = 72 + (1.6)(15) = 96$

7.16. Se informó a dos estudiantes que habían recibido puntuaciones estándar de 0.8 y -0.4, respectivamente, en una prueba de inglés. Si sus puntuaciones fueron 88 y 64, respectivamente, hallar la media y la desviación típica de las puntuaciones de esa prueba.

Solución

Usando la ecuación $X = \bar{X} + zs$, tenemos $88 = \bar{X} + 0.8s$ para el primer estudiante y $64 = \bar{X} - 0.4s$ para el segundo. Resolviendo esas ecuaciones se obtiene $\bar{X} = 72$ y $s = 20$.

7.17. Hallar el área bajo la curva normal en cada uno de los casos siguientes: (*a*) a (*g*), que corresponden a las Figuras 7.2(*a*) a 7.2(*g*), respectivamente. Usar el Apéndice II.

(*a*) Entre $z = 0$ y $z = 1.2$
(*b*) Entre $z = -0.68$ y $z = 0$
(*c*) Entre $z = -0.46$ y $z = 2.21$
(*d*) Entre $z = 0.81$ y $z = 1.94$
(*e*) A la izquierda de $z = -0.6$
(*f*) A la derecha de $z = -1.28$
(*g*) A la derecha de $z = 2.05$, y a la izquierda de $z = -1.44$

Solución

(*a*) En el Apéndice II miramos en la columna marcada z hasta ver la entrada 1.2; entonces nos desplazamos a la derecha a la columna marcada 0. El resultado, 0.3849, es el área pedida y representa la probabilidad de que z esté entre 0 y 1.2, denotada $\Pr\{0 \leq z \leq 1.2\}$.

(*b*) Por simetría, el área solicitada es la que hay entre $z = 0$ y $z = 0.68$. Para hallarla, buscamos en la columna marcada z en el Apéndice II hasta localizar 0.6; entonces a la derecha hasta la columna 8. El resultado, 0.2517, es el área buscada y representa la probabilidad de que z esté entre -0.68 y 0, denotada $\Pr\{-0.68 \leq z \leq 0\}$.

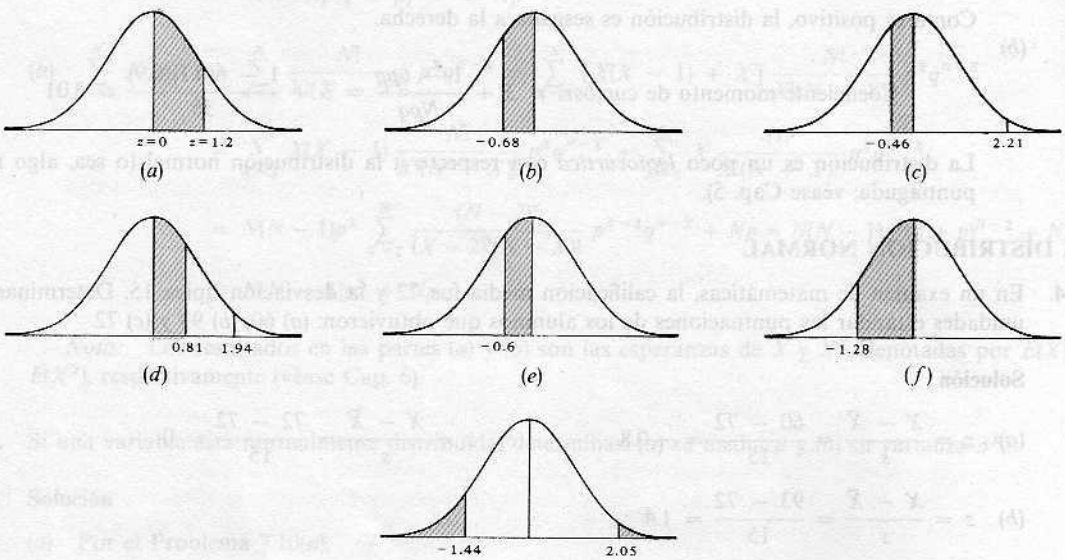


Figura 7.2.

(c) Area pedida = (área entre $z = -0.46$ y $z = 0$) + (área entre $z = 0$ y $z = 2.21$)
$= (\text{área entre } z = 0 \text{ y } z = 0.46) + (\text{área entre } z = 0 \text{ y } z = 2.21)$
$= 0.1772 + 0.4864 = 0.6636$

(d) Area pedida = (área entre $z = 0$ y $z = 1.94$) − (área entre $z = 0$ y $z = 0.81$)
$= 0.4738 - 0.2910 = 0.1828$

(e) Area pedida = (área a la izquierda de $z = 0$) − (área entre $z = -0.6$ y $z = 0$)
$= (\text{área a la izquierda de } z = 0) - (\text{área entre } z = 0 \text{ y } z = 0.6)$
$= 0.5 - 0.2258 = 0.2742$

(f) Area pedida = (área entre $z = -1.28$ y $z = 0$) + (área a la derecha de $z = 0$)
$= 0.3997 + 0.5 = 0.8997$

(g) Area pedida = área total − (área entre $z = -1.44$ y $z = 0$) − (área entre $z = 0$ y $z = 2.05$)
$= 1 - 0.4251 - 0.4798 = 1 - 0.9049 = 0.0951$

7.18. Determinar el valor o valores de z en los casos: (a), (b) y (c), que corresponden a las Figuras 7.3(a) a 7.3(c), respectivamente. La palabra «área» se refiere al área bajo la curva normal.

(a) El área entre 0 y z es 0.3770.
(b) El área a la izquierda de z es 0.8621.
(c) El área entre -1.5 y z es 0.0217.

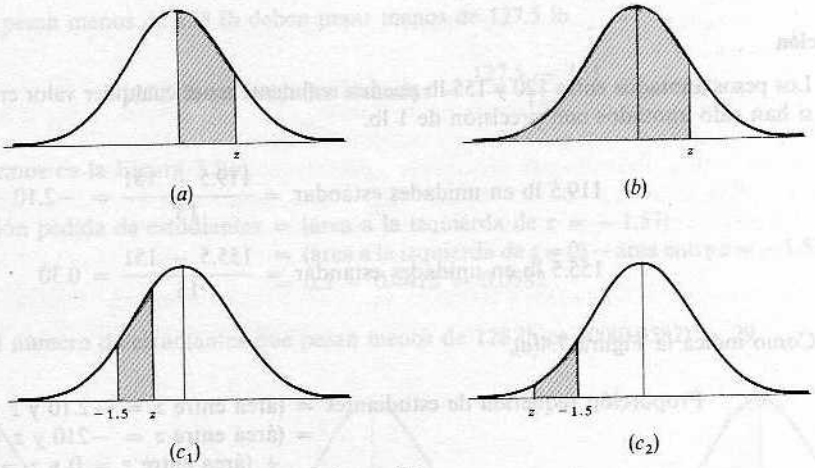


Figura 7.3.

Solución

(a) En el Apéndice II la entrada 0.3770 está a la derecha de la fila marcada 1.1 y bajo la columna 6; así pues, el z pedido es $z = 1.16$. Por simetría, $z = -1.16$ es otro valor solución de z, con lo que $z = \pm 1.16$.

(b) Como el área es mayor que 0.5, z debe ser positivo. El área entre 0 y $z = 0.8621 - 0.5 = 0.3621$, de donde $z = 1.09$.

(c) Si z fuera positivo, el área sería mayor que el área entre -1.5 y 0, que es 0.4332; luego z es negativo.

Caso 1 [z negativo, pero a la derecha de -1.5; véase Fig. 7.3(c_1)]

El área entre -1.5 y z = (área entre -1.5 y 0) − (área entre 0 y z), y $0.0217 = 0.4332 -$ −(área entre 0 y z). Así pues, el área entre 0 y $z = 0.4332 - 0.0217 = 0.4115$, de donde $z = -1.35$.

Caso 2 [z negativo, pero a la izquierda de -1.5; véase Fig. 7.3(c_2)]

El área entre z y -1.5 = (área entre z y 0) − (área entre -1.5 y 0), y 0.0217 = (área entre 0 y z) − 0.4332. Luego el área entre 0 y $z = 0.0217 + 0.4332 = 0.4549$, y $z = -1.694$ por interpolación lineal; o sea, con menos precisión, $z = -1.69$.

7.19. Hallar las ordenadas de la curva normal en: (*a*) $z = 0.84$, (*b*) $z = -1.27$ y (*c*) $z = -0.05$.

Solución

(*a*) En el Apéndice I, buscamos la entrada 0.8 en la columna de z y luego nos movemos a la derecha hasta la columna 4. La entrada 0.2803 es la ordenada pedida.
(*b*) Por simetría: (ordenada en $z = -1.27$) = (ordenada en $z = 1.27$) = 0.1781.
(*c*) (Ordenada en $z = -0.05$) = (ordenada en $z = 0.05$) = 0.3984.

7.20. El peso medio de 500 estudiantes varones de cierta Universidad es 151 libras (lb), y la desviación típica es 15 lb. Supuesto que los pesos están normalmente distribuidos, hallar cuántos estudiantes pesan: (*a*) entre 120 y 155 lb y (*b*) más de 185 lb.

Solución

(*a*) Los pesos anotados entre 120 y 155 lb pueden realmente tener cualquier valor entre 119.5 a 155.5 lb, si han sido anotados con precisión de 1 lb.

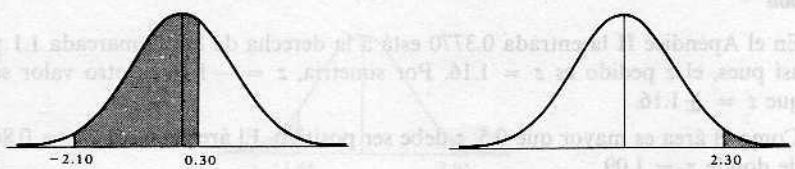
$$119.5 \text{ lb en unidades estándar} = \frac{119.5 - 151}{15} = -2.10$$

$$155.5 \text{ lb en unidades estándar} = \frac{155.5 - 151}{15} = 0.30$$

Como indica la Figura 7.4(*a*),

$$\begin{aligned}\text{Proporción requerida de estudiantes} &= (\text{área entre } z = -2.10 \text{ y } z = 0.30)\\ &= (\text{área entre } z = -210 \text{ y } z = 0)\\ &\quad + (\text{área entre } z = 0 \text{ y } z = 0.30)\\ &= 0.4821 + 0.1179 = 0.6000\end{aligned}$$

Luego el número de estudiantes que pesan entre 120 y 155 lb es $500(0.6000) = 300$.



(*a*) (*b*)

Figura 7.4.

(*b*) Los estudiantes que pesan más de 185 lb han de pesar al menos 185.5 lb.

$$185.5 \text{ lb en unidades estándar} = \frac{185.5 - 151}{15} = 2.30$$

Como se ve en la Figura 7.4(*b*),

$$\begin{aligned}\text{Proporción de estudiantes requerida} &= (\text{área a la derecha de } z = 2.30)\\ &= (\text{área a la derecha de } z = 0) - (\text{área entre } z = 0 \text{ y } z = 2.30)\\ &= 0.5 - 0.4893 = 0.0107\end{aligned}$$

Así que el número de estudiantes que pesan más de 185 lb es 500(0.0107) = 5.

Si W denota el peso de un estudiante al azar, podemos resumir los resultados precedentes en términos de probabilidad escribiendo

$$\Pr\{119.5 \leqslant W \leqslant 155.5\} = 0.6000 \quad \text{y} \quad \Pr\{W \geqslant 185.5\} = 0.0107$$

7.21. Determinar cuántos de los 500 estudiantes del problema anterior pesan: (*a*) menos de 128 lb, (*b*) 128 lb, y (*c*) no más de 128 lb.

Solución

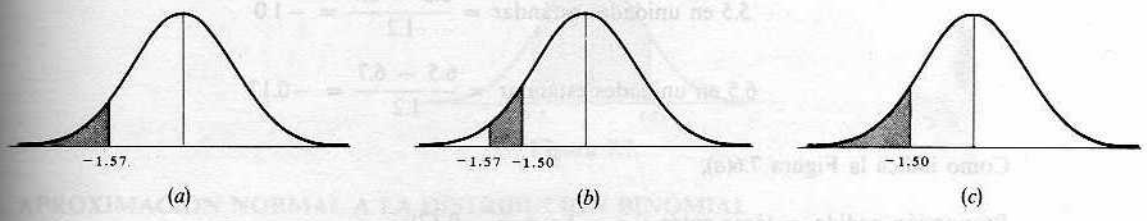
(*a*) Los que pesan menos de 128 lb deben pesar menos de 127.5 lb

$$127.5 \text{ lb en unidades estándar} = \frac{127.5 - 151}{15} = -1.57$$

Como vemos en la Figura 7.5(*a*),

$$\begin{aligned}\text{Proporción pedida de estudiantes} &= (\text{área a la izquierda de } z = -1.57)\\ &= (\text{área a la izquierda de } z = 0) - \text{área entre } z = -1.57 \text{ y } z = 0)\\ &= 0.5 - 0.4418 = 0.0582\end{aligned}$$

Luego el número de estudiantes que pesan menos de 128 lb es 500(0.0582) = 29.



(*a*) (*b*) (*c*)

Figura 7.5.

(*b*) Los que pesan 128, en realidad pesan entre 127.5 y 128.5 lb

$$127.5 \text{ lb en unidades estándar} = \frac{127.5 - 151}{15} = -1.57$$

$$128.5 \text{ lb en unidades estándar} = \frac{128.5 - 151}{15} = -1.50$$

Como muestra la Figura 7.5(*b*),

$$\begin{aligned}\text{Proporción requerida de estudiantes} &= (\text{área entre } z = -1.57 \text{ y } z = -1.50)\\ &= (\text{área entre } z = -1.57 \text{ y } z = 0)\\ &\quad - (\text{área entre } z = -1.50 \text{ y } z = 0)\\ &= 0.4418 - 0.4332 = 0.0086\end{aligned}$$

Por tanto, el número de estudiantes que pesan 128 lb es 500(0.0086) = 4.

(*c*) Los que no pasan de 128 lb deben pesar 128.5 lb

$$128.5 \text{ lb en unidades estándar} = \frac{128.5 - 151}{15} = -1.50$$

Como muestra la Figura 7.5(*c*),

$$\begin{aligned}\text{Proporción requerida de estudiantes} &= (\text{área a la izquierda de } z = -1.50)\\ &= (\text{área a la izquierda de } z = 0) - (\text{área entre } z = -1.50 \text{ y } z = 0)\\ &= 0.5 - 0.4332 = 0.0668\end{aligned}$$

Luego el número de estudiantes que no sobrepasan las 128 lb es 500(0.0668) = 33.

Otro método [usando las partes (*a*) y (*b*)]

El número de los que no pasan de 128 lb es (los que pesan menos de 128 lb) + (los que pesan 128 lb) = 29 + 4 = 33.

7.22. Las puntuaciones en un test de biología eran 0, 1, 2, ..., 10 puntos, según el número de respuestas correctas de entre las 10 cuestiones. La nota media fue 6.7 y la desviación típica 1.2. Supuesto que las notas estuvieran normalmente distribuidas, determinar: (*a*) el porcentaje de estudiantes que tuvo 6 puntos, (*b*) la nota máxima del 10% más bajo y (*c*) la nota mínima del 10% más alto de la clase.

Solución

(*a*) Para aplicar la distribución normal a datos discretos es necesario tratar los datos como si fueran continuos. Así que una nota de 6 puntos se considera que está entre 5.5 y 6.5 puntos

$$5.5 \text{ en unidades estándar} = \frac{5.5 - 6.7}{1.2} = -1.0$$

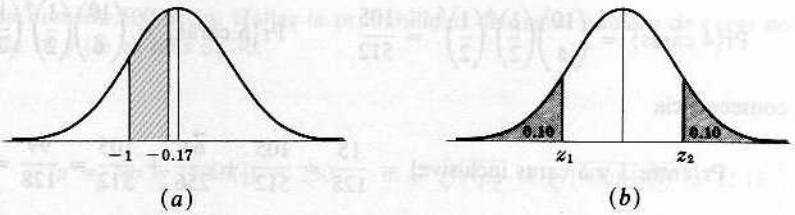
$$6.5 \text{ en unidades estándar} = \frac{6.5 - 6.7}{1.2} = -0.17$$

Como indica la Figura 7.6(*a*),

$$\begin{aligned}\text{Proporción pedida} &= (\text{área entre } z = -1 \text{ y } z = -0.17)\\ &= (\text{área entre } z = -1 \text{ y } z = 0) - (\text{área entre } z = -0.17 \text{ y } z = 0)\\ &= 0.3413 - 0.0675 = 0.2738 = 27\%\end{aligned}$$

(*b*) Sea X_1 la nota máxima y z_1 la nota en unidades estándar. De la Figura 7.6(*b*) se ve que el área a la izquierda de z_1 es 10% = 0.10; por tanto: (área entre z_1 y 0) = 0.40, y $z_1 = -1.28$ (muy aproximadamente). Luego $z_1 = (X_1 - 6.7)/1.2 = -1.28$; y $X_1 = 5.2$, o sea 5 redondeando.

(*c*) Sea X_2 la nota mínima y z_2 la nota en unidades estándar. De la parte (*b*), por simetría, $z_2 = 1.28$. Luego $(X_2 = 6.7)/1.2 = 1.28$; y $X_2 = 8.2$, o sea 8 redondeando.



(*a*) (*b*)

Figura 7.6.

7.23. El diámetro medio interior de una muestra de 200 tubos producidos por una máquina es 0.502 pulgadas (in) y la desviación típica es 0.005 in. El uso de los tubos permitirá una tolerancia en el diámetro de 0.496 a 0.508 in; de otro modo, se considerarán defectuosos. Determinar el porcentaje de tubos defectuosos, supuesto que los tubos producidos por esa máquina están normalmente distribuidos.

Solución

$$0.496 \text{ en unidades estándar} = \frac{0.496 - 0.502}{0.005} = -1.2$$

$$0.508 \text{ en unidades estándar} = \frac{0.508 - 0.502}{0.005} = 1.2$$

Como muestra la Figura 7.7,

$$\begin{aligned}\text{Proporción de tubos defectuosos} &= (\text{área bajo la curva normal entre } z = -1.2 \text{ y } z = 1.2)\\ &= (\text{doble del área entre } z = 0 \text{ y } z = 1.2)\\ &= 2(0.3849) = 0.7698 \quad \text{o sea} \quad 77\%\end{aligned}$$

Luego el porcentaje de tubos defectuosos es 100% − 77% = 23%.

Nótese que si pensamos que el intervalo de 0.496 a 0.508 representa diámetros desde 0.4955 hasta 0.5085 in, el resultado anterior cambia ligeramente. Con dos cifras significativas, sin embargo, el resultado se mantiene.

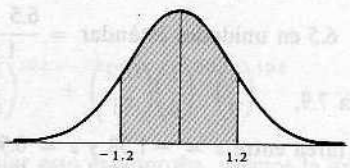


Figura 7.7.

APROXIMACION NORMAL A LA DISTRIBUCION BINOMIAL

7.24. Hallar la probabilidad de obtener entre 3 y 6 caras inclusive en 10 tiradas de una moneda, usando: (*a*) la distribución binomial y (*b*) la aproximación normal a la distribución binomial.

Solución

(*a*)

$$\Pr\{3 \text{ caras}\} = \binom{10}{3}\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right)^{15} = \frac{15}{128} \qquad \Pr\{5 \text{ caras}\} = \binom{10}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^5 = \frac{63}{256}$$

$$\Pr\{4 \text{ caras}\} = \binom{10}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^6 = \frac{105}{512} \qquad \Pr\{6 \text{ caras}\} = \binom{10}{6}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^4 = \frac{105}{512}$$

En consecuencia

$$\Pr\{\text{entre 3 y 6 caras inclusive}\} = \frac{15}{128} + \frac{105}{512} + \frac{63}{256} + \frac{105}{512} = \frac{99}{128} = 0.7734$$

(*b*) La distribución de Poisson para el número de caras en 10 tiradas está representada en las Figuras 7.8(*a*) y (*b*), donde esta última trata los datos como si fueran continuos. La probabilidad pedida es la suma de las áreas de los rectángulos sombreados de la Figura 7.8(*b*) y se puede aproximar por el área correspondiente bajo la curva normal, en sombra en la figura.

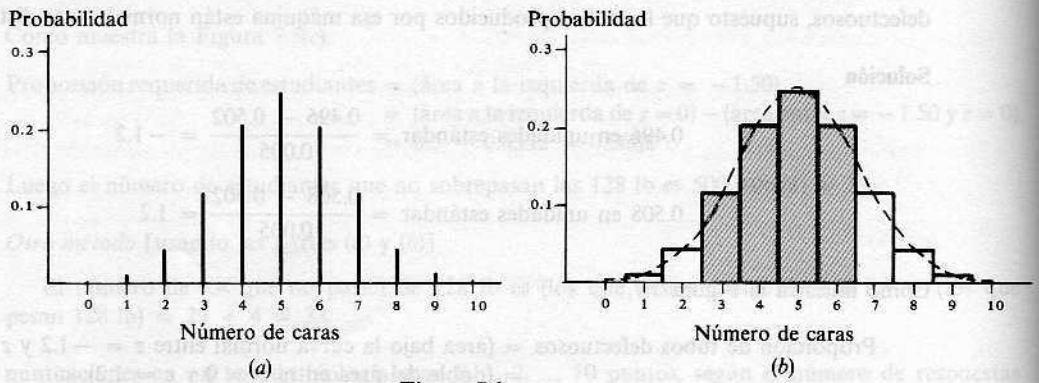


Figura 7.8.

Considerando los datos como continuos, se sigue que 3 a 6 caras es como decir de 2.5 a 6.5 caras. Además, la media y la varianza de la distribución binomial vienen dados por $\mu = Np = 10(\frac{1}{2}) = 5$ y $\sigma = \sqrt{Npq} = \sqrt{(10)(\frac{1}{2})(\frac{1}{2})} = 1.58$

$$2.5 \text{ en unidades estándar} = \frac{2.5 - 5}{1.58} = -1.58$$

$$6.5 \text{ en unidades estándar} = \frac{6.5 - 5}{1.58} = 0.95$$

Como se ve en la Figura 7.9,

$$\begin{aligned}\text{Probabilidad pedida} &= (\text{área entre } z = -1.58 \text{ y } z = 0.95)\\ &= (\text{área entre } z = -1.58 \text{ y } z = 0) + (\text{área entre } z = 0 \text{ y } z = 0.95)\\ &= 0.4429 + 0.3289 = 0.7718\end{aligned}$$

que encaja muy bien con el verdadero valor 0.7734 obtenido en la parte (*a*). La precisión es aún mayor para grandes N.

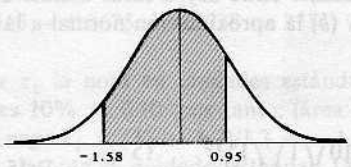


Figura 7.9.

7.25. Se lanza una moneda 500 veces. Hallar la probabilidad de que el número de caras no difiera de 250: (*a*) en más de 10 y (*b*) en más de 30.

Solución

$$\mu = Np = (500)(\tfrac{1}{2}) = 250 \qquad \sigma = \sqrt{Npq} = \sqrt{(500)(\tfrac{1}{2})(\tfrac{1}{2})} = 11.18$$

(*a*) Se nos pide la probabilidad de que el número de caras esté entre 240 y 260, o sea, considerando los datos como continuos, entre 239.5 y 260.5. Como 239.5 en unidades estándar es $(239.5 - 250)/11.18 = -0.94$, y 260.5 en unidades estándar es 0.94, tenemos

$$\begin{aligned}\text{Probabilidad pedida} &= (\text{área bajo la curva normal entre } z = -0.94 \text{ y } z = 0.94)\\ &= (\text{doble del área entre } z = 0 \text{ y } z = 0.94) = 2(0.3264) = 0.6528\end{aligned}$$

(*b*) Se pide la probabilidad de que el número de caras esté entre 220 y 280, o considerados los datos como continuos, entre 219.5 y 280.5. Como 219.5 en unidades estándar es $(219.5 - 250)/11.18 = -2.73$, y 280.5 en unidades estándar es 2.73, tenemos

$$\begin{aligned}\text{Probabilidad pedida} &= (\text{el doble del área bajo la curva normal entre } z = 0 \text{ y } z = -2.73)\\ &= 2(0.4968) = 0.9936\end{aligned}$$

Se sigue que, con gran confianza, el número de caras no diferirá del esperado (250) en más de 30. Así pues, si resultase que el número real de caras fuera 280, tendríamos derecho a sospechar que la moneda estaba trucada o era falsa.

7.26. Se lanza un dado 120 veces. Hallar la probabilidad de que salga el 4: (*a*) 18 veces o menos y (*b*) 14 veces o menos, supuesto como siempre que el dado no está trucado.

Solución

El 4 tiene probabilidad $p = \frac{1}{6}$ de salir y probabilidad $q = \frac{5}{6}$ de no salir.

(*a*) Queremos calcular la probabilidad de que el número de cuatros esté entre 0 y 18, y eso es exactamente

$$\binom{120}{18}\left(\frac{1}{6}\right)^{18}\left(\frac{5}{6}\right)^{102} + \binom{120}{17}\left(\frac{1}{6}\right)^{17}\left(\frac{5}{6}\right)^{103} + \cdots + \binom{120}{0}\left(\frac{1}{6}\right)^{0}\left(\frac{5}{6}\right)^{120}$$

pero como la tarea de calcular esto es ímproba, usemos la aproximación normal.

Considerando los datos como continuos, de 0 a 18 significa de -0.5 a 18.5. Además,

$$\mu = Np = 120(\tfrac{1}{6}) = 20 \qquad \text{y} \qquad \sigma = \sqrt{Npq} = \sqrt{(120)(\tfrac{1}{6})(\tfrac{5}{6})} = 4.08$$

Como -0.5 en unidades estándar es $(-0.5 - 20)/4.08 = -5.02$, y 18.5 en unidades estándar es -0.37, se tiene

$$\begin{aligned}\text{Probabilidad pedida} &= (\text{área bajo la curva normal entre } z = -5.02 \text{ y } z = -0.37)\\ &= (\text{área entre } z = 0 \text{ y } z = -5.02)\\ &\quad - (\text{área entre } z = 0 \text{ y } z = -0.37)\\ &= 0.5 - 0.1443 = 0.3557\end{aligned}$$

(*b*) Procedemos como en (*a*), sustituyendo 18 por 14. Como -0.5 en unidades estándar es -5.02, y 14.5 en unidades estándar es $(14.5 - 20)/4.08 = -1.35$, tenemos

$$\begin{aligned}\text{Probabilidad pedida} &= (\text{área bajo la curva normal entre } z = -5.02 \text{ y } z = -1.35)\\ &= (\text{área entre } z = 0 \text{ y } z = -5.02)\\ &\quad - (\text{área entre } z = 0 \text{ y } z = -1.35)\\ &= 0.5 - 0.4115 = 0.0885\end{aligned}$$

Se desprende que si tomamos repetidas muestras de 120 lanzamientos de un dado, el 4 saldría 14 veces o menos en aproximadamente un 10% de esas muestras.

LA DISTRIBUCION DE POISSON

7.27. Un 10% de las herramientas producidas en una fábrica son defectuosas. Hallar la probabilidad de que en una muestra de 10 herramientas tomadas al azar exactamente 2 sean defectuosas, usando: (*a*) la distribución binomial y (*b*) la aproximación de Poisson a la distribución binomial.

Solución

La probabilidad de una herramienta defectuosa es $p = 0.1$.

(*a*)

$$\Pr\{2 \text{ objetos defectuosos en } 10\} = \binom{10}{2}(0.1)^2(0.9)^8 = 0.1937 \qquad \text{o sea} \qquad 0.19$$

(*b*) Con $\lambda = Np = 10(0.1) = 1$ y usando $e = 2.718$,

$$\Pr\{2 \text{ objetos defectuosos en } 10\} = \frac{\lambda^X e^{-\lambda}}{X!} = \frac{(1)^2 e^{-1}}{2!} = \frac{e^{-1}}{2} = \frac{1}{2e} = 0.1839 \qquad \text{o sea} \qquad 0.18$$

En general, la aproximación de Poisson es buena si $p \leqslant 0.1$ y $\lambda = Np \leqslant 5$.

7.28. Si la probabilidad de que un individuo sufra una reacción negativa ante una inyección de cierto suero es 0.001, hallar la probabilidad de que entre 2000 individuos: (*a*) exactamente 3 y (*b*) más de 2 de ellos reaccionen negativamente.

Solución

$$\Pr\{X \text{ individuos reaccionen negativamente}\} = \frac{\lambda^X e^{-\lambda}}{X!} = \frac{2^X e^{-2}}{X!}$$

donde $\lambda = Np = (2000)(0.001) = 2$.

(*a*)

$$\Pr\{3 \text{ individuos reaccionen negativamente}\} = \frac{2^3 e^{-2}}{3!} = \frac{4}{3e^2} = 0.180$$

(*b*)

$$\Pr\{0 \text{ la sufran}\} = \frac{2^0 e^{-2}}{0!} = \frac{1}{e^2} \qquad \Pr\{1 \text{ la sufra}\} = \frac{2^1 e^{-2}}{1!} = \frac{2}{e^2} \qquad \Pr\{2 \text{ la sufran}\} = \frac{2^2 e^{-2}}{2!} = \frac{2}{e^2}$$

$$\Pr\{\text{más de 2 la sufran}\} = 1 - \Pr\{0 \text{ ó } 1 \text{ ó } 2 \text{ la sufran}\}$$

$$= 1 - \left(\frac{1}{e^2} + \frac{2}{e^2} + \frac{2}{e^2}\right) = 1 - \frac{5}{e^2} = 0.323$$

Nótese que de acuerdo con la distribución binomial las probabilidades solicitadas en (*a*) y (*b*) son, respectivamente,

(*a*) $\binom{2000}{3}(0.001)^3(0.999)^{1997}$

(*b*) $1 - \left\{\binom{2000}{0}(0.001)^0(0.999)^{2000} + \binom{2000}{1}(0.001)^1(0.999)^{1999} + \binom{2000}{2}(0.001)^2(0.999)^{1998}\right\}$

mucho más difíciles de evaluar directamente.

7.29. Una distribución de Poisson viene dada por

$$p(X) = \frac{(0.72)^X e^{-0.72}}{X!}$$

Calcular: (*a*) $p(0)$, (*b*) $p(1)$, (*c*) $p(2)$ y (*d*) $p(3)$.

Solución

(*a*)

$$p(0) = \frac{(0.72)^0 e^{-0.72}}{0!} = \frac{(1)e^{-0.72}}{1} = e^{-0.72} = 0.4868 \quad \text{usando el Apéndice VIII}$$

(*b*)

$$p(1) = \frac{(0.72)^1 e^{-0.72}}{1!} = 0.72e^{-0.72} = (0.72)(0.4868) = 0.3505$$

(*c*)

$$p(2) = \frac{(0.72)^2 e^{-0.72}}{2!} = \frac{(0.5184)e^{-0.72}}{2} = (0.2592)(0.4868) = 0.1262$$

Otro método

$$p(2) = \frac{0.72}{2}\,p(1) = (0.36)(0.3505) = 0.1262$$

(*d*)

$$p(3) = \frac{(0.72)^3 e^{-0.72}}{3!} = \frac{0.72}{3}\,p(2) = (0.24)(0.1262) = 0.0303$$

DISTRIBUCION MULTINOMIAL

Una caja contiene 5 bolas rojas, 4 blancas y 3 azules. Se saca al azar una bola de la caja, se anota su color y se vuelve a meter en la caja. Hallar la probabilidad de que entre 6 bolas así seleccionadas, 3 sean rojas, 2 blancas y 1 azul.

Solución

Pr{roja en cualquier extracción} = $\frac{5}{12}$, Pr{blanca en cualquier extracción} = $\frac{4}{12}$, Pr{azul en cualquier extracción} = $\frac{3}{12}$; luego

$$\Pr\{3 \text{ son rojas, } 2 \text{ son blancas, } 1 \text{ es azul}\} = \frac{6!}{3!2!1!}\left(\frac{5}{12}\right)^3\left(\frac{4}{12}\right)^2\left(\frac{3}{12}\right)^1 = \frac{625}{5184}$$

AJUSTE DE DATOS MEDIANTE DISTRIBUCIONES TEORICAS

7.31. Ajustar una distribución binomial a los datos del Problema 2.17.

Solución

$\Pr\{X \text{ caras en una tirada de 5 monedas}\} = p(X) = \binom{5}{X} p^X q^{5-X}$, donde p y q son las respectivas probabilidades de cara y cruz en una sola tirada. Por el Problema 7.11(*a*), el número medio de caras es $\mu = Np = 5p$. Para la distribución de frecuencias realmente observada, el número medio de caras es

$$\frac{\sum fX}{\sum f} = \frac{(38)(0) + (144)(1) + (342)(2) + (287)(3) + (164)(4) + (25)(5)}{1000} = \frac{2470}{1000} = 2.47$$

Igualando la media teórica con la observada, $5p = 2.47$, o sea $p = 0.494$. Luego la distribución binomial de ajuste viene dada por $p(X) = \binom{5}{X}(0.494)^X(0.506)^{5-X}$.

La Tabla 7.4 recoge las probabilidades así como las frecuencias esperadas (teóricas) y observadas. Se ve que el ajuste es bueno. Su bondad se investigará en el Problema 12.12.

Tabla 7.4

Número de caras (X)	$\Pr\{X \text{ caras}\}$	Frecuencia esperada	Frecuencia observada
0	0.0332	33.2, o sea 33	38
1	0.1619	161.9, o sea 162	144
2	0.3162	316.2, o sea 316	342
3	0.3087	308.7, o sea 309	287
4	0.1507	150.7, o sea 151	164
5	0.0294	29.4, o sea 29	25

7.32. Usar papel gráfico de probabilidad para determinar si la distribución de frecuencias de la Tabla 2.1 puede aproximarse bien por una distribución normal.

Solución

Primero se convierte la distribución de frecuencias dada en una distribución de frecuencias relativas acumuladas, como indica la Tabla 7.5. Entonces, las frecuencias relativas acumuladas, expresadas en porcentajes, se marcan en el gráfico del papel especial citado (Fig. 7.10). El grado en que tales puntos caen sobre una recta determina la precisión del ajuste de la distribución dada a una distribución normal. De lo anterior vemos que hay una distribución normal que ajusta muy bien los datos (véase el Problema 7.33).

Tabla 7.5

Altura (in)	Frecuencia relativa acumulada (%)
Menor que 62.5	5.0
Menor que 65.5	23.0 (c_1)
Menor que 68.5	65.0
Menor que 71.5	92.0
Menor que 74.5	100.0

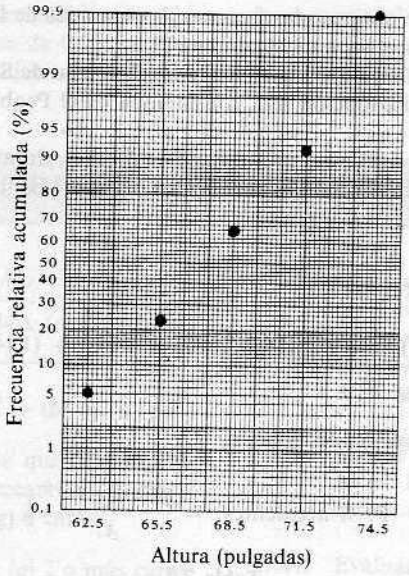


Figura 7.10.

7.33. Ajustar con una curva normal los datos de la Tabla 2.1.

Solución

El método lo esboza la Tabla 7.6. Al calcular z para las fronteras de clase, usamos $z = (X - \bar{X})/s$, donde la media $\bar{X}$ y la desviación típica s se han obtenido, respectivamente, en los Problemas 3.22 y 4.17.

Tabla 7.6

Alturas (in)	Fronteras de clase (X)	z para fronteras de clase	Area bajo la curva normal desde 0 a z	Area para cada clase	Frecuencia esperada	Frecuencia observada
	59.5	−2.72	0.4967			
60-62				0.0413	4.13, o sea 4	5
	62.5	−1.70	0.4554			
63-65				0.2068	20.68, o sea 21	18
	65.5	−0.67	0.2486			
66-68			} Suma	→0.3892	38.92, o sea 39	42
	68.5	0.36	0.1406			
69-71				0.2771	27.71, o sea 28	27
	71.5	1.39	0.4177			
72-74				0.0743	7.43, o sea 7	8
	74.5	2.41	0.4920			
Suma						

$\bar{X} = 67.45$ in $\quad s = 2.92$ in

En la columna 4 de la Tabla 7.6, las áreas bajo la curva normal entre 0 y z se han obtenido del Apéndice II. De ahí hallamos las áreas bajo la curva normal entre sucesivos valores de z, como muestra la columna 5. Se obtienen sin más que restar las áreas sucesivas de la columna 4 cuando las correspondientes z tienen el mismo signo, y sumando si son de signo opuesto (lo que ocurre sólo una vez en la tabla).

Multiplicando las entradas de la columna 5 (que representan frecuencias relativas) por la frecuencia

total N (en este caso $N = 100$) se obtienen las frecuencias esperadas de la columna 6. Veamos que hay buen acuerdo con las frecuencias observadas (columna 7).

Si se desea, puede emplearse la desviación típica con corrección de Sheppard [véase Prob. 4.21(*a*)]. La bondad del ajuste de la distribución será considerada en el Problema 12.13.

7.34. La Tabla 7.7 muestra el número f de días, en un plazo de 50 días, durante los cuales se produjeron X accidentes de automóvil en una cierta ciudad. Ajustar los datos mediante una distribución de Poisson.

Solución

El número medio de accidentes es

$$\lambda = \frac{\sum fX}{\sum f} = \frac{(21)(0) + (18)(1) + (7)(2) + (3)(3) + (1)(4)}{50} = \frac{45}{50} = 0.90$$

Luego, de acuerdo con la distribución de Poisson,

$$\Pr\{X \text{ accidentes}\} = \frac{(0.90)^X e^{-0.90}}{X!}$$

Tabla 7.7

Número de accidentes (X)	Número de días (f)
0	21
1	18
2	7
3	3
4	1
	Total 50

La Tabla 7.8 da las probabilidades de 0, 1, 2, 3 y 4 accidentes que predice la distribución de Poisson y el número esperado o teórico en los cuales se producen X accidentes (obtenidos multiplicando las respectivas probabilidades por 50). Para facilitar la comparación, la columna 4 repite el número real de días de la Tabla 7.7.

Tabla 7.8

Número de accidentes (X)	$\Pr\{X \text{ accidentes}\}$	Número esperado de días	Número real de días
0	0.4066	20.33, o sea 20	21
1	0.3659	18.30, o sea 18	18
2	0.1647	8.24, o sea 8	7
3	0.0494	2.47, o sea 2	3
4	0.0111	0.56, o sea 1	1

Nótese que el ajuste es bueno.

Para una verdadera distribución de Poisson, la varianza $\sigma^2 = \lambda$. El cálculo de la varianza de la distribución propuesta nos da 0.97, que se compara favorablemente con el valor 0.90 para λ, lo que añade más evidencia a lo adecuado de la distribución de Poisson como aproximación de nuestros datos.

PROBLEMAS SUPLEMENTARIOS

LA DISTRIBUCION BINOMIAL

7.35. Evaluar: (*a*) 7!, (*b*) 10!/(6!4!), (*c*) $\binom{9}{5}$, (*d*) $\binom{11}{8}$ y (*e*) $\binom{6}{1}$.

7.36. Desarrollar: (*a*) $(q + p)^7$ y (*b*) $(q + p)^{10}$

7.37. Hallar la probabilidad de que al lanzar 6 veces una moneda aparezcan: (*a*) 0, (*b*) 1, (*c*) 2, (*d*) 3, (*e*) 4, (*f*) 5 y (*g*) 6 caras.

7.38. Hallar la probabilidad de: (*a*) 2 o más caras, y (*b*) menos de 4 caras, en una tirada de 6 monedas.

7.39. Si X denota el número de caras en una sola tirada de 4 monedas, hallar: (*a*) $\Pr\{X = 3\}$, (*b*) $\Pr\{X < 2\}$, (*c*) $\Pr\{X \leq 2\}$ y (*d*) $\Pr\{1 < X \leq 3\}$.

7.40. Entre 800 familias con 5 hijos, ¿cuántas cabe esperar que tengan: (*a*) 3 chicos, (*b*) 5 chicas y (*c*) 2 ó 3 chicos? Se suponen probabilidades iguales para chicos y chicas.

7.41. Hallar la probabilidad de obtener una suma de 11 puntos (*a*) una vez y (*b*) dos veces, en dos lanzamientos de un par de dados.

7.42. ¿Cuál es la probabilidad de sacar 9 exactamente una vez en 3 lanzamientos de un par de dados?

7.43. Hallar la probabilidad de acertar al azar la respuesta de al menos 6 de entre 10 cuestiones en un test verdadero-falso.

7.44. Un agente de seguros contrata 5 pólizas con personas de la misma edad y de buena salud. Según las tablas en uso, la probabilidad de que un hombre de esa edad esté vivo dentro de 30 años es $\frac{2}{3}$. Hallar la probabilidad de que dentro de 30 años vivan: (*a*) los 5, (*b*) al menos 3, (*c*) sólo 2 y (*d*) al menos uno.

7.45. Calcular: (*a*) la media, (*b*) la desviación típica, (*c*) el coeficiente momento de sesgo y (*d*) el coeficiente momento de curtosis, para una distribución binomial en la que $p = 0.7$ y $N = 60$. Interpretar los resultados.

7.46. Probar que si una distribución binomial con $N = 100$ es simétrica, su coeficiente momento de curtosis es 2.98.

7.47. Evaluar: (*a*) $\sum (X - \mu)^3 p(X)$ (*b*) $\sum (X - \mu)^4 p(X)$ para la distribución binomial.

7.48. Probar las fórmulas (1) y (2) del comienzo de este capítulo para los coeficientes momento de sesgo y curtosis.

LA DISTRIBUCION NORMAL

7.49. En un examen de estadística, la media fue 78 y la desviación típica 10.

(*a*) Determinar las puntuaciones estándar de dos estudiantes que obtuvieron 93 y 62 puntos.

(*b*) Hallar las puntuaciones de dos estudiantes cuyas puntuaciones estándar fueron -0.6 y 1.2.

7.50. Hallar: (*a*) la media y (*b*) la desviación típica en un examen en el que las notas 70 y 88 correspondieron a puntuaciones estándar de -0.6 y 1.4, respectivamente.

7.51. Hallar el área bajo la curva normal entre: (*a*) $z = -1.20$ y $z = 2.40$, (*b*) $z = 1.23$ y $z = 1.87$, (*c*) $z = -2.35$ y $z = -0.50$.

7.52. Hallar el área bajo la curva normal: (*a*) a la izquierda de $z = -1.78$, (*b*) a la izquierda de $z = 0.56$, (*c*) a la derecha de $z = -1.45$, (*d*) correspondiente a $z \geq 2.16$, (*e*) corres-

pondiente a $-0.80 \leqslant z \leqslant 1.53$ y (f) a la izquierda de $z = -2.52$ y a la derecha de $z = 1.83$.

7.53. Si z está normalmente distribuida con media 0 y varianza 1, hallar: (a) $\Pr\{z \geqslant -1.64\}$, (b) $\Pr\{-1.96 \leqslant z \leqslant 1.96\}$, ($c$) $\Pr\{|z| \geqslant 1\}$.

7.54. Hallar el valor de z tal que: (a) el área a su derecha sea 0.2266, (b) el área a su izquierda sea 0.0314, (c) el área entre -0.23 y z sea 0.5722, (d) el área entre 1.15 y z sea 0.0730 y (e) el área entre $-z$ y z sea 0.9000.

7.55. Hallar z_1 si $\Pr\{z \geqslant z_1\} = 0.84$, donde z está normalmente distribuida con media 0 y varianza 1.

7.56. Hallar las ordenadas de la curva normal en: (a) $z = 2.25$, (b) $z = -0.32$ y (c) $z = -1.18$.

7.57. Si las alturas de 300 estudiantes están normalmente distribuidas con media 68.0 in y desviación típica 3.0 in, ¿cuántos estudiantes tienen altura: (a) mayor que 72 in, (b) menor o igual que 64 in, (c) entre 65 y 71 in inclusive y (d) de 68 in? Se supone que las alturas se han medido con precisión de 1 in.

7.58. Si los diámetros de las bolas de cojinetes están normalmente distribuidas con media 0.6140 in y desviación típica 0.0025 in, determinar el porcentaje de ellas con diámetros: (a) entre 0.610 y 0.618 inclusive, (b) mayores que 0.617 in, (c) menores que 0.608 in y (d) iguales a 0.615 in.

7.59. La nota media en un examen es 72 y la desviación típica 9. El 10% del curso recibirá grado A. ¿Cuál es la nota mínima para optar a él?

7.60. Si un conjunto de medidas está normalmente distribuida, ¿qué porcentaje de ellas difiere de la media: (a) más de 0.5 desviaciones típicas y (b) menos de 0.75 desviaciones típicas?

7.61. Si $\bar{X}$ es la media y s la desviación típica de un conjunto de medidas normalmente distribuidas, ¿qué porcentaje de ellas: (a) cae en el rango $\bar{X} \pm 2s$, (b) fuera del rango $\bar{X} \pm 1.2s$ y (c) son mayores que $\bar{X} - 1.5s$?

7.62. En el Problema 7.61, hallar a de manera que el porcentaje de casos: (a) en el rango $\bar{X} \pm as$ sea el 75% y (b) menor que $\bar{X} - as$ sea 22%.

APROXIMACION NORMAL A LA DISTRIBUCION BINOMIAL

7.63. Hallar la probabilidad de que en 200 lanzamientos de una moneda haya: (a) entre 80 y 120 caras inclusive, (b) menos de 90 caras, (c) menos de 85 o más de 115 caras y (d) 100 caras exactamente.

7.64. Hallar la probabilidad de que en un test verdadero-falso un estudiante conjeture acertadamente: (a) 12 o más de 20 y (b) 24 o más de 40 cuestiones.

7.65. El 10% de las piezas producidas en una máquina son defectuosas. Hallar la probabilidad de que en una muestra aleatoria de 400 piezas sean defectuosas: (a) a lo sumo 30, (b) entre 30 y 50, (c) entre 35 y 45 y (d) 55 o más.

7.66. Hallar la probabilidad de obtener más de 25 veces 7 en 100 tiradas de un par de dados.

LA DISTRIBUCION DE POISSON

7.67. Si el 3% de las válvulas manufacturadas por una compañía son defectuosas, hallar la probabilidad de que en una muestra de 100 válvulas: (a) 0, (b) 1, (c) 2, (d) 3, (e) 4 y (f) 5 sean defectuosas.

7.68. En el Problema 7.67, hallar la probabilidad de que sean defectuosas: (a) más de 5, (b) entre 1 y 3, (c) no más de 2 válvulas.

7.69. Una bolsa contiene 1 ficha roja y 7 blancas. Se saca una al azar, se anota su color y se devuelve a la bolsa, tras lo cual se remueven de nuevo. Usando: (a) la distribución binomial y (b) la aproximación de Poisson a la distribución binomial, hallar la probabilidad de que en 8 de esas extracciones salga la roja 3 veces exactamente.

7.70. De acuerdo con la National Office of Vital Statistics of the U.S. Department of Health,

Education, and Welfare, el número medio de ahogados por accidente al año en EE.UU. es 3.0 por cada 100,000 habitantes. Hallar la probabilidad de que en una ciudad de 200,000 habitantes haya: (*a*) 0, (*b*) 2, (*c*) 6, (*d*) 8, (*e*) entre 4 y 8 y (*f*) menos de 3 ahogados por accidente al año.

7.71. Entre las 2 y las 4 P.M., el número medio de llamadas telefónicas por minuto que recibe una centralita es 2.5. Hallar la probabilidad de que durante un minuto concreto se produzcan: (*a*) 0, (*b*) 1, (*c*) 2, (*d*) 3, (*e*) 4 o menos y (*f*) más de 6 llamadas.

LA DISTRIBUCION MULTINOMIAL

7.72. Se lanza un dado 6 veces. Hallar la probabilidad de que: (*a*) salgan 1 uno, 2 doses y 3 treses y (*b*) que salga cada número una vez.

7.73. Una caja contiene un gran número de fichas rojas, blancas, azules y amarillas, en la proporción 4 : 3 : 2 : 1, respectivamente. Hallar la probabilidad de que en 10 extracciones salgan: (*a*) 4 rojas, 3 blancas, 2 azules y 1 amarilla y (*b*) 8 rojas y 2 amarillas.

7.74. Hallar la probilidad de no sacar ni 1, ni 2, ni 3 en cuatro tiradas de un dado.

AJUSTE DE DATOS MEDIANTE DISTRIBUCIONES TEORICAS

7.75. Ajustar una distribución binomial a los datos de la Tabla 7.9.

Tabla 7.9

X	0	1	2	3	4
f	30	62	46	10	2

7.76. Determinar, usando papel gráfico de probabilidad, si los datos del Problema 3.59 se pueden aproximar bien por una distribución normal.

7.77. Ajustar una distribución normal a los datos del Problema 3.59.

7.78. Ajustar una distribución normal a los datos del Problema 3.61.

7.79. Ajustar una distribución de Poisson a los datos del Problema 7.75 y comparar este ajuste con el obtenido mediante la distribución binomial.

7.80. La Tabla 7.10 muestra el número de muertos al año por unidad, a causa de coces de los caballos, entre 10 unidades del ejército prusiano en un período de 20 años (1875 a 1894). Ajustar una distribución de Poisson a esos datos.

Tabla 7.10

X	0	1	2	3	4
f	109	65	22	3	1

CAPITULO 8

Teoría elemental del muestreo

TEORIA DEL MUESTREO

La *teoría del muestreo* estudia la relación entre una población y las muestras tomadas de ella. Es de gran utilidad en muchos campos. Por ejemplo, para *estimar* magnitudes desconocidas de una población, tales como media y varianza, llamadas a menudo *parámetros* de la población o simplemente parámetros, a partir del conocimiento de esas magnitudes sobre muestras, que se llaman *estadísticos de la muestra* o simplemente *estadísticos*. Los problemas de estimación se consideran en el Capítulo 9.

La teoría del muestreo es también útil para determinar si las diferencias observadas entre dos muestras son debidas a variaciones fortuitas o si son realmente significativas. Tales cuestiones aparecen, por ejemplo, al probar un nuevo suero como tratamiento de una enfermedad o al decidir si un proceso de producción es mejor que otro. Las respuestas implican el uso de los llamados *contrastes (o tests) de hipótesis y de significación*, importantes en la *teoría de las decisiones*, considerada en el Capítulo 10.

En general, un estudio de las inferencias hechas sobre una población a partir de muestras suyas, con indicación de la precisión de tales inferencias, se llama *inferencia estadística*.

MUESTRAS ALEATORIAS Y NUMEROS ALEATORIOS

Para que las conclusiones de la teoría del muestreo y de la inferencia estadística sean válidas, las muestras deben escogerse *representativas* de la población. El análisis de los métodos de muestreo y problemas relacionados se llama el *diseño del experimento*.

Una forma de obtener una muestra representativa es mediante *muestreo aleatorio*, de acuerdo con el cual, cada miembro de la población tiene la misma probabilidad de ser incluido en la muestra. Un método para lograrlo es asignarles a cada uno un número, escribir cada número en una papeleta, y realizar en una urna un sorteo justo con ellas. Un método alternativo consiste en recurrir a una tabla de *números aleatorios* (véase Apéndice IX) especialmente construida al efecto. Véase Problema 8.6.

MUESTREO CON Y SIN REPOSICION

Si sacamos un número de una urna, podemos volverlo a poner en ella o no, antes de la siguiente extracción. En el primer caso, ese número puede salir de nuevo más veces, mientras que en el

segundo sólo puede salir cada número una vez. Estos dos tipos de muestreo se llaman, respectivamente, *muestreo con reposición* y *muestreo sin reposición*.

Las poblaciones son finitas o infinitas. Si, por ejemplo, sacamos 10 bolas sucesivamente, sin reposición, de una urna que contiene 100 bolas, estamos tomando muestra en una población finita; mientras que si lanzamos 50 veces una moneda y contamos el número de caras, estamos ante una muestra de una población infinita.

Una población finita en la que se efectúa muestreo con reposición, puede considerarse infinita teóricamente, ya que se puede tomar cualquier número de muestras sin agotarla. Para muchos efectos prácticos, una población muy grande se puede considerar como si fuera infinita.

DISTRIBUCIONES DE MUESTREO

Consideremos todas las posibles muestras de tamaño N en una población dada (con o sin reposición). Para cada muestra, podemos calcular un estadístico (tal como la media o la desviación típica) que variará de muestra a muestra. De esta manera obtenemos una distribución del estadístico que se llama su *distribución de muestreo*.

Si, por ejemplo, el estadístico utilizado es la media muestral, entonces la distribución se llama la *distribución de muestreo de medias*, o *distribución de muestreo de la media*. Análogamente, podríamos tener distribución de muestreo de la desviación típica, de la varianza, de la mediana, de las proporciones, etcétera.

Para cada distribución de muestreo podemos calcular la media, la desviación típica, etc. Así pues, podremos hablar de la media y la desviación típica de la distribución de muestreo de medias, etcétera.

DISTRIBUCION DE MUESTREO DE MEDIAS

Supongamos que se toman todas las posibles muestras de tamaño N, sin reposición, de una población finita de tamaño $N_p > N$. Si denotamos la media y la desviación típica de la distribución de muestreo de medias por $\mu_{\bar{X}}$ y $\sigma_{\bar{X}}$ y las de la población por μ y σ, respectivamente, entonces

$$\mu_{\bar{X}} = \mu \quad \text{y} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}\sqrt{\frac{N_p - N}{N_p - 1}} \tag{1}$$

Si la población es infinita o si el muestreo es con reposición, los resultados anteriores se reducen a

$$\mu_{\bar{X}} = \mu \quad \text{y} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} \tag{2}$$

Para valores grandes de N ($N \geq 30$), la distribución de muestreo de medias es aproximadamente normal con media $\mu_{\bar{X}}$ y desviación típica $\sigma_{\bar{X}}$, independientemente de la población (en tanto en cuanto la media poblacional y la varianza sean finitas y el tamaño de la población sea al menos doble que el de la muestra). Este resultado para una población infinita es un caso especial del *teorema del límite central* de teoría avanzada de probabilidades, que afirma que la precisión de la

aproximación mejora al crecer N. Esto se indica en ocasiones diciendo que la distribución de muestreo es *asintóticamente normal.*

En caso de que la población esté normalmente distribuida, la distribución de muestreo de medias también lo está, incluso para pequeños valores de N (o sea, $N < 30$).

DISTRIBUCION DE MUESTREO DE PROPORCIONES

Supongamos que una población es infinita y que la probabilidad de ocurrencia de un suceso (su éxito) es p, mientras la probabilidad de que no ocurra es $q = 1 - p$. Por ejemplo, la población puede ser la de todas las posibles tiradas de una moneda, en la que la probabilidad del suceso «cara» es $p = \frac{1}{2}$. Consideremos todas las posibles muestras de tamaño N de tal población, y para cada una de ellas determinemos la proporción de éxitos P. En el caso de una moneda, P sería la proporción de caras en N tiradas. Obtenemos así una *distribución de muestreo de proporciones* cuya media μ_P y cuya desviación típica σ_P vienen dadas por

$$\mu_P = p \qquad \text{y} \qquad \sigma_P = \sqrt{\frac{pq}{N}} = \sqrt{\frac{p(1-p)}{N}} \tag{3}$$

que se pueden obtener de (2) poniendo $\mu = p$ y $\sigma = \sqrt{pq}$. Para valores grandes de N ($N \geq 30$), la distribución de muestreo está, muy aproximadamente, normalmente distribuida. Nótese que la población está *binomialmente distribuida.*

Las ecuaciones (3) son válidas también para una población finita en la que se hace muestreo con reposición. Para poblaciones finitas en que se haga muestreo sin reposición, las ecuaciones (3) quedan sustituidas por las ecuaciones (1) con $\mu = p$ y $\sigma = \sqrt{pq}$.

Notemos que (3) se deducen fácilmente dividiendo la media y la desviación típica (Np y $\sqrt{Npq}$) de la distribución binomial por N (véase Cap. 7).

DISTRIBUCION DE MUESTREO DE DIFERENCIAS Y SUMAS

Sean dadas dos poblaciones. Para cada muestra de tamaño N_1 de la primera, calculamos un estadístico S_1; eso da una distribución de muestreo para S_1, cuya media y desviación típica denotaremos por μ_{S1} y σ_{S1}. Del mismo modo, para cada muestra de tamaño N_2 de la segunda población, calculamos un estadístico S_2; eso nos da una distribución de muestreo para S_2, cuya media y desviación típica denotaremos por μ_{S2} y σ_{S2}. De todas las posibles combinaciones de estas muestras de las dos poblaciones podemos obtener una distribución de las diferencias, $S_1 - S_2$, que se llama *distribución de muestreo de las diferencias de los estadísticos.* La media y la desviación típica de esta distribución de muestreo, denotadas respectivamente por μ_{S1-S2} y σ_{S1-S2}, vienen dadas por

$$\mu_{S1-S2} = \mu_{S1} - \mu_{S2} \qquad \text{y} \qquad \sigma_{S1-S2} = \sqrt{\sigma_{S1}^2 + \sigma_{S2}^2} \tag{4}$$

supuesto que las muestras escogidas no dependan en absoluto una de otra (o sea, que sean *independientes*).

Si S_1 y S_2 son las medias muestrales de ambas poblaciones, cuyas medias denotaremos por $\bar{X}_1$ y $\bar{X}_2$, respectivamente, entonces la distribución de muestreo de las diferencias de medias viene dada para poblaciones infinitas con medias y desviaciones típicas (μ_1, σ_1) y (μ_2, σ_2), respectivamente por

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2 \qquad \text{y} \qquad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} \tag{5}$$

usando las ecuaciones (2). El resultado es válido también para poblaciones finitas si el muestreo es con reposición. Análogos resultados pueden alcanzarse para poblaciones finitas en que el muestreo sea sin reposición, usando (1).

Resultados correspondientes se pueden obtener para las distribuciones de muestreo de diferencias de proporciones de dos poblaciones binomialmente distribuidas con parámetros (p_1, q_1) y (p_2, q_2), respectivamente. En este caso, S_1 y S_2 corresponden a la proporción de éxitos P_1 y P_2, y las ecuaciones (4) llevan a

$$\mu_{P_1 - P_2} = \mu_{P_1} - \mu_{P_2} = p_1 - p_2 \qquad \text{y} \qquad \sigma_{P_1 - P_2} = \sqrt{\sigma^2_{P_1} + \sigma^2_{P_2}} = \sqrt{\frac{p_1 q_1}{N_1} + \frac{p_2 q_2}{N_2}} \tag{6}$$

Si N_1 y N_2 son grandes (N_1, $N_2 \geq 30$), la distribución de muestreo de diferencias de medias o proporciones están casi normalmente distribuidas.

A veces es útil hablar de la *distribución de muestreo de la suma de estadísticos*. La media y la desviación típica de tal distribución son

$$\mu_{S_1 + S_2} = \mu_{S_1} + \mu_{S_2} \qquad \text{y} \qquad \sigma_{S_1 + S_2} = \sqrt{\sigma^2_{S_1} + \sigma^2_{S_2}} \tag{7}$$

supuesto que las muestras sean independientes.

ERRORES TIPICOS

La desviación típica de una distribución de muestreo de un estadístico se suele llamar su *error típico*. La Tabla 8.1 presenta errores típicos de distribución de muestreo para varios estadísticos bajo las condiciones de muestreo aleatorio de una población infinita (o muy grande) o de muestreo con reposición de una finita. También recoge observaciones particulares que garantizan la validez de estos resultados y otras notas pertinentes.

Las cantidades μ, σ, p, μ_r y $\bar{X}$, s, P, m_r denotan, respectivamente, las medias de la población y de la muestra, las desviaciones típicas, proporciones y r-ésimos momentos respecto de la media.

Hay que hacer notar que si el tamaño de la muestra es lo bastante grande, las distribuciones de muestreo son normales o casi normales. Por ello, los métodos se conocen como *métodos de grandes muestras*. Cuando $N < 30$, las muestras se llaman pequeñas. La teoría de *pequeñas muestras* o *teoría exacta del muestreo*, como se le llama a veces, se trata en el Capítulo 11.

Cuando los parámetros de la población, tales como σ, p o μ_r son desconocidos, pueden ser estimados con precisión por sus correspondientes estadísticos muestrales, a saber, s (o sea $\hat{s} = \sqrt{N/(N-1)}\,s$), P y m_r si las muestras son suficientemente grandes.

Tabla 8.1. Errores típicos para algunas distribuciones de muestreo

Distribución de muestreo	Error típico	Observaciones
Medias	$\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{N}}$	Esto es cierto para muestras grandes y pequeñas. La distribución muestral de medias es casi normal para $N \geqslant 30$, incluso cuando la población no es normal. $\mu_{\bar{X}} = \mu$, la media de la población, en todos los casos.
Proporciones	$\sigma_P = \sqrt{\dfrac{p(1-p)}{N}} = \sqrt{\dfrac{pq}{N}}$	La nota precedente para las medias se aplica aquí también. $\mu_P = p$ en todos los casos.
Desviaciones típicas	(1) $\sigma_s = \dfrac{\sigma}{\sqrt{2N}}$ (2) $\sigma_s = \sqrt{\dfrac{\mu_4 - \mu_2^2}{4N\mu_2}}$	Para $N \geqslant 100$, la distribución muestral de s es casi normal. σ_s viene dada por (1) sólo si la población es normal (o aproximadamente normal). Si la población no es normal, se puede usar (2). Nótese que (2) se reduce a (1) cuando $\mu_2 = \sigma^2$ y $\mu_4 = 3\sigma^4$, lo cual es cierto para poblaciones normales. Para $N \geqslant 100$, $\mu_s = \sigma$ muy aproximadamente.
Medianas	$\sigma_{\text{med}} = \sigma\sqrt{\dfrac{\pi}{2N}} = \dfrac{1.2533\sigma}{\sqrt{N}}$	Para $N \geqslant 30$, la distribución de muestreo de la mediana es muy aproximadamente normal. El resultado dado es válido sólo si la población es normal (o casi normal). $\mu_{\text{med}} = \mu$
Primer y tercer cuartiles	$\sigma_{Q1} = \sigma_{Q3} = \dfrac{1.3626\sigma}{\sqrt{N}}$	Los comentarios hechos para las medianas se aplican aquí también. μ_{Q1} y μ_{Q3} son casi iguales al primer y tercer cuartiles de la población. Nótese que $\sigma_{Q2} = \sigma_{\text{med}}$
Deciles	$\sigma_{D1} = \sigma_{D9} = \dfrac{1.7094\sigma}{\sqrt{N}}$ $\sigma_{D2} = \sigma_{D8} = \dfrac{1.4288\sigma}{\sqrt{N}}$ $\sigma_{D3} = \sigma_{D7} = \dfrac{1.3180\sigma}{\sqrt{N}}$ $\sigma_{D4} = \sigma_{D6} = \dfrac{1.2680\sigma}{\sqrt{N}}$	De nuevo son aplicables aquí las observaciones hechas en el caso de las medianas. μ_{D1}, μ_{D2}, ... son casi iguales al primer, segundo, ... deciles de la población. Nótese que $\sigma_{D5} = \sigma_{\text{med}}$

Tabla 8.1. (*Continuación*)

Distribución de muestreo	Error típico	Observaciones
Rangos semi-intercuartiles	$\sigma_Q = \dfrac{0.7867\sigma}{\sqrt{N}}$	Las observaciones hechas acerca de las medianas se aplican de nuevo aquí. μ_Q es casi igual al rango semi-intercuartil de la población.
Varianzas	(1) $\sigma_{S^2} = \sigma^2\sqrt{\dfrac{2}{N}}$ (2) $\sigma_{S^2} = \sqrt{\dfrac{\mu_4 - \mu_2^2}{N}}$	Las observaciones hechas sobre la desviación típica son aplicables también aquí. Hagamos notar que (2) da (1) en el caso de poblaciones normales. $\mu_{S^2} = \sigma^2(N-1)/N$, que es casi igual a σ^2 para N grandes.
Coeficientes de varianza	$\sigma_V = \dfrac{v}{\sqrt{2N}}\sqrt{1+2v^2}$	Aquí $v = \sigma/\mu$ es el coeficiente de variación de la población. El resultado dado es válido para poblaciones normales (o casi normales) y $N \geqslant 100$.

PROBLEMAS RESUELTOS

DISTRIBUCION DE MUESTREO DE MEDIAS

8.1. Una población consta de los números 2, 3, 6, 8 y 11. Consideremos todas las posibles muestras de tamaño 2 que pueden tomarse con reposición de esa población. Hallar (*a*) la media de la población, (*b*) la desviación típica de la población, (*c*) la media de la distribución de muestreo de medias y (*d*) la desviación típica de la distribución de muestreo de medias (o sea, el error típico de medias).

Solución

(*a*) $$\mu = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6.0$$

(*b*) $$\sigma^2 = \frac{(2-6)^2+(3-6)^2+(6-6)^2+(8-6)^2+(11-6)^2}{5} = \frac{16+9+0+4+25}{5} = 10.8$$

y $\sigma = 3.29$.

(*c*) Hay $5(5) = 25$ muestras de tamaño 2 que se pueden tomar, con reposición de la población (porque cualquiera de los 5 números de la primera extracción puede asociarse con uno cualquiera de la segunda). Y son

(2, 2)	(2, 3)	(2, 6)	(2, 8)	(2, 11)
(3, 2)	(3, 3)	(3, 6)	(3, 8)	(3, 11)
(6, 2)	(6, 3)	(6, 6)	(6, 8)	(6, 11)
(8, 2)	(8, 3)	(8, 6)	(8, 8)	(8, 11)
(11, 2)	(11, 3)	(11, 6)	(11, 8)	(11, 11)

Las correspondientes medias muestrales son

$$\begin{array}{ccccc} 2.0 & 2.5 & 4.0 & 5.0 & 6.5 \\ 2.5 & 3.0 & 4.5 & 5.5 & 7.0 \\ 4.0 & 4.5 & 6.0 & 7.0 & 8.5 \\ 5.0 & 5.5 & 7.0 & 8.0 & 9.5 \\ 6.5 & 7.0 & 8.5 & 9.5 & 11.0 \end{array} \tag{8}$$

y la media de la distribución de muestreo de medias es

$$\mu_{\bar{X}} = \frac{\text{suma de todas las medias muestrales en (8)}}{25} = \frac{150}{25} = 6.0$$

ilustrando el hecho de que $\mu_{\bar{X}} = \mu$.

(*d*) La varianza $\sigma_{\bar{X}}^2$ de la distribución de muestreo de medias se obtiene restando la media 6 de cada número en (8), elevando al cuadrado el resultado, sumando los 25 números así obtenidos y dividiendo por 25. El resultado final es $\sigma_{\bar{X}}^2 = 135/25 = 5.40$, y por tanto $\sigma_{\bar{X}} = \sqrt{5.40} = 2.32$. Ello ilustra el que para poblaciones finitas y muestreo con reposición (o para poblaciones infinitas), $\sigma_{\bar{X}}^2 = \sigma^2/N$ porque el lado derecho es $10.8/2 = 5.40$, que coincide con el anterior valor.

8.2. Resolver el Problema 8.1 para el caso de muestreo sin reposición.

Solución

Como en las partes (*a*) y (*b*) del Problema 8.1, $\mu = 6$ y $\sigma = 3.29$.

(*c*) Hay $\binom{5}{2}$ muestras de tamaño 2 que se pueden elegir sin reposición (eso significa que sacamos un número y luego otro distinto del anterior) de la población: (2, 3), (2, 6), (2, 8), (2, 11), (3, 6), (3, 8), (3, 11), (6, 8), (6, 11) y (8, 11). La selección (2, 3), por ejemplo, se considera la misma que la (3, 2).

Las correspondientes medias de la muestra son 2.5, 4.0, 5.0, 6.5, 4.5, 5.5, 7.0, 7.0, 8.5 y 9.5, y la media de la distribución de muestreo de medias es

$$\mu_{\bar{X}} = \frac{2.5 + 4.0 + 5.0 + 6.5 + 4.5 + 5.5 + 7.0 + 7.0 + 8.5 + 9.5}{10} = 6.0$$

ilustrando el hecho de que $\mu_{\bar{X}} = \mu$.

(*d*) La varianza de la distribución de muestreo de medias es

$$\sigma_{\bar{X}}^2 = \frac{(2.5 - 6.0)^2 + (4.0 - 6.0)^2 + (5.0 - 6.0)^2 + \cdots + (9.5 - 6.0)^2}{10} = 4.05$$

y $\sigma_{\bar{X}} = 2.01$. Esto ilustra que

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{N}\left(\frac{N_p - N}{N_p - 1}\right)$$

ya que el lado derecho es igual a

$$\frac{10.8}{2}\left(\frac{5 - 2}{5 - 1}\right) = 4.05$$

como se había obtenido antes.

8.3. Las alturas de 3000 estudiantes varones de una Universidad están normalmente distribuidas con media 68.0 in y desviación típica 3.0 in. Si se toman 80 muestras de 25 estudiantes cada una, ¿cuáles serán la media y la desviación típica esperadas de la resultante distribución de muestreo de medias, si el muestreo se hizo (*a*) y con (*b*) sin reposición?

Solución

El número de muestras de tamaño 25 que podrían elegirse de un grupo de 3000 estudiantes con y sin reposición son $(3000)^{25}$ y $\binom{3000}{25}$, que son mucho mayores que 80. Por tanto no obtenemos una verdadera distribución de muestreo de medias, sino sólo una distribución de muestreo *experimental*. No obstante, como el número de muestras es grande, debiera haber buen acuerdo entre ambas distribuciones de muestreo. Así que la media y la desviación típica esperadas deben estar próximas a las de la distribución teórica. Por tanto, tenemos

(*a*) $$\mu_{\bar{X}} = \mu = 68.0 \text{ in} \quad \text{y} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = \frac{3}{\sqrt{25}} = 0.6 \text{ in}$$

(*b*) $$\mu_{\bar{X}} = 68.0 \text{ in} \quad \text{y} \quad \delta_{\bar{X}} = \frac{\sigma}{\sqrt{N}}\sqrt{\frac{N_p - N}{N_p - 1}} = \frac{3}{\sqrt{25}}\sqrt{\frac{3000 - 25}{3000 - 1}}$$

que es sólo muy ligeramente menor que 0.6 in y puede ser considerada, a todos los efectos prácticos, la misma que en muestreo con reposición.

Así pues, esperaríamos que la distribución de muestreo experimental de medias esté casi normalmente distribuida con media 68.0 in y desviación típica 0.6 in.

8.4. ¿En cuántas muestras del Problema 8.3 esperaríamos encontrar una media (*a*) entre 66.8 y 68.3 in y (*b*) menor que 66.4 in?

Solución

La media $\bar{X}$ de una muestra en unidades estándar viene dada aquí por

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - 68.0}{0.6}$$

(*a*) $$66.8 \text{ en unidades estándar} = \frac{66.8 - 68.0}{0.6} = -2.0$$

$$68.3 \text{ en unidades estándar} = \frac{68.3 - 68.0}{0.6} = 0.5$$

Como muestra la Figura 8.1(*a*),

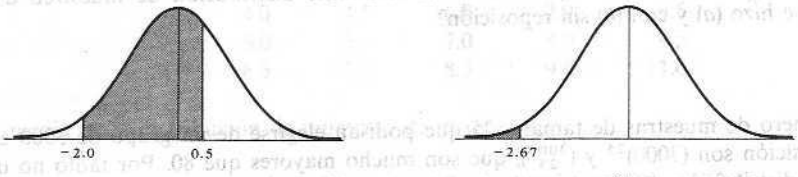
Proporción de muestras con medias entre 66.8 y 68.3 in =

= (área bajo la curva normal entre $z = -2.0$ y $z = 0.5$)

= (área entre $z = -2$ y $z = 0$) + (área entre $z = 0$ y $z = 0.5$)

$= 0.4772 + 0.1915 = 0.6687$

Así pues, el número esperado de muestras es (80)(0.6687) = 53.496, o 53.



(*a*) (*b*)

Figura 8.1.

(*b*) $$66.4 \text{ in en unidades estándar} = \frac{66.4 - 68.0}{0.6} = -2.67$$

Como muestra la Figura 8.1(*b*),

Proporción de muestras con media menor que 66.4 in =

= (área bajo la curva normal a la izquierda de $z = -2.67$)

= (área a la izquierda de $z = 0$) − (*área entre* $z = -2.67$ y $z = 0$)

= 0.5 − 0.4962 = 0.0038

Luego el número esperado de muestras es (80)(0.0038) = 0.304, o cero.

8.5. 500 bolas de cojinete tienen un peso medio de 5.02 gramos (g) y una desviación típica de 0.30 g. Hallar la probabilidad de que una muestra al azar de 100 bolas de ese conjunto tengan un peso total (*a*) entre 496 y 500 g y (*b*) más de 510 g.

Solución

Para la distribución de muestreo de medias, $\mu_{\bar{X}} = \mu = 5.02$ g, y

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}\sqrt{\frac{N_p - N}{N_p - 1}} = \frac{0.30}{\sqrt{100}}\sqrt{\frac{500 - 100}{500 - 1}} = 0.027 \text{ g}$$

(*a*) El peso total estaría entre 496 y 500 g si el peso medio de las 100 bolas está entre 4.96 y 5.00 g.

$$4.96 \text{ en unidades estándar} = \frac{4.96 - 5.02}{0.027} = -2.22$$

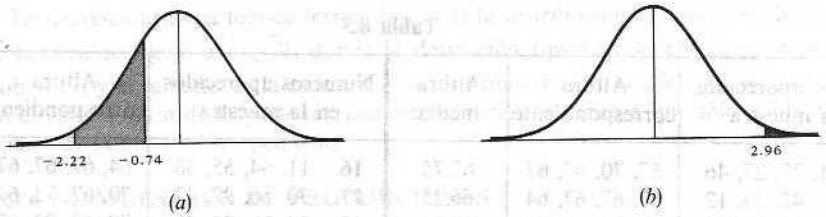
$$5.00 \text{ en unidades estándar} = \frac{5.00 - 5.02}{0.027} = -0.74$$

Como muestra la Figura 8.2(*a*),

Probabilidad pedida = (área entre $z = -2.22$ y $z = -0.74$)

= (área entre $z = -2.22$ y $z = 0$) − (área entre $z = -0.74$ y $z = 0$)

= 0.4868 − 0.2704 = 0.2164



(*a*) (*b*)

Figura 8.2.

(*b*) El peso total excecerá de 510 g si el peso medio de las 100 bolas excede de 5.10 g.

$$5.10 \text{ en unidades estándar} = \frac{5.10 - 5.02}{0.027} = 2.96$$

Como enseña la Figura 8.2(*b*),

$$\begin{aligned}\text{Probabilidad pedida} &= (\text{área a la derecha de } z = 2.96)\\ &= (\text{área a la derecha de } z = 0) - (\text{área entre } z = 0 \text{ y } z = 2.96)\\ &= 0.5 - 0.4985 = 0.0015\end{aligned}$$

Luego sólo hay 3 oportunidades en 2000 de tomar una muestra de 100 bolas que supere los 510 g.

8.6. (*a*) Indicar cómo se seleccionarían al azar 30 muestras de 4 estudiantes cada una (con reposición) de la Tabla 2.1, usando números aleatorios.
(*b*) Hallar la media y la desviación típica de la distribución de muestreo de medias en la parte (*a*).
(*c*) Comparar los resultados de (*b*) con los valores teóricos, explicando cualquier discrepancia.

Solución

(*a*) Usamos dos dígitos para numerar a los 100 estudiantes: 00, 01, 02, ..., 99 (véase Tabla 8.2). Así pues, los 5 estudiantes con pesos 60-62 in están numerados 00-04, los 18 con pesos 63-65 con 05-22, etc. Cada número de estudiante es un *número de muestreo*.

Ahora sacamos números de muestreo de la tabla de números aleatorios (Apéndice IX). En la primera línea vemos 51, 77, 27, 46, 40, etc., que tomamos como números aleatorios de muestreo, cada uno de los cuales da la altura de un estudiante particular. Así, 51 corresponde a un estudiante de 66-68 in, que tomamos como 67 in (la marca de clase). Análogamente, 77, 27 y 46 dan alturas 70, 67 y 67 respectivamente.

Por este proceso se obtiene la Tabla 8.3, que recoge los números de muestreo extraídos, las alturas correspondientes y la altura media para cada una de las 30 muestras. Debemos mencionar que aunque hemos entrado en la tabla de números aleatorios por su primera fila, se podía haber entrado de *cualquier* otra forma.

Tabla 8.2

Altura (in)	Frecuencia	Número de muestreo
60-62	5	00-04
63-65	18	05-22
66-68	42	23-64
69-71	27	65-91
72-74	8	92-99

Tabla 8.3

Números aparecidos en la muestra	Altura correspondiente	Altura media	Números aparecidos en la muestra	Altura correspondiente	Altura media
1. 51, 77, 27, 46	67, 70, 67, 67	67.75	**16.** 11, 64, 55, 58	64, 67, 67, 67	66.25
2. 40, 42, 33, 12	67, 67, 67, 64	66.25	**17.** 70, 56, 97, 43	70, 67, 73, 67	69.25
3. 90, 44, 46, 62	70, 67, 67, 67	67.75	**18.** 74, 28, 93, 50	70, 67, 73, 67	69.25
4. 16, 28, 98, 93	64, 67, 73, 73	69.25	**19.** 79, 42, 71, 30	70, 67, 70, 67	68.50
5. 58, 20, 41, 86	67, 64, 67, 70	67.00	**20.** 58, 60, 21, 33	67, 67, 64, 67	66.25
6. 19, 64, 08, 70	64, 67, 64, 70	66.25	**21.** 75, 79, 74, 54	70, 70, 70, 67	69.25
7. 56, 24, 03, 32	67, 67, 61, 67	65.50	**22.** 06, 31, 04, 18	64, 67, 61, 64	64.00
8. 34, 91, 83, 58	67, 70, 70, 67	68.50	**23.** 67, 07, 12, 97	70, 64, 64, 73	67.75
9. 70, 65, 68, 21	70, 70, 70, 64	68.50	**24.** 31, 71, 69, 88	67, 70, 70. 70	69.25
10. 96, 02, 13, 87	73, 61, 64, 70	67.00	**25.** 11, 64, 21, 87	64, 67, 64, 70	66.25
11. 76, 10, 51, 08	70, 64, 67, 64	66.25	**26.** 03, 58, 57, 93	61, 67, 67, 73	67.00
12. 63, 97, 45, 39	67, 73, 67, 67	68.50	**27.** 53, 81, 93, 88	67, 70, 73, 70	70.00
13. 05, 81, 45, 93	64, 70, 67, 73	68.50	**28.** 23, 22, 96, 79	67, 64, 73, 70	68.50
14. 96, 01, 73, 52	73, 61, 70, 67	67.75	**29.** 98, 56, 59, 36	73, 67, 67, 67	68.50
15. 07, 82, 54, 24	64, 70, 67, 67	67.00	**30.** 08, 15, 08, 84	64, 64, 64, 70	65.50

(*b*) La Tabla 8.4 da la distribución de frecuencias de las alturas medias de las muestras en la parte (*a*). Eso es una *distribución de muestreo de medias*. La media y la desviación típica se obtienen como de costumbre por métodos de compilación (Caps. 3 y 4):

$$\text{Media} = A + c\bar{u} = A + \frac{c\sum fu}{N} = 67.00 + \frac{(0.75)(23)}{30} = 67.58 \text{ in}$$

$$\text{Desviación típica} = c\sqrt{\bar{u}^2 - \bar{u}^2} = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = 0.75\sqrt{\frac{123}{30} - \left(\frac{23}{30}\right)^2} = 1.41 \text{ in}$$

Tabla 8.4

Media muestral	Recuento	f	u	fu	fu^2
64.00	/	1	−4	−4	16
64.75		0	−3	0	0
65.50	//	2	−2	−4	8
66.25	/𝍸 /	6	−1	−6	6
$A \rightarrow$67.00	////	4	0	0	0
67.75	////	4	1	4	4
68.50	/𝍸 //	7	2	14	28
69.25	/𝍸	5	3	15	45
70.00	/	1	4	4	16
		$\sum f = N = 30$		$\sum fu = 23$	$\sum fu^2 = 123$

(*c*) La media teórica de la distribución de muestreo de medias, dada por $\mu_{\bar{x}}$, debiera ser igual a la media μ de la población, que es 67.45 in (véase Prob. 3.22), de acuerdo con el valor 67.58 in de la parte (*b*).

La desviación típica teórica (error típico) de la distribución de muestreo de medias, dada por $\sigma_{\bar{X}}$, debiera ser igual a $\sigma/\sqrt{N}$, donde la desviación típica de la población $\sigma = 2.92$ in (véase Prob. 4.17) y el tamaño de la muestra $N = 4$. Como $\sigma/\sqrt{N} = 2.92/\sqrt{4} = 1.46$ in, hay acuerdo con el valor 1.41 in de la parte (*b*). Las discrepancias se deben a que sólo había 30 muestras y el tamaño de la muestra era pequeño.

DISTRIBUCION DE MUESTREO DE PROPORCIONES

8.7. Hallar la probabilidad de que en 120 lanzamientos de una moneda (*a*) entre el 40% y 60% sean caras y (*b*) $\frac{5}{8}$ o más sean caras.

Solución

Primer método

Consideremos los 120 lanzamientos como una muestra de la población infinita de todos los posibles lanzamientos de la moneda. En esa población, la probabilidad de cara es $p = \frac{1}{2}$ y la de cruz es $q = 1 - p = \frac{1}{2}$.

(*a*) Se pide la probabilidad de que el número de caras en 120 lanzamientos esté entre (40% de 120) = 48 y (60% de 120) = 72. Procederemos como en el Capítulo 7, usando la aproximación normal a la distribución binomial. Puesto que el número de caras es una variable discreta, nos preguntamos por la probabilidad de que el número de caras esté entre 47.5 y 72.5.

$$\mu = \text{números esperados de caras} = Np = 120(\tfrac{1}{2}) = 60 \qquad \text{y} \qquad \sigma = \sqrt{Npq} = \sqrt{(120)(\tfrac{1}{2})(\tfrac{1}{2})} = 5.48$$

$$47.5 \text{ en unidades estándar} = \frac{47.5 - 60}{5.48} = -2.28$$

$$72.5 \text{ en unidades estándar} = \frac{72.5 - 60}{5.48} = 2.28$$

Como indica la Figura 8.3,

$$\begin{aligned}\text{Probabilidad pedida} &= (\text{área bajo la curva normal entre } z = -2.28 \quad \text{y} \quad z = 2.28) \\ &= 2(\text{área entre } z = 0 \quad \text{y} \quad z = 2.28) \\ &= 2(0.4887) = 0.9774\end{aligned}$$

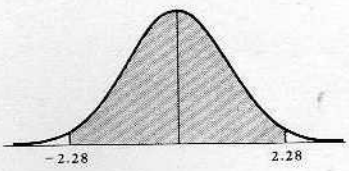


Figura 8.3.

Segundo método

$$\mu_P = p = \tfrac{1}{2} = 0.50 \qquad \sigma_P = \sqrt{\frac{pq}{N}} = \sqrt{\frac{(\frac{1}{2})(\frac{1}{2})}{120}} = 0.0456$$

$$40\% \text{ en unidades estándar} = \frac{0.40 - 0.50}{0.0456} = -2.19$$

$$60\% \text{ en unidades estándar} = \frac{0.60 - 0.50}{0.0456} = 2.19$$

$$\text{Probabilidad pedida} = (\text{área bajo la curva normal entre } z = -21.9 \text{ y } z = 2.19)$$
$$= 2(0.4857) = 0.9714$$

Aunque este resultado es correcto en dos cifras significativas, no coincide exactamente ya que no hemos hecho uso de que la proporción es en realidad una variable discreta. Para tenerlo en cuenta, restamos $1/2N = 1/2(120)$ de 0.40 y sumamos $1/2N = 1/2(120)$ a 0.60; así pues, como $1/240 = 0.00417$, las proporciones pedidas en unidades estándar son

$$\frac{0.40 - 0.00417 - 0.50}{0.0456} = -2.28 \quad \text{y} \quad \frac{0.60 + 0.00417 - 0.50}{0.0456} = 2.28$$

logrando ya el acuerdo con el primer método.

Nótese que $(0.40 - 0.00417)$ y $(0.60 + 0.00417)$ corresponde a las proporciones 47.5/120 y 72.5/120 en el primer método.

(*b*) Usando el segundo método de la parte (*a*), vemos que como $\frac{5}{8} = 0.6250$,

$$(0.6250 - 0.00417) \text{ en unidades estándar} = \frac{0.6250 - 0.00417 - 0.50}{0.0456} = 2.65$$

$$\text{Probabilidad requerida} = (\text{área bajo la curva normal a la derecha de } z = 2.65)$$
$$= (\text{área a la derecha de } z = 0) - (\text{área entre } z = 0 \text{ y } z = 2.65)$$
$$= 0.5 - 0.4960 = 0.0040$$

8.8. Cada persona de un grupo de 500 lanza una moneda 120 veces. ¿Cuántas personas se espera que (*a*) saquen entre 40% y 60% de caras y (*b*) $\frac{5}{8}$ de sus lanzamientos o más de caras?

Solución

Este problema está muy relacionado con el Problema 8.7. Aquí consideramos 500 muestras de tamaño 120 cada una de una población infinita (todos los posibles lanzamientos de la moneda).

(*a*) La parte (*a*) del Problema 8.7 establece que de todas las posibles muestras, consistentes cada una en 120 lanzamientos, podemos esperar un 97.74% con un porcentaje de caras entre 40% y 60%. Luego en 500 muestras cabe esperar unas (97.74% de 500) = 489 muestras con esa propiedad. Por tanto, unas 489 personas verán aparecer entre un 40% y un 60% de caras.

Es interesante notar que 500 − 489 = 11 personas se espera que den porcentajes de caras que no caen entre 40% y 60%. Tales personas pueden razonablemente concluir que sus monedas estaban trucadas, aunque fueran buenas. Este tipo de error es un riesgo omnipresente al tratar con probabilidades.

(*b*) Argumentando como en (*a*), deducimos que unas (500)(0.0040) = 2 personas verían salir $\frac{5}{8}$ o más de sus lanzamientos con cara.

8.9. Se ha encontrado que el 2% de las piezas fabricadas en una cierta máquina son defectuosas. ¿Cuál es la probabilidad de que en un envío de 400 piezas (*a*) el 3% o más y (*b*) el 2% o menos, sean defectuosas?

Solución

$$\mu_P = p = 0.02 \quad \text{y} \quad \sigma_P = \sqrt{\frac{pq}{N}} = \sqrt{\frac{(0.02)(0.98)}{400}} = \frac{0.14}{20} = 0.007$$

(*a*) *Primer método*

Usando la corrección por variables discretas, $1/2N = 1/800 = 0.00125$, tenemos

$$(0.03 - 0.00125) \text{ en unidades estándar} = \frac{0.03 - 0.00125 - 0.02}{0.007} = 1.25$$

Probabilidad requerida = (área bajo la curva normal a la derecha de $z = 1.25$) = 0.1056

Sin corrección se hubiera llegado al valor 0.0764.

Otro método

(3% de 400) = 12 piezas defectuosas. Sobre base continua, 12 o más significa 11.5 o más.

$$\bar{X} = (2\% \text{ de } 400) = 8 \quad \text{y} \quad \sigma = \sqrt{Npq} = \sqrt{(400)(0.02)(0.98)} = 2.8$$

Entonces, 11.5 en unidades estándar = $(11.5 - 8)/2.8 = 1.25$, y como antes la probabilidad pedida es 0.1056.

(*b*) $$(0.02 + 0.00125) \text{ en unidades estándar} = \frac{0.02 + 0.00125 - 0.02}{0.007} = 0.18$$

Probabilidad requerida = (área bajo la curva normal a la izquierda de $z = 0.18$)

$$= 0.5000 + 0.0714 = 0.5714$$

Sin corrección se obtendría 0.5000. El segundo método de la parte (*a*) también es aplicable.

8.10. En unas elecciones uno de los candidatos obtuvo el 46% de los votos. Hallar la probabilidad de qu en un muestreo de (*a*) 200 y (*b*) 1000 votantes elegidos al azar saliera mayoría a su favor.

Solución

(*a*) $$\mu_P = p = 0.46 \quad \text{y} \quad \sigma_P = \sqrt{\frac{pq}{N}} = \sqrt{\frac{(0.46)(0.54)}{200}} = 0.0352$$

Como $1/2N = 1/400 = 0.0025$, la muestra daría una mayoría si la proporción en favor de tal candidato fuese $0.50 + 0.0025 = 0.5025$ o más. (Esta proporción se puede obtener también recordando que 101 o más es mayoría, pero como variable continua eso es 100.5, y por tanto la proporción es $100.5/200 = 0.5025$.)

$$0.5025 \text{ en unidades estándar} = \frac{0.5025 - 0.46}{0.0352} = 1.21$$

Probabilidad requerida = (área bajo la curva normal a la derecha de $z = 1.21$)

$$= 0.5000 - 0.3869 = 0.1131$$

(*b*) $$\mu_P = p = 0.46 \quad \text{y} \quad \sigma_P = \sqrt{\frac{pq}{N}} = \sqrt{\frac{(0.46)(0.54)}{1000}} = 0.0158$$

$$0.5025 \text{ en unidades estándar} = \frac{0.5025 - 0.46}{0.0158} = 2.69$$

$$\text{Probabilidad requerida} = (\text{área bajo la curva normal a la derecha de } z = 2.69)$$
$$= 0.5000 - 0.4964 = 0.0036$$

DISTRIBUCION DE MUESTREO DE DIFERENCIAS Y SUMAS

8.11. Sea U_1 una variable que recorre los elementos de la población 3, 7, 8 y U_2 una variable que recorre los de la población 2, 4. Calcular (*a*) μ_{U1}, (*b*) μ_{U2}, (*c*) μ_{U1-U2}, (*d*) σ_{U1}, (*e*) σ_{U2} y (*f*) σ_{U1-U2}.

Solución

(*a*) μ_{U1} = media de la población $U_1 = \frac{1}{3}(3 + 7 + 8) = 6$.
(*b*) μ_{U2} = media de la población $U_2 = \frac{1}{2}(2 + 4) = 3$.
(*c*) La población consistente de las diferencias de cualquier elemento de U_1 y cualquiera de U_2, es

$$\begin{array}{lll} 3-2 & 7-2 & 8-2 \\ 3-4 & 7-4 & 8-4 \end{array} \quad \text{o} \quad \begin{array}{lll} 1 & 5 & 6 \\ -1 & 3 & 4 \end{array}$$

Luego $$\mu_{U1-U2} = \text{media de } (U_1 - U_2) = \frac{1 + 5 + 6 + (-1) + 3 + 4}{6} = 3$$

Eso ilustra el resultado general $\mu_{U1-U2} = \mu_{U1} - \mu_{U2}$, como se ve de las partes (*a*) y (*b*).

(*d*) $$\sigma_{U1}^2 = \text{varianza de la población } U_1 = \frac{(3-6)^2 + (7-6)^2 + (8-6)^2}{3} = \frac{14}{3}$$

es decir $\sigma_{U1} = \sqrt{\frac{14}{3}}$

(*e*) $$\sigma_{U2}^2 = \text{varianza de la población } U_2 = \frac{(2-3)^2 + (4-3)^2}{2} = 1 \quad \text{o sea.} \quad \sigma_{U2} = 1$$

(*f*) $$\sigma_{U1-U2}^2 = \text{varianza de la población } (U_1 - U_2)$$
$$= \frac{(1-3)^2 + (5-3)^2 + (6-3)^2 + (-1-3) + (3-3)^2 + (4-3)^2}{6} = \frac{17}{3}$$

es decir

$$\sigma_{U1-U2} = \sqrt{\frac{17}{3}}$$

Esto ilustra el resultado general, $\sigma_{U1-U2} = \sqrt{\sigma_{U1}^2 + \sigma_{U2}^2}$, para muestras independientes, como se ve de las partes (*d*) y (*e*).

8.12. Las lámparas de un fabricante A tienen vida media de 1400 horas (h) con desviación típica de 200 h, mientras que las de otro fabricante B tienen vida media de 1200 h con desviación típica de 100 h. Si se toma una muestra de 125 lámparas de cada clase, ¿cuál es la probabilidad de que las de A tengan una vida media que sea al menos (a) de 160 h y (b) 250 h, más que las de B?

Solución

Denotemos por $\bar{X}_A$ y $\bar{X}_B$ las vidas medias de las muestras A y B, respectivamente. Entonces

$$\mu_{\bar{X}_A - \bar{X}_B} = \mu_{\bar{X}_A} - \mu_{\bar{X}_B} = 1400 - 1200 = 200 \text{ h}$$

y

$$\sigma_{\bar{X}_A - \bar{X}_B} = \sqrt{\frac{\sigma_A^2}{N_A} + \frac{\sigma_B^2}{N_B}} = \sqrt{\frac{(100)^2}{125} + \frac{(200)^2}{125}} = 20 \text{ h}$$

La variable tipificada para la diferencia en medias es

$$z = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_{\bar{X}_A - \bar{X}_B})}{\sigma_{\bar{X}_A - \bar{X}_B}} = \frac{(\bar{X}_A - \bar{X}_B) - 200}{20}$$

y está casi normalmente distribuida.

(a) La diferencia 160 h en unidades estándar es $(160 - 200)/20 = -2$. Luego

$$\text{Probabilidad requerida} = (\text{área bajo la curva normal a la derecha de } z = -2)$$
$$= 0.5000 + 0.4772 = 0.9772$$

(b) La diferencia 250 h en unidades estándar es $(250 - 200)/20 = 2.50$. Por tanto

$$\text{Probabilidad requerida} = (\text{área bajo la curva normal a la derecha de } z = 2.50)$$
$$= 0.5000 - 0.4938 = 0.0062$$

8.13. Las bolas de rodamientos de cierto fabricante pesan 0.50 g de media, con desviación típica de 0.02 g. ¿Cuál es la probabilidad de que dos lotes de 1000 bolas cada uno difieran en peso en más de 2 g?

Solución

Sean $\bar{X}_1$ y $\bar{X}_2$ los pesos medios de las bolas de ambos lotes. Entonces

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = 0.50 - 0.50 = 0$$

y

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{(0.02)^2}{1000} + \frac{(0.02)^2}{1000}} = 0.000895$$

La variable tipificada para la diferencia en medias es

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{0.000895}$$

y es casi normalmente distribuida.

Una diferencia de 2 g en los lotes equivale a una diferencia de $2/1000 = 0.002$ g en las medias.

Esto puede suceder si $\bar{X}_1 - \bar{X}_2 \geqslant 0.002$ o $\bar{X}_1 - \bar{X}_2 \leqslant -0.002$; esto es

$$z \geqslant \frac{0.002 - 0}{0.000895} = 2.23 \quad \text{o} \quad z \leqslant \frac{-0.002 - 0}{0.00895} = -2.23$$

Entonces $\Pr\{z \geqslant 2.23 \text{ o } z \leqslant -2.23\} = \Pr\{z \geqslant 2.23\} + \Pr\{z \leqslant -2.23\} = 2(0.5000 - 0.4871) = 0.0258$.

8.14. *A* y *B* juegan a «cara o cruz» tirando 50 monedas. *A* ganará el juego si consigue 5 o más caras que *B*; de lo contrario, es *B* quien gana. Determinar las apuestas en contra de que *A* gane un juego.

Solución

Sean P_A y P_B las proporciones de caras logradas por *A* y *B*. Si suponemos que las monedas son buenas, como siempre, la probabilidad de cara es $p = \frac{1}{2}$. Así que

$$\mu_{P_A - P_B} = \mu_{P_A} - \mu_{P_B} = 0$$

y

$$\sigma_{P_A - P_B} = \sqrt{\sigma_{P_A}^2 + \sigma_{P_B}^2} = \sqrt{\frac{pq}{N_A} + \frac{pq}{N_B}} = \sqrt{\frac{2(\frac{1}{2})(\frac{1}{2})}{50}} = 0.10$$

La variable tipificada para la diferencia en proporciones es $z = (P_A - P_B - 0)/0.10$.

Sobre una base continua, 5 o más quiere decir 4.5 o más, de modo que la diferencia en proporciones debería ser $4.5/50 = 0.09$ o más; esto es, z mayor o igual que $(0.09 - 0)/0.10 = 0.9$ (o sea $z \geqslant 0.9$). La probabilidad de esto es el área bajo la curva normal a la derecha de $z = 0.9$, que es $(0.5000 - 0.3159) = 0.1841$.

Por tanto, las apuestas contra *A* están $(1 - 0.1841):0.1841 = 0.8159:0.1841$, o sea 4.43 a 1.

8.15. Dos distancias se han medido como 27.3 cm y 15.6 cm con desviación típica (error típico) de 0.16 cm y 0.08 cm, respectivamente. Hallar la media y la desviación típica de (*a*) la suma y (*b*) la diferencia, de esas distancias.

Solución

Si denotamos las distancias por D_1 y D_2, entonces:

(*a*)
$$\mu_{D1+D2} = \mu_{D1} + \mu_{D2} = 27.3 + 15.6 = 42.9 \text{ cm}$$
$$\sigma_{D1+D2} = \sqrt{\sigma_{D1}^2 + \sigma_{D2}^2} = \sqrt{(0.16)^2 + (0.08)^2} = 0.18 \text{ cm}$$

(*b*)
$$\mu_{D1-D2} = \mu_{D1} - \mu_{D2} = 27.3 - 15.6 = 11.7 \text{ cm}$$
$$\sigma_{D1-D2} = \sqrt{\sigma_{D1}^2 + \sigma_{D2}^2} = \sqrt{(0.16)^2 + (0.08)^2} = 0.18 \text{ cm}$$

8.16. Un cierto tipo de lámparas tiene una vida media de 1500 h y una desviación típica de 150 h. Se conectan tres de ellas de manera que en cuanto una falle se encenderá otra. Suponiendo que las vidas medias están normalmente distribuidas, ¿cuál es la probabilidad de que den luz durante (*a*) al menos 500 h y (*b*) a lo sumo 4200 h?

Solución

Supongamos que las vidas medias sean L_1, L_2 y L_3. Entonces

$$\mu_{L1+L2+L3} = \mu_{L1} + \mu_{L2} + \mu_{L3} = 1500 + 1500 + 1500 = 4500 \text{ h}$$

$$\sigma_{L1+L2+L3} = \sqrt{\sigma_{L1}^2 + \sigma_{L2}^2 + \sigma_{L3}^2} = \sqrt{3(150)^2} = 260 \text{ h}$$

(*a*) $$5000 \text{ h en unidades estándar} = \frac{5000 - 4500}{260} = 1.92$$

Probabilidad pedida = (área bajo la curva normal a la derecha de $z = 1.92$)

$$= 0.5000 - 0.4726 = 0.0274$$

(*b*) $$4200 \text{ h en unidades estándar} = \frac{4200 - 4500}{260} = -1.15$$

Probabilidad pedida = (área bajo la curva normal a la izquierda de $z = -1.15$)

$$= 0.5000 - 0.3749 = 0.1251$$

PROBLEMAS DIVERSOS

8.17. Con referencia al Problema 8.1, hallar (*a*) la media de la distribución de muestreo de varianzas y (*b*) la desviación típica de la distribución de muestreo de varianza (o sea, el error típico de varianzas).

Solución

(*a*) Las varianzas muestrales correspondientes a cada una de las 25 muestras del Problema 8.1 son

0	0.25	4.00	9.00	20.25
0.25	0	2.25	6.25	16.00
4.00	2.25	0	1.00	6.25
9.00	6.25	1.00	0	2.25
20.25	16.00	6.25	2.25	0

La media de la distribución de muestreo de varianzas es

$$\mu_{s^2} = \frac{\text{suma de todas las varianzas en la tabla anterior}}{25} = \frac{135}{25} = 5.40$$

Eso pone de relieve el hecho de que $\mu_{s^2} = (N-1)(\sigma^2)/N$, ya que para $N = 2$ y $\sigma^2 = 10.8$ [véase Prob. 8.1(*b*)], el lado derecho es $\frac{1}{2}(10.8) = 5.4$.

El resultado dice que es deseable definir una varianza corregida para las muestras como

$$\hat{s}^2 = \frac{N}{N-1} s^2$$

Se seguiría entonces que $\mu_{\hat{s}^2} = \sigma^2$. Debemos hacer constar que las varianzas de la población se definirían igual que antes y que sólo las varianzas muestrales serían corregidas.

(*b*) La varianza de la distribución de muestreo de varianzas $\sigma_{s^2}^2$ se obtiene restando la media 5.40 de cada uno de los 25 números en la tabla anterior, elevando al cuadrado, sumándolos y dividiendo el resultado por 25. Así pues $\sigma_{s^2}^2 = 575.75/25 = 23.03$, o sea $\sigma_{s^2} = 4.80$.

8.18. Rehacer el Problema 8.17 sin reposición.

Solución

(*a*) Hay 10 muestras cuyas varianzas vienen dadas por los números de encima (o debajo) de la diagonal de la tabla del Problema 8.17(*a*). Luego

$$\mu_{s^2} = \frac{0.25 + 4.00 + 9.00 + 20.25 + 2.25 + 6.25 + 16.00 + 1.00 + 6.25 + 2.25}{10} = 6.75$$

Esto es un caso especial del resultado general

$$\mu_{s^2} = \left(\frac{N_p}{N_p - 1}\right)\left(\frac{N - 1}{N}\right)\sigma^2$$

como se comprueba poniendo $N_p = 5$, $N = 2$ y $\sigma^2 = 10.8$ en el lado derecho para llegar a que $\mu_{s^2} = (\frac{5}{4})(\frac{1}{2})(10.8) = 6.75$.

(*b*) Restando 6.75 de cada uno de los 10 números sobre la diagonal de ceros de la tabla del Problema 8.17(*a*), elevando al cuadrado, sumando los resultados y dividiendo por 10, se ve que $\sigma_{s^2}^2 = 39.675$, o sea $\sigma_{s^2} = 6.30$.

8.19. La desviación típica de los pesos de una población muy numerosa de estudiantes es 10.0 lb. Se toman muestras de 200 estudiantes de dicha población y se calculan sus desviaciones típicas en altura. Hallar (*a*) la media y (*b*) la desviación típica de la distribución de muestreo de desviación típicas.

Solución

Podemos considerar que el muestreo es o bien de una población infinita o de una finita con reposición. De la Tabla 8.1 se tiene:

(*a*) La media de la distribución de muestreo de desviación típicas es $\mu_s = \sigma = 10.0$ lb.

(*b*) La desviación típica de la distribución de muestreo de desviaciones típicas es $\sigma_s = \sigma/\sqrt{2N} = 10/\sqrt{400} = 0.50$ lb.

8.20. ¿Qué porcentaje de las muestras del Problema 8.19 tendrían desviación típicas (*a*) mayores que 11.0 lb y (*b*) menores que 8.8 lb?

Solución

La distribución de muestreo de desviación típicas está casi normalmente distribuida con media 10.0 lb y desviación típica 0.50 lb.

(*a*) 11.10 lb en unidades estándar es $(11.0 - 10.0)/0.50 = 2.0$. El área bajo la curva normal a la derecha de $z = 2.0$ es $(0.5 - 0.4772) = 0.0228$; por tanto el porcentaje pedido es 2.3%.

(*b*) 8.8 lb en unidades estándar es $(8.8 - 10.0)/0.50 = -2.4$. El área bajo la curva normal a la izquierda de $z = -2.4$ es $(0.5 - 0.4918) = 0.0082$; luego el requerido porcentaje es 0.8%.

PROBLEMAS SUPLEMENTARIOS

DISTRIBUCION DE MUESTREO DE MEDIAS

8.21. Una población consiste en los números 3, 7, 11 y 15. Consideremos todas las posibles muestras de tamaño 2 que se pueden tomar de esa población con reposición. Hallar (*a*) la media de la población, (*b*) la desviación típica de la población, (*c*) la media de la distribución de muestreo de medias y (*d*) la desviación típica de la distribución de muestreo de medias. Verificar las partes (*c*) y (*d*) directamente de (*a*) y (*b*) usando fórmulas adecuadas.

8.22. Resolver el Problema 8.21 si el muestreo se hace con reposición.

8.23. Las masas de 1500 bolas de rodamientos están normalmente distribuidas, con media 22.40 g y desviación típica 0.048 g. Si se toman 300 muestras aleatorias de tamaño 36 en esa población, determinar la media esperada y la desviación típica esperada de la distribución de muestreo de medias, si el muestreo se hace (*a*) con, y (*b*) sin reposición.

8.24. Resolver el Problema 8.23 si la población consiste en 72 bolas.

8.25. ¿Cuántas de las muestras aleatorias del Problema 8.23 tendrían sus medias (*a*) entre 22.39 y 22.41 g, (*b*) mayor que 22.42 g, (*c*) menor que 22.37 g, y (*d*) menor que 22.38 g y más de 22.41 g?

8.26. Las lámparas que fabrica cierta empresa tienen una vida media de 800 h y una desviación típica de 60 h. Hallar la probabilidad de que una muestra aleatoria de 16 lámparas tenga una vida media (*a*) entre 790 y 810 h, (*b*) menor que 785 h, (*c*) más de 820 h y (*d*) entre 770 y 830 h.

8.27. Repetir el Problema 8.26 si se toma una muestra de 64 lámparas. Explicar la diferencia.

8.28. Los paquetes recibidos en un almacén tienen un peso medio de 300 lb y una desviación típica de 50 lb. ¿Cuál es la probabilidad de que 25 de esos paquetes, elegidos al azar y metidos en un montacargas, excedan el límite de carga de éste, que es de 8200 lb?

NUMEROS ALEATORIOS

8.29. Rehacer el Problema 8.6 usando un conjunto diferente de números aleatorios y seleccionando (*a*) 15, (*b*) 30, (*c*) 45 y (*d*) 60 muestras de tamaño 4 con reposición. Comparar en cada caso con los resultados teóricos.

8.30. Repetir el Problema 8.29 seleccionando muestras de tamaño (*a*) 2 y (*b*) 8 con reposición, en lugar de tamaño 4 con reposición.

8.31. Resolver el Problema 8.6 sin reposición. Comparar con los resultados teóricos.

8.32. (*a*) Mostrar cómo seleccionar 30 muestras de tamaño 2 de la distribución del Problema 3.61.

(*b*) Calcular la media y la desviación típica de la distribución de muestreo resultante de medias, y comparar con los resultados teóricos.

8.33. Resolver el Problema 8.32 usando muestras de tamaño 4.

DISTRIBUCION DE MUESTREO DE PROPORCIONES

8.34. Hallar la probabilidad de que en los 200 próximos nacimientos (*a*) menos del 40% sean niños, (*b*) entre 43% y 57% sean niñas y (*c*) más del 54% sean niños. Suponemos probabilidades de nacimiento iguales para niño y niña.

8.35. De 1000 muestras de 200 niños cada una, ¿en cuántas cabe esperar encontrar (*a*) menos del 40% de niños, (*b*) entre 40% y 60% son niñas y (*c*) el 53% o más son niñas?

8.36. Rehacer el Problema 8.34 si se consideran 100 niños en vez de 200, y explicar las diferencias en los resultados.

8.37. En una urna hay 80 fichas, de las que el 60% son rojas y el 40% blancas. De entre 50 muestras de 20 fichas cada una seleccionadas al azar, ¿cuántas es de esperar que tengan (*a*) tantas rojas como blancas, (*b*) 12 rojas y 8 blancas, (*c*) 8 rojas y 12 blancas y (*d*) 10 o más blancas?

8.38. Diseñar un experimento que ilustre los resultados del Problema 8.37. En vez de fichas rojas y blancas, puede usar papeletas en las que se han escrito R y B en las proporciones adecuadas. ¿Qué errores podrían introducirse al usar dos conjuntos diferentes de piezas?

8.39. Un fabricante envía 1000 lotes de 100 bombillas cada uno. Si el 5% de las bombillas son defectuosas, ¿en cuántos de los lotes se puede esperar que haya (*a*) menos de 90 bombillas buenas y (*b*) 98 o más buenas?

DISTRIBUCION DE MUESTREO DE DIFERENCIAS Y SUMAS

8.40. A y B producen dos tipos de cables que soportan cargas máximas medias de 4000 lb y 4500 lb, con desviación típica respectivas de 300 lb y 200 lb. Si se analizan 100 cables A y 50 cables B, ¿cuál es la probabilidad de que la carga máxima que soporta B sea (*a*) al menos 600 lb mayor que la de A y (*b*) al menos 450 lb mayor que la de A.

8.41. ¿Cuáles son las probabilidades en el Problema 8.40 si se analizan 100 cables de cada tipo? Razonar las diferencias.

8.42. La puntuación media en una prueba de aptitud es de 72 puntos con una desviación típica de 8 puntos. ¿Cuál es la probabilidad de que dos grupos de 28 y 36 estudiantes respectivamente, difieran en su puntuación media (*a*) 3 o más puntos, (*b*) 6 o más puntos y (*c*) entre 2 y 5 puntos?

8.43. Una urna contiene 60 piezas rojas y 40 blancas. Se sacan dos conjuntos de 30 con reposición y se anotan sus colores. ¿Cuál es la probabilidad de que los dos conjuntos difieran en 8 o más piezas rojas?

8.44. Resolver el Problema 8.43 sin reposición.

8.45. Un candidato recibe en unas elecciones el 65% de los votos. Hallar la probabilidad de que dos muestras aleatorias de 200 votantes indicasen una diferencia de más del 10% de votos a su favor.

8.46. Si U_1 y U_2 son los conjuntos de números del Problema 8,11, comprobar que (*a*) $\mu_{U1+U2} = \mu_{U1} + \mu_{U2}$ y (*b*) $\sigma_{U1+U2} = \sqrt{\sigma_{U1}^2 + \sigma_{U2}^2}$.

8.47. Se han medido tres masas como 20.48, 35.97, y 62.34 g con desviaciones típicas de 0.21, 0.46 y 0.54 g, respectivamente. Hallar (*a*) la media y (*b*) la desviación típica de la suma de las masas.

8.48. El voltaje medio de unas baterías es 15.0 voltios (V) y la desviación típica es 0.2 V. ¿Cuál es la probabilidad de que 4 de ellas, conectadas en serie, tengan un voltaje combinado de 60.8 V o más?

PROBLEMAS DIVERSOS

8.49. Una población de 7 números tiene una media de 40 y una desviación típica de 3. Si se toman muestras de tamaño 5 de esa población, y se calcula la varianza de cada muestra, hallar la media de la distribución de muestreo de varianzas si el muestreo se hace (*a*) con y (*b*) sin reposición.

8.50. Los tubos fabricados en cierta empresa tienen una vida media de 900 h y una desviación típica de 80 h. Se envían 1000 lotes de 100 tubos cada uno. ¿En cuántos de esos lotes se puede esperar que (*a*) la vida media exceda de 900 h y (*b*) la desviación típica de las vidas medias exceda de 95 h? ¿Qué hipótesis hay que hacer?

8.51. Si la mediana de las vidas medias del Problema 8.50 es 900 h, ¿en cuántos lotes cabe esperar que la mediana de las vidas medias sea mayor que 910 h? Comparar la respuesta

con el Problema 8.50(*a*) y explicar los resultados.

8.52. En un examen las notas estuvieron normalmente distribuidas con media 72 y desviación típica 8.

(*a*) Hallar la nota mínima del 20% de estudiantes mejores.

(*b*) Hallar la probabilidad de que en una muestra aleatoria de 100 estudiantes, la nota más baja sea inferior a 76.

CAPITULO 9

Teoría de la estimación estadística

ESTIMACION DE PARAMETROS

En el último capítulo vimos cómo se puede emplear la teoría del muestreo para recabar información acerca de muestras aleatorias tomadas de una población conocida. Desde un punto de vista práctico, no obstante, suele resultar más importante ser capaz de inferir información sobre la población a partir de muestras suyas. Con tal situación trata la *inferencia estadística*, que usa los principios de la teoría del muestreo.

Un problema importante de la inferencia estadística es la estimación de *parámetros de la población*, o brevemente *parámetros* (tales como la media o la varianza de la población), de los correspondientes *estadísticos muestrales*, o simplemente *estadísticos* (tales como la media y la varianza de la muestra). Consideramos este problema en el presente capítulo.

ESTIMACIONES SIN SESGO

Si la media de las distribuciones de muestreo de un estadístico es igual que la del correspondiente parámetro de la población, el estadístico se llama un *estimador sin sesgo* del parámetro; si no, se llama un estimador *sesgado*. Los correspondientes valores de tales estadísticos se llaman *estimaciones sin sesgo y sesgadas*, respectivamente.

EJEMPLO 1. La media de las distribuciones de muestreo de medias $\mu_{\bar{X}}$ e μ, la media de la población. Por tanto, la media muestral $\bar{X}$ es una estimación sin sesgo de la media de la población μ.

EJEMPLO 2. La media de las distribuciones de muestreo de varianza es

$$\mu_{s^2} = \frac{N-1}{N}\sigma^2$$

donde σ^2 es la varianza de la población y N es el tamaño de la muestra (véase Tabla 8.1). Así pues, la varianza de la muestra s^2 es una estimación sesgada de la varianza de la población σ^2. Usando la varianza modificada

$$\hat{s}^2 = \frac{N}{N-1}s^2$$

encontramos $\mu_{\hat{s}^2} = \sigma^2$, de manera que $\hat{s}^2$ es una estimación sin sesgo de σ^2. Sin embargo, $\hat{s}$ es una estimación sesgada de σ.

En términos de esperanzas (Cap. 6) podríamos decir que un estadístico es insesgado si su esperanza es igual al correspondiente parámetro de población. Así, $\bar{X}$ y $\hat{s}^2$ son insesgados porque $E\{\bar{X}\} = \mu$ y $E\{\hat{s}^2\} = \sigma^2$.

ESTIMACION EFICIENTE

Si las distribuciones de muestreo de dos estadísticos tienen la misma media (o esperanza), el de menor varianza se llama un *estimador eficiente* de la media, mientras que el otro se llama un *estimador ineficiente*. Los valores correspondientes de los estadísticos se llaman *estimación eficiente* e *estimación ineficiente*, respectivamente.

Si consideramos todos los posibles estadísticos cuyas distribuciones de muestreo tienen la misma media, aquel de varianza mínima se llama a veces el *estimador de máxima eficiencia*, o sea, *el mejor estimador*.

EJEMPLO 3. Las distribuciones de muestreo de media y mediana tienen ambas la misma media, a saber, la media de la población. Sin embargo, la varianza de la distribución de muestreo de medias es menor que la varianza de la distribución de muestreo de medianas (véase Tabla. 8.1). Por tanto, la media muestral da una estimación eficiente de la media de la población, mientras la mediana de la muestra da una estimación ineficiente de ella.

De todos los estadísticos que estiman la media de la población, la media muestral proporciona la mejor (la más eficiente) estimación.

En la práctica, estimaciones ineficientes se usan con frecuencia a causa de la relativa sencillez con que se obtienen algunas de ellas.

ESTIMACIONES DE PUNTO Y ESTIMACIONES DE INTERVALO; SU FIABILIDAD

Una estimación de un parámetro de la población dada por un solo número se llama una *estimación de punto* del parámetro. Una estimación de un parámetro de la población dada por dos números, entre los cuales se puede considerar encajado al parámetro, se llama una *estimación de intervalo* del parámetro.

Las estimaciones de intervalo indican la precisión de una estimación y son por tanto preferibles a las estimaciones de punto.

EJEMPLO 4. Si decimos que una distancia se ha medido como 5.28 metros (m), estamos dando una estimación de punto. Por otra parte, si decimos que la distancia es 5.28 ± 0.03 m (o sea, que está entre 5.25 y 5.31 m), estamos dando una estimación de intervalo.

El margen de error (o la precisión) de una estimación nos informa de su *fiabilidad*.

ESTIMACIONES DE INTERVALO DE CONFIANZA PARA PARAMETROS DE POBLACION

Sean μ_S y σ_S la media y la desviación típica (error típico) de la distribución de muestreo de un estadístico S. Entonces, si la distribución de muestreo de S es aproximadamente normal (que como

hemos visto es cierto para muchos estadísticos si el tamaño de la muestra $N \geqslant 30$), podemos esperar hallar un estadístico muestral real S que esté en los intervalos $\mu_S - \sigma_S$ a $\mu_S + \sigma_S$, $\mu_S - 2\sigma_S$ a $\mu_S + 2\sigma_S$, o $\mu_S - 3\sigma_S$ a $\mu_S + 3\sigma_S$ alrededor del 68.27%, 95.45% y 99.73% del tiempo, respectivamente.

Equivalentemente, podemos esperar hallar (o sea, podemos estar *confiados* en encontrar) μ_S en los intervalos $S - \sigma_S$ a $S + \sigma_S$, $S - 2\sigma_S$ a $S + 2\sigma_S$, o $S - 3\sigma_S$ a $S + 3\sigma_S$ alrededor del 68.27%, 95.45% y 99.73% del tiempo, respectivamente. Por esa razón, llamamos a esos respectivos intervalos los *intervalos de confianza* 68.27%, 95.45% y 99.73% para estimar μ_S. Los números extremos de estos intervalos ($S \pm \sigma_S$, $S \pm 2\sigma_S$ y $S \pm 3\sigma_S$) se llaman entonces los *límites de confianza* 68.27%, 95.45% y 99.73% o *límites fiduciales*.

Análogamente, $S \pm 1.96\sigma_S$ y $S \pm 2.58\sigma_S$ son los límites de confianza 95% y 99% (o sea, 0.95 y 0.99) para S. El porcentaje de confianza se suele llamar *nivel de confianza*. Los números 1.96, 2.58, etc. en los límites de confianza se llaman *coeficientes de confianza* o *valores críticos*, y se denotan por z_c. De los niveles de confianza podemos deducir los coeficientes de confianza y viceversa.

La Tabla 9.1 muestra los valores de z_c correspondientes a varios niveles de confianza usados en la práctica. Para niveles de confianza que no aparecen en la tabla, los valores de z_c se pueden encontrar gracias a las tablas de áreas bajo la curva normal (Apéndice II).

Tabla 9.1

Nivel de confianza	99.73%	99%	98%	96%	95.45%	95%	90%	80%	68.27%	50%
z_c	3.00	2.58	2.33	2.05	2.00	1.96	1.645	1.28	1.00	0.6745

Intervalos de confianza para las medias

Si el estadístico S es la media $\bar{X}$ de la muestra, entonces los límites de confianza 95% y 99% para estimar la media μ de la población vienen dados por $\bar{X} \pm 1.96\sigma_{\bar{X}}$ y $\bar{X} \pm 2.58\sigma_{\bar{X}}$, respectivamente. Más en general, los límites de confianza para estimar la media de la población μ vienen dados por $\bar{X} \pm z_c\sigma_{\bar{X}}$, donde z_c (que depende del nivel particular de confianza deseado) se puede leer en la Tabla 9.1. Usando los valores de $\sigma_{\bar{X}}$ obtenidos en el Capítulo 8, vemos que los límites de confianza para la media de la población están dados por

$$\bar{X} \pm z_c \frac{\sigma}{\sqrt{N}} \tag{1}$$

si el muestreo es de una población infinita o de una finita con reposición, y vienen dados por

$$\bar{X} \pm z_c \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} \tag{2}$$

si el muestreo es sin reposición de una población finita de tamaño N_p.

Generalmente, la desviación típica σ de la población no es conocida; así pues, para obtener los anteriores límites de confianza usamos la estimación muestral $\hat{s}$ o s. Esto se verá que es satisfactorio

para $N \geq 30$. Para $N < 30$, la aproximación es pobre y debe emplearse la teoría de pequeñas muestras (Cap. 11).

Intervalos de confianza para proporciones

Si el estadístico S es la proporción de «éxitos» en una muestra de tamaño N sacada de una población binomial en la que p es la proporción de éxitos (o sea, la probabilidad de éxito), entonces los límites de confianza para p vienen dados por $P \pm z_c\sigma_P$, donde P es la proporción de éxitos en la muestra de tamaño N. Usando los valores de σ_P obtenidos en el Capítulo 8, vemos que los límites de confianza para la proporción en la población vienen dados por

$$P \pm z_c\sqrt{\frac{pq}{N}} = P \pm z_c\sqrt{\frac{p(1-p)}{N}} \tag{3}$$

si el muestreo es de una población infinita o finita con reposición, y por

$$P \pm z_c\sqrt{\frac{pq}{N}}\sqrt{\frac{N_p - N}{N_p - 1}} \tag{4}$$

si el muestreo es de una población finita de tamaño N_p y sin reposición.

Para calcular estos límites de confianza, podemos usar la estimación muestral P para p, que generalmente resultará satisfactoria si $N \geq 30$. Un método más exacto para obtener los límites de confianza se presenta en el Problema 9.12.

Intervalos de confianza para diferencias y sumas

Si S_1 y S_2 son dos estadísticos muestrales con distribuciones de muestreo aproximadamente normales, los límites de confianza para la diferencia de los parámetros de población correspondientes a S_1 y S_2 vienen dados por

$$S_1 - S_2 \pm z_c\sigma_{S1-S2} = S_1 - S_2 \pm z_c\sqrt{\sigma_{S1}^2 + \sigma_{S2}^2} \tag{5}$$

mientras que los límites de confianza para la suma de los parámetros de población vienen dados por

$$S_1 + S_2 \pm z_c\sigma_{S1+S2} = S_1 + S_2 \pm z_c\sqrt{\sigma_{S1}^2 + \sigma_{S2}^2} \tag{6}$$

supuesto que las muestras sean independientes (véase Cap. 8).

Por ejemplo, los límites de confianza para la diferencia de dos medias poblacionales, en el caso de poblaciones infinitas, se calculan como

$$\bar{X}_1 - \bar{X}_2 \pm z_c\sigma_{\bar{X}1-\bar{X}2} = \bar{X}_1 - \bar{X}_2 \pm z_c\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} \tag{7}$$

donde $\bar{X}_1$, σ_1, N_1 y $\bar{X}_2$, σ_2, N_2 son las respectivas medias, desviaciones típicas y tamaños de las dos muestras sacadas de las poblaciones.

De forma similar, los límites de confianza para la diferencia de dos proporciones poblacionales con poblaciones infinitas, están dados por

$$P_1 - P_2 \pm z_c\sigma_{P1-P2} = P_1 - P_2 \pm z_c\sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}} \tag{8}$$

donde P_1 y P_2 son las dos proporciones muestrales, N_1 y N_2 los tamaños de las dos muestras, y p_1 y p_2 las proporciones en las dos poblaciones (estimadas por P_1 y P_2).

Intervalos de confianza para desviaciones típicas

Los límites de confianza para la desviación típica σ de una población normalmente distribuida, estimados con una muestra con desviación típica s, vienen dados por

$$s \pm z_c\sigma_s = s \pm z_c\frac{\sigma}{\sqrt{2N}} \tag{9}$$

usando la Tabla 8.1. Al calcular estos límites de confianza, usamos s o $\hat{s}$ para estimar σ.

ERROR PROBABLE

Los límites de confianza 50% de los parámetros de población correspondientes a un estadístico S vienen dados por $S \pm 0.6745\sigma_S$. La cantidad $0.6745\sigma_S$ se conoce como el *error probable de la estimación*.

PROBLEMAS RESUELTOS

ESTIMACIONES SIN SESGO Y EFICIENTES

9.1. Dar un ejemplo de estimadores (o estimaciones) que sean (*a*) sin sesgo y eficiente, (*b*) sin sesgo e ineficiente y (*c*) sesgado e ineficiente.

Solución

(*a*) La media muestral $\bar{X}$ y la varianza muestral modificada

$$\hat{s}^2 = \frac{N}{N-1}s^2$$

son dos ejemplos.

(*b*) La mediana muestral y el estadístico muestral $\frac{1}{2}(Q_1 + Q_3)$, donde Q_1 y Q_3 son los cuartiles muestrales inferior y superior, son dos ejemplos. Ambos son estimaciones sin sesgo de la media

de la población, pues la media de sus distribuciones de muestreo es la media de la población.

(*c*) La desviación típica muestral s, la desviación típica modificada $\hat{s}$, la desviación media y el rango semi-intercuartil son cuatro ejemplos.

9.2. En una muestra de cinco medidas, un científico anotó 6.33, 6.37, 6.36, 6.32 y 6.37 centímetros (cm). Determinar estimaciones insesgadas y eficientes de (*a*) la verdadera media y (*b*) la varianza.

Solución

(*a*) La estimación sin sesgo y eficiente de la media verdadera (o sea, la de la población) es

$$\bar{X} = \frac{\sum X}{N} = \frac{6.33 + 6.37 + 6.36 + 6.32 + 6.37}{5} = 6.35 \text{ cm}$$

(*b*) La estimación sin sesgo y eficiente de la media verdadera (o sea, la de la población) es

$$\hat{s}^2 = \frac{N}{N-1} s^2 = \frac{\sum (X - \bar{X})^2}{N-1}$$

$$= \frac{(6.33 - 6.35)^2 + (6.37 - 6.35)^2 + (6.36 - 6.35)^2 + (6.32 + 6.35)^2 + (6.37 - 6.35)^2}{5 - 1} =$$

$$= 0.00055 \text{ cm}^2$$

Nótese que aunque $\hat{s} = \sqrt{0.00055} = 0.023$ cm es una estimación de la verdadera desviación típica, esta estimación no es ni eficiente ni insesgada.

9.3. Supongamos que las alturas de 100 estudiantes varones de la Universidad XYZ representan una muestra aleatoria de las de los 1546 estudiantes de esa Universidad. Determinar estimaciones sin sesgo y eficientes de (*a*) la media verdadera y (*b*) la varianza verdadera.

Solución

(*a*) Por el Problema 3.22, la estimación sin sesgo y eficiente de la verdadera media es $\bar{X} = 67.45$ in.

(*b*) Del Problema 4.17 se sigue que la estimación sin sesgo y eficiente de la verdadera varianza es

$$\hat{s}^2 = \frac{N}{N-1} s^2 = \frac{100}{99} (8.5275) = 8.6136$$

Así pues, $\hat{s} = \sqrt{8.6136} = 2.93$ in. Notemos que ya que N es grande, no hay diferencia casi entre s^2 y $\hat{s}^2$, o sea entre s y $\hat{s}$.

No hemos usado la corrección de Sheppard para el agrupamiento. Para tener esto en cuenta, usaríamos $s = 2.79$ in (véase Prob. 4.21).

9.4. Dar una estimación sin sesgo e ineficiente para la verdadera media del diámetro de la esfera del Problema 9.2.

Solución

La mediana es un ejemplo. Para las cinco medidas, ordenadas por magnitud, la mediana es 6.36 cm.

donde $\bar{X}_1$, σ_1, N_1 y $\bar{X}_2$, σ_2, N_2 son las respectivas medias, desviaciones típicas y tamaños de las dos muestras sacadas de las poblaciones.

De forma similar, los límites de confianza para la diferencia de dos proporciones poblacionales, con poblaciones infinitas, están dados por

$$P_1 - P_2 \pm z_c\sigma_{P_1-P_2} = P_1 - P_2 \pm z_c\sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}} \tag{8}$$

donde P_1 y P_2 son las dos proporciones muestrales, N_1 y N_2 los tamaños de las dos muestras, y p_1 y p_2 las proporciones en las dos poblaciones (estimadas por P_1 y P_2).

Intervalos de confianza para desviaciones típicas

Los límites de confianza para la desviación típica σ de una población normalmente distribuida, estimados con una muestra con desviación típica s, vienen dados por

$$s \pm z_c\sigma_s = s \pm z_c\frac{\sigma}{\sqrt{2N}} \tag{9}$$

usando la Tabla 8.1. Al calcular estos límites de confianza, usamos s o $\hat{s}$ para estimar σ.

ERROR PROBABLE

Los límites de confianza 50% de los parámetros de población correspondientes a un estadístico S vienen dados por $S \pm 0.6745\sigma_S$. La cantidad $0.6745\sigma_S$ se conoce como el *error probable de la estimación*.

PROBLEMAS RESUELTOS

ESTIMACIONES SIN SESGO Y EFICIENTES

9.1. Dar un ejemplo de estimadores (o estimaciones) que sean (*a*) sin sesgo y eficiente, (*b*) sin sesgo e ineficiente y (*c*) sesgado e ineficiente.

Solución

(*a*) La media muestral $\bar{X}$ y la varianza muestral modificada

$$\hat{s}^2 = \frac{N}{N-1}s^2$$

son dos ejemplos.

(*b*) La mediana muestral y el estadístico muestral $\frac{1}{2}(Q_1 + Q_3)$, donde Q_1 y Q_3 son los cuartiles muestrales inferior y superior, son dos ejemplos. Ambos son estimaciones sin sesgo de la media

de la población, pues la media de sus distribuciones de muestreo es la media de la población.

(*c*) La desviación típica muestral s, la desviación típica modificada $\hat{s}$, la desviación media y el rango semi-intercuartil son cuatro ejemplos.

9.2. En una muestra de cinco medidas, un científico anotó 6.33, 6.37, 6.36, 6.32 y 6.37 centímetros (cm). Determinar estimaciones insesgadas y eficientes de (*a*) la verdadera media y (*b*) la varianza.

Solución

(*a*) La estimación sin sesgo y eficiente de la media verdadera (o sea, la de la población) es

$$\bar{X} = \frac{\sum X}{N} = \frac{6.33 + 6.37 + 6.36 + 6.32 + 6.37}{5} = 6.35 \text{ cm}$$

(*b*) La estimación sin sesgo y eficiente de la media verdadera (o sea, la de la población) es

$$\hat{s}^2 = \frac{N}{N-1} s^2 = \frac{\sum (X - \bar{X})^2}{N-1}$$

$$= \frac{(6.33 - 6.35)^2 + (6.37 - 6.35)^2 + (6.36 - 6.35)^2 + (6.32 + 6.35)^2 + (6.37 - 6.35)^2}{5 - 1} =$$

$$= 0.00055 \text{ cm}^2$$

Nótese que aunque $\hat{s} = \sqrt{0.00055} = 0.023$ cm es una estimación de la verdadera desviación típica, esta estimación no es ni eficiente ni insesgada.

9.3. Supongamos que las alturas de 100 estudiantes varones de la Universidad XYZ representan una muestra aleatoria de las de los 1546 estudiantes de esa Universidad. Determinar estimaciones sin sesgo y eficientes de (*a*) la media verdadera y (*b*) la varianza verdadera.

Solución

(*a*) Por el Problema 3.22, la estimación sin sesgo y eficiente de la verdadera media es $\bar{X} = 67.45$ in.

(*b*) Del Problema 4.17 se sigue que la estimación sin sesgo y eficiente de la verdadera varianza es

$$\hat{s}^2 = \frac{N}{N-1} s^2 = \frac{100}{99} (8.5275) = 8.6136$$

Así pues, $\hat{s} = \sqrt{8.6136} = 2.93$ in. Notemos que ya que N es grande, no hay diferencia casi entre s^2 y $\hat{s}^2$, o sea entre s y $\hat{s}$.

No hemos usado la corrección de Sheppard para el agrupamiento. Para tener esto en cuenta, usaríamos $s = 2.79$ in (véase Prob. 4.21).

9.4. Dar una estimación sin sesgo e ineficiente para la verdadera media del diámetro de la esfera del Problema 9.2.

Solución

La mediana es un ejemplo. Para las cinco medidas, ordenadas por magnitud, la mediana es 6.36 cm.

INTERVALOS DE CONFIANZA PARA MEDIAS

9.5. Hallar los intervalos de confianza (*a*) 95% y (*b*) 99% para estimar la altura media de los estudiantes del Problema 9.3.

Solución

(*a*) Los límites de confianza 95% son $\bar{X} \pm 1.96\sigma/\sqrt{N}$. Usando $\bar{X} = 67.45$ in y $\hat{s} = 2.93$ in como estimación de σ (véase Prob. 9.3), los límites de confianza son $67.45 \pm 1.96(2.93/\sqrt{100})$, o 67.45 ± 0.57 in. Luego el intervalo de confianza 95% para la media de la población μ es 66.88 a 68.02 in, que denotamos por $66.88 < \mu < 68.02$.

Podemos decir, por tanto, que la probabilidad de que la altura media de la población esté entre 66.88 y 68.02 in es del 95%, o sea 0.95. En símbolos escribimos $\Pr\{66.88 < \mu < 68.02\} = 0.95$. Esto equivale a decir que tenemos 95% de *confianza* de que la media de la población (o media verdadera) esté entre 66.88 y 68.02 in.

(*b*) Los límites de confianza 99% son $\bar{X} \pm 2.58\sigma/\sqrt{N} = \bar{X} \pm 2.58\hat{s}/\sqrt{N} = 67.45 \pm 2.58(2.93/\sqrt{100}) =$ $= 67.45 \pm 0.76$ in. Luego el intervalo de confianza 99% para la media de la población μ es 66.69 a 68.21 in, que se denota por $66.69 < \mu < 68.21$.

Al hallar los anteriores intervalos de confianza, hemos supuesto que la población era infinita o tan grande que podíamos considerarla como con reposición. Para poblaciones finitas y muestreo sin reposición, debe usarse

$$\frac{\sigma}{\sqrt{N}}\sqrt{\frac{N_p - N}{N_p - 1}} \quad \text{en lugar de} \quad \frac{\sigma}{\sqrt{N}}$$

No obstante, podemos considerar el factor

$$\sqrt{\frac{N_p - N}{N_p - 1}} = \sqrt{\frac{1546 - 100}{1546 - 1}} = 0.967$$

como esencialmente 1.0, y por tanto no es necesario usarlo. Si se usa, los límites de confianza anteriores se convierten en 67.45 ± 0.56 in y 67.45 ± 0.73 in, respectivamente.

9.6. Las medidas de los diámetros de una muestra aleatoria de 200 bolas de rodamientos producidas por una máquina en una semana, dieron una media de 0.824 cm y una desviación típica de 0.042 cm. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para el diámetro medio de todas las bolas.

Solución

(*a*) Los límites de confianza 95% son

$$\bar{X} \pm \frac{1.96\sigma}{\sqrt{N}} = \bar{X} \pm \frac{1.96\hat{s}}{\sqrt{N}} = 0.824 \pm 1.96\,\frac{0.042}{\sqrt{200}} = 0.824 + 0.0058 \text{ cm} \quad \text{o sea} \quad 0.824 \pm 0.006 \text{ cm}$$

(*b*) Los límites de confianza 99% son

$$\bar{X} \pm \frac{2.58\sigma}{\sqrt{N}} = \bar{X} \pm \frac{2.58\hat{s}}{\sqrt{N}} = 0.824 \pm 2.58\,\frac{0.042}{\sqrt{200}} = 0.824 \pm 0.0077 \text{ cm} \quad \text{o sea} \quad 0.824 \pm 0.008 \text{ cm}$$

Nótese que hemos supuesto la desviación típica dada como la desviación típica *modificada* $\hat{s}$. Si la desviación típica hubiera sido s, hubiéramos usado $\hat{s} = \sqrt{N/(N-1)}s = \sqrt{200/199}s$, que puede ser

tomada como s a efectos prácticos. En general, para $N \geq 30$ podemos suponer que s y $\hat{s}$ son prácticamente iguales.

9.7. Hallar los límites de confianza (*a*) 98%, (*b*) 90% y (*c*) 99.73% para el diámetro medio de las bolas del Problema 9.6.

Solución

(*a*) Sea $z = z_c$ tal que el área bajo la curva normal a su derecha es 1%. Entonces, por simetría, el área a la izquierda de $z = -z_c$ es también 1%, así que el área sombreada es el 98% del total; véase Figura 9.1(*a*). Como el área total bajo la curva es 1, el área desde $z = 0$ hasta $z = z_c$ es 0.49; por tanto, $z_c = 2.33$. Luego los límites de confianza 98% son $\bar{X} \pm 2.33\sigma/\sqrt{N} = 0.824 \pm \pm 2.33(0.042/\sqrt{200}) = 0.824 \pm 0.0069$ cm.

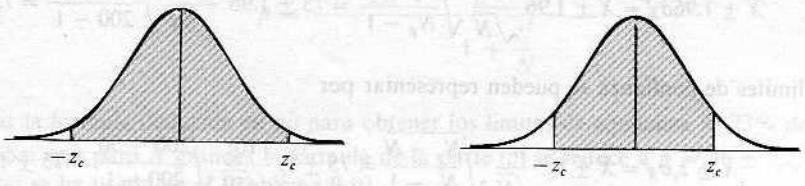


Figura 9.1.

(*b*) Deseamos un z_c tal que el área desde $z = 0$ hasta $z = z_c$ es 0.45, como muestra la Figura 9.1(*b*); entonces $z_c = 1.645$. Así pues, los límites de confianza 90% son $\bar{X} \pm 1.645\sigma/\sqrt{N} = 0.824 \pm \pm 1.645(0.042/\sqrt{200}) = 0.824 \pm 0.0049$ cm.

(*c*) Los límites de confianza del 99.73% son

$$\bar{X} \pm 3\sigma/\sqrt{N} = 0.824 \pm 3(0.042/\sqrt{200}) = 0.824 \pm 0.0089 \text{ cm}$$

9.8. Al medir el tiempo de reacción, un psicólogo estima que la desviación típica es 0.05 segundos. ¿De qué tamaño ha de tomarse una muestra de medidas para tener una confianza del (*a*) 95% y (*b*) 99% de que el error de la estimación no supera 0.01 segundos?

Solución

(*a*) Los límites de confianza 95% son $\bar{X} \pm 1.96\sigma/\sqrt{N}$, siendo el error de la estimación $1.96\sigma/\sqrt{N}$. Tomando $\sigma = s = 0.05$ seg, vemos que este error será igual a 0.01 seg si $(1.96)(0.05)/\sqrt{N} = 0.01$; esto es, $\sqrt{N} = (1.96)(0.05)/0.01 = 9.8$, o sea $N = 96.04$. Luego podemos estar confidentes al 95% de que el error de la estimación será menor que 0.01 seg si N es 97 o mayor.

Otro método

$$\frac{(1.96)(0.05)}{\sqrt{N}} \leq 0.01 \quad \text{si} \quad \frac{\sqrt{N}}{(1.96)(0.05)} \geq \frac{1}{0.01} \quad \text{o sea} \quad \sqrt{N} \geq \frac{(1.96)(0.05)}{0.01} = 9.8$$

Entonces $N \geq 96.04$, o sea $N \geq 97$.

(*b*) Los límites de confianza 99% son $\bar{X} \pm 2.58\sigma/\sqrt{N}$. Entonces $(2.58)(0.05)/\sqrt{N} = 0.01$, es decir

$N = 166.4$. Luego podemos tener confianza al 99% de que el error de la estimación será menor que 0.01 seg si N es 167 o mayor.

9.9. Una muestra al azar de 50 notas de matemáticas de entre un total de 200, revela una media de 75 y una desviación típica de 10.

(*a*) ¿Cuáles son los límites de confianza 95% para estimaciones de la media de las 200 notas?
(*b*) ¿Con qué grado de confianza podríamos decir que la media de las 200 es 75 ± 1?

Solución

(*a*) Como la población no es muy grande comparada con el tamaño de la muestra, debemos tenerlo en cuenta. Por tanto, los límites de confianza 95% son

$$\bar{X} \pm 1.96\sigma_{\bar{X}} = \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} = 75 \pm 1.96 \frac{10}{\sqrt{50}} \sqrt{\frac{200 - 50}{200 - 1}} = 75 \pm 2.4$$

(*b*) Los límites de confianza se pueden representar por

$$\bar{X} \pm z_c\sigma_{\bar{X}} = \bar{X} \pm z_c \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} = 75 \pm z_c \frac{10}{\sqrt{50}} \sqrt{\frac{200 - 50}{200 - 1}} = 75 \pm 1.23z_c$$

Como esto ha de ser igual a 75 ± 1, tenemos $1.23\, z_c = 1$, o sea $z_c = 0.81$. El área bajo la curva normal entre $z = 0$ y $z = 0.81$ es 0.2910; luego el requerido grado de confianza es $2(0.2910) = 0.582$, o sea 58.2%.

INTERVALOS DE CONFIANZA PARA PROPORCIONES

9.10. Un sondeo de 100 votantes elegidos al azar en un distrito indica que el 55% de ellos estaban a favor de un cierto candidato. Hallar los límites de confianza (*a*) 95%, (*b*) 99% y (*c*) 99.73% para la proporción de todos los votantes favorables a ese candidato.

Solución

(*a*) Los límites de confianza 95% para la población p son $P \pm 1.96\sigma_P = P \pm 1.96\sqrt{p(1-p)/N} = 0.55 \pm 1.96\sqrt{(0.55)(0.45)/100} = 0.55 \pm 0.10$, donde hemos usado la proporción muestral P para estimar p.
(*b*) Los límites de confianza 99% para p son $0.55 \pm 2.58\sqrt{(0.55)(0.45)/100} = 0.55 \pm 0.13$.
(*c*) Los límites de confianza 99.73% para p son $0.55 \pm 3\sqrt{(0.55)(0.45)/100} = 0.55 \pm 0.15$.

9.11. ¿De qué tamaño hay que tomar el sondeo del Problema 9.10 para tener confianza al (*a*) 95% y (*b*) 99.73% de que el candidato saldrá elegido?

Solución

Los límites de confianza para p son $P \pm z_c\sqrt{p(1-p)/N} = 0.55 \pm z_c\sqrt{(0.55)(0.45)/N} = 0.55 \pm 0.50z_c/\sqrt{N}$, donde hemos usado la estimación $P = p = 0.55$ basados en el Problema 9.10. Como el candidato ganará sólo si recibe más del 50% de los votos de la población, exigimos que $0.50z_c/\sqrt{N}$ sea menor que 0.05.

(*a*) Para 95% de confianza, $0.50z_c/\sqrt{N} = 0.50(1.96)/\sqrt{N} = 0.05$ cuando $N = 384.2$. Luego N debe ser al menos 385.

(*b*) Para 99.73 de confianza, $0.50z_c/\sqrt{N} = 0.50(3)/\sqrt{N} = 0.05$ cuando $N = 900$. Luego N debe ser al menos 901.

Otro método

$1.50/\sqrt{N} < 0.05$ cuando $\sqrt{N}/1.50 > 1/0.05$ o sea $\sqrt{N} > 1.50/0.05$. Entonces $\sqrt{N} > 30$, es decir, $N > 900$, así que N ha de ser al menos 901.

9.12. (*a*) Si P es la proporción observada de éxitos en una muestra de tamaño N, probar que los límites de confianza para estimar la proporción de éxitos p de la población en el nivel de confianza determinado por z_c vienen dados por

$$p = \frac{P + \frac{z_c^2}{2N} \pm z_c\sqrt{\frac{P(1-P)}{N} + \frac{z_c^2}{4N^2}}}{1 + \frac{z_c^2}{N}}$$

(*b*) Usar la fórmula deducida en (*a*) para obtener los límites de confianza 99.73% del Problema 9.10.

(*c*) Probar que para N grandes la fórmula de la parte (*a*) se reduce a $p = P \pm z_c\sqrt{P(1-P)/N}$, tal como se ha usado en el Problema 9.10.

Solución

(*a*) La proporción muestral P en unidades estándar es

$$\frac{P - p}{\sigma_P} = \frac{P - p}{\sqrt{p(1-p)/N}}$$

Los valores máximo y mínimo de esta variable tipificada son $\pm z_c$, donde z_c determina el valor de confianza. En estos valores extremos debemos tener en consecuencia

$$P - p = \pm z_c\sqrt{\frac{p(1-p)}{N}}$$

Elevando al cuadrado $$P^2 - 2pP + p^2 = \frac{z_c^2 p(1-p)}{N}$$

Multiplicando ambos lados por N y simplificando, encontramos que

$$(N + z_c^2)p^2 - (2NP + z_c^2)p + NP^2 = 0$$

Si $a = N + z_c^2$, $b = -(2NP + z_c^2)$ y $c = NP^2$, esta ecuación pasa a ser $ap^2 + bp + c = 0$ cuya solución para p viene dada por la fórmula cuadrática

$$p = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{2NP + z_c^2 \pm \sqrt{(2NP + z_c^2)^2 - 4(N + z_c^2)(NP^2)}}{2(N + z_c^2)}$$

$$= \frac{2NP + z_c^2 \pm z_c\sqrt{4NP(1-P) + z_c^2}}{2(N + z_c^2)}$$

Dividiendo el numerador y el denominador por $2N$, eso se convierte en

$$p = \frac{P + \frac{z_c^2}{2N} \pm z_c\sqrt{\frac{P(1-P)}{N} + \frac{z_c^2}{4N^2}}}{1 + \frac{z_c^2}{N}}$$

(*b*) Para límites de confianza 99.73%, $z_c = 3$. Entonces, usando $P = 0.55$ y $N = 100$ en la fórmula deducida en (*a*), vemos que $p = 0.40$ y 0.69, de acuerdo con el Problema 9.10(*c*).

(*c*) Si N es grande, entonces $z_c^2/(2N)$, $z_c^2/(4N^2)$ y z_c^2/N son todos despreciables y pueden tomarse esencialmente como cero, así que se llega al resultado deseado.

9.13. En 40 lanzamientos de una moneda, han salido 24 caras. Hallar los límites de confianza (*a*) 95% y (*b*) 99.73% para la proporción de caras que se obtendrían en un número ilimitado de lanzamientos de esa moneda.

Solución

(*a*) Al nivel 95%, $z_c = 1.96$. Haciendo $P = 24/40 = 0.6$ y $N = 40$ en la fórmula del Problema 9.12(*a*), hallamos $p = 0.45$ y 0.74. Luego podemos decir que, con 95% de confianza, p está entre 0.45 y 0.74.

Usando la fórmula aproximada $p = P \pm z_c\sqrt{P(1-P)/N}$, deducimos $p = 0.60 \pm 0.15$, que da al intervalo de 0.45 a 0.75.

(*b*) Al nivel 99.73%, $z_c = 3$. Usando la fórmula del Problema 9.12(*a*), hallamos $p = 0.37$ y 0.79.

Mediante la fórmula aproximada $p = P \pm z_c\sqrt{P(1-P)/N}$, hallamos $p = 0.60 \pm 0.23$, que da el intervalo de 0.37 a 0.83.

INTERVALOS DE CONFIANZA PARA DIFERENCIAS Y SUMAS

9.14. Una muestra de 150 lámparas del tipo A ha dado una vida media de 1400 horas (h) y una desviación típica de 120 h. Una muestra de 200 lámparas del tipo B dan vida media de 1200 h y desviación típica de 80 h. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para la diferencia de las vidas medias de las poblaciones de ambos tipos.

Solución

Los límites de confianza para la diferencia en medias de los dos tipos A y B vienen dados por

$$\bar{X}_A - \bar{X}_B \pm z_c\sqrt{\sigma_A^2/N_A + \sigma_B^2/N_B}$$

(*a*) Los límites de confianza 95% son $1400 - 1200 \pm 1.96\sqrt{(120)^2/150 + (80)^2/100} = 200 \pm 24.8$. Luego tenemos 95% de confianza de que la diferencia de las medias de las poblaciones está entre 175 y 225 h.

(*b*) Los límites de confianza 99% son $1400 - 1200 \pm 2.58\sqrt{(120)^2/150 + (80)^2/100} = 200 \pm 32.6$. Por tanto, tenemos 99% de confianza de que la diferencia de las medias de las poblaciones esté entre 167 y 233 h.

9.15. En una muestra aleatoria de 400 adultos y 600 jóvenes que vieron un cierto programa de televisión, 100 adultos y 300 jóvenes reconocieron que les había gustado. Determinar los límites de confianza (*a*) 95% y (*b*) 99% para la diferencia en proporciones de todos los adultos y jóvenes que vieron con agrado el programa.

Solución

Los límites de confianza para las diferencias en proporciones de los dos grupos vienen dados por

$$P_1 - P_2 \pm z_c\sqrt{p_1q_1/N_1 + p_2q_2/N_2}$$

donde los subindices 1 y 2 se refieren a jóvenes y adultos, respectivamente. Aquí, $P_1 = 300/600 = 0.50$ y $P_2 = 100/400 = 0.25$ son, respectivamente, las proporciones de jóvenes y de adultos a quienes agradó el programa.

(*a*) Los límites de confianza 95% son $0.50 - 0.25 \pm 1.96\sqrt{(0.50)(0.50)/600 + (0.25)(0.75)/400} = 0.25 \pm 0.06$. Luego tenemos 95% de confianza de que la verdadera diferencia en proporciones está entre 0.19 y 0.31.

(*b*) Los límites de confianza 99% son $0.50 - 0.25 \pm 2.58\sqrt{(0.50)(0.50)/600 + (0.25)(0.75)/400} = 0.25 \pm 0.08$. Luego tenemos 99% de confianza de que la verdadera diferencia en proporciones está entre 0.17 y 0.33.

9.16. La fuerza electromotriz media (fem) de las baterías producidas por una empresa es 45.1 voltios (V) y su desviación típica 0.04 V. Si se conectan en serie cuatro de ellas, hallar (*a*) 95%, (*b*) 99%, (*c*) 99.73% y (*d*) 50%.

Solución

Si E_1, E_2, E_3 y E_4 representa la fem de las cuatro baterías, tenemos

$$\mu_{E1+E2+E3+E4} = \mu_{E1} + \mu_{E2} + \mu_{E3} + \mu_{E4} \quad \text{y} \quad \sigma_{E1+E2+E3+E4} = \sqrt{\sigma_{E1}^2 + \sigma_{E2}^2 + \sigma_{E3}^2 + \sigma_{E4}^2}$$

Entonces, como $\mu_{E1} = \mu_{E2} = \mu_{E3} = \mu_{E4} = 45.1$ V y $\sigma_{E1} = \sigma_{E2} = \sigma_{E3} = \sigma_{E4} = 0.04$ V, tenemos $\mu_{E1+E2+E3+E4} = 4(45.1) = 180.4$ y $\sigma_{E1+E2+E3+E4} = \sqrt{4(0.04)^2} = 0.08$.

(*a*) Los límites de confianza 95% son $180.4 \pm 1.96(0.08) = 180.4 \pm 0.16$ V.
(*b*) Los límites de confianza 99% son $180.4 \pm 2.58(0.08) = 180.4 \pm 0.21$ V.
(*c*) Los límites de confianza 99.73% son $180.4 \pm 3(0.08) = 180.4 \pm 0.24$ V.
(*d*) Los límites de confianza 50% son $180.4 \pm 0.6745(0.08) = 180.4 \pm 0.054$ V. El valor 0.054 V se llama el *error probable*.

INTERVALOS DE CONFIANZA PARA DESVIACION TIPICA

9.17. La desviación típica de las vidas medias de una muestra de 200 bombillas es de 100 h. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para la desviación típica de ese tipo de bombillas.

Solución

Los límites de confianza para la desviación típica de la población σ vienen dados por $s \pm z_c\sigma/\sqrt{2N}$, donde z_c indica el nivel de confianza. Usamos la desviación típica muestral para estimar σ.

(*a*) Los límites de confianza 95% son $100 \pm 1.96(100)/\sqrt{400} = 100 \pm 9.8$. Luego tenemos 95% de confianza de que la desviación típica de la población está entre 90.2 y 109.8 h.

(*b*) Los límites de confianza 99% son $100 \pm 2.58(100)/\sqrt{400} = 100 \pm 12.9$. Luego tenemos 99% de confianza de que la desviación típica de la población está entre 87.1 y 112.9 h.

9.18. ¿De qué tamaño ha de tomarse una muestra de las bombillas del Problema 9.17 para tener 99.73% de confianza de que la verdadera desviación típica de la población no difiere de la desviación típica muestral en más de (*a*) 5% y (*b*) 10%?

Solución

Los límites de confianza 99% para σ son $s \pm 3\sigma/\sqrt{2N} = s \pm 3s/\sqrt{2N}$, usando s como estimación de σ. Luego el porcentaje de error en la desviación típica es

$$\frac{3s/\sqrt{2N}}{s} = \frac{300}{\sqrt{2N}}\%$$

(*a*) Si $300/\sqrt{2N} = 5$, entonces $N = 1800$. Luego la muestra ha de ser de al menos 1800 bombillas.
(*b*) Si $300/\sqrt{2N} = 10$, entonces $N = 450$. Por tanto, es necesaria una muestra de 450 o más bombillas.

ERROR PROBABLE

9.19. Los voltajes de 50 baterías del mismo tipo tienen una media de 18.2 V y una desviación típica de 0.5 V. Hallar (*a*) el error probable de la media y (*b*) los límites de confianza 50%.

Solución

(*a*) Error probable de la media $= 0.674\sigma_{\bar{X}} = 0.6745\dfrac{\sigma}{\sqrt{N}} = 0.6745\dfrac{\hat{s}}{\sqrt{N}}$

$$= 0.6745\frac{s}{\sqrt{N-1}} = 0.6745\frac{0.5}{\sqrt{49}} = 0.048 \text{ V}$$

Nótese que si la desviación típica de 0.5 V se toma como $\hat{s}$, el error probable es $0.6745(0.5/\sqrt{50}) = 0.048$ también, de modo que cualquier estimación puede utilizarse cuando N es lo bastante grande.

(*b*) Los límites de confianza 50% son 18 ± 0.048 V.

9.20. Se ha anotado una medida como 216.480 gramos (g) con un error probable de 0.272 g. ¿Cuáles son los límites de confianza 95% para esa medida?

Solución

El error probable es $0.272 = 0.6745\sigma_{\bar{X}}$, es decir, $\sigma_{\bar{X}} = 0.272/0.6745$. Luego los límites de confianza 95% son $\bar{X} \pm 1.96\sigma_{\bar{X}} = 216.480 \pm 1.96(0.272/0.6745) = 216.480 \pm 0.790$ g.

PROBLEMAS SUPLEMENTARIOS

ESTIMACIONES SIN SESGO Y EFICIENTES

9.21. Mediciones de una muestra de masas dieron 8.3, 10.6, 9.7, 8.8, 10,2 y 9.4 kilogramos (kg), respectivamente. Determinar estimaciones sin sesgo y eficientes de (*a*) la media de la población y (*b*) la varianza de la población, y comparar la desviación típica de la muestra con la estimada para la población.

9.22 Una muestra de 10 tubos de televisión procedentes de una cierta empresa dieron una vida media de 1200 h y una desviación típica de 100 h. Estimar (*a*) la media y (*b*) la desviación típica de la población de todos los tubos de esa clase.

9.23. (*a*) Rehacer el Problema 9.22 si los mismos

resultados se hubiesen dado con 30, 50, y 100 tubos.

(*b*) ¿Qué se puede concluir sobre la relación entre desviaciones típicas muestrales y estimaciones de las desviaciones típicas de la población para diferentes tamaños de las muestras?

INTERVALOS DE CONFIANZA PARA MEDIAS

9.24. La media y la desviación típica de las cargas máximas soportadas por 60 cables (véase Prob. 3.59) son 11.09 y 0.73 toneladas, respectivamente. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para la media de las cargas máximas soportadas por los cables de ese tipo.

9.25. La media y la desviación típica de los diámetros de una muestra de 250 remaches manufacturados por una empresa, son 0.72642 y 0.00058 in, respectivamente (véase Problema 3.61). Hallar los límites de confianza (*a*) 99%, (*b*) 98%, (*c*) 95% y (*d*) 90% para el diámetro medio de los remaches allí producidos.

9.26. Hallar (*a*) los límites de confianza 50% y (*b*) el error probable de los diámetros del Problema 9.25.

9.27. Si la desviación típica de las vidas medias de los tubos de televisión se estima en 100 h, ¿cómo de grande ha de ser una muestra para tener confianza del (*a*) 95%, (*b*) 90%, (*c*) 99% y (*d*) 99.73% de que el error en la vida media estimada no supera 20 h?

9.28. Idem si el error no debe superar 10 h.

9.29. Una empresa dispone de 500 cables, de los que una muestra de 40 elegidos al azar revela una tensión de ruptura media de 2400 lb y una desviación típica de 150 lb.

(*a*) Hallar los límites de confianza 95% y 99% para la estimación de la tensión media de ruptura de los 460 cables restantes.

(*b*) ¿Con qué grado de confianza se puede decir que la tensión media de ruptura de los 460 restantes es 2400 ± 35 lb?

INTERVALOS DE CONFIANZA PARA PROPORCIONES

9.30. Una urna contiene una proporción desconocida de fichas rojas y blancas. Una muestra aleatoria de 60 fichas, seleccionada con reposición, indicó que el 70% de ellas eran rojas. Hallar los límites de confianza (*a*) 95%, (*b*) 99% y (*c*) 99.73% para la proporción real de fichas rojas en la urna. Presentar los resultados usando tanto la fórmula aproximada como la más exacta del Problema 9.12.

9.31. ¿De qué tamaño ha de ser una muestra de las fichas del Problema 9.30 para tener confianza del (*a*) 95%, (*b*) 99% y (*c*) 99.73% de que la verdadera proporción no difiere de la muestral en más del 5%?

9.32. Se espera que una elección entre dos candidatos sea muy reñida. ¿Cuál es el mínimo número de votantes a sondear si se quiere tener un (*a*) 80%, (*b*) 90%, (*c*) 95% y (*d*) 99% de confianza sobre la decisión a favor de uno u otro?

INTERVALOS DE CONFIANZA PARA DIFERENCIAS Y SUMAS

9.33. De dos grupos similares de pacientes, *A* y *B*, con 50 y 100 individuos respectivamente, se suministró al *A* un nuevo tipo de somnífero y al *B* uno convencional. Para los del grupo *A* el número medio de horas de sueño fue 7.82 con desviación típica de 0.24 h. Para los del grupo *B*, 6.75 h y 0.30 h, respectivamente. Hallar los límites de confianza (*a*) 95% y (*b*) 99%, para la diferencia en media de las horas de sueño inducidas por ambos somníferos.

9.34. Una muestra de 200 tuercas de una cierta máquina probó que 15 eran defectuosas, mientras una muestra de 100 tuercas de otra máquina dio 12 defectuosas. Hallar los límites de confianza (*a*) 95%, (*b*) 99% y (*c*) 99.73% para la diferencia en proporciones de tuercas defectuosas de las dos máquinas. Discutir los resultados obtenidos.

9.35. Una compañía produce bolas de cojinetes de peso medio 0.638 lb y desviación típica de 0.012 lb. Hallar los límites de confianza (*a*)

95% y (*b*) 99% para los pesos de lotes de 100 bolas cada uno.

INTERVALOS DE CONFIANZA PARA DESVIACION TIPICA

9.36. La desviación típica de las tensiones de ruptura de 100 cables probados por una empresa era de 180 lb. Hallar los límites de confianza (*a*) 95%, (*b*) 99% y (*c*) 99.73% para la desviación típica de todos los cables de ese tipo.

9.37. Hallar el error probable de la desviación típica en el Problema 9.36.

9.38. ¿Cómo ha de ser de grande una muestra para tener confianza del (*a*) 95%, (*b*) 99% y (*c*) 99.73% de que la desviación típica de una población no diferirá de la desviación típica muestral en más del 2%?

CAPITULO 10

Teoría estadística de las decisiones

DECISIONES ESTADISTICAS

En la práctica nos vemos obligados con frecuencia a tomar decisiones relativas a una población sobre la base de información proveniente de muestras. Tales decisiones se llaman *decisiones estadísticas*. Por ejemplo, podemos querer decidir, basados en datos muestrales, si un método pedagógico es mejor que otro, o si una moneda está trucada o no.

HIPOTESIS ESTADISTICAS

Al intentar alcanzar una decisión, es útil hacer hipótesis (o conjeturas) sobre la población implicada. Tales hipótesis, que pueden ser o no ciertas, se llaman *hipótesis estadísticas*. Son, en general, enunciados acerca de las distribuciones de probabilidad de las poblaciones.

Hipótesis nula

En muchos casos formulamos una hipótesis estadística con el único propósito de rechazarla o invalidarla. Así, si queremos decidir si una moneda está trucada, formulamos la hipótesis de que la moneda es buena (o sea, $p = 0.5$, donde p es la probabilidad de cara). Análogamente, si deseamos decidir si un procedimiento es mejor que otro, formulamos la hipótesis de que *no* hay *diferencia* entre ellos (o sea, que cualquier diferencia observada se debe simplemente a fluctuaciones en el muestreo de la *misma* población). Tales hipótesis se suelen llamar *hipótesis nula* y se denotan por H_0.

Hipótesis alternativa

Toda hipótesis que difiera de una dada se llamará una *hipótesis alternativa*. Por ejemplo, si una hipótesis es $p = 0.5$, hipótesis alternativas podrían ser $p = 0.7$, $p \neq 0.5$ o $p > 0.5$. Una hipótesis alternativa a la hipótesis nula se denotará por H_1.

CONTRASTES DE HIPOTESIS Y SIGNIFICACION, O REGLAS DE DECISION

Si suponemos que una hipótesis particular es cierta pero vemos que los resultados hallados en una muestra aleatoria difieren notablemente de los esperados bajo tal hipótesis (o sea, esperados sobre la base del puro azar, por teoría de muestreo), entonces diremos que las diferencias observadas son *significativas* y nos veríamos inclinados a rechazar la hipótesis (o al menos a no aceptarla ante la evidencia obtenida). Así, si en 20 tiradas de una moneda salen 16 caras, estaríamos inclinados a rechazar la hipótesis de que la moneda es buena, aunque cabe la posibilidad de equivocarnos.

Los procedimientos que nos capacitan para determinar si las muestras observadas difieren significativamente de los resultados esperados, y por tanto nos ayudan a decidir si aceptamos o rechazamos hipótesis, se llaman *contrastes* (o *tests*) de *hipótesis* o *de significación* o *reglas de decisión*.

ERRORES DE TIPO I Y DE TIPO II

Si rechazamos una hipótesis cuando debiera ser aceptada, diremos que se ha cometido un *error de Tipo I*. Por otra parte, si aceptamos una hipótesis que debiera ser rechazada, diremos que se ha cometido un *error de Tipo II*. En ambos casos, se ha producido un juicio erróneo.

Para que las reglas de decisión (o contrastes de hipótesis) sean buenas, deben diseñarse de modo que minimicen los errores de la decisión. Y no es una cuestión sencilla, porque para cualquier tamaño de la muestra, un intento de disminuir un tipo de error suele ir acompañado de un crecimiento del otro tipo. En la práctica, un tipo de error puede ser más grave que el otro, y debe alcanzarse un compromiso que disminuya el error más grave. La única forma de disminuir ambos a la vez es aumentar el tamaño de la muestra, que no siempre es posible.

NIVEL DE SIGNIFICACION

Al contrastar una cierta hipótesis, la máxima probabilidad con la que estamos dispuestos a correr el riesgo de cometer un error de Tipo I se llama *nivel de significación* del contraste. Esta probabilidad, denotada a menudo por α, se suele especificar antes de tomar la muestra, de manera que los resultados obtenidos no influyan en nuestra elección.

En la práctica, es frecuente un nivel de significación de 0.05 ó 0.01, si bien se usan otros valores. Si, por ejemplo, se escoge el nivel de significación 0.05 (o 5%) al diseñar una regla de decisión, entonces hay unas 5 oportunidades entre 100 de rechazar la hipótesis cuando debiera haberse aceptado; es decir, tenemos un 95% *de confianza* de que hemos adoptado la decisión correcta. En tal caso decimos que la hipótesis ha sido rechazada al nivel de significación 0.05, lo cual quiere decir que la hipótesis tiene una probabilidad 0.05 de ser falsa.

CONTRASTES MEDIANTE LA DISTRIBUCION NORMAL

Para ilustrar las ideas presentadas hasta este momento, supongamos que bajo cierta hipótesis la distribución de muestreo de un estadístico S es una distribución normal con media μ_S y desviación

típica σ_S. Así pues, la distribución de la variable tipificada z, dada por $z = (S - \mu_S)/\sigma_S$, es la distribución normal canónica (media 0, varianza 1), como indica la Figura 10.1.

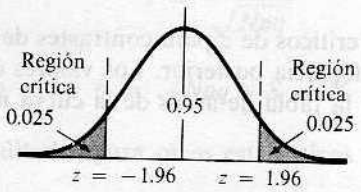


Figura 10.1.

Como se ve en la Figura 10.1, podemos tener 95% de confianza de que si la hipótesis es verdadera, entonces el valor de z para un estadístico muestral S estará entre -1.96 y 1.96 (porque el área bajo la curva normal entre esos valores es 0.95). Sin embargo, si al escoger una sola muestra al azar hallamos que el valor de z de su estadístico está *fuera* de ese rango, debemos concluir que tal suceso podría ocurrir con una probabilidad de sólo 0.05 (el área total sombreada en la figura) si la hipótesis dada fuera cierta. Diremos entonces que esta z difiere *de forma significativa* de lo que sería de esperar bajo la hipótesis, y nos veríamos empujados a rechazar la hipótesis.

El área total sombreada 0.05 es el nivel de significación del contraste. Representa la probabilidad de equivocarnos al rechazar la hipótesis (o sea, la probabilidad de un error de Tipo I). Así pues, decimos que la hipótesis *se rechaza a un nivel de significación 0.05*, o que el valor de z del estadístico muestral dado es *significativo al nivel 0.05*.

El conjunto de z fuera del rango -1.96 a 1.96 se llama *la región crítica de la hipótesis región de rechazo de la hipótesis*, o *región de significación*. El conjunto de z en el rango -1.96 a 1.96 se conoce como *región de aceptación de la hipótesis* o *región de no significación*.

Basados en las anteriores observaciones, podemos formular la siguiente regla de decisión (o contraste de hipótesis o significación):

Rechazar la hipótesis al nivel de significación 0.05 si el valor de z para el estadístico S está fuera del rango -1.96 a 1.96 (o sea, si $z > 1.96$ o $z < -1.96$). Esto equivale a decir que el estadístico muestral observado es significativo al nivel 0.05.

Aceptar la hipótesis en caso contrario (o, si se desea, no tomar decisión alguna).

Dado que z juega tan importante papel en el contraste de hipótesis, se le llama un *estadístico de contraste*.

Hay que hacer notar que se utilizan también otro nivel de significación. Por ejemplo, si se usa el nivel 0.01, debe sustituirse el 1.96 de antes por 2.58 (véase Tabla 10.1). Cabe utilizar asimismo la Tabla 9.1, ya que la suma de los niveles de significación y de confianza es 100%.

CONTRASTES DE UNA Y DE DOS COLAS

En el test precedente estábamos interesados en los valores extremos del estadístico S o en su correspondiente valor de z a *ambos* lados de la media (o sea, en las dos colas de la distribución). Tales tests se llaman *contrastes de dos colas* o *bilaterales*.

Con frecuencia, no obstante, estaremos interesados tan sólo en valores extremos a un lado de la media (o sea, en una de las colas de la distribución), tal como sucede cuando se contrasta la

hipótesis de que un proceso es mejor que otro (lo cual no es lo mismo que contrastar si un proceso es mejor o peor que el otro). Tales contrastes se llaman *unilaterales*, o *de una cola*. En tales situaciones, la región crítica es una región situada a un lado de la distribución, con área igual al nivel de significación.

La Tabla 10.1, que da valores críticos de z para contrastes de una o dos colas en varios niveles de significación, será útil como referencia posterior. Los valores críticos de z para otros niveles de significación se hallan a partir de la tabla de áreas de la curva normal (Apéndice II).

Tabla 10.1

Nivel de significación, α	0.10	0.05	0.01	0.005	0.002
Valores críticos de z para tests unilaterales	-1.28 o 1.28	-1.645 o 1.645	-2.33 o 2.33	-2.58 o 2.58	-2.88 o 2.88
Valores críticos de z para tests bilaterales	-1.645 y 1.645	-1.96 y 1.96	-2.58 y 2.58	-2.81 y 2.81	-3.08 y 3.08

CONTRASTES ESPECIALES

Para grandes muestras, las distribuciones de muestreo de muchos estadísticos son distribuciones normales (o casi normales), y los contrastes anteriores pueden aplicarse a los z correspondientes. Los siguientes casos especiales, tomados de la Tabla 8.1, no son sino unos pocos de los estadísticos de interés práctico. En cada caso los resultados son válidos para poblaciones infinitas o para muestreos con reposición. Para muestreos sin reposición en poblaciones finitas, esos resultados requieren modificación (véase pág. 186).

1. **Medias.** Aquí $S = \bar{X}$, la media muestral; $\mu_S = \mu_{\bar{X}} = \mu$, la media de la población; y $\sigma_S = \sigma_{\bar{X}} = \sigma/\sqrt{N}$, donde σ es la desviación típica de la población y N el tamaño de la muestra. El valor z viene dado por

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

Cuando sea necesario, se utilizará la desviación muestral s o $\hat{s}$ como estimación de σ.

2. **Proporciones.** Ahora $S = P$, la proporción de «éxitos» en una muestra; $\mu_S = \mu_P = p$, donde p es la proporción de éxitos de la población y N el tamaño de la muestra; y $\sigma_S = \sigma_P = \sqrt{pq/N}$, donde $q = 1 - p$.

El valor de z viene dado por

$$z = \frac{P - p}{\sqrt{pq/N}}$$

En el caso $P = X/N$, donde X es el número real de éxitos en una muestra, z es

$$z = \frac{X - N_p}{\sqrt{Npq}}$$

Esto es, $\mu_X = \mu = Np$, $\sigma_X = \sigma = \sqrt{Npq}$ y $S = X$.

Análogamente se obtienen los resultados para otros estadísticos.

CURVAS DE OPERACION CARACTERISTICAS; POTENCIA DE UN CONTRASTE

Hemos visto cómo limitar el error de Tipo I eligiendo adecuadamente el nivel de significación. Es posible evitar el riesgo de cometer error de Tipo II simplemente no aceptando nunca hipótesis, pero en muchas aplicaciones prácticas esto es inviable. En tales casos, se suele recurrir a *curvas de operación características*, o *curvas OC*, que son gráficos que muestran las probabilidades de error de Tipo II bajo diversas hipótesis. Proporcionan indicaciones de hasta qué punto un test dado nos permitirá evitar un error de Tipo II; es decir, nos indicará la *potencia de un test* a la hora de prevenir decisiones erróneas. Son útiles en el diseño de experimentos porque sugieren entre otras cosas el tamaño de muestra a manejar.

GRAFICOS DE CONTROL

A menudo adquiere importancia práctica saber cuándo un proceso ha variado tanto que deben adoptarse medidas para remediar la situación. Tales problemas aparecen, por ejemplo, en el control de calidad. Los supervisores del control de calidad han de decidir frecuentemente si los cambios observados se deben simplemente a fluctuaciones de azar o a cambios reales en un proceso de producción por deterioro de la maquinaria, descuidos de los empleados, etc. Los *gráficos de control* ponen a nuestra disposición un método sencillo y eficaz para enfrentarnos a esa clase de problemas (véase Prob. 10.16).

CONTRASTES MEDIANTE DIFERENCIAS MUESTRALES

Diferencias de medias

Sean $\bar{X}_1$ y $\bar{X}_2$ las medias muestrales obtenidas en grandes muestras de tamaños N_1 y N_2 tomadas de poblaciones con respectivas medias μ_1 y μ_2, y desviaciones típicas σ_1 y σ_2. Consideremos la hipótesis nula de que *no hay diferencia* entre las medias de las poblaciones (o sea, $\mu_1 = \mu_2$), que es como afirmar que las muestras se han tomado en dos poblaciones que tienen la misma media.

Poniendo $\mu_1 = \mu_2$ en la ecuación (5) del Capítulo 8, vemos que la distribución de muestreo de diferencia en medias está casi normalmente distribuida, con media y desviación típica dadas por

$$\mu_{\bar{X}1-\bar{X}2} = 0 \qquad \text{y} \qquad \sigma_{\bar{X}1-\bar{X}2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} \tag{1}$$

donde podemos, si es necesario, usar las desviaciones típicas muestrales s_1 y s_2 (o $\hat{s}_1$ y $\hat{s}_2$) como estimaciones de σ_1 y σ_2.

Usando la variable tipificada z dada por

$$z = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sigma_{\bar{X}1-\bar{X}2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}1-\bar{X}2}} \tag{2}$$

podemos contrastar la hipótesis nula frente a hipótesis alternativas (o la significación de una diferencia observada) a un nivel de significación apropiado.

Diferencias de proporciones

Sean P_1 y P_2 las proporciones muestrales obtenidas en grandes muestras de tamaños N_1 y N_2 tomadas de respectivas poblaciones que tienen proporciones p_1 y p_2. Consideremos la hipótesis nula de que *no hay diferencia* entre los parámetros de las poblaciones (o sea, $p_1 = p_2$) y por tanto que las muestras se han tomado de una misma población.

Poniendo $p_1 = p_2 = p$ en la ecuación (6) del Capítulo 8, vemos que la distribución de muestreo de diferencias en proporciones está casi normalmente distribuida, con media y desviación típica dadas por

$$\mu_{P1-P2} = 0 \quad \text{y} \quad \sigma_{P1-P2} = \sqrt{pq\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \tag{3}$$

donde

$$p = \frac{N_1P_1 + N_2P_2}{N_1 + N_2}$$

se usa como estimación para la proporción poblacional y donde $q = 1 - p$.

Mediante la variable tipificada

$$z = \frac{P_1 - P_2 - 0}{\sigma_{P1-P2}} = \frac{P_1 - P_2}{\sigma_{P1-P2}} \tag{4}$$

podemos contrastar diferencias observadas a un nivel de significación apropiado y, en consecuencia, contrastar la hipótesis nula.

Contrastes que involucran a otros estadísticos se diseñan de manera similar.

CONTRASTES MEDIANTE LA DISTRIBUCION BINOMIAL

También cabe diseñar contrastes mediante distribuciones binomiales (u otras distribuciones) de forma parecida a como se ha hecho con la distribución normal; los principios básicos son esencialmente los mismos. Véanse Problemas 10.23 a 10.28.

PROBLEMAS RESUELTOS

CONTRASTES DE MEDIAS Y PROPORCIONES USANDO DISTRIBUCIONES NORMALES

10.1. Hallar la probabilidad de sacar entre 40 y 60 caras inclusive en 100 tiradas de una moneda buena.

Solución

De acuerdo con la distribución binomial, la probabilidad pedida es

$$\binom{100}{40}\left(\frac{1}{2}\right)^{40}\left(\frac{1}{2}\right)^{60} + \binom{100}{41}\left(\frac{1}{2}\right)^{41}\left(\frac{1}{2}\right)^{59} + \cdots + \binom{100}{60}\left(\frac{1}{2}\right)^{60}\left(\frac{1}{2}\right)^{40}$$

Como $Np = 100(\frac{1}{2})$ y $Nq = 100(\frac{1}{2})$ son ambos mayores que 5, la aproximación normal a la distribución binomial es correcta a la hora de evaluar esa suma. La media y la desviación típica del número de caras en 100 tiradas son

$$\mu = Np = 100(\tfrac{1}{2}) = 50 \qquad \text{y} \qquad \sigma = \sqrt{Npq} = \sqrt{(100)(\tfrac{1}{2})(\tfrac{1}{2})} = 5$$

En una escala continua, decir entre 40 y 60 inclusive es como decir entre 39.5 y 60.5 caras. Luego

$$39.5 \text{ en unidades estándar} = \frac{39.5 - 50}{5} = -2.10 \qquad 60.5 \text{ en unidades estándar} = \frac{60.5 - 50}{5} = 2.10$$

$$\text{Probabilidad pedida} = \text{área bajo la curva normal entre } z = -2.10 \text{ y } z = 2.10$$
$$= 2(\text{área entre } z = 0 \text{ y } z = 2.10) = 2(0.4821) = 0.9642$$

10.2. Para contrastar la hipótesis de que una moneda es buena, adoptemos la siguiente regla de decisión:

Aceptarla si el número de caras en una sola muestra de 100 tiradas está entre 40 y 60 inclusive.

Rechazarla en caso contrario.

(*a*) Hallar la probabilidad de rechazar la hipótesis cuando en verdad sea correcta.
(*b*) Representar gráficamente la regla de decisión y el resultado de la parte (*a*).
(*c*) ¿Qué conclusiones se desprenden si resultan 53 caras en la muestra de 100 tiradas? ¿Y si salieran 60 caras?
(*d*) ¿Podría ser equivocada su conclusión sobre (*c*)? Explicar la respuesta.

Solución

(*a*) Del Problema 10.1, la probabilidad de no obtener entre 40 y 60 caras inclusive si la moneda es buena, es $1 - 0.9642 = 0.0358$. Luego la probabilidad de rechazar la hipótesis cuando sea correcta es 0.0358.

(*b*) La regla de decisión se ilustra en la Figura 10.2, que muestra las distribuciones de probabilidad de caras en 100 tiradas de una moneda buena. Si una sola muestra de 100 tiradas arroja un z entre -2.10 y 2.10, aceptamos la hipótesis; en caso contrario, la rechazamos y decidimos que la moneda está trucada.

El error de rechazar la hipótesis siendo correcta es el *error de Tipo I* de la regla de decisión; y su probabilidad, 0.0358 según (*a*), está representada por el área sombreada total en la figura. Si una sola muestra de 100 tiradas da un número de caras cuyo z está en las zonas sombreadas, diremos que ese valor de z difiere *de forma significativa* del esperado si la hipótesis fuese verdadera. Por tal razón, el área total sombreada (o sea, la probabilidad de un error de Tipo I) se llama el *nivel de significación* de la regla de decisión y vale 0.0358 en este caso. Así que podemos hablar de que rechazamos la hipótesis al nivel de significación 0.0358 (o sea al 3.58%).

(*c*) De acuerdo con la regla de decisión, tendremos que aceptar la hipótesis de que la moneda es buena en ambos casos. Cabe argumentar que con sólo una cara más ya la hubiésemos rechazado. ¡Siempre tiene uno que enfrentarse a una línea brusca de división al tomar decisiones!

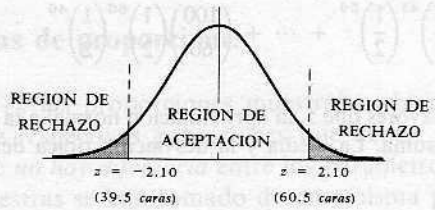


Figura 10.2.

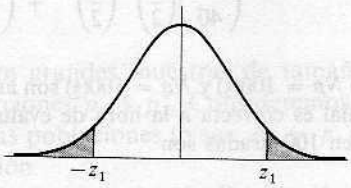


Figura 10.3.

(*d*) Sí. Podríamos aceptar la hipótesis cuando en realidad es rechazable, como sería el caso por ejemplo si la probabilidad de caras es 0.7 en vez de 0.5. El error cometido al aceptar la hipótesis que debiera ser rechazada es el *error de Tipo II* de la decisión. (Para más detalles, véanse Problemas 10.10 a 10.12).

10.3. Diseñar una regla de decisión para contrastar la hipótesis de que una moneda es buena y usar nivel de significación de (*a*) 0.05 y (*b*) 0.01.

Solución

(*a*) *Primer método*

Si el nivel de significación es 0.05, cada área sombreada en la Figura 10.3 es 0.025 por simetría. Entonces el área entre 0 y z_1 es $0.5000 - 0.0250 = 0.4750$, y $z_1 = 1.96$; los valores críticos -1.96 y 1.96 pueden leerse también en la Tabla 10.1. Así pues, una posible regla de decisión es:

Aceptar la hipótesis de que la moneda es buena si z está entre -1.96 y 1.96.

Rechazarla en caso contrario.

Para expresar la regla de decisión en términos del número de caras que se obtendrán en 64 tiradas de la moneda, nótese que la media y la desviación típica de la distribución de caras vienen dadas por:

$$\mu = Np = 64(0.5) = 32 \qquad \text{y} \qquad \sigma = \sqrt{Npq} = \sqrt{64(0.5)(0.5)} = 4$$

bajo la hipótesis de que la moneda es buena. Entonces $z = (X - \mu)/\sigma = (X - 32)/4$. Si $z = 1.96$, entonces $(X - 32)/4 = 1.96$ y $X = 39.84$; si $z = -1.96$, entonces $(X - 32)/4 = -1.96$ y $X = 24.16$. Luego la regla de decisión se convierte en:

Aceptar la hipótesis de que la moneda es buena si el número de caras está entre 24.16 y 39.84 (o sea, entre 25 y 39 inclusive).

Rechazarla en caso contrario.

Segundo método

Con probabilidad 0.95, el número de caras estará entre $\mu - 1.96\sigma$ y $\mu + 1.96\sigma$ (o sea, entre $Np - 1.96\sqrt{Npq}$ y $Np + 1.96\sqrt{Npq}$, es decir, entre $32 - 1.96(4) = 24.16$ y $32 + 1.96(4) = 39.84$, lo que conduce a la regla de decisión precedente.

Tercer método

Como $-1.96 < z < 1.96$ es equivalente a $-1.96 < \frac{1}{4}(X - 32) < 1.96$, entonces $-1.96(4) < (X - 32) < 1.96(4)$, o sea $32 - 1.96(4) < X < 32 + 1.96(4)$ (o sea, $24.16 < X < 39.84$), que también conduce a la anterior regla de decisión.

(*b*) Si el nivel de significación es 0.01, cada área sombreada en la Figura 10.3 es 0.005. Luego el área entre 0 y z_1 es $0.5000 - 0.0050 = 0.4950$ y $z_1 = 2.58$ (más exactamente 2.575); esto puede leerse en la Tabla 10.1. Siguiendo el procedimiento del segundo método de la parte (*a*), vemos que con probabilidad 0.99 el número de caras estará entre $\mu - 2.58\sigma$ y $\mu + 2.58\sigma$, que son $32 - 2.58(4) = 21.68$ y $32 + 2.58(4) = 42.32$. Luego la regla de decisión es:

Aceptar la hipótesis si el número de caras está entre 22 y 42 inclusive.

Rechazarla en caso contrario.

10.4. ¿Cómo diseñaría una regla de decisión en el Problema 10.3 de modo que se evite el error de Tipo II?

Solución

Un error de Tipo II consiste en aceptar una hipótesis falsa, y se puede evitar como sigue: en vez de aceptar la hipótesis, simplemente no la rechazamos, lo que quiere decir que estamos rehusando tomar decisión en ese caso. Por ejemplo, podríamos enunciar la regla de decisión del Problema 10.3(*b*) así:

No rechazar la hipótesis si el número de caras está entre 22 y 42 inclusive.

Rechazarla en caso contrario.

En muchas situaciones prácticas, es importante decidir si una hipótesis dada debe ser aceptada o rechazada. Una discusión completa de tales casos requiere considerar los errores de Tipo II (véanse Probs. 10.10 a 10.12).

10.5. En un experimento sobre percepción extrasensorial (PES), un individuo en una habitación es invitado a adivinar el color (rojo o azul) de una carta elegida de un mazo de 50 cartas bien mezcladas por otro individuo en otra habitación. El no sabe cuántas rojas y cuántas azules hay en el mazo. Si el sujeto identifica 32 cartas correctamente, determinar si el resultado es significativo al nivel (*a*) 0.05 y (*b*) 0.01.

Solución

Si p es la probabilidad de que el sujeto acierte el color de una carta, hemos de decidir entre dos hipótesis:

H_0: $p = 0.5$, y el sujeto está simplemente diciendo colores al azar.

H_1: $p > 0.5$, y el sujeto tiene poderes de PES.

Como no estamos interesados en el caso de que obtenga muy pocos aciertos, sino en el de que

consiga muchos, escogemos un contraste de una cola. Si la hipótesis H_0 es verdadera, la media y la desviación típica del número de cartas acertadas vienen dadas por

$$\mu = Np = 50(0.5) = 25 \quad \text{y} \quad \sigma = \sqrt{Npq} = \sqrt{50(0.5)(0.5)} = \sqrt{12.5} = 3.54$$

(*a*) Para un contraste unilateral al nivel de significación 0.05, debemos tomar z_1 en la Figura 10.4 de modo que el área en la región crítica sea 0.05. Entonces, el área entre 0 y z_1 es 0.4500 y $z_1 = 1.645$; lo que puede verse también en la Tabla 10.1. Luego nuestra regla de decisión (o contraste de significación) es:

Si el z observado es mayor que 1.645, el resultado es significativo al nivel 0.05 y el individuo tiene poderes PES.

En caso contrario, el resultado se debe al azar (no es significativo al nivel 0.05) y el sujeto no tiene PES.

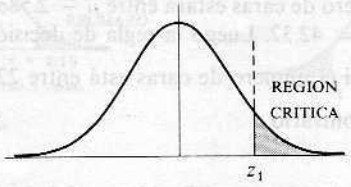


Figura 10.4.

Como 32 en unidades estándar es $(32 - 25)/3.54 = 1.98$, que es mayor que 1.645, concluimos que, al nivel 0.05, el individuo tiene poderes de PES.

Nótese que en realidad deberíamos aplicar una corrección de continuidad, porque 32 en escala continua está entre 31.5 y 32.5. Sin embargo, 31.5 tiene un valor estándar de $(31.5 - 25)/3.54 = 1.84$, y por tanto se alcanza idéntica conclusión.

(*b*) Si el nivel de significación es 0.01, el área entre 0 y z_1 es 0.4900, y $z_1 = 2.33$. Como 32 (o 31.5) en unidades estándar es 1.98 (o 1.84), que es menor que 2.33, concluimos que el resultado *no es significativo* al nivel 0.01.

Algunos estadísticos adoptan la terminología de que los resultados significativos al nivel 0.01 son *altamente significativos*, los que lo son al 0.05 pero no al 0.01 son *probablemente significativos*, y los que ni lo son al 0.05 se dicen *no significativos*. De modo que en el anterior experimento, el resultado es *probablemente significativo*, de manera que sería conveniente una investigación adicional.

Como los niveles de significación sirven de guía al tomar decisiones, algunos estadísticos citan las probabilidades implicadas. Así, como $\Pr\{z \geq 1.84\} = 0.0322$, en este problema, dirían que sobre la base del experimento, la probabilidad de equivocarnos al concluir que el sujeto tiene PES es de alrededor de un 3%. La probabilidad obtenida (0.0322 en este caso) se suele llamar *nivel de significación experimental o descriptivo*.

10.6. Un laboratorio de farmacia sostiene que uno de sus productos es 90% efectivo para reducir una alergia en 8 horas. En una muestra de 200 personas con esa alergia, el medicamento dio buen resultado en 160. Determinar si la afirmación del laboratorio es legítima.

Solución

Sea p la probabilidad de curación mediante ese fármaco. Hemos de decidir entre dos hipótesis:

H_0: $p = 0.9$, y la afirmación es correcta. H_1: $p < 0.9$, y la afirmación es falsa.

Como estamos interesados en determinar si la proporción de personas curadas es demasiado baja, elegimos un contraste de una cola. Si tomamos como nivel de significación el 0.01 (o sea, si el área sombreada en la Figura 10.5 es 0.01), entonces $z_1 = -2.33$, como se ve del Problema 10.5(*b*) por simetría de la curva o de la Tabla 10.1. Por tanto, adoptamos como regla de decisión:

No es legítima si z es menor que -2.33 (en cuyo caso rechazamos H_0).

En caso contrario, es legítima y los resultados observados se deben al azar (en cuyo caso aceptamos H_0).

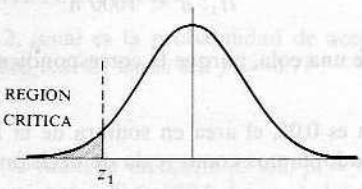


Figura 10.5.

Si H_0 es verdadera, entonces $\mu = Np = 200(0.9) = 180$ y $\sigma = \sqrt{Npq} = \sqrt{(200)(0.9)(0.1)} = 4.24$. Ahora bien, 160 en unidades estándar es $(160 - 180)/4.24 = -4.72$, que es mucho menor que -2.33. Luego, de acuerdo con nuestra regla de decisión, concluimos que la afirmación no es legítima y que los resultados del muestreo son altamente significativos (véase el final del Prob. 10.5).

10.7. La vida media de una muestra de 100 tubos fluorescentes producidos en una empresa es de 1570 h con una desviación típica de 120 h. Si μ es la vida media de todos los productos en esa empresa, contrastar la hipótesis de que $\mu = 1600$ h contra la hipótesis alternativa $\mu \neq 1600$ h, usando nivel de significación de (*a*) 0.05 y (*b*) 0.01.

Solución

Debemos decidir entre dos hipótesis:

$$H_0: \mu = 1600 \text{ h}$$
$$H_1: \mu \neq 1600 \text{ h}$$

Puesto que $\mu \neq 1600$ incluye valores mayores y menores que 1600, usaremos un contraste de dos colas.

(*a*) Para un contraste de dos colas al nivel de significación de 0.05, tenemos la siguiente regla de decisión:

Rechazar H_0 si el z de la media muestral está fuera del rango -1.96 a 1.96.

Aceptar H_0 en caso contrario.

El estadístico bajo consideración es la media muestral $\bar{X}$. La distribución de muestreo de $\bar{X}$ tiene media $\mu_{\bar{X}} = \mu$ y desviación típica $\sigma_{\bar{X}} = \sigma/\sqrt{N}$, donde μ y σ son la media y la desviación típica de toda la población de tubos producidos por la empresa. Bajo la hipótesis H_0, tenemos $\mu = 1600$ y $\sigma_{\bar{X}} = \sigma/\sqrt{N} = 120/\sqrt{100} = 12$, usando la desviación típica muestral como estimación de σ. Como $z = (\bar{X} - 1600)/12 = (1570 - 1600)/12 = -2.50$ está fuera del rango -1.96 a 1.96, rechazamos H_0 al nivel de significación 0.05.

(*b*) Si el nivel de significación es 0.01, el rango pasa a ser -2.58 a 2.58. Así pues, como el valor -2.50 de z cae dentro de ese rango, aceptamos H_0 (o rehusamos tomar decisión al nivel de significación 0.01.

10.8. En el Problema 10.7, contrastar la hipótesis $\mu = 1600$ h frente a la hipótesis alternativa $\mu < 1600$ h con nivel de significación de (*a*) 0.05 y (*b*) 0.01.

Solución

Tenemos que decidir entre las hipótesis:

$$H_0: \mu = 1600 \text{ h}$$
$$H_1: \mu < 1600 \text{ h}$$

Habrá que usar un contraste de una cola, porque la correspondiente figura es idéntica a la Figura 10.5 del Problema 10.6.

(*a*) Si el nivel de significación es 0.05, el área en sombra de la Figura 10.5 es 0.05, y hallamos que $z_1 = -1.645$. Por tanto, adoptamos como regla de decisión:

Rechazar H_0 si z es menor que -1.645.

Aceptarla en caso contrario (o declinar cualquier decisión).

Ya que [como en el Prob. 10.7(*a*)] z es -2.50, menor que -1.645, rechazamos H_0 al nivel 0.05. Nótese que esta decisión es idéntica a la alcanzada en el Problema 10.7(*a*) por medio de un contraste bilateral.

(*b*) Si el nivel de significación es 0.01, el valor z_1 en la Figura 10.5 es -2.33. Por consiguiente, adoptamos la regla de decisión siguiente:

Rechazar H_0 si z es menor que -2.33.

Aceptar H_0 en caso contrario (o declinar cualquier decisión).

Ya que [como en el Prob. 10.7(*a*)] z es -2.50, menor que -2.33, rechazamos H_0 al nivel 0.01. Nótese que esta decisión no es la alcanzada en el Problema 10.7(*b*) por medio de un contraste bilateral.

Se deduce que las decisiones relativas a una cierta hipótesis H_0 que están basadas en contrastes de una o dos colas no siempre concuerdan. Lo cual era de esperar, naturalmente, pues estamos contrastando H_0 frente a alternativas diferentes según el caso.

10.9. Las tensiones de ruptura de los cables fabricados por una empresa tienen media de 1800 lib y una desviación típica de 100 lb. Se desea comprobar si un nuevo proceso de fabricación aumenta dicha tensión media. Para ello se toma una muestra de 50 cables y se encuentra que su tensión media de ruptura es 1850 lb. ¿Se puede afirmar la mejoría del nuevo proceso al nivel de significación 0.01?

Solución

Tenemos que decidir entre dos hipótesis:

H_0: $\mu = 1800$ lb, y no hay realmente cambio en la tensión de ruptura.

H_1: $\mu > 1800$ lb, y hay realmente cambio en la tensión de ruptura.

Hay que usar un contraste de una cola; el diagrama asociado con él es idéntico a la Figura 10.4. Al nivel de significación 0.01, la regla de decisión es:

Si el z observado es mayor que 2.33, el resultado es significativo al nivel 0.01 y rechazamos H_0.

En caso contrario, se acepta H_0 (o se aplaza la decisión).

Bajo la hipótesis de que H_0 es verdadera, vemos que

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} = \frac{1850 - 1800}{100/\sqrt{50}} = 3.55$$

que es mayor que 2.33. Así que el resultado es altamente significativo y la afirmación puede mantenerse.

CURVAS DE OPERACION CARACTERISTICAS

10.10. Refiriendo al Problema 10.2, ¿cuál es la probabilidad de aceptar la hipótesis de que la moneda es buena cuando la probabilidad real de caras sea $p = 0.7$?

Solución

La hipótesis H_0 de que la moneda es buena (o sea, $p = 0.5$), es aceptada cuando el número de caras en 100 lanzamientos está entre 39.5 y 60.5. La probabilidad de rechazar H_0 cuando debería ser aceptada (o sea, la probabilidad de un error de Tipo I) viene representada por el área total α de la región sombreada de la izquierda en la Figura 10.6. Como calculamos en el Problema 10.2(*a*), esa área, que representa el nivel de significación del contraste de H_0, es igual a 0.0358.

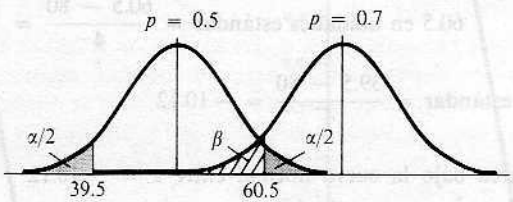


Figura 10.6.

Si $p = 0.7$, la distribución de caras en 100 lanzamientos está representada por la curva normal a la derecha en la Figura 10.6. Del diagrama es claro que la probabilidad de aceptar H_0 cuando en verdad $p = 0.7$ (es decir, la probabilidad de un error de Tipo II) viene dada por el área rayada β de la figura. Para calcularla, observamos que la distribución bajo la hipótesis $p = 0.7$ tiene media y desviación típica dadas por

$$\mu = Np = (100)(0.7) = 70 \quad \text{y} \quad \sigma = \sqrt{Npq} = \sqrt{(100)(0.7)(0.3)} = 4.58$$

$$60.5 \text{ en unidades estándar} = \frac{60.5 - 70}{4.58} = -2.07$$

$$39.5 \text{ en unidades estándar} = \frac{39.5 - 70}{4.58} = -6.66$$

Entonces $\beta = (\text{área bajo la curva normal entre } z = -6.66 \text{ y } z = -2.07) = 0.0192$

Luego hay poca opción, con la regla de decisión adoptada, de aceptar la hipótesis de que la moneda es buena si tiene en verdad $p = 0.7$.

Nótese que en este problema se nos da la regla de decisión, de la que calculamos α y β. En la práctica, aparecen otras dos posibilidades:

(1) Acordamos un α (tal como 0.05 o 0.01), llegamos a una decisión y entonces calculamos β.
(2) Acordamos α y β, y entonces llegamos a una regla de decisión.

10.11. Resolver el Problema 10.10 si (*a*) $p = 0.6$, (*b*) $p = 0.8$, (*c*) $p = 0.9$ y (*d*) $p = 0.4$.

Solución

(*a*) Si $p = 0.6$, la distribución de caras tiene su media y su desviación típica dadas por

$$\mu = Np = (100)(0.6) = 60 \quad \text{y} \quad \sigma = \sqrt{Npq} = \sqrt{(100)(0.6)(0.4)} = 4.90$$

$$60.5 \text{ en unidades estándar} = \frac{60.5 - 60}{4.90} = 0.102$$

$$39.5 \text{ en unidades estándar} = \frac{39.5 - 60}{4.90} = -4.18$$

Entonces β = (área bajo la curva normal entre $z = -4.18$ y $z = 0.102$) = 0.5406

Así que con la regla de decisión dada existen muchas posibilidades de aceptar la hipótesis de que la moneda es buena aunque en realidad tiene $p = 0.6$.

(*b*) Si $p = 0.8$, entonces

$$\mu = Np = (100)(0.8) = 80 \quad \text{y} \quad \sigma = \sqrt{Npq} = \sqrt{(100)(0.8)(0.2)} = 4$$

$$60.5 \text{ en unidades estándar} = \frac{60.5 - 80}{4} = -4.88$$

$$39.5 \text{ en unidades estándar} = \frac{39.5 - 80}{4} = -10.12$$

Entonces β = (área bajo la curva normal entre $z = -10.12$ y $z = -4.88$) = 0.0000 muy aproximadamente.

(*c*) Comparando con la parte (*b*) o por cálculo, vemos que si $p = 0.9$, entonces $\beta = 0$ a efectos prácticos.

(*d*) Por simetría, $p = 0.4$ da el mismo valor de β que $p = 0.6$ (es decir, $\beta = 0.5040$).

10.12. Representar los resultados de los Problemas 10.10 y 10.11 construyendo un gráfico de (*a*) β versus p y (*b*) $(1 - \beta)$ versus p. Interpretar los gráficos obtenidos.

Solución

La Tabla 10.2 muestra los valores de β correspondientes a valores dados de p, tal como se obtienen en el Problema 10.10 y en el 10.11. Aquí β representa la probabilidad de aceptar la hipótesis $p = 0.5$ cuando p es algún otro valor; si en verdad es $p = 0.5$, podemos interpretar β como la probabilidad de aceptar $p = 0.5$ cuando de hecho debía ser aceptada. Esta propiedad es $1 - 0.0358 = 0.9642$ y se ha incluido en la Tabla 10.2.

Tabla 10.2

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
β	0.0000	0.0000	0.0192	0.5040	0.9642	0.5040	0.0192	0.0000	0.0000

(*a*) El gráfico de β versus p, que se ve en la Figura 10.7(*a*), se llama la *curva de operación característica*, o *curva OC*, de la regla de decisión (o contraste de hipótesis). La distancia de su máximo a la recta $\beta = 1$ es igual a $\alpha = 0.0358$, el nivel de significación del test.

En general, cuanto más agudo el pico de la curva OC, mejor es la regla de decisión a la hora de rechazar hipótesis incorrectas.

(*b*) El gráfico de $(1 - \beta)$ versus p, Figura 10.7(*b*), se llama la *curva de potencia* de la regla de decisión. Se obtiene sin más que invertir la curva OC; luego ambos gráficos son equivalentes.

La cantidad $(1 - \beta)$ se suele llamar una *función de potencia*, porque indica la *potencia de un test* (o *contraste*) para rechazar hipótesis falsas, rechazables en consecuencia. La cantidad β se llama *función de operación característica* de un test.

10.13. Una compañía produce sogas cuya tensión de ruptura tiene media de 300 lb y desviación típica de 24 lb. Se espera que un nuevo proceso de fabricación haga crecer la media.

(*a*) Diseñar una regla de decisión para rechazar el proceso antiguo al nivel de significación 0.01 con una muestra de 64 sogas.

(*b*) Con esa regla de decisión, ¿cuál es la probabilidad de aceptar el antiguo procedimiento cuando de hecho el nuevo ha aumentado la tensión media de las sogas a 310 lb? Suponemos que la desviación típica es todavía de 24 lb.

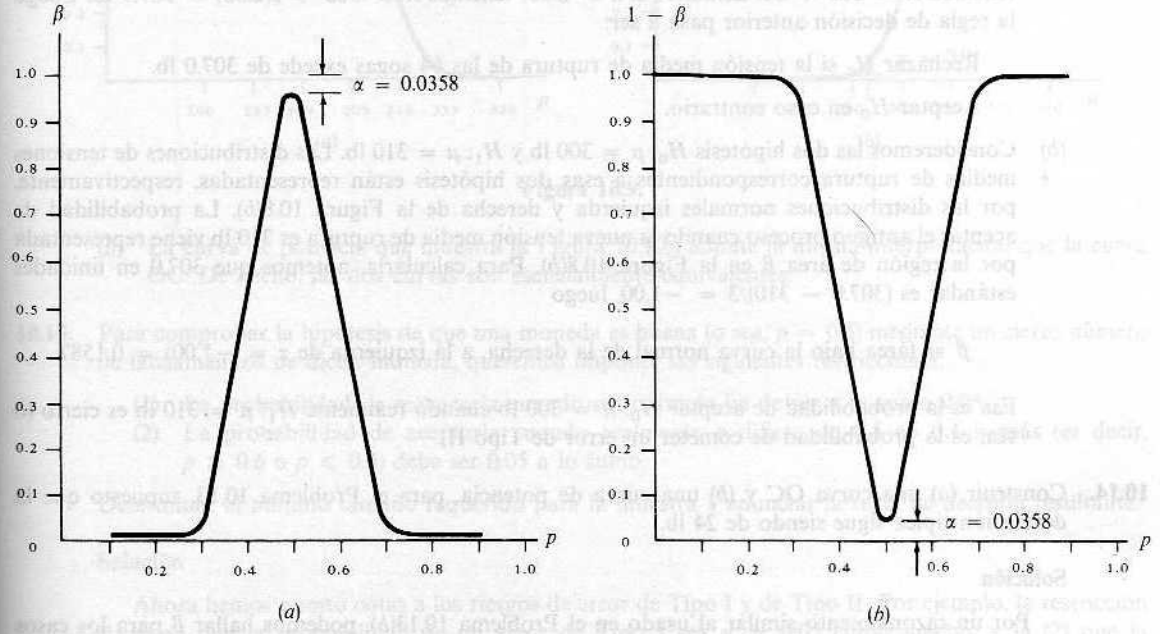


Figura 10.7.

Solución

(*a*) Si μ es la tensión media de ruptura, queremos decidir entre dos hipótesis:

H_0: $\mu = 300$ lb, y el nuevo proceso es como el antiguo.

H_1: $\mu > 300$ lb, y el nuevo proceso es mejor que el antiguo.

Para un contraste de una cola al nivel de significación 0.01, tenemos la siguiente regla de decisión [véase Fig. 10.8(*a*)]:

Rechazar H_0 si el valor z para la tensión media de ruptura es mayor que 2.33.

Aceptar H_0 en caso contrario.

Como $$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} = \frac{\bar{X} - 300}{24/\sqrt{64}}$$

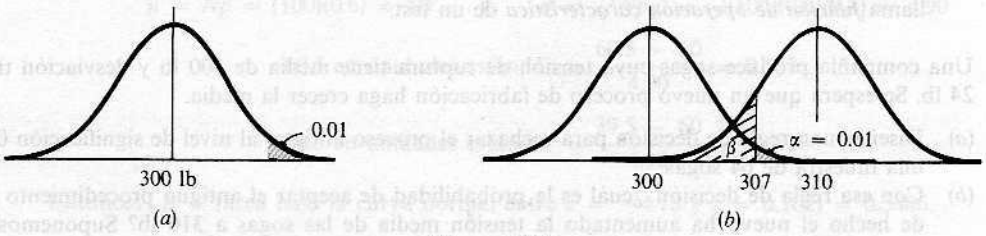


Figura 10.8.

tenemos $\bar{X} = 300 + 3z$. Entonces si $z > 2.33$, tenemos $\bar{X} > 300 + 3(2.33) = 307.7$ lb. Luego la regla de decisión anterior pasa a ser:

Rechazar H_0 si la tensión media de ruptura de las 64 sogas excede de 307.0 lb.

Aceptar H_0 en caso contrario.

(*b*) Consideremos las dos hipótesis H_0: $\mu = 300$ lb y H_1: $\mu = 310$ lb. Las distribuciones de tensiones medias de ruptura correspondientes a esas dos hipótesis están representadas, respectivamente, por las distribuciones normales izquierda y derecha de la Figura 10.8(*b*). La probabilidad de aceptar el antiguo proceso cuando la nueva tensión media de ruptura es 310 lb viene representada por la región de área β en la Figura 10.8(*b*). Para calcularla, notemos que 307.0 en unidades estándar es $(307.0 - 310)/3 = -1.00$, luego

$$\beta = (\text{área bajo la curva normal de la derecha, a la izquierda de } z = -1.00) = 0.1587$$

Esa es la probabilidad de aceptar H_0: $\mu = 300$ lb cuando realmente H_1: $\mu = 310$ lb es cierto (o sea, es la probabilidad de cometer un error de Tipo II).

10.14. Construir (*a*) una curva OC y (*b*) una curva de potencia, para el Problema 10.13, supuesto que la desviación típica sigue siendo de 24 lb.

Solución

Por un razonamiento similar al usado en el Problema 10.13(*b*), podemos hallar β para los casos en que el nuevo proceso de tensiones medias de ruptura μ iguales a 305 lb, 315 lb, etc. Por ejemplo, si $\mu = 305$ lb, entonces 307.0 lb en unidades estándar es $(307.0 - 305)/3 = 0.67$, y por tanto

$$\beta = (\text{área bajo la curva normal de la derecha, a la izquierda de } z = 0.67) = 0.7486$$

De esta forma se obtiene la Tabla 10.3.

Tabla 10.3

μ	290	295	300	305	310	315	320
β	1.0000	1.0000	0.9900	0.7486	0.1587	0.0038	0.0000

(*a*) La curva OC se ve en la Figura 10.9(*a*). En ella apreciamos que la probabilidad de conservar el antiguo proceso si la nueva tensión media de ruptura es menor que 300 lb es casi (excepto para el nivel de significación 0.01 cuando el nuevo proceso da una media de 300 lb). Cae rápidamente a cero, lo cual quiere decir que no hay prácticamente opción de mantener el antiguo proceso cuando la tensión media de ruptura es mayor que 315 lb.

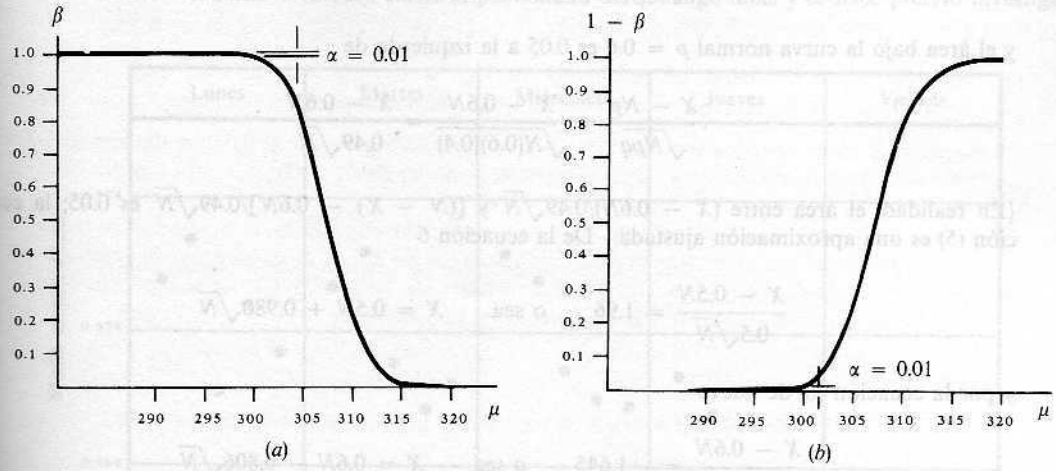


Figura 10.9.

(*b*) La curva de potencia que muestra la Figura 10.9(*b*) admite la misma interpretación que la curva OC. De hecho, las dos curvas son esencialmente equivalentes.

10.15. Para comprobar la hipótesis de que una moneda es buena (o sea, $p = 0.5$) mediante un cierto número de lanzamientos de dicha moneda, queremos imponer las siguientes restricciones:

(1) La probabilidad de rechazarla cuando sea correcta ha de ser a lo sumo 0.05.
(2) La probabilidad de aceptarla cuando realmente p difiera de 0.5 en 0.1 o más (es decir, $p \geqslant 0.6$ o $p \leqslant 0.4$) debe ser 0.05 a lo sumo.

Determinar el mínimo tamaño requerido para la muestra y enunciar la regla de decisión resultante.

Solución

Ahora hemos puesto cotas a los riesgos de error de Tipo I y de Tipo II. Por ejemplo, la restricción (1) exige que la probabilidad de un error de Tipo I sea $\alpha = 0.05$ como mucho, y la (2) que la probabilidad de un error de Tipo II sea $\beta = 0.05$ a lo más. La situación se refleja en la Figura 10.10.

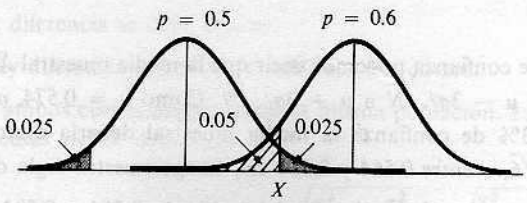


Figura 10.10.

Sea N el tamaño requerido para la muestra y X el número de caras en N tiradas, por encima del cual rechazamos la hipótesis de que $p = 0.5$. De la Figura 10.10, el área bajo la curva normal $p = 0.5$ es 0.025 a la derecha de

$$\frac{X - Np}{\sqrt{Npq}} = \frac{X - 0.5N}{\sqrt{N(0.5)(0.5)}} = \frac{X - 0.5N}{0.5\sqrt{N}} \tag{5}$$

y el área bajo la curva normal $p = 0.6$ es 0.05 a la izquierda de

$$\frac{X - Np}{\sqrt{Npq}} = \frac{X - 0.6N}{\sqrt{N(0.6)(0.4)}} = \frac{X - 0.6N}{0.49\sqrt{N}} \tag{6}$$

{En realidad, el área entre $(X - 0.6N)/0.49\sqrt{N}$ y $[(N - X) - 0.6N]/0.49\sqrt{N}$ es 0.05; la ecuación (5) es una aproximación ajustada.} De la ecuación 6

$$\frac{X - 0.5N}{0.5\sqrt{N}} = 1.96 \quad \text{o sea} \quad X = 0.5N + 0.980\sqrt{N} \tag{7}$$

y por la ecuación (6) de nuevo

$$\frac{X - 0.6N}{0.49\sqrt{N}} = -1.645 \quad \text{o sea} \quad X = 0.6N - 0.806\sqrt{N} \tag{8}$$

Y de (7) y (8) deducimos $N = 318.98$, luego la muestra ha de ser de 319 al menos (o sea, hay que lanzar al menos 319 veces la moneda). Poniendo $N = 319$ en la ecuación (7) u (8), $X = 177$.

Para $p = 0.5$ se tiene por tanto $X - Np = 177 - 159.5 = 17.5$. En consecuencia, adoptamos la siguiente regla de decisión:

Aceptar la hipótesis de que $p = 0.5$ si el número de caras en 319 lanzamientos está en el rango 159.5 ± 17.5 (o sea, entre 142 y 177).

Rechazarla en caso contrario.

GRAFICOS DE CONTROL

10.16. Se construye una máquina para fabricar bolas de rodamiento con diámetro medio de 0.574 cm y desviación típica de 0.008 cm. Para determinar si funciona correctamente, se toma una muestra de 6 bolas cada 2 horas y se halla para cada una de las muestras el diámetro medio.

(*a*) Diseñar una regla de decisión con la que se esté muy seguro de que la calidad del producto cumple los propósitos exigidos.

(*b*) Ilustrar gráficamente la regla de decisión de (*a*).

Solución

(*a*) Con el 99.73% de confianza podemos decir que la media muestral $\bar{X}$ debe estar entre $\mu_{\bar{X}} - 3\sigma_{\bar{X}}$ y $\mu_{\bar{X}} + 3\sigma_{\bar{X}}$, o sea $\mu - 3\sigma/\sqrt{N}$ a $\mu + 3\sigma/\sqrt{N}$. Como $\mu = 0.574$, $\sigma = 0.008$ y $N = 6$, se sigue que con el 99.73% de confianza la media muestral debería estar entre $0.574 - 0.024/\sqrt{6}$ y $0.574 + 0.024/\sqrt{6}$, o entre 0.564 y 0.584 cm. Luego nuestra regla de decisión es como sigue:

Si una media muestral cae dentro del rango de 0.564 a 0.584, aceptamos que la máquina funciona bien.

Si no, concluimos que no funciona bien e investigamos la razón.

(*b*) Se pueden anotar las observaciones en un gráfico como el de la Figura 10.11, llamado un *gráfico de control de calidad*. Cada vez que se toma una muestra, se representa por un punto concreto. En tanto que los puntos están entre el límite inferior (0.564 cm) y el superior (0.584 cm), el proceso está bajo control. Cuando un punto se sale de esos límites de control (como sucede con la tercera muestra tomada el jueves), existe la posibilidad de que algo falle, y se hace preciso investigarlo.

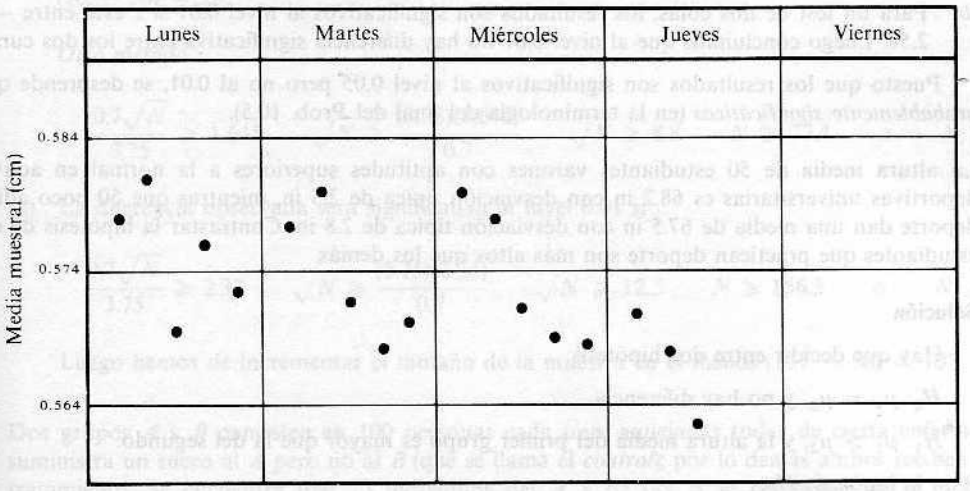


Figura 10.11.

Los límites de control antes especificados se llaman los límites de confianza 9.73%, o más brevemente, los límites 3σ. Otros límites de confianza (tales como 99% o 95%) se determinan del mismo modo. La elección en cada caso depende de las circunstancias particulares.

CONTRASTES MEDIANTE DIFERENCIAS DE MEDIAS Y PROPORCIONES

10.17. En un mismo examen realizado en dos cursos, la nota media del primero fue 74 con desviación típica 8, y en el otro fue 78 con desviación típica 7. ¿Hay diferencia significativa entre las calificaciones de ambos cursos al nivel de significación (*a*) 0.05 y (*b*) 0.01?

Solución

Supongamos que los dos cursos provienen de dos poblaciones con medias respectivas μ_1 y μ_2. Hemos de decidir entre las dos hipótesis:

H_0: $\mu_1 = \mu_2$, y la diferencia se debe al azar.

H_1: $\mu_1 \neq \mu_2$, y hay diferencia significativa entre los dos cursos.

Bajo la hipótesis H_0. ambos cursos provienen de la misma población. La media y la desviación típica de la diferencia en medias vienen dadas por

$$\mu_{\bar{X}_1 - \bar{X}_2} = 0 \quad \text{y} \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{8^2}{40} + \frac{7^2}{50}} = 1.606$$

donde hemos usado las desviaciones típicas muestrales como estimaciones de σ_1 y σ_2. Así pues

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}1-\bar{X}2}} = \frac{74 - 78}{1.606} = -2.49$$

(*a*) Para un test de dos colas, los resultados son significativos al nivel 0.05 si z está entre -1.96 y 1.96. Luego concluimos que al nivel 0.05 hay diferencia significativa y probablemente es mejor el segundo de los cursos.

(*b*) Para un test de dos colas, los resultados son significativos al nivel 0.01 si z está entre -2.58 y 2.58. Luego concluimos que al nivel 0.01 no hay diferencia significativa entre los dos cursos.

Puesto que los resultados son significativos al nivel 0.05 pero no al 0.01, se desprende que son *probablemente significativos* (en la terminología del final del Prob. 10.5).

10.18. La altura media de 50 estudiantes varones con aptitudes superiores a la normal en actividades deportivas universitarias es 68.2 in con desviación típica de 2.5 in, mientras que 50 poco adictos al deporte dan una media de 67.5 in con desviación típica de 2.8 in. Contrastar la hipótesis de que los estudiantes que practican deporte son más altos que los demás.

Solución

Hay que decidir entre dos hipótesis:

H_0: $\mu_1 = \mu_2$, y no hay diferencia.

H_1: $\mu_1 > \mu_2$, y la altura media del primer grupo es mayor que la del segundo.

Bajo la hipótesis H_0,

$$\mu_{\bar{X}1-\bar{X}2} = 0 \quad \text{y} \quad \sigma_{\bar{X}1-\bar{X}2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{(2.5)^2}{50} + \frac{(2.8)^2}{50}} = 0.53$$

donde hemos usado las desviaciones típicas muestrales como estimaciones de σ_1 y σ_2. Luego

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}1-\bar{X}2}} = \frac{68.2 - 67.5}{0.53} = 1.32$$

Con un contraste de una cola al nivel de significación 0.05, rechazaríamos H_0 si z fuera mayor que 1.645. Así que no podemos rechazarla a este nivel de significación.

Hay que hacer notar, no obstante, que la hipótesis puede ser rechazada al nivel 0.01 si estamos dispuestos a correr el riesgo de equivocarnos con una probabilidad de 0.10 (un 10%).

10.19. ¿Cuánto hay que aumentar el tamaño de la muestra en cada uno de los grupos del Problema 10.18 al objeto de que la diferencia observada de 0.7 in en las alturas medias sea significativa al nivel (*a*) 0.05 y (*b*) 0.01?

Solución

Sea N el tamaño de la muestra en cada grupo y supongamos que la desviación típica de los grupos sigue siendo la misma. Entonces, bajo la hipótesis H_0 tenemos

$$\mu_{\bar{X}1-\bar{X}2} = 0 \quad \text{y} \quad \sigma_{\bar{X}1-\bar{X}2} = \sqrt{\frac{\sigma_1^2}{N} + \frac{\sigma_2^2}{N}} = \sqrt{\frac{(2.5)^2 + (2.8)^2}{N}} = \sqrt{\frac{14.09}{N}} = \frac{3.75}{\sqrt{N}}$$

Para una diferencia observada en alturas medias de 0.7 in, tenemos pues

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}1-\bar{X}2}} = \frac{0.7}{3.75/\sqrt{N}} = \frac{0.7\sqrt{N}}{3.75}$$

(*a*) La diferencia observada será significativa al nivel 0.05 si $0.7\sqrt{N}/3.75 = 1.645$ al menos, de modo que N ha de ser al menos 78. Por tanto debemos aumentar el tamaño de la muestra en al menos $(78 - 50) = 28$.

Otro método

$$\frac{0.7\sqrt{N}}{3.75} \geq 1.645 \qquad \sqrt{N} \geq \frac{(3.75)(1.645)}{0.7} \qquad \sqrt{N} \geq 8.8 \qquad N \geq 77.4 \qquad \text{o} \qquad N \geq 78$$

(*b*) La diferencia observada será significativa al nivel 0.01 si

$$\frac{0.7\sqrt{N}}{3.75} \geq 2.33 \qquad \sqrt{N} \geq \frac{(3.75)(2.33)}{0.7} \qquad \sqrt{N} \geq 12.5 \qquad N \geq 156.3 \qquad \text{o} \qquad N \geq 157$$

Luego hemos de incrementar el tamaño de la muestra en el menos $(157 - 50) = 107$.

10.20. Dos grupos A y B consisten en 100 personas cada uno, aquejadas todas de cierta enfermedad. Se suministra un suero al A pero no al B (que se llama el *control*); por lo demás ambos reciben idéntico tratamiento. Se encuentra que 75 individuos del A y 65 del B se recuperan de la enfermedad. Contrastar la hipótesis de que el suero cura la enfermedad al nivel de significación (*a*) 0.01, (*b*) 0.05 y (*c*) 0.10.

Solución

Sean p_1 y p_2 las proporciones de población curadas (1) con, y (2) sin ese suero. Hemos de decidir entre dos hipótesis:

H_0: $p_1 = p_2$, y la diferencia observada se debe al azar (el suero es ineficaz).

H_1: $p_1 > p_2$, y el suero es eficaz.

Bajo la hipótesis H_0,

$$\mu_{P1-P2} = 0 \qquad \text{y} \qquad \sigma_{P1-P2} = \sqrt{pq\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} = \sqrt{(0.70)(0.30)\left(\frac{1}{100} + \frac{1}{100}\right)} = 0.0648$$

donde hemos usado como estimación de p la proporción media de curaciones en los dos grupos muestra, dadas por $(75 + 65)/200 = 0.70$, donde $q = 1 - p = 0.30$. Por tanto

$$z = \frac{P_1 - P_2}{\sigma_{P1-P2}} = \frac{0.750 - 0.650}{0.0648} = 1.54$$

(*a*) Con contraste de una cola al nivel de significación 0.01, debemos rechazar H_0 sólo si el valor z es mayor que 2.33. Como z es 1.54, concluimos que los resultados se deben al azar, a este nivel de significación.

(*b*) Con contraste de una cola al nivel de significación 0.05, debemos rechazar H_0 sólo si el valor z

es mayor que 1.645. Por tanto, concluimos que los resultados se deben al azar a este nivel de significación también.

(*c*) Con contraste de una cola al nivel de significación 0.10, debemos rechazar H_0 sólo si el valor z es mayor que 1.28. Como z es 1.54, concluimos que el suero es eficaz a este nivel de significación.

Nótese que estas conclusiones dependen de cuánto estamos dispuestos a arriesgar en equivocarnos. Si los resultados fuesen realmente debidos al azar, pero concluyésemos que el suero es eficaz (error de Tipo I), podríamos proceder a suministrarlo a grupos más grandes de enfermos, y nos convenceríamos finalmente de su ineficacia. Es un riesgo que no siempre se está dispuesto a correr.

Por otro lado, podríamos concluir que el suero no es efectivo, cuando en verdad lo fuese (error de Tipo II). Tal conclusión es muy peligrosa, especialmente si hay vidas en juego.

10.21. Resolver el Problema 10.20 si cada grupo consta de 300 enfermos y se curan 225 del *A* y 195 del *B*.

Solución

En este caso las proporciones de curación son 225/300 = 0.750 y 195/300 = 0.650, iguales que en el Problema 10.20. Bajo la hipótesis H_0,

$$\mu_{P1-P2} = 0 \quad \text{y} \quad \sigma_{P1-P2} = \sqrt{pq\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} = \sqrt{(0.70)(0.30)\left(\frac{1}{300} + \frac{1}{300}\right)} = 0.0374$$

donde (225 + 195)/600 = 0.70 se usa como estimación de p. Luego

$$z = \frac{P_1 - P_2}{\sigma_{P1-P2}} = \frac{0.750 - 0.650}{0.0374} = 2.67$$

Como este valor de z es mayor que 2.33, podemos rechazar la hipótesis al nivel de significación 0.01; es decir, concluimos que el suero es efectivo con sólo un 1% de probabilidad de equivocarnos.

Esto enseña la importancia del tamaño de la muestra en la fiabilidad de las decisiones. En muchos casos, sin embargo, puede no ser factible aumentar el tamaño. En tal circunstancia, estamos obligados a tomar decisiones sobre la base de la información disponible y arrostrar, por tanto, mayores riesgos de equivocación.

10.22. Un sondeo de 300 votantes del distrito *A* y 200 del *B* dan 56% y 48% respectivamente de votos en favor de un cierto candidato. Al nivel de significación 0.05, contrastar la hipótesis de que (*a*) hay diferencia entre los distritos y (*b*) ese candidato es el preferido en el distrito *A*.

Solución

Sean p_1 y p_2 las proporciones de todos los votantes en los distritos *A* y *B*, respectivamente, que son favorables a ese candidato. Bajo la hipótesis H_0: $p_1 = p_2$, tenemos

$$\mu_{P1-P2} = 0 \quad \text{y} \quad \sigma_{P1-P2} = \sqrt{pq\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} = \sqrt{(0.528)(0.472)\left(\frac{1}{300} + \frac{1}{200}\right)} = 0.0456$$

donde hemos usado como estimaciones para p y q los valores[(0.56)(300) + (0.48)(200)]/500 = 0.528 y (1 − 0.528) = 0.472, respectivamente. Luego

$$z = \frac{P_1 - P_2}{\sigma_{P1-P2}} = \frac{0.560 - 0.480}{0.0456} = 1.75$$

(*a*) Si sólo deseamos averiguar si hay diferencia entre los dos distritos, hemos de decidir entre las hipótesis H_0: $p_1 = p_2$ y H_1: $p_1 \neq p_2$, que implican un test de dos colas. Con él, rechazaríamos H_0 al nivel de significación 0.05 si z cae fuera del intervalo -1.96 a 1.96. Como $z = 1.75$ cae dentro de ese intervalo, no podemos rechazar H_0 a este nivel; esto es, no hay diferencia significativa entre los distritos.

(*b*) Si queremos determinar si el candidato es preferido en el distrito A, debemos decidir entre H_0: $p_1 = p_2$ y H_1: $p_1 > p_2$, lo cual implica un contraste de una cola. Usándolo al nivel de significación 0.05, rechazaremos H_0 si z es mayor que 1.645. Ya que tal es el caso, podemos rechazar H_0 a este nivel y concluir que el candidato es preferido en el distrito A.

CONTRASTES MEDIANTE LA DISTRIBUCION BINOMIAL

10.23. Un profesor propone a sus alumnos 10 cuestiones verdadero-falso. Para comprobar la hipótesis de que los estudiantes contestan al azar, adopta la siguiente regla de decisión:

Si al menos 7 respuestas son acertadas, el estudiante no ha contestado al azar.
Si hay menos de 7 correctas, ha contestado al azar.

Hallar la probabilidad de rechazar la hipótesis cuando sea correcta.

Solución

Sea p la probabilidad de que una cuestión sea acertada correctamente. La probabilidad de lograr X correctas de las 10 es $\binom{10}{X}p^X q^{10-X}$, con $q = 1 - p$. Bajo la hipótesis $p = 0.5$ (o sea, el estudiante responde al azar),

$$\Pr\{7 \text{ o más correctas}\} = \Pr\{7 \text{ correctas}\} + \Pr\{8 \text{ correctas}\} + \Pr\{9 \text{ correctas}\} + \Pr\{10 \text{ correctas}\}$$

$$= \binom{10}{7}\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right)^3 + \binom{10}{8}\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^2 + \binom{10}{9}\left(\frac{1}{2}\right)^9\left(\frac{1}{2}\right) + \binom{10}{10}\left(\frac{1}{2}\right)^{10} = 0.1719$$

Así que la probabilidad de concluir que no contestaban al azar cuando realmente sí lo hacían, es 0.1719. Nótese que esta es la probabilidad de un error de Tipo I.

10.24. En el Problema 10.23, hallar la probabilidad de aceptar la hipótesis $p = 0.5$ cuando en realidad $p = 0.7$

Solución

Bajo la hipótesis $p = 0.7$.

$$\Pr\{\text{menos de 7 correctas}\} = 1 - \Pr\{7 \text{ o más correctas}\} =$$

$$= 1 - \left[\binom{10}{7}(0.7)^7(0.3)^3 + \binom{10}{8}(0.7)^8(0.3)^2 + \binom{10}{9}(0.7)^9(0.3) + \binom{10}{10}(0.3)^{10}\right] =$$

$$= 0.3504$$

10.25. En el Problema 10.23, hallar la probabilidad de aceptar la hipótesis $p = 0.5$ cuando (*a*) $p = 0.6$, (*b*) $p = 0.8$, (*c*) $p = 0.9$, (*d*) $p = 0.4$, (*e*) $p = 0.3$, (*f*) $p = 0.2$ y (*g*) $p = 0.1$.

Solución

(*a*) Si $p = 0.6$,

$$\text{Probabilidad pedida} = 1 - [\Pr\{7 \text{ correctas}\} + \Pr\{8 \text{ correctas}\} + \Pr\{9 \text{ correctas}\} + \Pr\{10 \text{ correctas}\}]$$

$$= 1 - \left[\binom{10}{7}(0.6)^7(0.4)^3 + \binom{10}{8}(0.6)^8(0.4)^2 + \binom{10}{9}(0.6)^9(0.4) + \binom{10}{10}(0.6)^{10}\right] = 0.618$$

Los resultados de las partes (*b*) hasta (*g*) se pueden obtener de manera análoga, y se recogen en la Tabla 10.4, junto con los valores correspondientes a $p = 0.5$ y $p = 0.7$. Nótese que la probabilidad en la Tabla 10.4 se denota por β (probabilidad de un error de Tipo II); la entrada β para $p = 0.5$ viene dada por $\beta = 1 - 0.1719 = 0.828$ (del Prob. 10.23), y para $p = 0.7$ del Problema 10.24.

Tabla 10.4

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
β	1.000	0.999	0.989	0.945	0.828	0.618	0.350	0.121	0.013

10.26. Con ayuda del Problema 10.25, construir el gráfico de β versus p, obteniendo así las curvas de operación características de la regla de decisión del Problema 10.23.

Solución

El gráfico requerido es el de la Figura 10.12; obsérvese el parecido con la curva OC del Problema 10.14. Si hubiésemos representado $(1 - \beta)$ versus p, hubiéramos obtenido la *curva de potencia* de la regla de decisión. El gráfico indica que la regla de decisión es *potente* para rechazar $p = 0.5$ cuando realmente $p \leqslant 0.4$ o $p \geqslant 0.8$.

10.27. Una moneda da 6 caras en 6 tiradas. ¿Podemos concluir el nivel de significación (*a*) 0.05 y (*b*) 0.01 que está trucada? Considerar tanto contraste de una como de dos colas.

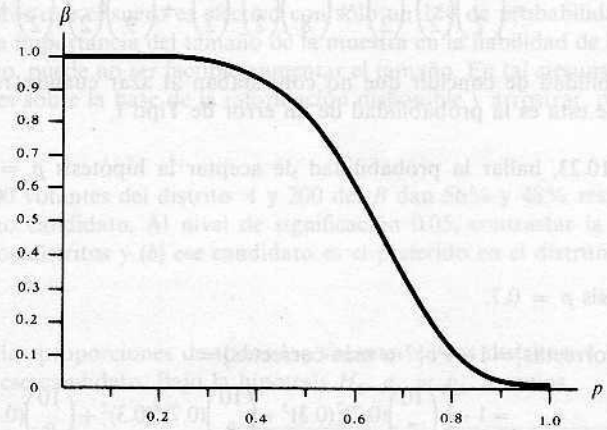


Figura 10.12.

Solución

Sea p la probabilidad de cara en una sola tirada de esa moneda. Bajo la hipótesis H_0: $p = 0.5$ (o sea, la moneda es buena),

$$p(X) = \Pr\{X \text{ caras en 6 tiradas}\} = \binom{6}{X}\left(\frac{1}{2}\right)^X\left(\frac{1}{2}\right)^{6-X} = \binom{6}{X}\left(\frac{1}{64}\right)$$

Así pues, las probabilidades de 0, 1, 2, 3, 4, 5 y 6 caras vienen dadas, respectivamente, por $\frac{1}{64}$, $\frac{6}{64}$, $\frac{15}{64}$, $\frac{20}{64}$, $\frac{15}{64}$, $\frac{6}{64}$ y $\frac{1}{64}$, representadas en la distribución de probabilidad de la Figura 10.13.

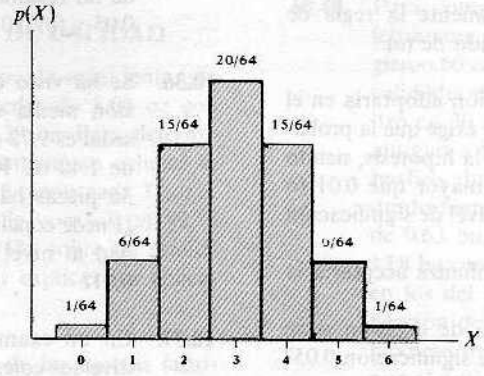


Figura 10.13.

Contraste de una cola

Aquí hay que decidir entre las hipótesis H_0: $p = 0.5$ y H_1: $p > 0.5$. Como Pr{6 caras} $= \frac{1}{64} =$ $= 0.01562$ y Pr{5 ó 6 caras} $= \frac{6}{64} + \frac{1}{64} = 0.1094$, podemos rechazar H_0 al nivel 0.05, pero no al 0.01 (es decir, el resultado observado es significativo al nivel 0.05 pero no al 0.01).

Contraste de dos colas

Ahora hemos de decidir entre H_0: $p = 0.5$ y H_1: $p \neq 0.5$. Ya que Pr{0 ó 6 caras} $= \frac{1}{64} + \frac{1}{64} =$ $= 0.03125$, podemos rechazar H_0 al nivel 0.05 pero no al 0.01.

10.28. Resolver el Problema 10.27 si la moneda diese 5 caras.

Solución

Contraste de una cola

Como Pr{5 ó 6 caras} $= \frac{6}{64} + \frac{1}{64} = \frac{7}{64} = 0.1094$, no podemos rechazar H_0 al nivel 0.05 ni al 0.01.

Contraste de dos colas

Como Pr{0 ó 1 ó 5 ó 6 caras} $= 2(\frac{7}{64}) = 0.2188$, no podemos rechazar H_0 al nivel 0.05 ni al 0.01.

PROBLEMAS SUPLEMENTARIOS

CONTRASTES DE MEDIAS Y PROPORCIONES USANDO LA DISTRIBUCION NORMAL

10.29. Una urna contiene fichas rojas y azules. Para comprobar la hipótesis de que hay tantas de un color como del otro, tomamos una muestra de 64 fichas con reposición y adoptamos la siguiente regla de decisión:

Aceptar la hipótesis si se sacan entre 28 y 36 rojas.

Rechazarla en caso contrario.

(*a*) Hallar la probabilidad de rechazar la hipótesis, siendo ésta verdadera.
(*b*) Representar gráficamente la regla de decisión y el resultado de (*a*).

10.30. (*a*) ¿Qué regla de decisión adoptaría en el Problema 10.29 si se exige que la probabilidad de rechazar la hipótesis, siendo ésta cierta, no sea mayor que 0.01 (o sea, si se desea un nivel de significación 0.01)?
(*b*) ¿A qué nivel de confianza aceptaría la hipótesis?
(*c*) ¿Cuál sería la regla de decisión si se adoptara el nivel de significación 0.05?

10.31. Supongamos que en el Problema 10.29 queremos comprobar la hipótesis de que hay mayor proporción de rojas que de azules.
(*a*) ¿Qué tomaría como hipótesis nula y como hipótesis alternativa?
(*b*) ¿Usaría un contraste de una o de dos colas? ¿Por qué?
(*c*) ¿Qué regla de decisión adoptaría para un nivel de significación de 0.05?
(*d*) ¿Cuál es la regla de decisión si el nivel de significación es 0.01?

10.32. Se tira un par de dados 100 veces y se ve que aparece suma 7 en 23 ocasiones. Contrastar la hipótesis de que los dados son buenos al nivel de significación 0.05 mediante un contraste de (*a*) una cola y (*b*) dos colas. Discutir las razones, si las hay, para preferir uno de ellos.

10.33. Rehacer el Problema 10.32 si el nivel de significación es 0.01.

10.34. Un fabricante afirma que al menos el 95% del equipamiento que ha suministrado a un cliente es acorde a las especificaciones. El examen de una muestra de 200 piezas revela que 18 eran defectuosas. Contrastar su afirmación al nivel de significación (*a*) 0.01 y (*b*) 0.05.

10.35. El porcentaje de grados A en un curso de Física de cierta Universidad en un largo período de tiempo fue del 10%. Durante un curso particular hubo 40 grados A entre 300 estudiantes. Contrastar la significación de tal resultado al nivel de significación (*a*) 0.05 y (*b*) 0.01.

10.36. Se ha visto experimentalmente que la tensión media de ruptura de cierta clase de sedal es 9.72 onzas (oz) con desviación típica de 1.40 oz. Recientemente, una muestra de 36 piezas ha dado una media de 8.93 oz. ¿Puede concluirse que ha empeorado la calidad al nivel de significación (*a*) 0.05 y (*b*) 0.01?

10.37. En un examen de muchos estudiantes de diversos colegios, la nota media ha sido 74.5 con desviación típica de 8.0. En un colegio particular, con 200 estudiantes, la nota media es 75.9. Discutir la significación de tal resultado al nivel de significación 0.05 desde el punto de vista de un contraste de (*a*) una cola y (*b*) de dos colas, explicando cuidadosamente qué conclusiones se desprenden de ellos.

10.38. Resolver el Problema 10.37 al nivel de significación 0.01.

CURVAS DE OPERACION CARACTERISTICAS

10.39. Refiriéndonos al Problema 10.29, hallar la probabilidad de aceptar la hipótesis de que haya igual proporción de rojas y azules cuando la proporción real p de fichas rojas es (*a*) 0.6, (*b*) 0.7, (*c*) 0.8, (*d*) 0.9 y (*e*) 0.3.

10.40. Representar los resultados del Problema 10.39 en un gráfico de (*a*) β versus p y (*b*) $1 - \beta$ versus p. Compararlos con los del Problema 10.12, considerando la analogía de fichas rojas y azules con cara y cruz, respectivamente.

10.41. (*a*) Resolver los Problemas 10.13 y 10.14 si se acuerda tomar una muestra de 400 sogas.
(*b*) ¿Qué conclusión se desprende acerca de los riesgos de error de Tipo II cuando se aumenta el tamaño de la muestra?

10.42. Construir (*a*) una curva OC y (*b*) una curva de potencia, para el Problema 10.31. Compararlas con las del Problema 10.14.

GRAFICOS DE CONTROL DE CALIDAD

10.43. En el pasado, cierto tipo de sedal tenía una tensión de ruptura media de 8.64 oz con desviación típica de 1.28 oz. Para determinar si el producto mantiene su calidad se toma una muestra de 16 piezas cada 3 horas. Registrar los límites de control (*a*) 99.73 (o 3σ), (*b*) 99% y (*c*) 95% sobre un gráfico de control de calidad y explicar sus aplicaciones.

10.44. En promedio, un 3% de las tuercas fabricadas por una empresa son defectuosas. Para mantener esa calidad de producción, se toma una muestra de 200 tuercas cada 4 horas. Determinar los límites de control (*a*) 99% y (*b*) 95% para el número de tuercas defectuosas en cada muestra. Nótese que sólo se necesitan *límites superiores de control* en este caso.

CONTRASTES MEDIANTE DIFERENCIAS DE MEDIAS Y PROPORCIONES

10.45. Una muestra de 100 bombillas de la marca *A* dan vida media de 1190 h y desviación típica de 90 h. Una muestra de 75 bombillas de la marca *B* dan vida media de 1230 h y desviación típica de 120 h. ¿Hay diferencia entre las vidas medias de esas dos marcas de bombillas al nivel de significación (*a*) 0.05 y (*b*) 0.01?

10.46. En el Problema 10.45, contrastar la hipótesis de que las bombillas de la marca *B* son de más calidad que las del *A*, usando nivel de significación (*a*) 0.05 y (*b*) 0.01. Explicar las diferencias entre estos resultados y los citados en la última parte del Problema 10.45. ¿Contradicen estos resultados a los del Problema 10.45?

10.47. En un examen de ortografía, la nota media de 32 niños ha sido 72 con una desviación típica de 8, mientras que la nota media de 36 niñas ha sido 75 con una desviación típica de 6. Contrastar la hipótesis de que al nivel de significación (*a*) 0.05 y (*b*) 0.01, las niñas superan a los niños en ortografía.

10.48. Para comprobar los efectos de un nuevo fertilizante en la producción de trigo, se escogieron 60 campos cuadrados de iguales áreas, calidades de tierra, horas de sol, etc. Se utilizó en 30 de ellos el nuevo fertilizante y el antiguo a los demás. El número medio de bushels (bu) de trigo cosechados por cuadrado fueron 18.2 bu con desviación típica de 0.63 bu, en los del nuevo fertilizante, y 17.8 bu con una desviación típica de 0.54 bu, en los del antiguo. Usando nivel de significación de (*a*) 0.05 y (*b*) 0.01, contrastar la hipótesis de que el nuevo fertilizante es mejor que el antiguo.

10.49. Muestras aleatorias de 200 piezas producidas por una máquina *A* y 100 fabricadas por otra *B* dieron 19 y 5 piezas defectuosas, respectivamente. Contrastar las hipótesis de que (*a*) las dos máquinas tienen distinta calidad de producción y (*b*) la *B* es mejor que la *A*. Usar el nivel de significación 0.05.

10.50. Dos urnas *A* y *B* contienen el mismo número de fichas, pero la proporción de rojas y blancas es desconocida en ambas. Una muestra de 50 fichas tomada con reposición en cada una de ellas dio 32 rojas en la urna *A* y 23 en la *B*. Con el nivel de significación 0.05, contrastar las hipótesis de que (*a*) la proporción de rojas es la misma en las dos urnas y (*b*) *A* tiene mayor proporción de rojas que *B*.

CONTRASTES MEDIANTE LA DISTRIBUCION BINOMIAL

10.51. Con referencia al Problema 10.23, hallar el número mínimo de cuestiones que un estudiante debe contestar correctamente para que el profesor esté seguro con nivel de significación de (*a*) 0.05, (*b*) 0.01, (*c*) 0.001 y (*d*) 0.06 de que no ha sido por azar. Discutir los resultados.

10.52. Construir gráficos similares a los del Problema 10.10 para el Problema 10.24.

10.53. Resolver los Problemas 10.23 al 10.25 cambiando en la regla de decisión el 7 por 8.

10.54. En 8 tiradas una moneda ha dado 7 caras. ¿Podemos rechazar la hipótesis de que la moneda es buena al nivel de significación (*a*) 0.05, (*b*) 0.10 y (*c*) 0.01? Usar un contraste bilateral.

10.55. Repetir el Problema 10.54 con contraste unilateral.

10.56. Repetir el Problema 10.54 si la moneda diera cara las 8 veces.

10.57. Repetir el Problema 10.54 si la moneda diera cara 6 veces.

10.58. Una bolsa contiene un gran número de bolas rojas y blancas. Una muestra de 8 bolas da 6 blancas y 2 rojas. Mediante contrastes y nivel de significación adecuados, discutir la proporción de rojas y blancas en la bolsa.

10.59. Discutir cómo se puede recurrir a la teoría del muestreo para investigar las proporciones de distintos tipos de peces en un lago.

CAPITULO 11

Teoría de pequeñas muestras

PEQUEÑAS MUESTRAS

En capítulos precedentes hemos hecho uso de que para muestras de tamaño $N > 30$, llamadas *grandes muestras*, las distribuciones de muestreo de muchos estadísticos son aproximadamente normales, siendo la aproximación tanto mejor cuanto mayor sea N. Para muestras de tamaño menor que 30, llamadas *pequeñas muestras*, esa aproximación no es buena y empeora al decrecer N, de modo que son precisas ciertas modificaciones.

El estudio de la distribución de muestreo de estadísticos para pequeñas muestras se llama *teoría de pequeñas muestras*. Sin embargo, un nombre más apropiado sería *teoría exacta del muestreo*, pues sus resultados son válidos tanto para pequeñas muestras como para grandes. En ese capítulo analizamos tres distribuciones importantes: la distribución de Student, la distribución ji-cuadrado y la distribución F.

DISTRIBUCION t DE STUDENT

Definamos el estadístico

$$t = \frac{\bar{X} - \mu}{s}\sqrt{N-1} = \frac{\bar{X} - \mu}{\hat{s}/\sqrt{N}} \tag{1}$$

que es análogo al estadístico z dado por

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

(véase pág. 225).

Si consideramos muestras de tamaño N tomadas de una población normal (o casi normal) con media μ y si para cada una calculamos t, usando la media muestral $\bar{X}$ y la desviación típica muestral s o $\hat{s}$, puede obtenerse la distribución de muestreo para t. Esta distribución (véase Figura 11.1) viene dada por

$$Y = \frac{Y_0}{\left(1 + \dfrac{t^2}{N-1}\right)^{N/2}} = \frac{Y_0}{\left(1 + \dfrac{t^2}{\nu}\right)^{(\nu+1)/2}} \tag{2}$$

donde Y_0 es una constante que depende de N tal que el área total bajo la curva es 1, y donde la constante $\nu = (N - 1)$ se llama el *número de grados de libertad* (ν es la letra griega nu). Para una definición de grados de libertad, véase página 255.

La distribución (2) se llama *distribución t de Student* en honor de su descubridor, W. S. Gossett, quien publicó su obra bajo el pseudónimo de «Student» («estudiante») a principios de este siglo.

Para grandes valores de ν o de N (ciertamente $N \geqslant 30$), las curvas (2) se ajustan mucho a la curva normal canónica

$$Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$

como se muestra en la Figura 11.1.

INTERVALOS DE CONFIANZA

Al igual que se hizo con la distribución normal, se pueden definir los intervalos de confianza 95%, 99%, u otros, usando la tabla de la distribución t en el Apéndice III. De esta forma podemos estimar la media de la población dentro de límites especificados.

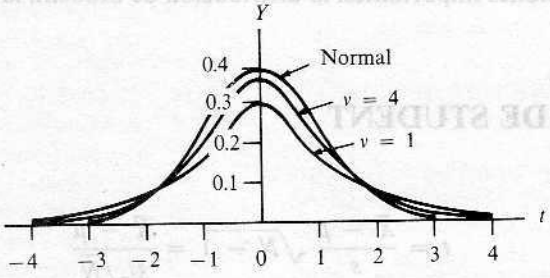


Figura 11.1. Distribución t de Student para varios valores de ν.

Por ejemplo, si $-t_{.975}$ y $t_{.975}$ son los valores de t para los que el 2.5% del área está en cada cola de la distribución t, entonces el intervalo de confianza 95% para t es

$$-t_{.975} < \frac{\bar{X} - \mu}{s}\sqrt{N-1} < t_{.975} \tag{3}$$

de donde vemos que μ se estima que estará en el intervalo

$$\bar{X} - t_{.975}\frac{S}{\sqrt{N-1}} < \mu < \bar{X} + t_{.975}\frac{S}{\sqrt{N-1}} \tag{4}$$

con el 95% de confianza (o sea, probabilidad 0,95).

Nótese que $t_{.975}$ representa el valor 97.5 percentil, mientras que $t_{.025} = -t_{.975}$ representa el valor 2.5 percentil.

En general, podemos representar límites de confianza para medias poblacionales por

$$\bar{X} \pm t_c \frac{s}{\sqrt{N-1}} \tag{5}$$

donde los valores $\pm t_c$, llamados *valores críticos* o *coeficientes de confianza*, dependen del nivel de confianza deseado y del tamaño de la muestra. Pueden verse en el Apéndice III.

Comparando las ecuaciones (5) con los límites de confianza ($\bar{X} \pm z_c\sigma/\sqrt{N}$) del Capítulo 9, página 211, vemos que para pequeñas muestras debemos sustituir z_c (obtenido de la distribución normal) por t_c (obtenido de la distribución de Student) y σ con $\sqrt{N/(N-1)}s = \hat{s}$, que es la estimación muestral de σ. Cuando N crece, ambos métodos tienden a coincidir.

CONTRASTES DE HIPOTESIS Y SIGNIFICACION

Los contrastes de hipótesis y significación o reglas de decisión (discutidos en el Capítulo 10), se extienden fácilmente a pequeñas muestras. La única diferencia consiste en que el *estadístico z* queda sustituido por el *estadístico t*.

1. **Medias.** Para constrastar la hipótesis H_0 de que una población normal tiene medida μ, usamos el estadístico t

$$t = \frac{\bar{X} - \mu}{s}\sqrt{N-1} = \frac{\bar{X} - \mu}{\hat{s}}\sqrt{N} \tag{6}$$

donde $\bar{X}$ es la media de una muestra de tamaño N. Esto es análogo al uso de

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

para grandes N, excepto que se usa $\hat{s} = \sqrt{N/(N-1)}s$ en lugar de σ. La diferencia es que mientras z está normalmente distribuida, t sigue la distribución de Student. Al crecer N, ambas tienden a coincidir.

2. **Diferencias de medias.** Supongamos que se toman dos muestras aleatorias de tamaños N_1 y N_2 de poblaciones normales cuyas desviaciones típicas son iguales ($\sigma_1 = \sigma_2$). Y supongamos además que estas dos muestras tienen medias $\bar{X}_1$ y $\bar{X}_2$ y desviaciones típicas s_1 y s_2, respectivamente. Para contrastar la hipótesis H_0 de que las muestras provienen de la misma población (o sea, $\mu_1 = \mu_2$ y también $\sigma_1 = \sigma_2$),

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{1/N_1 + 1/N_2}} \quad \text{donde} \quad \sigma = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}} \tag{7}$$

Su distribución es una distribución de Student con $\nu = N_1 + N_2 - 2$ grados de libertad. El uso de (7) aparece como plausible si se hace $\sigma_1 = \sigma_2 = \sigma$ en el z de la ecuación (2) del Capítulo 10, y se usa entonces como estimación de σ^2 la media ponderada

$$\frac{(N_1 - 1)\hat{s}_1^2 + (N_2 - 1)\hat{s}_2^2}{(N_1 - 1) + (N_2 - 1)} = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}$$

donde $\hat{s}_1^2$ y $\hat{s}_2^2$ son las estimaciones sin sesgo de σ_1^2 y σ_2^2 (véase Propiedad 3 en la página 95).

DISTRIBUCION JI-CUADRADO

Definamos el estadístico

$$\chi^2 = \frac{Ns^2}{\sigma^2} = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_N - \bar{X})^2}{\sigma^2} \tag{8}$$

donde χ es la letra griega ji y χ^2 se lee «ji-cuadrado».

Si consideramos muestras de tamaño N tomadas de una población normal con desviación típica σ, y si para cada muestra calculamos χ^2, se obtiene para χ^2 una distribución de muestreo, llamada *distribución ji-cuadrado*, que viene dada por

$$Y = Y_0(\chi^2)^{\frac{1}{2}(\nu-2)} e^{-\frac{1}{2}\chi^2} = Y_0 \chi^{\nu-2} e^{-\frac{1}{2}\chi^2} \tag{9}$$

donde $\nu = N - 1$ es el *número de grados de libertad*, e Y_0 es una constante que depende de ν tal que el área total bajo la curva es 1. La distribución ji-cuadrado correspondientes a varios valores ν se muestran en la Figura 11.2. El máximo de Y ocurre en $\chi^2 = \nu - 2$ para $\nu \geq 2$.

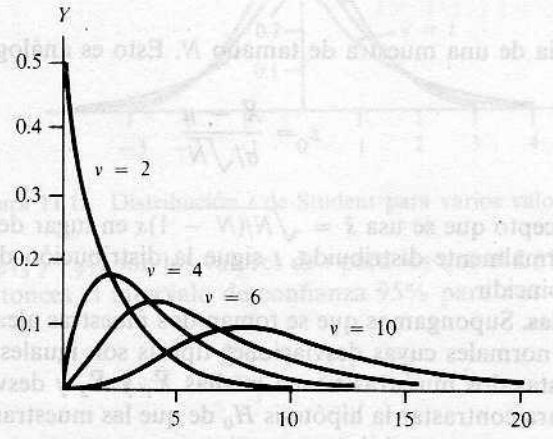


Figura 11.2. Distribuciones ji-cuadrado para varios valores de ν.

INTERVALOS DE CONFIANZA PARA LA DISTRIBUCION JI-CUADRADO

Como se hizo con la distribución normal y con la distribución de Student, podemos definir los intervalos y límites de confianza 95%, 99%, u otros, usando la tabla de la distribución ji-cuadrado

en el Apéndice IV. De este modo podemos estimar, dentro de límites especificados, la desviación típica de la población en términos de una desviación típica muestral s.

Por ejemplo, si $\chi^2_{.025}$ y $\chi^2_{.975}$ son los valores de χ^2 (llamados *valores críticos*) para los que el 2.5% del área está en cada cola de la distribución, entonces el intervalo de confianza 95% es

$$\chi^2_{.025} < \frac{Ns^2}{\sigma^2} < \chi^2_{.975} \tag{10}$$

del cual vemos que σ se estima que estará en el intervalo

$$\frac{s\sqrt{N}}{\chi_{.975}} < \sigma < \frac{s\sqrt{N}}{\chi_{.025}} \tag{11}$$

con el 95% de confianza. Otros intervalos de confianza se hallan de forma parecida. Los valores de $\chi_{.025}$ y $\chi_{.975}$ representan, respectivamente, los valores 2.5 y 97.5 percentil.

El Apéndice IV da los valores percentiles correspondientes al número de grados de libertad ν. Para grandes $\nu(\nu \geq 30)$, podemos utilizar el hecho de que $(\sqrt{2\chi^2} - \sqrt{2\nu - 1})$ está casi normalmente distribuida con media 0 y desviación típica 1; luego se pueden usar tablas de la distribución normal si $\nu \geq 30$. Entonces, si χ^2_p y z_p son los p-ésimos percentiles de la distribución ji-cuadrado y de la distribución normal, respectivamente, tenemos

$$\chi^2_p = \tfrac{1}{2}(z_p + \sqrt{2\nu - 1})^2 \tag{12}$$

En esos casos, hay muy buen acuerdo con los resultados obtenidos en los Capítulos 8 y 9

Para otras aplicaciones de la distribución ji-cuadrado, véase el Capítulo 12.

GRADOS DE LIBERTAD

Para el cálculo de un estadístico tal como (1) u (8), es necesario emplear tanto observaciones de muestras como propiedades de ciertos parámetros de la población. Si estos parámetros son desconocidos, hay que estimarlos a partir de la muestra.

El *número de grados de libertad* de un estadístico, generalmente denotado por ν, se define como el número N de observaciones independientes en la muestra (o sea, el tamaño de la muestra) menos el número k de parámetros de la población, que debe ser estimado a partir de observaciones muestrales. En símbolos, $\nu = N - k$.

En el caso del estadístico (1), el número de observaciones independientes en la muestra es N, de donde podemos calcular $\bar{X}$ y s. Sin embargo, como debemos estimar μ, $k = 1$ y $\nu = N - 1$).

En el caso del estadístico (8), el número de observaciones independientes en la muestra es N, de donde podemos calcular s. Sin embargo, como debemos estimar σ, $k = 1$ y $\nu = N - 1$.

LA DISTRIBUCION F

Como hemos visto, es importante en algunas aplicaciones conocer la distribución de muestreo de la diferencia en medias $(\bar{X}_1 - \bar{X}_2)$ de dos muestras. De la misma manera, podemos necesitar la

distribución de muestreo de la diferencia en varianzas $(S_1^2 - S_2^2)$. Resulta, sin embargo, que esta distribución es complicada, por lo que en lugar de eso, consideramos el estadístico S_1^2/S_2^2, ya que un cociente grande o pequeño indicará una gran diferencia, mientras un cociente cercano a 1 indica una pequeña diferencia. Su distribución de muestreo se llama *distribución F*, en honor de R. A. Fisher.

Más concretamente, sean dos muestras, 1 y 2, de tamaños N_1 y N_2, respectivamente, tomadas de dos poblaciones normales (o casi) con varianzas σ_1^2 y σ_2^2. Definamos el estadístico

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2} = \frac{N_1 S_1^2/(N_1 - 1)\sigma_1^2}{N_2 S_2^2/(N_2 - 1)\sigma_2^2} \tag{13}$$

donde

$$\hat{S}_1^2 = \frac{N_1 S_1^2}{N_1 - 1} \qquad \hat{S}_2^2 = \frac{N_2 S_2^2}{N_2 - 1} \tag{14}$$

(véase pág. 208). Entonces la distribución de muestreo de F se llama distribución F de Fisher, o en breve, distribución F, con $\nu_1 = N_1 - 1$ y $\nu_2 = N_2 - 1$ grados de libertad. Esta distribución viene dada por

$$Y = \frac{CF^{(\nu_1/2)-1}}{(\nu_1 F + \nu_2)^{(\nu_1+\nu_2)/2}} \tag{15}$$

donde C es una constante que depende de ν_1 y ν_2 tal que el área total bajo la curva es 1. La curva tiene una forma del tipo que indica la Figura 11.3, aunque esa forma puede variar considerablemente según los valores de ν_1 y ν_2.

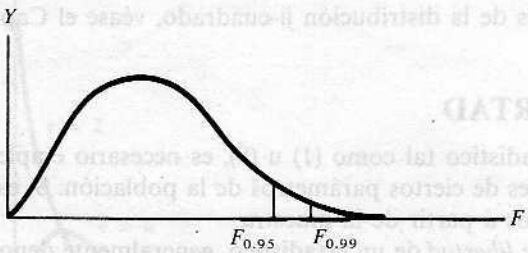


Figura 11.3.

Los Apéndices V y VI dan valores percentiles de F para los que las áreas en la cola de la derecha son 0.05 y 0.01, denotadas $F_{.95}$ y $F_{.99}$, respectivamente. Representando los niveles de significación 5% y 1%, éstos se pueden usar para determinar si la varianza S_1^2 es significativamente mayor que S_2^2, o no. En la práctica, la muestra con mayor varianza se elige como muestra 1.

PROBLEMAS RESUELTOS

DISTRIBUCION t DE STUDENT

11.1. La Figura 11.4 recoge el gráfico de la distribución de Student con 9 grados de libertad. Hallar el valor de t_1 para el que (*a*) el área sombreada de la derecha es 0.05, (*b*) el área total sombreada es 0.05, (*c*) el

área total sin sombrear es 0.99, (*d*) el área en sombra de la izquierda es 0.01 y (*e*) el área a la izquierda de t_1 es 0.90.

Solución

(*a*) Si el área sombreada de la derecha es 0.05, el área a la izquierda de t_1 es $(1 - 0.05) = 0.95$ y t_1 es el 95 percentil, $t_{.95}$. En el Apéndice III, buscamos el 9 en la columna encabezada con v, y después nos desplazamos a la derecha hasta la columna $t_{.95}$; el resultado, 1.83, es el valor pedido de t.

(*b*) Si el área total sombreada es 0.05, la de la derecha es 0.025 por simetría. Luego el área a la izquierda de t_1 es $(1 - 0.025) = 0.975$ y t_1 representa el 97.5 percentil, $t_{.975}$. En el Apéndice III encontramos que el valor requerido de t es 2.26.

(*c*) Si el área total sin sombrear es 0.99, el área en sombra es $(1 - 0.99) = 0.01$, y su mitad derecha es 0.005. En el Apéndice III vemos que $t_{.995} = 3.25$.

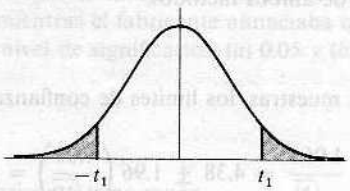


Figura 11.4.

(*d*) Si el área sombreada de la izquierda es 0.01, por simetría la de la derecha es igual. El Apéndice III da $t_{.99} = 2.82$. Luego el valor crítico de t para el cual el área sombreada de la izquierda es 0.01 es -2.82.

(*e*) Si el área a la izquierda de t_1 es 0.90, el t_1 corresponde al 90 percentil $t_{.90}$, que según el Apéndice III es igual a 1.38.

11.2. Hallar los valores críticos de t para los que el área de la cola derecha de la distribución t es 0.05 si el número de grados de libertad, v, es (*a*) 16, (*b*) 27 y (*c*) 200.

Solución

En el Apéndice III, columna $t_{.95}$, hallamos los valores (*a*) 1.75 para $v = 16$; (*b*) 1.70 para $v = 27$; y (*c*) 1.645 para $v = 200$. (Este último es el valor que se obtendría de la curva normal; en el Apéndice III corresponde a la entrada marcada ∞ en la última fila).

11.3. Los coeficientes de confianza 95% (con dos colas) para la distribución normal vienen dados por ± 1.96. ¿Cuáles son los correspondientes coeficientes para la distribución t si (*a*) $v = 9$, (*b*) $v = 20$, (*c*) $v = 30$ y (*d*) $v = 60$?

Solución

Para los coeficientes de confianza 95% (con dos colas), el área total sombreada en la Figura 11.4 ha de ser 0.05. Así que el área de la cola derecha es 0.025 y el correspondiente valor crítico de t es $t_{.975}$. Entonces los coeficientes de confianza pedidos son $\pm t_{.975}$; para los valores dados de v, son (*a*) ± 2.26, (*b*) ± 2.09, (*c*) ± 2.04 y (*d*) ± 2.00.

11.4. Una muestra de 10 medidas del diámetro de una esfera dan una media $\bar{X} = 4.38$ cm y una desviación típica $s = 0.06$ cm. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para el diámetro verdadero.

Solución

(*a*) Los límites de confianza 95% vienen dados por $\bar{X} \pm t_{.975}(s/\sqrt{N-1})$.

Como $\nu = N - 1 = 10 - 1 = 9$, encontramos $t_{.975} = 2.26$ [véase también el Problema 11.3(*a*)]. Entonces, usando $\bar{X} = 4.38$ y $s = 0.06$, los requeridos límites de confianza 95% son $4.38 \pm 2.26(0.06/\sqrt{10-1}) = 4.38 \pm 0.0452$ cm. Luego podemos tener 95% de confianza de que la verdadera media está entre $(4.38 - 0.045) = 4.335$ cm y $(4.38 + 0.045) = 4.425$ cm.

(*b*) Los límites de confianza 99% están dados por $\bar{X} \pm t_{.995}(s/\sqrt{N-1})$.

Para $\nu = 9$, $t_{.995} = 3.25$. Entonces los límites de confianza 99% son $4.38 \pm 3.25(0.06/\sqrt{10-1}) = 4.38 \pm 0.0650$ cm, y el intervalo de confianza 99% es 4.315 a 4.445 cm.

11.5. (*a*) Repetir el Problema 11.4 suponiendo que son válidos los métodos de la teoría de grandes muestras.

(*b*) Comparar los resultados de ambos métodos.

Solución

(*a*) En el método de grandes muestras, los límites de confianza 95% son

$$\bar{X} \pm \frac{1.96\sigma}{\sqrt{N}} = 4.38 \pm 1.96\left(\frac{0.06}{\sqrt{10}}\right) = 4.38 \pm 0.037 \text{ cm}$$

donde se ha usado la desviación típica muestral 0.06 como estimación de σ. Análogamente, los límites de confianza 99% son

$$\bar{X} \pm \frac{2.58\sigma}{\sqrt{N}} = 4.38 \pm 2.58\left(\frac{0.06}{\sqrt{10}}\right) = 4.38 \pm 0.049 \text{ cm}$$

(*b*) En cada caso, los límites de confianza obtenidos usando la teoría exacta (pequeñas muestras) son mayores que los obtenidos por métodos de grandes muestras. Era de esperar, porque la precisión disponible con pequeñas muestras es menor que con muestras grandes.

11.6. Hace tiempo, una máquina producía arandelas de 0.05 pulgadas(in) de espesor. Para determinar si sigue en buen estado, se toma una muestra de 10 arandelas, que dan un espesor medio de 0.053 in con desviación típica de 0.003 in. Contrastar la hipótesis de que la máquina sigue funcionando bien, con nivel de significación (*a*) 0.05 y (*b*) 0.01.

Solución

Queremos decidir entre las hipótesis:

H_0: $\mu = 0.050$, y la máquina sigue en buen estado.

H_1: $\mu \neq 0.050$, y la máquina está deteriorada.

Por tanto, se precisa un contraste de dos colas. Bajo la hipótesis H_0, tenemos

$$t = \frac{\bar{X} - \mu}{s}\sqrt{N-1} = \frac{0.053 - 0.050}{0.003}\sqrt{10-1} = 3.00$$

(*a*) Para un test de dos colas al nivel de significación 0.05, adoptamos la siguiente regla de decisión:

Aceptar H_0 si t está en el intervalo $-t_{.975}$ a $t_{.975}$, que para $10 - 1 = 9$ grados de libertad es desde -2.26 a 2.26.

Rechazarla en caso contrario.

Como $t = 3.00$, rechazamos H_0 al nivel 0.05.

(*b*) Para un test de dos colas al nivel de significación 0.01, adoptamos la siguiente regla de decisión:

Aceptar H_0 si t está en el intervalo $-t_{.995}$ a $t_{.995}$, que para $10 - 1 = 9$ grados de libertad es desde -3.25 a 3.25.

Rechazarla en caso contrario.

Como $t = 3.00$, aceptamos H_0 al nivel 0.01.

Como podemos rechazar H_0 al nivel 0.05 pero no al 0.01, decimos que el resultado de la muestra es *probablemente significativo* (véase final del Problema 10.5). Sería recomendable revisar la máquina o al menos tomar otra muestra.

11.7. Una prueba con 6 sogas de un cierto fabricante dio una tensión media de ruptura de 7750 lb y una desviación típica de 145 lb, mientras el fabricante anunciaba que era de 8000 lb. ¿Puede sostenerse la afirmación del fabricante al nivel de significación (*a*) 0.05 y (*b*) 0.01?

Solución

Hemos de decidir entre:

H_0: $\mu = 8000$ lb, y el fabricante tiene razón.

H_1: $\mu < 8000$ lb, y el fabricante no tiene razón.

Hay que aplicar un contraste de una cola. Bajo la hipótesis H_0, tenemos

$$t = \frac{\bar{X} - \mu}{s}\sqrt{N - 1} = \frac{7750 - 8000}{145}\sqrt{6 - 1} = -3.86$$

(*a*) Para un contraste de una cola al nivel de significación 0.05, adoptamos la siguiente regla de decisión:

Aceptar H_0 si t es mayor que $-t_{.95}$, que para $6 - 1 = 5$ grados de libertad quiere decir $t > -2.01$.

Rechazar H_0 en caso contrario.

Como $t = -3.86$, rechazamos H_0.

(*b*) Para un contraste de una cola al nivel de significación 0.01, adoptamos la siguiente regla de decisión:

Aceptar H_0 si t es mayor que $-t_{.99}$, que para 5 grados de libertad quiere decir $t > -3.36$.

Rechazar H_0 en caso contrario.

Como $t = -3.86$, rechazamos H_0.

Deducimos que es muy improbable que el fabricante tuviese razón.

11.8. Los cocientes de inteligencia(IQ) de 16 estudiantes de un barrio dieron una media de 107 con desviación típica 10, y 14 estudiantes de otro barrio dieron media 112 con desviación típica 8. ¿Hay diferencia significativa entre los IQ de los dos grupos al nivel de significación (*a*) 0.01 y (*b*) 0.05?

Solución

Si μ_1 y μ_2 denotan los IQ medios de la población de ambos barrios, respectivamente, tenemos que decidir entre:

H_0: $\mu_1 = \mu_2$, y no hay diferencia esencial entre los dos barrios.

H_1: $\mu_1 \neq \mu_2$, y hay diferencia significativa entre ellos.

Bajo la hipótesis H_0,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{1/N_1 + 1/N_2}} \qquad \text{donde } \sigma = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}}$$

Luego

$$\sigma = \sqrt{\frac{16(10)^2 + 14(8)^2}{16 + 14 - 2}} = 9.44 \qquad \text{y} \qquad t = \frac{112 - 107}{9.44\sqrt{1/16 + 1/14}} = 1.45$$

(*a*) Con un contraste bilateral al nivel de significación 0.01, rechazaríamos H_0 si t estuviera fuera del rango $-t_{.995}$ a $t_{.995}$, que para $(N_1 + N_2 - 2) = (16 + 14 - 2) = 28$ grados de libertad es el rango -2.76 a 2.76. Así pues, no podemos rechazar H_0 al nivel de significación 0.01.

(*b*) Con un contraste bilateral al nivel de significación 0.05, rechazaríamos H_0 si t estuviera fuera del rango $-t_{.975}$ a $t_{.975}$, que para 28 grados de libertad es el rango -2.05 a 2.05. Así pues, no podemos rechazar H_0 al nivel de significación 0.01.

Concluimos que no hay diferencia significativa entre los dos grupos.

11.9. Con el fin de probar un fertilizante, se tomaron 24 parcelas de la misma área, de las que la mitad se trataron con ese fertilizante y las otras no (el grupo de control); por lo demás, las condiciones fueron idénticas para todas ellas. La producción media de trigo en las parcelas sin tratar fue de 4.8 bushels(bu) con desviación típica de 0.40 bu, y en las tratadas fue 5.1 bu con desviación típica de 0.36 bu. ¿Podemos concluir que se produjo mejora a causa del fertilizante de significación (*a*) 1% y (*b*) 5 %?

Solución

Si μ_1 y μ_2 denotan las producciones medias de trigo de las poblaciones tratada y sin tratar, respectivamente, hemos de decidir entre:

H_0: $\mu_1 = \mu_2$, y la diferencia es fortuita.

H_1: $\mu_1 > \mu_2$, y el fertilizante mejora la cosecha.

Bajo la hipótesis H_0,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{1/N_1 + 1/N_2}} \qquad \text{donde} \qquad \sigma = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}}$$

Así pues $\sigma = \sqrt{\dfrac{12(0.40)^2 + 12(0.36)^2}{12 + 12 - 2}} = 0.397$ y $t = \dfrac{5.1 - 4.8}{0.397\sqrt{1/12 + 1/12}} = 1.85$

(*a*) Con un contraste de una cola al nivel de significación 0.01, rechazaremos H_0 si t es mayor que $t_{.99}$, que para $(N_1 + N_2 - 2) = (12 + 12 - 2) = 22$ grados de libertad es 2.51. Luego no podemos rechazar H_0 al nivel de significación 0.01.

(*b*) Con un contraste de una cola al nivel de significación 0.05, rechazaremos H_0 si t es mayor que $t_{.95}$, que para 22 grados de libertad es 1.72. Luego podemos rechazar H_0 al nivel de significación 0.05.

Concluimos que la mejora causada por el fertilizante es *probablemente significativa*. No obstante, antes de sacar conclusiones definitivas sería deseable una evidencia más nítida.

DISTRIBUCION JI-CUADRADO

11.10. El gráfico de la distribución ji-cuadrado con 5 grados de libertad se muestra en la Figura 11.5. Hallar los valores críticos de χ^2 para los que (*a*) el área sombreada a la derecha es 0.05, (*b*) el área total en sombra es 0.05, (*c*) el área sombreada de la izquierda es 0.10 y (*d*) el área sombreada a la derecha es 0.01.

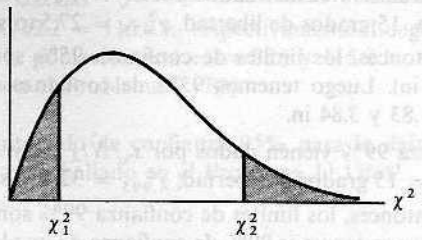


Figura 11.5.

Solución

(*a*) Si el área sombreada de la derecha es 0.05, el área a la izquierda de χ_2^2 es $(1 - 0.05) = 0.95$ y χ_2^2 representa el 95 percentil, χ_{95}^2. Buscando en el Apéndice IV el 5 bajo la columna v, y entonces desplazándonos a la derecha hasta la columna $\chi_{.95}^2$, resulta 11.1, que es el requerido valor crítico de χ^2.

(*b*) Como la distribución no es simétrica, hay muchos valores críticos para los que el área total sombreada es 0.05. Por ejemplo, la de la derecha podría ser 0.04 y la de la izquierda 0.01. Es costumbre, sin embargo, salvo que se especifique lo contrario, escoger ambas iguales. En este caso, cada área será de 0.025.

Si el área sombreada a la derecha es 0.025, el área a la izquierda de χ_2^2 es $1-0.025 = 0.975$ y χ_2^2 representa el 95 percentil, $\chi_{.975}^2$, que por el Apéndice IV es 12.8. Análogamente, si el área sombreada de la izquierda es 0.025, el área a la izquierda de χ_1^2 es 0.025 y χ_1^2 representa el 2.5 percentil, $\chi_{.025}^2$, que es 0.831. Luego los valores críticos son 0.831 y 12.8.

(*c*) Si el área sombreada de la derecha es 0.10, χ_1^2 representa el 10° percentil, $\chi_{.10}^2$, que es 1.61.

(*d*) Si el área sombreada de la derecha es 0.01, el área a la izquierda de χ_2^2 es 0.99 y χ_2^2 representa el 99 percentil, $\chi_{.99}^2$, que es 15.1.

11.11. Hallar los valores críticos de χ^2 para los cuales el área de la cola derecha de la distribución ji-cuadrado es 0.05, siendo el número de grados de libertad, v, igual a (*a*) 15, (*b*) 21 y (*c*) 50.

Solución

Usando el Apéndice IV, se ven en la columna encabezada por $\chi_{.95}^2$ los valores (*a*) 25.0 para $v = 15$, (*b*) 32.7 para $v = 21$ y (*c*) 67.5 para $v = 50$.

11.12. Hallar la mediana de χ^2 correspondiente a (*a*) 9, (*b*) 28 y (*c*) 40 grados de libertad.

Solución

Usando el Apéndice IV, vemos en la columna encabezada por $\chi_{.50}^2$ (ya que la mediana es el 50 percentil) el valor (*a*) 8.34 para $v = 9$; (*b*) 27.3 para $v = 28$; y (*c*) 39.3 para $v = 40$.

Conviene fijarse en que las medianas son casi iguales al número de grados de libertad. De hecho, para $v > 10$, los valores de la mediana son $(v - 0.7)$, como se ve en la tabla.

11.13. La desviación típica de las alturas de 16 estudiantes varones tomados al azar en un colegio de 1000 alumnos es 2.40 in. Hallar los límites de confianza (*a*) 95% y (*b*) 99% de la desviación típica para todos los estudiantes de ese colegio.

Solución

(*a*) Los límites de confianza 95% vienen dados por $s\sqrt{N}/\chi_{.975}$ y $s\sqrt{N}/\chi_{.025}$.

Para $\nu = 16 - 1 = 15$ grados de libertad, $\chi^2_{.975} = 27.5$ (o sea $\chi_{.975} = 5.24$) y $\chi^2_{.025} = 6.26$ (o sea $\chi_{.005} = 2.50$). Entonces, los límites de confianza 95% son $2.40\sqrt{16}/5.24$ y $2.40\sqrt{16}/2.50$ (es decir, 1.83 y 3.84 in). Luego tenemos 95% de confianza de que la desviación típica de la población está entre 1.83 y 3.84 in.

(*b*) Los límites de confianza 99% vienen dados por $s\sqrt{N}/\chi_{.995}$ y $s\sqrt{N}/\chi_{.005}$

Para $\nu = 16 - 1 = 15$ grados de libertad, $\chi^2_{.995} = 32.8$ (o sea $\chi_{.995} = 5.73$) y $\chi^2_{.005} = 4.60$, es decir $\chi_{.025} = 2.14$). Entonces, los límites de confianza 99% son $2.40\sqrt{16}/5.73$ y $2.40\sqrt{16}/2.14$ (es decir, 1.68 y 4.49 in). Luego tenemos 99% de confianza de que la desviación típica de la población está entre 1.68 y 4.49 in.

11.14. Hallar $\chi^2_{.95}$ para (*a*) $\nu = 50$ y (*b*) $\nu = 100$ grados de libertad.

Solución

Para $\nu > 30$ podemos usar el que $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ está casi normalmente distribuida con media 0 y desviación típica 1. Así que si z_p es el valor z percentil de la distribución normal canónica, podemos escribir, con muy buena aproximación,

$$\sqrt{2\chi_p^2} - \sqrt{2\nu - 1} = z_p \quad \text{o sea} \quad \sqrt{2\chi_p^2} = z_p + \sqrt{2\nu - 1}$$

de donde $\chi_p^2 = \frac{1}{2}(z_p + \sqrt{2\nu - 1})^2$.

(*a*) Si $\nu = 50$, $\chi^2_{.95} = \frac{1}{2}(z_{.95} + \sqrt{2(50) - 1})^2 = \frac{1}{2}(1.64 + \sqrt{99})^2 = 67.2$, que está en buen acuerdo con el valor 67.5 dado en el Apéndice IV.

(*b*) Si $\nu = 100$, $\chi^2_{.95} = \frac{1}{2}(z_{.95} + \sqrt{2(100) - 1})^2 = \frac{1}{2}(1.64 + \sqrt{199})^2 = 124.0$ (valor real = 124.3).

11.15. La desviación típica de las vidas medias de una muestra de 200 lámparas es 100 h. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para la desviación típica de todas las lámparas de ese tipo.

Solución

(*a*) Los límites de confianza 95% están dados por $s\sqrt{N}/\chi_{.975}$ y $s\sqrt{N}/\chi_{.025}$.

Para $\nu = 200 - 1 = 199$ grados de libertad, encontramos (como en el Problema 11.14)

$$\chi^2_{.975} = \tfrac{1}{2}(z_{.975} + \sqrt{2(199) - 1})^2 = \tfrac{1}{2}(1.96 + 19.92)^2 = 239$$
$$\chi^2_{.025} = \tfrac{1}{2}(z_{.025} + \sqrt{2(199) - 1})^2 = \tfrac{1}{2}(-1.96 + 19.92)^2 = 161$$

de donde $\chi_{.975} = 15.5$ y $\chi_{.025} = 12.7$. Entonces los límites de confianza 95% son $100\sqrt{200}/15.5 = 91.2$ h y $100\sqrt{200}/12.7 = 111.3$ h, respectivamente. Luego estamos 95% confiados de que la desviación típica de la población está entre 91.2 y 111.3 h.

Comparar esto con el Problema 9.17(*a*).

(*b*) Los límites de confianza 99% están dados por $s\sqrt{N}/\chi_{.995}$ y $s\sqrt{N}/\chi_{.005}$.

Para $\nu = 200 - 1 = 199$ grados de libertad,

$$\chi^2_{.995} = \tfrac{1}{2}(z_{.995} + \sqrt{2(199) - 1})^2 = \tfrac{1}{2}(2.58 + 19.92)^2 = 253$$
$$\chi^2_{.005} = \tfrac{1}{2}(z_{.005} + \sqrt{2(199) - 1})^2 = \tfrac{1}{2}(-2.58 + 19.92)^2 = 150$$

de donde $\chi_{.995} = 15.9$ y $\chi_{.005} = 12.2$. Entonces los límites de confianza 99% son $100\sqrt{200}/15.9 = 88.9$ h y $100\sqrt{200}/12.2 = 115.9$ h, respectivamente. Luego estamos 99% confiados de que la desviación típica de la población está entre 88.9 y 115.9 h.

Comparar esto con el Problema 9.17(*b*).

11.16. ¿Es posible obtener un intervalo de confianza 95% para la desviación típica de la población cuya anchura sea menor que la del hallado en el Problema 11.15(*a*)?

Solución

Los límites de confianza para la desviación típica de la población hallados en el Problema 11.15(*a*) se obtuvieron escogiendo valores críticos de χ^2 tales que el área en cada cola era 2.5%. Es posible hallar otros límites de confianza eligiendo valores críticos de χ^2 para los que la suma de las áreas en las dos colas sea 5%, pero con áreas desiguales en las colas.

En la Tabla 11.1 se han recogido varios de tales valores críticos (obtenidos por los métodos del Problema 11.14), y los correspondientes intervalos de confianza 95%. De ahí vemos que un intervalo 95% con anchura de sólo 19.8 es el que va desde 91.0 a 110.8. Se puede lograr otro con menor anchura todavía continuando de esa forma, usando valores críticos como $\chi_{.031}$ y $\chi_{.981}$, $\chi_{.032}$ y $\chi_{.982}$, etc. En general, sin embargo, el decrecimiento que se consigue en el intervalo es despreciable y no merece la pena el trabajo exigido.

Tabla 11.1

Valores críticos	Intervalo de confianza del 95%	Anchura
$\chi_{.01} = 12.44$, $\chi_{.96} = 15.32$	92.3 a 113.7	21.4
$\chi_{.02} = 12.64$, $\chi_{.97} = 15.42$	91.7 a 111.9	20.2
$\chi_{.03} = 12.76$, $\chi_{.98} = 15.54$	91.0 a 110.8	19.8
$\chi_{.04} = 12.85$, $\chi_{.99} = 15.73$	89.9 a 110.0	20.1

11.17. Tiempo atrás, la desviación típica de los pesos de ciertos envases llenados por una máquina era 0.25 onzas(oz). Una muestra aleatoria de 20 envases ha dado una desviación típica de 0.32 oz. ¿Es significativo el aparente aumento en la variabilidad al nivel de significación (*a*) 0.05 y (*b*) 0.01?

Solución

Hemos de decidir entre las hipótesis:

H_0: $\sigma = 0.25$ oz, y el resultado observado es fortuito.

H_1: $\sigma > 0.25$ oz, y la variabilidad ha aumentado realmente.

El valor de χ^2 para la muestra es

$$\chi^2 = \frac{Ns^2}{\sigma^2} = \frac{20(0.32)^2}{(0.25)^2} = 32.8$$

(*a*) Usando un contraste unilateral, rechazaríamos H_0 al nivel de significación 0.05 si el valor de χ^2 para que la muestra fuese mayor que $\chi^2_{.95}$, que es igualar a 30.1 para $\nu = 20 - 1 = 19$ grados de libertad. Así pues, rechazaríamos H_0 al nivel de significación 0.05.

(*b*) Usando un contraste unilateral, rechazaríamos H_0 al nivel de significación 0.01 si el valor de χ^2 para la muestra fuese mayor que $\chi^2_{.99}$, que es igual a 36.2 para 19 grados de libertad. Así pues, no rechazaríamos H_0 al nivel de significación 0.01.

Concluimos que la variabilidad ha crecido probablemente. Debiera hacerse una revisión de esa máquina.

DISTRIBUCION F

11.18. Dos muestras de tamaños 9 y 12 se han tomado en dos poblaciones normalmente distribuidas con varianzas respectivas 16 y 25. Si las varianzas muestrales son 20 y 8, determinar si la primera muestra tiene una varianza significativamente mayor que la segunda al nivel de significación (*a*) 0.05 y (*b*) 0.01.

Solución

Para las dos muestras, 1 y 2, tenemos $N_1 = 9$, $N_2 = 12$, $\sigma_1^2 = 16$, $\sigma_2^2 = 25$, $S_1^2 = 20$ y $S_2^2 = 8$. Luego

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2} = \frac{N_1 S_1^2/(N_1 - 1)\sigma_1^2}{N_2 S_2^2/(N_2 - 1)\sigma_2^2} = \frac{(9)(20)/(9 - 1)(16)}{(12)(8)/(12 - 1)(25)} = 4.03$$

(*a*) Los grados de libertad para el numerador y el denominador de F son $\nu_1 = N_1 - 1 = 9 - 1 = 8$ y $\nu_2 = N_2 - 1 = 12 - 1 = 11$. Entonces del Apéndice V vemos que $F_{.95} = 2.95$. Como la $F = 4.03$ calculada es mayor que 2.95, concluimos que la varianza de la muestra 1 es significativamente mayor que la de la muestra 2 al nivel de significación 0.05.

(*b*) Para $\nu_1 = 8$ y $\nu_2 = 11$, hallamos en el Apéndice VI que $F_{.01} = 4.74$. Luego no podemos concluir que la muestra 1 tenga varianza mayor que la muestra 2 al nivel de significación 0.01.

11.19. Se toman dos muestras de tamaños 8 y 10 de dos poblaciones normalmente distribuidas con varianzas respectivas 20 y 36. Hallar la probabilidad de que la varianza de la primera sea doble que la de la segunda.

Solución

Tenemos $N_1 = 8$, $N_2 = 10$, $\sigma_1^2 = 20$ y $\sigma_2^2 = 36$. Por tanto,

$$F = \frac{8S_1^2/(7)(20)}{10S_2^2/(9)(36)} = 1.85\,\frac{S_1^2}{S_2^2}$$

El número de grados de libertad para el numerador y el denominador son $\nu_1 = N_1 - 1 = 8 - 1 = 7$ y $\nu_2 = N_2 - 1 = 10 - 1 = 9$. Ahora bien, si S_1^2 es más del doble que S_2^2, entonces

$$F = 1.85\,\frac{S_1^2}{S_2^2} > (1.85)(2) = 3.70$$

Buscando 3.70 en los Apéndices V y VI, hallamos que la probabilidad es menor que 0.05 pero mayor que 0.01. Valores más precisos requieren una tabulación más exhaustiva de la distribución F.

PROBLEMAS SUPLEMENTARIOS

DISTRIBUCION t DE STUDENT

11.20. Para una distribución de Student con 15 grados de libertad, hallar el valor de t_1 tal que (*a*) el área a su derecha sea 0.01, (*b*) el área a su izquierda sea 0.95, (*c*) el área a su derecha sea 0.10, (*d*) la suma de áreas a la derecha de t_1 y a la izquierda de $-t_1$ sea 0.01 y (*e*) el área entre $-t_1$ y t_1 sea 0.95.

11.21. Hallar los valores críticos de t para los que el área de la cola derecha de la distribución t es 0.01 si el número de grados de libertad ν, es igual a (*a*) 4, (*b*) 12, (*c*) 25, (*d*) 60 y (*e*) 150.

11.22. Hallar los valores de t_1 para la distribución de Student que satisfacen cada una de las condiciones siguientes:

(*a*) El área entre $-t_1$ y t_1 es 0.90 y $\nu = 25$.
(*b*) El área a la izquierda de $-t_1$ es 0.025 y $\nu = 20$.
(*c*) La suma de áreas a la derecha de t_1 y a la izquierda de $-t_1$ es 0.01 y $\nu = 5$.
(d) El área a la derecha de t_1 es 0.55 y $\nu = 16$.

11.23. Si una variable U tiene una distribución de Student con $\nu = 10$, hallar la constante C tal que (*a*) $\Pr\{U > C\} = 0.05$, (*b*) $\Pr\{-C \leq U \leq C\} = 0.98$, (*c*) $\Pr\{U \leq C\} = 0.20$ y (*d*) $\Pr\{U \geq C\} = 0.90$.

11.24. Los coeficientes de confianza 99% (con dos colas) para la distribución normal vienen dados por $\pm$ 2.58. ¿Cuáles son los correspondientes coeficientes para la distribución t de Student si (*a*) $\nu = 4$, (*b*) $\nu = 12$, (*c*) $\nu = 25$, (*d*) $\nu = 30$ y (*e*) $\nu = 40$?

11.25. Una muestra de 12 medidas de la tensión de ruptura de hilos de algodón da una media de 7.38 gramos (g) y una desviación típica de 1.24 g. Hallar los limites de confianza (*a*) 95% y (*b*) 99% para la verdadera tensión de ruptura.

11.26. Repetir el Problema 11.25 en el supuesto de que los métodos de grandes muestras fuesen aplicables, y comparar los resultados obtenidos.

11.27. Cinco medidas del tiempo de reacción de un individuo ante cierto estímulo se han registrado como 0.28, 0.30, 0.27, 0.33 y 0.31 segundos. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para el tiempo real de reacción.

11.28. La vida media de las lámparas producidas por una empresa era, en tiempos, de 1120 h con desviación típica de 125 h. Una muestra reciente de 8 lámparas da una vida media de 1070 h. Contrastar la hipótesis de que la vida media de esas lámparas no ha cambiado, con nivel de significación (*a*) 0.05 y (*b*) 0.01.

11.29. En el Problema 11.28, contrastar la hipótesis $\mu = 1120$ h frente a la hipótesis alternativa $\mu < 1120$ h, usando nivel de significación (*a*) 0.05 y (*b*) 0.01.

11.30. Las especificaciones para la fabricación de cierta aleación exigen un 23.2% de cobre. Una muestra de 10 análisis del producto ha revelado un contenido medio de cobre del 23.5% con desviación típica de 0.24%. ¿Podemos concluir que el producto cumple las especificaciones al nivel de significación (*a*) 0.01 y (*b*) 0.05?

11.31. En el Problema 11.30, contrastar la hipótesis de que el contenido medio de cobre es mayor de lo especificado, usando nivel de significación (*a*) 0.01 y (*b*) 0.05.

11.32. Un técnico sostiene que introduciendo un nuevo tipo de maquinaria en un proceso de producción se puede disminuir sustancialmente el tiempo requerido en la producción. A causa del alto costo de mantenimiento, el empresario piensa que salvo que se reduzca ese tiempo en al menos un 8%, no vale la pena tal inversión. Seis experiencias arrojan una disminución media del

tiempo de producción del 8.4% con desviación típica de 0.32%. Con nivel de significación (*a*) 0.01 y (*b*) 0.05, contrastar la hipótesis de que el proceso merece ser renovado.

11.33. Con gasolina de la marca *A*, el número medio de millas por galón que recorren 5 automóviles similares en igualdad de condiciones es 22.6 con desviación típica 0.48. Con gasolina de otra marca *B*, el resultado es 21.4 con desviación típica 0.54. Usando un nivel de significación 0.05, investigar si la marca *A* es de mejor calidad que la *B*.

11.34. Dos tipos de soluciones químicas, *A* y *B*, han sido probadas para ver su pH (grado de acidez de la solución). El análisis de 6 muestras de *A* arroja un pH medio 7.52 con desviación típica 0.024, mientras que 5 muestras de *B* dan un pH medio 7.49 con desviación típica 0.032. Usando el nivel de significación 0.05, determinar si los dos tipos de soluciones tienen distinto pH.

11.35. En un examen de psicología, 12 estudiantes de una clase obtuvieron media de 78 con desviación típica 6, y 15 de otra clase consiguieron media de 74 con desviación típica 8. Mediante un nivel de significación 0.05, determinar si el primer grupo es superior al segundo.

DISTRIBUCION JI-CUADRADO

11.36. Para una distribución ji-cuadrado con 12 grados de libertad, hallar el valor de χ_c^2 tal que (*a*) el área a la derecha de χ_c^2 es 0.05, (*b*) el área a la izquierda de χ_c^2 es 0.99 y (*c*) el área a la derecha de χ_c^2 es 0.025.

11.37. Hallar los valores críticos de χ^2 para los cuales el área de la cola derecha de la distribución ji-cuadrado es 0.05 si el número de grados de libertad, ν, es igual (*a*) 8, (*b*) 19, (*c*) 28 y (*d*) 40.

11.38. Repetir el Problema 11.37 si el área de la cola de la derecha es 0.01.

11.39. (*a*) Hallar χ_1^2 y χ_2^2 tales que el área bajo la distribución ji-cuadrado correspondiente a $\nu = 20$ entre χ_1^2 y χ_2^2 es 0.95, suponiendo áreas iguales a la derecha de χ_2^2 y a la izquierda de χ_1^2.

(*b*) Probar que si la suposición de áreas iguales en (*a*) se omite, los valores χ_1^2 y χ_2^2 no son únicos.

11.40. Si la variable U tiene una distribución ji-cuadrado con $\nu = 7$, hallar χ_1^2 y χ_2^2 tales que (*a*) $\Pr\{U > \chi_2^2\} = 0.025$, (*b*) $\Pr\{U < \chi_1^2\} = 0.50$, (*c*) $\Pr\{\chi_1^2 \leq U \leq \chi_2^2\} = 0.90$.

11.41. La desviación típica de las vidas medias de 10 bombillas es 120 h. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para la desviación típica de las bombillas de esa clase.

11.42. Rehacer el Problema 11.41 si 25 bombillas diesen esa misma desviación típica de 120 h.

11.43. Hallar (*a*) $\chi_{.05}^2$ y (*b*) $\chi_{.95}^2$ para $\nu = 150$.

11.44. Hallar (*a*) $\chi_{.025}^2$ y (*b*) $\chi_{.975}^2$ para $\nu = 250$.

11.45. Probar que para grandes valores de ν, una buena aproximación de χ^2 viene dada por $(\nu + z_p\sqrt{2\nu})$, donde z_p es el p-ésimo percentil de la distribución normal canónica.

11.46. Resolver el Problema 11.39 usando la distribución ji-cuadrado si una muestra de 100 bombillas da la misma desviación típica de 120 h. Comparar los resultados con los obtenidos por los métodos del Capítulo 9.

11.47. ¿Cuál es el intervalo de confianza 95% del Problema 11.44 que tiene anchura mínima?

11.48. La desviación típica de las tensiones de ruptura de ciertos cables producidos por una empresa es 240 lb. Tras un cambio en el proceso de producción, una muestra de 8 cables dio una desviación típica de 300 lb. Investigar si es significativo ese crecimiento en variabilidad, usando nivel de significación (*a*) 0.05, y (*b*) 0.01.

11.49. La desviación típica de las temperaturas anuales en una ciudad a lo largo de 100 años es 16 °F. Usando la temperatura media del día 15 de cada mes durante los últimos 15 años, ha resultado una desviación

típica de 10 °F. Contrastar la hipótesis de que las temperaturas en esa ciudad son menos variables que en el pasado, con nivel de significación (*a*) 0.05 y (*b*) 0.01.

DISTRIBUCION F

11.50. Hallar los valores de F en cada caso:

(*a*) $F_{.95}$ con $\nu_1 = 8$ y $\nu_2 = 10$
(*b*) $F_{.99}$ con $\nu_1 = 24$ y $\nu_2 = 11$
(*c*) $F_{.95}$ con $N_1 = 16$ y $N_2 = 25$
(*d*) $F_{.99}$ con $N_1 = 21$ y $N_2 = 23$

11.51. Calcular $F_{.95}$ con $\nu_1 = 22$ y $\nu_2 = 27$.

11.52. En dos poblaciones normalmente distribuidas con varianzas 40 y 60, se toman muestras respectivas de tamaño 10 y 15. Si las varianzas muestrales son 90 y 50, determinar si la muestra 1 tiene varianza significativamente mayor que la muestra 2, al nivel de significación (*a*) 0.05 y (*b*) 0.01.

11.53. Dos empresas *A* y *B* producen lámparas eléctricas, cuyas vidas medias están muy normalmente distribuidas, con desviaciones típicas de 20 y 27 h, respectivamente. Si seleccionamos 16 lámparas de *A* y 20 de *B* y las desviaciones típicas de sus vidas medias resultan ser 15 y 40 h respectivamente, ¿podemos concluir a los niveles de significación (*a*) 0.05 y (*b*) 0.01 que la variabilidad de las de *A* es significativamente menor que la de las de *B*?

CAPITULO 12

Test ji-cuadrado

FRECUENCIAS OBSERVADAS Y TEORICAS

Como ya hemos visto repetidamente, los resultados obtenidos por muestreo no siempre coinciden exactamente con los esperados teóricamente de acuerdo con las leyes de las probabilidades. Por ejemplo, aunque consideraciones teóricas conducen a esperar 50 caras y 50 cruces en 100 tiradas de una moneda (buena), es raro que ocurra eso exactamente.

Supongamos que en una muestra particular un conjunto de sucesos posibles $E_1, E_2, E_3, \ldots, E_k$ (véase Tabla 12.1) se observa que ocurren con frecuencias $o_1, o_2, o_3, \ldots, o_k$, llamadas *frecuencias observadas*, y que según las leyes de las probabilidades, se espera que sucedan con frecuencias $e_1, e_2, e_3, \ldots, e_k$, llamadas *frecuencias esperadas* o *teóricas*.

Tabla 12.1

Suceso	E_1	E_2	E_3	$\cdots$	E_k
Frecuencia observada	o_1	o_2	o_3	$\cdots$	o_k
Frecuencia esperada	e_1	e_2	e_3	$\cdots$	e_k

A menudo deseamos saber si las frecuencias observadas difieren significativamente de las esperadas. Para el caso en que sólo son posibles dos sucesos E_1 y E_2 (llamado a veces una *dicotomía* o *clasificación dicotómica*), como es el caso de cara o cruz, piezas defectuosas o no, etc., el problema se resuelve satisfactoriamente por los métodos de los anteriores capítulos. En este capítulo consideramos el problema general.

DEFINICION DE χ^2

Una medida de la discrepancia existente entre las frecuencias observadas y esperadas viene proporcionada por el estadístico χ^2 (léase ji-cuadrado) dado por

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \cdots + \frac{(o_k - e_k)^2}{e_k} = \sum_{j=1}^{k} \frac{(o_j - e_j)^2}{e_j} \qquad (1)$$

donde si la frecuencia total es N,

$$\sum o_j = \sum e_j = N \tag{2}$$

Una expresión equivalente a la fórmula (1) es (véase Prob. 12.11)

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N \tag{3}$$

Si $\chi^2 = 0$, las frecuencias observadas y teóricas coinciden completamente; mientras que si $\chi^2 > 0$, no coinciden exactamente. A valores más grandes de χ^2, mayor discrepancia entre las frecuencias observadas y esperadas.

La distribución muestral de χ^2 se aproxima muy bien por la distribución ji-cuadrado

$$Y = Y_0(\chi^2)^{\frac{1}{2}(\nu-2)} e^{-\frac{1}{2}\chi^2} = Y_0 \chi^{\nu-2} e^{-\frac{1}{2}\chi^2} \tag{4}$$

(ya considerada en el Capítulo 11) si las frecuencias esperadas son al menos iguales a 5, y mejora para valores más grandes.

El número de grados de libertad, ν, viene dado por

1. $\nu = k - 1$ si las frecuencias esperadas se pueden calcular sin tener que estimar los parámetros de la población a partir de estadísticos muestrales. Nótese que hemos restado 1 de k a causa de la ligadura (2), que establece que si conocemos $k - 1$ de las frecuencias esperadas, la restante puede determinarse ya.
2. $\nu = k - 1 - m$ si las frecuencias esperadas se pueden calcular sólo estimando m parámetros de la población a partir de estadísticos de la muestra.

CONTRASTES DE SIGNIFICACION

En la práctica, las frecuencias esperadas se calculan sobre la base de una hipótesis H_0. Si bajo tal hipótesis el valor calculado para χ^2 dado por (1) o (3) es mayor que algún valor crítico (tal como $\chi^2_{.95}$ o $\chi^2_{.99}$, que son los valores críticos de los niveles de significación 0.05 y 0.01 respectivamente), debemos concluir que las frecuencias observadas difieren *significativamente* de las frecuencias esperadas y rechazaremos H_0 al correspondiente nivel de significación; en caso contrario, la aceptaremos (o al menos no la rechazaremos). Este procedimiento se llama el *test o contraste ji-cuadrado* de hipótesis o significación.

Hay que hacer constar que debe mirarse con suspicacia en circunstancias en las que χ^2 sea *demasiado próximo a cero*, pues es raro que las frecuencias observadas coincidan *demasiado bien* con las frecuencias esperadas. Para examinar tales situaciones, podemos determinar si el valor calculado de χ^2 es menor que $\chi^2_{.05}$ o $\chi^2_{.01}$, en cuyo caso hablaremos de decidir que el acuerdo es *demasiado bueno* al nivel de significación 0.05 ó 0.01, respectivamente.

EL TEST JI-CUADRADO PARA LA BONDAD DE AJUSTE

El test ji-cuadrado puede utilizarse para determinar la calidad del ajuste mediante distribuciones teóricas (como la distribución normal o la distribución binomial) de distribuciones empiricas (o sea, las obtenidas de los datos de la muestra). Véanse Problemas 12.12 y 12.13.

TABLAS DE CONTINGENCIA

La Tabla 12.1, en la que las frecuencias observadas ocupan una sola fila, se llama una *tabla de clasificación de entrada única*. Como el número de columnas es k, también se le llama una tabla $1 \times k$ (leído «1 por k»). Extendiendo estas ideas, podemos llegar a *tablas de doble entrada*, o *tablas* $h \times k$, en las que las frecuencias observadas ocupan h filas y k columnas. Tales tablas se suelen llamar *tablas de contingencia*.

Correspondiendo a cada frecuencia observada en una tabla de contingencia $h \times k$, hay una *frecuencia esperada* (o *teórica*) que se calcula sujeta a ciertas hipótesis de acuerdo con las leyes de las probabilidades. Estas frecuencias, que ocupan las celdas de una tabla de contingencia, se llaman *frecuencias de celda*. La frecuencia total en cada fila o en cada columna se llama la *frecuencia marginal*.

Para investigar el acuerdo entre las frecuencias observadas y las frecuencias esperadas, calculamos el estadístico

$$\chi^2 = \sum_j \frac{(o_j - e_j)^2}{e_j} \tag{5}$$

donde la suma se toma sobre todas las celdas de una tabla de contingencia y donde los símbolos o_j y e_j representan, respectivamente, las frecuencias observadas y frecuencias esperadas de la j-ésima celda. Esta suma, análoga a la ecuación (1), contiene hk términos. La suma de todas las frecuencias observadas se denota por N y es igual a la suma de todas las frecuencias esperadas [comparar con la ecuación (2)].

Como antes, el estadístico (5) tiene una distribución muestral dada muy aproximadamente por (4), supuesto que las frecuencias esperadas no sean demasiado pequeñas. El número de grados de libertad, ν, de esta distribución ji-cuadrado viene dado por $h > 1$ y $k > 1$ por:

1. $\nu = (h - 1)(k - 1)$ si las frecuencias esperadas se pueden calcular sin recurrir a estimaciones muestrales de los parámetros de la población. Para una demostración de esto, véase el Problema 12.18.
2. $\nu = (h - 1)(k - 1) - m$ si las frecuencias esperadas sólo se pueden calcular mediante estimación de m parámetros de la población a partir de estadísticos de la muestra.

Los contrastes de significación para las tablas $h \times k$ son similares a los de las tablas $1 \times k$. Las frecuencias esperadas se hallan sujetas a una hipótesis particular H_0. Una hipótesis común es suponer que las dos clasificaciones son mutuamente independientes.

Las tablas de contingencia se pueden generalizar a más dimensiones. Así, por ejemplo, podemos tener tablas $h \times k \times l$, donde están presentes tres clasificaciones.

CORRECCION DE YATES A LA CONTINUIDAD

Cuando se aplican resultados de distribuciones continuas a datos discretos, pueden hacerse ciertas correcciones a la continuidad, como se ha visto en capítulos precedentes. Una corrección similar existe cuando se usa la distribución ji-cuadrado. La corrección consiste en reformular la ecuación (1) como

$$\chi^2 \text{ (corregido)} = \frac{(|o_1 - e_1| - 0.5)^2}{e_1} + \frac{(|o_2 - e_2| - 0.5)^2}{e_2} + \cdots + \frac{(|o_k - e_k| - 0.5)^2}{e_k} \tag{6}$$

y se llama *corrección de Yates*. Una modificación análoga existe para (5).

En general, la corrección se hace sólo cuando el número de grados de libertad es $\nu = 1$. Para grandes muestras, esto da prácticamente los mismos resultados que el χ^2 sin corregir, pero pueden surgir dificultades cerca de los valores críticos (véase Prob. 12.8). Para pequeñas muestras donde cada frecuencia esperada está entre 5 y 10, es quizás mejor comparar ambos valores de χ^2, corregido y sin corregir. Si ambos llevan a la misma conclusión acerca de la hipótesis, tal como el rechazo al nivel de significación 0.05, rara vez surgen dificultades. Si conducen a diferente conclusión, uno debe pensar en aumentar el tamaño de la muestra o, si ello no es factible, en emplear métodos de probabilidad que involucren la *distribución multinomial* del Capítulo 6.

FORMULAS SIMPLES PARA CALCULAR

Existen fórmulas sencillas para calcular χ^2 que implican tan sólo las frecuencias observadas. Lo que sigue da los resultados para tablas de contingencia 2×2 y 2×3 (véanse Tablas 12.2 y 12.3, respectivamente).

Tablas 2×2

$$\chi^2 = \frac{N(a_1b_2 - a_2b_1)^2}{(a_1 + b_1)(a_2 + b_2)(a_1 + a_2)(b_1 + b_2)} = \frac{N\Delta^2}{N_1N_2N_AN_B} \tag{7}$$

Tabla 12.2

	I	II	Total
A	a_1	a_2	N_A
B	b_1	b_2	N_B
Total	N_1	N_2	N

Tabla 12.3

	I	II	III	Total
A	a_1	a_2	a_3	N_A
B	b_1	b_2	b_3	N_B
Total	N_1	N_2	N_3	N

donde $\Delta = a_1b_2 - a_2b_1$, $N = a_1 + a_2 + b_1 + b_2$, $N_1 = a_1 + b_1$, $N_2 = a_2 + a_2$, $N_A = a_1 + a_2$, y $N_B = b_1 + b_2$ (véase Prob. 12.19). Con corrección de Yates esto se convierte en

$$\chi^2 \text{ (corregido)} = \frac{N(|a_1b_2 - a_2b_1| - \frac{1}{2}N)^2}{(a_1 + b_1)(a_2 + b_2)(a_1 + a_2)(b_1 + b_2)} = \frac{N(|\Delta| - \frac{1}{2}N)^2}{N_1N_2N_AN_B} \tag{8}$$

Tablas 2×3

$$\chi^2 = \frac{N}{N_A}\left[\frac{a_1^2}{N_1} + \frac{a_2^2}{N_2} + \frac{a_3^2}{N_3}\right] + \frac{N}{N_B}\left[\frac{b_1^2}{N_1} + \frac{b_2^2}{N_2} + \frac{b_3^2}{N_3}\right] - N \tag{9}$$

donde hemos usado el resultado general válido para todas las tablas de contingencia (véase Problema 12.43):

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N \tag{10}$$

El resultado (9) para tablas de contingencia $2 \times k$, con $k > 3$, admite generalización (véase Problema 12.46).

COEFICIENTE DE CONTINGENCIA

Una medida del grado de interrelación, asociación o dependencia de las clasificaciones en una tabla de contingencia viene dada por

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \tag{11}$$

que se llama el *coeficiente de contingencia*. Cuanto mayor es C, mayor es el grado de asociación. El número de filas y de columnas en la tabla de contingencia determina el máximo valor de C, que nunca es mayor que 1. Si el número de filas y columnas de una tabla de contingencia es igual a k, el máximo valor de C está dado por $\sqrt{(k-1)/k}$ (véanse Problemas 12.22, 12.52 y 12.53).

CORRELACION DE ATRIBUTOS

Ya que las clasificaciones en una tabla de contingencia describen a menudo características de individuos u objetos, se les conoce como *atributos*, y el grado de dependencia, asociación o interrelación se llama la *correlación de atributos*. Para tablas $k \times k$, definimos

$$r = \sqrt{\frac{\chi^2}{N(k-1)}} \tag{12}$$

como el coeficiente de contingencia entre atributos (o clasificaciones). Este coeficiente está entre 0 y 1 (véase Prob. 12.24). Para tablas 2×2 en las que $k = 2$, la *correlación se llama tetracórica*.

El problema general de correlación de variables numéricas se considera en el Capítulo 14.

PROPIEDAD ADITIVA DE χ^2

Supongamos que los resultados de experimentos repetidos dan valores muestrales de χ^2 dados por $\chi_1^2, \chi_2^2, \chi_3^2, \ldots$ con $\nu_1, \nu_2, \nu_3, \ldots$ grados de libertad, respectivamente. Entonces el resultado de todos esos experimentos puede considerarse equivalente a un valor de χ^2 dado por $\chi_1^2 + \chi_2^2 + \chi_3^2 + \cdots$ con $\nu_1 + \nu_2 + \nu_3 + \cdots$ grados de libertad (véase Prob. 12.25).

PROBLEMAS RESUELTOS

EL TEST JI-CUADRADO

12.1. En 200 tiradas de una moneda, han salido 115 caras y 85 cruces. Contrastar la hipótesis de que la moneda es buena, con nivel de significación (*a*) 0.05 y (*b*) 0.01.

Solución

Las frecuencias observadas de caras y cruces son $o_1 = 115$ y $o_2 = 85$, respectivamente, y las frecuencias esperadas (si la moneda es buena) son $e_1 = 100$ y $e_2 = 100$, respectivamente. Entonces

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} = \frac{(115 - 100)^2}{100} + \frac{(85 - 100)^2}{100} = 4.50$$

Como el número de categorías, o clases (caras, cruces) es $k = 2$, $\nu = k - 1 = 2 - 1 = 1$.

(*a*) El valor crítico $\chi^2_{.95}$ para 1 grado de libertad es 3.84. Así pues, como $4.50 > 3.84$, rechazamos la hipótesis de que la moneda es buena al nivel de significación 0.05.

(*b*) El valor crítico $\chi^2_{.99}$ para 1 grado de libertad es 6.63. Así pues, como $4.50 < 6.63$, no podemos rechazar la hipótesis de que la moneda es buena al nivel de significación 0.01.

Concluimos que los resultados observados son *probablemente significativos* y que la moneda es *probablemente falsa*. Para comparar este método con los usados previamente, véase el Problema 12.3.

12.2. Rehacer el Problema 12.1 usando la corrección de Yates.

Solución

$$\chi^2(\text{corregido}) = \frac{(|o_1 - e_1| - 0.5)^2}{e_1} + \frac{(|o_2 - e_2| - 0.5)^2}{e_2} = \frac{(|105 - 100| - 0.5)^2}{100} + \frac{(|85 - 100| - 0.5)^2}{100}$$

$$= \frac{(14.5)^2}{100} + \frac{(14.5)^2}{100} = 4.205$$

Como $4.205 > 3.84$ y $4.205 < 6.63$, las conclusiones alcanzadas en el Problema 12.1 son válidas. Para comparar con métodos previos, ver el Problema 12.3.

12.3. Resolver el Problema 12.1 usando la aproximación normal a la distribución binomial.

Solución

Bajo la hipótesis de que la moneda es buena, la media y la desviación típica del número de caras esperadas en 200 tiradas son $\mu = Np = (200)(0.5) = 100$ y $\sigma = \sqrt{Npq} = \sqrt{(200)(0.5)(0.5)} = 7.07$, respectivamente.

Primer método

$$115 \text{ caras en unidades estándar} = \frac{115 - 100}{7.07} = 2.12$$

Usando el nivel de significación 0.05 y un contraste de dos colas, rechazaríamos la hipótesis de que la moneda es buena si z cae fuera de intervalo -1.96 a 1.96. Con nivel de significación 0.01, el

intervalo correspondiente sería de -2.58 a 2.58. Se sigue que (como en el Problema 12.1) podemos rechazarla al nivel 0.05 pero no al 0.01.

Nótese que el cuadrado del recuento estándar anterior $(2.12)^2 = 4.50$, es lo mismo que el valor de χ^2 obtenido en el Problema 12.1. Este es siempre el caso para un test ji-cuadrado que involucre dos categorías (véase Prob. 12.10).

Segundo método

Usando corrección de continuidad, 115 o más caras es equivalente a 114.5 o más caras. Pero 114.5 en unidades estándar $= (114.5 - 100)/7.07 = 2.05$. Eso lleva a las mismas conclusiones que el primer método.

Nótese que el cuadrado de ese valor estándar es $(2.05)^2 = 4.20$, que coincide con el valor de χ^2 corregido por continuidad con la corrección de Yates del Problema 12.2. Esto sucede siempre para un test ji-cuadrado que implique a dos categorías a las que se ha aplicado la corrección de Yates.

12.4. La Tabla 12.4 muestra las frecuencias observadas y las frecuencias esperadas al lanzar un dado 120 veces. Contrastar la hipótesis de que el dado es bueno, con un nivel de significación de 0.05.

Tabla 12.4

Cara del dado	1	2	3	4	5	6
Frecuencia observada	25	17	15	23	24	16
Frecuencia esperada	20	20	20	20	20	20

Solución

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \frac{(o_3 - e_3)^2}{e_3} + \frac{(o_4 - e_4)^2}{e_4} + \frac{(o_5 - e_5)^2}{e_5} + \frac{(o_6 - e_6)^2}{e_6}$$

$$= \frac{25 - 20)^2}{20} + \frac{(17 - 20)^2}{20} + \frac{(15 - 20)^2}{20} + \frac{(23 - 20)^2}{20} + \frac{(24 - 20)^2}{20} + \frac{(16 - 20)^2}{20} = 5.00$$

Como el número de categorías, o clases (caras 1, 2, 3, 4, 5 y 6), es $k = 6$, $\nu = k - 1 = 6 - 1 = 5$. El valor crítico $\chi^2_{.95}$ para 5 grados de libertad es 11.1. Así que $5.00 < 11.1$ y no podemos rechazar la hipótesis de que el dado es bueno.

Para 5 grados de libertad, $\chi^2_{.05} = 1.15$, así que $\chi^2 = 5.00 > 1.15$. Se deduce que el acuerdo no es excepcionalmente bueno, y debemos mirarlo con recelo.

12.5. La Tabla 12.5 recoge la distribución de los dígitos 0, 1, 2, ..., 9 en una tabla de números aleatorios de 250 dígitos. ¿Difiere la distribución observada de la esperada de forma significativa?

Tabla 12.5

Dígito	0	1	2	3	4	5	6	7	8	9
Frecuencia observada	17	31	29	18	14	20	35	30	20	36
Frecuencia esperada	25	25	25	25	25	25	25	25	25	25

Solución

$$\chi^2 = \frac{(17-25)^2}{25} + \frac{(31-25)^2}{25} + \frac{(29-25)^2}{25} + \frac{(18-25)^2}{25} + \cdots + \frac{(36-25)^2}{25} = 23.3$$

El valor $\chi^2_{.99}$ para $v = k - 1 = 9$ grados de libertad es 21.7 y $23.3 > 21.7$. Por tanto, concluimos que la distribución observada difiere significativamente de la esperada al nivel de significación 0.01. Luego dicha tabla de números aleatorios merece cierto recelo.

12.6. En su experimento con guisantes. Gregor Mendel observó que 315 eran redondos y amarillos, 108 redondos y verdes, 101 rugosos y amarillos y 32 rugosos y verdes. De acuerdo con su teoría de la herencia, esos números debían estar en la proporción 9:3:3:1. ¿Hay alguna evidencia para dudar de su teoría al nivel de significación (*a*) 0.01 y (*b*) 0.05?

Solución

El número total de guisantes es $315 + 108 + 101 + 32 = 556$. Como los números esperados están en la proporción 9:3:3:1 (y $9 + 3 + 3 + 1 = 16$), esperaríamos

$\frac{9}{16}(556) = 312.75$ lisos y amarillos $\quad \frac{3}{16}(556) = 104.25$ rugosos y amarillos

$\frac{3}{16}(556) = 104.25$ lisos y verdes $\quad \frac{1}{16}(556) = 34.75$ rugosos y verdes

$$\text{Luego } \chi^2 = \frac{(315-312.75)^2}{312.75} + \frac{(108-104.25)^2}{104.25} + \frac{(101-104.25)^2}{104.25} + \frac{(32-34.75)^2}{34.75} = 0.470$$

Como hay 4 categorías, $k = 4$ y el número de grados de libertad es $v = 4 - 1 = 3$.

(*a*) Para $v = 3$, $\chi^2_{.99} = 11.3$, y, por tanto, no podemos rechazar la teoría al nivel 0.01.
(*b*) Para $v = 3$, $\chi^2_{.95} = 7.81$, y, por tanto, no podemos rechazar al nivel 0.05.

Concluimos que teoría y experimentos están en buen acuerdo.

Nótese que para 3 grados de libertad, $\chi^2_{.05} = 0.352$ y $\chi^2 = 0.470 > 0.352$. Así pues, aunque el acuerdo es bueno, los resultados obtenidos están sujetos a un error de muestreo razonable.

12.7. Una urna contiene un gran número de fichas de 4 colores diferentes: rojo, naranja, amarillo y verde. Una muestra de 12 fichas ha dado 2 rojas, 5 naranjas, 4 amarillas y 1 verde. Contrastar la hipótesis de que la urna contiene iguales proporciones de los cuatro colores.

Solución

Bajo la hipótesis de proporciones idénticas, se esperarían 3 fichas de cada color. Como estos números esperados son menores que 5, la aproximación ji-cuadrado será errónea. Para evitar eso, combinamos categorías de modo que el número esperado en cada una sea al menos 5.

Si deseamos rechazar la hipótesis, debemos combinarlas de manera tal que la evidencia en contra de la hipótesis sea más nítida. Ello se logra en nuestro caso considerando las categorías «rojo o verde» y «naranja o amarillo», para las cuales la muestra daba 3 y 9 fichas, respectivamente. Como el número esperado en cada categoría bajo la hipótesis de proporciones iguales es 6, tenemos

$$\chi^2 = \frac{(3-6)^2}{6} + \frac{(9-6)^2}{6} = 3$$

Para $v = 2 - 1 = 1$, $\chi^2_{.95} = 3.84$. Luego no podemos rechazarla al nivel de significación 0.05 (aunque sí al 0.01). Cabe concebir que los resultados observados pudieran ser fruto del azar, aunque haya igual proporción presente de cada color.

Otro método

Con corrección de Yates se obtiene

$$\chi^2 = \frac{(|3 - 6| - 0.5)^2}{6} + \frac{(|9 - 6| - 0.5)^2}{6} = \frac{(2.5)^2}{6} + \frac{(2.5)^2}{6} = 2.1$$

que conduce a las mismas conclusiones que antes. Era de esperar, claro está, pues la corrección de Yates siempre *reduce* el valor de χ^2.

Hay que hacer notar que si se hubiera usado la aproximación χ^2 a pesar de que las frecuencias son demasiado pequeñas, se hubiera obtenido

$$\chi^2 = \frac{(2 - 3)^2}{3} + \frac{(5 - 3)^2}{3} + \frac{(4 - 3)^2}{3} + \frac{(1 - 3)^2}{3} = 3.33$$

Como para $\nu = 4 - 1 = 3$, $\chi^2_{.95} = 7{,}81$, llegaríamos a la misma conclusión de antes. Desgraciadamente, la aproximación χ^2 para pequeñas frecuencias es pobre; por tanto, cuando no sea aconsejable combinar frecuencias, debemos recurrir a los métodos exactos de probabilidad del Capítulo 6.

12.8. En 360 tiradas de un par de dados, han salido 74 sietes y 24 onces. Con nivel de significación 0.05, contrastar la hipótesis de que los dados son buenos.

Solución

Un par de dados puede caer de 36 formas. Un 7 ocurre de 6 formas y un 11 en 2 formas. Luego $\Pr\{\text{siete}\} = \frac{6}{36} = \frac{1}{6}$ y $\Pr\{\text{once}\} = \frac{2}{36} = \frac{1}{18}$. Por tanto, en 360 tiradas esperaríamos $360/6 = 60$ sietes y $360/18 = 20$ onces, de modo que

$$\chi^2 = \frac{(74 - 60)^2}{60} + \frac{(24 - 20)^2}{20} = 4.07$$

Para $\nu = 2 - 1 = 1$, $\chi^2_{.95} = 3.84$. Luego, como $4.07 > 3.84$, estaríamos inclinados a rechazar la hipótesis de que los dados son buenos. Usando la corrección de Yates, sin embargo, encontramos

$$\chi^2(\text{corregido}) = \frac{(|74 - 60| - 0.5)^2}{60} + \frac{(|24 - 20| - 0.5)^2}{20} = \frac{(13.5)^2}{60} + \frac{(3.5)^2}{20} = 3.65$$

Así que sobre la base del χ^2 corregido no podemos rechazarla al nivel de significación 0.05.

En general, para grandes muestras como las de este ejemplo, los resultados usando la corrección de Yates son más fiables. No obstante, como incluso los valores corregidos de χ^2 están tan cerca del valor crítico, dudamos en tomar decisiones en un sentido u otro. En tales casos es quizás mejor aumentar el tamaño de la muestra si estamos interesados especialmente en el nivel de significación 0.05 por alguna razón; de otro modo, podríamos rechazar la hipótesis a algún otro nivel (tal como 0.01) si ello es satisfactorio.

12.9. Un estudio sobre 320 familias con 5 hijos reveló la distribución de la Tabla 12.6. ¿Es consistente el resultado con la hipótesis de que los nacimientos de chicos y chicas son igualmente probables?

Tabla 12.6

Número de chicos y chicas	5 chicos 0 chicas	4 chicos 1 chica	3 chicos 2 chicas	2 chicos 3 chicas	1 chico 4 chicas	0 chicos 5 chicas	Total
Número de familias	18	56	110	88	40	8	320

Solución

Sea p = probabilidad de que nazca un chico y $q = 1 - p$ la de una chica. Entonces, las probabilidades de (5 chicos), (4 chicos y 1 chica), ... (5 chicas) vienen dadas por los términos del desarrollo del binomio

$$(p + q)^5 = p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5$$

Si $p = q = \frac{1}{2}$, tenemos

$$\Pr\{5 \text{ chicos y } 0 \text{ chicas}\} = (\tfrac{1}{2})^5 = \tfrac{1}{32} \qquad \Pr\{2 \text{ chicos y } 3 \text{ chicas}\} = 10(\tfrac{1}{2})^2(\tfrac{1}{2})^3 = \tfrac{10}{32}$$
$$\Pr\{4 \text{ chicos y } 1 \text{ chica}\} = 5(\tfrac{1}{2})^4(\tfrac{1}{2}) = \tfrac{5}{32} \qquad \Pr\{1 \text{ chico y } 4 \text{ chicas}\} = 5(\tfrac{1}{2})(\tfrac{1}{2})^4 = \tfrac{5}{32}$$
$$\Pr\{3 \text{ chicos y } 2 \text{ chicas}\} = 10(\tfrac{1}{2})^3(\tfrac{1}{2})^2 = \tfrac{10}{32} \qquad \Pr\{0 \text{ chicos y } 5 \text{ chicas}\} = (\tfrac{1}{2})^5 = \tfrac{1}{32}$$

Así que el número esperado de familias con 5, 4, 3, 2, 1 y 0 chicos se obtiene multiplicando las probabilidades anteriores por 320, y los resultados son 10, 50, 100, 100, 50 y 10, respectivamente. Por tanto,

$$\chi^2 = \frac{(18-10)^2}{10} + \frac{(56-50)^2}{50} + \frac{(110-100)^2}{100} + \frac{(88-100)^2}{100} + \frac{(40-50)^2}{50} + \frac{(8-10)^2}{10} = 12.0$$

Como $\chi^2_{.95} = 11.1$ y $\chi^2_{.99} = 15.1$ para $\nu = 6 - 1 = 5$ grados de libertad, podemos rechazar la hipótesis al nivel de significación 0.05 pero no al 0.01. Así pues, concluimos que los resultados son probablemente significativos, y los nacimientos de chicos y chicas no son equiprobables.

12.10. Probar que un test ji-cuadrado con sólo dos categorías es equivalente al contraste de significación para proporciones (o sea, el test 2) de la página 226.

Solución

Si P es la proporción muestral para la categoría I, p la proporción de la población y N la frecuencia total, podemos describir la situación por medio de la Tabla 12.7. Entonces, por definición,

$$\chi^2 = \frac{(NP - Np)^2}{Np} + \frac{[N(1-P) - N(1-p)]^2}{Nq} = \frac{N^2(P-p)^2}{Np} + \frac{N^2(P-p)^2}{Nq}$$
$$= N(P-p)^2\left(\frac{1}{p} + \frac{1}{q}\right) = \frac{N(P-p)^2}{pq} = \frac{(P-p)^2}{pq/N}$$

que es el cuadrado del estadístico z de la página 226.

Tabla 12.7

	I	II	Total
Frecuencia observada	NP	$N(1-P)$	N
Frecuencia esperada	Np	$N(1-p) = Nq$	N

12.11. (*a*) Probar que la fórmula (1) de este capítulo se puede escribir

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N$$

(*b*) Usar el resultado de la parte (*a*) para verificar el valor de χ^2 calculado en el Problema 12.6.

Solución

(*a*) Por definición,

$$\chi^2 = \sum \frac{(o_j - e_j)^2}{e_j} = \sum \left(\frac{o_j^2 - 2o_j e_j + e_j^2}{e_j} \right)$$

$$= \sum \frac{o_j^2}{e_j} - 2\sum o_j + \sum e_j = \sum \frac{o_j^2}{e_j} - 2N + N = \sum \frac{o_j^2}{e_j} - N$$

donde se ha usado la fórmula (2) de este capítulo.

(*b*) $$\chi^2 = \sum \frac{o_j^2}{e_j} - N = \frac{(315)^2}{312.75} + \frac{(108)^2}{104.25} + \frac{(101)^2}{104.25} + \frac{(32)^2}{34.75} - 556 = 0.470$$

BONDAD DEL AJUSTE

12.12. Usar el test ji-cuadrado para determinar la bondad del ajuste de los datos de la Tabla 7.4 del Problema 7.31.

Solución

$$\chi^2 = \frac{(38-33.2)^2}{33.2} + \frac{(144-161.9)^2}{161.9} + \frac{(342-316.2)^2}{316.2} + \frac{(287-308.7)^2}{308.7} + \frac{(164-150.7)^2}{150.7} + \frac{(25-29.4)^2}{29.4}$$
$$= 7.54$$

Como el número de parámetros utilizados en la estimación de las frecuencias esperadas es $m = 1$ (a saber, el parámetro p de la distribución binomial), $\nu = k - 1 - m = 6 - 1 - 1 = 4$.

Para $\nu = 4$, $\chi^2_{.95} = 9.49$. Así que el ajuste de los datos es bueno.

Para $\nu = 4$, $\chi^2_{.05} = 0.711$. Así pues, como $\chi^2 = 7.54 > 0.711$. El acuerdo no es tan extremadamente bueno como para ser increíble.

12.13. Determinar la bondad del ajuste de los datos en la Tabla 7.6 del Problema 7.33.

Solución

$$\chi^2 = \frac{(5 - 4.13)^2}{4.13} + \frac{(18 - 20.68)^2}{20.68} + \frac{(42 - 38.92)^2}{38.92} + \frac{(27 - 27.71)^2}{27.71} + \frac{(8 - 7.43)^2}{7.43} = 0.959$$

Como el número de parámetros utilizados en la estimación de las frecuencias esperadas es $m = 2$ (a saber la media μ y la desviación σ de la distribución normal), $\nu = k - 1 - m = 5 - 1 - 2 = 2$.

Para $\nu = 2$, $\chi^2_{.95} = 5.99$. Luego concluimos que el ajuste es muy bueno.

Para $\nu = 2$, $\chi^2_{.05} = 0.103$. Así pues, como $\chi^2 = 0.959 > 0.103$, el ajuste no es «demasiado bueno».

TABLA DE CONTINGENCIA

12.14. Resolver el Problema 10.20 usando el test ji-cuadrado.

Solución

Las condiciones del Problema se presentan en la Tabla 12.8(*a*). Bajo la hipótesis nula H_0 de que el suero no tiene efecto, esperaríamos 70 personas curadas en cada grupo, como indica la Tabla 12.8(*b*). Nótese que H_0 equivale a decir que la recuperación es independiente del uso del suero (o sea, las clasificaciones son independientes).

Tabla 12.8(*a*). Frecuencias observadas

	Curados	No curados	Total
Grupo A (usando suero)	75	25	100
Grupo B (sin suero)	65	35	100
Total	140	60	200

Tabla 12.8(*b*). Frecuencias esperadas bajo H_0

	Curados	No curados	Total
Grupo A (usando suero)	70	30	100
Grupo B (sin suero)	70	30	100
Total	140	60	200

$$\chi^2 = \frac{(75-70)^2}{70} + \frac{(65-70)^2}{70} + \frac{(25-30)^2}{30} + \frac{(35-30)^2}{30} = 2.38$$

Para determinar el número de grados de libertad, consideremos la Tabla 12.9, que es la misma que la 12.8 excepto que sólo muestra los totales. Es claro que somos libres de colocar sólo un número en cualquiera de las 4 celdas vacías, ya que una vez hecho eso los números en las restantes celdas vacías quedan fijados por los totales indicados. Luego hay 1 grado de libertad.

Tabla 12.9

	Curados	No curados	Total
Grupo A			100
Grupo B			100
Total	140	60	200

Otro método

Por la fórmula (véase Problema 12.18), $\nu = (h-1)(k-1) = (2-1)(2-1) = 1$. Como $\chi^2_{.95} = 3.84$ para 1 grado de libertad y como $\chi^2 = 2.38 < 3.84$, concluimos que los resultados *no son significativos* al nivel 0.05. Somos incapaces, en consecuencia, de rechazar H_0 a este nivel, y o bien concluimos que el suero no es efectivo o aplazamos la decisión, a la espera de más observaciones.

Nótese que $\chi^2 = 2.38$ es el cuadrado del z, $z = 1.54$, obtenido en el Problema 10.20. En general, el test ji-cuadrado que involucra proporciones muestrales en una tabla de contingencia 2×2 es equivalente a un contraste de significación de diferencias en proporciones usando la aproximación normal, como en la página 228. (Véase Prob. 12.20).

Hacemos notar también que un contraste de una cola usando χ^2 es equivalente a uno de dos colas usando χ ya que, por ejemplo, $\chi^2 > \chi^2_{.95}$ corresponde a $\chi > \chi_{.95}$ o $\chi < -\chi_{.95}$. Como para tablas de contingencia 2×2, χ^2 es el cuadrado de z, se sigue que χ es lo mismo que z para este caso. Así pues, un rechazo de la hipótesis al nivel 0.05 usando χ^2 equivale a un rechazo en un contraste de dos colas al nivel 0.10 usando z.

12.15. Repetir el Problema 12.14 haciendo la corrección de Yates.

Solución

$$\chi^2(\text{corregido}) = \frac{(|75-70|-0.5)^2}{70} + \frac{(|65-70|-0.5)^2}{70} + \frac{(|25-30|-0.5)^2}{30} + \frac{(|35-30|-0.5)^2}{30} = 1.93$$

Luego las conclusiones del Problema 12.14 son válidas. Lo cual se podía haber visto de golpe recordando que la corrección de Yates siempre decrece el valor de χ^2.

12.16. La Tabla 12.10 muestra los números de estudiantes aprobados y suspendidos por tres profesores: Mr. X, Mr. Y y Mr. Z. Contrastar la hipótesis de que las proporciones de suspendidos por los tres profesores son iguales.

Tabla 12.10. Frecuencias observadas

	Mr. X	Mr. Y	Mr. Z	Total
Aprobados	50	47	56	153
Suspensos	5	14	8	27
Total	55	61	64	180

Solución

Bajo la hipótesis H_0 de que las proporciones de estudiantes suspendidos por los tres profesores son iguales, hubieran suspendido 27/180 = 15% de los estudiantes y aprobado el 85%. En ese caso Mr. X, por ejemplo, hubiera suspendido al 15% de 55 estudiantes y hubiera aprobado al 85% de esos 55. Las frecuencias esperadas bajo H_0 se recogen en la Tabla 12.11. Tenemos pues

$$\chi^2 = \frac{(50-46.75)^2}{46.75} + \frac{(47-51.85)^2}{51.85} + \frac{(56-54.40)^2}{54.40} + \frac{(5-8.25)^2}{8.25} + \frac{(14-9.15)^2}{9.15} + \frac{(8-9.60)^2}{9.60} = 4.84$$

Para determinar el número de grados de libertad, consideremos la Tabla 12.12, que es la misma que las Tablas 12.10 y 12.11 excepto que sólo muestra los totales. Es claro que tenemos la libertad de

sólo un número en una celda vacía de la primera columna y uno en una celda vacía de la segunda o tercera columna, tras lo cual todos los demás números de las otras casillas quedan fijados unívocamente por los totales indicados. Luego hay 2 grados de libertad en este caso.

Tabla 12.11. Frecuencias esperadas bajo H_0

	Mr. X	Mr. Y	Mr. Z	Total
Aprobados	88% de 55 = 46.75	85% de 61 = 51.85	85% de 64 = 54.40	153
Suspensos	15% de 55 = 8.25	15% de 61 = 9.15	15% de 64 = 9.60	27
Total	55	61	64	180

Tabla 12.12

	Mr. X	Mr. Y	Mr. Z	Total
Aprobados				153
Suspensos				27
Total	55	61	64	180

Otro método

Por la fórmula, $\nu = (h - 1)(k - 1) = (2 - 1)(3 - 1) = 2$. Como $\chi^2_{.95} = 5.99$, no podemos rechazar H_0 al nivel 0.05. Nótese, no obstante, que como $\chi^2_{.90} = 4.61$, podemos rechazar H_0 al nivel 0.10 si estamos dispuestos a correr el riesgo de uno entre 10 de equivocarnos.

12.17. Usar la fórmula (9) de este capítulo para calcular el valor de χ^2 para el Problema 12.16.

Solución

Tenemos $a_1 = 50$, $a_2 = 47$, $a_3 = 56$, $b_1 = 5$, $b_2 = 14$, $b_3 = 8$, $N_A = a_1 + a_2 + a_3 = 153$, $N_B = b_1 + b_2 + b_3 = 27$, $N_1 = a_1 + b_1 = 55$, $N_2 = a_2 + b_2 = 61$, $N_3 = a_3 + b_3 = 64$ y $N = N_A + N_B = N_1 + N_2 + N_3 = 180$. Luego

$$\chi^2 = \frac{N}{N_A}\left[\frac{a_1^2}{N_1} + \frac{a_2^2}{N_2} + \frac{a_3^2}{N_3}\right] + \frac{N}{N_B}\left[\frac{b_1^2}{N_1} + \frac{b_2^2}{N_2} + \frac{b_3^2}{N_3}\right] - N$$

$$= \frac{180}{153}\left[\frac{(50)^2}{55} + \frac{(47)^2}{61} + \frac{(56)^2}{64}\right] + \frac{180}{27}\left[\frac{(5)^2}{55} + \frac{(14)^2}{61} + \frac{(8)^2}{64}\right] - 180 = 4.84$$

12.18. Probar que para una tabla de contingencia $h \times k$ el número de grados de libertad es $(h - 1) \times (k - 1)$, donde $h > 1$ y $k > 1$.

Solución

En una tabla con h filas y k columnas, podemos dejar de lado un número en cada columna, porque tales números se pueden recuperar por el conocimiento de los totales de filas y columnas. Se sigue que tenemos la libertad de colocar sólo $(h-1)(k-1)$ números en la tabla, ya que los demás se determinan univocamente. Luego el número de grados de libertad es $(h-1)(k-1)$. Este resultado vale si se conocen los parámetros de la población necesarios para obtener las frecuencias esperadas.

12.19. (*a*) Probar que para la tabla de contingencia recogida en la Tabla 12.13(*a*).

$$\chi^2 = \frac{N(a_1b_2 - a_2b_1)^2}{N_1N_2N_AN_B}$$

(*b*) Ilustrar el resultado de la parte (*a*) con los datos del Problema 12.14.

Tabla 12.13(*a*). Resultados observados

	I	II	Total
A	a_1	a_2	N_A
B	b_1	b_2	N_B
Total	N_1	N_2	N

Tabla 12.13(*b*). Resultados esperados

	I	II	Total
A	N_1N_A/N	N_2N_A/N	N_A
B	N_1N_B/N	N_2N_B/N	N_B
Total	N_1	N_2	N

Solución

(*a*) Como en el Problema 12.14, los resultados esperados bajo una hipótesis nula se muestran en la Tabla 12.13(*b*). Entonces

$$\chi^2 = \frac{(a_1 - N_1N_A/N)^2}{N_1N_A/N} + \frac{(a_2 - N_2N_A/N)^2}{N_2N_A/N} + \frac{(b_1 - N_1N_B/N)^2}{N_1N_B/N} + \frac{(b_2 - N_2N_B/N)^2}{N_2N_B/N}$$

Pero $$a_1 - \frac{N_1N_A}{N} = a_1 - \frac{(a_1+b_1)(a_1+a_2)}{a_1+b_1+a_2+b_2} = \frac{a_1b_2 - a_2b_1}{N}$$

Análogamente, $$a_2 - \frac{N_2N_A}{N} \quad \text{y} \quad b_1 - \frac{N_1N_B}{N} \quad \text{y} \quad b_2 - \frac{N_2N_B}{N}$$

son también iguales a $$\frac{a_1b_2 - a_2b_1}{N}$$

Así que podemos escribir

$$\chi^2 = \frac{N}{N_1N_A}\left(\frac{a_1b_2 - a_2b_1}{N}\right)^2 + \frac{N}{N_2N_A}\left(\frac{a_1b_2 - a_2b_1}{N}\right)^2 + \frac{N}{N_1N_B}\left(\frac{a_1b_2 - a_2b_1}{N}\right)^2 + \frac{N}{N_2N_B}\left(\frac{a_1b_2 - a_2b_1}{N}\right)^2$$

que se simplifica a $$\chi^2 = \frac{N(a_1b_2 - a_2b_1)^2}{N_1N_2N_AN_B}$$

(*b*) En el Problema 12.14, $a_1 = 75$, $a_2 = 25$, $b_1 = 65$, $b_2 = 35$, $N_1 = 140$, $N_2 = 60$, $N_A = 100$, $N_B = 100$ y $N = 200$; entonces, como se ha obtenido antes.

$$\chi^2 = \frac{200[(75)(35) - (25)(65)]^2}{(140)(60)(100)(100)} = 2.38$$

Usando la corrección de Yates, el resultado es el mismo que en el Problema 12.15:

$$\chi^2 \text{ (corregido)} = \frac{N(|a_1b_2 - a_2b_1| - \frac{1}{2}N)^2}{N_1N_2N_AN_B} = \frac{200[|(75)(35) - (25)(65)| - 100]^2}{(140)(60)(100)(100)} = 1.93$$

12.20. Probar que un test ji-cuadrado que implique a dos proporciones muestrales es equivalente a un contraste de significación de diferencias en proporciones mediante la aproximación normal (véase página 228).

Solución

Sean P_1 y P_2 dos proporciones muestrales, y sea p la proporción de la población. Con referencia al Problema 12.19, se tiene

$$P_1 = \frac{a_1}{N_1} \quad P_2 = \frac{a_2}{N_2} \quad 1 - P_1 = \frac{b_2}{N_1} \quad 1 - P_2 = \frac{b_2}{N_2} \tag{13}$$

y

$$p = \frac{N_A}{N} \quad 1 - p = q = \frac{N_B}{N} \tag{14}$$

Por tanto,

$$a_1 = N_1P_1 \quad a_2 = N_2P_2 \quad b_1 = N_1(1 - P_1) \quad b_2 = N_2(1 - P_2) \tag{15}$$

y

$$N_A = Np \quad N_B = N_q \tag{16}$$

Usando las ecuaciones (15) y (16), del Problema 12.19 deducimos

$$\chi^2 = \frac{N(a_1b_2 - a_2b_1)^2}{N_1N_2N_AN_B} = \frac{N[N_1P_1N_2(1 - P_2) - N_2P_2N_1(1 - P_1)]^2}{N_1N_2NpNq}$$

$$= \frac{N_1N_2(P_1 - P_2)^2}{Npq} = \frac{(P_1 - P_2)^2}{pq(1/N_1 + 1/N_2)} \quad (\text{porque } N = N_1 + N_2)$$

que es el cuadrado del estadístico z dado en la página 228.

TABLA DE CONTINGENCIA

12.21. Hallar el coeficiente de contingencia para los datos de la tabla de contingencia del Problema 12.14.

Solución

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{2.38}{2.38 + 200}} = \sqrt{0.01176} = 0.1084$$

12.22. Hallar el máximo valor de C para la tabla 2 × 2 del Problema 12.14.

Solución

El máximo de C ocurre cuando las dos clasificaciones son perfectamente dependientes o asociadas. En tal caso, todos los que toman el suero se recuperan y todos los que no lo toman siguen enfermos. La tabla de contingencia aparece en la Tabla 12.14.

Tabla 12.14

	Curados	No curados	Total
Grupo A (usando suero)	100	0	100
Grupo B (sin suero)	0	100	100
Total	100	100	200

Como las frecuencias esperadas de celda, supuesta completa independencia, son todas 50,

$$\chi^2 = \frac{(100-50)^2}{50} + \frac{(0-50)^2}{50} + \frac{(0-50)^2}{50} + \frac{(100-50)^2}{50} = 200$$

Así que el máximo de C es $\sqrt{\chi^2/(\chi^2+N)} = \sqrt{200/(200+200)} = 0.7071$.

En general, para dependencia perfecta en una tabla de contingencia donde los números de filas y columnas son ambos k, las únicas frecuencias de celda no nulas se producen en la diagonal desde la esquina superior izquierda hasta la inferior derecha. Para tales casos, $C_{\text{máx}} = \sqrt{(k-1)/k}$. (Véase Problemas 12.52 y 12.53.)

CORRELACION DE ATRIBUTOS

12.23. Para la Tabla 12.8 del Problema 12.14, hallar el coeficiente de contingencia (*a*) sin y (*b*) con la corrección de Yates.

Solución

(*a*) Como $\chi^2 = 2.38$, $N = 200$, y $k = 2$, se tiene

$$r = \sqrt{\frac{\chi^2}{N(k-1)}} = \sqrt{\frac{2.38}{200}} = 0.1091$$

lo que indica poca correlación entre recuperación y uso del suero.

(*b*) Por el Problema 12.15, r (corregido) $= \sqrt{1.93/200} = 0.0982$.

12.24. Probar que el coeficiente de contingencia para tablas de contingencia, como se definió en la ecuación (12) de este capítulo, está entre 0 y 1.

Solución

Por el problema 12.53, el máximo valor de $\sqrt{\chi^2/(\chi^2+N)}$ es $\sqrt{(k-1/k}$. Luego

$$\frac{\chi^2}{\chi^2+N} \leqslant \frac{k-1}{k} \quad k\chi^2 \leqslant (k-1)(\chi^2+N) \quad k\chi^2 \leqslant k\chi^2 - \chi^2 + kN - N$$

$$\chi^2 \leqslant (k-1)N \quad \frac{\chi^2}{N(k-1)} \leqslant 1 \quad \text{y} \quad r = \sqrt{\frac{\chi^2}{N(k-1)}} \leqslant 1$$

Puesto que $\chi^2 \geqslant 0$, $r \geqslant 0$. Así que, $0 \leqslant r \leqslant 1$, como deseábamos probar.

PROPIEDAD ADITIVA DE χ^2

12.25. Para contrastar una hipótesis se ha realizado tres veces un experimento. Los valores resultantes de χ^2 son 2.37, 2.86 y 3.54, cada uno de los cuales corresponde a un grado de libertad. Probar que mientras H_0 no se puede rechazar al nivel 0.05 sobre la base de uno sólo de esos experimentos, sea cual sea, sí se puede rechazar cuando se combinan los tres.

Solución

Los valores de χ^2 obtenidos al combinar los tres experimentos es, de acuerdo con la *propiedad aditiva*, $\chi^2 = 2.37 + 2.86 + 3.54 = 8.77$ con $1 + 1 + 1 = 3$ grados de libertad. Como $\chi^2_{.95}$ para 3 grados de libertad es 7.81, podemos rechazar H_0 al nivel de significación 0.05. Pero como $\chi^2_{.95} = 3.84$ para 1 grado de libertad, no se puede rechazarla sobre la base de un solo experimento.

Al combinar experimentos en los que se obtienen valores de χ^2 correspondientes a 1 grado de libertad, la corrección de Yates se omite debido a que tiene tendencia a corregir en exceso.

PROBLEMAS SUPLEMENTARIOS

EL TEST JI-CUADRADO

12.26. En 60 lanzamientos de una moneda han salido 37 caras y 23 cruces. Usando nivel de significación (*a*) 0.05 y (*b*) 0.01, contrastar la hipótesis de que la moneda es buena.

12.27. Repetir el Problema 12.26 usando la corrección de Yates.

12.28. En un largo período de tiempo, los grados dados por un grupo de profesores en un curso particular han dado como promedio 12% Aes, 18% Bes, 40% Ces, 18% Des y 12% Efes. Un nuevo profesor da 22 Aes, 34 Bes, 66 Ces, 16 Des y 12 Efes en dos semestres. Determinar al nivel de significación 0.05 si el profesor nuevo sigue la norma de grados de los otros.

12.29. Se lanzan tres monedas 240 veces con el número de caras que recoge, junto con los resultados esperados bajo la hipótesis de que las monedas son buenas, la Tabla 12.15. Contrastar la hipótesis al nivel de significación.

Tabla 12.15

	Fr. observada	F. esperada
Caras 0	24	30
Caras 1	108	90
Caras 2	95	90
Caras 3	23	30

12.30. La Tabla 12.16 indica el número de libros prestados en una biblioteca pública durante una semana concreta. Contrastar la hipótesis de que el número de libros prestados no depende del día de la semana, usando nivel de significación (*a*) 0.05 y (*b*) 0.01.

Tabla 12.16

	N.º de libros prestados
Lunes	135
Martes	108
Miércoles	120
Jueves	114
Viernes	146

12.31. Una urna contiene 6 fichas rojas y 3 blancas. Se sacan dos al azar, se anotan sus colores y se devuelven a la urna. Este proceso se realiza 120 veces, y los resultados los presenta la Tabla 12.17.

(*a*) Calcular las frecuencias esperadas.
(*b*) Determinar al nivel de significación 0.05 si los resultados obtenidos son consistentes con los esperados.

Tabla 12.17

	Número de extracciones
0 Rojas 2 Blancas	6
1 Roja 1 Blanca	53
2 Rojas 0 Blancas	61

12.32. Se toman al azar 200 tuercas de las producidas por cada una de 4 máquinas. Las defectuosas encontradas fueron 2, 9, 10 y 3. Determinar si hay una diferencia significativa entre las máquinas, usando nivel de significación 0.05.

BONDAD DEL AJUSTE

12.33. (*a*) Usando el test ji-cuadrado, determinar la bondad del ajuste de los datos de la Tabla 7.9 del Problema 7.75. (*b*) ¿Es «demasiado bueno» el ajuste? Trabajar al nivel de significación 0.05.

12.34. Usar el test ji-cuadrado para juzgar la bondad del ajuste de los datos en (*a*) la Tabla 3.8 del Problema 3.59 y (*b*) la Tabla 3.10 del Problema 3.61. Usar un nivel de significación de 0.05 y determinar en cada caso si el ajuste es «demasiado bueno».

12.35. Usar el test ji-cuadrado para determinar la bondad del ajuste de los datos en (*a*) la Tabla 7.9 del Problema 7.79 y (*b*) la Tabla 7.10 del Problema 7.80. ¿Es consistente el resultado de la parte (*a*) con el del Problema 12.33?

TABLA DE CONTINGENCIA

12.36. La Tabla 12.18 recoge el resultado de un experimento para investigar el efecto de la vacunación de animales de laboratorio contra una cierta enfermedad. Con nivel de significación (*a*) 0.01 y (*b*) 0.05 contrastar la hipótesis de que no hay diferencia entre los grupos con y sin vacuna (o sea, que vacuna y enfermedad son independientes).

Tabla 12.18

	Enfermaron	No enfermaron
Vacunados	9	42
No vacunados	17	28

Tabla 12.19

	Aprobados	Suspensos
Clase *A*	72	17
Clase *B*	64	23

12.37. Rehacer el Problema 12.36 usando la corrección de Yates.

12.38. La Tabla 12.19 muestra el número de estudiantes en las clases *A* y *B* que aprobaron y suspendieron un examen propuesto a ambos grupos. Al nivel de significación (*a*) 0.05 y (*b*) 0.01 contrastar la hipótesis de que no hay diferencia entre las dos clases. Resolver el problema con y sin corrección de Yates.

12.39. A una parte de los pacientes con insomnio se les administró un tipo de píldoras inductoras del sueño y a los demás píldoras de azúcar (aunque ellos *creían* tomar un somnífero). Se les preguntó más tarde si las píldoras hacían efecto, con las respuestas que contiene la Tabla 12.20. Supuesto que los pacientes contestaron con sinceridad, contrastar la hipótesis de que no hay diferencia entre ambos tipos de píldoras al nivel de significación 0.05.

Tabla 12.20

	Durmieron bien	No durmieron bien
Tomaron píldoras somníferas	44	10
Tomaron píldoras inocuas	81	35

12.40. Ante una propuesta de política exterior, demócratas y republicanos adjudicaron sus votos como muestra la Tabla 12.21. Al nivel de significación (*a*) 0.01 y (*b*) 0.05, contrastar la hipótesis de que no hay diferencia entre los dos partidos en lo que a dicha propuesta se refiere.

Tabla 12.21

	Demócratas	Republicanos
A favor	85	118
En contra	78	61
Indecisos	37	25

12.41. La Tabla 12.22 presenta la relación entre las notas de estudiantes en matemáticas y física. Contrastar la hipótesis de que ambas son independientes, usando nivel de significación (*a*) 0.05 y (*b*) 0.01

Tabla 12.22

Física	Matemáticas		
	Calific. altas	Calific. bajas	Calific. medias
Calific. altas	56	71	12
Calific. medias	47	163	38
Calific. bajas	14	42	85

12.42. La Tabla 12.23 recoge los resultados de un estudio sobre si la edad de los conductores, de 21 años o más, afecta al número de accidentes que sufren (incluidos pequeños percances). Al nivel de significación (*a*) 0.05 y (*b*) 0.01, contrastar la hipótesis de que el número de accidentes es independiente de la edad del conductor. ¿Qué posibles dificultades en las técnicas de muestreo, o qué otras consideraciones, podrían afectar a las conclusiones?

Tabla 12.23

Edad del conductor	Número de accidentes			
	0	1	2	>2
21-30	748	74	31	9
31-40	821	60	25	10
41-50	786	51	22	6
51-60	720	66	16	5
61-70	672	50	15	7

12.43. (*a*) Probar que $\chi^2 = \sum(o_j^2/e_j) - N$ para todas las tablas de contingencia, donde N es la frecuencia total de todas las celdas.

(*b*) Usando el resultado de la parte (*a*), resolver el Problema 12.41.

12.44. Si N_i y N_j denotan, respectivamente, la suma de frecuencias de la i-ésima fila y de la j-ésima columna de una tabla de contingencia (las *frecuencias marginales*), probar que la frecuencia esperada para la celda que está en la i-ésima fila y en la j-ésima columna es N_iN_j/N, donde N es la frecuencia total de todas las celdas.

12.45. Demostrar la fórmula (9) de este capítulo. (*Ayuda:* Usar los Problemas 12.43 y 12.44.)

12.46. Extender el resultado de la fórmula (9) a las tablas de contingencia $2 \times k$, con $k > 3$.

12.47. Probar la fórmula (8) de este capítulo.

12.48. Por analogía con las ideas desarrolladas para tablas de contingencia $h \times k$, discutir las tablas de contingencia $h \times k \times l$ citando sus posibles aplicaciones.

COEFICIENTE DE CONTINGENCIA

12.49. La Tabla 12.24 presenta la relación entre el color del pelo y el de los ojos en una muestra de 200 estudiantes.

(*a*) Hallar el coeficiente de contingencia sin y con corrección de Yates.

(*b*) Comparar el resultado de (*a*) con el coeficiente de contingencia máximo.

Tabla 12.24

Color de los ojos	Color del cabello	
	Rubio	No rubio
Azul	49	25
No azul	30	96

12.50. Hallar el coeficiente de contingencia para los datos de (*a*) el Problema 12.36 y (*b*) el Problema 12.38, sin y con corrección de Yates.

12.51. Hallar el coeficiente de contingencia para los datos del Problema 12.41.

12.52. Probar que el coeficiente de contingencia máximo para una tabla de contingencia 3×3 es $\sqrt{\frac{2}{3}} = 0.8165$ aproximadamente.

12.53. Probar que el coeficiente de contingencia máximo de una tabla de contingencia $k \times k$ es $\sqrt{(k-1)/k}$.

CORRELACION DE ATRIBUTOS

12.54. Hallar el coeficiente de correlación para los datos de la Tabla 12.24.

12.55. Hallar el coeficiente de correlación para los datos de la (*a*) Tabla 12.18 y (*b*) Tabla 12.19 sin y con corrección de Yates.

12.56. Hallar el coeficiente de correlación entre las notas de matemáticas y física de la Tabla 12.22.

12.57. Si C es el coeficiente de contingencia para una tabla de contingencia $k \times k$ y r es el correspondiente coeficiente de correlación, probar que $r = C/\sqrt{(1 - C^2)(k - 1)}$.

PROPIEDAD ADITIVA DE χ^2

12.58. Para contrastar una hipótesis, se ha realizado cinco veces un experimento. Los valores resultantes de χ^2, cada uno correspondiendo a 4 grados de libertad, son 8.3, 9.1, 8.9, 7.8 y 8.6, respectivamente. Probar que mientras H_0 no puede ser rechazada al nivel 0.05 sobre la base de cada experimento por separado, puede rechazarse al nivel 0.005 atendiendo al resultado combinado de los cuatro experimentos.

CAPITULO 13

Ajuste de curvas y el método de mínimos cuadrados

RELACIONES ENTRE VARIABLES

En la práctica encontramos a menudo que existen relaciones entre dos (o más) variables. Por ejemplo, los pesos de las personas dependen en cierta medida de sus alturas, las circunferencias de los círculos dependen de los radios, y la presión de una masa de gas dada depende de su volumen y de su temperatura.

Suele ser deseable expresar tales relaciones en forma matemática determinando una ecuación que conecte a las variables.

AJUSTE DE CURVAS

Para hallar una ecuación que relacione las variables, el primer paso es recoger datos que muestren valores correspondientes de las variables bajo consideración. Así por ejemplo, supongamos que X e Y denotan, respectivamente, la altura y el peso de personas adultas; entonces una muestra de N individuos revelaría las alturas $X_1, X_2, ..., X_N$ y los pesos correspondientes $Y_1, Y_2, ..., Y_N$.

El próximo paso es marcar los puntos (X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N) sobre un sistema de coordenadas rectangulares. El conjunto de puntos resultante se llama a veces un *diagrama de dispersión*.

A partir del diagrama de dispersión es posible, con frecuencia visualizar una curva suave que aproxima los datos. Tal curva se llama una *curva aproximante*. En la Figura 13.1, por ejemplo, los datos parecen aproximarse bien a una línea recta, y decimos que hay una *relación lineal* entre las variables. En la Figura 13.2, sin embargo, aunque existe una relación entre las variables, no es lineal, y se dice que es una *relación no lineal*.

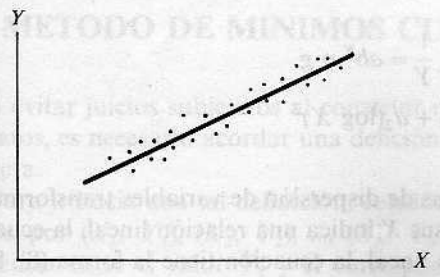


Figura 13.1.

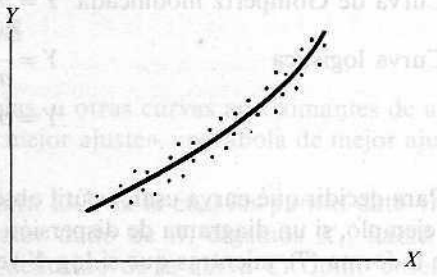


Figura 13.2.

El problema general de hallar ecuaciones de curvas aproximantes que se ajusten a un conjunto de datos se llama *ajuste de curvas*.

ECUACIONES DE CURVAS APROXIMANTES

Varios tipos comunes de curvas aproximantes y sus ecuaciones se presentan en la lista adjunta para facilitar posteriores referencias. Todas las letras excepto X e Y representan constantes. Las variables X e Y se llaman *variable independiente y dependiente*, respectivamente, aunque estos papeles se pueden intercambiar.

Línea recta	$Y = a_0 + a_1X$	(1)
Parábola, o curva cuadrática	$Y = a_0 + a_1X + a_2X^2$	(2)
Curva cúbica	$Y = a_0 + a_1X + a_2X^2 + a_3X^3$	(3)
Curva cuártica	$Y = a_0 + a_1X + a_2X^2 + a_3X^3 + a_4X^4$	(4)
Curva de grado n	$Y = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n$	(5)

Los lados derechos de las ecuaciones anteriores se llaman *polinomios* de grado uno, dos, tres, cuatro y n, respectivamente. Las funciones definidas por las cuatro primeras ecuaciones se llaman a veces funciones *lineal*, *cuadrática*, *cúbica* y *cuártica*, respectivamente.

He aquí algunas otras de las muchas ecuaciones que se utilizan frecuentemente en la práctica:

Hipérbola	$Y = \dfrac{1}{a_0 + a_1X}$ ó $\dfrac{1}{Y} = a_0 + a_1X$	(6)
Curva exponencial	$Y = ab^X$ ó $\log Y = \log a + (\log b)X = a_0 + a_1X$	(7)
Curva geométrica	$Y = aX^b$ ó $\log Y = \log a + b(\log X)$	(8)
Curva exponencial modificada	$Y = ab^X + g$	(9)
Curva geométrica modificada	$Y = aX^b + g$	(10)
Curva de Gompertz	$Y = pq^{b^X}$ ó $\log Y = \log p + b^X(\log q) = ab^X + g$	(11)
Curva de Gompertz modificada	$Y = pq^{b^X} + h$	(12)
Curva logística	$Y = \dfrac{1}{ab^X + g}$ ó $\dfrac{1}{Y} = ab^X + g$	(13)
	$Y = a_0 + a_1(\log X) + a_2(\log X)^2$	(14)

Para decidir qué curva usar, es útil obtener diagramas de dispersión de variables transformadas. Por ejemplo, si un diagrama de dispersión de $\log Y$ versus X indica una relación lineal, la ecuación tiene la forma (7), mientras que si $\log Y$ versus $\log X$ es lineal, la ecuación tiene la forma (8). Suele usarse papel gráfico especial para facilitar la decisión sobre qué curva usar. El *papel gráfico* que tiene sólo una escala calibrada logarítmicamente se llama *semilogarítmico* (o *semilog*), y el que tiene las dos escalas logarítmicas se llama *papel log-log*.

AJUSTE DE CURVAS A MANO

A menudo puede recurrirse a la intuición personal a la hora de dibujar una curva que ajuste un conjunto de datos. Esto se conoce como *método de ajuste de curvas a mano*. Si el tipo de ecuación de esa curva es conocido, es posible obtener las constantes de la ecuación eligiendo tantos puntos de la curva como constantes haya en la ecuación. Por ejemplo, si la curva es una recta, son necesarios dos puntos; si es una parábola, son precisos tres puntos. El método tiene la desventaja de que diferentes observadores obtendrán distintas curvas y ecuaciones.

LA RECTA

El tipo más sencillo de curva aproximante es una línea recta, cuya ecuación puede escribirse

$$Y = a_0 + a_1 X \tag{15}$$

Dados cualesquiera dos puntos (X_1, Y_1) y (X_2, Y_2) sobre la recta, se pueden determinar las contantes a_0 y a_1. La ecuación así obtenida se puede expresar

$$Y - Y_1 = \left(\frac{Y_2 - Y_1}{X_2 - X_1}\right)(X - X_1) \quad \text{o sea} \quad Y - Y_1 = m(X - X_1) \tag{16}$$

donde

$$m = \frac{Y_2 - Y_1}{X_2 - X_1}$$

se llama la *pendiente* de la recta y representa el cambio en Y dividido por el correspondiente cambio en X.

Cuando la ecuación se escribe en la forma (15), la constante a_1 es la pendiente m. La constante a_0, que es el valor de Y cuando $X = 0$, se llama la *Y-intersección*.

EL METODO DE MINIMOS CUADRADOS

Para evitar juicios subjetivos al construir rectas, parábolas, u otras curvas aproximantes de ajuste de datos, es necesario acordar una defición de «recta de mejor ajuste», «parábola de mejor ajuste», etcétera.

Para ir hacia una tal definición, consideremos la Figura 13.3, en la cual los puntos dato vienen dados por (X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N). Para un valor dado de X, digamos X_1, habrá una diferencia entre el valor Y_1 y el correspondiente valor deducido de la curva C. Como enseña la figura, denotamos esta diferencia por D_1, que se llama a veces *desviación*, *error* o *residual*, y puede ser positiva, negativa o nula. Análogamente, asociadas a los datos X_2, ..., X_N se obtienen desviaciones D_2, ..., D_N.

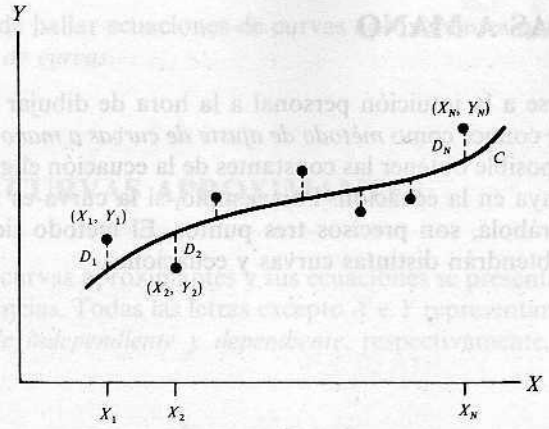


Figura 13.3.

Una medida de la «bondad del ajuste» de la curva C a los datos dados viene proporcionada por la cantidad $D_1^2 + D_2^2 + \cdots + D_N^2$. Si es pequeña, el ajuste es bueno; si es grande, el ajuste es malo. Hacemos, por tanto, la siguiente

Definición. De todas las curvas que aproximan un conjunto dado de datos, la que tiene la propiedad de que $D_1^2 + D_2^2 + \cdots + D_N^2$ es mínimo se llama una *curva de ajuste óptimo*.

Una tal curva se dice que ajusta los datos en el sentido de *mínimos cuadrados* y se llama una *curva de mínimos cuadrados*. Así pues, una recta con esa propiedad se llama *recta de mínimos cuadrados*, una parábola con esa propiedad se llama *parábola de mínimos cuadrados*, etc.

Es habitual emplear la definición precedente cuando X es la variable independiente e Y la dependiente. Si la variable dependiente es X, la definición se modifica considerando desviaciones horizontales en lugar de verticales, lo que viene a ser como intercambiar los ejes X e Y. Estas dos definiciones conducen, en general, a curvas distintas de mínimos cuadrados. Salvo que se especifique lo contrario, consideraremos a Y como la variable dependiente y a X como la independiente.

Es posible definir otras curvas de mínimos cuadrados considerando distancias perpendiculares desde cada uno de los puntos a la curva, en vez de distancias verticales u horizontales, pero no son de uso común.

LA RECTA DE MINIMOS CUADRADOS

La recta de mínimos cuadrados que aproxima el conjunto de puntos $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ tiene por ecuación

$$Y = a_0 + a_1 X \tag{17}$$

donde las constantes a_0 y a_1 quedan fijadas al resolver simultáneamente las ecuaciones

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 \end{aligned} \tag{18}$$

que se llaman las *ecuaciones normales para la recta de mínimos cuadrados* (17). Las constantes a_0 y a_1 de las ecuaciones (18) se pueden hallar, si se desea, de las fórmulas

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N\sum X^2 - (\sum X)^2} \qquad a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} \tag{19}$$

Las ecuaciones normales (18) son fáciles de recordar sin más que observar que la primera se obtiene formalmente sumando en ambos lados de (17) [o sea, $\sum Y = \sum (a_0 + a_1 X) = a_0 N + a_1 \sum X$], mientras la segunda se obtiene formalmente multiplicando primero ambos lados de (17) por X y sumando después [o sea, $\sum XY = \sum X(a_0 + a_1 X) = a_0 \sum X + a_1 \sum X^2$]. Nótese que esto no es una deducción de las ecuaciones normales, sino sólo una forma de recordarlas. Nótese además que en las ecuaciones (18) y (19) hemos usado la notación abreviada $\sum X$, $\sum XY$, etc., en lugar de $\sum_{j=1}^{N} X_j$, $\sum_{j=1}^{N} X_j Y_j$, etc.

El trabajo requerido para hallar una recta de mínimos cuadrados se puede aliviar en ocasiones transformando los datos de manera que $x = X - \bar{X}$ e $y = Y - \bar{Y}$. La ecuación de la recta de mínimos cuadrados se puede escribir entonces (véase Prob. 13.15).

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \qquad \text{o} \qquad y = \left(\frac{\sum xY}{\sum x^2}\right)x \tag{20}$$

En particular, si X es tal que $\sum X = 0$ (es decir, $\bar{X} = 0$), esto se convierte en

$$Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X \tag{21}$$

La ecuación (20) implica que $y = 0$ cuando $x = 0$; así que la recta de mínimos cuadrados pasa por el punto $(\bar{X}, \bar{Y})$, llamado *centroide* o *centro de gravedad*, de los datos.

Si se toma X como variable dependiente, escribimos (17) como $X = b_0 + b_1 Y$. Entonces los resultados anteriores son válidos si se intercambian X e Y, y se sustituyen a_0 y a_1 por b_0 y b_1, respectivamente. La recta de mínimos cuadrados resultante, sin embargo, no es generalmente la misma que la obtenida antes [véanse Probs. 13.11 y 13.15(*d*)].

RELACIONES NO LINEALES

Las relaciones no lineales pueden reducirse en ocasiones a relaciones lineales por un apropiado cambio de variables (véase Prob. 13.21).

LA PARABOLA DE MINIMOS CUADRADOS

La parábola de mínimos cuadrados que aproxima el conjunto de puntos (X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N) tiene ecuación dada por

$$Y = a_0 + a_1 X + a_2 X^2 \tag{22}$$

donde las constantes a_0, a_1 y a_2 se determinan al resolver simultáneamente las ecuaciones

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X + a_2 \sum X_2 \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum X^3 \\ \sum X^2 Y &= a_0 \sum X^2 + a_1 \sum X^3 + a_2 \sum X^4 \end{aligned} \tag{23}$$

llamadas *ecuaciones normales de la parábola de mínimos cuadrados* (22).

Las ecuaciones (23) se recuerdan fácilmente observando que se pueden obtener formalmente multiplicando (22) por 1, X y X^2, respectivamente, y sumando en ambos lados de las ecuaciones resultantes. Esta técnica puede extenderse para obtener ecuaciones normales para curvas cúbicas de mínimos cuadrados curvas cuárticas de mínimos cuadrados, y en general cualquiera de las curvas de mínimos cuadrados correspondientes a la ecuación (5).

Como en el caso de la recta de mínimos cuadra dos, las ecuaciones (23) se simplifican si se elige X de modo $\sum X = 0$. También se produce simplificación tomando como nuevas variables $x = X - \bar{X}$ e $y = Y - \bar{Y}$.

REGRESION

A menudo deseamos estimar, basados en datos de una muestra, el valor de una variable Y correspondiente a un valor dado de la variable X. Ello se puede hacer estimando el valor de Y mediante una curva de mínimos cuadrados que ajuste los datos. La curva resultante se llama una *curva de regresión de Y sobre X*, ya que Y se estima a partir de X.

Si queremos estimar el valor de X a paritr de un valor dado de Y, hemos de usar una *curva de regresión de X sobre Y*, que viene a ser un intercambio de las variables en el diagrama de dispersión de modo que X sea la variable dependiente e Y la independiente. Eso equivale a sustituir las desviaciones verticales en la definición de la curva de mínimos cuadrados en la página 291 por desviaciones horizontales.

En general, la recta o curva de regresión de Y sobre X no es la misma que la de X sobre Y.

APLICACIONES A SERIES EN EL TIEMPO

Si la variable independiente X es el tiempo, los datos muestran los valores de Y en varios instantes. Datos ordenados en el tiempo se llaman *series en el tiempo*. La recta o curva de regresión de Y sobre X en este caso se suele llamar una *recta o curva de tendencia*, y se utilizan en estimación y predicción.

PROBLEMAS EN MAS DE DOS VARIABLES

Los problemas que involucran a más de dos variables pueden tratarse de manera análoga a los de dos variables. Por ejemplo, puede haber una relación entre tres variables X, Y y Z descrita por la ecuación

$$Z = a_0 + a_1 X + a_2 Y \tag{24}$$

que se llama una *ecuación lineal en las variables X, Y y Z.*

En un sistema de coordenadas rectangulares tridimensional esa ecuación representa un plano, y los puntos (X_1, Y_1, Z_1), (X_2, Y_2, Z_2), ..., (X_N, Y_N, Z_N) de la muestra pueden «dispersarse» no lejos de ese plano, que se llama un *plano aproximante*.

Por extensión del método de mínimos cuadrados, podemos hablar de un *plano de mínimos cuadrados* que aproxima los datos. Si estamos estimando Z a partir de valores de X e Y, se le llama un *plano de regresión de Z sobre X e Y*. Las ecuaciones normales correspondientes al plano de mínimos cuadrados (24) vienen dadas por

$$\begin{aligned} \sum Z &= a_0 N + a_1 \sum X + a_2 \sum Y \\ \sum XZ &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum XY \\ \sum YZ &= a_0 \sum Y + a_1 \sum XY + a_2 \sum Y^2 \end{aligned} \tag{25}$$

y se pueden memorizar como obtenidas de (24) multiplicándola por 1, X, Y sucesivamente, y sumando después.

Cabe considerar también ecuaciones más complicadas que (24), que representan *superficies de regresión*. Si el número de variables es mayor que tres, se pierde la intuición geométrica ya que se requieren espacios de 4, 5, ... dimensiones.

Los problemas de estimación de una variable a partir de dos o más variables se llaman problemas de *regresión múltiple* y se considerarán con más detalle en el Capítulo 15.

PROBLEMAS RESUELTOS

RECTAS

13.1. (*a*) Construir una recta que aproxime los datos de la Tabla 13.1.
(*b*) Hallar una ecuación para esa recta.

Tabla 13.1

X	2	3	5	7	9	10
Y	1	3	7	11	15	17

Solución

(*a*) Marcar los puntos (2, 1), (3, 3), (5, 7), (7, 11), (9, 15) y (10, 17) en un sistema rectangular de coordenadas, como indica la Figura 13.4. Es claro de esa figura que todos los puntos están en una recta (dibujada a trazos); así que una recta ajusta esos datos *exactamente*.

(*b*) Para hallar la ecuación de la recta dada por

$$Y = a_0 + a_1 X \tag{26}$$

sólo se necesitan dos puntos. Escogemos los puntos (2, 1) y (3, 3), por ejemplo. Para el punto (2, 1), $X = 2$ y $Y = 1$; sustituyendo esos valores en (26) se ve que

$$1 = a_0 + 2a_1 \tag{27}$$

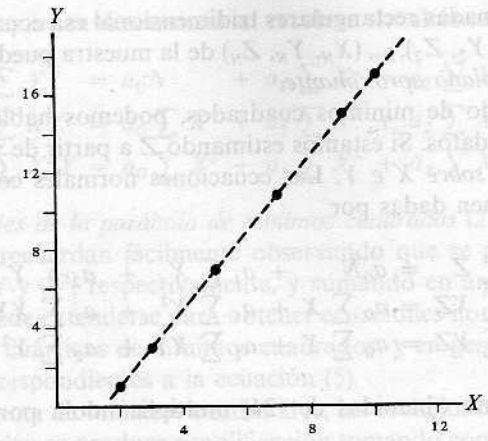


Figura 13.4.

Análogamente, para los puntos (3, 3), $X = 3$ e $Y = 3$; sustituyendo esos valores en (26) se obtiene

$$3 = a_0 + 3a_1 \tag{28}$$

Resolviendo (27) y (28) simultáneamente, $a_0 = -3$ y $a_1 = 2$, y la requerida ecuación es

$$Y = -3 + 2X \quad \text{o sea} \quad Y = 2X - 3$$

Como comprobación, véase que los puntos (5, 7), (7, 11), (9, 15) y (10, 17) están también sobre esa recta.

13.2. En el Problema 13.1 hallar (*a*) Y cuando $X = 4$, (*b*) Y cuando $X = 15$, (*c*) Y cuando $X = 0$, (*d*) X cuando $Y = 7.5$, (*e*) X cuando $Y = 0$ y (*f*) el crecimiento en Y correspondiente a un crecimiento unidad en X.

Solución

Suponemos que para otros valores de X e Y distintos de los especificados en la Tabla 13.1 es válida la misma relación $Y = 2X - 3$.

(*a*) Si $X = 4$, $Y = 2(4) - 3 = 8 - 3 = 5$. Como estamos hallando el valor de Y correspondiente a un valor de X incluido *entre* dos valores dados de X, este proceso se llama *interpolación lineal.*

(*b*) Si $X = 15$, $Y = 2(15) - 3 = 30 - 3 = 27$. Como estamos hallando el valor de Y correspondiente a un valor de X *exterior* a los valores dados de X, este proceso se llama *extrapolación lineal.*

(*c*) Si $X = 0$, $Y = 2(0) - 3 = 0 - 3 = -3$. El valor de Y cuando $X = 0$ se llama *Y-intersección.* Es el valor de Y en el punto donde la recta (extendida si es preciso) corta al eje Y.

(*d*) Si $Y = 7.5$, $7.5 = 2X - 3$; entonces $2X = 7.5 + 3 = 10.5$ y $X = 10.5/2 = 5.25$.

(*e*) Si $Y = 0$, $0 = 2X - 3$; entonces $2X = 3$ y $X = 1{,}5$. El valor de X cuando $Y = 0$ se llama X-intersección. Es el valor de X en el punto donde la recta (extendida si es preciso) corta al eje X.

(*f*) Si X crece una unidad de 2 a 3, Y crece de 1 a 3, un cambio de dos unidades. Si X crece de 2 a 10, o sea $(10 - 2) = 8$ unidades, Y crece de 1 a 17, un cambio de $(17 - 1) = 16$ unidades; esto es, Y crece 2 unidades por cada unidad que crece X.

En general, si ΔY denota el cambio en Y debido a un cambio en X de ΔX entonces el cambio en Y por unidad de cambio en X viene dado por $\Delta Y/\Delta X = 2$. Esto se llama la pendiente de la

recta y es siempre igual a a_1 en la ecuación $Y = a_0 + a_1X$. La constante a_0 es la *Y-intersección* de la recta [véase parte (*c*)].

Las cuestiones anteriores se pueden contestar también directamente del gráfico, Figura 13.4.

13.3. (*a*) Probar que la ecuación de una recta que pasa por los puntos (X_1, Y_1) y (X_2, Y_2) viene dada por

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1)$$

(*b*) Hallar la ecuación de una recta que pasa por los puntos $(2, -3)$ y $(4, 5)$.

Solución

(*a*) La ecuación de la recta es

$$Y = a_0 + a_1X \tag{29}$$

Como (X_1, Y_1) está en la recta,

$$Y_1 = a_0 + a_1X_1 \tag{30}$$

Como (X_2, Y_2) está en la recta,

$$Y_2 = a_0 + a_1X_2 \tag{31}$$

Restando la ecuación (30) de (29),

$$Y - Y_1 = a_1(X - X_1) \tag{32}$$

Restando la ecuación (30) de (31),

$$Y_2 - Y_1 = a_1(X_2 - X_1) \qquad \text{o sea} \qquad a_1 = \frac{Y_2 - Y_1}{X_2 - X_1}$$

Sustituyendo este valor de a_1 en la ecuación (32), obtenemos

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1)$$

como se pedía. La cantidad

$$\frac{Y_2 - Y_1}{X_2 - X_1}$$

abreviada usualmente como m, representa el cambio en Y dividido por el correspondiente cambio en X y es la pendiente de la recta. La ecuación pedida puede escribirse $Y - Y_1 = m(X - X_1)$.

(*b*) *Primer método* [usando el resultado de la parte (*a*)]

Correspondiendo al primer punto $(2, -3)$, tenemos $X_1 = 2$ e $Y_1 = -3$; para el segundo, $(4, 5)$, tenemos $X_2 = 4$ e $Y_2 = 5$. Luego la pendiente es

$$m = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{5 - (-3)}{4 - 2} = \frac{8}{2} = 4$$

y la ecuación pedida es

$$Y - Y_1 = m(X - X_1) \qquad \text{o sea} \qquad Y - (-3) = 4(X - 2)$$

que se puede expresar $Y + 3 = 4(X - 2)$, o sea $Y = 4X - 11$.

Segundo método [usando el método del Problema 13.1(*b*)]

La ecuación de una recta es $Y = a_0 + a_1X$. Como el punto $(2, -3)$ está en la recta, $-3 = a_0 + 2a_1$, y como el punto $(4, 5)$ está en la recta, $5 = a_0 + 4a_1$; resolviendo esas dos ecuaciones simultáneamente, obtenemos $a_1 = 4$ y $a_0 = -11$. Luego la ecuación pedida es

$$Y = -11 + 4X \qquad \text{o sea} \qquad Y = 4X - 11$$

13.4. Dar una interpretación gráfica de la parte (*a*) del Problema 13.3.

Solución

La Figura 13.5 muestra la recta que pasa por los puntos P y Q, de coordenadas (X_1, Y_1) y (X_2, Y_2), respectivamente. El punto R, con coordenadas (X, Y), representa cualquier otro punto sobre esa recta.

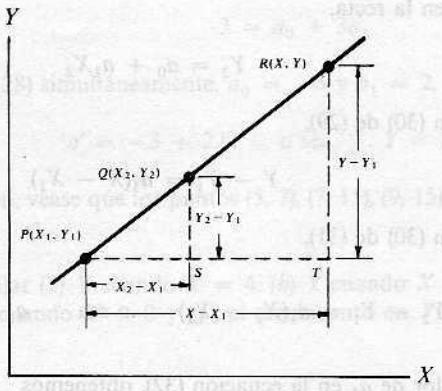


Figura 13.5.

Por semejanza de los triángulos PRT y PQS

$$\frac{RT}{TP} = \frac{QS}{SP} \qquad \text{o sea} \qquad \frac{Y - Y_1}{X - X_1} = \frac{Y_2 - Y_1}{X_2 - X_1} \tag{33}$$

Entonces, multiplicando ambos lados por $X - X_1$,

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1)$$

que es la ecuación solicitada para la recta.

Nótese que cada uno de los cocientes en la ecuación (33) es la pendiente m; eso puede escribirse $Y - Y_1 = m(X - X_1)$.

13.5. Hallar (*a*) la pendiente, (*b*) la ecuación, (*c*) la Y-intersección, y (*d*) la X-intersección de la recta que pasa por los puntos $(1, 5)$ y $(4, -1)$.

Solución

(*a*) $(X_1 = 1, Y_1 = 5)$ y $(X_2 = 4, Y_2 = -1)$. Luego

$$m = \text{pendiente} = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{-1 - 5}{4 - 1} = \frac{-6}{3} = -2$$

El signo negativo de la pendiente indica que cuando X crece, Y decrece, tal como se ve en la Figura 13.6.

(*b*) La ecuación de la recta es

$$Y - Y_1 = m(X - X_1) \quad \text{o sea} \quad Y - 5 = -2(X - 1)$$

Es decir, $$Y - 5 = -2X + 2 \quad \text{o sea} \quad Y - 7 - 2X$$

Esto puede obtenerse también por el segundo método del Problema 13.3(*b*).

(*c*) La Y-intersección, que es el valor de Y cuando $X = 0$, viene dada por $Y = 7 - 2(0) = 7$. Eso puede verse directamente en la Figura 13.6.

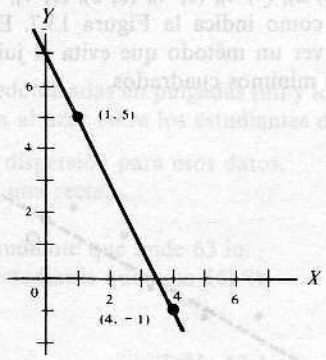


Figura 13.6.

(*d*) La X-intersección es el valor de X cuando $Y = 0$. Sustituyendo $Y = 0$ en la ecuación $Y = 7 - -2X$, o sea $0 = 7 - 2X$, o sea $2X = 7$ y $X = 3.5$. Eso puede verse directamente en la Figura 13.6.

13.6. Hallar las ecuaciones de una recta que pase por el punto (4, 2) y sea paralela a la recta $2X + 3Y = 6$.

Solución

Si dos rectas son paralelas, sus pendientes son iguales. De $2X + 3Y = 6$ tenemos $3Y = 6 - 2X$, o sea $Y = 2 - \frac{2}{3}X$, así que la pendiente de la recta es $m = -\frac{2}{3}$. Luego la ecuación de la recta pedida es

$$Y - Y_1 = m(X - X_1) \quad \text{o sea} \quad Y - 2 = -\tfrac{2}{3}(X - 4)$$

que también se puede escribir $2X + 3Y = 14$.

Otro método

Cualquier recta paralela a $2X + 3Y = 6$ tiene ecuación $2X + 3Y = c$. Para hallar c, hacemos $X = 4$ e $Y = 2$. Entonces $2(4) + 3(2) = c$, o sea $c = 14$, y la ecuación buscada es $2X + 3Y = 14$.

13.7. Hallar la ecuación de una recta cuya pendiente es -4 y cuya Y-intersección es 16.

Solución

En la ecuación $Y = a_0 + a_1X$, $a_0 = 16$ es la Y-intersección y $a_1 = -4$ es la pendiente. Así pues, la ecuación buscada es $Y = 16 - 4X$.

13.8. (*a*) Construir una recta que aproxime los datos de la Tabla 13.2.
(*b*) Hallar la ecuación de esa recta.

Tabla 13.2

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

Solución

(*a*) Marcar los puntos (1, 1), (3, 2), (4, 4), (6, 4), (8, 5), (9, 7), (11, 8) y (14, 9) sobre un sistema de coordenadas rectangulares, como indica la Figura 13.7. En la figura se ha trazado una recta aproximante *a mano*. Para ver un método que evita el juicio subjetivo, consultar el Problema 13.11, que usa el método de mínimos cuadrados.

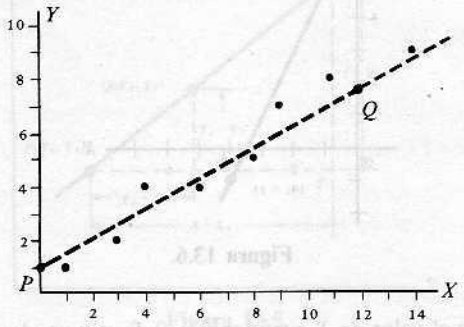


Figura 13.7.

(*b*) Para obtener una ecuación de esa recta, escojamos dos puntos en ella, tales como P y Q; las coordenadas de P y Q, según el gráfico, son aproximadamente (0, 1) y (12, 7.5). La ecuación de la recta es $Y = a_0 + a_1X$. Luego para que (0, 1) esté en ella, ha de ser $1 = a_0 + a_1(0)$, y para que esté el punto (12, 7.5), ha de ser $7.5 = a_0 + 12a_1$; como la primera de estas ecuaciones da $a_0 = 1$, la segunda nos dice que $a_1 = 6.5/12 = 0.542$. Por tanto, la requerida ecuación es $Y = 1 + 0.542X$.

Otro método

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1) \quad \text{e} \quad Y - 1 = \frac{7.5 - 1}{12 - 0}(X - 0) = 0.542X$$

Así pues $Y = 1 + 0.542X$.

13.9. (*a*) Comparar los valores de Y obtenidos de la recta aproximante con los de la Tabla 13.2.
(*b*) Estimar el valor de Y cuando $X = 10$.

Solución

(*a*) Para $X = 1$, $Y = 1 + 0.542(1) = 1.542$, o sea 1.5. Para $X = 3$, $Y = 1 + 0.542(3) = 2.626$, o 2,6. Los valores de Y correspondientes a otros valores de X se obtienen de la misma manera. Los valores de Y estimados por la ecuación $Y = 1 + 0.542X$ se denotan por Y_{est}. Estos valores estimados, junto con los verdaderos datos de la Tabla 13.2, se recogen en la Tabla 13.3.
(*b*) El valor estimado de Y cuando $X = 10$ es $Y = 1 + 0.542(10) = 6.42$ o sea 6.4.

Tabla 13.3

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9
Y_{est}	1.5	2.6	3.2	4.3	5.3	5.9	7.0	8.6

13.10. La Tabla 13.4 da las alturas redondeadas en pulgadas (in) y los pesos en libras (lb) de una muestra de 12 estudiantes varones tomada al azar entre los estudiantes de primer año del State College.

(*a*) Obtener un diagrama de dispersión para esos datos.
(*b*) Aproximar los datos con una recta.
(*c*) Hallar su ecuación.
(*d*) Estimar el peso de un estudiante que mide 63 in.
(*e*) Estimar la altura de un estudiante que pesa 168 lb.

Tabla 13.4

Altura X (in)	70	63	72	60	66	70	74	65	62	67	65	68
Peso Y (lb)	155	150	180	135	156	168	178	160	132	145	139	152

Solución

(*a*) El diagrama de dispersión, véase Figura 13.8, se obtiene marcando los puntos (70, 155), (63, 150), (72, 180), ..., (68, 152).

(*b*) Una recta que aproxima a los datos se ve en trazos en la Figura 13.8. No es sino una de las muchas posibles rectas que se podían haber construido.

(*c*) Escoger un par de puntos arbitrarios P y Q de esa recta. Sus coordenadas según el gráfico vienen a ser (60, 130) y (72, 170). Por tanto

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1) \qquad Y - 130 = \frac{170 - 130}{72 - 60}(X - 60) \qquad Y = \frac{10}{3}X - 70$$

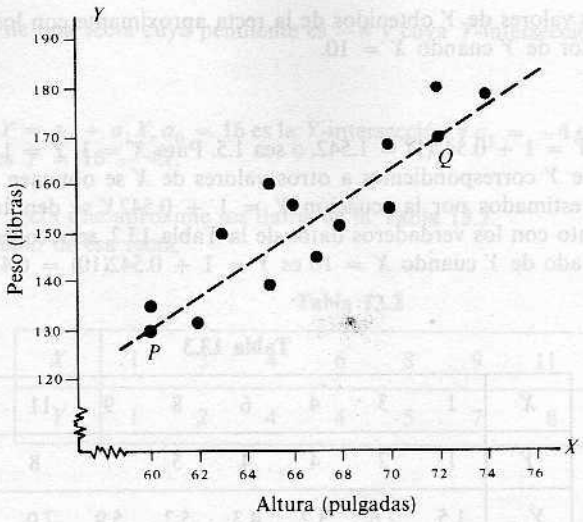


Figura 13.8.

(*d*) Si $X = 63$, entonces $Y = \frac{10}{3}(63) - 70 = 140$ lb.

(*e*) Si $Y = 168$, entonces $168 = \frac{10}{3}X - 70$, $\frac{10}{3}X = 238$ y $X = 71.4$, o sea 71 in.

LA RECTA DE MINIMOS CUADRADOS

13.11. Ajustar una recta de mínimos cuadrados a los datos del Problema 13.8 usando (*a*) X como variable independiente y (*b*) X como variable dependiente.

Solución

(*a*) La ecuación de la recta es $Y = a_0 + a_1 X$. Las ecuaciones normales son

$$\sum Y = a_0 N + a_1 \sum X$$
$$\sum XY = a_0 \sum X + a_1 \sum X^2$$

El trabajo exigido para calcular las sumas se puede ordenar como en la Tabla 13.5. Si bien la columna de la derecha no es necesaria para esta parte del problema, la usaremos en (*b*).

Tabla 13.5

X	Y	X^2	XY	Y^2
1	1	1	1	1
3	2	9	6	4
4	4	16	16	16
6	4	36	24	16
8	5	64	40	25
9	7	81	63	49
11	8	121	88	64
14	9	196	126	81
$\sum X = 56$	$\sum Y = 40$	$\sum X^2 = 524$	$\sum XY = 364$	$\sum Y^2 = 256$

Puesto que hay ocho pares de valores de X e Y, $N = 8$ y las ecuaciones normales se convierten en

$$\begin{aligned} 8a_0 + 56a_1 &= 40 \\ 56a_0 + 524a_1 &= 364 \end{aligned}$$

Resolviendo simultáneamente, $a_0 = \frac{6}{11}$, o sea 0.545; $a_1 = \frac{7}{11}$, o sea 0.636; y la recta de mínimos cuadrados pedida es $Y = \frac{6}{11} + \frac{7}{11}X$, o sea $Y = 0.545 + 0.636X$.

Otro método

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N\sum X^2 - (\sum X)^2} = \frac{(40)(524) - (56)(364)}{(8)(524) - (56)^2} = \frac{6}{11} \quad \text{o sea} \quad 0.545$$

$$a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = \frac{(8)(364) - (56)(40)}{(8)(524) - (56)^2} = \frac{7}{11} \quad \text{o sea} \quad 0.636$$

Luego $Y = a_0 + a_1X$, o sea $Y = 0.545 + 0.636X$, como antes.

(*b*) Si se considera X como variable dependiente e Y como independiente, la ecuación de la recta de mínimos cuadrados es $X = b_0 + b_1Y$ y las ecuaciones normales son

$$\begin{aligned} \sum X &= b_0N + b_1\sum Y \\ \sum XY &= b_0\sum Y + b_1\sum Y^2 \end{aligned}$$

Por la Tabla 13.5 las ecuaciones normales se convierten en

$$\begin{aligned} 8b_0 + 40b_1 &= 56 \\ 40b_0 + 256b_1 &= 364 \end{aligned}$$

de donde $b_0 = -\frac{1}{2}$, o sea $-0{,}50$ y $b_1 = \frac{3}{2}$, o sea 1.50. Estos valores pueden deducirse también de

$$b_0 = \frac{(\sum X)(\sum Y^2) - (\sum Y)(\sum XY)}{N\sum Y^2 - (\sum Y)^2} = \frac{(56)(256) - (40)(364)}{(8)(256) - (40)^2} = -0.50$$

$$b_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum Y^2 - (\sum Y)^2} = \frac{(8)(364) - (56)(40)}{(8)(256) - (40)^2} = 1.50$$

Luego la ecuación solicitada de la recta de mínimos cuadrados es $X = b_0 + b_1Y$, o sea $X = -0.50 + 1.50Y$.

Nótese que resolviendo esa ecuación obtenemos $Y = \frac{1}{3} + \frac{2}{3}X$, o sea $Y = 0{,}333 + 0{,}667X$, que es distinta de la recta a la que llegamos en la parte (*a*).

13.12. Dibujar las dos rectas del Problema 13.11.

Solución

Los gráficos de las rectas $Y = 0.545 + 0.636X$ y $X = -0.500 + 1.50Y$, se muestran en la Figura 13.9. Hagamos notar que en este caso son casi coincidentes, lo cual indica que los datos están muy bien descritos por una relación lineal.

La recta de la parte (*a*) del Problema 13.11 se suele llamar la *recta de regresión de Y sobre X*, y se

usa para estimar Y en valores dados de X. La recta de la parte (b) del Problema 13.11 se suele llamar la *recta de regresión de X sobre Y*, y se usa para estimar X en valores dados de Y.

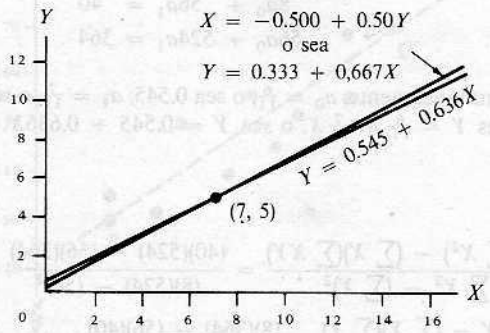


Figura 13.9.

13.13. (a) Probar que las rectas de mínimos cuadrados obtenidas en el Problema 13.11 se cortan en el punto $(\bar{X}, \bar{Y})$.
(b) Estimar el valor de Y cuando $X = 12$.
(c) Estimar el valor de X cuando $Y = 3$.

Solución

$$\bar{X} = \frac{\sum X}{N} = \frac{56}{8} = 7 \qquad \bar{Y} = \frac{\sum Y}{N} = \frac{40}{8} = 5$$

Luego el punto $(\bar{X}, \bar{Y})$, llamado el *centroide*, es (7, 5).

(a) El punto (7,5) está en la recta $Y = 0{,}545 + 0{,}636X$; o, más exactamente, $Y = \frac{6}{11} + \frac{7}{11}X$, pues $5 = \frac{6}{11} + \frac{7}{11}(7)$. El punto (7, 5) está en la recta $X = -\frac{1}{2} + \frac{3}{2}Y$, ya que $7 = -\frac{1}{2} + \frac{3}{2}(5)$.

Otro método

Las ecuaciones de las dos rectas son $Y = \frac{6}{11} + \frac{7}{11}X$ y $X = -\frac{1}{2} + \frac{3}{2}Y$. Resolviendo simultáneamente se encuentra $X = 7$ e $Y = 5$. Luego las rectas se cortan en el punto (7, 5).

(b) Haciendo $X = 12$ en la recta de regresión de Y(Problema 13.11), $Y = 0.545 + 0636(12) = 8.2$.
(c) Haciendo $Y = 3$ en la recta de regresión de X(Problema 13.11), $X = -0.50 + 1.50(3) = 4.0$.

13.14. Probar que una recta de mínimos cuadrados siempre pasa por el punto $(\bar{X}, \bar{Y})$.

Solución

Caso 1: (X es la variable independiente)

La ecuación de la recta de mínimos cuadrados es

$$Y = a_0 + a_1 X \tag{34}$$

Una ecuación normal para la recta de mínimos cuadrados es

$$\sum Y = a_0 N + a_1 \sum X \tag{35}$$

Dividiendo la ecuación (35) a ambos lados por N tenemos

$$\bar{Y} = a_0 + a_1\bar{X} \tag{36}$$

Restando (36) de (34), la recta de mínimos cuadrados se puede expresar

$$Y - \bar{Y} = a_1(X - \bar{X}) \tag{37}$$

que demuestra que la recta pasa por el punto $(\bar{X}, \bar{Y})$.

Caso 2: (X es la variable dependiente)

Procediendo como en el caso 1, pero intercambiando X e Y y sustituyendo las constantes a_0 y a_1 por b_0 y b_1, respectivamente, vemos que la recta de mínimos cuadrados se puede escribir

$$X - \bar{X} = b_1(Y - \bar{Y}) \tag{38}$$

lo cual indica que la recta pasa por el punto $(\bar{X}, \bar{Y})$.

Nótese que las rectas (37) y (38) no son coincidentes; se cortan en $(\bar{X}, \bar{Y})$.

13.15. (*a*) Considerando X como variable independiente, probar que la ecuación de la recta de mínimos cuadrados se puede escribir como

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \qquad \text{es decir} \qquad y = \left(\frac{\sum xY}{\sum x^2}\right)x$$

donde $x = X - \bar{X}$ donde $y = Y - \bar{Y}$.

(*b*) Si $\bar{X} = 0$, demostrar que la recta de mínimos cuadrados de la parte (*a*) se escribe

$$Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X$$

(*c*) Escribir la ecuación de la recta de mínimos cuadrados correspondiente a la de la parte (*a*) si Y es la variable independiente.

(*d*) Verificar que las rectas en las partes (*a*) y (*c*) no son necesariamente la misma.

Solución

(*a*) La ecuación (37) se puede escribir $y = a_1x$, donde $x = X - \bar{X}$ e $y = Y - \bar{Y}$. Además, de la solución simultánea de las ecuaciones normales (18) tenemos

$$a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = \frac{N\sum (x + \bar{X})(y + \bar{Y}) - [\Sigma (x + \bar{X})][\sum (y + \bar{Y})]}{N\sum (x + \bar{X})^2 - [\sum (x + \bar{X})]^2}$$

$$= \frac{N\sum (xy + x\bar{Y} + \bar{X}y + \bar{X}\bar{Y}) - (\sum x + N\bar{X})(\sum y + N\bar{Y})}{N\sum (x^2 + 2x\bar{X} + \bar{X}^2) - (\sum x + N\bar{X})^2}$$

$$= \frac{N\sum xy + N\bar{Y}\sum x + N\bar{X}\sum y + N^2\bar{X}\bar{Y} - (\sum x + N\bar{X})(\sum y + N\bar{Y})}{N\sum x^2 + 2N\bar{X}\sum x + N^2\bar{X}^2 - (\sum x + N\bar{X})^2}$$

Pero $\sum x = \sum (X - \bar{X}) = 0$ y $\sum y = \sum (Y - \bar{Y}) = 0$; por tanto, lo anterior se reduce a

$$a_1 = \frac{N\sum xy + N^2\bar{X}\bar{Y} - N^2\bar{X}\bar{Y}}{N\sum x^2 + N^2\bar{X}^2 - N^2\bar{X}^2} = \frac{\sum xy}{\sum x^2}$$

Esto puede escribirse como

$$a_1 = \frac{\sum xy}{\sum x^2} = \frac{\sum x(Y - \bar{Y})}{\sum x^2} = \frac{\sum xY - \bar{Y}\sum x}{\sum x^2} = \frac{\sum xY}{\sum x^2}$$

Asi que la recta de mínimos cuadrados es $y = a_1 x$; es decir,

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \quad \text{o sea} \quad y = \left(\frac{\sum xY}{\sum x^2}\right)x$$

(*b*) Si $\bar{X} = 0$, $x = X - \bar{X} = X$. Entonces de

$$y = \left(\frac{\sum xY}{\sum x^2}\right)$$

se tiene $$y = \left(\frac{\sum XY}{\sum X^2}\right)X \quad \text{o sea} \quad Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X$$

Otro método

Las ecuaciones normales de la recta de mínimos cuadrados $Y = a_0 + a_1 X$ son

$$\sum Y = a_0 N + a_1 \sum X \quad \text{y} \quad \sum XY = a_0 \sum X + a_1 \sum X^2$$

Si $\bar{X} = (\sum X)/N = 0$, entonces $\sum X = 0$ y las ecuaciones normales pasan a ser

$$\sum Y = a_0 N \quad \text{y} \quad \sum XY = a_1 \sum X^2$$

de donde $$a_0 = \frac{\sum Y}{N} = \bar{Y} \quad \text{y} \quad a_1 = \frac{\sum XY}{\sum X^2}$$

Luego la ecuación pedida de la recta de mínimos cuadrados es

$$Y = a_0 + a_1 X \quad \text{o sea} \quad Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X$$

(*c*) Intercambiando X e Y, o sea x e y, podemos ver como en (*a*) que

$$x = \left(\frac{\sum xy}{\sum y^2}\right)y$$

(*d*) Por la parte (*a*), la recta de mínimos cuadrados es

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \tag{39}$$

Por la parte (*c*), la recta de mínimos cuadrados es

$$x = \left(\frac{\sum xy}{\sum y^2}\right)y$$

o sea

$$y = \left(\frac{\sum y^2}{\sum xy}\right) x \tag{40}$$

Como en general $$\frac{\sum xy}{\sum x^2} \neq \frac{\sum y^2}{\sum xy}$$

la recta de mínimos cuadrados (39) y (40) son diferentes en general. Sin embargo, intersectan en $x = 0$ e $y = 0$ [o sea, en el punto $(\bar{X}, \bar{Y})$].

13.16. Si $X' = X + A$ e $Y' = Y + B$, donde A y B son constantes, probar que

$$a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = \frac{N\sum X'Y' - (\sum X')(\sum Y')}{N\sum X'^2 - (\sum X')^2} = a_1'$$

Solución

$$x' = X' - \bar{X}' = (X + A) - (\bar{X} + A) = X - \bar{X} = x$$
$$y' = Y' - \bar{Y}' = (Y + B) - (\bar{Y} + B) = Y - \bar{Y} = y$$

Entonces $$\frac{\sum xy}{\sum x^2} = \frac{\sum x'y'}{\sum x'^2}$$

y el resultado se sigue del Problema 13.15. Un resultado similar se aplica a b_1.

Este resultado es útil porque nos capacita para simplificar cálculos al obtener la recta de regresión restando constantes adecuadas de las variables X e Y (véase el segundo método del Problema 13.17).

Nota: El resultado no es válido si $X' = c_1X + A$ e $Y' = c_2Y + B$ a menos que $c_1 = c_2$.

13.17. Ajustar una recta de mínimos cuadrados a los datos del Problema 13.10 usando (*a*) X como variable independiente y (*b*) X como variable dependiente.

Solución

Primer método

(*a*) Del Problema 13.15(*a*) sabemos que la recta requerida es

$$y = \left(\frac{\sum xy}{\sum x^2}\right) x$$

donde $x = X - \bar{X}$ e $y = Y - \bar{Y}$. El trabajo de calcular las sumas se puede organizar como sugiere la Tabla 13.6. De sus dos primeras columnas hallamos $\bar{X} = 802/12 = 66.8$ e $\bar{Y} = 1850/12 = 154.2$. La última columna se utilizará en la parte (*b*).

La recta de mínimos cuadrados pedida es

$$y = \left(\frac{\sum xy}{\sum x^2}\right) x = \frac{616.32}{191.68} x = 3.32x$$

o sea $Y - 154.2 = 3.22(X - 66.8)$, que se puede escribir $Y = 3.22X - 60.9$ Esta ecuación se llama la *recta de regresión de Y sobre X*, y se usa para estimar Y para valores dados de X.

(*b*) Si X es la variable dependiente, la recta en cuestión es

$$x = \left(\frac{\sum xy}{\sum y^2}\right) y = \frac{616.32}{2659.68} y = 0.232y$$

Tabla 13.6

Altura X	Peso Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	xy	x^2	y^2
70	155	3.2	0.8	2.56	10.24	0.64
63	150	−3.8	−4.2	15.96	14.44	17.64
72	180	5.2	25.8	134.16	27.04	665.64
60	135	−6.8	−19.2	130.56	46.24	368.64
66	156	−0,8	1.8	−1.44	0,64	3.24
70	168	3.2	13.8	44.16	10.24	190.44
74	178	7.2	23.8	171.36	51.84	566.44
65	160	−1.8	5.8	−10.44	3.24	33.64
62	132	−4.8	−22.2	106.52	23.04	492.84
67	145	0.2	−9.2	− 1.84	0.04	84.64
65	139	−1.8	−15.2	27.36	3.24	231.04
68	152	1.2	−2.2	−2.64	1.44	4.84
$\sum X = 802$	$\sum Y = 1850$			$\sum xy = 616.32$	$\sum x^2 = 191.68$	$\sum y^2 = 2659.68$
$\bar{X} = 66.8$	$\bar{Y} = 154.2$					

que se puede expresar como $X - 66.8 = 0.232(Y - 154.2)$, o sea $X = 31.0 + 0.232\ Y$. Esta es la *recta de regresión de X sobre Y*, usada para estimar X para valores de Y dados.

Nótese que el método del Problema 13.11 es también aplicable si se desea.

Segundo método

Usando el resultado del Problema 13.16, podemos sustraer constantes adecuadas de X e Y. Escogemos sustraer 65 de X y 150 de Y. Con ello los resultados se muestran en la Tabla 13.7.

Tabla 13.7

X'	Y'	X'^2	$X'Y'$	Y'^2
5	5	25	25	25
−2	0	4	0	0
7	30	49	210	900
−5	−15	25	75	225
1	6	1	6	36
5	18	25	90	324
9	28	81	252	784
0	10	0	0	100
−3	−18	9	54	324
2	−5	4	−10	25
0	−11	0	0	121
3	2	9	6	4
$\sum X' = 22$	$\sum Y' = 50$	$\sum X'^2 = 232$	$\sum X'Y' = 708$	$\sum Y'^2 = 2868$

$$a_1 = \frac{N\sum X'Y' - (\sum X')(\sum Y')}{N\sum X'^2 - (\sum X')^2} = \frac{(12)(708) - (22)(50)}{(12)(232) - (22)^2} = 3.22$$

$$b_1 = \frac{N\sum X'Y' - (\sum Y')(\sum X')}{N\sum Y'^2 - (\sum Y')^2} = \frac{(12)(708) - (50)(22)}{(12)(2868) - (50)^2} = 0{,}232$$

Como $\bar{X} = 65 + 22/12 = 66.8$ e $\bar{Y} = 150 + 50/12 = 154.2$, las ecuaciones de regresión son $Y - 154.2 = 3.2(X - 66.8)$ y $X - 66.8 = 0.232\,(Y - 154.2)$; esto es, $Y = 3.22\,X - 60.9$ y $X = 0.232\,Y + 31.0$, de acuerdo con el primer método.

13.18. (*a*) Dibujar, en un mismo par de ejes, los gráficos de las dos rectas del Problema 13.17.
(*b*) Estimar el peso de un estudiante que mide 63 in.
(*c*) Estimar la altura de un estudiante que pesa 168 lb.

Solución

(*a*) Las dos rectas se muestran en la Figura 13.10, junto a los puntos dato originales. Obsérvese que se cortan en $(\bar{X}, \bar{Y})$, o sea (66.8, 154.2).
(*b*) Para estimar Y a partir de X usaremos la recta de regresión de Y sobre X, dada en el Problema 13.17 por $Y = 3.22X - 60.9$. Entonces, si $X = 63$, $Y = 3.22(63) - 60.9 = 142$ lb.
(*c*) Para estimar X a partir de Y usaremos la recta de regresión de X sobre Y, dada en el Problema 13.17 por $X = 31.0 + 0.232Y$. Entonces, si $Y = 168$, $X = 31.0 + 0.232(168) = 70.0$ in.

Los resultados de las partes (*b*) y (*c*) deben compararse con los del Problema 13.10, partes (*d*) y (*c*).

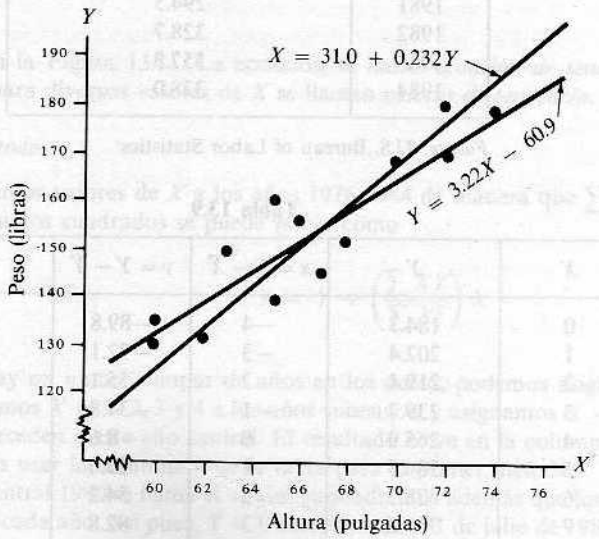


Figura 13.10.

APLICACIONES A SERIES EN EL TIEMPO

13.19. El índice de costes sanitarios en EE.UU. para los años 1976-1984, tomando como 100 el del año 1967, se da en la Tabla 13.8.

(*a*) Representar los datos gráficamente.
(*b*) Hallar la ecuación de una recta de mínimos cuadrados que ajuste esos datos.

(*c*) Estimar el índice para el año 1985 y comparar con el valor real, 396.1.
(*d*) Estimar el índice para 1975 y comparar con el valor verdadero, 168.6.

Solución

(*a*) Véase Figura 13.11.
(*b*) *Primer método*

Usar las ecuaciones $y = (\sum xy/\sum x^2)x$, donde $x = X - \bar{X}$ e $y = Y - \bar{Y}$. La Tabla 13.9 resume la tarea. La requerida ecuación es $y = (1511.3/60)x$, o sea $y = 25.19x$, que se puede escribir

$$Y - 274.5 = 25.19(X - 4) \quad \text{o sea} \quad Y = 173.7 + 25.19X$$

Tabla 13.8

Año	Indice de costes sanitarios en EE. UU. (1967 = 100)
1976	184.7
1977	202.4
1978	219.4
1979	239.7
1980	265.9
1981	294.5
1982	328.7
1983	357.3
1984	378.0

Fuente: U.S. Bureau of Labor Statistics.

Tabla 13.9

Año	X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy
1976	0	184.7	−4	−89.8	16	359.2
1977	1	202.4	−3	−72.1	9	216.3
1978	2	219.4	−2	−55.1	4	110.2
1979	3	239.7	−1	−34.8	1	34.8
1980	4	265.9	0	−8.6	0	0.0
1981	5	284.5	1	20.0	1	20.0
1982	6	328.7	2	54.2	4	108.4
1983	7	357.3	3	82.8	9	248.4
1984	8	378.0	4	103.5	16	414.0
	$\sum X = 36$	$\sum Y = 2470.6$			$\sum x^2 = 60$	$\sum xy = 1511.3$
	$\bar{X} = 4$	$\bar{Y} = 274.5$				

donde el origen $X = 0$ es el año 1976 (se suele tomar la mitad del año, el 1 de julio de 1976) y la unidad de X es 1 año. El gráfico de esta recta, llamada a veces una *recta de tendencia*, se muestra

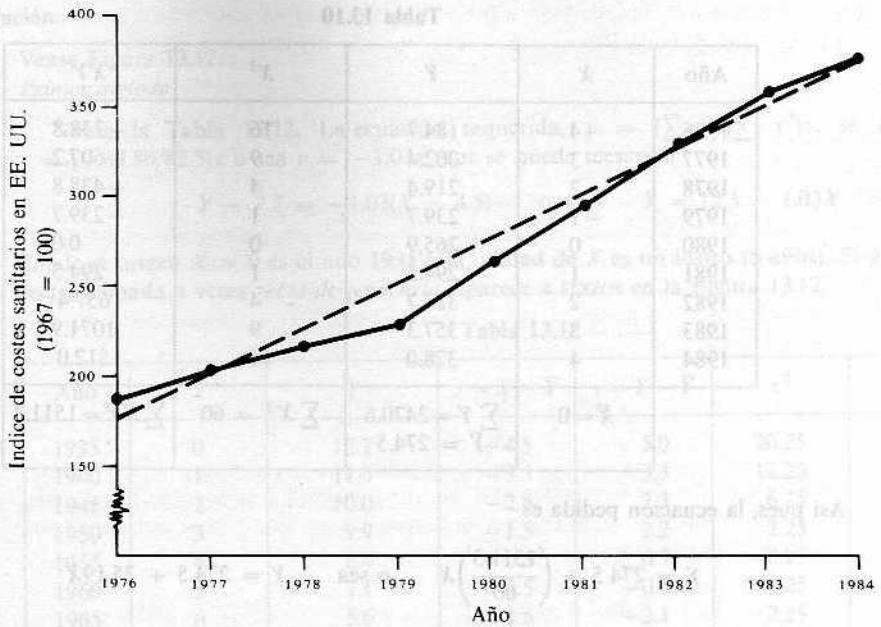


Figura 13.11.

en trazos en la Figura 13.11. La ecuación se llama *ecuación de tendencia*, y los valores de Y calculados para diversos valores de X se llaman *valores de tendencia*.

Segundo método

Si asignamos valores de X a los años 1976-1984 de manera que $\sum X = 0$, la ecuación de la recta de mínimos cuadrados se puede poner como

$$Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right) X$$

Como hay un número impar de años en los datos, podemos asignar $X = 0$ al año central, 1980; asignamos $X = 1, 2, 3$ y 4 a los años sucesivos; y asignamos $X = -1, -2, -3$ y -4 a los años que preceden a este año central. El resultado se ve en la columna 2 de la Tabla 13.10 y es equivalente a usar la columna 4 de la tabla para el primer método.

El año central 1980 se llama el *origen*; supondremos además que los valores de Y se refieren al 1 de julio de cada año. Así pues, $X = 0$ corresponde al 1 de julio de 1980; $X = -1$ al 1 de julio de 1979, etc. Los cálculos se resumen en la Tabla 13.10.

Tabla 13.10

Año	X	Y	X^2	XY
1976	−4	184.7	16	−738.8
1977	−3	202.4	9	−607.2
1978	−2	219.4	4	−438.8
1979	−1	239.7	1	−239.7
1980	0	265.9	0	0.0
1981	1	294.5	1	294.5
1982	2	328.7	4	657.4
1983	3	357.3	9	1071.9
1984	4	378.0	16	1512.0
	$\bar{X}=0$	$\sum Y=2470.6$ $\bar{Y}=274.5$	$\sum X^2=60$	$\sum XY=1511.3$

Así pues, la ecuación pedida es

$$Y = 274.5 + \left(\frac{1511.3}{60}\right)X \quad \text{o sea} \quad Y = 274.5 + 25.19X$$

donde el origen $X = 0$ es el año 1980 y la unidad de X es 1 año. Para desplazar el origen a 1976, 4 años antes, sustituimos X por $X - 4$, con lo que se llega a la ecuación $Y = 274.5 + 25.19(X - 4)$, o $Y = 173.7 + 25.19X$, como en el primer método.

El segundo método es mejor que el primero porque reduce el trabajo de cálculo. Sin embargo, mientras el primer método es aplicable en todos los casos, el segundo exige modificaciones en el caso de un número de años par en los datos. Para tal modificación, ver el segundo método del Problema 13.20(*b*).

(*c*) Usar la ecuación de tendencia $Y = 173.7 + 25.19X$, donde $X = 0$ corresponde al año 1976. Entonces el año 1985 corresponde a $X = 9$, luego el valor de Y para 1985 es $Y = 173.7 + 25.19(9) = 400.4$.

El mismo resultado se puede obtener de la ecuación de tendencia $Y = 274.5 + 25.19X$, donde el origen $X = 0$ corresponde al año 1980, haciendo $X = 5$.

(*d*) Usando la ecuación de tendencia $Y = 173.7 + 25.19X$, con $X = -1$, encontramos el valor $Y = 173.7 + 25.19(-1) = 148.5$.

13.20. La Tabla 13.11 indica el censo de trabajadores agrícolas en EE. UU. los años 1935, 1940, 1945, ..., 1980, en millones.

(*a*) Representar los datos gráficamente.

(*b*) Hallar una ecuación para la recta de mínimos cuadrados que ajuste esos datos.

(*c*) Predecir el censo de trabajadores agrícolas en los años 1990 y 2000, suponiendo que la tendencia se mantenga.

Tabla 13.11

Año	1935	1940	1945	1950	1955	1960	1965	1970	1975	1980
Trabajadores agrícolas en EE. UU. (millones)	12.7	11.0	10.0	9.9	8.4	7.1	5.6	4.5	4.3	3.7

Fuente: U.S. Department of Agriculture.

Solución

(*a*) Véase Figura 13.12.

(*b*) *Primer método*

Véase la Tabla 13.12. La ecuación requerida, $y = (\sum xy/\sum x^2)x$, se convierte en $y = (-84.80/82.5)x$ o sea $y = -1{,}03x$, que se puede reescribir

$$Y - 7.7 = -1.03(X - 4.5) \qquad \text{o sea} \qquad Y = 12.3 - 1.03X$$

donde el origen $X = 0$ es el año 1935 y la unidad de X es un lustro (5 años). El gráfico de esta recta, llamada a veces *recta de tendencia*, aparece a trazos en la Figura 13.12.

Tabla 13.12

Año	X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy
1935	0	12.7	−4.5	5.0	20.25	−22.50
1940	1	11.0	−3.5	3.3	12.25	−11.55
1945	2	10.0	−2.5	2.3	6.25	−5.75
1950	3	9.9	−1.5	2.2	2.25	−3.30
1955	4	8.4	−0.5	0.7	0.25	−0.35
1960	5	7.1	0.5	−0.6	0.25	−0.30
1965	6	5.6	1.5	−2.1	2.25	−3.15
1970	7	4.5	2.5	−3.2	6.25	−8.00
1975	8	4.3	3.5	−3.4	12.25	−11.90
1980	9	3.7	4.5	−4.0	20.25	−18.00
	$\sum X = 45$	$\sum Y = 77.2$			$\sum x^2 = 82.5$	$\sum xy = -84.80$
	$\bar{X} = 4.5$	$\bar{Y} = 7.7$				

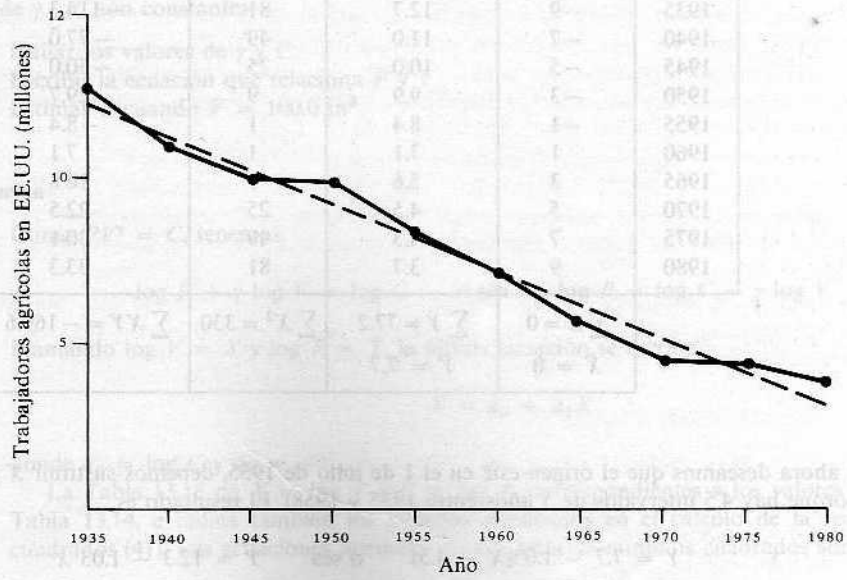


Figura 13.12.

Segundo método

En este método queremos asignar valores de X a los años de modo que $\sum X = 0$. Como hay un número par de años, no hay año central y no se puede usar el segundo método del Problema 13.19(*b*). No obstante, podemos asociar los números -0.5 y 0.5 a los años centrales, 1955 y 1960, de manera que 1965, 1970, 1975 y 1980 están representados por 1.5, 2.5, 3.5 y 4.5 y 1950, 1945, 1940 y 1935 lo están por -1.5, -2.5, -3.5 y -4.5. Esto viene a ser esencialmente la columna 4 de la Tabla 13.12.

Además, para evitar fracciones, doblamos esos valores, obteniendo la columna 2 de la Tabla 13.13. Nótese que con estos valores de X el origen $X = 0$ está a medio camino entre el 1 de julio de 1955 y el 1 de julio de 1960, que es el 1 de enero de 1958 o el 31 de diciembre de 1957. La unidad de X es medio lustro, o sea 2.5 años. Como $\bar{X} = 0$, la ecuación pedida tiene la forma $Y = \bar{Y} + (\sum XY/\sum X^2)X$, que da (véase Tabla 13.13)

$$Y = 7.7 + \left(\frac{-169.6}{330}\right)X \qquad \text{o sea} \qquad Y = 7.7 - 0.514X$$

donde el origen $X = 0$ corresponde al 1 de enero de 1958, y X se mide en unidades de 2.5 años.

Si queremos medir X en intervalos de 5 años en vez de 2.5 años, debemos reemplazar X por $2X$, con lo que la ecuación es

$$Y = 7.7 - 1.028X \qquad \text{o sea} \qquad Y = 7.7 - 1.03X$$

donde el origen es el 1 de enero de 1958, y X se mide en unidades de 5 años.

Tabla 13.13

Año	X	Y	X^2	XY
1935	-9	12.7	81	-114.3
1940	-7	11.0	49	-77.0
1945	-5	10.0	25	-50.0
1950	-3	9.9	9	-29.7
1955	-1	8.4	1	-8.4
1960	1	7.1	1	7.1
1965	3	5.6	9	16.8
1970	5	4.5	25	22.5
1975	7	4.3	49	30.1
1980	9	3.7	81	33.3
	$\sum X = 0$ $\bar{X} = 0$	$\sum Y = 77.2$ $\bar{Y} = 7.7$	$\sum X^2 = 330$	$\sum XY = -169.6$

Si ahora deseamos que el origen esté en el 1 de julio de 1935, debemos sustituir X por $X - 4.5$ (porque hay 4.5 intervalos de 5 años entre 1935 y 1958). El resultado es

$$Y = 7.7 - 1.03(X - 4.5) \qquad \text{o sea} \qquad Y = 12.3 - 1.03X$$

Esto coincide con la ecuación obtenida en el primer método.

(*c*) Usando el primer método en la parte (*b*), los años 1990 y 2000 corresponden a $X = 11$ y $X = 13$, respectivamente. Entonces

$$Y = 12.3 - 1.03X = 12.3 - 1.03(11) = 0.97 \text{ millones en } 1990$$
$$Y = 12.3 - 1.03X = 12.3 - 1.03(13) = -1.09 \text{ millones en } 2000$$

Mientras el primer resultado de un millón de trabajadores agrícolas en 1990 es posible, especialmente a la vista de las nuevas tecnologías y de las importaciones agrícolas, el segundo resultado es claramente imposible. Hemos de concluir que la tendencia que muestra la Tabla 13.13 no se mantendrá por mucho tiempo.

ECUACIONES NO LINEALES REDUCIBLES A FORMA LINEAL

13.21. La Tabla 13.14 presenta valores experimentales de la presión P de una masa dada de gas correspondiente a varios valores del volumen V.

Tabla 13.14

Volumen V en pulgadas cúbicas (in^3)	54.3	61.8	72.4	88.7	118.6	194.0
Presión P en libras por pulgada cuadrada (lb/in^2)	61.2	49.5	37.6	28.4	19.2	10.1

De acuerdo con la Termodinámica, existe una relación del tipo $PV^\gamma = C$ entre las variables P y V, donde γ y C son constantes.

(*a*) Hallar los valores de γ y C.
(*b*) Escribir la ecuación que relaciona P y V.
(*c*) Estimar P cuando $V = 100.0$ in^3.

Solución

Como $PV^\gamma = C$, tenemos

$$\log P + \gamma \log V = \log C \quad \text{o sea} \quad \log P = \log C - \gamma \log V$$

Llamando $\log V = X$ y $\log P = Y$, la última ecuación se escribe

$$Y = a_0 + a_1 X \tag{41}$$

donde $a_0 = \log C$ y $a_1 = -\gamma$.

La Tabla 13.15 da $X = \log V$ e $Y = \log P$, correspondientes a los valores de V y P de la Tabla 13.14, e indica también los cálculos implicados en el cálculo de la recta de mínimos cuadrados (41). Las ecuaciones normales de esa recta de mínimos cuadrados son

$$\sum Y = a_0 N + a_1 \sum X \quad \text{y} \quad \sum XY = a_0 \sum X + a_1 \sum X^2$$

Tabla 13.15

$X = \log V$	$Y = \log P$	X^2	XY
1.7348	1.7868	3.0095	3.0997
1.7910	1.6946	3.2077	3.0350
1.8597	1.5752	3.4585	2.9294
1.9479	1.4533	3.7943	2.8309
2.0741	1.2833	4.3019	2.6617
2.2878	1.0043	5.2340	2.2976
$\sum X = 11.6953$	$\sum Y = 8.7975$	$\sum X^2 = 23.0059$	$\sum XY = 16.8543$

de donde

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N\sum X^2 - (\sum X)^2} = 4.20 \qquad a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = -1.40$$

Luego $Y = 4.20 - 1.40\,X$.

(*a*) Como $a_0 = 4.20 = \log C$ y $a_1 = -1.40 = -\gamma$, $C = 1.60 \times 10^4$ y $\gamma = 1.40$.
(*b*) La ecuación requerida en términos de P y V puede escribirse $PV^{1.40} = 16{,}000$.
(*c*) Cuando $V = 100$, $X = \log V = 2$ e $Y = \log P = 4.20 - 1.40(2) = 1.40$. Entonces $P =$ antilog $1.40 = 25.1$ lb/in^2.

13.22. Resolver el Problema 13.21 representando los datos en papel log-log.

Solución

Obtenemos primero un punto para cada par de valores de la presión P y del volumen V en la Tabla 13.14, y marcamos esos puntos en papel log-log, como indica la Figura 13.13. Entonces trazamos una recta que aproxime esos puntos (la recta de la figura esté trazada «a mano»). El gráfico resultante muestra que hay una relación lineal entre $\log P$ y $\log V$ representable por la ecuación

$$\log P = a_0 + a_1 \log V \qquad \text{o sea} \qquad Y = a_0 + a_1 X$$

La pendiente a_1, que es negativa en este caso, viene dada numéricamente por el cociente de longitudes de AB y AC (usando una unidad de longitud apropiada). La medida en este caso da $a_1 = -1.4$.

Para hallar a_0, se necesita un punto sobre la recta. Por ejemplo, cuando $V = 100$, $P = 25$ en el gráfico; por tanto, $a_0 = \log P - a_1 \log V = \log 25 + 1.4 \log 100 = 1.4 + (1.4)(2) = 4.2$, y en consecuencia tenemos $\log P + 1.4 \log V = 4.2$ $\log PV^{1.4} = 4.2$ y $PV^{1.4} = 16{,}000$.

LA PARABOLA DE MINIMOS CUADRADOS

13.23. La Tabla 13.16 da la población de EE. UU. en los años 1880-1980 en intervalos de 10 años.

(*a*) Hallar la ecuación de una parábola de mínimos cuadrados que ajuste los datos.
(*b*) Calcular los valores de tendencia para los años dados en la Tabla 13.16, y compararlos con los verdaderos.
(*c*) Estimar la población en 1990 y 2000.
(*d*) Estimar la población en 1870 y 1860, y comparar con los valores reales (véase página 18).

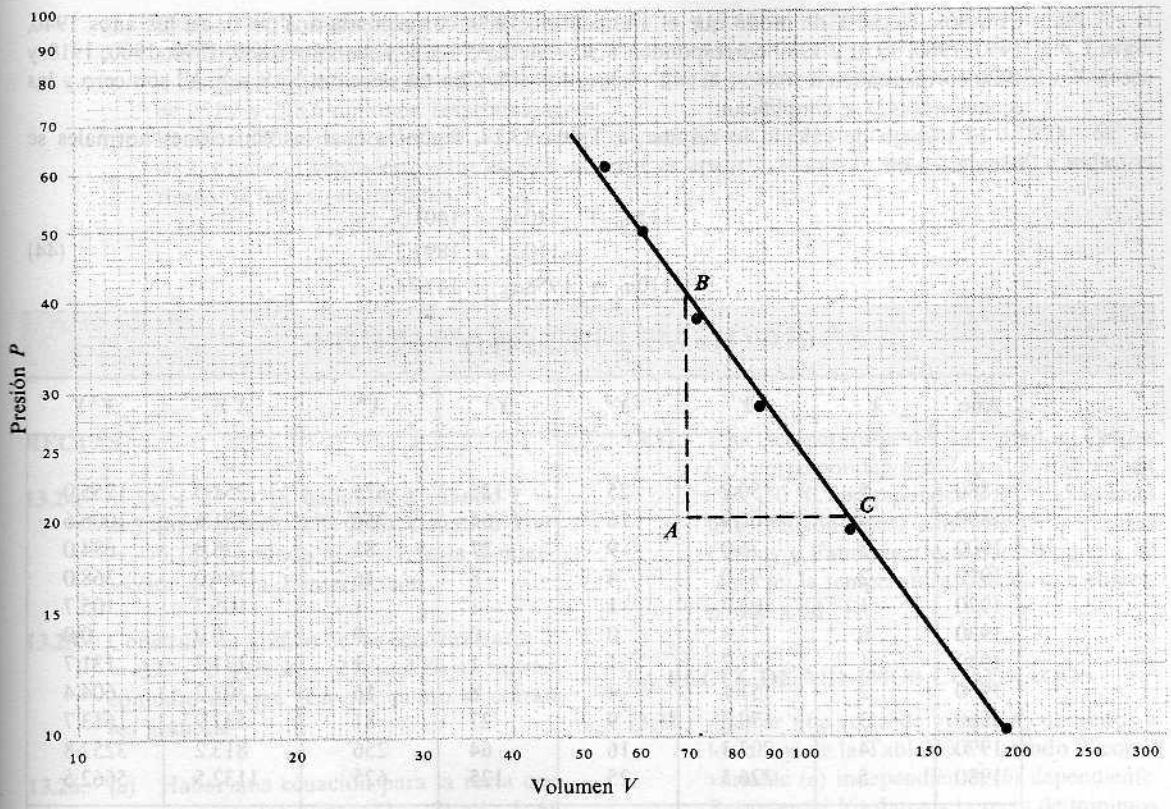


Figura 13.13.

Tabla 13.16

Año	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980
Población de EE. UU. (millones)	50.2	62.9	76.0	92.0	105.7	122.8	131.7	151.1	179.3	203.3	226.5

Fuente: U.S. Bureau of the Census.

Solución

(*a*) Sean X, Y, respectivamente, el año y la población en ese año. La ecuación de una parábola de mínimos cuadrados que ajuste los datos es

$$Y = a_0 + a_1 X + a_2 X^2 \tag{42}$$

donde a_0, a_1 y a_2 se deducen de las ecuaciones normales

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X + a_2 \sum X_2 \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum X^3 \\ \sum X^2 Y &= a_0 \sum X^2 + a_1 \sum X^3 + a_2 \sum X^4 \end{aligned} \tag{43}$$

Conviene elegir X de modo que el año central, 1930, corresponda a $X = 0$; así los años 1940, 1950, 1960, 1970 y 1980 corresponden a $X = 1, 2, 3, 4$ y 5; y los años 1880, 1890, 1900, 1910 y 1920 corresponden a $X = -1, -2, -3, -4$ y -5. Con tal elección, $\sum X$ y $\sum X^3$ son cero y las ecuaciones (43) se simplifican.

El trabajo de cálculo lo resume la Tabla 13.17, según la cual las ecuaciones normales se convierten en

$$\begin{aligned} 11a_0 + 110a_2 &= 1401.5 \\ 110a_1 &= 1897.2 \\ 110a_0 + 1958a_2 &= 14{,}684.2 \end{aligned} \tag{44}$$

Tabla 13.17

Año	X	Y	X^2	X^3	X^4	XY	X^2Y
1880	−5	50.2	25	−125	625	−251.0	1255.0
1890	−4	62.9	16	−64	256	−251.6	1006.6
1900	−3	76.0	9	−27	81	−228.0	684.0
1910	−2	92.0	4	−8	16	−184.0	368.0
1920	−1	105.7	1	−1	1	−105.7	105.7
1930	0	122.8	0	0	0	0.0	0.0
1940	1	131.7	1	1	1	131.7	131.7
1950	2	151.1	4	8	16	302.2	604.4
1960	3	179.3	9	27	81	537.9	1613.7
1970	4	203.3	16	64	256	813.2	3252.8
1980	5	226.5	25	125	625	1132.5	5662.5
	$\sum X = 0$	$\sum Y = 1401.5$	$\sum X^2 = 110$	$\sum X^3 = 0$	$\sum X^4 = 1958$	$\sum XY = 1897.2$	$\sum X^2Y = 14{,}684.2$

De la segunda ecuación en (44), $a_1 = 17.25$; de la primera, $a_0 = 119.61$; y de la tercera, $a_2 = 0.7800$. Luego la ecuación buscada es

$$Y = 119.61 + 17.25X + 0.7800X^2 \tag{45}$$

donde el origen $X = 0$ es el 1 de julio de 1930 y la unidad de X son 10 años.

(*b*) Los valores de tendencia se obtienen haciendo $X = -5, -4, -3, -2, -1, 0, 1, 2, 3, 4$ y 5 en la ecuación (45). Estos valores de tendencia, junto con los valores reales, se recogen en la Tabla 13.18. Vemos que el acuerdo es bueno.

Tabla 13.18

Año	$X=-5$ 1880	$X=-4$ 1890	$X=-3$ 1900	$X=-2$ 1910	$X=-1$ 1920	$X=0$ 1930	$X=1$ 1940	$X=2$ 1950	$X=3$ 1960	$X=4$ 1970	$X=5$ 1980
Valor de tendencia	52.9	63.1	74.9	88.2	103.1	119.6	137.6	157.2	178.4	201.1	225.4
Valor real	50.2	62.9	76.0	92.0	105.7	122.8	131.7	151.1	179.3	203.3	226.5

(*c*) El año 1990 corresponde a $X = 6$, para el que $Y = 119{,}61 + 17.25(6) + 0.7800(6)^2 = 251.2$, y el año 2000 corresponde a $X = 7$, para el que $Y = 119.61 + 17.25(7) + 0.7800(7)^2 = 278.6$. Luego, si continúa la tendencia actual, podemos esperar que la población de EE. UU. en 1990 y 2000 sea de 251.2 y 278.6 millones, respectivamente.

(*d*) El año 1870 corresponde a $X = -6$, para el cual $Y = 119.61 + 17.25(-6) + 0.7800(-6)^2 = 44.2$. Como el verdadero valor es 39.8, el error es aproximadamente del 10 por 100 e indica el riesgo de las extrapolaciones.

PROBLEMAS SUPLEMENTARIOS

RECTAS

13.24. Si $3X + 2Y = 18$, hallar (*a*) X cuando $Y = 3$, (*b*) Y cuando $X = 2$, (*c*) X cuando $Y = -5$, (*d*) Y cuando $X = -1$, (*e*) la X-intersección y (*f*) la Y-intersección.

13.25. Construir un gráfico de las ecuaciones (*a*) $Y = 3X - 5$ y (*b*) $X + 2Y = 4$ en un mismo conjunto de ejes. ¿En qué punto se cortan los gráficos?

13.26. (*a*) Hallar una ecuación para la recta que pasa por los puntos $(3, -2)$ y $(-1, 6)$.
(*b*) Determinar sus intersecciones con los ejes.
(*c*) Hallar el valor de Y correspondiente a $X = 3$ y $X = 5$.
(*d*) Verificar directamente sobre el gráfico las respuestas de (*a*), (*b*) y (*c*).

13.27. Hallar una ecuación para la recta de pendiente $2/3$ y cuya Y-intersección es -3.

13.28. (*a*) Hallar la pendiente y la Y-intersección de la recta $3X - 5Y = 20$.
(*b*) ¿Cuál es la ecuación de una recta paralela a la de la parte (*a*) y que pasa por el punto $(2, -1)$?

13.29. Hallar (*a*) la pendiente, (*b*) la Y-intersección y (*c*) la ecuación de la recta que pasa por los puntos (5, 4) y (2, 8).

13.30. Hallar la ecuación de una recta cuyas intersecciones X e Y son 3 y -5, respectivamente.

13.31. Una temperatura de 100 grados Celsius (°C) corresponden a 212 grados Fahrenheit (°F), y 0 °C corresponden a 32 °F. Supuesta una relación lineal entre las temperaturas Celsius y Fahrenheit que corresponde a 80 C, y (*c*) la temperatura Celsius que corresponde a 68 °F.

LA RECTA DE MINIMOS CUADRADOS

13.32. Ajustar una recta de mínimos cuadrados a los datos de la Tabla 13.19 usando X como variable (*a*) independiente, (*b*) dependiente. Representar los datos y la recta de mínimos cuadrados sobre unos mismos ejes de coordenadas.

Tabla 13.19

X	3	5	6	8	9	11
Y	2	3	4	6	5	8

13.33. Para los datos del Problema 13.32, hallar (*a*) los valores de Y cuando $X = 5$ y $X = 12$ y (*b*) el valor de X cuando $Y = 7$.

13.34. (*a*) Obtener una ecuación, por el método «a mano», para una recta que ajuste los datos del Problema 13.32.
(*b*) Usando el resultado de (*a*), resolver el Problema 13.33.

13.35. La Tabla 13.20. presenta las notas en Algebra y Física de 10 estudiantes elegidos al azar entre un grupo muy numeroso.

Tabla 13.20

Algebra (X)	Física (Y)
75	82
80	78
93	86
65	72
87	91
71	80
98	95
68	72
84	89
77	74

(a) Representar los datos.
(b) Hallar una recta de mínimos cuadrados que ajuste los datos, usando X como variable independiente.
(c) Hallar una recta de mínimos cuadrados que ajuste los datos, usando Y como variable independiente.
(d) Si un estudiante tiene 75 en Algebra, ¿cuál es su nota esperada en Física?
(e) Si un estudiante tiene 95 en Física, ¿cuál es su nota esperada en Algebra?

13.36. La Tabla 13.21 muestra la tasa de natalidad en EE. UU. durante 1920-1980, en intervalos de 10 años.
(a) Representar los datos.
(b) Hallar una recta de mínimos cuadrados que ajuste esos datos.
(c) Calcular los valores de tendencia y compararlos con los verdaderos.

Tabla 13.21

Año	Tasa de natalidad por cada 1000 habitantes
1920	27.7
1930	21.3
1940	19.4
1950	24.1
1960	23.7
1970	18.4
1980	15.9

Fuente: National Center for Health Statistics.

(d) Predecir la tasa de natalidad en los años 1990 y 2000, y discutir las posibles causas de error en tal predicción.

13.37. La Tabla 13.22 recoge los porcentajes de la población de EE. UU. de 65 años o más, para los años 1890-1980.
(a) Representar los datos.
(b) Ajustar los datos con una recta de mínimos cuadrados.
(c) Calcular los valores de tendencia y compararlos con los verdaderos.
(d) Predecir el porcentaje de esas edades para los años 1990 y 2000, y discutir las posibles causas de error en esa predicción.
(e) ¿Cuándo se esperaría que el porcentaje alcance 25, 35 y 50 % y qué hipótesis hay que hacer para responder?

Tabla 13.22

Año	Porcentaje
1890	3.84
1900	4.05
1910	4.29
1920	4.67
1930	5.40
1940	6.85
1950	8.12
1960	9.30
1970	9.89
1980	11.35

Fuente: U.S. Bureau of the Census.

CURVAS DE MINIMOS CUADRADOS

13.38. Ajustar una parábola de mínimos cuadrados, $Y = a_0 + a_1X + a_2X^2$, a los datos de la Tabla 13.23.

Tabla 13.23

X	Y
0	2.4
1	2.1
2	3.2
3	5.6
4	9.3
5	14.6
6	21.9

13.39. El tiempo necesario para detener un coche tras percibir un peligro es el tiempo de reacción (el tiempo entre la percepción del peligro y la aplicación de los frenos) más el tiempo de frenada (lo que tarda en detenerse bajo la acción de los frenos). La Tabla 13.24. da la distancia D (en pies) que recorre antes de pararse un coche que circula a V millas por hora, a partir del instante en que se ha percibido el peligro.

(*a*) Representar los datos.
(*b*) Ajustar una parábola de mínimos cuadrados de la forma $D = a_0 + a_1 V + a_2 V^2$ a los datos.
(*c*) Estimar D cuando $V = 45$ mi/h y 80 mi/h.

Tabla 13.24

Velocidad V (mi/h)	Distancia de frenado D (pies)
20	54
30	90
40	138
50	206
60	292
70	396

13.40. La Tabla 13.25 presenta las poblaciones masculina y femenina de EE. UU. durante 1920-1980.

(*a*) Representar las diferencias entre esas dos poblaciones.
(*b*) Usando una ecuación apropiada, hallar una curva de mínimos cuadrados que ajuste esos datos.
(*c*) Estimar la diferencia para los años 1990 y 2000.
(*d*) Determinar en qué año habrá una proporción 2:1 de mujeres a hombres. Al determinar esto, ¿qué hipótesis hay que hacer?

Tabla 13.25

Año	Población masculina	Población femenina
1920	53.90	51.81
1930	62.14	60.64
1940	66.06	65.61
1950	75.19	76.14
1960	88.33	90.99
1970	98.93	104.31
1980	110.05	116.49

Fuente: U.S. Bureau of the Census.

13.41. Resolver el Problema 13.40 usando el cociente en vez de la diferencia entre poblaciones.

13.42. Resolver el Problema 13.37 con una parábola de mínimos cuadrados y comparar los resultados.

13.43. El número de bacterias por unidad de volumen en un cultivo tras X horas viene dado en la Tabla 13.26.

(*a*) Representar los datos en papel semilog, usando escala logarítmica para Y y escala aritmética para X.
(*b*) Ajustar una curva de mínimos cuadrados de la forma $Y = ab^X$ a los datos y explicar por qué esa ecuación particular debe dar buenos resultados.
(*c*) Comparar los valores de Y obtenidos de esa ecuación con los valores reales.
(*d*) Estimar el valor de Y cuando $X = 7$.

Tabla 13.26

Número de horas (X)	Numero de bacterias por unidad de volumen (Y)
0	32
1	47
2	65
3	92
4	132
5	190
6	275

13.44. En el Problema 13.43, mostrar cómo un gráfico en papel semilog puede ser utilizado para la obtención de la ecuación requerida sin recurrir al método de mínimos cuadrados.

CAPITULO 14

Teoría de la correlación

CORRELACION Y REGRESION

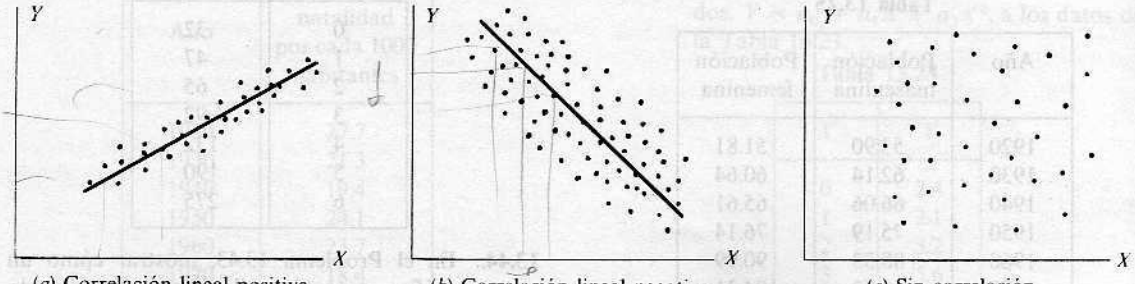
En el último capítulo hemos considerado el problema de la *regresión* o *estimación* de una variable (la variable dependiente) de una o más variables relacionadas (las variables independientes). En este capítulo tratamos el problema cercano de la *correlación*, o grado de interconexión entre variables, que intenta determinar *con qué precisión* describe o explica la relación entre variables una ecuación lineal o de cualquier otro tipo.

Si todos los valores de las variables satisfacen una ecuación exactamente, decimos que las variables están *perfectamente correlacionadas* o que hay *correlación perfecta* entre ellas. Así, las circunferencias C y los radios r de todos los círculos están perfectamente correlacionados porque $C = 2\pi r$. Si se lanzan dos dados 100 veces, no hay relación entre las puntuaciones de ambos dados (a menos que estén trucados) es decir, *no están en correlación*. Variables tales como el peso y la altura de las personas tienen una cierta correlación.

Cuando sólo están en juego dos variables, hablamos de *correlación simple* y *regresión simple*. En otro caso, se habla de *correlación múltiple* y *regresión múltiple*. Este capítulo considera sólo correlación simple. La correlación y regresión múltiples se analizarán en el Capítulo 15.

CORRELACION LINEAL

Si X e Y son las dos variables en cuestión, un *diagrama de dispersión* muestra la localización de los puntos (X, Y) sobre un sistema rectangular de coordenadas. Si todos los puntos del diagrama de dispersión parecen estar en una recta, como en las Figuras 14.1(*a*) y 14.1(*b*), la correlación se llama *lineal*. En tales casos, como ya hemos visto en el Capítulo 13, una ecuación lineal es adecuada a efectos de regresión (o estimación).



(*a*) Correlación lineal positiva (*b*) Correlación lineal negativa (*c*) Sin correlación

Figura 14.1.

Si Y tiende a crecer cuando X crece, como en la Figura 14.1(a), la *correlación* se dice *positiva*, o *directa*. Si Y tiende a decrecer cuando X crece, como en la Figura 14.1(b), la *correlación* se dice *negativa*, o *inversa*.

Si todos los puntos parecen estar sobre una cierta curva, la correlación se llama *no lineal*, y una ecuación no lineal será apropiada para la regresión, como hemos visto en el Capítulo 13. Es claro que la correlación no lineal puede ser positiva o negativa.

Si no hay relación entre las variables, como en la Figura 14.1(c), decimos que *no hay correlación* entre ellas.

MEDIDAS DE CORRELACION

Podemos determinar de forma *cualitativa* con qué precisión describe una curva dada la relación entre variables por observación directa del propio diagrama de dispersión. Por ejemplo, se ve que una recta es mucho más conveniente para describir la relación entre X e Y para los datos de la Figura 14.1(a) que para los de la Figura 14.1(b), porque hay menos dispersión relativa a la recta en la Figura 14.1(a).

Si hemos de enfrentarnos al problema de la dispersión de datos muestrales respecto de rectas o curvas de modo *cuantitativo*, será necesario definir *medidas de correlación*.

LA RECTA DE REGRESION DE MINIMOS CUADRADOS

Consideremos primero el problema de ver con qué calidad explica una recta la relación entre dos variables. Para ello, necesitaremos las ecuaciones de la recta de regresión de mínimos cuadrados obtenidas en el Capítulo 13. Tal como vimos, la recta de regresión de mínimos cuadrados de Y sobre X es

$$Y = a_0 + a_1 X \tag{1}$$

donde a_0 y a_1 se obtienen de las ecuaciones normales

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 \end{aligned} \tag{2}$$

de las que se deduce

$$\begin{aligned} a_0 &= \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N\sum X^2 - (\sum X)^2} \\ a_1 &= \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} \end{aligned} \tag{3}$$

Análogamente, la recta de regresión de X sobre Y es

$$X = b_0 + b_1 Y \tag{4}$$

donde b_0 y b_1 se obtienen de las ecuaciones normales

$$\begin{aligned} \sum X &= b_0 N + b_1 \sum Y \\ \sum XY &= b_0 \sum X + b_1 \sum Y^2 \end{aligned} \tag{5}$$

obteniéndose

$$b_0 = \frac{(\sum X)(\sum Y^2) - (\sum Y)(\sum XY)}{N\sum Y^2 - (\sum Y)^2}$$

$$b_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum Y^2 - (\sum Y)^2} \tag{6}$$

Las ecuaciones (1) y (4) se pueden escribir, respectivamente, como

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \quad \text{y} \quad x = \left(\frac{\sum xy}{\sum y^2}\right)y \tag{7}$$

donde $x = X - \bar{X}$ e $y = Y - \bar{Y}$.

Las ecuaciones de regresión son idénticas si y sólo si todos los puntos del diagrama de dispersión están en una recta. En tal caso hay una *correlación lineal perfecta* entre X e Y.

ERROR TIPICO DE ESTIMACION

Si denotamos por Y_{est} el valor de Y para valores dados de X, tal como se estima a partir de la ecuación (1), una medida de la dispersión respecto de la recta de regresión de Y sobre X viene proporcionada por la cantidad

$$s_{Y.X} = \sqrt{\frac{\sum (Y - Y_{est})^2}{N}} \tag{8}$$

que se llama el *error típico de estimación* de Y sobre X.

Si se usa la recta de regresión (4), un error típico de estimación análogo de la estimación de X sobre Y se define como

$$s_{X.Y} = \sqrt{\frac{\sum (X - X_{est})^2}{N}} \tag{9}$$

En general, $s_{Y.X} \neq s_{X.Y}$.

La ecuación (8) se puede formular

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N} \tag{10}$$

que puede ser más conveniente para el cálculo (véase Prob. 14.3). Existe una expresión similar para (9).

El error típico de estimación tiene propiedades análogas a las de la desviación típica. Por ejemplo, si construimos rectas paralelas a la de regresión de Y sobre X a distancias verticales respectivas $s_{Y.X}$, $2s_{Y.X}$, y $3s_{Y.X}$ de ella, encontraremos, si N es lo bastante grande, que estarían incluidos entre esas rectas aproximadamente el 68%, 95% y 99.7% de los puntos muestrales.

Igual que la desviación típica modificada

$$\hat{s} = \sqrt{\frac{N}{N-1}}\, s$$

era útil para pequeñas muestras, será útil un error típico de estimación modificado dado por

$$\hat{s}_{Y.X} = \sqrt{\frac{N}{N-2}}\, s_{Y.X}$$

Por esta razón, algunos estadísticos prefieren definir (8) ó (9) con $N - 2$ en lugar de N en el denominador.

VARIACION EXPLICADA Y VARIACION INEXPLICADA

La *variación total* de Y se define como $\sum(Y - \bar{Y})^2$: esto es, la suma de los cuadrados de las desviaciones de los valores de Y respecto de la media Y. Como se ve en el Problema 14.7, eso se puede escribir

$$\sum(Y - \bar{Y})^2 = \sum(Y - Y_{est})^2 + \sum(Y_{est} - \bar{Y})^2 \tag{11}$$

El primer término de la derecha en la ecuación (11) se llama la *variación explicada*, mientras que el segundo se llama la *variación inexplicada* (porque las desviaciones $Y_{est} - \bar{Y}$ tienen un esquema definido mientras las desviaciones $Y - Y_{est}$ se comportan de modo caótico, impredictible). Resultados similares son válidos para la variable X.

COEFICIENTE DE CORRELACION

El coeficiente entre la variación explicada y la variación total se llama *coeficiente de determinación*. Si la variación explicada es cero (o sea, toda la variación es variación inexplicada), ese cociente es 0. Si la variación inexplicada es cero (o sea, toda la variación es explicada), el cociente es 1. En los demás casos, está entre 0 y 1. Como nunca es negativo, denotaremos ese cociente por r. La cantidad r, llamada *coeficiente de correlación*, viene dada por

$$r = \pm\sqrt{\frac{\text{variación explicada}}{\text{variación total}}} = \pm\sqrt{\frac{\sum(Y_{est} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}} \tag{12}$$

y varía entre -1 y $+1$. Se usan los signos $+$ y $-$ para las correlaciones positivas y negativas, respectivamente. Nótese que r es una cantidad adimensional, es decir, no depende de las unidades empleadas.

Usando las ecuaciones (8) y (11) y el hecho de que la desviación típica de Y es

$$s_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}} \tag{13}$$

encontramos que la ecuación (12) se puede escribir, independientemente del signo, como

$$r = \sqrt{1 - \frac{s_{Y.X}^2}{s_Y^2}} \quad \text{o sea} \quad s_{Y.X} = s_Y\sqrt{1 - r^2} \tag{14}$$

Ecuaciones similares existen cuando se intercambian X e Y.

Para el caso de correlación lineal, la cantidad r es la misma tanto si es X como Y la variable independiente. Así pues, r es una buena medida de la correlación lineal entre dos variables.

OBSERVACIONES SOBRE EL COEFICIENTE DE CORRELACION

Las definiciones del coeficiente de correlación en (12) y (14) son completamente generales y se pueden usar tanto para relaciones lineales como no lineales, con la única diferencia de que Y se calcula de una ecuación de regresión no lineal en lugar de una lineal, y que se omiten los signos $+$ y $-$. En tal caso, la ecuación (8), que define el error típico de estimación, es perfectamente general. La (10), sin embargo, que sólo se aplica a regresión lineal, debe ser modificada. Si, por ejemplo, la ecuación de estimación es

$$Y = a_0 + a_1X + a_2X^2 + \cdots + a_{n-1}X^{n-1} \tag{15}$$

la ecuación (10) queda sustituida por

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0\sum Y - a_1\sum XY - \cdots - a_{n-1}\sum X^{n-1}Y}{N} \tag{16}$$

En tal caso el *error típico de estimación modificado* (discutido previamente en este capítulo) es

$$\hat{s}_{Y.X} = \sqrt{\frac{N}{N-n}}\, s_{Y.X}$$

donde la cantidad $N - n$ se llama el número de *grados de libertad*.

Hay que insistir en que en todo caso el valor calculado de r mide el grado de relación con referencia al tipo de ecuación que se adopta. Así pues, si se supone una ecuación lineal y (12) o (14) dan un valor de r próximo a cero, eso significa que no hay apenas correlación lineal entre las variables. No obstante, no quiere decir que no haya correlación en absoluto, pues puede haber una fuerte *correlación no lineal* entre ellas. En otras palabras, el coeficiente de correlación mide la

bondad del ajuste entre: (1) la ecuación adoptada y (2) los datos. A menos que se especifique lo contrario, el término *coeficiente de correlación* se usará para el *coeficiente de correlación lineal.*

Hemos de hacer constar que un coeficiente de correlación alto (o sea, cercano a 1 ó -1) no indica necesariamente una dependencia directa de las variables. Puede haber una alta correlación entre el número de libros publicados cada año y el número de tormentas cada año. Tales ejemplos constituyen lo que se llama *correlaciones sin sentido*, o *espúreas.*

FORMULAS MOMENTO-PRODUCTO PARA EL COEFICIENTE DE CORRELACION LINEAL

Si se supone una relación lineal entre dos variables, la ecuación (12) se convierte en

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} \tag{17}$$

donde $x = X - \bar{X}$ e $y = Y - \bar{Y}$ (véase Prob. 14.10). Esta fórmula, que da automáticamente el signo apropiado de r, se llama la *fórmula momento-producto* y muestra claramente la simetría entre X e Y.

Si escribimos

$$s_{XY} = \frac{\sum xy}{N} \qquad s_X = \sqrt{\frac{\sum x^2}{N}} \qquad s_Y = \sqrt{\frac{\sum y^2}{N}} \tag{18}$$

entonces s_X y s_Y se reconocen como la desviación típica de las variables X e Y, mientras que s_X^2 y s_Y^2 son sus varianzas. La nueva cantidad s se llama la *covarianza* de X e Y. En términos de símbolos de (18), la fórmula (17) se reescribe

$$r = \frac{s_{XY}}{s_X s_Y} \tag{19}$$

Nótese que r no es sólo independiente de la elección de unidades de X e Y, sino también de la elección del origen.

FORMULAS CORTAS DE CALCULO

La fórmula (17) se puede escribir en la forma equivalente

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}} \tag{20}$$

que se usa con frecuencia al calcular r (véanse Probs. 14.15 y 14.16).

Para datos agrupados como en una *tabla de frecuencias de dos variables*, o en una *distribución de frecuencias de dos variables* (véase Prob. 14.17), conviene usar un *método de compilación* como en los capítulos previos. En tal caso, la fórmula (20) se escribe

$$r = \frac{N\sum f_X u_X u_Y - (\sum f_X u_X)(\sum f_Y u_Y)}{\sqrt{[N\sum f_X u_X^2 - (\sum f_X u_X)^2][N\sum f_Y u_Y^2 - (\sum f_Y u_Y)^2]}} \tag{21}$$

(véase Prob. 14.18). Por conveniencia en los cálculos cuando se recurre a esa fórmula, se usa una *tabla de correlación* (véase Prob. 14.19).

Para datos agrupados, las (18) se expresan

$$s_{XY} = c_X c_Y \left[\frac{\sum f_X u_X u_Y}{N} - \left(\frac{\sum f_X u_X}{N}\right)\left(\frac{\sum f_Y u_Y}{N}\right) \right] \tag{22}$$

$$s_X = c_X \sqrt{\frac{\sum f_X u_X^2}{N} - \left(\frac{\sum f_X u_X}{N}\right)^2} \tag{23}$$

$$s_Y = c_Y \sqrt{\frac{\sum f_Y u_Y^2}{N} - \left(\frac{\sum f_Y u_Y}{N}\right)^2} \tag{24}$$

donde c_X y c_Y son las anchuras de intervalos de clase (supuestas constantes) de las variables X e Y. Nótese que (23) y (24) son equivalentes a la fórmula (11) del Capítulo 4.

La fórmula (19) es equivalente a (21), como se ve sin más que usar (22) a (24).

RECTAS DE REGRESION Y EL COEFICIENTE DE CORRELACION LINEAL

La ecuación de la recta de mínimos cuadrados $Y = a_0 + a_1 X$, la recta de regresión de Y sobre X, se puede escribir

$$Y - \bar{Y} = \frac{rs_Y}{s_X}(X - \bar{X}) \qquad \text{o sea} \qquad y = \frac{rs_Y}{s_X} x \tag{25}$$

Análogamente, la recta de regresión de X sobre Y, $X = b_0 + b_1 Y$, puede expresarse como

$$X - \bar{X} = \frac{rs_X}{s_Y}(Y - \bar{Y}) \qquad \text{o sea} \qquad x = \frac{rs_X}{s_Y} y \tag{26}$$

Las pendientes de las rectas en las ecuaciones (25) y (26) son iguales si y sólo si $r = \pm 1$. En tal caso las dos rectas son idénticas y hay correlación lineal perfecta entre X e Y. Si $r = 0$, las rectas son perpendiculares y no hay correlación lineal entre X e Y. Así pues, el coeficiente de correlación lineal mide la separación de ambas rectas de regresión.

Obsérvese que si (25) y (26) se escriben como $Y = a_0 + a_1 X$ y $X = b_0 + b_1 Y$, respectivamente, entonces $a_1 b_1 = r^2$ (véase Prob. 14.22).

CORRELACION DE SERIES EN EL TIEMPO

Si las variables X e Y dependen del tiempo, es posible que pueda existir una relación entre X e Y aun cuando no sea una dependencia directa y pueda producir «correlación espúrea». El coeficiente

de correlación se obtiene simplemente considerando los pares de valores (X, Y) correspondientes a varios tiempos y procediendo como de costumbre, haciendo uso de las fórmulas anteriores (véase Problema 14.28).

Es posible intentar correlacionar valores de una variable X en ciertos tiempos con valores correspondientes de X en tiempos anteriores. Tales correlaciones se llaman *autocorrelaciones*.

CORRELACION DE ATRIBUTOS

Los métodos descritos en este capítulo no nos capacitan para considerar la correlación de variables que sean de naturaleza no numérica, tales como los *atributos* de individuos (color del pelo, de los ojos, etc.). Para una discusión de la correlación de atributos, véase el Capítulo 12

TEORIA MUESTRAL DE LA CORRELACION

Los N pares de valores (X, Y) de dos variables pueden verse como muestras de una población de todos los pares posibles. Como están en juego dos variables, se llama una *población de dos variables*, que supondremos tiene una *distribución normal de dos variables*.

Podemos pensar en un coeficiente de correlación de población teórico, denotado por ρ, que se estima por el coeficiente de correlación r de la muestra. Contrastes de hipótesis o significación relativos a varios valores de ρ exigen conocer la distribución muestral de r. Para $\rho = 0$ esta distribución es simétrica, y se puede usar un estadístico con distribución de Student. Para $\rho \neq 0$, la distribución es sesgada y en tal caso una transformación debida a Fisher produce un estadístico que es aproximadamente normal. Los siguientes contrastes resumen los procedimientos implicados:

1. **Contraste de hipótesis** $\rho = 0$. Aquí usamos el hecho de que el estadístico

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{27}$$

tiene una distribución de Student con $\nu = N - 2$ grados de libertad (véanse Probs. 14.31 y 14.32).

2. **Contraste de hipótesis** $\rho = \rho_0 \neq 0$. Aquí usamos el hecho de que el estadístico

$$Z = \tfrac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) = 1.1513\,\log_{10}\left(\frac{1+r}{1-r}\right) \tag{28}$$

donde $e = 2.71828...$, está casi normalmente distribuido con media y desviación típica dadas por

$$\mu_Z = \tfrac{1}{2}\log_e\left(\frac{1+\rho_0}{1-\rho_0}\right) = 1.1513\,\log_{10}\left(\frac{1+\rho_0}{1-\rho_0}\right) \quad \sigma_Z = \frac{1}{\sqrt{N-3}} \tag{29}$$

Las ecuaciones (28) y (29) se pueden utilizar también para hallar límites de confianza para el coeficiente de correlación (véanse Probs. 14.33 y 14.34). La ecuación (28) se llama *transformación Z de Fisher*.

3. **Significación de una diferencia entre coeficiente de correlación.** Para determinar si dos coeficientes de correlación, r_1 y r_2, sacados de muestras de tamaños N_1 y N_2, respectivamente,

difieren significativamente uno de otro, calculamos Z_1 y Z_2 correspondientes a r_1 y r_2 usando (28). Y utilizamos entonces el hecho de que el estadístico de contraste

$$z = \frac{Z_1 - Z_2 - \mu_{Z1-Z2}}{\sigma_{Z1-Z2}} \tag{30}$$

donde

$$\mu_{Z1-Z2} = \mu_{Z1} - \mu_{Z2}$$

y

$$\sigma_{Z1-Z2} = \sqrt{\sigma_{Z1}^2 + \sigma_{Z2}^2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$$

está normalmente distribuido (véase Prob. 14.35).

TEORIA MUESTRAL DE LA REGRESION

La ecuación de regresión $Y = a_0 + a_1 X$ se obtiene a partir de los datos de la muestra. A menudo estamos interesados en la correspondiente ecuación de regresión para la población de la que procede el muestreo. He aquí tres contrastes relativos a dicha población:

1. **Contraste de hipótesis** $a_1 = A_1$. Para contrastar la hipótesis de que el coeficiente de regresión a_1 es igual a cierto valor A_1 especificado, usamos el hecho de que el estadístico

$$t = \frac{a_1 - A_1}{s_{Y.X}/s_X}\sqrt{N - 2} \tag{31}$$

tiene distribución de Student con $N - 2$ grados de libertad. Esto se puede también utilizar para hallar intervalos de confianza para los coeficientes de regresión de la población a partir de los valores de la muestra (véanse Probs. 14.36 y 14.37).

2. **Contraste de hipótesis para valores de predicción.** Sea Y_0 la predicción para el valor de Y correspondiente a $X = X_0$ tal como se estima a partir de la ecuación de regresión muestral (o sea $Y_0 = a_0 + a_1 X_0$). Sea Y_p la predicción del valor de Y correspondiente a $X = X_0$ para la población. Entonces el estadístico

$$t = \frac{Y_0 - Y_p}{s_{Y.X}\sqrt{N + 1 + (X_0 - \bar{X})^2/s_X^2}}\sqrt{N - 2} = \frac{Y_0 - Y_p}{\hat{s}_{Y.X}\sqrt{1 + 1/N + (X_0 - \bar{X})^2/(Ns_X^2)}} \tag{32}$$

tiene distribución de Student con $N - 2$ grados de libertad. De donde pueden hallarse límites de confianza para las predicciones de los valores poblacionales (véase Prob. 14.38).

3. **Contraste de hipótesis para predicciones de valores medios.** Sea Y_0 el valor de predicción de Y correspondiente a $X = X_0$ estimado a partir de la ecuación de regresión muestral (o sea, $Y_0 = a_0 + a_1 X_0$). Denotemos por $\bar{Y}_p$ la predicción del *valor medio* de Y correspondiente a $X = X_0$ para la población. Entonces el estadístico

$$t = \frac{Y_0 - \bar{Y}_p}{s_{Y.X}\sqrt{1 + (X_0 - \bar{X})^2/s_X^2}}\sqrt{N - 2} = \frac{Y_0 - \bar{Y}_p}{\hat{s}_{Y.X}\sqrt{1/N + (X_0 - \bar{X})^2/(Ns_X^2)}} \tag{33}$$

tiene distribución de Student con $N - 2$ grados de libertad. De ahí se pueden reducir límites de confianza para las predicciones de los valores medios de la población (véase Prob. 14.39).

PROBLEMAS RESUELTOS

DIAGRAMA DE DISPERSION Y RECTAS DE REGRESION

14.1. La Tabla 14.1 da en pulgadas las respectivas alturas X e Y de una muestra de 12 padres y sus hijos mayores.

(*a*) Construir un diagrama de dispersión.
(*b*) Hallar la recta de regresión de mínimos cuadrados de Y sobre X.
(*c*) Hallar la recta de regresión de mínimos cuadrados de X sobre Y.

Tabla 14.1

Altura X del padre	65	63	67	64	68	62	70	66	68	67	69	71
Altura Y del hijo	68	66	68	65	69	66	68	65	71	67	68	70

Solución

(*a*) El diagrama de dispersión se obtiene marcando los puntos (X, Y) en un sistema rectangular de coordenadas, como ilustra la Figura 14.2.

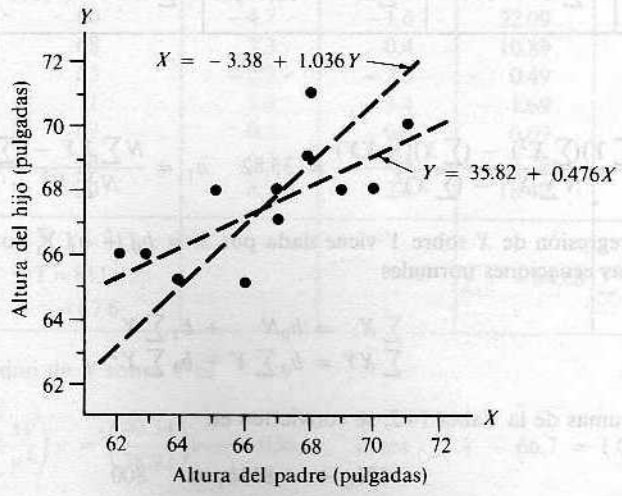


Figura 14.2.

(*b*) La recta de regresión de Y sobre X viene dada por $Y = a_0 + a_1X$, donde a_0 y a_1 se obtienen resolviendo las ecuaciones normales

$$\sum Y = a_0 N + a_1 \sum X$$
$$\sum XY = a_0 \sum X + a_1 \sum X^2$$

Las sumas se indican en la Tabla 14.2, de la que las ecuaciones normales pasan a ser

$$12a_0 + 800a_1 = 811$$
$$800a_0 + 53{,}418a_1 = 54{,}107$$

y de aquí concluimos que $a_0 = 35.82$ y $a_1 = 0.476$, y por tanto $Y = 35.82 + 0.476X$. El gráfico de esta ecuación aparece en la Figura 14.2.

Tabla 14.2

X	Y	X^2	XY	Y^2
65	68	4225	4420	4624
63	66	3969	4158	4356
67	68	4489	4556	4624
64	65	4096	4160	4225
68	69	4624	4692	4761
62	66	3844	4092	4356
70	68	4900	4760	4624
66	65	4356	4290	4225
68	71	4624	4828	5041
67	67	4489	4489	4489
69	68	4761	4692	4624
71	70	5041	4970	4900
$\sum X = 800$	$\sum Y = 811$	$\sum X^2 = 53{,}418$	$\sum XY = 54{,}107$	$\sum Y^2 = 54{,}849$

Otro método

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N\sum X^2 - (\sum X)^2} = 35.82 \qquad a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = 0.476$$

(*c*) La recta de regresión de X sobre Y viene dada por $X = b_0 + b_1 Y$, donde b_0 y b_1 se obtienen resolviendo las ecuaciones normales

$$\sum X = b_0 N + b_1 \sum Y$$
$$\sum XY = b_0 \sum Y + b_1 \sum Y^2$$

Usando las sumas de la Tabla 14.2, se convierten en

$$12b_0 + 811b_1 = 800$$
$$811b_0 + 54{,}849b_1 = 54{,}107$$

y de ahí deducimos $b_0 = -3.38$ y $b_1 = 1.036$, y por tanto, $X = -3.38 + 1.036Y$. El gráfico de estas ecuaciones se ve en la Figura 14.2.

Otro método

$$b_0 = \frac{(\sum X)(\sum Y^2) - (\sum Y)(\sum XY)}{N\sum Y^2 - (\sum Y)^2} = -3.38 \qquad b_1 = \frac{N\sum XY - (\sum Y)(\sum X)}{N\sum Y^2 - (\sum Y)^2} = 1.036$$

14.2. Rehacer los Problemas 14.1(*b*) y 14.1(*c*) usando las rectas de regresión

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \quad \text{y} \quad x = \left(\frac{\sum xy}{\sum y^2}\right)y$$

donde $x = X - \bar{X}$ e $y = Y - \bar{Y}$.

Solución

Primer método

La Tabla 14.3 resume la tarea. La recta de regresión de Y sobre X es

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x = \left(\frac{40.34}{84.68}\right)x = 0.476x \qquad \text{o sea} \qquad Y - 67.6 = 0.476(X - 66.7)$$

Tabla 14.3

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy	y^2
65	68	−1.7	0.4	2.89	−0.68	0.16
63	66	−3.7	−1.6	13.69	5.92	2.56
67	68	0.3	0.4	0.09	0.12	0.16
64	65	−2.7	−2.6	7.29	7.02	6.76
68	69	1.3	1.4	1.69	1.82	1.96
62	66	−4.7	−1.6	22.09	7.52	2.56
70	68	3.3	0.4	10.89	1.32	0.16
66	65	−0.7	−2.6	0.49	1.82	6.76
68	71	1.3	3.4	1.69	4.42	11.56
67	67	0.3	−0.6	0.09	−0.18	0.36
69	68	2.3	0.4	5.29	0.92	0.16
71	70	4.3	2.4	18.49	10.32	5.76
$\sum X = 800$	$\sum Y = 811$					
$\bar{X} = 800/12$	$\bar{Y} = 811/12$			$\sum x^2 = 84.68$	$\sum xy = 40.34$	$\sum y^2 = 38.92$
$= 66.7$	$= 67.6$					

La recta de regresión de X sobre Y es

$$x = \left(\frac{\sum xy}{\sum y^2}\right)y = \left(\frac{40.34}{38.92}\right)y = 1.036y \qquad \text{o sea} \qquad X - 66.7 = 1.036(Y - 67.6)$$

Coinciden con los resultados del Problema 14.1.

Segundo método

Restar una constante adecuada, 60, por ejemplo, de cada valor de X e Y. Los resultados se pueden ordenar como en la Tabla 14.4. Procedamos con el segundo método del Problema 13.17. Así pues.

$$a_1 = \frac{N\sum X'Y' - (\sum X')(\sum Y')}{N\sum X'^2 - (\sum X')^2} = 0.476 \qquad b_1 = \frac{N\sum X'Y' - (\sum Y')(\sum X')}{N\sum Y'^2 - (\sum Y')^2} = 1.036$$

Como $\bar{X} = 60 + 80/12 = 66.7$ e $\bar{Y} = 60 + 91/12 = 67.6$, las requeridas ecuaciones de regresión son las de antes.

Nótese que si a_0 y b_0 se hallasen por este método, *no* obtendríamos los mismos resultados que antes, ya que a_0 y b_0 dependen de la elección del origen. De manera que este método se usa *sólo* para hallar a_1 y b_1, que son independientes de la elección del origen.

Tabla 14.4

X'	Y'	X'^2	$X'Y'$	Y'^2
5	8	25	40	64
3	6	9	18	36
7	8	49	56	64
4	5	16	20	25
8	9	64	72	81
2	6	4	12	36
10	8	100	80	64
6	5	36	30	25
8	11	64	88	121
7	7	49	49	49
9	8	81	72	64
11	10	121	110	100
$\sum X' = 80$	$\sum Y' = 91$	$\sum X'^2 = 618$	$\sum X'Y' = 647$	$\sum Y'^2 = 729$

ERROR TIPICO DE ESTIMACION

14.3. Si la recta de regresión de Y sobre X viene dada por $Y = a_0 + a_1X$, probar que el error típico de estimación $s_{Y.X}$ viene dado por

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0\sum Y - a_1\sum XY}{N}$$

Solución

Los valores de Y estimados por la recta de regresión están dados por $Y_{est} = a_0 + a_1X$. Luego

$$s_{Y.X}^2 = \frac{\sum(Y - Y_{est})^2}{N} = \frac{\sum(Y - a_0 - a_1X)^2}{N}$$
$$= \frac{\sum Y(Y - a_0 - a_1X) - a_0\sum(Y - a_0 - a_1X) - a_1\sum X(Y - a_0 - a_1X)}{N}$$

Ahora bien $$\sum(Y - a_0 - a_1X) = \sum Y - a_0N - a_1\sum X = 0$$

y $$\sum X(Y - a_0 - a_1X) = \sum XY - a_0\sum X - a_1\sum X^2 = 0$$

ya que de las ecuaciones normales

$$\sum Y = a_0N + a_1\sum X$$
$$\sum XY = a_0\sum X + a_1\sum X^2$$

Por tanto $$s_{Y.X}^2 = \frac{\sum Y(Y - a_0 - a_1X)}{N} = \frac{\sum Y^2 - a_0\sum Y - a_1\sum XY}{N}$$

Este resultado puede ser extendido a ecuaciones de regresión no lineales.

14.4. Si $x = X - \bar{X}$ e $y = Y - \bar{Y}$, probar que el resultado del Problema 14.3 puede expresarse

$$s_{Y.X}^2 = \frac{\sum y^2 - a_1\sum xy}{N}$$

Solución

Del Problema 14.3, con $X = x + \bar{X}$ e $Y = y + \bar{Y}$, tenemos

$$\begin{aligned} Ns_{Y.X}^2 &= \sum Y^2 - a_0\sum Y - a_1\sum XY = \sum(y + \bar{Y})^2 - a_0\sum(y + \bar{Y}) - a_1\sum(x + \bar{X})(y + \bar{Y}) \\ &= \sum(y^2 + 2y\bar{Y} + \bar{Y}^2) - a_0(\sum y + N\bar{Y}) - a_1\sum(xy + \bar{X}y + x\bar{Y} + \bar{X}\bar{Y}) \\ &= \sum y^2 + 2\bar{Y}\sum y + N\bar{Y}^2 - a_0N\bar{Y} - a_1\sum xy - a_1\bar{X}\sum y - a_1\bar{Y}\sum x - a_1N\bar{X}\bar{Y} \\ &= \sum y^2 + N\bar{Y}^2 - a_0N\bar{Y} - a_1\sum xy - a_1N\bar{X}\bar{Y} \\ &= \sum y^2 - a_1\sum xy + N\bar{Y}(\bar{Y} - a_0 - a_1\bar{X}) \\ &= \sum y^2 - a_1\sum xy \end{aligned}$$

donde hemos usado los resultados $\sum x = 0$, $\sum y = 0$ e $\bar{Y} = a_0 + a_1\bar{X}$ (que se siguen al dividir ambos lados de la ecuación normal $\sum Y = a_0N + a_1\sum X$ por N).

14.5. Calcular el error típico de estimación, $s_{Y.X}$, para los datos del Problema 14.1, usando (*a*) la definición y (*b*) el resultado del Problema 14.4.

Solución

(*a*) Según el Problema 14.1(*b*) la recta de regresión de Y sobre X es $Y = 35.82 + 0.476X$. La Tabla 14.5 da los valores reales de Y(de la Tabla 14.1) y los valores estimados de Y, denotados por Y_{est}, que se obtienen de la recta de regresión; por ejemplo, correspondiente a $X = 65$ tenemos $Y_{est} = 35.82 + 0.476(65) = 66.76$. También se recogen los valores $Y - Y_{est}$, que se necesitan al calcular $s_{Y.X}$:

$$s_{Y.X}^2 = \frac{\sum(Y - Y_{est})}{N} = \frac{(1.24)^2 + (0.19)^2 + \cdots + (0.38)^2}{12} = 1.642$$

y $s_{Y.X} = \sqrt{1.642} = 1.28$ in.

Tabla 14.5

X	65	63	67	64	68	62	70	66	68	67	69	71
Y	68	66	68	65	69	66	68	65	71	67	68	70
Y_{est}	66.76	65.81	67.71	66.28	68.19	65.33	69.14	67.24	68.19	67.71	68.66	69.62
$Y - Y_{est}$	1.24	0.19	0.29	−1.28	0.81	0.67	−1.14	−2.24	2.81	−0.71	−0.66	0.38

(*b*) De los Problemas 14.1, 14.2 y 14.4

$$s_{Y.X}^2 = \frac{\sum y^2 - a_1 \sum xy}{N} = \frac{38.92 - 0.476(40.34)}{12} = 1.643$$

y $s_{Y.X} = \sqrt{1.643} = 1.28$ in.

14.6. (*a*) Construir dos rectas paralelas a la recta de regresión del Problema 14.1 y que estén a una distancia vertical $s_{Y.X}$ de ella.

(*b*) Determinar el porcentaje de puntos dato que caen entre esas dos rectas.

Solución

(*a*) La recta de regresión $Y = 35.82 + 0.476X$, obtenida en el Problema 14.1, es la de trazo grueso en la Figura 14.3. Las paralelas a distancia vertical $s_{Y.X} = 1.28$ de ella (véase Prob. 14.5), son las de trazo discontinuo en esa figura.

(*b*) De la Figura 14.3 se ve que mientras 7 de los 12 puntos dato caen entre esas rectas, 3 aparecen sobre ellas. Un examen más detallado (usando la fila inferior de la Tabla 14.5, por ejemplo) revela que dos de ellos están entre esas dos rectas. Luego el porcentaje requerido es 9/12 = 75%.

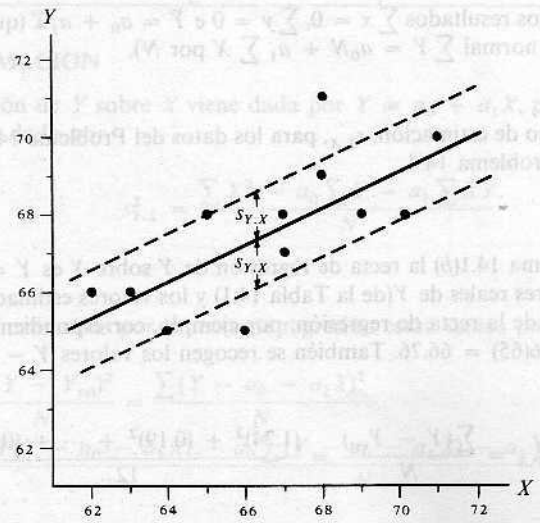


Figura 14.3.

Otro método

Según la fila de abajo en la Tabla 14.5, $Y - Y_{est}$ está entre -1.28 y 1.28 (por ejemplo, $+s_{Y.X}$) para nueve puntos (X, Y). Luego el porcentaje pedido es $9/12 = 75\%$.

Si los puntos están normalmente distribuidos respecto de la recta de regresión, la teoría predice que alrededor del 68% de los puntos están entre las dos rectas. Ello sería más preciso si el tamaño de la muestra fuese grande.

Nota: Una estimación mejor del error típico de estimación de la población de la que procede la muestra viene dada por $\hat{s}_{Y.X} = \sqrt{N/(N-2)}s_{Y.X} = \sqrt{12/10}(1.28) = 1.40$ in.

VARIACION EXPLICADA Y VARIACION INEXPLICADA

14.7. Probar que $\sum(Y - \bar{Y})^2 = \sum(Y - Y_{est})^2 + \sum(Y_{est} - \bar{Y})^2$.

Solución

Elevando al cuadrado ambos mienbros de $Y - \bar{Y} = (Y - Y_{est}) + (Y_{est} - \bar{Y})$ y sumando, tenemos

$$\sum(Y - \bar{Y})^2 = \sum(Y - Y_{est})^2 + \sum(Y_{est} - \bar{Y})^2 + 2\sum(Y - Y_{est})(Y_{est} - \bar{Y})$$

El resultado buscado se sigue inmediatamente si conseguimos ver que la última suma es cero; en el caso de regresión lineal, eso es cierto, porque

$$\begin{aligned}\sum(Y - Y_{est})(Y_{est} - \bar{Y}) &= \sum(Y - a_0 - a_1X)(a_0 + a_1X - \bar{Y}) \\ &= a_0\sum(Y - a_0 - a_1X) + a_1\sum X(Y - a_0 - a_1X) - \bar{Y}\sum(Y - a_0 - a_1X) = 0\end{aligned}$$

a causa de las ecuaciones normales $\sum(Y - a_0 - a_1X) = 0$ y $\sum X(Y - a_0 - a_1X) = 0$.

Análogamente se ve que el resultado es válido para regresión no lineal usando una curva de mínimos cuadrados dada por $Y_{est} = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n$.

14.8. Calcular (*a*) la variación total, (*b*) la variación inexplicada y (*c*) la variación explicada para los datos del Problema 14.1.

Solución

(*a*) La variación total (Prob. 14.2) es $\sum(Y - \bar{Y})^2 = \sum y^2 = 38.92$.

(*b*) La variación inexplicada (Prob. 14.5) es $\sum(Y - Y_{est})^2 = Ns_{Y.X}^2 = 19.70$.

(*c*) La variación explicada (Prob. 14.7) es $\sum(Y_{est} - \bar{Y})^2 = 38.92 - 19.70 = 19.22$.

Otro método

Como $\bar{Y} = 811/12 = 67.58$, podemos construir la Tabla 14.6 usando los valores Y_{est} de la Tabla 14.5; entonces $\sum(Y_{est} - \bar{Y})^2 = (-0.82)^2 + (-1.77)^2 + \cdots + (2.04)^2 = 19.21$. Los resultados de las partes (*a*) y (*b*) se pueden deducir también directamente.

Tabla 14.6

$Y_{est} - \bar{Y} =$ $Y_{est} - 67.58$	-0.82	-1.77	0.13	-1.30	0.61	-2.25	1.56	-0.34	0.61	0.13	1.08	2.04

COEFICIENTE DE CORRELACION

14.9. Hallar (*a*) el coeficiente de determinación y (*b*) el coeficiente de correlación para los datos del Problema 14.1. Usar los resultados del Problema 14.8.

Solución

(*a*) Coeficiente de determinación $= r^2 = \dfrac{\text{variación explicada}}{\text{variación total}} = \dfrac{19.22}{38.92} = 0.4938.$

(*b*) Coeficiente de correlación $= r = \pm\sqrt{0.4938} = \pm 0.7027.$

Como la variable Y crece al crecer X, la correlación es positiva y por tanto escribimos $r = 0.7027$, o sea 0.70 con dos cifras significativas.

14.10. Probar que para regresión lineal el coeficiente de correlación entre las variables X e Y se puede escribir

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

donde $x = X - \bar{X}$ e $y = Y - \bar{Y}$.

Solución

La recta de regresión de mínimos cuadrados de Y sobre X es $Y_{est} = a_0 + a_1 X$ ó $y_{est} = a_1 x$, donde [véase Prob. 13.15(*a*)]

$$a_1 = \frac{\sum xy}{\sum x^2} \quad \text{e} \quad y_{est} = Y_{est} - \bar{Y}$$

Entonces

$$r^2 = \frac{\text{variación explicada}}{\text{variación total}} = \frac{\sum (Y_{est} - \bar{Y})^2}{\sum (Y - \bar{Y})^2} = \frac{\sum y_{est}^2}{\sum y^2}$$

$$= \frac{\sum a_1^2 x^2}{\sum y^2} = \frac{a_1^2 \sum x^2}{\sum y^2} = \left(\frac{\sum xy}{\sum x^2}\right)^2 \frac{\sum x^2}{\sum y^2} = \frac{(\sum xy)^2}{(\sum x^2)(\sum y^2)}$$

y

$$r = \pm \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

Sin embargo, como la cantidad

$$\frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

es positiva cuando y_{est} crece al crecer x (o sea, correlación lineal positiva) y negativa cuando y decrece al crecer x (o sea, correlación lineal negativa), *automáticamente* tiene el signo correcto. Por tanto, definimos el coeficiente de correlación lineal como

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

Esto se suele llamar la *fórmula momento-producto* para el coeficiente de correlación lineal.

FORMULA MOMENTO-PRODUCTO PARA EL COEFICIENTE DE CORRELACION LINEAL

14.11. Hallar el coeficiente de correlación lineal entre las variables X e Y presentadas en la Tabla 14.7.

Tabla 14.7

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

Solución

Los cálculos se resumen en la Tabla 14.8.

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{84}{\sqrt{(132)(56)}} = 0.977$$

De ahí observamos que hay una correlación lineal muy alta entre las variables, como ya se comprobó en los Problemas 13.8 y 13.12.

Tabla 14.8

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy	y^2
1	1	−6	−4	36	24	16
3	2	−4	−3	16	12	9
4	4	−3	−1	9	3	1
6	4	−1	−1	1	1	1
8	5	1	0	1	0	0
9	7	2	2	4	4	4
11	8	4	3	16	12	9
14	9	7	4	49	28	16
$\sum X = 56$	$\sum Y = 40$			$\sum x^2 = 132$	$\sum xy = 84$	$\sum y^2 = 56$
$\bar{X} = 56/8 = 7$	$\bar{Y} = 40/8 = 5$					

14.12. Para los datos del Problema 14.11, hallar (*a*) la desviación típica de X, (*b*) la desviación típica de Y, (*c*) la varianza de X, (*d*) la varianza de Y y (*e*) la covarianza de X e Y.

Solución

(*a*) Desviación típica de $X = xs_X = \sqrt{\dfrac{\sum (X - \bar{X})^2}{N}} = \sqrt{\dfrac{\sum x^2}{N}} = \sqrt{\dfrac{132}{8}} = 4.06$

(*b*) Desviación típica de $Y = s_Y = \sqrt{\dfrac{\sum (Y - \bar{Y})^2}{N}} = \sqrt{\dfrac{\sum y^2}{N}} = \sqrt{\dfrac{56}{8}} = 2.65$

(*c*) Varianza de $X = s_X^2 = 16.50$

(*d*) Varianza de $Y = s_Y^2 = 7.00$

(*e*) Covarianza de X e $Y = s_{XY} = \dfrac{\sum xy}{N} = \dfrac{84}{8} = 10.50.$

14.13. Para los datos del Problema 14.11, verificar la fórmula

$$r = \frac{s_{XY}}{s_X s_Y}$$

Solución

Del Problema 14.12

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{10.50}{(4.06)(2.65)} = 0.976$$

que, salvo por errores de redondeo, coincide con el resultado del Problema 14.11.

14.14. Obtener, mediante la fórmula momento-producto, el coeficiente de correlación lineal para los datos del Problema 14.1.

Solución

Se puede organizar el trabajo como en la Tabla 14.3 del Problema 14.2. Entonces

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{40.34}{\sqrt{(84.68)(38.92)}} = 0.7027$$

que está de acuerdo con el método más largo del Problema 14.9.

14.15. Demostrar que el coeficiente de correlación lineal viene dado por

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

Solución

Haciendo $x = X - \bar{X}$ e $y = Y - \bar{Y}$ en el resultado del Problema 14.10, tenemos

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{[\sum (X - \bar{X})^2][\sum (Y - \bar{Y}]}} \tag{34}$$

Pero $\sum (X - \bar{X})(Y - \bar{Y}) = \sum (XY - \bar{X}Y - X\bar{Y} + \bar{X}\bar{Y}) = \sum XY - \bar{X}\sum Y - \bar{Y}\sum X + N\bar{X}\bar{Y}$

$$= \sum XY - N\bar{X}\bar{Y} - N\bar{Y}\bar{X} + N\bar{X}\bar{Y} = \sum XY - N\bar{X}\bar{Y}$$

$$= \sum XY - \frac{(\sum X)(\sum Y)}{N}$$

ya que $\bar{X} = (\sum X)/N$ e $\bar{Y} = (\sum Y)/N$. Análogamente,

$$\sum (X - \bar{X})^2 = \sum (X^2 - 2X\bar{X} + \bar{X}^2) = \sum X^2 - 2\bar{X}\sum X + N\bar{X}^2$$

$$= \sum X^2 - \frac{2(\sum X)^2}{N} + \frac{(\sum X)^2}{N} = \sum X^2 - \frac{(\sum X)^2}{N}$$

y

$$\sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{N}$$

Así pues, la ecuación (34) se convierte en

$$r = \frac{\sum XY - (\sum X)(\sum Y)/N}{\sqrt{[\sum X^2 - (\sum X)^2/N][\sum Y^2 - (\sum Y)^2/N]}} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

14.16. Mediante la fórmula del Problema 14.15, hallar el coeficiente de correlación lineal para los datos del Problema 14.1.

Solución

Según la Tabla 14.2 del Problema 14.1 se tiene

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

$$= \frac{(12)(54{,}107) - (800)(811)}{\sqrt{[(12)(53{,}418) - (800)^2][(12)(54{,}849) - (811)^2]}} = 0.7027$$

como en los Problemas 14.9 y 14.14.

Otro método

El valor de r es independiente de la elección del origen de X e Y. Así pues, podemos usar los resultados del segundo método del Problema 14.2, con lo que se obtiene

$$r = \frac{N\sum X'Y' - (\sum X')(\sum Y')}{\sqrt{[N\sum X'^2 - (\sum X')^2][N\sum Y'^2 - (\sum Y')^2]}} = \frac{(12)(647) - (80)(91)}{\sqrt{[(12)(618) - (80)^2][(12)(729) - (91)^2]}} = 0.7027$$

COEFICIENTE DE CORRELACION PARA DATOS AGRUPADOS

14.17. La Tabla 14.9 da las distribuciones de frecuencias de las notas finales de 100 estudiantes en Matemáticas y Física. Con referencia a esa tabla, determinar:

(*a*) El número de estudiantes que sacó notas entre 70-79 en Matemáticas y entre 80-89 en Física.
(*b*) El porcentaje de estudiantes con nota de Matemáticas menor que 70.
(*c*) El número de estudiantes que obtuvo 70 o más en Física y menos de 80 en Matemáticas.
(*d*) El porcentaje de estudiantes que aprobó al menos una de las dos materias, si se exigían 60 puntos para aprobar.

Tabla 14.9

		Calificación en Matemáticas						
		40-49	50-59	60-69	70-79	80-89	90-99	Total
Calificación en Física	90-99				2	4	4	10
	80-89			1	4	6	5	16
	70-79			5	10	8	1	24
	60-69	1	4	9	5	2		21
	50-59	3	6	6	2			17
	40-49	3	5	4				12
	Total	7	15	25	23	20	10	100

Solución

(*a*) En la Tabla 14.9, miramos hacia abajo en la columna encabezada con 70-79 (nota de Matemáticas) a la fila con rótulo 80-89 (nota de Física), donde la entrada es 4, que es el número de estudiantes pedido.

(*b*) El número total de estudiantes con nota de Matemáticas inferior a 70 es la suma de los que tienen 40-49, 50-59 y 60-69 = 7 + 15 + 25 = 47. Luego el porcentaje pedido es 47/100 = 47%.

(*c*) El número pedido es el total de las entradas de la Tabla 14.10 (que representa parte de la Tabla 14.9). Por tanto, el número de estudiantes requerido es 1 + 5 + 2 + 4 + 10 = 22.

(*d*) La Tabla 14.11 (sacada de la Tabla 14.9), dice que el número de estudiantes con notas menores que 60 en ambas asignaturas es 3 + 3 + 6 + 5 = 17. Luego el número de los que tienen al menos una nota de 60 o más es 100 − 17 = 83, y el porcentaje requerido es 83/100 = 83%.

Tabla 14.10

		Calificación en Matemáticas	
		60-69	70-79
Calificación en Física	90-99		2
	80-89	1	4
	70-79	5	10

Tabla 14.11

		Calificación en Matemáticas	
		40-49	50-59
Calificación en Física	50-59	3	6
	40-49	3	5

La Tabla 14.9 se llama a veces una *tabla de frecuencias de dos variables*. Cada cuadrado de ese tabla se llama una celda y corresponde a un par de clases o intervalos de confianza. El número indicado en la celda se llama *frecuencia de celda*. Así, en la parte (*a*) el número 4 es la frecuencia de la celda correspondiente al par de intervalos de confianza 70-79 en Matemáticas y 80-89 en Física.

Los totales indicados en la última fila y en la última columna se llaman *totales marginales* o *frecuencias marginales*. Corresponden, respectivamente, a las frecuencias de clase de las distribuciones de frecuencias separadas de las notas de Matemáticas y Física.

14.18. Mostrar cómo modificar la fórmula del Problema 14.15 para el caso de datos agrupados como en la tabla de frecuencias de dos variables (Tabla 14.9).

Solución

Para datos agrupados, podemos considerar los diversos valores de las variables X e Y como coincidentes con las marcas de clase, mientras f_X y f_Y son las correspondientes frecuencias de clase, o frecuencias marginales, que se recogen en la última fila y columna de la tabla de frecuencias de dos variables. Si denotamos por f las diversas frecuencias de celda asociadas a los pares de marcas de clase (X, Y), podemos sustituir la fórmula del Problema 14.15 por

$$r = \frac{N\sum fXY - (\sum f_X X)(\sum f_Y Y)}{\sqrt{[N\sum f_X X^2 - (\sum f_X X)^2][N\sum f_Y Y^2 - (\sum f_Y Y)^2]}} \tag{35}$$

Si hacemos $X = A + c_X u_X$ e $Y = B + c_Y u_Y$, donde c_X y c_Y son las anchuras de intervalos de clase

(supuestas constantes) y A y B son marcas de clase arbitrarias correspondientes a las variables, la fórmula (35) se convierte en la (21):

$$r = \frac{N\sum f u_X u_Y - (\sum f_X u_X)(\sum f_Y u_Y)}{\sqrt{[N\sum f_X u_X^2 - (\sum f_X u_X)^2][N\sum f_Y u_Y^2 - (\sum f_Y u_Y)^2]}} \tag{21}$$

Este es el *método de compilación* empleado en capítulos precedentes como método abreviado para calcular medias, desviaciones típicas y momentos superiores.

Tabla 14.12

Calificación en Física		Calificación en Matemáticas X									
	X	44.5	54.5	64.5	74.5	84.5	94.5				
	Y \ u_X / u_Y	−2	−1	0	1	2	3	f_Y	$f_Y u_Y$	$f_Y u_Y^2$	Suma de los números de las esquinas en cada fila
	94.5 2				2 [4]	4 [16]	4 [24]	10	20	40	44
	84.5 1			1 [0]	4 [4]	6 [12]	5 [15]	16	16	16	31
	74.5 0			5 [0]	10 [0]	8 [0]	1 [0]	24	0	0	0
	64.5 −1	1 [2]	4 [4]	9 [0]	5 [−5]	2 [−4]		21	−21	21	−3
	54.5 −2	3 [12]	6 [12]	6 [0]	2 [−4]			17	−34	68	20
	44.5 −3	3 [18]	5 [15]	4 [0]				12	−36	108	33
	f_X	7	15	25	23	20	10	$\sum f_X = \sum f_Y = N = 100$	$\sum f_Y u_Y = -55$	$\sum f_Y u_Y^2 = 253$	$\sum f u_X u_Y = 125$
	$f_X u_X$	−14	−15	0	23	40	30	$\sum f_X u_X = 64$			
	$f_X u_X^2$	28	15	0	23	80	90	$\sum f_X u_X^2 = 236$			
	Suma de los números de las esquinas en cada columna	32	31	0	−1	24	39	$\sum f u_X u_Y = 125$			

Comprobación

14.19. Hallar el coeficiente de correlación lineal de las notas del Problema 14.17.

Solución

Usamos la fórmula (21). El proceso se resume en la Tabla 14.12, que se llama una tabla de correlación. Las sumas $\sum f_X$, $\sum f_X u_X$, $\sum f_X u_X^2$, $\sum f_Y$, $\sum f_Y u_Y$ y $\sum f_Y u_Y^2$ se obtienen mediante el método de compilación, como en capítulos anteriores.

El número en la esquina de cada celda en la Tabla 14.12 representa el producto $fu_X u_Y$, donde f es la frecuencia de celda. Su suma en cada fila se indica en la fila correspondiente de la última columna. Y su suma en cada columna se indica en la correspondiente columna de la última fila. Los totales finales de la última fila y columna son iguales y representan

$$r = \frac{N\sum fu_X u_Y - (\sum f_X u_X)(\sum f_Y u_Y)}{\sqrt{[N\sum f_X u_X^2 - (\sum f_X u_X)^2][N\sum f_Y u_Y^2 - (\sum f_Y u_Y)^2]}}$$

$$= \frac{(100)(125) - (64)(-55)}{\sqrt{[(100)(236) - (64)^2][(100)(253) - (-55)^2]}} = \frac{16{,}020}{\sqrt{(19{,}504)(22{,}275)}} = 0.7686$$

14.20. Usar la Tabla 14.12 para calcular (*a*) s_X, (*b*) s_Y y (*c*) s_{XY} y así verificar la fórmula $r = s_{XY}/(s_X s_Y)$.

Solución

(*a*) $$s_X = c_X\sqrt{\frac{\sum f_X u_X^2}{N} - \left(\frac{\sum f_X u_X}{N}\right)^2} = 10\sqrt{\frac{236}{100} - \left(\frac{64}{100}\right)^2} = 13.966$$

(*b*) $$s_Y = c_Y\sqrt{\frac{\sum f_Y u_Y^2}{N} - \left(\frac{\sum f_Y u_Y}{N}\right)^2} = 10\sqrt{\frac{253}{100} - \left(\frac{-55}{100}\right)^2} = 14.925$$

(*c*) $$s_{XY} = c_X c_Y\left[\frac{\sum fu_X u_Y}{N} - \left(\frac{\sum f_X u_X}{N}\right)\left(\frac{\sum f_Y u_Y}{N}\right)\right] = (10)(10)\left[\frac{125}{100} - \left(\frac{64}{100}\right)\left(\frac{-55}{100}\right)\right] = 160.20$$

Luego las desviaciones típicas de las notas de Matemáticas y Física son 14.0 y 14.9 respectivamente, mientras que su covarianza es 160.2. El coeficiente de correlación r es, por tanto,

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{160.20}{(13.966)(14.925)} = 0.7686$$

en coincidencia con el Problema 14.19.

RECTAS DE REGRESION Y EL COEFICIENTE DE CORRELACION

14.21. Probar que las rectas de regresión de Y sobre X y de X sobre Y tienen ecuaciones respectivas (*a*) $Y - \bar{Y} = (rs_Y/s_X)(X - \bar{X})$ y (*b*) $X - \bar{X} = (rs_X/s_Y)(Y - \bar{Y})$.

Solución

(*a*) Del Problema 13.15(*a*) sabemos que la recta de regresión de Y sobre X es

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \qquad \text{o sea} \qquad Y - \bar{Y} = \left(\frac{\sum xy}{\sum x^2}\right)(X - \bar{X})$$

Entonces, como $r = \dfrac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$ (véase Prob. 14.10)

tenemos $$\frac{\sum xy}{\sum x^2} = \frac{r\sqrt{(\sum x^2)(\sum y^2)}}{\sum x^2} = \frac{r\sqrt{\sum y^2}}{\sqrt{\sum x^2}} = \frac{rs_Y}{s_X}$$

y el resultado es el deseado.

(*b*) Esto se deduce intercambiando X e Y en la parte (*a*).

14.22. Si las rectas de regresión de Y sobre X y de X sobre Y son, respectivamente, $Y = a_0 + a_1 X$ y $X = b_0 + b_1 Y$, probar que $a_1 b_1 = r^2$.

Solución

Del Problema 14.21, partes (*a*) y (*b*),

$$a_1 = \frac{rs_Y}{s_X} \quad \text{y} \quad b_1 = \frac{rs_X}{s_Y}$$

Luego $$a_1 b_1 = \left(\frac{rs_Y}{s_X}\right)\left(\frac{rs_X}{s_Y}\right) = r^2$$

Cabe tomar este resultando como punto de partida para la definición del coeficiente de correlación lineal.

14.23. Usar el resultado del Problema 14.22 para hallar el coeficiente de correlación lineal para los datos del Problema 14.1.

Solución

Del Problema 14.1 [partes (*b*) y (*c*), respectivamente] $a_1 = 484/1016 = 0.476$ y $b_1 = 484/467 = 1.036$. Así que $r^2 = a_1 b_1 = (484/1016)(484/467)$ y $r = 0.7027$, de acuerdo con los Problemas 14.9, 14.14 y 14.16.

14.24. Para los datos del Problema 14.19, escribir las ecuaciones de las rectas de regresión de (*a*) Y sobre X y (*b*) X sobre Y.

Solución

De la tabla de correlación (Tabla 14.12) del Problema 14.19, tenemos

$$\bar{X} = A + c_X \frac{\sum f_X u_X}{N} = 64.5 + \frac{(10)(64)}{100} = 70.9$$

$$\bar{Y} = B + c_Y \frac{\sum f_Y u_Y}{N} = 74.5 + \frac{(10)(-55)}{100} = 69.0$$

Por el Problema 14.20, $s_X = 13.966$, $s_Y = 14.925$ y $r = 0.7686$. Ahora usamos el Problema 14.21, partes (*a*) y (*b*), para obtener las ecuaciones de las rectas de regresión.

(*a*) $Y - \bar{Y} = \dfrac{rs_Y}{s_X}(X - \bar{X})$ $\quad Y - 69.0 = \dfrac{(0.7686)(14.925)}{13.966}(X - 70.9) = 0.821(X - 70.9)$

(*b*) $X - \bar{X} = \dfrac{rs_X}{s_Y}(Y - \bar{Y})$ $\quad X - 70.9 = \dfrac{(0.7686)(13.966)}{14.925}(Y - 69.0) = 0.719(Y - 69.0)$

14.25. Calcular, para los datos del Problema 14.19, los errores típicos de estimación (*a*) $s_{Y.X}$ y (*b*) $s_{X.Y}$. Usar los resultados del Problema 14.20.

Solución

(*a*) $s_{Y.X} = s_Y\sqrt{1 - r^2} = 14.925\sqrt{1 - (0.7686)^2} = 9.548$

(*b*) $s_{X.Y} = s_X\sqrt{1 - r^2} = 13.966\sqrt{1 - (0.7686)^2} = 8.934$

14.26. La Tabla 14.13 muestra los índices de precios al consumo de alimentación y de asistencia sanitaria durante los años 1975-1983 comparados con los precios en un año base, 1967 (tomados como 100). Calcular el coeficiente de correlación entre esos dos índices.

Tabla 14.13

Año	1975	1976	1977	1978	1979	1980	1981	1982	1983
Alimentación	175	181	192	211	235	255	275	286	292
Asistencia sanitaria	169	185	202	219	240	266	295	329	357

Fuente: Survey of Current Business.

Solución

(*a*) Denotando por X e Y los índices de alimentación y de asistencia sanitaria, respectivamente, el cálculo del coeficiente de correlación procede como sugiere la Tabla 14.14. (Nótese que el año se emplea sólo para especificar los valores correspondientes de X e Y). Entonces, por la fórmula momento-producto,

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{23.442}{\sqrt{(16,774)(34,107)}} = 0.98$$

Luego existe una correlación lineal muy buena entre ambos índices de costo. Hay que hacer constar, no obstante, que eso no quiere decir que los costes hayan aumentado *lo mismo* a lo largo de los años: así, por ejemplo, de 1975 a 1983 los alimentos han subido un 67% mientras que la asistencia sanitaria lo ha hecho en un 111%.

Tabla 14.14

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy	y^2
175	169	−59	−82	3,481	4,838	6,724
181	185	−53	−66	2,809	3,498	4,356
192	202	−42	−49	1,764	2,058	2,401
211	219	−23	−32	529	736	1,024
235	240	1	−11	1	−11	121
255	266	21	15	441	315	225
275	295	41	44	1,681	1,804	1,936
286	329	52	78	2,704	4,056	6,084
292	357	58	106	3,364	6,148	11,236
$\sum X = 2,102$ $\bar{X} = 234$	$\sum Y = 2,262$ $\bar{Y} = 251$			$\sum x^2 = 16,774$	$\sum xy = 23,442$	$\sum y^2 = 34,107$

CORRELACION NO LINEAL

14.27. Ajustar una parábola de mínimos cuadrados de la forma $Y = a_0 + a_1X + a_2X^2$ al conjunto de datos de la Tabla 14.15.

Tabla 14.15

X	1.2	1.8	3.1	4.9	5.7	7.1	8.6	9.8
Y	4.5	5.9	7.0	7.8	7.2	6.8	4.5	2.7

Solución

Las ecuaciones normales (23) del Capítulo 13 son

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X + a_2 \sum X^2 \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum X^3 \\ \sum X^2 Y &= a_0 \sum X^2 + a_1 \sum X^3 + a_2 \sum X^4 \end{aligned} \tag{36}$$

El proceso de cálculo de las sumas se presenta en la Tabla 14.16. Como $N = 8$, las ecuaciones normales (36) pasan a ser

$$\begin{aligned} 8a_0 + 42.2a_1 + 291.20a_2 &= 46.4 \\ 42.2a_0 + 291.20a_1 + 2275.35a_2 &= 230.42 \\ 291.20a_0 + 2275.35a_1 + 18971.92a_2 &= 1449.00 \end{aligned} \tag{37}$$

Resolviendo, $a_0 = 2.588$, $a_1 = 2.065$, y $a_2 = -0.2110$; por tanto, la parábola de mínimos cuadrados buscada es

$$Y = 2.588 + 2.065X - 0.2110X^2$$

14.28. Estimar, mediante la parábola de mínimos cuadrados del Problema 14.27, los valores de Y a partir de los valores de X dados.

Solución

Para $X = 1.2$, $Y_{est} = 2.588 + 2.065(1.2) - 0.2110(1.2)^2 = 4.762$. Otros valores estimados se obtienen análogamente. Los resultados, junto con los valores reales de Y, se muestran en la Tabla 14.17.

Tabla 14.16

X	Y	X^2	X^3	X^4	XY	X^2Y
1.2	4.5	1.44	1.73	2.08	5.40	6.48
1.8	5.9	3.24	5.83	10.49	10.62	19.12
3.1	7.0	9.61	29.79	92.35	21.70	67.27
4.9	7.8	24.01	117.65	576.48	38.22	187.28
5.7	7.2	32.49	185.19	1055.58	41.04	233.93
7.1	6.8	50.41	357.91	2541.16	48.28	342.79
8.6	4.5	73.96	636.06	5470.12	38.70	332.82
9.8	2.7	96.04	941.19	9223.66	26.46	259.31
$\sum X$ = 42.2	$\sum Y$ = 46.4	$\sum X^2$ = 291.20	$\sum X^3$ = 2275.35	$\sum X^4$ = 18,971.92	$\sum XY$ = 230.42	$\sum X^2Y$ = 1449.00

Tabla 14.17

Y_{est}	4.762	5.621	6.962	7.640	7.503	6.613	4.741	2.561
Y	4.5	5.9	7.0	7.8	7.2	6.8	4.5	2.7

14.29. (*a*) Hallar el coeficiente de correlación lineal entre las variables X e Y del Problema 14.27.
(*b*) Hallar el coeficiente de correlación no lineal entre estas variables, suponiendo la relación parabólica obtenida en el Problema 14.27.
(*c*) Explicar la diferencia entre los coeficientes de correlación obtenidos en las partes (*a*) y (*b*).
(*d*) ¿Qué porcentaje de la variación total queda inexplicada al suponer una relación parabólica entre X e Y?

Solución

(*a*) Haciendo uso de los cálculos ya realizados en la Tabla 14.16 y el hecho añadido de que $\sum Y^2 = 290.52$, vemos que

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}} = \frac{(8)(230.42) - (42.2)(46.4)}{\sqrt{[(8)(291.20) - (42.2)^2][(8)(290.52) - (46.4)^2]}} = -0.3743$$

(*b*) De la Tabla 14.16, $\bar{Y} = (\sum Y)/N = 46.4/8 = 5.80$; luego la variación total es $\sum (Y - \bar{Y})^2 = 21.40$. De la Tabla 14.17 vemos que la variación explicada es $\sum (Y_{est} - \bar{Y})^2 = 21.02$. Luego

$$r^2 = \frac{\text{variación explicada}}{\text{variación total}} = \frac{21.02}{21.40} = 0.9822 \quad \text{y} \quad r = 0.9911 \quad \text{o sea} \quad 0.99$$

(*c*) El que (*a*) haya dado un coeficiente de correlación lineal de sólo -0.3743 indica que *no hay* prácticamente *relación lineal* entre X e Y. Sin embargo, hay una *relación no lineal* muy fuerte dada por la parábola del Problema 14.27, como ratifica el hecho de que el coeficiente de correlación en (*b*) es 0.9.

(*d*) $$\frac{\text{Variación inexplicada}}{\text{Variación total}} = 1 - r^2 = 1 - 0.9822 = 0.0178$$

Luego el 1.78% de la variación total queda inexplicada. Ello podría ser debido a fluctuaciones aleatorias o a una variable adicional que no se ha tenido en cuenta.

14.30. Hallar (*a*) s_Y y (*b*) $s_{Y.X}$ para los datos del Problema 14.27.

Solución

(*a*) Del Problema 14.29(*a*), $\sum (Y - \bar{Y})^2 = 21.40$. Así pues, la desviación típica de Y es

$$s_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}} = \sqrt{\frac{21.40}{8}} = 1.636 \quad \text{o sea} \quad 1.64$$

(*b*) *Primer método*

Usando la parte (*a*) y el Problema 14.29(*b*), el error típico de estimación de Y sobre X es

$$s_{Y.X} = s_Y\sqrt{1 - r^2} = 1.636\sqrt{1 - (0.9911)^2} = 0.218 \quad \text{o sea} \quad 0.22$$

Segundo método

Usando el Problema 14.29,

$$s_{Y.X} = \sqrt{\frac{\sum(Y - Y_{est})^2}{N}} = \sqrt{\frac{\text{variación inexplicada}}{N}} = \sqrt{\frac{21.40 - 21.02}{8}} = 0.218 \quad \text{o sea} \quad 0.22$$

Tercer método

Por el Problema 14.27 y el cálculo adicional $\sum Y^2 = 290.52$, tenemos

$$s_{Y.X} = \sqrt{\frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY - a_2 \sum X^2 Y}{N}} = 0.218 \quad \text{o sea} \quad 0.22$$

TEORIA MUESTRAL DE LA CORRELACION

14.31. Al calcular el coeficiente de correlación de una muestra de tamaño 18, ha dado el valor 0.32. ¿Podemos concluir al nivel de significación (*a*) 0.05 y (*b*) 0.01 que el coeficiente de correlación de la población correspondiente difiere de cero?

Solución

Queremos decidir entre las hipótesis H_0: $\rho = 0$ y H_1: $\rho > 0$.

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{0.32\sqrt{18-2}}{\sqrt{1-(0.32)^2}} = 1.35$$

(*a*) Usando un contraste unilateral con la distribución de Student en el nivel 0.05, rechazaríamos la hipótesis H_0 si $t > t_{.95} = 1.75$ para $(18 - 2) = 16$ grados de libertad. Luego no podemos rechazar H al nivel 0.05.

(*b*) Puesto que no podemos rechazar H al nivel 0.05, ciertamente, tampoco al 0.01.

14.32. ¿Cuál es el mínimo tamaño de muestra necesario para poder concluir que un coeficiente de correlación de 0.32 difiere significativamente de cero al nivel 0.05?

Solución

Con un contraste de una cola de la distribución de Student en el nivel 0.05, el mínimo valor de N debe ser tal que

$$\frac{0.32\sqrt{N-2}}{\sqrt{1-(0.32)^2}} = t_{.95}$$

para $N - 2$ grados de libertad. Para un número infinito de grados de libertad, $t_{95} = 1.64$ y por tanto, $N = 25.6$.

Para $N = 26$: $\nu = 24$ $t_{.95} = 1.71$ $t = 0.32\sqrt{24}/\sqrt{1-(0.32)^2} = 1.65$

Para $N = 27$: $\nu = 25$ $t_{.95} = 1.71$ $t = 0.32\sqrt{25}/\sqrt{1-(0.32)^2} = 1.69$

Para $N = 28$: $\nu = 26$ $t_{.96} = 1.71$ $t = 0.32\sqrt{26}/\sqrt{1-(0.32)^2} = 1.72$

Así que el tamaño mínimo de la muestra es $N = 28$.

14.33. Un coeficiente de correlación de una muestra de tamaño 24 resulta ser $r = 0.75$. Al nivel de significación 0.05, ¿podemos rechazar la hipótesis de que el coeficiente de correlación de la población es tan pequeño como (*a*) $\rho = 0.60$ y (*b*) $\rho = 0.50$?

Solución

(*a*) $$Z = 1.1513 \log\left(\frac{1 + 0.75}{1 - 0.75}\right) = 0.9730 \qquad \mu_Z = 1.1513 \log\left(\frac{1 + 0.60}{1 - 0.60}\right) = 0.6932$$

y $$\sigma_Z = \frac{1}{\sqrt{N - 3}} = \frac{1}{\sqrt{21}} = 0.2182$$

Por tanto $$z = \frac{Z - \mu_Z}{\sigma_Z} = \frac{0.9730 - 0.6932}{0.2182} = 1.28$$

Usando un contraste de una cola con la distribución normal al nivel 0.05, rechazaríamos la hipótesis sólo si z fuera mayor que 1.64. Luego no podemos rechazar la hipótesis de que el coeficiente de correlación de la población es tan pequeño como 0.60.

(*b*) Si $\rho = 0.50$, entonces $\mu_Z = 1.1513 \log 3 = 0.5493$ y $z = (0.9730 - 0.5493)/0.2182 = 1.94$. Luego podemos rechazar la hipótesis de que el coeficiente de correlación de la población sea tan pequeño como $\rho = 0.50$, al nivel 0.05.

14.34. El coeficiente de correlación entre las notas en Física y Matemáticas para un grupo de 21 estudiantes resulta ser 0.80. Hallar los límites de confianza 95% para este coeficiente.

Solución

Como $r = 0.80$ y $N = 21$, los límites de confianza 95% para μ_Z vienen dados por

$$Z \pm 1.96\sigma_Z = 1.1513 \log\left(\frac{1 + r}{1 - r}\right) \pm 1.96\left(\frac{1}{\sqrt{N - 3}}\right) = 1.0986 \pm 0.4620$$

Así pues, μ_Z tiene el intervalo de confianza 95% desde 0.5366 a 1.5606. Ahora bien, si

$$\mu_Z = 1.1513 \log\left(\frac{1 + \rho}{1 - \rho}\right) = 0.5366 \qquad \text{entonces} \qquad \rho = 0.4904$$

y si $$\mu_Z = 1.1513 \log\left(\frac{1 + \rho}{1 - \rho}\right) = 1.5606 \qquad \text{entonces} \qquad \rho = 0.9155$$

Luego los límites de confianza 95% para ρ son 0.49 y 0.92.

14.35. Dos coeficientes de correlación obtenidos de muestras de tamaños $N_1 = 28$ y $N_3 = 35$ han resultado ser $r_1 = 0.50$ y $r_2 = 0.30$, respectivamente. ¿Hay diferencia significativa entre los dos coeficientes al nivel 0.05?

Solución

$$Z_1 = 1.1513 \log\left(\frac{1 + r_1}{1 - r_1}\right) = 0.5493 \qquad Z_2 = 1.1513 \log\left(\frac{1 + r_2}{1 - r_2}\right) = 0.3095$$

y $$\sigma_{Z1\ Z2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} = 0.2669$$

Queremos decidir entre dos hipótesis H_0: $\mu_{Z1} = \mu_{Z2}$ y H_1: $\mu_{Z1} \neq \mu_{Z2}$. Bajo la hipótesis H_0,

$$z = \frac{Z_1 - Z_2 - (\mu_{Z1} - \mu_{Z2})}{\sigma_{Z1-Z2}} = \frac{0.5493 - 0.3095 - 0}{0.2669} = 0.8985$$

Con un contraste bilateral mediante la distribución normal, rechazaríamos H sólo si $z > 1.96$ o si $z < -1.96$. Por tanto, no podemos rechazar H, y concluimos que los resultados no son significativamente diferentes al nivel 0.05.

TEORIA MUESTRAL DE LA REGRESION

14.36. En el Problema 14.1 hallamos como ecuación de regresión de Y sobre X la que sigue: $Y = 35.82 + 0.476X$. Contrastar la hipótesis, al nivel de significación 0.05, de que el coeficiente de correlación de la ecuación de regresión de la población es 0.180.

Solución

$$t = \frac{a_1 - A_1}{s_{Y.X}/s_X}\sqrt{N-2} = \frac{0.476 - 0.180}{1.28/2.66}\sqrt{12-2} = 1.95$$

como $s_{Y.X} = 1.28$ (calculado en el Problema 14.5) y $s_X = \sqrt{(\sum x^2)/N} = \sqrt{84.68/12} = 2.66$ (del Problema 14.2). Usando un contraste de una cola con la distribución de Student al nivel 0.05, rechazaríamos la hipótesis de que el coeficiente de regresión es tan bajo como 0.180 si $t > t_{.95} = 1.81$ para $(12 - 2) = 10$ grados de libertad. Luego no podemos rechazar la hipótesis.

14.37. Hallar los límites de confianza 95% para el coeficiente de regresión del Problema 14.36.

Solución

$$A_1 = a_1 - \frac{t}{\sqrt{N-2}}\left(\frac{s_{Y.X}}{s_X}\right)$$

Luego los límites de confianza para A (obtenidos haciendo $t = \pm t_{.975} = \pm 2.23$ para $12 - 2 = 10$ grados de libertad) vienen dados por

$$a_1 \pm \frac{2.23}{\sqrt{12-2}}\left(\frac{s_{Y.X}}{s_X}\right) = 0.476 \pm \frac{2.23}{\sqrt{10}}\left(\frac{1.28}{2.66}\right) = 0.476 \pm 0.340$$

Es decir, tenemos 95% de confianza de que A está entre 0.136 y 0.816.

14.38. En el Problema 14.1, hallar los límites de confianza 9% para las alturas de los hijos cuyos padres miden (*a*) 65.0 y (*b*) 70.0 in.

Solución

Como $t_{.975} = 2.23$ para $(12 - 2) = 10$ grados de libertad, los límites de confianza 95% para Y_p (véase pág. 330) vienen dados por

$$Y_0 \pm \frac{2.23}{\sqrt{N-2}} s_{Y.X}\sqrt{N + 1 + \frac{(X_0 - \bar{X})^2}{s_X^2}}$$

donde $Y_0 = 35.82 + 0.476X_0$ (Problema 14.1), $s_{Y.X} = 1.28$, $s_X = 2.66$ (Problema 14.36) y $N = 12$.

(*a*) Si $X_0 = 65.0$, entonces $Y_0 = 66.76$ in. Además $(X_0 - \bar{X})^2 = (65.0 - 800/12)^2 = 2.78$. Así pues los límites de confianza al 95% son

$$66.76 \pm \frac{2.23}{\sqrt{10}}(1.28)\sqrt{12 + 1 + \frac{2.78}{(2.66)^2}} = 66.76 \pm 3.31 \text{ in}$$

Esto es, podemos tener un 95% de confianza de que las alturas de los hijos están entre 63.4 y 70.1.

(*b*) Si $X_0 = 70.0$, entonces $Y_0 = 69.14$ in. Además, $(X_0 - \bar{X})^2 = (70.0 - 800/12)^2 = 11.11$. Luego los límites de confianza 95% resultan ser 69.14 ± 3.45 in; es decir, con un 95% de confianza las alturas de los hijos están entre 65.7 y 72.6 in.

Nótese que para los valores grandes de N, los límites de confianza 95% vienen dados aproximadamente por $Y_0 \pm 1.96s_{Y.X}$ o sea $Y_0 \pm 2s_{Y.X}$, supuesto que $(X_0 - \bar{X})$ no sea demasiado grande. Eso coincide con los resultados aproximados mencionados en la página 210. Los métodos de este problema son válidos con independencia del valor de N o de $(X_0 - \bar{X})$; esto es, los métodos de muestreo son exactos.

14.39. En el Problema 14.1 hallar los límites de confianza 95% para las alturas medias de los hijos cuyos padres miden (*a*) 65.0 in y (*b*) 70.0 in.

Solución

Ya que $t_{.975} = 2.23$ para 10 grados de libertad, los límites de confianza 95% para $\bar{Y}_p$ (véase página 330) vienen dados por

$$Y_0 \pm \frac{2.23}{\sqrt{10}} s_{Y.X}\sqrt{1 + \frac{(X_0 - \bar{X})^2}{s_X^2}}$$

donde $Y_0 = 35.82 + 0.476X_0$ (Problema 14.1), $s_{Y.X} = 1.28$ y $s_X = 2.66$ (Problema 14.36).

(*a*) Si $X_0 = 65.0$, vemos que los límites de confianza 95% son 66.76 ± 1.07 in [comparar con el Problema 14.38(*a*)]. Es decir, podemos tener 95% de confianza de que la altura media de todos los hijos cuyos padres miden 65.0 in está entre 65.7 y 67.8 in.

(*b*) Si $X_0 = 70.0$, vemos que los límites de confianza 95% son 69.14 ± 1.45 in [comparar con el Problema 14.38(*b*)]. Es decir, podemos tener 95% de confianza de que la *altura media* de todos los hijos cuyos padres miden 70.0 in estará entre 67.7 y 70.6 in.

PROBLEMAS SUPLEMENTARIOS

REGRESION LINEAL Y CORRELACION LINEAL

14.40. La Tabla 14.18 presenta las notas de dos exámenes de Biología, X e Y, de 10 estudiantes.

(*a*) Construir un diagrama de dispersión.

(*b*) Hallar la recta de regresión de mínimos cuadrados de Y sobre X.

(*c*) Hallar la recta de regresión de mínimos cuadrados de X sobre Y.

(*d*) Representar las dos rectas de las partes (*b*) y (*c*) en el diagrama de dispersión de la parte (*a*).

Tabla 14.18

Calificaciones en el primer examen (X)	Calificaciones en el segundo examen (Y)
6	8
5	7
8	7
8	10
7	5
6	8
10	0
4	6
9	8
7	6

14.41. Hallar (*a*) $s_{Y.X}$ y (*b*) $s_{X.Y}$ para los datos de la Tabla 14.18.

14.42. Calcular (*a*) la variación total en Y, (*b*) la variación inexplicada en Y y (*c*) la variación explicada en Y, para los datos del Problema 14.40.

14.43. Usar los resultados del Problema 14.42 para hallar el coeficiente de correlación entre los dos conjuntos de notas del Problema 14.40.

14.44. (*a*) Hallar el coeficiente de correlación entre los dos conjuntos de notas del Problema 14.40 usando la fórmula momento-producto, y comparar con el resultado del Problema 14.45.

(*b*) Obtener el coeficiente de correlación directamente a partir de las pendientes de las rectas de regresión del Problema 14.42, partes (*b*) y (*c*).

14.45. Hallar la covarianza para los datos del Problema 14.40 (*a*) directamente y (*b*) usando la fórmula $s_{XY} = rs_X s_Y$ y el resultado del Problema 14.43 ó 14.44.

14.46. La Tabla 14.19 da las edades X y las presiones sanguíneas (en sístole) Y de 12 mujeres.

(*a*) Hallar el coeficiente de correlación entre X e Y.

(*b*) Determinar la ecuación de regresión de mínimos cuadrados de Y sobre X

(*c*) Estimar la presión sanguínea de una mujer de 45 años.

Tabla 14.19

Edad (X)	Presión sanguínea
56	147
42	125
72	160
36	118
63	149
47	128
55	150
49	145
38	115
42	140
68	152
60	155

14.47. Hallar el coeficiente de correlación para los datos del (*a*) Problema 13.32 y (*b*) Problema 13.35.

14.48. El coeficiente de correlación entre las variables X e Y es $r = 0.60$. Si $s_X = 1.50$, $s_Y = 2.00$, $\bar{X} = 10$ e $\bar{Y} = 20$, hallar la ecuación de la recta de regresión de (*a*) Y sobre X y (*b*) X sobre Y.

14.49. Calcular (*a*) $s_{Y.X}$ y (*b*) $s_{X.Y}$ para los datos del Problema 14.48.

14.50. Si $s_{Y.X} = 3$ y $s_Y = 5$, calcular r

14.51. Si el coeficiente de correlación entre X e Y es 0.50, ¿qué porcentaje de la variación total queda inexplicado por la ecuación de regresión?

14.52. (*a*) Probar que la ecuación de la recta de regresión de Y sobre X puede escribirse

$$Y - \bar{Y} = \frac{s_{XY}}{s_X^2}(X - \bar{X})$$

(*b*) Escribir una ecuación análoga para la recta de regresión de X sobre Y

14.53. (*a*) Calcular el coeficiente de correlación entre los valores correspondientes de X e Y dados en la Tabla 14.20.

(*b*) Multiplicar cada valor de X en la tabla por 2 y sumar 6. Multiplicar cada valor de Y en la tabla por 3 y restar 15. Hallar el coeficiente de correlación entre los dos nuevos conjuntos de valores, explicando por qué se obtiene o por qué no se obtiene el mismo resultado que en (*a*).

Tabla 14.20

X	Y
2	18
4	12
5	10
6	8
11	5

14.54. (*a*) Hallar las ecuaciones de regresión de Y sobre X para los datos considerados en el Problema 14.53, partes (*a*) y (*b*).

(*b*) Discutir la relación entre estas ecuaciones de regresión.

14.55. (*a*) Probar que el coeficiente de correlación entre X e Y puede expresarse

$$r = \frac{\overline{XY} - \bar{X}\bar{Y}}{\sqrt{[\overline{X^2} - \bar{X}^2][\overline{Y^2} - \bar{Y}^2]}}$$

(*b*) Usando ese método, resolver el Problema 14.1.

14.56. Probar que un coeficiente de correlación es independiente de la elección de origen de las variables o de las unidades en que se expresan. (*Ayuda:* Supóngase que $X' = c_1X + A$ e $Y' = c_2 Y + B$, donde c_1, c_2, A y B son constantes arbitrarias, y pruébese que el coeficiente de correlación entre X' e Y' es el mismo que entre X e Y).

14.57. (*a*) Probar que, para regresión lineal,

$$\frac{s_{Y.X}^2}{s_Y^2} = \frac{s_{X.Y}^2}{s_X^2}$$

(*b*) ¿Es válido el resultado para regresión no lineal?

COEFICIENTE DE CORRELACION PARA DATOS AGRUPADOS

14.58. Hallar el coeficiente de correlación entre las alturas y pesos de los 300 hombres adultos de EE.UU recogidos en la tabla de frecuencias dada en la Tabla 14.21.

Tabla 14.21

Pesos	Alturas X (in)				
Y (lb)	59-62	63-66	67-70	71-74	75-78
90-109	2	1			
110-129	7	8	4	2	
130-149	5	15	22	7	1
150-169	2	12	63	19	5
170-189		7	28	32	12
190-209		2	10	20	7
210-229			1	4	2

14.59. (*a*) Hallar la recta de regresión de mínimos cuadrados de Y sobre X para los datos del Problema 14.58.

(*b*) Estimar los pesos de dos hombres cuyas alturas son 64 y 72 in.

14.60. Hallar (*a*) $s_{Y.X}$ y (*b*) $s_{X.Y}$ para los datos del Problema 14.58.

14.61. Establecer la fórmula (21) de este capítulo para el coeficiente de correlación de datos agrupados.

CORRELACION DE SERIES EN EL TIEMPO

14.62. La Tabla 14.22 muestra los precios al por menor del cinc en EE.UU. y los correspondientes índices de precios al consumo en los

años 1978-1985. Hallar el coeficiente de correlación.

14.63. La Tabla 14.23 da la temperatura media y la precipitación en una ciudad durante el mes de julio de los años 1975-1984. Hallar el coeficiente de correlación.

TEORIA MUESTRAL DE LA CORRELACION

14.64. Un coeficiente de correlación basado en una muestra de tamaño 27 resultó ser 0.40. ¿Se puede concluir que el coeficiente de correlación de la población correspondiente, al nivel de significación (*a*) 0.05 y (*b*) 0.01, difiere de cero?

Tabla 14.22

Año	Precio de cinc (centavos por libra)	Indice de precios al consumo (1967 = 100)
1978	31.0	195.4
1979	37.3	217.4
1980	37.4	246.8
1981	44.6	272.4
1982	38.5	289.1
1983	41.4	298.4
1984	48.6	311.1
1985	40.3	322.2

Fuente: U.S. Bureau of Labor Statistics and Bureau of Mines.

Tabla 14.23

Año	Temperatura (°F)	Precipitación (in)
1975	78.1	6.23
1976	71.8	3.64

Tabla 14.23. (*Continuación*)

Año	Temperatura (°F)	Precipitación (in)
1977	75.6	3.42
1978	72.7	2.84
1979	75.3	1.83
1980	73.6	2.82
1981	75.1	4.04
1982	75.3	2.56
1983	73.8	1.18
1984	70.4	4.19

14.65. Un coeficiente de correlación basado en una muestra de tamaño 35 ha dado 0.50. Al nivel de significación 0.05, ¿podemos rechazar la hipótesis de que el coeficiente de correlación de la población es (*a*) tan pequeño como 0.30 y (*b*) tan grande como 0.70?

14.66. Hallar los límites de confianza (*a*) 95% y (*b*) 99% para un coeficiente de correlación que se ha calculado como 0.60 a partir de una muestra de tamaño 28.

14.67. Resolver el Problema 14.66 con una muestra de tamaño 52.

14.68. Hallar los límites de confianza 95% para el coeficiente de correlación calculado en (*a*) el Problema 14.46 y (*b*) el Problema 14.58.

14.69. Dos coeficientes de correlación obtenidos de muestras de tamaños 23 y 28 resultan ser 0.80 y 0.95 respectivamente. ¿Podemos concluir a nivel de significación (*a*) 0.05 y (*b*) 0.01 que hay una diferencia significativa entre ellos?

TEORIA MUESTRAL DE LA REGRESION

14.70. Con una muestra de tamaño 27 se ha encontrado una ecuación de regresión de Y

sobre X dada por $Y = 25.0 + 2.00X$. Si $s_{Y.X} = 1{,}50$, $s_X = 3.00$ y $\bar{X} = 7.50$, hallar los límites de confianza (*a*) 95% y (*b*) 99% para el coeficiente de regresión.

14.71. En el Problema 14.70, contrastar la hipótesis de que el coeficiente de regresión de la población al nivel de significación 0.01 es (*a*) tan bajo como 1.70 y (*b*) tan alto como 2.20.

14.72. En el Problema 14.70, hallar los límites de confianza (*a*) 95% y (*b*) 99% para Y cuando $X = 6.00$.

14.73. En el Problema 14.70, hallar los límites de confianza (*a*) 95% y (*b*) 99% para la media de todos los valores de Y correspondientes a $X = 6.00$.

14.74. Con referencia al Problema 14.46, hallar los límites de confianza del 95% para (*a*) el coeficiente de regresión de Y sobre X, (*b*) las presiones sanguíneas de las mujeres de 45 años y (*c*) la media de las presiones sanguíneas de las mujeres de 45 años.

CAPITULO 15

Correlación múltiple y parcial

CORRELACION MULTIPLE

El grado de correlación existente entre tres o más variables se llama *correlación múltiple*. Los principios fundamentales implicados en los problemas de correlación múltiple son análogos a los de la correlación simple, tratados en el Capítulo 14.

NOTACION DE SUBINDICES

Para permitir generalizaciones a números grandes de variables, conviene adoptar una notación de subíndices.

Denotaremos por X_1, X_2, X_3, ... las variables bajo consideración. Entonces denotaremos por $X_{11}, X_{12}, X_{13}, \ldots$ los valores que toma la variable X_1, y $X_{21}, X_{22}, X_{23}, \ldots$ los que toma la variable X_2, etcétera. Con esta notación, una suma tal como $X_{21} + X_{22} + X_{23} + \cdots + X_{2N}$ se escribirá $\sum_{j=1}^{N} X_{2j}$, $\sum_j X_{2j}$, o simplemente $\sum X_2$. Cuando no haya ambigüedad, usaremos la última notación. En tal caso, la media de X_2 se escribe $\bar{X}_2 = \sum X_2/N$.

ECUACIONES DE REGRESION Y PLANOS DE REGRESION

Una *ecuación de regresión* es una ecuación para estimar una variable dependiente, digamos X_1, a partir de las variables independientes X_2, X_3, ... y se llama una *ecuación de regresión de* X_1 *sobre* X_2, X_3, ... En notación funcional eso se escribe a veces brevemente como $X_1 = F(X_2, X_3, \ldots)$ (léase «X_1 es una función de X_2, X_3, etc.»).

Para el caso de tres variables, la ecuación de regresión más simple de X_1 sobre X_2 y X_3 tiene la forma

$$X_1 = b_{1.23} + b_{12.3}X_2 + b_{13.2}X_3 \tag{1}$$

donde $b_{1.23}$ $b_{12.3}$, y $b_{13.2}$ son constantes. Si mantenemos X_3 constante en la ecuación (1), el gráfico de X_1 versus X_2 es una recta con pendiente $b_{12.3}$. Si mantenemos constante X_2, el gráfico de X_1 versus X_3 es una recta con pendiente $b_{13.2}$. Es claro que los subíndices tras el punto indican las variables que se mantienen constantes en cada caso.

Debido al hecho de que X_1 varía parcialmente a causa de la variación en X_2 y parcialmente a

causa de la de X_3, se llama a $b_{12.3}$ y $b_{13.2}$ los *coeficientes de regresión parcial* de X_1 sobre X_2 dejando X_3 constante, y de X_1 sobre X_3 dejando X_2 constante, respectivamente.

La ecuación (1) se llama una *ecuación de regresión lineal* de X_1 sobre X_2 y X_3. En un sistema rectangular tridimensional de coordenadas representa un plano llamado *plano de regresión* y es generalización de la recta de regresión en dos variables, tal como se consideró en el Capítulo 13.

ECUACIONES NORMALES PARA EL PLANO DE REGRESION DE MINIMOS CUADRADOS

Así como existen rectas de regresión de mínimos cuadrados que aproximan un conjunto de N puntos dato (X, Y) en un diagrama de dispersión, existen también *planos de regresión de mínimos cuadrados* que ajustan un conjunto de N puntos dato (X_1, X_2, X_3) en un diagrama de dispersión tridimensional.

El plano de regresión de mínimos cuadrados de X_1 sobre X_2 y X_3 tiene ecuación (1) donde $b_{1.23}$, $b_{12.3}$ y $b_{13.2}$ se determinan resolviendo simultáneamente las *ecuaciones normales*

$$\begin{aligned} \sum X_1 &= b_{1.23}N + b_{12.3}\sum X_2 + b_{13.2}\sum X_3 \\ \sum X_1X_2 &= b_{1.23}\sum X_2 + b_{12.3}\sum X_2^2 + b_{13.2}\sum X_2X_3 \\ \sum X_1X_3 &= b_{1.23}\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum X_3^2 \end{aligned} \tag{2}$$

Estas pueden obtenerse formalmente multiplicando ambos lados de la ecuación (1) por 1, X_2 y X_3 sucesivamente y sumando en ambos lados.

A menos que se especifique lo contrario, siempre que nos refiramos a una ecuación de regresión se supondrá que se habla de la ecuación de regresión de mínimos cuadrados.

Si $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$ y $x_3 = X_3 - \bar{X}_3$, la ecuación de regresión de X_1 sobre X_2 y X_3 pueden escribirse más sencillamente como

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \tag{3}$$

donde $b_{12.3}$ y $b_{13.2}$ se obtienen resolviendo simultáneamente las ecuaciones

$$\begin{aligned} \sum x_1x_2 &= b_{12.3}\sum x_2^2 + b_{13.2}\sum x_2x_3 \\ \sum x_1x_3 &= b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2 \end{aligned} \tag{4}$$

Estas ecuaciones que son equivalentes a las ecuaciones normales (2) se pueden obtener formalmente multiplicando (3) por x_2 y x_3 sucesivamente y sumando (véase Prob. 15.8).

PLANOS DE REGRESION Y COEFICIENTES DE CORRELACION

Si los coeficientes de correlación entre variables X_1 y X_2, X_1 y X_3 y X_2 y X_3, tal como se calculaban en el Capítulo 14, se denotan respectivamente por r_{12}, r_{13} y r_{23} (llamados a veces *coeficientes de correlación de orden cero*), entonces el plano de regresión de mínimos cuadrados tiene la ecuación

$$\frac{x_1}{s_1} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\frac{x_2}{s_2} + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\frac{x_3}{s_3} \tag{5}$$

donde $x_1 = X - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$ y $x_3 = X_3 - \bar{X}_3$ y donde s_1, s_2 y s_3 son la desviación típica de X_1, X_2 y X_3, respectivamente (véase Prob. 15.9).

Nótese que si la variable X_3 no existiese y si $X_1 = Y$ y $X_2 = X$, entonces la ecuación (5) se reduce a la ecuación (25) del Capítulo 14.

ERROR TIPICO DE ESTIMACION

Por una generalización obvia de la ecuación 8 del Capítulo 14, podemos definir el *error típico de estimación de* X_1 sobre X_2 y X_3 como

$$s_{1.23} = \sqrt{\frac{\sum (X_1 - X_{1,\,\mathrm{est}})^2}{N}} \tag{6}$$

donde $X_{1,\,\mathrm{est}}$ indica los valores estimados de X_1 tal como se calculan mediante las ecuaciones de regresión (1) o (5).

En términos de los coeficientes de correlación r_{12}, r_{13} y r_{23}, el error típico de estimación se puede calcular también a partir del resultado

$$s_{1.23} = s_1 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \tag{7}$$

La interpretación muestral del error típico de estimación para dos variables, vista en la página 324 para el caso en que N es grande, puede extenderse a tres dimensiones sustituyendo las rectas paralelas a la de regresión por planos paralelos al plano de regresión. Una estimación mejor del error típico de estimación de la población viene dada por $\hat{s}_{1.23} = \sqrt{N/(N-3)}\, s_{1.23}$.

COEFICIENTE DE CORRELACION MULTIPLE

El *coeficiente de correlación múltiple* se define por extensión de la ecuación (12) o (14) del Capítulo 14. En el caso de dos variables independientes, por ejemplo, el coeficiente de correlación múltiple viene dado por

$$R_{1.23} = \sqrt{1 - \frac{s_{1.23}^2}{s_1^2}} \tag{8}$$

donde s_1 es la desviación típica de X_1 y $s_{1.23}$ viene dado por la ecuación (6) o (7). La cantidad $R_{1.23}^2$ se llama *coeficiente de determinación múltiple*.

Cuando se usa una ecuación de regresión lineal, el coeficiente de correlación múltiple se llama *coeficiente de correlación múltiple lineal*. Salvo que se especifique lo contrario, siempre que nos refiramos a correlación múltiple querremos decir correlación múltiple lineal.

En términos de r_{12}, r_{13} y r_{23}, la ecuación (8) se puede expresar

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \tag{9}$$

Un coeficiente de correlación múltiple, tal como $R_{1.23}$, está entre 0 y 1. Cuanto más cerca de 1, más precisa es la relación lineal entre las variables. Cuanto más cerca de 0, peor es la relación lineal. Si el coeficiente de correlación múltiple es 1, la correlación se dice *perfecta*. Aunque un coeficiente de correlación igual a 0 indica que no hay relación lineal entre las variables, puede haber una *relación no lineal*.

CAMBIO DE VARIABLE DEPENDIENTE

Los resultados anteriores son válidos cuando se considera a X_1 como variable dependiente. Sin embargo, si queremos considerar a X_3 (por ejemplo) como la variable dependiente en vez de X_1, sólo tendríamos que reemplazar los subíndices 1 por 3 y 3 por 1 en las fórmulas ya obtenidas. Por ejemplo, la ecuación de regresión de X_3 sobre X_1 y X_2 sería

$$\frac{x_3}{s_3} = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2}\right)\frac{x_2}{s_2} + \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right)\frac{x_1}{s_1} \tag{10}$$

que se deduce de (5) haciendo uso de $r_{32} = r_{23}$, $r_{31} = r_{13}$ y $r_{21} = r_{12}$.

GENERALIZACIONES A MAS DE TRES VARIABLES

Estas se obtienen por analogía con los resultados precedentes. Así, las ecuaciones de regresión lineales de X_1 sobre X_2, X_3 y X_4 pueden escribirse

$$X_1 = b_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 \tag{11}$$

y representan un *hiperplano en el espacio de cuatro dimensiones*. Multiplicando ambos miembros de (11) por 1, X_2, X_3 y X_4 sucesivamente y sumando, se llega a las ecuaciones normales para determinar $b_{1.234}$, $b_{12.34}$, $b_{13.24}$ y $b_{14.23}$; sustituyendo estas en la ecuación (11) nos da la *ecuación de regresión de mínimos cuadrados* de X_1 sobre X_2, X_3 y X_4. Esta ecuación de regresión de mínimos cuadrados se puede escribir de modo similar a la (5). (Véase Prob. 15.41.)

CORRELACION PARCIAL

A menudo es importante medir la correlación entre una variable dependiente y una variable independiente particular, cuando todas las demás variables se suprimen (indicado con frecuencia con la frase «quedando iguales las restantes»). Esto se consigue definiendo un *coeficiente de correlación parcial*, como en la ecuación (12) del Capítulo 14, excepto que hemos de considerar la variación explicada y la variación inexplicada que aparecen tanto con como sin la variable independiente particular.

Si denotamos por $r_{12.3}$ el coeficiente de correlación parcial entre X_1 y X_2 manteniendo X_3 constante encontramos que

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \tag{12}$$

De la misma manera, si $r_{12.34}$ es el coeficiente de correlación parcial entre X_1 y X_2 manteniendo X_3 y X_4 constante, entonces

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{(1 - r_{13.4}^2)(1 - r_{23.4}^2)}} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}} \tag{13}$$

Estos resultados son útiles porque por su mediación cualquier coeficiente de correlación parcial se puede hacer depender en última instancia de los coeficientes de correlación r_{12}, r_{23}, etc. (o sea, los *coeficientes de correlación de orden cero*).

En el caso de dos variables X e Y, si las dos rectas de regresión tienen ecuaciones $Y = a_0 + a_1X$ y $X = b_0 + b_1Y$, hemos visto que $r^2 = a_1b_1$ (véase Prob. 14.22). Este resultado admite generalización. Así, si

$$X_1 = b_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 \tag{14}$$

y

$$X_4 = b_{4.123} + b_{41.23}X_1 + b_{42.13}X_2 + b_{43.12}X_3 \tag{15}$$

son ecuaciones de regresión lineales de X_1 sobre X_2, X_3 y X_4 y de X_4 sobre X_1, X_2 y X_3, respectivamente, entonces

$$r_{14.23}^2 = b_{14.23}b_{41.23} \tag{16}$$

(véase Prob. 15.18). Esto se puede adoptar como punto de partida para una definición de los coeficientes de correlación parcial lineales.

RELACIONES ENTRE COEFICIENTES DE CORRELACION PARCIAL Y MULTIPLE

Hay interesantes resultados que conectan los coeficientes de correlación múltiple. Como ejemplo,

$$1 - R_{1.23}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2) \tag{17}$$

$$1 - R_{1.234}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2) \tag{18}$$

Es fácil generalizar estos resultados.

REGRESION MULTIPLE NO LINEAL

Los resultados anteriores para regresión múltiple lineal se pueden extender a la regresión múltiple no lineal. Se pueden definir coeficientes de correlación parcial y múltiple por métodos similares a los ya vistos.

PROBLEMAS RESUELTOS

ECUACION DE REGRESION EN TRES VARIABLES

15.1. Usando notación de subíndices adecuada, escribir la ecuación de regresión de (*a*) X_2 sobre X_1 y X_3; (*b*) X_3 sobre X_1, X_2 y X_4, y (*c*) X_5 sobre X_1, X_2, X_3 y X_4.

Solución

(*a*) $X_2 = b_{2.13} + b_{21.3}X_1 + b_{23.1}X_3$

(*b*) $X_3 = b_{3.124} + b_{31.24}X_1 + b_{32.14}X_2 + b_{34.12}X_4$

(*c*) $X_5 = b_{5.1234} + b_{51.234}X_1 + b_{52.134}X_2 + b_{53.124}X_3 + b_{54.123}X_4$

15.2. Escribir las ecuaciones normales correspondientes a la ecuación de regresión (*a*) $X_3 = b_{3.12} + b_{31.2}X_1 + b_{32.1}X_2$ y (*b*) $X_1 = b_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$.

Solución

(*a*) Multiplicar la ecuación sucesivamente por 1, X_1 y X_2, y sumar en ambos lados. Las ecuaciones normales son

$$\sum X_3 = b_{3.12}N + b_{31.2}\sum X_1 + b_{32.1}\sum X_2$$

$$\sum X_1X_3 = b_{3.12}\sum X_1 + b_{31.2}\sum X_1^2 + b_{32.1}\sum X_1X_2$$

$$\sum X_2X_3 = b_{3.12}\sum X_2 + b_{31.2}\sum X_1X_2 + b_{32.1}\sum X_2^2$$

(*b*) Multiplicar la ecuación sucesivamente por 1, X_2, X_3 y X_4, y sumar en ambos lados. Las ecuaciones normales son

$$\sum X_1 = b_{1.234}N + b_{12.34}\sum X_2 + b_{13.24}\sum X_3 + b_{14.23}\sum X_4$$

$$\sum X_1X_2 = b_{1.234}\sum X_2 + b_{12.34}\sum X_2^2 + b_{13.24}\sum X_2X_3 + b_{14.23}\sum X_2X_4$$

$$\sum X_1X_3 = b_{1.234}\sum X_3 + b_{12.34}\sum X_2X_3 + b_{13.24}\sum X_3^2 + b_{14.23}\sum X_3X_4$$

$$\sum X_1X_4 = b_{1.234}\sum X_4 + b_{12.34}\sum X_2X_4 + b_{13.24}\sum X_3X_4 + b_{14.23}\sum X_4^2$$

Nótese que esto no es una demostración de las ecuaciones normales, sino sólo un medio de acordarse de ellas.

El número de ecuaciones normales es igual al número de constantes desconocidas.

15.3. La Tabla 15.1 da los pesos X_1 redondeados en libras (lb), las alturas X_2 redondeadas en pulgadas (in), y las edades X_3 redondeadas en años, de niños.

(*a*) Hallar la ecuación de regresión de mínimos cuadrados de X_1 sobre X_2 y X_3.

(*b*) Determinar los valores estimados de X_1 a partir de los valores dados de X_2 y X_3.

(*c*) Estimar el peso de un niño de 9 años que mide 54 in.

Tabla 15.1

Peso (X_1)	64	71	53	67	55	58	77	57	56	51	76	68
Altura (X_2)	57	59	49	62	51	50	55	48	52	42	61	57
Edad (X_3)	8	10	6	11	8	7	10	9	10	6	12	9

Solución

(*a*) La ecuación de regresión lineal de X_1 sobre X_2 y X_3 puede expresarse

$$X_1 = b_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

Las ecuaciones normales de la ecuación de regresión de mínimos cuadrados son

$$\begin{aligned} \sum X_1 &= b_{1.23}N + b_{12.3}\sum X_2 + b_{13.2}\sum X_3 \\ \sum X_1X_2 &= b_{1.23}\sum X_2 + b_{12.3}\sum X_2^2 + b_{13.2}\sum X_2X_3 \\ \sum X_1X_3 &= b_{1.23}\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum X_3^2 \end{aligned} \tag{19}$$

El camino a seguir se indica en la Tabla 15.2. (Aunque la columna encabezada por X_1^2 no se necesita ahora, se ha añadido para referencia posterior.)

Tabla 15.2

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3
64	57	8	4096	3249	64	3648	512	456
71	59	10	5041	3481	100	4189	710	590
53	49	6	2809	2401	36	2597	318	294
67	62	11	4489	3844	121	4154	737	682
55	51	8	3025	2601	64	2805	440	408
58	50	7	3364	2500	49	2900	406	350
77	55	10	5929	3025	100	4235	770	550
57	48	9	3249	2304	81	2736	513	432
56	52	10	3136	2704	100	2912	560	520
51	42	6	2601	1764	36	2142	306	252
76	61	12	5776	3721	144	4636	912	732
68	57	9	4624	3249	81	3876	612	513
$\sum X_1 = 753$	$\sum X_2 = 643$	$\sum X_3 = 106$	$\sum X_1^2 = 48{,}139$	$\sum X_2^2 = 34{,}843$	$\sum X_3^2 = 976$	$\sum X_1X_2 = 40{,}830$	$\sum X_1X_3 = 6796$	$\sum X_2X_3 = 5779$

Usando la Tabla 15.2, las ecuaciones normales (19) pasan a ser

$$\begin{aligned} 12b_{1.23} + 643b_{12.3} + 106b_{13.2} &= 753 \\ 643b_{1.23} + 34{,}843b_{12.3} + 5.779b_{13.2} &= 40.830 \\ 106b_{1.23} + 5{,}779b_{12.3} + 976b_{13.2} &= 6{,}796 \end{aligned} \tag{20}$$

Resolviendo, $b_{1.23} = 3.6512$, $b_{12.3} = 0.8546$ y $b_{13.2} = 1.5063$, y la ecuación de regresión pedida será

$$X_1 = 3.6512 + 0.8546X_2 + 1.5063X_3 \quad \text{o sea} \quad X_1 = 3.65 + 0.855X_2 + 1.506\ X_3 \tag{21}$$

Para otro método, que evita resolver ecuaciones simultáneas, véase el Problema 15.6.

(*b*) Usando la ecuación de regresión (21), obtenemos los valores estimados de X_1, denotados por

$X_{1,\text{est}}$, sustituyendo los valores correspondientes de X_2 y X_3. Por ejemplo, sustituyendo $X_2 = 57$ y $X_3 = 8$ en (21), vemos que $X_{1,\text{est}} = 64.414$.

Los otros valores estimados de X_1 se obtienen del mismo modo. Se recogen en la Tabla 15.3 junto con los valores muestrales de X_1.

(*c*) Poniendo $X_2 = 54$ y $X_3 = 9$ en la ecuación (21), el peso estimado es $X_{1,\text{est}} = 63.356$, es decir, unas 63 lb.

Tabla 15.3

$X_{1,\text{est}}$	64.414	69.136	54.564	73.206	59.286	56.925	65.717	58.229	63.153	48.582	73.857	65.920
X_1	64	71	53	67	55	58	77	57	56	51	76	68

15.4. Calcular las derivaciones estándar (*a*) s_1, (*b*) s_2 y (*c*) s_3 para los datos del Problema 15.3.

Solución

(*a*) La cantidad s_1 es la desviación típica de la variable X_1. Entonces, usando la Tabla 15.2 del Problema 15.3(*a*) y los métodos del Capítulo 4, se ve que

$$s_1 = \sqrt{\frac{\sum X_1^2}{N} - \left(\frac{\sum X_1}{N}\right)^2} = \sqrt{\frac{48{,}139}{12} - \left(\frac{753}{12}\right)^2} = 8.6035 \quad \text{o sea} \quad 8.6 \text{ lb}$$

(*b*)
$$s_2 = \sqrt{\frac{\sum X_2^2}{N} - \left(\frac{\sum X_2}{N}\right)^2} = \sqrt{\frac{34{,}843}{12} - \left(\frac{643}{12}\right)^2} = 5.6930 \quad \text{o sea} \quad 5.7 \text{ in}$$

(*c*)
$$s_3 = \sqrt{\frac{\sum X_3^2}{N} - \left(\frac{\sum X_3}{N}\right)^2} = \sqrt{\frac{976}{12} - \left(\frac{106}{12}\right)^2} = 1.8181 \quad \text{o sea} \quad 1.8 \text{ años}$$

15.5. Calcular (*a*) r_{12}, (*b*) r_{13} y (*c*) r_{23} para los datos del Problema 15.3.

Solución

(*a*) La cantidad r_{12} es el coeficiente de correlación lineal entre las variables X_1 y X_2, ignorando la variable X_3. Entonces, usando los métodos del Capítulo 14, se tiene

$$r_{12} = \frac{N\sum X_1X_2 - (\sum X_1)(\sum X_2)}{\sqrt{[N\sum X_1^2 - (\sum X_1)^2][N\sum X_2^2 - (\sum X_2)^2]}} =$$

$$= \frac{(12)(40{,}830) - (753)(643)}{\sqrt{[(12)(48.139) - (753)^2][(12)(34.843) - (643)^2]}} = 0.8196 \quad \text{o sea} \quad 0.82$$

(*b*) y (*c*) Usando las fórmulas correspondientes, se obtiene $r_{12} = 0.7698$, o sea 0.77 y $r_{23} = 0.7984$, ó 0.80.

15.6. Resolver el Problema 15.3(*a*) usando la ecuación (5) y los resultados de los Problemas 15.4 y 15.5.

Solución

La ecuación de regresión de X_1 sobre X_2 y X_3 es, multiplicando cada miembro de la ecuación (5) por s_1,

$$x_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_2}\right)x_2 + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_3}\right)x_3 \tag{22}$$

donde $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$ y $x_3 = X_3 - \bar{X}_3$. Usando los resultados de los Problemas 15.4 y 15.5, la (22) se convierte en

$$x_1 = 0.8546x_2 + 1.5063x_3$$

Como $\bar{X}_1 = \dfrac{\sum X_1}{N} = \dfrac{753}{12} = 62.750$ $\bar{X}_2 = \dfrac{\sum X_2}{N} = 53.583$ y $\bar{X}_3 = 8.833$

(por la Tabla 15.2 del Prob. 15.3), la requerida ecuación se puede expresar

$$X_1 - 62.750 = 0.8546(X_2 - 53.583) + 1.506(X_3 - 8.833)$$

que coincide con el resultado del Problema 15.3(*a*).

15.7. Para los datos del Problema 15.3, determinar (*a*) el crecimiento promedio en peso por pulgada de crecimiento en altura, para niños de la misma edad y (*b*) el crecimiento promedio en peso por año, para niños de la misma altura.

Solución

De la ecuación de regresión obtenida en el Problema 15.3(*a*) o en el 15.6 vemos que la respuesta a (*a*) es 0.8546, o sea unas 0.9 lb, y la de (*b*) es 1.5063 lb, o sea unas 1.5 lb.

15.8. Probar que las ecuaciones (3) y (4) de este capítulo se siguen de las ecuaciones (1) y (2).

Solución

De la primera de las ecuaciones (2), dividiendo ambos lados por N, se tiene

$$\bar{X}_1 = b_{1.23} + b_{12.3}\bar{X}_2 + b_{13.2}\bar{X}_3 \tag{23}$$

Restando (23) de (1) vemos que

$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

o

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \tag{24}$$

que no es sino la ecuación (3).

Sean $X_1 = x_1 + \bar{X}_1$, $X_2 = x_2 + \bar{X}_2$ y $X_3 = x_3 + \bar{X}_3$ en la segunda y tercera ecuaciones (2). Entonces, tras algunas manipulaciones algebraicas, usando los resultados $\sum x_1 = \sum x_2 = \sum x_3 = 0$, pasan a ser

$$\sum x_1x_2 = b_{12.3}\sum x_2^2 + b_{13.2}\sum x_2x_3 + N\bar{X}_2[b_{1.23} + b_{12.3}\bar{X}_2 + b_{13.2}\bar{X}_3 - \bar{X}_1] \tag{25}$$

$$\sum x_1x_3 = b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2 + N\bar{X}_3[b_{1.23} + b_{12.3}\bar{X}_2 + b_{13.2}\bar{X}_3 - \bar{X}_1] \tag{26}$$

que se reducen a (4) pues las cantidades entre corchetes de la derecha en las ecuaciones (25) y (26) son cero debido a la ecuación (1).

Otro método

Véase Problema 15.30.

15.9. Establecer la ecuación (5), que copiamos aquí:

$$\frac{x_1}{s_1} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\frac{x_2}{s_2} + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\frac{x_3}{s_3} \tag{5}$$

Solución

De las ecuaciones (25) y (26)

$$\begin{aligned} b_{12.3}\sum x_2^2 \quad + b_{13.2}\sum x_2x_3 &= \sum x_1x_2 \\ b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2 \quad &= \sum x_1x_3 \end{aligned} \tag{27}$$

Como $s_2^2 = \dfrac{\sum x_2^2}{N}$ y $s_3^2 = \dfrac{\sum x_3^2}{N}$

$\sum x_2^2 = Ns_2^2$ y $\sum x_3^2 = Ns_3^2$. Puesto que

$$r_{23} = \frac{\sum x_2x_3}{\sqrt{(\sum x_2^2)(\sum x_3^2)}} = \frac{\sum x_2x_3}{Ns_2s_3}$$

$\sum x_2x_3 = Ns_2s_3r_{23}$. Análogamente, $\sum x_1x_2 = Ns_1s_2r_{12}$ y $\sum x_1x_3 = Ns_1s_3r_{13}$. Sustituyendo en (27) y simplificando, hallamos

$$\begin{aligned} b_{12.3}s_2 \quad + b_{13.2}s_3r_{23} &= s_1r_{12} \\ b_{12.3}s_2r_{23} + b_{13.2}s_3 \quad &= s_1r_{13} \end{aligned} \tag{28}$$

Resolviendo simultáneamente, tenemos

$$b_{12.3} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_2}\right) \quad \text{y} \quad b_{13.2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_3}\right)$$

que sustituidas en la ecuación $x_1 = b_{12.3}x_2 + b_{13.2}x_3$ [ecuación (24)] y dividiendo por s_1, dan el resultado anunciado.

ERROR TIPICO DE ESTIMACION

15.10. Calcular el error típico de estimación de X_1 sobre X_2 y X_3 para los datos del Problema 15.3.

Solución

De la Tabla 15.3 del Problema 15.3(*b*) vemos que

$$s_{1.23} = \sqrt{\frac{\sum (X_1 - X_{1.\text{est}})^2}{N}} = \sqrt{\frac{(64 - 64.414)^2 + (71 - 69.136)^2 + \cdots + (68 - 65.920)^2}{12}} = 4.6447 \text{ o sea } 4.6 \text{ lb}$$

El error típico de estimación de la población se estima como $\hat{s}_{1.23} = \sqrt{N/(N-3)}s_{1.23} = 5.3$ lb en este caso.

15.11. Deducir el resultado del Problema 15.10, usando

$$s_{1.23} = s_1\sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

Solución

Por los Problemas 15.4(*a*) y 15.5 tenemos

$$s_{1.23} = 8.6035\sqrt{\frac{1 - (0.8196)^2 - (0.7698)^2 - (0.7984)^2 + 2(0.8196)(0.7698)(0.7984)}{1 - (0.7984)^2}} = 4.6 \text{ lb}$$

Nótese que con el método de este problema el error típico de estimación se puede encontrar sin recurrir a la ecuación de regresión.

COEFICIENTE DE CORRELACION MULTIPLE

15.12. Calcular el coeficiente de correlación múltiple lineal de X_1 sobre X_2 y X_3 para los datos del Problema 15.3.

Solución

Primer método

De los resultados de los Problemas 15.4(*a*) y 15.10 tenemos

$$R_{1.23} = \sqrt{1 - \frac{s_{1.23}^2}{s_1^2}} = \sqrt{1 - \frac{(4.6447)^2}{(8.6035)^2}} = 0.8418$$

Segundo método

De los resultados del Problema 15.5 tenemos

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} = \sqrt{\frac{(0.8169)^2 + (0.7698)^2 - 2(0.8196)(0.7698)(0.7984)}{1 - (0.7984)^2}} = 0.8418$$

Obsérvese que el coeficiente de correlación múltiple, $R_{1.23}$, es mayor que cualquiera de los coeficientes r_{12} o r_{13} (véase Prob. 15.5). Esto ocurre siempre y era de esperar, de hecho, ya que teniendo en cuenta variables independientes relevantes adicionales llegaríamos a una relación más exacta entre las variables.

15.13. Calcular el coeficiente de determinación múltiple de X_1 sobre X_2 y X_3 para los datos del Problema 15.3.

Solución

El coeficiente de determinación múltiple de X_1 sobre X_2 y X_3 es

$$R_{1.23}^2 = (0.8418)^2 = 0.7086$$

usando el Problema 15.12. Así pues, alrededor del 71% de la variación total de X es explicada por la ecuación de regresión.

15.14. Para los datos del Problema 15.3, calcular (*a*) $R_{2.13}$ y (*b*) $R_{3.12}$ y comparar sus valores con el valor de $R_{1.23}$.

Solución

(*a*) $$R_{2.13}=\sqrt{\frac{r_{12}^2+r_{23}^2-2r_{12}r_{13}r_{23}}{1-r_{13}^2}}=\sqrt{\frac{(0.8196)^2+(0.7984)^2-2(0.8196)(0.7698)(0.7984)}{1-(0.7698)^2}}=0.8606$$

(*b*) $$R_{3.12}=\sqrt{\frac{r_{13}^2+r_{23}^2-2r_{12}r_{13}r_{23}}{1-r_{12}^2}}=\sqrt{\frac{(0.7698)^2+(0.7984)^2-2(0.8196)(0.7698)(0.7984)}{1-(0.8196)^2}}=0.8234$$

Este problema ilustra el hecho de que, en general, $R_{2.13}$, $R_{3.12}$ y $R_{1.23}$ no son necesariamente iguales, como se ve comparando con el Problema 15.12.

15.15. Si $R_{1.23} = 1$, probar que (*a*) $R_{2.13} = 1$ y (*b*) $R_{3.12} = 1$.

Solución

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \tag{29}$$

y

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}} \tag{30}$$

(*a*) En la ecuación (29), poniendo $R_{1.23} = 1$ y elevando al cuadrado ambos lados, $r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23} = 1 - r_{23}^2$. Entonces

$$r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23} = 1 - r_{13}^2 \quad \text{o sea} \quad \frac{r_{12}^2 + r_{23}^2 - 2r_1r_{13}r_{23}}{1 - r_{13}^2} = 1$$

Esto es, $R_{2.13}^2 = 1$ o sea $R_{2.13} = 1$, ya que el coeficiente de correlación múltiple se considera no negativo.

(*b*) $R_{3.12} = 1$ se sigue de la parte (*a*) intercambiando los subíndices 2 y 3 en el resultado $R_{2.13} = 1$.

15.16. Si $R_{1.23} = 0$, ¿se deduce necesariamente que $R_{2.13} = 0$?

Solución

De la ecuación (29), $R_{1.23} = 0$ si y sólo si

$$r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23} = 0 \quad \text{o sea} \quad 2r_{12}r_{13}r_{23} = r_{12}^2 + r_{13}^2$$

Entonces, de la ecuación (30) tenemos

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - (r_{12}^2 + r_{13}^2)}{1 - r_{13}^2}} = \sqrt{\frac{r_{23}^2 - r_{13}^2}{1 - r_{13}^2}}$$

que no es necesariamente cero.

CORRELACION PARCIAL

15.17. Para los datos del Problema 15.3, calcular los coeficientes de correlación parcial lineal (*a*) $r_{12.3}$, (*b*) $r_{13.2}$ y (*c*) $r_{23.1}$.

Solución

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \qquad r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}} \qquad r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}}$$

Para los resultados del Problema 15.5 sabemos que $r_{12.3} = 0.5334$, $r_{13.2} = 0.3346$ y $r_{23.1} = 0.4580$. Se sigue que para niños de la misma edad, el coeficiente de correlación entre peso y edad es 0.53; para niños del mismo peso, el coeficiente de correlación entre peso y edad es sólo 0.33. Como estos resultados se basan en una muestra pequeña de sólo 12 niños, no son, claro está, tan fiables como los que se obtendrían con una muestra grande.

15.18. Si $X_1 = b_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$ y $X_3 = b_{3.12} + b_{32.1}X_2 + b_{31.2}X_1$ son la ecuación de regresión de X_1 sobre X_2 y X_3 y de X_3 sobre X_2 y X_1, respectivamente, probar que $r_{13.2}^2 = b_{13.2}b_{31.2}$.

Solución

La ecuación de regresión de X_1 sobre X_2 y X_3 se puede escribir [véase ecuación (5) de este capítulo]

$$X_1 - \bar{X}_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_3}\right)(X_3 - \bar{X}_3) \tag{31}$$

La ecuación de regresión de X_3 sobre X_2 y X_1 se puede escribir [véase ecuación (10)]

$$X_3 - \bar{X}_3 = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_3}{s_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_3}{s_1}\right)(X_1 - \bar{X}_1) \tag{32}$$

De (31) y (32) los coeficientes de X_3 y X_1 son, respectivamente,

$$b_{13.2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_3}\right) \quad \text{y} \quad b_{31.2} = \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_3}{s_1}\right)$$

Luego
$$b_{13.2}b_{31.2} = \frac{(r_{13} - r_{12}r_{23})^2}{(1 - r_{23}^2)(1 - r_{12}^2)} = r_{13.2}^2$$

15.19. Si $r_{12.3} = 0$, demostrar que

$$(a)\ r_{13.2} = r_{13}\sqrt{\frac{1 - r_{23}^2}{1 - r_{12}^2}} \qquad (b)\ r_{23.1} = r_{23}\sqrt{\frac{1 - r_{13}^2}{1 - r_{12}^2}}$$

Solución

Si
$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = 0$$

tenemos $r_{12} = r_{13}r_{23}$

$$(a)\quad r_{13.2}=\frac{r_{13}-r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}=\frac{r_{13}-(r_{13}r_{23})r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}=\frac{r_{13}(1-r_{23}^2)}{\sqrt{(1-r_{12}^2)(1-r_{23}^3)}}=r_{13}\sqrt{\frac{1-r_{23}^2}{1-r_{12}^2}}$$

(*b*) Intercambiar los subíndices 1 y 2 en el resultado de la parte (*a*).

CORRELACION MULTIPLE Y PARCIAL EN CUATRO O MAS VARIABLES

15.20. Un examen de ingreso en cierta universidad consistía de tres partes; matemáticas, inglés y cultura general. Para analizar la capacidad del examen a la hora de predecir el rendimiento en un curso de estadística, se estudiaron los datos de 200 estudiantes. Llamando

X_1 = nota en estadística X_3 = nota en inglés
X_2 = nota en matemáticas X_4 = nota en cultura general

se han obtenido los siguientes resultados:

$$\bar{X}_1 = 75 \quad s_1 = 10 \quad \bar{X}_2 = 24 \quad s_2 = 5$$
$$\bar{X}_3 = 15 \quad s_3 = 3 \quad \bar{X}_4 = 36 \quad s_4 = 6$$
$$r_{12} = 0.90 \quad r_{13} = 0.75 \quad r_{14} = 0.80 \quad r_{23} = 0.70 \quad r_{24} = 0.70 \quad r_{34} = 0.85$$

Hallar la ecuación de regresión de mínimos cuadrados de X_1 sobre X_2, X_3 y X_4.

Solución

Generalizando el resultado del Problema 15.8, podemos escribir la ecuación de regresión de mínimos cuadrados de X_1 sobre X_2, X_3 y X_4 en la forma

$$x_1 = b_{12.34}x_2 + b_{13.24}x_3 + b_{14.23}x_4 \tag{33}$$

donde $b_{12.34}$, $b_{13.24}$ y $b_{14.23}$ pueden obtenerse de las ecuaciones normales

$$\begin{aligned}
\sum x_1x_2 &= b_{12.34}\sum x_2^2 + b_{13.24}\sum x_2x_3 + b_{14.23}\sum x_2x_4 \\
\sum x_1x_3 &= b_{12.34}\sum x_2x_3 + b_{13.24}\sum x_3^2 + b_{14.23}\sum x_3x_4 \\
\sum x_1x_4 &= b_{12.34}\sum x_2x_4 + b_{13.24}\sum x_3x_4 + b_{14.23}\sum x_4^2
\end{aligned} \tag{34}$$

y donde $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$, $x_3 = X_3 - \bar{X}_3$ y $x_4 = X_4 - \bar{X}_4$.

De los datos, deducimos

$$\sum x_2^2 = Ns_2^2 = 5000 \quad \sum x_1x_2 = Ns_1s_2r_{12} = 9000 \quad \sum x_2x_3 = Ns_1s_3r_{23} = 2100$$
$$\sum x_3^2 = Ns_3^2 = 1800 \quad \sum x_1x_3 = Ns_1s_3r_{13} = 4500 \quad \sum x_2x_4 = Ns_2s_4r_{24} = 4200$$
$$\sum x_4^2 = Ns_4^2 = 7200 \quad \sum x_1x_4 = Ns_1s_4r_{14} = 9600 \quad \sum x_3x_4 = Ns_3s_4r_{34} = 3060$$

Poniendo esos resultados en las ecuaciones (34), obtenemos

$$b_{12.34} = 1.3333 \quad b_{13.24} = 0.0000 \quad b_{14.23} = 0.5556 \tag{35}$$

que, al ser sustituidos en (33), dan la ecuación de regresión pedida

$$x_1 = 1.3333x_2 + 0.0000x_3 + 0.5556x_4$$

o sea $$X_1 - 75 = 1.3333(X_2 - 24) + 0.5556(X_4 - 27) \tag{36}$$

es decir $$X_1 = 22.9999 + 1.3333X_2 + 0.5556X_4$$

Una solución exacta de las ecuaciones (34) da $b_{12.34} = \frac{4}{3}$, $b_{13.24} = 0$ y $b_{14.23} = \frac{5}{9}$, así que la ecuación de regresión se puede también escribir como

$$X_1 = 23 + \tfrac{4}{3}X_2 + \tfrac{5}{9}X_4 \tag{37}$$

Es interesante observar que la ecuación de regresión no involucra la nota de inglés X_3. Ello no quiere decir que el conocimiento del inglés no tenga peso en el rendimiento en estadística. Más bien, significa que la necesidad del inglés, en lo que concierne a la predicción del rendimiento en estadística, queda ampliamente reflejada en las notas de las restantes materias.

15.21. Dos estudiantes obtuvieron en el examen del Problema 15.20 notas respectivas de (*a*) 30 en matemáticas, 18 en inglés y 32 en cultura general y (*b*) 18 en matemáticas, 20 en inglés y 36 en cultura general. ¿Cuál sería la predicción para sus notas en estadística?

Solución

(*a*) Sustituyendo $X_2 = 30$, $X_3 = 18$ y $X_4 = 32$ en (37), la predicción de la nota en estadística es $X_1 = 81$.

(*b*) Procediendo como en la parte (*a*) con $X_2 = 18$, $X_3 = 20$ y $X_4 = 36$, vemos que $X_1 = 67$.

15.22. Para los datos del Problema 15.20, hallar los coeficientes de correlación parcial (*a*) $r_{12.34}$, (*b*) $r_{13.24}$ y (*c*) $r_{14.23}$.

Solución

(*a*) y (*b*)

$$r_{12.4} = \frac{r_{12} - r_{14}r_{24}}{\sqrt{(1-r_{14}^2)(1-r_{24}^2)}} \qquad r_{13.4} = \frac{r_{13} - r_{14}r_{34}}{\sqrt{(1-r_{14}^2)(1-r_{34}^2)}} \qquad r_{23.4} = \frac{r_{23} - r_{24}r_{34}}{\sqrt{(1-r_{24}^2)(1-r_{34}^2)}}$$

Sustituyendo los valores del Problema 15.20, obtenemos $r_{12.4} = 0.7935$, $r_{13.4} = 0.2215$ y $r_{23.4} = 0.2791$. Luego

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{(1-r_{13.4}^2)(1-r_{23.4}^2)}} = 0.7814 \qquad \text{y} \qquad r_{13.24} = \frac{r_{13.4} - r_{12.4}r_{23.4}}{\sqrt{(1-r_{12.4}^2)(1-r_{23.4}^2)}} = 0.0000$$

(*c*)

$$r_{14.3} = \frac{r_{14} - r_{13}r_{34}}{\sqrt{(1-r_{13}^2)(1-r_{34}^2)}} \qquad r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \qquad r_{24.3} = \frac{r_{24} - r_{23}r_{34}}{\sqrt{(1-r_{23}^2)(1-r_{34}^2)}}$$

Sustituyendo los valores del Problema 15.20, obtenemos $r_{14.3} = 0.4664$, $r_{12.3} = 0.7939$ y $r_{24.3} = 0.2791$. Por tanto

$$r_{14.23} = \frac{r_{14.3} - r_{12.3}r_{24.3}}{\sqrt{1-r_{12.3}^2)(1-r_{24.3}^2)}} = 0.4193$$

15.23. Interpretar los coeficientes de correlación parcial (*a*) $r_{12.4}$, (*b*) $r_{13.4}$, (*c*) $r_{12.34}$, (*d*) $r_{14.3}$ y (*e*) $r_{14.23}$.

Solución

(*a*) $r_{12.4} = 0.7935$ representa el coeficiente de correlación (lineal) entre las notas de estadística y matemáticas para estudiantes con iguales notas en cultura general. Al obtener este coeficiente, las notas en inglés (así como otros factores que no se han tenido en cuenta) no se consideran, como lo evidencia el hecho de que el subíndice 3 se ha omitido.

(*b*) $r_{13.4} = 0.2215$ representa el coeficiente de correlación entre las notas de estadística e inglés para estudiantes con la misma nota en cultura general. Ahora, las notas en matemáticas no se han considerado.

(*c*) $r_{12.34} = 0.7814$ representada el coeficiente de correlación entre las notas de estadística y matemáticas para estudiantes con la misma nota en inglés y en cultura general.

(*d*) $r_{14.3} = 0.4664$ representa el coeficiente de correlación entre las notas de estadística y cultura general para estudiantes con la misma nota en inglés.

(*e*) $r_{14.23} = 0.4193$ representa el coeficiente de correlación entre las notas de estadística y cultura general para estudiantes con iguales notas en matemáticas e inglés.

15.24. (*a*) Para los datos del Problema 15.20, mostrar que

$$\frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{(1 - r_{13.4}^2)(1 - r_{23.4}^2)}} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}} \tag{38}$$

(*b*) Explicar el significado de la igualdad en la parte (*a*).

Solución

(*a*) El lado izquierdo de (38) se calcula en el Problema 15.22(*a*), con el resultado 0.7814. Para calcular el lado derecho, usamos el Problema 15.22(*c*); de nuevo, resulta 0.7814. Luego la igualdad es válida en este caso especial. Se puede demostrar, por métodos algebraicos directos, que la igualdad es válida en general.

(*b*) El lado izquierdo de (38) es $r_{12.34}$, y el lado derecho es $r_{12.43}$. Como $r_{12.34}$ es la correlación entre X_1 y X_2 dejando X_3 y X_4 constantes, mientras que $r_{12.43}$ es la correlación entre X_1 y X_2 dejando X_4 y X_3 constantes, salta a la vista por qué es cierta la igualdad.

15.25. Para los datos del Problema 15.20, hallar (*a*) el coeficiente de correlación múltiple $R_{1.234}$ y (*b*) el error típico de estimación $S_{1.234}$.

Solución

(*a*) $$1 - R_{1.234}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2) \quad \text{o sea} \quad R_{1.234} = 0.9310$$

como $r_{12} = 0.90$ por el Problema 15.20, $r_{14.23} = 0.4193$ por el Problema 15.22(*c*), y

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}} = \frac{0.75 - (0.90)(0.70)}{\sqrt{[1 - (0.90)^2][1 - (0.70)^2]}} = 0.3855$$

Otro método

Intercambiando los subíndices 2 y 4 en la primera ecuación se deduce

$$1 - R_{1.234}^2 = (1 - r_{14}^2)(1 - r_{13.4}^2)(1 - r_{12.34}^2) \quad \text{o sea} \quad R_{1.234} = 0.9319$$

donde se ha hecho uso directo de los resultados del Problema 15.22(*a*).

(*b*) $R_{1.234}=\sqrt{\dfrac{1-s_{1.234}^2}{s_1^2}}$ o sea $s_{1.234}=s_1\sqrt{1-R_{1.234}^2}=10\sqrt{1-(0.9310)^2}=3.659$

Comparar con la ecuación (8) de este capítulo.

PROBLEMAS SUPLEMENTARIOS

ECUACION DE REGRESION EN TRES VARIABLES

15.26. Usando notación de subíndices adecuada, escribir las ecuaciones de regresión (*a*) X_3 sobre X_1 y X_2 y (*b*) X_4 sobre X_1, X_2, X_3 y X_5.

15.27. Escribir las ecuaciones normales correspondientes a la ecuación de regresión de (*a*) X_2 sobre X_1 y X_3 y (*b*) X_5 sobre X_1, X_2, X_3 y X_4.

15.28. La Tabla 15.4 muestra los valores correspondientes de tres variables: X_1, X_2 y X_3.

(*a*) Hallar la ecuación de regresión de mínimos cuadrados de X_3 sobre X_1 y X_2.
(*b*) Estimar X_3 cuando $X_1 = 10$ y $X_2 = 6$.

Tabla 15.4

X_1	3	5	6	8	12	14
X_2	16	10	7	4	3	2
X_3	90	72	54	42	30	12

15.29. Un profesor de matemáticas desea determinar la relación de las notas del examen final con las de dos parciales anteriores. Llamando X_1, X_2 y X_3 a las notas en el primer parcial, segundo parcial y examen final, efectuó los siguientes cálculos para un total de 120 estudiantes:

$$\bar{X}_1 = 6.8 \qquad \bar{X}_2 = 7.0 \qquad \bar{X}_3 = 74$$
$$s_1 = 1.0 \qquad s_2 = 0.80 \qquad s_2 = 9.0$$
$$r_{12} = 0.60 \qquad r_{13} = 0.70 \qquad r_{23} = 0.65$$

(*a*) Hallar la ecuación de regresión de mínimos cuadrados de X_3 sobre X_1 y X_2.
(*b*) Estimar las notas finales de dos estudiantes cuyas respectivas notas en los parciales fueron (1) 9 y 7 y (2) 4 y 8.

15.30. Resolver el Problema 15.8, enunciado anteriormente, escogiendo las variables X_2 y X_3 tales que $\sum X_2 = \sum X_3 = 0$.

ERROR TIPICO DE ESTIMACION

15.31. Para los datos del Problema 15.28, hallar el error típico de estimación de X_3 sobre X_1 y X_2.

15.32. Para los datos del Problema 15.29, hallar el error típico de estimación de (*a*) X_3 sobre X_1 y X_2 y (*b*) X_1 sobre X_2 y X_3.

COEFICIENTE DE CORRELACION MULTIPLE

15.33. Para los datos del Problema 15.28, calcular el coeficiente de correlación múltiple de X_3 sobre X_1 y X_2.

15.34. Para los datos del Problema 15.29, calcular (*a*) $R_{3.12}$, (*b*) $R_{1.23}$ y (*c*) $R_{2.13}$.

15.35. (*a*) Si $r_{12} = r_{13} = r_{23} = r \neq 1$, mostrar que

$$R_{1.23} = R_{2.31} = R_{3.12} = \frac{r\sqrt{2}}{\sqrt{1+r}}$$

(*b*) Discutir el caso $r = 1$.

15.36. Si $R_{1.23} = 0$, probar que $|r_{23}| \geq |r_{12}|$ y $|r_{23}| \geq |r_{13}|$ e interpretar

CORRELACION PARCIAL

15.37. Calcular los coeficientes de correlación parcial lineal (*a*) $r_{12.3}$, (*b*) $r_{13.2}$ y (*c*) $r_{23.1}$ para los datos del Problema 15.28 e interpretar la respuesta.

15.38. Rehacer el Problema 15.37 para los datos del Problema 15.29.

15.39. Si $r_{12} = r_{13} = r_{23} = r \neq 1$, probar que $r_{12.3} = r_{13.2} = r_{23.1} = r/(1 + r)$. Discutir el caso $r = 1$.

15.40. Si $r_{12.3} = 1$, probar que (*a*) $|r_{13.2}| = 1$, (*b*) $|r_{23.1}| = 1$, (*c*) $R_{1.23} = 1$ y (*d*) $s_{1.23} = 0$.

CORRELACION MULTIPLE Y PARCIAL EN CUATRO O MAS VARIABLES

15.41. Probar que la ecuación de regresión de X_4 sobre X_1, X_2 y X_3 puede escribirse

$$\frac{x_4}{s_4} = a_1\left(\frac{x_1}{s_1}\right) + a_2\left(\frac{x_2}{s_2}\right) + a_3\left(\frac{x_3}{s_3}\right)$$

donde a_1, a_2 y a_3 vienen determinados al resolver simultáneamente las ecuaciones

$$a_1 r_{11} + a_2 r_{12} + a_3 r_{13} = r_{14}$$
$$a_1 r_{21} + a_2 r_{22} + a_3 r_{23} = r_{24}$$
$$a_1 r_{31} + a_2 r_{32} + a_3 r_{33} = r_{34}$$

y donde $x_j = X_j - \bar{X}_j$, $r_{jj} = 1$ y $j = 1, 2, 3$ y 4. Generalizar al caso de más de cuatro variables.

15.42. Dados $\bar{X}_1 = 20$, $\bar{X}_2 = 36$, $\bar{X}_3 = 12$, $\bar{X}_4 = 80$, $s_1 = 1.0$, $s_2 = 2.0$, $s_3 = 1.5$, $s_4 = 6.0$, $r_{12} = -0.20$, $r_{13} = 0.40$, $r_{23} = 0.50$, $r_{14} = 0.40$, $r_{24} = 0.30$ y $r_{34} = -0.10$, (*a*) hallar la ecuación de regresión de X_4 sobre X_1, X_2 y X_3 y (*b*) estimar X_4 cuando $X_1 = 15$, $X_2 = 40$ y $X_3 = 14$.

15.43. Hallar (*a*) $r_{41.23}$, (*b*) $r_{42.13}$ y (*c*) $r_{43.12}$ para los datos del Problema 15.42 e interpretar el resultado.

15.44. Para los datos del Problema 15.42, hallar (*a*) $R_{4.123}$ y (*b*) $s_{4.123}$.

15.45. Un científico ha coleccionado datos relativos a cuatro variables T, U, V y W. Piensa que una ecuación de la forma $W = aT^bU^cV^d$, donde a, b, c y d son constantes desconocidas, podría ser válida para determinar W a partir del conocimiento de T, U y V. Describir un procedimiento por el cual se pueda lograr ese objetivo. [*Ayuda:* Tomar logaritmos en ambos lados de esa ecuación.]

CAPITULO 16

Análisis de varianza

OBJETIVO DEL ANALISIS DE VARIANZA

En el Capítulo 8 hemos usado la teoría del muestreo para contrastar la significación de diferencias entre dos medias muestrales, en el supuesto de que las dos poblaciones de las que se tomaban las muestras tenían la misma varianza. En muchas situaciones es necesario hacer eso mismo con tres o más medias muestrales, o sea, equivalentemente, contrastar la hipótesis de que todas las medias son iguales.

EJEMPLO 1. Supongamos que en un experimento agrario, cuatro tratamientos químicos con abonos distintos han producido cosechas medias de trigo de 28, 22, 18 y 24 bushels por acre. ¿Hay diferencia significativa en esas medias o la dispersión se debe simplemente al azar?

Problemas como éste se pueden resolver usando una importante técnica conocida como *análisis de varianza*, desarrollada por Fisher. Hace uso de la distribución F ya considerada en el Capítulo 11.

EXPERIMENTOS DE FACTOR UNICO

En un *experimento de un factor*, las medidas (u observaciones) se obtienen para a grupos independientes de muestras, donde el número de medidas en cada grupo es b. Hablamos de *a tratamientos*, cada uno de los cuales tiene b *repeticiones* o *réplicas*. En el Ejemplo 1, $a = 4$.

Los resultados de un experimento de un factor se pueden presentar en una tabla con a filas y b columnas, como indica la Tabla 16.1. Aquí X_{jk} denota la medida en la j-ésima fila y en la k-ésima columna, donde $j = 1, 2, ..., a$ y donde $k = 1, 2, ..., b$. Por ejemplo, X_{35} se refiere a la quinta medida para el tercer tratamiento.

Tabla 16.1

Tratamiento 1	$X_{11}, X_{12}, ..., X_{1b}$	$\bar{X}_1$
Tratamiento 2	$X_{21}, X_{22}, ..., X_{2b}$	$\bar{X}_2$
⋮	⋮	⋮
Tratamiento a	$X_{a1}, X_{a2}, ..., X_{ab}$	$\bar{X}_a$

Denotaremos por $\bar{X}_{j.}$ la media de las medidas en la fila j-ésima. Tenemos

$$\bar{X}_{j.} = \frac{1}{b}\sum_{k=1}^{b} X_{jk} \qquad j = 1, 2, ..., a \tag{1}$$

El punto en $\bar{X}_{j.}$ se usa para anunciar que el índice k se ha sumado. Los valores $\bar{X}_{j.}$ se llaman *medias de grupo, medias de tratamiento* o *medias de fila*. La *media global* es la media de todas las medidas en todos los grupos y se denota por $\bar{X}$:

$$\bar{X} = \frac{1}{ab}\sum_{j=1}^{a}\sum_{k=1}^{b} X_{jk} \tag{2}$$

VARIACION TOTAL, VARIACION DENTRO DE LOS TRATAMIENTOS Y VARIACION ENTRE TRATAMIENTOS

Definimos la *variación total*, denotada por V, como la suma de los cuadrados de las desviaciones de cada medida respecto de la media global $\bar{X}$:

$$\text{Variación total} = V = \sum_{j,k} (X_{jk} - \bar{X})^2 \tag{3}$$

Escribiendo la identidad

$$X_{jk} - \bar{X} = (X_{jk} - \bar{X}_{j.}) + (\bar{X}_{j.} - \bar{X}) \tag{4}$$

elevando al cuadrado y sumando en j y k, se tiene (Prob. 16.1)

$$\sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 + \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 \tag{5}$$

o sea

$$\sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 + b\sum_{j} (\bar{X}_{j.} - \bar{X})^2 \tag{6}$$

Llamamos a la primera suma de la derecha de (5) y (6) la *variación dentro de los tratamientos* (puesto que implica a los cuadrados de las desviaciones de X_{jk} respecto de las medias de tratamientos $\bar{X}_{j.}$) y la denotamos por V_W. Luego

$$V_W = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 \tag{7}$$

La segunda suma del lado derecho de (5) y (6) se llama la *variación entre tratamientos* (ya que involucra a los cuadrados de las desviaciones de las diversas medias de tratamientos $\bar{X}_{j.}$ respecto de la media global $\bar{X}$) y se denota por V_B. Así pues,

$$V_B = \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 = b\sum_{j} (\bar{X}_{j} - \bar{X})^2 \tag{8}$$

Las ecuaciones (5) y (6) se pueden expresar, por tanto, como

$$V = V_W + V_B \tag{9}$$

METODOS ABREVIADOS PARA CALCULAR VARIACIONES

Para minimizar la tarea de calcular las variaciones precedentes, son convenientes las formas siguientes:

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} \tag{10}$$

$$V_B = \frac{1}{b}\sum_j T_{j\cdot}^2 - \frac{T^2}{ab} \tag{11}$$

$$V_W = V - V_B \tag{12}$$

donde T es el total de los valores X_{jk} y $T_{j\cdot}$ es el total de los valores en el tratamiento j-ésimo:

$$T = \sum_{j,k} X_{jk} \qquad T_{j\cdot} = \sum_k X_{jk} \tag{13}$$

En la práctica es conveniente restar alguna cantidad fija de todos los datos de la tabla para simplificar los cálculos; tal operación no tiene efecto alguno sobre el resultado final.

MODELOS MATEMATICOS PARA EL ANALISIS DE VARIANZA

Podemos considerar cada fila de la Tabla 16.1 como una muestra aleatoria de tamaño b de la población para un tratamiento particular. Los X_{jk} diferirán de la media poblacional μ_j para el tratamiento j-ésimo por un *error de azar* o *error aleatorio*, que denotamos por ε_{jk}; así pues

$$X_{jk} = \mu_j + \varepsilon_{jk} \tag{14}$$

Estos errores se suponen normalmente distribuidos con media 0 y varianza σ^2. Si μ es la media de la población para todos los tratamientos y hacemos $\alpha_j = \mu_j - \mu$, de manera que $\mu_j = \mu + \alpha_j$, entonces la ecuación (14) se convierte en

$$X_{jk} = \mu + \alpha_j + \varepsilon_{jk} \tag{15}$$

donde $\sum_j \alpha_j = 0$ (véase Prob. 16.9). De la ecuación (15) y de la hipótesis de que los ε_{jk} están normalmente distribuidos con media 0 y varianza σ^2, concluimos que los X_{jk} se pueden considerar como variables aleatorias normalmente distribuidas con media μ y varianza σ^2.

La hipótesis nula de que todas las medias de los tratamientos son iguales viene dada por (H_0: $\alpha_j = 0$; $j = 1, 2, \ldots, a$), o lo que es equivalente, por (H_0: $\mu_j = \mu$; $j = 1, 2, \ldots, a$). Si H_0 es verdadera, las poblaciones de los tratamientos tendrán todas la misma distribución normal (o sea,

con la misma media y varianza). En tales casos hay sólo una población de tratamiento (o sea, todos los tratamientos son estadísticamente idénticos); en otras palabras, no hay diferencia significativa entre los tratamientos.

VALORES ESPERADOS DE LAS VARIACIONES

Se puede demostrar (véase Prob. 16.10) que los valores esperados de V_W, V_B y V vienen dados por

$$E(V_W) = a(b - 1)\sigma^2 \tag{16}$$

$$E(V_B) = (a - 1)\sigma^2 + b\sum_j \alpha_j^2 \tag{17}$$

$$E(V) = (ab - 1)\sigma^2 + b\sum_j \alpha_j^2 \tag{18}$$

De la ecuación (16) se deduce que

$$E\left[\frac{V_W}{a(b - 1)}\right] = \sigma^2 \tag{19}$$

luego

$$\hat{S}_W^2 = \frac{V_W}{a(b - 1)} \tag{20}$$

es siempre una estimación óptima (no sesgada) de σ^2 independientemente de que H_0 sea verdadera o no. Por otro lado, vemos de (16) y (18) que sólo si H_0 es verdadera (o sea, $\alpha_j = 0$) tendremos

$$E\left(\frac{V_B}{a - 1}\right) = \sigma^2 \quad \text{y} \quad E\left(\frac{V}{ab - 1}\right) = \sigma^2 \tag{21}$$

así que sólo en tal circunstancia proporcionan

$$\hat{S}_B^2 = \frac{V_B}{a - 1} \quad \text{y} \quad \hat{S}^2 = \frac{V}{ab - 1} \tag{22}$$

estimaciones sin sesgo de σ^2. Si H_0 es falsa, sin embargo, tenemos de la ecuación (16) que

$$E(\hat{S}_B^2) = \sigma^2 + \frac{b}{a - 1}\sum_j \alpha_j^2 \tag{23}$$

DISTRIBUCIONES DE LAS VARIACIONES

Usando la propiedad aditiva de *ji*-cuadrado (página 272, podemos probar los siguientes teoremas fundamentales sobre las distribuciones de las variaciones V_W, V_B y V:

TEOREMA 1. V_W/σ^2 tiene distribución *ji*-cuadrado con $a(b-1)$ grados de libertad.

TEOREMA 2. Bajo la hipótesis nula H_0, V_B/σ^2 y V/σ^2 tiene distribución *ji*-cuadrado con $a-1$ y $ab-1$ grados de libertad, respectivamente.

Es importante recalcar que el Teorema 1 es válido independientemente de que se suponga H_0 o no, mientras que el Teorema 2 es válido sólo cuando se supone H_0.

EL CONTRASTE F PARA LA HIPOTESIS NULA DE IGUALDAD DE MEDIAS

Si la hipótesis nula H_0 es falsa (o sea, si las medias de los tratamientos no son iguales), vemos de (23) que cabe esperar que $\hat{S}_B^2$ sea mayor que σ^2, con el efecto tanto más pronunciado cuanto mayor sea la discrepancia entre las medias. Por otra parte, de (19) y (20) cabe esperar que $\hat{S}_W^2$ sea igual a σ^2 independientemente de que las medias sean o no iguales. Deducimos que un buen estadístico para contrastar H_0 viene dado por $\hat{S}_B^2/\hat{S}_W^2$. Si este estadístico es significativamente grande, podemos concluir que hay una diferencia significativa entre las medias de los tratamientos y podemos, por tanto, rechazar H_0; en caso contrario, podemos ya sea aceptar H_0 o reservar la decisión, pendiente de posteriores análisis adicionales.

Para usar el estadístico $\hat{S}_B^2/\hat{S}_W^2$, debemos conocer su distribución muestral. Esto lo proporciona el Teorema 3.

TEOREMA 3. El estadístico $F = \hat{S}_B^2/\hat{S}_W^2$ tiene distribución F con $a-1$ y $a(b-1)$ grados de libertad.

El Teorema 3 nos capacita para contrastar la hipótesis nula a algún nivel de significación especificado mediante un contraste unilateral con la distribución F (Cap. 11).

TABLAS DE ANALISIS DE VARIANZA

Los cálculos que requiere el contraste anterior se resumen en la Tabla 16.2, que se llama una *tabla de análisis de varianza*. En la práctica, calcularíamos V y V_B por el método largo [ecuaciones (3) y (8)] o por el método corto [ecuaciones (10) y (11)], calculando después $V_W = V - V_B$. Hagamos notar que los grados de libertad para la variación total (o sea, $ab-1$) son igual a la suma de los grados de libertad para las variaciones dentro de los tratamientos y las variaciones entre tratamientos.

Tabla 16.2

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, $V_B = b\sum_j (\bar{X}_{j\cdot} - \bar{X})^2$	$a-1$	$\hat{S}_B^2 = \dfrac{V_B}{a-1}$	$\dfrac{\hat{S}_B^2}{\hat{S}_W^2}$ con $a-1$ y $a(b-1)$ grados de libertad
Dentro de los tratamientos, $V_W = V - V_B$	$a(b-1)$	$\hat{S}_W^2 = \dfrac{V_W}{a(b-1)}$	
Total, $V = V_B + V_W = \sum_{j,k} (X_{jk} - \bar{X})^2$	$ab-1$		

MODIFICACIONES PARA NUMEROS DISTINTOS DE OBSERVACIONES

Si los tratamientos 1, ..., a tienen diferentes números de observaciones, iguales a N_1, ..., N_a, respectivamente, los resultados anteriores se modifican sin dificultad y se obtiene

$$V = \sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} X_{jk}^2 - \frac{T^2}{N} \tag{24}$$

$$V_B = \sum_{j,k} (\bar{X}_{j\cdot} - \bar{X})^2 = \sum_{j} N_j(\bar{X}_{j\cdot} - \bar{X})^2 = \sum_{j} \frac{T_{j\cdot}^2}{N_j} - \frac{T^2}{N} \tag{25}$$

$$V_W = V - V_B \tag{26}$$

donde $\sum_{j,k}$ denota la suma sobre k desde 1 hasta N_j y después la suma sobre j desde 1 hasta a. La Tabla 16.3 es la tabla del análisis de varianza para este caso.

Tabla 16.3

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, $V_B = \sum_j N_j(\bar{X}_{j\cdot} - \bar{X})^2$	$a - 1$	$\hat{S}_B^2 = \dfrac{V_B}{a-1}$	$\dfrac{\hat{S}_B^2}{\hat{S}_W^2}$ con $a - 1$ y $N - a$ grados de libertad
Dentro de los tratamientos, $V_W = V - V_B$	$N - a$	$\hat{S}_W^2 = \dfrac{V_W}{N-a}$	
Total, $V = V_B + V_W = \sum_{j,k} (X_{jk} - \bar{X})^2$	$N - 1$		

EXPERIMENTOS DE DOS FACTORES

Las ideas del análisis de varianza para un solo factor, pueden generalizarse a *experimentos de dos factores*, tal como ilustra el Ejemplo 2.

EJEMPLO 2. Supongamos que en un experimento agrario se examina la producción por acre de 4 variedades de trigo, cada una sembrada en 5 parcelas de terreno. Se necesitan en total 20 parcelas. Conviene, en tal caso, combinarlas en bloques, digamos 4 por bloque, con una variedad distinta de trigo en cada una de ellas dentro de un bloque. Eso requiere 5 bloques.

En este caso hay dos factores, ya que puede haber diferencias en la producción por acre debidas a (1) la variedad de trigo elegida y (2) el bloque particular usado (por distinta fertilidad del terreno, etc.).

Por analogía con el Ejemplo 2, nos referimos con frecuencia a los dos factores de un experimento como *tratamientos* y *bloques*, pero naturalmente podíamos llamarlos simplemente factor 1 y factor 2.

NOTACION PARA EXPERIMENTOS DE DOS FACTORES

Si hay a tratamientos y b bloques, construimos la Tabla 16.4, donde se supone que hay un valor experimental (tal como producción por acre) correspondiente a cada tratamiento y bloque. Para el tratamiento j y el bloque k, lo denotamos por X_{jk}. La media de las entradas de la fija j-ésima se denota por $\bar{X}_{j.}$, donde $j = 1, ..., a$, mientras la media de las entradas de la columna k-ésima se denota $\bar{X}_{.k}$, donde $k = 1, ..., b$. La media global se denota por $\bar{X}$. En símbolos,

$$\bar{X}_{j.} = \frac{1}{b}\sum_{k=1}^{b} X_{jk} \qquad \bar{X}_{.k} = \frac{1}{a}\sum_{j=1}^{a} X_{jk} \qquad \bar{X} = \frac{1}{ab}\sum_{j,k} X_{jk} \tag{27}$$

Tabla 16.4

	Bloque				
	1	2	...	b	
Tratamiento 1	X_{11}	X_{12}	...	X_{1b}	$\bar{X}_{1.}$
Tratamiento 2	X_{21}	X_{22}	...	X_{2b}	$\bar{X}_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
Tratamiento a	X_{a1}	X_{a2}	...	X_{ab}	$\bar{X}_{a.}$
	$\bar{X}_{.1}$	$\bar{X}_{.2}$		$\bar{X}_{.b}$	

VARIACIONES PARA EXPERIMENTOS DE DOS FACTORES

Como en el caso de experimentos de un factor, podemos definir variaciones para experimentos de dos factores. Definimos primero la *variación total*, como en la ecuación (3), a saber

$$V = \sum_{j,k} (X_{jk} - \bar{X})^2 \tag{28}$$

Escribiendo la identidad

$$X_{jk} - \bar{X} = (X_{jk} - \bar{X}_{j.} - \bar{X}_{.k} + \bar{X}) + (\bar{X}_{j.} - \bar{X}) + (\bar{X}_{.k} - \bar{X}) \tag{29}$$

elevando ahora al cuadrado y sumando sobre j y k, se ve que

$$V = V_E + V_R + V_C \tag{30}$$

donde

V_E = variación debida a error o azar $= \sum_{j,k} (X_{jk} - \bar{X}_{j.} - \bar{X}_{.k} + \bar{X})^2$

V_R = variación entre filas (tratamientos) $= b\sum_{j=1}^{a} (\bar{X}_{j.} - \bar{X})^2$

V_C = variación entre columnas (bloques) $= a\sum_{k=1}^{b} (\bar{X}_{.k} - \bar{X})^2$

La variación debida al error aleatorio se conoce como *variación residual* o *aleatoria*.

Las que siguen, análogas a las ecuaciones (10), (11) y (12), son fórmulas abreviadas para el cálculo:

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} \tag{31}$$

$$V_R = \frac{1}{b} \sum_{j=1}^{a} T_{j.}^2 - \frac{T^2}{ab} \tag{32}$$

$$V_C = \frac{1}{a} \sum_{k=1}^{b} T_{.k}^2 - \frac{T^2}{ab} \tag{33}$$

$$V_E = V - V_R - V_C \tag{34}$$

donde $T_{j.}$ es el total de las entradas en la fija j-ésima, $T_{.k}$ es el total de entradas en la columna k-ésima, y T el total de las entradas.

ANALISIS DE VARIANZA PARA EXPERIMENTOS DE DOS FACTORES

La generalización del modelo matemático para experimentos de un factor dado por (15) nos lleva a suponer para experimentos de dos factores que

$$X_{jk} = \mu + \alpha_j + \beta_k + \varepsilon_{jk} \tag{35}$$

donde $\sum \alpha_j = 0$ y $\sum \beta_k = 0$. Aquí μ es la media global de la población, α_j es la parte de X_{jk} debida a los diferentes tratamientos (llamados *efectos de los tratamientos*), β_k la parte de X_{jk} debida a los diferentes bloques (*efectos de los bloques*) y ε_{jk} es la parte debida a error o azar. Como antes, suponemos que los ε_{jk} están normalmente distribuidos con media 0 y varianza σ^2, así que los X_{jk} también están normalmente distribuidos con media μ y varianza σ^2.

Correspondientes a los resultados (16), (17) y (18), podemos probar que las esperanzas de las variaciones vienen dadas por

$$E(V_E) = (a - 1)(b - 1)\sigma^2 \tag{36}$$

$$E(V_R) = (a - 1)\sigma^2 + b \sum_j \alpha_j^2 \tag{37}$$

$$E(V_C) = (b - 1)\sigma^2 + a \sum_k \beta_k^2 \tag{38}$$

$$E(V) = (ab - 1)\sigma^2 + b \sum_j \alpha_j^2 + a \sum_k \beta_k^2 \tag{39}$$

Hay dos hipótesis nulas que querríamos contrastar:

$H_0^{(1)}$: Todos los tratamientos (fila) tienen la misma media; o sea, $\alpha_j = 0$ y $j = 1, \ldots, a$.

$H_0^{(2)}$: Todos los bloques (columna) tienen la misma media; es decir, $\beta_k = 0$ y $k = 1, \ldots, b$.

Vemos de (38) que, independientemente de $H_0^{(1)}$ o $H_0^{(2)}$, una estimación óptima (sin sesgo) de σ^2 la da

$$\hat{S}_E^2 = \frac{V_E}{(a-1)(b-1)} \quad \text{es decir,} \quad E(\hat{S}_E^2) = \sigma^2 \tag{40}$$

Además, si las hipótesis $H_0^{(1)}$ y $H_0^{(2)}$ son verdaderas, entonces

$$\hat{S}_R^2 = \frac{V_R}{a-1} \qquad \hat{S}_C^2 = \frac{V_C}{b-1} \qquad \hat{S}^2 = \frac{V}{ab-1} \tag{41}$$

serán estimaciones sin sesgo de σ^2. Si $H_0^{(1)}$ y $H_0^{(2)}$ son falsas, no obstante, de las ecuaciones (36) y (37), respectivamente, tendremos

$$E(\hat{S}_R^2) = \sigma^2 + \frac{b}{a-1}\sum_j \alpha_j^2 \tag{42}$$

$$E(\hat{S}_C^2) = \sigma^2 + \frac{a}{b-1}\sum_k \beta_k^2 \tag{43}$$

Los siguientes teoremas son similares a los Teoremas 1 y 2:

TEOREMA 4. V_E/σ^2 tiene una distribución *ji*-cuadrado con $(a-1)(b-1)$ grados de libertad, independientemente de $H_0^{(1)}$ o $H_0^{(2)}$.

TEOREMA 5. Bajo la hipótesis $H_0^{(1)}$, V_R/σ^2 tiene una distribución *ji*-cuadrado con $a-1$ grados de libertad. Bajo $H_0^{(2)}$, V_C/σ^2 tiene una distribución *ji*-cuadrado con $b-1$ grados de libertad. Bajo ambas hipótesis, $H_0^{(1)}$ y $H_0^{(2)}$, V/σ^2 tiene una distribución *ji*-cuadrado con $ab-1$ grados de libertad.

Para contrastar la hipótesis $H_0^{(1)}$, es natural considerar el estadístico $\hat{S}_R^2/\hat{S}_E^2$ ya que podemos ver de la ecuación (42) que $\hat{S}_R^2$ se espera que difiera significativamente de σ^2 si las medias de fila (tratamiento) son significativamente diferentes. Análogamente, para contrastar $H_0^{(2)}$, consideramos el estadístico $\hat{S}_C^2/\hat{S}_E^2$. Las distribuciones de $\hat{S}_R^2/\hat{S}_E^2$ y $\hat{S}_C^2/\hat{S}_E^2$ vienen dadas por el Teorema 6. que es análogo al Teorema 3.

TEOREMA 6. Bajo la hipótesis $H_0^{(1)}$, el estadístico $\hat{S}_R^2/\hat{S}_E^2$ tiene una distribución F con $a-1$ y $(a-1)(b-1)$ grados de libertad. Bajo la hipótesis $H_0^{(2)}$, el estadístico $\hat{S}_C^2/\hat{S}_E^2$ tiene una distribución F con $b-1$ y $(a-1)(b-1)$ grados de libertad.

El Teorema 6 nos capacita para aceptar o rechazar $H_0^{(1)}$ o $H_0^{(2)}$ a niveles de significación específicos. Por conveniencia, como en el caso de experimentos de un factor, se puede construir una tabla de análisis de varianza, como indica la Tabla 16.5.

EXPERIMENTOS DE DOS FACTORES CON REPETICION

En la Tabla 16.4 hay sólo una entrada correspondiente a un tratamiento y un bloque dados. Se puede obtener más información acerca de los factores repitiendo el experimento, un proceso

Tabla 16.5

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, $V_R = b\sum_j (\bar{X}_{j.} - \bar{X})^2$	$a - 1$	$\hat{S}_R^2 = \dfrac{V_R}{a - 1}$	$\hat{S}_R^2/\hat{S}_E^2$ con $a - 1$ y $(a - 1)(b - 1)$ grados de libertad
Entre bloques, $V_C = a\sum_k (\bar{X}_{.k} - \bar{X})^2$	$b - 1$	$\hat{S}_C^2 = \dfrac{V_C}{b - 1}$	$\hat{S}_C^2/\hat{S}_E^2$ con $b - 1$ y $(a - 1)(b - 1)$ grados de libertad
Residual o aleatoría, $V_E = V - V_R - V_C$	$(a - 1)(b - 1)$	$\hat{S}_E^2 = \dfrac{V_E}{(a - 1)(b - 1)}$	
Total, $V = V_R + V_C + V_E = \sum_{j,k} (X_{jk} - \bar{X})^2$	$ab - 1$		

llamado *repetición*. En tal caso habrá más de una entrada correspondiente a un tratamiento y a un bloque dados. Supondremos que hay c entradas para toda posición; cuando los números de repeticiones no son iguales han de hacerse las modificaciones pertinentes.

A causa de la repetición, se debe usar un modelo apropiado para sustituir el dado por la ecuación (35). Usaremos

$$X_{jkl} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{jkl} \tag{44}$$

donde los subíndices j, k y l de X_{jkl} corresponden a la fila j-ésima (o tratamiento), la k-ésima columna (o bloque) y la l-ésima repetición, respectivamente. En la ecuación (44) los μ, α_j y β_k se definen como antes; ε_{jkl} es un término de azar o error, mientras que los γ_{jk} denotan los *efectos de interacción* fila-columna (o sea, tratamiento-bloque), llamados a menudo *interacciones*. Tenemos las restricciones

$$\sum_j \alpha_j = 0 \qquad \sum_k \beta_k = 0 \qquad \sum_j \gamma_{jk} = 0 \qquad \sum_k \gamma_{jk} = 0 \tag{45}$$

y los X_{jkl} se suponen normalmente distribuidos con media μ y varianza σ^2.

Como antes, la variación total V de todos los datos se puede romper en variaciones debidas a filas V_R, columnas V_C, interacción V_I y error residual o aleatorio V_E:

$$V = V_R + V_C + V_I + V_E \tag{46}$$

donde

$$V = \sum_{j,k,l} (X_{jkl} - \bar{X})^2 \tag{47}$$

$$V_R = bc \sum_{j=1}^{a} (\bar{X}_{j..} - \bar{X})^2 \tag{48}$$

$$V_C = ac \sum_{k=1}^{b} (\bar{X}_{.k.} - \bar{X})^2 \tag{49}$$

$$V_I = c \sum_{j,k} (\bar{X}_{jk.} - \bar{X}_{j..} - \bar{X}_{.k.} + \bar{X})^2 \tag{50}$$

$$V_E = \sum_{j,k,l} (X_{jkl} - \bar{X}_{jk.})^2 \tag{51}$$

En estos resultados los puntos en los subíndices tienen significados análogos a los antes citados (página 375); así, por ejemplo,

$$\bar{X}_{j..} = \frac{1}{bc} \sum_{k,l} X_{jkl} = \frac{1}{b} \sum_{k} \bar{X}_{jk.} \tag{52}$$

Los valores esperados de las variaciones se hallan como antes. Usando el número apropiado de grados de libertad para cada fuente de variación, podemos establecer la tabla del análisis de varianza como indica la Tabla 16.6. Los F-cocientes en la última columna de esa tabla se pueden utilizar para contrastar las hipótesis nula:

$H_0^{(1)}$: Todas las medias de tratamiento (fila) son iguales; esto es, $\alpha_j = 0$.

$H_0^{(2)}$: Todas las medias de bloque (columna) son iguales; o sea, $\beta_k = 0$.

$H_0^{(3)}$: No hay interacciones entre tratamientos y bloques, es decir, $\gamma_{jk} = 0$.

Tabla 16.6

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, V_R	$a - 1$	$\hat{S}_R^2 = \dfrac{V_R}{a - 1}$	$\hat{S}_R^2/\hat{S}_E^2$ con $a - 1$ y $ab(c - 1)$ grados de libertad
Entre bloques, V_C	$b - 1$	$\hat{S}_C^2 = \dfrac{V_C}{b - 1}$	$\hat{S}_C^2/\hat{S}_E^2$ con $b - 1$ y $ab(c - 1)$ grados de libertad
Interacción, V_I	$(a - 1)(b - 1)$	$\hat{S}_I^2 = \dfrac{V_I}{(a - 1)(b - 1)}$	$\hat{S}_I^2/\hat{S}_E^2$ con $(a - 1)(b - 1)$ y $ab(c - 1)$ grados de libertad
Residual o aleatoria, V_E	$ab(c - 1)$	$\hat{S}_E^2 = \dfrac{V_E}{ab(c - 1)}$	
Total, V	$abc - 1$		

Desde un punto de vista práctico debemos decidir primero si $H_0^{(3)}$ puede ser rechazada o no a un nivel de significación apropiado, usando el F-cociente $\hat{S}_I^2/\hat{S}_E^2$ de la Tabla 16.6. Dos casos son posibles:

1. $H_0^{(3)}$ *no se puede rechazar*. En este caso podemos concluir que las interacciones no son demasiado grandes. Podemos entonces contrastar $H_0^{(1)}$ y $H_0^{(2)}$ usando los F-cocientes $\hat{S}_R^2/\hat{S}_E^2$ y $\hat{S}_C^2/\hat{S}_E^2$, respectivamente, como se muestra en la Tabla 16.6. Algunos estadísticos recomiendan tomar el total de $V_I + V_E$ y dividirlo por el total correspondiente de grados de libertad $(a-1)(b-1) + ab(c-1)$ y usar este valor como sustituto del denominador $\hat{S}_E^2$ en F test.
2. $H_0^{(3)}$ *puede ser rechazada*. En este caso podemos concluir que las interacciones son significativamente grandes. Diferencias en los factores serían entonces importantes sólo si fueran grandes comparadas con tales interacciones. Por esta razón muchos estadísticos recomiendan contrastar $H_0^{(1)}$ y $H_0^{(2)}$ mediante los F-cocientes $\hat{S}_R^2/\hat{S}_I^2$ y $\hat{S}_C^2/\hat{S}_I^2$ más bien que con los de la Tabla 16.6. Nosotros usaremos también aquí este procedimiento alternativo.

El análisis de varianza con repetición se realiza de forma sencilla totalizando primero los valores de repetición que corresponden a tratamientos (filas) y bloques (columnas) particulares. Esto produce una tabla de dos factores con entradas únicas, que puede analizarse como en la Tabla 16.5. Este procedimiento se ilustra en el Problema 16.16.

DISEÑO EXPERIMENTAL

Las técnicas del análisis de varianza discutidas hasta ahora se emplean una vez que se han obtenido los resultados de un experimento. Sin embargo, con el fin de adquirir cuanta información sea posible, el diseño de un experimento debe planificarse cuidadosamente; eso se conoce como el *diseño del experimento*. He aquí varios ejemplos importantes de diseño experimental:

1. **Aleatorización completa.** Supongamos que tenemos un experimento agrario como el del Ejemplo 1. Para su diseño, debemos dividir el campo en $4 \times 4 = 16$ parcelas (indicadas en la Figura 16.1 por cuadrados, aunque se puede usar cualquier forma) y asignar cada tratamiento (indicado por A, B, C y D) a cuatro bloques elegidos completamente al azar. El objetivo de la aleatorización completa es eliminar varias fuentes de error, tales como la fertilidad del suelo.

D	A	C	C
B	D	B	A
D	C	B	D
A	B	C	A

Aleatorización completa

Figura 16.1.

I	C	B	A	D
II	A	B	D	C
III	B	C	D	A
IV	A	D	C	B

Bloques aleatorizados

Figura 16.2.

D	B	C	A
B	D	A	C
C	A	D	B
A	C	B	D

Cuadrado latino

Figura 16.3.

B_γ	A_β	D_δ	C_α
A_δ	B_α	C_γ	D_β
D_α	C_δ	B_β	A_γ
C_β	D_γ	A_α	B_δ

Cuadrado greco-latino

Figura 16.4.

2. **Bloques aleatorios.** Cuando, como en el Ejemplo 2, es necesario tener un conjunto completo de tratamientos para cada bloque, los tratamientos A, B, C y D se introducen en orden aleatorio dentro de cada bloque: I, II, III y IV (o sea, las filas en la Fig. 16.2), y por esa razón se habla de los bloques como *bloques aleatorios*. Este tipo de diseño se usa cuando se desea controlar *una fuente de error o variabilidad*: a saber, la diferencia en bloques.
3. **Cuadrados latinos.** Para algunos propósitos es preciso controlar *dos fuentes de error o variabilidad* al mismo tiempo, tales como la diferencia en filas y la diferencia en columnas. Así, en el experimento del Ejemplo 1, errores en diferentes filas y columnas podrían ser debidos a cambios en la fertilidad en diferentes partes del campo. En tal caso es deseable que cada tratamiento ocurra una vez en cada fila y una vez en cada columna, como en la Figura 16.3. Esa disposición se llama un *cuadrado latino* por cuanto se usan las letras latinas A, B, C y D.
4. **Cuadrados greco-latinos.** Si es necesario controlar *tres fuentes de error o variabilidad*, se usa un *cuadrado greco-latino* como el que muestra la Figura 16.4. Tal cuadrado es esencialmente como un par de cuadrados latinos unidos, con letras unidas A, B, C y D para uno y griegas β, γ y δ para el otro. El requisito adicional que deben satisfacer es que cada letra latina ha de usarse una y sólo una vez con cada letra griega; cuando ese requisito se cumple, el cuadrado se dice *ortogonal*.

PROBLEMAS RESUELTOS

EXPERIMENTOS DE UN FACTOR

16.1. Probar que $V = V_W + V_B$; esto es

$$\sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 + \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2$$

Solución

Tenemos $$X_{jk} - \bar{X} = (X_{jk} - \bar{X}_{j.}) + (\bar{X}_{j.} - \bar{X})$$

Entonces, elevando al cuadrado y sumando en j y k, obtenemos

$$\sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 + \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 + 2\sum_{j,k} (X_{jk} - \bar{X}_{j.})(\bar{X}_{j.} - \bar{X})$$

Para probar el resultado pedido, debemos mostrar que la última suma es cero. Para ello, procedemos como sigue:

$$\sum_{j,k} (X_{jk} - \bar{X}_{j.})(\bar{X}_{j.} - \bar{X}) = \sum_{j=1}^{a} (\bar{X}_{j.} - \bar{X}) \left[\sum_{k=1}^{b} (X_{jk} - \bar{X}_{j.}) \right]$$

$$= \sum_{j=1}^{a} = (\bar{X}_{j.} - \bar{X}) \left[\left(\sum_{k=1}^{b} X_{jk} \right) - b\bar{X}_{j.} \right] = 0$$

ya que $$\bar{X}_{j.} = \frac{1}{b} \sum_{k=1}^{b} X_{jk}$$

16.2. Comprobar que (*a*) $T = ab\bar{X}$, (*b*) $T_{j.} = b\bar{X}_{j.}$ y (*c*) $\sum_j T_{j.} = ab\bar{X}$, usando la notación de la página 376.

Solución

(*a*)
$$T = \sum_{j,k} X_{jk} = ab\left(\frac{1}{b}\sum_{j,k} X_{jk}\right) = ab\bar{X}$$

(*b*)
$$T_{j.} = \sum_{k} X_{jk} = b\left(\frac{1}{b}\sum_{k} X_{jk}\right) = b\bar{X}_{j.}$$

(*c*) Como $T_{j.} = \sum_k X_{jk}$, por la parte (*a*) se tiene

$$\sum_j T_{j.} = \sum_j \sum_k X_{jk} = T = ab\bar{X}$$

16.3. Verificar las fórmulas (10), (11) y (12) de este capítulo.

Solución

Tenemos

$$\begin{aligned} V &= \sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk}^2 - 2\bar{X}X_{jk} + \bar{X}^2) \\ &= \sum_{j,k} X_{jk}^2 - 2\bar{X}\sum_{j,k} X_{jk} + ab\bar{X}^2 \\ &= \sum_{j,k} X_{jk}^2 - 2\bar{X}(ab\bar{X}) + ab\bar{X}^2 \\ &= \sum_{j,k} X_{jk}^2 - ab\bar{X}^2 \\ &= \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} \end{aligned}$$

usando el Problema 16.2(*a*) en la tercera y en la última línea. De igual modo,

$$\begin{aligned} V_B &= \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 = \sum_{j,k} (\bar{X}_{j.}^2 - 2\bar{X}\bar{X}_{j.} + \bar{X}^2) \\ &= \sum_{j,k} \bar{X}_{j.}^2 - 2\bar{X}\sum_{j,k} \bar{X}_{j.} + ab\bar{X}^2 \\ &= \sum_{j,k} \left(\frac{T_{j.}}{b}\right)^2 - 2\bar{X}\sum_{j,k} \frac{T_{j.}}{b} + ab\bar{X}^2 \\ &= \frac{1}{b^2}\sum_{j=1}^{a}\sum_{k=1}^{b} T_{j.}^2 - 2\bar{X}(ab\bar{X}) + ab\bar{X}^2 \\ &= \frac{1}{b}\sum_{j=1}^{a} T_{j.}^2 - ab\bar{X}^2 \\ &= \frac{1}{b}\sum_{j=1}^{a} T_{j.}^2 - \frac{T^2}{ab} \end{aligned}$$

usando el Problema 16.2(*b*) en la tercera línea y el Problema 16.2(*a*) en la última. Finalmente, la ecuación (12) se sigue de que $V = V_W + V_B$, o sea $V_W = V - V_B$.

16.4. La Tabla 16.7 da las producciones por acre de una cierta variedad de trigo que crece en terrenos tratados con fertilizantes A, B y C. Hallar (*a*) las producciones medias para los diferentes tratamientos, (*b*) la media global para todos los tratamientos, (*c*) la variación total, (*d*) la variación entre tratamientos y (*e*) la variación dentro de los tratamientos. Usar el método largo.

Tabla 16.7

A	48	49	50	49
B	47	49	48	48
C	49	51	50	50

Tabla 16.8

3	4	5	4
2	4	3	3
4	6	5	5

Solución

Para simplificar la aritmética, podemos restar 45 a todos los datos sin que ello afecte a los valores de las variaciones. Entonces obtenemos los datos de la Tabla 16.8.

(*a*) Las medias de tratamiento (fila) para la Tabla 16.8 vienen dadas por

$$\bar{X}_{1.} = \tfrac{1}{4}(3+4+5+4) = 4 \qquad \bar{X}_{2.} = \tfrac{1}{4}(2+4+3+3) = 3 \qquad \bar{X}_{3.} = \tfrac{1}{4}(4+6+5+5) = 5$$

Luego las producciones medias, obtenidas añadiendo 45 a éstas, son de 49, 48 y 50 bushels por acre para A, B y C, respectivamente.

(*b*) La media global para todos los tratamientos es

$$\bar{X} = \tfrac{1}{12}(3+4+5+4+2+4+3+3+4+6+5+5) = 4$$

Así que la media global para los datos originales es 45 + 4 = 49 bushels por acre.

(*c*) La variación es

$$V = \sum_{j,k} (X_{jk} - X)^2 = (3-4)^2 + (4-4)^2 + (5-4)^2 + (4-4)^2 + (2-4)^2 + (4-4)^2 +$$
$$+ (3-4)^2 + (3-4)^2 + (4-4)^2 + (6-4)^2 + (5-4)^2 + (5-4)^2 = 14$$

(*d*) La variación entre tratamientos es

$$V_B = b \sum_{j} (\bar{X}_{j.} - \bar{X})^2 = 4[(4-4)^2 + (3-4)^2 + (5-4)^2] = 8$$

(*e*) La variación dentro de los tratamientos es

$$V_W = V - V_B = 14 - 8 = 6$$

Otro método

$$V_W = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 = (3-4)^2 + (4-4)^2 + (5-4)^2 + (4-4)^2 + (2-3)^2 + (4-3)^2 +$$
$$+ (3-3)^2 + (3-3)^2 + (4-5)^2 + (6-5)^2 + (5-5)^2 + (5-5)^2 = 6$$

Nota: La Tabla 16.9 es la tabla de análisis de varianza para los Problemas 16.4, 16.5 y 16.6.

Tabla 16.9

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, $V_B = 8$	$a - 1 = 2$	$\hat{S}_B^2 = \frac{8}{2} = 4$	$\frac{\hat{S}_B^2}{\hat{S}_W^2} = \frac{4}{2/3} = 6$ con 2 y 9 grados de libertad
Dentro de los tratamientos, $V_W = V - V_B = 14 - 8 = 6$	$a(b-1) = (3)(3) = 9$	$\hat{S}_W^2 = \frac{6}{9} = \frac{2}{3}$	
Total, $V = 14$	$ab - 1 = (3)(4) - 1 = 11$		

16.5. Con referencia al Problema 16.4, hallar una estimación sin sesgo de la varianza de la población σ^2 de (*a*) la variación entre tratamientos bajo la hipótesis nula de medias de tratamiento iguales y (*b*) la variación entre tratamientos.

Solución

(*a*) $$\hat{S}_B^2 = \frac{V_B}{a-1} = \frac{8}{3-1} = 4$$

(*b*) $$\hat{S}_W^2 = \frac{V_W}{a(b-1)} = \frac{6}{3(4-1)} = \frac{2}{3}$$

16.6. En el Problema 16.4, ¿podemos rechazar la hipótesis nula de medias iguales al nivel de significación (*a*) 0.05 y (*b*) 0.01?

Solución

Se tiene $$F = \frac{\hat{S}_B^2}{\hat{S}_W^2} = \frac{4}{2/3} = 6$$

con $a - 1 = 3 - 1 = 2$ grados de libertad y $a(b-1) = 3(4-1) = 9$ grados de libertad.

(*a*) En el Apéndice V, con $\nu_1 = 2$ y $\nu_2 = 9$, vemos que $F_{.95} = 4.26$. Como $F = 6 > F_{.95}$, podemos rechazar la hipótesis nula de medias iguales al nivel 0.05.

(*b*) En el Apéndice VI, con $\nu_1 = 2$ y $\nu_2 = 9$, vemos que $F_{.99} = 8.02$. Puesto que $F = 6 < F_{.99}$, no podemos rechazar la hipótesis nula de medias iguales al nivel 0.01.

16.7. Usar las fórmulas abreviadas (10), (11) y (12) para llegar a los resultados del Problema 16.4.

Solución

Conviene disponer los datos como en la Tabla 16.10.

Tabla 16.10

		$T_{j.}$	$T_{j.}^2$
A	3 4 5 4	16	256
B	2 4 3 3	12	144
C	4 6 5 5	20	400
	$\sum_{j,k} X_{jk}^2 = 206$	$T = \sum_j T_{j.} = 48$	$\sum_j T_{j.}^2 = 800$

(*a*) Usando la fórmula (10), vemos que

$$\sum_{j,k} X_{jk}^2 = 9 + 16 + 25 + 16 + 4 + 16 + 9 + 9 + 16 + 36 + 25 + 25 = 206$$

y

$$T = 3 + 4 + 5 + 4 + 2 + 4 + 3 + 3 + 4 + 6 + 5 + 5 = 48$$

Luego

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} = 206 - \frac{(48)^2}{(3)(4)} = 206 - 192 = 14$$

(*b*) Los totales de las filas son

$$T_{1.} = 3+4+5+4 = 16 \qquad T_{2.} = 2+4+3+3 = 12 \qquad T_{3.} = 4+6+5+5 = 20$$

y

$$T = 16 + 12 + 20 = 48$$

Así que, por la fórmula (11), se deduce

$$V_B = \frac{1}{b}\sum_j T_{j.}^2 - \frac{T^2}{ab} = \frac{1}{4}(16^2 + 12^2 + 20^2) - \frac{(48)^2}{(3)(4)} = 200 - 192 = 8$$

(*c*) Mediante la fórmula (12), se obtiene

$$V_W = V - V_B = 14 - 8 = 6$$

Los resultados coinciden con los obtenidos en el Problema 16.4, y desde este punto en adelante el análisis es como antes.

16.8. Una empresa quiere comprar una de entre cinco máquinas diferentes: A, B, C, D o E. En un experimento diseñado para comprobar si hay diferencia entre ellas, cada máquina fue manejada por un operario experto distinto en cada una, durante tiempos iguales. La Tabla 16.11 muestra los números de unidades producidas por las máquinas. Contrastar la hipótesis de que no hay diferencia entre las máquinas al nivel de significación (*a*) 0.05 y (*b*) 0.01.

Solución

Restar un número adecuado, 60 por ejemplo, a todos los datos de la Tabla 16.12. Entonces

$$V = 2658 - \frac{(54)^2}{(5)(4)} = 2658 - 145.8 = 2512.2$$

y

$$V_B = \frac{1}{5}(3874) - \frac{(54)^2}{(5)(4)} = 774.8 - 145.8 = 629.0$$

Ahora formamos la Tabla 16.13. Para 4 y 20 grados de libertad tenemos $F_{.95} = 2.87$. Luego no podemos rechazar la hipótesis nula al nivel 0.05 y por tanto con menos motivo al 0.01.

Tabla 16.11

A	68	72	77	42	53
B	72	53	63	53	48
C	60	82	64	75	72
D	48	61	57	64	50
E	64	65	70	68	53

Tabla 16.12

						$T_{j.}$	$T_{j.}^2$
A	8	12	17	−18	−7	12	144
B	12	−7	3	−7	−12	−11	121
C	0	22	4	15	12	53	2809
D	−12	1	−3	4	−10	−20	400
E	4	5	10	8	−7	20	400
	$\sum X_{jk}^2 = 2658$					54	3874

Tabla 16.13

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, $V_B = 629.0$	$a-1=4$	$\hat{S}_B^2 = \frac{629.0}{4} = 157.25$	$\frac{\hat{S}_B^2}{\hat{S}_W^2} = 1.67$
Dentro de los tratamientos, $V_W = 1883.8$	$a(b-1)=(5)(4)=20$	$\hat{S}_W^2 = \frac{1883.2}{(5)(4)} = 94.16$	
Total, $V = 2512.2$	$ab-1=24$		

MODIFICACIONES PARA NUMEROS DISTINTOS DE OBSERVACIONES

16.9. La Tabla 16.4 da las vidas medias, en horas, de muestras de tres tipos distintos de tubos de televisión producidos por cierta empresa. Usando el método largo, determinar si hay diferencia entre ellos al nivel de significación (*a*) 0.05 y (*b*) 0.01.

Tabla 16.14

Muestra 1	407	411	409		
Muestra 2	404	406	408	405	402
Muestra 3	410	408	406	408	

Solución

Conviene restar a los datos un número apropiado, digamos 400, con lo que se obtiene la Tabla 16.15. Esta muestra los totales de fila, las medias muestrales (o de grupo) y la media global. Así pues, se tiene

$$V = \sum_{j,k} (X_{jk} - \bar{X})^2 = (7 - 7)^2 + (11 - 7)^2 + \cdots + (8 - 7)^2 = 72$$

$$V_B = \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 = \sum_{j} N_j(\bar{X}_{j.} - \bar{X})^2 = 3(9 - 7)^2 + 5(7 - 5)^2 + 4(8 - 7)^2 = 36$$

$$V_W = V - V_B = 72 - 36 = 36$$

Tabla 16.15

						Total	Media
Muestra 1	7	11	9			27	9
Muestra 2	4	6	8	5	2	25	5
Muestra 3	10	8	6	8		32	8
	$\bar{X}$ = media final = $\frac{84}{12} = 7$						

Podemos también obtener V_W directamente observando que es igual a

$$(7 - 9)^2 + (11 - 9)^2 + (9 - 9)^2 + (4 - 5)^2 + (6 - 5)^2 + (8 - 5)^2 + (5 - 5)^2 +$$
$$+ (2 - 5)^2 + (10 - 8)^2 + (8 - 8)^2 + (6 - 8)^2 + (8 - 8)^2$$

Los datos se resumen en la Tabla 16.16, la tabla del análisis de varianza. Para 2 y 9 grados de libertad, vemos en el Apéndice V que $F_{.95} = 4.26$ y en el Apéndice VI vemos que $F_{.99} = 8.02$. Luego podemos rechazar la hipótesis de medias iguales (o sea, no hay diferencia entre los tres tipos de tubos) al nivel 0.05, pero no al 0.01.

Tabla 16.16

Variación	Grados de libertad	Cuadrado medio	F
$V_B = 36$	$a - 1 = 2$	$\hat{S}_B^2 = \frac{36}{2} = 18$	$\frac{\hat{S}_B^2}{\hat{S}_W^2} = \frac{18}{4}$
$V_W = 36$	$N - a = 9$	$\hat{S}_W^2 = \frac{36}{9} = 4$	$= 4.5$

16.10. Resolver el Problema 16.9 usando las fórmulas abreviadas incluidas en las ecuaciones (24), (25) y (26).

Solución

De la Tabla 16.15 se sigue $N_1 = 3$, $N_2 = 5$, $N_3 = 4$, $N = 12$, $T_1 = 27$, $T_2 = 25$, $T_3 = 32$ y $T = 84$. En consecuencia,

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{N} = 7^2 + 11^2 + \cdots + 6^2 + 8^2 - \frac{(84)^2}{12} = 72$$

$$V_B = \sum_j \frac{T_{j.}^2}{N_j} - \frac{T^2}{N} = \frac{(27)^2}{3} + \frac{(25)^2}{5} + \frac{(32)^2}{4} - \frac{(84)^2}{12} = 36$$

$$V_W = V - V_B = 36$$

Usando esto, el análisis de varianza se hace ya como en el Problema 16.9.

EXPERIMENTOS DE DOS FACTORES

16.11. La Tabla 16.17 muestra las producciones por acre de cuatro semillas sembradas en campos tratados con tres fertilizantes distintos. Por el método largo, determinar el nivel de significación 0.01 si hay diferencia en producción por acre (*a*) debida a los fertilizantes y (*b*) debida a las semillas.

Tabla 16.17

	Semilla I	Semilla II	Semilla III	Semilla IV
Fertilizante *A*	4.5	6.4	7.2	6.7
Fertilizante *B*	8.8	7.8	9.6	7.0
Fertilizante *C*	5.9	6.8	5.7	5.2

Solución

Calcular los totales de fila, de columna, las medias de columna, el total global y la media global, como indica la Tabla 16.18. De esa tabla se obtiene:

Tabla 16.18

	Cosecha I	Cosecha II	Cosecha III	Cosecha IV	Total de fila	Media de fila
Fertilizante *A*	4.5	6.4	7.2	6.7	24.8	6.2
Fertilizante *B*	8.8	7.8	9.6	7.0	33.2	8.3
Fertilizante *C*	5.9	6.8	5.7	5.2	23.6	5.9
Total de columna	19.2	21.0	22.5	18.9	Total final = 81.6	
Media de columna	6.4	7.0	7.5	6.3	Media final = 6.8	

La variación de las medias de fila respecto de la media global es

$$V_R = 4[(6.2 - 6.8)^2 + (8.3 - 6.8)^2 + (5.9 - 6.8)^2] = 13.68$$

La variación de las medias de columna respecto de la media global es

$$V_C = 3[(6.4 - 6.8)^2 + (7.0 - 6.8)^2 + (7.5 - 6.8)^2 + (6.3 - 6.8)^2] = 2.82$$

La variación total es

$$\begin{aligned} V = {} & (4.5 - 6.8)^2 + (6.4 - 6.8)^2 + (7.2 - 6.8)^2 + (6.7 - 6.8)^2 + \\ & + (8.8 - 6.8)^2 + (7.8 - 6.8)^2 + (9.6 - 6.8)^2 + (7.0 - 6.8)^2 + \\ & + (5.9 - 6.8)^2 + (6.8 - 6.8)^2 + (5.7 - 6.8)^2 + (5.2 - 6.8)^2 = 23.08 \end{aligned}$$

La variación aleatoria es

$$V_E = V - V_R - V_C = 6.58$$

Eso conduce al análisis de varianza de la Tabla 16.19.

Al nivel de significación 0.05 con 2 y 6 grados de libertad, $F_{.95} = 5.14$. Por tanto, desde $6.24 > 5.14$, podemos rechazar la hipótesis de que las medias de fila son iguales y concluir que hay diferencia significativa en producción debida a los fertilizantes.

Como el valor F correspondiente a la diferencia en medias de columna es menor que 1, concluimos que no hay diferencia significativa debida a las semillas en la producción

Tabla 16.19

Variación	Grados de libertad	Cuadrado medio	F
$V_R = 13.68$	2	$\hat{S}_R^2 = 6.84$	$\hat{S}_R^2/\hat{S}_E^2 = 6.24$ con 2 y 6 grados de libertad
$V_C = 2.82$	3	$\hat{S}_C^2 = 0.94$	$\hat{S}_C^2/\hat{S}_E^2 = 0.86$ con 3 y 6 grados de libertad
$V_E = 6.58$	6	$\hat{S}_E^2 = 1.097$	
$V = 23.08$	11		

16.12. Usar las fórmulas abreviadas para llegar a los resultados del Problema 16.11.

Solución

De la Tabla 16.18 tenemos

$$\sum_{j,k} X_{jk}^2 = (4.5)^2 + (6.4)^2 + \cdots + (5.2)^2 = 577.96$$

$$T = 24.8 + 33.2 + 23.6 = 81.6$$

$$\sum T_{j.}^2 = (24.8)^2 + (33.2)^2 + (23.6)^2 = 2274.24$$

$$\sum T_{.k}^2 = (19.2)^2 + (21.0)^2 + (22.5)^2 + (18.9)^2 = 1673.10$$

Entonces

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} = 577.96 - 554.88 = 23.08$$

$$V_R = \frac{1}{b}\sum T_{j.}^2 - \frac{T^2}{ab} = \frac{1}{4}(2274.24) - 554.88 = 13.68$$

$$V_C = \frac{1}{a}\sum T_{.k}^2 - \frac{T^2}{ab} = \frac{1}{3}(1673.10) - 554.88 = 2.82$$

$$V_E = V - V_R - V_C = 23.08 - 13.68 - 2.82 = 6.58$$

de acuerdo con el Problema 16.11.

EXPERIMENTOS DE DOS FACTORES CON REPETICION

16.13. Un empresario desea determinar la eficacia de cuatro tipos distintos de máquinas (A, B, C y D) en la producción de tornillos. Para ello, anota el número de tornillos defectuosos cada día de una semana en dos turnos de trabajo, con los resultados que recoge la Tabla 16.20. Hacer un análisis de varianza para determinar al nivel de significación 0.05 si hay diferencia (a) entre las máquinas y (b) entre los turnos.

Solución

Los datos se organizan de modo equivalente en la Tabla 16.21, en la que los dos factores, máquinas y turnos, quedan indicados. Hay dos turnos para cada máquina. Los días de la semana pueden considerarse como repeticiones del trabajo de cada máquina. La variación total para todos los datos de la Tabla 16.21 es

$$V = 6^2 + 4^2 + 5^2 + \cdots + 7^2 + 10^2 - \frac{(268)^2}{40} = 1946 - 1795.6 = 150.4$$

Tabla 16.20

Máquina	Primer turno					Segundo turno				
	L.	Mar.	Miér.	J.	V.	L.	Mar.	Miér.	J.	V.
A	6	4	5	5	4	5	7	4	6	8
B	10	8	7	7	9	7	9	12	8	8
C	7	5	6	5	9	9	7	5	4	6
D	8	4	6	5	5	5	7	9	7	10

Tabla 16.21

Factor I: Máquina	Factor II: Ensayo	Réplicas					Total
		L.	Mar.	Miér.	J.	V.	
A	1	6	4	5	5	4	24
	2	5	7	4	6	8	30
B	1	10	8	7	7	9	41
	2	7	9	12	8	8	44
C	1	7	5	6	5	9	32
	2	9	7	5	4	6	31
D	1	8	4	6	5	5	28
	2	5	7	9	7	10	38
	Total	57	51	54	47	59	268

Con el fin de considerar los dos factores, limitamos nuestra atención al total de valores de repetición correspondientes a cada combinación de factores. Recogidos en la Tabla 16.22 hacen de ésta una tabla de dos factores con entrada única. La variación total para la Tabla 16.22, que llamaremos *variación subtotal* V_S, viene dada por

$$V_S = \frac{(24)^2}{5} + \frac{(41)^2}{5} + \frac{(32)^2}{5} + \frac{(28)^2}{5} + \frac{(30)^2}{5} + \frac{(44)^2}{5} + \frac{(31)^2}{5} + \frac{(38)^2}{5} - \frac{(268)^2}{40} =$$

$$= 1861.2 - 1795.6 = 65.6$$

La variación entre filas es

$$V_R = \frac{(54)^2}{10} + \frac{(85)^2}{10} + \frac{(63)^2}{10} + \frac{(66)^2}{10} - \frac{(268)^2}{40} = 1846.6 - 1795.6 = 51.0$$

La variación entre columnas viene dada por

$$V_C = \frac{(125)^2}{20} + \frac{(143)^2}{20} - \frac{(268)^2}{40} = 1803.7 - 1795.6 = 8.1$$

Tabla 16.22

Máquina	Primer ensayo	Segundo ensayo	Total
A	24	30	54
B	41	44	85
C	32	31	63
D	28	38	66
Total	125	143	268

Tabla 16.21

Factor I: Máquina	Factor II: Ensayo	Réplicas L.	Mar.	Miér.	J.	V.	Total
A	1	6	4	5	5	4	24
	2	5	7	4	6	8	30
B	1	10	8	7	7	9	41
	2	7	9	12	8	8	44
C	1	7	5	6	5	9	32
	2	9	7	5	4	6	31
D	1	8	4	6	5	5	28
	2	5	7	9	7	10	38
	Total	57	51	54	47	59	268

Con el fin de considerar los dos factores, limitamos nuestra atención al total de valores de repetición correspondientes a cada combinación de factores. Recogidos en la Tabla 16.22 hacen de ésta una tabla de dos factores con entrada única. La variación total para la Tabla 16.22, que llamaremos *variación subtotal* V_S, viene dada por

$$V_S = \frac{(24)^2}{5} + \frac{(41)^2}{5} + \frac{(32)^2}{5} + \frac{(28)^2}{5} + \frac{(30)^2}{5} + \frac{(44)^2}{5} + \frac{(31)^2}{5} + \frac{(38)^2}{5} - \frac{(268)^2}{40} =$$

$$= 1861.2 - 1795.6 = 65.6$$

La variación entre filas es

$$V_R = \frac{(54)^2}{10} + \frac{(85)^2}{10} + \frac{(63)^2}{10} + \frac{(66)^2}{10} - \frac{(268)^2}{40} = 1846.6 - 1795.6 = 51.0$$

La variación entre columnas viene dada por

$$V_C = \frac{(125)^2}{20} + \frac{(143)^2}{20} - \frac{(268)^2}{40} = 1803.7 - 1795.6 = 8.1$$

Tabla 16.22

Máquina	Primer ensayo	Segundo ensayo	Total
A	24	30	54
B	41	44	85
C	32	31	63
D	28	38	66
Total	125	143	268

Si restamos ahora de V_S la suma de las variaciones entre filas y columnas ($V_R + V_C$), obtenemos la variación debida a la *interacción* entre filas y columnas, que está dada por

$$V_I = V_S - V_R - V_C = 65.6 - 51.0 - 8.1 = 6.5$$

Finalmente, la variación residual, que se puede ver como la variación de error o azar V_E (supuesto que creemos que los diversos días de la semana no producen diferencias relevantes), se halla restando la variación subtotal (o sea, la suma de las variaciones de fila, columna e interacción) de la variación total V. Eso da

$$V_E = V - (V_R + V_C + V_I) = V - V_S = 150.4 - 65.6 = 84.8$$

Estas variaciones se recogen en la Tabla 16.23, el análisis de varianza. La tabla da también el número de grados de libertad correspondiente a cada tipo de variación. Así pues, como hay cuatro filas en la Tabla 16.22, la variación debida a filas tiene $4 - 1 = 3$ grados de libertad, mientras que la variación debida a las dos columnas tiene $2 - 1 = 1$ grados de libertad. Para hallar los grados de libertad debidos a la interacción, notemos que hay ocho entradas en la Tabla 16.22; luego los grados de libertad totales son $8 - 1 = 7$. Puesto que 3 de ellos se deben a las filas y 1 a las columnas, los restantes $[7 - (3 + 1) = 3]$ se deben a la interacción. Puesto que hay 40 entradas en la tabla original 16.21, el total de grados de libertad es $40 - 1 = 39$. De modo que los grados de libertad debidos a la variación residual o de azar son $39 - 7 = 32$.

Tabla 16.23

Variación	Grados de libertad	Cuadrado medio	F
Filas (máquinas), $V_R = 51.0$	3	$\hat{S}_R^2 = 17.0$	$\frac{17.0}{2.65} = 6.42$
Columnas (turnos), $V_C = 8.1$	1	$\hat{S}_C^2 = 8.1$	$\frac{8.1}{2.65} = 3.06$
Interacción, $V_I = 6.5$	3	$\hat{S}_I^2 = 2.167$	$\frac{2.167}{2.65} = 0.817$
Subtotal, $V_S = 65.6$	7		
Aleatorio o residual, $V_E = 84.8$	32	$\hat{S}_E^2 = 2.65$	
Total, $V = 150.4$	39		

Para continuar, hemos de determinar primero si hay interacción significativa entre los factores básicos (o sea, las filas y columnas de la Tabla 16.22). De la Tabla 16.23 vemos que para la interacción es $F = 0.817$, lo cual nos dice que la interacción no es significativa; esto es, no podemos rechazar la hipótesis $H_0^{(3)}$ de la página 385. Siguiendo las reglas de la misma página, vemos que la F calculada para filas es 6.42. Como $F_{.95} = 2.90$ para 3 y 32 grados de libertad, podemos rechazar la hipótesis $H_0^{(1)}$

de que las filas tienen medias iguales. Ello equivale a decir que al nivel 0.05 podemos concluir que las máquinas no son igualmente eficaces.

Para 1 y 32 grados de libertad, $F_{.95} = 4.15$. Entonces, ya que la F calculada para columnas es 3.06, no podemos rechazar la $H_0^{(2)}$ de que las columnas tienen medias iguales. Lo que equivale a decir que al nivel 0.05 no hay diferencia significativa entre los turnos.

Si podemos optar por analizar los resultados uniendo las variaciones residual y de interacción, como propugnan algunos estadísticos, encontramos que $V_I + V_E = 6.5 + 84.8 = 91.3$ para la variación conjunta y $V_I + V_E = 3 + 32 = 35$ para los grados de libertad conjuntos, que nos da una varianza conjunta de $91.3/35 = 2.61$. Usar este valor en lugar de 2.65 para el denominador de F en la Tabla 16.23 no afecta a las conclusiones antes alcanzadas.

16.14. Rehacer el Problema 16.13 al nivel de significación 0.01.

Solución

A este nivel no hay todavía interacción apreciable, así que podemos continuar.

Como $F_{.99} = 4.47$ para 3 y 32 grados de libertad y el F calculado para filas es 6.42, podemos concluir que incluso al nivel 0.01 las máquinas no son igualmente efectivas.

Como $F_{.99} = 7.51$ para 1 y 32 grados de libertad y la F para columnas es 3.06, podemos concluir que al nivel de significación 0.01 no hay diferencia significativa entre turnos.

CUADRADOS LATINOS

16.15. Un labrador quiere contrastar los efectos de cuatro fertilizantes (A, B, C y D) en la producción de trigo. Con el fin de eliminar fuentes de error debidas a la variabilidad en la fertilidad del suelo, los utiliza en una disposición de cuadrado latino, tal como indica la Tabla 16.24, donde los números están en bushels por unidad de área. Hacer un análisis de varianza para determinar si hay diferencia entre los fertilizantes al nivel de significación (*a*) 0.05 y (*b*) 0.01.

Solución

Primero obtenemos totales de filas y columnas (véase Tabla 16.25). También obtenemos las producciones totales de cada uno de los fertilizantes (véase Tabla 16.26). La variación total y las variaciones para filas, columnas y tratamientos se deducen de ahí del modo usual. Encontramos:

La variación total es

$$V = (18)^2 + (21)^2 + (25)^5 + \cdots + (10)^2 + (17)^2 - \frac{(295)^2}{16} =$$

$$= 5769 - 5439.06 = 329.94$$

Tabla 16.24

A 18	C 21	D 25	B 11
D 22	B 12	A 15	C 19
B 15	A 20	C 23	D 24
C 22	D 21	B 10	A 17

Tabla 16.25

					Total
	A 18	C 21	D 25	B 11	75
	D 22	B 12	A 15	C 19	68
	B 15	A 20	C 23	D 24	82
	C 22	D 21	B 10	A 17	70
Total	77	74	73	71	295

Tabla 16.26

	A	*B*	*C*	*D*	
Total	70	48	85	92	295

La variación entre filas es

$$V_R = \frac{(75)^2}{4} + \frac{(68)^2}{4} + \frac{(82)^2}{4} + \frac{(70)^2}{4} - \frac{(295)^2}{16} =$$

$$= 5468.25 - 5439.06 = 29.19$$

La variación entre columnas es

$$V_C = \frac{(77)^2}{4} + \frac{(74)^2}{4} + \frac{(73)^2}{4} + \frac{(71)^2}{4} - \frac{(295)^2}{16} =$$

$$= 5443.75 - 5439.06 = 4.69$$

La variación entre tratamientos es

$$V_B = \frac{(70)^2}{4} + \frac{(48)^2}{4} + \frac{(85)^2}{4} + \frac{(92)^2}{4} - \frac{(295)^2}{16} =$$

$$= 5723.25 - 5439.06 = 284.19$$

La Tabla 16.27 muestra el análisis de la varianza.

Tabla 16.27

Variación	Grados de libertad	Cuadrado medio	*F*
Filas, 29.19	3	9.73	4.92
Columnas, 4.69	3	1.563	0.79
Tratamientos, 284.19	3	94.73	47.9
Residuales, 11.87	6	1.978	
Total, 329.94	15		

(*a*) Como $F_{.95,\,3,\,6} = 4.76$, podemos rechazar al nivel 0.05 la hipótesis de medias de fila iguales. Se sigue que al nivel 0.05 hay diferencia en fertilidad del terreno de una fila a otra.

Como el valor F para columnas es menor que 1, no hay diferencia en fertilidad en las columnas.

Ya que el valor F para tratamientos es $47.9 > 4.76$, concluimos que hay diferencia entre los fertilizantes.

(*b*) Puesto que $F_{99,3,6} = 9.78$, podemos aceptar la hipótesis de que no hay diferencia en fertilidad en las filas (o en las columnas) al nivel 0.01. Sin embargo, debemos concluir todavía que hay diferencia entre los fertilizantes al nivel 0.01.

CUADRADOS GRECO-LATINOS

16.16. Interesa saber si hay diferencia en millas recorridas por galón entre las gasolinas *A*, *B*, *C* y *D*. Diseñar un experimento con cuatro conductores distintos, cuatro coches distintos y cuatro carreteras distintas.

Solución

Como se usa el mismo número de cada uno de los factores, podemos recurrir a un cuadrado greco-latino. Supongamos que los diferentes coches se representan por filas y los diferentes conductores por columnas, como en la Tabla 16.28. Ahora asignamos las diferentes gasolinas (*A*, *B*, *C* y *D*) a las filas y columnas al azar, con el único requisito de que cada letra aparezca una vez en cada fila y en cada columna. Así pues, cada conductor conducirá una vez cada coche y usará una vez cada gasolina, y ningún coche será conducido dos veces con la misma gasolina.

Ahora asignamos al azar las cuatro carreteras, denotadas por α, β, γ y δ, con el mismo requisito impuesto sobre los cuadrados latinos. Así que cada conductor tendrá oportunidad de conducir por cada una de ellas. La Tabla 16.28 muestra una de las posibles disposiciones.

Tabla 16.28

	Conductor			
	1	2	3	4
Coche 1	B_γ	A_β	D_δ	C_α
Coche 2	A_δ	B_α	C_γ	D_β
Coche 3	D_α	C_δ	B_β	A_γ
Coche 4	C_β	D_γ	A_α	B_δ

16.17. Supongamos que al realizar el experimento del Problema 16.16, el número de millas por galón resulta ser el que indica la Tabla 16.29. Determinar por análisis de varianza si hay diferencias al nivel de significación 0.05.

Tabla 16.29

	Conductor			
	1	2	3	4
Coche 1	B_γ 19	A_β 16	D_δ 16	C_α 14
Coche 2	A_δ 15	B_α 18	C_γ 11	D_β 15
Coche 3	D_α 14	C_δ 11	B_β 21	A_γ 16
Coche 4	C_β 16	D_γ 16	A_α 15	B_δ 23

Solución

Primero obtenemos los totales de filas y columnas (véase Tabla 16.30) y a continuación los totales para cada letra latina y para cada letra griega, como sigue:

$$\begin{aligned}
A \text{ total:} &\quad 15 + 16 + 15 + 16 = 62 \\
B \text{ total:} &\quad 19 + 18 + 21 + 23 = 81 \\
C \text{ total:} &\quad 16 + 11 + 11 + 14 = 52 \\
D \text{ total:} &\quad 14 + 16 + 16 + 15 = 61 \\
\alpha \text{ total:} &\quad 14 + 18 + 15 + 14 = 61 \\
\beta \text{ total:} &\quad 16 + 16 + 21 + 15 = 68 \\
\gamma \text{ total:} &\quad 19 + 16 + 11 + 16 = 62 \\
\delta \text{ total:} &\quad 15 + 11 + 16 + 23 = 65
\end{aligned}$$

Tabla 16.30

					Total
	B_γ 19	A_β 16	D_δ 16	C_α 14	65
	A_δ 15	B_α 18	C_γ 11	D_β 15	59
	D_α 14	C_δ 11	B_β 21	A_γ 16	62
	C_β 16	D_γ 16	A_α 15	B_δ 23	70
Total	64	61	63	68	256

Ahora calculamos las variaciones correspondientes a todas éstas, mediante el método abreviado:

$$\text{Filas:} \quad \frac{(65)^2}{4} + \frac{(59)^2}{4} + \frac{(62)^2}{4} + \frac{(70)^2}{4} - \frac{(256)^2}{16} = 4112.50 - 4096 = 16.50$$

$$\text{Columnas:} \quad \frac{(64)^2}{4} + \frac{(61)^2}{4} + \frac{(63)^2}{4} + \frac{(68)^2}{4} - \frac{(256)^2}{16} = 4102.50 - 4096 = 6.50$$

$$\text{Gasolinas } (A, B, C, D): \quad \frac{(62)^2}{4} + \frac{(81)^2}{4} + \frac{(52)^2}{4} + \frac{(61)^2}{4} - \frac{(256)^2}{16} = 4207.50 - 4096 = 111.50$$

$$\text{Carreteras } (\alpha, \beta, \gamma, \delta): \quad \frac{(61)^2}{4} + \frac{(68)^2}{4} + \frac{(62)^2}{4} + \frac{(65)^2}{4} - \frac{(256)^2}{16} = 4103.50 - 4096 = 7.50$$

La variación total es

$$(19)^2 + (16)^2 + (16)^2 + \cdots + (15)^2 + (23)^2 - \frac{(256)^2}{16} = 4244 - 4096 = 148.00$$

de manera que la variación debida a error es

$$148.00 - 16.50 - 6.50 - 111.50 - 7.50 = 6.00$$

Los resultados del análisis de varianza se recogen en la Tabla 16.31. El número total de grados de libertad es $N^2 - 1$ para un cuadrado $N \times N$. Cada fila, columna, letra latina y letra griega tiene $N - 1$ grados de libertad. Así pues, los grados de libertad para el error son $N^2 - 1 - 4(N - 1) = (N - 1)(N - 3)$. En nuestro caso, $N = 4$.

Tenemos $F_{.95,3,3} = 9.28$ y $F_{.99,3,3} = 29.5$. Luego podemos rechazar la hipótesis de que las gasolinas son iguales al nivel 0.05 pero no al 0.01.

PROBLEMAS DIVERSOS

16.18. Probar [como en la ecuación (15) de este capítulo] que $\sum_j \alpha_j = 0$.

Solución

Las medias de tratamiento de la población μ_j y la media de la población están relacionadas por

$$\mu = \frac{1}{a}\sum_j \mu_j \tag{53}$$

Entonces, como $\alpha_j = \mu_j - \mu$, tenemos, usando la ecuación (53),

$$\sum_j \alpha_j = \sum_j (\mu_j - \mu) = \sum_j \mu_j - a\mu = 0 \tag{54}$$

Tabla 16.31

Variación	Grados de libertad	Cuadrado medio	F
Filas (coches), 16.50	3	5.500	$\frac{5.500}{2.000} = 2.75$
Columnas (conductores), 6.50	3	2.167	$\frac{2.167}{2.000} = 1.08$
Gasolinas (A, B, C, D), 111.50	3	37.167	$\frac{37.167}{2.000} = 18.6$
Carreteras (α, β, γ, δ), 7.50	3	2.500	$\frac{2.500}{2.000} = 1.25$
Error, 6.00	3	2.000	
Total, 148.00	15		

16.19. Deducir (*a*) la ecuación (16) y (*b*) la ecuación (17) de este capítulo.

Solución

(*a*) Por definición se tiene

$$V_W = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 = b \sum_{j=1}^{a} \left[\frac{1}{b} \sum_{k=1}^{b} (X_{jk} - \bar{X}_{j.})^2 \right] = b \sum_{j=1}^{a} S_j^2$$

donde S_j^2 es la varianza de la muestra para el j-ésimo tratamiento. Entonces, como el tamaño de la muestra es b,

$$E(V_W) = b \sum_{j=1}^{a} E(S_j^2) = b \sum_{j=1}^{a} \left(\frac{b-1}{b} \sigma^2 \right) = a(b-1)\sigma^2$$

(*b*) Por definición,

$$V_B = b \sum_{j=1}^{a} (\bar{X}_{j.} - \bar{X})^2 = b \sum_{j=1}^{a} \bar{X}_{j.}^2 - 2b\bar{X} \sum_{j=1}^{a} \bar{X}_{j.} + ab\bar{X}^2 = b \sum_{j=1}^{a} \bar{X}_{j.}^2 - ab\bar{X}^2$$

ya que $\bar{X} = (\sum_j X_{jk})/a$. Omitiendo el índice de suma, se tiene

$$E(V_B) = b \sum E(\bar{X}_{j.}^2) - abE(\bar{X}^2) \tag{55}$$

Ahora bien, para cualquier variable aleatoria U, $E(U^2) = \text{var}\,(U) + [E(U)]^2$, donde var (U) denota la varianza de U. Así pues,

$$E(\bar{X}_{j.}^2) = \text{var}\,(\bar{X}_{j.}) + [E(\bar{X}_{j.})]^2 \tag{56}$$

$$E(\bar{X}^2) = \text{var}\,(\bar{X}) + [E(\bar{X})]^2 \tag{57}$$

Pero como las poblaciones de los tratamientos son normales con medias $\mu_j = \mu + \alpha$, tenemos que

$$\text{var}\,(\bar{X}_{j.}) = \frac{\sigma^2}{b} \tag{58}$$

$$\text{var}\,(\bar{X}) = \frac{\sigma^2}{ab} \tag{59}$$

$$E(\bar{X}_{j.}) = \mu_j = \mu + \alpha_j \tag{60}$$

$$E(\bar{X}) = \mu \tag{61}$$

Los resultados (56) a (61) junto con (53) nos dan

$$\begin{aligned} E(V_B) &= b \sum \left[\frac{\sigma^2}{b} + (\mu + \alpha_j)^2 \right] - ab \left[\frac{\sigma^2}{ab} + \mu^2 \right] = \\ &= a\sigma^2 + b \sum (\mu + \alpha_j)^2 - \sigma^2 - ab\mu^2 = \\ &= (a-1)\sigma^2 + ab\mu^2 + 2b\mu \sum \alpha_j + b \sum \alpha_j^2 + ab\mu^2 = \\ &= (a-1)\sigma^2 + b \sum \alpha_j^2 \end{aligned}$$

16.20. Demostrar el Teorema 1 de este capítulo.

Solución

Como muestra el Problema 16.19,

$$V_W = b\sum_{j=1}^{a} S_j^2 \qquad \text{o sea} \qquad \frac{V_W}{\sigma^2} = \sum_{j=1}^{a} \frac{bS_j^2}{\sigma^2}$$

donde S_j^2 es la varianza de la muestra para muestras de tamaño b tomadas en la población del tratamiento j. De la página 254 vemos que bS_j^2/σ^2 tiene una distribución ji-cuadrado con $b - 1$ grados de libertad. Luego, como las varianzas S_j^2 son independientes, concluimos de la página 272 que V_W/σ^2 tiene una distribución ji-cuadrado con $a(b - 1)$ grados de libertad.

PROBLEMAS SUPLEMENTARIOS

EXPERIMENTOS DE UN FACTOR

16.21. Se realiza un experimento para determinar las producciones de 5 variedades de trigo: A, B, C, D y E. Se asignan 4 parcelas a cada variedad. Las producciones (en bushels por acre) se dan en la Tabla 16.32. Supuesto que las parcelas son de la misma fertilidad y que las variedades se asignan al azar a las parcelas, determinar si hay diferencia entre las producciones al nivel de significación (a) 0.05 y (b) 0.01.

Tabla 16.32

A	20	12	15	19
B	17	14	12	15
C	23	16	18	14
D	15	17	20	12
E	21	14	17	18

16.22. Una empresa quiere comparar cuatro tipos de llantas: A, B, C y D. Sus vidas medias en rodaje (en miles de millas) se dan en la Tabla 16.33, donde cada tipo ha sido probado en seis coches similares asignados al azar a las llantas. Determinar si hay diferencia significativa al nivel de significación (a) 0.05 y (b) 0.01 entre las llantas.

Tabla 16.33

A	33	38	36	40	31	35
B	32	40	42	38	30	34
C	31	37	35	33	34	30
D	29	34	32	30	33	31

16.23. Un profesor quiere contrastar tres tipos distintos de enseñanza: I, II y III. Para ello, escoge al azar tres grupos de 5 estudiantes cada uno, y aplica a cada uno un método distinto. Tras proponer, al final del curso, el mismo examen a todos ellos, se obtienen las notas que indica la Tabla 16.34. Determinar si hay diferencia significativa entre los tres métodos al nivel de significación (a) 0.05 y (b) 0.01.

Tabla 16.34

Método I	75	62	71	58	73
Método II	81	85	68	92	90
Método III	73	79	60	75	81

MODIFICACIONES PARA NUMEROS DISTINTOS DE OBSERVACIONES

16.24. La Tabla 16.35 da el número de millas por galón recorridas por coches similares usando cinco tipos distintos de gasolina. Determinar si hay diferencia significativa entre las gasolinas al nivel de significación (*a*) 0.05 y (*b*) 0.01.

Tabla 16.35

Tipo *A*	12	15	14	11	15
Tipo *B*	14	12	15		
Tipo *C*	11	12	10	14	
Tipo *D*	15	18	16	17	14
Tipo *E*	10	12	14	12	

16.25. Durante un curso, un estudiante obtuvo las calificaciones que figuran en la Tabla 16.36. Determinar si hay diferencia significativa entre esas calificaciones al nivel de significación.

Tabla 16.36

Matemáticas	72	80	83	75	
Ciencias	81	74	77		
Inglés	88	82	90	87	80
Economía	74	71	77	70	

EXPERIMENTOS DE DOS FACTORES

16.26. Los artículos manufacturados por una compañía se producen en 3 máquinas distintas manejadas por 3 operarios diferentes. El dueño desea saber si hay diferencia (*a*) entre los operarios y (*b*) entre las máquinas. Se realiza un experimento para conocer el número de artículos producidos al día, con los resultados que recoge la Tabla 16.37. Establecer la deseada información al nivel de significación 0.05.

Tabla 16.37

	Operador		
	1	2	3
Máquina *A*	23	27	24
Máquina *B*	34	30	28
Máquina *C*	28	25	27

16.27. Rehacer el Problema 16.26 al nivel de significación 0.01.

16.28. Se siembran semillas de maíz de 4 tipos distintos en 5 bloques, cada bloque dividido en 4 parcelas que se asignan al azar a dichos 4 tipos de semillas. Determinar el nivel de significación 0.05 si las producciones en bushels por acre, dadas en la Tabla 16.38, varían significativamente con diferentes (*a*) terrenos (o sea, los 5 bloques) y (*b*) tipos de maíz.

Tabla 16.38

	Tipo de maíz			
	I	II	III	IV
Bloque *A*	12	15	10	14
Bloque *B*	15	19	12	11
Bloque *C*	14	18	15	12
Bloque *D*	11	16	12	16
Bloque *E*	16	17	11	14

16.29. Resolver el Problema 16.28 al nivel de significación 0.01.

16.30. Supongamos que en el Problema 16.22 se hace la primera observación para cada tipo de llanta usando un tipo particular de coche, la segunda con otro tipo de coche, etc. Determinar si hay diferencia significativa al nivel de significación 0.05 entre (*a*) los tipos de llantas y (*b*) las clases de coches usados.

16.31. Rehacer el Problema 16.30 al nivel de significación 0.01.

16.32. Supongamos que en el Problema 16.23 la primera entrada para cada método de enseñanza corresponde a un estudiante de un colegio concreto, la segunda a uno de otro colegio, etc. Contrastar la hipótesis, al nivel de significación 0.05, de que hay diferencia entre (*a*) los métodos de enseñanza y (*b*) los colegios.

16.33. Se realiza un experimento para saber si el color del cabello y la altura de mujeres adultas en EE.UU. tienen alguna influencia sobre el rendimiento escolar. Los resultados figuran en la Tabla 16.39, donde los números indican individuos en el 10% más alto de entre los que se gradúan. Analizar el experimento al nivel de significación 0.05.

Tabla 16.39

	Pelirroja	Rubia	Castaña
Alta	75	78	80
Media	81	76	79
Baja	73	75	77

16.34. Repetir el Problema 16.33 al nivel de significación 0.01.

EXPERIMENTOS DE DOS FACTORES CON REPETICION

16.35. Supongamos que el experimento del Problema 16.21 se realizó en el sur de EE.UU. y que las columnas de la Tabla 16.32 indican ahora 4 tipos de fertilizantes, mientras que un experimento similar se llevó a cabo en el Oeste con los resultados de la Tabla 16.40. Determinar al nivel de significación 0.05 si hay diferencia en producción debida a (*a*) los fertilizantes y (*b*) la localización.

Tabla 16.40

A	16	18	20	23
B	15	17	16	19
C	21	19	18	21
D	18	22	21	23
E	17	18	24	20

16.36. Rehacer el Problema 16.35 al nivel de significación 0.01.

16.37. La Tabla 16.41 da el número de artículos producidos por 4 trabajadores en dos máquinas distintas, I y II, en diferentes días de la semana. Determinar si hay diferencia significativa al nivel 0.05 entre (*a*) los trabajadores y (*b*) las máquinas.

Tabla 16.41

	Máquina I				
	L.	Mar.	Miér.	J.	V.
Operador *A*	15	18	17	20	12
Operador *B*	12	16	14	18	11
Operador *C*	14	17	18	16	13
Operador *D*	19	16	21	23	18

	Máquina II				
	L.	Mar.	Miér.	J.	V.
Operador *A*	14	16	18	17	15
Operador *B*	11	15	12	16	12
Operador *C*	12	14	16	14	11
Operador *D*	17	15	18	20	17

CUADRADOS LATINOS

16.38. Se lleva a cabo un experimento para comprobar los efectos en la producción de maíz de 4 fertilizantes (*A*, *B*, *C* y *D*) y de las variaciones del terreno en dos direcciones perpendiculares. El cuadrado latino de la Tabla 16.42 da los resultados obtenidos, donde los números muestran la producción de maíz por unidad de área. Contrastar al nivel de significación 0.01 la hipótesis de que no hay diferencia entre (*a*) los fertilizantes y (*b*) las variaciones del terreno.

Tabla 16.42

C 8	A 10	D 12	B 11
A 14	C 12	B 11	D 15
D 10	B 14	C 16	A 10
B 7	D 16	A 14	C 12

16.39. Resolver el Problema 16.38 al nivel de significación 0.05.

16.40. Con referencia al Problema 16.33, supongamos que introducimos un factor adicional, dando la parte E, M o W de los EE.UU. en que nació un estudiante, como muestra la Tabla 16.43. Determinar si hay diferencia significativa al nivel 0.05 en los rendimientos escolares debidas a diferencias en (*a*) altura, (*b*) color del cabello y (*c*) lugar de nacimiento.

Tabla 16.43

E 75	W 78	M 80
M 81	E 76	W 79
W 73	M 75	E 77

CUADRADOS GRECO-LATINOS

16.41. Con objeto de lograr mejorar la calidad de un pienso para gallinas, se han añadido dos productos químicos a sus ingredientes básicos. Las distintas cantidades del primero se indican por A, B, C y D, y las del segundo por α, β, γ y δ. Se da el pienso a animales ordenados en grupos de acuerdo con cuatro pesos iniciales diferentes (W_1, W_2, W_3 y W_4) y cuatro especies diferentes (S_1, S_2, S_3 y S_4). Los aumentos de peso por unidad de tiempo vienen dados en el cuadrado greco-latino de la Tabla 16.44. Hacer un análisis de varianza del experimento al nivel de significación 0.05, sacando las conclusiones pertinentes.

Tabla 16.44

	W_1	W_2	W_3	W_4
S_1	C_γ 8	B_β 6	A_α 5	D_δ 6
S_2	A_δ 4	D_α 3	C_β 7	B_γ 3
S_3	D_β 5	A_γ 6	B_δ 5	C_α 6
S_4	B_α 6	C_δ 10	D_γ 10	A_β 8

16.42. Cuatro tipos de cables (T_1, T_2, T_3 y T_4) se fabrican en cada una de las empresas (C_1, C_2, C_3 y C_4). Cuatro operarios (A, B, C y D) usando cuatro máquinas distintas (α, β, γ y δ) miden las tensiones de ruptura de esos cables, obteniendo los valores promedio que indica el cuadrado greco-latino de la Tabla 16.45. Hacer un análisis de varianza al nivel de significación 0.05 para llegar a las conclusiones pertinentes.

Tabla 16.45

	C_1	C_2	C_3	C_4
T_1	A_β 164	B_γ 181	C_α 193	D_δ 160
T_2	C_δ 171	D_α 162	A_γ 183	B_β 145
T_3	D_γ 198	C_β 212	B_δ 207	A_α 188
T_4	B_α 157	A_δ 172	D_β 166	C_γ 136

PROBLEMAS DIVERSOS

16.43. La Tabla 16.46 proporciona datos sobre la herrumbre acumulada sobre el hierro tra-

Tabla 16.46

A	3	5	4	4
B	4	2	3	3
C	6	4	5	5

tado con productos químicos A, B o C, respectivamente. Determinar al nivel de significación (*a*) 0.05 y (*b*) 0.01 si hay diferencia significativa entre esos tratamientos.

16.44. Un experimento mide los coeficientes de inteligencia (IQ) de estudiantes varones adultos de estatura alta, media y baja, con los resultados que figuran en la Tabla 16.47. Determinar si hay diferencia significativa al nivel de significación (*a*) 0.05 y (*b*) 0.01 en los IQ por efecto de las diferencias en altura.

Tabla 16.47

Alto	110	105	118	112	90	
Bajo	95	103	115	107		
Medio	108	112	93	104	96	102

16.45. Probar los resultados (10), (11) y (12) de este capítulo.

16.46. Se hace una prueba para saber si responden mejor los veteranos o los no veteranos de diversos IQ. Las calificaciones obtenidas son las de la Tabla 16.48. Determinar si hay diferencia significativa al nivel de significación 0.05, debida a diferencias en (*a*) ser o no veterano y (*b*) IQ.

Tabla 16.48

	Resultado del test		
	Alto IQ	Medio IQ	Bajo IQ
Veterano	90	81	74
No veterano	85	78	70

16.47. Repetir el Problema 16.46 al nivel de significación 0.01.

16.48. La Tabla 16.49 muestra las notas de una muestra de estudiantes procedentes de diferentes partes del país y con diferentes IQ. Analizar los datos de la tabla al nivel de significación 0.05 y establecer conclusiones.

Tabla 16.49

	Resultado del test		
	Alto IQ	Medio IQ	Bajo IQ
Este	88	80	72
Oeste	84	78	75
Sur	86	82	70
Norte y central	80	75	79

16.49. Resolver el Problema 16.48 al nivel de significación 0.01.

16.50. En el Problema 16.37, ¿puede determinar si hay diferencia significativa en el número de artículos producidos en distintos días de la semana? Explíquese.

16.51. En cálculos de análisis de varianza se sabe que puede añadirse o restarse una constante adecuada a cada entrada sin que ello afecte a las conclusiones. ¿Es eso cierto también si cada entrada se multiplica por una constante? Justificar la respuesta.

16.52. Deducir los resultados (24), (25) y (26) para números distintos de observaciones.

16.53. Supongamos que los resultados de la Tabla 16.46 del Problema 16.43 son válidos para la parte nordeste de los EE.UU., mientras que los de la Tabla 16.50 lo son para la parte oeste. Determinar al nivel de significación 0.05 si hay diferencias debidas a (*a*) los productos químicos y (*b*) la localización.

Tabla 16.50

A	5	4	6	3
B	3	4	2	3
C	5	7	4	6

16.54. Refiriéndonos a los Problemas 16.21 y 16.35, supongamos que se realiza un experimento adicional en la parte nordeste de EE.UU. y produce los resultados de la Tabla 16.51. Determinar al nivel 0.05 si hay diferencia en la producción debida (*a*) a los fertilizantes, y (*b*) a las tres localizaciones.

Tabla 16.51

A	17	14	18	12
B	20	10	20	15
C	18	15	16	17
D	12	11	14	11
E	15	12	19	14

16.55. Repetir el Problema 16.54 al nivel de significación 0.01.

16.56. Hacer un análisis de varianza del cuadrado latino de la Tabla 16.52 al nivel de significación 0.05 y establecer las conclusiones pertinentes.

16.57. Describir un experimento que conduzca al cuadrado latino de la Tabla 16.52.

Tabla 16.52

Factor 1

Factor 2			
	B 16	*C* 21	*A* 15
	A 18	*B* 23	*C* 14
	C 15	*A* 18	*B* 12

16.58. Hacer un análisis de varianza del cuadrado greco-latino de la Tabla 16.53 al nivel de significación 0.05 y sacar las conclusiones.

Tabla 16.53

Factor 1

Factor 2				
	A_γ 6	B_β 12	C_δ 4	D_α 18
	B_δ 3	A_α 8	D_γ 15	C_β 14
	D_β 15	C_γ 20	B_α 9	A_δ 5
	C_α 16	D_δ 6	A_β 17	B_γ 7

16.59. Describir un experimento que conduzca al cuadrado greco-latino de la Tabla 16.53.

16.60. Describir cómo usar el análisis de varianza para experimentos de tres factores con repetición.

16.61. Enunciar y resolver un problema que ilustre el procedimiento del Problema 16.60.

16.62. Probar (*a*) la ecuación (30) y (*b*) los resultados (31) a (34) de este capítulo.

16.63. En la práctica, ¿cabe esperar hallar (*a*) un cuadrado latino 2×2 y (*b*) un cuadrado greco-latino 3×3? Explicar la razón.

CAPITULO 17

Contrastes no paramétricos

INTRODUCCION

La mayor parte de los contrastes de hipótesis y significación (o reglas de decisión) considerados en los capítulos precedentes requieren varias suposiciones acerca de la distribución de la población cuyas muestras se analizan. Por ejemplo, en la página 187 las distribuciones de la población se exigían normales o casi normales.

En la práctica aparecen situaciones en las que tales requisitos no están justificados, como es el caso de una población fuertemente asimétrica. A causa de ello, los estadísticos han creado varios contrastes y métodos que son independientes de las distribuciones de la población y de los parámetros asociados. Estos se llaman *contrastes* o *tests no paramétricos*.

Los tests no paramétricos se pueden usar como abreviaciones de contrastes más complicados. Son especialmente útiles cuando se trata con datos no numéricos, por ejemplo, cuando los consumidores colocan productos por orden de preferencia.

EL TEST DE LOS SIGNOS

Consideremos la Tabla 17.1, que indica los números de tuercas defectuosas producidas por dos tipos de máquinas, I y II, en 12 días consecutivos y que supone que ambas máquinas tienen la misma producción diaria. Deseamos contrastar la hipótesis H_0 de que no hay diferencia entre las máquinas: que las diferencias observadas se deben simplemente al azar, lo que equivale a decir que las muestras proceden de la misma población.

Un sencillo test no paramétrico en este caso de muestras emparejadas la proporciona el *test de los signos*, que consiste en tomar la diferencia entre los números de tuercas defectuosas cada día y escribir sólo el *signo* de esa diferencia; por ejemplo, para el primer día se tiene 47-71, que es negativo. De este modo se obtiene de la Tabla 17.1 la secuencia de signos

$$- \quad - \quad + \quad - \quad - \quad - \quad + \quad - \quad + \quad - \quad - \quad - \tag{1}$$

(o sea, tres + y nueve −). Ahora bien, si fuese tan probable obtener + como −, esperaríamos seis + + y seis −. El contraste de H equivale al de si una moneda es buena sabiendo que en 12 tiradas han salido 3 caras (+) y 9 cruces (−). Ello involucra a la distribución binomial del Capítulo 7. El Problema 17.1 muestra que mediante un contraste de dos colas con la distribución binomial al nivel de significación 0.05, no podemos rechazar H_0; esto es, no hay diferencia entre las máquinas a ese nivel.

Tabla 17.1

Día	1	2	3	4	5	6	7	8	9	10	11	12
Máquina I	47	56	54	49	36	48	51	38	61	49	56	52
Máquina II	71	63	45	64	50	55	42	46	53	57	75	60

Nota 1: Si un día las máquinas producen el mismo número de tuercas defectuosas, aparecerá una diferencia *cero* en la secuencia (1). En tal caso podemos omitir ese par de valores muestrales y utilizar 11 en vez de 12 observaciones.

Nota 2: Se puede usar también una aproximación normal a la distribución binomial, mediante corrección por continuidad (véase Prob. 17.2).

Aunque el test de los signos es particularmente útil para muestras emparejadas, como en la Tabla 17.1, se puede usar también en problemas con una sola muestra (véase Probs. 17.3 y 17.4).

EL *U*-TEST DE MANN-WHITNEY

Consideremos la Tabla 17.2, que da las resistencias de cables fabricados con dos aleaciones distintas, I y II. En esa tabla tenemos dos muestras: 8 cables de la aleación I y 10 de la II. Queremos decidir si hay o no diferencia entre las muestras, o sea, si proceden o no de una misma población. Si bien este problema se puede atacar con el contraste t del Capítulo 11, es conveniente un test no paramétrico llamado el *U-test de Mann-Whitney*, o abreviadamente, *U-test*. Consiste en los siguientes pasos:

Tabla 17.2

Aleación I				Aleación II				
18.3	16.4	22.7	17.8	12.6	14.1	20.5	10.7	15.9
18.9	25.3	16.1	24.2	19.6	12.9	15.2	11.8	14.7

Paso 1. Combinar todos los valores muestrales en una ordenación del menor al mayor, y asignar rangos (en este caso de 1 a 8) a todos esos valores. Si dos o más valores muestrales son idénticos (o sea, son coincidencias), se les asigna a cada uno un rango que es la media de los rangos que les hubieran correspondido sin tal coincidencia. Si la entrada 18.9 en la Tabla 17.2 fuese 18.3, dos valores idénticos 18.3 ocuparían los rangos 12 y 13 en la ordenación, de modo que se asignaría a cada uno el rango $\frac{1}{2}(12 + 13) = 12.5$.

Paso 2. Hallar la suma de los rangos para cada muestra. Las denotamos R_1 y R_2, donde N_1 y N_2 son los respectivos tamaños muestrales. Por conveniencia elegimos N_1 que es el menor si son desiguales tales que $N_1 \leqslant N_2$. Una diferencia significativa entre las sumas de rangos R_1 y R_2 implica una diferencia significativa entre las muestras.

Paso 3. Para contrastar la diferencia entre las sumas de rangos, usamos el estadístico

$$U = N_1N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 \tag{2}$$

correspondiente a la muestra 1. La distribución muestral U es simétrica y tiene una media y una varianza dadas por

$$\mu_U = \frac{N_1N_2}{2} \qquad \sigma_U^2 = \frac{N_1N_2(N_1 + N_2 + 1)}{12} \tag{3}$$

Si N_1 y N_2 son ambos al menos iguales a 8, resulta que la distribución de U es aproximadamente normal, de manera que

$$z = \frac{U - \mu_U}{\sigma_U} \tag{4}$$

está normalmente distribuido con media 0 y varianza 1. Usando el Apéndice II, podemos entonces decidir si las muestras son significativamente diferentes. El Problema 17.5 enseña que hay diferencia significativa entre los cables al nivel 0.05.

Nota 3: Un valor correspondiente a la muestra 2 viene dado por el estadístico

$$U = N_1N_2 + \frac{N_2(N_2 + 1)}{2} - R_2 \tag{5}$$

y tiene la misma distribución muestral que el (2), con la media y la varianza de las fórmulas (3). El estadístico (5) está relacionado con el (2), porque si U_1 y U_2 son los valores correspondientes a los estadísticos (2) y (5), respectivamente, se tiene

$$U_1 + U_2 = N_1N_2 \tag{6}$$

Se tiene además

$$R_1 + R_2 = \frac{N(N + 1)}{2} \tag{7}$$

donde $N = N_1 + N_2$. El resultado (7) proporciona una comprobación para los cálculos.

Nota 4: El estadístico U en (2) es el número total de veces que los valores de la muestra 1 preceden a los de la muestra 2 cuando todos los valores se ordenan de modo creciente. Ello proporciona un *método* alternativo *de recuento* para hallar U.

EL *H*-TEST DE KRUSKAL-WALLIS

El U-test es un test no paramétrico para decidir si dos muestras provienen o no de la misma población. Una generalización para k muestras la da el *H-test de Kruskal-Wallis*, o simplemente *H-test*.

El *H-test* puede describirse como sigue: Sean k muestras de tamaños $N_1, N_2, ..., N_k$, con tamaño suma total $N = N_1 + N_2 + \cdots + N_k$. Supongamos que los datos de todas las muestras se ordenan y que las sumas de rangos para las k muestras son $R_1, R_2, ..., R_k$, respectivamente. Si definimos el estadístico

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{N_j} - 3(N+1) \tag{8}$$

se puede demostrar que su distribución de muestreo es muy próxima a una *distribución ji-cuadrado* con $k - 1$ grados de libertad, supuesto que $N_1, N_2, ..., N_k$ son al menos 5 todos ellos.

El *H*-test nos da un test no paramétrico en el *análisis de varianza* para experimentos de un factor, y admite generalización.

EL *H*-TEST CORREGIDO POR COINCIDENCIAS

En caso de haber demasiadas coincidencias entre las observaciones en los datos muestrales, el valor de H dado por (8) es menor de lo que debiera. El valor corregido de H, denotado H, se obtiene dividiendo el valor dado en (8) por el factor de corrección

$$1 - \frac{\sum(T^3 - T)}{N^3 - N} \tag{9}$$

donde T es el número de coincidencias correspondientes a cada observación y donde la suma se toma sobre todas las observaciones. Si no hay coincidencias, $T = 0$ y el factor (9) se reduce a 1, así que no se precisa corrección. En la práctica, la corrección suele ser despreciable (o sea, no suficiente para cambiar la decisión).

EL TEST DE LAS RACHAS PARA EL CARACTER ALEATORIO

Aunque la palabra «aleatorio» ha sido utilizada con frecuencia en este libro (por ejemplo en «muestreo aleatorio»), no hemos visto ningún criterio de aleatoriedad. Un test no paramétrico a tal fin lo proporciona la *teoría de rachas*.

Para entender qué son las rachas (o escalones) consideremos una secuencia con dos símbolos, a y b tal como

$$a \;\; a \mid b \;\; b \;\; b \mid a \mid b \;\; b \mid a \;\; a \;\; a \;\; a \;\; a \mid b \;\; b \;\; b \mid a \;\; a \;\; a \;\; a \mid \tag{10}$$

Al tirar una moneda, por ejemplo, a sería «cara» y b «cruz»; en el muestreo de tuercas defectuosas, a sería «defectuosa» y b «no defectuosa».

Una *racha* se define como un conjunto de símbolos idénticos (o relacionados) contenido entre dos símbolos diferentes o uno sólo si estamos al comienzo o al final de la secuencia. Leyendo de izquierda a derecha en la secuencia (10) la primera racha, indicada por una barra vertical, consiste de dos aes, la segunda de tres bes, la tercera de una a, etc. Hay siete rachas en total.

Parece claro que existe relación entre aleatoriedad y el número de rachas. Así, para la secuencia

$$a \mid b \mid a \mid b \mid a \mid b \mid a \mid b \mid a \mid b \mid a \mid b \mid \tag{11}$$

hay un *esquema cíclico*, en el que vamos de a a b, vuelta al a, etc, que difícilmente puede ser aleatorio. En ese caso tenemos *demasiadas* rachas (de hecho, hay el máximo posible con ese número de letras a y b).

Por otra parte, para la secuencia

$$a\ a\ a\ a\ a\ a \mid b\ b\ b\ b \mid a\ a\ a\ a\ a \mid b\ b\ b \mid \tag{12}$$

parece haber un *esquema de tendencia* o de inercia, en el que las *aes* y las *bes* están agrupadas. En este caso hay *demasiado pocas* rachas, y no consideraríamos tampoco aleatoria a esa secuencia.

Así pues, una secuencia se considera no aleatoria si hay demasiadas o demasiado pocas rachas, y aleatoria en los demás casos. Para cuantificar esa idea, supongamos que formamos todas las posibles secuencias con N_1 *aes* y N_2 *bes*, para un total de N símbolos, $(N_1 + N_2 = N)$. La colección de todas esas secuencias nos da una distribución muestral. Cada secuencia tiene asociado un número de rachas, denotado por V. De este modo nos vemos conducidos a la distribución muestral del estadístico V. Se demuestra que esta distribución tiene media y varianza dadas por

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1 \qquad \sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)} \tag{13}$$

Mediante las fórmulas (13), podemos contrastar la hipótesis de aleatoriedad a niveles de significación apropiados. Resulta que si N_1 y N_2 son ambos al menos iguales a 8, entonces la distribución muestral de V es muy próxima una distribución normal. Luego

$$z = \frac{V - \mu_V}{\sigma_V} \tag{14}$$

está normalmente distribuido con media 0 y varianza 1, y se puede utilizar el Apéndice II.

OTRAS APLICACIONES DEL TEST DE LAS RACHAS

He aquí otras aplicaciones del test de las rachas en problemas de estadística:

1. **Test sobre- y bajo-mediana para la aleatoriedad de datos numéricos.** Para determinar si unos datos numéricos (como los tomados en una muestra) son aleatorios, los colocamos primero en el *mismo orden* en que fueron tomados, hallamos la mediana y sustituimos cada entrada por la letra a o b según que ese valor esté *sobre* o *bajo* la mediana. Si un valor coincide con la mediana, lo suprimimos. La muestra es aleatoria o no según lo sea la secuencia de *aes* y *bes* así obtenida. (Véase Prob. 17.20).
2. **Diferencias en poblaciones de las que se toman muestras.** Sean dos muestras de tamaños m y n, denotadas por $a_1, a_2, \ldots, a_m$ y $b_1\ b_2, \ldots, b_n$. Para decidir si las muestras proceden o no de una misma población, colocamos los $m + n$ valores en orden creciente. Si varios valores coinciden, se ordenan por algún procedimiento de azar (usando números aleatorios, por

ejemplo). Si la secuencia resultante es aleatoria, concluimos que las dos muestras no son realmente diferentes y provienen, por tanto, de una misma población; si no es aleatoria, no podemos sacar esa conclusión. Este test proporciona una alternativa al U-test de Mann-Whitney (véase Prob. 17.21).

CORRELACION DE RANGO DE SPEARMAN

Se pueden usar también métodos no paramétricos para medir la correlación de dos variables X e Y. En lugar de usar valores precisos de las variables, o cuando tal precisión no es alcanzable, a los datos se les pueden asignar un rango de 1 a N ordenándolos por su tamaño, importancia, etc. Si X e Y tienen asignado un rango así, el *coeficiente de correlación de rango*, o *fórmula de Spearman para la correlación de rango* (como se suele llamar), viene dado por

$$r_S = 1 - \frac{6\sum D^2}{N(N^2 - 1)} \tag{15}$$

donde D denota la diferencia entre los rangos de valores correspondientes de X e Y, y donde N es el número de pares de valores (X, Y) en los datos.

PROBLEMAS RESUELTOS

EL TEST DE LOS SIGNOS

17.1. Con referencia a la Tabla 17.1, contrastar la hipótesis H_0 de que no hay diferencia entre las máquinas I y II frente a la hipótesis alternativa H_1 de que sí la hay, al nivel de significación 0.05

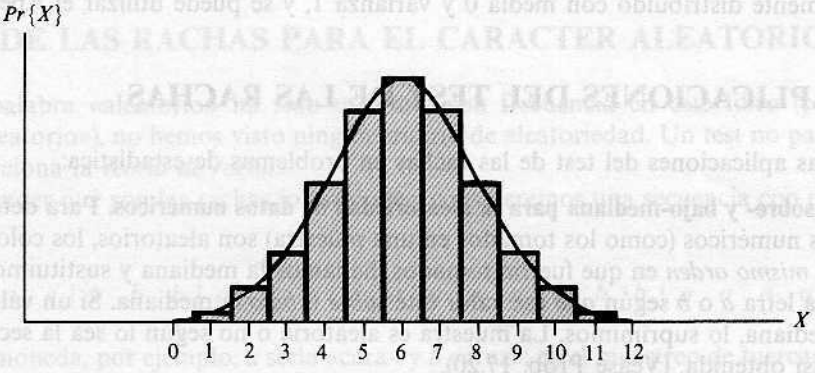


Figura 17.1.

Solución

La Figura 17.1 es un gráfico de la distribución binomial (y una aproximación normal a ella) que da

las probabilidades de X caras en 12 tiradas de una moneda buena, donde $X = 0, 1, 2, \ldots, 12$. Del Capítulo 7 sabemos que la probabilidad de X caras es

$$\Pr\{X\} = \binom{12}{X}\left(\frac{1}{2}\right)^X\left(\frac{1}{2}\right)^{12-X} = \binom{12}{X}\left(\frac{1}{2}\right)^{12}$$

de donde $\Pr\{0\} = 0.00024$, $\Pr\{1\} = 0.00293$, $\Pr\{2\} = 0.01611$ y $\Pr\{3\} = 0.05371$.

Como H_1 es la hipótesis de que hay *diferencia* entre las máquinas, no que la I sea *mejor* que la II, usamos un contraste de dos colas. Al nivel de significación 0.05 cada cola tiene asociada la probabilidad $\frac{1}{2}(0.05) = 0.025$. Ahora sumamos las probabilidades de la cola izquierda hasta que la suma sobrepase 0.025. Luego

$$\Pr\{0, 1 \text{ ó } 2 \text{ caras}\} = 0.00024 + 0.00293 + 0.01611 = \boldsymbol{0.01928}$$

$$\boldsymbol{Pr}\{0, 1, 2 \text{ ó } 3 \text{ caras}\} = 0.00024 + 0.00293 + 0.01611 + 0.05371 = 0.07299$$

Como 0.025 es mayor que 0.01928 pero menor que 0.07299, podemos rechazar H_0 si el número de caras es 2 o menor (o por simetría, si es 10 o mayor); no obstante, el número de caras [los signos + en la secuencia (1)] es 3. Luego no podemos rechazar H_0 al nivel de significación 0.05 y debemos concluir que no hay diferencia entre las máquinas a ese nivel.

17.2. Rehacer el Problema 17.1 usando una aproximación normal a la distribución binomial.

Solución

Para lograr una aproximación normal a la distribución binomial, usaremos el hecho de que el recuento z correspondiente al número de caras es

$$z = \frac{X - \mu}{\sigma} = \frac{X - Np}{\sqrt{Npq}}$$

(véase pág. 161). Como la variable X para la distribución binomial es discreta mientras que para una distribución normal es continua, hacemos una *corrección por continuidad* (por ejemplo, 3 caras es realmente un valor entre 2,5 y 3,5 caras). Eso equivale a disminuir X en 0,5 si $X > Np$ y a aumentar X en 0,5 si $X < Np$. Ahora bien, $N = 12$, $\mu = Np = (12)(0.5) = 6$ y $\sigma = \sqrt{Npq} = \sqrt{(12)(0.5)(0.5)} = 1.73$, de modo que

$$z = \frac{(3 + 0.5) - 6}{1.73} = -1.45$$

Como esto es mayor que -1.96 (el valor de z para el cual el área en la cola izquierda es 0.025), llegamos a la misma conclusión que en el Problema 17.1.

Nótese que $\Pr\{z \leq -1.45\} = 0.0735$, que está en buen acuerdo con la $\Pr\{X \leq 3 \text{ caras}\} = 0.07299$ del Problema 17.1.

17.3. La empresa PQR afirma que la vida media de un tipo de baterías que fabrica es superior a 250 horas(h). Un defensor de los consumidores desea saber si tal afirmación está justificada, y para ello mide las vidas medias de 24 baterías, con los resultados que figuran en la Tabla 17.3. Supuesto que la muestra era aleatoria, determinar si la empresa tiene razón al nivel de significación 0.05.

Solución

Sea H_0 la hipótesis de que las baterías de esa empresa tienen vida media igual a 250 h, y sea H_1 la hipótesis de que la vida media es mayor que 250 h. Para contrastar H_0, podemos usar el test de los signos. Para ello, restamos 250 a cada entrada de la Tabla 17.3 y anotamos los signos de las diferencias, tal como indica la Tabla 17.4. Vemos que hay 15 signos + y 9 signos −.

Tabla 17.3

271	230	198	275	282	225	284	219
253	216	262	288	236	291	253	224
264	295	211	252	294	243	272	268

Tabla 17.4

+	−	−	+	+	−	+	−
+	−	+	+	−	+	+	−
+	+	−	+	+	−	+	+

Usando un contraste unilateral al nivel de significación 0.05, rechazaríamos H_0 si el recuento z fuese mayor que 1.645 (Fig. 17.2). Como el z, usando corrección por continuidad, es

$$z = \frac{(15 - 0.5) - (24)(0.5)}{\sqrt{(24)(0.5)(0.5)}} = 1.02$$

la afirmación de la empresa no estaba justificada al nivel 0.05.

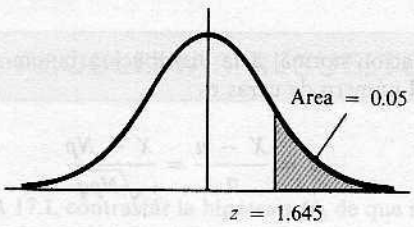


Figura 17.2.

17.4. La Tabla 17.5 recoge una muestra de 40 notas en un examen de ámbito nacional. Contrastar al nivel de significación 0.05 la hipótesis de que la nota mediana de todos los participantes es (*a*) 66 y (*b*) 75.

Solución

(*a*) Restando 66 de las entradas de la Tabla 17.5 y reteniendo sólo los signos de las diferencias, se obtiene la Tabla 17.6, en la que hay 23 signos +, 15 signos − y 2 ceros. Descartados los ceros, quedan 23 + y 15 −. Usando un contraste bilateral con la distribución normal con probabilidades $\frac{1}{2}(0.05) = 0.025$ en cada cola (Fig. 17.3), adoptamos la siguiente regla de decisión:

Tabla 17.5

71	67	55	64	82	66	74	58	79	61
78	46	84	93	72	54	78	86	48	52
67	95	70	43	70	73	57	64	60	83
73	40	78	70	64	86	76	62	95	66

Tabla 17.6

+	+	−	−	+	0	+	−	+	−
+	−	+	+	+	−	+	+	−	−
+	+	+	−	+	+	−	−	−	+
+	−	+	+	−	+	+	−	+	0

Tabla 17.7

−	−	−	−	+	−	−	−	+	−
+	−	+	+	−	−	+	+	−	−
−	+	−	−	−	−	−	−	−	+
−	−	+	−	−	+	+	−	+	−

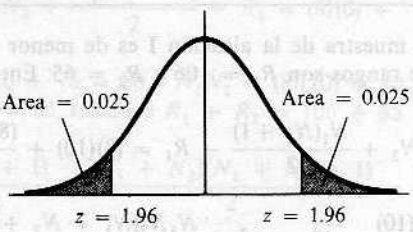


Figura 17.3.

Aceptar la hipótesis si $-1.96 \leqslant z \leqslant 1.96$.
Rechazarla en caso contrario.

Como
$$z = \frac{X - Np}{\sqrt{Npq}} = \frac{(23 - 0.5) - (38)(0.5)}{\sqrt{(38)(0.5)(0.5)}} = 1.14$$

aceptamos la hipótesis de que la mediana es 66, al nivel 0.05.

Nótese que podríamos haber usado 15, el número de signos −. En ese caso,

$$z = \frac{(15 + 0.5) - (38)(0.5)}{\sqrt{(38)(0.5)(0.5)}} = -1.14$$

con la misma conclusión.

(*b*) Restando 75 de las entradas de la Tabla 17.5 se llega a la Tabla 17.7, con 13 + y 27 −. Como

$$z = \frac{(13 + 0.5) - (40)(0.5)}{\sqrt{(40)(0.5)(0.5)}} = -2.06$$

rechazamos la hipótesis de que la mediana es 75, al nivel 0.05.

Por este método, podemos llegar al intervalo de confianza del 95% para la nota mediana del examen. (Véase Prob. 17.30.)

EL U-TEST DE MANN-WHITNEY

17.5. Con referencia a la Tabla 17.2, determinar si hay diferencia entre los cables de aleaciones I y II, al nivel de significación 0.05.

Solución

Seguimos los pasos 1, 2 y 3 descritos antes en este capítulo.

Paso 1. Combinando los 18 valores de la muestra en una ordenación de menor a mayor tenemos la primera fila de la Tabla 17.8. La segunda fila les asigna rango de 1 a 18.

Paso 2. Para hallar la suma de los rangos de cada muestra, reescribimos la Tabla 17.2 usando los rangos asociados de la Tabla 17.8, lo que nos da la Tabla 17.9. La suma de los rangos es 106 para la aleación I y 65 para la aleación II.

Tabla 17.8

10.7	11.8	12.6	12.9	14.1	14.7	15.2	15.9	16.1	16.4	17.8	18.3	18.9	19.6	20.5	22.7	24.2	25.3
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Paso 3. Puesto que la muestra de la aleación I es de menor tamaño, $N_1 = 8$ y $N_2 = 10$. Las correspondientes sumas de rangos son $R_1 = 106$ y $R_2 = 65$. Entonces

$$U = N_1N_2 + \frac{N_1(N_1+1)}{2} - R_1 = (8)(10) + \frac{(8)(9)}{2} - 106 = 10$$

$$\mu_U = \frac{N_1N_2}{2} = \frac{(8)(10)}{2} = 40 \qquad \sigma_U^2 = \frac{N_1N_2(N_1+N_2+1)}{12} = \frac{(8)(10)(19)}{12} = 126.67$$

Tabla 17.9

Aleación I		Aleación II	
Resistencia del cable	Rango	Resistencia del cable	Rango
18.3	12	12.6	3
16.4	10	14.1	5
22.7	16	20.5	15
17.8	11	10.7	1
18.9	13	15.9	8
25.3	18	19.6	14
16.1	9	12.9	4
24.2	17	15.2	7
	Suma 106	11.8	2
		14.7	6
			Suma 65

Así pues $\sigma_U = 11.25$ y

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{10 - 40}{11.25} = -2.67$$

Como la hipótesis H_0 que estamos estudiando es que *no* hay diferencia entre las aleaciones, se requiere un contraste de dos colas. Al nivel de significación 0.05, tenemos como regla de decisión:

Aceptar H_0 si $-1.96 \leqslant z \leqslant 1.96$.
Rechazarla en caso contrario.

Como $z = -2.67$, rechazamos H_0 y concluimos que hay diferencia entre las dos aleaciones al nivel 0.05.

17.6. Comprobar los resultados (6) y (7) de este capítulo para los datos del Problema 17.5.

Solución

(*a*) Dado que las muestras 1 y 2 resultan valores para U dados por

$$U_1 = N_1N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 = (8)(10) + \frac{(8)(9)}{2} - 106 = 10$$

$$U_2 = N_1N_2 + \frac{N_2(N_2 + 1)}{2} - R_2 = (8)(0) + \frac{(10)(11)}{2} - 65 = 70$$

tenemos $U_1 + U_2 = 10 + 70 = 80$ y $N_1N_2 = (8)(10) = 80$.

(*b*) Como $R_1 = 106$ y $R_2 = 65$, tenemos $R_1 + R_2 = 106 + 65 = 171$ y

$$\frac{N(N + 1)}{2} = \frac{(N_1 + N_2)(N_1 + N_2 + 1)}{2} = \frac{(18)(19)}{2} = 171$$

17.7. Resolver el Problema 17.5 usando el estadístico U para la muestra de la aleación II.

Solución

Para la muestra de la aleación II,

$$U = N_1N_2 + \frac{N_2(N_2 + 1)}{2} - R_2 = (8)(10) + \frac{(10)(11)}{2} - 65 = 70$$

así que

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{70 - 40}{11.25} = 2.67$$

Este valor de z es el *negativo* del z del Problema 17.5, y se usa la cola derecha de la distribución normal en vez de la izquierda. Ya que este valor de z también cae fuera de $-1.96 \leqslant z \leqslant 1.96$, la conclusión es la misma que en el Problema 17.5.

17.8. Un profesor de psicología tiene dos clases, una matinal de 9 estudiantes y otra vespertina de 12. En un examen común a todos ellos, las notas fueron las que recoge la Tabla 17.10. ¿Podemos concluir al nivel de significación 0.05 que la clase de la mañana es peor que la de la tarde?

Tabla 17.10

Clase matinal	73	87	79	75	82	66	95	75	70			
Clase vespertina	86	81	84	88	90	85	84	92	83	91	53	84

Solución

Paso 1. La Tabla 17.11 muestra la ordenación de notas y rangos. Nótese que el rango para las dos notas de 75 es $\frac{1}{2}(5 + 6) = 5.5$, mientras que para las tres de 84 es $\frac{1}{3}(11 + 12 + 13) = 12$.

Paso 2. Reescribiendo la Tabla 17.10 en términos de rangos obtenemos la Tabla 17.12.

Comprobación: $R_1 = 73$, $R_2 = 158$ y $N = N_1 + N_2 = 9 + 12 = 21$; luego $R_1 + R_2 = 73 + 158 = 231$ y

$$\frac{N(N+1)}{2} = \frac{(21)(22)}{2} = 231 = R_1 + R_2$$

Tabla 17.11

57	66	70	73	75	75	79	81	82	83	84	84	84	85	86	87	88	90	91	92	95
1	2	3	5.5	7		8	9	10	12		14		15	16	17	18	19	20	21	

Tabla 17.12

													Suma de rangos
Clase matinal	4	16	7	5.5	9	2	21	5.5	3				73
Clase vespertina	15	8	12	17	18	14	12	20	10	19	1	12	158

Paso 3.

$$U = N_1N_2 + \frac{N_1(N_1+1)}{2} - R_1 = (9)(12) + \frac{(9)(10)}{2} - 73 = 80$$

$$\mu_U = \frac{N_1N_2}{2} = \frac{(9)(12)}{2} = 54 \qquad \sigma_U^2 = \frac{N_1N_2(N_1+N_2+1)}{12} = \frac{(9)(12)(22)}{12} = 198$$

Por tanto,

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{80 - 54}{14.07} = 1.85$$

Puesto que deseamos contrastar la hipótesis H_0 de que la clase de la mañana es peor que la otra frente a la H_0 de que no hay diferencia al nivel 0.05, necesitamos un contraste unilateral. Con referencia a la Figura 17.2, que se aplica aquí, tenemos la regla de decisión:

Aceptar H_0 si $z \leq 1.645$.
Rechazar H_0 si $z > 1.645$.

Como el valor real es $z = 1.85 > 1.645$, rechazamos H_0 y concluimos que la clase matinal es peor al nivel 0.05. Esa conclusión no se mantiene, sin embargo, al nivel de significación 0.01 (véase Problema 17.33).

17.9. Hallar U para los datos de la Tabla 17.13, usando (*a*) la fórmula (2) de este capítulo y (*b*) el método de recuentos descrito en la Nota 4 de este capítulo.

Solución

(*a*) Ordenando los datos de ambas muestras en orden de magnitud creciente y asignándoles rangos

de 1 a 5, se llega a la Tabla 17.14. Sustituyendo los datos de la Tabla 17.13 por los rangos correspondientes se obtiene la Tabla 17.15, en la cual las sumas de rangos son $R_1 = 5$ y $R_2 = 10$. Como $N_1 = 2$ y $N_2 = 3$, el valor de U para la muestra 1 es

$$U = N_1N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 = (2)(3) + \frac{(2)(3)}{2} - 5 = 4$$

El valor de U para la muestra 2 se halla de forma similar y es $U = 2$.

Tabla 17.13

Muestra 1	22	10	
Muestra 2	17	25	14

Tabla 17.14

Datos	10	14	17	22	25
Rango	1	2	3	4	5

Tabla 17.15

				Suma de rangos
Muestra 1	4	1		5
Muestra 2	3	5	2	10

(*b*) Sustituyamos los valores muestrales en la Tabla 17.14 por I o II, según la muestra a la que el valor pertenezca. Entonces la primera línea de la Tabla 17.14 pasa a ser

Datos	I	II	II	I	I	II

De ahí vemos que

Número de valores de la muestra 1 que preceden al primero de la muestra 2 = 1
Número de valores de la muestra 1 que preceden al segundo de la muestra 2 = 1
Número de valores de la muestra 1 que preceden al tercero de la muestra 2 = 2
Total = 4

Luego el valor de U correspondiente a la muestra 1 es 4.

Análogamente se tiene

Número de valores de la muestra 2 que preceden al primero de la muestra 1 = 0
Número de valores de la muestra 2 que preceden al segundo de la muestra 1 = 2
Total = 2

Luego el valor de U para la muestra 2 es 2.

Nótese que como $N_1 = 2$ y $N_2 = 3$, estos valores satisfacen $U_1 + U_2 = N_1N_2$; es decir, $4 + 2 = (2)(3) = 6$.

17.10. Se toman dos muestras sin reposición de una población que consiste en los valores 7, 12 y 15: la primera muestra consta de un solo valor y la segunda de dos valores. [Entre ambas muestras cubren toda la población.]

(*a*) Hallar la distribución de muestreo de U y su gráfico.
(*b*) Hallar la media y la varianza de esa distribución.
(*c*) Comprobar los resultados de la parte (*b*) mediante las fórmulas (3) de este capítulo.

Solución

(*a*) Escogemos muestreo sin reposición para evitar coincidencias, que ocurrirían si, por ejemplo, el valor 12 apareciese en ambas muestras.

Hay $3 \cdot 2 = 6$ posibilidades para escoger las muestras, como indica la Tabla 17.16. Debemos observar que podríamos usar los rangos 1, 2 y 3 en vez de 7, 12 y 15. El valor de U en la Tabla 17.16, es el hallado para la muestra 1, pero si se usara el U para la muestra 2, la distribución sería la misma.

Tabla 17.16

Muestra 1	Muestra 2	U
7	12 15	2
7	15 12	2
12	7 15	1
12	15 7	1
15	7 12	0
15	12 7	0

El gráfico de esta distribución aparece en la Figura 17.4, donde f es la frecuencia. La distribución de probabilidad de U también puede representarse; en este caso $\Pr\{U = 0\} = \Pr\{U = 1\} = \Pr\{U = 2\} = \frac{1}{3}$. El gráfico pedido es el mismo que el de la Figura 17.4, pero con las ordenadas 1 y 2 sustituidas por $\frac{1}{6}$ y $\frac{1}{3}$, respectivamente.

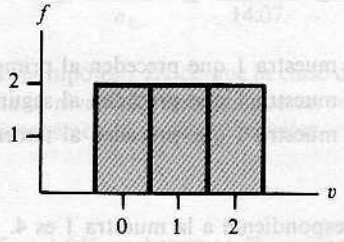


Figura 17.4.

(*b*) La media y la varianza halladas a partir de la Tabla 17.15 vienen dadas por

$$\mu_U = \frac{2 + 2 + 1 + 1 + 0 + 0}{6} = 1$$

$$\sigma_U^2 = \frac{(2 - 1)^2 + (2 - 1)^2 + (1 - 1)^2 + (1 - 1)^2 + (0 - 1)^2 + (0 - 1)^2}{6} = \frac{2}{3}$$

(*c*) Por las fórmulas (3),

$$\mu_U = \frac{N_1 N_2}{2} = \frac{(1)(12)}{2} = 1$$

$$\sigma_U^2 = \frac{N_1 N_2 (N_1 + N_2 + 1)}{12} = \frac{(1)(2)(1 + 2 + 1)}{12} = \frac{2}{3}$$

en buen acuerdo con la parte (*a*).

17.11. (*a*) Hallar la distribución muestral de U en el Problema 17.9 y representarla gráficamente.
(*b*) Representar la correspondiente distribución de probabilidad de U.
(*c*) Obtener la media y la varianza de U directamente de los resultados de la parte (*a*).
(*d*) Verificar la parte (*c*) usando las fórmulas (3) de este capítulo.

Solución

(*a*) En este caso hay $5 \cdot 4 \cdot 3 \cdot 2 = 120$ posibilidades para escoger valores en las dos muestras y el método del Problema 17.9 es demasiado laborioso. Para simplificar el proceso, vamos a concentrarnos en la muestra menor (de tamaño $N_1 = 2$) y las posibles sumas de rangos, R. La suma de los rangos para la muestra 1 es *mínima* cuando la muestra consiste en los dos números de rango más bajo (1, 2); entonces $R_1 = 1 + 2 = 3$. Análogamente, es *máxima* cuando la muestra 1 consta de los números de rango más alto (4, 5); entonces $R_1 = 4 + 5 = 9$. Luego R_1 varía de 3 a 9.

La columna 1 de la Tabla 17.17 da esos valores de R_1 (desde 3 hasta 9), y la columna 2 da los correspondientes valores en la muestra 1 cuya suma es R_1. La columna 3 da la frecuencia (o número) de muestras con suma R_1; por ejemplo, hay $f = 2$ muestras con $R_1 = 5$. Como $N_1 = 2$ $N_2 = 3$, tenemos

$$U = N_1 N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 = (2)(3) + \frac{(2)(3)}{2} - R_1 = 9 - R_1$$

Tabla 17.17

R_1	Valores de la muestra 1	f	U	$\Pr\{U = R_1\}$
3	(1, 2)	1	6	0.1
4	(1, 3)	1	5	0.1
5	(1, 4), (2, 3)	2	4	0.2
6	(1, 5), (2, 4)	2	3	0.2
7	(2, 5), (3, 4)	2	2	0.2
8	(3, 5)	1	1	0.1
9	(4, 5)	1	0	0.1

Hallamos los correspondientes valores de U en la columna 4; nótese que cuando R_1 varía de 3 a 9, U varía de 6 a 0. La distribución muestral viene dada por las columnas 3 y 4, y su gráfico por la Figura 17.5

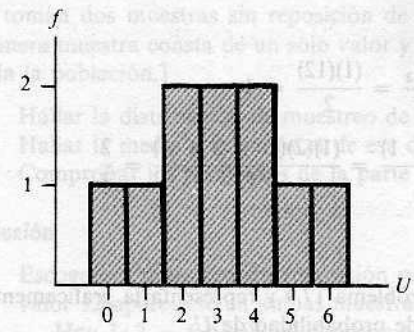


Figura 17.5.

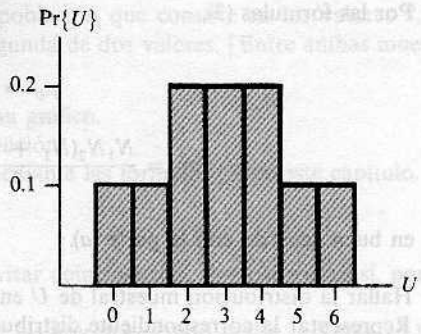


Figura 17.6.

(*b*) La probabilidad de que $U = R_1$ (es decir, $\Pr\{U = R_1\}$) aparece en la columna 5 de la Tabla 17.17 y se obtiene hallando la frecuencia relativa, cociente de cada frecuencia f por la suma de todas las frecuencias, o sea 10; así, $\Pr\{U = 5\} = \frac{2}{10} = 0.2$. El gráfico de la distribución de probabilidad se muestra en la Figura 17.6.

(*c*) De las columnas 3 y 4 de la Tabla 17.17, se deduce

$$\mu_U = \bar{U} = \frac{\sum fU}{\sum f} = \frac{(1)(6)+(1)(5)+(2)(4)+(2)(3)+(2)(2)+(1)(1)+(1)(0)}{1+1+2+2+2+1+1} = 3$$

$$\sigma_U^2 = \frac{\sum f(U-\bar{U})^2}{\sum f}$$

$$= \frac{(1)(6-3)^2+(1)(5-3)^2+(2)(4-3)^2+(2)(3-3)^2+(2)(2-3)^2+(1)(1-3)^2+(1)(0-3)^2}{10} = 3$$

Otro método

$$\sigma_U^2 = \overline{U^2} - \bar{U}^2 = \frac{(1)(6)^2+(1)(5)^2+(2)(4)^2+(2)(3)^2+(2)(2)^2+(1)(1)^2+(1)(0)^2}{10} - (3)^2 = 3$$

(*d*) Por las fórmulas (3), usando $N_1 = 2$ y $N_2 = 3$, tenemos

$$\mu_U = \frac{N_1N_2}{2} = \frac{(2)(3)}{2} = 3 \qquad \sigma_U^2 = \frac{N_1N_2(N_1+N_2+1)}{12} = \frac{(2)(3)(6)}{12} = 3$$

17.12. Si N números en un conjunto se enumeran con rangos de 1 a N, probar que la suma de rangos es $[N(N+1)]/2$.

Solución

Si llamamos R a la suma de rangos, tenemos

$$R = 1 + 2 + 3 + \cdots + (N-1) + N \tag{16}$$

$$R = N + (N-1) + (N-2) + \cdots + 2 + 1 \tag{17}$$

donde la suma en (17) se obtiene escribiendo la de (16) hacia atrás. Sumando las ecuaciones (16) y (17) resulta

$$2R = (N + 1) + (N + 1) + (N + 1) + \cdots + (N + 1) + (N + 1) = N(N + 1)$$

ya que $(N + 1)$ aparece N veces en la suma; así pues, $R = [N(N + 1)]/2$. Esto se puede obtener también recurriendo al álgebra elemental de progresiones aritméticas.

17.13. Si R_1 y R_2 son las respectivas sumas de rangos para las muestras 1 y 2 en el U-test, probar que $R_1 + R_2 = [N(N + 1)]/2$.

Solución

Suponemos que no hay coincidencias en los datos muestrales. Entonces R_1 ha de ser la suma de los rangos (números) del conjunto 1, 2, 3, .., N y R_2, la suma de los restantes rangos. Así que la suma $R_1 + R_2$ debe ser la suma de todos los rangos del conjunto; es decir, $R_1 + R_2 = 1 + 2 + 3 + \cdots + N = [N(N + 1)]/2$ por el Problema 17.12.

EL H-TEST DE KRUSKAL-WALLIS

17.14. Una empresa desea comprar una de las cinco máquinas distintas A, B, C, D y E. En un experimento diseñado para saber si hay diferencia en la eficacia de tales máquinas, cinco operarios expertos trabajaron cada uno con las máquinas un mismo tiempo en cada una. Los resultados se recogen en la Tabla 17.18, en número de unidades producidas. Contrastar la hipótesis de que no hay diferencia entre ellas al nivel de significación (a) 0.05 y (b) 0.01.

Tabla 17.18

A	68	72	77	42	53
B	72	53	63	53	48
C	60	82	64	75	72
D	48	61	57	64	50
E	64	65	70	68	53

Tabla 17.19

						Suma de rangos
A	17.5	21	24	1	6.5	70
B	21	6.5	12	6.5	2.5	48.5
C	10	25	14	23	21	93
D	2.5	11	9	14	4	40.5
E	14	16	19	17.5	6.5	73

Solución

Como hay 5 muestras (A, B, C, D y E), $k = 5$. Y como cada muestra consta de 5 valores, tenemos $N_1 = N_2 = N_3 = N_4 = N_5 = 5$ y $N = N_1 + N_2 + N_3 + N_4 + N_5 = 25$. Ordenando todos los valores en orden creciente de magnitud y asignando rangos apropiados a las coincidencias, cambiamos la Tabla 17.18 por la 17.19, cuya columna de la derecha da la suma de rangos. Vemos de la Tabla 17.19 que $R_1 = 70$, $R_2 = 48.5$, $R_3 = 93$, $R_4 = 40.5$ y $R_5 = 73$. Luego

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{N_j} - 3(N + 1)$$

$$= \frac{12}{(25)(26)} \left[\frac{(70)^2}{5} + + \frac{(48.5)^2}{5} + \frac{(93)^2}{5} + \frac{(40.5)^2}{5} + \frac{(73)^2}{5} \right] - 3(26) = 6.44$$

Para $k - 1 = 4$ grados de libertad al nivel de significación 0.05, por el Apéndice IV sabemos que $\chi^2_{.95} = 9.49$. Puesto que $6.44 < 9.49$, no podemos rechazar la hipótesis de igualdad entre la máquinas al nivel 0.05 y, por tanto, tampoco al 0.01. En otras palabras, podemos aceptar la hipótesis de que no hay diferencia entre las máquinas a ambos niveles (o reservar nuestra opinión).

Nótese que ya hemos resuelto este problema mediante análisis de varianza (véase Prob. 16.8) y llegamos a la misma conclusión.

17.15. Repetir el Problema 17.14 haciendo corrección por coincidencias.

Solución

La Tabla 17.20 da el número de coincidencias correspondientes a cada una de las observaciones con coincidencias. Por ejemplo, 48 aparece dos veces, de donde $T = 2$, y 53 aparece cuatro veces, luego $T = 4$. Calculando $T^3 - T$ para cada valor de T y sumando, encontramos que $\sum(T^3 - T) = 6 + 60 + 24 + 6 + 24 = 120$, como indica la Tabla 17.20. Entonces, como $N = 25$, el factor de corrección es

$$1 - \frac{\sum(T^3 - T)}{N^3 - N} = 1 - \frac{120}{(25)^3 - 25} = 0.9923$$

y el valor corregido de H es

$$H_c = \frac{6.44}{0.9923} = 6.49$$

Esta corrección no es suficiente para cambiar la decisión del Problema 17.14.

Tabla 17.20

Observación	48	53	64	68	72	
Número de coincidencias (T)	2	4	3	2	3	
$T^3 - T$	6	60	24	6	24	$\sum(T^3 - T) = 120$

17.16. Se toman al azar tres muestras de una población. Al ordenar los datos de acuerdo con el rango se obtiene la Tabla 17.21. Determinar si hay diferencia entre las muestras al nivel de significación (*a*) 0.05 y (*b*) 0.01.

Solución

Aquí $k = 3$, $N_1 = 4$, $N_2 = 3$, $N_3 = 5$, $N = N_1 + N_2 + N_3 = 12$, $R_1 = 7 + 4 + 6 + 10 = 27$, $R_2 = 11 + 9 + 12 = 32$ y $R_3 = 5 + 1 + 3 + 8 + 19$. Por tanto,

$$H = \frac{12}{N(N+1)}\sum_{j=1}^{k}\frac{R_j^2}{N_j} - 3(N+1) = \frac{12}{(12)(13)}\left[\frac{(27)^2}{4} + \frac{(32)^2}{3} + \frac{(19)^2}{5}\right] - 3(13) = 6.83$$

(*a*) Para $k - 1 = 3 - 1 = 2$ grados de libertad, $\chi^2_{.95} = 5.99$. Luego, como $6.83 > 5.99$, concluimos que hay diferencia significativa entre las muestras al nivel 0.05.

(*b*) Para 2 grados de libertad, $\chi^2_{.95} = 9.21$. Luego, como $6.83 < 9.21$ no podemos concluir que haya diferencia al nivel 0.01.

Tabla 17.21

Muestra 1	7 4 6 10
Muestra 2	11 9 12
Muestra 3	5 1 3 8 2

EL TEST DE LAS RACHAS PARA EL CARACTER ALEATORIO

17.17. En 30 tiradas de una moneda se ha obtenido la siguiente secuencia de caras (H) y cruces (T):

H T T H T H H H T H H T T H T
H T H H T H T T H T H H T H T

(*a*) Hallar el número de rachas, V.
(*b*) Decidir al nivel de significación 0.05 si la secuencia es aleatoria.

Solución

(*a*) Separando las rachas con barras verticales, vemos en

H | T T | H | T | H H H | T | H H | T T | H | T |
H | T | H H | T | H | T T | H | T | H H | T | H | T |

que el número de rachas es $V = 22$.

(*b*) Hay $N_1 = 16$ caras y $N_2 = 14$ cruces en la muestra dada, y por la parte (*a*) sabemos que el número de rachas es $V = 22$. Luego de (13) se deduce

$$\mu_V = \frac{2(16)(14)}{16 + 14} + 1 = 15.93 \qquad \sigma_V^2 = \frac{2(16)(14)[2(16)(14) - 16 - 14]}{(16 + 14)^2(16 + 14 - 1)} = 7.175$$

o sea $\sigma_V = 2.679$. El z correspondiente a $V = 22$ es, en consecuencia,

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{22 - 15.93}{2.679} = 2.27$$

Ahora bien, para un contraste bilateral al nivel de significación 0.05, aceptaríamos la hipótesis H_0 de aleatoriedad si $-1.96 \leq z \leq 1.96$ y la rechazaríamos en caso contrario (véase Fig. 17.7). Como el valor calculado de z es $2.27 > 1.96$, concluimos que los lanzamientos no son aleatorios al nivel 0.05. El test nos ha hecho ver que hay *demasiadas* rachas, sugiriendo un *esquema cíclico*.

Si se hace corrección por continuidad, el z anterior pasa a ser

$$z = \frac{(22 - 0.5) - 15.93}{2.679} = 2.08$$

y se obtiene la misma conclusión.

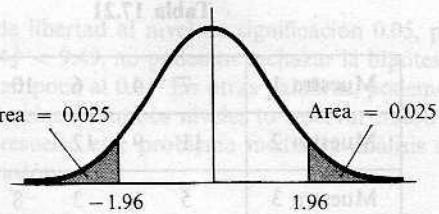


Figura 17.7.

17.18. Una muestra de 48 piezas producidas por una máquina ha dado la siguiente secuencia de piezas correctas (G) y defectuosas (D):

G G G G G G D D G G G G G G G G
G G D D D D G G G G G G D G G G
G G G G G G D D G G G G G D G G

Contrastar la aleatoriedad de esa secuencia al nivel de significación 0.05.

Solución

Los números de Des y Ges son $N_1 = 10$ y $N_2 = 38$, respectivamente, y el número de rachas es $V = 11$. Luego la media y la varianza vienen dadas por

$$\mu_V = \frac{2(10)(38)}{10 + 38} + 1 = 16.83 \qquad \sigma_V^2 = \frac{2(10)(38)[2(10)(38) - 10 - 38]}{(10 + 38)^2(10 + 38 - 1)} = 4.997$$

así que $\sigma_V = 2.235$.

Para un contraste bilateral al nivel de significación 0.05, aceptaríamos la hipótesis H_0 de aleatoriedad si $-1.96 \leqslant z \leqslant 1.96$ (véase Fig. 17.7) y la rechazaríamos en caso contrario. Como el z correspondiente a $V = 11$ es

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{11 - 16.83}{2.235} = -2.16$$

y $-2.61 < -1.96$, podemos rechazar H_0 al nivel 0.05.

El test pone de manifiesto que hay *demasiado pocas* rachas, indicando un hacinamiento de piezas defectuosas. En otras palabras, parece haber un *esquema de tendencia* en la producción de piezas defectuosas. Debe examinarse con más profundidad el proceso de fabricación.

17.19. (*a*) Formar todas las posibles secuencias consistentes en tres *aes* y dos *bes*, y dar los números de rachas V para cada una de ellas.
(*b*) Obtener la distribución muestral de V y su gráfico.
(*c*) Obtener la distribución de probabilidad de V y su gráfico.

Solución

(*a*) El número de posibles secuencias de ese tipo es

$$\binom{5}{2} = \frac{5!}{2!3!} = 10$$

Estas secuencias se recogen en la Tabla 17.22, junto con el número de rachas de cada una.

(*b*) La distribución muestral de V viene dada en la Tabla 17.23 (deducida de la Tabla 17.21), donde V denota el número de rachas y f la frecuencia. Por ejemplo, la Tabla 17.23 dice que hay 1 cinco, 4 cuatros, etc. El gráfico correspondiente se puede ver en la Figura 17.8.

Tabla 17.22

Secuencia					Rachas (V)
a	*a*	*a*	*b*	*b*	2
a	*a*	*b*	*a*	*b*	4
a	*a*	*b*	*b*	*a*	3
a	*b*	*a*	*b*	*a*	5
a	*b*	*b*	*a*	*a*	3
a	*b*	*a*	*a*	*b*	4
b	*b*	*a*	*a*	*a*	2
b	*a*	*b*	*a*	*a*	4
b	*a*	*a*	*a*	*b*	3
b	*a*	*a*	*b*	*a*	4

Tabla 17.23

V	f
2	2
3	3
4	4
5	1

(*c*) La distribución de probabilidad de V, dibujada en la Figura 17.9, se obtiene de la Tabla 17.23 dividiendo cada frecuencia por la frecuencia total $2 + 33 + 4 + 1 = 10$. Por ejemplo, $\Pr\{V = 5\} = \frac{1}{10} = 0.1$.

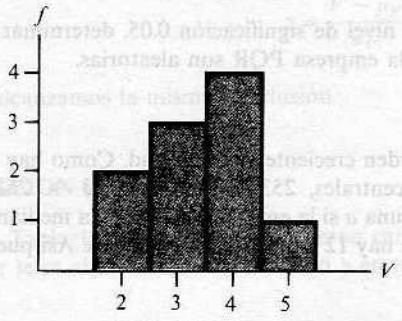


Figura 17.8.

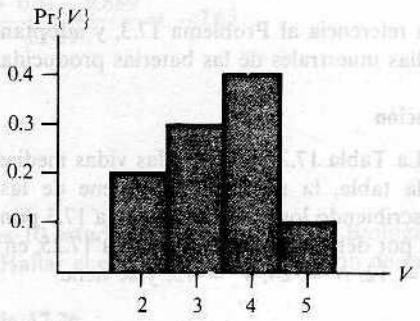


Fig. 17.9.

17.20. Hallar (*a*) la media y (*b*) la varianza del número de rachas en el Problema 17.19 directamente de los resultados allí obtenidos.

Solución

(*a*) De la Tabla 17.22 tenemos

$$\mu_V = \frac{2 + 4 + 3 + 5 + 3 + 4 + 2 + 4 + 3 + 4}{10} = \frac{17}{5}$$

Otro método

De la Tabla 17.22 el método de datos agrupados da

$$\mu_V = \frac{\sum fV}{\sum f} = \frac{(2)(2) + (3)(3) + (4)(4) + (1)(5)}{2 + 3 + 4 + 1} = \frac{17}{5}$$

(*b*) Usando el método de datos agrupados para calcular la varianza, se sigue de la Tabla 17.23 que

$$\sigma_V^2 = \frac{\sum f(V-\bar{V})^2}{\sum f} = \frac{1}{10}\left[(2)\left(2-\frac{17}{5}\right)^2 + (3)\left(3-\frac{17}{5}\right)^2 + (4)\left(4-\frac{17}{5}\right)^2 + (1)\left(5-\frac{17}{5}\right)^2\right] = \frac{21}{25}$$

Otro método

Como en el Capítulo 3, la varianza viene dada por

$$\sigma_V^2 = \overline{V^2} - \bar{V}^2 = \frac{(2)(2)^2 + (3)(3)^2 + (4)(4)^2 + (1)(5)^2}{10} - \left(\frac{17}{5}\right)^2 = \frac{21}{25}$$

17.21. Resolver el Problema 17.20 con las fórmulas (13) de este capítulo.

Solución

Puesto que hay tres *aes* y dos *bes*, se tiene $N_1 = 3$ y $N_2 = 2$. Así pues

(*a*) $$\mu_V = \frac{2N_1N_2}{N_1+N_2} + 1 = \frac{2(3)(2)}{3+2} + 1 = \frac{17}{5}$$

(*b*) $$\sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1+N_2)^2(N_1+N_2-1)} = \frac{2(3)(2)[2(3)(2)-3-2]}{(3+2)^2(3+2-1)} = \frac{21}{25}$$

OTRAS APLICACIONES DEL TEST DE LAS RACHAS

17.22. Con referencia al Problema 17.3, y adoptando un nivel de significación 0.05, determinar si las vidas medias muestrales de las baterías producidas por la empresa PQR son aleatorias.

Solución

La Tabla 17.24 presenta las vidas medias en orden creciente de magnitud. Como hay 24 entradas en la tabla, la mediana se obtiene de las dos centrales, 253 y 262, es $\frac{1}{2}(253 + 262) = 257.5$. Reescribiendo los datos de la Tabla 17.3 poniendo una *a* si la entrada está sobre la mediana y una *b* si está por debajo, se llega a la Tabla 17.25, en la que hay 12 *aes*, 12 *bes* y 15 rachas. Así pues, $N_1 = 12$, $N_2 = 12$, $N = 24$, $V = 15$, y se tiene

$$\mu_V = \frac{2N_1N_2}{N_1+N_2} + 1 = \frac{2(12)(12)}{12+12} + 1 = 13 \qquad \sigma_V^2 = \frac{2(12)(12)(264)}{(24)^2(23)} = 5.739$$

luego $$z = \frac{V-\mu_V}{\sigma_V} = \frac{15-13}{2.396} = 0.835$$

Con un contraste de dos colas al nivel de significación 0.05, aceptaríamos la hipótesis de aleatoriedad si $-1.96 \leq z \leq 1.96$. Como 0.835 cae dentro de ese intervalo, concluimos que la muestra es aleatoria.

Tabla 17.24

198	211	216	219	224	225	230	236
243	252	253	253	262	264	268	271
272	275	282	284	288	291	294	295

Tabla 17.25

a	*b*	*b*	*a*	*a*	*b*	*a*	*b*
b	*b*	*a*	*a*	*b*	*a*	*b*	*b*
a	*a*	*b*	*b*	*a*	*b*	*a*	*a*

17.23. Resolver el Problema 17.5 aplicando el test de las rachas para decidir sobre la aleatoriedad.

Solución

La ordenación de todos los valores de ambas muestras aparece en la línea 1 de la Tabla 17.8. Usando los símbolos respectivos a y b para los datos de las muestras I y II, se convierte en

$$b \quad b \quad b \quad b \quad b \quad b \quad b \quad b \quad a \quad a \quad a \quad a \quad a \quad b \quad b \quad a \quad a \quad a$$

Como hay 4 rachas, tenemos $V = 4$, $N_1 = 8$ y $N_2 = 10$. Entonces

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1 = \frac{2(8)(10)}{18} + 1 = 9.889$$

$$\sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)} = \frac{2(8)(10)(142)}{(18)^2(17)} = 4.125$$

así que

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{4 - 9.889}{2.031} = -2.90$$

Si H_0 es la hipótesis de que no hay diferencia entre las aleaciones, esa es también la hipótesis de que la secuencia anterior es aleatoria. La aceptaríamos si $-1.96 \leqslant z \leqslant 1.96$ y la rechazaríamos en caso contrario. Puesto que $z = -2.90$ está fuera de ese intervalo, rechazamos H_0 y llegamos a la misma conclusión que en el Problema 17.5.

Nótese que si se hace corrección por continuidad,

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{(4 + 0.5) - 9.889}{2.031} = -2.65$$

y alcanzamos la misma conclusión.

CORRELACION DE RANGO

17.24. La Tabla 17.26 muestra cómo fueron calificados 10 estudiantes de un curso de Biología, ordenados por letra alfabética, en laboratorio y en teoría. Hallar el coeficiente de correlación de rango.

Tabla 17.26

Laboratorio	8	3	9	2	7	10	4	6	1	5
Teoría	9	5	10	1	8	7	3	4	2	6

Solución

La diferencia en rangos, D, en laboratorio y en teoría, para cada estudiante se da en la Tabla 17.27, que da también D^2 y $\sum D^2$. Luego

$$r_S = 1 - \frac{6\sum D^2}{N(N^2 - 1)} = 1 - \frac{6(24)}{10(10^2 - 1)} = 0.8545$$

indicando que hay una marcada relación entre las calificaciones de laboratorio y de teoría.

Tabla 17.27

Diferencia de rangos (D)	-1	-2	-1	1	-1	3	1	2	-1	-1	
D^2	1	4	1	1	1	9	1	4	1	1	$\sum D^2 = 24$

17.25. En la Tabla 17.28 aparecen las alturas de una muestra de 12 padres y sus hijos mayores. Hallar el coeficiente de correlación de rango.

Tabla 17.28

Altura del padre (pulgadas)	65	63	67	64	68	62	70	66	68	67	69	71
Altura del hijo (pulgadas)	68	66	68	65	69	66	68	65	71	67	68	70

Solución

Ordenados de menor a mayor, las alturas de los padres son

$$62 \quad 63 \quad 64 \quad 65 \quad 66 \quad 67 \quad 67 \quad 68 \quad 68 \quad 69 \quad 71 \tag{18}$$

Como el sexto y el séptimo lugares representan la misma altura (67 in), asignamos a esos lugares un rango medio $\frac{1}{2}(6 + 7) = 6.5$. Análogamente, al octavo y noveno lugar se les asigna rango $\frac{1}{2}(8 + 9) = 8.5$. Así que las alturas de los padres tienen asignados los rangos

$$1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6.5 \quad 6.5 \quad 8.5 \quad 8.5 \quad 10 \quad 11 \quad 12 \tag{19}$$

De la misma manera, ordenadas de menor a mayor, las alturas de los hijos son

$$65 \quad 65 \quad 66 \quad 66 \quad 67 \quad 68 \quad 68 \quad 68 \quad 68 \quad 69 \quad 70 \quad 71 \tag{20}$$

y como los lugares del sexto al noveno tienen la misma altura anotada (68 in), les asignamos el rango medio $\frac{1}{4}(6 + 7 + 8 + 9) = 7.5$. Así pues, a las alturas de los hijos se les asignan los rangos

$$1.5 \quad 1.5 \quad 3.5 \quad 3.5 \quad 5 \quad 7.5 \quad 7.5 \quad 7.5 \quad 7.5 \quad 10 \quad 11 \quad 12 \tag{21}$$

Usando las correspondencia(18) y (19), y (20) y (21), podemos sustituir la Tabla 17.28 por la Tabla 17.29. La Tabla 17.30 da las diferencias en rangos, D, y los cálculos de D^2 y $\sum D^2$, de donde

$$r_S = 1 - \frac{6\sum D^2}{N(N^2 - 1)} = 1 - \frac{6(72.50)}{12(12^2 - 1)} = 0.7465$$

Este resultado está en buen acuerdo con el coeficiente de correlación obtenido por otros métodos (véanse Probs. 14.9, 14.14, 14.16 y 14.23).

Tabla 17.29

Rango del padre	4	2	6.5	3	8.5	1	11	5	8.5	6.5	10	12
Rango del hijo	7.5	3.5	7.5	1.5	10	3.5	7.5	1.5	12	5	7.5	11

Tabla 17.30

D	−3.5	−1.5	−1.0	1.5	−1.5	−2.5	3.5	3.5	−3.5	1.5	2.5	1.0	
D^2	12.25	2.25	1.00	2.25	2.25	6.25	12.25	12.25	12.25	2.25	6.25	1.00	$\sum D^2 = 72.50$

PROBLEMAS SUPLEMENTARIOS

EL TEST DE LOS SIGNOS

17.26. Una empresa afirma que si se añade su producto en el depósito de gasolina de un automóvil, las millas recorridas por galón aumentan. Para contrastar tal afirmación, se toman 15 automóviles distintos y se miden las millas por galón recorridas con y sin ese producto, con los resultados de la Tabla 17.31. Suponiendo que las condiciones de conducción son las mismas, determinar si hay diferencia debida a ese producto, al nivel de significación (*a*) 0.05 y (*b*) 0.01.

Tabla 17.31

Con aditivo	Sin aditivo
34.7	31.4
28.3	27.2
19.6	20.4
25.1	24.6
15.7	14.9
24.5	22.3
28.7	26.8
23.5	24.1
27.7	26.2
32.1	31.4
29.6	28.8
22.4	23.1
25.7	24.0
28.1	27.3
24.3	22.9

17.27. ¿Se puede concluir al nivel de significación 0.05 que las millas recorridas por galón en el Problema 17.26 aumentan al añadir ese producto?

17.28. Un club de adelgazamiento anuncia que ha preparado un programa especial que producirá pérdidas de peso de al menos un 6% en un mes, si se sigue rigurosamente. Para comprobar esa afirmación, 36 adultos siguen el programa. De ellos, 25 perdieron lo anunciado, 6 engordaron y el resto no sufrió cambio esencialmente. Determinar al nive de significación 0.05 si el programa era eficaz.

17.29. Un director de personal sostiene que con un curso especial para el personal de la sección de ventas, una empresa aumentará sus ventas. Para comprobarlo, se impartió el curso a 24 personas, de las que 16 vieron las ventas aumentadas, 6 las vieron decrecer y las de 2 quedaron sin cambio. Contrastar al nivel de significación 0.05 la hipótesis de que el curso hizo crecer las ventas de la empresa.

17.30. Una empresa fabricante de refrescos hizo degustaciones en 27 localidades del país para saber hacia qué refresco de cola, *A* o *B*, se inclinaban la preferencias del público. En 8 localidades se prefirió el *A*, en 17 el *B* y en las restantes no hubo preferencia por ninguno sobre el otro. ¿Se puede concluir que, al nivel de significación 0.05, el *B* es el preferido?

17.31. Las tensiones de ruptura de una muestra aleatoria de 25 sogas de un cierto fabricante se dan en la Tabla 17.32. Contrastar con esa muestra, al nivel de significación 0.05, la

afirmación del fabricante de que tal tensión es (*a*) 25, (*b*) 30, (*c*) 35 y (*d*) 40.

Tabla 17.32

41	28	35	38	23
37	32	24	46	30
25	36	22	41	37
43	27	34	27	36
42	33	28	31	24

17.32. Indicar cómo se pueden obtener los límites de confianza 95% para los datos del Problema 17.4.

17.33. Plantear y resolver un problema que utilice el test de los signos.

EL *U*-TEST DE MANN-WHITNEY

17.34. Los profesores *A* y *B* dan cursos de química en la Universidad XYZ. En un examen común, sus estudiantes recibieron las calificaciones que aparecen en la Tabla 17.33. Contrastar al nivel de significación 0.05 la hipótesis de que no hay diferencia entre las calificaciones de ambos profesores.

Tabla 17.33

A	*B*
88	72
75	65
92	84
71	53
63	76
84	80
55	51
64	60
82	57
96	85
	94
	87
	73
	61

17.35. Refiriéndonos al Problema 17.34, ¿puede concluirse al nivel de significación 0.01, que las notas de la clase matinal son peores que las de la vespertina?

17.36. Un agricultor quiere saber si hay diferencia entre las producciones de dos variedades de trigo, I y II. La Tabla 17.34 indica las producciones de trigo por unidad de área con ambas variedades. ¿Puede concluirse que existe diferencia al nivel de significación (*a*) 0.05 y (*b*) 0.01?

Tabla 17.34

Trigo I	Trigo II
15.9	16.4
15.3	16.8
16.4	17.1
14.9	16.9
15.3	18.0
16.0	15.6
14.6	18.1
15.3	17.2
14.5	15.4
16.6	
16.0	

17.37. Puede el agricultor del Problema 17.36 concluir al nivel de significación 0.05 que la variedad II da mayor producción que la I?

17.38. Se desea averiguar si hay diferencia entre dos clases de gasolina, *A* y *B*. La Tabla 17.35 da las distancias recorridas por galón para cada clase. ¿Se puede concluir al nivel de significación 0.05 que (*a*) hay diferencia entre ambas y (*b*) la *B* es mejor que la *A*?

Tabla 17.35

A	*B*
30.4	33.5
28.7	29.8
29.2	30.1
32.5	31.4
31.7	33.8
29.5	30.9
30.8	31.3
31.1	29.6
30.7	32.8
31.8	33.0

17.39. ¿Puede usarse el U-test para determinar si hay diferencia entre las máquinas I y II de la Tabla 17.1? Explicar la respuesta.

17.40. Proponer y resolver un problema que utilice el U-test.

17.41. Hallar U para los datos de la Tabla 17.36, usando (*a*) el método de la fórmula y (*b*) el método de recuento.

Tabla 17.36

Muestra 1	15	25
Muestra 2	20	32

17.42. Resolver el Problema 17.41 para los datos de la Tabla 17.37.

Tabla 17.37

Muestra 1	40	27	30	56
Muestra 2	10	35		

17.43. Una población consta de los valores 2, 5, 9 y 12. Se toman dos muestras, la primera de uno de esos valores y la segunda de los tres restantes.

(*a*) Obtener la distribución muestral de U y su gráfico.

(*b*) Hallar la media y la varianza de esa distribución, directamente y por la fórmula.

17.44. Probar que $U_1 + U_2 = N_1N_2$.

17.45. Probar que $R_1 + R_2 = [N(N + 1)]/2$ para el caso en que el número de coincidencias es (*a*), 1, (*b*) 2 y (*c*) cualquier número.

17.46. Si $N_1 = 14$, $N_2 = 12$ y $R_1 = 105$, hallar (*a*) R_2, (*b*) U_1 y (*c*) U_2.

17.47. Si $N_1 = 10$, $N_2 = 16$, y $U_2 = 60$, hallar (*a*) R_1, (*b*) R_2 y (*c*) U_1.

17.48. ¿Cuál es el mayor número de valores N_1, N_2, R_1, R_2, U_1 y U_2 que puede calcularse a partir de los restantes? Demostrar la respuesta.

EL H-TEST DE KRUSKAL-WALLIS

17.49. Se realiza un experimento para determinar las producciones de cinco variedades de trigo: A, B, C, D y E. Se asignan a cada variedad cuatro parcelas. La producción (en bushels por acre) se indica en la Tabla 17.38. Suponiendo que las parcelas tienen igual fertilidad y que las variedades se asignan a las parcelas de modo aleatorio, determinar si hay diferencia significativa entre las producciones al nivel de significación (*a*) 0.05 y (*b*) 0.01.

Tabla 17.38

A	20	12	15	19
B	17	14	12	15
C	23	16	18	14
D	15	17	20	12
E	21	14	17	18

17.50. Las vidas medias de cuatro tipos de llantas A, B, C y D, vienen dadas en la Tabla 17.39 (en miles de millas de rodaje); cada tipo se ha probado con seis automóviles similares asignados a las llantas al azar. Determinar si hay diferencia significativa entre las llantas al nivel de significación (*a*) 0.05 y (*b*) 0.01.

Tabla 17.39

A	33	38	36	40	31	35
B	32	40	42	38	30	34
C	31	37	35	33	34	30
D	27	33	32	29	31	28

17.51. Un pedagogo quiere probar tres métodos de enseñanza: I, II y III. Para ello, escoge al azar tres grupos de 5 estudiantes cada uno y les aplica métodos diferentes. Se da el mismo examen a todos ellos y se producen las notas que figuran en la Tabla 17.40. Determinar si hay diferencia entre esos métodos de enseñanza al nivel de significación (*a*) 0.05 y (*b*) 0.01.

Tabla 17.40

Método I	78	62	71	58	73
Método II	76	85	77	90	87
Método III	74	79	60	75	80

17.52. En la Tabla 17.41 se ven las notas de un alumno. Al nivel de significación (*a*) 0.05 y (*b*) 0.01 decidir si hay diferencia entre las notas en las diversas materias.

Tabla 17.41

Matemáticas	72	80	83	75	
Ciencias	81	74	77		
Inglés	88	82	90	87	80
Economía	74	71	77	70	

17.53. Usando el *H* test, resolver (*a*) Problema 16.9, (*b*) Problema 16.21 y (*c*) Problema 16.25.

17.54. Usando el *H* test, resolver (*a*) Problema 16.23, (*b*) Problema 16.24 y (*c*) Problema 16.25.

EL TEST DE LAS RACHAS PARA EL CARACTER ALEATORIO

17.55. Determinar el número de rachas, *V*, para cada una de estas secuencias:

(*a*) A B A B B A A A B B A B

(*b*) H H T H H H T T T T H H T H H T H T

17.56. Se ha preguntado a 25 individuos si les gusta (Y) o no (N) un cierto producto, y se ha obtenido la secuencia de respuestas siguiente:

Y Y N N N N Y Y Y N Y N N
Y N N N N N Y Y Y Y N N

(*a*) Hallar el número de rachas.

(*b*) Decidir al nivel de significación 0.05 si las respuestas son aleatorias.

17.57. Usar el test de las rachas en las secuencias (10) y (11) de este capítulo, y establecer las conclusiones acerca de su aleatoriedad.

17.58. (*a*) Formar todas las posibles secuencias con dos *aes* y una *b*, y dar el número *V*, de rachas en cada una.

(*b*) Hallar la distribución muestral de *V* y su gráfico.

(c) Obtener la distribución de probabilidad de *V* y su gráfico.

17.59. En el Problema 17.58, hallar la media y la varianza de *V* (*a*) directamente de la distribución muestral y (*b*) por la fórmula.

17.60. Resolver los Problemas 17.58 y 17.59 para los casos en que hay (*a*) dos *aes* y dos *bes* (*b*) una *a* y tres *bes*, y (*c*) una *a* y cuatro *bes*.

17.61. Resolver los Problemas 17.58 y 17.59 con (*a*) dos *aes* y cuatro *bes* y (*b*) tres *aes* y tres *bes*.

OTRAS APLICACIONES DEL TEST DE LAS RACHAS

17.62. Determinar, al nivel de significación 0.05, si la muestra de 40 calificaciones de la Tabla 17.5 es aleatoria.

17.63. Las cotizaciones de ciertas acciones en 25 días sucesivos vienen dadas en la Tabla 17.42. Determinar al nivel de significación 0.05 si son aleatorias.

Tabla 17.42

10.375	11.125	10.875	10.625	11.500
11.625	11.250	11.375	10.750	11.000
10.875	10.750	11.500	11.250	12.125
11.875	11.375	11.875	11.125	11.750
11.375	12.125	11.750	11.500	12.250

17.64. Los primeros dígitos de $\sqrt{2}$ son 1.41421 35623 73095 0488 ··· . ¿Qué conclusiones se pueden sacar sobre su aleatoriedad?

17.65. ¿Qué conclusiones se pueden sacar sobre el carácter aleatorio de los siguientes dígitos?

(*a*) $\sqrt{3} = 1.73205\ 08075\ 68877\ 2935...$

(*b*) $\pi = 3.14159\ 26535\ 89793\ 2643...$

17.66 En el Problema 17.30, aplicar el test de las rachas para decidir sobre su aleatoriedad.

17.67. En el Problema 17.32, aplicar el test de las rachas para decidir sobre su aleatoriedad.

17.68. En el Problema 17.34, aplicar el test de las rachas para decidir sobre su aleatoriedad.

CORRELACION DE RANGO

17.69. En un concurso, dos jueces hubieron de colocar a ocho candidatos (numerados de 1 a 8) por orden de preferencia, con el resultado que recoge la Tabla 17.43.

(*a*) Hallar el coeficiente de correlación de rango.

(*b*) Decidir cuántos coincidentes fueron las elecciones de ambos jueces.

Tabla 17.43

Primer juez	Segundo juez
5	4
2	5
8	7
1	3
4	2
6	8
3	1
7	6

17.70. Aplicar correlación de rango al (*a*) Problema 14.26, (*b*) Problema 14.42, (*c*) Problema 14.46 y (*d*) Problema 14.63.

17.71. El coeficiente de correlación de rango se deduce usando los datos con rango en la fórmula momento-producto del Capítulo 14. Ilustrar esto resolviendo algún problema por ambos métodos.

17.72. ¿Puede hallarse el coeficiente de correlación de rango para datos agrupados? Explicar la respuesta e ilustrarla con un ejemplo.

CAPITULO 18

Análisis de series en el tiempo

SERIES EN EL TIEMPO

Una *serie en el tiempo* es un conjunto de observaciones tomadas en instantes específicos, generalmente a intervalos iguales. Ejemplos de tales series en el tiempo son la producción anual total de acero en EE.UU. durante un cierto número de años, la cotización diaria al cierre de la sesión bursátil de ciertas acciones, las temperaturas anunciadas cada hora por el instituto meteorológico para una ciudad o el total de ventas mensuales en una empresa.

Matemáticamente, una serie en el tiempo se define por los valores Y_1, Y_2, ... de una variable Y (temperatura, cotización, etc.) en tiempos t_1, t_2, ... Así pues, Y es una función de t; esto se denota por $Y = F(t)$.

GRAFICOS DE SERIES EN EL TIEMPO

Una serie en el tiempo que involucra a una variable Y se representa por un gráfico de Y respecto de t, como se ha hecho ya muchas veces en capítulos anteriores. Por ejemplo, la Figura 18.1 es el gráfico de una serie en el tiempo que muestra el número de cabezas de ganado en EE.UU. durante los años 1870-1980.

MOVIMIENTOS CARACTERISTICOS DE SERIES EN EL TIEMPO

Es interesante pensar en el gráfico de una serie en el tiempo (tal como el de la Fig. 18.1) como un gráfico que describe un punto moviéndose con el paso del tiempo, análogo en muchos aspectos a la trayectoria de una partícula física que se mueve bajo la influencia de fuerzas físicas. Claro está que, en lugar de fuerzas físicas, aquí cabe pensar en el resultado de una combinación de fuerzas económicas, sociológicas, psicológicas o de otros tipos.

La experiencia con muchos ejemplos de series en el tiempo ha revelado ciertos *movimientos* o *variaciones características* que aparecen a menudo, y cuyo análisis es de gran interés por muchas razones, una de ellas el problema de *predicción* de futuros movimientos. No puede sorprendernos, en consecuencia, que muchas empresas y gobiernos estén preocupados por este importante tema.

CLASIFICACION DE MOVIMIENTOS DE SERIES EN EL TIEMPO

Los movimientos característicos de series en el tiempo se pueden clasificar en cuatro tipos principales, a menudo llamados *componentes* de una serie en el tiempo:

1. **Movimientos a largo plazo o seculares.** Se refieren a la dirección general en la que el gráfico de una serie en el tiempo parece progresar en un largo período de tiempo. En la Figura 18.1, este movimiento secular (o *variación secular* o *tendencia secular*, como se llama a veces) se indica por una *curva de tendencia*, en trazo discontinuo. Para algunas series en el tiempo puede ser apropiada una *recta de tendencia*. La determinación de tales curvas o rectas de tendencia por mínimos cuadrados se ha considerado en el Capítulo 13. Otros métodos se discutirán más adelante en este capítulo.
2. **Movimientos característicos o variaciones cíclicas.** Estas se refieren a las oscilaciones a largo término en torno a una recta o curva de tendencia. Estos *ciclos*, como se les llama, pueden ser *periódicos* o no; es decir, pueden seguir o no esquemas repetidos en intervalos iguales de tiempo. En actividades de negocios o financieras, los movimientos se consideran cíclicos sólo si son recurrentes en un período de tiempo de al menos un año. Un importante ejemplo de movimientos característicos lo constituyen los llamados *ciclos económicos*, que representan intervalos de prosperidad, recesión, depresión y recuperación. Los movimientos característicos en torno a las curvas de tendencia son muy nítidos en la Figura 18.1

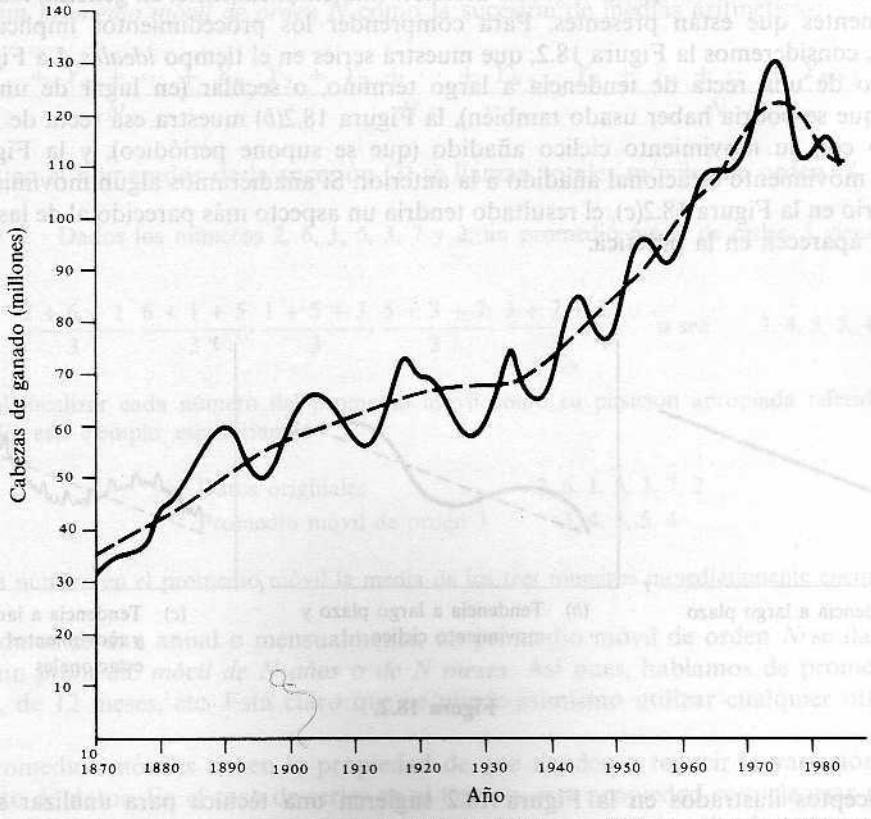


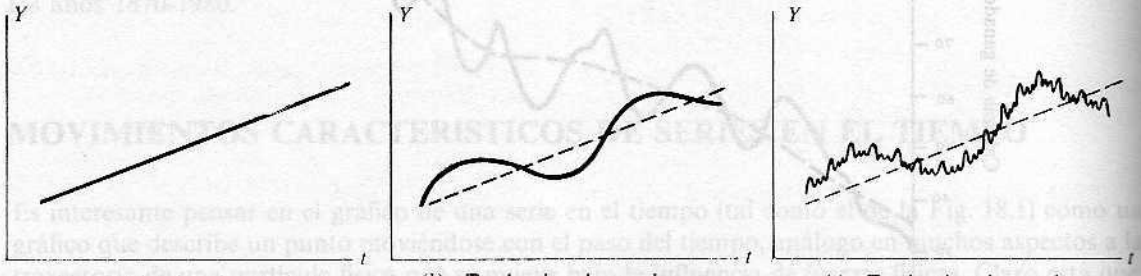
Figura 18.1. Censo de ganado en EE.UU., 1870-1980 (*Fuente:* U.S. Department of Agriculture).

3. **Movimientos estacionales o variaciones estacionales.** Estos se refieren a los esquemas idénticos o casi idénticos que una serie en el tiempo parece seguir durante meses correspondientes en años sucesivos. Tales movimientos se deben a sucesos recurrentes que tienen lugar anualmente, tales como el brusco aumento de precios al consumo antes de la Navidad. En la Figura 18.1 no se aprecian movimientos estacionales, pues el gráfico fue obtenido mediante datos anuales.

   Aunque los movimientos estacionales se refieren generalmente en teoría económica a periodicidad *anual*, las ideas en juego admiten extensión a intervalos cualesquiera de periodicidad (días, horas o semanas), según el tipo de datos de que disponemos.
4. **Movimientos irregulares o aleatorios.** Estos se refieren a los movimientos esporádicos de las series en el tiempo debidos a sucesos de azar, tales como inundaciones, huelgas o elecciones. Si bien se suele suponer que tales sucesos producen variaciones que pierden su influencia tras poco tiempo, cabe la posibilidad de que sean tan intensos que den lugar a nuevos movimientos cíclicos o de otro tipo.

ANALISIS DE SERIES EN EL TIEMPO

El análisis de series en el tiempo consiste en describir (matemáticamente, en general) los movimientos componentes que están presentes. Para comprender los procedimientos implicados en tal descripción, consideremos la Figura 18.2, que muestra series en el tiempo *ideales*. La Figura 18.2(*a*) es el gráfico de una recta de tendencia a largo término, o secular (en lugar de una curva de tendencia, que se podría haber usado también), la Figura 18.2(*b*) muestra esa recta de tendencia a largo plazo con su movimiento cíclico añadido (que se supone periódico), y la Figura 18.2(*c*) muestra un movimiento estacional añadido a la anterior. Si añadiéramos algún movimiento irregular o aleatorio en la Figura 18.2(*c*), el resultado tendría un aspecto más parecido al de las series en el tiempo que aparecen en la práctica.



(*a*) Tendencia a largo plazo

(*b*) Tendencia a largo plazo y movimiento cíclico

(*c*) Tendencia a largo plazo y movimientos cíclicos y estacionales

Figura 18.2.

Los conceptos ilustrados en la Figura 18.2 sugieren una técnica para analizar series en el tiempo. Supongamos que la serie en el tiempo tiene por variable Y el producto de varias variables

T, C, S e I que producen los movimientos de tendencia, cíclicos, estacionales e irregulares, respectivamente. En símbolos,

$$Y = T \times C \times S \times I = TCSI \tag{1}$$

El análisis de series en el tiempo requiere investigar los factores T, C, S e I, y se conoce a menudo como una *descomposición* de una serie en el tiempo en movimientos componentes básicos.

Hay que hacer constar que algunos estadísticos prefieren considerar Y como la suma $T + C + S + I$ de las variables básicas involucradas. Aunque supondremos la descomposición dada por la ecuación (1) cuando examinemos los métodos discutidos en este capítulo, procedimientos análogos entran en juego cuando se trata con una suma. En la práctica, la decisión sobre cuál de los métodos de descomposición se adopta depende del grado de éxito a que conduce la aplicación de cada uno.

PROMEDIOS MOVILES; SUAVIZACION DE SERIES EN EL TIEMPO

Dado un conjunto de números

$$Y_1, Y_2, Y_3, \ldots \tag{2}$$

definimos un *promedio móvil de orden N* como la sucesión de medias aritméticas:

$$\frac{Y_1 + Y_2 + \cdots + Y_N}{N}, \frac{Y_2 + Y_3 + \cdots + Y_{N+1}}{N}, \frac{Y_3 + Y_4 + \cdots + Y_{N+2}}{N}, \ldots \tag{3}$$

Las sumas en el numerador de la sucesión (3) se llaman totales móviles de orden N.

EJEMPLO 1. Dados los números 2, 6, 1, 5, 3, 7 y 2, un promedio móvil de orden 3 viene dado por la sucesión

$$\frac{2+6+1}{3}, \frac{6+1+5}{3}, \frac{1+5+3}{3}, \frac{5+3+7}{3}, \frac{3+7+2}{3} \quad \text{o sea} \quad 3, 4, 3, 5, 4$$

Es usual localizar cada número del promedio móvil como su posición apropiada referida a los datos originales. En este ejemplo escribiríamos

Datos originales	2, 6, 1, 5, 3, 7, 2
Promedio móvil de orden 3	3, 4, 3, 5, 4

siendo cada número en el promedio móvil la media de los tres números inmediatamente encima de él.

Si los datos se dan anual o mensualmente, un promedio móvil de orden N se llama, respectivamente, un *promedio móvil de N años* o *de N meses*. Así pues, hablamos de promedios móviles de 5 años, de 12 meses, etc. Está claro que se puede asimismo utilizar cualquier otra unidad de tiempo.

Los promedios móviles tienen la propiedad de que tienden a reducir la variación presente en un conjunto de datos. En el caso de series en el tiempo, esta propiedad se suele usar para eliminar fluctuaciones indeseables, en un proceso que se conoce como *suavización de series en el tiempo*.

Si se usan medias aritméticas ponderadas en la sucesión (3), con pesos especificados de antemano, la sucesión resultante se llama un *promedio móvil ponderado de orden N.*

EJEMPLO 2. Si se usan pesos 1, 4 y 1 en el Ejemplo 1, un promedio móvil ponderado de orden 3 viene dado por la sucesión

$$\frac{1(2)+4(6)+1(1)}{1+4+1}, \frac{1(6)+4(1)+1(5)}{1+4+1}, \frac{1(1)+4(5)+1(3)}{1+4+1}, \frac{1(5)+4(3)+1(7)}{1+4+1}, \frac{1(3)+4(7)+1(2)}{1+4+1}$$

o sea, 4.5, 2.5, 4.0, 4.0, 5.5

ESTIMACION DE LA TENDENCIA

1. **Método de los mínimos cuadrados.** Este método, descrito en el Capítulo 13, se puede utilizar para hallar la ecuación de la recta o curva de tendencia adecuada. De esta ecuación se podrán calcular los valores de tendencia T.
2. **Método «a mano».** Este método, que consiste en ajustar una curva o recta de tendencia por simple inspección del gráfico, también se puede usar para estimar T. No obstante, tiene la desventaja evidente de depender muy fuertemente del criterio personal de cada cual.
3. **Método del promedio móvil.** Usando promedios móviles de órdenes apropiados, podemos eliminar esquemas cíclicos, estacionales e irregulares, dejando así tan sólo el movimiento de tendencia.

   Una desventaja de este método es que los datos al comienzo y al final de una serie se pierden: así, en el Ejemplo 1 comenzamos con siete números, y con un promedio móvil de orden 3 llegamos a cinco números. Otra desventaja es que los promedios móviles pueden generar ciclos u otros movimientos que no estaban presentes en los datos originales. Una tercera desventaja es que los promedios móviles se ven muy afectados por los valores extremos. Para obviar esto último en cierta medida, se usa a veces un promedio móvil ponderado con pesos adecuados; en tal caso, al valor o valores centrales se les asigna peso máximo, y a los valores extremos, pesos pequeños.
4. **Método de semipromedios.** Consiste en separar los datos en dos partes (preferible que sean iguales) y promediar los datos de cada parte, obteniendo con ello dos puntos en el gráfico de la serie en el tiempo. Entonces se traza una recta de tendencia entre esos dos puntos, y los valores de tendencia se determinan de esa recta de tendencia. Los valores de tendencia se pueden determinar también directamente, sin gráfico (véase Prob. 18.6).

   Aunque el método es sencillo de aplicar, puede conducir a resultados pobres si se usa indiscriminadamente. Además es sólo aplicable cuando la tendencia es lineal o aproximadamente lineal, si bien puede extenderse a casos en que los datos pueden agruparse en varias partes, en cada una de las cuales la tendencia es lineal.

ESTIMACION DE LAS VARIACIONES ESTACIONALES; EL INDICE ESTACIONAL

Para determinar el factor estacional S en la ecuación (1), debemos estimar cómo varían los datos de la serie en el tiempo de mes a mes en un año típico. Un conjunto de números que muestra los

valores relativos de una variable durante los meses del año se llama un *índice estacional* para la variable. Por ejemplo, si sabemos que las ventas durante enero, febrero, marzo, etc., son el 50, 120, 90, ... % del promedio de ventas mensual en el total del año, entonces los números 50, 120, 90, ... dan el índice estacional de ese año, y se llaman *números índice estacionales*. El índice estacional medio del año ha de ser 100%; esto es, la suma de los números índice de los 12 meses ha de ser 1200%.

Se dispone de varios métodos para calcular un índice estacional:

1. **Método de porcentaje medio.** En este método expresamos los datos de cada mes como porcentajes del promedio anual. Los porcentajes para meses correspondientes en distintos años se promedian entonces, usando una media o una mediana; si se usa la media, es mejor evitar valores extremos que puedan aparecer. Los 12 porcentajes resultantes dan el índice estacional. Si su media no es el 100% (o sea, si su suma no es 1200%), deben ser ajustados, lo que se logra multiplicándolos por un factor adecuado.
2. **Método del porcentaje de tendencia.** En este método expresamos los datos para cada mes como porcentajes de valores de tendencia mensuales. Un promedio apropiado de los porcentajes para meses correspondientes da entonces el índice requerido. Como en el método 1, los ajustamos si no tienen promedio 100%.

   Nótese que al dividir cada valor mensual Y por el correspondiente valor de tendencia T resulta $Y/T = CSI$, de la ecuación (1), y que el subsiguiente promedio de Y/T produce los índices estacionales. En tanto en cuanto estos índices incluyen variaciones cíclicas e irregulares, puede ser una desventaja del método, especialmente si las variaciones son grandes.
3. **Método del promedio móvil en porcentaje.** En este método calculamos un promedio móvil de 12 meses. Como los resultados obtenidos así caen entre meses sucesivos en lugar de en el centro del mes (que es donde caen los datos originales), calculamos un promedio móvil de 2 meses de ese promedio móvil de 12 meses. El resultado se llama a veces un *promedio móvil de 12 meses centrado*.

   Tras hacer eso, expresamos los datos originales de cada mes como un porcentaje del promedio móvil centrado de 12 meses que corresponde a los datos originales. Los porcentajes de los meses correspondientes se promedian a continuación, dando el índice buscado. Como antes, los ajustamos si no promedian 100%.

   Obsérvese que el razonamiento lógico que subyace a este método se sigue de la ecuación (1). Un promedio móvil centrado de 12 meses de Y sirve para eliminar los movimientos estacionales e irregulares S e I, y es por tanto equivalente a los valores dados por TC. Al dividir los datos originales por TC nos da SI. Los promedios subsiguientes sobre meses correspondientes sirven para eliminar la irregularidad I y en consecuencia producen un índice S adecuado.
4. **Método de la relación de enlace.** En este método expresamos los datos para cada mes como un porcentaje de los datos para los meses previos; estos porcentajes mensuales se llaman *relaciones de enlace* porque relacionan cada mes con el precedente. Entonces tomamos un promedio adecuado de los enlaces relativos para los meses correspondientes. De estas 12 relaciones de enlace promedio obtenemos los porcentajes relativos de cada mes respecto a enero, que se adopta como el 100%.

   Tras hacer eso, encontraremos que el siguiente enero tiene un porcentaje asociado que es mayor o menor que 100%, según haya habido un crecimiento o decrecimiento en la tendencia. Usando este porcentaje del próximo enero, ajustamos los diversos porcentajes relativos mensuales (antes obtenidos) para esta tendencia. Estos porcentajes finales, ajustados de modo que promedien 100%, dan el índice estacional requerido.

DATOS AJUSTADOS A LA VARIACION ESTACIONAL

Si los datos mensuales originales se dividen por los correspondientes números índice estacionales, los datos resultantes se llaman *desestacionalizados* o *ajustados a la variación estacional*. Tales datos incluyen todavía movimientos de tendencia, cíclicos e irregulares.

ESTIMACION DE LAS VARIACIONES CICLICAS

Una vez ajustados los datos a la variación estacional, pueden ser ajustados también a la tendencia sin más que dividirlos por los correspondientes valores de tendencia. De acuerdo con la ecuación (1), el proceso de ajustar a la variación estacional y a la tendencia corresponde a dividir Y por ST, lo que da CI (las variaciones cíclicas e irregulares). Un promedio móvil apropiado de unos pocos meses de duración (digamos 3, 5 ó 7 meses, de manera que el centrado subsiguiente no sea necesario) sirve entonces para suavizar las variaciones irregulares I y para dejar sólo las variaciones cíclicas C. Una vez que estas variaciones cíclicas han sido aisladas de esa forma, se pueden estudiar en detalle. Si ocurre una periodicidad, exacta o aproximada, de ciclos, se pueden construir *índices cíclicos* de manera parecida a como se ha hecho para los índices estacionales.

ESTIMACION DE LAS VARIACIONES IRREGULARES

Las variaciones irregulares (o aleatorias) se pueden estimar ajustando los datos a las variaciones de tendencia, estacionales y cíclicas. Eso significa tener que dividir los datos originales Y por T, S y C, que [por la ecuación (1)] da I. En la práctica se encuentra que las variaciones irregulares tienden a tener pequeña magnitud y con frecuencia tienden a seguir el esquema de una distribución normal; es decir, las pequeñas desviaciones ocurren con gran frecuencia y grandes desviaciones ocurren con pequeña frecuencia.

COMPARACION DE DATOS

Al comparar datos, hay que tener siempre mucho cuidado de que tal comparación esté justificada. Por ejemplo, al comparar datos de marzo con datos de febrero, debemos tener bien presente que febrero tiene 28 ó 29 días y marzo tiene 31; y al comparar datos de febrero de años diferentes, hay que recordar que en un año bisiesto febrero tiene 29 días en lugar de 28. Para poner otro ejemplo, el número de días laborables durante varios meses del mismo año o de años diferentes, pueden ser distintos a causa de las vacaciones, huelgas, etc.

En la práctica, no se sigue una regla definida para ajustar tales variaciones. La necesidad de tales ajustes queda a voluntad del investigador.

PREDICCION

Los métodos y principios anteriores se usan en la importante tarea de predecir series en el tiempo. Hay que ser conscientes de que, naturalmente, el tratamiento matemático de los datos no resuelve por sí mismo todos los problemas. No obstante, acoplado al sentido común del investigador, a su

experiencia, su ingenio y buen juicio, el análisis matemático ha demostrado su utilidad tanto en predicciones de largo como de corto alcance.

RESUMEN DE LOS PASOS FUNDAMENTALES EN EL ANALISIS DE SERIES EN EL TIEMPO

1. Recoger datos para la serie en el tiempo, procurando que esos datos sean fiables. Tener siempre presente el objetivo eventual de la serie en el tiempo; por ejemplo, si uno quiere predecir una serie en el tiempo dada, puede ser útil obtener series en el tiempo relacionadas (así como cualquier otra información). Si es necesario, se ajustan los datos que han de ser comparados, teniendo en cuenta años bisiestos, vacaciones, etc.
2. Representar gráficamente la serie en el tiempo, observando cualitativamente la presencia de variaciones y de variaciones de tendencia a largo término y cíclicas.
3. Construir la curva o recta de tendencia a largo término, y obtener los valores de tendencia apropiados usando los métodos de mínimos cuadrados, «a mano», promedios móviles o semipromedios.
4. Si hay variaciones estacionales, obtener un índice estacional y desestacionalizar los datos (o sea, ajustar los datos a las variaciones).
5. Ajustar los datos desestacionalizados a la tendencia. Los datos resultantes contienen (teóricamente) sólo variaciones cíclicas e irregulares. Un promedio móvil de 3, 5 ó 7 meses servirá para remover las variaciones irregulares, revelando las variaciones cíclicas.
6. Representar gráficamente las variaciones cíclicas obtenidas en el paso 5, observando cualquier periodicidad, exacta o aproximada, que pueda estar presente.
7. Si se desea una predicción, hágase combinando los resultados de los pasos 1 a 6 y utilizando toda otra información disponible. Identificar y evaluar todas las posibles fuentes de error y su magnitud.

PROBLEMAS RESUELTOS

MOVIMIENTOS CARACTERISTICOS DE SERIES EN EL TIEMPO

18.1. ¿Con qué movimientos característicos de una serie en el tiempo asociaría principalmente (*a*) un incendio en una fábrica que retrasa 3 semanas su producción, (*b*) una época de prosperidad, (*c*) las ventas prenavideñas en un establecimiento, (*d*) la necesidad de aumentar la producción de trigo a causa de un constante crecimiento de la población y (*e*) la lluvia caída mensualmente sobre una ciudad en un período de 5 años?

Solución

Los movimientos característicos son (*a*) irregular (*b*) cíclico, (*c*) estacional, (*d*) a largo término y (*e*) estacional.

PROMEDIOS MOVILES; SUAVIZACION DE SERIES EN EL TIEMPO

18.2. La Tabla 18.2 muestra la población agricultora (en millones) en EE.UU. durante los años 1973-1983. Construir (*a*) un promedio móvil de 5 años y (*b*) un promedio móvil de 4 años.

Tabla 18.1

Año	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
Población agricultora (millones)	9.47	9.26	8.86	8.25	7.81	8.01	7.55	7.24	7.01	6.88	7.03

Fuente: U.S. Department of Agriculture.

Solución

(*a*) Referimos a la Tabla 18.2. En la columna 3 el primer total móvil, 43.65, es la suma de las entradas primera a quinta de la columna 2; el segundo total móvil, 42.19, es la suma de la segunda a sexta entradas de la columna 2; etc.

En la práctica, tras obtener el primer total móvil (43.65), podemos obtener fácilmente el segundo restando de él 9.47 (la primera entrada de la columna 2) y sumando 8.01 (la sexta entrada de la columna 2), con lo que se llega a 42.19. Los sucesivos totales móviles se hallan del mismo modo.

Dividiendo cada total móvil por 5 se obtiene el promedio móvil requerido (columna 4).

(*b*) Referimos a la Tabla 18.3. Los totales móviles de 4 años se hallan como en la parte (*a*), excepto que ahora se suman sólo cuatro entradas de la columna 2. Nótese que, a diferencia del método de la parte (*a*), los totales móviles están centrados *entre* años sucesivos. Este es siempre el caso cuando se toma el promedio móvil sobre un número *par* de años. Así, si consideramos que 1974 significa el 1 de julio de 1974, entonces el total móvil de los cuatro primeros años está centrado en el 1 de enero de 1975, o el 31 de diciembre de 1974.

Los promedios móviles de 4 años se obtienen dividiendo los totales móviles de 4 años por 4.

18.3. Construir un promedio móvil centrado de 4 años para los datos del Problema 18.2.

Tabla 18.2

Año	Datos	Total móvil de 5 años	Promedio móvil de 5 años
1973	9.47		
1974	9.26		
1975	8.86	43.65	8.73
1976	8.25	42.19	8.44
1977	7.81	40.48	8.10
1978	8.01	38.86	7.77
1979	7.55	37.62	7.52
1980	7.24	36.69	7.34
1981	7.01	35.71	7.14
1982	6.88		
1983	7.03		

Tabla 18.3

Año	Datos	Total móvil de 4 años	Promedio móvil de 4 años
1973	9.47		
1974	9.26		
1975	8.86	35.84	8.96
1976	8.25	34.18	8.55
1977	7.81	32.93	8.23
1978	8.01	31.62	7.91
1979	7.55	30.61	7.65
1980	7.24	29.81	7.45
1981	7.01	28.68	7.17
1982	6.88	28.16	7.04
1983	7.03		

Solución

Primer método

Primero calculamos un promedio móvil de 4 años, como en el Problema 18.2(*b*); estos valores están centrados entre años sucesivos, como muestra la Tabla 18.4. Si ahora calculamos un total móvil

de 2 años para esos promedios móviles de 4 años, los resultados están centrados en los años requeridos. Dividiendo los resultados de la columna 4 por 2 obtenemos el promedio móvil *centrado* requerido (columna 5).

Tabla 18.4

Año	Datos	Promedio móvil de 4 años	Total móvil de 2 años para la columna 3	Promedio móvil centrado de 4 años (Columnas 4 ÷ 2)
1973	9.47			
1974	9.26			
1975	8.86	8.96	17.51	8.76
1976	8.25	8.55	16.78	8.39
1977	7.81	8.23	16.14	8.07
1978	8.01	7.91	15.56	7.78
1979	7.55	7.65	15.10	7.55
1980	7.24	7.45	14.62	7.31
1981	7.01	7.17	14.21	7.11
1982	6.88	7.04		
1983	7.03			

Segundo método

Primero calculamos un promedio móvil de 4 años, como en el Problema 18.2(*b*); estos valores están centrados entre años sucesivos, como muestra la Tabla 18.5. Si ahora calculamos un total móvil de 2 años para esos promedios móviles de 4 años, los resultados están centrados en los años requeridos. Dividiendo los resultados de la columna 4 por 8 (2 × 4) obtenemos el promedio móvil *centrado* requerido. Para el año 1975, la ligera diferencia entre 8.76 y 8.75 en las Tablas 18.4 y 18.5 se debe a errores de redondeo.

Tabla 18.5

Año	Datos	Total móvil de 4 años	Total móvil de 2 años para la columna 3	Promedio móvil centrado de 4 años (Columnas 4 ÷ 8)
1973	9.47			
1974	9.26			
1975	8.86	35.84	70.02	8.75
1976	8.25	34.18	67.11	8.39
1977	7.81	32.93	64.55	8.07
1978	8.01	31.62	62.23	7.78
1979	7.55	30.61	60.42	7.55
1980	7.24	29.81	58.49	7.31
1981	7.01	28.68	56.84	7.11
1982	6.88	28.16		
1983	7.03			

18.4. Probar que el promedio móvil centrado de 4 años del Problema 18.3 es equivalente a un promedio móvil ponderado de 5 años con pesos 1, 2, 2, 2 y 1, respectivamente.

Solución

Denotemos por Y_1, Y_2, ..., Y_{11} los valores correspondientes a los años 1973, 1974, ..., 1983, respectivamente. Entonces, procediendo como en el segundo método del Problema 18.3, obtenemos la Tabla 18.6, de cuya columna de la derecha vemos que el promedio móvil centrado de 4 años es un promedio móvil ponderado de 5 años con pesos 1, 2, 2, 2 y 1, respectivamente. Nótese que la suma de esos pesos es $1 + 2 + 2 + 2 + 1 = 8$.

Este método se puede utilizar para llegar a los resultados del Problema 18.3. Por ejemplo, la primera entrada (correspondiente a 1975) es

$$\frac{(1)(9.47) + 2(9.26) + 2(8.86) + 2(8.25) + 1(7.81)}{8} = 8.75$$

Tabla 18.6

Año	Y	Total móvil de 4 años	Total móvil de 2 años para la columna 3	Promedio móvil centrado de 4 años (Columnas 4 ÷ 8)
1948	Y_1			
1949	Y_2	$Y_1+Y_2+Y_3+Y_4$		
1950	Y_3	$Y_2+Y_3+Y_4+Y_5$	$Y_1+2Y_2+2Y_3+2Y_4+2Y_5$	$\frac{1}{8}(Y_1+2Y_2+2Y_3+2Y_4+Y_5)$
1951	Y_4	$Y_3+Y_4+Y_5+Y_6$	$Y_2+2Y_3+2Y_4+2Y_5+2Y_6$	$\frac{1}{8}(Y_2+2Y_3+2Y_4+2Y_5+Y_6)$
1952	Y_5	$Y_4+Y_5+Y_6+Y_7$	$Y_3+2Y_4+2Y_5+2Y_6+Y_7$	$\frac{1}{8}(Y_3+2Y_4+2Y_5+2Y_6+Y_7)$
1953	Y_6			
...	...	.....	.....	.....
1958	Y_{11}			

18.5. Representar gráficamente el promedio móvil del Problema 18.2(*a*) y los datos originales (de la Tabla 18.1).

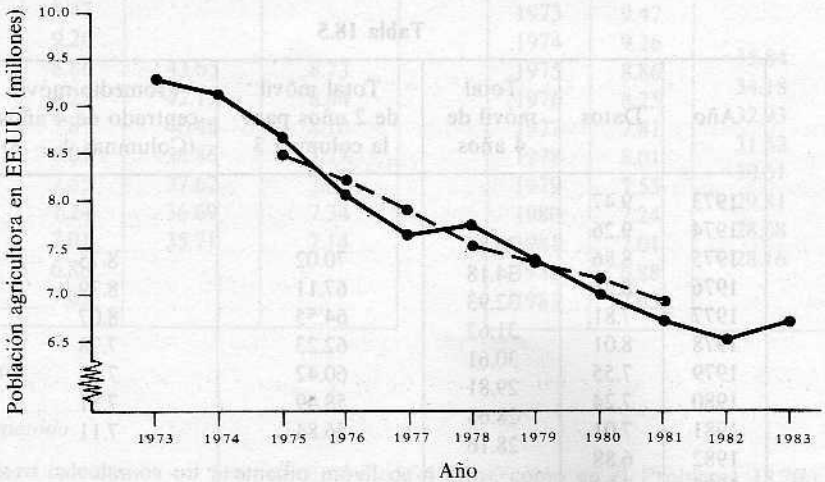


Figura 18.3.

Solución

El gráfico de los datos originales se indica con trazo continuo en la Figura 18.3, y el gráfico del promedio móvil se indica en trazo discontinuo. Observemos que el promedio móvil ha suavizado el gráfico de los datos originales, mostrando con nitidez la recta de tendencia.

Una desventaja del promedio móvil es que se pierden datos al comienzo y al final de la serie en el tiempo, lo cual puede ser grave si se dispone de un número escaso de datos.

ESTIMACION DE LA TENDENCIA

18.6. Usando el método de los semipromedios, hallar los valores de tendencia para los datos del Problema 18.2 tomando el promedio como (*a*) media y (*b*) mediana.

Solución

(*a*) Dividimos los datos en dos partes iguales (omitiendo el año central, 1978), como muestra la Tabla 18.7. Entonces calculamos la media para los datos de cada parte. De los resultados obtenidos se deduce que en 6 años (1975-1981) ha habido un *decrecimiento* de 8.73 − 7.14 = = 1.59 millones, con un decrecimiento de 1.59/6 = 0.265 anual. Sabiendo esto, podemos calcular los valores de tendencia. Así pues, los valores de tendencia para 1976 y 1977 son, respectivamente, 8.73 − 0.265 = 8.47 y 8.73 − 2(0.265) = 8.20; los valores de tendencia para 1974 y 1973 son, respectivamente, 8.73 + 0.265 = 9.00 y 8.73 + 2(0.265) = 9.26; etc. como recoge la Tabla 18.8.

Los resultados se pueden obtener también dibujando el gráfico de una recta que conecte los puntos (1975, 8.73) y (1981, 7.14) y leyendo los valores de tendencia de ese gráfico.

Tabla 18.7

1973	9.47
1974	9.26
1975	8.86
1976	8.25
1977	7.81
Total	43.65

Media = 43.65/5 = 8.73
(correspondiente a 1975)

1979	7.55
1980	7.24
1981	7.01
1982	6.88
1983	7.03
Total	35.71

Media = 35.71/5 = 7.14
(correspondiente a 1981)

Tabla 18.8

Año	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
Valor de tendencia	9.26	9.00	8.73	8.47	8.20	7.94	7.67	7.41	7.14	6.88	6.61

(*b*) Las medianas de las dos partes iguales en (*a*) son 8.86 y 7.03, respectivamente. Luego hay un decrecimiento de (8.86 − 7.03)/6 = 0.305 al año, y los valores de tendencia se muestran en la Tabla 18.9.

Cuando se usan medianas, el método se suele llamar método de *semimedianas*. Si no se especifica el tipo de promedio, se sobreentiende la media.

Tabla 18.9

Año	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
Valor de tendencia	9.47	9.17	8.86	8.56	8.25	7.95	7.64	7.34	7.03	6.73	6.42

18.7. Describir cómo utilizar (*a*) el método «a mano» y (*b*) el método de promedios móviles, para calcular los valores de tendencia para los datos del Problema 18.2.

Solución

(*a*) Con el método a mano, simplemente construimos una recta o curva que se aproxime al gráfico de la Figura 18.3, y a continuación leemos los valores de tendencia de esa recta o curva.

(*b*) Vimos en el Problema 18.5 que el promedio móvil de 5 años suavizaba los datos de la serie en el tiempo considerablemente. Podemos usar los promedios obtenidos como valores de tendencia para los años 1975-1981. Así pues, vemos del Problema 18.2(*a*) que los valores de tendencia correspondientes a 1975, 1976, 1977, etc., son 8.73, 8.44, 8.10, etc. Sin embargo, este método hace que no podamos disponer de los valores de tendencia para 1973, 1974, 1982 y 1983; si se desea, se pueden obtener por extrapolación de la Figura 18.3 (el gráfico del Prob. 18.5).

18.8. (*a*) Usar el método de mínimos cuadrados para ajustar una recta a los datos del Problema 18.2.

(*b*) Del resultado de la parte (*a*), hallar los valores de tendencia.

Solución

(*a*) Puesto que los datos se refieren a un número impar de años, usamos el segundo método del Problema 13.19, de donde obtenemos la Tabla 18.10. Luego la recta de mínimos cuadrados pedida es

$$Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X = \frac{87.37}{11} + \left(\frac{-29.55}{110}\right)X \qquad \text{o sea} \qquad Y = 7.94 - 0.269X$$

donde el origen $X = 0$ es el año 1978 y la unidad de Y es 1 año.

(*b*) Haciendo $X = -5, -4, -3, \ldots, 5$ en la ecuación de mínimos cuadrados de la parte (*a*), se deducen los valores de tendencia que recoge la Tabla 18.11. Los resultados están en buen acuerdo con los del Problema 18.6.

ESTIMACION DE LAS VARIACIONES ESTACIONALES; EL INDICE ESTACIONAL

18.9. La Tabla 18.12 muestra la producción de energía eléctrica mensual de consumo no industrial, en miles de millones de kilovatios-hora (kwh), en EE.UU. durante los años 1976-1981.

(*a*) Construir un gráfico con los datos.

(*b*) Hallar un índice estacional por medio del método de porcentaje promedio.

Tabla 18.10

Año	X	Y	X^2	XY
1973	−5	9.47	25	−47.35
1974	−4	9.26	16	−37.04
1975	−3	8.86	9	−26.58
1976	−2	8.25	4	−16.50
1977	−1	7.81	1	−7.81
1978	0	8.01	0	0
1979	1	7.55	1	7.55
1980	2	7.24	4	14.48
1981	3	7.01	9	21.03
1982	4	6.88	16	27.52
1983	5	7.03	25	35.15
		$\sum Y = 87.37$	$\sum X^2 = 110$	$\sum XY = -29.55$

Tabla 18.11

Año	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
Valor de tendencia	9.28	9.01	8.74	8.47	8.20	7.93	7.66	7.39	7.12	6.85	6.58

Tabla 18.12

	En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
1976	178.2	156.7	164.2	153.2	157.5	172.6	185.9	185.8	165.0	163.6	169.0	183.1
1977	196.3	162.8	168.6	156.9	168.2	180.2	197.9	195.9	176.0	166.4	166.3	183.9
1978	197.3	173.7	173.2	159.7	175.2	187.4	202.6	205.6	185.6	175.6	176.3	191.7
1979	209.5	186.3	183.0	169.5	178.2	186.7	202.4	204.9	180.6	179.8	177.4	188.9
1980	200.0	188.7	187.5	168.6	175.7	189.4	216.1	215.4	191.5	178.5	178.6	195.6
1981	205.2	179.6	185.4	172.4	177.7	202.7	220.2	210.2	186.9	181.4	175.6	195.6

Fuente: Survey of Current Business.

Solución

(*a*) Véase la Figura 18.4.

(*b*) La Tabla 18.13 muestra los promedios (medias) totales y mensuales para 1976-1981. Dividiendo los datos mensuales de la Tabla 18.12 por los correspondientes promedios mensuales para cada año de la Tabla 18.13 y expresando el resultado como porcentaje, nos da las entradas de la Tabla 18.14; por ejemplo, la primera entrada viene dada por 178.2/169.6 = 105.1%. La fila de abajo en la Tabla 18.14 da el porcentaje medio para cada mes; como el total de esos porcentajes es 1200%, no es necesario ajustarlos, y en consecuencia los números de esa fila inferior representan el índice estacional pedido.

Este índice estacional muestra que, en promedio, la producción de energía eléctrica requerida

es mínima en abril y máxima en los meses de verano, julio y agosto (cuando el aire acondicionado provoca demanda extra). Las cifras de producción real para varias regiones del país ayudan a las compañías eléctricas a cubrir adecuadamente las necesidades de los usuarios.

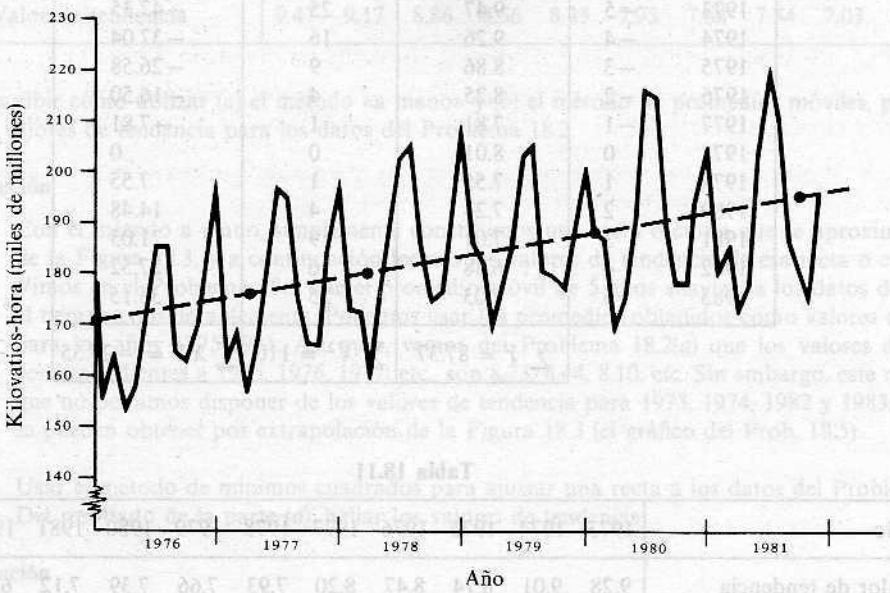


Figura 18.4. Producción de energía eléctrica no industrial en EE. UU., 1976-1981.

Tabla 18.13

Año	1976	1977	1978	1979	1980	1981
Total	2034.8	2119.4	2203.9	2247.2	2285.6	2292.9
Promedio mensual	169.6	176.6	183.7	187.3	190.5	191.1

Tabla 18.14

	En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
1976	105.1	92.4	96.8	90.3	92.9	101.8	109.6	109.6	97.3	96.5	99.6	108.0
1977	111.2	92.2	95.5	88.8	95.2	102.0	112.1	110.9	99.7	94.2	94.2	104.1
1978	107.4	94.6	94.3	86.9	95.4	102.0	110.3	111.9	101.0	95.6	96.0	104.4
1979	111.9	99.5	97.7	90.5	95.1	99.7	108.1	109.4	96.4	96.0	94.7	100.9
1980	105.0	99.1	98.4	88.5	92.2	99.4	113.4	113.1	100.5	93.7	93.8	102.7
1981	107.4	94.0	97.0	90.2	93.0	106.1	115.2	110.0	97.8	94.9	91.9	102.4
Total	648.0	571.8	579.7	535.2	563.8	611.0	668.7	664.9	592.7	570.9	570.2	622.5
Media	108.0	95.3	96.6	89.2	94.0	101.8	111.5	110.8	98.8	95.2	95.0	103.8

18.10. Hallar el índice estacional para el Problema 18.9 usando la mediana en vez de la media.

Solución

Los números en la columna de enero de la Tabla 18.14, cuando se colocan por orden creciente de magnitud, son 105.0, 105.1, 107.4, 107.4, 111.2 y 111.9, luego la mediana es $\frac{1}{2}(107.4 + 107.4) = 107.4$. Las medianas para los otros meses se hallan del mismo modo y se recogen en la segunda fila de la Tabla 18.15. Como estas medianas suman 1198.2, las ajustamos multiplicando cada número por 1200/1198.2. Eso produce los números de la tercera fila de la Tabla 18.15, que ya nos da el índice estacional buscado. Los resultados están en buen acuerdo con los obtenidos usando la media (Problema 18.9). En la práctica, siempre que los resultados con media y mediana difieren, se opta por usar la mediana para eliminar los valores extremos.

Tabla 18.15

	En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
Mediana	107.4	94.3	96.9	90.0	94.0	101.9	111.2	110.5	98.8	95.3	94.5	103.4
Indice estacional	107.6	94.4	97.0	90.1	94.1	102.1	111.4	110.7	98.9	95.4	94.6	103.6

18.11. Hallar un índice estacional para los datos del Problema 18.9 usando el método del porcentaje de tendencia. Al aplicar este método, obténgase los valores de tendencia mensuales por mínimos cuadrados.

Solución

A la vista del gráfico de los datos reales (Fig. 18.4) se desprende que la tendencia a largo término se puede aproximar convenientemente por una recta. En vez de hallar esta recta a partir de los datos mensuales de la Tabla 18.12, la hallaremos de los promedios mensuales de los años 1976-1981, como muestra la Tabla 18.16 (que se ha tomado de la Tabla 18.13). Supongamos que las cifras mensuales de la Tabla 18.12 corresponden a la mitad del mes; así pues, los promedios de la Tabla 18.16 corresponden al 30 de junio o al 1 de julio del año en cuestión.

Como hay un número par de años, usamos el segundo método del Problema 13.20, de donde obtenemos la Tabla 18.17. La recta de mínimos cuadrados pedida es

$$Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X = \frac{1098.8}{6} + \left(\frac{152.8}{70}\right)X = 183.13 + 2.183X$$

Tabla 18.16

Año	1976	1977	1978	1979	1980	1981
Promedio mensual	169.6	176.6	183.7	187.3	190.5	191.1

Tabla 18.17

Año	X	Y	X^2	XY
1976	−5	169.6	25	−848.0
1977	−3	176.6	9	−529.8
1978	−1	183.7	1	−183.7
1979	1	187.3	1	187.3
1980	3	190.5	9	571.5
1981	5	191.1	25	955.5
		$\sum Y = 1098.8$	$\sum X^2 = 70$	$\sum XY = 152.8$

donde X se mide en semestres y el origen es del 31 de diciembre de 1978 o el 1 de enero de 1979. De esta ecuación se deduce que los valores de Y crecen 2.183 cada semestre, o sea $2.183/6 = 0.3638$ cada mes. Así pues, cuando $X = 0$ (1 enero 1978), $Y = 183.13$. Medio mes después (15 enero 1978) el valor de Y es $183.13 + \frac{1}{2}(0.3638) = 183.31$. Añadiendo sucesivamente 0.3638 a 183.31, hallamos los valores de tendencia para febrero de 1978, marzo de 1978, etc., que son $183.31 + 0.3638 = 183.7$, $183.31 + 2(0.3638) = 184.0$, etc. Análogamente, restando sucesivamente 0.3638 de 183.31, hallamos los valores de tendencia para diciembre de 1977, noviembre de 1977, etc., que son $183.13 - 0.3638 = 182.8$, $183.13 - 2(0.3638) = 182.4$, etc. De esta manera obtenemos los valores de tendencia mensuales que se indican en la Tabla 18.18. La recta de mínimos cuadrados se muestra en trazo discontinuo en la Figura 18.4.

Tabla 18.18

	En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
1976	170.2	170.6	170.9	171.3	171.7	172.0	172.4	172.8	173.1	173.5	173.9	174.2
1977	174.6	174.9	175.3	175.7	176.0	176.4	176.8	177.1	177.5	177.9	178.2	178.6
1978	178.9	179.3	179.7	180.0	180.4	180.8	181.1	181.5	181.9	182.9	182.6	182.9
1979	183.3	183.7	184.0	184.4	184.8	185.1	185.5	185.9	186.2	186.6	186.9	187.3
1980	187.7	188.0	188.4	188.8	189.1	189.5	189.9	190.2	190.6	190.9	191.3	191.7
1981	192.0	192.4	192.8	193.1	193.5	193.9	194.2	194.6	195.0	195.3	195.7	196.0

Tabla 18.19

	En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
1976	104.7	91.9	96.1	89.4	91.7	100.3	107.8	107.5	95.3	94.3	97.2	105.1
1977	112.4	93.1	96.2	89.3	95.6	102.2	111.9	110.8	99.2	93.5	93.3	103.0
1978	110.3	96.9	96.4	88.7	97.1	103.7	111.9	113.3	102.0	96.4	96.5	104.8
1979	114.6	101.4	99.5	91.9	96.4	100.9	109.1	110.2	97.0	96.4	94.9	100.9
1980	106.6	100.4	99.7	89.4	92.9	99.9	113.8	113.2	100.5	93.5	93.4	102.0
1981	106.9	93.3	96.2	89.3	91.8	104.5	113.4	108.0	95.8	92.9	89.7	99.8
Media	109.3	96.2	97.4	89.7	94.3	101.9	111.3	110.5	98.3	94.5	94.2	102.6
Mediana	108.6	95.1	96.3	89.3	94.3	101.6	111.9	110.5	98.1	93.9	94.2	102.5
Mediana ajustada	108.9	95.4	96.6	89.6	94.6	101.9	112.2	110.8	98.4	94.2	94.5	102.8

Ahora dividimos cada uno de los valores mensuales de la Tabla 18.12 por los correspondientes valores de tendencia en la Tabla 18.18. Los resultados, expresados en porcentajes, se recogen en la Tabla 18.19; por ejemplo, la primera entrada de la tabla viene dada por 178.2/170.2 = 104.7%.

Como el total de las medias en la Tabla 18.19 es 1200.2, que es muy próximo a 1200%, no es necesario ajustar; en consecuencia, la tercera fila por abajo de esa tabla representa el índice estacional determinado por la media. Ya que el total de las medianas es 1196.6, debemos ajustarlas; para ello, las multiplicamos por 1200/1196.6, obteniendo así la fila de más abajo de la Tabla 18.19, que muestra el deseado índice estacional determinado esta vez por la mediana. Vemos que hay buen acuerdo entre las medias y las medianas ajustadas en la Tabla 18.19. Estos resultados coinciden asimismo con los de los Problemas 18.9(*b*) y 18.10.

18.12. Obtener un índice estacional para los datos del Problema 18.9 usando el método del promedio móvil en porcentajes.

Solución

Usando el segundo método del Problema 18.3, obtenemos primero un promedio móvil centrado de 12 meses, como muestra la Tabla 18.20. Estos resultados los recoge el gráfico de la Figura 18.5; nótese que el esquema estacional ha desaparecido, lo cual suaviza el gráfico considerablemente.

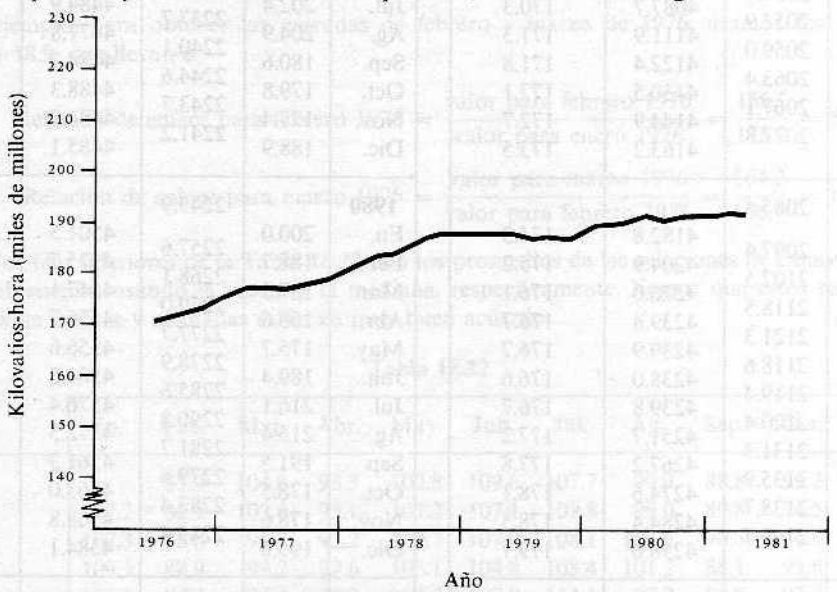


Figura 18.5. Promedio móvil centrado de 12 meses.

Ahora dividimos cada uno de los valores mensuales reales por el correspondiente promedio móvil centrado de 12 meses y expresamos cada resultado como porcentaje; para julio de 1976, por ejemplo, obtenemos 185.9/170.3 = 109.2%. Los resultados figuran en la Tabla 18.21. Obsérvese que las entradas de los 6 primeros meses de 1976 y los 6 últimos de 1981 no se obtienen por este método.

La Tabla 18.21 da el porcentaje promedio para cada mes en términos tanto de la media como de la mediana. Mientras las medias (que totalizan 1200.2) no han sido ajustadas, las medianas (que totalizan 1198.9) sí lo han sido. Así pues, la primera y tercera filas por abajo en la tabla representan los índices estacionales obtenidos usando la media y la mediana, respectivamente. Estos índices coinciden entre sí y con los obtenidos por otros métodos en problemas anteriores.

18.13. Hallar un índice estacional para los datos del Problema 18.9 por medio del método de relación de enlace.

Solución

Expresamos primero los datos de cada mes como un porcentaje de los datos del mes anterior, como muestra la Tabla 18.22. Cada uno de estos porcentajes se llama *relación de enlace*.

Tabla 18.20

Año y mes	Datos	Total móvil de 12 meses	Total móvil de 2 meses para la columna 3	Promedio móvil centrado de 12 meses (Col. 4 ÷ 24)
1976				
En.	178.2			
Feb.	156.7			
Mar.	164.2			
Abr.	153.2			
May.	157.5			
Jun.	172.6	2034.8		
Jul.	185.9	2052.9	4087.7	170.3
Ag.	185.8	2059.0	4111.9	171.3
Sep.	165.0	2063.4	4122.4	171.8
Oct.	163.6	2067.1	4130.5	172.1
Nov.	169.0	2077.8	4144.9	172.7
Dic.	183.1		4163.2	173.5
1977		2085.4		
En.	196.3	2097.4	4182.8	174.3
Feb.	162.8	2107.5	4204.9	175.2
Mar.	168.6	2118.5	4226.0	176.1
Abr.	156.9	2121.3	4239.8	176.7
May.	168.2	2118.6	4239.9	176.7
Jun.	180.2	2119.4	4238.0	176.6
Jul.	197.9	2120.4	4239.8	176.7
Ag.	195.9	2131.3	4251.7	177.2
Sep.	176.0	2135.9	4267.2	177.8
Oct.	166.4	2138.7	4274.6	178.1
Nov.	166.3	2145.7	4284.4	178.5
Dic.	183.9		4298.6	179.1
1978		2152.9		
En.	197.3	2157.6	4310.5	179.6
Feb.	173.7	2167.3	4324.9	180.2
Mar.	173.2	2176.9	4344.2	181.0
Abr.	159.7	2186.1	4363.0	181.8
May.	175.2	2196.1	4382.2	182.6
Jun.	187.4	2203.9	4400.0	183.3
Jul.	202.6	2216.1	4420.0	184.2
Ag.	205.6	2228.7	4444.8	185.2
Sep.	185.6	2238.5	4467.2	186.1
Oct.	175.6	2248.3	4486.8	187.0
Nov.	176.3	2251.3	4499.6	187.5
Dic.	191.7		4501.9	187.6

Año y mes	Datos	Total móvil de 12 meses	Total móvil de 2 meses para la columna 3	Promedio móvil centrado de 12 meses (Col. 4 ÷ 24)
1979		2250.6		
En.	209.5	2250.4	4501.0	187.5
Feb.	186.3	2249.7	4500.1	187.5
Mar.	183.0	2244.7	4494.4	187.3
Abr.	169.5	2248.9	4493.6	187.2
May.	178.2	2250.0	4498.9	187.5
Jun.	186.7	2247.2	4497.2	187.4
Jul.	202.4	2237.7	4484.9	186.9
Ag.	204.9	2240.1	4477.8	186.6
Sep.	180.6	2244.6	4484.7	186.9
Oct.	179.8	2243.7	4488.3	187.0
Nov.	177.4	2241.2	4484.9	186.9
Dic.	188.9		4485.1	186.9
1980		2243.9		
En.	200.0	2257.6	4501.5	187.6
Feb.	188.7	2268.1	4525.7	188.6
Mar.	187.5	2279.0	4547.1	189.5
Abr.	168.6	2277.7	4556.7	189.9
May.	175.7	2278.9	4556.6	189.9
Jun.	189.4	2285.6	4564.5	190.2
Jul.	216.1	2290.8	4576.4	190.7
Ag.	215.4	2281.7	4572.5	190.5
Sep.	191.5	2279.6	4561.3	190.1
Oct.	178.5	2283.4	4563.0	190.1
Nov.	178.6	2285.4	4568.8	190.4
Dic.	195.6		4584.1	191.0
1981		2298.7		
En.	205.2	2302.8	4601.5	191.7
Feb.	179.6	2297.6	4600.4	191.7
Mar.	185.4	2293.0	4590.6	191.3
Abr.	172.4	2295.9	4588.9	191.2
May.	177.7	2292.9	4588.8	191.2
Jun.	202.7	2292.9	4585.8	191.1
Jul.	202.2			
Ag.	210.2			
Sep.	186.9			
Oct.	181.4			
Nov.	175.6			
Dic.	195.6			

Tabla 18.21

	En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
1976							109.2	108.5	96.0	95.1	97.9	105.5
1977	112.6	92.9	95.7	88.8	95.2	102.0	112.0	110.6	99.0	93.4	93.2	102.7
1978	109.9	96.4	95.7	87.8	95.9	102.2	110.0	111.0	99.7	93.9	94.0	102.2
1979	111.7	99.4	97.7	90.5	95.0	99.6	108.3	109.8	96.6	96.1	94.9	101.1
1980	106.6	100.1	98.9	88.8	92.5	99.6	113.3	113.1	100.7	93.9	93.8	102.4
1981	107.0	93.7	96.9	90.2	92.9	106.1						
Media	109.6	96.5	97.0	89.2	94.3	101.9	110.6	110.6	98.4	94.5	94.8	102.8
Mediana	109.9	96.4	96.9	88.8	95.0	102.0	110.0	110.6	99.0	93.9	94.0	102.4
Media ajustada	110.0	96.5	97.0	88.9	95.1	102.1	110.1	110.7	99.1	94.0	94.1	102.5

Por ejemplo, para obtener las entradas de febrero y marzo de 1976, usamos los datos del Problema 18.9, que llevan a

$$\text{Relación de enlace para febrero 1976} = \frac{\text{valor para febrero 1976}}{\text{valor para enero 1976}} = \frac{156.7}{178.2} = 87.9\%$$

$$\text{Relación de enlace para marzo 1976} = \frac{\text{valor para marzo 1976}}{\text{valor para febrero 1976}} = \frac{164.2}{156.7} = 104.8\%$$

Las dos filas inferiores de la Tabla 18.22 dan los promedios de las relaciones de enlace para cada mes que obtenemos usando la media y la mediana, respectivamente. Vemos que estos resultados procedentes de medias y medianas están en muy buen acuerdo.

Tabla 18.22

	En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
1976		87.9	104.8	93.3	102.8	109.6	107.7	99.9	88.8	99.2	103.3	108.3
1977	107.2	82.9	103.6	93.1	107.2	107.1	109.8	99.0	89.8	94.5	99.9	110.6
1978	107.3	88.0	99.7	92.2	109.7	107.0	108.1	101.5	90.3	94.6	100.4	108.7
1979	109.3	88.9	98.2	92.6	105.1	104.8	108.4	101.2	88.1	99.6	98.7	106.5
1980	105.9	94.4	99.4	89.9	104.2	107.8	114.1	99.7	88.9	93.2	100.1	109.5
1981	104.9	87.5	103.2	93.0	103.1	114.1	108.6	95.5	88.9	97.1	96.8	111.4
Media	106.9	88.3	101.5	92.4	105.4	108.4	109.5	99.5	89.1	96.4	99.9	109.2
Mediana	107.2	88.0	101.5	92.8	104.7	107.5	108.5	99.8	88.9	95.9	100.0	109.1

Consideremos el de enero como valor 100% (véase Tabla 18.23). Como la relación de enlace promedio para febrero es 88.0 (usando el valor de la mediana en la Tabla 18.23), los datos para febrero son, en promedio, el 88.0% de los datos de enero (o sea, 88.0% de 100.0 = 88.0); análogamente, la relación de enlace promedio para marzo es 101.5% del de febrero (o sea, 101.5% de 88.0 = 89.3); etc. Continuando de este modo llegamos a la Tabla 18.23, cuyas entradas se llaman a veces *relaciones en cadena*.

En el lado derecho de la Tabla 18.23, el resultado para el segundo enero es 100.7, un crecimiento de 0.7 sobre el primer enero. Este crecimiento se debe a la tendencia a largo plazo en los datos. Con el fin de ajustar a dicha tendencia, hemos de restar (12/12)(0.7) = .0.7 del 100.7 del segundo enero (para lograr 100.0), restar (11/12)(0.7) = 0.64 del valor de diciembre, (10/12)(0.7) = 0.58 del de noviembre, etc. Los valores ajustados a la tendencia se muestran en la Tabla 18.24. [Estrictamente hablando, habría que *multiplicar* las entradas de derecha a izquierda, respectivamente, por $(100.0/100.7)^{12/12}$, $(100.0/100.7)^{11/12}$, $(100.0/100.7)^{10/12}$, etc. esto, sin embargo, conduce prácticamente a los mismos resultados que los de la Tabla 18.24.] Como los porcentajes de la Tabla 18.24 suman 1094.5 en total, los ajustamos multiplicando cada porcentaje por 1200/1094.5, con lo que ya se obtiene el índice estacional recogido en la Tabla 18.25.

Tabla 18.23

En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.	En.
100.0	88.0	89.3	82.9	86.8	93.3	101.2	101.0	89.8	86.1	86.1	93.9	100.7

Tabla 18.24

En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
100.0	87.9	89.2	82.7	86.6	93.0	100.8	100.6	89.3	85.6	85.5	93.3

Tabla 18.25

	En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
Indice estacional	109.6	96.4	97.8	90.7	94.9	102.0	110.5	110.3	97.9	93.9	93.7	102.3

18.14. Construir una tabla de comparación para los índices estacionales hallados por los diversos métodos de los Problemas 18.9, 18.11, 18.12 y 18.14.

Solución

Véase la Tabla 18.26 que muestra los índices estacionales obtenidos usando la mediana.

Tabla 18.26

Método	En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
Porcentaje promedio (Problema 18.9)	107.6	94.4	97.0	90.1	94.1	102.1	111.4	110.7	98.9	95.4	94.6	103.6
Relación con la tendencia (Problema 18.11)	108.9	95.4	96.6	89.6	94.6	101.9	112.2	110.8	98.4	94.2	94.5	102.8
Relación con el promedio anual (Problema 18.12)	110.0	96.5	97.0	88.9	95.1	102.1	110.1	110.7	99.1	94.0	94.1	102.5
Relación de enlace (Problema 18.13)	109.6	96.4	97.8	90.7	94.9	102.0	110.5	110.3	97.9	93.9	93.7	102.3

DATOS AJUSTADOS A LA VARIACION ESTACIONAL

18.15. Ajustar los datos del Problema 18.9 a la variación estacional, es decir, desestacionalizar los datos.

Solución

Para ajustar los datos a la variación estacional, hemos de dividir todas las entradas en los datos originales del Problema 18.9 por el índice estacional del mes correspondiente, hallado por alguno de los métodos expuestos. Por ejemplo, si se usa el índice estacional del Problema 18.12, hay que dividir todos los valores de enero por 110.0% (o sea, 1.100), todos los valores de febrero por 96.5% (o sea, 0.965), etc. Los datos ajustados que resultan se recogen en la Tabla 18.27.

Tabla 18.27

	En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
1976	160.5	162.4	169.3	172.3	165.6	169.0	168.8	167.8	166.5	174.0	179.6	178.6
1977	178.5	168.7	173.8	176.5	176.9	176.5	179.7	177.0	177.6	177.0	176.7	179.4
1978	179.4	180.0	178.6	179.6	184.2	183.5	184.0	185.7	187.3	186.8	187.4	187.0
1979	190.5	193.1	188.7	190.7	187.4	182.9	183.8	185.1	182.2	191.3	188.5	184.3
1980	181.8	195.5	193.3	189.7	184.8	185.5	196.3	194.6	193.2	189.9	189.8	190.8
1981	186.5	186.1	191.1	193.9	186.9	198.5	200.0	189.9	188.6	193.0	186.6	190.8

18.16. (*a*) Representar en un gráfico los datos desestacionalizados del Problema 18.15.
(*b*) Comparar este gráfico con la Figura 18.4 del Problema 18.9(*a*).

Solución

(*a*) Vésea Figura 18.6.
(*b*) El gráfico de los datos ajustados a la variación estacional muestra la tendencia a largo término, que, aparte sus fluctuaciones, se aproxima a una recta. Si denotamos los datos del Problema 18.9 por $Y = TCSI$, el gráfico de la Figura 18.6 es el de la variable $Y/S = TCI$ en función del tiempo t y contiene la tendencia a largo término, los movimientos cíclicos y los irregulares. Como el gráfico indica una tendencia a largo término con escasa influencia de tipo cíclico e irregular, parece que el producto CI de los factores cíclico e irregular debe ser cercano al 100%. (El Problema 18.18 confirma esta sospecha.)

ESTIMACION DE VARIACIONES CICLICAS E IRREGULARES

18.17. Ajustar los datos del Problema 18.16 a la tendencia.

Solución

Para eliminar la tendencia de los datos del Problema 18.16, dividimos cada entrada por el valor de tendencia mensual correspondiente, calculado por cualquiera de los métodos precedentes. Usemos los valores hallados en el Problema 18.12 por el método de promedios móviles. Los resultados se indican en la Tabla 18.28. Para obtener la entrada de julio de 1976, por ejemplo, dividimos la correspondiente entrada de la Tabla 18.27, que es 168.8, por el valor 170.3 (véase la primera entrada de la columna 5 en la Tabla 18.20), lo cual da $168.8/170.3 = 99.1\%$. Las entradas restantes se hallan de forma similar. Una desventaja de este método, y de cuantos métodos manejan promedios móviles, es que los datos de los extremos de la serie en el tiempo se pierden.

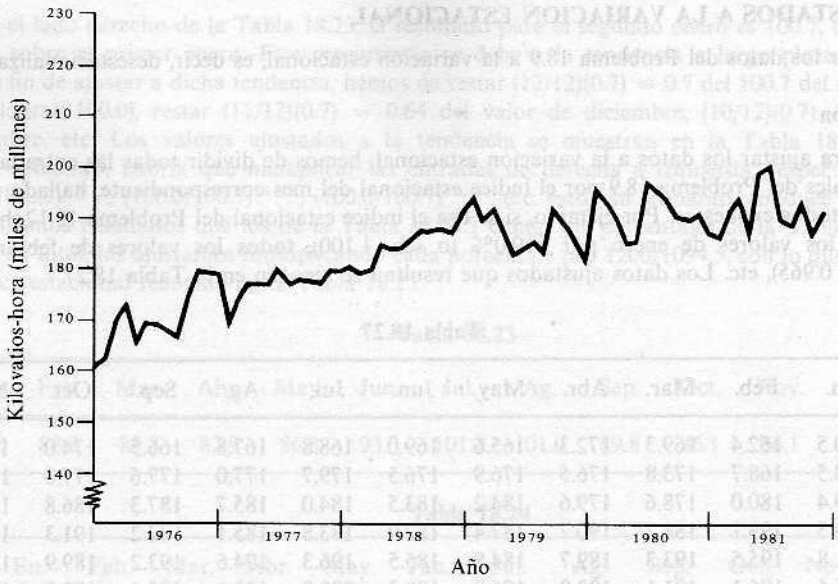


Figura 18.6. Datos ajustados a la estación.

18.18. (*a*) Representar en un gráfico los datos obtenidos en el Problema 18.17.
(*b*) Explicar el significado de ese gráfico.

Solución

(*a*) Conviene restar 100(%) de los datos del Problema 18.17 y hacer el gráfico de las desviaciones resultantes. Este gráfico, en una escala vertical muy aumentada, se puede ver en la Figura 18.7.

(*b*) Los datos originales se representan por $Y = TCSI$. Ajustando a la variación estacional (como en el Prob. 18.15) mediante división de ambos lados por el índice estacional S, se obtiene $Y/S = TCI$. El ajuste subsiguiente a la tendencia exige dividir por T, con lo que se obtiene $Y/ST = CI$. Restando 100(%) queda $(Y/ST) - 100 = CI - 100$. Así pues, la variable dependiente en la Figura 18.7 es $(Y/ST) - 100$, y la independiente es el tiempo t.

El gráfico de la Figura 18.7 consta teóricamente sólo de los movimientos cíclicos e irregulares C e I. Nótese que el producto CI varía entre 96% y 104%, confirmando la conclusión a la que habíamos llegado en el Problema 18.16(*b*).

Tabla 18.28

	En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
1976							99.1	98.0	96.9	101.1	104.0	102.9
1977	102.4	96.3	98.7	99.9	100.1	100.0	101.7	99.9	99.9	99.4	99.0	100.2
1978	99.9	99.9	98.7	98.8	100.9	100.1	99.9	100.3	100.6	99.6	99.9	99.7
1979	101.6	103.0	100.7	101.9	99.9	97.6	98.3	99.2	97.5	102.3	100.9	98.6
1980	96.9	103.7	102.0	99.9	97.3	97.5	102.9	102.2	101.6	99.9	99.7	99.9
1981	97.3	97.1	99.9	101.4	97.8	103.9						

18.19. (*a*) Hallar los promedios móviles de 3 y 7 meses para los datos del Problema 18.17.
(*b*) Construir los gráficos de los promedios móviles de la parte (*a*).
(*c*) Interpretar los gráficos.

Solución

(*a*) Los promedios móviles pedidos se muestran en la Tabla 18.29.
(*b*) Los gráficos de los promedios móviles de 3 y 7 meses pueden verse en las Figuras 18.8 y 18.9, respectivamente.

Tabla 18.29

Año y mes	Datos	Total móvil de 3 meses	Promedio móvil de 3 meses	Total móvil de 7 meses	Promedio móvil de 7 meses
1976					
Jul.	91.1				
Ag.	98.0	294.0	98.0		
Sep.	96.9	296.0	98.7		
Oct.	101.1	302.0	100.7	704.4	100.6
Nov.	104.0	308.0	102.7	701.6	100.2
Dic.	102.9	309.3	103.1	702.3	100.3
1977					
En.	102.4	301.6	100.5	705.3	100.8
Feb.	96.3	297.4	99.1	704.3	100.6
Mar.	98.7	294.9	98.3	700.3	100.0
Abr.	99.9	298.7	99.6	699.1	99.9
May.	100.1	300.0	100.0	704.7	100.7
Jun.	100.0	301.8	100.6	700.2	100.0
Jul.	101.7	301.6	100.5	700.9	100.1
Ag.	99.9	301.5	100.5	700.0	100.0
Sep.	99.9	299.2	99.7	700.1	100.0
Oct.	99.4	298.3	99.4	700.0	100.0
Nov.	99.0	298.6	99.5	698.2	99.7
Dic.	100.2	299.1	99.7	697.0	99.6
1978					
En.	99.9	300.0	100.0	695.9	99.4
Feb.	99.9	298.5	99.5	697.4	99.6
Mar.	98.7	297.4	99.1	698.5	99.8
Abr.	98.8	298.4	99.5	698.2	99.7
May.	100.9	299.8	99.9	698.6	99.8
Jun.	100.1	300.9	100.3	699.3	99.9
Jul.	99.9	300.3	100.1	700.5	100.1
Ag.	100.3	300.8	100.3	701.6	100.2
Sep.	100.6	300.8	100.3	700.4	100.1
Oct.	99.9	300.4	100.1	701.9	100.3
Nov.	99.9	299.5	99.8	705.0	100.7
Dic.	99.7	301.2	100.4	705.4	100.8

Tabla 18.29. (*Continuación*)

Año y mes	Datos	Total móvil de 3 meses	Promedio móvil de 3 meses	Total móvil de 7 meses	Promedio móvil de 7 meses
1979					
En.	101.6	304.3	101.4	706.7	101.0
Feb.	103.0	305.3	101.8	706.7	101.0
Mar.	100.7	305.6	101.9	704.4	100.6
Abr.	101.9	302.5	100.8	703.0	100.4
May.	99.9	299.4	99.8	700.6	100.1
Jun.	97.6	295.8	98.6	695.1	99.3
Jul.	98.3	295.1	98.4	696.7	99.5
Ag.	99.2	295.0	98.3	695.7	99.4
Sep.	97.5	299.0	99.7	694.4	99.2
Oct.	102.3	300.7	100.2	693.7	99.1
Nov.	100.9	301.8	100.6	699.1	99.9
Dic.	98.6	296.4	98.8	701.9	100.3
1980					
En.	96.9	299.2	99.7	704.3	100.6
Feb.	103.7	302.6	100.9	699.3	99.9
Mar.	102.0	305.6	101.9	695.9	99.4
Abr.	99.9	299.2	99.7	700.2	100.0
May.	97.3	294.7	98.2	705.5	100.8
Jun.	97.5	297.7	99.2	703.4	100.5
Jul.	102.9	302.6	100.9	701.3	100.2
Ag.	102.2	306.7	102.2	701.1	100.2
Sep.	101.6	303.7	101.2	703.7	100.5
Oct.	99.9	301.2	100.4	703.5	100.5
Nov.	99.7	299.5	99.8	697.7	99.7
Dic.	99.9	296.9	99.0	695.4	99.3
1981					
En.	97.3	294.3	98.1	695.2	99.3
Feb.	97.1	294.3	98.1	693.1	99.0
Mar.	99.9	298.4	99.5	696.8	99.5
Abr.	101.4	299.1	99.7		
May.	97.8	303.1	101.0		
Jun.	103.9				

(*c*) Tal como era de esperar, los promedios móviles sirven para suavizar las irregularidades de los datos del Problema 18.17, como se ve sin más que comparar las Figuras 18.8 y 18.9 con la Figura 18.7. Es claro además de los gráficos que el promedio móvil de 7 meses proporciona un mejor suavizamiento de los datos en este caso que el de 3 meses. Todas las fluctuaciones para el promedio móvil de 3 meses son menores que un 3% aproximadamente, mientras que para el caso del de 7 meses están por debajo del 1%.

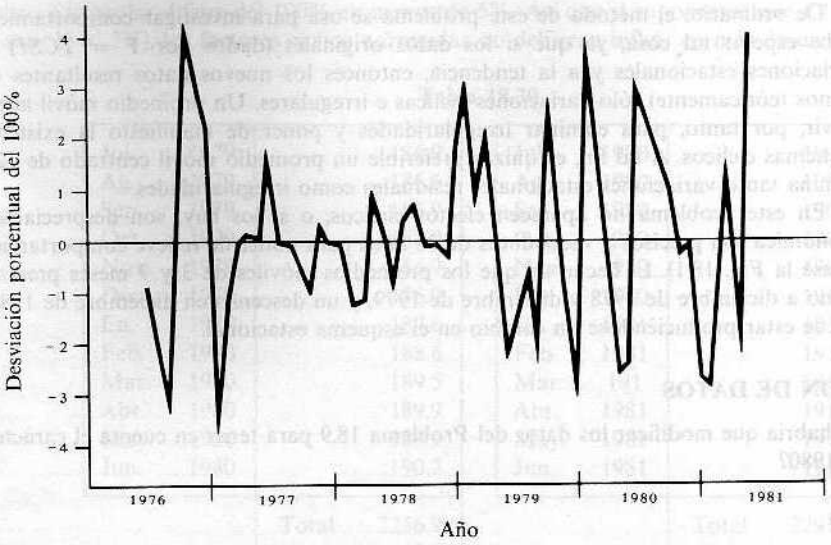


Figura 18.7. Variaciones cíclicas e irregulares.

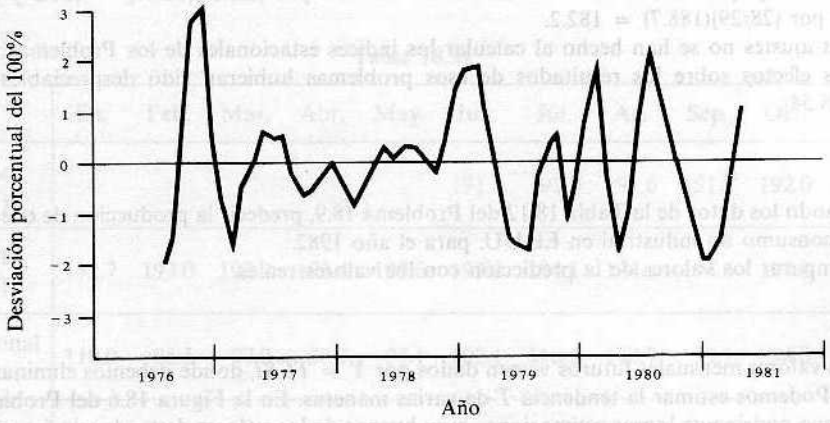


Figura 18.8. Promedio móvil de 3 meses.

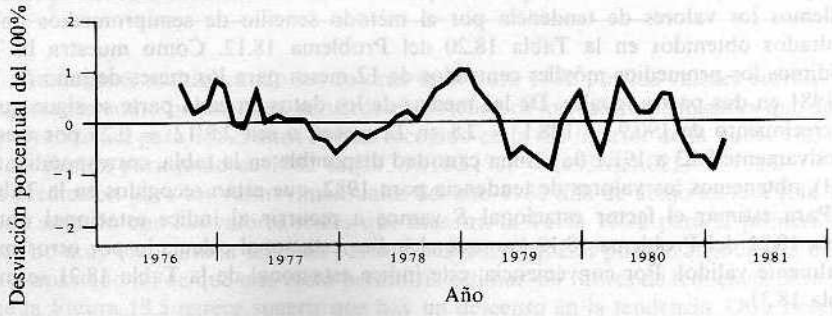


Figura 18.9. Promedio móvil de 7 meses.

De ordinario, el método de este problema se usa para investigar comportamientos cíclicos. Cabe esperar tal cosa, ya que si los datos originales (dados por $Y = TCSI$) se ajustan a variaciones estacionales y a la tendencia, entonces los nuevos datos resultantes contienen (al menos teóricamente) sólo variaciones cíclicas e irregulares. Un promedio móvil adecuado puede servir, por tanto, para eliminar irregularidades y poner de manifiesto la existencia o no de esquemas cíclicos. A tal fin, es quizás preferible un promedio móvil centrado de 12 meses, pues elimina tanto variaciones estacionales residuales como irregularidades.

En este problema no aparecen efectos cíclicos; o si los hay, son despreciables. En teoría económica son precisos a veces datos de 20 años para poner de relieve comportamientos cíclicos (véase la Fig. 18.1). El hecho de que los promedios móviles de 3 y 7 meses presenten picos en torno a diciembre de 1978 y diciembre de 1979, y un descenso en diciembre de 1980, indica que puede estar produciéndose un cambio en el esquema estacional.

COMPARACION DE DATOS

18.20. ¿Cómo habría que modificar los datos del Problema 18.9 para tener en cuenta el carácter bisiesto de 1976 y 1980?

Solución

En un año bisiesto, febrero tiene 29 días en vez de los 28 habituales. Para lograr la comparabilidad, debemos multiplicar los datos de un año bisiesto por 28/29. Así pues, en la Tabla 18.12 del Problema 18.9 hay que sustituir el valor de febrero de 1976 por (28/29)(156.7) = 151.3 y el de febrero de 1980 por (28/29)(188.7) = 182.2.

Estos ajustes no se han hecho al calcular los índices estacionales de los Problemas 18.9 a 18.13, pero sus efectos sobre los resultados de esos problemas hubieran sido despreciables (véase Problema 18.54).

PREDICCION

18.21. (*a*) Usando los datos de la Tabla 18.12 del Problema 18.9, predecir la producción de energía eléctrica de consumo no industrial en EE.UU. para el año 1982.

(*b*) Comparar los valores de la predicción con los valores reales.

Solución

(*a*) Los valores mensuales futuros vienen dados por $Y = TCSI$, donde debemos eliminar T, C, S e I.

Podemos estimar la tendencia T de varias maneras. En la Figura 18.6 del Problema 18.16 se ve que podríamos lograr estimaciones muy buenas de los valores de tendencia futuros ajustando una recta a los valores de tendencia de los dos últimos años, por ejemplo. Podríamos hacer tal cosa por mínimos cuadrados o por cualquier otro de los métodos discutidos en este capítulo. Hallemos los valores de tendencia por el método sencillo de semipromedios aplicado a los resultados obtenidos en la Tabla 18.20 del Problema 18.12. Como muestra la Tabla 18.30, dividimos los promedios móviles centrados de 12 meses para los meses de julio de 1979 a junio de 1981 en dos partes iguales. De las medias de los datos en cada parte se sigue que ha habido un crecimiento de 190.9 − 188.1 = 2.8 en 12 meses, o sea 2.8/12 = 0.23 por mes; añadiendo sucesivamente 0.23 a 191.1 (la última cantidad disponible en la tabla, correspondiente a junio de 1981), obtenemos los valores de tendencia para 1982, que están recogidos en la Tabla 18.31.

Para estimar el factor estacional S vamos a recurrir al índice estacional obtenido en la Tabla 18.21 del Problema 18.12 (aunque el índice estacional calculado por otros métodos sería igualmente válido). Por conveniencia, este índice estacional de la Tabla 18.21 se muestra en la Tabla 18.31.

Vemos de la Figura 18.8 del Problema 18.19 que el producto estimado CI de los factores

cíclico e irregular difiere del 100% en menos de 5%. Así que si suponemos que $CI = 100\% = 1$ (o sea, $Y = TS$), los factores cíclico e irregular no debieran influir en más de un 5% en Y.

Tabla 18.30

Jul.	1979		186.9	Jul.	1980		190.7
Ag.	1979		186.6	Ag.	1980		190.5
Sep.	1979		186.9	Sep.	1980		190.1
Oct.	1979		187.0	Oct.	1980		190.1
Nov.	1979		186.9	Nov.	1980		190.4
Dic.	1979		186.9	Dic.	1980		191.0
En.	1980		187.6	En.	1981		191.7
Feb.	1980		188.6	Feb.	1981		191.7
Mar.	1980		189.5	Mar.	191		191.3
Abr.	1980		189.9	Abr.	1981		191.2
May.	1980		189.9	May.	1981		191.2
Jun.	1980		190.2	Jun.	1981		191.1
		Total	2256.9			Total	2291.0
		Media	188.1			Media	190.9

Tabla 18.31

	En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
1981 valor de tendencia (T)						191.1	191.3	191.6	191.8	192.0	192.3	192.5
1982 valor de tendencia (T)	192.7	193.0	193.2	193.4	193.6	193.9	194.1	194.3	194.6	194.8	195.0	195.2
Indice estacional ($S\%$)	110.0	96.5	97.0	88.9	95.1	102.1	110.1	110.7	99.1	94.0	94.1	102.5
Predicción de energía para 1982 (TS)	212.0	186.2	187.4	171.9	184.1	198.0	213.7	215.1	192.8	183.1	183.5	200.1

Finalmente, multiplicando los valores de T para 1982 por los valores correspondientes de S (expresados como porcentajes, recuérdese), obtenemos los valores mensuales que da la predicción, o *proyecciones*, para 1982; estos se han recogido en la fila inferior de la Tabla 18.31. Por ejemplo, la predicción para junio de 1982 es (193.9)(102.1%) = (193.9)(1.021) = 198.0

(*b*) La predicción para los valores mensuales del año 1982 (fila de abajo en la Tabla 18.31) están en buen acuerdo con los valores reales que muestra la Tabla 18.32 para la primera parte de 1982, pero no muy bien para la segunda parte. Estas discrepancias pueden atribuirse a nuestra hipótesis en el apartado (*a*) de que una recta permitiría estimar los vaores de tendencia para 1982, mientras que la Figura 18.5 parece sugerir que hay un descenso en la tendencia. Otra potencial fuente de error es un posible cambio en el esquema estacional (véase la nota al final del Prob. 18.19).

Tabla 18.32

	En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
1982 energía real	210.1	180.3	187.7	172.6	177.1	186.1	210.6	205.7	180.7	173.0	173.4	184.7

Fuente: Survey of Current Business.

Podemos mejorar la precisión de la predicción usando una *parábola de mínimos cuadrados* (véanse Probs. 18.40 y 18.67) para ajustar los promedios mensuales en la Tabla 18.13 del Problema 18.9. La Tabla 18.33 presenta los valores de predicción obtenidos mediante una parábola de mínimos cuadrados y también los valores reales para 1982. Los resultados son mejores que los dados en la Tabla 18.31, ya que como enseña la fila inferior de la Tabla 18.33, los errores no superan, en porcentaje, el 5%.

Tabla 18.33

	En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
1982 potencia real	210.1	180.3	187.7	172.6	177.1	186.1	210.6	205.7	180.7	173.0	173.4	184.7
1982 predicción de potencia	210.1	184.3	185.0	169.6	181.3	194.5	209.6	210.6	188.4	178.2	178.3	194.0
Porcentaje de error	0.0%	2.2%	1.4%	1.7%	2.4%	4.5%	0.5%	2.4%	4.3%	3.2%	3.0%	5.0%

PROBLEMAS SUPLEMENTARIOS

MOVIMIENTOS CARACTERISTICOS DE SERIES EN EL TIEMPO

18.22. ¿Con qué movimientos característicos de una serie en el tiempo están asociados (*a*) una recesión, (*b*) un decrecimiento estival del paro, (*c*) el descenso de la mortalidad debido a los avances de la Medicina, (*d*) una huelga en la metalurgia y (*e*) una demanda continuamente creciente de automóviles utilitarios?

PROMEDIOS MOVILES

18.23. Dados los números 1, 0, −1, 0, 1, 0, −1, 0 y 1, determinar un promedio móvil de orden (*a*) 2, (*b*) 3, (*c*) 4 y (*d*) 5.

18.24. Probar que si una sucesión de números tiene período N (es decir, la sucesión se repite tras N términos), todo promedio móvil de orden menor que N tiene período N. Ilustrar esto haciendo referencia al Problema 18.23.

18.25. (*a*) En el Problema 18.24, ¿qué ocurre en el caso de un promedio móvil de orden N?

(*b*) ¿Qué ocurre si el orden es mayor que N? Ilustrar esto mediante el Problema 18.23.

18.26. Probar que si todos los números de una sucesión se aumentan (o disminuyen) en una constante, el promedio móvil también aumenta (o disminuye) en esa misma constante.

18.27. Demostrar que si todo número de una sucesión se multiplica (o divide) por una constante no nula, el promedio móvil queda también multiplicado (dividido) por esa constante.

18.28. Hallar el promedio móvil ponderado de los números en el Problema 18.23, partes (*b*), (*c*) y (*d*), con pesos respectivos (*b*) 1, 2 y 1; (*c*) 1, 2, 2 y 1; y (*d*) 1, 2, 2, 2 y 1. Comparar los resultados con los del Problema 18.23.

18.29. (*a*) Probar las propiedades de los Problemas 18.26 y 18.27 para promedios móviles ponderados.

(*b*) ¿Es válido el resultado del Problema 18.24 para promedios móviles ponderados?

18.30. Una sucesión tiene (*a*) 24, (*b*) 25 y (*c*) 200 números. ¿Cuántos números habrá en un promedio móvil de orden 5?

18.31. Una sucesión tiene M números.

(*a*) Probar que en un promedio móvil de orden N habrá $M - N + 1$ números. Ilustrar esto con varios ejemplos, usando distintos valores de M y N.

(*b*) Discutir el caso $M = N$.

18.32. La Tabla 18.34 muestra la producción mensual media (en miles) en EE.UU. de automóviles para los años 1976-1985. Construir (*a*) un promedio móvil de 2 años, (*b*) un promedio móvil centrado de 2 años, (*c*) un promedio móvil de 3 años, (*d*) un promedio móvil centrado de 4 años y (*e*) un promedio móvil centrado de 6 años.

Tabla 18.34

Año	Promedio mensual de producción de automóviles en EE.UU. (miles)
1976	708
1977	767
1978	764
1979	702
1980	533
1981	521
1982	421
1983	562
1984	635
1985	667

Fuente: Survey of Current Business.

18.33. Representar en un gráfico los promedios móviles del Problema 18.32 junto con los datos originales, y discutir los resultados obtenidos.

18.34. (*a*) Probar que el promedio móvil centrado de 2 años del Problema 18.32(*b*) es equivalente a un promedio móvil ponderado de 3 años con pesos respectivos 1, 2 y 1. Ilustrar esto mediante cálculos directos.

(*b*) Probar que el promedio móvil centrado de 6 años del Problema 18.32(*e*) equivale a un promedio móvil ponderado apropiado.

18.35. (*a*) Para los datos del Problema 18.32, determinar un promedio móvil ponderado de orden 3 con los pesos 1, 4 y 1.

(*b*) Representar este promedio móvil y comparar con los resultados del Problema 18.32(*c*).

18.36. La Tabla 18.35 presenta la producción total (en millones de libras) de todos los tipos de queso en los años 1983-1985. Construir (*a*) un promedio móvil de 12 meses, (*b*) un promedio móvil centrado de 12 meses y (*c*) un promedio móvil centrado de 6 meses. En las partes (*b*) y (*c*), representar en un gráfico el promedio móvil junto con los datos originales, y comparar los resultados.

Tabla 18.35

	1983	1984	1985
En.	375	387	391
Feb.	353	369	355
Mar.	417	413	412
Abr.	408	415	430
May.	429	437	456
Jun.	436	420	442
Jul.	401	388	439
Ag.	381	369	423
Sep.	375	349	400
Oct.	392	378	428
Nov.	388	372	412
Dic.	420	392	438

Fuente: Survey of Current Business.

ESTIMACION DE LA TENDENCIA

18.37. Obtener, por el método de semipromedios, los valores de tendencia para los datos del Problema 18.32, tomando como promedio (*a*) la media y (*b*) la mediana. Dibujar un gráfico que ilustre los resultados alcanzados.

18.38. Rehacer el Problema 18.32 usando (*a*) el método «a mano» y (*b*) un promedio móvil de orden adecuado. Comparar los resultados con los del Problema 18.37.

18.39. (*a*) Usar el método de mínimos cuadrados para ajustar una recta a los datos del Problema 18.32.

(*b*) A partir del resultado de la parte (*a*), hallar los valores de tendencia y compararlos con los de los Problemas 18.35 y 18.36.

18.40. (*a*) Ajustar una parábola $Y = a_0 + a_1X + a_2X^2$ a los datos del Problema 18.9, usando los promedios mensuales de la Tabla 18.13 del Problema 18.10.

(*b*) Comparar los resultados del apartado (*a*) con los de la recta de mínimos cuadrados del Problema 18.10, y calcular los valores de tendencia.

18.41. Hallar valores de tendencia para los datos del Problema 18.36 usando (*a*) el método de semipromedios, (*b*) el método a mano, (*c*) un promedio móvil centrado de 12 meses y (*d*) una curva adecuada de mínimos cuadrados (para determinarla, úsese el gráfico de los datos originales construido en el Problema 18.36). Discutir las ventajas y desventajas de cada método.

ESTIMACION DE LAS VARIACIONES ESTACIONALES; EL INDICE ESTACIONAL

18.42. La Tabla 18.36 muestra la producción mensual en EE.UU. (en miles) de acondicionadores de aire durante los años 1980-1985.

(*a*) Hacer un gráfico de los datos.

(*b*) Hallar un índice estacional por medio del método del porcentaje medio. Antes de hallar ese índice, ajustar los datos para tener en cuenta los años bisiestos.

Tabla 18.36

	1980	1981	1982	1983	1984	1985
En.	203	222	191	88	168	209
Feb.	342	366	361	130	262	277
Mar.	434	623	572	309	504	530
Abr.	416	603	517	259	459	524
May.	344	477	419	300	588	632
Jun.	316	653	289	265	648	416
Jul.	566	283	145	306	187	171
Ag.	94	64	61	108	47	68
Sep.	66	52	17	58	40	49
Oct.	90	90	31	32	36	24
Nov.	125	94	71	52	51	39
Dic.	203	163	84	98	113	113

Fuente: Survey of Current Business.

18.43. Hallar un índice estacional para los datos del Problema 18.42 usando el método del porcentaje de tendencia. Para obtener los valores de tendencia, ajustar una curva adecuada de mínimos cuadrados a los promedios mensuales de los años dados.

18.44. Obtener un índice estacional para los datos del Problema 18.42 mediante el método del promedio móvil en porcentaje.

18.45. Hallar un índice estacional para los datos del Problema 18.42 por el método del enlace relativo.

18.46. Comparar los resultados obtenidos en los Problemas 18.42 al 18.45.

18.47. La Tabla 18.37 presenta la producción en EE.UU. de papel de prensa (en miles de toneladas cortas) durante los años 1980-1985.

(*a*) Construir un gráfico de los datos.

(*b*) Hallar un índice estacional por el método del porcentaje medio.

Tabla 18.37

	1980	1981	1982	1983	1984	1985
En.	343	379	415	403	417	425
Feb.	334	356	378	378	410	406
Mar.	358	399	420	406	434	443

Tabla 18.37. (*Continuación*)

	1980	1981	1982	1983	1984	1985
Abr.	339	391	396	364	422	387
May.	368	402	385	399	436	418
Jun.	356	404	383	372	424	408
Jul.	341	405	363	378	409	416
Ag.	374	426	372	419	426	414
Sep.	353	400	353	377	415	405
Oct.	377	420	406	406	426	407
Nov.	358	412	373	414	417	397
Dic.	338	359	330	372	389	398

Fuente: Survey of Current Business.

18.48. Rehacer el Problema 18.47 por el método del porcentaje de tendencia.

18.49. Rehacer el Problema 18.47 por el método del promedio móvil en porcentaje.

18.50. Rehacer el Problema 18.47 por el método del enlace relativo.

18.51. Comparar los índices estacionales obtenidos en los Problemas 18.47 al 18.50.

18.52. Obtener un índice estacional para el Problema 18.36 usando dos métodos y comparar los resultados.

18.53. (*a*) Para los datos del Problema 18.9, calcular un índice estacional para los 3 últimos años, usando el método que se desee.
(*b*) Comparar los dos índices obtenidos en el apartado (*a*).

18.54. Ajustando sus datos para tener en cuenta los años bisiestos, rehacer los Problemas 18.42 al 18.45. Determinar si el ajuste tiene influencia significativa sobre el índice estacional finalmente obtenido.

DATOS AJUSTADOS A LA VARIACION ESTACIONAL

18.55. (*a*) Ajustar los datos del Problema 18.42 a las variaciones estacionales, usando cualquiera de los índices estacionales calculados en los Problemas 18.42 a 18.45.
(*b*) Dibujar un gráfico con los datos así ajustados y explicar los resultados obtenidos.

18.56. (*a*) Ajustar los datos del Problema 18.47 a las variaciones estacionales, usando cualquiera de los resultados de los Problemas 18.47 a 18.51.
(*b*) Dibujar un gráfico con los datos así ajustados y explicar los resultados obtenidos.

18.57. (*a*) Ajustar los datos del Problema 18.36 a las variaciones estacionales, usando los dos índices estacionales obtenidos en el Problema 18.52.
(*b*) Dibujar un gráfico con los datos así ajustados y explicar los resultados obtenidos.

ESTIMACION DE VARIACIONES CICLICAS E IRREGULARES

18.58. (*a*) Ajustar a la tendencia los datos del Problema 18.55, usando cualquier método.
(*b*) Dibujar en un gráfico los datos así obtenidos.
(*c*) Tomar promedios móviles de 3 y 7 meses para los datos de la parte (*a*).
(*d*) Representar en un gráfico los resultados de la parte (*c*) y explicar la variación observada. En particular, identificar cualquier movimiento cíclico que esté presente.

18.59. Rehacer el Problema 18.58 para los datos del Problema 18.56.

18.60. Rehacer el Problema 18.58 para los datos del Problema 18.57.

18.61. En la Tabla 18.38 puede verse la producción media mensual en EE.UU. de papel de prensa (en miles de toneladas cortas) en los años 1960-1985.
(*a*) Hacer un gráfico con los datos.
(*b*) Una vez analizados los datos, discutir si hay evidencia de ciclos en ellos.

Tabla 18.38

Año	Promedio mensual
1960	170
1961	174
1962	179
1963	185
1964	188
1965	182
1966	201
1967	218
1968	245
1969	269
1970	276
1971	275
1972	285

Fuente: Survey of Current Business.

Año	Promedio mensual
1973	286
1974	290
1975	297
1976	307
1977	323
1978	317
1979	342
1980	353
1981	396
1982	381
1983	391
1984	419
1985	410

Fuente: Survey of Current Business.

18.62. Al ajustar los datos a la tendencia y a las variaciones estacionales, ¿importa cuál de esos ajustes se hace primero? Incluir en la respuesta (*a*) una discusión teórica y (*b*) una ilustración que emplee la serie en el tiempo de los Problemas 18.42, 18.47 ó 18.53.

18.63. (*a*) Resolver el Problema 18.19 usando un promedio móvil centrado de 12 meses y construir el gráfico.

(*b*) ¿Qué conclusiones se sacan de los resultados del apartado (*a*)?

18.64. (*a*) Obtener una distribución de frecuencias para las magnitudes de las variaciones irregulares halladas en los Problemas 18.17 y 18.18.

(*b*) ¿Se aproxima la distribución hallada en (*a*) a una distribución normal? En caso afirmativo, dar una razón de que tal cosa suceda.

PREDICCION

18.65. (*a*) Predecir, a la vista de los resultados del Problema 18.42, la producción de acondicionadores de aire para 1986.

(*b*) Discutir posibles fuentes de error.

(*c*) Comparar la predicción con los valores reales para 1986 que se recogen en la Tabla 18.39.

Tabla 18.39

En.	131
Feb.	175
Mar.	422
Abr.	456
May.	451
Jun.	427
Jul.	361
Ag.	89
Sep.	89
Oct.	53
Nov.	56
Dic.	77

Fuente: Survey of Current Business.

18.66. (*a*) Predecir, a la vista de los resultados del Problema 18.47, la producción de papel de prensa para 1986.

(*b*) Discutir posibles fuentes de error.

(*c*) Comparar la predicción con los valores reales para 1986 que se recogen en la Tabla 18.40.

(*d*) ¿Ayuda el uso de los datos extra del Problema 18.61? Explicar la respuesta.

Tabla 18.40

En.	420
Feb.	394
Mar.	444
Abr.	409
May.	446
Jun.	420
Jul.	433
Ag.	441
Sep.	420
Oct.	426
Nov.	429
Dic.	428

Fuente: Survey of Current Business.

18.67. Usar la parábola de mínimos cuadrados del Problema 18.40 para obtener los datos para 1982 en el Problema 18.9, y comparar los valores dados por la predicción con los valores reales que se ven en la Tabla 18.32 del Problema 18.21.

18.68. La Tabla 18.41 muestra la producción (en millones de libras) de mantequilla en EE.UU. durante los años 1979-1983. En 1982, sin embargo, los datos se recogieron trimestralmente desde abril, no mensualmente. El total de cada uno de esos trimestres aparece en negrita en la tabla. Usar métodos de análisis de series en el tiempo para estimar los valores mensuales que faltan. Discutir posibles fuentes de error.

18.69. Omitir algunos datos de la Tabla 18.12 del Problema 18.9 y ver si se logran recuperar mediante técnicas de análisis de series en el tiempo.

18.70. Rehacer el problema anterior con los datos de los Problemas 18.42 y 18.47.

Tabla 18.41

	1979	1980	1981	1982	1983
En.	97.4	103.8	121.3	128.3	133.9
Feb.	86.6	99.1	110.1	116.8	120.7
Mar.	89.3	101.7	116.7	123.4	126.1
Abr.	92.4	112.3	116.9		126.5
May.	99.2	116.6	115.5	**332.9**	121.1
Jun.	83.0	93.9	95.9		104.6
Jul.	72.5	83.7	82.7		94.7
Ag.	64.3	75.3	82.3	**262.2**	83.9
Sep.	60.5	77.0	85.2		84.2
Oct.	78.0	91.4	99.5		98.3
Nov.	75.8	84.7	93.4	**295.1**	98.8
Dic.	84.0	103.6	109.5		108.5

Fuente: Survey of Current Business.

PROBLEMAS DIVERSOS

18.71. Analizar cada una de las series en el tiempo, (*a*) hasta (*e*), en las Tablas 18.42 y 18.43, que dan datos de EE.UU. para los años 1960-1986 y 1979-1986, respectivamente. Si se desea úsense sólo los datos hasta 1985 y hágase la predicción de los de 1986, que podrán así ser comparados con los datos reales. Nótese que la Tabla 18.42 muestra los *promedios mensuales* para cada año, mientras la Tabla 18.43 contiene los *valores mensuales* para cada año.

18.72. En la Tabla 18.44 se presentan las ventas mensuales totales (en millones de dólares) de los fabricantes de maquinaria eléctrica en EE.UU. durante los años 1979-1986.

(*a*) Analizar los esquemas estacional y cíclico de la serie en el tiempo.

(*b*) Identificar y discutir las dificultades que implica el análisis a causa de la inflación de los precios.

Tabla 18.42

Años	(*a*) Viviendas construidas (miles)	(*b*) Producción de hulla (millones de toneladas cortas)	(*c*) Automóviles nuevos vendidos (miles)	(*d*) Producción de tablazón para la construcción (millones de pies de tabla)	(*e*) Producción de aluminio (miles de toneladas cortas)
1960	106.6	32.72	556		167.9
1961	113.8	33.58	462	2654	158.6
1962	123.5	35.25	578	2740	176.5
1963	136.7	38.24	637	2879	192.7
1964	132.1	40.17	646	2951	212.7
1965	128.6	42.67	776	3075	229.6
1966	104.3	44.33	717	3011	247.3
1967	110.2	46.05	620	2940	272.4
1968	129.0	44.99	735	3089	271.3
1969	125.0	46.71	685	3162	316.1
1970	122.1	49.17	546	3050	331.3
1971	173.7	46.02	715	3051	327.1
1972	198.2	49.62	735	3239	343.5
1973	171.5	49.17	805	3191	377.5
1974	112.7	49.26	611	2872	408.6
1975	97.7	53.19	560	2654	323.3
1976	129.0	55.42	708	3045	354.3
1977	165.8	56.00	767	3125	378.1
1978	152.6	53.94	764	3128	400.2
1979	145.8	64.17	702	3084	418.5
1980	109.3	69.17	533	2613	427.5
1981	91.7	68.67	521	2435	412.3
1982	89.3	69.15	421	2224	300.8
1983	142.7	64.85	562	2610	281.6
1984	146.3	74.15	635	2830	341.6
1985	145.4	73.52	667	3050	291.6
1986	150.6	73.70	626	3363	253.0

Fuente: Survey of Current Business.

Tabla 18.43

Años	(*a*) Viviendas construidas (miles)	(*b*) Producción de hulla (millones de toneladas cortas)	(*c*) Automóviles nuevos vendidos (miles)	(*d*) Producción de tablazón para la construcción (millones de pies de tabla)	(*e*) Producción de aluminio (miles de toneladas cortas)
1979					
En.	88.4	56.49	737	2877	418
Feb.	84.7	53.63	709	2877	379
Mar.	153.3	65.49	883	3306	419
Abr.	161.3	62.79	761	3119	402
May.	189.1	67.93	922	3219	423
Jun.	192.0	69.40	820	3143	410
Jul.	165.0	54.50	587	3018	429
Ag.	171.4	72.10	449	3355	430
Sep.	163.8	63.90	630	3131	419
Oct.	169.0	75.91	787	3412	435
Nov.	119.2	67.56	641	2914	423
Dic.	91.8	60.32	494	2631	435
1980					
En.	73.4	67.81	513	2798	431
Feb.	80.6	64.33	619	2835	406
Mar.	86.1	69.87	649	2879	434
Abr.	96.6	69.87	572	2257	421
May.	93.0	70.40	518	2307	438
Jun.	117.8	71.36	544	2486	425
Jul.	121.5	60.70	432	2479	427
Ag.	131.7	70.24	299	2783	426
Sep.	147.0	72.06	529	2818	419
Oct.	153.7	75.75	675	2903	437
Nov.	113.5	65.51	560	2480	427
Dic.	96.3	72.12	490	2329	439
1981					
En.	85.2	66.16	439	2523	445
Feb.	72.5	69.79	475	2542	404
Mar.	108.9	77.27	620	2818	448
Abr.	124.0	38.02	645	2780	431
May.	110.6	37.28	670	2651	441
Jun.	107.0	61.90	712	2588	420
Jul.	101.0	73.35	513	2483	426
Ag.	87.3	78.20	345	2554	416
Sep.	90.9	81.30	522	2307	393
Oct.	88.1	84.78	520	2379	396
Nov.	64.9	76.03	425	1831	364
Dic.	59.7	79.97	370	1765	364

Tabla 18.43. (*Continuación*)

Años	(*a*) Viviendas construidas (miles)	(*b*) Producción de hulla (millones de toneladas cortas)	(*c*) Automóviles nuevos vendidos (miles)	(*d*) Producción de tablazón para la construcción (millones de pies de tabla)	(*e*) Producción de aluminio (miles de toneladas cortas)
1982					
En.	47.6	65.72	273	1810	351
Feb.	52.0	69.62	320	1891	311
Mar.	78.7	82.93	469	2148	336
Abr.	85.1	73.16	488	2281	319
May.	99.2	70.66	510	2251	321
Jun.	91.9	71.23	561	2338	300
Jul.	107.2	59.87	439	2376	297
Ag.	97.2	72.09	356	2560	287
Sep.	108.4	67.60	429	2445	271
Oct.	111.5	70.48	431	2333	275
Nov.	109.9	63.68	407	2247	266
Dic.	83.4	62.73	366	2004	275
1983					
En.	92.9	61.85	457	2484	279
Feb.	96.7	60.26	474	2481	223
Mar.	135.8	68.13	575	2682	248
Abr.	136.4	61.27	529	2623	245
May.	175.5	62.94	587	2645	265
Jun.	173.8	62.23	644	2718	261
Jul.	162.0	55.03	461	2585	284
Ag.	177.7	73.11	492	2714	297
Sep.	156.8	70.44	627	2748	299
Oct.	159.9	71.34	678	2787	320
Nov.	136.4	68.27	636	2504	318
Dic.	108.5	63.35	581	2345	340
1984					
En.	109.2	67.87	647	2740	342
Feb.	130.4	73.68	682	2678	324
Mar.	138.1	81.59	772	3104	350
Abr.	173.0	71.71	665	2983	348
May.	182.2	79.83	699	2828	365
Jun.	184.3	75.29	676	2968	351
Jul.	163.1	73.92	517	2685	349
Ag.	147.8	90.37	519	2933	344
Sep.	149.6	78.54	538	2776	329
Oct.	152.7	69.42	686	3154	338
Nov.	126.5	64.04	668	2814	325
Dic.	99.0	63.48	553	2295	334

Tabla 18.43. (*Continuación*)

Años	(*a*) Viviendas construidas (miles)	(*b*) Producción de hulla (millones de toneladas cortas)	(*c*) Automóviles nuevos vendidos (miles)	(*d*) Producción de tablazón para la construcción (millones de pies de tabla)	(*e*) Producción de aluminio (miles de toneladas cortas)
1985					
En.	105.4	67.98	733	2727	329
Feb.	95.8	67.04	659	2718	289
Mar.	145.2	77.66	736	3085	312
Abr.	176.0	76.54	744	3296	295
May.	170.5	78.24	760	3256	304
Jun.	163.4	73.02	677	3101	288
Jul.	161.0	69.01	565	3034	292
Ag.	161.1	79.48	554	3299	289
Sep.	148.6	73.82	638	3196	280
Oct.	173.2	80.12	739	3387	285
Nov.	124.1	69.29	658	2851	265
Dic.	120.5	70.01	540	2649	271
1986					
En.	115.9	78.29	713	3092	273
Feb.	107.2	72.69	675	3046	251
Mar.	151.1	77.57	655	3347	281
Abr.	188.3	74.89	713	3362	275
May.	186.7	73.14	685	3405	284
Jun.	183.6	72.67	706	3355	241
Jul.	172.2	67.82	505	2967	231
Ag.	163.8	76.55	426	3441	235
Sep.	154.3	75.02	637	3397	231
Oct.	154.9	76.83	684	3820	243
Nov.	115.7	68.67	556	3496	239
Dic.	113.1	70.26	561	3623	252

Fuente: Survey of Current Business.

Tabla 18.44

	En.	Feb.	Mar.	Abr.	May.	Jun.	Jul.	Ag.	Sep.	Oct.	Nov.	Dic.
1979	8.128	9.107	9.562	8.873	8.990	9.851	8.178	9.029	9.877	9.790	9.614	9.720
1980	9.204	10.617	10.778	9.909	9.838	10.714	9.150	10.263	11.169	11.459	11.201	10.596
1981	9.986	11.293	11.812	11.301	11.338	12.452	10.463	11.465	12.397	11.988	11.725	11.125
1982	10.410	11.689	12.094	11.831	11.949	12.588	10.843	11.327	12.301	11.908	11.496	11.421
1983	11.042	12.214	13.028	12.462	12.526	13.890	11.481	12.416	14.398	14.066	14.059	14.330
1984	13.129	14.435	15.791	14.646	14.980	16.549	13.700	15.009	16.718	15.605	15.372	16.572
1985	13.557	15.288	16.352	14.612	14.796	16.844	13.586	15.064	16.565	16.104	16.509	16.237
1986	13.614	15.887	17.024	15.549	15.504	17.537	14.643	16.375	18.362	17.240	17.614	17.845

Fuente: Survey of Current Business.

CAPITULO 19

Números índice

NUMERO INDICE

Un *número índice* es una medida estadística diseñada para poner de relieve cambios en una variable o en un grupo de variables relacionadas con respecto al tiempo, situación geográfica, ingresos, o cualquier otra característica. Una colección de números índice para diferentes años, lugares, etc., se llama a veces una *serie de índices*.

APLICACIONES DE LOS NUMEROS INDICE

Los números índice se usan para hacer comparaciones. Por ejemplo, con números índice podemos comparar los costes de alimentación o de otros servicios en una ciudad durante un año con los del año anterior, o la producción de acero en un año en una zona del país con la de otra zona. Aunque se usan principalmente en economía e industria, los números índice son aplicables en muchos otros campos. En educación, por ejemplo, se pueden usar los números índice para comparar la inteligencia relativa de estudiantes en sitios diferentes o en años diferentes.

Muchos gobiernos y agencias privadas se ocupan de elaborar números índice (o índices, como se les llama a veces) con el propósito de predecir condiciones económicas o industriales, tales como índices de paro, de producción, salariales y tantos otros. Tal vez el más conocido sea el *índice de coste de la vida* o *índice de precios al consumo*, que prepara el Instituto de Estadística. En muchos contratos aparecen ciertas *cláusulas de revisión* que producen aumentos salariales automáticos correspondientes a los aumentos del índice de precios al consumo.

En este capítulo estaremos interesados sobre todo en números índice que muestran cambios respecto del tiempo, si bien los métodos descritos en este capítulo son aplicables ciertamente en otros casos.

RELACIONES DE PRECIOS

Uno de los ejemplos más simples de un número índice es una *relación de precios*, que no es sino el cociente entre el precio de un artículo en un período dado y su precio en otro período, conocido como *período base* o *período de referencia*. Supondremos, por sencillez, que los precios en cada período son constantes. Si no lo son, podemos tomar un promedio adecuado para el período de modo que la suposición sea esencialmente válida.

Si p_n y p_o denotan los precios de un artículo durante el período dado y el período base, respectivamente, entonces, por definición,

$$\text{Relación de precios} = \frac{p_n}{p_o} \tag{1}$$

La relación de precios se expresa habitualmente como un porcentaje multiplicándola por 100.

Más en general, si p_a y p_b son los precios de un artículo durante los períodos a y b, respectivamente, *la relación de precios en el período b con respecto al período a se define como* p_b/p_a y se denota por $p_{a|b}$, notación que resultará de utilidad. Con esta notación, la relación de precios en la ecuación (1) se denota por $p_{o|n}$.

EJEMPLO 1. Supongamos que los precios al consumo de un cuarto de galón de leche en los años 1970 y 1980 eran de 45¢ y 54¢, respectivamente. Tomando 1970 como *año base* y 1980 como el *año dado*, tenemos

$$\text{Relación de precios} = p_{1970|1980} = \frac{\text{precio en 1980}}{\text{precio en 1970}} = \frac{54¢}{45¢} = 1.2 = 120\%$$

o brevemente 120, omitiendo el signo % (como se hace con frecuencia en la literatura estadística). Este resultado simplemente significa que en 1980 el precio de la leche era el 120% del de 1970; es decir, aumentó un 20%.

EJEMPLO 2. Con 1980 como año base y 1970 como año dado en el Ejemplo 1, se tiene

$$\text{Relación de precios} = p_{1980|1970} = \frac{\text{precio en 1970}}{\text{precio en 1980}} = \frac{45¢}{54¢} = \frac{5}{6} = 83\tfrac{1}{3}\%$$

o sea $83\frac{1}{3}$. Esto quiere decir que en 1970 el precio de la leche era el $83\frac{1}{3}\%$ del de 1980; esto es, era $16\frac{2}{3}\%$ menor que en 1980.

Nótese que la relación de precios para un período dado con respecto al *mismo* período es siempre 100%, o sea 100. En particular, la relación de precios correspondiente al período base es siempre 100. Esto da cuenta de la notación (frecuente en la literatura estadística) de escribir, por ejemplo, «1970 = 100» para indicar que se ha tomado 1970 como período base.

PROPIEDADES DE LAS RELACIONES DE PRECIOS

Si p_a, p_b, p_c, ... denotan los precios en los períodos a, b, c, ..., respectivamente, tenemos las siguientes propiedades para las relaciones de precios asociadas. Las demostraciones son consecuencia inmediata de las definiciones.

1. **Propiedad identidad:** $p_{a|a} = 1$ Esto dice simplemente que la relación de precios para un período respecto de él mismo es 1, o sea 100%.
2. **Propiedad de inversión temporal:** $p_{a|b}p_{b|a} = 1$, o sea $p_{a|b} = 1/p_{b|a}$. Esto afirma que si dos períodos se intercambian, las correspondientes relaciones de precios son cada una la inversa de la otra (véase Ejemplos 1 y 2).
3. **Propiedad cíclica o circular:** $p_{a|b}p_{b|c}p_{c|a} = 1$, $p_{a|b}p_{b|c}p_{c|d}p_{d|a} = 1$, etc.

4. **Propiedad cíclica (o circular) modificada:** $p_{a|b}p_{b|c} = p_{a|c}$, $p_{a|b}p_{b|c}p_{c|d} = p_{a|d}$, etc. Esta propiedad se sigue directamente de las Propiedades 2 y 3.

RELACIONES DE CANTIDAD O DE VOLUMEN

En vez de comparar los precios de un artículo, podemos estar interesados en comparar las cantidades (o volúmenes) de producción, consumo o exportación. En tales casos hablamos de *relaciones de cantidad* o *relaciones de volumen*. Por sencillez, como en el caso de los precios, suponemos que las cantidades son constantes en cada período. Si no lo son, se pueden tomar promedios adecuados de forma que esencialmente la hipótesis sea válida.

Si q denota la cantidad (o volumen) de un artículo que se ha producido, consumido, exportado, etcétera durante un período base, y q la correspondiente cantidad producida, consumida, exportada, etcétera durante un período dado, definimos

$$\text{Relación de cantidad o de volumen} = \frac{q_n}{q_0} \tag{2}$$

que se suele exprexar como porcentaje.

Al igual que para las relaciones de precios, usamos la notación $q_{a|b} = q_b/q_a$ para denotar la relación de cantidad en el período b respecto al período a. Las mismas observaciones y propiedades comentadas para las relaciones de precios son válidad para las relaciones de cantidad.

RELACIONES DE VALOR

Si p es el precio de un artículo durante un período y q es la cantidad (o volumen) producida, vendida, etc., durante ese período, entonces pq se llama el *valor total*. Así, si 1000 cuartos (de galón de leche se venden a 56¢ el cuarto, el valor total es pq = (\$0.56)(1000) = \$560.

Si p_o y q_o son el precio y la cantidad de un artículo durante un período base, y p_n y q_n el precio y la cantidad correspondientes a un período dado, los valores totales durante esos períodos vienen dados por v_o y v_n, respectivamente, y definimos

$$\text{Relación de valor} = \frac{v_n}{v_o} = \frac{p_n q_n}{p_o q_o} = \left(\frac{p_n}{p_o}\right)\left(\frac{q_n}{q_o}\right) = \text{relación de precios} \times \text{relación de cantidad} \tag{3}$$

Las mismas observaciones, notación y propiedades aplicables a las relaciones de precios y a las relaciones de cantidad lo son a las relaciones de valor. En particular, si $p_{a|b}$, $q_{a|b}$ y $v_{a|b}$ denotan las relaciones de precios, cantidad y valor del período b respecto al período a, entonces, como en la ecuación (3),

$$v_{a|b} = p_{a|b}q_{a|b}$$

que se llama la *propiedad de inversión de factores*.

RELACIONES DE ENLACE Y EN CADENA

Si p_1, p_2, p_3, ... representan los precios durante intervalos sucesivos de tiempo 1, 2, 3, ..., entonces $p_{1|2}$, $p_{2|3}$, $p_{3|4}$, ... representan las relaciones de precios de cada intervalo respecto al intervalo de tiempo precedente, y se llaman *relaciones de enlace*.

EJEMPLO 3. Si los precios de un artículo durante 1983, 1984, 1985 y 1986 fueron 8¢, 12¢ 15¢ y 18¢, respectivamente, entonces las relaciones de enlace son $p_{1983|1984} = \frac{12}{8} = 150(\%)$, $p_{1984|1985} = \frac{15}{22} = 125(\%)$ y $p_{1985|1986} = \frac{18}{15} = 120(\%)$.

La relación de precios para un período dado con respecto a otro tomado como base, se puede siempre expresar en términos de relaciones de enlace. Esto es una consecuencia de la propiedad *cíclica*, o *circular*, de las relaciones. Así, $p_{5|2} = p_{5|4}p_{4|3}p_{3|2}$.

EJEMPLO 4. Por ejemplo 3, la relación de precios para 1986 con respecto al año base 1983 es

$$p_{1983|1986} = p_{1983|1984}p_{1984|1985}p_{1985|1986} = \frac{12}{8}\cdot\frac{15}{12}\cdot\frac{18}{15} = \frac{18}{8} = 225(\%)$$

Las relaciones de precios con respecto a un período base fijo, que como hemos visto se pueden hallar mediante relaciones de enlace, se llaman en ocasiones *relaciones en cadena* con respecto a esa base.

EJEMPLO 5. En los Ejemplos 3 y 4, la colección de relaciones en cadena para los años 1984, 1985 y 1986 con respecto a la base 1983 viene dada por

$$p_{1983|1984} = \frac{12}{8} = 150(\%)$$

$$p_{1983|1985} = p_{1983|1984}p_{1984|1985} = \frac{12}{8}\cdot\frac{15}{12} = 187.5(\%)$$

$$p_{1983|1986} = p_{1983|1984}p_{1984|1985}p_{1985|1986} = \frac{12}{8}\cdot\frac{15}{12}\cdot\frac{18}{15} = 225(\%)$$

Las ideas anteriores son también aplicables a las relaciones de cantidad y a las relaciones de valor.

PROBLEMAS IMPLICITOS EN EL CALCULO DE NUMEROS INDICE

A la hora de las aplicaciones prácticas estamos menos interesados en comparar precios, cantidades o valores de artículos aislados que en comparar los precios (etc.) de grandes grupos de artículos. Por ejemplo, al calcular un índice de precios al consumo no sólo queremos comparar los precios de la leche en dos períodos, sino también el precio de los huevos, de la carne, del calzado, de la vivienda, etc., de modo que se consiga una visión general. Naturalmente, podríamos simplemente hacer una lista con *todos* esos precios, pero eso no sería muy satisfactorio. Lo deseable es disponer de *un solo* número índice de precios que compare los precios en ambos períodos *en promedio*.

No es difícil ver que los cálculos de números índice que afecten a un grupos de artículos conllevan muchos problemas que hay que solventar. Al calcular un índice de precios al consumo, por ejemplo, debemos decidir qué artículos o servicios deben incluirse, así como su peso de importancia relativa; hemos de recolectar datos referentes a precios y cantidades de tales artículos; hemos de decidir qué hacer con las distintas *calidades* dentro de un mismo artículo, o con ciertos artículos o servicios que están disponibles un año pero no en el año base; por fin, hemos de decidir cómo reunir toda esa información y sacar un solo número índice del coste de la vida que tenga significado práctico.

EL USO DE PROMEDIOS

Ya que hemos de llegar a un solo número índice resumiendo una gran cantidad de información, es fácil comprender que los promedios (discutidos en el Capítulo 3) juegan un papel importante en el cálculo de números índice.

Así como existen muchos métodos para calcular promedios, también hay muchos para calcular los números índice, cada uno con sus ventajas y desventajas propias.

En lo que sigue examinaremos unos pocos métodos comúnmente empleados en la práctica, usando varios procedimientos para promediar. Aunque nos restringimos a índices de precios al principio, veremos cómo modificar adecuadamente las cosas para el caso de índice de valor o de cantidad.

CRITERIOS TEORICOS PARA NUMEROS INDICE

Desde un punto de vista teórico es deseable que los números índice para grupos de artículos tengan las propiedades que cumplían las relaciones (números índice para un solo artículo). Todo número índice que tenga tal o cual propiedad se dice que satisface el criterio asociado con ella. Por ejemplo, los números índice que tengan la propiedad de inversión temporal se dirá que satisfacen el *criterio de inversión temporal*, etc.

No se conoce ningún número índice que cumpla todos los criterios, si bien en muchos casos se satisfacen aproximadamente. El índice ideal de Fisher (pág. 484), que en particular verifica el criterio *de inversión temporal* y el de *inversión de factores*, es mejor que cualquier otro número índice útil en cuanto a satisfacer las propiedades consideradas importantes (de ahí el apelativo de «ideal»).

Desde una perspectiva práctica, no obstante, otros números índice sirven también, y examinaremos algunos de ellos.

NOTACION

Es habitual denotar por $p_n^{(1)}, p_n^{(2)}, p_n^{(3)}, \ldots$ los precios de un primer, segundo, tercer, ... artículo durante un período dado n, mientras los precios respectivos en el período base se denotan por $p_o^{(1)}, p_o^{(2)}, p_o^{(3)}$, etcétera. Los números 1, 2, 3, ... son *superíndices* y no deben ser confundidos con exponentes. Con esa notación, el precio del artículo j durante el período n es $p_n^{(j)}$.

Como en capítulos anteriores, podemos usar la notación de sumatorio al sumar sobre el índice j. Por ejemplo, supuesto que haya un total de N artículos, la suma de sus precios durante el período n se puede expresar como $\sum_{j=1}^{N} p_n^{(j)}$ o $\sum p_n^{(j)}$. Sin embargo, es más sencillo omitir el superíndice y

escribir $\sum p_n$, cosa que haremos cuando no haya riesgo de confusión; recuérdese que el simbolismo completo está sobreentendido. Con esta notación, $\sum p_o$ denotará la suma de los precios de todos los artículos durante el período base.

Análoga notación se usa para cantidades y valores.

EL METODO DE AGREGACION SIMPLE

En este método de calcular un índice de precios, expresamos el precio total de los artículos en el año dado como porcentaje del precio total de los artículos en el año base. En símbolos,

$$\text{Indice de precios por agregación simple} = \frac{\sum p_n}{\sum p_o} \tag{4}$$

donde $\sum p_o$ = suma de todos los precios de los artículos en el año base

$\sum p_n$ = suma de todos los precios de los artículos en el año dado

y donde el resultado se expresa como porcentaje, al igual que se hace con los números índice en general.

Aunque este método es fácil de aplicar, tiene dos grandes desventajas que lo convierten en insatisfactorio:

1. No tiene en cuenta la importancia relativa de los diversos artículos. Así pues, asigna igual peso a la leche que a la crema de afeitar a la hora de calcular el índice de precios al consumo.
2. Las unidades escogidas al anotar los precios (galones, bushels, libras, ...) afectan al índice. (Véase Prob. 19.12.)

EL METODO DEL PROMEDIO SIMPLE DE RELACIONES

El índice producido por este método depende del procedimiento utilizado para promediar las relaciones de precios; los procedimientos incluyen la media aritmética, la geométrica, la armónica y la mediana. Con la media aritmética, por ejemplo, tendríamos

$$\text{Indice de la media aritmética simple de relaciones de precios} = \frac{\sum p_n/p_0}{N} \tag{5}$$

donde $\sum p_n/p_o$ = suma de todas las relaciones de precios de los artículos.

N = número de relaciones de precios de artículos utilizados.

Para índices basados en otros tipos de promedios, véanse Problemas 19.14 y 19.15.

Si bien este método no tiene ya la segunda desventaja antes citada, todavía mantiene la primera.

EL METODO DE AGREGACION PONDERADA

Con el fin de evitar las desventajas del método de agregación simple, asignamos un *peso* al precio de cada artículo, en general la cantidad (o volumen) vendida durante el año base, durante el año dado o durante algún año típico (que puede ser un promedio de varios años). Tales pesos indican la importancia del artículo en cuestión. Dependiendo de que se use el año base, el año dado o un año típico (denotados respectivamente por q_0, q_n y q_t, usamos una de las siguientes fórmulas:

1. **Indice de Laspeyres o método del año base:**

$$\text{Indice de precios por agregación ponderada con pesos de cantidad en el año base} = \frac{\sum p_n q_0}{\sum p_0 q_0} \tag{6}$$

2. **Indice de Paasche o método del año dado:**

$$\text{Indice de precios por agregación ponderada con pesos de cantidad en el año dado} = \frac{\sum p_n q_n}{\sum p_0 q_n} \tag{7}$$

3. **El método del año típico:** Si q denota la cantidad durante algún período típico t, definimos

$$\text{Indice de precios por agregación ponderada con pesos de cantidad en el año típico} = \frac{\sum p_n q_t}{\sum p_0 q_t} \tag{8}$$

Para $t = o$ y $t = n$, esto se reduce a las ecuaciones (6) y (7), respectivamente.

INDICE IDEAL DE FISHER

Definimos

$$\text{Indice ideal de Fisher} = \sqrt{\left(\frac{\sum p_n q_0}{\sum p_0 q_0}\right)\left(\frac{\sum p_n q_n}{\sum p_0 q_n}\right)} \tag{9}$$

Este índice de precios es la media geométrica de los números índice de Laspeyres y de Paasche dados por las ecuaciones (6) y (7). Como ya hemos comentado, el índice ideal de Fisher satisface los *criterios de inversión temporal* y de *inversión de factores*, lo que le confiere una cierta ventaja teórica sobre otros números índice.

EL INDICE DE MARSHALL-EDGEWORTH

El índice de Marshall-Edgeworth usa el método de agregación ponderada con año típico, en el que los pesos se toman como la media aritmética de las cantidades del año base y del año dado; es decir, $q_t = \frac{1}{2}(q_0 + q_n)$. Sustituyendo este valor de q en la ecuación (8) resulta

$$\text{Indice de Marshall-Edgeworth} = \frac{\sum p_n(q_0 + q_n)}{\sum p_o(q_o + q_n)} \tag{10}$$

EL METODO DEL PROMEDIO PONDERADO DE RELACIONES

Para paliar las desventajas del método del promedio simple de relaciones se puede usar un *promedio ponderado de relaciones*. El promedio ponderado más utilizado es la *media aritmética ponderada*, aunque tambien se utilizan otros, como la media geométrica ponderada (véase Cap. 3).

En este método asignamos a cada relación de precios un peso dado por el valor total del artículo en términos de alguna unidad monetaria, digamos el dolar. Como el valor de un artículo se obtiene multiplicando su precio p por la cantidad q, los pesos vienen dados por pq.

Según se use el año base, el año dado o el año típico para calcular tales pesos (denotados respectivamente por p_oq_o, p_nq_n, y p_tq_t), usamos una u otra de las fórmulas siguientes:

Media aritmética ponderada de relaciones de precios, usando pesos del año base:

$$\frac{\sum (p_n/p_o)(p_oq_o)}{\sum p_oq_o} = \frac{\sum p_nq_o}{\sum p_oq_o} \tag{11}$$

Media aritmética ponderada de relaciones de precios, usando pesos de un año típico:

$$\frac{\sum (p_n/p_o)(p_nq_n)}{\sum p_nq_n} \tag{12}$$

Media aritmética ponderada de relaciones de precios, usando pesos de un año típico:

$$\frac{\sum (p_n/p_o)(p_tq_t)}{\sum p_tq_t} \tag{13}$$

Nótese que la fórmula (11) da el mismo resultado que la (6) de Laspeyres.

NUMEROS INDICE DE CANTIDAD O VOLUMEN

Las fórmulas descritas previamente para la obtención de números índice de precios se modifican fácilmente para hallar números índice de cantidad (o volumen) intercambiando simplemente p y q. Por ejemplo, sustituyendo p por q en la ecuación (5) resulta

$$\text{Indice de media aritmética simple de relaciones de volumen} = \frac{\sum q_n/q_o}{N} \tag{14}$$

donde $\sum q_n/q_o$ = suma de relaciones de cantidad de todos los artículos
N = número de relaciones de cantidad usadas

Análogamente, las fórmulas (6) y (7) se convierten en

$$\text{Indice de agregación ponderada de volumen con pesos del año base} = \frac{\sum q_np_o}{\sum q_op_o} \tag{15}$$

$$\text{Indice de agregación ponderada de volumen con pesos del año dado} = \frac{\sum q_n p_n}{\sum q_o p_n} \tag{16}$$

La fórmula (15) se llama a veces un *índice de volumen de Laspeyres*, y la (10) un *índice de volumen de Paasche*. En estas fórmulas se toman los precios como pesos. No obstante, cabe utilizar cualquier otro peso apropiado.

De forma parecida se modifican las fórmulas (8) a (13).

NUMEROS INDICE DE VALOR

Exactamente igual que hemos hecho con los números índice de precios o de cantidad, se pueden definir *índices de valor*. El más sencillo de ellos es

$$\text{Indice de valor} = \frac{\sum p_n q_n}{\sum p_o q_o} \tag{17}$$

donde $\sum p_o q_o$ = valor total de todos los artículos en el período base

$\sum p_n q_n$ = valor total de todos los artículos en el período dado

Este es un *índice de agregación simple*, ya que los valores no han recibido pesos relativos. Se pueden enunciar fórmulas que les asignen pesos para tener en cuenta la importancia relativa de los artículos.

CAMBIO DEL PERIODO BASE EN LOS NUMEROS INDICE

En la práctica es deseable que el período base elegido para la comparación sea un período de estabilidad económica no muy alejado en el pasado. De cuando en cuando puede ser necesario, por tanto, cambiar el período base.

Una posibilidad es recalcular todos los números índice en términos del nuevo período base. Un método aproximado más simple consiste en dividir todos los números índice para los diversos años correspondientes al período base antiguo por los números índice correspondientes al nuevo período base, expresando los resultados como porcentajes. Estos resultados representan los nuevos números índice, siendo el número índice para el nuevo período base 100(%), como debe ser.

Matemáticamente hablando, este método es estrictamente aplicable sólo si los números índice satisfacen el *criterio circular* (véase Prob. 19.37). Sin embargo, para muchos tipos de índices el método, afortunadamente, da resultados que en la práctica son suficientemente próximos a los que se obtendrían teóricamente.

DEFLACION DE SERIES EN EL TIEMPO

Aunque los ingresos de las personas pueden estar creciendo teóricamente durante un cierto número de años, sus *ingresos reales* pueden en verdad estar disminuyendo debido al aumento del coste de la vida, en tanto en cuanto este aumento del coste de la vida hace que disminuya su *poder adquisitivo*.

Calculamos los ingresos reales dividiendo los *ingresos aparentes* de cada año por el número índice del coste de la vida en ese año, usando un período base adecuado. Por ejemplo, si los ingresos de un individuo en 1980 son el 150% de sus ingresos en 1970 (o sea, han crecido un 50%) y el coste de la vida se ha doblado en ese mismo período de tiempo, entonces sus ingresos reales en 1980 son sólo del $\frac{150}{2} = 75\%$ de lo que eran en 1970.

El párrafo anterior describe el proceso de deflación de una serie en el tiempo referida a ingresos de una persona. Un procedimiento análogo se sigue para la deflación de otras series en el tiempo. Así, en el Capítulo 18 usamos un procedimiento similar para *desestacionalizar* datos mediante números índice estacionales.

En términos matemáticos, este método de deflación de series en el tiempo es estrictamente aplicable sólo si los números índice cumplen el *criterio de inversión de factores*, y por esta razón el índice ideal de Fisher es adecuado. No obstante, otros números índice dan también resultados correctos a efectos prácticos.

PROBLEMAS RESUELTOS

RELACIONES DE PRECIOS

19.1. Los precios al por menor (en centavos por libra) del cinc en EE. UU. durante 1978-1984 se ven en la Tabla 19.1.

(*a*) Con 1978 como base, hallar las relaciones de precios correspondientes a los años 1983 y 1984.
(*b*) Con 1980 como base, hallar las relaciones de precios correspondientes a los años dados.
(*c*) Usando 1978-1980 como período base, hallar las relaciones de precios correspondientes a los años dados.

Tabla 19.1

Año	1978	1979	1980	1981	1982	1983	1984
Precio promedio del cinc al por menor	31.0	37.3	37.4	44.6	38.5	41.4	48.6

Fuente: U.S. Bureau of Mines.

Solución

(*a*) La relación de precios para 1982 con 1978 como base es

$$p_{1978|1982} = \frac{\text{precio en 1982}}{\text{precio en 1978}} = \frac{38.5}{31.0} = 1.242 = 124.2\%$$

La relación de precios para 1984 con 1978 como base es

$$p_{1978|1984} = \frac{\text{precio en 1984}}{\text{precio en 1978}} = \frac{48.6}{31.0} = 1.568 = 156.8\%$$

En la literatura estadística es usual omitir los símbolos % al citar los números índice, quedando sobreentendidos. Usando ese convenio, citamos las relaciones de precios anteriores como 124.2 y 156.8 respectivamente.

(*b*) Dividimos cada precio al por menor en la Tabla 19.1 por 37.4 (centavos por libra), el precio del año 1980; así pues, las relaciones de precios pedidas, expresadas en porcentajes, se indican en la Tabla 19.2. Representan los *números índice* de los precios del cinc al por menor para los años 1978-1984, y la colección completa se llama una *serie de índices*. Obsérvese que la relación de precios (o número índice de precios) del año 1980 es en porcentaje 100.0, como ocurre siempre para el período base. Esto se suele escribir simbólicamente en estadística como 1980 = 100.

Tabla 19.2

Año	1978	1979	1980	1981	1982	1983	1984
Relación de precios (1980 = 100)	82.9	99.7	100.0	119.3	102.9	110.7	129.9

(*c*) La media aritmética de los precios para los años 1978-1980 es $\frac{1}{3}(31.0 + 37.3 + 37.4) = 35.2$. Dividamos cada precio al por menor de la Tabla 19.1 por ese precio promedio del período base de 35.2 (centavos por libra). Las requeridas relaciones de precios, en forma de porcentajes, se recogen en la Tabla 19.3. Representan los números índice de precios del cinc para los años 1978-1984 con 1978-1980 como período base. Nótese que la media aritmética de los números índice correspondientes al período base 1978-1980 es $\frac{1}{3}(88.1 + 106.0 + 106.3) = 100.1$, o sea 100.0 (la ligera discrepancia se debe a errores de redondeo), como ocurre siempre para el período base. Esto se escribe a veces simbólicamente como 1978-1980 = 100.

Tabla 19.3

Año	1978	1979	1980	1981	1982	1983	1984
Relación de precios (1978-1980 = 100)	88.1	106.0	106.3	126.7	109.4	117.6	138.1

19.2. Probar (*a*) que $p_{a|b}p_{b|c} = p_{a|c}$ y (*b*) que $p_{a|b}p_{b|a} = 1$.

Solución

Por la definición, basta ver que

(*a*)
$$p_{a|b}p_{b|c} = \frac{p_b}{p_a} \cdot \frac{p_c}{p_b} = \frac{p_c}{p_a} = p_{a|c}$$

(*b*)
$$p_{a|b}p_{b|a} = \frac{p_b}{p_a} \cdot \frac{p_a}{p_b} = 1$$

19.3. Con la Tabla 19.3, que usa 1978-1980 como período base, hallar las relaciones de precios con 1980 como base.

Solución

Dividimos cada relación de precios de la Tabla 19.3 por la relación de precios 106.3. Los números resultantes, expresados como porcentajes, son las relaciones de precios requeridas, y se muestran en la Tabla 19.2 (aparte errores de redondeo).

Esta solución demuestra que, dada una serie de índices correspondiente a un período base, podemos hallar la serie de índices correspondiente a otro período base sin hacer uso de los datos originales sobre precios. El método implicado se llama *cambio de período base*, o *desplazamiento de la base*. Para una demostración de este método, ver el Problema 19.36.

19.4. En 1986 el precio medio de un artículo era un 20% más que en 1985, 20% menos que en 1984 y 50% más que en 1987. Reducir los datos a relaciones de precios usando (*a*) 1985, (*b*) 1986 y (*c*) 1984-1985 como base.

Solución

(*a*) La relación de precios (o número índice) con 1985 como base es 100 (simbólicamente, 1985 = = 100, o sea 100%.)

Como el precio en 1986 es 20% más que en 1985, la relación de precios correspondiente a 1986 es 100 + 20 = 120; esto es, el precio en 1986 es 120% del de 1985.

Como el precio en 1986 es 20% menor que en 1984, debe ser el 100-20 = 80% del precio de 1984. Así pues, el precio de 1984 es $1/0.80 = \frac{5}{4} = 125\%$ del de 1986; es decir,

Relación de precios 1984 = 125% de la relación de precios 1986 = 125% de 120 = 150

Ya que el precio en 1986 es 50% más que en 1987, debe ser 100 + 50 = 150% del de 1987. Luego el precio de 1987 es $1/1.50 = \frac{2}{3}$ del de 1986; esto es,

Relación de precios 1987 = $\frac{2}{3}$ de la relación de precios 1986
= $\frac{2}{3}$ de 120 = 80

Luego las relaciones de precios pedidas con 1985 como base son las que recoge la Tabla 19.4.

(*b*) Usamos el método de cambio del período base descrito en el Problema 19.3. Dividimos cada relación de precios de la Tabla 19.4 por 120 (la relación de precios del nuevo año base 1986) y expresamos el resultado como porcentaje. Así pues, las relaciones de precios deseadas con 1986 como base, las muestra la Tabla 19.5.

(*c*) *Primer método* [usando la parte (*a*)]

De la Tabla 19.4 vemos que la media aritmética de las relaciones de precios para 1984 y 1985 es $\frac{1}{2}(150 + 100) = 125$. Dividiendo cada relación de precios en la Tabla 19.4 por 125, obtenemos las relaciones de precios requeridas, que se muestran en la Tabla 19.6.

Segundo método [usando la parte (*b*)]

Según la Tabla 19.5, la media aritmética de las relaciones de precios para 1984 y 1985 es $\frac{1}{2}(125 + 83.3) = 104.2$. Dividiendo cada relación de precios en la Tabla 19.5 por 104.2, obtenemos los mismos resultados que con el primer método.

Tabla 19.4

Año	1984	1985	1986	1987
Relación de precios (1985 = 100)	150	100	120	80

Tabla 19.5

Año	1984	1985	1986	1987
Relación de precios (1986 = 100)	125	83.3	100	66.7

Tabla 19.6

Año	1984	1985	1986	1987
Relación de precios (1984-1985 = 100)	120	80	96	64

RELACIONES DE CANTIDAD O VOLUMEN

19.5. La Tabla 19.7 presenta la producción de trigo (en millones de bushels) en EE. UU. durante 1977-1985. Reducir los datos de la tabla a relaciones de cantidad usando (*a*) 1982 y (*b*) 1977-1980 como base.

Solución

(*a*) Dividir las cifras de producción de cada año por 2765 (la producción del año base 1982). Así las requeridas relaciones de cantidad (o números índice de cantidad), expresadas en porcentajes, se muestran en la Tabla 19.8.

Tabla 19.7

Año	1977	1978	1979	1980	1981	1982	1983	1984	1985
Producción de trigo	2046	1776	2134	2380	2785	2765	2420	2595	2425

Fuente: U.S. Department of Agriculture.

Tabla 19.8

Año	1977	1978	1979	1980	1981	1982	1983	1984	1985
Relación de cantidad (1982 = 100)	74.0	64.2	77.2	86.1	100.7	100.0	87.5	93.9	87.7

(*b*) La media aritmética de producción en los años 1977-1980 es $\frac{1}{4}(2046 + 1776 + 2134 + 2380) = 2084$. Dividiendo la producción de cada año por 2084 obtenemos las relaciones de cantidad deseadas, expresadas en porcentajes, que figuran en la Tabla 19.9. Nótese que $\frac{1}{4}(98.2 + 85.2 + 102.4 + 114.2) = 100.0$, lo cual sirve de comprobación.

Tabla 19.9

Año	1977	1978	1979	1980	1981	1982	1983	1984	1985
Relación de cantidad (1977-1980 = 100)	98.2	85.2	102.4	114.2	133.6	132.7	116.1	124.5	116.4

19.6. Mientras la relación de cantidad para 1986 es 105 cuando se toma 1977 como base, es 140 cuando el año base es 1980. Hallar la relación de cantidad para 1980 tomando 1977 como base.

Solución

Primer método

Por las propiedades de las relaciones de cantidad, tenemos $q_{a|b}q_{b|c} = q_{a|c}$. Poniendo $a = 1977$, $b = 1980$ y $c = 1986$, tenemos $q_{1977|1980} = q_{1977|1986}q_{1986|1980} = (1.05)(1/1.40) = 0.75 = 75\%$. Luego la relación de cantidad pedida es 75.

Segúndo método

Sean q_{1977}, q_{1980} y q_{1986} las cantidades reales en 1977, 1980 y 1986, respectivamente; así pues,

$$\text{Relación de cantidad para 1986 con base 1977} = \frac{q_{1986}}{q_{1977}} = 105\% = 1.05$$

$$\text{Relación de cantidad para 1986 con base 1980} = \frac{q_{1986}}{q_{1980}} = 140\% = 1.40$$

Luego la relación de cantidad para 1980 con 1977 como base es

$$\frac{q_{1980}}{q_{1977}} = \frac{q_{1980}/q_{1986}}{q_{1977}/q_{1986}} = \frac{1/1.40}{1/1.05} = \frac{1.05}{1.40} = 75\%$$

Tercer método

Como $q_{1986} = 1.05q_{1977} = 1.40q_{1980}$, tenemos $q_{1980}/q_{1977} = 1.05/1.40 = 75\%$. Por tanto, la relación de cantidad requerida es 75.

RELACIONES DE VALOR

19.7 En enero de 1980 una empresa pagó un total de \$80,000 a 120 empleados en nómina. En julio de ese mismo año, la empresa tenía 30 trabajadores más en nómina y pagó \$12,000 más que en enero.

(*a*) Con enero de 1980 como base, hallar el número índice de empleo (la relación de cantidad) para julio.

(*b*) Con enero de 1980 como base, hallar el número índice (relación de valor) trabajo-gasto para julio.

(*c*) Usando el resultado *relación de precios* × *relación de cantidad* = *relación de valor*, ¿qué interpretación se puede dar a la relación de precios en este caso?

Solución

(*a*) El número índice de empleo es

$$\text{Relación de cantidad} = \frac{120 + 30}{120} = 1.25 = 125\% \quad \text{o sea} \quad 125$$

(*b*) El número índice trabajo-gasto es

$$\text{Relación de valor} = \frac{\$80{,}000 + \$12{,}000}{\$80{,}000} = 1.15 = 115\% \quad \text{o sea} \quad 115$$

(*c*) La relación de precios es

$$\frac{\text{Relación de valor}}{\text{Relación de cantidad}} = \frac{115}{125} = 0.92 = 92\% \qquad \text{o sea} \qquad 92$$

Podemos interpretar esto como un *número índice de costo por empleado*. El significado es que en julio de 1980 el costo por empleado era el 92% del de enero de 1980, el período base. A veces se llama a esto un número índice de costo laboral *per capita*.

19.8. Una compañía espera que sus ventas de un producto crezcan un 50% el año próximo. ¿En qué porcentaje debe aumentar su precio de venta para doblar los ingresos brutos provenientes de ese producto?

Solución

Dado que Relación de precios × relación de cantidad = relación de valor

tenemos Relación de precios × 150% = 200%

Luego Relación de precios = $\frac{200}{150} = \frac{4}{3} = 133\frac{1}{3}\%$

La compañía debe aumentar por tanto el precio de ese producto en un $133\frac{1}{3} - 100 = 33\frac{1}{3}\%$.

RELACIONES DE ENLACE Y EN CADENA

19.9. Sabiendo que las relaciones de enlace para los precios en los años 1981, 1982, ..., 1985 son 125, 120, 135, 150 y 175, respectivamente, (*a*) hallar la relación de precios para 1982 con 1980 como base y (*b*) las relaciones de enlace y en cadena con 1981 como base.

Solución

Tenemos $p_{1980|1981} = 1.25$, $p_{1981|1982} = 1.20$, $p_{1982|1983} = 1.35$, $p_{1983|1984} = 1.50$ y $p_{1984|1985} = 1.75$. Por tanto:

(*a*) $p_{1980|1982} = p_{1980|1981}p_{1981|1982} = (1.25)(1.20) = 1.50 = 15\%$

(*b*) $p_{1981|1980} = \dfrac{1}{p_{1980|1981}} = \dfrac{1}{1.25} = 80\%$

$p_{1981|1981} = 100\%$

$p_{1981|1982} = 120\%$

$p_{1981|1983} = p_{1981|1982}p_{1982|1983} = (1.20)(1.35) = 1.62 = 162\%$

$p_{1981|1984} = p_{1981|1982}p_{1982|1983}p_{1983|1984} = (1.20)(1.35)(1.50) = 2.43 = 243\%$

$p_{1981|1985} = p_{1981|1982}p_{1982|1983}p_{1983|1984}p_{1984|1985} = (1.20)(1.35)(1.50)(1.75) = 425\%$

NUMEROS INDICE; EL METODO DE AGREGACION SIMPLE

19.10. La Tabla 19.10 muestra los precios al por mayor y las producciones en EE. UU. de leche, mantequilla y queso para los años 1980, 1981 y 1985. Calcular un índice de precios al por mayor por agregación de esos productos lácteos para el año 1985, tomando como base (*a*) 1980 y (*b*) 1980-1981

Tabla 19.10

	Precio (centavos por libra)			Cantidad (millones de libras)		
	1980	1981	1985	1980	1981	1985
Leche	13.23	13.95	12.90	128,500	132,800	143,700
Mantequilla	139.3	148.0	141.1	1,145	1,228	1,248
Queso	156.2	167.2	162.0	2,381	2,664	2,854

Fuente: Survey of Current Business.

Solución

(*a*) El índice de precios por agregación simple es

$$\frac{\sum p_n}{\sum p_o} = \frac{\text{suma de precios en el año prefijado (1985)}}{\text{suma de precios en el año base (1980)}} = \frac{12.90 + 141.1 + 162.0}{13.23 + 139.3 + 156.2} = 102.4(\%)$$

Es decir, el precio promedio al por mayor de esos tres productos en 1985 es el 102.4% del de 1980 (o sea, 2.4% mayor).

(*b*) El precio promedio (medio) de la leche en el período base 1980-1981 es

$$\tfrac{1}{2}(13.23 + 13.95) = 13.59¢/\text{lb}$$

el precio promedio (medio) de la leche en el período base 1980-1981 es

$$\tfrac{1}{2}(139.3 + 148.0) = 143.7¢/\text{lb}$$

el precio promedio (medio) de la leche en el período base 1980-1981 es

$$\tfrac{1}{2}(156.2 + 167.2) = 161.7¢/\text{lb}$$

y por tanto el índice de precios por agregación simple es

$$\frac{\sum p_n}{\sum p_o} = \frac{\text{suma de precios en el año prefijado (1985)}}{\text{suma de precios en el año base (1980-1981)}} = \frac{12.90 + 141.1 + 162.0}{13.59 + 143.7 + 161.7} = 99.1(\%)$$

Nótese que este método no hace uso de las *cantidades* producidas, sino sólo de los precios de los artículos.

A efectos ilustrativos, sólo se han considerado aquí tres artículos, pero en la práctica se incluyen mucho más.

19.11. Explicar por qué los números índice obtenidos en el Problema 19.10 pueden ser medidas inapropiadas de los cambios de precios en los artículos en cuestión.

Solución

El índice calculado en el Problema 19.10 no tiene en cuenta la importancia relativa de los productos, tal como quedaría determinada por ejemplo por cuánto los usan los consumidores o cuánto se produce para el consumo. Estas consideraciones se incorporan en problemas posteriores.

19.12. La Tabla 19.11 muestra los precios y la producción, en promedio, de algodón y trigo en EE. UU. durante los años 1980 y 1986. Explicar por qué un índice de precios por agregación simple para 1986 con 1980 como base es inapropiado como medida del cambio de precios en esos dos productos.

Tabla 19.11

	Precio		Cantidad*	
	1980	1986	1980	1986
Algodón	74.4¢ (por libra)	56.8¢ (por libra)	11.122 (millones de balas)	13.432 (millones de balas)
Trigo	$3.91 (por bushel)	$3.16 (por bushel)	511.4 (millones de bushels)	487.1 (millones de bushels)

* 1 bala = 480 lb; 1 bushel = 60 lb.
Fuente: Survey of Current Business.

Solución

Si se usa un índice de precios de agregación simple, el resultado es

$$\frac{\sum p_n}{\sum p_o} = \frac{\text{suma de precios en el año prefijado (1986)}}{\text{suma de precios en el año base (1980)}} = \frac{56.8¢ + 316¢}{74.4¢ + 391¢} = 0.801 = 80.1(\%)$$

indicando que el precio medio de esos productos era en 1986 del orden de un 80% respecto al de 1980.

Si expresamos el precio del trigo en centavos por libra, es $3.91/60 = 6.52¢ lb para 1980 y $3.16/60 = = 5.27¢ para 1986. En este caso el índice de precios por agregación simple es

$$\frac{\sum p_n}{\sum p_o} = \frac{56.8¢ + 5.27¢}{74.4¢ + 6.52¢} = 76.7(\%)$$

Esto ilustra el hecho de que el índice de precios por agregación simple puede ser muy sensible a las unidades utilizadas al anotar los precios; en consecuencia, está claro que da una medida inapropiada en tales casos. Este hecho, junto con la desventaja comentada en el Problema 19.11, dan buenas razones para abandonar su uso en la práctica.

La nota al final del Problema 19.10 se aplica también a este problema.

EL METODO DEL PROMEDIO SIMPLE DE RELACIONES

19.13. Usar el método del promedio simple de relaciones para calcular un índice de precios al por mayor para los productos de la Tabla 19.10 para el año 1985, usando 1980 como base.

Solución

Las relaciones de precios para la leche, la mantequilla y el queso en 1985 con 1980 como base son como siguen:

$$\text{Relación de precios para la leche} = \frac{\text{precio de la leche en 1985}}{\text{precio de la leche en 1980}} = \frac{12.90}{13.23} = 97.5(\%)$$

$$\text{Relación de precios para la mantequilla} = \frac{\text{precio de la mantequilla en 1985}}{\text{precio de la mantequilla en 1980}} = \frac{141.1}{139.3} = 101.3(\%)$$

$$\text{Relación de precios para el queso} = \frac{\text{precio del queso en 1985}}{\text{precio del queso en 1980}} = \frac{162.0}{156.2} = 103.7(\%)$$

$$\text{Promedio (media) de relaciones de precios} = \frac{\sum p_n/p_o}{N} = \frac{97.5 + 101.3 + 103.7}{3} = 100.8(\%)$$

19.14. Rehacer el Problema 19.13 usando la mediana en lugar de la media.

Solución

(*a*) Número índice solicitado = mediana de relaciones de precios 97.5, 101.3 y 103.7 = 101.3.
(*b*) Número índice solicitado = mediana de relaciones de precios 94.9, 98.2 y 100.2 = 98.2.

19.15. Resolver el Problema 19.13 con la media geométrica en lugar de la media

Solución

(*a*) Número índice pedido = media geométrica de las relaciones de precios 97,5, 101.3, y 103.7 = $= \sqrt[3]{(97.5)(101.3)(103.7)} = 100.8$.
(*b*) Número índice pedido = media geométrica de las relaciones de precios 94.9, 98.2 y 100.2 = $= \sqrt[3]{(94.9)(98.2)(100.2)} = 97.7$.

19.16. Usar el promedio simple (media) de las relaciones de precios para obtener un número índice de precios para los artículos de la Tabla 19.11, con 1980 como año base y 1986 como año dado.

Solución

$$\text{Relación de precios para el algodón} = \frac{\text{precio del algodón en 1986}}{\text{precio del algodón en 1980}} = \frac{56.8¢}{74.4¢} = 76.3(\%)$$

$$\text{Relación de precios para el trigo} = \frac{\text{precio del trigo en 1986}}{\text{precio del trigo en 1980}} = \frac{\$3.16}{\$3.91} = 80.8(\%)$$

$$\text{Promedio simple (media) de relaciones de precios} = \frac{\sum p_n/p_o}{N} = \frac{76.3 + 80.8}{2} = 78.6(\%)$$

Nótese que el resultado es *independiente* de las unidades usadas al anotar los precios (comparar con el Problema 19.12).

19.17. Resolver el Problema 19.16 usando la media geométrica.

Solución

Número índice requerido = media geométrica de relaciones de precios 76.3 y 78.6 = $= \sqrt{(76.3)(78.6)} = 77.4(\%)$.

EL METODO DE AGREGACION PONDERADA; INDICES DE LASPEYRES Y PAASCHE

19.18. Calcular, con los datos de la Tabla 19.10, un número índice de precios de Laspeyres para 1985 con (*a*) 1980 y (*b*) 1980-1981 como base.

Solución

(*a*) El índice de Laspeyres, el índice de precios por agregación ponderada con las cantidades del período base como pesos, es

$$\frac{\sum p_n q_o}{\sum p_o q_o} = \frac{\sum (\text{precios en } 1985)(\text{cantidades en } 1980)}{\sum (\text{precios en } 1980)(\text{cantidades en } 1980)}$$

$$= \frac{(12.90)(128{,}500) + (141.1)(1145) + (162.0)(2381)}{(13.23)(128{,}500) + (139.3)(1145) + (156.2)(2381)} = 0.9881 = 98.8(\%)$$

(*b*) Las cantidades promedio de leche, mantequilla y queso producidas en 1980-1981 son $\frac{1}{2}(128{,}500 + 132{,}800) = 130{,}650$, $\frac{1}{2}(1145 + 1228) = 1186.5$ y $\frac{1}{2}(2381 + 2664) = 2522.5$, respectivamente; los precios promedio en 1980-1981 se indican en el Problema 19.10(*b*). Luego el índice de Laspeyres es

$$\frac{\sum p_n q_o}{\sum p_o q_o} = \frac{\sum (\text{precios en } 1985)(\text{cantidades promedio en } 1980\text{-}1981)}{\sum (\text{precios en } 1980\text{-}1981)(\text{cantidades promedio en } 1980\text{-}1981)}$$

$$= \frac{(12.90)(130{,}650) + (141.1)(1186.5) + (162.0)(2522.5)}{(13.59)(130{,}650) + (143.7)(1186.5) + (161.7)(2522.5)} = 0.9607 = 96.1(\%)$$

19.19. Usando los datos de la Tabla 19.10, calcular un número índice de precios de Paasche para 1985 con (*a*) 1980 y (*b*) 1980-1981 como base.

Solución

(*a*) El índice de Paasche, el índice de precios por agregación ponderada con las cantidades del año dado como pesos, es

$$\frac{\sum p_n q_n}{\sum p_o q_n} = \frac{\sum (\text{precios en } 1985)(\text{cantidades en } 1985)}{\sum (\text{precios en } 1980)(\text{cantidades en } 1985)}$$

$$= \frac{(12.90)(130{,}650) + (141.1)(1186.5) + (162.0)(2522.5)}{(13.59)(130{,}650) + (143.7)(1186.5) + (161.7)(2522.5)} = 0.9607 = 96.1(\%)$$

(*b*) El índice de Paasche es

$$\frac{\sum p_n q_n}{\sum p_o q_n} = \frac{\sum (\text{precios en } 1985)(\text{cantidades en } 1985)}{\sum (\text{precios en } 1980\text{-}1981)(\text{cantidades en } 1985)}$$

$$= \frac{(12.90)(143{,}700) + (141.1)(1248) + (162.0)(2854)}{(13.59)(143{,}700) + (143.65)(1248) + (161.7)(2854)} = 0.9609 = 96.1(\%)$$

19.20. (*a*) Hallar los números índice de Laspeyres para los datos de la Tabla 19.11.
(*b*) Hallar los números índice de Paasche para los datos de la Tabla 19.11.
(*c*) En la hipótesis de que deban revisarse los números índice cada año, apuntar una ventaja del índice de Laspeyres sobre el de Paasche.

Solución

(*a*) El índice de Laspeyres es

$$\frac{\sum p_n q_o}{\sum p_o q_o} = \frac{\sum (\text{precios en } 1986)(\text{cantidades en } 1980)}{\sum (\text{precios en } 1980)(\text{cantidades en } 1980)}$$

$$= \frac{(56.8¢/\text{lb})(11.122 \times 500 \text{ millones de lb}) + (316¢/\text{bu})(511.4 \text{ millones de bu})}{(74.4¢/\text{lb})(11.122 \times 500 \text{ millones de lb}) + (391¢/\text{bu})(511.4 \text{ millones de bu})} = 87.2(\%)$$

(*b*) El índice de Paasche es

$$\frac{\sum p_n q_n}{\sum p_o q_n} = \frac{\sum (\text{precios en } 1986)(\text{cantidades en } 1986)}{\sum (\text{precios en } 1980)(\text{cantidades en } 1986)}$$

$$= \frac{(56.8¢/\text{lb})(13.432 \times 500 \text{ millones de lb}) + (316¢/\text{bu})(487.1 \text{ millones de bu})}{(74.4¢/\text{lb})(13.432 \times 500 \text{ millones de lb}) + (391¢/\text{bu})(487.1 \text{ millones de bu})} = 77.6(\%)$$

Nota: En la práctica, donde ha de calcularse un número índice para muchos artículos, es aconsejable tabular de forma adecuada el cálculo (véase Prob. 19.31 por ejemplo).

(*c*) Al calcular el índice de Laspeyres, los pesos (o sea, las cantidades producidas o consumidas en el año base, si se calcula un índice de precios) no cambian de año en año, así que la única información que uno precisa es una lista de los últimos precios. Al calcular un índice de Paasche, uno necesita esa información tanto sobre los precios como sobre los pesos (o cantidades); por tanto, calcular un índice de Paasche es más laborioso en cuanto a recolección de datos.

19.21. Interpretar los números índice de (*a*) Laspeyres y (*b*) Paasche, en términos del valor total (o coste total) de los artículos.

Solución

(*a*) Al calcular un índice de precios de Laspeyres, $\sum p_o q_o$ representa el valor total (o coste total) de un conjunto de artículos (a veces llamado la *cesta de la compra*) en el año o período base. La cantidad $\sum p_n q_o$ representa lo mismo en el año o período dado. Así pues, un índice de Laspeyres sirve para medir los costes totales en cualquier año dado de una *cesta de la compra fija* adquirida en el año base.

(*b*) Al calcular un índice de precios de Paasche, $\sum p_o q_n$ es el valor total (o coste total) de artículos adquiridos en el año dado, suponiendo precios del año base, mientras $\sum p_n q_n$ es el valor total de artículos adquiridos en el año dado a los precios del año dado. Luego un índice de Paasche sirve para medir el coste total de una cesta de la compra del año dado respecto a cuál sería su coste si su adquisición se hubiera efectuado en el año base.

19.22. Se dice a veces que el índice de precios de Laspeyres tiende a *sobreestimar* los cambios de precios, mientras el de Paasche tiende a *subestimarlos*. Dar posibles argumentos que apoyen tal afirmación.

Solución

De acuerdo con la *ley económica de la oferta y la demanda*, la gente tiende a comprar menos cuando los precios son altos y más cuando son bajos. Esta es la llamada *demanda elástica*, válida si se trata de un artículo que no es de primera necesidad.

En el caso del índice de Laspeyres, $\sum p_n q_o$ será algo mayor de lo que debiera ser, pues por la ley de la oferta y la demanda la gente tenderá a adquirir menos artículos de alto precio y más de bajo precio,

de modo que el coste total sería menor que el que predice $\sum p_n q_o$. Así pues, el índice de Laspeyres $(\sum p_n q_o)/\sum p_o q_o)$ tiende a ser mayor de lo que debiera.

En el índice de Paasche, los papeles del año base y del año dado se intercambian respecto del que jugaban en el de Laspeyres. Ello hace que el índice de Paasche tienda a ser menor de lo que debiera.

Los razonamientos anteriores no implican que el índice de Laspeyres sea *siempre* mayor que el de Paasche, sino sólo que *tiende a* ser mayor. En la práctica, el índice de Laspeyres puede ser mayor, igual o menor que el de Paasche. (Véanse Probs. 19.18 y 19.19, en los que el índice de Laspeyres es de hecho menor que el de Paasche.)

19.23. Probar que los números índice de precios de agregación ponderada con pesos (cantidades) fijos satisfacen el criterio circular.

Solución

Denotando por q_0 los pesos fijos, tenemos para cualesquiera períodos a, b y c, los números índice

$$I_{a|b} = \frac{\sum p_b q_o}{\sum p_a q_o} \quad \text{y} \quad I_{b|c} = \frac{\sum p_c q_o}{\sum p_b q_o}$$

Entonces

$$I_{a|b} I_{b|c} = \frac{\sum p_b q_o}{\sum p_a q_o} \cdot \frac{\sum p_c q_o}{\sum p_b q_o} = \frac{\sum p_c q_o}{\sum p_a q_o} = I_{a|c}$$

que demuestra que el criterio circular se verifica.

Los números índice de Laspeyres y de Paasche no cumplen el criterio circular.

INDICE IDEAL DE FISHER

19.24. Probar que el índice ideal de Fisher es la media geométrica de los números índice de Laspeyres y de Paasche.

Solución

Si F, L y P denotan respectivamente los índices de Fisher, Laspeyres y Paasche, tenemos

$$F = \sqrt{\left(\frac{\sum p_n q_o}{\sum p_o q_o}\right)\left(\frac{\sum p_n q_n}{\sum p_o q_n}\right)} = \sqrt{LP}$$

según la definición de L y P. Como $\sqrt{LP}$ es la media geométrica de L y P, eso concluye la demostración.

19.25. Probar que el índice ideal de Fisher está entre los números índice de Laspeyres y de Paasche.

Solución

Esto se sigue directamente de que $F = \sqrt{LP}$ está entre L y P, pues L y P son positivos. Nótese que si $L = P$, entonces $F = L = P$.

Como por el Problema 19.22 L tiende a *sobreestimar* los cambios de precios y P tiende a *subestimarlos*, se deduce que F, que está entre ambos, debe producir una estimación más correcta que L o P.

19.26. Hallar el índice ideal de Fisher para los productos de la Tabla 19.10 para el año 1985, con (*a*) 1980 y (*b*) 1980-1981 como base.

Solución

(*a*) Por los Problemas 19.18(*a*) y 19.19(*a*),

$$F = \sqrt{LP} = \sqrt{(0.9881)(0.9886)} = 0.9883 = 98.8(\%)$$

(*b*) Por los problemas 19.18(*b*) y 19.19(*b*),

$$F = \sqrt{LP} = \sqrt{(0.9607)(0.9609)} = 0.9608 = 96.1(\%)$$

19.27. Hallar el índice ideal de Fisher para los datos de la Tabla 19.11.

Solución

Del Problema 19.20, $F = \sqrt{LP} = \sqrt{(0.9881)(0.776)} = 0.823 = 82.3(\%)$. Nótese que una buena aproximación a $\sqrt{LP}$, cuando L y P son casi iguales, viene dada por $(L+P)/2$. Esta media aritmética de L y P puede usarse como definición de un nuevo número índice que está entre L y P.

19.28. Demostrar que el índice ideal de Fisher satisface el criterio de inversión temporal.

Solución

Denotemos por $F_{o|n}$ el índice ideal de Fisher para un año dado con respecto a un año base, y sea $F_{n|o}$ el índice ideal de Fisher cuando el año dado y el año base se intercambian. Entonces el criterio de inversión temporal se satisface si

$$F_{o|n} = \sqrt{\left(\frac{\sum p_n q_o}{\sum p_o q_o}\right)\left(\frac{\sum p_n q_n}{\sum p_o q_n}\right)}$$

Entonces

$$F_{n|o} = \sqrt{\left(\frac{\sum p_o q_n}{\sum p_n q_n}\right)\left(\frac{\sum p_o q_o}{\sum p_n q_o}\right)}$$

y

$$F_{o|n}F_{n|o} = \sqrt{\left(\frac{\sum p_n q_o}{\sum p_o q_o}\right)\left(\frac{\sum p_n q_n}{\sum p_o q_n}\right)\left(\frac{\sum p_o q_n}{\sum p_n q_n}\right)\left(\frac{\sum p_o q_o}{\sum p_n q_o}\right)} = 1$$

EL INDICE DE MARSHALL-EDGEWORTH

19.29. Calcular el índice de precios de Marshall-Edgeworth para los datos del Problema 19.10.

Solución

El índice de Marshall-Edgeworth es

$$\frac{\sum p_n(q_o + q_n)}{\sum p_o(q_o + q_n)} = \frac{\sum (\text{precios en 1985})(\text{suma de cantidades en 1980 y 1985})}{\sum (\text{precios en 1980})(\text{suma de cantidades en 1980 y 1985})}$$

$$= \frac{(12.90)(128{,}500 + 143{,}700) + (141.1)(1145 + 1228) + (162.0)(2381 + 2854)}{(13.23)(128{,}500 + 143{,}700) + (139.3)(1145 + 1228) + (156.2)(2381 + 2854)}$$

$$= 0.9884 = 98.8(\%)$$

Obsérvese que está entre los números índice de Laspeyres y Paasche (véase Prob. 19.20). Para una demostración de que ese es siempre el caso, véase el Problema 19.30.

19.30. (*a*) Demostrar que si $X_1/X_2 < Y_1/Y_2$, entonces $X_1/X_2 < (X_1 + Y_1)/(X_2 + Y_2) < Y_1/Y_2$, donde X_1, X_2 e Y_1, Y_2 son números positivos arbitrarios.

(*b*) Usar el resultado de la parte (*a*) para probar que el índice de Marshall-Edgeworth está entre los de Laspeyres y Paasche.

Solución

(*a*) Si $X_1/X_2 < Y_1/Y_2$, entonces
$$X_1Y_2 < X_2Y_1 \tag{18}$$

Sumando X_1X_2 a ambos lados de la ecuación (18), se tiene

$$X_1X_2 + X_1Y_2 < X_1X_2 + X_2Y_1 \quad \text{o sea} \quad X_1(X_2 + Y_2) < X_2(X_1 + Y_1)$$

es decir
$$\frac{X_1}{X_2} < \frac{X_1 + Y_1}{X_2 + Y_2} \tag{19}$$

dividiendo ambos lados por $X_2(X_2 + Y_2)$. Sumando Y_1Y_2 a ambos lados de la ecuación (18), tenemos

$$X_1Y_2 + Y_1Y_2 < X_2Y_1 + Y_1Y_2 \quad \text{o sea} \quad Y_2(X_1 + Y_1) < Y_1(X_2 + Y_2)$$

es decir
$$\frac{X_1 + Y_1}{X_2 + Y_2} < \frac{Y_1}{Y_2} \tag{20}$$

dividiendo ambos miembros por $Y_1(X_1 + Y_1)$.

El resultado anunciado se sigue de las ecuaciones (19) y (20).

(*b*) **Caso 1** (el índice de Laspeyres es menor que el de Paasche)

Sean $X_1 = \sum p_nq_o$, $X_2 = \sum p_oq_o$, $Y_1 = \sum p_nq_n$ e $Y_2 = \sum p_oq_n$. Luego $X_1/X_2 < Y_1/Y_2$, y por tanto de la parte (*a*) se sigue

$$\frac{\sum p_nq_o}{\sum p_oq_o} < \frac{\sum p_nq_o + \sum p_nq_n}{\sum p_oq_o + \sum p_oq_n} < \frac{\sum p_nq_n}{\sum p_oq_n}$$

o sea
$$\frac{\sum p_nq_o}{\sum p_oq_o} < \frac{\sum p_n(q_o + q_n)}{\sum p_o(q_o + q_n)} < \frac{\sum p_nq_n}{\sum p_oq_n}$$

es decir índice de Laspeyres < índice de Marshall-Edgeworth < índice de Paasche

Caso 2 (el índice de Paasche es menor que el de Laspeyres)

Sean $X_1 = \sum p_nq_n$, $X_2 = \sum p_oq_n$, $Y_1 = \sum p_nq_o$ e $Y_2 = \sum p_oq_o$. Luego $X_1/X_2 < Y_1/Y_2$, luego por la parte (*a*) tenemos

$$\frac{\sum p_nq_n}{\sum p_oq_n} < \frac{\sum p_nq_n + \sum p_nq_o}{\sum p_oq_n + \sum p_oq_o} < \frac{\sum p_nq_o}{\sum p_oq_o}$$

es decir
$$\frac{\sum p_nq_n}{\sum p_oq_n} < \frac{\sum p_n(q_o + q_n)}{\sum p_o(q_o + q_n)} < \frac{\sum p_nq_n}{\sum p_oq_n}$$

o sea índice de Paasche < índice de Marshall-Edgeworth < índice de Laspeyres

Se sigue de los Casos 1 y 2 que, independientemente de que el índice de Laspeyres sea mayor que el de Paasche, el índice de Marshall-Edgeworth está entre ellos dos.

EL METODO DEL PROMEDIO PONDERADO DE RELACIONES

19.31. Calcular una media aritmética ponderada de las relaciones de precios para los datos de la Tabla 19.11, usando como pesos (*a*) los valores del año dado y (*b*) los valores del año base, siendo 1985 el año dado y 1980 el año base.

Solución

(*a*) Usando como pesos los valores para el año dado, la media aritmética ponderada de las relaciones de precios es

$$\frac{\sum (p_n/p_o)(p_n q_n)}{\sum p_n q_n} = \frac{\sum (\text{relaciones de precios})(\text{valores del año prefijado})}{\sum (\text{valores del año prefijado})}$$

El cálculo lo resume la Tabla 19.12, donde el subíndice n se refiere al año dado 1985, y el subíndice o al año base 1980; y donde p y q denotan precios y cantidades, respectivamente.

Tabla 19.12

	p_o	p_n	q_n	$p_n q_n$	$p_n q_n$ (millones de dólares)	$(p_n/p_o)(p_n q_n)$ (millones de dólares)
Algodón	74.4¢ (por libra)	56.8¢ (por libra)	13.432×480 (millones de libras)	0.7634	3662.1	2795.6
Trigo	\$3.91 (por bushel)	\$3.16 (por bushel)	487.1 (millones de bushels)	0.8082	1539.2	1244.0
					$\sum p_n q_n = 5201.3$	$\sum (p_n/p_o)(p_n q_n) = 4039.6$

Así pues, el número índice pedido es

$$\frac{\sum (p_n/p_o)(p_n q_n)}{\sum p_n q_n} = \frac{4039.6}{5201.3} = 77.7(\%)$$

(*b*) Con los valores del año base como pesos, la media aritmética ponderada de las relaciones de precios es

$$\frac{\sum (p_n/p_o)(p_o q_o)}{\sum p_o q_o} = \frac{\sum p_n q_o}{\sum p_o q_o} = \text{índice de Laspeyres del Problema 19.20(}a\text{)} = 87.2\ (\%)$$

Se puede presentar el cálculo en una tabla como en (*a*).

NUMEROS INDICE DE CANTIDAD O VOLUMEN

19.32. Usando los datos de la Tabla 19.11, calcular un índice de volumen para 1986 con año base 1980 mediante (*a*) una media aritmética simple de relaciones de volumen, (*b*) un índice de volumen de agregación ponderada con los precios del año base como pesos y (*c*) ídem con los precios del año dado como pesos.

Solución

(*a*) El índice de una media aritmética simple de relaciones de volumen es

$$\frac{\sum (q_n/q_o)}{N} = \frac{13.432/11.122 + 487.1/511.4}{2} = \frac{120.8(\%) + 95.2(\%)}{2} = 108.0(\%)$$

(*b*) El índice de volumen de agregación ponderada con los precios del año base como pesos es

$$\frac{\sum q_n p_o}{\sum q_o p_o} = \frac{\sum (\text{cantidades en 1986})(\text{precios en 1980})}{\sum (\text{cantidades en 1980})(\text{precios en 1980})}$$

$$= \frac{(13.432 \times 480 \text{ millones de lb})(74.4¢/\text{lb}) + (487.1 \text{ millones de bu})(391¢/\text{bu})}{(11.122 \times 480 \text{ millones de lb})(74.4¢/\text{lb}) + (511.4 \text{ millones de bu})(391¢/\text{bu})} = 112.2(\%)$$

Esto se llama a veces *número índice de cantidad (o volumen) de Laspeyres.*

(*c*) El índice de volumen de agregación ponderada con los precios del año dado como pesos es

$$\frac{\sum q_n p_n}{\sum q_o p_n} = \frac{\sum (\text{cantidades en 1986})(\text{precios en 1986})}{\sum (\text{cantidades en 1980})(\text{precios en 1986})}$$

$$= \frac{(13.432 \times 480 \text{ millones de lb})(56.8¢/\text{lb}) + (487.1 \text{ millones de bu})(316¢/\text{bu})}{(11.122 \times 480 \text{ millones de lb})(56.8¢/\text{lb}) + (511.4 \text{ millones de bu})(316¢/\text{bu})} = 111.9(\%)$$

19.33. A partir de los resultados del Problema 19.32, determinar el índice ideal de Fisher de volumen (o cantidad).

Solución

Al igual que el de precios, el índice ideal de Fisher de volumen viene dado por la media geométrica de los números índice de volumen de Laspeyres y Paasche. Luego, por el Problema 19.32,

$$\text{índice ideal de Fisher de volumen} = \sqrt{(112.2)(111.9)} = 112.0(\%)$$

NUMEROS INDICE DE VALOR

19.34. Probar que el índice ideal de Fisher satisface el criterio de inversión de factores.

Solución

Dicho criterio se satisface si el índice es tal que

$$\text{Indice de valor} = (\text{índice de precios})(\text{índice de cantidad})$$

Sean F_P y F_Q índices ideales de Fisher de precios y de cantidad, respectivamente. Entonces

$$\text{Indice de valor} = F_P F_Q = \sqrt{\left(\frac{\sum p_n q_o}{\sum p_o q_o}\right)\left(\frac{\sum p_n q_n}{\sum p_o q_n}\right)}\sqrt{\left(\frac{\sum q_n p_o}{q_o p_o}\right)\left(\frac{\sum q_n p_n}{\sum q_o p_n}\right)} = \frac{\sum p_n q_n}{\sum p_o q_o}$$

y por tanto el índice ideal de Fisher verifica el criterio de inversión de factores.

19.35. Calcular el índice de valor en el Problema 19.34 para los datos de la Tabla 19.11.

Solución

Como el resultado

$$\text{Indice de valor} = (\text{índice de precios})(\text{índice de cantidad})$$

vale exactamente cuando se usa el índice ideal de Fisher, de los Problemas 19.27 y 19.33 obtenemos

$$\text{Indice de valor} = (82.3\%)(112.0\%) = 92.2\%$$

Este resultado se puede obtener también por sustitución directa en $\sum p_n q_n / \sum p_o q_o$.

CAMBIO DEL PERIODO BASE EN LOS NUMEROS INDICE

19.36. Establecer la validez del método del Problema 19.3 para hallar relaciones de precios para un nuevo período base.

Solución

Numeramos los períodos sucesivamente de 1 a N, como en la primera fila de la Tabla 19.13, y denotamos por $p_1, p_2, \ldots, p_N$ los precios en esos períodos, como en la segunda fila de la tabla. Las relaciones de precios para los períodos j y k, que llamaremos período viejo y nuevo respectivamente, se indican en las filas tercera y cuarta de esa tabla; aquí $p_{j|1} = p_1/p_j$, $p_{j|2} = p_2/p_j$, etc. Es claro que la cuarta fila se puede obtener de la tercera dividiendo cada entrada de la tercera por $p_{j|k}$ (o sea, la relación de precios del período k respecto al período j tomando como base); por ejemplo,

$$\frac{p_{j|1}}{p_{j|k}} = \frac{p_1/p_j}{p_k/p_j} = \frac{p_1}{p_k} = p_{k|1} \qquad \text{etc.}$$

Los resultados se aplican a relaciones de cantidad y de valor además de a las de precios.

Tabla 19.13

Período	1	2	3	...	j	...	k	...	N
Precio	p_1	p_2	p_3	...	p_j	...	p_k	...	$p_{j\|N}$
Relación de precios correspondiente al antiguo período j	$p_{j\|1}$	$p_{j\|2}$	$p_{j\|3}$	...	100%	...	$p_{j\|k}$	...	$p_{j\|N}$
Relación de precios correspondiente al nuevo período k	$p_{k\|1}$	$p_{k\|2}$	$p_{k\|3}$	...	$p_{k\|j}$	...	100%	...	$p_{k\|N}$

19.37. Demostrar que el método del Problema 19.36 para cambiar el período base de los números índice es aplicable si y sólo si los números índice satisfacen el criterio circular.

Solución

Si denotamos los números índice para los diversos períodos, con el período j como base, por

$$I_{j|1}, I_{j|2}, \ldots, I_{j|N} \tag{21}$$

y los correspondientes números índice, con el período k como base, por

$$I_{k|1}, I_{k|2}, \ldots, I_{k|N} \tag{22}$$

obtendremos la sucesión (22) dividiendo cada miembro de (21) por $I_{j|k}$ si y sólo si

$$\frac{I_{j|1}}{I_{j|k}} = I_{k|1}, \frac{I_{j|2}}{I_{j|k}} = I_{k|2}, \ldots$$

o sea $$I_{j|1} = I_{j|k}I_{k|1}, I_{j|2} = I_{j|k}I_{k|2}, \ldots$$

lo cual implica que los números índice satisfacen el criterio circular.

Como los índices de Laspeyres, Paasche, Fisher y Marshall-Edgeworth no lo satisfacen, el método en cuestión para cambiar el período base no se les aplica exactamente. Sin embargo, se aplica con buena aproximación en la práctica.

Los números índice por agregación ponderada con pesos fijos satisfacen el criterio circular (véase Problema 19.23). Para ellos sí se aplica exactamente el método expuesto de cambio de base.

19.38. La Tabla 19.14 muestra el índice de producción industrial para todos los productos manufacturados en EE. UU. en los años 1974-1985 con 1977 como período base. Hallar un nuevo índice con (*a*) 1979 y (*b*) 1974-1976 como base.

Tabla 19.14

Año	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
Indice de producción industrial (1977 = 100)	93	85	93	100	107	111	109	111	103	109	122	125

Fuente: Survey of Current Business.

Solución

(*a*) Dividimos cada índice de la Tabla 19.14 por 111 (el índice correspondiente a 1979) y expresamos el resultado como porcentaje. Los números índice requeridos, con 1979 como base, se muestran en la Tabla 19.15.

(*b*) El índice promedio para los años 1974-1985 con 1974-1976 como base es $\frac{1}{3}(93 + 85 + 93) = 90.33$. Dividiendo cada índice de la Tabla 19.14 por 90.33 pedidos, que recoge la Tabla 19.16 Nótese que el índice promedio para el nuevo período base, 1974-1976, es $\frac{1}{3}(103 + 94 + 103) = 100$, como tenía que ser.

Tabla 19.15

Año	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
Indice de producción industrial (1979 = 100)	84	77	84	90	96	100	98	100	93	98	100	113

Tabla 19.16

Año	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
Indice de producción industrial (1974-1976 = 100)	103	94	103	111	118	123	121	123	114	121	135	138

DEFLACION DE SERIES EN EL TIEMPO

19.39. La Tabla 19.17 muestra el salario semanal medio de los trabajadores en el comercio minorista de EE.UU. durante 1973-1983. También contiene el índice de precios al consumo para esos años, con 1972 como base. En términos del salario medio de 1973, determinar sus salarios reales en los años 1973-1983.

Tabla 19.17

Año	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
Salario semanal promedio de los trabajadores (dólares)	96.32	102.68	108.86	114.60	121.66	130.20	138.62	147.38	158.03	163.85	171.05
Indice de precios al consumo (1972 = 100)	106.2	117.9	128.7	136.1	144.9	155.9	173.5	197.0	217.4	230.7	238.1

Fuente: U.S. Department of Labor.

Solución

Hallamos primero un número índice de precios al consumo con 1973 como base, dividiendo todos los números de la fila de abajo en la Tabla 19.17 por 106.2 y expresando el resultado en porcentajes. Así se llega a la fila central de la Tabla 19.18. Ahora dividimos cada salario medio para los años dados (fila central de la Tabla 19.17) por el correspondiente número índice (fila central de la Tabla 19.18) para obtener los salarios reales (fila inferior de la Tabla 19.18).

Así, por ejemplo, el salario real correspondiente a 1983 es 171.05/224.2(%) = \$76.29. Se sigue que aunque los salarios aparentes casi se doblaron desde 1973 hasta 1983, los salarios reales han ido decreciendo con los años; de hecho, el salario real en 1983 venía a ser 20\$ menor que el de 1973. En otras palabras, el poder adquisitivo de los trabajadores decreció aproximadamente en un 20/96,32 = = 21%.

Tabla 19.18

Año	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
Indice de precios al consumo (1973 = 100)	100.0	111.0	121.2	128.2	136.4	146.8	163.4	185.5	204.7	217.2	224.2
Salario semanal real de los trabajadores (dólares)	96.32	92.50	89.82	89.39	89.19	88.69	84.83	79.45	77.20	75.44	76.29

19.40. Usar el índice de precios al consumo de la Tabla 19.18 para determinar el poder adquisitivo del dólar en los diversos años, respecto del valor adquisitivo de un dólar en 1973.

Solución

Dividiendo \$1.00 por cada índice de precios de la fila central de la Tabla 19.18, se deduce la Tabla 19.19, que muestra el poder adquisitivo de un dólar de 1973 en los años siguientes. En 1983, por ejemplo, la entrada 0.45 significa que un dólar de 1983 permitía comprar sólo un 45% de lo que permitía uno de 1973; esto es, el dólar valía \$0.45 en términos del dólar de 1973.

Los datos expresados en términos del valor del dólar en algún período específico de tiempo se dicen expresados en *dólares constantes* (con el período dado como base o referencia).

Tabla 19.19

Año	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
Poder adquisitivo del dolar en dólares de 1973	1.00	0.90	0.83	0.78	0.73	0.68	0.61	0.54	0.49	0.46	0.45

Por cada dólar cobrado en 1973, un trabajador debiera haber cobrado \$1.00/0.45 = \$2.22 en 1983 para compensar la inflación. Dicho de otro modo, en dólares constantes de 1973 los trabajadores cobraron \$96.32 en 1973, \$92.50 en 1974, \$89.82 en 1975, ..., y \$76.29 en 1983, como se ve en la Tabla 19.18.

PROBLEMAS SUPLEMENTARIOS

RELACIONES DE PRECIOS

19.41. La Tabla 19.20 muestra los precios medios al por mayor de los huevos en EE.UU. durante 1978-1985. Hallar la relación de precios (*a*) para 1984 con 1978 como base, (*b*) para 12 como base y (*c*) para

Tabla 19.20

Año	Precio medio de los huevos (centavos por docena)
1978	60.3
1979	66.2
1980	62.8
1981	69.0
1982	66.8
1983	72.7
1984	78.6
1985	63.4

Fuente: U.S. Department of Agriculture.

19.42. Probar (*a*) que $p_{a|b}p_{b|c}p_{c|a} = 1$ y (*b*) que $p_{a|b}p_{b|c}p_{c|d} = p_{a|d}$.

19.43. Probar que $p_{o|n} = p_{o|1}p_{1|2}p_{2|3}\cdots p_{(n-1)|n}$.

19.44. Demostrar que la propiedad circular modificada se sigue directamente de la propiedad circular y de la de inversión temporal.

19.45. La tabla 19.21 recoge las relaciones de precios de un artículo con 1977-1979 = 100. Determinar las relaciones de precios con (*a*) 1980 = 100 y (*b*) 1983-1984 = 100.

Tabla 19.21

Año	Relación de precios (1977-1979 = 100)
1980	127
1981	134
1982	118
1983	125
1984	137
1985	141

19.46. La relación de precios para 1984 con 1986 como base es $62\frac{1}{2}$, mientras que la de 1985 con 1984 como base es $133\frac{1}{3}$. Hallar la relación de precios para el año 1986 con (*a*) 1985 y (*b*) 1984-1985 como base.

19.47. En 1980 el precio medio de un producto decreció un 25% de su valor en 1976, pero creció un 50% de su valor en 1972. Hallar la relación de precios para (*a*) 1976 y (*b*) 1980 con 1972 como base.

RELACIONES DE CANTIDAD O VOLUMEN

19.48. La Tabla 19.22 muestra la energía eléctrica, en miles de millones de kilovatios-hora (kwh) de consumo doméstico, durante los años 1975-1986. Reducir los datos a relaciones de cantidad con (*a*) 1981 y (*b*) 1975-1977 como base.

Tabla 19.22

Año	Energía electrica (miles de millones de kWh)
1975	1.918
1976	2.038
1977	2.124
1978	2.206
1979	2.247
1980	2.286
1981	2.295
1982	2.241
1983	2.310
1984	2.416
1985	2.470
1986	2.512

Fuente: Survey of Current Business.

19.49. En 1984 la producción de un mineral creció un 40% sobre la de 1983. En 1985 la producción estaba un 20% por debajo de la de 1984, pero un $16\frac{2}{3}$% por encima de la de 1986. Hallar las relaciones de precios para los años 1983-1986 con (*a*) 1983, (*b*) 1986 y (*c*) 1983-1986 como base.

19.50. Si la producción del mineral del Problema 19.49 era de 3.20 millones de toneladas cortas en 1985, determinar la producción en (*a*) 1983, (*b*) 1984 y (*c*) 1986.

RELACIONES DE VALOR

19.51. En 1985 el precio de un producto creció un 50% sobre el de 1978 y su producción decreció un 30%. En 1985, ¿en qué porcentaje creció o decreció el valor total en dólares de ese producto con respecto a su valor en 1978?

19.52. La Tabla 19.23 muestra las relaciones de valor y de precios de un artículo en los años 1982-1986, con los periodos indicados como base. Hallar las relaciones de cantidad para ese artículo con (*a*) 1982 y (*b*) 1982-1984 como base. Interpretar los resultados.

Tabla 19.23

Año	Relación de precios (1982 = 100)	Relación de valor (1973-1975 = 100)
1982	100	150
1983	125	180
1984	150	207
1985	175	231
1986	200	252

RELACIONES DE ENLACE Y EN CADENA

19.53. Las relaciones de enlace para el consumo de un producto en los años 1982-1985 fueron 90, 120, 125 y 80, respectivamente.

(*a*) Hallar la relación de precios para 1983 con 1985 como base.

(*b*) Encadenar las relaciones de enlace a una base 1984.

(*c*) Encadenar las relaciones de enlace a una base 1982-1983.

19.54. Al final del primero de n años sucesivos, la producción de un artículo de consumo era de A unidades. En cada año sucesivo la producción aumentó un r% sobre la del año precedente.

(*a*) Probar que la producción durante el n-ésimo año fue de $A(1 + r/100)^{n-1}$ unidades.

(*b*) Probar que la producción total de los n años fue de $(100A/r)[(1+r/100)^n - 1]$ unidades.

NUMEROS INDICE; EL METODO DE AGREGACION SIMPLE

19.55. La Tabla 19.24 muestra los precios y cantidades de consumo en EE. UU. de varios metales para los años 1975 y 1984. Tomando 1975 como base, calcular un índice de precios para el año 1984 por el método de agregación simple.

Tabla 19.24

	Precio (centavos por libra)		Cantidad (millones de libras)	
	1975	1984	1975	1984
Cobre	64.2	66.8	3440	2406
Plomo	21.5	25.6	1144	710
Estaño	339.8	623.8	49.4	42.8
Cinc	39.0	48.6	1068	558

Fuente: U.S. Department of the Interior, Bureau of Mines.

19.56. Demostrar que el número índice por agregación simple satisface el criterio de inversión temporal y circular, pero no el de inversión de factores.

EL METODO DEL PROMEDIO SIMPLE DE RELACIONES

19.57. De los datos de la Tabla 19.24 del Problema 19.55, obtener un índice de precios de esos metales para 1984 con 1975 como año base, mediante un promedio simple (media) de las relaciones de precios. Comparar los resultados con los del Problema 19.55.

19.58. Rehacer el Problema 19.57 usando la mediana.

19.59. Rehacer el Problema 19.57 usando la media geométrica.

19.60. Rehacer el Problema 19.57 usando la media armónica.

EL METODO DE AGREGACION PONDERADA; INDICES DE LASPEYRES Y PAASCHE

19.61. De los datos de la Tabla 19.24, obtener un índice de precios de Laspeyres para 1984 con 1975 como año base.

19.62. De los datos de la Tabla 19.24, obtener un índice de precios de Paasche para 1984 con 1975 como año base.

19.63. Probar que los índices de (*a*) Laspeyres y (*b*) Paasche no satisfacen los criterios de inversión temporal y de inversión de factores.

INDICE IDEAL DE FISHER

19.64. Obtener, de los datos de la Tabla 19.24, un índice ideal de Fisher para 1984 con 1975 como año base.

19.65. Probar que el índice ideal de Fisher no cumple el criterio circular.

EL INDICE DE MARSHALL-EDGEWORTH

19.66. A partir de los datos de la Tabla 19.24, obtener el índice de precios de Marshall-Edgeworth para 1984 con 1975 como base.

19.67. Probar que el índice de Marshall-Edgeworth satisface el criterio de inversión temporal pero no el de inversión de factores.

EL METODO DEL PROMEDIO PONDERADO DE RELACIONES

19.68. Obtener, de los datos de la Tabla 19.24, el número índice de promedio ponderado de relaciones para 1984 con 1975 como año base, usando como pesos (*a*) valores del año dado y (*b*) valores del año base.

NUMEROS INDICE DE CANTIDAD O VOLUMEN

19.69. De los datos de la Tabla 19.24, calcular índices de cantidad para 1984 con 1975 como base, usando (*a*) la media aritmética simple de las relaciones de cantidad, (*b*) la media geométrica simple de las relaciones de cantidad, (*c*) un índice de cantidad de

promedio ponderado con los precios del año base como pesos (número índice de cantidad de Laspeyres), (*d*) un índice de cantidad de promedio ponderado con los precios del año dado como pesos (número índice de cantidad de Paasche), (*e*) el índice ideal de Fisher de cantidad y (*f*) el índice de cantidad de Marshall-Edgeworth.

NUMEROS INDICE DE VALOR

19.70. (*a*) Con 1975 como año base en los datos de la Tabla 19.24, calcular el índice de valor para 1984.

(*b*) Comprobar que el índice de valor en la parte (*a*) es el mismo que el obtenido del producto de los índices ideales de Fisher de precios y de cantidad.

19.71. Con 1975 como año base en los datos de la Tabla 19.24, calcular el índice de precios × el índice de cantidad para 1984 usando el número índice de (*a*) Laspeyres y (*b*) Paasche. Comparar los resultados con el índice de valor real.

19.72. Probar que los números índice de valor por agregación simple satisfacen los criterios circular y de inversión temporal.

CAMBIO DE PERIODO BASE EN LOS NUMEROS INDICE

19.73. La Tabla 19.25 muestra el índice de precios de fábrica en EE. UU. en los años 1973-1983 con base 1967. Hallar el índice de precios con (*a*) 1973 y (*b*) 1976-1978 como base.

Tabla 19.25

Año	Indice de precios en fábrica (1967 = 100)
1973	134.7
1974	160.1
1975	174.9
1976	183.0
1977	194.2
1978	209.3
1979	235.6
1980	268.8

Tabla 19.25. (*Continuación*)

Año	Indice de precios en fábrica (1967 = 100)
1981	293.4
1982	299.3
1983	303.1

Fuente: Survey of Current Business.

19.74. Comparar el índice de precios del Problema 19.73 con el índice de precios al consumo de la Tabla 19.18 del Problema 19.39, y discutir las semejanzas y diferencias entre ambos.

19.75. La Tabla 19.26 presenta los índices de precios al consumo de alimentación, vivienda y atención sanitaria en EE. UU. durante 1973-1983 con 1967 como año base.

(*a*) Obtener los índices de precios con 1973 como base.

(*b*) ¿Se hubieran podido obtener los resultados de la parte (*a*) si no se hubiese dado el año base 1967? Explicar la respuesta.

Tabla 19.26

Año	Alimentación	Vivienda	Atención sanitaria
1973	141.4	133.7	137.7
1974	161.7	148.8	150.5
1975	175.4	164.5	168.6
1976	180.8	174.6	184.7
1977	192.2	186.5	202.4
1978	211.2	202.8	219.4
1979	234.7	227.6	239.7
1980	255.3	263.3	265.9
1981	274.9	293.5	294.5
1982	285.8	314.7	328.7
1983	291.8	323.1	357.3

Fuente: U.S. Bureau of Labor Statistics.

19.76. Con referencia al Problema 19.74, determinar (*a*) el porcentaje de crecimiento de los

costes de la vivienda sobre los de alimentación, (*b*) el porcentaje de crecimiento de la asistencia sanitaria sobre la alimentación, (*c*) el porcentaje de crecimiento o decrecimiento de los costes de asistencia sanitaria sobre los de vivienda y (*d*) el primer año en que el crecimiento de los costes de vivienda sobrepasó al de alimentación.

DEFLACION DE SERIES EN EL TIEMPO

19.77. (*a*) De los datos de la Tabla 19.25 del Problema 19.73, determinar el poder adquisitivo al por mayor de un dólar en cada uno de los años 1973-1983.

(*b*) Comparar los resultados obtenidos en la parte (*a*) con el poder adquisitivo de un dólar en el Problema 19.40, y discutir las razones de sus semejanzas y diferencias.

19.78. ¿Cuánto tendrían que cobrar los trabajadores del Problema 19.39 semanalmente en 1983 para mantener exactamente el mismo nivel de vida que en 1973? Comparar la respuesta con los salarios reales.

19.79. (*a*) Una familia compró una casa en 1975 por $45,000. Suponiendo que no hicieran mejoras en ella, usar la Tabla 19.17 del Problema 19.39 para calcular un precio de reventa justo en 1982.

(*b*) ¿Qué otros factores habría que tener en cuenta al estimar ese precio de reventa?

19.80. Resolver el Problema 19.79 si la familia invirtió en la casa $6000 y $4000 en 1978 y 1980, respectivamente.

19.81. Una serie en el tiempo dada muestra el valor total en dólares de un conjunto de artículos.

(*a*) Describir cómo se podría ajustar la serie en el tiempo para eliminar el efecto del cambio de valor del dólar de año en año.

(*b*) Ilustrar el método de la parte (*a*) con un ejemplo.

19.82. Aplicar el proceso de deflación a la serie en el tiempo de la Tabla 18.44 y explicar el significado de los datos que resultan.

19.83. Probar que el método de deflación de series en el tiempo (tal como se ha utilizado en el Problema 19.39) es estrictamente aplicable sólo si los números índice satisfacen el criterio de inversión de factores.

19.84. En la Tabla 19.27 pueden verse los precios y cantidades de venta al por mayor de varios productos agrícolas en EE. UU. en 1978 y 1985. Todas las cantidades están en millones de bushels excepto las de algodón, que están en millones de balas. Tomando como base 1978, calcular un número índice de precios al por mayor para 1985 usando (*a*) el método de agregación simple, (*b*) un promedio simple (media) de relaciones, (*c*) el índice de Laspeyres, (*d*) el índice de Paasche, (*e*) el índice ideal de Fisher, (*f*) el índice de Marshall-Edgeworth y (*g*) una media aritmética ponderada con los valores del año dado como pesos.

Tabla 19.27

	1978		1985	
	Precio (dólares)	Cantidad	Precio (dólares)	Cantidad
Cebada	2.32	454.8	2.00	589.2
Maíz	2.53	7268	2.41	8865
Algodón	0.592	10.9	0.548	13.4
Avena	1.34	581.7	1.25	518.6
Soja	5.96	1869	5.16	2099
Trigo	3.71	1776	3.16	2425

Fuente: U.S. Department of Agriculture.

19.85. Con 1978 como año base, calcular un número índice de cantidad para 1985 usando los datos de la Tabla 19.27 y cada uno de los métodos del Problema 19.84: (*a*) hasta (*g*).

19.86. Probar que si los números índice de Laspeyres y de Paasche son iguales, entonces coinciden con el índice de Marshall-Edgeworth y con el índice ideal de Fisher.

19.87. Construir una tabla de los diversos tipos de números índice, especificando en cada caso si satisface o no los criterios de inversión temporal, de inversión de factores y circular.

Soluciones a los problemas suplementarios

CAPITULO 1

1.46. (*a*) Continua; (*b*) continua; (*c*) discreta; (*d*) discreta; (*e*) discreta.

1.47. (*a*) De cero en adelante; continua. (*b*) 2, 3, ...; discreta.
(*c*) Soltero, casado, divorciado, separado, viudo; discreta. (*d*) De cero en adelante; continua.
(*e*) 0, 1, 2, ...; discreta.

1.48. (*a*) 3300; (*b*) 5.8; (*c*) 0.004; (*d*) 46.74; (*e*) 126.00; (*f*) 4,000,000; (*g*) 148; (*h*) 0.000099; (*i*) 2180; (*j*) 43.88.

1.49. (*a*) 1,325,000; (*b*) 0.0041872; (*c*) 0.0000280; (*d*) 7,300,000,000; (*e*) 0.0003487; (*f*) 18.50.

1.50. (*a*) 3; (*b*) 4; (*c*) 7; (*d*) 3; (*e*) 8; (*f*) ilimitada; (*g*) 3; (*h*) 3; (*i*) 4; (*j*) 5.

1.51. (*a*) 0.005 millones de bu, o sea 5000 bu; tres. (*b*) 0.000000005 cm, o sea 5×10^{-9} cm; cuatro.
(*c*) 0.5 pies, cuatro. (*d*) 0.05×10^{8} m, o sea 5×10^{6} m; dos. (*e*) 0.5 mi/seg; seis.
(*f*) 500 mi/seg; tres.

1.52. (*a*) 3.17×10^{-4}; (*b*) 4.280×10^{8}; (*c*) 2.160000×10^{4}; (*d*) 9.810×10^{-6}; (*e*) 7.32×10^{5};
(*f*) 1.80×10^{-3}.

1.53. (*a*) 374; (*b*) 14.0.

1.54. (*a*) 280 (dos cifras significativas), 2.8 cientos, o 2.8×10^{2}; (*b*) 178.9;
(*c*) 250,000 (tres cifras significativas). 250 (miles, o 2.50×10^{5}; (*d*) 53.0; (*e*) 5.461; (*f*) 9.05;
(*g*) 11.54; (*h*) 5,745,000 (cuatro cifras significativas), 5745 miles, 5.745 millones, o 5.745×10^{6};
(*i*) 1.2; (*j*) 4157.

1.55. (*a*) -11; (*b*) 2; (*c*) $\frac{35}{8}$, o sea 4.375; (*d*) 21; (*e*) 3; (*f*) -16; (*g*) $\sqrt{98}$, o sea 9.89961 aproximadamente;
(*h*) $-7/\sqrt{34}$, o sea -1.20049 aproximadamente; (*i*) 32; (*j*) $10/\sqrt{17}$, o sea 2.42536 aproximadamente.

1.56. (*a*) 22, 18, 14, 10, 6, 2, -2, -6 y -10; (*b*) 19.6, 16.4, 13.2, 2.8, -0.8, -4 y 8.4;
(*c*) -1.2, 30, $10 - 4\sqrt{2} = 4.34$ aproximadamente y $10 + 4\pi = 22.57$ aproximadamente;
(*d*) 3, 1, 5, 2.1, -1.5, 2.5 y 0; (*e*) $X = \frac{1}{4}(10 - Y)$.

1.57. (*a*) -5; (*b*) -24; (*c*) 8.

1.58. (*a*) -8; (*b*) 4; (*c*) -16.

1.76. (*a*) -4; (*b*) 2; (*c*) 5; (*d*) $\frac{3}{4}$; (*e*) 1; (*f*) -7.

1.77. *a*) $a = 3, b = 4$; (*b*) $a = -2, b = 6$; (*c*) $X = -0.2, Y = -1.2$;
(*d*) $A = \frac{184}{7} = 26.28571$ aproximadamente, $B = \frac{110}{7} = 15.71429$ aproximadamente;
(*e*) $a = 2, b = 3, c = 5$; (*f*) $X = -1, Y = 3, Z = -2$; (*g*) $U = 0.4, V = -0.8, W = 0.3$.

1.78. (*b*) $(2, -3)$; es decir, $X = 2, Y = -3$.

1.79. (*a*) 2, -2.5; (*b*) 2.1 y -0.8 aproximadamente.

1.80. (*a*) $\frac{4 \pm \sqrt{76}}{6}$, o sea 2.12 y -0.79 aproximadamente.
(*b*) 2 y -2.5.
(*c*) 0.549 y -2.549 aproximadamente.
(*d*) $\frac{-8 \pm \sqrt{-36}}{2} = \frac{-8 \pm \sqrt{36}\sqrt{-1}}{2} = \frac{-8 \pm 6i}{2} = -4 \pm 3i$, donde $i = \sqrt{-1}$.

Estas raíces son *números complejos* y no aparecerán cuando se emplee un procedimiento gráfico.

1.81. (*a*) $-6.15 < -4.3 < -1.5 < 1.52 < 2.37$; (*b*) $2.37 > 1.52 > -1.5 > -4.3 > -6.15$.

1.82. (*a*) $30 \leqslant N \leqslant 50$; (*b*) $S \geqslant 7$; (*c*) $-4 \leqslant X < 3$; (*d*) $P \leqslant 5$; (*e*) $X - Y > 2$.

1.83. (*a*) $X \geqslant 4$; (*b*) $X > 3$; (*c*) $N < 5$; (*d*) $Y \leqslant 1$; (*e*) $-8 \leqslant X \leqslant 7$; (*f*) $-1.8 \leqslant N < 3$; (*g*) $2 \leqslant a < 22$.

1.84. (*a*) 2.5877; (*b*) 9.5877 $-$ 10; (*c*) 8.8987 $-$ 10; (*d*) 4.1653; (*e*) 9.7812 $-$ 10; (*f*) 7.4464 $-$ 10; (*g*) 2.6779; (*h*) 0.0030; (*i*) 0.8541; (*j*) 1.8541; (*k*) 6.9912 $-$ 10; (*l*) 7.9275.

1.85. (*a*) 3640; (*b*) 0.675; (*c*) 50.64; (*d*) 0.08445; (*e*) 295.1; (*f*) 0.0002951; (*g*) 0.06314; (*h*) 5096; (*i*) 1202; (*j*) 2,422,000, o sea 2.422×10^6.

1.86. (*a*) 1,296,000, o sea 1.296×10^6; (*b*) 0.05739, o sea 0.0574 con tres cifras significativas; (*c*) 556.0; (*d*) 804.4; (*e*) 40,820; (*f*) 0.03438; (*g*) 15.51; (*h*) 45.67;
(*i*) $0.0004519 = 4.519 \times 10^{-4}$, o sea 4.52×10^{-4} con tres cifras significativas; (*j*) 3096.

1.88. (*a*) $X^2 = 100Y^3$; (*b*) $Y = 3 \times 10^{-2X}$

1.89. (*a*) 3; (*b*) $\frac{3}{2}$; (*c*) -2; (*d*) -5; (*e*) 0.

CAPITULO 2

2.19. (*b*) 62.

2.20. (*a*) 799; (*b*) 1000; (*c*) 949.5; (*d*) 1099.5 y 1199.5; (*e*) 100 horas; (*f*) 76; (*g*) $\frac{62}{400} = 0.155$, o sea, 15.5%; (*h*) 29.5%; (*i*)19.0%; (*j*) 78.0%.

2.25. (*a*) 24%; (*b*) 11%; (*c*) 46%.

2.26. (*a*) 0.003 in; (*b*) 0.3195, 0.3225, 0.3255, ..., 0.3375 in;
(*c*) 0.320-0.322, 0.323-0.325, 0.326-0.328, ..., 0.335-0.337 in.

2.31. (*a*) \$2500 y \$150,000;
(*b*) siete (aunque estrictamente hablando la última clase no tiene tamaño especificado);
(*c*) una (aunque la primera clase parece ser abierta, sustituye realmente a \$0 – \$2499.9);
(d) \$0 – \$2499;
(*e*) \$3749.50 y \$74,999.50 (para la mayor parte de los supuestos prácticos, se pueden presentar como \$3750 y \$75,000, respectivamente.
(*f*) \$9999.50 y \$19,999.50; (*g*) 36.3% y 52.1%; (*h*) 22.7%; (*i*) 8.6%;
(*j*) debido a los errores de redondeo al calcular porcentajes.

2.23. (*a*) 492,100; (*b*) 1,455,000; (*c*) 153,700.

2.34. (*b*) 0.30; (*d*) 0.008 para 4 horas diarias, y 0.52 para 8 horas diarias.

CAPITULO 3

3.47. (*a*) $X_1 + X_2 + X_3 + X_4 + 8$
(*b*) $f_1X_1^2 + f_2X_2^2 + f_3X_3^2 + f_4X_4^2 + f_5X_5^2$
(*c*) $U_1(U_1 + 6) + U_2(U_2 + 6) + U_3(U_3 + 6)$
(*d*) $Y_1^2 + Y_2^2 + \cdots + Y_N^2 - 4N$
(*e*) $4X_1Y_1 + 4X_2Y_2 + 4X_3Y_3 + 4X_4Y_4$

3.48. (*a*) $\sum_{j=1}^{3}(X_j + 3)^3$; (*b*) $\sum_{j=1}^{15} f_j(Y_j - a)^2$; (*c*) $\sum_{j=1}^{N}(2X_j - 3Y_j)$; (*d*) $\sum_{j=1}^{8}\left(\frac{X_j}{Y_j} - 1\right)^2$; (*e*) $\frac{\sum_{j=1}^{12} f_j a_j^2}{\sum_{j=1}^{12} f_j}$

3.51. (*a*) 20; (*b*) – 37; (*c*) 53; (*d*) 6; (*e*) 226; (*f*) – 62; (*g*) $\frac{25}{12}$.

3.52. (*a*) -1; (*b*) 23.

3.53. 86.

3.54. 0.50 seg.

3.55. 8.25.

3.56. (*a*) 82; (*b*) 79.

3.57. 78.

3.58. 80% hombres y 20% mujeres.

3.59. 11.09 ton.

3.60. 501.0.

3.61. 0.72642 cm.

3.62. 26.2.

3.63. 715 horas.

3.64. (*b*) 1.7349 cm.

3.65. (*a*) Media = 5.4, mediana = 5; (*b*) media = 19.91; mediana = 19.85.

3.66. 85.

3.67. 0.51 seg.

3.68. 8.

3.69. 11.07 ton.

3.70. 490.6.

3.71. 0.72638 cm.

3.72. 25.4.

3.73. (*a*) 33.1 para hombres y 30.6 para mujeres.

3.74. $9192.

3.75. 708.3 horas.

3.76. (*a*) Media = 8.9, mediana = 9, moda = 7.
(*b*) Media = 6.4, mediana = 6. Como los números 4, 5, 6, 8 y 10 aparecen dos veces cada uno, podemos considerarlos a todos como modas; no obstante, es más razonable concluir en este caso que no existe moda.

3.77. No existe.

3.78. 0.53 seg.

3.79. 10.

3.80. 11.06 ton.

3.81. 462.

3.82. 0.72632 cm.

3.83. 23.5

3.84. 668.7 horas.

3.88. (*a*) 8.4; (*b*) 4.23.

3.89. (*a*) $G = 8$; (*b*) $\bar{X} = 12.4$.

3.90. (*a*) 4.14; (*b*) 45.8.

3.91. (*a*) 11.07 ton; (*b*) 499.5.

3.92. 18.9%.

3.93. (*a*) 1.086%; (*b*) 212.3 millones; (*c*) 252.3 millones.

3.94. $1586.87.

3.95. $1608.44.

3.96. 3.6 y 14.4.

3.97. (*a*) 3.0; (*b*) 4.48.

3.98. (*a*) 3; (*b*) 0; (*c*) 0.

3.100. (*a*) 11.04; (*b*) 498.2.

3.101. 38.3 mi/h.

3.102. (*b*) 420 mi/h.

3.104. (*a*) 25; (*b*) 3.55.

3.107. (*a*) Cuartil inferior = $Q_1 = 67$, cuartil medio = Q_2 = mediana = 75 y cuartil superior = $Q_3 = 83$. (*b*) El 25% tuvo 67 o menos (o sea, el 75% obtuvo 67 o más), el 50% tuvo 75 o más (luego el 50% tuvo 75 o más), y el 75% tuvo 83 o menos (o sea, el 25% tuvo 83 o más).

3.108. (*a*) $Q_1 = 10.55$ ton, $Q_2 = 11.07$ ton y $Q_3 = 11.57$ ton; (*b*) $Q_1 = 469.3$, $Q_2 = 490.6$ y $Q_3 = 523.3$.

3.109. (*a*) 31.1 y 29.1; (*b*) 39.7 y 36.9; (*c*) 68.8 y 66.2; (*d*) 54.7 y 51.2.

3.110. (*a*) 10.15 ton; (*b*) 11.78 ton; (*c*) 10.55 ton; (*d*) 11.57 ton.

3.112. (*a*) 83; (*b*) 64.

CAPITULO 4

4.33. (*a*) 9; (*b*) 4.273.

4.34. 4.0 ton.

4.35. 0.0036 cm.

4.36. 7.88 kg.

4.37. (*a*) 35; (*b*) indeterminado; (*c*) 900 horas.

4.38. (*a*) 18.2; (*b*) 3.58; (*c*) 6.21; (*d*) 0; (*e*) $\sqrt{2} = 1.414$ aproximadamente; (*f*) 1.88.

4.39. (*a*) 2; (*b*) 0.85.

4.40. (*a*) 2.2; (*b*) 1.317.

4.41. 0.576 ton.

4.42. (*a*) 0.00437 cm; (*b*) 60.0%, 85.2% y 96.4%.

4.43. (*a*) 3.0; (*b*) 2.8.

4.44. (*a*) 31.2; (*b*) 30.6.

4.45. (*a*) 6.0; (*b*) 6.0.

4.48. (*a*) 0.51 ton; (*b*) 27.0; (*c*) 12.

4.49. (*a*) $1801; (*b*) 10.8 años.

4.52. (*a*) 1.63 ton; (*b*) 33.6 o sea 34.

4.53. (*a*) $136,650; (*b*) 42.4 años para hombres y 41.2 años para mujeres.

4.56. (*a*) 2.16; (*b*) 0.90; (*c*) 0.484.

4.58. 45.

4.59. (*a*) 0.733 ton; (*b*) 38.60; (*c*) 12.1.

4.61. (*a*) $\bar{X} = 2.47$; (*b*) $s = 1.11$.

4.63. (*a*) 0.00576 cm; (*b*) 72.1%, 93.3% y 99.76%.

4.64. (*a*) 0.719 ton; (*b*) 38.24; (*c*) 11.8.

4.65. (*a*) 0.000569 cm; (*b*) 71.6% y 99.68%.

4.66. (*a*) 146.8 lb y 12.9 lb.

4.67. (*a*) 1.7349 cm y 0.00495 cm.

4.74. (*a*) 15; (*b*) 12.

4.75. (*a*) estadística; (*b*) álgebra.

4.76. (*a*) 6.6%; (*b*) 19.0%.

4.77. 51.9%.

4.79. Algebra.

4.80. 0.19, −1.75, 1.17, 0.68, −0.29.

CAPITULO 5

5.15. (*a*) 6; (*b*) 40; (*c*) 288; (*d*) 2188.

5.16. (*a*) 0; (*b*) 4; (*c*) 0; (*d*) 25.86.

5.17. (*a*) -1; (*b*) 5; (*c*) -91; (*d*) 53.

5.19. 0, 26.25, 0, 1193.1.

5.21. 7.

5.22. (*a*) 0, 6, 19, 42; (*b*) -4, 22, -117, 560; (*c*) 1, 7, 38, 155.

5.23. 0, 02344, -0.0586, 0.0696.

5.25. (*a*) $m_1 = 0$; (*b*) $m_2 = pq$; (*c*) $m_3 = pq(q - p)$; (*d*) $m_4 = pq(p^2 - pq + q^2)$.

5.27. $m_1 = 0$, $m_2 = 5.97$, $m_3 = -0.397$, $m_4 = 89.22$.

5.29. m_1 (corregido) $= 0$, m_2 (corregido) $= 5.440$, m_3 = (corregido) $= -0.5920$, m_4 (corregido) $= 76.2332$.

5.30. (*a*) $m_1 = 0$, $m_2 = 0.53743$, $m_3 = 0.36206$, $m_4 = 0.84914$;
(*b*) m_2 (corregido) $= 0.51660$, m_4 (corregido) $= 0.78378$.

5.31. (*a*) 0; (*b*) 52.95; (*c*) 92.35; (*d*) 7158.20; (*e*) 26.2; (*f*) 7.28; (*g*) 739.58; (*h*) 22.247; (*i*) 706.428; (*j*) 24,545.

5.32. (*a*) -0.2464; (*b*) -0.2464.

5.33. 0.9190.

5.34. Primera distribución.

5.35. (*a*) 0.040; (*b*) 0.074.

5.36. (*a*) -0.02; (*b*) -0.13.

5.37. (*b*) -0.078,

5.38. (*a*) 2.62; (*b*) 2.58.

5.39. (*a*) 2.94; (*b*) 2.94.

5.40. (*a*) Segunda; (*b*) primera.

5.41. (*a*) Segunda; (*b*) ninguna de ellas; (*c*) primera.

5.42. (*a*) Mayor que 1875; (*b*) igual a 1875; (*c*) menor que 1875.

5.43. (*a*) 0.313.

CAPITULO 6

6.40. (*a*) $\frac{5}{26}$; (*b*) $\frac{5}{36}$; (*c*) 0.98; (*d*) $\frac{2}{9}$; (*e*) $\frac{7}{8}$.

6.41. (*a*) Probabilidad de rey en la primera y no rey en la segunda.
(*b*) Probabilidad de rey en la primera o en la segunda o en ambas.
(*c*) No rey en la primera, no rey en la segunda, o ambas cosas (es decir, ni rey en una ni rey en otra).
(*d*) Probabilidad de rey en la tercera, supuesto que ha salido rey en la primera pero no en la segunda.
(*e*) Ningún rey en las tres extracciones.
(*f*) Probabilidad de rey en la primera y segunda o no rey en la segunda y rey en la tercera.

6.42. (*a*) $\frac{1}{3}$, (*b*) $\frac{3}{5}$; (*c*) $\frac{11}{15}$; (*d*) $\frac{2}{5}$; (*e*) $\frac{4}{5}$.

6.43. (*a*) $\frac{4}{25}$; (*b*) $\frac{4}{75}$; (*c*) $\frac{16}{25}$; (*d*) $\frac{64}{225}$; (*e*) $\frac{11}{15}$; (*f*) $\frac{1}{5}$; (*g*) $\frac{104}{225}$; (*h*) $\frac{221}{225}$; (*i*) $\frac{6}{25}$; (*j*) $\frac{52}{225}$.

6.44. (*a*) $\frac{29}{185}$; (*b*) $\frac{2}{37}$; (*c*) $\frac{118}{185}$; (*d*) $\frac{52}{185}$; (*e*) $\frac{11}{15}$; (*f*) $\frac{1}{5}$; (*g*) $\frac{86}{185}$; (*h*) $\frac{182}{185}$; (*i*) $\frac{9}{37}$; (*j*) $\frac{26}{111}$.

6.45. (*a*) $\frac{5}{18}$; (*b*) $\frac{11}{36}$; (*c*) $\frac{1}{36}$.

6.46. (*a*) $\frac{47}{52}$; (*b*) $\frac{16}{221}$; (*c*) $\frac{15}{34}$; (*d*) $\frac{13}{17}$; (*e*) $\frac{210}{221}$; (*f*) $\frac{10}{13}$; (*g*) $\frac{40}{51}$; (*h*) $\frac{77}{442}$.

6.47. $\frac{5}{18}$.

6.48. (*a*) 81:44, (*b*) 21:4.

6.49. $\frac{19}{42}$.

6.50. (*a*) $\frac{2}{5}$; (*b*) $\frac{1}{5}$; (*c*) $\frac{4}{15}$; (*d*) $\frac{13}{15}$.

6.51. (*a*) 37.5%; (*b*) 93.75%; (*c*) 6.25%; (*d*) 68.75%.

6.52. (*a*)

X	0	1	2	3	4
$p(X)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

6.53. (*a*) $\frac{1}{48}$; (*b*) $\frac{7}{24}$; (*c*) $\frac{3}{4}$; (*d*) $\frac{1}{6}$.

6.54. (*a*)

X	0	1	2	3
$p(X)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{30}$

6.55. (*a*) $\frac{3}{10}$; es la probabilidad de sacar un total de 2 fichas rojas.
(*b*) $\frac{5}{6}$; es la probabilidad de sacar 1, 2 ó 3 fichas rojas (o sea, de extraer al menos una roja).

6.56. \$9.

6.57. \$4.80 diarios.

6.58. A contribuye \$12.50; B contribuye \$7.50.

6.59. (*a*) 7; (*b*) 590; (*c*) 541; (*d*) 10,900.

6.60. (*a*) 1.2; (*b*) 0.56; (*c*) $\sqrt{0.56} = 0.75$ aproximadamente.

6.64. (*a*) 12; (*b*) 2520; (*c*) 720.

6.65. $n = 5$.

6.66. 60.

6.67. (*a*) 5040; (*b*) 720; (*c*) 240.

6.68. (*a*) 8400; (*b*) 2520.

6.69. (*a*) 32,805; (*b*) 11,664.

6.70. 26.

6.71. (*a*) 120; (*b*) 72; (*c*) 12.

6.72. (*a*) 35; (*b*) 70; (*c*) 45.

6.73. $n = 6$.

6.74. 210.

6.75. 840.

6.76. (*a*) 42,000; (*b*) 7000.

6.77. (*a*) 120; (*b*) 12,600.

6.78. (*a*) 150; (*b*) 45; (*c*) 100.

6.79. (*a*) 17; (*b*) 163

6.81. 2.95×10^{25}.

6.83. (*a*) $\frac{6}{5525}$; (*b*) $\frac{22}{425}$; (*c*) $\frac{169}{425}$; (*d*) $\frac{73}{5525}$.

6.84. $\frac{171}{1296}$.

6.85. (*a*) 0.59049; (*b*) 0.32805; (*c*) 0.08866,

6.86. (*b*) $\frac{3}{4}$; (*c*) $\frac{7}{8}$.

6.87. (*a*) 8; (*b*) 78; (*c*) 86; (*d*) 102; (*e*) 20; (*f*)142.

6.90. $\frac{1}{3}$.

CAPITULO 6

6.40. (*a*) $\frac{5}{26}$; (*b*) $\frac{5}{36}$; (*c*) 0.98; (*d*) $\frac{2}{5}$; (*e*) $\frac{7}{8}$.

6.41. (*a*) Probabilidad de rey en la primera y no rey en la segunda.
(*b*) Probabilidad de rey en la primera o en la segunda o en ambas.
(*c*) No rey en la primera, no rey en la segunda, o ambas cosas (es decir, ni rey en una ni rey en otra).
(*d*) Probabilidad de rey en la tercera, supuesto que ha salido rey en la primera pero no en la segunda.
(*e*) Ningún rey en las tres extracciones.
(*f*) Probabilidad de rey en la primera y segunda o no rey en la segunda y rey en la tercera.

6.42. (*a*) $\frac{1}{3}$, (*b*) $\frac{3}{5}$; (*c*) $\frac{11}{15}$; (*d*) $\frac{2}{5}$; (*e*) $\frac{4}{5}$.

6.43. (*a*) $\frac{4}{25}$; (*b*) $\frac{4}{75}$; (*c*) $\frac{16}{25}$; (*d*) $\frac{64}{225}$; (*e*) $\frac{11}{15}$; (*f*) $\frac{1}{5}$; (*g*) $\frac{104}{225}$; (*h*) $\frac{221}{225}$; (*i*) $\frac{6}{25}$; (*j*) $\frac{52}{225}$.

6.44. (*a*) $\frac{29}{185}$; (*b*) $\frac{2}{37}$; (*c*) $\frac{118}{185}$; (*d*) $\frac{52}{185}$; (*e*) $\frac{11}{15}$; (*f*) $\frac{1}{5}$; (*g*) $\frac{86}{185}$; (*h*) $\frac{182}{185}$; (*i*) $\frac{9}{37}$; (*j*) $\frac{26}{111}$.

6.45. (*a*) $\frac{5}{18}$; (*b*) $\frac{11}{36}$; (*c*) $\frac{1}{36}$.

6.46. (*a*) $\frac{47}{52}$; (*b*) $\frac{16}{221}$; (*c*) $\frac{15}{34}$; (*d*) $\frac{13}{17}$; (*e*) $\frac{210}{221}$; (*f*) $\frac{10}{13}$; (*g*) $\frac{40}{51}$; (*h*) $\frac{77}{442}$.

6.47. $\frac{5}{18}$.

6.48. (*a*) 81:44, (*b*) 21:4.

6.49. $\frac{19}{42}$.

6.50. (*a*) $\frac{2}{5}$; (*b*) $\frac{1}{5}$; (*c*) $\frac{4}{15}$; (*d*) $\frac{13}{15}$.

6.51. (*a*) 37.5%; (*b*) 93.75%; (*c*) 6.25%; (*d*) 68.75%.

6.52. (*a*)

X	0	1	2	3	4
$p(X)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

6.53. (*a*) $\frac{1}{48}$; (*b*) $\frac{7}{24}$; (*c*) $\frac{3}{4}$; (*d*) $\frac{1}{6}$.

6.54. (*a*)

X	0	1	2	3
$p(X)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{30}$

6.55. (*a*) $\frac{3}{10}$; es la probabilidad de sacar un total de 2 fichas rojas.
(*b*) $\frac{5}{6}$; es la probabilidad de sacar 1, 2 ó 3 fichas rojas (o sea, de extraer al menos una roja).

6.56. \$9.

6.57. \$4.80 diarios.

6.58. A contribuye \$12.50; B contribuye \$7.50.

6.59. (*a*) 7; (*b*) 590; (*c*) 541; (*d*) 10,900.

6.60. (*a*) 1.2; (*b*) 0.56; (*c*) $\sqrt{0.56} = 0.75$ aproximadamente.

6.64. (*a*) 12; (*b*) 2520; (*c*) 720.

6.65. $n = 5$.

6.66. 60.

6.67. (*a*) 5040; (*b*) 720; (*c*) 240.

6.68. (*a*) 8400; (*b*) 2520.

6.69. (*a*) 32,805; (*b*) 11,664.

6.70. 26.

6.71. (*a*) 120; (*b*) 72; (*c*) 12.

6.72. (*a*) 35; (*b*) 70; (*c*) 45.

6.73. $n = 6$.

6.74. 210.

6.75. 840.

6.76. (*a*) 42,000; (*b*) 7000.

6.77. (*a*) 120; (*b*) 12,600.

6.78. (*a*) 150; (*b*) 45; (*c*) 100.

6.79. (*a*) 17; (*b*) 163

6.81. 2.95×10^{25}.

6.83. (*a*) $\frac{6}{5525}$; (*b*) $\frac{22}{425}$; (*c*) $\frac{169}{425}$; (*d*) $\frac{73}{5525}$.

6.84. $\frac{171}{1296}$.

6.85. (*a*) 0.59049; (*b*) 0.32805; (*c*) 0.08866,

6.86. (*b*) $\frac{3}{4}$; (*c*) $\frac{7}{8}$.

6.87. (*a*) 8; (*b*) 78; (*c*) 86; (*d*) 102; (*e*) 20; (*f*)142.

6.90. $\frac{1}{3}$.

6.91. 1/3,838,380 (es decir, las apuestas están en contra en relación 3,838,379 a 1).

6.92. (*a*) 658,007 a 1; (*b*) 91,389 a 1; (*c*) 9879 a 1.

6.93. (*a*) 649,739 a 1; (*b*) 71,192 a 1; (*c*) 4164 a 1; (*d*) 693 a 1.

6.94. $\frac{11}{36}$.

6.95. $\frac{1}{4}$.

CAPITULO 7

7.35. (*a*) 5040; (*b*) 210; (*c*) 126; (*d*) 165; (*e*) 6.

7.36. (*a*) $q^7+7q^6p+21q^5p^2+35q^4p^3+35q^3p^4+21q^2p^5+7qp^6+p^7$

(*b*) $q^{10}+10q^9p+45q^8p^2+120q^7p^3+210q^6p^4+252q^5p^5+210q^4p^6+120q^3p^7+45q^2p^8+10qp^9+p^{10}$

7.37. (*a*) $\frac{1}{64}$; (*b*) $\frac{3}{32}$; (*c*) $\frac{15}{64}$; (*d*) $\frac{5}{16}$; (*e*) $\frac{15}{64}$; (*f*) $\frac{3}{32}$; (*g*) $\frac{1}{64}$.

7.38. (*a*) $\frac{57}{64}$; (*b*) $\frac{21}{32}$.

7.39. (*a*) $\frac{1}{4}$; (*b*) $\frac{5}{16}$; (*c*) $\frac{11}{16}$; (*d*) $\frac{5}{8}$.

7.40. (*a*) 250; (*b*) 25; (*c*) 500.

7.41. (*a*) $\frac{17}{162}$; (*b*) $\frac{1}{324}$.

7.42. $\frac{64}{243}$.

7.43. $\frac{193}{512}$.

7.44. (*a*) $\frac{32}{243}$; (*b*) $\frac{192}{243}$; (*c*) $\frac{40}{243}$; (*d*) $\frac{242}{243}$.

7.45. (*a*) 42; (*b*) 3.550; (*c*) -0.1127; (*d*) 2.927.

7.47. (*a*) $Npq(q-p)$; (*b*) $Npq(1-6pq)+3N^2p^2q^2$.

7.49. (*a*) 1.5 y -1.6; (*b*) 72 y 90.

7.50. (*a*) 75.4; (*b*) 9.

7.51. (*a*) 0.8767; (*b*) 0.0786; (*c*) 0.2991.

7.52. (*a*) 0.0375; (*b*) 0.7123; (*c*) 0.9265; (*d*) 0.0154; (*e*) 0.7251; (*f*) 0.0395.

7.53. (*a*) 0.9495; (*b*) 0.9500; (*c*) 0.6826.

7.54. (*a*) 0.75; (*b*) -1.86; (*c*) 2.08; (*d*) 1.625 o sea 0.849; (*e*) ± 1.645.

7.55. -0.995.

7.56. (*a*) 0.0317; (*b*) 0.3790; (*c*) 0.1989.

7.57. (*a*) 20; (*b*) 36; (*c*) 227; (*d*) 40.

7.58. (*a*) 93%; (*b*) 8.1%; (*c*) 0.47%; (*d*) 15%.

7.59. 84.

7.60. (*a*) 61.7%; (*b*) 54.7%.

7.61. (*a*) 95.4%; (*b*) 23.0%; (*c*) 93.3%.

7.62. (*a*) 1.15; (*b*) 0.77.

7.63. (*a*) 0.9962; (*b*) 0.0687; (*c*) 0.0286; (*d*) 0.0558.

7.64. (*a*) 0.2511; (*b*) 0.1342.

7.65. (*a*) 0.0567; (*b*) 0.9198; (*c*) 0.6404; (*d*) 0.0079.

7.66. 0.0089.

7.67. (*a*) 0.04979; (*b*) 0.1494; (*c*) 0.2241; (*d*) 0.2241; (*e*) 0.1680; (*f*) 0.1008.

7.68. (*a*) 0.0838; (*b*) 0.5976; (*c*) 0.4232.

7.69. (*a*) 0.05610; (*b*) 0.06131.

7.70. (*a*) 0.00248; (*b*) 0.04462; (*c*) 0.1607; (*d*) 0.1033; (*e*) 0.6964; (*f*) 0.0620.

7.71. (*a*) 0.08208; (*b*) 0.2052; (*c*) 0.2565; (*d*) 0.2138; (*e*) 0.8911; (*f*) 0.0142.

7.72. (*a*) $\frac{5}{3888}$; (*b*) $\frac{5}{324}$.

7.73. (*a*) 0.0348; (*b*) 0.000295.

7.74. $\frac{1}{16}$.

7.75. $p(X)=\binom{4}{X}(0.32)^X(0.68)^{4-X}$. Las frecuencias esperadas son 32, 60, 43, 13 y 2, respectivamente.

7.77. Las frecuencias esperadas son 1.7, 5.5, 12.0, 15.9, 13.7, 7.6, 2.7 y 0.6, respectivamente.

7.78. Las frecuencias esperadas son 1.1, 4.0, 11.1, 23.9, 39.5, 50.2, 49.0, 36.6, 21.1, 9.4, 3.1 y 1.0, respectivamente.

7.79. Las frecuencias esperadas son 41.7, 53.4, 34.2, 14.6 y 4.7, respectivamente.

7.80. $p(X) = \dfrac{(0.61)^X e^{-0.61}}{X!}$. Las frecuencias esperadas son 108.7, 66.3, 20.2, 4.1 y 0.7, respectivamente.

CAPITULO 8

8.21. (*a*) 9.0; (*b*) 4.47; (*c*) 9.0; (*d*) 3.16.

8.22. (*a*) 9.0; (*b*) 4.47; (*c*) 9.0; (*d*) 2.58.

8.23. (*a*) $\mu_{\bar{X}} = 22.40$ g, $\sigma_{\bar{X}} = 0.008$ g; (*b*) $\mu_{\bar{X}} = 22.40$ g, $\sigma_{\bar{X}}$ = ligeramente menor que 0.008 g.

8.24. (*a*) $\mu_{\bar{X}} = 22.40$ g, $\sigma_{\bar{X}} = 0.008$ g; (*b*) $\mu_{\bar{X}} = 22.40$ g, $\sigma_{\bar{X}} = 0.0057$ g.

8.25. (*a*) 237; (*b*) 2; (*c*) ninguna, (*d*) 34.

8.26. (*a*) 0.4972; (*b*) 0.1587; (*c*) 0.0918; (*d*) 0.9544.

8.27. (*a*) 0.8164; (*b*) 0.0228; (*c*) 0.0038; (*d*) 1.0000.

8.28. 0.0026.

8.34. (*a*) 0.0029; (*b*) 0.9596; (*c*) 0.1446.

8.35. (*a*) 2; (*b*) 996; (*c*) 218.

8.36. (*a*) 0.0179; (*b*) 0.8664; (*c*) 0.1841.

8.37. (*a*) 6; (*b*) 9; (*c*) 2; (*d*) 12.

8.39. (*a*) 19; (*b*) 125.

8.40. (*a*) 0.0077; (*b*) 0.8869.

8.41. (*a*) 0.0028; (*b*) 0.9172.

8.42. (*a*) 0.2150; (*b*) 0.0064, 0.4504.

8.43. 0.0482.

8.44. 0.0188.

8.45. 0.0410.

8.47. (*a*) 118.79 g; (*b*) 0.74 g.

8.48. 0.0228.

8.49. (*a*) 7.2; (*b*) 8.4.

8.50. (*a*) 106; (*b*) 4.

8.51. 159.

8.52. (*a*) 78.7; (*b*) 0.0090.

CAPITULO 9

9.21. (*a*) 9.5 kg; (*b*) 0.74 kg^2; (*c*) 0.78 kg y 0.86 kg, respectivamente.

9.22. (*a*) 1200 h; (*b*) 105.4 h.

9.23. (*a*) Las estimaciones de desviaciones típicas de la población para muestras de 30, 50 y 100 tubos son 101.7 h, 101.0 h y 100.5 h, respectivamente; las estimaciones de medias de la población son 1200 h en todos los casos.

9.24. (*a*) 11.09 $\pm$ 0.18 ton; (*b*) 11.09 $\pm$ 0.24 ton.

9.25. (*a*) 0.72642 $\pm$ 0.000095 in; (*b*) 0.72642 $\pm$ 0.000085 in; (*c*) 0.72642 $\pm$ 0.000072 in; (*d*) 0.72642 $\pm$ 0.000060 in.

9.26. (*a*) 0.72642 $\pm$ 0.000025 in; (*b*) 0.000025 in.

9.27. (*a*) Al menos 97; (*b*) al menos 68; (*c*) al menos 167; (*d*) al menos 225.

9.28. (*a*) Al menos 385; (*b*) al menos 271; (*c*) al menos 666; (*d*) al menos 900.

9.29. (*a*) 2400 $\pm$ 45 lb, 2400 $\pm$ 59 lb; (*b*) 87.6%.

9.30. (*a*) 0.70 $\pm$ 0.12, 0.69 $\pm$ 0.11; (*b*) 0.70 $\pm$ 0.15, 0.68 $\pm$ 0.15; (*c*) 0.70 $\pm$ 0.18, 0.67 $\pm$ 0.17.

9.31. (*a*) Al menos 323; (*b*) al menos 560; (*c*) al menos 756.

9.32. (*a*) 16,400; (*b*) 27,100; (*c*) 38,420; (*d*) 66,000.

9.33. (*a*) 1.07 $\pm$ 0.09 h; (*b*) 1.07 $\pm$ 0.12 h.

9.34. (*a*) 0.045 $\pm$ 0.073; (*b*) 0.045 $\pm$ 0.097; (*c*) 0.045 $\pm$ 0.112.

9.35. (*a*) 63.8 $\pm$ 0.24 lb; (*b*) 63.8 $\pm$ 0.31 lb.

9.36. (*a*) 180 $\pm$ 24.9 lb; (*b*) 180 $\pm$ 32.8 lb; (*c*) 180 $\pm$ 38.2 lb.

9.37. 8.6 lb.

9.38. (*a*) Al menos 4802; (*b*) al menos 8321; (*c*) al menos 11,250.

CAPITULO 10

10.29. (*a*) 0.2606.

10.30. (*a*) Aceptar la hipótesis si se sacan entre 22 y 42 rojas, y rechazarla en caso contrario; (*b*) 0.99; (*c*) aceptar la hipótesis si se sacan entre 24 y 40 rojas, y rechazarla en caso contrario.

10.31. (*a*) H_0: $p = 0.5$, H_1: $p > 0.5$; (*b*) criterio de una cola; (*c*) rechazar H_0 si se sacan más de 39 rojas, y aceptarla en caso contrario (o aplazar la decisión); (*d*) rechazar H_0 si se sacan más de 41 rojas, y aceptarla en caso contrario (o aplazar la decisión).

10.32. (*a*) No se puede rechazar la hipótesis al nivel 0.05; (*b*) se puede rechazar la hipótesis al nivel 0.05.

10.33. No se puede rechazar al nivel 0.01, ni con criterio unilateral ni con bilateral.

10.34. Usando un criterio unilateral, la podemos rechazar a ambos niveles.

10.35. Con criterio de una cola, el resultado es significativo al nivel 0.05, pero no al 0.01.

10.36. Sí, es significativo a ambos niveles, usando en cada caso criterio unilateral.

10.37. Tanto con criterio de una como de dos colas, el resultado es significativo al nivel 0.05.

10.38. El resultado es significativo al nivel 0.01 usando un criterio de una cola, pero no con uno de dos colas.

10.39. (*a*) 0.3112; (*b*) 0.0118; (*c*) 0; (*d*) 0; (*e*) 0.0118.

10.43. (*a*) 8.64 ± 0.96 oz; (*b*) 8.64 ± 0.83 oz; (*c*) 8.64 ± 0.63 oz.

10.44. Los límites superiores de control son, respectivamente, (*a*) 6 y (*b*) 4 defectuosos.

10.45. (*a*) Sí; (*b*) no.

10.46. Un criterio unilateral a ambos niveles de significación muestra que B es superior a A.

10.47. Un criterio de una cola muestra que la diferencia es significativa al nivel 0.05, pero no al 0.01.

10.48. Un criterio de una cola demuestra que el nuevo fertilizante es superior a ambos niveles de significación.

10.49. (*a*) Un criterio bilateral muestra que no hay diferencia al nivel 0.05.
(*b*) Un criterio unilateral muestra que B no es mejor que A al nivel 0.05.

10.50. (*a*) Un criterio de dos colas al nivel 0.05 no rechaza la hipótesis de proporciones iguales.
(*b*) Un criterio de una cola al nivel 0.05 muestra que A tiene mayor proporción de rojas que B.

10.51. (*a*) 9; (*b*) 10; (*c*) 10; (*d*) 8.

10.54. (*a*) No; (*b*) sí; (*c*) no.

10.55. (*a*) Sí; (*b*) sí; (*c*) no.

10.56. (*a*) Sí, (*b*) sí; (*c*) sí.

10.57. (*a*) No; (*b*) no; (*c*) no.

CAPITULO 11

11.20. (*a*) 2.60; (*b*) 1.75; (*c*) 1.34; (*d*) 2.95; (*e*) 2.13.

11.21. (*a*) 3.75; (*b*) 2.68; (*c*) 2.48; (*d*) 2.39; (*e*) 2.33.

11.22. (*a*) 1.71; (*b*) 2.09; (*c*) 4.03; (*d*) -0.128.

11.23. (*a*) 1.81; (*b*) 2.76; (*c*) -0.879; (*d*) -1.37.

11.24. (*a*) ± 4.60; (*b*) ± 3.06; (*c*) ± 2.79; (*d*) ± 2.75; (*e*) ± 2.70.

11.25. (*a*) 7.38 ± 0.82 g; (*b*) 7.38 ± 1.16 g.

11.26. (*a*) 7.38 ± 0.73 g; (*b*) 7.38 ± 0.96 g.

11.27. (*a*) 0.298 ± 0.030 seg; (*b*) 0.298 ± 0.049 seg.

11.28. Un criterio de dos colas enseña que no hay evidencia ni al nivel 0.05 ni al 0.01 de que la vida media haya cambiado.

11.29. Un criterio de una cola no pone de manifiesto decrecimiento en la media ni al nivel 0.05 ni al 0.01.

11.30. Un criterio de dos colas a ambos niveles muestra que el producto no cumple las especificaciones requeridas.

11.31. Un criterio unilateral a ambos niveles muestra que el contenido medio de cobre es mayor que lo que las especificaciones exigen.

11.32. Un criterio de una cola muestra que el proceso no debe ser introducido si el nivel adoptado es el 0.01 pero sí en caso de adoptar el nivel 0.05.

11.33. Un criterio unilateral muestra que A es menor que B al nivel 0.05 de significación.

11.34. Con un criterio bilateral al nivel 0.05 no concluimos, a la vista de las muestras, que haya diferencia en acidez entre los dos tipos.

11.35. Con un criterio de una cola al nivel 0.05, concluimos que el primer grupo no es superior al segundo.

11.36. (*a*) 21.0; (*b*) 26.2; (*c*) 23.3.

11.37. (*a*) 15.5; (*b*) 30.1; (*c*) 41.3; (*d*) 55.8.

11.38. (*a*) 20.1; (*b*) 36.2; (*c*) 48.3; (*d*) 63.7.

11.39. (*a*) $\chi_1^2 = 9.59$ y $\chi_2^2 = 34.2$.

11.40. (*a*) 16.0; (*b*) 6.35; (*c*) suponiendo áreas iguales en ambas colas, $\chi_1^2 = 2.17$ y $\chi_2^2 = 14.1$.

11.41. (*a*) 87.0 a 230.9 h; (*b*) 78.1 a 288.5 h.

11.42. (*a*) 95.6 a 170.4 h; (*b*) 88.9 a 190.8 h.

11.43. (*a*) 122.5; (*b*) 179.2.

11.44. (*a*) 207.7; (*b*) 295.2.

11.46. (*a*) 106.1 a 140.5 h; (*b*) 102.1 a 148.1 h.

11.47. 105.5 a 139.6 h.

11.48. Sobre la base de la muestra dada, el aparente crecimiento en variabilidad no es significativo en esos dos niveles.

11.49. El aparente decrecimiento en variabilidad es significativo al nivel 0.05, pero no al 0.01.

11.50. (*a*) $F_{.95} = 3.07$; (*b*) $F_{.99} = 4.02$; (*c*) $F_{.95} = 2.11$; (*d*) $F_{.99} = 2.83$.

11.51. $F_{.95} = 1.95$, usando interpolación.

11.52. La varianza de la muestra 1 es significativamente mayor al nivel 0.05, pero no al 0.01.

11.53. (*a*) Sí; (*b*) no.

CAPITULO 12

12.26. La hipótesis no es rechazable en ninguno de los dos niveles.

12.27. Misma conclusión que antes.

12.28. El nuevo no sigue el esquema de los otros. (El hecho de que las calificaciones sean mejores que la media *puede* ser debido a una especial habilidad para la enseñanza o a menor exigencia, o a ambas cosas a la vez.)

12.29. No hay razón para rechazar la hipótesis de que las monedas son buenas.

12.30. No hay razón para rechazar la hipótesis a ninguno de los niveles.

12.31. (*a*) 10, 60 y 50, respectivamente;
(*b*) la hipótesis de que los resultados son los esperados no se puede rechazar al nivel 0.05.

12.32. La diferencia es significativa al nivel 0.05.

12.33. (*a*) El ajuste es bueno; (*b*) no.

12.34. (*a*) El ajuste es «demasiado bueno»; (*b*) el ajuste es pobre al nivel 0.05.

12.35. (*a*) El ajuste es muy malo al nivel 0.05; como la distribución binomial da un buen ajuste de los datos, esto es consistente con el Problema 12.33.
(*b*) El ajuste es bueno, pero no «demasiado bueno».

12.36. La hipótesis se puede rechazar al nivel 0.05 pero no al 0.01.

12.37. Misma conclusión que antes.

12.38. La hipótesis no se puede rechazar a esos niveles.

12.39. La hipótesis no se puede rechazar al nivel 0.05.

12.40. La hipótesis se puede rechazar a ambos niveles.

12.41. La hipótesis se puede rechazar a ambos niveles.

12.42. La hipótesis no se puede rechazar ni a un nivel ni al otro.

12.49. (*a*) 0.3863 (sin corregir) y (*b*) 0.3779 (con la corrección de Yates).

12.50. (*a*) 0.2205, 0.1985 (corregidos); (*b*) 0.0872, 0.0738 (corregido).

12.51. 0.4651.

12.54. (*a*) 0.4188, 0.4082 (corregido).

12.55. (*a*) 0.2261, 0.2026 (corregido); (*b*) 0.0875, 0.0740 (corregido).

12.56. 0.3715.

CAPITULO 13

13.24. (*a*) 4; (*b*) 6; (*c*) $\frac{28}{3}$, (*d*) 10.5; (*e*) 6; (*f*) 9.

13.25. (2, 1).

13.26. (*a*) $2X + Y = 4$; (*b*) X intersección $= 2$, Y intersección $= 4$; (*c*) $-2, -6$.

13.27. $Y = \frac{2}{3}X - 3$, o sea $2X - 3Y = 9$.

13.28. (*a*) Pendiente $= \frac{3}{5}$, Y intersección $= -4$; (*b*) $3X - 5Y = 11$.

13.29. (*a*) $-\frac{4}{3}$; (*b*) $\frac{32}{3}$; (*c*) $4X + 3Y = 32$.

13.30. $X/3 + Y/(-5) = 1$, o sea $5X - 3Y = 15$.

13.31. (*a*) °F $= \frac{9}{5}$ °C $+ 32$; (*b*) 176 °F; (*c*) 20 °C.

13.32. (*a*) $Y = -\frac{1}{3} + \frac{5}{7}X$, o sea $Y = -0.333 + 0.714X$: (*b*) $X = 1 + \frac{9}{7}Y$, o sea $X = 1.00 + 1.29Y$.

13.33. (*a*) 3.24, 8.24; (*b*) 10.00.

13.35. (*b*) $Y = 29.13 + 0.661X$; (*c*) $X = -14.39 + 1.15Y$; (*d*) 79; (*e*) 95.

13.38. $Y = 5.51 + 3.20(X - 3) + 0.733(X - 3)^2$, o sea $Y = 2.51 - 1.20X + 0.733X^2$.

13.39. (*b*) $D = 41.77 - 1.096V + 0.08786V^2$; (*c*) 170 pies, 516 pies.

13.43. (*b*) $Y = 32.14(1.427)^X$, o sea $Y = 32.14(10)^{0.1544X}$, o sea $Y = 32.14\ e^{0.3556X}$, donde $e = 2.718...$ es la base de los logaritmos naturales.
(*d*) 387.

CAPITULO 14

14.40. (*b*) $Y = 4.000 + 0.500X$; (*c*) $X = 2.408 + 0.612Y$.

14.41. (*a*) 1.304; (*b*) 1.443.

14.42. (*a*) 24.50; (*b*) 17.00; (*c*) 7.50.

14.43. 0.5533.

14.45. 1.5.

14.46. (*a*) 0.8961; (*b*) $Y = 80.78 + 1.138X$; (*c*) 132.

14.47. (*a*) 0.958; (*b*) 0.872.

14.48. (*a*) $Y = 0.8X + 12$; (*b*) $X = 0.45Y + 1$.

14.49. (*a*) 1.60; (*b*) 1.20.

14.50. ± 0.80.

14.51. 75%.

14.53. (*a*) -0.9203.

14.54. (*a*) $Y = 18.04 - 1.34X$, $Y = 51.18 - 2.01X$.

14.58. 0.5440.

14.59. (*a*) $Y = 4.44X - 142.22$; (*b*) 141.9 lb y 177.5 lb, respectivamente.

14.60. (*a*) 16.92 lb; (*b*) 2.07 in.

14.62. 0.754.

14.63. 0.22.

14.64. (*a*) Sí; (*b*) no.

14.65. (*a*) No; (*b*) sí.

14.66. (*a*) 0.2923 y 0.7951; (*b*) 0.1763 y 0.8361.

14.67. (*a*) 0.3912 y 0.7500; (*b*) 0.3146 y 0.7861

14.68. (*a*) 0.7096 y 0.9653; (*b*) 0.4961 y 0.7235.

14.69. (*a*) Sí; (*b*) no.

14.70. (*a*) 2.00 ± 0.21; (*b*) 2.00 ± 0.28,

14.71. (*a*) Usando un criterio de una cola podemos rechazarla.
(*b*) Usando un criterio de una cola no podemos rechazarla.

14.72. (*a*) 37.0 ± 3.28; (*b*) 37.0 ± 4.45.

14.73. (*a*) 37.0 ± 0.69; (*b*) 37.0 ± 0.94.

14.74. (*a*) 1.138 ± 0.398; (*b*) 132.0 ± 16.6; (*c*) 132.0 ± 5.4.

CAPITULO 15

15.26. (*a*) $X_3 = b_{3.12} + b_{31.2}X_1 + b_{32.1}X_2$; (*b*) $X_4 = b_{4.1235} + b_{41.235}X_1 + b_{42.135}X_2 + b_{43.125}X_3$.

15.28. (*a*) $X_3 = 61.40 - 3.65X_1 + 2.54X_2$; (*b*) 40.

15.29. (*a*) $X_3 - 74 = 4.36(X_1 - 6.8) + 4.04(X_2 - 7.0)$, o sea $X_3 = 16.07 + 4.36X_1 + 4.04X_2$; (*b*) 84 y 66.

15.31. 3.12.

15.32. (*a*) 5.883; (*b*) 0.6882.

15.33. 0.9927.

15.34. (*a*) 0.7567; (*b*) 0.7255; (*c*) 0.6810.

15.37. (*a*) 0.5950; (*b*) -0.8995; (*c*) 0.8727.

15.38. (*a*) 0.2672; (*b*) 0.5099; (*c*) 0.4026.

15.42. (*a*) $X_4 = 6X_1 + 3X_2 - 4X_3 - 100$; (*b*) 54.

15.43. (*a*) 0.8710; (*b*) 0.8587; (*c*) -0.8426.

15.44. (*a*) 0.8947; (*b*) 2.680.

CAPITULO 16

16.21. Hay diferencia significativa a ambos niveles.

16.22. No hay diferencia significativa a ambos niveles.

16.23. Hay diferencia significativa entre los métodos de enseñanza al nivel 0.05 pero no al 0.01.

16.24. Hay diferencia significativa al nivel 0.05 pero no al 0.01.

16.25. Hay diferencia significativa entre las calificaciones a ambos niveles.

16.26. No hay diferencia significativa entre operarios o entre máquinas.

16.27. Misma respuesta que en el Problema 16.26.

16.28. Al nivel 0.05 hay diferencia significativa en términos del tipo de maíz, pero no en términos del terreno.

16.29. Al nivel 0.01 no hay diferencia significativa según el tipo de maíz ni del tipo de terreno.

16.30. Al nivel 0.05 hay diferencia significativa entre los neumáticos y entre los automóviles.

16.31. Al nivel 0.01 no hay diferencia significativa entre los neumáticos ni entre los automóviles.

16.32. Al nivel 0.01 hay diferencia significativa entre los métodos, pero no entre los colegios.

16.33. No hay diferencia significativa ni en el color del cabello ni en la altura.

16.34. Misma respuesta que en el Problema 16.33.

16.35. Al nivel 0.05 hay diferencia significativa debida a los lugares, pero no debida a los fertilizantes.

16.36. Al nivel 0.01 no hay diferencia significativa debida a los lugares ni a los fertilizantes.

16.37. Hay diferencia significativa entre operarios, no entre máquinas.

16.38. No hay diferencia significativa ni entre terrenos ni entre fertilizantes.

16.39. Misma respuesta que en el Problema 16.38.

16.40. No hay diferencia significativa debida a diferencias en altura, color del cabello o lugar de nacimiento.

16.41. Hay diferencia significativa en términos de las especies de gallinas y de las cantidades del primer producto, pero no en el segundo ni en el peso inicial de las gallinas.

16.42. Hay diferencia significativa debida al tipo de cable, pero no debida a los operarios, a las máquinas o a las empresas.

16.43. No hay diferencia significativa a ninguno de los niveles.

16.44. No hay diferencia significativa a ninguno de los niveles.

16.46. Al nivel 0.05 hay una diferencia en los resultados debida tanto al grado de veteranía como al IQ.

16.47. Al nivel 0.01 la diferencia en los resultados debida a la veteranía no es significativa, pero sí lo es la debida al IQ.

16.48. No hay diferencias significativas en términos de los lugares de procedencia de los estudiantes, pero sí en términos del IQ.

16.49. Misma respuesta que en el Problema 16.48.

16.53. Al nivel 0.05 hay una diferencia debida tanto a los productos como a los lugares.

16.54. Al nivel 0.05 hay diferencia en los resultados debida a los lugares, pero no a los fertilizantes.

16.55. Al nivel 0.01 no hay diferencia debida a los lugares ni a los fertilizantes.

16.56. No hay diferencia significativa debida a los factores 1 y 2, ni a los tratamientos A, B y C.

16.58. No hay diferencia significativa debida a los factores ni a los tratamientos.

CAPITULO 17

17.26. Hay diferencia al nivel 0.05, no al 0.01.

17.27. Sí.

17.28. El programa es eficaz al nivel 0.05.

17.29. Podemos rechazar la hipótesis de crecimiento en las ventas al nivel 0.05.

17.30. No.

17.31. (*a*) Rechazar; (*b*) aceptar; (*c*) aceptar; (*d*) rechazar.

17.34. No hay diferencia al nivel 0.05.

17.35. No.

17.36. (*a*) Sí; (*b*) sí.

17.37. Sí.

17.38. (*a*) Sí; (*b*) sí.

17.41. 3.

17.42. 6.

17.49. No hay diferencia significativa en ninguno de los niveles.

17.50. La diferencia es significativa al nivel 0.05, pero no al 0.01.

17.51. La diferencia es significativa al nivel 0.05, pero no al 0.01.

17.52. Hay diferencia significativa entre las calificaciones en ambos niveles.

17.55. (*a*) 8; (*b*) 10.

17.56. (*a*) 10; (*b*) las respuestas son aleatorias al nivel 0.05.

17.62. La muestra no es aleatoria al nivel 0.05. Hay *demasiadas* rachas, que indican un esquema cíclico.

17.63. La muestra no es aleatoria al nivel 0.05. Hay *demasiado pocas* rachas, lo que indica un esquema de tendencia.

17.64. Los dígitos son aleatorios al nivel 0.05.

17.65. (*a*) Los digitos son aleatorios al nivel 0.05; (*b*) los digitos son aleatorios al nivel 0.05.

17.69. (*a*) 0.67; (*b*) los jueces no coincidieron demasiado bien en sus elecciones.

CAPITULO 18

18.22. (*a*) Cíclico; (*b*) estacional; (*c*) a largo término; (*d*) irregular; (*e*) a largo término.

18.23. (*a*) 0.5, −0.5, −0.5, 0.5, 0.5, −0.5, −0.5, 0.5; (*b*) 0, $-\frac{1}{3}$, 0, $\frac{1}{3}$, 0, $-\frac{1}{3}$, 0; (*c*) 0, 0, 0, 0, 0, 0; (*d*) $\frac{1}{5}$, 0, $-\frac{1}{5}$, 0, $\frac{1}{5}$.

18.28. (*b*) 0, −0.5, 0, 0.5, 0, −0.5, 0; (*c*) $-\frac{1}{6}$, $-\frac{1}{6}$, $\frac{1}{6}$ $\frac{1}{6}$, $-\frac{1}{6}$, $-\frac{1}{6}$; (*d*) 0, 0, 0, 0, 0.

18.30. (*a*) 20; (*b*) 21; (*c*) 196.

CAPITULO 19

19.41. (*a*) 130.3; (*b*) 105.4, 91.9; (*c*) 95.6, 104.9, 99.5, 109.4, 105.9, 115.2, 124.6, 100.5.

19.45. (*a*) 100, 106, 93, 98, 108, 111; (*b*) 97, 102, 90, 95, 105, 108.

19.46. (*a*) 120; (*b*) 137.

19.47. (*a*) 200; (*b*) 150.

19.48. (*a*) 83.6, 88.8, 92.5, 96.1, 97.9, 99.6, 100.0, 97.6, 100.7, 105.3, 107.6, 109.5.
(*b*) 94.6, 100.6, 104.8, 108.8, 110.9, 112.8, 113.2, 110.6, 114.0, 119.2, 121.9, 123.9.

19.49. (*a*) 100, 140, 112, 96; (*b*) 104, 146, 117, 100; (*c*) 89.3, 125, 100, 85.7.

19.50. (*a*) 2.86; (*b*) 4.00; (*c*) 2.74 millones de toneladas cortas.

19.51. 5% de crecimiento.

19.52. (*a*) 100, 96, 92, 88, 84; (*b*) 104, 100, 96, 92, 98.

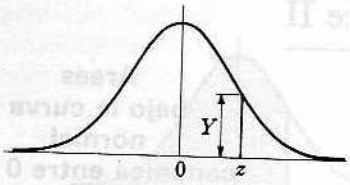
19.53. (*a*) 100; (*b*) 74.1, 66.7, 80.0, 100 y 80.0, correspondientes a los años 1981-1985, respectivamente; (*c*) 101, 90.9, 109, 136 y 109, correspondientes a los años 1981-1985, respectivamente.

19.78. $214.04.

Apéndices

Apéndice I

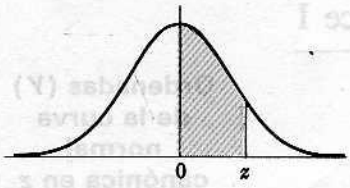
Ordenadas (Y) de la curva normal canónica en z



z	0	1	2	3	4	5	6	7	8	9
0.0	.3989	.3989	.3989	.3988	.3986	.3984	.3982	.3980	.3977	.3973
0.1	.3970	.3965	.3961	.3956	.3951	.3945	.3939	.3932	.3925	.3918
0.2	.3910	.3902	.3894	.3885	.3876	.3867	.3857	.3847	.3836	.3825
0.3	.3814	.3802	.3790	.3778	.3765	.3752	.3739	.3725	.3712	.3697
0.4	.3683	.3668	.3653	.3637	.3621	.3605	.3589	.3572	.3555	.3538
0.5	.3521	.3503	.3485	.3467	.3448	.3429	.3410	.3391	.3372	.3352
0.6	.3332	.3312	.3292	.3271	.3251	.3230	.3209	.3187	.3166	.3144
0.7	.3123	.3101	.3079	.3056	.3034	.3011	.2989	.2966	.2943	.2920
0.8	.2897	.2874	.2850	.2827	.2803	.2780	.2756	.2732	.2709	.2685
0.9	.2661	.2637	.2613	.2589	.2565	.2541	.2516	.2492	.2468	.2444
1.0	.2420	.2396	.2371	.2347	.2323	.2299	.2275	.2251	.2227	.2203
1.1	.2179	.2155	.2131	.2107	.2083	.2059	.2036	.2012	.1989	.1965
1.2	.1942	.1919	.1895	.1872	.1849	.1826	1.804	.1781	.1758	.1736
1.3	.1714	.1691	.1669	.1647	.1626	.1604	.1582	.1561	.1539	.1518
1.4	.1497	.1476	.1456	.1435	.1415	.1394	.1374	.1354	.1334	.1315
1.5	.1295	.1276	.1257	.1238	.1219	.1200	.1182	.1163	.1145	.1127
1.6	.1109	.1092	.1074	.1057	.1040	.1023	.1006	.0989	.0973	.0957
1.7	.0940	.0925	.0909	.0893	.0878	.0863	.0848	.0833	.0818	.0804
1.8	.0790	.0775	.0761	.0748	.0734	.0721	.0707	.0694	.0681	.0669
1.9	.0656	.0644	.0632	.0620	.0608	.0596	.0584	.0573	.0562	.0551
2.0	.0540	.0529	.0519	.0508	.0498	.0488	.0478	.0468	.0459	.0449
2.1	.0440	.0431	.0422	.0413	.0404	.0396	.0387	.0379	.0371	.0363
2.2	.0355	.0347	.0339	.0332	.0325	.0317	.0310	.0303	.0297	.0290
2.3	.0283	.0277	.0270	.0264	.0258	.0252	.0246	.0241	.0235	.0229
2.4	.0224	.0219	.0213	.0208	.0203	.0198	.0194	.0189	.0184	.0180
2.5	.0175	.0171	.0167	.0163	.0158	.0154	.0151	.0147	.0143	.0139
2.6	.0136	.0132	.0129	.0126	.0122	.0119	.0116	.0113	.0110	.0107
2.7	.0104	.0101	.0099	.0096	.0093	.0091	.0088	.0086	.0084	.0081
2.8	.0079	.0077	.0075	.0073	.0071	.0069	.0067	.0065	.0063	.0061
2.9	.0060	.0058	.0056	.0055	.0053	.0051	.0050	.0048	.0047	.0046
3.0	.0044	.0043	.0042	.0040	.0039	.0038	.0037	.0036	.0035	.0034
3.1	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026	.0025	.0025
3.2	.0024	.0023	.0022	.0022	.0021	.0020	.0020	.0019	.0018	.0018
3.3	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014	.0013	.0013
3.4	.0012	.0012	.0012	.0011	.0011	.0010	.0010	.0010	.0009	.0009
3.5	.0009	.0008	.0008	.0008	.0008	.0007	.0007	.0007	.0007	0.006
3.6	.0006	.0006	.0006	.0005	.0005	.0005	.0005	.0005	.0005	.0004
3.7	.0004	.0004	.0004	.0004	.0004	.0004	.0003	.0003	.0003	.0003
3.8	.0003	.0003	.0003	.0003	.0003	.0002	.0002	.0002	.0002	.0002
3.9	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001

Apéndice II

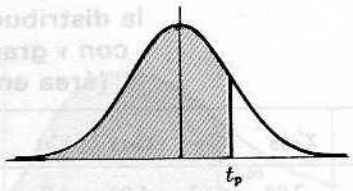
Áreas bajo la curva normal canónica entre 0 y z



z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

Apéndice III

Valores percentiles (t_p) para la distribución t de Student con ν grados de libertad (área en sombra = p)

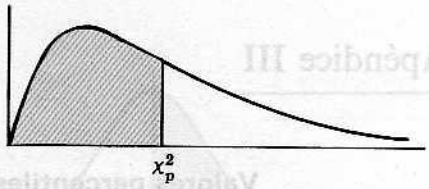


ν	$t_{.995}$	$t_{.99}$	$t_{.975}$	$t_{.95}$	$t_{.90}$	$t_{.80}$	$t_{.75}$	$t_{.70}$	$t_{.60}$	$t_{.55}$
1	63.66	31.82	12.71	6.31	3.08	1.376	1.000	.727	.325	.158
2	9.92	6.96	4.30	2.92	1.89	1.061	.816	.617	.289	.142
3	5.84	4.54	3.18	2.35	1.64	.978	.765	.584	.277	.137
4	4.60	3.75	2.78	2.13	1.53	.941	.741	.569	.271	.134
5	4.03	3.36	2.57	2.02	1.48	.920	.727	.559	.267	.132
6	3.71	3.14	2.45	1.94	1.44	.906	.718	.553	.265	.131
7	3.50	3.00	2.36	1.90	1.42	.896	.711	.549	.263	.130
8	3.36	2.90	2.31	1.86	1.40	.889	.706	.546	.262	.130
9	3.25	2.82	2.26	1.83	1.38	.883	.703	.543	.261	.129
10	3.17	2.76	2.23	1.81	1.37	.879	.700	.542	.260	.129
11	3.11	2.72	2.20	1.80	1.36	.876	.697	.540	.260	.129
12	3.06	2.68	2.18	1.78	1.36	.873	.695	.539	.259	.128
13	3.01	2.65	2.16	1.77	1.35	.870	.694	.538	.259	.128
14	2.98	2.62	2.14	1.76	1.34	.868	.692	.537	.258	.128
15	2.95	2.60	2.13	1.75	1.34	.866	.691	.536	.258	.128
16	2.92	2.58	2.12	1.75	1.34	.865	.690	.535	.258	.128
17	2.90	2.57	2.11	1.74	1.33	.863	.689	.534	.257	.128
18	2.88	2.55	2.10	1.73	1.33	.862	.688	.534	.257	.127
19	2.86	2.54	2.09	1.73	1.33	.861	.688	.533	.257	.127
20	2.84	2.53	2.09	1.72	1.32	.860	.687	.533	.257	.127
21	2.83	2.52	2.08	1.72	1.32	.859	.686	.532	.257	.127
22	2.82	2.51	2.07	1.72	1.32	.858	.686	.532	.256	.127
23	2.81	2.50	2.07	1.71	1.32	.858	.685	.532	.256	.127
24	2.80	2.49	2.06	1.71	1.32	.857	.685	.531	.256	.127
25	2.79	2.48	2.06	1.71	1.32	.856	.684	.531	.256	.127
26	2.78	2.48	2.06	.171	1.32	.856	.684	.531	.256	.127
27	2.77	2.47	2.05	1.70	1.31	.855	.684	.531	.256	.127
28	2.76	2.47	2.05	1.70	1.31	.855	.683	.530	.256	.127
29	2.76	2.46	2.04	1.70	1.31	.854	.683	.530	.256	.127
30	2.75	2.46	2.04	1.70	1.31	.854	.683	.530	.256	.127
40	2.70	2.42	2.02	1.68	1.30	.851	.681	.529	.255	.126
60	2.66	2.39	2.00	1.67	1.30	.848	.679	.527	.254	.126
120	2.62	2.36	1.98	1.66	1.29	.845	.677	.526	.254	.126
∞	2.58	2.33	1.96	1.645	1.28	.842	.674	.524	.253	.126

Fuente: R. A. Fisher y F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research* (5.ª edición), Tabla III, Oliver y Boyd Ltd., Edinburgh, con autorización de los autores y editores.

Apéndice IV

Valores percentiles (χ_p^2) para la distribución ji-cuadrado con ν grados de libertad (área en sombra = p)

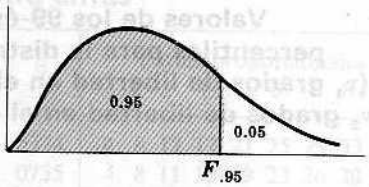


ν	$\chi^2_{.995}$	$\chi^2_{.99}$	$\chi^2_{.975}$	$\chi^2_{.95}$	$\chi^2_{.90}$	$\chi^2_{.75}$	$\chi^2_{.50}$	$\chi^2_{.25}$	$\chi^2_{.10}$	$\chi^2_{.05}$	$\chi^2_{.025}$	$\chi^2_{.01}$	$\chi^2_{.005}$
1	7.88	6.63	5.02	3.84	2.71	1.32	.455	.102	.0158	.0039	.0010	.0002	.0000
2	10.6	9.21	7.38	5.99	4.61	2.77	1.39	.575	.211	.103	.0506	.0201	.0100
3	12.8	11.3	9.35	7.81	6.25	4.11	2.37	1.21	.584	.352	.216	.115	.072
4	14.9	13.3	11.1	9.49	7.78	5.39	3.36	1.92	1.06	.711	.484	.297	.207
5	16.7	15.1	12.8	11.1	9.24	6.63	4.35	2.67	1.61	1.15	.831	.554	.412
6	18.5	16.8	14.4	12.6	10.6	7.84	5.35	3.45	2.20	1.64	1.24	.872	.676
7	20.3	18.5	16.0	14.1	12.0	9.04	6.35	4.25	2.83	2.17	1.69	1.24	.989
8	22.0	20.1	17.5	15.5	13.4	10.2	7.34	5.07	3.49	2.73	2.18	1.65	1.34
9	23.6	21.7	19.0	16.9	14.7	11.4	8.34	5.90	4.17	3.33	2.70	2.09	1.73
10	25.2	23.2	20.5	18.3	16.0	12.5	9.34	6.74	4.87	3.94	3.25	2.56	2.16
11	26.8	24.7	21.9	19.7	17.3	13.7	10.3	7.58	5.58	4.57	3.82	3.05	2.60
12	28.3	26.2	23.3	21.0	18.5	14.8	11.3	8.44	6.30	5.23	4.40	3.57	3.07
13	29.8	27.7	24.7	22.4	19.8	16.0	12.3	9.30	7.04	5.89	5.01	4.11	3.57
14	31.3	29.1	26.1	23.7	21.1	17.1	13.3	10.2	7.79	6.57	5.63	4.66	4.07
15	32.8	30.6	27.5	25.0	22.3	18.2	14.3	11.0	8.55	7.26	6.26	5.23	4.60
16	34.3	32.0	28.8	26.3	23.5	19.4	15.3	11.9	9.31	7.96	6.91	5.81	5.14
17	35.7	33.4	30.2	27.6	24.8	20.5	16.3	12.8	10.1	8.67	7.56	6.41	5.70
18	37.2	34.8	31.5	28.9	26.0	21.6	17.3	13.7	10.9	9.39	8.23	7.01	6.26
19	38.6	36.2	32.9	30.1	27.2	22.7	18.3	14.6	11.7	10.1	8.91	7.63	6.84
20	40.0	37.6	34.2	31.4	28.4	23.8	19.3	15.5	12.4	10.9	9.59	8.26	7.43
21	41.4	38.9	35.5	32.7	29.6	24.9	20.3	16.3	13.2	11.6	10.3	8.90	8.03
22	42.8	40.3	36.8	33.9	30.8	26.0	21.3	17.2	14.0	12.3	11.0	9.54	8.64
23	44.2	41.6	38.1	35.2	32.0	27.1	22.3	18.1	14.8	13.1	11.7	10.2	9.26
24	45.6	43.0	39.4	36.4	33.2	28.2	23.3	19.0	15.7	13.8	12.4	10.9	9.89
25	46.9	44.3	40.6	37.7	34.4	29.3	24.3	19.9	16.5	14.6	13.1	11.5	10.5
26	48.3	45.6	41.9	38.9	35.6	30.4	25.3	20.8	17.3	15.4	13.8	12.2	11.2
27	49.6	47.0	43.2	40.1	36.7	31.5	26.3	21.7	18.1	16.2	14.6	12.9	11.8
28	51.0	48.3	44.5	41.3	37.9	32.6	27.3	22.7	18.9	16.9	15.3	13.6	12.5
29	52.3	49.6	45.7	42.6	39.1	33.7	28.3	23.6	19.8	17.7	16.0	14.3	13.1
30	53.7	50.9	47.0	43.8	40.3	34.8	29.3	24.5	20.6	18.5	16.8	15.0	13.8
40	66.8	63.7	59.3	55.8	51.8	45.6	39.3	33.7	29.1	26.5	24.4	22.2	20.7
50	79.5	76.2	71.4	67.5	63.2	56.3	49.3	42.9	37.7	34.8	32.4	29.7	28.0
60	92.0	88.4	83.3	79.1	74.4	67.0	59.3	52.3	46.5	43.2	40.5	37.5	35.5
70	104.2	100.4	95.0	90.5	85.5	77.6	69.3	61.7	55.3	51.7	48.8	45.4	43.3
80	116.3	112.3	106.6	101.9	96.6	88.1	79.3	71.1	64.3	60.4	57.2	53.5	51.2
90	128.3	124.1	118.1	113.1	107.6	98.6	89.3	80.6	73.3	69.1	65.6	61.8	59.2
100	140.2	135.8	129.6	124.3	118.5	109.1	99.3	90.1	82.4	77.9	74.2	70.1	67.3

Fuente: Catherine M. Thompson, *Table of percentage points of the χ^2 distribution*, Biometrika, Vol. 32 (1941), con autorización del autor y del editor.

Apéndice V

Valores de los 95-ésimos percentiles para la distribución F (ν_1 grados de libertad en el numerador) (ν_2 grados de libertad en el denominador)

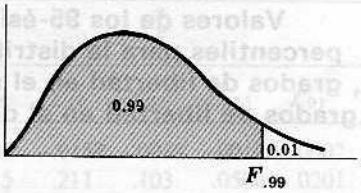


ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Fuente: E. S. Pearson y H. O. Hartley, *Biometrika Tables fos Statisticians*, Vol. 2 (1972), Tabla 5, página 178, reproducción autorizada.

Apéndice VI

Valores de los 99-ésimos percentiles para la distribución F (v_1 grados de libertad en el numerador) (v_2 grados de libertad en el denominador)



v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	5000	5403	5625	5764	5859	5928	5981	6023	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.82	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Fuente: E. S. Pearson y H. O. Hartley, *Biometrika Tables for Statisticians*, Vol. 2 (1972), Tabla 5, página 180, reproducción autorizada.

Apéndice VII

Logaritmos decimales con cuatro cifras

N	0	1	2	3	4	5	6	7	8	9	Partes proporcionales 1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	16	19	21
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
N	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

Logaritmos decimales con cuatro cifras (*continuación*)

N	0	1	2	3	4	5	6	7	8	9	Partes proporcionales 1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	3	4	5	6	7	8
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1	2	2	3	4	5	6	7	7
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	1	2	3	4	4	5	6	7
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	1	2	3	4	4	5	6	7
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	5	5	6
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	3	3	4	4	5	6
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1	1	2	2	3	4	4	5	6
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1	1	2	2	3	4	4	5	6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	5
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	5
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1	2	2	3	4	4	5	5
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	5
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	5
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
N	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

Logaritmos decimales con cuatro cifras (*Continuación*)

N	0	1	2	3	4	5	6	7	8	9	Partes proporcionales 1	2	3	4	5	6	7	8	9
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	0	1	1	2	2	3	3	4	4
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0	1	1	2	2	3	3	4	4
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0	1	1	2	2	3	3	4	4
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0	1	1	2	2	3	3	3	4
N	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

Apéndice VIII

Valores de $e^{-\lambda}$

$(0 < \lambda < 1)$

λ	0	1	2	3	4	5	6	7	8	9
0.0	1.0000	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9231	.9139
0.1	.9048	.8958	.8869	.8781	.8694	.8607	.8521	.8437	.8353	.8270
0.2	.8187	.8106	.8025	.7945	.7866	.7788	.7711	.7634	.7558	.7483
0.3	.7408	.7334	.7261	.7189	.7118	.7047	.6977	.6907	.6839	.6771
0.4	.6703	.6636	.6570	.6505	.6440	.6376	.6313	.6250	.6188	.6126
0.5	.6065	.6005	.5945	.5886	.5827	.5770	.5712	.5655	.5599	.5543
0.6	.5488	.5434	.5379	.5326	.5273	.5220	.5169	.5117	.5066	.5016
0.7	.4966	.4916	.4868	.4819	.4771	.4724	.4677	.4630	.4584	.4538
0.8	.4493	.4449	.4404	.4360	.4317	.4274	.4232	.4190	.4148	.4107
0.9	.4066	.4025	.3985	.3946	.3906	.3867	.3829	.3791	.3753	.3716

$(\lambda = 1, 2, 3, ..., 10)$

λ	1	2	3	4	5	6	7	8	9	10
$e^{-\lambda}$	.36788	.13534	.04979	.01832	.006738	.002479	.000912	.000335	.000123	.000045

Nota: Para obtener valores de $e^{-\lambda}$ para otros valores de λ, usar las leyes de la función exponencial.

Ejemplo: $e^{-3.48} = (e^{-3.00})(e^{-0.48}) = (0.04979)(0.6188) = 0.03081$.

Apéndice IX

Números aleatorios

51772	74640	42331	29044	46621	62898	93582	04186	19640	87056
24033	23491	83587	06568	21960	21387	76105	10863	97453	90581
45939	60173	52078	25424	11645	55870	56974	37428	93507	94271
30586	02133	75797	45406	31041	86707	12973	17169	88116	42187
03585	79353	81938	82322	96799	85659	36081	50884	14070	74950
64937	03355	95863	20790	65304	55189	00745	65253	11822	15804
15630	64759	51135	98527	62586	41889	25439	88036	24034	67283
09448	56301	57683	30277	94623	85418	68829	06652	41982	49159
21631	91157	77331	60710	52290	16835	48653	71590	16159	14676
91097	17480	29414	06829	87843	28195	27279	47152	35683	47280
50532	25496	95652	42457	73547	76552	50020	24819	52984	76168
07136	40876	79971	54195	25708	51817	36732	72484	94923	75936
27989	64728	10744	08396	56242	90985	28868	99431	50995	20507
85184	73949	36601	46253	00477	25234	09908	36574	72139	70185
54398	21154	97810	36764	32869	11785	55261	59009	38714	38723
65544	34371	09591	07839	58892	92843	72828	91341	84821	63886
08263	65952	85762	64236	39238	18776	84303	99247	46149	03229
39817	67906	48236	16057	81812	15815	63700	85915	19219	45943
62257	04077	79443	95203	02479	30763	92486	54083	23631	05825
53298	90276	62545	21944	16530	03878	07516	95715	02526	33537

Indice

Abscisa, 5
Agrupados, datos, 38
  métodos de compilación (*véase* Método de compilación)
Ajuste de curvas, a mano, 291, 444, 452
  ecuaciones especiales usadas en el, 290
  métodos de mínimos cuadrados, 289-321
Ajuste de datos, 163, 180-183 (*véase* Ajuste de curvas)
  por distribución binominal, 180
  por distribución de Poisson, 182, 183
  por distribución normal, 180, 182
  usando papel gráfico, 163, 180
Ajuste de datos a las variaciones estacionales, 446, 461
Aleatorio, 56, 414
Aleatorización completa, 386
Aleatorizados, bloques, 387
Análisis combinatorio, 134, 148-152
  probabilidad y, 148-152
Análisis de series en el tiempo, 440-477 (*véase* Series en el tiempo)
  pasos fundamentales en el, 447
Análisis de varianza, 375-410
  con réplicas, 383-387
  cuadrados greco-latinos, 387, 401-403
  cuadrados latinos, 387
  experimentos de dos factores usando, 380-387, 394-399
  experimentos de un factor usando, 375, 387-394
  F test, 379, 380, 383-387, 390-396
  modelo matemático, 377, 384
  propósito del, 375
  tablas, 379, 383, 385
Antilogaritmos, 8, 27-30 (*véase* Logaritmos)
Aproximación normal a la distribución binomial, 162, 175-178
Apuestas, 129
Areas de la distribución, ji-cuadrado, 254, 538
  F, 256, 539, 540
  normal, 160, 161, 169-172, 536
  t, 251, 252, 537
Aritmética, media, 61-65, 68-75
  calculada mediante datos agrupados, 62, 64, 72-75
  comprobación Charlier, 95, 106
  de distribuciones de probabilidad, 143
  de medias aritméticas, 62, 69, 70
  efecto de los valores extremos sobre la, 68, 70, 76
  intervalo de confianza para la, 209, 210, 214-216
  método de compilación, 63, 74, 75
  métodos largos y cortos para su cálculo, 63, 74
  poblacional y muestral, 133
  ponderada, 62, 68-70
  propiedades, 62, 63, 71
  relación con las medias geométricas y armónica, 65, 66, 80, 81
  relación con mediana y moda, 64, 65, 80
  supuesta o conjeturada, 63, 72
Armónica, media, 61, 65, 82, 83
  ponderada, 83
  relación con las medias aritmética y geométrica, 65, 81
Asimetría (sesgo), 42, 118-120, 125
  coeficiente cuartil de, 118, 125
  coeficiente de, 118
  coeficiente percentil 10-90 de, 118, 125
  coeficiente de Pearson, 118, 125
  negativa (a la izquierda), 42, 118
  para la distribución binomial, 160, 161
  para la distribución de Poisson, 162
  positiva (a la derecha), 42, 118
Asimétricas, curvas de frecuencia, 42
Asintóticamente normal, 188
Atributos, correlación de, 272, 284, 328
Autocorrelación, 328

Base, 3
  de logaritmos comunes, 7
  de logaritmos naturales, 36
Bayes, teorema o regla de, 158
Bernoulli, James, 160
Binomial, desarrollo o fórmula, 159, 165

Binomial, distribución, 159, 161, 163-170
ajuste de datos, 180
propiedades, 159, 161
relación con la distribución de Poisson, 162, 178
relación con la distribución normal, 161, 175-178
test (contraste) de hipótesis usando la, 228, 245-248
Binomiales, coeficientes, 159, 165
triángulo de Pascal, 165, 166
Bivariable:
distribución de frecuencia o tabla, 329, 342
distribución normal, 329
población, 329
Bloques, 370-387
aleatorizados, 386
Bondad del ajuste, 163 (*véase* Ajuste de datos)
test ji-cuadrado, 269, 278, 279

Canónica, curva normal (*véase* Normal, curva)
Cantidad (o volumen), números índice de, 485, 502, 503
Característica, 7, 27, 28
Categorías, 37
Centro de gravedad, 293
Centroide, 293
Cero, punto, 5
Cesta de la compra, 497
Charlier, comprobación de, 95, 106, 117, 124
para la media y la varianza, 95, 106
para momentos, 118, 124
Cíclicos, movimientos o variaciones, 441, 445, 446, 461-466
Ciclos financieros, 441
Clase, 37 (*véase* Intervalos de clase)
Clase modal, 45, 64
frecuencia, 45, 64
Claúsulas de revisión, 478
Cociente de inteligencia, 105, 106
Coeficiente de correlación de orden cero, 358
Coeficiente de correlación de rango, 416, 433-435
Coeficiente de curtosis, 119, 126 (*véase* Curtosis)
de asimetría, 119 (*véase* Asimetría)
Coeficiente de determinación múltiple, 359, 367, 368
Coeficiente percentil de curtosis, 119, 126
Coeficiente de confianza, 210, 252
Coeficiente de correlacción, 325-329, 338-350 (*véase* Correlación)
de tablas de contingencia, 272, 284
fórmula momento-producto, 327, 338-345
líneas de regresión y, 328, 344-347
para datos agrupados, 328, 341-345
series en el tiempo y, 328, 346
teoría del muestreo y, 329, 349, 350
Coincidencias, en el *H* test de Kruskal-Wallis, 413
en el *U* test de Mann-Whitney, 412
Combinaciones, 135, 146-149
Comparación de datos, 446, 466
Compuesto, interés, 82
Comunes (decimales), logaritmos, 7, 8, 27-30
tabla, 541, 542
Conjunto vacío, 136
Constantes, 1
dólares, 505, 506
Contingencia, coeficiente de, 272, 283, 284
Contingencia, tablas de, 270-272, 279-285
coeficiente de correlación, 272, 284
fórmulas para ji-cuadrado, 270, 272, 282, 283
Continua, variable, 1, 9
Contraste (test) de hipótesis y significación, 223-250, 253, 254, 257, 260, 263, 264
con la distribución binomial, 228, 245-248
con la distribución *F* (*véase* Distribución *F*)
con la distribución ji-cuadrado (*véase* Distribución ji-cuadrado)
con la distribución normal, 224, 226 228-234
con la distribución *t* (*véase* Distribución *t*)
en relación con correlación y regresión, 328-331
para diferencias de medias y proporciones, 227, 228, 241-245
para medias y proporciones, 225-234
Control de calidad, gráficos de, 227, 240, 241
Coordenadas rectangulares, 5, 15-19
Correlación, 322-356 (*véase* Regresión)
auto-, 329
coeficiente de (*véase* Coeficientes de correlación)
de atributos, 272, 284, 329
lineal, 322
medidas de, 323
múltiple (*véase* Correlación múltiple)
parcial, 357-374
positiva y negativa, 323
rango, 416, 433-435
simple, 322, 355
sin sentido o espúrea, 326
tetracórica, 272
Correlación múltiple, 357-374
coeficiente de, 359, 367, 368
Correlación parcial, 357-374
coeficientes de, 361, 368, 369
Correlación sin sentido, 326
Correlación, tabla de, 327, 344

Covarianza, 327, 339
coeficiente de correlación en términos de, 327
Crítica(os):
región, 225
valores, 210, 253
Cuadrantes, 5
Cuadrática, curva, 289
Cuadrática, ecuación, 35
fórmula para la solución, 35
Cuadrática, función, 17, 290
mínimo de la, 110
Cuadrática, media, o raíz de la media del cuadrado, 65, 83
relación con la media geométrica, 84
Cuantiles, 66
Cuártica, curva, 289
función, 290
Cuartil, coeficiente, de dispersión relativa, 110
de asimetría, 119, 125
Cuartil, desviación (*véase* Semi-intercuartil, rango)
Cuartiles, 66, 84-86
de datos agrupados, 66
errores típicos para, 191
Cúbica, curva, 289
función, 290
Curtosis (o aplastamiento), 119, 120, 126
coeficiente de, 119, 125
coeficiente percentil de, 119, 125
de la distribución binomial, 160
de la distribución de Poisson, 162
de la distribución normal, 119, 161
Curva de frecuencia bimodal, 42
Curva de frecuencias en forma de J, 42
invertida, 42
Curva de Gompertz, 290
Curva de potencia, 237 (*véase* Curvas de operación características)
Curvas de aproximación, ecuaciones de las, 289, 290
Curvas de frecuencia, 41, 42, 55, 56
relativa, 41
tipos de, 42
Curvas de frecuencia asimétricas, 42, 64
Curvas de operación características, 227, 235-240, 245, 247

Datos ajustados estacionalmente, 446, 461
Datos continuos, 1, 8, 9
representación gráfica, 54, 55
Datos, continuos (*véase* Datos continuos)
agrupados, 38
ajustados a las variaciones estacionales, 446, 461
comparación de, 446, 466
discretos (*véase* Discretos, datos)
dispersión de (*véase* Dispersión)
extensión o variación de, 70 (*véase* Dispersión; Variación)
fila, 37
redondeo de (*véase* Redondeo de datos)
Deciles, 66, 84-86
errores típicos, 191
para datos agrupados, 52, 83-86
Decisión, reglas de, 223 (*véase* Decisiones estadísticas)
Decisiones estadísticas, 223-225
hipótesis (*véase* Hipótesis)
inferencia, 1, 186, 208
Deductiva, estadística, 1
Deflación de series en el tiempo, 486, 505, 506
Descriptiva, estadística, 1
Desigualdad, símbolos de, 6
Desigualdades, 6, 7
Desviación de la media aritmética, 61, 71
cuartil (*véase* Semi-intercuartil, rango)
curva de mínimos cuadrados, 291
media (*véase* Media, desviación)
típica (*véase* Desviación típica)
Desviación típica, 91-96, 100-112
corregida (*véase* Sheppard, corrección de)
de datos agrupados, 92, 94, 101-108
de distribuciones de muestreo, 187-192 (*véase* Errores típicos)
de una distribución de probabilidad, 143
intervalo de confianza, 212, 219, 221
método de compilación, 94, 105, 106
métodos breves para su cálculo, 94, 101-106
propiedad de mínimo, 94, 110
propiedades de, 94, 108-110
relación con la desviación media y el rango semi-intercuartil, 96, 108
relación de población y muestreo, 92
Determinación, coeficientes de, 325, 338
múltiple, 359, 367, 368
Diagrama de dispersión, 289, 331-335
tridimensional, 358
Diagrama de Euler, 135, 152-155
Diagramas (*véase* Gráfico)
Dicotomía, clasificación por, 268
10-90, rango percentil, 92, 99
Discreta, variable, 1, 8, 9
Discretas, distribuciones de probabilidad, 132
Discretos, datos, 1
representación gráfica de, 54

Diseño de experimentos, 186, 386
Dispersión, 66 (*véase* Variación)
absoluta, 96, 110, 111
coeficiente de, 96, 110, 111
medidas de, 91-115
relativa, 96, 110, 111
Dispersión absoluta, 96, 110, 111 (*véase* Dispersión)
Distribución de Bernoulli (*véase* Binomial, distribución)
Distribución de Poisson, 162, 178-180
ajuste de datos con la, 182, 183
propiedades, 162
relación con las distribuciones binomial y normal, 162
Distribución de probabilidad acumulada, 132
Distribución *F*, 255, 256 (*véase* Análisis de varianza)
Distribución gaussiana (*véase* Distribución normal)
Distribución ji-cuadrado, 254, 261-264 (*véase* Ji-cuadrado)
contraste de hipótesis y significación, mediante la, 268-288
intervalos de confianza usando, 254, 255, 262, 263
tabla de percentiles para la, 538
Distribución modelo o teórica, 163
Distribución normal, 95, 108, 109, 159-162, 169-178 (*véase* Normal, curva)
ajuste de datos con la, 181, 182
contraste de hipótesis y significación usando la, 224-241
forma canónica, 160
proporciones de, 160, 162
relación con la binomial, 162, 175-178
relación con la de Poisson, 162
Distribución *t*, 251, 256-260
contraste de hipótesis y significación, usando la, 256-260
en teoría muestral de correlación y regresión, 328-332, 349, 351
intervalos de confianza usando la, 252, 253, 257
tabla de valores percentiles, 537
Distribuciones continuas de probabilidad, 133, 141, 143
Distribuciones de frecuencias (*véase* Frecuencias, distribuciones de)
muestreo (*véase* Muestreo, distribuciones de)
probabilidad, 56, 132, 133, 141-143
unimodal, 64
Dominio de una variable, 1, 9

Ecuaciones, 5, 25, 26
cuadráticas, 35
de curvas aproximantes, 289, 290
de regresión, 357-366
equivalentes, 6, 26
miembros izquierdo y derecho de, 5
normales (*véase* Normales, ecuaciones)
simultáneas, 6, 25, 26
solución de, 6
transposición en, 25
Ecuaciones simultáneas, 5, 25, 26
Edad cronológica, 105
Edad mental, 105
Eficientes, estimadores y estimaciones, 209, 213
Ejes *X*, *Y* del sistema de coordenadas rectangulares, 5
Elástica, demanda, 497
Eliminación de incógnitas (*véase* Ecuaciones simultáneas)
Empírica, relación, entre media, mediana y moda, 64, 80
entre medidas de dispersión, 96, 108-110
Entrada simple, tabla de, 44
Enumeración, 2, 3
Error de agrupamiento, 39, 50
Error típico de estimación, 324, 334-337, 359, 366-367
modificado, 325
Errores, de agrupamiento, 39, 50
de tipos I y II (*véase* Errores de tipo I y II)
probables, 212, 220
redondeo de, 2, 9
típicos (*véase* Errores típicos)
Errores aleatorios, 377, 382, 384
Errores de redondeo acumulados, 2, 9
Errores de tipo I y II, 224, 230-232, 235, 240, 243, 245
curva de operaciones característica y, 227, 235-240, 246
Errores típicos, de distribuciones de muestreo, 189-191
tabla, para diversos estadísticos, 191
ESP (*véase* Percepción extrasensorial)
Espacio 4-dimensional, 360
Esperanza matemática, 133, 143, 144
Espúrea, correlación, 326
Esquema cíclico, en el test de peldaños (o rachas), 415, 429
Estacionales, variaciones, 442, 444, 446
Estadística, 1, 186, 208
deductiva o descriptiva, 1
definición, 1
inductiva, 1

muestral, 186, 208
Estadístico *t*, 252
Estadístico *H* (*H* test), 374
Estimación, 186, 208-222, 294 (*véase* Estimaciones)
de la tendencia, 444, 451, 452
de variaciones cíclicas, 446, 461-466
de variaciones estacionales, 442, 444, 446, 452-461
de variaciones irregulares, 446, 461-466
y regresión (*véase* Regresión)
y teoría del muestreo, 208-222
Estimación óptima, 210
Estimaciones de intervalo, 209
Estimaciones puntuales, 210
Estimaciones sesgadas, 208, 212
Estimaciones, sesgadas y sin sesgo, 208, 210, 212, 213 (*véase* Estimación)
eficiente e ineficiente, 209, 210, 212, 213
intervalo de confianza, 209-210 (*véase* Intervalos de confianza)
punto e intervalo, 209
Estimadores (*véase* Estimaciones)
Estocástica, variable (*véase* Variable aleatoria)
Excluyentes, sucesos mutuamente, 131
Exito, 129, 159
Experimental, diseño, 151, 386
Experimentos de dos factores, 380-386, 394-399
Experimentos de factor único, 375, 387-394
Explicada, variación, 325, 337, 338, 348
Exponencial, curva, 290
Exponenciales, tabla de, 544
Exponente, 3
Extrapolación, 296

F test (*véase* Análisis de varianza)
Factorial, 134
fórmula de Stirling, 135
Fiabilidad, 209
Fila de datos, 37
Fila, medias de, 375
Fracaso, 129, 159
Frecuencia acumulada, 40, 41
distribución o tabla, 40, 51-55
polígono, 41, 52 (*véase* Ogiva)
Frecuencia de clase, 37, 39
acumulada, 40, 41
modal, 45
relativa, 40, 41
Frecuencia, distribución de, 37, 59
acumulada, 40, 41, 51-55
de porcentajes o relativa, 40, 49
reglas para formar, 39
Frecuencias de celda, 270, 342
Frecuencias esperadas o teóricas, 268
Frecuencias, histograma de (*véase* Histogramas)
relativas, 40, 49
Fronteras de clase, inferior y superior, 38, 39
Función, 4, 13-15
cuadrática (*véase* Cuadrática, función)
de distribución, 132
de frecuencias, 132
de probabilidad, 132
lineal, 17, 290
multivaluada, 4, 14
univaluada, 4, 14
Función densidad, probabilidad, 132
Función de distribución, 132
Función de frecuencia, 132
Función de operación característica, 237
Función de potencia, 237
Funciones univaluadas, 4, 14

Geométrica, curva, 290
Geométrica, media, 61, 65, 80-82
de datos agrupados, 65, 80
ponderada, 80
relación con las medias aritmética y armónica, 65
Gossett, 252
Grado *n*, curva de, 290
Grados de libertad, 252, 254, 255
Gráfico, 5, 15-24
circular, 5, 23, 24
de barras (*véase* Gráfico de barras)
de varillas, 54
lineal, 18, 20
Gráfico circular, 5, 23, 24
Gráfico de barras, 5, 20, 22, 23
de componentes, 19, 22
Gráficos de control, 227, 240, 241
grupo, 243, 260
Greco-latinos, cuadrados, 386, 401

H test de Kruskal-Wallis, 413, 427, 428
Hipérbola, 290
Hiperplano, 360
Hipótesis alternativa, 223
contraste (test) de, 186, 224 (*véase* Contraste de hipótesis y significación)
nula, 223

probabilidades de, usando la regla de Bayes, 158
Histogramas, 39, 44-51
cálculo de medianas mediante, 64, 77, 78
frecuencia relativa o de porcentajes, 40, 49, 50
probabilidad, 141
Hoja de recuentos, 39, 48

Identidad, 6
propiedad de relaciones de precios, 479
Independiente, variable, 4, 14, 15
Independientes, sucesos, 130
Indice del coste de vida, 478
Indice de precios al consumo (IPC), 346, 478, 505
Indice estacional, 445, 446, 452-460
Indice ideal de Fisher, 484, 498, 499
transformación Z, 329, 350
Indices (*véase* Números índice)
Indices, notación de, 482
Inductiva, Estadística, 1
Ineficientes, estimadores y estimaciones, 209, 212, 213
Ingresos reales o salarios, 486, 505
Interacción, 384
Interés compuesto, fórmula del, 82
Interpolación, 7, 28, 296
en logaritmos y antilogaritmos, 7, 28
Intersección de conjuntos, 136
Intersecciones, X e Y, 291, 296, 298
Intervalos de clase, 38, 39
abierto, 38
anchura o tamaño, 39
desiguales, 50
mediana, 63, 76, 77
modal, 45
Intervalos de confianza:
en correlación y regresión, 307-331, 349-352
para desviaciones típicas, 212, 219, 220
para medias, 210, 214-216
para proporciones, 211, 216-218
para sumas y diferencias, 211, 212, 218, 219
usando la distribución ji-cuadrado, 254, 255, 263
usando la distribución normal, 209-212, 214-220
usando la distribución t, 252, 253, 257, 258
Inversión temporal, propiedad de, 479
IQ (*véase* Cociente de inteligencia)

Ji-cuadrado, propiedad aditiva de, 272
análisis de varianza usando, 378, 379, 382
corrección de Yates, 271, 280, 283
definición, 268, 269
distribución (*véase* Distribución ji-cuadrado)
fórmulas para, en tablas de contingencia, 271
para bondad del ajuste, 269
test, 163, 268-288
J invertida, distribución en forma de, 42

Laspeyres, índice de, 484, 495-498
Latinos, cuadrados, 386, 399, 401
ortogonales, 387
Leptocúrtica, 119
Límites de clase, 38
inferior y superior, 38
verdaderos, 38
Límites de confianza, 210
Límites fiduciales (*véase* Límites de confianza)
Lineal, correlación (*véase* Correlación)
Lineal, función, 17, 290
Lineal, gráfico, 18, 19
Logaritmos, 27-30
base de, 7, 27, 36
cálculos con, 8
característica, 27
decimales (comunes), 27
tabla, 541-543
interpolación en, 7, 28
mantisa, 7, 8, 28
naturales, 36
Logística, curva, 290
Log-log, papel gráfico, 290, 316
Longitud de clase, anchura o tamaño, 38

Mantisa, 7, 8, 28
Marca de clase, 38, 39
Marginales, frecuencias, 270, 342
Marshall-Edgeworth, índice de, 484, 499, 500
Media aritmética (*véase* Aritmética, media)
armónica (*véase* Armónica, media)
cuadrática (*véase* Cuadrática, media)
geométrica (*véase* Geométrica, media)
Media cuadrática, 66, 83
desviación, 92
Media del grupo, 376
Media, desviación, 91, 92, 97-100
de la distribución normal, 161
para datos agrupados, 91, 98
Media final, 376, 377, 381
Mediana, 63, 64, 75-78

cálculo por histogramas, 64, 77, 78
efecto de los valores extremos, 76
para datos agrupados, 62, 76, 77
relación con la media aritmética y la moda, 64, 66, 80
Medidas, 2
Medidas de tendencia central, 60-90
«Menor que», distribución acumulada, 51, 52
Mesocúrtica, 119
Método «a mano» de ajuste de curvas, 291, 444, 452
Método de agregación, simple, 483, 492, 494
ponderada, 484, 495-498
Método de compilación, para coeficientes de correlación, 327, 342
para el momento, 117, 122, 124
para la desviación típica, 93, 105, 106
para la media, 62, 74
Método de promedio de relativos, simple, 483, 494, 495
ponderado, 484, 501
Método de promedio ponderado de relaciones, 485, 501
Método de recuento en el *U* test de Mann-Whitney, 414, 423
Método de relación a la tendencia, 445, 455, 456
Método de relación al promedio móvil, 445, 457
Método del año base, 484
Método del año prefijado, 484
Método del año típico, 484
Mínimos cuadrados:
curva, 292
parábola, 293, 294, 316-319
plano, 295
recta, 292, 293, 302-309
Mínimos cuadrados, método de, 291 (*Véase* Ajuste de curvas)
Moda, fórmula para la, 78, 80
para datos agrupados, 64, 78, 80
relación con la media aritmética y la mediana, 64, 65, 80
Momento-producto, fórmula para el coeficiente de correlación, 327, 338-340
Momentos, 116-128
adimensionales, 118
comprobación Charlier para su cálculo, 117, 124
correcciones de Sheppard para, 117, 126
definición, 116
método de compilación para su cálculo, 117, 122, 124
para datos agrupados, 116, 117, 122
relaciones entre, 117
Momentos adimensionales, 118
Movimientos a largo plazo, 441
Movimientos característicos de serie en el tiempo, 441, 447-451
clasificación, 441, 442
MQ (*véase* Media cuadrática)
Muestra, 1, 55, 186
aleatoria, 56, 186
Muestra aleatoria, 56, 186, 195-197
Muestral, espacio, 135, 136, 152-155
Muestral, estadística, 186, 208
Muestreo, con sustitución, 186
sin sustitución, 186
Muestreo, distribuciones de, 187-204
de diversos estadísticos, 189
de medias, 187, 190, 191
de proporciones, 188, 190, 197-200
de sumas y diferencias, 188, 200-203
de varianzas, 191
experimental, 193
Muestreo, números de, 195, 196
Muestreo, teoría del, 186-207, 251-267
de correlación, 329, 349, 350
de regresión, 330, 351
grandes muestras, 189
pequeñas muestras, 189, 251-267
uso en contraste de hipótesis y significación, 223-250
uso en la estimación, 208-222
Multimodal, curva de frecuencia, 43
Multinomial, desarrollo, 163
Multinomial, distribución, 163, 179, 271

Naturales, base de logaritmos, 36
Negativa, correlación, 323
asimetría, 42, 118
Nivel de significación, 224
descriptivo o experimental, 232
Niveles de confianza, tabla de, 210
No aleatorio, 415
No lineal(es):
correlación y regresión, 323, 326, 347-349, 361
ecuaciones reducibles a forma lineal, 293, 315-316
regresión múltiple, 361
relaciones entre las variables, 289, 293
Normal, curva, 42, 160 (*véase* Distribución normal)

área bajo la, 160, 169-175, 536
forma canónica (o estándar), 160
ordenadas de las, 161, 172, 535
papel gráfico, 163, 180
Normales, ecuaciones, de la recta de mínimos cuadrados, 292, 293, 302-309
de la parábola de mínimos cuadrados, 293, 294, 316-319
de plano de mínimos cuadrados, 295, 358
Normas de cálculo, 4, 10-13
usando logaritmos, 7, 29, 30
Notación científica, 2, 10
Nula, hipótesis, 223, 377, 379
Números aleatorios, 186, 195, 196
tabla de, 545
uso de, 195, 196
Números índice, 478-510
aplicaciones, 478
cíclicos, 446
definición, 478
de cantidad o volumen, 485, 502
de precios, 478, 482-484, 492-501
estacionales, 445, 452-461
problemas en su cálculo, 481, 482
test teóricos para, 482
valor, 486, 502, 503
Números índice cíclicos, 446

«O más», distribución acumulada, 41, 51-53
Observadas, frecuencias, 268
OC curvas (*véase* Curvas de operación características)
Ogivas, 40-42, 51-57
de porcentajes, 41, 51, 52
deciles, percentiles y cuartiles obtenidos de las, 84-86
mediana obtenida de las, 77, 78
«menor que», 41, 51-53
«o más», 41, 53
suavizadas, 42, 55, 56
Ogivas de porcentajes, 41, 51, 52
suavizadas, 55, 56
Ordenaciones, 37, 43, 44
Ordenadas, 5
de la curva normal, 161, 172
Origen, del sistema rectangular de coordenadas, 5
de series en el tiempo, 311
Ortogonal, cuadrado latino, 386

Paasche, índice de, 484-486, 495-501, 504
Papel gráfico, log-log, 290, 316
probabilidad, 163, 180
semilog, 290
Parábola, 17, 290
de mínimos cuadrados, 293, 294, 316-319
Parámetros, estimación de, 208, 209 (*véase* Estimación)
de población, 208, 209
Pascal, triángulo de, 165, 166
Pearson, coeficiente de asimetría de, 118, 125
Pendiente de una recta, 291, 298
Percentil, rango (10-90), 92, 99
Percentiles, 66, 84-86
de datos agrupados, 66, 84-86
de la distribución *F*, 256, 539, 540
de la distribución ji-cuadrado, 255, 258
de la distribución *t*, 252, 256, 257, 537
Percepción extrasensorial, 232
Período base de números índice, 478
cambio de, 486, 503, 504
Permutaciones, 134, 135, 144-146
circulares, 146
Pictogramas, 5, 18
Plano de mínimos cuadrados, 295
XY-, 5
Platicúrtica, 119
Población, 1, 186
finita o infinita, 187
parámetros de, 208, 209
Poder adquisitivo, 486, 505, 506
Polígonos de frecuencia, 39, 44-51
de porcentajes o relativa, 40, 49
suavizadas, 41, 55, 56
Polinomios, 290
Ponderación, factores de, 62
Ponderada(o):
media aritmética, 61, 68-71
media armónica, 82
media geométrica, 80
promedio móvil, 444
Ponderada, método de agregación, 484, 495-498
Porcentaje(s):
distribución acumulada, 40, 52
distribución, 40
frecuencia acumulada, 40, 52
gráfico de componentes, 19
histograma, 39
método de la tendencia, 445, 455, 456
Positiva, correlación, 322
asimetría, 42, 118
Potencia de un test, 227
Precios, índice de (*véase* Números índice)

Precios, relaciones de, 478, 479, 486-490
notación, 478
propiedades, 479
Predicción meteorológica, 294, 446, 466-468
Probabilidad, 56, 129-158
análisis combinatorio y, 134, 148-152
axiomática, 130
condicional, 130
curvas, 56
definición clásica, 129
definición como frecuencia relativa, 129
distribuciones, 56, 132
empírica, 130
papel gráfico, 163, 180
reglas fundamentales, 136-140
relación con la teoría de conjuntos, 135, 136
Probabilidad condicional, 130
Probabilidad, distribuciones de, 132, 141-142
acumuladas, 132
continuas, 130
discreta, 132
Probabilidad empírica, 130
Probabilidad, función de, 132
Probable, error, 212, 220
Progresión aritmética, momentos de una, 127
varianza de, 114
Promedio, 60, 68 (*véase* Semi-promedios)
desviación (*véase* Desviación media)
método de porcentajes, 445, 452-454
móvil, 443, 444 (*véase* Móvil, promedio)
Promedio móvil centrado, 445, 449, 450
Promedios móviles, 443, 444, 447-451
centrados, 445, 447-451
método de porcentajes, 445, 457
ponderados, 444
Propiedad cíclica de las relaciones de precios, 479
Propiedad circular de las relaciones de precios, 479
Propiedad del factor inverso, 480, 502, 503
Proporciones, 188, 189, 191, 197-200, 211, 212, 216-218, 226-233
contraste de hipótesis, 226-233
distribución muestral, 187, 188
intervalo de confianza para, 211, 212, 216-218

Quintiles, 85

Rachas (o peldaños), aplicaciones de, 415, 416, 432
definición, 414
test del carácter aleatorio, 414, 415, 429-432
Rango, 38
10-90 percentil, 92, 99
intercuartil, 92
semi-intercuartil, 92, 99, 119
Rango intercuartil, 92
semi-, 92, 99, 119
Recta, 289-293, 295-303
ecuación, 289, 290, 295, 297
de mínimos cuadrados, 291, 292
de regresión, 295
pendiente, 291, 297
Recuentos o enumeraciones, 2, 3
Redondeo de datos, 2, 9
Redondeo, errores de, 2, 9
Región de aceptación, 225, 230
Regresión, 294, 322, 323, 328, 357-366
curva de, 294
múltiple, 322, 357
plano de, 294, 357, 358
recta de, 294, 303, 307, 308, 323, 324 (*véase* Recta de mínimos cuadrados)
simple, 322
superficie de, 295
teoría de muestreo, 330
Relaciones de cantidad, 480, 490, 491
Relaciones de enlace, 445, 457-460, 481, 492
Relaciones de precios (*véase* Precios, relaciones de)
Relativa, dispersión o variación, 96, 110, 111
Relativa, frecuencia, 40, 49
curvas, 41
definición de probabilidad, 129
distribución, 39
tabla, 39
Réplicas, 375, 382
Residual, 291
variación, 382
Riesgo, 198

Secular, tendencia o variación, 441
Semi-intercuartil, rango, 92, 99, 119
Semilog, papel gráfico, 290
Semimedianas, 451
Semipromedios, método de los, 444, 451, 452
Serie de índices, 478-487 (*véase* Números índice)
Series en el tiempo, 440-477
ajuste de curvas para, 294, 309-315
análisis de, 440-477
correlación de, 328, 346
deflación, 486, 505, 506
gráficos, 19

movimientos característicos, 440, 447-451
predicción meteorológica, 294, 446, 466-468
suavización, 443
Sheppard, corrección de, para los momentos, 117, 124, 126
para la varianza, 95, 106-108
Significación, nivel de (*véase* Nivel de significación)
test de (*véase* Contraste de hipótesis y significación)
Significativos, dígitos o cifras, 3, 10
Signos, test de los, 411, 412, 416-419
Simétrica o en forma de campana, curva, 42
Simple, correlación, 322-356
Simple, método de agregación, 483, 492, 493
Simple, método de promedio de relaciones, 483, 494, 495
Solución de ecuaciones, 6
Spearman, fórmula de, para correlación de rango, 416, 433-435
Standard, recuentos, 96, 111, 112, 169
unidades, 96, 169, 172-176
Stirling, aproximación de $n!$, 135
«Student», distribución de (*véase* Distribución t)
Suavización de series en el tiempo, 443
Subíndices, notación de, 60, 357, 362
Suceso compuesto, 130
Sucesos, 129, 130
compuestos, 130
dependientes, 130
independientes, 130
mutuamente excluyentes, 131
probabilidad de (*véase* Probabilidad)
Sucesos dependientes, 130
Suma, notación de, 60, 66, 67
Superficie de regresión, 295
Superíndices, notación de, 482

Tabla de entrada simple, 43
Tabla de frecuencias (*véase* Distribuciones de frecuencias)
acumuladas, 41, 52, 53
relativas, 40
Tabla de partes proporcionales para logaritmos, 541-543
Tendencia, curva o línea de, 294, 310, 441-443
Tendencia, esquema de la, 415, 430
Tendencia, estimación de la, 446, 451, 452
secular, 441
Tendencia, valores de, 311, 451-453
Teorema del límite central, 187
Teoría de pequeñas muestras, 190, 251-267
Teóricas, frecuencias, 268
Test (contrastes), bilateral o de dos colas, 225
Test estadístico, 225
Test unilateral o de una cola, 226
Tests no paramétricos, 411-439
H test de Kruskal-Wallis, 413, 427
para la correlación, 416, 433-435
test de los signos, 411, 412, 416-419
test de rachas (o peldaños), 414-416, 429-433
U test de Mann-Whitney, 412, 416, 420-427
Tetracórica, correlación, 272
Totales móviles, 443, 447-449
Transposición, en ecuaciones, 25
en desigualdades, 27
Tratamiento, 376, 380
medias, 376

U test de Mann-Whitney, 412, 416, 420-427

Valor absoluto, 91
Variable, 1, 2, 4, 8, 9
aleatoria (*véase* Variable aleatoria)
continua, 1, 9
dependiente, 4, 14, 15
discreta, 1, 8, 9
distribuida normalmente, 161
dominio, 1, 9
estocástica (*véase* Variable aleatoria)
independiente, 4, 14, 15
tipificada, 96, 111
Variable aleatoria, 132, 136, 141, 142
continua, 132, 142
discreta, 132, 141
Variable independiente, 4, 14, 15
cambio de, en la ecuación de regresión, 360
Variable tipificada, 96, 111
Variables, relación entre, 289, 357 (*véase* Ajuste de curvas; Correlación; Regresión)
Variación, 70 (*véase* Dispersión)
aleatoria, 442-446
cíclica, 441, 446, 461-466
coeficiente cuartil de, 111
coeficiente de, 96, 110
estacional, 442, 444, 445
explicada e inexplicada, 325, 337, 338, 348
residual, 382
secular, 441

Variación total, 325, 337, 348, 368, 376, 381, 384
Variaciones aleatorias, 442, 446
Variaciones irregulares, 422, 443, 446 461-466
Varianza, 93 (*véase* Desviación típica)
  análisis de (*véase* Análisis de varianza)
  combinada, 95
  comprobación Charlier, 95, 106
  corrección de Sheppard, 95, 106-108
  de distribuciones de probabilidad, 143
  de distribuciones muestrales, 187-203
  muestral modificada, 208, 212
  relación entre población y muestra, 133
Varillas, gráficos de, 54
Venn, diagramas de (*véase* Diagramas de Euler)
Volumen, números índice de, 485, 502
  relaciones de, 480

X-intersección, 291, 296, 298
XY-plano, 5

Y-intersección, 291, 297
Yates, corrección de, para la continuidad, 270-276
  en tablas de contingencia, 270, 271, 280, 283

[redacted]

Schaum

3 000102 528744

Los textos de la serie Schaum se han convertido en clásicos, por estar a la vanguardia en el estudio, y por ser una inestimable ayuda para el alumno a la hora de adquirir un conocimiento y pericia completos en la materia que se aborda.

Cada capítulo está estructurado de la siguiente manera:

- **Teoría**: resumen de las definiciones, principios y teoremas pertinentes, que sirve al estudiante como repaso.
- **Problemas resueltos**: completamente desarrollados, y en grado creciente de dificultad.
- **Problemas propuestos**: con la solución indicada; y que permiten al estudiante afianzar los conocimientos adquiridos.

9 788476 155622

ISBN: 84-7615-5