CAPÍTULO DIEZ

Pruebas de bondad de ajuste y análisis de tablas de contingencia

10.1 Introducción

Recuérdese que una hipótesis estadística es una afirmación con respecto a una característica que se desconoce de una población de interés. En el capítulo 9 fue, en forma exclusiva, el valor de algún parámetro θ. En este capítulo se examinarán las pruebas de hipótesis estadísticas en las que la característica que se desconoce es alguna propiedad de la forma funcional de la distribución que se muestrea. Además, se discutirán las pruebas de independiencia entre dos variables aleatorias en las cuales la evidencia muestral se obtiene mediante la clasificación de cada variable aleatoria en un cierto número de categorías.

En forma tradicional, este tipo de prueba recibe el nombre de *bondad del ajuste* ya que ésta compara los resultados de una muestra aleatoria con aquéllos que se espera observar si la hipótesis nula es correcta. La comparación se hace mediante la clasificación de los datos que se observan en cierto número de categorías y entonces comparando las frecuencias observadas con las esperadas para cada categoría. Para un tamaño específico del error de tipo I, la hipótesis nula será rechazada si existe una diferencia suficiente entre las frecuencias observadas y las esperadas.

Vale la pena notar que para situaciones de este tipo la hipótesis alternativa es compuesta y, en muchas ocasiones, no se encuentra identificada en forma explícita. El resultado es que la función de potencia es muy difícil de obtener en forma analítica. En consecuencia, una prueba de bondad de ajuste no debe usarse por sí misma para aceptar la afirmación de la hipótesis nula. La decisión es no rechazar H_0 (más que aceptarla) si la diferencia que existe entre las frecuencias observadas y esperadas es, en forma relativa, pequeña.

10.2 La prueba de bondad de ajuste chi-cuadrada

Una prueba de bondad de ajuste se emplea para decidir cuándo un conjunto de datos se apega a una distribución de probabilidad dada. Considérese una muestra aleatoria de tamaño n de la distribución de una variable aleatoria X dividida en k clases exhaustivas y mutuamente excluyentes, y sea N_i, $i = 1, 2, \ldots, k$, el número de observaciones en la i-ésima clase. Considérese la verificación de la hipótesis nula

$$H_0: F(x) = F_0(x), \tag{10.1}$$

en donde el modelo de probabilidad propuesto $F_0(x)$ se encuentra especificado, de manera completa, con respecto a todos los parámetros. De esta forma la hipótesis nula es sencilla. Dado que se especifica $F_0(x)$ de manera completa, se puede obtener la probabilidad p_i de obtener una observación en la i-ésima clase bajo H_0, en donde necesariamente $\sum_{i=1}^{k} p_i = 1$.

Sea n_i la realización de N_i para $i = 1, 2 \ldots k$ de manera tal que $\sum_{i=1}^{k} n_i = n$. La probabilidad de tener, de manera exacta, n_i observaciones en la i-ésima clase es $p_i^{n_i}$ para $i = 1, 2 \ldots k$. Dado que existen k categorías mutuamente excluyentes con probabilidades $p_1, p_2, \ldots, p_k$, entonces bajo la hipótesis nula la probabilidad de la muestra agrupada es igual a la función de probabilidad de una distribución multinomial determinada (6.3).

Para deducir una prueba estadística adecuada para H_0, considérese el caso en el que $k = 2$. Este es la distribución binomial con una función de probabilidad dada por (4.1) y en la que $x = n_1$, $p = p_1$, $n - x = n_2$, y $1 - p = p_2$. Considérese la variable aleatoria estandarizada

$$Y = \frac{N_1 - np_1}{\sqrt{np_1(1 - p_1)}}.$$

Del capítulo 5, recuérdese que para un valor de n suficientemente grande, la distribución de Y es aproximadamente igual a la normal estándar. Además, del ejemplo 5.14 se sabe que el cuadrado de una variable aleatoria normal estándar tiene una distribución chi-cuadrada con un grado de libertad. Entonces, la estadística

$$\begin{aligned}
\frac{(N_1 - np_1)^2}{np_1(1-p_1)} &= \frac{(N_1 - np_1)^2}{np_1} + \frac{(N_1 - np_1)^2}{np_2} \\
&= \frac{(N_1 - np_1)^2}{np_1} + \frac{[n - N_2 - n(1 - p_2)]^2}{np_2} \\
&= \frac{(N_1 - np_1)^2}{np_1} + \frac{(N_2 - np_2)^2}{np_2} \\
&= \sum_{i=1}^{2} \frac{(N_i - np_i)^2}{np_i}
\end{aligned}$$

tiene aproximadamente una distribución chi-cuadrada con un grado de libertad conforme n va tomando valores cada vez más grandes.

Si se sigue este tipo de razonamiento, puede demostrarse que para $k \geq 2$ categorías distintas, la estadística

$$\sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \tag{10.2}$$

tiene una distribución, en forma aproximada, chi-cuadrada con $k - 1$ grados de libertad, si n tiene un valor suficientemente grande. Nótese que N_i es la frecuencia observada en la i-ésima clase, y np_i es la frecuencia correspondiente que se esperaba bajo la hipótesis nula. De acuerdo con lo anterior, la estadística es la suma sobre todas las k clases de los cocientes de los cuadrados de las diferencias entre las frecuencias observada y esperada, y la frecuencia esperada. La estadística dada por (10.2) recibe el nombre de *prueba de bondad de ajuste chi-cuadrada* de Pearson. Si existe una concordancia perfecta entre las frecuencias que se observaban y las que se esperaban, la estadística tendrá un valor igual a cero; por otro lado, si existe gran discrepancia entre estas frecuencias, la estadística tomará un valor muy grande. Por ello se desprende que para un tamaño dado del error de tipo I, la región crítica es el extremo superior de una distribución chi-cuadrada con $k - 1$ grados de libertad.

Ejemplo 10.1 El gerente de una planta industrial pretende determinar si el número de empleados que asisten al consultorio médico de la planta se encuentra distribuido, en forma equitativa, durante los cinco días de trabajo de la semana. Con base en una muestra aleatoria de cuatro semanas completas de trabajo, se observó el siguiente número de consultas:

Lunes	*Martes*	*Miércoles*	*Jueves*	*Viernes*
49	35	32	39	45

Con $\alpha = 0.05$, ¿existe alguna razón para creer que el número de empleados que asisten al consultorio médico, no se encuentra distribuido en forma equitativa durante los días de trabajo de la semana?

Una distribución uniforme implicaría que las proporciones para cada día de la semana sean iguales. Por lo tanto, deberá probarse la hipótesis nula

$$H_0: p_i = 0.2, \qquad i = 1, 2, \ldots, 5.$$

Dado que el tamaño de la muestra es $n = 200$, la frecuencia esperada para cada día es $np_i = 40$. Entonces, el valor de la estadística de prueba es

$$\chi^2 = \frac{(49 - 40)^2}{40} + \frac{(35 - 40)^2}{40} + \frac{(32 - 40)^2}{40} + \frac{(39 - 40)^2}{40} + \frac{(45 - 40)^2}{40} = 4.9.$$

Para $k = 5$ clases, se observa que el valor crítico es $\chi^2_{0.95,\,4} = 9.49$. Ya que $\chi^2 = 4.9 < \chi^2_{0.95,\,4} = 9.49$, no puede rechazarse la hipótesis nula. Con base en esta evidencia, no existe ninguna razón para creer que el número de empleados que acuden al

consultorio no se encuentre distribuido en forma uniforme a lo largo de la semana de trabajo.

Una ventaja de la prueba de bondad de ajuste chi-cuadrada es que para valores grandes de n, la distribución límite chi-cuadrada de la estadística, es independiente a la forma de la distribución propuesta $F_0(x)$ bajo H_0. Como resultado se tiene que la prueba de bondad de ajuste chi-cuadrada también se emplea en situaciones en las que $F_0(x)$ es continua. Sin embargo, debe hacerse énfasis en que la naturaleza de la prueba de bondad de ajuste chi-cuadrada es discreta en el sentido en el que ésta compara las frecuencias que se observan y se esperan para un número finito de categorías. De acuerdo con lo anterior, si $F_0(x)$ es continua, la prueba no compara las frecuencias que se observan alisadas con la función de densidad propuesta tal como lo implica la hipótesis nula. Más bien, la comparación se lleva a cabo aproximando la distribución continua bajo H_0 con un número finito de intervalo de clase. A pesar de esta limitación, la prueba de bondad de ajuste chi-cuadrada es un procedimiento razonablemente adecuado para probar suposiciones de normalidad siempre y cuando el tamaño de la muestra sea, en forma moderada, grande. Con respecto a la pregunta de qué tan grande debe ser el tamaño de la muestra, se ha encontrado que con n igual a cinco veces el número de clases, los resultados son aceptables. Una regla conservadora a seguir es el seleccionar un muestra de manera tal que toda frecuencia esperada no sea menor que cinco. Lo anterior puede lograrse combinando clases vecinas pero, para cada par de clases que se combina, el número de grados de libertad debe reducirse en uno.

A menos que pueda especificarse una hipótesis alternativa que consista en un modelo alternativo $F_1(x)$ particular, la potencia de la prueba de bondad de ajuste chi-cuadrada es muy difícil de determinar en forma analítica. Sin embargo, puede demostrarse que la potencia tiende a 1 conforme n tiende a ∞. Este resultado implica que para muestras de gran tamaño es casi seguro el rechazar la hipótesis nula debido a que es muy difícil especificar una H_0 lo suficientemente cercana a la verdadera distribución. De esta forma, la aplicabilidad de la prueba de bondad de ajuste chi-cuadrada es cuestionable cuando se tienen muestras de tamaño muy grande.

Ejemplo 10.2 En la tabla 5.2 se proporcionan los datos que se agrupan para el número de respuestas correctas para la prueba SAT de matemáticas, de los alumnos del tercer año de preparatoria. Recuérdese que en el ejemplo 5.5 se compararon las frecuencias que se observaron con las que se esperaron, en donde estas últimas se obtuvieron con base en una distribución normal con media 491 y desviación estándar igual a 120. Con base en la prueba de bondad de ajuste chi-cuadrada, ¿existe alguna razón para creer que el número de respuestas correctas para la prueba de matemáticas SAT no se encuentran distribuidas normalmente con media 491 y desviación estándar igual a 120 a un nivel de $\alpha = 0.01$?

Considérese la prueba de la siguiente hipótesis nula

$$H_0\colon F(x) = F_0(x),$$

en donde $F_0(x)$ es el modelo de probabilidad normal con media 491 y desviación estándar 120. Bajo la hipótesis nula, las frecuencias esperadas para las 12 clases se

encuentran en la última columna de la tabla 5.2. Éstas se determinaron primero convirtiendo cada intervalo de cada clase al correspondiente intervalo normal estándar, empleando para esto $\mu = 491$ y $\sigma = 120$. Después se determinó la probabilidad de cada intervalo bajo H_0. Finalmente, para cada clase, el valor de probabilidad se multiplicó por el tamaño de la muestra $n = 478\,193$ para obtener la frecuencia esperada. Nótese que las probabilidades que aparecen en la penúltima columna de la tabla 5.2 no suman uno. Pero bajo la hipótesis nula las clases deben ser exhaustivas, de manera tal que $\Sigma_{i=1}^{k} p_i = 1$. Lo anterior puede lograrse mediante el ajuste de las clases primera y última de manera tal que la primera no tenga límite inferior y la última no tenga límite superior. Dado que bajo H_0, $X \sim N(491, 120)$,

$$P(X \leq 250) = P(Z \leq -2.01) = 0.0222,$$

y la frecuencia modificada para la primera clase es (478 193) (0.0222) = 10 615.88. De manera similar para la última clase

$$P(X \geq 750) = P(Z \geq 2.16) = 0.0154,$$

lo cual da como resultado una frecuencia esperada de 7 364.17.

Con base en las 12 clases, el valor de la estadística chi-cuadrada es

$$\chi^2 = \frac{(3\,423 - 10\,615.88)^2}{10\,615.88} + \frac{(18\,434 - 16\,115.10)^2}{16\,115.10} + \cdots + \frac{(6\,414 - 7\,364.17)^2}{7\,364.17}$$

$$= 13\,067.02,$$

el cual se encuentra, en forma clara, más allá del valor crítico $\chi^2_{0.99,\,11} = 24.75$. De acuerdo con lo anterior, la hipótesis nula de que el número de respuestas correctas para la prueba SAT se encuentra normalmente distribuido con media 491 y desviación estándar de 120, debe rechazarse. Este ejemplo ilustra el comentario formulado con anterioridad con respecto a muestras de gran tamaño, en donde la hipótesis nula casi seguramente resulta rechazada.

Recuérdese que la hipótesis nula dada por (10.1) es simple ya que el modelo de probabilidad propuesto $F_0(x)$ se especificó de manera completa con respecto a todos sus parámetros. Sin embargo, para muchas aplicaciones que toman en cuenta la bondad del ajuste, sólo puede especificarse la forma de $F_0(x)$. Por ejemplo, supóngase que se desea probar la hipótesis nula de que un conjunto de observaciones de una medida de interés X se ajustan a una distribución normal, pero no puede especificarse el valor de la media o el de la variaza. Lo anterior da como resultado que la hipótesis nula

$$H_0: F(x) = F_0(x)$$

es compuesta. En consecuencia, se tiene que las frecuencias esperadas np_i para las $i = 1, 2 \ldots k$ clases no pueden determinarse, ya que éstas son funciones de los parámetros desconocidos de $F_0(x)$.

Supóngase que T es una estadística para un parámetro desconocido θ de $F_0(x)$. En el contexto de la prueba de bondad de ajuste, tanto las frecuencias observables

N_i como las frecuencias esperadas $np_i(T)$ son variables aleatorias, en donde $p_i(T)$ indica que las probabilidades bajo la hipótesis nula son funciones de la estadística T de θ. Puede demostrarse que si para cualquier parámetro desconocido θ la estadística T es el estimador de máxima verosimilitud de θ, y si las frecuencias esperadas se determinan como funciones de los estimadores de máxima verosimilitud, entonces

$$\sum_{i=1}^{k} \frac{[N_i - np_i(T)]^2}{np_i(T)} \tag{10.3}$$

tiene aproximadamente una distribución chi-cuadrada con $k - 1 - r$ grados de libertad, para valores de n grandes, en donde r es el número de parámetros que se está tratando de estimar.

Al igual que en el caso previo en el que se tenía una H_0, sencilla, la región crítica es el extremo superior de la distribución chi-cuadrada. Pero, a diferencia del caso anterior, el número de grados de libertad se reduce por una cantidad igual al número de parámetros que se están estimando. Como consecuencia, existe un corrimiento hacia la izquierda en el valor crítico para el mismo tamaño del error de tipo I, y la hipótesis nula puede rechazarse para un valor observado más pequeño de (10.3) que en el caso previo. Lo anterior es lógico ya que el ajuste deberá ser mejor debido a que los parámetros desconocidos se estiman con base en las observaciones de la muestra.

Las características importantes para la aplicación de la prueba de bondad de ajuste chi-cuadrada para el caso compuesto son idénticas a las que tienen para la hipótesis nula simple. Surge un problema relativamente pequeño al decidir si los parámetros desconocidos deberán estimarse con base en los datos que se agruparon en los que no. En forma teórica, ninguno de los dos enfoques puede ser el correcto debido a que los estimados de máxima verosimilitud deben obtenerse maximizando la función de verosimilitud con base en la distribución multinomial. En forma afortunada, resulta que la mayoría de las veces el error que se comete no es serio. De esta forma, se pueden utilizar los estimados de máxima verosimilitud obtenidos, ya sea de los datos agrupados o de los no agrupados, en forma segura.

Ejemplo 10.3 Recuérdese el ejemplo 4.5 en el que se compararon el número de anotaciones de seis puntos por equipo y por juego en la NFL con el número que esperaban de éstos, si el número de anotaciones de seis puntos tiene una distribución de Poisson. Con base en la información contenida en la tabla 4.3, ¿existe alguna razón para creer, a un nivel de 0.05, que el número de anotaciones no es variable aleatoria de Poisson?

Dado que el valor del parámetro de Poisson λ no se especifica, el estimado de máxima verosimilitud de λ con base en la información que se proporcionó en la tabla 4.3 es $\hat{\lambda} = 2.435$. Bajo la hipótesis nula de una distribución de Poisson, la probabilidad de tener cero anotaciones es

$$p(0) = (2.435)^0 \exp(-2.435)/0! = 0.0876.$$

Para $n = 448$, el número esperado de cero anotaciones es $(448)(0.0876) = 39.24$. Si se sigue este procedimiento, pueden obtenerse las demás frecuencias esperadas. En la tabla 10.1, se presenta el cálculo de la estadística chi-cuadrada.

TABLA 10.1 Cálculo de la estadística chi-cuadrada para el ejemplo 10.3

Número de anotaciones	*Frecuencia observada*	*Frecuencia esperada*	$\frac{[n_i - np_i(\hat{\lambda})]^2}{np_i(\hat{\lambda})}$
0	35	39.24	0.458
1	99	95.56	0.124
2	104	116.34	1.309
3	110	94.44	2.564
4	62	57.48	0.355
5	25	28.00	0.321
6	10	11.38	0.167
7	3	5.56	1.179
Totales	448	448	6.477

Para $k = 8$ categorías y con un parámetro estimado, el número de grados de libertad es 6. Para $\alpha = 0.05$ el valor crítico es $\chi^2_{0.95,6} = 12.60$. Dado que $\chi^2 = 6.477 < \chi^2_{0.95,6} = 12.60$, no puede rechazarse la hipótesis nula de que el número de anotaciones de seis puntos por equipo en la NFL es una variable aleatoria de Poisson.

10.3 La estadística de Kolmogorov-Smirnov

Recuérdese que para aplicar la prueba de bondad de ajuste chi-cuadrada cuando el modelo propuesto bajo H_0 es continuo, es necesario aproximar $F_0(x)$ mediante el agrupamiento de los datos observados en un número finito de intervalos de clase. Este requisito de agrupar los datos implica tener una muestra de tamaño más o menos grande. De esta manera, la prueba de bondad de ajuste chi-cuadrada se encuentra limitada cuando $F_0(x)$ es continua y la muestra aleatoria disponible tiene un tamaño pequeño. Una prueba de bondad de ajuste más apropiada que la chi-cuadrada cuando $F_0(x)$ es continua, es la basada en la estadística de Kolmogorov-Smirnov. La prueba de Kolmogorov-Smirnov no necesita que los datos se encuentren agrupados y es aplicable a muestras de tamaño pequeño. Ésta se basa en una comparación entre las funciones de distribución acumulativa que se observan en la muestra ordenada y la distribución propuesta bajo la hipótesis nula. Si esta comparación revela una diferencia suficientemente grande entre las funciones de distribución muestral y propuesta, entonces la hipótesis nula de que la distribución es $F_0(x)$, se rechaza.

Considérese la hipótesis nula por (10.1), en donde $F_0(x)$ se especifica en forma completa. Denótense por $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ a las observaciones ordenadas de una muestra aleatoria de tamaño n y definase la función de distribución acumulativa muestral como

$$S_n(x) = \begin{cases} 0 & x < x_{(1)}, \\ k/n & x_{(k)} \leq x < x_{(k+1)}, \\ 1 & x \geq x_n. \end{cases} \tag{10.4}$$

En otras palabras, para cualquier valor ordenado x de la muestra aleatoria, $S_n(x)$ es la proporción del número de valores en la muestra que son iguales o menores a x. Ya que $F_0(x)$ se encuentra completamente especificada, es posible evaluar a $F_0(x)$ para algún valor deseado de x, y entonces comparar este último con el valor correspondiente de $S_n(x)$. Si la hipótesis nula es verdadera, entonces es lógico esperar que la diferencia sea relativamente pequeña. La estadística de Kolmogorov-Smirnov se define como

$$D_n = \max_x | S_n(x) - F_0(x) |. \tag{10.5}$$

La estadística D_n tiene una distribución que es independiente del modelo propuesto bajo la hipótesis nula. Por esta razón, se dice D_n es una estadística independiente de la distribución. Lo anterior da como resultado que la función de distribución de D_n pueda evaluarse sólo en función del tamaño de la muestra y después usarse para cualquier $F_0(x)$. En la tabla J del apéndice, se proporcionan los valores cuantiles superiores de D_n para varios tamaños de la muestra. El lector debe notar que los valores asintóticos de d_n que se encuentran en la parte inferior de la tabla proporcionan una adecuada aproximación para valores de n mayores de 50.

Para un tamaño α del error de tipo I, la región crítica es de la forma

$$P\left(D_n > \frac{c}{\sqrt{n}}\right) = \alpha.$$

De acuerdo con lo anterior, la hipótesis H_0 se rechaza si para algún valor x observado el valor de D_n se encuentra dentro de la región crítica de tamaño α.

Como se hizo notar anteriormente, la estadística de Kolmogorov-Smirnov es, en general, superior a la prueba de bondad de ajuste chi-cuadrada cuando los datos involucran una variable aleatoria continua, debido a que no es necesario agrupar los datos. Además, la prueba de Kolmogorov-Smirnov tiene la atractiva propiedad de ser aplicable a muestras de tamaño pequeño. Por otro lado, la estadística se encuentra limitada, ya que el modelo propuesto bajo H_0 debe especificarse en forma completa. La estadística de Kolmogorov-Smirnov no se aplica a todos aquellos casos para los que las observaciones no son inherentemente cuantitativas a consecuencia de las ambigüedades que pueden surgir cuando se ordenan las observaciones.

Ejemplo 10.4 A continuación se proporcionan los valores ordenados de una muestra aleatoria del número de respuestas correctas para la SAT que se aplicó a todos los estudiantes que ingresaron a una universidad: 852, 875, 910, 933, 957, 963, 981, 998, 1007, 1010, 1015, 1018, 1023, 1035, 1048, 1063. En años anteriores, el número de respuestas correctas estaba representado, en forma adecuada, por una distribución normal con media 985 y desviación estándar 50. Con base en esta muestra, ¿existe alguna razón para creer que ha ocurrido un cambio en la distribución de respuestas correctas para la prueba SAT en esta universidad? Empléese un nivel $\alpha = 0.05$.

Sea X la variable aleatoria que representa el número de respuestas correctas para la prueba SAT. Considérese la prueba de la siguiente hipótesis nula

$$H_0: F(x) = F_0(x),$$

donde $F_0(x)$ es la función de distribución normal con media 985 y desviación estándar 50. Dado que X es una variable aleatoria continua y el tamaño de la muestra de X es pequeño, se usará la estadística de Kolmogorov-Smirnov para probar a H_0. La función de distribución muestral se obtiene mediante el empleo de (10.4) para los valores ordenados. Lo anterior involucra un incremento de $1/6 = 0.0625$ al valor previo de la distribución muestral. Los valores correspondientes del modelo normal propuesto se obtienen estandarizando primero a $N(0, 1)$ y empleando la tabla D del apéndice. En la tabla 10.2 se encuentra la información más importante.

Se observa que la máxima desviación es de 0.1207. De la tabla J del apéndice, el valor crítico de D_{16} para $\alpha = 0.05$ es 0.328. Dado que $0.1207 < 0.328$, no puede rechazarse la hipótesis nula. De acuerdo con ello no es posible detectar un cambio en la distribución para el número de respuestas correctas de la prueba SAT de la ya establecida $N(985, 50)$.

10.4 La prueba chi-cuadrada para el análisis de tablas de contingencia con dos criterios de clasificación

Muchas veces surge la necesidad de determinar si existe alguna relación entre dos rasgos diferentes en los que una población ha sido clasificada y en donde cada rasgo se encuentra subdividido en cierto número de categorías. Por ejemplo, ¿existe una relación entre el fumar cigarrillos y la predisposición a desarrollar cáncer pulmonar?, o también ¿existe una relación entre la filiación política y la opinión con respecto a incrementar el presupuesto armamentista? En ambos ejemplos, se ha clasificado a la población en dos características y en donde se supone que cada una de

TABLA 10.2 Cálculo de la estadística de Kolmogorov-Smirnov para el ejemplo 10.4

Valores ordenados	$S_n(x)$	$F_0(x)$	$\lvert S_n(x) - F_0(x)\rvert$
852	0.0625	0.0039	0.0586
875	0.1250	0.0139	0.1111
910	0.1875	0.0668	0.1207
933	0.2500	0.1492	0.1008
957	0.3125	0.2877	0.0248
963	0.3750	0.3300	0.0450
981	0.4375	0.4681	0.0306
998	0.5000	0.6026	0.1026
1007	0.5625	0.6700	0.1075
1010	0.6250	0.6915	0.0665
1015	0.6875	0.7257	0.0382
1018	0.7500	0.7454	0.0046
1023	0.8125	0.7764	0.0361
1035	0.8750	0.8413	0.0337
1048	0.9375	0.8962	0.0413
1063	1.0000	0.9406	0.0594

éstas tiene por lo menos dos categorías exhaustivas y mutuamente excluyentes. En el primer ejemplo las dos características son, si se es fumador, y si desarrolla cáncer pulmonar. Las categorías para estas dos características podrían ser si se es fumador crónico, moderado o no fumador, para la primera, y el si se desarrolla o no cáncer pulmonar para la segunda.

Cuando una muestra aleatoria que se obtiene de una población se clasifica de esta manera, el resultado recibe el nombre de *tabla de contingencia con dos criterios de clasificación*. Esta tabla se forma por las frecuencias relativas que se observaron para las dos clasificaciones y sus correspondientes categorías. A pesar de que sólo se analizarán tablas de contingencia con dos clasificaciones, es posible analizar tablas que contengan más de dos clasificaciones.

El análisis de una tabla de este tipo supone que las dos clasificaciones son independientes. Esto es, bajo la hipótesis nula de independencia se desea saber si existe una diferencia suficiente entre las frecuencias que se observan y las correspondientes frecuencias que se esperan, tal que la hipótesis nula se rechace. La prueba chi-cuadrada, discutida en la sección 10.2, proporciona los medios apropiados para analizar este tipo de tablas.

Sea n una muestra aleatoria de una población que se clasifica de acuerdo con dos características A y B, cada una de las cuales contiene un número r y c de categorías, respectivamente. Además, sea N_{ij} el número de observaciones en la categoría (i, j), de las características A y B, respectivamente, para $i = 1, 2 \ldots r$ y $j = 1, 2 \ldots c$. Entonces una tabla de contingencia es un arreglo matricial de $r \times c$, dado en la tabla 10.3, en donde las entradas representan las realizaciones de las variables aleatorias N_{ij}.

Nótese que el total del i-ésimo renglón es la frecuencia de la i-ésima categoría de característica A, sumando sobre todas las categorías de la característica B. De manera similar, el total de la j-ésima columna es la frecuencia observada de la j-ésima categoría de B sumada sobre todas las categorías de A. Sean

$$n_{i\cdot} = \sum_{j=1}^{c} n_{ij} \qquad i = 1, 2, \ldots, r,$$

$$n_{\cdot j} = \sum_{i=1}^{r} n_{ij} \qquad j = 1, 2, \ldots, c,$$

TABLA 10.3 Tabla de contingencia con dos clasificaciones

		Característica B				
	Categorías	1	2	⋯	c	*Totales*
Característica A	1	n_{11}	n_{12}	⋯	n_{1c}	$n_{1\cdot}$
	2	n_{21}	n_{22}	⋯	n_{2c}	$n_{2\cdot}$
	·	·	·	⋯	·	·
	·	·	·	⋯	·	·
	·	·	·	⋯	·	·
	r	n_{r1}	n_{r2}		n_{rc}	$n_{r\cdot}$
	Totales	$n_{\cdot 1}$	$n_{\cdot 2}$	⋯	$n_{\cdot c}$	n

los símbolos para denotar las sumas de los renglones y de las columnas, respectivamente, en donde la notación "punto" indica el subscripto sobre el cual se lleva a cabo la sumatoria.

Sea p_{ij} la probabilidad de que un objeto seleccionado al azar de una población de interés se encuentre en la categoría *(i, j)* de la tabla de contingencia. Sea $p_{i\cdot}$ la probabilidad (marginal) de que un objeto se encuentre en la categoría *i* de la característica A, y sea $p_{\cdot j}$ la probabilidad de que un objeto se encuentre en la categoría *j* de la característica B. Si las dos características son independientes, la probabilidad conjunta debe ser igual al producto de las probabilidades marginales. De esta forma puede establecerse la hipótesis nula de la siguiente manera:

$$H_0: p_{ij} = p_{i\cdot} p_{\cdot j} \qquad i = 1, 2, \ldots, r; \; j = 1, 2, \ldots, c. \tag{10.6}$$

Si pueden especificarse las probabilidades marginales $p_{i\cdot}$ y $p_{\cdot j}$, entonces, bajo la hipótesis nula, la estadística

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(N_{ij} - np_{i\cdot}p_{\cdot j})^2}{np_{i\cdot}p_{\cdot j}} \tag{10.7}$$

tiene en forma aproximada una distribución chi-cuadrada con $rc - 1$ grados de libertad para valores grandes de n. Sin embargo, la mayoría de las veces pueden no conocerse los valores de las probabilidades marginales y, de esta forma, se estiman con base en la muestra. Afortunadamente, la prueba de bondad de ajuste chi-cuadrada permanece como la estadística apropiada para probar (10.6), siempre que se empleen los estimados de máxima verosimilitud y se reste un grado de libertad del total para cada parámetro que se esté estimando. Dado que $\sum_{i=1}^{r} p_{i\cdot} = 1$ y $\sum_{j=1}^{c} p_{\cdot j} = 1$, existen $r - 1$ parámetros de renglón y $c - 1$ de columna a ser estimados. De esta forma, el número de grados de libertad será $rc - 1 - (r - 1) - (c - 1) = rc - r - c + 1 = (r - 1)(c - 1)$.

Puede demostrarse que los estimados de máxima verosimilitud de $p_{i\cdot}$ y $p_{\cdot j}$ están dados por

$$\hat{p}_{i\cdot} = n_{i\cdot}/n, \tag{10.8}$$

y

$$\hat{p}_{\cdot j} = n_{\cdot j}/n, \tag{10.9}$$

respectivamente. Al sustituir (10.8) y (10.9) en (10.7), se obtiene la estadística

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(N_{ij} - \dfrac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\dfrac{n_{i\cdot}n_{\cdot j}}{n}}, \tag{10.10}$$

que para valores grandes de n es, en forma aproximada, una variable aleatoria chi-cuadrada con $(r - 1) \times (c - 1)$ grados de libertad.

Ejemplo 10.5 Una compañía evalúa una propuesta para fusionarse con una corporación. El consejo de directores desea muestrear la opinión de los accionistas para determinar si ésta es independiente del número de acciones que cada uno posee. Una muestra aleatoria de 250 accionistas proporciona la información que se muestra en la tabla 10.4. Con base en esta información, ¿existe alguna razón para dudar de que la opinión con respecto a la propuesta es independiente del número de acciones que posee el accionista? Úsese $\alpha = 0.10$.

La hipótesis nula se establece de la siguiente forma

$$H_0: p_{ij} = p_{i\cdot}p_{\cdot j}, \qquad i = 1, 2, 3; \quad j = 1, 2, 3.$$

En ésta, p_{ij} es la probabilidad de que un accionista seleccionado al azar se encuentre en la categoría *(i, j)*; $p_{i\cdot}$ es la probabilidad marginal de que el número de acciones que posee un accionista seleccionado al azar se encuentre en la categoría *i;* y $p_{\cdot j}$ es la probabilidad marginal de que un accionista seleccionado al azar tenga una opinión *j*. Por la expresión (10.10) la frecuencia esperada de la celda *(i, j)* es el producto del total de *i*-ésimo renglón por el total de la *j*-ésima columna dividido por el tamaño de la muestra $n = 250$. Por ejemplo, el número esperado de accionistas que están a favor de la propuesta y que poseen más de 1 000 acciones, es (95)(100)/250 = 38. Al continuar este proceso, se determinan las frecuencias esperadas para cada combinación. En cada celda de la tabla 10.5, la primera línea representa la frecuencia observada, la segunda la frecuencia esperada y la tercera la contribución de cada celda al valor de la estadística, de acuerdo con (10.10).

De esta manera, el valor de la estadística es

$$\chi^2 = \frac{(38 - 30.4)^2}{30.4} + \frac{(29 - 39.52)^2}{39.52} + \cdots + \frac{(4 - 7.6)^2}{7.6} = 10.80.$$

Dado que $r = c = 3$, el número de grados de libertad es 4. Para $\alpha = 0.1$, el valor crítico es $\chi^2_{0.9,4} = 7.78$. De esta forma, el valor que se observa de la estadística de prueba se encuentra dentro de la región crítica, y la hipótesis nula debe rechazarse. De acuerdo con lo anterior, existe una razón para creer que la opinión con respecto a la propuesta y el número de acciones que cada accionista posee, no son independientes.

TABLA 10.4 Datos muestrales para el ejemplo 10.5

Número de acciones	*Opinión*			*Totales*
	A favor	*En contra*	*Indecisos*	
Menos de 200	38	29	9	76
200–1000	30	42	7	79
Más de 1000	32	59	4	95
Totales	100	130	20	250

TABLA 10.5 Frecuencias esperadas y observadas para el ejemplo 10.5

Número de acciones	*A favor*	*En contra*	*Indecisos*	*Totales*
	38	29	9	76
Menos de 200	30.40	39.52	6.08	76
	1.90	2.80	1.40	6.10
	30	42	7	79
200–1000	31.60	41.08	6.32	79
	0.08	0.02	0.07	0.17
	32	59	4	95
Más de 1000	38	49.40	7.60	95
	0.95	1.87	1.71	4.53
	100	130	20	250
Totales	100	130	20	250
	2.93	4.69	3.18	10.80

Referencias

1. P. G. Hoel, *Introduction to mathematical statistics*, 4th ed., Wiley, New York, 1971.
2. B. W. Lindgren, *Statistical theory*, 3rd ed., Macmillan, New York, 1976.

Ejercicios

10.1. Con base en los registros de una tienda de modas, el 50% de los vestidos adquiridos por ésta para la temporada se venderán a precio de menudeo, el 25% a un 20% menos del precio de menudeo, 15% se venderán después de una reducción en su precio del 40% y los restantes con una disminución en su precio del 60%. Para esta temporada, se adquirieron 300 vestidos y su venta fue en la siguiente forma:

Precio de venta	*20% de*	*40% de*	*60% de*
140	90	30	40

¿Existe alguna razón para creer que la disminución en ventas fue diferente en esta temporada con respecto a las anteriores? Úsese $\alpha = 0.05$. ¿Cuál es el valor de p?

10.2. En un hospital, el número de nacimientos observados para cada mes de cierto año, fueron los siguientes:

Ene	Feb	Marzo	Abril	Mayo	Jun	Julio	Ago	Sept	Oct	Nov	Dic
95	105	95	105	90	95	105	110	105	100	95	100

Si $\alpha = 0.01$, ¿existe alguna razón para creer que el número de nacimientos no se encuentra distribuido en forma uniforme durante todos los meses del año? ¿Cuál es el valor de p?

10.3. En el ejercicio 10.2, supóngase que el número de nacimientos que se observaron cada mes durante un periodo de 10 años es simplemente igual a diez veces los números observados en el ejercicio 10.2 para un año.

a) ¿Cambiará esto la conclusión del ejercicio 10.2?
b) ¿Qué puede concluirse con respecto al empleo de prueba de bondad de ajuste chi-cuadrada para valores grandes de n?

10.4. Un fabricante asegura que produce sólo el 5% de unidades defectuosas. Un comprador de grandes cantidades de estas unidades selecciona 100 y encuentra diez defectuosas.

a) Mediante el empleo de la prueba de bondad de ajuste chi-cuadrada, determinar si existe una razón para dudar de la afirmación del fabricante. Úsese $\alpha = 0.05$.
b) Compárese la respuesta con la parte *a*, que se obtiene al utilizar el método aproximado que se discutió en el capítulo 9 para probar la hipótesis nula de que la verdadera proporción de artículos defectuosos es 0.05.
c) ¿Existe alguna relación entre los valores de las estadísticas de prueba obtenidos en las partes *a* y *b*? ¿Existe alguna condición para esta relación?

10.5. Una organización de seguridad vial desea determinar si el número de accidentes fatales se encuentra distribuido de igual forma para el color de los automóviles involucrados en los accidentes. La organización obtuvo una muestra aleatoria de 600 accidentes automovilísticos en los cuales ocurrió por lo menos una muerte y anotó el color del automóvil. Se obtuvo la siguiente información:

Rojo	*Café*	*Amarillo*	*Blanco*	*Gris*	*Azul*
75	125	70	80	135	115

¿Existe alguna razón para creer que las proporciones de color no son idénticas? Úsese $\alpha = 0.01$.

10.6. Durante un periodo de 30 años se llevó a cabo un estudio médico para determinar, entre otras cosas, si los hábitos de fumador pueden influenciar en el desarrollo de la enfermedad cardiaca. Durante este periodo, 160 hombres desarrollaron alguna enfermedad cardiaca. Estos hombres fueron clasificados como fumadores agudos (más de dos cajetillas de cigarros al día), fumadores moderados (una a dos cajetillas al día), fumadores ocasionales (menos de una cajetilla al día) o no fumadores. El número de hombres en cada categoría que desarrolló alguna enfermedad cardiaca es el siguiente:

Fumador agudo	*Fumador moderado*	*Fumador ocasional*	*No fumador*
58	54	36	12

a) Si se supone que al comienzo del estudio había una cantidad igual de hombres en cada una de las cuatro categorías, ¿existe alguna razón a un nivel de $\alpha = 0.01$ para creer que las proporciones en estas categorías no son las mismas?
b) ¿Cómo se podría prevenir al investigador médico del uso de la prueba de bondad de ajuste chi-cuadrada en esta situación?

10.7. En un proceso de producción se toma una muestra aleatoria diaria de 100 artículos y se inspecciona para encontrar artículos defectuosos. Para una semana dada y para los cinco días de operación, se observó el siguiente número de unidades defectuosas:

Lunes	*Martes*	*Miércoles*	*Jueves*	*Viernes*
12	7	6	5	10

Si el porcentaje total de artículos defectuosos es del 8%, ¿puede concluirse que a un nivel de $\alpha = 0.05$ existe una diferencia discernible en el porcentaje diario de artículos defectuosos?

10.8. Con referencia a los datos del ejercicio 1.1, empleando la prueba de bondad de ajuste chi-cuadrada, ¿puede concluirse que los lapsos de tiempo no se encuentran exponencialmente distribuidos con $\theta = 3.2$ minutos? Úsese $\alpha = 0.01$.

10.9. Considere los datos del ejercicio 1.7.

a) Para $\alpha = 0.05$, empléese la prueba de bondad de ajuste chi-cuadrada para probar la hipótesis nula de que la distribución del número de anotaciones de seis puntos por equipo y por juego en la NFL, es una distribución de Poisson con parámetro $\lambda = 2.7$.

b) Supóngase que se estima el valor de λ a partir de los datos. ¿Cómo podría este cambio efectuar la respuesta a la parte *a*?

10.10. Úsese la estadística de Kolmogorov-Smirnov en los datos del ejercicio 1.1 y compare el resultado con el que se obtiene en el ejercicio 10.8.

10.11. Úsese lá estadística de Kolmogorov-Smirnov para probar la hipótesis nula de que los datos del ejercicio 1.2 se encuentran normalmente distribuidos con media 50 y desviación estándar 10. Úsese $\alpha = 0.05$.

10.12. Como se notó con anterioridad, una limitación de la estadística de Kolmogorov-Smirnov es que debe especificarse el modelo propuesto bajo H_0. A pesar de que no se encuentra disponible ningún método cuando algunos de los parámetros no se especifica, Lilliefors* obtuvo los límites de rechazo a través de un estudio de simulación para el problema específico de probar la normalidad. Si la media y la desviación estándar muestral se emplean como parámetros de la distribución normal bajo la hipótesis nula, la estadística D_n tiene una distribución cuyos cuantiles también obtuvo Lilliefors. De manera específica, para $\alpha = 0.05$ los valores del 95avo, percentil de la distribución de esta estadística bajo H_0 fueron los siguientes:

n	10	12	14	15	16	18	20	25	>25
95avo. percentil	0.258	0.242	0.227	0.220	0.213	0.200	0.190	0.173	$0.886/\sqrt{n}$

Empléese la modificación de Lilliefors a la estadística de Kolmogorov-Smirnov para probar la normalidad de los datos del ejercicio 1.2. Compárese el resultado con el del ejercicio 10.11.

10.13. Úsese el procedimiento de la prueba de bondad de ajuste chi-cuadrada para probar la hipótesis nula de que los datos del ejercicio 1.2 se encuentran distribuidos, normalmente, a un nivel de $\alpha = 0.01$.

10.14. Se toma una muestra aleatoria de 25 hombres casados y se les pregunta la edad que tenían cuando se casaron. Se obtienen los siguientes datos: 24, 19, 20, 22, 50, 23, 23,

On the Kolmogorov–Smirnov test for normality with mean and variance unknown*, J. Amer. Statistical Assoc. **64 (1967), 399–402. 1967.

21, 25, 27, 45, 27, 26, 26, 35, 29, 28, 30, 31, 32, 31, 33, 34, 38, 41. Úsese la estadística de Kolmogorov-Smirnov para probar la hipótesis nula de que la distribución de las edades de los hombres cuando contrajeron sus primeras nupcias es una distribución gama con $\theta = 2$ y $\alpha = 16$. Úsese $\alpha = 0.05$. (Sugerencia: Para calcular las probabilidades gama, véase una tabla de la función gama incompleta determinada por (5.55).)

10.15. En el ejemplo 4.10, úsese la prueba de bondad de ajuste chi-cuadrada para demostrar que la hipótesis nula de una distribución binomial negativa para el número de anotaciones de seis puntos, no puede ser rechazada a un nivel $\alpha = 0.05$.

10.16. Con la prueba de bondad de ajuste chi-cuadrada determínese si la hipótesis nula de los datos del accidente del ejercicio 8.14 sigue una distribución binomial negativa, que se puede remitir al nivel $\alpha = 0.05$

10.17. Los totales de los renglones y columnas de una tabla de contingencia de dos características son los siguientes:

				10
				12
				15
8	14	10	5	37

Bajo la hipótesis nula de independencia, determinar la tabla de frecuencias esperadas.

10.18. Un proceso de producción emplea cinco máquinas en sus tres operaciones de desplazamiento. Se clasificó una muestra aleatoria de 164 fallas de acuerdo con la máquina y la operación de desplazamiento en la que ocurrió la falla, y los resultados se muestran en la tabla 10.6. Con base en esta información, ¿existe alguna razón para dudar acerca de la independencia entre la operación de desplazamiento y la falla de la máquina? Úsese $\alpha = 0.01$.

TABLA 10.6 Fallas por máquina y desplazamiento

	Máquinas				
Desplazamiento	*A*	*B*	*C*	*D*	*E*
1	10	12	8	14	8
2	15	8	13	8	11
3	12	9	14	12	10

10.19. Se condujo una encuesta aleatoria entre los ciudadanos en edad de votar para determinar si existía alguna relación entre la afiliación partidista y la opinión con respecto al control de armas. Se obtuvo la información proporcionada en la tabla 10.7. Para $\alpha = 0.01$, ¿existe alguna razón para creer que existe una dependencia entre la opinión y la afiliación partidista?

TABLA 10.7 Filiación partidaria y opiniones sobre el control de armas

	A favor	*En contra*	*Sin decisión*
Demócratas	110	64	26
Republicanos	90	116	14
Independientes	55	35	10

10.20. En una muestra aleatoria de recién egresados de la preparatoria se registraron dos características (la calificación promedio y el número de respuestas correctas para la prueba SAT). Esta información se clasificó como se muestra en la tabla **10.8**

TABLA 10.8 Calificaciones promedio y número de respuestas correctas para la prueba SAT

	Número de respuestas correctas para la prueba SAT		
GPA	900–1100	1100–1300	1300–1500
>3.5	50	65	38
3.0–3.5	78	72	42
2.5–3.0	97	80	25
2.0–2.5	105	25	18

a) ¿Existe una dependencia entre el número de respuestas correctas en la prueba SAT y el promedio de clasificaciones. discernible estadísticamente a un nivel $\alpha = 0.01$?

b) ¿Se tiene alguna reserva con respecto a esta clasificación? ¿Se puede pensar en otras características que deban considerarse?

10.21. En un estudio reciente que involucró una muestra aleatoria de 300 accidentes automovilísticos, se clasificó la información de acuerdo con el tamaño del automóvil.

	Pequeño	*Mediano*	*Grande*
Por lo menos un muerto	42	35	20
Ningún muerto	78	65	60

Con estos datos, ¿depende la frecuencia de accidentes del tamaño del automóvil? Úsese $\alpha = 0.05$.

10.22. Se llevó a cabo una encuesta con respecto a la preferencia del consumidor para determinar si existía alguna predilección para tres marcas competitivas (A, B y C) dependiendo de la región geográfica en la que habita el consumidor. Con base en una muestra aleatoria de consumidores, se obtuvo la siguiente información para tres distintas regiones.

	Región 1	*Región 2*	*Región 3*
Marca A	40	52	25
Marca B	52	70	35
Marca C	68	78	60

Con base en esta información, ¿la preferencia por una determinada marca depende de la región geográfica a un nivel $\alpha = 0.05$?

CAPÍTULO ONCE

Métodos para el control de calidad y muestreo para aceptación

11.1 Introducción

En los últimos años ha aumentado el interés que se tiene, por parte de los productores así como de los consumidores, en la calidad de los productos manufacturados. Un fabricante que desea mantener cierto nivel de calidad en su producto terminado debe implantar un procedimiento para detectar cualquier desviación seria del estándar de calidad deseado. En el logro de este fin, las tablas estadísticas de control de calidad y el muestreo periódico han demostrado ser medios muy efectivos para controlar la calidad de los productos manufacturados.

Por otro lado, el consumidor desea asegurarse de que el producto que adquiere reúne ciertos estándares de calidad. Lo anterior es especialmente cierto si el consumidor, como muchas veces ocurre en la práctica, compra lotes muy grandes de cierto producto. En estos casos es necesario establecer un procedimiento para inspeccionar una muestra relativamente pequeña del producto proveniente del lote para decidir si reúne los estándares de calidad deseados. Un procedimiento de esta naturaleza incluye la noción del muestreo para aceptación.

En este capítulo se analizarán los principios básicos y métodos de las tablas de control estadístico y los procedimientos del muestreo para aceptación. El lector debe considerar el material de este capítulo sólo como introducción al control estadístico de calidad y a los procedimientos del muestreo para aceptación, pero éste debe ser útil como antecedente para un estudio posterior. Con este propósito se sugieren las referencias [2] y [3].

11.2 Tablas de control estadístico

Una tabla de control estadístico es un procedimiento inferencial basado en un muestreo repetitivo para estudiar un proceso. De acuerdo con su creador, W.A.

Shewhart, una tabla de control se emplea para definir un estándar de calidad para un proceso de fabricación y para determinar si éste se mantiene por el proceso.

En el desarrollo de tablas de control, el factor clave es la variabilidad en la calidad del producto terminado. Para cualquier proceso, es inherente cierta cantidad de variabilidad en la calidad, sin importar cuántos esfuerzos se encaminen para lograr su control. Este tipo de variabilidad es una función de factores aleatorios que, de manera común, se encuentran más allá del control. Esta variación aleatoria generalmente es aceptable y no compromete en modo alguno el estándar de calidad deseado. La variabilidad también se puede deber a causas no aleatorias o fijas; éstas pueden tomar la forma de un mal funcionamiento en una máquina, indiferencia del trabajador, variabilidad en la calidad de las materias primas y otras. De esta forma, una tabla de control estadístico es el procedimiento inferencial con el cual se decide si una desviación observada de la norma deseada se debe sólo al azar o a alguna causa fija. Si la decisión es que la variación es aleatoria, entonces se dice que el proceso de interés se encuentra bajo control. De otro modo, se juzga como fuera de control y en este caso lo que se hace, en forma general, es detener el proceso y llevar a cabo todos los esfuerzos necesarios para detectar la causa del problema.

Dado que la inferencia se basa en la probabilidad, es posible que un proceso se juzgue fuera de control cuando, de hecho, se encuentra bajo control o viceversa. Las consecuencias de estos errores pueden ser severas; por ejemplo si se declara a un proceso como fuera de control, cuando en realidad está bajo control, se tratará de determinar una causa inexistente. Por otro lado, si el proceso en realidad está fuera de control y se permite que éste continúe, el estándar de calidad deseado no se alcanzará. Debe notarse que estos errores son facsímiles de los errores de tipo I y II analizados en el capítulo 9.

Usualmente, la determinación de una tabla de control depende de la toma periódica de muestras aleatorias de tamaño n del proceso de interés, con lo que se obtiene, para cada una de éstas, un valor de alguna estadística de importancia como la media o la varianza muestral. Por lo tanto, la tabla de control es una gráfica de los valores de la estadística observada, contra el número de la muestra o contra el periodo durante el cual se obtuvo ésta. La tabla contiene límites de control superior e inferior, los cuales constituyen los criterios de decisión para el proceso, es decir, el proceso será juzgado como bajo control mientras los valores de la estadística se encuentren dentro de estos límites. Si un valor de la estadística se encuentra fuera de los límites de control, se considerará al proceso como fuera de control. También se encuentra una línea central que define la norma prescrita para el proceso.

El usuario decide cuáles deben ser los valores de los límites de control, cuántas veces es necesario muestrear, cuál debe ser el tamaño de la muestra que se toma y qué acción realizar una vez que se juzga al proceso como fuera de control. Sin embargo, existen algunos principios generales que el usuario puede seguir. Shewhart argumentaba que podía alcanzarse un balance apropiado entre el costo del muestreo y la exactitud del estimador, si las muestras tienen un tamaño de cuatro o cinco observaciones cada vez. También los límites de control "tres-sigma" han demostrado ser muy satisfactorios y son los que se emplean en Estados Unidos, así como en muchos otros países.

Considérense las tablas de control para la media y la desviación estándar. La primera se conoce como tabla $\bar{X}$ y la segunda como tabla S. Debe notarse que, de ma-

nera tradicional, se emplea el rango R para determinar tablas para la variabilidad de un proceso debido a su cálculo fácil. Pero es mejor la tabla S, la cual no ofrece ningún problema de cálculo con los paquetes para computadora disponibles en la actualidad. Para la determinación de las tablas $\bar{X}$ y S se supondrá que se muestrea una distribución normal; en un caso, se dará por hecho que se conoce el valor de la media o el de la varianza y, para el otro, que ambos valores son desconocidos.

11.2.1 Tablas $\bar{X}$ (media conocida de la población)

Se puede construir una tabla de control con base en la media muestral cuando la medición de interés se encuentra normalmente distribuida con media μ y desviación estándar o σ conocidas. El conocimiento que se tiene sobre μ y σ se puede deber a la naturaleza particular del proceso de interés, el cual puede proporcionar la suficiente información con respecto a la media y a la desviación estándar. Para este caso, una tabla $\bar{X}$ proporciona el procedimiento inferencial por medio del cual se puede decidir si la media del proceso es la que se afirma.

Sea $X_1, X_2, \ldots, X_n$ una muesta aleatoria de tamaño n del proceso de interés. Dado que por hipótesis $X_i \sim N(\mu,\sigma)$, la media muestral es $\bar{X} \sim N(\mu,\sigma/\sqrt{n})$, la probabilidad de que $|\bar{X} - \mu|$ sea menor que $3\sigma/\sqrt{n}$, es

$$P(|\bar{X} - \mu| < 3\sigma/\sqrt{n}) = 0.9974.$$

De esta forma, los límites de control tres-sigma son $\mu \pm 3\sigma/\sqrt{n}$, es decir, cuando se toma una muestra de tamaño n se calcula y se grafica un valor de la media muestral. Si éste se encuentra dentro de los límites de control $\mu \pm 3\sigma/\sqrt{n}$, se supone que el proceso se encuentra bajo control; de otra forma, está fuera de control. Por lo tanto, cada vez que se toma una muestra se está probando la hipótesis nula de que la media del proceso es igual a μ contra la alternativa de que ha ocurrido un corrimiento en la media del proceso. El rechazo de la hipótesis nula implica que el proceso se encuentra fuera de control.

Ejemplo 11.1 En un proceso de llenado se tiene una máquina que vacía una cantidad promedio de 500 g en cada recipiente, con una desviación estándar de 2 g. Se toman 10 muestras diarias, cada una de cinco recipientes, y se mide el peso de cada recipiente. Los pesos promedio para las 10 muestras en una semana dada son los siguientes:

Número de muestra	1	2	3	4	5
Promedio de la muestra	498.37	499.49	501.25	498.63	502.97
Número de muestra	6	7	8	9	10
Promedio de la muestra	500.56	499.23	498.76	501.05	500.27

Para los límites de control 3σ, ¿se encontró el proceso bajo control durante esta semana? Con estos límites, ¿cuál es la probabilidad de no detectar un corrimiento de 500 a 503 g en la media?

Dado que $n = 5$, $\mu = 500$, y $\sigma = 2$, los límites de control 3σ son $500 \pm 3(2/\sqrt{5}) = 500 \pm 2.6833$ o (497.3167, 502.5833). En la figura 11.1 se muestra la tabla de control para las medias muestrales. Nótese que la quinta media muestral se encuentra por encima del límite superior de control; de esta forma, durante este tiempo el proceso se juzgó como fuera de control en relación con el promedio. La probabilidad de observar un valor de $\bar{X}$ fuera de los límites de control, si el proceso se encuentra realmente bajo control, es

$$P(|\bar{X} - 500| > 2.6833) = 0.0026.$$

La probabilidad de no detectar un corrimiento de 500 a 503 gramos en la media es

$$\begin{aligned} P(497.3167 < \bar{X} < 502.6833 \mid \mu = 503) &= P\left(\frac{497.3167 - 503}{2/\sqrt{5}} < Z < \frac{502.6833 - 503}{2/\sqrt{5}}\right) \\ &= P(-6.35 < Z < -0.35) \\ &= 0.3632. \end{aligned}$$

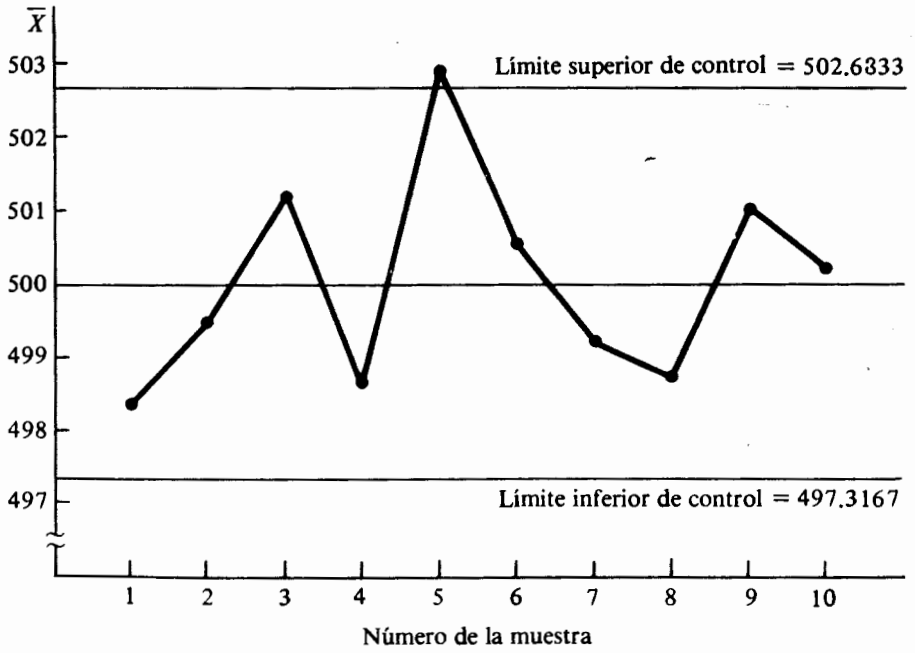


FIGURA 11.1 Tabla $\bar{X}$ para los datos del ejemplo 11.1

11.2.2 Tablas S (desviación estándar conocida de la población)

En muchas ocasiones la variabilidad de un proceso es, por lo menos, tan importante como la media de éste; por ejemplo, en la fabricación de instrumentos de precisión, mantener la variación en las mediciones a un nivel aceptable es, probablemente, tan importante como el promedio.

Se considerarán las tablas de control para la variabilidad de un proceso mediante el empleo de la desviación estándar de la muestra

$$S = \left[\sum (X_i - \bar{X})^2/(n-1)\right]^{1/2}.$$

Los límites de control 3σ son $E(S) \pm 3\, d.e.(S)$. Para obtener $E(S)$ y $Var(S)$, recuérdese de la sección 7.5 que la variable aleatoria

$$Y = \frac{(n-1)S^2}{\sigma^2}$$

tiene una distribución chi-cuadrada con $n-1$ grados de libertad, en donde la función de densidad de probabilidad de Y está dada por (7.16). Dado que

$$S^2 = \frac{\sigma^2 Y}{n-1},$$

entonces

$$S = \frac{\sigma Y^{1/2}}{(n-1)^{1/2}},$$

y

$$E(S) = \frac{\sigma}{(n-1)^{1/2}} E(Y^{1/2}).$$

Pero

$$E(Y^{1/2}) = c\int_0^\infty y^{1/2}\, y^{(n-3)/2} \exp(-y/2)\mathrm{d}y, \tag{11.1}$$

en donde

$$c = \frac{1}{\Gamma[(n-1)/2]\, 2^{(n-1)/2}}.$$

En (11.1) sea $u = y/2$; entonces $dy = 2du$ y

$$E(Y^{1/2}) = 2^{n/2}\, c\int_0^\infty u^{(n-2)/2} \exp(-u)du = 2^{n/2} c\Gamma(n/2).$$

Entonces

$$\begin{aligned} E(S) &= \frac{\sigma}{(n-1)^{1/2}}\, 2^{n/2}\, c\Gamma(n/2) \\ &= \sigma \frac{2^{1/2}\Gamma(n/2)}{(n-1)^{1/2}\Gamma[(n-1)/2]}. \end{aligned} \tag{11.2}$$

Es preferible utilizar una notación para el control de calidad y escribir

$$E(S) = c_4\sigma,$$

en donde

$$c_4 = \frac{2^{1/2}\Gamma(n/2)}{(n-1)^{1/2}\Gamma[(n-1)/2]}. \tag{11.3}$$

Para la varianza de S, por definición

$$Var(S) = E(S^2) - E^2(S).$$

Pero en la sección 7.5 se demostró que $E(S^2) = \sigma^2$, en consecuencia

$$Var(S) = \sigma^2 - c_4^2\sigma^2 = \sigma^2(1 - c_4^2),$$

o en la notación preferible,

$$Var(S) = c_5^2\sigma^2.$$

Por lo tanto, $d.e.(S) = c_5\sigma$, y los límites de control 3σ son

$$c_4\sigma \pm 3c_5\sigma, \tag{11.4}$$

en donde c_4 está dada por (11.3) y $c_5 = (1 - c_4^2)^{1/2}$. Nótese que, dado que se supone que el valor de σ se conoce, los límites de control sólo son funciones del tamaño de cada muestra. En la tabla 11.1 se determinan los valores de c_4 y c_5 para distintos valores usuales del tamaño n de las muestras.

Como ilustración, si $\sigma = 2$, los límites de control 3σ para la desviación estándar muestral, con base en $n = 5$, son $(0.94)(2) \pm (3)(0.3412)(2)$ o (0, 3.9272). Para este ejemplo, en la tabla S el límite inferior de control es cero, la línea central se encuentra en 1.88 y el límite superior de control es 3.9272. Para $n = 5$ y $\sigma = 2$, la variabilidad del proceso se considera bajo control, siempre que el valor de la desviación estándar muestral se encuentre dentro de los límites de control ya establecidos.

11.2.3 Tablas $\bar{X}$ y S (media y varianza desconocidas de la población)

Se considerarán las tablas de control para aquellos casos en los que la distribución de la población es normal, pero no se conocen los valores de la media y la desviación estándar. Para esta situación, los límites de control se basan en los valores estimados para μ y σ.

Dado que no se conoce el valor promedio del proceso, tampoco se conoce la línea central de la tabla de control. Si la línea central es un valor estimado basado en un gran número de muestras, los límites de control que se obtienen de esta manera de-

TABLA 11.1 Valores de c_4 y c_5 para tamaños n normales de la muestra

n	4	5	6	7	8	9	10
c_4	0.9213	0.9400	0.9515	0.9594	0.9650	0.9693	0.9727
c_5	0.3889	0.3412	0.3076	0.2820	0.2622	0.2459	0.2321

ben considerarse sólo como *límites tentativos*, ya que quizá se necesite una modificación antes de que se puedan utilizar para medir la calidad de un producto en futuras operaciones de producción. Lo anterior significa que los límites de control tentativos son apropiados para determinar si las operaciones pasadas de un proceso de producción estuvieron bajo control. Para extenderlos a la producción futura, el procedimiento usual es eliminar todos aquellos puntos que se encuentren fuera de los límites tentativos de control y recalcular el valor de éstos con base en el resto de la información muestral. Se continúa este proceso hasta que todos los puntos se encuentren dentro de los límites de control, tanto para la tabla $\bar{X}$ como para S. La razón para este procedimiento es que los límites de control para la futura producción deben ser funciones de las observaciones que se recabaron mientras el proceso de producción estaba bajo control.

De acuerdo con Shewhart, los límites tentativos de control deben estar basados, por lo menos, en 20 muestras, cada una con cuatro o cinco observaciones. Shewhart denominó a estas muestras *subgrupos racionales*. Éstos deben seleccionarse de manera tal que cada subgrupo sea prácticamente homogéneo y proporcione la máxima oportunidad de variación de un subgrupo a otro. Para un proceso de producción esto implica que las observaciones para un subgrupo deben tomarse en un momento que sea diferente al de otro subgrupo. Se emplea un tamaño relativamente pequeño de la muestra de cuatro o cinco observaciones, no sólo para mantener el balance entre el costo del muestreo y la exactitud del estimado, sino también para dar una mínima oportunidad de variación dentro de cada subgrupo.

Sea m el número de muestras y supóngase que $n_i = n$ para toda $i = 1, 2, \ldots, m$. Además, sean $\bar{X}_i$ y S_i la media y desviación muestral de la i-ésima muestra. Para todas las m muestras, defínanse las estadísticas.

$$\bar{\bar{X}} = \frac{1}{m}\sum_{i=1}^{m} \bar{X}_i \tag{11.5}$$

y

$$\bar{S} = \frac{1}{m}\sum_{i=1}^{m} S_i. \tag{11.6}$$

Es evidente que $E(\bar{\bar{X}}) = \mu$; de esta forma, el promedio de todas las m muestras en un estimador no sesgado de μ. De manera similar,

$$E(\bar{S}) = \frac{1}{m}\sum E(S_i) = \frac{1}{m}(mc_4\sigma) = c_4\sigma,$$

lo cual sugiere que un estimador de σ es $\bar{S}/c_4$. Los límites tentativos 3σ para la media muestral cuando no se conocen los valores de μ y σ son

$$\bar{\bar{X}} \pm 3\frac{\bar{S}}{c_4\sqrt{n}}, \tag{11.7}$$

y los correspondientes a la desviación estándar de muestra son

$$\bar{S} \pm 3\frac{c_5\bar{S}}{c_4}, \tag{11.8}$$

en donde los valores de c_4 y c_5 son los ya definidos.

Ejemplo 11.2 Los datos en la tabla 11.2 son 20 muestras, cada una con cinco observaciones tomadas en intervalos de dos horas, de la resistencia a la tensión en libras de un hilo. Para cada muestra se proporcionan los valores de la media y la desviación estándar. Constrúyanse las tablas de control $\overline{X}$ y S con base en estos datos.

Al promediar las 20 medias muestrales se obtiene $\bar{\bar{x}} = 47.12$, y si se promedian las desviaciones estándar muestrales, se tiene $\bar{s} = 2.326$. Para $n = 5$, $c_4 = 0.94$ y $c_5 = 0.3412$. Entonces, por (11.7) y (11.8), los límites tentativos de control 3σ para las medias muestrales son

$$47.12 \pm \frac{(3)(2.326)}{(0.94)\sqrt{5}} = (43.80, 50.44),$$

y los límites para las desviaciones estándar muestrales son

$$2.326 \pm \frac{(3)(0.3412)(2.326)}{0.94} = (0, 4.8589).$$

En la figura 11.2 se proporcionan las tablas de control. Nótese que la variabilidad del proceso parece estar bajo control, pero la media muestral para la vigésima muestra se encuentra fuera de los límites tentativos. Debido a lo anterior, se obtienen nuevos valores para los límites después de omitir esta muestra. Éstos son

$$47.31 \pm \frac{(3)(2.368)}{(0.94)\sqrt{5}} = (43.93, 50.69)$$

TABLA 11.2 Datos de la muestra de la resistencia a la tensión de un hilo en libras

Número de la muestra	*Valores de la muestra*					$\overline{X}$	S
1	44	46	48	52	49	47.8	3.03
2	44	47	49	46	44	46.0	2.12
3	47	49	47	43	44	46.0	2.45
4	45	47	51	46	48	47.4	2.30
5	44	41	50	46	50	46.2	3.90
6	49	46	45	46	49	47.0	1.87
7	47	48	50	46	47	47.6	1.52
8	49	46	51	48	46	48.0	2.12
9	47	42	48	44	46	45.4	2.41
10	46	48	45	51	50	48.0	2.55
11	45	47	51	48	46	47.4	2.30
12	52	51	48	48	45	48.8	2.77
13	45	45	47	49	44	46.0	2.00
14	46	47	43	48	45	45.8	1.92
15	48	49	52	46	51	49.2	2.39
16	44	46	45	47	52	46.8	3.11
17	48	50	47	46	49	48.0	1.58
18	48	52	51	47	46	48.8	2.59
19	47	51	50	46	49	48.6	2.07
20	44	43	42	43	46	43.6	1.52

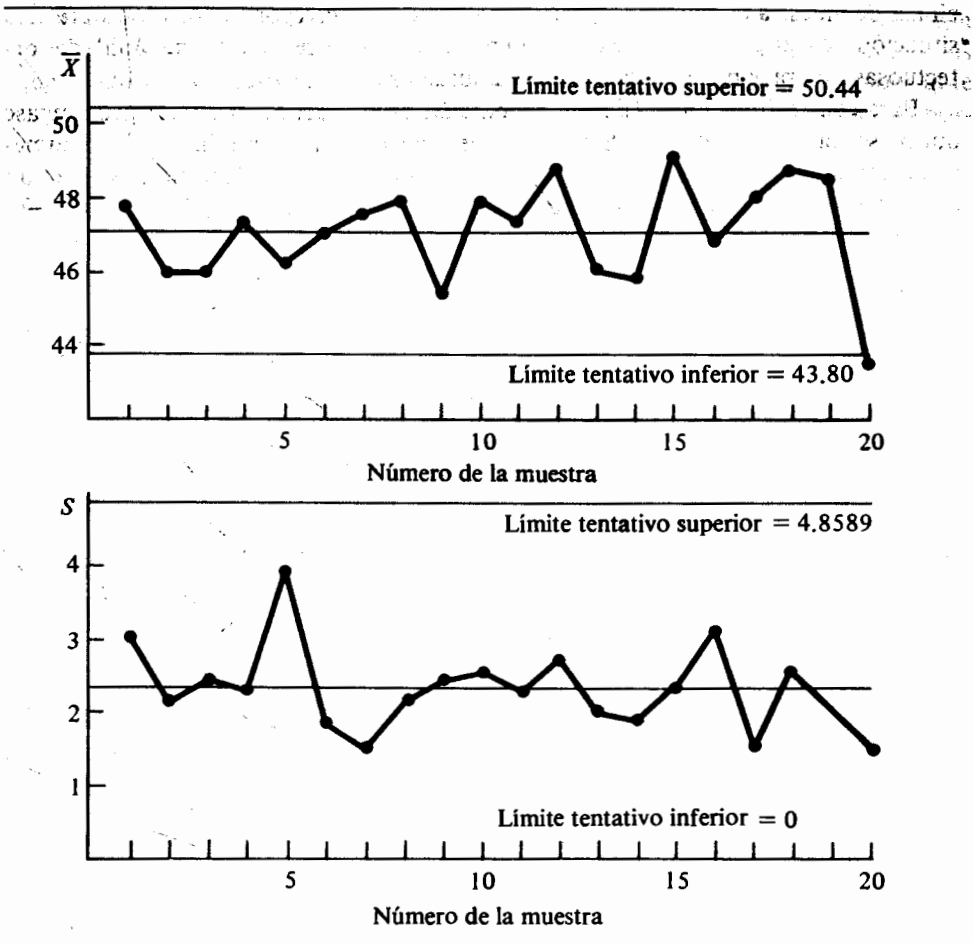


FIGURA 11.2 Tablas $\bar{X}$ y S para los datos del ejemplo 11.2

para $\bar{X}$ y los límites

$$2.368 \pm \frac{(3)(0.3412)(2.368)}{0.94} = (0 \quad 4.9466)$$

para S. Se observa que todos los puntos se encuentran dentro de los nuevos límites tentativos, tanto en la tabla $\bar{X}$, como en la S.

La construcción de las tablas $\bar{X}$ y S se basa en la distribución normal. La tabla $\bar{X}$ es, relativamente, insensible a la hipótesis de normalidad debido al teorema del límite central. Sin embargo, la tabla S es mucho más sensible a la hipótesis de normalidad.

Vale la pena mencionar la existencia de la tabla p. La tabla p puede construirse cuando se supone que el muestreo se lleva a cabo sobre una distribución binomial con parámetro de proporción p. Los límites de control se obtienen para las propor-

ciones de muestra de unidades que caen en una de dos categorías posibles. Para esta situación, lo que generalmente es de interés, es vigilar la proporción de unidades defectuosas que produce un proceso de manufactura.

Para construir los límites de control para las proporciones muestrales, supóngase que no se conoce el valor de p. Sea m el número de muestras disponible, y X_i el número de unidades defectuosas en la i-ésima muestra de tamaño n. Entonces X_i/n es un estimador de p basado en la i-ésima muestra, y $\overline{P} = (1/mn) \sum_{i=1}^{m} X_i$ es un estimador de p basado en todas las m muestras. De acuerdo con lo anterior, los límites tentativos 3σ para las proporciones muestrales X_i/n son

$$\overline{P} \pm 3\sqrt{\frac{\overline{P}(1 - \overline{P})}{n}}. \tag{11.9}$$

11.3 Procedimientos del muestreo para aceptación

Un consumidor puede escoger uno de los tres caminos siguientes para verificar la calidad de los artículos de un embarque que ha recibido: inspeccionar todos los artículos en el lote; inspeccionarlos en una muestra aleatoria proveniente del lote, o aceptar el lote sin llevar a cabo ninguna inspección. La primera opción tiene generalmente un precio prohibitivo y la última es poco probable que sea aceptada por un consumidor serio, con respecto a la calidad de los artículos que adquiere. Por lo tanto, la opción que tiene un balance adecuado entre el costo de la inspección y el que implica aceptar un lote y usar artículos defectuosos, es la de inspeccionar los artículos en una muestra aleatoria proveniente del lote que se acaba de adquirir. Con base en el proceso de inspección, la decisión usual es aceptar el lote, rechazarlo o tomar otra muestra aleatoria. Si la decisión de aceptar o rechazar se toma con base en los valores medidos de los artículos, con respecto a una medición física continua, entonces se dice que la inspección se lleva a cabo *por variables*. Si los artículos que se inspeccionan se clasifican como defectuosos o no defectuosos, y el lote se acepta o rechaza con base en el número de artículos defectuosos en la muestra, se dice que la inspección se lleva a cabo *por características*.

En esta sección se considerarán los fundamentos para desarrollar planes sencillos de muestreo con base en características para decidir si se acepta o se rechaza un lote. Posteriormente se examinará en forma breve el muestreo para aceptación por variables. Sea N el tamaño del lote. Entonces un plan básico de muestreo para aceptación es seleccionar n artículos del lote de tamaño N y aceptar el lote si el número de artículos defectuosos en la muestra es menor o igual a un número de aceptación c, previamente estipulado. De otra forma, el lote se rechaza. Por ejemplo, un plan de muestreo puede definirse de la siguiente forma $N = 10\,000$, $n = 100$, y $c = 1$. Lo anterior significa que se seleccionarán, en forma aleatoria, 100 artículos de los 10 000 que contiene el lote, y si se encuentra cuando mucho un artículo defectuoso, se aceptará el lote de $N = 10\,000$ artículos. Si hay más de un artículo defectuoso, el lote será rechazado. El consumidor puede escoger entre regresar el lote rechazado al fabricante o someterlo a una inspección del 100%. El primero constituye lo que se conoce como un procedimiento de *inspección no verificable*, y el segundo como proceso de *inspección verificable*.

Supóngase que la información disponible para el consumidor con respecto a la calidad de los artículos en el lote, es la proporción promedio de artículos defectuosos que produce el proceso de manufactura que los fabrica. Un criterio muy importante en un plan de muestreo es la probabilidad de aceptar el lote *P(A)*, dada una proporción de artículos defectuosos p. Bajo las hipótesis adecuadas y para algún valor de p y de c, la probabilidad de que el lote sea aceptado con base en una muestra de tamaño n, es la probabilidad binomial acumulativa

$$P(A) \equiv P(X \leq c) = \sum_{x=0}^{c} \binom{n}{x} p^x (1-p)^{n-x}, \tag{11.10}$$

en donde la variable aleatoria X representa el número de artículos defectuosos encontrados en la muestra. Si np tiene un tamaño moderado, la probabilidad binomial dada por (11.10) se puede aproximar en forma adecuada por la probabilidad acumulativa de Poisson

$$P(A) = \sum_{x=0}^{c} \frac{\lambda^x}{x!} \exp(-\lambda), \tag{11.11}$$

en donde $\lambda = np$.

Una gráfica de la probabilidad de aceptación contra p, es la curva de operación característica (CO). Como ilustración se analizará el plan de muestreo $n = 100$ y $c = 2$. Mediante el empleo de la aproximación de Poisson dada por (11.11) se obtiene la probabilidad de aceptar para valores de p en un intervalo de 0.01 a 0.09. Las probabilidades de aceptación se dan en la tabla 11.3 y están graficadas contra p en la figura 11.3.

La naturaleza de una curva *CO* es afectada por el tamaño n de la muestra y por el número de aceptación c. Como ilustración, considérense los planes de muestreo $n = 50$, $c = 1$; $n = 100$, $c = 2$; y $n = 200$, $c = 4$. En la figura 11.4 se muestran las curvas *CO* para estos planes. Nótese que aunque el cociente c a n es constante, las curvas *CO* son algo diferentes. De hecho, las curvas son más sensibles al tamaño de la muestra. Conforme n aumenta, la pendiente de la curva se torna más pronunciada. De esta forma, para tamaños grandes de la muestra, la probabilidad de aceptación disminuye muy rápidamente conforme el valor de p aumenta. Si el valor de n es fijo, un aumento en el número de aceptación c tenderá a desplazar a la curva hacia la derecha. Esto implica que para una p dada, la probabilidad de aceptación es alta conforme c aumenta. En consecuencia, puede pensarse que entre más cercano a cero se encuentre el valor de c, mejor es el plan de muestreo. Pero la figura 11.4 indica que los planes con valores grandes de c son mejores siempre que el tamaño de la muestra sea, apreciablemente, grande.

TABLA 11.3 Probabilidades de aceptación para el plan de muestreo $n = 100$, $c = 2$

p	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$P(A)$	0.9197	0.6767	0.4232	0.2381	0.1247	0.0620	0.0296	0.0138	0.0062

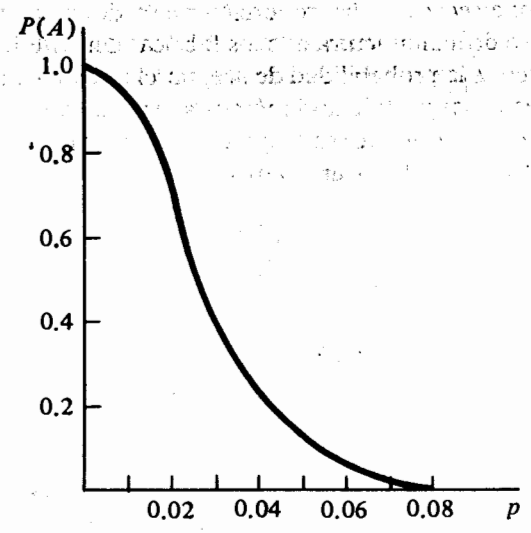


FIGURA 11.3 Curva característica de operación para el plan de muestreo $n = 100$, $c = 2$

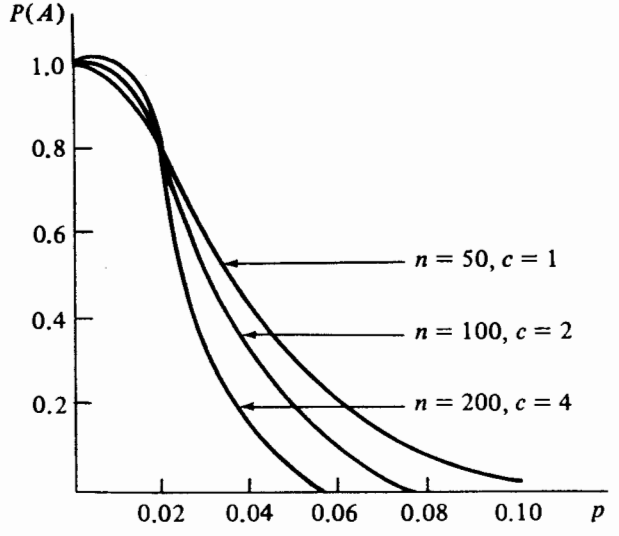


FIGURA 11.4 Curvas características de operación para los tres planes de muestreo

El desarrollo de buenos planes de muestreo incluye tanto al productor como al comprador del lote. De manera normal el productor es el vendedor y el consumidor el comprador. Un productor ciertamente desearía que el consumidor rechazara un porcentaje muy pequeño de los lotes vendidos y que son, en general, buenos; el consumidor desearía aceptar un porcentaje muy pequeño de los lotes que son malos. De esta forma los dos experimentan cierto riesgo. Supóngase que ambos están de acuerdo en que un lote es aceptable si la proporción de artículos defectuosos es $p \leq p_1$, y no aceptable si $p \geq p_2$. Se dan las siguientes definiciones que implican riesgos.

Definición 11.1 El riesgo del productor α es la probabilidad de que el consumidor rechace un lote cuya proporción de artículos defectuosos no es mayor que p_1.

Definición 11.2 El riesgo del consumidor β es la probabilidad de aceptar un lote cuya proporción de artículos defectuosos es mayor o igual a p_2.

Con base en estas definiciones, el riesgo del productor es la probabilidad del error de tipo I, dado que éste representa la probabilidad de rechazar un lote aceptable. De manera similar, el riesgo del consumidor es la probabilidad del error de tipo II, ya que éste representa la probabilidad de equivocarse al no rechazar un lote inaceptable. En otras palabras, la situación anterior es análoga a probar la hipótesis nula H_0: $p = p_1$ contra la alternativa H_1: $p = p_2$.

Los riesgos del productor y del consumidor pueden representarse por dos puntos sobre una curva característica de operación, como se ilustra en la figura 11.5. En

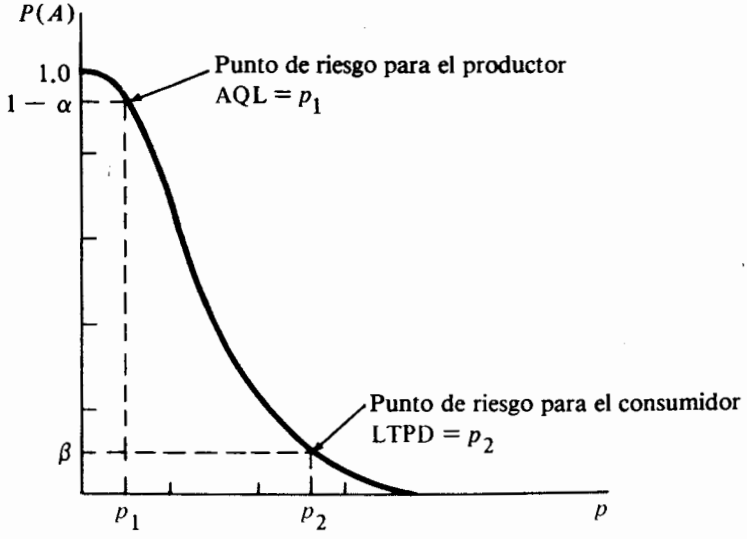


FIGURA 11.5 Curva CO para los puntos de riesgo especificados para el productor y el consumidor

este contexto, p_1 recibe el nombre de *nivel aceptable de calidad* (NAC), y p_2 el de *tolerancia de la proporción de defectuosos en el lote* (TPDL). La práctica usual ha sido la de escoger la probabilidad de aceptación $P(A) = 1 - \alpha$ en NAC cercano al punto 0.95 de la curva, y la probabilidad de aceptación $P(A) = \beta$ en TPDL cercano al punto 0.10 sobre la curva. Entonces, el 95% de los lotes que provienen de un proceso cuya proporción de artículos defectuosos se encuentra en NAC, o por encima de éste, se aceptará, mientras que sólo el 10% de los que provienen de un proceso cuya proporción de artículos defectuosos se encuentra en TPDL o más, será aceptada.

11.3.1 El desarrollo de planes de muestreo sencillos para riesgos estipulados del productor y del consumidor

Se examinará un procedimiento para obtener planes de muestreo sencillos para valores especificados de los riesgos del productor y del consumidor. La esencia del procedimiento está en determinar el tamaño de la muestra n y el número de aceptación c, dadas las probabilidades de aceptación en el NAC y el TPDL. Por ejemplo, supóngase que se desea un plan sencillo de muestreo para el que la curva característica de operación pasa a través de un riesgo del productor $\alpha = 0.05$ en un NAC de 0.01, y de un riesgo del consumidor $\beta = 0.1$ en un TPDL de 0.05. De esta forma, las probabilidades de aceptación al NAC = 0.01 y TPDL = 0.05 son 0.95 y 0.1, respectivamente.

Supóngase que las condiciones son tales, que la distribución de Poisson proporcionará una aproximación adecuada. Sea X la variable aleatoria que representa el número de artículos defectuosos en una muestra de tamaño n. Entonces para el riesgo del productor, se desea obtener n y c, tales que

$$P(A) \equiv P(X \leq c) = \sum_{x=0}^{c} \frac{\lambda^x \exp(-\lambda)}{x!} = 1 - \alpha, \tag{11.12}$$

en donde $\lambda = np_1$. De manera similar, para el riesgo del consumidor, se desea obtener n y c, tales que

$$P(A) \equiv P(X \leq c) = \sum_{x=0}^{c} \frac{\lambda^x \exp(-\lambda)}{x!} = \beta, \tag{11.13}$$

en donde ahora $\lambda = np_2$. Dado que se conocen los valores de α, β, p_1 y p_2, el procesamiento se reduce a la solución simultánea de (11.12) y (11.13) para n y c. No existe ningún método directo para resolver estas dos ecuaciones; en otras palabras, es virtualmente imposible determinar un plan de muestreo cuya curva CO pasa en forma exacta a través de dos puntos $(p_1, 1 - \alpha)$ y (p_2, β) debido a que los valores de n y c deben ser números enteros. Lo que se hace en forma general, es obtener cuatro planes, dos de los cuales tendrán el valor dado de α pero diferirán muy poco para el valor de β, mientras que los otros dos tendrán el valor de β dado, pero diferirán muy poco del valor de α.

Dados $\alpha = 0.05$, $\beta = 0.1$, $p_1 = 0.01$, y $p_2 = 0.05$, el procedimiento es el siguiente: sea $\lambda_1 = np_1$ y $\lambda_2 = np_2$ y fórmese el cociente de λ_2 a λ_1. Para el ejem-

plo se observa que el valor de éste es 5. En forma ideal, lo que se busca es obtener el valor de c cuando λ_2/λ_1 es exactamente 5. Dado que no es probable tener este valor de manera precisa, lo que se desea es determinar los dos valores de c que se encuentran relacionados con el valor de 5. Lo anterior puede lograrse si se inicia con $c = 0$ y se interpola, para encontrar valores λ_1, tales que $P(A) = 1 - \alpha$, y para λ_2, tales que $P(A) = \beta$, mediante el empleo de la distribución acumulativa de Poisson (tabla B del apéndice). Entonces se aumenta el valor de c, y se continúa el proceso hasta que se encuentren los valores de c que estén relacionados con el cociente deseado. Los tamaños correspondientes de las muestras se obtienen, primero, al fijar la probabilidad de aceptación del riesgo del productor dado, y después al hacer lo mismo para el riesgo del consumidor, este procedimiento dará como resultado cuatro planes de muestreo diferentes.

Dado que $P(A) = 0.95$ y $c = 0$, se obtiene que $\lambda_1 = 0.05$. De manera similar, para $P(A) = 0.1$ y $c = 0$, λ_2 tiene un valor de 2.30, y para el cociente $\lambda_2/\lambda_1 = 46$. Ahora, para $P(A) = 0.95$ y $c = 1$, $\lambda_1 = 0.36$, y para $P(A) = 0.10$, $\lambda_2 = 3.9$. De esta forma $\lambda_2/\lambda_1 = 10.83$. El proceso continúa y se obtienen los resultados que se muestran en la tabla 11.4. Los dos valores de c que se relacionan con el cociente ideal de 5 son 2 y 3.

Para obtener n, supóngase que se mantiene el riesgo del productor en $\alpha = 0.05$. Entonces para $c = 2$, $np_1 = 0.82$; pero $p_1 = 0.01$ y $n = 82$. Para el plan $n = 82$ y $c = 2$, la probabilidad de aceptar a un nivel TPDL $= 0.05$ se obtiene mediante $\lambda_2 = (82)(0.05) = 4.1$. De cuerdo con lo anterior $P(A) \equiv P(X \leq 2) = 0.2238$.

Si se fija el riesgo del consumidor en $\beta = 0.1$, entonces para $c = 2$, $np_2 = 5.32$, y $n = 107$. Como resultado se tiene que $\lambda_1 = (107)(0.01) = 1.07$, y la probabilidad de aceptar en un NAC $= 0.01$ es $P(A) \equiv P(X \leq 2) \approx 0.91$. Se pueden establecer los otros dos planes si se repite el proceso anterior con $c = 3$. En la tabla 11.5 se resumen los cuatro planes; de éstos, el que parece tener la menor importancia con respecto al riesgo especificado del consumidor es $n = 82$ y $c = 2$. Los otros tres, en especial los últimos dos, se encuentran cercanos a los riesgos especificados, tanto del productor como del consumidor. La decisión final sobre cuál adoptar se toma con base en las circunstancias de la situación.

11.3.2 Muestreo para aceptación por variables

La mayoría de los planes de muestreo para aceptación se llevan a cabo por características, debido a dos razones fundamentales: la inspección por características es

TABLA 11.4 Determinación de los valores de c que se encuentran relacionados con $\lambda_2/\lambda_1 = 5$.

Número de aceptación c	*Valor de* $\lambda_1 = np_1$ *para* $P(A) = 0.95$	*Valor de* $\lambda_2 = np_2$ *para* $P(A) = 0.1$	λ_2/λ_1
0	0.05	2.30	46.00
1	0.36	3.90	10.83
2	0.82	5.32	6.49
3	1.37	6.68	4.88

TABLA 11.5 Cuatro planes de muestro para $\alpha = 0.05$, $\beta = 0.1$, NAC = 0.01, y TPDL = 0.05.

Plan de muestreo	*Probabilidad de aceptación para* NAC = 0.01	*Probabilidad de aceptación para* TPDL = 0.05
$n = 82$, $c = 2$	0.95	0.2238
$n = 107$, $c = 2$	0.91	0.10
$n = 137$, $c = 3$	0.95	0.09
$n = 134$, $c = 3$	0.95	0.10

muy económica y muchas de las características de calidad sólo son observables como atributos. Sin embargo, en algunos casos puede hacerse una medición física de la calidad de un producto dado. Cuando la aceptación se hace con base en mediciones físicas se dice que el muestreo se lleva a cabo por variables. Cuando éste es posible, se convierte en el tipo de muestreo más popular, ya que una medición física es probable que proporcione mucho más información útil con respecto a la calidad de un producto que la dada por característica. Además, pueden obtenerse curvas CO más pronunciadas para el mismo tamaño de la muestra. La inspección por variables en general es más costosa que la inspección por características, debido a que, principalmente, tiene que aplicarse el criterio de aceptación por separado para cada medición de calidad cuando se muestrea por variables.

En el caso sencillo en el que la aceptación de un lote se hace con base en las medias de la muestra, se supone que la medición de la calidad es una variable aleatoria normalmente distribuida y con varianza conocida. Sean α el riesgo del productor y μ_α el promedio del lote para el que la probabilidad de aceptación es $1 - \alpha$. En forma similar, sea β el riesgo del consumidor y μ_β el promedio del lote para el cual la probabilidad de aceptación es β. Es decir, si el lote tiene una media μ_α, se desea aceptar el lote con una probabilidad $1 - \alpha$, y si éste tiene una media μ_β ($\mu_\alpha > \mu_\beta$) se desea aceptar el lote con una probabilidad β. Dados α, β, μ_α, y μ_β, el plan de muestreo por variables es una muestra de tamaño n y un valor de aceptación $\bar{x}_a$, tales que, cuando el valor observado de la media de la muestra $\bar{X}$ es mayor que $\bar{x}_a$, el lote será aceptado.

Para obtener $\bar{x}_a$ y n, considérese lo siguiente. Para el riesgo del productor

$$P(\bar{X} \leq \bar{x}_a) = \alpha$$

o

$$P\left(Z \leq \frac{\bar{x}_a - \mu_\alpha}{\sigma/\sqrt{n}}\right) = \alpha,$$

en donde

$$\frac{\bar{x}_a - \mu_\alpha}{\sigma/\sqrt{n}} = z_\alpha. \tag{11.14}$$

Para el riesgo del consumidor

$$P(\bar{X} > \bar{x}_a) = \beta$$

o

$$P\left(Z > \frac{\bar{x}_a - \mu_\beta}{\sigma/\sqrt{n}}\right) = \beta,$$

en donde

$$\frac{\bar{x}_a - \mu_\beta}{\sigma/\sqrt{n}} = z_{1-\beta}. \tag{11.15}$$

Las ecuaciones dadas por (11.14) y (11.15) dependen de las incógnitas $\bar{x}_a$ y n. Al resolver (11.14) y (11.15) para $\bar{x}_a$, se tiene

$$\bar{x}_a = \frac{\sigma}{\sqrt{n}} z_\alpha + \mu_\alpha \tag{11.16}$$

y

$$\bar{x}_a = \frac{\sigma}{\sqrt{n}} z_{1-\beta} + \mu_\beta. \tag{11.17}$$

Al igualar (11.16) y (11.17) y resolver para n, se tiene

$$n = \left[\frac{\sigma(z_{1-\beta} - z_\alpha)}{\mu_\alpha - \mu_\beta}\right]^2 \tag{11.18}$$

Cuando se emplea (11.18) para obtener el tamaño de la muestra, el valor de aceptación $\bar{x}_a$ se obtiene, ya sea de (11.16) o de (11.17).

Ejemplo 11.3 La compañía constructora de un gran edificio de oficinas se interesa en la resistencia a la compresión del concreto que se empleará en la construcción del edificio. El proceso a través del cual se fabrica el concreto con una resistencia promedio de 350 kilogramos por centímetro cuadrado es bueno. El concreto adquirido en este proceso debe aceptarse el 95% de las veces. Un proceso que ofrece una resistencia promedio de 347 kilogramos por centímetro cuadrado no es efectivo, y al ser adquirido será rechazado el 90% de las veces. Si el fabricante de cemento asegura a la compañía que la desviación estándar de su proceso no es mayor de 5 kilogramos por centímetro cuadrado, ¿cuántas muestras de concreto debe inspeccionar el contratista con respecto a su resistencia, y cuál debe ser el valor de aceptación para la media de la muestra bajo las condiciones dadas? Supóngase que la resistencia a la compresión del concreto se encuentra normalmente distribuida.

Los riesgos del productor y del consumidor están dados como $\alpha = 0.05$ para $\mu_\alpha = 350$ y $\beta = 0.10$ para $\mu_\beta = 347$, respectivamente. Para $\alpha = 0.05$ y $1 - \beta = 0.9$, los valores cuantiles normales estandarizados correspondientes son $z_{0.05} = -1.645$ y $z_{0.9} = 1.282$. Entonces, mediante el empleo de (11.18), el tamaño necesario de la muestra es

$$n = \left[\frac{5(1.282 + 1.645)}{350 - 347}\right]^2 = 24.$$

Para el riesgo del productor (11.16)

$$\bar{x}_a = \frac{5}{\sqrt{24}}(-1.645) + 350 = 348.32,$$

y para el del consumidor (11.17)

$$\bar{x}_a = \frac{5}{\sqrt{24}}(1.282) + 347 = 348.31.$$

Para $\bar{x}_a = 348.32$, el plan de muestreo consiste en probar la resistencia de 24 muestras de concreto provenientes del proceso v aceptar el concreto siempre que la resistencia promedio sea mayor de 348.32 kilogramos por centímetro cuadrado.

11.3.3 Sistemas de planes de muestreo

Desde la Segunda Guerra Mundial, los planes de muestreo para aceptación se han convertido en procedimientos estándar para asegurar la calidad de los productos manufacturados y con este propósito se ha desarrollado una gran variedad de sistemas de planes de muestreo para aceptación. Tres de los sistemas más empleados son MIL-STD-105D*, MIL-STD-414, y el Dodge-Romig Sampling Inspection Tables. En las referencias [4], [5] y [1] se encuentra información detallada de estos sistemas. Los primeros dos fueron desarrollados por el Departamento de la Defensa y se aplican bajo un procedimiento de inspección no verificable. MIL-STD-105D contiene planes para el muestreo por características y MIL- STD-414 para el muestreo por variables. Los planes de muestreo Dodge-Romig se basan en un programa de inspección con verificación; éstos suponen un porcentaje de unidades defectuosas del proceso conocido, y los planes de muestreo sencillos se encuentran *indexados* por TPDL para riesgo del consumidor de 0.10. Estos tres sistemas se encuentran descritos en [3].

Referencias

1. H. F. Dodge and H. G. Romig, *Sampling inspection tables — Single and double sampling*, 2nd ed. Wiley, New York, 1959.
2. A. J. Duncan, *Quality control and industrial statistics*, 4th ed., Richard D. Irwin, Homewood, Ill., 1974.
3. E. L. Grant and R. S. Leavenworth, *Statistical quality control*, 4th ed., McGraw-Hill, New York, 1972.
4. *Military standard 105D, Sampling procedures and tables for inspection by attributes*, Superintendent of Documents, Government Printing Office, Washington, D.C., 1963.

* Fuera de Estados Unidos el sistema se conoce como ABC-STD-105D.

5. *Military standard 414, Sampling procedures and tables for inspection by variables for percent defective*, Superintendent of Documents, Government Printing Office, Washington, D.C., 1957.

Ejercicios

11.1. El consejo estatal formado para controlar la calidad del agua selecciona cada semana cinco muestras de agua de una fuente de abastecimiento y determina la concentración promedio de una sustancia tóxica. Los siguientes datos son las cantidades promedio en partes por millón durante 12 semanas.

Semana	1	2	3	4	5	6	7	8	9	10	11	12
Media de la muestra	5.2	4.9	5.5	5.4	4.8	4.6	5.5	4.7	5.1	4.5	5.8	5.6

a) Si los valores de la concentración promedio y de la desviación estándar son 5 y 0.5 ppm, respectivamente, obténganse los límites de control 3σ para la concentración promedio. Para este periodo, ¿existió alguna razón para alarmarse?

b) Si se considera como peligrosa una concentración de 6 ppm, ¿que tan probable es tener un resultado como el anterior, con base en cinco muestras de agua, si la concentración real promedio es de 5 ppm?

c) Mediante el uso de los límites de control de la parte *a*, ¿cuál es la probabilidad de detectar un desplazamiento en el valor de la concentración media de 5 ppm a 5.25 ppm?

11.2. Mediante el empleo de la información proporcionada en el ejercicio 11.1, obténganse los límites de control 3σ para la desviación estándar de la muestra.

11.3. Los siguientes datos son las tensiones de ruptura promedio de seis muestras de metal tomadas en forma periódica:

Muestra	1	2	3	4	5	6	7	8	9	10
Media de la muestra	498.6	508.3	484.6	505.7	491.7	495.4	482.6	515.2	510.8	503.7

Se sabe que los valores de la tensión de ruptura promedio y de la desviación estándar son 500 y 20 libras, respectivamente.

a) Obténganse los límites de control 3σ para la tensión de ruptura media de la muestra y hágase una gráfica de la tabla de control. ¿Existe alguna media muestral que se encuentre fuera de los límites de control?

b) Obténgase la probabilidad de no detectar un corrimiento en el valor real de la tensión de ruptura promedio de 500 a 494 libras.

c) Obténganse los límites de control 3σ para la desviación estándar muestral.

11.4. Los datos que se encuentran en la tabla 11.6 consisten en 20 muestras, cada una con cuatro observaciones, de los diámetros de cojinetes producidos por un proceso de manufactura.

a) Constrúyanse los límites tentativos 3σ para las tablas de control $\bar{X}$ y S.

b) Si se detecta que el proceso no se encuentra bajo control, con base en alguna muestra, recalcúlense los límites tentativos.

TABLA 11.6 Datos de la muestra para el ejercicio 11.4

Número de la muestra	*Valores de la muestra (en centímetros)*			
1	4.01	4.03	3.98	4.04
2	3.97	3.99	3.99	4.02
3	4.06	4.05	3.97	4.02
4	3.96	3.98	4.07	4.03
5	3.98	3.99	3.99	4.00
6	4.01	4.02	3.96	3.99
7	3.95	3.98	4.02	4.03
8	4.03	4.00	3.96	4.04
9	4.07	3.96	3.98	4.05
10	3.98	3.97	4.02	4.04
11	3.92	4.03	4.05	3.99
12	3.97	4.05	4.04	4.01
13	4.04	4.04	3.96	3.99
14	4.03	4.00	4.02	4.05
15	3.95	3.96	3.95	4.02
16	4.05	4.09	4.07	4.02
17	3.98	4.06	4.04	4.03
18	4.01	4.02	4.00	3.97
19	4.02	4.01	4.05	3.99
20	3.99	3.99	4.01	4.00

11.5. Las tablas de control $\bar{X}$ y S de un proceso de llenado de recipientes se conservan por algún tiempo. Con base en 25 muestras periódicas, cada una con cinco recipientes, se obtiene que $\bar{\bar{X}} = 400.2$ g y $\bar{S} = 15.3$ g.

a) Si se supone que el proceso de llenado se encuentra bajo control ¿cuáles son los límites de control de la media y la desviación estándar muestral?

b) Obténgase un estimado de la desviación estándar del proceso.

11.6. En el ejercicio 11.5, supóngase que cada muestra contenía seis recipientes. ¿Cómo puede afectar este cambio a las respuestas de las partes *a* y *b*?

11.7. En un proceso de manufactura, cada día se seleccionan al azar 100 unidades y se envían para su inspección. Los siguientes datos son el número de unidades defectuosas en la muestra durante 25 días.

Día	1	2	3	4	5	6	7	8	9	10	11	12	13
Número de unidades defectuosas	2	1	4	3	2	2	5	3	4	2	1	5	2
Día	14	15	16	17	18	19	20	21	22	23	24	25	
Número de unidades defectuosas	3	2	1	0	6	4	5	2	1	8	3	2	

a) Con base en esta información, obténgase una tabla *p*.

b) Revísense los límites de control si algún día el proceso se juzgó como fuera de control.

c) Si se supone que el proceso se encuentra bajo control con un porcentaje de unidades defectuosas, igual al obtenido en la parte *b*, ¿cuál es la probabilidad de que, en un día determinado el proceso se considere como fuera de control?

11.8. Se supone que el porcentaje de unidades defectuosas para un proceso de manufactura es de 4%. El proceso se vigila diariamente mediante la toma de muestras de $n = 80$ unidades. Éste se detiene cada vez que se encuentran cinco o más unidades defectuosas en la muestra. Si el verdadero porcentaje de unidades defectuosas es de 5.5%, ¿cuál es la probabilidad de detener el proceso?

11.9. Supóngase que la calidad de un lote muy grande es de sólo 5% de unidades defectuosas. Un plan de muestreo para aceptación requiere una muestra de 40 unidades y un número de aceptación igual a 2 unidades.

a) ¿Cuál es la probabilidad de que el lote sea aceptado?
b) Si la calidad real del lote es de 6.25% de unidades defectuosas, ¿cuál es la probabilidad de que el lote sea aceptado?

11.10. Para el ejercicio 11.9, supóngase que el tamaño de la muestra es de $n = 80$ unidades y el número de aceptación es igual a cuatro unidades. ¿Cómo afectarán estos cambios a las respuestas de las partes *a* y *b*?

11.11. La calidad de un lote de $N = 20$ unidades es del 10% defectuosas. Si se toma una muestra aleatoria de cinco unidades y no se encuentra ninguna defectuosa se aceptará el lote. ¿Cuál es la probabilidad de aceptar el lote?

11.12. Hágase una gráfica de las curvas características de operación para los planes de muestreo $n = 25$, $c = 1$ y $n = 50$, $c = 2$. Compárense las curvas características de operación.

11.13. Para el plan de muestreo $n = 25$, $c = 1$, empléese la curva CO para obtener el TPDL para un riesgo del consumidor de 0.05.

11.14. Para el plan de muestreo $n = 50$, $c = 2$, empléese la curva CO para obtener el NAC para un riesgo del productor de 0.05.

11.15. Obténganse los cuatro planes de muestreo que relacionarán los riesgos del productor y del consumidor de $\alpha = 0.05$ para NAC = 0.02 y $\beta = 0.1$ para TPDL = 0.08, respectivamente.

11.16. Obténganse los cuatro planes de muestreo que relacionarán los riesgos del productor y del consumidor de $\alpha = 0.10$ para NAC = 0.01 y $\beta = 0.1$ para TPDL = 0.05.

11.17. En muchas ocasiones se emplea un plan de muestreo doble para el muestreo de aceptación; este plan requiere una muestra aleatoria de n_1 unidades de un lote de N unidades. Si el número de unidades defectuosas no es mayor que c_1, el lote se acepta; si se encuentra una cantidad de unidades defectuosas $c_2 > c_1$ el lote se rechaza. Si el número de unidades defectuosas en la primera muestra es mayor que c_1, pero menor que c_2, se toma otra muestra aleatoria de tamaño n_2. El lote se acepto si el número de unidades defectuosas en ambas muestras no es mayor que c_2; de otra forma el lote se rechaza. Mediante el empleo de este procedimiento determínense las siguientes probabilidades para el doble plan de muestreo $N = 5000, n_1 = 50, n_2 = 80, c_1 = 0, c_2 = 3$ si la calidad del lote es de 2% de unidades defectuosas.

a) La probabilidad de aceptar el lote con base en la primera muestra.

b) La probabilidad de rechazar el lote con base en la primera muestra.
c) La probabilidad de aceptar el lote después de tomar la segunda muestra.
d) La probabilidad de rechazar el lote después de tomar la segunda muestra.

11.18. Una agencia estatal se encarga de vigilar el nivel de concentración de cierto contaminante químico, el cual ha sido derramado en grandes cantidades en uno de los ríos más grandes del estado. La agencia debe decidir en forma periódica cuándo el nivel de concentración se encuentra entre límites seguros para permitir la pesca con fines comerciales. La agencia desea obtener un plan de muestreo por variables de tal manera que cuando el nivel de concentración promedio real sea de 5.6 ppm decidirá el 95% de las veces que la pesca continúe. Pero desea prohibir la pesca el 99% de las veces que se observe una concentración hasta de 6.0 ppm. Si la desviación estándar no es mayor de una parte por millón, determínese el plan de muestreo. Supóngase que la concentración de este contaminante se encuentra normalmente distribuida.

11.19. Un comprador de grandes cantidades de hilo desea desarrollar un plan de muestreo por variables para la tensión de ruptura del hilo. El hilo será aceptado por el comprador si su tensión de ruptura es mayor de 60 libras. Si se sabe que la desviación estándar del hilo es de 8 libras y dados $\alpha = 0.05$, $\beta = 0.05$, NAC = 0.05 y TPDL = 0.1, obténgase el plan de muestreo. Supóngase que la tensión del hilo se encuentra normalmente distribuida.

CAPÍTULO DOCE

Diseño y análisis de experimentos estadísticos

12.1 Introducción

En las secciones 9.6.3 y 9.6.4 se introdujeron algunas ideas básicas con respecto a la planeación y adquisición de datos experimentales, con el propósito de alcanzar el máximo beneficio de la aplicación de la inferencia estadística. En este capítulo se estudiará la noción de experimentos diseñados estadísticamente y se extenderán algunos de los métodos del capítulo 9 mediante la introducción de una técnica estadística importante conocida como análisis de varianza.

12.2 Experimentos estadísticos

Para cualquier fenómeno en el que existe la incertidumbre, el procedimiento apropiado para investigarlo es experimentar con él, de manera que puedan identificarse las características de interés. Por ejemplo, supóngase que se desea identificar el comportamiento óptimo de un sistema con respecto a su funcionamiento y costo en distintas condiciones; entonces debe pensarse en un experimento como medio para que el sistema sea observado bajo las condiciones de interés, de tal manera que su comportamiento pueda conocerse.

El elemento más importante de un experimento, y que muchas veces se subestima, es la formulación del problema por resolver. No puede esperarse una oportunidad de éxito razonable sin alguna dirección con respecto al propósito del experimento. Una vez que éste se define, es necesario identificar la variable por medir o *respuesta* que se va a estudiar y el *factor* o *factores* potenciales que pueden influenciar la variabilidad de la respuesta. La respuesta también se conoce como *variable dependiente;* el factor recibe el nombre de *variable independiente;* se supone que este último se encuentra bajo el control del investigador. Por ejemplo, en una tienda el interés recae en el número de empleados disponible, de manera que el tiempo de espera del cliente no sea excesivo. En este caso, la respuesta es el tiempo de espera y el factor el número de empleados disponible.

Un *nivel* o *tratamiento* del factor es un valor o condición de éste bajo el cual se observará la respuesta medible. Por ejemplo, supóngase que se desea observar el tiempo de espera cuando la tienda tiene a su servicio dos, cuatro o seis empleados a la vez. Si un experimento consiste en varios factores, un tratamiento es una combinación de los niveles de cada factor; por ejemplo, si se desea estudiar el tiempo de espera como una función del número de empleados en un determinado momento del día, entonces un tratamiento es la combinación de un número particular de empleados en un momento dado del día. El proceso por medio del cual se seleccionan los tratamientos se encuentra dictado más o menos por las metas del experimento. Para experimentos preliminares, en los cuales el propósito primordial es aislar los principales factores, el investigador debe escoger mentalmente los tratamientos con una visión muy amplia, de manera que obtenga un conocimiento útil del mecanismo bajo estudio. En forma posterior, se puede conducir un experimento más preciso con el propósito de hacer hallazgos más específicos.

Una *unidad experimental* se define como el objeto (persona o cosa) que es capaz de producir una medición de la variable de respuesta después de aplicar un tratamiento dado. La selección de una unidad experimental o del tamaño de ésta descansa, de nuevo, enteramente en el experimentador. Por ejemplo, si un fabricante de focos desea comparar la duración de éstos con la de sus competidores, entonces los focos seleccionados son las unidades experimentales y el número de marcas diferentes los tratamientos. O si se tiene interés en determinar la concentración de un contaminante en un lago en función de la ubicación geográfica, entonces las localidades del lago que se seleccionan para medir la concentración del contaminante son los tratamientos y la pequeña área superficial de cada localidad, la unidad experimental.

En un ambiente de incertidumbre los experimentos son, en forma general, comparativos en el sentido de que, idealmente, miden y comparan las respuestas de unidades experimentales esencialmente idénticas, después de que éstas se exponen a los tratamientos seleccionados y aplicados por el investigador. Todos los factores externos que pueden influenciar la respuesta deben eliminarse o controlarse. Sin embargo, no siempre puede garantizarse el control de los factores externos; por ejemplo, en forma práctica, casi cualquier experimento que incluye alguna actividad financiera guardará alguna interrelación con las condiciones económicas prevalecientes que no pueden controlarse. Tal desviación del control experimental ideal necesita de la repetición del experimento en una muestra de unidades experimentales para determinar la variación aleatoria o *error experimental*. Esta es la variación extraña en la respuesta o la variación que no puede ser atribuible a un cambio de tratamiento. Por lo tanto, es posible la inferencia estadística al comparar el error experimental con las respuestas promedio que resultan de la aplicación de los diferentes tratamientos.

En algunas ciencias pueden llevarse a cabo experimentos de laboratorio ideales, pero en las ciencias socioeconómicas, las desviaciones de las condiciones experimentales ideales tienen un lugar común debido a que el medio no permite un control suficiente. Por ejemplo, puede ser interesante estudiar el efecto de un aumento en las tasas de interés (tramiento) en la actividad de construcción de casas (respuesta) por parte de los constructores (unidades experimentales). Los tratamientos no pueden aplicarse a las unidades experimentales, ni la respuesta puede medirse de acuerdo con un experimento planeado. Sólo puede registrarse la información conforme cambian las condi-

ciones en el mundo real. Aunque para un purista lo anterior no constituye un experimento, estos tipos de estudios merecen una considerable atención. Para el análisis de estos datos es más apropiado el empleo de los métodos de regresión que los que se estudiarán en este capítulo. En los capítulos 13 y 14 se examinará el análisis de regresión.

12.3 Diseños estadísticos

El proceso por medio del cual se miden las observaciones de la respuesta se centra en un diseño estadístico. En general, en los experimentos diseñados estadísticamente, las unidades experimentales deben seleccionarse en forma imparcial, así como los tratamientos asignados a éstas, mediante un proceso aleatorio, con el propósito de remover los posibles sesgos sistemáticos. Como ya se indicó en el capítulo 9, el proceso aleatorio no sólo protege contra el sesgo sistemático, sino también tiende a neutralizar los efectos de todos aquellos factores externos que no se encuentren bajo el control del investigador. Entonces las comparaciones entre los tratamientos se miden, en forma práctica, como si el efecto en la respuesta se debiera sólo a la diferencia entre los tratamientos.

En un experimento diseñado estadísticamente es de igual importancia el concepto de *repetición.* Como ya se ha notado con anterioridad, el propósito de la repetición es medir el error experimental. La magnitud de éste juega un papel muy importante en la toma de decisiones con respecto a la posibilidad de que las diferencias entre los tratamientos sean discernibles en forma estadística.

En el diseño de experimentos estadísticos, el interés primario recae en cómo asignar las unidades experimentales a los tratamientos (o viceversa), para asegurar un proceso imparcial. En este contexto surgen dos conceptos básicos: el proceso de asignación debe hacerse con base en un *diseño completamente aleatorio,* o en un *diseño en bloque completamente aleatorio.* Cualquiera de estos dos diseños puede emplearse en experimentos unifactoriales o en aquéllos en los que se desea investigar varios factores en forma simultánea. Con un diseño complementario aleatorio, la asignación de los tratamientos a cada unidad experimental se lleva a cabo en forma totalmente aleatoria y todas las unidades se suponen homogéneas. En forma general, se hace uso de un procedimiento aleatorio sencillo como la generación de números aleatorios para llevar a cabo el proceso de asignación. El uso de un diseño completamente aleatorio implica que las condiciones bajo las cuales será observada la respuesta (u otras que se encuentren bajo el control del investigador) serán las mismas a través de todo el experimento. Este tipo de diseño no debe usarse en aquellas situaciones en las que las observaciones se realizarán sobre factores potenciales como el tiempo, el espacio o efectos demográficos, a menos que éstos sean partes legítimas del experimento.

No obstante, muchas veces el investigador se da cuenta de que el experimento no se puede conducir en el mismo ambiente, debido, principalmente, a que no todas las unidades experimentales son homogéneas; por lo tanto, éstas se clasifican en *bloques* homogéneos y se asignan todos los tratamientos en forma aleatoria a las unidades de cada bloque, con lo que se crea lo que se conoce como un diseño en bloques completamente aleatorio. La palabra "completamente" indica que cada bloque contiene todos los

tratamientos, mientras que la palabra "aleatorio" significa que todos los tratamientos serán asignados, en forma aleatoria, a las unidades experimentales de cada bloque.

El investigador reconoce la necesidad de agrupar en bloques, mediante la identificación de los elementos potenciales de las unidades experimentales que no se han incluido en la definición de un tratamiento, pero que pueden causar una variación significativa en la respuesta. Muchas veces éstos guardan relación con efectos espaciales, temporales o demográficos. Por ejemplo, si las unidades experimentales son seres humanos, entonces el agrupamiento por bloques deberá hacerse tomando en cuenta sexo, edad, condiciones de salud, experiencia, etc., como lo dicta el experimento. Si éste se va a realizar en un lapso grande deberá considerarse como una variable para el agrupamiento por bloques. Si los datos experimentales se van a recolectar, ya sea en distintas localidades o en grupos, entonces éstos deberán considerarse como variables en bloque. Si se van a usar varios instrumentos para registrar los datos, se deberá considerar un agrupamiento de instrumentos por bloques, aun si éstos son del mismo modelo y con mayor razón si provienen de distintos fabricantes.

Por lo tanto, la necesidad de agrupar en bloques es evidente; entre más heterogéneas son las unidades experimentales, mayor es el error experimental y menor la oportunidad de detectar diferencias reales entre los diversos tratamientos. La razón de agrupar en bloques es tomar en cuenta, y de esta forma remover, la fuente de variación en la respuesta que no es de interés, con lo que se incrementa la sensitividad para detectar diferencias entre los tratamientos. Así, el principio general de un diseño estadístico radica en minimizar el error experimental mediante el control de las variaciones extrañas, de manera que pueda detectarse la variación sistemática en la respuesta.

12.4 Análisis de experimentos unifactoriales en un diseño completamente aleatorio

El tipo de experimento más sencillo es aquél que compara el efecto de $k \geq 2$ niveles de un solo factor sobre alguna variable de respuesta. Los niveles del factor son los tratamientos, y si éstos se aplican en forma aleatoria a un conjunto virtualmente homogéneo de unidades experimentales, el experimento tiene un diseño completamente aleatorio. Esta situación es una extensión natural del problema que surge cuando se comparan dos medias poblacionales en donde las variantes son desconocidas pero que se suponen iguales. La prueba t para dos muestras, la cual se estudió en el capítulo 9, se basa en un diseño completamente aleatorio.

Para $k \geq 2$ niveles, se desea probar la hipótesis nula

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k \tag{12.1}$$

contra la alternativa de que algunas de las medias de la población no son las mismas. Si es posible rechazar la hipótesis nula con base en k muestras independientes, entonces las medias de las k poblaciones no son todas iguales entre sí, o el efecto de los tratamientos sobre la respuesta es estadísticamente discernible. Si no puede rechazarse la hipótesis nula, cualquier desviación observada en la respuesta se debe sólo al error aleatorio y no a causa de un cambio en el tratamiento.

Se pueden manejar muchos problemas prácticos con un experimento unifactorial completamente aleatorio. Unos cuantos ejemplos son los siguientes: saber si tienen algún efecto sobre el consumo de energía ligeras diferencias en el aislamiento de los techos de las casas; si la media del llenado producido por máquinas en un proceso de llenado es la misma, o si los vendedores que reciben diferentes métodos de entrenamiento, incrementan su volumen de ventas en forma diferente. En estos casos, los tratamientos son el aislamiento de los techos, las diferentes máquinas y los diversos métodos de entrenamiento; las unidades experimentales son las causas seleccionadas, los recipientes llenos y los vendedores, respectivamente. En el primer caso los tratamientos son cuantitativos, ya que los distingue una escala bien definida (R). En los últimos dos casos los tratamientos son cualitativos, dado que representan cosas o sujetos diferentes y por lo tanto carecen de escalas numéricas.

La necesidad de tener unidades experimentales homogéneas esencialmente puede ilustrarse con el primer ejemplo. Si se seleccionan casas para el experimento que no sean del mismo tamaño, en ese caso no se tiene el mismo aislamiento en los techos y se tienen distintas calidades con respecto al clima, si éstas se localizan en distintas zonas geográficas; de esta forma las diferencias en el consumo de energía no se pueden atribuir sólo al aislamiento del techo. Así, para un diseño completamente aleatorio los resultados serán ambiguos, a menos que las unidades experimentales sean virtualmente homogéneas.

La técnica del *análisis de varianza* proporciona el procedimiento inferencial para probar la hipótesis nula dada por (12.1). Para desarrollar esta técnica, se analizará el problema del aislamiento. Supóngase que se tiene interés en k diferentes niveles de aislamiento en el techo, tales que para el j-ésimo nivel se observará el consumo de energía mensual del sistema de calentamiento en n_j casas diferentes pero muy similares. Las casas que se seleccionan para este experimento son homogéneas y los factores externos están controlados dentro de ciertos límites prácticos. La información de la muestra puede colocarse como se presenta en la tabla 12.1, donde la respuesta medible es el número de kilowats-hora mensuales utilizados por el sistema de calentamiento de cada casa.

TABLA 12.1 Arreglo común de los datos de la muestra de un experimento con sólo un factor completamente aleatorizado

Tratamientos					
1	2	⋯	j	⋯	k
Y_{11}	Y_{12}	⋯	Y_{1j}	⋯	Y_{1k}
Y_{21}	Y_{22}	⋯	Y_{2j}	⋯	Y_{2k}
⋮	⋮		⋮		⋮
Y_{i1}	Y_{i2}	⋯	Y_{ij}	⋯	Y_{ik}
⋮	⋮		⋮		⋮
$Y_{n_1 1}$	$Y_{n_2 2}$	⋯	$Y_{n_j j}$	⋯	$Y_{n_k k}$

Se supone que cada nivel de aislamiento térmico en los techos representa una población a partir de la cual se obtiene una muestra; también, que las distribuciones de las poblaciones para cada nivel de aislamiento son normales con varianzas iguales. De acuerdo con lo anterior, las columnas de la tabla 12.1 representan k muestras aleatorias independientes de tamaños n_j, $j = 1, 2, \ldots, k$. Si la hipótesis nula dada por (12.1) es cierta, la observación Y_{ij} es el uso promedio de energía de los sistemas de calentamiento para todos los k niveles de aislamiento térmico y cualquier desviación del promedio se debe a un error aleatorio. Si H_0 es falsa, entonces Y_{ij} está constituida por todos los promedios, más el efecto del j-ésimo tratamiento y el error aleatorio. El promedio matemático para un experimento unifactorial completamente aleatorio es

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij} \qquad \begin{matrix} j = 1, 2, \ldots, k, \\ i = 1, 2, \ldots, n_j, \end{matrix} \tag{12.2}$$

en donde Y_{ij} es la i-ésima observación del j-ésimo tratamiento, μ es la media sobre todas las k poblaciones, τ_j es el efecto sobre la respuesta debido al j-ésimo tratamiento, y ε_{ij}* es el error experimental para la i-ésima observación bajo el j-ésimo tratamiento.

Se supone que los errores son independientes y que se encuentran normalmente distribuidos con medias cero y varianzas iguales. En otras palabras, $\varepsilon_{ij} \sim N(0, \sigma^2)$ para toda i y j. La suposición sobre los τ_j' depende de cómo considere el investigador los niveles del factor. Si el investigador está interesado en lo que le pasa a la respuesta, sólo para ciertos niveles del factor que se seleccionan de antemano, entonces $\tau_1, \tau_2, \ldots, \tau_k$ se consideran como parámetros fijos tales, que

$$\sum_{j=1}^{k} n_j \tau_j = 0.$$

Por lo tanto, el modelo dado por (12.2) se conoce como *modelo de efectos fijos* y las inferencias estadísticas con respecto a los efectos de los tratamientos pertenecen, en forma exclusiva, a los niveles seleccionados.

Por otro lado, si los niveles empleados en el experimento se seleccionaron al azar, de una población de posibles niveles, entonces $\tau_1, \tau_2, \ldots, \tau_k$ son variables aleatorias independientes que $\tau_j \sim N(0, \sigma_\tau^2)$ para toda j. En este caso, el modelo dado por (12.2) se conoce como *modelo de efectos aleatorios,* y las inferencias estadísticas con respecto a los niveles de un factor pertenecen a la población de niveles.

En general, para factores cuantitativos es deseable escoger niveles fijos del intervalo de interés, debido a que no es probable que una selección aleatoria proporcione una amplia cobertura de éste. La interpolación de los niveles fijos previamente seleccionados también es una práctica muy segura para factores cuantitativos. Cuando los factores son cualitativos como seres humanos, localidades o grupos, su selección sólo es importante cuando puede revelar algo con respecto a la variabilidad de la población.

*En lugar de emplear una letra mayúscula para las variables aleatorias ε_{ij}, se seguirá la tradición de utilizar la letra griega minúscula épsilon.

Para un modelo de efectos fijos, una hipótesis nula equivalente a (12.2) es

$$H_0: \tau_j = 0, \quad \text{para toda } j. \tag{12.3}$$

La hipótesis nula (12.3) establece que no existe ningún efecto de los tratamientos sobre la respuesta, lo que a su vez implica que las k medias de la población son iguales entre sí. Entonces se tiene como resultado que cada observación consiste en una media común y cualquier desviación con respecto a ésta se debe a la variación inherente dentro de cada población.

Para un modelo de efectos aleatorios, la hipótesis nula consiste en la proposición de que la varianza entre los τ_j (o los efectos del tratamiento) es cero; es decir,

$$H_0: \sigma_\tau^2 = 0. \tag{12.4}$$

Así, al suponer independencia entre los errores y tratamientos aleatorios,

$$Var(Y_{ij}) = \sigma^2 + \sigma_\tau^2.$$

Para el modelo de efectos aleatorios, el interés recae en hacer una evaluación de cuánto de la varianza en las observaciones se debe a diferencias reales en las medias de los tratamientos y cuánto se debe a errores aleatorios con respecto a estas medias.

En este capítulo el principal interés se centra en el modelo de efectos fijos, pero se incluirá el caso de efectos aleatorios cuando sea necesario. El punto de vista empleado para desarrollar la técnica del análisis de varianza será, en gran parte, intuitivo. Para un tratamiento teórico de la materia, véase [6].

12.4.1 Análisis de varianza para un modelo de efectos fijos

Sean $\mu_1, \mu_2, \ldots, \mu_k$ las medias de las k poblaciones, y sea μ la media de todas las poblaciones. Se define el efecto τ_j del j-ésimo tratamiento como la desviación de la j-ésima población media μ_j respecto a la media global μ. De esta forma,

$$\tau_j = \mu_j - \mu, \qquad j = 1, 2, \ldots, k.$$

En el mismo sentido, el error aleatorio correspondiente ε_{ij} de la observación Y_{ij} es la desviación de Y_{ij} con respecto de la j-ésima media μ_j o

$$\varepsilon_{ij} = Y_{ij} - \mu_j, \qquad j = 1, 2, \ldots, k,$$
$$i = 1, 2, \ldots, n_j.$$

De acuerdo con lo anterior, el modelo dado por (12.2) puede escribirse de la siguiente manera

$$Y_{ij} = \mu + (\mu_j - \mu) + (Y_{ij} - \mu_j),$$

o

$$Y_{ij} - \mu = (\mu_j - \mu) + (Y_{ij} - \mu_j). \tag{12.5}$$

La igualdad dada por (12.5) establece, en forma explícita, que cualquier desviación de una observación con respecto a la media global se debe a dos posibles causas: a la diferencia en el tratamiento o a un error aleatorio. Si se rechaza la hipótesis nula dada por (12.3), los datos de la muestra deben demostrar que la desviación total que se debe a la diferencia en el tratamiento es, suficientemente, más grande que la desviación causada por el error aleatorio. De esta forma, la técnica del análisis de varianza es en realidad un análisis de la variación de las medias y éste se logra mediante la participación de la variación total en las observaciones en componentes especificados por el modelo matemático. Esto permite determinar una estadística apropiada de tal manera que pueda tomarse una decisión con respecto a la hipótesis H_0: $\tau_j = 0$

Los parámetros $\mu_1, \mu_2, \ldots, \mu_k$ y μ no son conocidos, pero pueden estimarse con base en las observaciones de las k muestras aleatorias. Para la información de la muestra dada en la tabla 12.1 se define lo siguiente:

$$T_{\cdot j} = \sum_{i=1}^{n_j} Y_{ij}, \qquad j = 1, 2, \ldots, k,$$

$$\overline{Y}_{\cdot j} = T_{\cdot j}/n_j, \qquad j = 1, 2, \ldots, k,$$

$$T_{\cdot\cdot} = \sum_{j=1}^{k} T_{\cdot j},$$

$$N = \sum_{j=1}^{k} n_j,$$

$$\overline{Y}_{\cdot\cdot} = T_{\cdot\cdot}/N.$$

De nuevo, se emplea la notación de punto para indicar que la suma se lleva a cabo sobre el correspondiente subíndice. En particular, $T_{\cdot j}$ es la suma de las n_j observaciones en el j-ésimo tratamiento, $\overline{Y}_{\cdot j}$ es la media de la muestra del j-ésimo tratamiento, $T_{\cdot\cdot}$ es la suma de todas las N observaciones y $\overline{Y}_{\cdot\cdot}$ es la media de la muestra de todas las observaciones.

Al sustituir las estadísticas $\overline{Y}_{\cdot j}$ y $\overline{Y}_{\cdot\cdot}$ en (12.5) para los parámetros μ_j y μ, respectivamente, se obtiene la correspondiente igualdad en la muestra

$$Y_{ij} - \overline{Y}_{\cdot\cdot} = (\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}) + (Y_{ij} - \overline{Y}_{\cdot j}). \tag{12.6}$$

La esencia de la identidad de la muestra (12.6) es la división de la desviación de una observación Y_{ij} del promedio de la muestra total $\overline{Y}_{\cdot\cdot}$ en dos componentes la desviación de la media de la muestra del tratamiento $\overline{Y}_{\cdot j}$ de $\overline{Y}_{\cdot\cdot}$, y la desviación de Y_{ij} de su propia media de tratamiento $\overline{Y}_{\cdot j}$. De acuerdo con lo anterior, puede argumentarse en forma lógica que entre mayor sea la desviación entre $\overline{Y}_{\cdot j}$ y $\overline{Y}_{\cdot\cdot}$, se tiene más inclinación a rechazar la hipótesis nula dada por (12.3).

Para determinar una estadística de prueba apropiada, supóngase que se toma el cuadrado de ambos miembros de (12.6) y se suman sobre todos los i y j. De esta

forma,

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_{..})^2=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{Y}_{.j}-\overline{Y}_{..})^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_{.j})^2$$
$$+2\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{Y}_{.j}-\overline{Y}_{..})(Y_{ij}-\overline{Y}_{.j}). \qquad (12.7)$$

Pero

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{Y}_{.j}-\overline{Y}_{..})(Y_{ij}-\overline{Y}_{.j})=\sum_{j=1}^{k}(\overline{Y}_{.j}-\overline{Y}_{..})\left[\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_{.j})\right]$$
$$=\sum_{j=1}^{k}(\overline{Y}_{.j}-\overline{Y}_{..})\left[\sum_{i=1}^{n_j}Y_{ij}-n_j\overline{Y}_{.j}\right]$$
$$=0,$$

dado que $\Sigma_{i=1}^{n_j} Y_{ij} = T_{.j} = n_j\overline{Y}_{.j}$.

Como resultado se tiene que la ecuación

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_{..})^2=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{Y}_{.j}-\overline{Y}_{..})^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_{.j})^2 \qquad (12.8)$$

establece que la suma total de los cuadrados de las desviaciones con respecto a la media global se descompone en la suma de los cuadrados de las desviaciones de las medias de los tratamientos en relación con la media global, y la suma de los cuadrados de las desviaciones de las observaciones con respecto a sus propias medias de tratamiento. La expresión (12.8) se conoce como la ecuación fundamental del análisis de varianza. El término en el lado izquierdo de (12.8) es la *suma total de cuadrados* y se denota por *STC*. El término en medio de (12.8) es la *suma de los cuadrados de los tratamientos* y se denota por *SCTR*. El último término es la *suma de los cuadrados de los errores,* denotada por *SCE*. Por lo tanto,

$$STC = SCTR + SCE \qquad (12.9)$$

SCE mide la cantidad de variación en las observaciones debida a un error aleatorio. Si todas las observaciones que se encuentran dentro de un mismo tratamiento son las mismas, y si este hecho es cierto para todos los k tratamientos, entonces $SCE = 0$. De acuerdo con lo anterior, entre más grande es *SCE*, mayor es la variación en las observaciones que puede atribuirse a un error aleatorio. *SCTR* mide la extensión de la variación, en las observaciones, que se debe a las diferencias entre los tratamientos. Si todas las medias de los tratamientos son iguales entre sí, entonces $SCTR = 0$. De esta forma, entre más grande es el valor de *SCTR,* mayor es la diferencia que existe entre las medias de los tratamientos y la media global.

Puede demostrarse que bajo la hipótesis nula H_0: $\tau_j = 0$ y la suposición de que $\varepsilon_{ij} \sim N(0, \sigma^2)$, SCTR/σ^2 y SCE/σ^2 son dos variables aleatorias independientes con una distribución chi-cuadrada. Los grados de libertad se obtienen al separar la suma

total de cuadros. *STC* tiene $N - 1$ grados de libertad debido a que se pierde un grado de libertad al ser necesario que la suma de las desviaciones $(Y_{ij} - \bar{Y}_{..})$ para toda k y j sea cero. La suma de los cuadrados de los tratamientos tiene $k - 1$ grados de libertad debido a que se impone la restricción $\sum_{j=1}^{k} n_j (\bar{Y}_{.j} - \bar{Y}_{..}) = 0$ para las k desviaciones $(\bar{Y}_{.j} - \bar{Y}_{..})$. Esta restricción surge del hecho de que $\sum_{j=1}^{k} n_j \tau_j = 0$. Entonces, con base en (12.9), el número de grados de libertad para *SCE* será igual a la diferencia entre el número de grados de libertad para *STC* y *SCTR*,

$$\begin{aligned} \text{gl(SCE)} &= \text{gl(STC)} - \text{gl(SCTR)} \\ &= N - 1 - (k - 1) \\ &= N - k. \end{aligned}$$

Una suma de cuadrados dividido entre sus grados de libertad da origen a lo que se conoce como *cuadrado medio*. De acuerdo con lo anterior, el cuadrado medio del tratamiento es

$$\text{CMTR} = \text{SCTR}/(k - 1),$$

y el cuadrado medio del error es

$$\text{CME} = \text{SCE}/(N - k).$$

Ahora se puede argumentar que, dado que SCTR/σ^2 y SCE/σ^2 son dos variables aleatorias independientes chi-cuadrada con $k - 1$ y $N - k$ grados de libertad, respectivamente, entonces el cociente de las medias cuadráticas de la sección 7.8 tiene una distribución F con $k - 1$ y $N - k$ grados de libertad. Este cociente es la estadística apropiada para probar la hipótesis nula

$$H_0 : \tau_j = 0.$$

Lo anterior puede verificarse al examinar los valores esperados de los cuadrados medios. Puede demostrarse que

$$E(\text{CME}) = \sigma^2$$

y

$$E(\text{CMTR}) = \sigma^2 + \frac{\sum_{j=1}^{k} n_j \tau_j^2}{k - 1},$$

en donde σ^2 es la varianza común de los errores. Como resultado se tiene que el cuadrado medio del error es un estimador no sesgado de σ^2 sin importar si la hipótesis nula es cierta. Por otro lado, si H_0 es cierta, $\tau_j = 0$ para toda j, y $\Sigma n_j \tau_j^2 = 0$ Entonces $E(\text{CMTR}) = \sigma^2$; es decir, bajo H_0, tanto *CME* como *CMTR* son estimadores no sesgados de la varianza del error. Pero si la hipótesis nula no es de cierta, *CMTR* tiende generalmente a ser mayor que *CME*, dado que el término $\Sigma n_j \tau_j^2$ será positivo. En otras palabras, entre más grande sea la diferencia entre las medias de

los tratamientos y la media global, mayor será *CMTR*. Pero una ocurrencia de este tipo sugiere que las medias de los k tratamientos no son todas iguales entre sí y de esta forma debe rechazarse la hipótesis nula. De acuerdo con lo anterior, la hipótesis nula será rechazada cuando el valor del cociente.

$$F = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2/(k-1)}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_{\cdot j})^2/(N-k)} \tag{12.10}$$

se encuentre dentro de una región crítica superior de tamaño α.

El análisis anterior constituye la técnica del análisis de varianza para un experimento con sólo un factor completamente aleatorizado. Las fuentes de variación, grados de libertad, sumas de cuadrados, cuadrados medios, y el cociente F juntos, constituyen lo que se conoce como tabla de análisis de varianza *(ANOVA)* que se presenta en la tabla 12.2.

Dadas las verificaciones y_{ij}, $j = 1, 2, \ldots, k$, $i = 1, 2, \ldots, n_j$, el cálculo de las cantidades que aparecen en la tabla 12.2 puede hacerse en forma fácil mediante el empleo de cualquier paquete estadístico estándar para computadora. Para llevar a cabo el cálculo a mano, las sumas de los cuadrados pueden calcularse mediante el empleo de fórmulas algebraicamente equivalentes, pero desde un punto de vista de computación, más convenientes

$$\text{STC} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij}^2 - \frac{T_{\cdot\cdot}^2}{N},$$

$$\text{SCTR} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2 = \sum_{j=1}^{k}\frac{T_{\cdot j}^2}{n_j} - \frac{T_{\cdot\cdot}^2}{N},$$

$$\text{SCE} = \text{STC} - \text{SCTR}$$

Debe notarse que la hipótesis nula H_0: $\mu_1 = \mu_2$ para el caso de dos muestras también puede manejarse con el método del análisis de varianza. En el capítulo 13 se mostrará la relación que existe entre las estadísticas F y t de Student para $k = 2$.

TABLA 12.2 Tabla de análisis de varianza para un experimento con sólo un factor completamente aleatorio

Fuente de variación	gl	SC	CM	*Estadística F*
Tratamientos	$k-1$	$\Sigma\Sigma(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2$	$\Sigma\Sigma(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2/(k-1)$	$F = \dfrac{\Sigma\Sigma(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2/(k-1)}{\Sigma\Sigma(Y_{ij} - \bar{Y}_{\cdot j})^2/(N-k)}$
Error	$N-k$	$\Sigma\Sigma(Y_{ij} - \bar{Y}_{\cdot j})^2$	$\Sigma\Sigma(Y_{ij} - \bar{Y}_{\cdot j})^2/(N-k)$	
Total	$N-1$	$\Sigma\Sigma(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$		

TABLA 12.3 Calor empleado para cinco niveles de aislamiento

Espesor del aislamiento del techo (pulgadas)				
4	6	8	10	12
14.4	14.5	13.8	13.0	13.1
14.8	14.1	14.1	13.4	12.8
15.2	14.6	13.7	13.2	12.9
14.3	14.2	13.6		13.2
14.6		14.0		13.3
				12.7

Ejemplo 12.1 Los datos que figuran en la tabla 12.3 son los resultados de un diseño completamente aleatorizado para el cual la respuesta son los kilowats hora, empleados por los sistemas de calentamiento (en cientos de kilowats hora) para casas muy similares en un mes dado, como función de cinco niveles de aislamiento térmico (en pulgadas). Con base en esta información, ¿existe alguna razón para creer que por lo menos algunos de los consumos de energía promedio para los cinco niveles de aislamiento son diferentes? Supóngase un error de tipo I con α igual a 0.01.

Se desea probar la hipótesis nula de que

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu,$$

o en forma equivalente

$$H_0: \tau_j = 0, \quad j = 1, 2, \ldots, 5.$$

Los tamaños de las muestras son $n_1 = 5$, $n_2 = 4$, $n_3 = 5$, $n_4 = 3$, y $n_5 = 6$; así que $N = 5 + 4 + \cdots + 6 = 23$. Las sumas de los tratamientos son $T_{.1} = 73.3$, $T_{.2} = 57.4$, $T_{.3} = 69.2$, $T_{.4} = 39.6$, y $T_{.5} = 78$. La suma total es $T = 73.3 + 57.4 + \cdots + 78 = 317.5$. Las sumas de los cuadrados son las siguientes:

$$\text{STC} = 14.4^2 + 14.8^2 + \cdots + 12.7^2 - \frac{317.5^2}{23} = 11.05,$$

$$\text{SCTR} = \frac{73.3^2}{5} + \frac{57.4^2}{4} + \frac{69.2^2}{5} + \frac{39.6^2}{3} + \frac{78^2}{6} - \frac{317.5^2}{23} = 9.836,$$

$$\text{SCE} = 11.05 - 9.836 = 1.214.$$

La información se ha agrupado en una tabla de análisis de varianza que se muestra en la tabla 12.4. Dado que $f = 36.48 > f_{0.99,\,4,\,18} = 4.58$ se rechaza la hipótesis nula de que no existe ningún efecto debido a los tratamientos. En relación con lo anterior, existe una razón para creer que parte de los consumos promedio de energía son diferentes para los cinco niveles de aislamiento.

TABLA 12.4 Tabla ANOVA para el ejemplo 12.1

Fuente de variación	*gl*	*SC*	*CM*	*Valor F*
Tratamientos	4	9.836	2.459	36.48
Error	18	1.214	0.0674	
Total	22	11.05	$f_{0.99,\,4,\,18} = 4.58$	

12.4.2 Método de Scheffé para comparaciones múltiples

Recuérdese que la hipótesis alternativa en el análisis de varianza no especifica qué medias son diferentes; lo que establece es que por lo menos una es diferente a las otras, así que el rechazo de la hipótesis nula con base en la estadística F no puede emplearse como fundamento para aceptar una alternativa en particular. Por ejemplo, supóngase que se rechaza la hipótesis nula H_0: $\mu_1 = \mu_2 = \mu_3$; lo anterior significa que μ_3 es diferente, pero que μ_1 y μ_2 son las mismas. O puede expresar que las tres medias son diferentes entre sí, o cualquier otra combinación posible de estos resultados. Por lo tanto, ésta es una razón muy fuerte para que el investigador necesite un análisis más completo para explorar las diferencias estadísticamente discernibles entre cierto número de medias de población.

Con este propósito se han propuesto varios métodos; entre éstos se encuentran el procedimiento de rangos estudentizados de Tukey, la prueba de rangos múltiples de Duncan y el métodos de Scheffé (véase [5]). Sólo se analizará el método de Scheffé para comparaciones múltiples debido a que tiene, en forma relativa, pocas restricciones y es preferido por muchos cuando se comparan combinaciones de las medias de los tratamientos. El método de Scheffé radica en la formulación de un *contraste* que es una comparación que escoge el investigador para representar una combinación lineal de cualquier número de medias de población. Un contraste es un método general de comparación que permite al investigador determinar, con base en la evidencia de la muestra, si el contraste dado es estadísticamente discernible.

Se define un contraste, denotado por L, como

$$L = \sum_{j=1}^{k} c_j \mu_j, \tag{12.11}$$

en donde μ_j es la media del j-ésimo nivel, y las c_j's son constantes tales que $\sum_{j=1}^{k} c_j = 0$. Por ejemplo, $L = \mu_3 - \mu_4$ es un contraste con $c_1 = 1$ y $c_2 = -1$. Este contraste es una comparación entre μ_3 y μ_4. Otro contraste es $L = 3\mu_1 - \mu_2 - \mu_3 - \mu_4$, con $c_1 = 3$, $c_2 = c_3 = c_4 = -1$. Este contraste es una comparación entre μ_1 y μ_2, μ_3, y μ_4. De esta forma el método de Scheffé permite que el investigador escoja las comparaciones basadas en las características de interés.

Un estimador no sesgado de L está dado por

$$\hat{L} = \sum_{j=1}^{k} c_j \bar{Y}_{\cdot j}, \tag{12.12}$$

cuya varianza se estima mediante

$$s^2(\hat{L}) = \text{CME} \sum_{j=1}^{k} \frac{c_j^2}{n_j}. \tag{12.13}$$

Scheffé demostró (véase [7]) que todos los posibles contrastes definidos por (12.11) se encuentran incluidos, con una probabilidad de $1 - \alpha$, en el conjunto de intervalos

$$\hat{L} - A\,s(\hat{L}) \leq L \leq \hat{L} + A\,s(\hat{L}), \tag{12.14}$$

en donde

$$A = \sqrt{(k-1)f_{1-\alpha,\,k-1,\,N-k}}$$

y $\hat{L}$ y $s^2(\hat{L})$ se definen mediante (12.12) y (12.13), respectivamente. Si para algún contraste L se obtiene un intervalo a partir de (12.14) que no incluye al cero, entonces el contraste es estadísticamente discernible. Por lo tanto, en realidad para cada contraste L se está probando la hipótesis nula

$$H_0\colon L = 0.$$

La esencia del conjunto de intervalos definidos por (12.14) es que para *todos* los intervalos el nivel de confianza es de $100(1 - \alpha)$. Si se va a repetir un experimento muchas veces, y para cada una se calculan los invervalos de confianza para todos los posibles contrastes mediante el empleo de (12.14), entonces en un $100(1 - \alpha)$ de las repeticiones, todos los intervalos de confianza serán correctos. Que el intervalo de confianza sea del $100(1 - \alpha)$ para *todos* los intervalos, es mejor a obtener un intervalo de confianza del $100(1 - \alpha)$ para cada par de medias de tratamientos, en cuyo caso el nivel de confianza sólo es para cada par individual y no para el conjunto entero de éstos.

Ejemplo 12.2 En el ejemplo 12.1, compárese μ_4 contra μ_5; μ_2, μ_3, y μ_4 contra μ_5; μ_1 contra μ_2; y μ_3 y μ_4 contra μ_5, empleando el método de Scheffé con $\alpha = 0.01$.

Aunque pueden efectuarse comparaciones entre diversas combinaciones de los tratamientos, ciertas comparaciones parecen razonables si el objetivo es el ordenar los tratamientos en subgrupos dentro de los cuales no aparezca ninguna diferencia apreciable. Por ejemplo, si no existe una diferencia discernible entre el empleo de energía promedio para aislamientos térmicos de 10 y 12 pulgadas, puede ser, desde un punto de vista económico, más razonable utilizar un aislamiento de 10 pulgadas que uno de 12. Los contrastes para las cuatro comparaciones son:

$$L_1 = \mu_4 - \mu_5, \qquad L_2 = \mu_2 + \mu_3 + \mu_4 - 3\mu_5,$$

$$L_3 = \mu_1 - \mu_2, \qquad L_4 = 2\mu_5 - \mu_3 - \mu_4.$$

Se ilustrará el cálculo del intervalo de confianza para L_2. Dado que $\bar{y}_2 = 14.35$.

$\bar{y}_{\cdot 3} = 13.84$, $\bar{y}_{\cdot 4} = 13.2$, y $\bar{y}_{\cdot 5} = 13$,

$$\hat{L}_2 = 14.35 + 13.84 + 13.2 - (3)(13) = 2.39.$$

La varianza estimada es

$$s^2(\hat{L}_2) = 0.0674\left[\frac{1^2}{4} + \frac{1^2}{5} + \frac{1^2}{3} + \frac{(-3)^2}{6}\right] = 0.1539,$$

y

$$s(\hat{L}_2) = 0.3923.$$

Dado que $f_{0.99,\,4,\,18} = 4.58$, $A = \sqrt{(4)(4.58)} = 4.28$, el intervalo de confianza para L_2 es

$$2.39 \pm (4.28)(0.3923) = (0.7109, 4.0691).$$

Al seguir el mismo procedimiento se obtiene que los intervalos de confianza para los otros contrastes son

$$L_1: \quad (-0.5857, 0.9857),$$

$$L_3: \quad (-0.4354, 1.0554),$$

$$L_4: \quad (-2.2572, 0.1772).$$

Nótese que de los cuatro intervalos de confianza para los contrastes de interés sólo el de L_2 no incluye el valor cero. Dado que la inclusión de este valor en estos intervalos de confianza es equivalente a la falta de significancia estadística en una prueba bilateral con respecto a la diferencia entre las medias, una comparación de los cuatro intervalos revela que no existe ninguna diferencia apreciable en el consumo de energía promedio para un grosor del aislamiento térmico de 8, 10 o 12 pulgadas. Se llega a esta conclusión debido a que los contrastes L_1 y L_4 no son estadísticamente discernibles, pero L_2 sí lo es. Dado que L_2 es igual que L_4 excepto que éste contiene a μ_2 (6 pulgadas de aislamiento), con base en los resultados de este experimento puede considerarse a un aislamiento de 8 pulgadas de espesor, como óptimo, desde un punto de vista económico.

Debe notarse que si se rechaza la hipótesis nula de medias iguales mediante el empleo de la estadística F, entonces el método de Scheffé dará por lo menos un contraste que es estadísticamente significativo.

12.4.3 Análisis de residuos y efectos de la violación de las suposiciones

De la sección 9.6.3. recuérdese que, para muestras de diferente tamaño, el efecto de violar la suposición de varianzas iguales cuando se comparan dos medias puede ser sustancial. Dado que esta misma suposición se formula cuando se comparan k medias, se desean examinar las formas en que lo anterior puede detectarse y analizar los efectos sobre la inferencia cuando no violan las suposiciones.

Una forma sencilla y útil para detectar la discrepancia con el modelo propuesto se basa en un análisis de residuos. Un *residuo* es un estimador del error aleatorio ε_{ij}. Dado que

$$\varepsilon_{ij} = Y_{ij} - \mu_j,$$

el residuo correspondiente denotado por e_{ij}, se define como

$$e_{ij} = y_{ij} - \bar{y}_{\cdot j}, \qquad j = 1, 2, \ldots, k, \qquad i = 1, 2, \ldots, n_j.$$

Los residuos no son estimados en el sentido de estimación de parámetros, sino como estimadores de los valores de las variables aleatorias no observables ε_{ij} con base en los estimadores $\bar{y}_{\cdot j}$ para los k medias de población.

Si es válida la suposición de que los errores aleatorios tienen las mismas varianzas para todos los niveles de k, entonces una gráfica de los residuos de cada tratamiento no revelará ninguna diferencia apreciable en la dispersión de los residuos alrededor del cero. Si esta dispersión es notablemente diferente para algunos tratamientos, entonces es posible que las varianzas no sean iguales para todos los tratamientos. Para normalizar la escala de magnitudes de los residuos es preferible emplear los *residuos estandarizados* $e_{ij}/\sqrt{\text{CME}}$. Entonces, dado que por hipótesis los errores aleatorios se encuentran normalmente distribuidos, un residuo estandarizado rara vez se encontrará más allá de un intervalo de ± 3

Se ilustrará el análisis de residuos empleando los datos del ejemplo 12.1. Dado que $\bar{y}_{\cdot 1} = 14.66$ y $\sqrt{\text{CME}} = 0.2596$, los residuos para el primer tratamiento son $14.4 - 14.66 = -0.26$, $14.8 - 14.66 = 0.14$, $15.2 - 14.66 = 0.54$, $14.3 - 14.66 = -0.36$, y $14.6. - 14.66 = -0.06$, y los residuos correspondientes estandarizados son –1.00, 0.54, 2.08, –1.39 y –0.23. Al seguir este procedimiento se obtienen todos los residuos estandarizados que aparecen en la tabla 12.5.

La figura 12.1 ilustra los residuos estandarizados para cada tratamiento. Se observa que no existe ninguna diferencia notable en la dispersión para cada uno de los cinco tratamientos excepto para uno de los residuos del primer tratamiento. De acuerdo con lo anterior, parece que la hipótesis de que las varianzas de los cinco tratamientos son las mismas, es razonable en este caso. También se encuentran disponibles en la literatura estadística procedimientos formales para verificar la hipótesis de igualdad entre las k varianzas. Dos de los usados con más frecuencia son la prueba de Bartlett y la prueba de Hartley. Se invita al lector a que consulte [5] para conocer los detalles.

TABLA 12.5 Residuos estandarizados para el ejemplo 12.1

4	6	8	10	12
−1.00	0.58	−0.15	−0.77	0.39
0.54	−0.96	1.00	0.77	−0.77
2.08	0.96	−0.54	0	−0.39
−1.39	−0.58	−0.92		0.77
−0.23		0.62		1.16
				−1.16

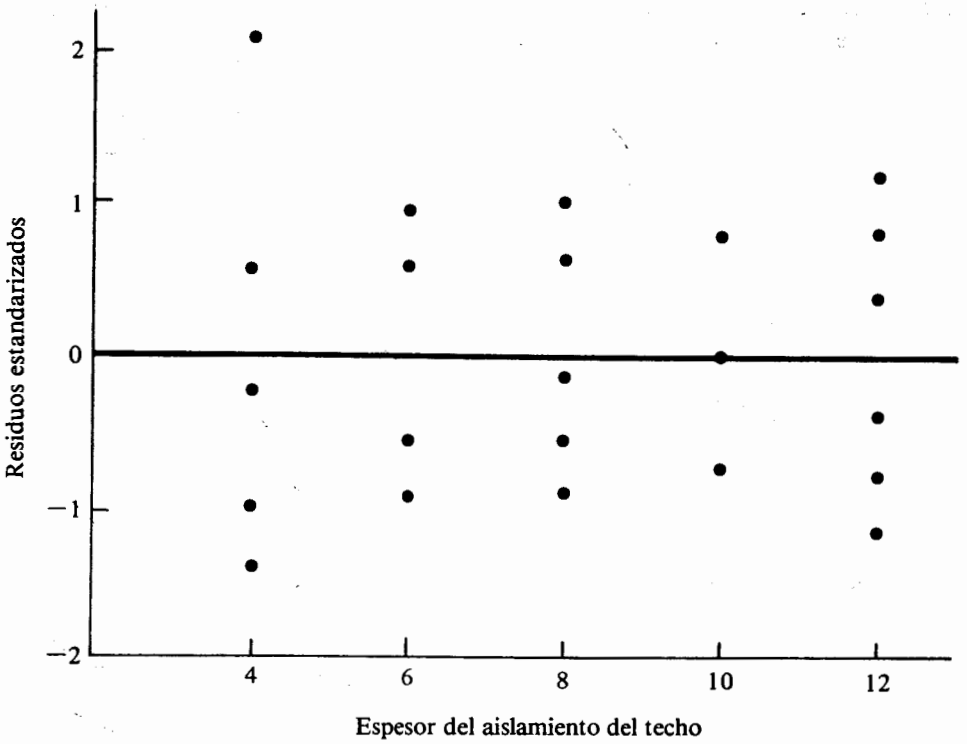


FIGURA 12.1 Gráfica de los residuos estandarizados para los cinco tratamientos del ejemplo 12.1

Como se examinó en el capítulo 9, el efecto sobre las inferencias con respecto a las medias, cuando los errores aleatorios no se encuentran normalmente distribuidos, es menor mientras el alejamiento de la normalidad no sea muy severo. De esta forma, la estadística F en el análisis de varianza es robusta con respecto a los alejamientos de la hipótesis de normalidad. Si las varianzas de todos los tratamientos no son iguales entre sí, puede aumentarse el tamaño de la región crítica de la estadística F para el caso de efectos fijos; pero, como se analizó en el capítulo 9, este efecto puede minimizarse mediante el empleo de muestras de igual tamaño para cada tratamiento. En otras palabras, en el análisis de varianza, la estadística F también es más robusta ante varianzas desiguales siempre y cuando los tamaños de la muestra de los tratamientos sean iguales. Desafortunadamente este resultado no se extiende al caso de efectos aleatorios en el que la violación de la hipótesis de varianzas iguales generalmente tendrá efectos considerables sobre las inferencias aun para muestras del mismo tamaño.

La hipótesis crucial en el desarrollo del análisis de varianza es que los errores aleatorios son independientes. Si los errores son interdependientes, el tamaño real de la región crítica puede ser, en forma substancial, más grande (cinco o más veces) que

el tamaño dictado al seleccionar la probabilidad del error de tipo I. Se invita al lector a que consulte [3], para una revisión de las consecuencias que surgen al violar las suposiciones en el análisis de varianza.

12.4.4 El caso de efectos aleatorios

Para introducir el caso de efectos aleatorios se utilizará el siguiente análisis breve. Para una presentación más completa se sugiere consultar [6]. Para el modelo de efectos aleatorios se formuló la suposición de que los niveles empleados en el experimento fueron seleccionados en forma aleatoria de una población de posibles niveles. Además se supondrá que $\tau_j \sim N(0,\sigma_\tau^2)$, en donde σ_τ^2 es la varianza de los tratamientos aleatorios τ_j. La descomposición de la suma total de cuadrados y el análisis de varianza es igual a la del caso de efectos fijos para un experimento con sólo un factor, pero en este caso el valor esperado del cuadrado medio de tratamiento es diferente. Dadas muestras de igual tamaño n para todos los niveles, se puede demostrar que

$$E(\text{CME}) = \sigma^2,$$

y (12.15)

$$E(\text{CMTR}) = \sigma^2 + n\sigma_\tau^2.$$

La región apropiada de rechazo sigue siendo la misma ya que un valor grande del cociente entre *CMTR* y *CME* sugiere que debe rechazarse la hipótesis nula H_0: $\sigma_\tau^2 = 0$

Ejemplo 12.3 Una planta de enlatado emplea un número muy grande de máquinas para su proceso de llenado. Se da por hecho que cada máquina vacía un peso especificado del producto en cada lata. El gerente de la planta sospecha que existe una gran variación en la cantidad del producto que se vacía entre las distintas máquinas. Para verificar su sospecha, escoge al azar cuatro máquinas y pesa el contenido de cinco latas, seleccionadas en forma aleatoria, llenadas por cada una de las cuatro máquinas. Los resultados se muestran en la tabla 12.6. ¿Qué proporción de la varianza en los pesos puede atribuirse a las diferencias que existen entre las máquinas?

Primero se llevará a cabo un análisis de varianza para saber si puede rechazarse H_0: $\sigma_\tau^2 = 0$. Los totales de las máquinas son $T_{.1} = 6.14$, $T_{.2} = 6.03$, $T_{.3} = 5.99$ y

TABLA 12.6 Contenido en peso para un proceso de llenado

Máquina			
1	*2*	*3*	*4*
1.24	1.20	1.19	1.18
1.22	1.20	1.20	1.18
1.22	1.21	1.19	1.19
1.23	1.22	1.20	1.18
1.23	1.20	1.21	1.20

TABLA 12.7 Tabla ANOVA para el ejemplo 12.3

Fuente de variación	gl	SC	CM	*Valor F*
Tratamientos	3	0.004695	0.001565	20.87
Error	16	0.0012	0.000075	
Total	19	0.005895	$f_{0.95,\,3,\,16} = 3.24$	

$T_{\cdot 4} = 5.93$. El total global es $T_{\cdot\cdot} = 24.09$, y los tamaños de todas las muestras son $n = 5$. Entonces

$$\text{STC} = 1.24^2 + 1.22^2 + \cdots + 1.20^2 - \frac{24.09^2}{20} = 0.005895,$$

$$\text{SCTR} = \frac{6.14^2 + 6.03^2 + 5.99^2 + 5.93^2}{5} - \frac{24.09^2}{20} = 0.004695,$$

$$\text{SCE} = 0.005895 - 0.004695 = 0.0012.$$

La tabla ANOVA se da en la tabla 12.7. Dado que $f = 20.87 > f_{0.95,\,3,\,16} = 3.24$, se rechaza la hipótesis nula de que no hay variación debida a las máquinas.

Para estimar la varianza en los pesos y qué proporción de ésta puede atribuirse a las diferencias entre las máquinas, recuérdese que para un modelo de efectos aleatorios

$$Var(Y_{ij}) = \sigma^2 + \sigma_\tau^2.$$

De (12.15), un estimado de σ^2 es $CME = 0.000075$, y un estimador de $\sigma^2 + 5\sigma_\tau^2$ es $CMTR = 0.001565$. En otras palabras,

$$0.000075 + 5s_\tau^2 = 0.001565$$

$$s_\tau^2 = \frac{0.001565 - 0.000075}{5}$$

$$= 0.000298$$

es un estimador de σ_τ^2. Entonces un estimador de la varianza en el peso es

$$s^2(Y_{ij}) = 0.000075 + 0.000298$$

$$= 0.000373,$$

de la cual 0.000298/0.000373, o el 79.89%, se debe a diferencias entre las máquinas.

12.5 Análisis de experimentos con sólo un factor en un diseño en bloque completamente aleatorizado

Recuérdese que cuando las unidades experimentales no son homogéneas, se introduce una fuente potencial de variación que, en general, puede afectar la inferencia con respecto al factor de interés. En estos casos es necesario emplear un diseño aleatorizado para remover la fuente externa de variación con lo que se incrementa la sensitividad para detectar diferencias entre los tratamientos de interés.

Ejemplo 12.4 La agencia de Protección del Medio Ambiente (APMA) anualmente clasifica de acuerdo con la eficiencia en el quemado de combustible a todos los automóviles disponibles para venta de Estados Unidos. Sin embargo, es un hecho muy conocido que las clasificaciones de la APMA se basan, principalmente, en pruebas de laboratorio y de esta forma se tiende a sobreestimar la eficiencia real en el quemado de combustible. Una empresa independiente desea determinar si existe una diferencia, estadísticamente discernible, en la eficiencia del quemado promedio de combustible bajo condiciones de rodamiento real para cinco automóviles compactos que tienen la misma clasificación APMA. La empresa tiene acceso a un recorrido de 400 millas que incluye tanto el manejo en ciudad como en carretera. Estúdiense los aspectos de diseño de este experimento.

Es claro que los tratamientos están constituidos por los cinco automóviles y que la respuesta medible es el número de millas por galón logradas por los automóviles durante el recorrido de 400 millas. Pero, ¿cuál es la unidad experimental?; ésta tiene que ser la persona que maneja el automóvil, pero no es común que una empresa que realiza pruebas utilice un conductor para todo el experimento. Supóngase que se escogen cuatro conductores para el experimento. Aunque la empresa explicará el propósito del experimento en forma breve, a los conductores ya se ha introducido otra fuente de posible variación. No importa qué tan similares sean los conductores entre sí; a pesar de todo existe un riesgo potencial de tener efectos por los conductores que pueden tomarse en cuenta mediante la creación de cuatro bloques, uno para cada conductor, de tal manera que los tratamientos dentro de cada bloque (los cinco automóviles) se apliquen a unidades experimentales homogéneas (el mismo conductor). La pregunta que surge en este momento es, ¿cómo asignar los automóviles a los conductores? El diseño aleatorizado especifica que la asignación de los tratamientos a las unidades experimentales dentro de cada bloque debe hacerse en forma aleatoria. De esta manera, para asignar el orden en el cual serán manejados los automóviles por cada conductor, se concibe un proceso de selección aleatorio simple. Por ejemplo, la asignación puede hacerse de acuerdo con la tabla 12.8, la cual constituye un diseño en bloque completamente aleatorizado.

A continuación se analizará un experimento con sólo un factor en un diseño en bloques completamente aleatorizado. Primero será necesario generalizar para después regresar al ejemplo de la eficiencia en consumo de combustible e ilustrar los pasos del cálculo. Las observaciones del experimento pueden colocarse como se muestran en la tabla 12.9.

TABLA 12.8 Diseño en bloque completamente aleatorizado para el ejemplo 12.4

		Automóvil				
Conductor	1	A_1	A_3	A_5	A_4	A_2
	2	A_5	A_3	A_4	A_2	A_1
	3	A_4	A_1	A_5	A_3	A_2
	4	A_2	A_5	A_4	A_1	A_3

Supóngase que se tienen k tratamientos y n bloques, el modelo matemático para un diseño con sólo un factor en bloques completamente aleatorizado es

$$Y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij} \qquad \begin{matrix} i = 1, 2, \ldots, n, \\ j = 1, 2, \ldots, k, \end{matrix} \tag{12.16}$$

en donde Y_{ij} es la observación de la respuesta en el i-ésimo bloque y bajo el j-ésimo tratamiento, μ es la media global, β_i es el efecto sobre la respuesta debido al i-ésimo bloque, τ_j es el efecto debido al j-ésimo tratamiento y ε_{ij} es el error aleatorio. Como en el caso anterior, se da por hecho que los errores son variables aleatorias independientes, tales que $\varepsilon_{ij} \sim N(0, \sigma^2)$ para toda i y j. Si tanto los tratamientos como los bloques son de efectos fijos, entonces las β_i'y los τ_j' son parámetros fijos que representan desviaciones de las medias de los bloques y los tratamientos de la media global, respectivamente. En otras palabras,

$$\begin{aligned} \beta_i &= \mu_{i\cdot} - \mu, \qquad i = 1, 2, \ldots, n, \\ \tau_j &= \mu_{\cdot j} - \mu, \qquad j = 1, 2, \ldots, k, \end{aligned} \tag{12.17}$$

en donde $\mu_{i\cdot}$ y $\mu_{\cdot j}$ son las medias de las población para el i-ésimo bloque y el j-ésimo tratamiento, respectivamente.

Al igual que en el diseño completamente aleatorizado, se supone que las varianzas de la población para todos los tratamientos son iguales. También debe suponerse que el efecto del tratamiento sobre la respuesta es el mismo para todos los bloques; en otras

TABLA 12.9 Arreglo común de las observaciones para un diseño con sólo un factor en bloque completamente aleatorizado

		Tratamiento					
		1	2	⋯	j	⋯	k
Bloque	1	Y_{11}	Y_{12}	⋯	Y_{1j}	⋯	Y_{1k}
	2	Y_{21}	Y_{22}	⋯	Y_{2j}	⋯	Y_{2k}
	⋮	⋮	⋮		⋮		⋮
	i	Y_{i1}	Y_{i2}	⋯	Y_{ij}	⋯	Y_{ik}
	⋮	⋮	⋮		⋮		⋮
	n	Y_{n1}	Y_{n2}	⋯	Y_{nj}	⋯	Y_{nk}

palabras, puede obtenerse la misma conclusión a partir de todos los bloques con respecto al efecto del tratamiento. Cuando esto ocurre se dice que los tratamientos y los bloques ***no interactúan,*** y sus efectos individuales sobre la respuesta son aditivos. La noción de ***interacción*** entre dos factores se examinará en la siguiente sección.

Para un diseño de un sólo factor en bloque completamente aleatorizado, el principal propósito es determinar si las diferencias en los tratamientos son estadísticamente significativas, es decir, para el caso de efectos fijos se desea probar la hipótesis nula

$$H_0: \tau_j = 0, \qquad j = 1, 2, \ldots, k.$$

El lector puede sorprenderse con respecto al efecto del bloque, pero el interés, en realidad, no recae en determinar si éste es estadísticamente apreciable. Todo lo que se desea hacer es aislar el efecto del bloque y removerlo del error experimental, de tal manera que se incremente la eficiencia para detectar diferencias reales entre los tratamientos, si es que éstas existen.

Para el procedimiento del análisis de varianza, puede escribirse el modelo dado por (12.16) como

$$\varepsilon_{ij} = Y_{ij} - \mu - \beta_i - \tau_j. \tag{12.18}$$

Al sustituir (12.17) para β_i y τ_j en (12.18), se tiene

$$\varepsilon_{ij} = Y_{ij} - \mu - \mu_{i\cdot} + \mu - \mu_{\cdot j} + \mu. \tag{12.19}$$

Ahora, al reemplazar (12.17) para β_i y τ_j y (12.19) para ε_{ij} en (12.16) se obtiene la siguiente identidad:

$$Y_{ij} - \mu = (\mu_{i\cdot} - \mu) + (\mu_{\cdot j} - \mu) + (Y_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu). \tag{12.20}$$

En otras palabras, la desviación de una observación con respecto a la media global tiene tres componentes (la desviación debida a los bloques, a los tratamientos y al error aleatorio).

Para las observaciones que se encuentran en la tabla 12.9 se definen las siguientes estadísticas:

$$T_{i\cdot} = \sum_{j=1}^{k} Y_{ij}, \qquad \overline{Y}_{i\cdot} = T_{i\cdot}/k, \qquad i = 1, 2, \ldots, n$$

$$T_{\cdot j} = \sum_{i=1}^{n} Y_{ij}, \qquad \overline{Y}_{\cdot j} = T_{\cdot j}/n, \qquad j = 1, 2, \ldots, k$$

$$T_{\cdot\cdot} = \sum_{i=1}^{n} \sum_{j=1}^{k} Y_{ij}, \qquad \overline{Y}_{\cdot\cdot} = T_{\cdot\cdot}/nk.$$

Por lo tanto, la identidad en términos de la muestra correspondiente a (12.20) es

$$Y_{ij} - \overline{Y}_{\cdot\cdot} = (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}) + (\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}) + (Y_{ij} - \overline{Y}_{i\cdot} - \overline{Y}_{\cdot j} + \overline{Y}_{\cdot\cdot}).$$

Al elevar al cuadrado ambos miembros y llevar a cabo la suma sobre i y j se tiene la relación

$$\sum_{i=1}^{n}\sum_{j=1}^{k}(Y_{ij}-\overline{Y}_{..})^2=\sum_{i=1}^{n}\sum_{j=1}^{k}(\overline{Y}_{i.}-\overline{Y}_{..})^2+\sum_{i=1}^{n}\sum_{j=1}^{k}(\overline{Y}_{.j}-\overline{Y}_{..})^2$$
$$+\sum_{i=1}^{n}\sum_{j=1}^{k}(Y_{ij}-\overline{Y}_{i.}-\overline{Y}_{.j}+\overline{Y}_{..})^2,$$

en donde puede demostrarse que los tres términos que contienen productos cruzados se reducen a cero. Ésta es la ecuación fundamental para el análisis de varianza, y establece que la suma total de los cuadrados *STC* se separa en la suma de los cuadrados de los bloques *SCB*, la suma de los cuadrados de los tratamientos SCTR y la suma de los cuadrados de los errores SCE.

Por causa de la restricción $\Sigma_{i=1}^{n}\Sigma_{j=1}^{k}(Y_{ij}-\overline{Y}_{..})=0$, el número de grados de libertad para STC es igual a $nk-1$. En forma similar, por causa de las restricciones $\Sigma_{i=1}^{n}(\overline{Y}_{i.}-\overline{Y}_{..})=0$ y $\Sigma_{j=1}^{k}(\overline{Y}_{.j}-\overline{Y}_{..})=0$, el número de grados de libertad para *SCB* y *SCTR* son iguales a $n-1$ y $k-1$, respectivamente. Se sigue que

$$\begin{aligned}\text{gl(SCE)} &= \text{gl(STC)} - \text{gl(SCB)} - \text{gl(SCTR)}\\ &= nk-1-(n-1)-(k-1)\\ &= (n-1)(k-1).\end{aligned}$$

Puede demostrarse que bajo las suposiciones del modelo y la hipótesis H_0: $\tau_j=0$, SCTR/σ^2 y SCE/σ^2 son dos variables aleatorias independientes con una distribución chi-cuadrada con $k-1$ y $(n-1)(k-1)$ grados de libertad, en forma correspondiente. También puede demostrarse que los valores esperados de los cuadrados medios del error y del tratamiento son

$$E(\text{CME})=\sigma^2$$

y

$$E(\text{CMTR})=\sigma^2+\frac{n\sum_{j=1}^{k}\tau_j^2}{k-1}.$$

Entonces, con base en el argumento previo, la estadística de prueba apropiada es el cociente de los cuadrados medios del tratamiento y del error, el cual tiene una distribución F con $k-1$ y $(n-1)(k-1)$ grados de libertad. Como antes, se sugiere una región crítica de tamaño α, ya que un valor grande del cociente tiende a implicar que no todas las medias de los tratamientos son las mismas. El análisis de varianza aparece en la tabla 12.10.

Debe notarse que es posible una prueba para el efecto de bloque al formar el cociente entre *CMB* y *CME* y compararlo con la región crítica que se encuentra en el extremo superior de una distribución F con $n-1$ y $(n-1)(k-1)$ grados de libertad.

TABLA 12.10 Tabla de análisis de varianza para un experimento con sólo un factor en bloque completamente aleatorizado

Fuente de variación	gl	SC	CM	*Estadística F*
Bloques	$n - 1$	$\Sigma\Sigma(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$		
Tratamientos	$k - 1$	$\Sigma\Sigma(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2$	$\text{CMTR} = \text{SCTR}/(k - 1)$	$F = \dfrac{\text{CMTR}}{\text{CME}}$
Error	$(n - 1)(k - 1)$	$\Sigma\Sigma(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2$	$\text{CME} = \text{SCE}/(n - 1)(k - 1)$	
Total	$nk - 1$	$\Sigma\Sigma(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$		

Lo anterior no constituye en realidad una parte integral del análisis. Después de todo, se escoge un bloque completamente aleatorizado para un experimento con sólo un factor para remover el efecto potencial de la fuente de variación extraña. Si tal efecto es estadísticamente significativo, realmente no es de gran interés.

Para realizar cálculos a mano, es preferible emplear las siguientes fórmulas que son equivalentes, en un sentido algebraico, para obtener las sumas de cuadrados.

$$\text{STC} = \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}^2 - \frac{T_{\cdot\cdot}^2}{nk}$$

$$\text{SCB} = \frac{1}{k}\sum_{i=1}^{n} T_{i\cdot}^2 - \frac{T_{\cdot\cdot}^2}{nk}$$

$$\text{SCTR} = \frac{1}{n}\sum_{j=1}^{k} T_{\cdot j}^2 - \frac{T_{\cdot\cdot}^2}{nk}$$

$$\text{SCE} = \text{STC} - \text{SCB} - \text{SCTR}$$

Para ilustrar los pasos de cálculo, supóngase que los resultados del experimento descrito en el ejemplo 12.4 son los que se muestran en la tabla 12.11 (las mediciones están dadas en millas por galón para un recorrido de 400 millas). Para probar la hipótesis nula

$$H_0\colon \tau_j = 0, \quad j = 1, 2, \ldots, 5,$$

las sumas de cuadrados dan

$$\text{STC} = 33.6^2 + 36.9^2 + \cdots + 32.8^2 - \frac{672.4^2}{20} = 102.212,$$

$$\text{SCB} = \frac{156.1^2 + \cdots + 172.4^2}{5} - \frac{672.4^2}{20} = 41.676,$$

$$\text{SCTR} = \frac{139.5^2 + \cdots + 133.3^2}{4} - \frac{672.4^2}{20} = 38.092,$$

TABLA 12.11 Datos experimentales para el ejemplo 12.4

	Automóvil					
Conductor	A_1	A_2	A_3	A_4	A_5	*Totales*
1	33.6	32.8	31.9	27.2	30.6	$T_{1\cdot} = 156.1$
2	36.9	36.1	32.1	34.4	35.3	$T_{2\cdot} = 174.8$
3	34.2	35.3	33.7	31.3	34.6	$T_{3\cdot} = 169.1$
4	34.8	37.1	34.8	32.9	32.8	$T_{4\cdot} = 172.4$
Totales	$T_{\cdot 1} = 139.5$	$T_{\cdot 2} = 141.3$	$T_{\cdot 3} = 132.5$	$T_{\cdot 4} = 125.8$	$T_{\cdot 5} = 133.3$	$T_{\cdot\cdot} = 672.4$

$$\text{SCE} = 102.212 - 41.676 - 38.092 = 22.444.$$

La tabla ANOVA se encuentra dada en la tabla 12.12. Dado que $f = 5.09 > f_{0.95,\,4,\,12} = 3.26$, se rechaza la hipótesis nula de igualdad de efecto de tratamiento. Por lo tanto, existe una razón para creer que las eficiencias en consumo medio de combustible de algunos de estos automóviles no son iguales.

La identificación y eliminación del efecto de los bloques de la variación total permite que se hagan comparaciones múltiples sobre los tratamientos, como ya se vio en la sección 12.4.2. Pueden definirse y probarse un gran número de contrastes para determinar si son estadísticamente apreciables al seguir el procedimiento delineado en la sección 12.4.2. La única excepción es que la cantidad denotada por A en (12.14) ahora está dada por

$$A = \sqrt{(k-1)f_{1-\alpha,\,k-1,\,(n-1)(k-1)}}.$$

A veces los bloques no son de efectos fijos, es decir, se eligen para el experimento en forma aleatoria de una población de posibles bloques. Si los tratamientos son de efectos fijos, la única diferencia con respecto al caso previo se encuentra en la suposición de β_i; i.e., $\beta_i \sim N(0, \sigma_\beta^2)$; pero el análisis sigue siendo el mismo, aun para comparaciones múltiples entre los tratamientos.

Además de la suposición de independencia, se hacen dos suposiciones clave para un diseño en bloques aleatorizados: las varianzas de cada tratamiento son iguales y

TABLA 12.12 Tabla ANOVA para el ejemplo 12.4

Fuente de variación	gl	SC	CM	*Valor F*
Bloques	3	41.676		
Tratamientos	4	38.092	9.523	5.09
Error	12	22.444	1.870	
Total	19	102.212	$f_{0.95,\,4,\,12} = 3.26$	

los bloques y tratamientos no interactúan. La presencia de interacción entre bloques y tratamientos implica que no es posible evaluar el efecto del tratamiento sobre todos los bloques, sino que éste se debe describir en forma individual para cada bloque. Si además los efectos del bloque y del tratamiento son aditivos, la estadística F no es sensitiva a la violación de la suposición de varianzas iguales; para éstas, si existe una interacción entre bloques y tratamientos, la estadística F se encuentra sesgada negativamente, es decir, si se rechaza la hipótesis nula de que no existe diferencia alguna entre los tratamientos, entonces puede confiarse en que existe una diferencia entre los tratamientos. Pero si la hipótesis nula no se rechaza, esto se puede deber, ya sea a un sesgo negativo (la presencia de interacción) o a la ausencia de diferencias entre los tratamientos. Puede emplearse un procedimiento desarrollado por Tukey, el cual se describe en [4], para probar la interacción entre bloques y tratamientos.

Si se violan tanto la suposición de varianzas iguales como la de aditividad, la estadística F para las diferencias en los tratamientos tiene un sesgo positivo; en otras palabras, si se rechaza la hipótesis nula de que no existe ninguna diferencia entre los tratamientos, esto no necesariamente implica que las diferencias entre los tratamientos sean estadísticamente significativas. Cuando existe preocupación sobre estas suposiciones, debe usarse una *prueba F conservadora* desarrollada por Geisser y Greenhouse (véase [4]). Los pasos de cálculo para esta prueba son iguales a los del método convencional ya descrito, excepto que el número de grados de libertad para este caso es de 1 y $n-1$ en lugar de $k-1$ y $(n-1)(k-1)$, para cada uno. Si para ambas pruebas se rechaza la hipótesis nula, puede tenerse la seguridad de que las diferencias entre los tratamientos son estadísticamente significativas. Si ambas pruebas no rechazan a H_0, entonces se puede proceder como si no existiese diferencia alguna entre los tratamientos.

12.6 Experimentos factoriales

Hasta este momento la presentación se ha dirigido hacia el análisis del efecto de un factor sobre la variable respuesta. Pero en muchas situaciones prácticas es necesario investigar, en forma simultánea, los efectos que tienen varios factores sobre la respuesta. Una forma muy eficiente de lograr lo anterior es mediante el uso de un experimento factorial en el que todos los niveles de un factor se combinan con todos los niveles de cualquier otro para formar los tratamientos. Por ejemplo, en un experimento factorial de dos factores en el que uno tiene tres niveles y el otro dos, existirán $3 \times 2 = 6$ tratamientos. En otras palabras, la respuesta será observada bajo seis tratamientos diferentes.

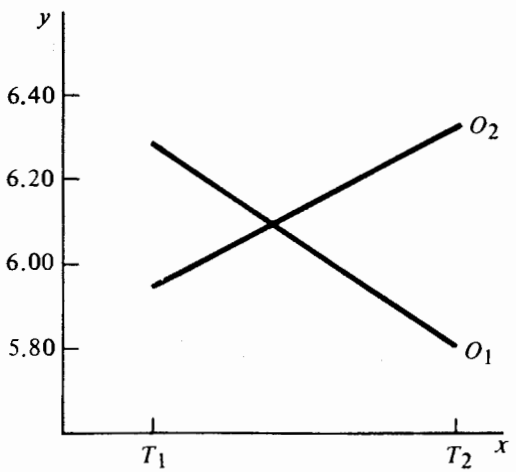
Con los experimentos factoriales no sólo es posible evaluar los efectos individuales de los factores sobre la respuesta, sino que también es posible determinar el efecto causado por sus interacciones. El efecto de un factor sobre una respuesta es simplemente el cambio en ésta, causado por un cambio en el nivel del factor. Pero si el efecto de un factor sobre la respuesta es diferente para distintos niveles de otro factor, entonces se dice que los dos factores interactúan entre sí. La presencia de interacción indica que el efecto de los factores sobre la respuesta es no lineal y de esta forma no puede asumirse un modelo aditivo.

Para ilustrar la interacción entre dos factores, considérese lo siguiente. Un fabricante de partes electrónicas emplea dos hornos y dos temperaturas con el propósito de probar la duración de cierto componente. Se seleccionan cuatro componentes de algún lote y se prueba su duración de acuerdo con las cuatro combinaciones posibles de hornos y temperaturas. El tiempo de duración de los componentes en horas es el siguiente:

	O_1	O_2
T_1	6.29	5.95
T_2	5.80	6.32

Los tratamientos para las cuatro posibles combinaciones de hornos y temperaturas son: O_1T_1, O_1T_2, O_2T_1, y O_2T_2. La diferencia en duración para los tratamientos O_1T_2 y O_1T_1 representa un estimador del efecto en la duración de los componentes en el primer horno, a consecuencia de un cambio en la temperatura. Se observa que este estimador es $5.80 - 6.29 = -0.49$. La diferencia en duración para los tratamientos O_2T_2 y O_2T_1 también es un estimador del efecto de la temperatura sobre la duración, pero ahora en el segundo horno. Esta diferencia es de $6.32 - 5.95 = 0.37$. Dado que estos dos estimadores son bastantes diferentes entre sí, el efecto de la temperatura en la duración del componente depende del horno en que éste se coloque. De esta forma, existe una interacción entre el horno y la temperatura. También se observa la misma ocurrencia al estimar el efecto del horno para $T_1(5.95 - 6.29 = -0.34)$ y $T_2(6.32 - 5.80 = 0.52)$. Estos resultados se ilustran en forma gráfica en la figura 12.2 en donde el eje y representa las observaciones de la respuesta; el eje x repre-

FIGURA 12.2 Efectos que interactúan



senta los niveles de un factor y los puntos graficados representan a cada nivel del otro factor. Si existe poca interacción entre el horno y la temperatura, las líneas que aparecen en la gráfica serían casi paralelas.

La determinación de si los efectos individuales o interacciones son estadísticamente apreciables puede hacerse sólo mediante inferencia estadística y no mediante el empleo de un analisis gráfico. En los siguientes párrafos se examinará un modelo no aditivo para un experimento factorial de dos factores en un diseño completamente aleatorizado. Se pueden analizar experimentos factoriales con más de dos factores mediante la extensión del procedimiento que a continuación se examina.

En un experimento factorial que incluye dos factores A y B con a y b niveles, respectivamente, el número de tratamientos es igual a $a \times b$. Si no se puede suponer un modelo aditivo (no interacción), sólo es posible una prueba para determinar si un efecto por interacción es estadísticamente apreciable, si se toma más de una observación de la respuesta para cada tratamiento. Lo anterior se debe a que no puede determinarse para cada estimador de la variación del error aleatorio a menos que la respuesta se observe más de una vez cada tratamiento, es decir, la evaluación de la variación del error aleatorio se basa en las diferencias en la respuesta observada bajo el mismo tratamiento. No está por demás notar que para un diseño completamente aleatorizado, los tratamientos deben aplicarse a unidades experimentales homogéneas sin importar cuántas veces se repita el proceso.

Si se suponen n aplicaciones de los ab tratamientos, el modelo matemático no aditivo para un factorial de dos factores es

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \qquad \begin{matrix} i = 1, 2, \ldots, a, \\ j = 1, 2, \ldots, b, \\ k = 1, 2, \ldots, n, \end{matrix} \qquad (12.21)$$

en donde Y_{ijk} es la k-ésima observación de la respuesta para el tratamiento (i, j), μ es la media global, α_i es el efecto principal causado por el i-ésimo nivel de A, β_j es el efecto principal causado por el j-ésimo nivel de B, $(\alpha\beta)_{ij}$ es el efecto de interacción para el i-ésimo nivel de A y el j-ésimo nivel de B y ε_{ijk} es el k-ésimo error aleatorio en el tratamiento (i, j). Como antes, se supone que las varianzas de la población para cada uno de los ab tratamientos son iguales, y que los errores aleatorios son variables aleatorias independientes, normalmente distribuidas, con medias iguales a cero y varianza común σ^2.

Si se supone que los factores A y B son de efectos fijos, entonces α_i, β_j, y $(\alpha\beta)_{ij}$ son parámetros fijos, tales que

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = 0$$

y

$$\sum_{i=1}^{a} (\alpha\beta)_{ij} = \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0.$$

Las siguientes hipótesis son de interés:

1. $H_0: (\alpha\beta)_{ij} = 0$ para toda i y j,
2. $H_0: \alpha_i = 0$ para toda i,
3. $H_0: \beta_j = 0$ para toda j.

Las últimas dos hipótesis incluyen los efectos (individuales) *principales* de los factores A y B, y la primera hipótesis pertenece a la posible interacción entre A y B. Si existe una fuerte interacción entre A y B, los resultados de las pruebas para demostrar un efecto principal causado por A o B pueden no ser significativos. Lo anterior es cierto debido a que los dos factores pueden interaccionar en tal forma (direcciones opuestas) que los efectos se compensen para uno o ambos factores. Este proceso de compensación puede evitar la detección de efectos principales significativos con base en una comparación entre las medias del nivel del factor.

Para desarrollar el procedimiento del análisis de varianza, puede escribirse el modelo (12.21) en términos de las desviaciones, al igual que en los casos previos.

$$Y_{ijk} - \mu = (\mu_{i\cdot\cdot} - \mu) + (\mu_{\cdot j\cdot} - \mu) + (\mu_{ij\cdot} - \mu_{i\cdot\cdot} - \mu_{\cdot j\cdot} + \mu) + (Y_{ijk} - \mu_{ij\cdot}), \quad (12.22)$$

en donde $\mu_{i\cdot\cdot}$ es la media real del i-ésimo nivel de A, $\mu_{\cdot j\cdot}$ es la media real del j-ésimo nivel de B y $\mu_{ij\cdot}$ es la media real del tratamiento (i, j). De esta forma, la igualdad dada por (12.22) establece que la desviación de una observación con respecto al promedio global está formada por cuatro componentes: las desviaciones causadas por el efecto principal de A; por el efecto principal de B; por el efecto de interacción entre A y B, por el error aleatorio.

Las observaciones de un factorial con dos factores en un experimento completamente aleatorizado pueden colocarse como se muestra en la tabla 12.13. De ésta se

TABLA 12.13 Arreglo común de las observaciones para un diseño factorial con dos factores y n observaciones por tratamiento

		A										
		Nivel 1				Nivel i				Nivel a		
B	Nivel 1	Y_{111}	$\cdots$	Y_{11k}	$\cdots$ Y_{11n}	$\cdots$	Y_{i11}	$\cdots$	Y_{i1k} $\cdots$ Y_{i1n}	$\cdots$	Y_{a11}	$\cdots$ Y_{a1k} $\cdots$ Y_{a1n}
	Nivel j	Y_{1j1}	$\cdots$	Y_{1jk}	$\cdots$ Y_{1jn}	$\cdots$	Y_{ij1}	$\cdots$	Y_{ijk} $\cdots$ Y_{ijn}	$\cdots$	Y_{aj1}	$\cdots$ Y_{ajk} $\cdots$ Y_{ajn}
	Nivel b	Y_{1b1}	$\cdots$	Y_{1bk}	$\cdots$ Y_{1bn}	$\cdots$	Y_{ib1}	$\cdots$	Y_{ibk} $\cdots$ Y_{ibn}	$\cdots$	Y_{ab1}	$\cdots$ Y_{abk} $\cdots$ Y_{abn}

definen las siguientes estadísticas:

$$T_{i\cdot\cdot} = \sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk}, \qquad T_{\cdot j\cdot} = \sum_{i=1}^{a}\sum_{k=1}^{n} Y_{ijk}, \qquad T_{\cdot\cdot k} = \sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ijk},$$

$$\overline{Y}_{i\cdot\cdot} = T_{i\cdot\cdot}/nb, \qquad \overline{Y}_{\cdot j\cdot} = T_{\cdot j\cdot}/na, \qquad \overline{Y}_{\cdot\cdot k} = T_{\cdot\cdot k}/ab,$$

$$T_{ij\cdot} = \sum_{k=1}^{n} Y_{ijk}, \qquad \overline{Y}_{ij\cdot} = T_{ij\cdot}/n,$$

$$T_{\cdots} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk}, \qquad \overline{Y}_{\cdots} = T_{\cdots}/nab.$$

Nótese que $T_{i\cdot\cdot}$ ($T_{\cdot j\cdot}$) es la suma de todas las observaciones en el i-ésimo (j-ésimo) nivel de A (B) y $T_{\cdot\cdot k}$ es la suma de todas las observaciones en la k-ésima repetición. En forma similar, $T_{ij\cdot}$ es la suma de todas las observaciones en el tratamiento (i, j). Las definiciones correspondientes para las medias de la muestra deben ser aparentes.

Al reemplazar los parámetros en (12.12) con sus correspondientes estimadores, se tiene

$$(Y_{ijk} - \overline{Y}_{\cdots}) = (\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}) + (\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}) + (\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}) + (Y_{ijk} - \overline{Y}_{ij\cdot}).$$

Si se eleva al cuadrado la identidad con base en la muestra anterior y se suman sobre i, j y k, todos los términos que contienen productos cruzados se reducen a cero, y se tiene el siguiente resultado:

$$\sum_i\sum_j\sum_k (Y_{ijk} - \overline{Y}_{\cdots})^2 = nb\sum_i (\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2 + na\sum_j (\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots})^2 + n\sum_i\sum_j (\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots})^2 + \sum_i\sum_j\sum_k (Y_{ijk} - \overline{Y}_{ij\cdot})^2. \quad (12.23)$$

En otras palabras, la suma total de cuadrados se separa en las sumas de cuadrados debidas: al factor *A (SCA)*, el factor *B (SCB)*, a la interacción entre *A* y *B (SCAB)* y a los errores *(SCE)*.

También puede escribirse el modelo (12.21) en términos de las desviaciones causadas por los tratamientos y el error aleatorio, es decir

$$(Y_{ijk} - \mu) = (\mu_{ij\cdot} - \mu) + (Y_{ijk} - \mu_{ij\cdot}). \quad (12.24)$$

En esta forma, la desviación debida a los tratamientos abarca los efectos debidos a A, B y la interacción A B. Al sustituir en (12.24) las correspondientes estadísticas, se tiene

$$(Y_{ijk} - \overline{Y}_{\cdots}) = (\overline{Y}_{ij\cdot} - \overline{Y}_{\cdots}) + (Y_{ijk} - \overline{Y}_{ij\cdot}),$$

las que, al elevarse al cuadrado y sumar sobre i, j y k, dan como resultado

$$\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{...})^2 = n \sum_i \sum_j (\bar{Y}_{ij\cdot} - \bar{Y}_{...})^2 + \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij\cdot})^2,$$

o

$$\text{STC} = \text{SCTR} + \text{SCE}. \tag{12.25}$$

De (12.23) se desprende que

$$\text{SCTR} = \text{SCA} + \text{SCB} + \text{SCAB}. \tag{12.26}$$

Puede demostrarse que, con base en (12.23), la descomposición del número de grados de libertad es la siguiente:

$$\text{gl(STC)} = \text{gl(SCA)} + \text{gl(SCB)} + \text{gl(SCAB)} + \text{gl(SCE)},$$

o

$$(nab - 1) = (a - 1) + (b - 1) + (a - 1)(b - 1) + ab(n - 1).$$

Para las suposiciones del modelo y la hipótesis de interés, SCA/σ^2, SCB/σ^2, SCAB/σ^2, y SCE/σ^2 son variables aleatorias independientes chi-cuadrada con $(a - 1)$, $(b - 1)$, $(a - 1)(b - 1)$ y $ab(n - 1)$ grados de libertad, para cada una. De acuerdo con lo anterior, la estadística de prueba para los efectos principales y de interacción son los cocientes entre los cuadrados medios, correspondientes y cuadrado medio del error y tienen una distribución F. Al igual que para los casos anteriores, una región crítica de tamaño α en el extremo superior de la región es la apropiada para cada caso. Puede observarse que el resultado anterior sigue siendo válido al examinar los valores esperados de los cuadrados medios. Para el caso de efectos fijos, estos valores son los siguientes:

$$E(\text{CME}) = \sigma^2,$$

$$E(\text{CMA}) = \sigma^2 + nb \frac{\sum \alpha_i^2}{a - 1},$$

$$E(\text{CMB}) = \sigma^2 + na \frac{\sum \beta_j^2}{b - 1},$$

$$E(\text{CMAB}) \quad ^2 + n \frac{\sum \sum (\alpha\beta)_{ij}^2}{(a - 1)(b - 1)}.$$

Si no existe ninguna interacción entre A y B (es decir, si $(\alpha\beta)_{ij} = 0$ para toda i y j), entonces $CMAB$ y CME tienen el mismo valor esperado y los efectos son aditivos. Pero si el cociente $CMAB/CME$ tiene un valor suficientemente grande, esto sugeriría una interacción estadísticamente apreciable entre A y B y, por lo tanto, debe rechazarse la hipótesis nula. De manera similar si $\alpha_i = 0$ para toda i, CMA y CME tienen valores esperados iguales y no existe un efecto principal causado por A. Pero un cociente grande entre CMA y CME tiende a implicar que el efecto principal

atribuible a A es estadísticamente significativo. El mismo argumento es válido para el efecto principal de B.

En la tabla 12.14 se encuentra un resumen del análisis de varianza para un diseño factorial con dos factores. Aunque en la tabla se proporcionan fórmulas de cálculo para cada fuente de variación, la práctica usual para realizarlos a mano es calcular *SCT* mediante la fórmula que aparece en la tabla 12.14 y *SCTR* de la fórmula

$$\text{SCTR} = \frac{1}{n}\sum_i \sum_j T_{ij\cdot}^2 - \frac{T_{\cdots}^2}{nab}.$$

Entonces puede obtenerse *SCE* al emplear (12.25). A su vez, mediante el empleo de las fórmulas que aparecen en la tabla 12.14 se calculan *SCA* y *SCB*, y se obtiene *SCAB* con base en (12.26).

Ejemplo 12.5 Se llevó a cabo una investigación para determinar si pueden encontrarse diferencias apreciables en los salarios iniciales para contadores graduados con base en el sexo, localidad del lugar de trabajo o la interacción de los dos. El estudio se llevó a cabo en grandes ciudades del noroeste, el oeste medio y el oeste. Se piensa que será suficiente un arreglo factorial en un diseño completamente aleatorizado. Se decide emplear los salarios iniciales de cuatro personas para cada una de las seis combinaciones de tratamientos. Para asegurar que las unidades experimentales son homogéneas, se seleccionaron personas con antecedentes muy similares en la medida de lo posible. Tienen la misma edad y el mismo promedio de calificaciones durante sus estudios; ninguno tenía experiencia profesional y todos se graduaron en

TABLA 12.14 Tabla ANOVA para un experimento factorial con dos factores completamente aleatorizados

Fuente de variación	gl	SC	CM	*Estadística F*
Factor A	$a-1$	$\frac{1}{nb}\sum_i T_{i\cdot\cdot}^2 - \frac{T_{\cdots}^2}{nab}$	$\text{SCA}/(a-1)$	CMA/CME
Factor B	$b-1$	$\frac{1}{na}\sum_j T_{\cdot j\cdot}^2 - \frac{T_{\cdots}^2}{nab}$	$\text{SCB}/(b-1)$	CMB/CME
Interacción AB	$(a-1)(b-1)$	$\frac{1}{n}\sum_i\sum_j T_{ij\cdot}^2 - \frac{1}{nb}\sum_i T_{i\cdot\cdot}^2 - \frac{1}{na}\sum_j T_{\cdot j\cdot}^2 + \frac{T_{\cdots}^2}{nab}$	$\text{SCAB}/(a-1)(b-1)$	CMAB/CME
Error	$ab(n-1)$	$\sum_i\sum_j\sum_k Y_{ijk}^2 - \frac{1}{n}\sum_i\sum_j T_{ij\cdot}^2$	$\text{SCE}/ab(n-1)$	
Total	$nab-1$	$\sum_i\sum_j\sum_k Y_{ijk}^2 - \frac{T_{\cdots}^2}{nab}$		

TABLA 12.15 Salarios iniciales para contadores graduados (miles de dólares)

	Noroeste	*Oeste medio*	*Oeste*	*Totales*
Mujeres	15.2	14.9	16.2	
	16.8	16.2	15.9	
	15.5	15.6	16.8	
	14.9	15.3	15.8	
	$T_{11\cdot} = 62.4$	$T_{21\cdot} = 62.0$	$T_{31\cdot} = 64.7$	$T_{\cdot 1\cdot} = 189.1$
Hombres	18.1	17.8	18.4	
	16.3	18.2	16.8	
	17.2	18.1	17.5	
	17.9	17.6	18.7	
	$T_{12} = 69.5$	$T_{22\cdot} = 71.7$	$T_{32\cdot} = 71.4$	$T_{\cdot 2\cdot} = 212.6$
Totales	$T_{1\cdot\cdot} = 131.9$	$T_{2\cdot\cdot} = 133.7$	$T_{3\cdot\cdot} = 136.1$	$T_{\cdot\cdot\cdot} = 401.7$

universidades del mismo nivel académico. Con base en la información de la muestra proporcionada en la tabla 12.15, determínense cuáles efectos son estadísticamente apreciables.

Las sumas de interés aparecen en la tabla. Entonces

$$\text{STC} = 15.2^2 + 16.8^2 + \cdots + 18.7^2 - \frac{401.7^2}{24} = 32.8563,$$

$$\text{SCTR} = \frac{62.4^2 + 69.5^2 + \cdots + 71.4^2}{4} - \frac{401.7^2}{24} = 24.7838,$$

$$\text{SCE} = 32.8563 - 24.7838 = 8.0725.$$

De manera similar,

$$\text{SC(SEX)} = \frac{189.1^2 + 212.6^2}{12} - \frac{401.7^2}{24} = 23.0104,$$

$$\text{SC(LOC)} = \frac{131.9^2 + 133.7^2 + 136.1^2}{8} - \frac{401.7^2}{24} = 1.11.$$

De esta forma

$$\text{SC(LOC} \times \text{SEX)} = 24.7838 - 23.0104 - 1.11 = 0.6634.$$

La tabla del análisis de varianza se encuentra en la tabla 12.16. Con base en esta información, puede concluirse que el único efecto discernible estadísticamente en el salario inicial se debe al sexo del graduado.

Debe notarse que el método de Scheffé para comparar las medias del nivel del factor se extiende, en forma directa, a experimentos factoriales. También puede

TABLA 12.16 Tabla ANOVA para el ejemplo 12.5

Fuente de variación	gl	SC	CM	*Valor F*
Localidad	2	1.11	0.555	1.24
Sexo	1	23.0104	23.0104	51.31
Localidad × sexo	2	0.6634	0.3317	0.74
Error	18	8.0725	0.4485	
Total	23	32.8563	$f_{0.99,\,1,\,18} = 8.29$; $f_{0.99,\,2,\,18} = 6.01$	

efectuarse un análisis de residuos para los niveles de cada factor para verificar, entre otras cosas, la hipótesis de varianzas iguales. Los residuos se obtienen mediante el empleo de la relación

$$e_{ijk} = y_{ijk} - \bar{y}_{ij}.$$

En los casos que se han examinado hasta este momento, siempre se empleó el cuadrado medio del error como el denominador del cociente F. Sin embargo, para experimentos estadísticos que incluyen dos o más factores, lo anterior no siempre es válido. La estadística F apropiada para un análisis de varianza depende, en forma directa, de las esperanzas de los cuadrados medios de las fuentes de variación, las que a su vez dependen de si se consideran a los efectos correspondientes como fijos o aleatorios.

Para experimentos factoriales con dos factores surgen tres situaciones distintas: *a*) los niveles de ambos factores son de efectos fijos; *b*) los niveles de ambos factores son de efectos aleatorios, o *c*) los niveles de un factor son de efectos fijos mientras que los del otro son de efectos aleatorios. Ya se ha analizado la primera posiblidad. Para las otras dos, los valores esperados de los cuadrados medios tanto para el modelo de efectos aleatorios como para el modelo de efectos mixtos se proporcionan en la tabla 12.17.

TABLA 12.17 Esperanzas de cuadrados medios para un factorial con dos factores: modelos de efectos aleatorios o de efectos mixtos

	Efectos aleatorios (A y B aleatorios)		*Efectos mixtos (A fijo, B aleatorio)*	
Fuente	ECM	*Estadística F*	ECM	*Estadística F*
A	$\sigma^2 + n\sigma^2_{\alpha\beta} + nb\sigma^2_{\alpha}$	CMA/CMAB	$\sigma^2 + n\sigma^2_{\alpha\beta} + nb\dfrac{\sum \alpha_i^2}{(a-1)}$	CMA/CMAB
B	$\sigma^2 + n\sigma^2_{\alpha\beta} + na\sigma^2_{\beta}$	CMB/CMAB	$\sigma^2 + na\sigma^2_{\beta}$	CMB/CME
AB	$\sigma^2 + n\sigma^2_{\alpha\beta}$	CMAB/CME	$\sigma^2 + n\sigma^2_{\alpha\beta}$	CMAB/CME
Error	σ^2		σ^2	

Con base en el material de este capítulo, el procedimiento que se ha empleado para construir la estadística de prueba es comparar dos cuadrados medios que, bajo la hipótesis nula, tengan el mismo valor esperado, y bajo la hipótesis alternativa, el cuadrado medio del numerador tenga un valor esperado mucho más grande que el del denominador correspondiente. Si la hipótesis nula es cierta, la estadística tiene una distribución F con un número apropiado de grados de libertad. Con esto en mente, los cocientes de cuadrados medios indicados en la tabla 12.17 deben ser ya evidentes. Por ejemplo, considérese el caso de efectos aleatorios y, en particular, la hipótesis nula de que no existe variación alguna entre todos los posibles niveles de A; esto es, H_0: $\sigma_\alpha^2 = 0$. Si H_0 es cierta, entonces $E(CMA) = \sigma^2 + n\sigma_{\alpha\beta}^2$, donde $\sigma_{\alpha\beta}^2$ denota la varianza causada por la interacción entre A y B. Pero este valor esperado es el mismo sólo para E *(CMAB)* y no para E *(CME)* bajo H_0. Por otro lado, si H_0 es falsa, E (*CMA*) es mayor que E (*CMAB*). De acuerdo con lo anterior, la estadística de prueba apropiada para H_0 es *CMA/CMAB*.

Debe recordarse que en experimentos factoriales, el cuadrado medio del error será el denominador en el cociente de cuadrados medios para todos los efectos principales y de interacción, sólo si los niveles de todos los factores son de efectos fijos. De esta forma, en la fase de diseño de un experimento estadístico es muy importante la selección de los niveles del factor, ya que tienen una influencia directa en el análisis.

Referencias

1. W. G. Cochran and G. M. Cox, *Experimental designs*, 2nd ed., Wiley, New York, 1957.
2. R. C. Hicks, *Fundamental concepts in the design of experiments*, 2nd ed., Holt, Rinehart and Winston, New York, 1973.
3. R. L. Horton, *The general linear model*, McGraw-Hill, New York, 1978.
4. R. E. Kirk, *Experimental design: Procedures for the behavioral sciences*, Brooks/Cole, Belmont, Calif., 1968.
5. J. Neter and W. Wasserman, *Applied linear statistical models*, Richard D. Irwin, Homewood, Ill., 1974.
6. H. Scheffé, *Analysis of variance*, Wiley, New York, 1953.
7. H. Scheffé, *A method for judging all contrasts in the analysis of variance*, Biometrika **40** (1953), 87–104.

Ejercicios

12.1. Suponga que se asigna al lector la responsabilidad de investigar en una fábrica el efecto que pueden tener diferentes cambios en la semana de 40 horas de trabajo, sobre la productividad promedio en una gran fábrica. En forma específica, se desean comparar cinco días a la semana, 4 días a la semana y 3⅓-días a la semana. Describa con gran detalle su propuesta de diseño estadístico. Asegúrese de identificar los tratamientos, las unidades experimentales y otros factores importantes para llevar a cabo la investigación.

12.2. Las estadísticas para accidentes indican que alrededor de dos terceras partes de los accidentes automovilísticos de consecuencias fatales en Estados Unidos son causados por conductores en estado de ebriedad. Supoga que usted es comisionado para investigar el

grado en el que el alcohol afecta la habilidad de las personas para desempeñar funciones de rutina al conducir un automóvil. Descríbase con gran detalle un diseño estadístico para lograr esta tarea e indíquese cómo debe llevarse a cabo este experimento.

12.3. Una compañía de seguros desea determinar si existen diferencias discernibles en el número de días promedio que los pacientes que padecen una misma enfermedad permanecen en cuatro grandes hospitales de cierta área metropolitana. La compañía también está interesada en detectar cualquier efecto debido al sexo de los pacientes. Descríbase con detalle un diseño estadístico para lograr este objetivo. Asegúrese de identificar la naturaleza de cada factor, ya sea como de efecto fijo o aleatorio; escríbase el modelo y establézcase la hipótesis por probar.

12.4. Una operación de llenado tiene tres máquinas idénticas que se ajustan para vaciar una cantidad específica de un producto en recipientes de igual tamaño. Con el propósito de verificar la igualdad de las cantidades promedio vaciadas por cada máquina, se toman muestras aleatorias, en forma periódica, de cada una. Para un periodo particular, se observaron los datos que aparecen en la tabla 12.18.

TABLA 12.18 Datos de la muestra para el ejercicio 12.4

Máquina		
A	B	C
16	18	19
15	19	20
15	19	18
14	20	20
	19	19
	19	

a) Calcúlese $y_{ij} - \bar{y}_{.j}$ y verifíquese que la suma de estas desviaciones para toda i y j es cero.
b) Estímese τ_j para toda j, y verifíquese que la suma de $n_j(\bar{y}_{.j} - \bar{y}_{..})$ sobre todas las j es cero.
c) Calcúlese, en forma directa, cada una de las tres sumas de cuadrados dadas en la expresión 12.8 para verificar que $STC = SCTR + SCE$.
d) ¿Existen algunas diferencias estadísticamente significativas en las cantidades promedio vaciadas por las tres máquinas? Empléese $\alpha = 0.05$.

12.5. En el ejercicio 12.4, supóngase que se divide cada observación entre 10. Demuéstrese si esta operación tiene algún efecto con las respuestas a las partes *c* y *d*.

12.6. Para el ejercicio 12.4, constrúyanse constrastes a su elección y empléese el método de Scheffé para determinar si éstos son estadísticamente significativos.

12.7. Se pide a un laboratorio de prueba independiente que compare la durabilidad de cuatro diferentes marcas de pelotas de golf. El laboratorio propone un experimento en el que se seleccionan, en forma aleatoria ocho pelotas por cada fabricante y se ponen en una máquina que golpea cada pelota con una fuerza constante. La medición de interés es el número de veces que la máquina golpea la pelota antes de que su recubrimiento externo se rompa. En la tabla 12.19 se encuentra la información que se obtuvo al llevar a cabo el experimento.

TABLA 12.19 Datos de la muestra para el ejercicio 12.7

Marca			
A	B	C	D
205	242	237	212
229	253	259	244
238	226	265	229
214	219	229	272
242	251	218	255
225	212	262	233
209	224	242	224
204	247	234	245

a) ¿Existe alguna razón para creer que la durabilidad promedio es diferente para cada una de las cuatro marcas? Úsese $\alpha = 0.05$.
b) ¿Existe alguna razón para dudar de la suposición de que las varianzas de los errores son iguales?

12.8. Para determinar si existen diferencias en la cosecha promedio de tres variedades de maíz, se dividió en tres partes iguales un área para siembra. A su vez, cada una de estas partes se subdivide en otras cinco iguales entre sí, y se siembra cada una con una variedad de maíz. En el momento de la cosecha, la medición de interés es el número de toneladas por acre. La tabla 12.20 es una tabla de análisis de varianza incompleta para este problema.

TABLA 12.20. Tabla parcial ANOVA para el ejercicio 12.8

Fuente	gl	SC	CM	*Valor F*
Tratamientos		64		
Error				
Total		100		

a) Escríbase el modelo para este problema.
b) ¿Se está satisfecho con las suposiciones? Hágase un comentario.
c) Establézcase la hipótesis nula por probar.
d) Complétese la tabla ANOVA y determínese si puede rechazarse la hipótesis nula para un nivel $\alpha = 0.01$.

12.9. Se desea determinar si la cantidad de carbón empleado en la fabricación de acero tiene algún efecto en la resistencia a la tensión de éste. Se investigaron cinco diferentes porcentajes de carbón: 0.2, 0.3, 0.4, 0.5 y 0.6%. Para cada porcentaje de carbón se seleccionaron, en forma aleatoria del mismo lote, cinco muestras de acero y se midieron las resistencias a la tensión. Se obtuvo la información que se muestra en la tabla 12.21, donde la tensión se encuentra en kilogramos por centímetro cuadrado.

a) Con base en esta información, determínese si el porcentaje de carbón tiene un efecto estadísticamente significativo sobre la resistencia a la tensión del acero. Úsese $\alpha = 0.01$.
b) Si la respuesta a la parte *a* es afirmativa, propónganse los contrastes relevantes y pruébese su significancia estadística.

TABLA 12.21 Datos de la muestra para el ejercicio 12.9

Contenido de carbón				
0.2%	0.3%	0.4%	0.5%	0.6%
1240	1420	1480	1610	1700
1350	1510	1470	1590	1790
1390	1410	1520	1580	1740
1280	1530	1540	1630	1810
1320	1470	1510	1560	1730

12.10. En el ejercicio 12.9, ¿existe alguna razón para dudar de la suposición de varianzas iguales?

12.11. Se seleccionó una muestra al azar de un número de presidentes de compañías, en cuatro diferentes áreas geográficas de Estados Unidos, con el propósito de determinar si el área tiene algún efecto sobre los ingresos anuales de estos altos ejecutivos. Se observaron los salarios anuales que se muestran en la tabla 12.22. Con la información dada, proporciónese un argumento, ya sea en contra o a favor, de si debe utilizarse la técnica del análisis de varianza para determinar si el área tiene algún efecto sobre el ingreso anual. Trátese de dar un apoyo sustancial en cualquiera de los dos casos.

TABLA 12.22 Datos de la muestra para el ejercicio 12.11 (miles de dólares)

Área			
Noreste	*Oeste medio*	*Sureste*	*Oeste*
140	93	78	85
125	135	112	72
95	68	57	97
110	53	97	105
59	115	52	62

12.12. En una planta industrial se desea determinar si diferentes trabajadores con el mismo nivel de habilidad tienen algún efecto sobre el número de unidades que se espera que produzcan durante un periodo fijo. Se lleva a cabo un experimento en el que se seleccionan al azar cinco trabajadores y se observa el número de unidades que cada uno produce en seis periodos con la misma duración, produciéndose los resultados que se encuentran en la tabla 12.23.

TABLA 12.23 Datos de la muestra para el ejercicio 12.12

Trabajador				
1	2	3	4	5
45	52	39	57	48
47	55	37	49	44
43	58	46	52	55
48	49	45	50	53
50	47	42	48	49
44	57	41	55	52

a) Escríbase el modelo para este problema y explíquese cada término.
b) Establézcase la hipótesis nula por probar.
c) Determínese si puede rechazarse la hipótesis nula para un nivel $\alpha = 0.05$.
d) ¿Qué fracción de la varianza en el número de unidades producidas es atribuible a diferencias entre los trabajadores?

12.13. Desde el incremento en los precios de la gasolina se han desarrollado varios dispositivos, los cuales se colocan en los carburadores de los automóviles, con el propósito de aumentar el rendimiento de éstos. Una empresa selecciona tres de los dispositivos más populares para someterlos a prueba. La empresa desea compararlos con los carburadores estándar, con el propósito de determinar si existe un incremento apreciable de millas por galón de gasolina con el uso de estos dispositivos. La compañía selecciona cinco tipos de automóviles para el experimento. Para controlar la variación, se planea utilizar el mismo conductor para todo el experimento.

TABLA 12.24 Datos de la muestra para el ejercicio 12.13 (millas por galón)

Automóvil	*Carburador estándar*	*Dispositivo* A	*Dispositivo* B	*Dispositivo* C
1	18.2	18.9	19.1	20.4
2	27.4	27.9	28.1	29.9
3	35.2	34.9	35.8	38.2
4	14.8	15.2	14.9	17.3
5	25.4	24.8	25.6	26.9

a) Hágase un bosquejo del plan específico para realizar este experimento.
b) Supóngase que se observan los datos que se encuentran en la tabla 12.24. Escríbase el modelo y establézcase la hipótesis nula por probar. ¿Puede rechazarse la hipótesis nula para un nivel $\alpha = 0.05$.
c) Si se rechaza la hipótesis nula de la parte *b*, constrúyanse por lo menos dos contrastes relevantes y pruébese su significancia estadística.

12.14. En el ejercicio 12.13, supóngase que no se ha considerado el automóvil como una fuente viable de variación en el rendimiento observado y muéstrese si esta omisión tiene algún efecto con la respuesta a la parte *b*.

12.15. Los cigarrillos producen cantidades apreciables de monóxido de carbono. Cuando se inhala el humo del cigarrillo, el monóxido de carbono se combina con la hemoglobina para formar carboxihemoglobina. En un estudio reciente,* los investigadores deseaban determinar si una concentración apreciable de carboxihemoglobina reduce la tolerancia al ejercicio en aquellos pacientes que sufren de bronquitis crónica y enfisema. Se seleccionaron siete** de estos pacientes y, en un ambiente controlado, se les pidió que caminaran durante 12 minutos respirando una de las siguientes cuatro mezclas gaseosas: aire, oxígeno, aire más monóxido de carbono (CO) u oxígeno más monóxido de carbono. La cantidad de monóxido de carbono respirado fue suficiente para elevar la concentración de carboxihemoglobina de cada sujeto en 9%. Para controlar el consumo de monóxido de carbono, se pidió a los siete fumadores que dejaran de fumar 12

*P. M. A. Calverly, R. J. E. Leggett, and D. C. Flenley, *Carbon monoxide and exercise tolerance in chronic bronchitis and emphysema*, Brit. Med. J. **283** (1981), 877–880.

** El estudio completo se llevó a cabo con 15 sujetos.

horas antes del experimento. Los datos que figuran el la tabla 12.25 representan las distancias caminadas por los sujetos en 12 minutos para cada condición.

TABLA 12.25 Datos de la muestra para el ejercicio 12.15 (en litros)

	Mezcla gaseosa			
Sujeto	*Aire*	*Oxígeno*	*Aire* + CO	*Oxígeno* + CO
1	835	874	750	854
2	787	827	755	829
3	724	738	698	726
4	336	378	210	279
5	252	315	168	336
6	560	672	558	642
7	336	341	260	336

a) Escríbase el modelo para este problema.
b) ¿Puede rechazarse la hipótesis nula de que no existe algún efecto, debido a la mezcla de gas, en la distancia caminada durante el lapso de 12 minutos para un nivel de $\alpha = 0.05$?
c) Llévese a cabo una prueba F conservadora para la hipótesis nula. ¿Es la conclusión diferente a la de la parte *b*?
d) Si la respuesta a la parte *b* es sí, constrúyanse los contrastes pertinentes y empléese el método de Scheffé para determinar si éstos son estadísticamente significativos.

12.16. Se desea determinar si existen diferencias apreciables en los precios promedio entre cuatro grandes supermercados en una ciudad dada. De los artículos de la misma marca que se venden con regularidad, se seleccionan al azar 10 y se observan sus precios unitarios en cada supermercado. Se obtiene la información que figura en la tabla 12.26.

a) Escríbase el modelo para este problema.
b) Establézcase una hipótesis nula apropiada y determínese si ésta puede rechazarse para un nivel de $\alpha = 0.01$.
c) Determínense todos los residuos y hágase la gráfica de éstos para cada tratamiento y para cada bloque. Hágase un comentario sobre sus resultados.

TABLA 12.26 Datos de la muestra para el ejercicio 12.16 (en dólares)

	Supermercado			
Artículo	A	B	C	D
1	3.29	3.42	3.27	3.35
2	0.59	0.65	0.59	0.60
3	1.25	1.29	1.25	1.27
4	4.35	4.59	4.29	4.49
5	0.89	0.95	0.89	0.89
6	1.85	1.79	1.89	1.89
7	0.95	0.89	0.89	0.90
8	0.75	0.79	0.69	0.79
9	2.35	2.35	2.39	2.39
10	1.49	1.55	1.55	1.49

12.17. En el ejemplo que sirvió como introducción en la sección 12.6, supóngase que se seleccionan en forma aleatoria 12 componentes del mismo lote y en grupos de tres se asig-

nan a las cuatro combinaciones de hornos y temperaturas. Los tiempos de duración de los componentes se encuentran en la tabla 12.27.

TABLA 12.27 Datos de la muestra para el ejercicio 12.17 (en horas)

	O_1	O_2
T_1	6.29	5.95
	6.38	6.05
	6.25	5.89
T_2	5.80	6.32
	5.92	6.44
	5.78	6.29

a) Escríbase el modelo apropiado para este problema.
b) Establézcase la hipótesis por probar.
c) Determínese la tabla del análisis de varianza y obténganse conclusiones apropiadas. Empléese $\alpha = 0.05$.

12.18. En el ejercicio 12.3, supóngase que se obtuvo la información proporcionada en la tabla 12.28 para pacientes seleccionados al azar, que padecen la misma enfermedad.

TABLA 12.28 Datos de la muestra para el ejercicio 12.18. Duración de la hospitalización en días en cuatro hospitales.

	Hospital A	*Hospital* B	*Hospital* C	*Hospital* D
Hombres	7	9	10	6
	10	9	8	7
	8	12	12	6
	11	14	13	9
Mujeres	9	11	13	8
	12	12	11	9
	12	14	14	8
	11	13	14	10

a) Determínese qué efectos son estadísticamente discernibles a un nivel de $\alpha = 0.01$.
b) Determínense todos los residuos y hágase la gráfica de éstos para cada hospital. ¿Qué conclusión puede dar?

12.19. El objetivo de un experimento de agricultura fue determinar si existían diferencias apreciables en la cantidad de trigo cosechado, de entre cuatro variedades y tres tipos de fertilizantes. Para el experimento se encontró una área muy grande de siembra en la que las condiciones del suelo eran, prácticamente, homogéneas. El área fue dividida en 12 zonas de igual tamaño para las 12 combinaciones de variedad de trigo y tipo de fertilizante. Para medir el error experimental, cada zona se dividió a su vez en cuatro y cada una de éstas recibió el mismo tratamiento. Las tres clases de fertilizante se seleccionaron, en forma aleatoria, de entre un número relativamente grande de fertilizantes, pero el interés no se extendió más allá de las cuatro variedades de trigo seleccionadas para el experimento. En el momento de la cosecha se observaron los datos que aparecen en la tabla 12.29.

TABLA 12.29 Datos de la muestra para el ejercicio 12.19 (toneladas por acre)

	Variedad de trigo			
Fertilizante	A	B	C	D
1	35	45	24	55
	26	39	23	48
	38	39	36	39
	20	43	29	49
2	55	64	58	68
	44	57	74	61
	68	62	49	60
	64	61	69	75
3	97	93	89	82
	89	91	98	78
	92	82	85	89
	99	98	87	92

a) Escríbase el modelo apropiado para este problema.
b) Establézcase la hipótesis nula por probar.
c) Determínese la tabla de análisis de varianza y obténganse las conclusiones apropiadas. Úsese $\alpha = 0.05$.

12.20. En el ejercicio 12.19, ¿Cómo puede cambiar la respuesta a la parte *c*, si

a) ¿Se supone que las variedades son de efectos aleatorios, y los tipos de fertilizante son de efectos fijos?
b) ¿Se supone que ambos son de efectos fijos?
c) ¿Se supone que ambos son de efectos aleatorios?

CAPÍTULO TRECE

Análisis de regresión: el modelo lineal simple

13.1 Introducción

En el capítulo anterior se desarrollaron los criterios básicos para el diseño estadístico de experimentos. En este capítulo se examinarán las asociaciones cuantitativas entre un número de variables, lo que en la terminología estadística se conoce como *análisis de regresión*.

Aunque en muchas disciplinas se están realizando experimentos diseñados en forma estadística, la precisión en la comparación que en forma general se requiere, evita el empleo de estos diseños en muchas situaciones. Investigar el efecto simultáneo de varios factores con base en las técnicas del análisis de varianza requiere de la suposición de que los datos se han colectado en arreglos balanceados y que se llevaron a efecto los procedimientos de aleatorización adecuados. En forma obvia, lo anterior es deseable si puede cumplirse, pero muchas veces es impráctico. En realidad, a lo que en general se enfrenta el experimento es a un conjunto de datos que de manera común, no espera que hayan sido observados bajo condiciones estrictamente controladas y los que, salvo en ciertas ocasiones, no contienen ninguna réplica real que permita una estimación apropiada del error experimental. Bajo estas condiciones, los métodos más apropiados son el de mínimos cuadrados y el análisis de regresión, y no los del análisis de varianza.

El propósito de este capítulo radica en proporcionar los conceptos y metodología básicos para extraer de grandes cantidades de datos las características principales de una relación que no es evidente. De manera específica, se examinarán técnicas que permitan ajustar una ecuación de algún tipo al conjunto de datos dado, con el propósito de obtener una ecuación empírica de predicción razonablemente precisa y que proporcione un modelo teórico que no está disponible. Se supondrá la existencia de un conjunto de n mediciones $y_1, y_2, \ldots, y_n$ de una variable respuesta Y, las cuales se han observado bajo un conjunto de condiciones experimentales $(x_1, x_2, \ldots, x_k)$ que representan los valores de k variables de predicción. El interés recae en determinar una función matemática sencilla, por ejemplo un polinomio que describa, en

forma razonable, el comportamiento de la variable respuesta, dados los valores de las variables de predicción. Nótese que la ecuación que se obtiene por esta forma puede tener algunas limitaciones con respecto a su interpretación física; sin embargo, en un medio empírico, será muy útil si puede proporcionar una adecuada capacidad de predicción para la respuesta en el interior de una región especificada de las variables de predicción.

A pesar de que no se encuentra problema alguno con las designaciones comunes de variable dependiente e independiente para Y y x, respectivamente, se preferirá denominarlas como variable de *respuesta* y de *predicción*, ya que en la regresión sólo puede *asociarse* un valor de Y con uno de predicción x; no es posible establecer una relación causa-efecto entre la Y y las x. Algunos ejemplos proporcionarán una idea del por qué obtener una relación causa-efecto se encuentra más allá del alcance del análisis de regresión. De manera obvia, existe una relación entre la altura y el peso de los seres humanos, pero ¿implica esta relación, por ejemplo, que pueda cambiar la altura de una persona si se modifica su peso? También se tiene una relación entre la cantidad de gas bruto que se consume en cierta área de alguna ciudad y la temperatura atmosférica promedio, pero ¿significa esto que es posible aumentar la temperatura mediante la reducción del consumo de gas? También puede existir alguna relación entre un factor económico en particular y un ciclo financiero, pero ¿implica lo anterior que el factor económico "causa" el ciclo financiero?

La esencia de los ejemplos anteriores está en el hecho de que el análisis de regresión sólo descubre una asociación entre la variable de respuesta y las variables de predicción, en lugar de detectar una relación causa-efecto. La causalidad implica que un cambio en las x causará uno correspondiente en la variable respuesta. Por ejemplo, cuando se calienta un metal éste se expande; en este caso no existe ninguna duda de que establecer una relación causa-efecto es muy importante. Pero en forma desafortunada, lo anterior generalmente no puede llevarse a cabo con base en un análisis estadístico, a menos que se efectúe un experimento rigurosamente controlado. Un ejemplo de lo anterior, es la relación que existe entre fumar y el cáncer pulmonar. La evidencia que se tiene resulta abrumadora con respecto a que el fumador crónico (predicción) está estadísticamente ascociado con una alta incidencia de cáncer pulmonar (respuesta). La industria cigarrera argumenta, en contra de estos hallazgos, que todavía no existe una relación de tipo causal entre fumar mucho y la incidencia de cáncer pulmonar.

El enfoque que se utilizará en este capítulo, así como en el siguiente, se limitará a establecer el grado de asociación que existe entre variables, sin tomar en cuenta la noción de causalidad. En este capítulo se examinarán los fundamentos del análisis de regresión para el modelo con una sola variable de predicción. En el capítulo 14 se estudiará lo que se conoce como el *modelo lineal general* en el que se supone que una respuesta dada es una función de varias variables de predicción.

13.2 El significado de la regresión y suposiciones básicas

Si los métodos de regresión son tan útiles en la práctica, debe comprenderse su significado y las suposiciones bajo las cuales se han desarrollado. Las técnicas de regre-

sión proporcionan medios legítimos a través de los cuales pueden establecerse asociaciones entre las variables de interés en las cuales la relación usual no es causal. La palabra "regresión" se usó por primera vez en este contexto por Francis Galton (1822-1911) en sus estudios biológicos sobre la herencia. En ellos se notó que las características promedio de la siguiente generación de un grupo en particular tendían a moverse en la dirección de las características promedio de la población general, más que hacia las de la generación previa de ese grupo. Esta tendencia fue referida como una regresión hacia la media de la población.

De manera básica, la regresión tiene dos significados: uno surge de la distribución conjunta de probabilidad de dos variables aleatorias; el otro es empírico y nace de la necesidad de ajustar alguna función a un conjunto de datos. Para ilustrar el primer significado se tratará de predecir el salario anual de un profesionista dado el número de años que han transcurrido desde su graduación. Sea X el número de años y Y el salario anual. Debe ser obvio que para un valor dado de x es imposible predecir, de manera exacta, el salario anual de una persona en particular. Sin embargo, es posible predecir el salario promedio de todos aquellos individuos para los que el número de años x que han transcurrido desde su graduación es el mismo. En otras palabras, para cada valor de x existe una distribución de ingresos anuales y lo que se busca es la media de esa distribución, dado x. La gráfica de la media condicional $E(Y|x)$ como una función de x recibe el nombre de *curva de regresión* de Y sobre X. De esta forma, si $f(x, y)$ es la función de densidad conjunta de probabilidad de X y Y, y si $f(y|x)$ es la función de densidad condicional de Y dado x, se define la curva de regresión como

$$E(Y|x) = \int_{-\infty}^{\infty} yf(y|x)\, dy.$$

Ejemplo 13.1 Considérese la función de densidad conjunta de probabilidad dada por

$$f(x, y) = \begin{cases} 2x & 0 < x < y < 1, \\ 0 & \text{para cualquier otro valor} \end{cases}$$

Obténgase la curva de regresión de Y sobre X.

Dado que

$$f(y|x) = f(x, y)/f_X(x),$$

entonces

$$f_X(x) = \int_y f(x, y)\, dy = \int_x^1 2x\, dy = 2x(1 - x),$$

y

$$f(y|x) = \frac{2x}{2x(1 - x)} = \frac{1}{1 - x}.$$

Por lo tanto, la curva de regresión es

$$E(Y|x) = \int_x^1 (1 - x)^{-1} y \, dy = (1 + x)/2,$$

la cual es una línea recta con pendiente e intersección igual a 1/2.

El segundo significado de la regresión es mucho más práctico que el primero. En él no se tienen los elementos necesarios para determinar la curva de regresión tal como se hizo en el ejemplo 13.1. No obstante, dado un conjunto de datos, puede asumirse una forma funcional para la curva de regresión y entonces tratar de ajustar ésta a los datos. En estas situaciones, la variable respuesta es una variable aleatoria cuyos valores se observan mediante la selección de los valores de las variables de predicción en un intervalo de interés. Por lo tanto, las variables de predicción no se consideran como variables aleatorias, sino que éstas son un conjunto de valores fijos que representan los puntos de observación para la variable respuesta. El modelo de regresión propuesto debe ser relativamente sencillo y deberá contener pocos parámetros. Un procedimiento muy útil para la selección inicial cuando se tiene sólo una variable de predicción es graficar la variable respuesta contra la variable de predicción. Si esta gráfica revela una tendencia lineal, deberá suponerse un modelo de regresión lineal. Si es evidente alguna curvatura, deberá suponerse un modelo cuadrático o de mayor grado para ajustarse a los datos.

Una vez que se ha seleccionado el modelo, la siguiente tarea es la de obtener estimaciones para los parámetros que intervienen en el mismo. Una técnica muy aceptada para este propósito es el *método de mínimos cuadrados (MC)*. Este método encuentra las estimaciones para los parámetros en la ecuación seleccionada mediante la minimización de la suma de los cuadrados de las diferencias entre los valores observados de la variable respuesta y de aquéllos proporcionados por la ecuación de predicción. Estos valores se conocen como los estimadores por mínimos cuadrados (*EMC*) de los parámetros. Los estimadores por mínimos cuadrados poseen ciertas propiedades deseables, pero para determinarlas es necesario formular las siguientes suposiciones:

1. Se ha seleccionado la forma correcta de la ecuación de regresión. Esto implica que cualquier variabilidad en la variable respuesta que no pueda explicarse mediante el empleo de la ecuación de regresión, se debe a un error aleatorio. Por ejemplo, se sabe que la distancia d que recorre un objeto en un tiempo t, está dada por la siguiente relación

$$d = \beta_0 + \beta_1 t,$$

donde β_1 es la velocidad promedio y β_0 es la posición del objeto para $t = 0$. Si no fuese posible medir d en forma precisa para un valor dado de t, pero se observó un valor

$$y = d + \varepsilon,$$

donde ε es el error aleatorio, se ha seleccionado la forma correcta de la ecuación de regresión y el problema se reduce a estimar los valores de β_0 y β_1. Sin embargo, rara es la vez que el problema resulta ser tan sencillo.

Por ejemplo, si se tiene interés en predecir la cantidad de ozono que se encuentra en la estratósfera, como una función de los niveles de concentración de los constituyentes químicos de ésta en cierto momento del día, la ecuación por seleccionar será, en primera instancia, una conjetura. El error no puede considerarse como puramente aleatorio ya que pueden existir variaciones sistemáticas por causa de errores en el modelo. Algunos de los valores de la variable de respuesta proporcionados por la ecuación de predicción estarán sesgados ya que las estimaciones de los parámetros también se encuentran sesgadas.

2. Los datos que se observan son comunes en el sentido en que constituyen una muestra representativa de un medio acerca del cual el investigador desea generalizar. Si el investigador sabe que los datos no son representativos, el comportamiento general del mecanismo puede encontrarse más allá del alcance de los datos.

3. Los valores observados de la variable respuesta no se encuentran estadísticamente correlacionados. Se supone que cada valor observado está constituido por un valor real y una componente aleatoria. La componente aleatoria consiste en una variable aleatoria no observable; entonces la covarianza entre cualesquiera dos observaciones Y_i y Y_j, o entre los correspondientes errores aleatorios ε_i y ε_j, es cero para toda $\neq j$.

4. Para toda $i = 1, 2 \ldots n$, la media de ε_i es cero y la varianza de ε_i es σ^2. Esta última recibe el nombre de *varianza del error* y, generalmente, no es conocida. Dado que las variables de predicción no son variables aleatorias, la varianza de Y_i también es σ^2 para toda i y de esta forma es independiente del punto de observación. Si no es posible formular la suposición de que la varianza es constante para las observaciones de la variable respuesta, generalmente se emplea el método de mínimos cuadrados con factores de peso. Este tema se estudiará con cierto detalle en el capítulo 14.

5. Los puntos de observación o los valores de las variables de predicción son fijos o se seleccionan con anticipación y se miden sin error. Para muchas situaciones prácticas, ambas condiciones no se cumplen. Afortunadamente, el método de mínimos cuadrados sigue siendo válido siempre y cuando los errores en los valores de las x sean pequeños al compararse con los errores aleatorios y dado que éstos no dependan de los parámetros del modelo.

A manera de comentario final sobre las suposiciones del procedimiento *MC*, se considerarán sólo mínimos cuadrados lineales, donde la palabra "lineal" significa que el modelo seleccionado es lineal en los parámetros. La frase "lineal en los parámetros" significa que ningún parámetro en el modelo aparece como un exponente o es multiplicado por o dividido entre cualquier otro parámetro. Por ejemplo, los modelos

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon,$$

$$Y = \beta_0 + \beta_1 \ln(x) + \varepsilon,$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

son lineales en los parámetros β_0, β_1, β_2, y β_3, pero el modelo

$$Y = \beta_0 \exp(\beta_1 x) + \varepsilon$$

no lo es debido a que el parámetro β_1 aparece como un exponente.

13.3 Estimación por mínimos cuadrados para el modelo lineal simple

En esta sección se estudiará la estimación por mínimos cuadrados para el modelo lineal simple en el que sólo se tiene una variable de predicción, y se supone una ecuación de regresión lineal. Por ejemplo, los estudiantes universitarios que aprenden más rápido tienen mejores calificaciones promedio *(CP)* y por lo tanto, mejores oportunidades de obtener buenos empleos después de graduarse. Supóngase que los datos que se encuentran en la tabla 13.1 representan las calificaciones promedio de 15 recién graduados y sus correspondientes salarios iniciales.

Para este ejemplo, la variable respuesta es el salario inicial y la variable de predicción potencial es la calificación promedio. Estas últimas se seleccionaron de tal manera que reflejen un amplio intervalo. Se desea determinar una ecuación de regresión para el salario inicial promedio como una función de la calificación promedio. Dado que se ha propuesto sólo una variable de predicción, graficar los datos puede ser útil en la selección inicial de un modelo de regresión. La gráfica de los salarios iniciales contra las calificaciones promedio se muestra en la figura 13.1. Debe notarse que esta gráfica fue realizada por un paquete estadístico para computadora conocido como "minitab". Aunque no es tan sofisticado como SAS, Minitab es muy fácil de usar y se recomienda para llevar a cabo análisis preliminares de regresión, entre otras aplicaciones.

TABLA 13.1 Datos de la muestra para un modelo lineal simple (miles de dólares)

CP	*Salario inicial*
2.95	18.5
3.20	20.0
3.40	21.1
3.60	22.4
3.20	21.2
2.85	15.0
3.10	18.0
2.85	18.8
3.05	15.7
2.70	14.4
2.75	15.5
3.10	17.2
3.15	19.0
2.95	17.2
2.75	16.8

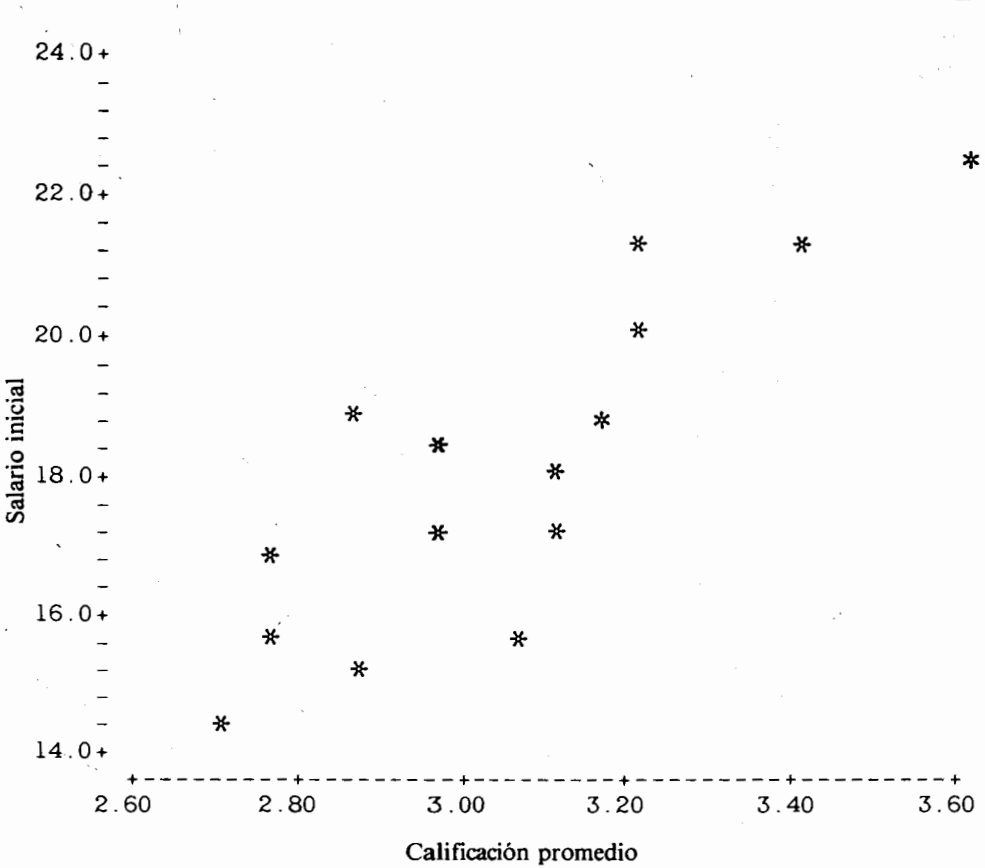


TABLA 13.1 Salario inicial contra calificación promedio

A pesar de que esta gráfica muestra una gran dispersión,* se observa una tendencia lineal. De acuerdo con lo anterior se supondrá un modelo de la forma

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, 2, \ldots, n, \tag{13.1}$$

donde Y_i es la i-ésima observación de la variable respuesta, la cual corresponde al i-ésimo valor x_i de la variable de predicción, ε_i es el error aleatorio no observable asociado con Y_i; y β_0 y β_1 son los parámetros desconocidos que representan la intersección y la pendiente, respectivamente. La expresión (13.1) se conoce como *modelo lineal simple,* debido a que es lineal en los parámetros y se tiene sólo una variable de predicción.

Cada observación Y_i es una variable aleatoria que es la suma de dos componentes; el término no aleatorio $\beta_0 + \beta_1 x_i$, y la componente aleatoria ε_i. Si ε_i fuera un

* Por esta razón, este tipo de gráfica se conoce como gráfica de dispersión.

valor igual a cero, la observación Y_i se encontraría precisamente sobre la línea de regresión $\beta_0 + \beta_1 x_i$. Por lo tanto, ε_i es la distancia vertical de la observación a la línea de regresión. Dado que se supone

$$E(\varepsilon_i) = 0, \quad Var(\varepsilon_i) = \sigma^2 \qquad i = 1, 2, \ldots, n,$$

y

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \qquad i \neq j;$$

entonces

$$E(Y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i,$$

$$Cov(Y_i, Y_j) = \sigma^2 \qquad i \neq j,$$

y

$$Var(Y_i) = Var(\beta_0 + \beta_1 x_i + \varepsilon_i) = Var(\varepsilon_i) = \sigma^2.$$

El último resultado surge del hecho de que la varianza de una variable aleatoria no varía con respecto a la localización; en este caso, el corrimiento en localización está proporcionado por el término no aleatorio $\beta_0 + \beta_1 x_i$. Por lo tanto, en términos reales, lo que se supone es que para cada calificación promedio x existe una distribución de probabilidad para los salarios iniciales cuya media es una función lineal de x y cuya varianza es la misma para toda x. El modelo proporcionado por (13.1) debe considerarse sólo como una selección inicial para la forma funcional de la curva de regresión. Con base en análisis más apropiados, los cuales se examinarán más adelante, puede ser necesario hacer ajustes y éstos a su vez pueden dar como resultado una ecuación final de predicción diferente de la del modelo inicial.

Para obtener los estimadores de mínimos cuadrados de β_0 y β_1, se generalizará un conjunto de datos consistente en n pares (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), donde los valores de y son las observaciones de la variable aleatoria respuesta. El método de mínimos cuadrados considera la desviación de la observación Y_i de su valor medio y determina los valores de β_0 y β_1 que minimizan la suma de los cuadrados de estas desviaciones. La i-ésima desviación o error es

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i), \tag{13.2}$$

y la suma de los cuadrados de los errores es

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2. \tag{13.3}$$

Los estimadores de mínimos cuadrados de β_0 y β_1 se obtienen mediante la diferenciación de (13.3) con respecto a β_0 y β_1 y después al igualar cada derivada parcial con cero, es decir

$$\frac{\partial \sum \varepsilon_i^2}{\partial \beta_0} = -2 \sum (Y_i - B_0 - B_1 x_i) = 0,$$

y

$$\frac{\partial \sum \varepsilon_i^2}{\partial \beta_1} = -2 \sum x_i(Y_i - B_0 - B_1 x_i) = 0,$$

donde B_0 y B_1 son los estimadores de mínimos cuadrados* de β_0 y β_1, respectivamente. Al simplificar y distribuir las sumas en estas ecuaciones, se tiene

$$\sum_{i=1}^{n} Y_i = nB_0 + B_1 \sum_{i=1}^{n} x_i$$

y (13.4)

$$\sum_{i=1}^{n} x_i Y_i = B_0 \sum_{i=1}^{n} x_i + B_1 \sum_{i=1}^{n} x_i^2.$$

Las dos ecuaciones dadas por (13.4) se conocen como *ecuaciones normales.*

Dadas las realizaciones $y_1, y_2, \ldots, y_n$, las ecuaciones pueden resolverse para los estimados de mínimos cuadrados b_0 y b_1. Si se dividen ambos mienbros de la primera ecuación entre n, se obtiene

$$\frac{\sum y_i}{n} = b_0 + b_1 \frac{\sum x_i}{n};$$

entonces el estimador de mínimos cuadrados de β_0 es

$$b_0 = \frac{\sum_{i=1}^{n} y_i}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}. \tag{13.5}$$

Al sustituir b_0 en la segunda ecuación de (13.4) se obtiene

$$\sum x_i y_i = \left(\frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n}\right) \sum x_i + b_1 \sum x_i^2,$$

la que, después de resolver para b_1, se reduce a

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{13.6}$$

* Muchos autores prefieren designar a los estimadores de mínimos cuadrados con letras cursivas minúsculas. Para mantener la consistencia de la notación con respecto a los capítulos anteriores se designará al estimador de mínimos cuadrados con una letra cursiva en mayúscula y el tipo en minúscula para el estimado *MC*.

Los valores dados por (13.5) y (13.6) son aquellos que minimizan la suma de los cuadrados de los errores.

Dados los estimadores de mínimos cuadrados B_0 y B_1 para la intersección y la pendiente, respectivamente, la recta de regresión estimada para el modelo (13.1) es

$$\hat{Y}_i = B_0 + B_1 x_i \tag{13.7}$$

donde $\hat{Y}_i$ es el estimador para la media de la observación Y_i, la cual corresponde al valor x_i de la variable de predicción. Nótese que si se sustituye (13.5) por B_0 en (13.7) se obtiene una forma alternativa para la recta de regresión estimada, la cual se encuentra dada por

$$\begin{aligned}\hat{Y}_i &= \overline{Y} - B_1\overline{x} + B_1 x_i \\ &= \overline{Y} + B_1(x_i - \overline{x}).\end{aligned} \tag{13.8}$$

Con base en (13.2), la diferencia entre la realización y_i y el valor estimado $\hat{y}_i$ es un estimador del correspondiente error. Este estimador se conoce como el i-ésimo residual y se denota por

$$e_i = y_i - \hat{y}_i. \tag{13.9}$$

De nuevo, nótese que los residuos no son estimados en el sentido clásico de la estimación de parámetros (fijos), sino que son estimadores de los valores de las variables aleatorias no observables ε_i, los cuales se obtienen de la recta de regresión estimada. Los residuos $e_1, e_2, \ldots, e_n$ son muy importantes debido a que proporcionan una abundante información sobre lo que puede faltar del modelo de regresión estimado. Más adelante se darán más detalles con respecto a lo anterior. En este momento se ilustrarán los pesos de cálculo para obtener la recta de regresión estimada para el modelo lineal simple empleando para ello los datos de los salarios. El propósito de esto radica en familiarizar al lector únicamente con el procedimiento de cálculo. De lo contrario, se puede hacer uso de algún paquete estadístico para computadora. Posteriormente, se presentará un listado de computadora para este ejemplo.

En la tabla 13.2, se incluyen los cálculos básicos necesarios para obtener los estimadores de mínimos cuadrados de la intersección y la pendiente. Las últimas cuatro columnas de esta tabla no son necesarias para la determinación de b_0 y b_1, éstas serán empleadas después en otro contexto.

Mediante el empleo de (13.5) y (13.6) el estimador de mínimos cuadrados para la pendiente es

$$b_1 = \frac{830.425 - \dfrac{(45.6)(270.8)}{15}}{139.51 - \dfrac{(45.6)^2}{15}} = 8.12,$$

y el correspondiente estimado de mínimos cuadrados para la intersección es

$$b_0 = \frac{270.8}{15} - (8.12)\frac{45.6}{15} = -6.63.$$

TABLA 13.2 Cálculos básicos para obtener los estimadores de mínimos cuadrados b_0 y b_1 (con base en los datos de salarios dados en la tabla 13.1)

	CP x_i	*Salario* y_i	x_iy_i	x_i^2	y_i^2	*Salario estimado* $\hat{y}_i$	*Residuo* $y_i - \hat{y}_i$	*Cuadrado del residuo* $(y_i - \hat{y}_i)^2$
	2.95	18.5	54.575	8.7025	342.25	17.32	1.18	1.3924
	3.20	20.0	64.000	10.2400	400.00	19.35	0.65	0.4225
	3.40	21.1	71.740	11.5600	445.21	20.98	0.12	0.0144
	3.60	22.4	80.640	12.9600	501.76	22.60	−0.20	0.0400
	3.20	21.2	67.840	10.2400	449.44	19.35	1.85	3.4225
	2.85	15.0	42.750	8.1225	225.00	16.51	−1.51	2.2801
	3.10	18.0	55.800	9.6100	324.00	18.54	−0.54	0.2916
	2.85	18.8	53.580	8.1225	353.44	16.51	2.29	5.2441
	3.05	15.7	47.885	9.3025	246.49	18.13	−2.43	5.9049
	2.70	14.4	38.880	7.2900	207.36	15.29	−0.89	0.7921
	2.75	15.5	42.625	7.5625	240.25	15.70	−0.20	0.0400
	3.10	17.2	53.320	9.6100	295.84	18.54	−1.34	1.7956
	3.15	19.0	59.850	9.9225	361.00	18.95	0.05	0.0025
	2.95	17.2	50.740	8.7025	295.84	17.32	−0.12	0.0144
	2.75	16.8	46.200	7.5625	282.24	15.70	1.10	1.2100
Totales	45.6	270.8	830.425	139.5100	4970.12	270.79	0.01	22.8671

De acuerdo con lo anterior, la ecuación estimada de regresión es

$$\hat{y}_i = -6.63 + 8.12\,x_i. \tag{13.10}$$

Al intentar interpretar esta ecuación se tiene que los valores $\hat{y}_i$ son los estimadores para las medias de las distribuciones de probabilidad de los salarios iniciales correspondientes a las calificaciones promedio x_i. Tener una intersección negativa resulta fastidioso, ya que, por ejemplo, si $x = 0.5$, $\hat{y} = -2.57$, lo cual es absurdo. Pero las calificaciones promedio en este conjunto de datos varían de 2.70 a 3.60, por lo tanto, cualquiera que sea la validez que tiene la ecuación estimada de regresión al predecir los salarios iniciales promedio se mantiene, para todos aquellos valores de x que se encuentren entre 2.70 y 3.60. En la práctica, muchas veces se desea predecir la respuesta más allá del intervalo de valores de x para los cuales se obtuvo la ecuación estimada de regresión. Si un valor de x se encuentra muy cercano a este intervalo, la predicción tendrá cierta validez. De otra forma, ésta debe verse con mucho cuidado, ya que la ecuación de regresión estimada puede no ser apropiada para un intervalo de valores más amplio de la variable de predicción.

La interpretación del valor estimado de la pendiente es directa. El incremento estimado en el salario inicial promedio para cada aumento igual a una unidad de la calificación promedio es de 8 120 dólares.

La tercera columna de la derecha en la tabla 13.2, contiene los salarios promedio estimados para cada calificación promedio dada por (13.10). Por ejemplo, si $x = 2.95$, el salario inicial estimado promedio es $\hat{y} = -6.63 + 8.12(2.95) = (13.9)$. miles de dólares. Dado que el correspondiente valor observado es 18.5, de (13.9), $e = 18.5 - 17.32 = 1.18$ es el residuo para $x = 2.95$. En otras palabras, el valor resi-

dual 1.18 es la distancia vertical que existe entre la observación 18.5 y el punto sobre la recta estimada de regresión para $x = 2.95$. Los otros residuos se obtienen de la misma manera y tienen significados similares. La figura 13.2 ilustra los residuos como distancias verticales desde la recta de regresión estimada. Dado que un residuo representa la cantidad en la que un valor estimado falla para predecir la media de la correspondiente observación aleatoria, entre más grandes son las magnitudes de los residuos, mayor tenderá a ser el efecto de la componente aleatoria en el modelo.

Recuérdese que la varianza σ^2 de la variable respuesta es igual a la varianza del error y ésta es constante para todos los valores de la variable de predicción. En general, dado que el valor de σ^2 no se conoce, puede obtenerse un estimador de éste a partir de los estimados de mínimos cuadrados b_0 y b_1. Dado que cada $\hat{y}_i$ estima la media de Y_i, la diferencia $y_i - \hat{y}_i$ representa la desviación de Y_i con respecto a su propia media. La suma de los cuadrados de estas diferencias, dividida entre una constante apropiada, es la forma en la que se determina una varianza. Pero estas diferencias son los residuos; por lo tanto, la suma de los cuadrados de los residuos dividida entre una constante apropiada es un estimador de σ^2. La constante apropiada es $n - 2$, ya que se pierden dos grados de libertad al tener que estimar los dos parámetros β_0 y β_1 antes de obtener $\hat{y}_i$. El estimador de σ^2 se denota como s^2 y está dado por

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - 2} = \frac{\sum_{i=1}^{n} e_i^2}{n - 2}. \tag{13.11}$$

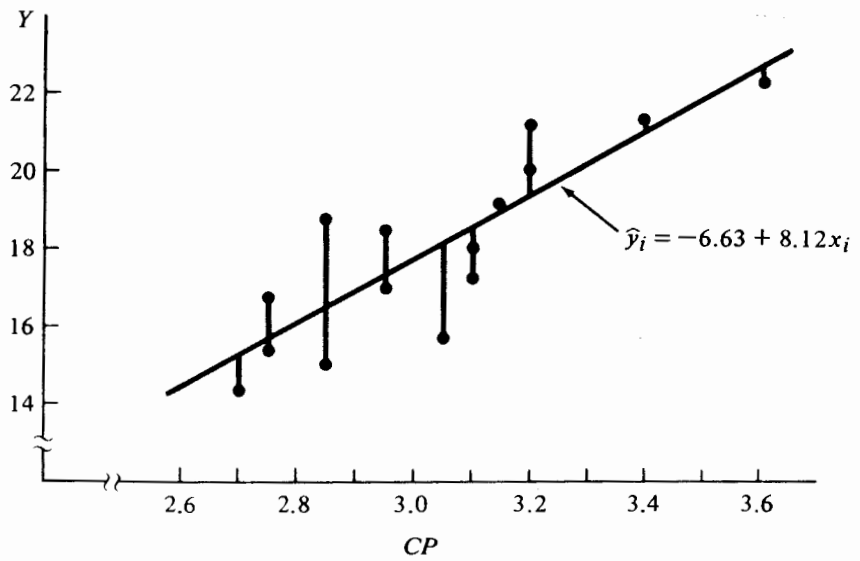


FIGURA 13.2 Residuos como distancias verticales desde la ecuación estimada de regresión

El estimador s^2 recibe el nombre de *varianza residual* o *CME,* y la raíz cuadrada positiva s se conoce como la *desviación estándar residual.* Para el ejemplo de los salarios iniciales, la varianza residual es $s^2 = 22.8671/13 = 1.759$. La varianza residual s^2 es una medida absoluta de qué tan bien se ajusta la recta estimada de regresión a las medias de las observaciones de la variable respuesta. Por lo tanto, en general entre más pequeño sea el valor de s^2, se ajustará mejor al modelo. Puede demostrarse que el estimador S^2 es un estimador no sesgado de σ^2 con tal de que la forma del modelo de regresión sea la correcta. De otra manera, S^2 estima σ^2 más una componente que es el sesgo causado por un error en el modelo.

Cuando se obtiene una recta de regresión por el método de mínimos cuadrados, surgen cierto número de propiedades. Algunas de éstas son las siguientes:

1. $\Sigma_{i=1}^{n} e_i = 0.$
2. $\Sigma_{i=1}^{n} y_i = \Sigma_{i=1}^{n} \hat{y}_i.$
3. $\Sigma_{i=1}^{n} x_i e_i = 0.$

Se demostrará la propiedad 1 y se dejan las correspondientes demostraciones de las propiedades restantes al lector. Debe notarse que la propiedad 2 se obtiene de la primer ecuación dada en (13.14) y la propiedad 3 de la segunda ecuación normal. Para la propiedad 1,

$$\begin{aligned}\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} (y_i - \hat{y}_i) \\ &= \sum (y_i - b_0 - b_1 x_i) \\ &= \sum y_i - nb_0 - b_1 \Sigma x_i \\ &= n\bar{y} - n(\bar{y} - b_1\bar{x}) - nb_1\bar{x} \\ &= 0.\end{aligned}$$

A causa de los errores de redondeo, la suma de los residuos dados en la tabla 13.2 no es exactamente igual a cero. Además, dado que los estimadores *MC* se obtienen mediante la minimización de la suma de los cuadrados de los errores, para este ejemplo el valor mínimo es 22.8671.

13.4 Estimación por máxima verosimilitud para el modelo lineal simple

Puede emplearse el principio de máxima verosimilitud para estimar los parámetros desconocidos en el modelo lineal simple dado por (13.1). Recuérdese que los estimadores de mínimos cuadrados se obtuvieron sin tener que especificar la distribución de probabilidad de los errores aleatorios ε_i. Si se supone que los ε_i son variables aleatorias independientes, normalmente distribuidas, con media cero y varianza σ^2 para toda $i = 1, 2, \ldots, n$, es posible obtener los estimadores de máxima verosimi-

litud de β_0, β_1, y σ^2,es decir, si además de las suposiciones previas se especifica que $\varepsilon_i \sim N(0, \sigma^2)$ para toda $i = 1, 2, \ldots, n$, entonces cada Y_i también se encuentra normalmente distribuida con media $\beta_0 + \beta_1 x_i$ y varianza σ^2, dado que ésta es una función lineal de una variable aleatoria con distribución normal. Los estimadores de máxima verosimilitud se obtienen mediante la maximización de la función de verosimilitud dada por

$$\begin{aligned} L(y_1, y_2, \ldots, y_n; \beta_0, \beta_1, \sigma^2) &= \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_1 - \beta_0 - \beta_1 x_1)^2 \right] \\ &\quad \cdots \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_n - \beta_0 - \beta_1 x_n)^2 \right] \\ &= \left(\frac{1}{\sqrt{2\pi}\,\sigma} \right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right], \end{aligned}$$

donde

$$\ln[L(\beta_0, \beta_1, \sigma^2)] = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum (y_i - \beta_0 - \beta_1 x_i)^2.$$

Al tomar las derivadas parciales con respecto a β_0, β_1, y σ^2, y después de igualarlas a cero, puede demostrarse que los estimadores de máxima verosimilitud de β_0 y β_1 son idénticos a los dados por (13.5) y (13.6), respectivamente, y el correspondiente a σ^2 está dado por

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}. \tag{13.12}$$

El estimador de máxima verosimilitud de σ^2 es sesgado pero, para valores grandes de n, la diferencia entre éste y el estimador de mínimos cuadrados no es importante.

El lector puede sorprenderse del por qué la necesidad de tratar con los estimadores de máxima verosimilitud, si éstos son iguales a los estimadores de mínimos cuadrados. Una de estas razones es que los estimadores de máxima verosimilitud tienen propiedades deseables de consistencia, suficiencia y varianza mínima. Además, éstos proporcionan los medios necesarios para el desarrollo de criterios de inferencia para β_0 y β_1.

La suposición de que los errores se encuentran normalmente distribuidos es justificable, debido a que la componente de error en el modelo es, en general, un efecto compuesto que representa muchas perturbaciones pequeñas pero aleatorias, las cuales son independientes de la variable de predicción y se deben a factores que no se encuentran incluidos en el modelo. En todo caso, las desviaciones de la suposición de normalidad para valores grandes de n no son, en general, serias.

13.5 Propiedades generales de los estimadores de mínimos cuadrados

En esta sección se desarrollarán algunas propiedades generales de los estimadores de mínimos cuadrados, por lo que se considerarán algunos criterios que permitan la construcción de intervalos de confianza y la realización de pruebas de hipótesis con respecto a los parámetros de regresión β_0 y β_1. Así mismo, se examinará la estimación de la respuesta media para una x dada y la predicción de una Y en particular para un valor dado de x. En gran medida, el enfoque de esta sección será de carácter teórico.

Considérense los estimadores no sesgados de β_0 y β_1 que son funciones lineales de las observaciones $Y_1, Y_2, \ldots, Y_n$. Si entre todos estos estimadores de β_0 y β_1 existen algunos cuyas varianzas son más pequeñas que las de todos los demás estimadores no sesgados de β_0 y β_1, entonces estos son los *mejores estimadores lineales no sesgados* (*MELI*) de β_0 y β_1. El siguiente teorema conocido generalmente como teorema de Gauss- Markov, garantiza que los estimadores de mínimos cuadrados de β_0 y β_1 son los *MELI* para β_0 y β_1.

Teorema 13.1 Sean las suposiciones para el modelo $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ las mismas que aquellas que se necesitan para la estimación de mínimos cuadrados de β_0 y β_1. Entonces los estimadores de mínimos cuadrados de B_0 y B_1 son los mejores estimadores lineales no sesgados de β_0 y β_1.

Mientras que la demostración del teorema 13.1 se encuentra más allá de los objetivos de este libro, se demostrará que B_0 y B_1 son combinaciones lineales de las observaciones $Y_1, Y_2, \ldots, Y_n$. Lo anterior permitirá demostrar que

$$E(B_1) = \beta_1$$

y

$$Var(B_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{13.13}$$

mientras que

$$E(B_0) = \beta_0$$

y

$$Var(B_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{13.14}$$

Para demostrar que B_1 es una combinación lineal de $Y_1, Y_2, \ldots, Y_n$, recuérdese la segunda expresión dada en (13.6). Primero, se desea demostrar que

$$\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} (x_i - \bar{x}) Y_i.$$

Lo anterior es verdadero, dado que

$$\sum (x_i - \bar{x})(Y_i - \bar{Y}) = \sum (x_i - \bar{x})Y_i - \bar{Y}\sum (x_i - \bar{x});$$

pero $\Sigma(x_i - \bar{x}) = 0$, y

$$\sum (x_i - \bar{x})(Y_i - \bar{Y}) = \sum (x_i - \bar{x})Y_i.$$

De acuerdo con lo anterior,

$$B_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

donde las x_i son fijas ya que son valores de una variable de predicción no aleatoria.

Sea

$$c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{13.15}$$

donde los c_i' son cantidades fijas, dado que las x_i' también son fijas. Entonces el estimador B_1 se expresa como

$$B_1 = \sum_{i=1}^{n} c_i Y_i,$$

la cual es una combinación lineal de las observaciones $Y_1, Y_2, \ldots, Y_n$.

Para demostrar que B_1 es un estimador no sesgado de β_1, se tiene

$$\begin{aligned} E(B_1) &= E\left(\sum_{i=1}^{n} c_i Y_i\right) \\ &= \sum c_i E(Y_i) \\ &= \sum c_i(\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i. \end{aligned}$$

Pero

$$\sum_{i=1}^{n} c_i = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 0,$$

y

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

De esta forma,

$$E(B_1) = \beta_1.$$

Dado que por hipótesis las observaciones Y_i no se encuentran correlacionadas por pares, $Cov(Y_i, Y_j) = 0$, $i \neq j$. Entonces, mediante el empleo de la segunda parte del teorema 6.1, se demuestra que la varianza de B_1 está dada por (13.13.) De esta forma se tiene

$$\begin{aligned} Var(B_1) &= Var\left(\sum_{i=1}^{n} c_i Y_i\right) \\ &= \sum c_i^2 Var(Y_i) \\ &= \sum c_i^2 \sigma^2 \\ &= \sigma^2 \sum c_i^2 \\ &= \sigma^2 \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}\right]^2 \\ &= \sigma^2 \left\{\sum (x_i - \bar{x})^2 \Big/ \left[\sum (x_i - \bar{x})^2\right]^2\right\} \\ &= \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \end{aligned}$$

La raíz cuadrada de $Var(B_1)$ es la desviación estándar* del estimador de mínimos cuadrados de la pendiente y está dada por

$$d.e.(B_1) = \frac{\sigma}{\left[\sum (x_i - \bar{x})^2\right]^{1/2}}.$$

Dado que, en general, la desviación estándar σ del error es desconocida, puede obtenerse un estimador de $d.e.(B_1)$ al reemplazar σ por la desviación estándar residual s, como está dada por (13.11). De esta forma, un estimado de la desviación estándar de B_1 es

* También conocida como error estándar.

$$s(B_1)^* = \frac{s}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^{1/2}}. \tag{13.16}$$

Ahora considérese el estimador de mínimos cuadrados de la intersección desconocida β_0. Dado que el estimador de mínimos cuadrados es

$$B_0 = \bar{Y} - B_1\bar{x},$$

y ya que el estimador de mínimos cuadrados de la pendiente es una combinación lineal de las observaciones $Y_1, Y_2, \ldots, Y_n$, entonces también B_0 es una combinación lineal de las observaciones. Para demostrar que B_0 es un estimador no sesgado de β_0, se tiene

$$\begin{aligned} E(B_0) &= E(\bar{Y} - B_1\bar{x}) \\ &= \frac{\sum_{i=1}^{n} E(Y_i)}{n} - \bar{x}\,E(B_1) \\ &= \frac{\sum (\beta_0 + \beta_1 x_i)}{n} - \beta_1\bar{x} \\ &= \frac{n\beta_0 + \beta_1 \sum x_i}{n} - \beta_1\bar{x} \\ &= \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} \\ &= \beta_0. \end{aligned}$$

Para demostrar que $Var(B_0)$ está dada por (13.14), de nuevo se empleará la segunda parte del teorema 6.1 y el hecho de que B_0 y B_1 son combinaciones lineales de variables aleatorias no correlacionadas. Dado que $B_0 = \bar{Y} - B_1\bar{x}$,

$$\begin{aligned} Var(B_0) &= Var(\bar{Y} - B_1\bar{x}) \\ &= Var\left(\frac{\sum Y_i}{n} - \bar{x}\sum c_i Y_i\right) \\ &= Var\left[\sum_{i=1}^{n}\left(\frac{Y_i}{n} - \bar{x}c_i Y_i\right)\right] \\ &= Var\left[\sum\left(\frac{1}{n} - \bar{x}c_i\right)Y_i\right] \end{aligned}$$

* Se empleará la notación más conveniente $s^2(T)$ y $s(T)$ para denotar, respectivamente, la varianza y la desviación estándar estimadas de un estimador. T.

$$= \sum \left(\frac{1}{n} - \bar{x}c_i\right)^2 Var(Y_i)$$

$$= \sigma^2 \sum \left(\frac{1}{n^2} - \frac{2\bar{x}c_i}{n} + \bar{x}^2 c_i^2\right)$$

$$= \sigma^2 \left(\frac{1}{n} - \frac{2\bar{x}}{n}\sum c_i + \bar{x}^2 \sum c_i^2\right).$$

Al sustituir (13.15) por c_i y al recordar que $\Sigma c_i = 0$, se tiene

$$Var(B_0) = \sigma^2 \left\{ \frac{1}{n} + \bar{x}^2 \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2} \right\}$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right]$$

Finalmente, si se sustituye $\bar{x}^2 = (\Sigma x_i)^2/n^2$, se obtiene

$$Var(B_0) = \sigma^2 \left[\frac{1}{n} + \frac{\left(\sum x_i\right)^2}{n^2 \sum (x_i - \bar{x})^2}\right]$$

$$= \sigma^2 \left[\frac{n \sum (x_i - \bar{x})^2 + \left(\sum x_i\right)^2}{n^2 \sum (x_i - \bar{x})^2}\right]$$

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

Entonces, un estimador de la desviación estándar de B_0 es

$$s(B_0) = s\left[\frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}\right]^{1/2} \qquad (13.17)$$

Es interesante notar que las varianzas de B_0 y B_1 son funciones de los valores x_i para los cuales se observa la variable respuesta. En particular, para el estimador de la pendiente B_1, $Var(B_1)$ tiene un valor máximo cuando $\Sigma(x_i - \bar{x})^2$ tiene un valor máximo. Pero $\Sigma(x_i - \bar{x})^2$ es máxima cuando la distancia entre los valores de x_i es la más grande. Esto ocurre cuando se escoge observar la respuesta sólo en los valores de los extremos del intervalo de variación de la variable de predicción, es decir, si verdaderamente el modelo de regresión es lineal, entonces deberán tomarse $n/2$ observaciones en un extremo y $n/2$ en el otro para obtener la mejor eficiencia posible al estimar la pendiente de la línea recta. Lo anterior es lógico ya que sólo se necesitan dos puntos para definir una línea recta; sin embargo, en la práctica, no es muy común el hecho de saber que la función de regresión es lineal de manera tal, que no sería prudente seleccionar los extremos del intervalo de x como puntos de observación y minimizar la varianza del estimador de la pendiente. Una alternativa más segura consiste en tener puntos de observación espaciados de igual forma sobre todo el intervalo de interés de la variable de predicción.

Para el modelo lineal simple, la recta de regresión estimada dada por (13.7) permite obtener un estimador para la media de la variable de respuesta para un valor específico de la variable de predicción. Sea x_p este valor en particular y para el cual se desea estimar la media de la variable respuesta Y_p. Entonces el estimador es $\hat{y}_p = b_0 + b_1 x_p$. Para el mismo conjunto de valores de x existe una variación muestra a muestra en el estimador $\hat{Y}_p$, dado que existe una variación del mismo tipo para los estimadores de mínimos cuadrados B_0 y B_1. Puede observarse que lo anterior es cierto para el ejemplo del salario inicial, ya que no se espera tener la misma recta de regresión estimada si se selecciona otro conjunto de estudiantes con las mismas *CP* que los primeros.

Considérese que la determinación de la media, y la varianza de $\hat{Y}_p$. $\hat{Y}_p$ es un estimador no sesgado de la media de Y_p, dado que

$$E(\hat{Y}_p) = E(B_0 + B_1 x_p) = \beta_0 + \beta_1 x_p = E(Y_p).$$

Para obtener la varianza de $\hat{Y}_p$, se hará uso de la misma técnica empleada para la varianza de B_0. De (13.8) se tiene

$$\begin{aligned} Var(\hat{Y}_p) &= Var[\bar{Y} + B_1(x_p - \bar{x})] \\ &= Var\left[\frac{\sum Y_i}{n} + (x_p - \bar{x})\sum c_i Y_i\right] \\ &= Var\left\{\sum_{i=1}^{n}\left[\frac{1}{n} + c_i(x_p - \bar{x})\right]Y_i\right\} \\ &= \sum\left[\frac{1}{n} + c_i(x_p - \bar{x})\right]^2 Var(Y_i) \\ &= \sigma^2\left[\frac{1}{n} + \frac{2(x_p - \bar{x})}{n}\sum c_i + (x_p - \bar{x})^2\sum c_i^2\right] \\ &= \sigma^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right] \end{aligned} \tag{13.18}$$

Por lo tanto, un estimador de la desviación estándar de $\hat{Y}_p$ está dado por

$$s(\hat{Y}_p) = s\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]^{1/2} \tag{13.19}$$

Supóngase que en lugar de estimar la media de Y_p en x_p, se desea predecir un valor particular de Y_p que se observaría si se impusiera un valor x_p para la variable de predicción. Por ejemplo, dada la ecuación de regresión estimada, ¿cuál podría ser el valor del salario inicial para un estudiante en particular con un *CP* conocido? Aunque se trata de un solo estudiante, puede ser razonable predecir el salario inicial promedio para un *CP* dado. De esta forma, si se desea estimar la media de Y_p o un valor particular de Y_p para x_p, el valor estimado es el mismo y está dado por (13.7). Pero es evidente que la varianza de la predicción para este último caso puede tener un valor más grande, ya que ésta no sólo considera la variación muestra a muestra de $\hat{Y}_p$, sino también la variación inherente de la distribución de probabilidad de Y_p. Si se supone que el valor predicho de Y_p para x_p es independiente de la muestra que proporciona la recta de regresión estimada, la covarianza de Y_p y $\hat{Y}_p$ es cero. Entonces

$$\begin{aligned} Var(\hat{Y}_{\text{part}}) &= Var(Y_p) + Var(\hat{Y}_p) \\ &= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right] \\ &= \sigma^2\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right], \end{aligned} \tag{13.20}$$

donde $\hat{Y}_{\text{part}}$ denota la predicción particular para Y_p en x_p. Del análisis anterior se obtiene que un estimador de la desviación estándar de $\hat{Y}_{\text{part}}$ está dado por

$$s(\hat{Y}_{\text{part}}) = s\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]^{1/2} \tag{13.21}$$

Mediante el empleo de los datos que aparecen en la tabla 13.2, se ilustra el cálculo de las varianzas y las desviaciones estándar de los estimadores de mínimos cuadrados B_1 y B_0. Dado que

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - \left(\sum x_i\right)^2/n = 0.886$$

y $s^2 = 1.759$,

$$s^2(B_1) = \frac{1.759}{0.886} = 1.985,$$

y

$$s(B_1) = 1.409.$$

De manera similar,

$$s^2(B_0) = \frac{(1.759)(139.51)}{(15)(0.886)} = 18.465$$

y

$$s(B_0) = 4.297.$$

Si se continúa con este ejemplo, supóngase que se desea estimar la media de la distribución de salarios inciales cuando el *CP* es $x_p = 3.25$. Nótese que este valor no es uno de los valores de x que dieron origen a la recta de regresión estimada, pero se encuentra dentro del intervalo definido por estos valores. De (13.10) y (13.18) la media y la varianza estimadas para $x_p = 3.25$, son

$$\hat{y}_p = -6.63 \pm 8.12(3.25) = 19.76$$

y

$$s^2(\hat{Y}_p) = 1.759\left[\frac{1}{15} + \frac{(3.25 - 3.04)^2}{0.886}\right] = 0.205,$$

respectivamente. De esta forma, la desviación estándar estimada es $\sqrt{0.205} = 0.453$ miles de dólares. Si se desea predecir el salario inicial real para un estudiante en particular con una *CP* de 3.25, el valor estimado sería aún de 19.76 miles de dólares, pero la varianza estimada sería de

$$1.759\left[1 + \frac{1}{15} + \frac{(3.25 - 3.04)^2}{0.886}\right] = 1.964,$$

o una desviación estándar de 1.401 miles de dólares.

En esta sección se han determinado las medias y las varianzas de los estimadores B_0, B_1, $\hat{Y}_p$ y $\hat{Y}_{\text{part}}$, pero aún no se han desarrollado sus distribuciones de muestreo. Para realizar esto es necesario suponer el caso de la teoría normal de la sección anterior, en el que se supone que cada error aleatorio ε_i tiene una distribución normal con media cero y varianza σ^2 para toda $i = 1, 2, \ldots, n$. Por lo tanto, las observaciones $Y_1, Y_2, \ldots, Y_n$ son variables aleatorias independientes y distribuidas en forma normal con medias $\beta_0 + \beta_1 x_i$ y varianza común σ^2, para $i = 1, 2, \ldots, n$.

Para obtener la distribución de la muestra para el estimador de la pendiente B_1, bajo el caso de la teoría normal, sólo necesita recordarse que B_1 es una combinación lineal de variables aleatorias normalmente distribuidas y, de esta forma, la combinación es una variable aleatoria con distribución normal, media β_1 y varianza dadas por (13.13). Al recordar la definición de una variable aleatoria t de Student puede demostrarse que la distribución de la cantidad

$$(B_1 - \beta_1)/s(B_1)$$

es la t de Student con $n - 2$ grados de libertad. El estimador B_0 también es una combinación lineal de variables aleatorias normalmente distribuidas. Así, B_0 también es normalmente distribuida, con media β_0 y varianza dadas por (13.14). Además, se puede mostrar que la cantidad

$$(B_0 - \beta_0)/s(B_0)$$

es una variable aleatoria de la t de Student con $n - 2$ grados de libertad. Como se verá en la siguiente sección, estos resultados permiten la formulación de inferencias estadísticas con respecto a los parámetros desconocidos β_0 y β_1.

Bajo el caso de la teoría normal, el estimador $\hat{Y}_p = B_0 + B_1 x_p$ de la media de Y_p para x_p también se encuentra normalmente distribuido con media $E(Y_p)$ y varianza dada por (13.18), ya que ésta es una combinación lineal de variables aleatorias normalmente distribuidas. Entonces, la distribución de

$$[\hat{Y}_p - E(Y_p)]/s(\hat{Y}_p)$$

es la t de Student con, de nuevo, $n - 2$ grados de libertad. También se obtiene un resultado similar para la predicción $\hat{Y}_{\text{part}}$ para una respuesta en particular Y_p correspondiente a x_p. Así, resulta comprensible el porqué $n - 2$ grados de libertad, ya que la determinación de la recta de regresión necesita la estimación de los dos parámetros de regresión β_0 y β_1.

13.6 Inferencia estadística para el modelo lineal simple

En la sección precedente se examinaron las propiedades teóricas de los estimadores para el modelo lineal simple. En esta sección se emplearán esas propiedades para llevar a cabo un analisis de regresión, es decir, se desarrollarán pruebas de hipótesis e intervalos de confianza para las cantidades de interés en este modelo.

El parámetro clave del modelo lineal simple

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

tiene que ser la pendiente β_1. Si la respuesta Y se encuentra relacionada en forma lineal con la variable de predicción x, la pendiente β_1 tiene que ser diferente de cero. De otra forma, no existe ninguna relación lineal entre Y y x. Un procedimiento inferencial natural para β_1 es construir un intervalo de confianza del $100(1 - \alpha)\%$ para β_1. Si este intervalo no contiene el valor cero, entonces es razonable concluir que β_1 es diferente de cero y que Y y x están, en algún grado, relacionados en forma lineal.

Recuérdese que bajo el caso de la teoría normal, la variable aleatoria $(B_1 - \beta_1)/s(B_1)$ tiene una distribución t de Student con $n - 2$ grados de libertad. Entonces

$$P[B_1 - t_{1-\alpha/2,\, n-2} s(B_1) < \beta_1 < B_1 + t_{1-\alpha/2,\, n-2} s(B_1)] = 1 - \alpha,$$

o la probabilidad de que el intervalo aleatorio $[B_1 - t_{1-\alpha/2,\, n-2} s(B_1),\ B_1 + t_{1-\alpha/2,\, n-2} s(B_1)]$ contenga el valor real de la pendiente β_1 es $1 - \alpha$. Al reemplazar el estimador de mínimos cuadrados B_1 por su estimador dado por (13.6), el intervalo de confianza del $100(1 - \alpha)\%$ para β_1 es

$$b_1 \pm t_{1-\alpha/2,\, n-2} s(B_1),$$

donde la desviación estándar estimada $s(B_1)$ está dada por (13.16). Como ejemplo, recuérdese la recta de regresión estimada $\hat{y}_i = -6.63 + 8.12 x_i$ para los datos de

los salarios iniciales. Dado que $b_1 = 8.12$ y $s(B_1) = 1.409$, entonces un intervalo del 95% de confianza para β_1 es

$$8.12 \pm (2.160)(1.409) = (5.08, 11.16),$$

donde $t_{0.975,\,13} = 2.160$. La interpretación de este intervalo es la siguiente: supóngase que se toman muestras repetidas, cada una del mismo tamaño n, de la variable de respuesta para algunos de los valores de x que producen la recta estimada $\hat{y}_i = -6.63 + 8.12x_i$, construyéndose para cada muestra un intervalo de confianza del 95% para β_1. Por lo tanto, el 95% de todos estos intervalos incluirá el valor real de la pendiente β_1.

Considérese la prueba de la hipótesis nula

$$H_0: \beta_1 = \beta_{1_0}$$

contra la alternativa

$$H_1: \beta_1 \neq \beta_{1_0}$$

donde β_{1_0} es el valor propuesto de la pendiente desconocida β_1. Bajo H_0, la estadística

$$T = \frac{B_1 - \beta_{1_0}}{s(B_1)}$$

tiene una distribución t de Student con $n - 2$ grados de libertad. De esta forma, para un tamaño dado del error de tipo I puede tomarse una decisión, en forma fácil, con base en la evidencia de la muestra. Nótese que también es posible tener hipótesis alternativas unilaterales.

Al igual que en los casos ya analizados, cualquier valor propuesto de β_1 que se encuentre en el correspondiente intervalo de confianza, causará una equivocación al rechazar a H_0. En general, el valor propuesto es el cero; es decir, la hipótesis nula establece que no existe ninguna asociación lineal entre x y Y, así que el valor de la estadística de prueba es

$$t = b_1/s(B_1).$$

Como ejemplo, considérese la prueba de la hipótesis nula

$$H_0: \beta_1 = 0$$

contra la alternativa

$$H_1: \beta_1 > 0$$

para el ejemplo de los salarios iniciales contra *CP*. Se ha seleccionado una hipótesis alternativa unilateral, ya que el sentido común dicta que si existe una relación lineal entre *CP* y el salario inicial, la pendiente deberá ser positiva. Para $\alpha = 0.01$; entonces $t_{0.99,\,13} = 2.650$, y

$$t = 8.12/1.409 = 5.76.$$

Por lo tanto, se rechaza la hipótesis nula de que la pendiente es cero. Este resultado, junto con el intervalo de confianza para β_1, sugiere que el salario inicial promedio se encuentra influenciado, en forma lineal, por la calificación promedio.

También puede construirse un intervalo de confianza para el parámetro de intersección β_0 en forma similar. Dado que $(B_0 - \beta_0)/s(B_0) \sim t$ de Student con $n-2$ grados de libertad,

$$P[B_0 - t_{1-\alpha/2,\, n-2}s(B_0) < \beta_0 < B_0 + t_{1-\alpha/2,\, n-2}s(B_0)] = 1 - \alpha$$

es un intervalo aleatorio para β_0 con probabilidad $1 - \alpha$. Por lo tanto, un intervalo de confianza del $100(1 - \alpha)\%$ para β_0 es

$$b_0 \pm t_{1-\alpha/2,\, n-2}s(B_0)$$

donde b_0 es el estimador de mínimos cuadrados y $s(B_0)$ es la desviación estándar estimada. De nuevo, para el ejemplo de los salarios iniciales, un intervalo de confianza del 99% para β_0 es

$$-6.63 \pm (3.012)(4.297) = (-19.57, 6.31).$$

El lector debe darse cuenta que el significado de un intervalo como el anterior no es del todo aparente, ya que el modelo de regresión no tiene sentido si la calificación promedio es cero. En general, deben evitarse las inferencias con respecto a la intersección, a menos que exista un valor de la respuesta para $x = 0$.

Ahora, considérese la estimación por intervalo de la media de Y_p para x_p. Recuérdese que bajo el caso de la teoría normal, el estimador $\hat{Y}_p = B_0 + B_1x_p$ tiene una distribución normal con media $E(Y_p)$ y varianza dada por (13.18) y la distribución de muestreo de $[\hat{Y}_p - E(Y_p)]/s(\hat{Y}_p)$ es la t de Student con $n-2$ grados de libertad. Entonces la probabilidad del intervalo aleatorio

$$\hat{Y}_p - t_{1-\alpha/2,\, n-2}s(\hat{Y}_p) < E(Y_p) < \hat{Y}_p + t_{1-\alpha/2,\, n-2}s(\hat{Y}_p)$$

es $1 - \alpha$, y un intervalo de confianza del $100(1 - \alpha)\%$ para $E(Y_p)$ es

$$\hat{y}_p \pm t_{1-\alpha/2,\, n-2}s(\hat{Y}_p).$$

Para el ejemplo de los salarios iniciales, supóngase que se desea construir un intervalo de confianza del 95% para la media de Y_p en $x_p = 2.80$. El valor estimado es

$$\hat{y}_p = -6.63 + 8.12(2.80) = 16.11$$

y la desviación estándar estimada es

$$s(\hat{Y}_p) = \left\{1.759\left[\frac{1}{15} + \frac{(2.80 - 3.04)^2}{0.886}\right]\right\}^{1/2} = 0.481.$$

Dado que $t_{0.975,\, 13} = 2.160$, un intervalo de confianza del 95% para $E(Y_p)$ es

$$16.11 \pm (2.160)(0.481) = (15.07, 17.15).$$

Al seguir este procedimiento, pueden obtenerse intervalos de confianza del 95% para $E(Y_p)$ para distintos valores de la variable de predicción. Los resultados se encuentran resumidos en la tabla 13.3.

TABLA 13.3 Intervalos de confianza para los salarios iniciales medios

x_p	$\hat{y}_p$	$s(\hat{Y}_p)$	*Intervalo de confianza del 95%*
2.60	14.48	0.708	(12.95, 16.01)
2.70	15.29	0.589	(14.02, 16.56)
2.80	16.11	0.481	(15.07, 17.15)
2.90	16.92	0.395	(16.07, 17.77)
3.00	17.73	0.347	(16.98, 18.48)
3.10	18.54	0.353	(17.78, 19.30)
3.20	19.35	0.410	(18.46, 20.24)
3.30	20.17	0.501	(19.09, 21.25)
3.40	20.98	0.612	(19.66, 22.30)
3.50	21.79	0.733	(20.21, 23.37)
3.60	22.60	0.860	(20.74, 24.46)

Para ilustrar la naturaleza de estos intervalos de confianza, cuando se comparan con los valores de la variable de predicción, se grafica la recta estimada de regresión y después los límites inferior y superior de cada intervalo contra x_p. El resultado se ilustra en la figura 13.3. Nótese que los límites inferior y superior forman dos hipérbolas con respecto a la recta de regresión estimada. La distancia vertical entre cada

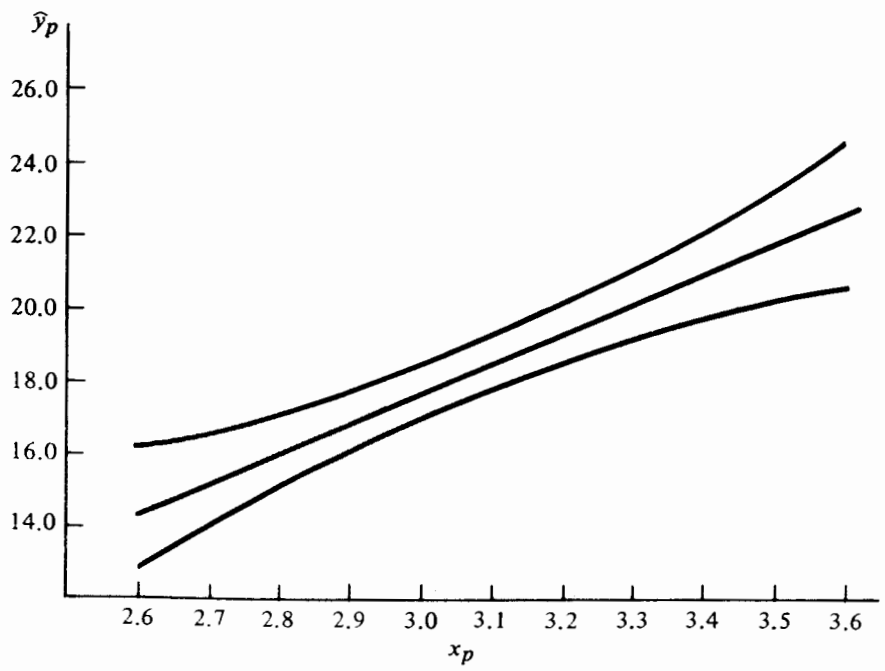


FIGURA 13.3 Intervalos de confianza y la recta de regresión estimada

curva y la recta de regresión es más pequeña para el punto $\bar{x} = 3.04$ y aumenta, en forma simétrica, en ambas direcciones al alejarse de $\bar{x}$. Si se plantea en forma sencilla, los resultados anteriores indican que la predicción de $E(Y_p)$ es más confiable (varianza más pequeña) alrededor de la mitad de los valores de x obtenidos por medio de la ecuación de regresión que en los extremos del intervalo de valores x.

Recuérdese que el usuario puede estar más interesado en predecir una respuesta particular para una x dada, que en estimar la respuesta media para ese mismo valor x. Mientras que el valor predicho puede ser el mismo en cualquier caso, la variabilidad del estimado con respecto a la respuesta en particular será decididamente más grande que la correspondiente a la respuesta media. Dado que, bajo el caso de la teoría normal, la cantidad $[\hat{Y}_{\text{part}} - Y_p]/s(\hat{Y}_{\text{part}})$ es una variable aleatoria t de Student con $n - 2$ grados de libertad entonces, para un α dado,

$$P[\hat{Y}_{\text{part}} - t_{1-\alpha/2,\, n-2}s(\hat{Y}_{\text{part}}) < Y_p < \hat{Y}_{\text{part}} + t_{1-\alpha/2,\, n-2}s(\hat{Y}_{\text{part}})] = 1 - \alpha.$$

Con base en este resultado puede obtenerse lo que, en general, recibe el nombre de *intervalo de predicción* para la observación Y_p. Un intervalo de predicción es el análogo del intervalo de confianza. Un intervalo de predicción del $100(1 - \alpha)\%$ para una observación particular Y_p, es

$$\hat{y}_{\text{part}} \pm t_{1-\alpha/2,\, n-2}s(\hat{Y}_{\text{part}}).$$

Como ejemplo, se construirá un intervalo de predicción del 95% para el salario inicial de un recién graduado con una calificación promedio de 2.80. El valor predicho puede ser el mismo que el de la respuesta media,

$$\hat{y}_{\text{part}} = -6.63 + 8.12(2.80) = 16.11;$$

pero la desviación estándar estimada es

$$s(\hat{Y}_{\text{part}}) = \left\{1.759\left[1 + \frac{1}{15} + \frac{(2.80 - 3.04)^2}{0.886}\right]\right\}^{1/2} = 1.411,$$

la cual es mucho más grande que el valor comparable de 0.481 para $\hat{Y}_p$. Por lo tanto, un intervalo de predicción del 95% para Y_p, es

$$16.11 \pm (2.160)(1.411) = (13.06,\ 19.16).$$

En la tabla 13.4 se proporcionan los intervalos de predicción del 95% para las observaciones de la respuesta correspondiente a cada uno de los valores de x que se encuentran en la tabla 13.3 y que no son parte del conjunto original que dio origen a la ecuación de regresión estimada. Como era de esperarse, los intervalos de predicción para las observaciones individuales de la respuesta son mucho más grandes que los correspondientes intervalos de confianza para la media de la misma.

Ya que el análisis de regresión se basa en la teoría normal, es apropiado formular un comentario con respecto a las consecuencias sobre la inferencia cuando las distribuciones de probabilidad de los errores aleatorios no son normales. Si la desviación con respecto a la normalidad no es muy grande, las distribuciones de muestreo de los

TABLA 13.4 Intervalos de predicción para los salarios iniciales individuales

x_p	$\hat{y}_{\text{part}}$	$s(\hat{Y}_{\text{part}})$	*Intervalo de predicción del 95%*
2.60	14.48	1.503	(11.23, 17.73)
2.80	16.11	1.411	(13.06, 19.16)
2.90	16.92	1.384	(13.93, 19.91)
3.00	17.73	1.371	(14.77, 20.69)
3.30	20.17	1.418	(17.11, 23.23)
3.50	21.79	1.515	(18.52, 25.06)

estimadores serán muy cercanas a la normalidad y se acercarán a ésta conforme aumente el tamaño de la muestra. Bajo estas condiciones, la distribución t de Student sigue siendo muy robusta y proporciona aproximaciones muy cercanas a los niveles de confianza propuestos.

13.7 El uso del análisis de varianza

El análisis de regresión para el modelo lineal sencillo también abarca la aplicación de la técnica del análisis de varianza analizada en el capítulo 12. En síntesis, la técnica del análisis de varianza proporciona sólo un medio alternativo al de la sección 13.6 para probar la hipótesis nula de que la pendiente es cero. Sin embargo, permite una comprensión natural del problema y por lo tanto es muy útil para el análisis de modelos más complicados, lo cual se hará más adelante.

Recuérdese que la técnica del análisis de varianza divide la variación total de las observaciones en sus partes componentes de acuerdo con el modelo propuesto. En esencia, para el modelo lineal simple la variación total es la suma de dos componentes: la causada por el término no aleatorio $\beta_1 x$, y la que se debe al error aleatorio ε. Dado que lo que se pretende es que la recta estimada de regresión explique la mayor cantidad posible de la variación total, la contribución del término $\beta_1 x$ debe ser sustancial. El resultado anterior implicaría que las variables respuesta y predicción están relacionadas en forma lineal. Si $\beta_1 = 0$, no existe una asociación lineal entre x y Y.

Para desarrollar el enfoque del análisis de varianza, se seguirá el procedimiento establecido en el capítulo 12. Considérese la desviación de la observación Y_i de la media de las observaciones $\overline{Y}$. Por el momento, supóngase que todas las observaciones Y_i son iguales entre sí, así que la pendiente β_1 debe ser cero, $\varepsilon_i = 0$, y $Y_i = \overline{Y}$ para toda i. Por otro lado, si la magnitud de la desviación $Y_i - \overline{Y}$ es mayor que cero, ésta deberá atribuirse a las componentes del modelo.

Para la desviación

$$Y_i - \overline{Y}$$

supóngase que se suma y se resta el estimador $\hat{Y}_i$ para la media de Y_i, tal como se obtiene de la ecuación de regresión. Entonces

$$\begin{aligned} Y_i - \overline{Y} &= Y_i - \overline{Y} + \hat{Y}_i - \hat{Y}_i \\ &= \hat{Y}_i - \overline{Y} + Y_i - \hat{Y}_i. \end{aligned}$$

De esta forma, la desviación total de la observación Y_i con respecto a la media $\overline{Y}$, es la suma de la desviación de la respuesta media estimada $\hat{Y}_i$ de $\overline{Y}$ y la desviación de Y_i con respecto a $\hat{Y}_i$. Nótese que la última diferencia es el estimador para el i-ésimo residuo, el cual representa la distancia vertical desde la respuesta observada al punto correspondiente sobre la recta de regresión estimada. Las desviaciones $Y_i - \hat{Y}_i$ representan la contribución a la componente de error a la variación total. Recuérdese que $\hat{Y}_i$ estima la media de Y_i para x_i. Si la variable de predicción no tiene ningún efecto lineal sobre la respuesta, entonces $\hat{Y}_i$ es virtualmente igual a $\overline{Y}$ para toda i; es decir, $\beta_1 = 0$, y el estimador de mínimos cuadrados de β_0 es $\overline{Y}$. Si la magnitud de la desviación de $\hat{Y}_i - \overline{Y}$ es grande, entonces se tiene un efecto lineal de x sobre Y ($\beta_1 \neq 0$).

Para proseguir con el enfoque del análisis de varianza se tomará el cuadrado de ambos miembros de la identidad

$$Y_i - \overline{Y} = \hat{Y}_i - \overline{Y} + Y_i - \hat{Y}_i,$$

y se sumarán para todas las observaciones $\Sigma(Y_i - \overline{Y})^2$. Entonces se tiene

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + 2 \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})(Y_i - \hat{Y}_i).$$

Para demostrar que los productos cruzados son cero, se vuelve a escribir la última suma como

$$\begin{aligned} \sum (\hat{Y}_i - \overline{Y})(Y_i - \hat{Y}_i) &= \sum [\hat{Y}_i(Y_i - \hat{Y}_i) - \overline{Y}(Y_i - \hat{Y}_i)] \\ &= \sum \hat{Y}_i(Y_i - \hat{Y}_i) - \overline{Y} \sum (Y_i - \hat{Y}_i). \end{aligned}$$

De acuerdo con la propiedad 1 de la recta de regresión estimada, examinada en la sección 13.3, la segunda suma es cero. La primera suma puede escribirse como

$$\begin{aligned} \sum \hat{Y}_i(Y_i - \hat{Y}_i) &= \sum \hat{Y}_i e_i \\ &= \sum (B_0 + B_1 x_i) e_i \\ &= B_0 \sum e_i + B_1 \sum x_i e_i \\ &= 0, \end{aligned}$$

Dado que Σe_i y $\Sigma x_i e_i$ son cero de las propiedades 1 y 3, respectivamente, por lo tanto la ecuación fundamental del análisis de regresión es

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \tag{13.22}$$

De acuerdo con la terminología dada en el capítulo 12, el término $\Sigma(Y_i - \overline{Y})^2$ es la suma total de cuadrados STC la cual toma en cuenta la variación total de las observaciones Y_i con respecto a su media sin considerar la variable de predicción. Las compo-

nentes de *STC* son la suma de los cuadrados de los errores $SCE = \Sigma(Y_i - \hat{Y}_i)^2$ y la *suma de los cuadrados de la regresión* $SCR = \Sigma(\hat{Y}_i - \bar{Y})^2$. *SCE* toma en cuenta la variación de las observaciones con respecto a la recta de regresión estimada. Si todas las observaciones se encuentran sobre la recta estimada, el valor de todos los residuos es cero y $SCE = 0$. Se desprende el hecho de que entre más grande es el valor de *SCE*, mayor es la contribución de la componente de error a la variación de las observaciones, o mayor es la incertidumbre cuando se estima la respuesta mediante el uso de la ecuación de regresión. Por otro lado, *SCR* representa la variación de la observación que es atribuible al efecto lineal de x sobre Y. Si la pendiente de la recta estimada de regresión es cero, entonces $SCR = 0$. De esta forma, entre más grande es la proporción de *SCR* con respecto a *SCT*, mayor será la cantidad de la variación en las observaciones que puede explicarse mediante el término lineal $\beta_1 x$

¿Cuál es el número de grados de libertad asociado con cada término de (13.22)? Recuérdese la definición del número de grados de libertad asociados con una suma de cuadrados dada en el capítulo 12. Para *STC* existen $n - 1$ grados de libertad ya que se pierde uno por causa de la restricción lineal $\Sigma(Y_i - \bar{Y}) = 0$ entre las observaciones Y_i. Nótese que *SCE* es el numerador de la expresión (13.11) para el cálculo de la varianza residual, así que el número de grados de libertad para *SCE* será de $n - 2$.* Dado que los grados de libertad son aditivos,

$$\text{gl(SCR)} = \text{gl(STC)} - \text{gl(SCE)},$$

y *SCR* tiene un grado de libertad. Como se observará posteriormente, cuando se traten modelos más complicados, el número de grados de libertad para *SCR* será siempre igual al número de parámetros de regresión en el modelo, sin contar a β_0.

Para el análisis de varianza se buscará una estadística para probar la hipótesis nula

$$H_0: \beta_1 = 0$$

contra la alternativa

$$H_1: \beta_1 \neq 0.$$

En general, H_0 se conoce como la hipótesis de regresión no lineal entre x y Y. Si se supone el caso de la teoría normal, entonces bajo la hipótesis nula las observaciones Y_i son n variables aleatorias independientes normalmente distribuidas con la misma media $\mu = \beta_0$ y varianza σ^2. Por lo tanto, puede demostrarse que SCR/σ^2 y SCE/σ^2 son dos variables aleatorias independientes con una distribución chi-cuadrada con 1 y $n - 2$ grados de libertad, respectivamente. Entonces, del teorema 7.8, la variable aleatoria

$$F = \frac{\dfrac{\text{SCR}/\sigma^2}{1}}{\dfrac{\text{SCE}/\sigma^2}{n-2}} = \frac{\text{SCR}/1}{\text{SCE}/(n-2)} = \text{CMR}/\text{CME} \qquad (13.23)$$

* Los dos grados de libertad que se pierden se deben a las dos restricciones lineales dadas por las propiedades 1 y 3 de la sección 13.3.

tiene una distribución F con 1 y $n-2$ grados de libertad, donde el cuadrado medio del error es igual a la varianza residual.

Para llegar a la región de rechazo apropiada para H_0 sugerida por (13.23), se empleará la intuición. Con base en un conjunto dado de datos, un valor grande de *CME* comparado con *CMR* implicará ajuste pobre y sugerirá la ausencia de una asociación lineal entre x y Y. Pero un valor relativamente pequeño para *CME* implicará el hecho de que una porción considerable de la variación en las observaciones es atribuible a un efecto lineal de x sobre Y. Por lo tanto, la hipótesis nula de regresión no lineal entre x y Y debe rechazarse siempre que el valor de (13.23) sea relativamente grande. De otro modo, la evidencia experimental no apoya el rechazo de H_0. Sobre una base más teórica, puede demostrarse que

$$E(\text{CMR}) = \sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2$$

y

$$E(\text{CME}) = \sigma^2.$$

Si H_0 es cierta, entonces el valor esperado de *CMR* también es σ^2. Pero si $\beta_1 \neq 0$, E(CMR) es mayor que σ^2, ya que el término $\beta_1^2 \Sigma(x_i - \bar{x})^2$ es positivo. Por lo tanto, la estadística apropiada está dada por (13.23) con el extremo superior de la distribución F como región crítica; es decir, para un tamaño dado del error de tipo I α se rechaza la hipótesis nula de no regresión lineal cuando un valor de $F = CMR/CME$ se encuentra dentro de la región crítica superior de la distribución F con 1 y n -2 grados de libertad. La tabla de análisis de varianza (ANOVA) para el modelo lineal simple se encuentra en la tabla 13.5.

Para calcular las sumas de cuadrados que aparecen en la tabla 13.5, se tiene

$$\text{STC} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n},$$

$$\text{SCE} = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2,$$

donde $y_1, y_2, \ldots, y_n$ son las verificaciones de las observaciones, y $e_1, e_2, \ldots, e_n$ son los residuos correspondientes. Entonces $SCR = STC - SCE$, o puede calcularse

TABLA 13.5 Tabla ANOVA para el modelo lineal simple

Fuente de variación	*gl*	SC	CM	*Estadística F*
Regresión	1	$\sum (\hat{Y}_i - \bar{Y})^2$	$\sum (\hat{Y}_i - \bar{Y})^2/1$	$\dfrac{\sum (\hat{Y}_i - \bar{Y})^2/1}{\sum (Y_i - \hat{Y}_i)^2/(n-2)}$
Error	$n-2$	$\sum (Y_i - \hat{Y}_i)^2$	$\sum (Y_i - \hat{Y}_i)^2/(n-2)$	
Total	$n-1$	$\sum (Y_i - \bar{Y})^2$		

en forma directa. Dado que la recta estimada es

$$\hat{y}_i = \bar{y} + b_1(x_i - \bar{x})$$

o

$$\hat{y}_i - \bar{y} = b_1(x_i - \bar{x}),$$

al elevar al cuadrado ambos miembros y si se suman para todas las $i = 1, 2 \ldots n$ se obtiene

$$\text{SCR} = \sum (\hat{y}_i - \bar{y})^2 = b_1^2 \sum (x_i - \bar{x})^2. \tag{13.24}$$

Como ejemplo, recuérdese de nuevo el problema de los salarios inciales. Supóngase que se desea probar la hipótesis nula de que no existe una regresión lineal entre el salario inicial y *CP*, contra la alternativa de que ésta existe, con $\alpha = 0.01$. Mediante el uso de la tabla 13.2 se calculan las siguientes cantidades:

$$\text{STC} = 4970.12 - \frac{(270.8)^2}{15} = 81.2773,$$

$$\text{SCE} = 22.8671,$$

$$\text{SCR} = 81.2773 - 22.8671 = 58.4102.$$

Para $n = 15$ se proporciona la tabla ANOVA en la tabla 13.6. Dado que $f = 33.21 > f_{0.99,\,1,\,13} = 9.07$, se rechaza la hipótesis nula de no regresión lineal y se concluye que el salario inicial promedio está influenciado, en forma lineal, por la calificación promedio.

Como es de esperarse, existe una relación entre la estadística F anterior con 1 y $n - 2$ grados de libertad y la correspondiente estadística t de Student (véase la sección 13.3) para una hipótesis alternativa bilateral. Puede establecerse la relación mediante lo siguiente: dado que

$$\text{SCR} = B_1^2 \sum (x_i - \bar{x})^2$$

y

$$s^2(B_1) = \text{CME} \Big/ \sum (x_i - \bar{x})^2,$$

TABLA 13.6 Tabla ANOVA para los salarios iniciales

Fuente de variación	gl	SC	CM	*Valor F*
Regresión	1	58.4102	58.4102	33.21
Error	13	22.8671	1.759	
Total	14	81.2773	$f_{0.99,\,1,\,13} = 9.07$	

entonces

$$F = \frac{\text{CMR}}{\text{CME}} = \frac{B_1^2 \sum (x_1 - \bar{x})^2/1}{s^2(B_1) \sum (x_i - \bar{x})^2} = [B_1/s(B_1)]^2.$$

De acuerdo con lo anterior, si una variable aleatoria tiene una distribución F con 1 y $n-2$ grados de libertad, entonces

$$F = T^2,$$

donde T es una variable aleatoria t de Student con $n-2$ grados de libertad. La relación entre los cuantiles es

$$f_{1-\alpha,\,1,\,n-2} = t^2_{1-\alpha/2,\,n-2}. \tag{13.25}$$

Hasta aquí se han examinado algunas maneras para probar la hipótesis nula de no regresión lineal entre x y Y. Ahora se presentará una cantidad numérica muy útil que es una medida relativa del grado de asociación lineal entre x y Y. Lo que se desea es tener una cantidad que mida la proporción de la variación total de las observaciones con respecto a su media la cual es atribuida a la recta estimada de regresión. Dado que *STC* representa la variación total con respecto a la media y *SCR* mide la porción de ésta, que es atribuible a un efecto lineal de x sobre Y, una medida apropiada es

$$r^2 = \frac{\text{SCR}}{\text{STC}} = \frac{\text{STC} - \text{SCE}}{\text{STC}} = 1 - \frac{\text{SCE}}{\text{STC}}. \tag{13.26}$$

r^2 recibe el nombre de *coeficiente de determinación*. Los valores que toma están siempre en el intervalo $0 \leq r^2 \leq 1$ ya que $0 \leq \text{SCE} \leq \text{STC}$. De manera ideal, se desea tener un $r^2 = 1$ ya que entonces $SCE = 0$, y toda la variación presente en las observaciones puede explicarse por la presencia lineal de x en la ecuación de regresión. De esta forma, entre más cercano se encuentre r^2 a uno, mayor es el grado de asociación lineal que existe entre x y Y. Como ilustración, el coeficiente de determinación para el ejemplo del salario inicial, es

$$r^2 = 1 - \frac{22.8671}{81.2773} = 0.7187.$$

Por lo tanto, la presencia lineal de *CP* en el modelo de regresión explica el 71.87% de la variación total en los salarios iniciales observados.

Ya que muchas veces se da una mala interpretación a r^2, debe hacerse un comentario sobre lo que r^2 no mide. r^2 no mide la validez del modelo de regresión propuesto, es decir, r^2 no puede verificar que la verdadera ecuación de regresión entre x y Y sea estrictamente lineal. Todo lo que puede medir es cuánto se explica de la variación total mediante la ecuación de regresión estimada. En realidad, el modelo verdadero de regresión entre x y Y puede contener términos no lineales en x, u otras variables de predicción, o ambos. Estas cuestiones serán examinadas en el capítulo 14.

A continuación se presenta una muestra de un listado de computadora para el análisis de regresión lineal de los datos de salarios. La lista de paquetes estadísticos disponible para computadora incluye a *SAS, SPSS, BMDP* y Minitab. El listado que se muestra en la figura 13.4 fue generado por Minitab. Nótese que incluye los coeficientes de la regresión estimados, sus desviaciones estándar, la prueba T para pendiente cero, la desviación estándar residual (o la desviación estándar de Y con respecto a la recta de regresión); el valor de r^2 las sumas de los cuadrados, los cuadrados medios para el análisis de varianza y los residuos estandarizados definidos en el capítulo 12.

LA ECUACIÓN DE REGRESIÓN ES
Y = − 6.63 + 8.12 X1

	COLUMNA	COEFICIENTE	DEV. EST. DEL COEF.	COCIENTE-T = COEF/D.E.
	−	− 6.627	4.298	−1.54
X1	C2 ·	8.118	1.409	5.76

LA DEV. EST. DE Y CON RESPECTO A LA RECTA DE REGRESIÓN ES
S = 1.327
CON (15− 2) = 13 GRADOS DE LIBERTAD
R-CUADRADO = 71.8%

ANÁLISIS DE VARIANZA

DEBIDA A	DF	SC	CM = SC/GL
REGRESIÓN	1	58.393	58.393
RESIDUO	13	22.880	1.760
TOTAL	14	81.274	

RENGLON	X1 C2	Y C1	VALOR PRED. Y	DEV. EST. PRED. Y	RESIDUO	RES. EST.
1	2.95	18.500	17.323	0.365	1.177	0.92
2	3.20	20.000	19.352	0.410	0.648	0.51
3	3.40	21.100	20.976	0.612	0.124	0.11
4	3.60	22.400	22.600	0.860	−0.200	−0.20
5	3.20	21.200	19.352	0.410	1.848	1.46
6	2.85	15.000	16.511	0.435	−1.511	−1.21
7	3.10	18.000	18.540	0.353	−0.540	−0.42
8	2.85	18.800	16.511	0.435	2.289	1.83
9	3.05	15.700	18.134	0.343	−2.434	−1.90
10	2.70	14.400	15.293	0.589	−0.893	−0.75
11	2.75	15.500	15.699	0.533	−0.199	−0.16
12	3.10	17.200	18.540	0.353	−1.340	−1.05
13	3.15	19.000	18.946	0.376	0.054	0.04
14	2.95	17.200	17.323	0.365	−0.123	−0.10
15	2.75	16.800	15.699	0.533	1.101	0.91

FIGURA 13.4 Listado de computadora para el análisis de regresión lineal (datos de los salarios iniciales)

13.8 Correlación lineal

En la sección 6.4 se definió el coeficiente de correlación ρ dado por (6.14), como una medida de la asociación lineal que existe entre las variables aleatorias X y Y. En esta sección se examinará el coeficiente de correlación de la muestra en el contexto del análisis de regresión.

Durante toda la presentación del análisis de regresión se ha asumido la disponibilidad de una muestra aleatoria de la variable respuesta $Y_1, Y_2, \ldots, Y_n$, correspondientes a n valores fijos $x_1, x_2, \ldots, x_n$ de una variable de predicción. Para definir el coeficiente de correlación de la muestra, se supondrá que tanto X como Y son variables aleatorias. Sea la distribución conjunta de X y Y la normal bivariada (véase la sección 6.8), y sean $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ una muestra aleatoria de tamaño n de esta distribución. Entonces puede demostrarse que el estimador de máxima verosimilitud de ρ (denominado *coeficiente de correlación de la muestra*), está dado por

$$r^*(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\left[\sum_{i=1}^{n} (X_i - \overline{X})^2\right]^{1/2} \left[\sum_{i=1}^{n} (Y_i - \overline{Y})^2\right]^{1/2}}. \tag{13.27}$$

Después de efectuar algunos cálculos algebraicos, puede obtenerse una expresión equivalente de la forma

$$r(X, Y) = \frac{\sum X_i Y_i - \dfrac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}}{\left[\sum X_i^2 - \dfrac{\left(\sum X_i\right)^2}{n}\right]^{1/2} \left[\sum Y_i^2 - \dfrac{\left(\sum Y_i\right)^2}{n}\right]^{1/2}}. \tag{13.28}$$

Al igual que el parámetro ρ, r se encuentra en el intervalo $-1 \leq r \leq 1$ y mide la relación lineal entre X y Y, si X se emplea para predecir Y o viceversa. Con base en una muestra aleatoria, un valor de $r = -1$ indica una relación lineal negativa perfecta entre X y Y, mientras que un valor de $r = 1$ señalará una asociación lineal positiva perfecta de X y Y. Si $r = 0$, entonces no existe ninguna relación lineal entre X y Y. En la figura 13.5 se muestran algunas gráficas de dispersión comunes para algunos valores de r.

A causa de varias interpretaciones injustificables que ha sufrido r, es imperioso que el lector comprenda que r por sí mismo no puede ni probar ni desmentir una relación causal entre X y Y, aun si $r = \pm 1$. Como ya se indicó al principio de este capítulo, la manifestación de una relación causa-efecto es posible sólo a través de la comprensión de la relación natural que existe entre X y Y, y ésta no debe manifestarse sólo por la existencia de una fuerte correlación entre X y Y.

* Se seguirá la norma de utilizar una r minúscula.

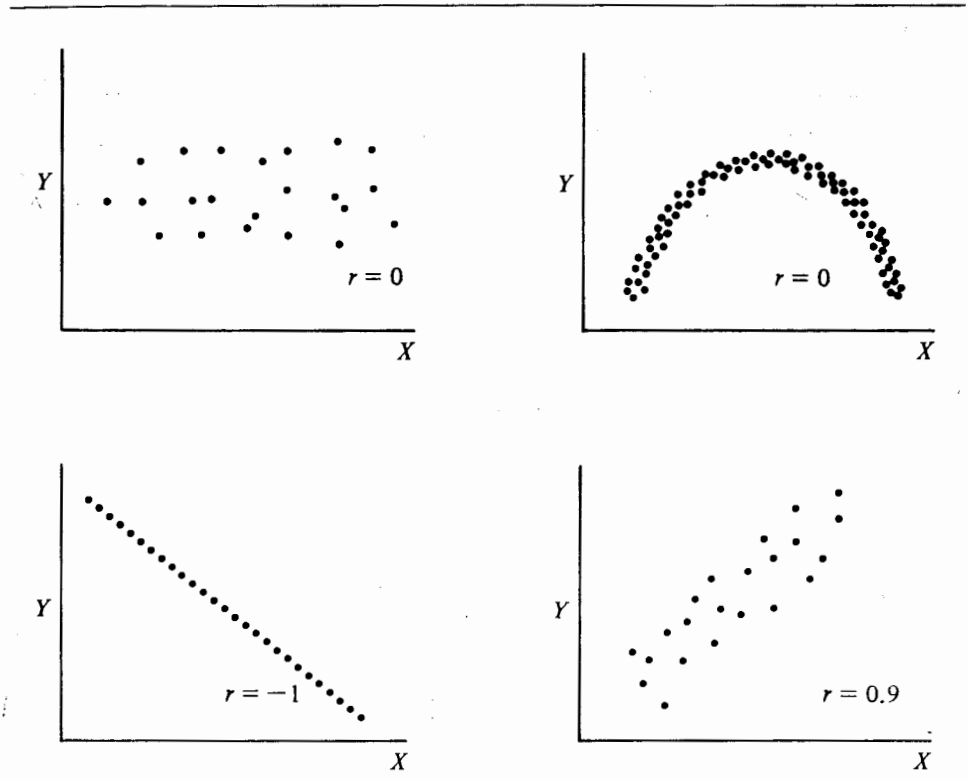


FIGURA 13.5 Gráficas de dispersión comunes para algunos valores de r

Mientras que en el análisis de regresión se supone que los valores de x son fijos, el coeficiente de correlación de la muestra definido por (13.27) o (13.28) es todavía un estimador de ρ. Dado que r mide el grado de asociación lineal entre x y Y, y ya que B_1 es el correspondiente estimador de mínimos cuadrados de la pendiente para el modelo lineal propuesto entre x y Y, entonces debe existir una relación entre r y B_1. Mediante el empleo de la segunda ecuación de (13.6) y (13.27), puede demostrarse que el estimador de mínimos cuadrados de la pendiente y el correspondiente valor del coeficiente de correlación de la muestra se encuentran relacionados por

$$b_1 = \left[\frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2}\right]^{1/2} r. \tag{13.29}$$

Nótese que si $r = 0$, $b_1 = 0$ y viceversa. Además, el signo de b_1 siempre es igual al de r. Finalmente, el cuadrado del coeficiente de correlación de la muestra es el coeficiente de determinación, es decir, si r^2 y b_1 son conocidos, entonces se sabe el valor de r y su signo; por lo tanto, se sigue que r no sólo es una medida del grado de aso-

ciación lineal entre dos variables, sino que puede emplearse una función de r como una medida de la bondad del ajuste para una ecuación estimada de regresión.

13.9 Series de tiempo y autocorrelación

En las secciones anteriores se han examinado los análisis de regresión y de correlación con base en una muestra aleatoria de la variable respuesta Y. En muchas situaciones, por ejemplo en economía y finanzas, la variable respuesta se mide en forma periódica con respecto al tiempo. Por ejemplo, puede escogerse examinar la tasa de desempleo para los pasados 24 meses, o puede observarse el volumen de ventas trimestral de alguna compañía y compararlo con el correspondiente volumen de ventas de toda la industria durante los pasados 12 trimestres. Dado que para ambos ejemplos las observaciones se registran de manera secuencial con el paso del tiempo, forman lo que se conoce como una *serie de tiempo*.

Aunque los metodos de regresión pueden ser útiles al analizar datos de series de tiempo, las observaciones de Y en una serie de tiempo no pueden considerarse como representativas de una muestra aleatoria. De hecho, pueden encontrarse correlacionadas entre sí. Por ejemplo, es probable que el cambio en la tasa de desempleo para este mes se encuentre relacionada con la que se observará para el siguiente mes. De esta forma, algunas de las suposiciones que son necesarias para el desarrollo de procedimientos inferenciales posiblemente no se verifiquen para los datos de una serie de tiempo.

En este contexto se desea considerar un procedimiento inferencial útil, conocido como estadística de Durbin-Watson, para determinar si los errores en un modelo* lineal sencillo se encuentran correlacionados en el tiempo. Los errores del mismo modelo de regresión que se encuentran correlacionados como funciones del tiempo reciben el nombre de *correlacionados serialmente* o *autocorrelacionados*. Antes de analizar el procedimiento de Durbin-Watson, se mencionarán en forma breve los componentes usuales de datos de una serie de tiempo.

13.9.1 Componentes de una serie de tiempo

Las fluctuaciones de la variable respuesta en una serie de tiempo de tipo económico se asignan, por lo general, a cuatro causas diferentes (componentes): la variación en la tendencia T, la variación por temporada S, la variación cíclica C y la variación aleatoria R. la *variación en la tendencia* es el movimiento a largo plazo en Y. Por ejemplo, la producción de automóviles en Estados Unidos ha mostrado una tendencia hacia el crecimiento durante los últimos 50 años, pero lo anterior no necesariamente implica que la producción aumentó todos los años durante este periodo. De esta forma, la tendencia refleja el movimiento general de Y a lo largo de un periodo. La *variación por temporada* representa el movimiento de Y que ocurre durante periodos específicos a lo largo de un año. Por ejemplo, el volumen de ventas al menudeo tiende a ser mayor en el último trimestre del año que durante el primero. La *va-*

* También puede emplearse este procedimiento para el modelo lineal general, el cual se estudiará en el capítulo 14.

riación cíclica muestra el movimiento de Y que se repite durante periodos que, en general, son mayores de un año. Los movimientos cíclicos se encuentran muchas veces relacionados con las condiciones económicas prevalecientes. Por ejemplo, la construcción de casas en Estados Unidos disminuyó durante el periodo de recesión de 1974-1975; aumentó durante el de recuperación de 1976-1979 y volvió a disminuir en la recesión de 1981-1982. La *variación aleatoria* en una serie de tiempo es la fluctuación de Y que no es posible asignar a una causa identificable. Por lo tanto, la fluctuación total de Y con respecto al tiempo se asigna a una variación sistemática (tendencia, temporada y ciclo) y a una variación aleatoria.

Al suponer cómo se encuentran relacionadas estas componentes, puede formularse un modelo de una serie de tiempo que ayudará a separar estas componentes y formular predicciones con respecto a Y. Los modelos de las series de tiempo usualmente son aditivos de la forma

$$Y = T + S + C + R$$

o multiplicativos de la forma

$$Y = T \times S \times C \times R.$$

Para un modelo aditivo se supone que los cuatro componentes son independientes entre sí, mientras que para el multiplicativo se encuentran relacionados entre sí. Para tratamientos completos del análisis de las series de tiempo se sugieren las referencias [1] y [4].

13.9.2 La estadística de Durbin-Watson

En esta sección el interés radicará, en forma exclusiva, en la detección de errores autocorrelacionados y en un análisis con respecto a medidas correctivas. Una de las razones de la existencia de la autocorrelación es que podrían no haberse tomado en cuenta en el modelo variables importantes de predicción. Por ejemplo, se mencionó que, en general, la producción de automóviles tuvo un incremento durante un periodo de 50 años. Si se supone algún modelo de regresión con el tiempo como la única variable de predicción, no es de dudar que se encontrarán correlaciones entre los errores. Pero durante el mismo periodo aumentó la población así como el nivel económico de los habitantes de Estados Unidos. Cuando variables de predicción como éstas están positivamente correlacionadas con la producción de automóviles, pero no se toman en cuenta en el modelo de regresión, entonces los errores tenderán a estar positivamente autocorrelacionados, ya que también reflejan los efectos de las variables de predicción faltantes. Este tipo de autocorrelación sólo es aparente y puede eliminarse mediante la inclusión de las variables omitidas en el modelo de regresión.

En las series de tiempo económicas, la autocorrelación también puede presentarse debido a que los residuos sucesivos tienden a estar positivamente correlacionados, es decir, los grandes residuos negativos siguen a grandes residuos negativos y los grandes residuos positivos siguen a grandes residuos positivos. Este tipo de autocorrelación es, en general, la clase que necesita algún ajuste. El interés recaerá en este tipo y se estudiarán las medidas correctivas tales como la transformación de los datos.

Recuérdese que por la suposición 3 de la sección 13.2, la covarianza entre los errores aleatorios ε_i y ε_j es cero para todo $i \neq j$. A pesar de que esta suposición no es necesaria para obtener los estimadores de mínimos cuadrados, su violación afecta las propiedades inferenciales de estos estimadores. Cuando se encuentra presente la autocorrelación, el análisis de regresión es afectado en tres formas.

1. Los estimadores *MC*, aunque son no sesgados ya no tienen varianza mínima.
2. Los estimados $s^2(B_i)$ pueden subestimar, en forma seria, las varianzas de los estimadores *MC* de B_i.
3. Los intervalos de confianza y las pruebas de hipótesis que incluyen, ya sea la distribución t de Student o la distribución F, no son teóricamente válidas.

Por ejemplo, supóngase que los datos que figuran más adelante representadas las ventas Y de alguna compañía (en millones de dólares) y las ventas x (también en millones de dólares) para toda la industria en los pasados 16 trimestres, donde los datos ya se han ajustado de acuerdo con la inflación.

t	1	2	3	4	5	6	7	8
x_t	270.36	258.38	254.96	259.70	265.40	274.98	281.86	285.78
Y_t	44.84	42.97	41.98	42.75	43.95	45.65	46.87	47.35
t	9	10	11	12	13	14	15	16
x_t	290.58	290.18	296.72	292.32	301.72	305.42	314.96	321.10
Y_t	48.13	47.95	49.10	48.52	50.22	51.15	52.78	53.91

Una gráfica de Y contra x revela una tendencia lineal, lo que a su vez sugiere que las acciones de la compañía se mantienen en el mercado. Supóngase un modelo lineal simple como el dado por (13.1). El listado de computadora producido por Minitab, se muestra en la figura 13.6

Nótese que parece que el modelo ajusta los datos en forma excelente, ya que $r^2 = 0.997$, y se rechaza la hipótesis nula de pendiente igual a cero para casi cualquier nivel α. Las desviaciones estádar estimadas para B_0 y B_1 son pequeñas, y en forma especial para el estimador de la pendiente. Pero al graficar los residuos estandarizados contra el tiempo, como se muestra en la figura 13.7, se nota que los residuos del mismo signo aparecen agrupados. Por ejemplo, los residuos 5-7 son positivos, 8-13 son negativos y 14-16 son positivos. Este tipo de patrón es característico cuando se tienen errores autocorrelacionados.

La estadística de Durbin-Watson constituye un enfoque más formal que al graficar los residuos para detectar los errores autocorrelacionados; se basa en la suposición de que los errores ε_t en el modelo de regresión

$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \tag{13.30}$$

forman una serie autorregresiva de primer orden dada por

$$\varepsilon_t = \rho \varepsilon_{t-1} + \eta_t \qquad t \geq 2, \tag{13.31}$$

LA ECUACION DE REGRESION ES
Y = - 2.97 + 0.177 X1

	COLUMNA	COEFICIENTE	DEV. EST. DEL COEF.	COCIENTE-T = COEF/D.E.
	–	-2.9716	0.7023	-4.23
X1	C2	0.176510	0.002456	71.86

LA DEV. EST. DE Y CON RESPECTO A LA RECTA DE REGRESION ES
S = 0.1919
CON (16 - 2) = 14 GRADOS DE LIBERTAD

R-CUADRADO = 99.7%

ANALISIS DE VARIANZA

DEBIDO A	DF	SC	CM = SC/GL
REGRESION	1	190.2330	190.2330
RESIDUO	14	0.5157	0.0368
TOTAL	15	190.7487	

RENGLON	X1 C2	Y C1	VALOR PRED. Y	DEV. EST. PRED. Y	RESIDUO	RES. EST.
1	270	44.8400	44.7497	0.0604	0.0903	0.50
2	258	42.9699	42.6350	0.0816	0.3349	1.93
3	255	41.9800	42.0314	0.0886	-0.0515	-0.30
4	260	42.7499	42.8680	0.0790	-0.1181	-0.68
5	265	43.9499	43.8742	0.0684	0.0758	0.42
6	275	45.6500	45.5651	0.0542	0.0848	0.46
7	282	46.8699	46.7795	0.0487	0.0904	0.49
8	286	47.3499	47.4714	0.0480	-0.1215	-0.65
9	291	48.1299	48.3187	0.0497	-0.1887	-1.02
10	290	47.9499	48.2481	0.0495	-0.2981	-1.61
11	297	49.0999	49.4025	0.0556	-0.3025	-1.65
12	292	48.5200	48.6258	0.0510	-0.1059	-0.57
13	302	50.2199	50.2850	0.0627	-0.0651	-0.36
14	305	51.1500	50.9381	0.0689	0.2119	1.18
15	315	52.7800	52.6220	0.0873	0.1579	0.92
16	321	53.9100	53.7058	0.1002	0.2042	1.25

FIGURA 13.6 Análisis de regresión lineal (datos del mercado de acciones)

donde $|\rho| < 1$ es la pendiente de la recta que pasa por el origen y η_t es el error aleatorio puro que no se encuentra correlacionado con cualquier otra componente. El término η_t se denomina de manera común como *ruido blanco*. Debe notarse que (13.31) es un modelo autorregresivo, ya que la variable de predicción ε_{t-1} es un término retardado en el tiempo de la variable respuesta ε_t. A pesar de que la estructura de correlación entre los errores puede ser más compleja que la implicada por (13.31), un modelo autorregresivo de primer orden es una aproximación razonable, debido a que muchas veces la autocorrelación entre ε_t y ε_{t+p} disminuye de manera rápida conforme la distancia entre los puntos en el tiempo t y $t + p$ aumenta.

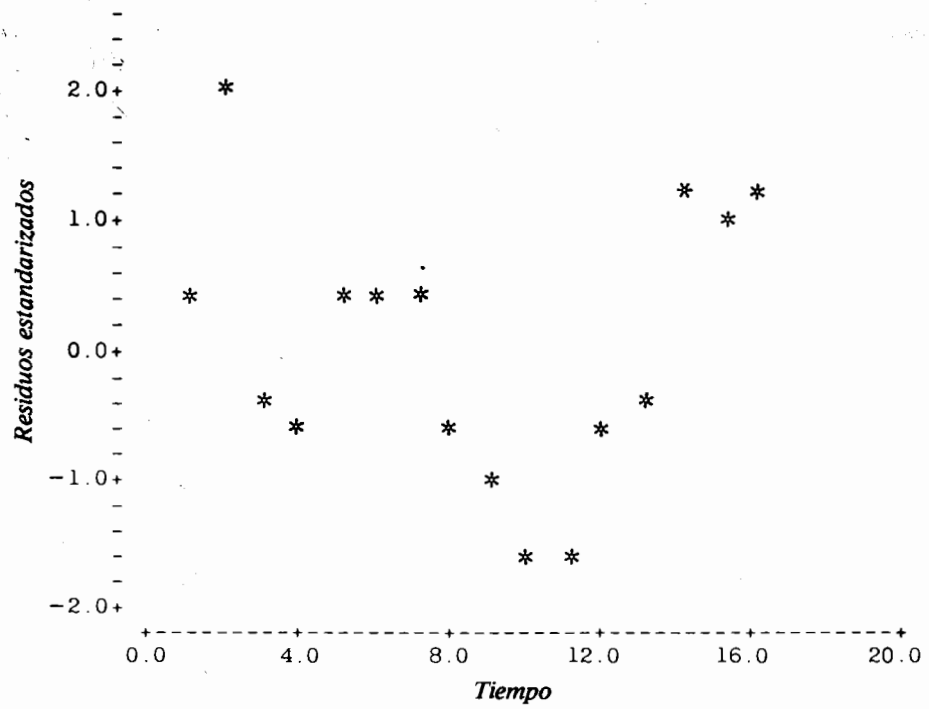


FIGURA 13.7 Residuos estandarizados contra tiempo para el ejemplo de las ventas

Para el modelo dado por (13.31), se desea emplear la estadística de Durbin-Watson para probar la hipótesis nula

$$H_0: \rho = 0$$

contra la alternativa

$$H_1: \rho > 0.$$

Nótese que H_1 es una hipótesis alternativa unilateral superior, ya que las series de tiempo económicas exhiben muchas veces una autocorrelación positiva. La estadística de Durbin-Watson se basa en los residuos que resultan después de obtener la ecuación de regresión estimada para (13.30). Se calcula un valor de esta estadística a partir de la expresión

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}, \tag{13.32}$$

donde el residuo es $e_t = y_t - \hat{y}_t$.

Si los errores se encuentran positivamente autocorrelacionados, es probable que los errores adyacentes tengan la misma magnitud. De esta forma, pequeñas diferencias entre los residuos adyacentes sugieren que ρ es mayor que cero; pero cuando las diferencias son pequeñas, el numerador de (13.32) también lo es. De acuerdo con lo anterior, se rechaza la hipótesis nula de autocorrelación cero siempre que d tiene un valor relativamente pequeño.

Durbin y Watson tabularon los límites inferior y superior d_L y d_U, respectivamente, para probar H_0. En la tabla K del apéndice se proporcionan los límites d_L y d_U para $\alpha = 0.05$ y 0.01 como funciones del tamaño n de la muestra y el número k de variables de predicción en el modelo de regresión. Dados los límites d_L y d_U, la decisión para H_0 se toma de la siguiente forma:

a Si $d < d_L$, rechazar H_0.

b Si $d > d_U$, no puede rechazarse H_0.

c Si $d_L < d < d_U$, la prueba no es concluyente.

Debe señalarse que la prueba para autocorrelación negativa (H_1: $\rho < 0$) también es posible con la estadística de Durbin-Watson. En este caso, el valor de la estadística es $4 - d$, donde d se calcula de acuerdo con (13.32). El procedimiento de decisión es igual al ya dado, comparando $4 - d$ con d_L o d_U. En cualquier caso, si la prueba es no concluyente, la alternativa que se sugiere es tomar más observaciones.

Para el ejemplo se calcula d primero, con lo que se obtienen las diferencias $e_t - e_{t-1}$, mediante el uso de la columna de residuos dada en el listado de computadora. Estas diferencias son las siguientes:

t	2	3	4	5	6
$e_t - e_{t-1}$	0.2446	−0.3864	−0.0666	0.1939	0.0090
t	7	8	9	10	11
$e_t - e_{t-1}$	0.0056	−0.2119	−0.0672	−0.1094	−0.0044
t	12	13	14	15	16
$e_t - e_{t-1}$	0.1966	0.0408	0.2770	−0.0540	0.0463

Mediante el empleo de (13.32) se obtiene

$$d = 0.434789/0.5157 = 0.843.$$

Por ejemplo, $\alpha = 0.05$; entonces para el modelo lineal simple (13.30) y $n = 16$, los límites son $d_L = 1.10$ y $d_U = 1.37$. Dado que $d < d_L$, se rechaza la hipótesis nula y se concluye que existe una razón para creer que los errores en (13.30) se encuentran autocorrelacionados.

13.9.3 Eliminación de la autocorrelación mediante la transformación de datos

Cuando se rechaza la hipótesis nula de autocorrelación cero, debe ajustarse la ecuación estimada de regresión para compensar la presencia de errores autocorrelacionados. A continuación se mostrará un enfoque debido a Cochrane y Orcutt.* Se basa en un método iterativo el cual incluye la transformación de las variables respuesta y predicción en el modelo original de regresión.

Para el modelo dado por (13.30), considérese la transformación

$$Y'_t = Y_t - \rho Y_{t-1}. \tag{13.33}$$

Al sustituir en (13.33) Y_t y Y_{t-1}, de acuerdo con (13.30), se tiene que

$$\begin{aligned} Y'_t &= (\beta_0 + \beta_1 x_t + \varepsilon_t) - \rho(\beta_0 + \beta_1 x_{t-1} + \varepsilon_{t-1}) \\ &= \beta_0(1 - \rho) + \beta_1(x_t - \rho x_{t-1}) + (\varepsilon_t - \rho\varepsilon_{t-1}). \end{aligned}$$

Pero de (13.31)

$$\varepsilon_t - \rho\varepsilon_{t-1} = \eta_t$$

donde η_t son errores aleatorios no correlacionados. Entonces

$$Y'_t = \beta_0(1 - \rho) + \beta_1(x_t - \rho x_{t-1}) + \eta_t,$$

o

$$Y'_t = \beta'_0 + \beta'_1 x'_t + \eta_t, \tag{13.34}$$

donde $\beta'_0 = \beta_0(1 - \rho)$, $\beta'_1 = \beta_1$, y $x'_t = x_t - \rho x_{t-1}$. De acuerdo con lo anterior, los errores en el modelo lineal simple transformado (13.34) no están correlacionados entre sí, y de esta forma este modelo satisface las suposiciones estándar.

Nótese que las observaciones transformadas $Y'_t = Y_t - \rho Y_{t-1}$ y $x'_t = x_t - \rho x_{t-1}$ son funciones de la autocorrelación desconocida ρ, así que antes de ajustar el modelo transformado debe obtenerse un estimador de ρ. Lo anterior puede hacerse mediante el empleo de los residuos obtenidos de la ecuación de regresión estimada originalmente para calcular un estimador *MC* de la pendiente ρ en el modelo autorregresivo de primer orden dado por (13.31). Ya que este modelo tiene una intersección igual a cero, el estimador *MC*, r de la pendiente ρ basado en el análisis de la sección 13.3, es

$$r = \frac{\sum_{t=2}^{n} e_{t-1}e_t}{\sum_{t=1}^{n} e_t^2}, \tag{13.35}$$

*D. Cochrane y G. H. Orcutt, *Application of least squares regression to relationships containing autocorrelated error terms*, J. Amer. Statistical Assoc. **44** (1949), 32-61.

y los valores transformados son

$$y'_t = y_t - ry_{t-1},$$
$$x'_t = x_t - rx_{t-1}. \tag{13.36}$$

Dados los valores transformados para las variables de respuesta y predicción, el procedimiento iterativo consiste en determinar la ecuación de regresión estimada para el modelo transformado y entonces volver a calcular la estadística de Durbin-Watson. Si no es posible rechazar la hipótesis nula de autocorrelación cero, el procedimiento llega a su fin. De otra forma, se repite hasta que H_0 no pueda rechazarse. Si se requiere más de una iteración, entonces se sugiere buscar otros procedimientos alternativos.

Como ejemplo, el estimado *MC* de ρ para el ejemplo de las ventas es

$$r = 0.2734/0.5157 = 0.53,$$

y los valores transformados son los siguientes:

t	2	3	4	5	6	7	8	9
x'_t	115.09	118.02	124.57	127.76	134.32	136.12	136.39	139.12
Y'_t	19.20	19.21	20.50	21.29	22.36	22.68	22.51	23.03

t	10	11	12	13	14	15	16
x'_t	136.17	142.92	135.06	146.79	145.51	153.09	154.17
Y'_t	22.44	23.69	22.50	24.50	24.53	25.67	25.94

El listado de computadoras que se obtiene mediante el empleo de Minitab* para el modelo transformado se muestra en la figura 13.8. Nótese que el listado también incluye el valor de $d = 1.61$ para la estadística de Durbin-Watson; Minitab proporciona este valor como parte del listado. Para $n = 15$ y $\alpha = 0.05$, se obtienen los límites $d_L = 1.08$ y $d_U = 1.36$ al consultar la tabla *K*. Dado que $d > d_U$, no es posible rechazar la hipótesis nula de autocorrelación cero.

Ahora, es necesario escribir la ecuación de regresión estimada en términos de las variables originales y ajustar las desviaciones estándar estimadas de B_0 y B_1 para reflejar la eliminación de los errores autocorrelacionados. Dado que $\beta'_0 = \beta_0(1 - \rho)$ y $\beta'_1 = \beta_1$, los estimadores *MC* de β_0 y β_1 son

$$b_0 = \frac{b'_0}{(1 - r)} = \frac{-1.5178}{(1 - 0.53)} = -3.2294,$$

y

$$b_1 = b'_1 = 0.1774.$$

* Se ha omitido una porción del listado que incluye los valores de las variables de respuesta y predicción, residuos, etc.

LA ECUACION DE REGRESION ES
Y = −1.52 + 0.177 X1

	COLUMNA	COEFICIENTE	DEV. EST. DEL COEF.	COCIENTE-T = COEF/D.E.
	−	−1.5178	0.5176	−2.93
X1	C2	0.177407	0.003784	46.88

LA DEV. EST. DE T CON RESPECTO A LA RECTA DE REGRESION ES
S = 0.1627
CON (15 − 2) = 13 GRADOS DE LIBERTAD

R-CUADRADO = 99.4%

ANALISIS DE VARIANZA

DEBIDA A	DF	SC	CM = SC/GL
REGRESION	1	58.16086	58.16086
RESIDUO	13	0.34401	0.02646
TOTAL	14	58.50485	

ESTADISTICA DE DURBIN-WATSON = 1.61

FIGURA 13.8 Análisis de regresión lineal después de la transformación de los datos por autocorrelación

Para los estimadores B_0' y B_1' del listado de la figura 13.8, se nota que sus desviaciones estándar estimadas son $s(B_0') = 0.5176$ y $s(B_1') = 0.003784$. Por lo tanto, para las desviaciones estándar estimadas de B_0 y B_1, se tiene

$$s(B_0) = s\left[\frac{B_0'}{(1-r)}\right] = s(B_0')/(1-r) = 1.1013,$$

$$s(B_1) = s(B_1') = 0.003784.$$

En la tabla 13.7 se encuentra un resumen de la información pertinente para las ecuaciones de regresión estimadas original y final para los datos de ventas. Nótese que a pesar de que el cambio en los valores estimados de los coeficientes es pequeño, existe un considerable aumento en las desviaciones estándar estimadas de B_0 y B_1, y en

TABLA 13.7 Resumen de la información para los datos de ventas

Ecuación original estimada	*Ecuación final estimada*
$\hat{y}_t = -2.9716 + 0.1765x_t$	$\hat{y}_t = -3.2294 + 0.1774x_t$
$s(B_0) = 0.7023, \quad s(B_1) = 0.002456$	$s(B_0) = 1.1013, \quad s(B_1) = 0.003784$
CME = 0.0368	CME = 0.0265
$r^2 = 0.997$	$r^2 = 0.994$

forma especial para B_1. Pero la varianza residual (*CME*) ha disminuido. En este ejemplo, la autocorrelación aparente no fue lo suficientemente fuerte como para causar diferencias sustanciales en la inferencia. Cuando ocurre lo contrario, es probable que se noten diferencias muy drásticas.

13.10 Enfoque matricial para el modelo lineal simple

El uso del álgebra de matrices proporciona un medio conveniente para el análisis de regresión de modelos lineales, en forma especial de aquellos que contienen más de una variable de predicción. Se ilustrará el uso del álgebra de matrices mediante el examen del modelo lineal simple. Para una breve revisión de los fundamentos del álgebra de matrices, se invita al lector a que consulte el apéndice que se encuentra al final de este capítulo.

Para los n pares $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$, el siguiente modelo lineal simple

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, 2, \ldots, n.$$

En otras palabras,

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 x_n + \varepsilon_n \end{aligned}$$

son n ecuaciones lineales para las que $Y_1, Y_2, \ldots, Y_n$ son las observaciones de la respuesta para los correspondientes valores fijos $x_1, x_2, \ldots, x_n$ de la variable de predicción, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ son los errores aleatorios no observables y β_0 y β_1 son los parámetros por estimarse. Si se definen las matrices

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

entonces

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_n + \varepsilon_n \end{bmatrix}.$$

Como resultado se tiene que el modelo lineal simple puede expresarse en la notación de matrices

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{13.37}$$

Si se supone el caso de la teoría normal, entonces $\boldsymbol{\varepsilon}$ es un vector de variables aleatorias normales, tal que

$$E(\boldsymbol{\varepsilon}) = \mathbf{0},$$

$$Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

donde σ^2 es la varianza del error, común a todos ellos, e $\mathbf{I}$ es la matriz de identidad correspondiente.

Ahora, considérese la estimación de mínimos cuadrados de β_0 y β_1. Recuérdese que las ecuaciones normales están dadas por (13.4). Dado que

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix} \tag{13.38}$$

y

$$\mathbf{X'Y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum x_i Y_i \end{bmatrix}, \tag{13.39}$$

entonces

$$(\mathbf{X'X})\mathbf{B} = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix} \begin{bmatrix} \mathrm{B}_0 \\ \mathrm{B}_1 \end{bmatrix} = \begin{bmatrix} nB_0 & + B_1\Sigma x_i \\ B_0\Sigma x_i & + B_1\Sigma x_i^2 \end{bmatrix} = \begin{bmatrix} \Sigma Y_i \\ \Sigma x_i Y_i \end{bmatrix} = \mathbf{X'Y}.$$

Por lo tanto, las ecuaciones normales en forma matricial son

$$(\mathbf{X'X})\mathbf{B} = \mathbf{X'Y}, \tag{13.40}$$

donde

$$\mathbf{B} = \begin{bmatrix} B_0 \\ B_1 \end{bmatrix}$$

es el vector que contiene los estimadores de mínimos cuadrados B_0 y B_1.

Si se supone que la matriz cuadrada $\mathbf{X'X}$ tiene inversa, entonces en (13.40)

$$(\mathbf{X'X})^{-1}(\mathbf{X'X})\mathbf{B} = (\mathbf{X'X})^{-1}\mathbf{X'Y},$$

o

$$\mathbf{IB} = (\mathbf{X'X})^{-1}\mathbf{X'Y},$$

y

$$\mathbf{B} = (\mathbf{X'X})^{-1}\mathbf{X'Y} \tag{13.41}$$

es la expresión matricial para obtener los estimadores de mínimos cuadrados B_0 y B_1.

Al emplear los datos correspondientes al ejemplo de salarios inciales, se ilustrará que la expresión dada por (13.41) proporciona los mismos estimadores de mínimos

cuadrados para β_0 y β_1 obtenidos con anterioridad. El vector **Y** de salarios iniciales y la matriz **X** correspondiente a las calificaciones promedio son

$$\mathbf{Y} = \begin{bmatrix} 18.5 \\ 20.0 \\ 21.1 \\ \vdots \\ 16.8 \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & 2.95 \\ 1 & 3.20 \\ 1 & 3.40 \\ \vdots & \vdots \\ 1 & 2.75 \end{bmatrix}.$$

El lector debe notar que los números uno que se encuentran en la primera columna de **X** representan la intersección β_0 definida de acuerdo con el modelo lineal simple propuesto. Al seguir con el cálculo se tiene

$$(\mathbf{X}'\mathbf{X}) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 2.95 & 3.20 & \cdots & 2.75 \end{bmatrix} \begin{bmatrix} 1 & 2.95 \\ 1 & 3.20 \\ \vdots & \vdots \\ 1 & 2.75 \end{bmatrix} = \begin{bmatrix} 15 & 45.6 \\ 45.6 & 139.51 \end{bmatrix}$$

y

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 2.95 & 3.20 & \cdots & 2.75 \end{bmatrix} \begin{bmatrix} 18.5 \\ 20.0 \\ \vdots \\ 16.8 \end{bmatrix} = \begin{bmatrix} 270.8 \\ 830.425 \end{bmatrix}$$

La inversa de la matriz de 2×2 es igual a

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{13.29} \begin{bmatrix} 139.51 & -45.6 \\ -45.6 & 15 \end{bmatrix}$$

Para evitar la posibilidad de graves errores por redondeo, lo mejor es no dividir cada elemento de $(\mathbf{X}'\mathbf{X})^{-1}$ por el valor 13.29 hasta que se efectúe el producto $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Entonces, de (13.41) los estimadores de mínimos cuadrados son

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \frac{1}{13.29} \begin{bmatrix} 139.51 & -45.6 \\ -45.6 & 15 \end{bmatrix} \begin{bmatrix} 270.8 \\ 830.425 \end{bmatrix}$$

$$= \frac{1}{13.29} \begin{bmatrix} -88.072 \\ 107.895 \end{bmatrix} = \begin{bmatrix} -6.6269 \\ 8.1185 \end{bmatrix},$$

o $b_0 = -6.6269$ y $b_1 = 8.1185$. Al redondear a dos dígitos significativos, estos valores son iguales a los ya obtenidos con anterioridad.

Referencias

1. G. E. P. Box and G. M. Jenkins, *Time series analysis: Forecasting and control*, 2nd ed., Holden–Day, San Francisco, 1977.
2. S. Chatterjee and B. Price, *Regression analysis by example*, Wiley, New York, 1977.
3. N. R. Draper and H. Smith, *Applied regression analysis*, 2nd ed., Wiley, New York, 1981.
4. C. R. Nelson, *Applied time series analysis for managerial forecasting*, Holden–Day, San Francisco, 1977.
5. J. Neter and W. Wasserman, *Applied linear statistical models*, Richard D. Irwin, Homewood, Ill., 1974.

Ejercicios

13.1. Formúlese un comentario con respecto a la causalidad para las siguientes situaciones:

a) Durante los pasados 12, de 15 años, el mercado de valores creció cuando el promedio global de la liga mayor de beisbol disminuyó y viceversa.

b) Desde el primer Super Tazón en 1967, el mercado de valores aumentó durante todos los años en los que el equipo que ganaba el Super Tazón provenía de la vieja Liga Nacional de futbol, y disminuyó durante los años en lo que el campeón era un equipo de la vieja Liga Americana de futbol.

13.2. De los siguientes modelos, ¿cuáles son lineales?

a. $Y = \beta \operatorname{sen}(x) + \varepsilon$
b. $Y = \beta_1 \operatorname{sen}(\beta_2 x) + \varepsilon$
c. $Y = \beta_0 + \beta_1 x_1^2 x_2 + \beta_2 x_2^3 + \varepsilon$
d. $Y = \beta_0 + \beta_1^2 x + \varepsilon$

13.3. Dado el modelo lineal $Y_i = \beta x_i + \varepsilon_i$, $i = 1, 2, \ldots, n$, supóngase que $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ para toda i y $Cov(\varepsilon_i, \varepsilon_j) = 0$ para toda $i \neq j$.

a) Obténgase el estimador **B** de mínimos cuadrados para β.

b) Detérmínese si **B** es un estimador no sesgado de β, y demuéstrese que $Var(B) = \sigma^2 / \Sigma x_i^2$.

13.4. Una compañía local de energía seleccionó una residencia típica para desarrollar un modelo empírico para el consumo de energía (en kilowatts por día) como una función de la temperatura promedio diaria durante los meses de invierno. Se obtuvo la siguiente información durante un periodo de 15 días.

Temperatura (°C)	0	8	7.5	13.5	14	8.5	4.5	−11
Consumo de energía	70	57	60	63	57	66	67	107
Temperatura (°C)	−7.5	−8.5	1.5	0.5	2	−6	−4	
Consumo de energía	96	88	80	64	79	82	97	

a) Grafíquense los datos. ¿Sugiere la gráfica una asociación lineal?

b) Para un modelo lineal simple, obténgase la ecuación estimada de regresión y grafíquese sobre la gráfica de la parte *a*.

c) Interprétense los coeficientes de regresión estimados.
d) ¿Qué se recomendaría a la compañía para mejorar el modelo empírico?

13.5. Una compañía de seguros desea determinar el grado de relación que existe entre el ingreso familiar x y el monto del seguro de vida Y del jefe de familia. Con base en una muestra aleatoria de 18 familias, se obtuvo la siguiente información (en miles de dólares).

Ingreso	45	20	40	40	47	30	25	20	15
Seguro de vida	70	50	60	50	90	55	55	35	40
Ingreso	35	40	55	50	60	15	30	35	45
Seguro de vida	65	75	105	110	120	30	40	65	80

Repítanse todos los incisos del ejercicio 13.4.

13.6. Dada la ecuación de regresión estimada para el ejercicio 13.4:

a) Calcúlense los residuos.
b) Verifíquese que se cumplen las propiedades 2 y 3 de la sección 13.3.
c) Obténgase la varianza residual.
d) Calcúlense los estimadores de las desviaciones estándar de B_0 y B_1.
e) Obténgase un intervalo estimado de confianza del 95% para el valor real de la pendiente.
f) Determínese si una relación lineal entre la temperatura atmosférica promedio y el consumo de energía es estadísticamente discernible para un nivel $\alpha = 0.05$
g) Para cada temperatura atmosférica, calcúlense los intervalos de confianza del 95% estimados para el uso medio de energía y grafíquense éstos contra la recta estimada de regresión.

13.7. Repítanse todos los incisos del ejercicio 13.6 para la ecuación de regresión estimada del ejercicio 13.5.

13.8. Con respecto al ejercicio 13.4, estímense los consumos individuales de energía para las siguientes temperaturas: -10, -8, -5, -2, 1, 4, 7, 10, y 13. Obténganse intervalos de predicción del 95% para las estimaciones.

13.9. Con respecto al ejercicio 13.5, estímense los montos individuales del seguro de vida para los ingresos anuales de 18, 28, 38, 48 y 58 y obténganse intervalos de predicción del 95% para sus estimaciones.

13.10. Mediante el empleo de los datos de los ejercicios 13.4 y 13.5

a) Llévese a cabo un análisis de varianza para cada conjunto de datos y determínese si se puede rechazar la hipótesis nula de no regresión lineal a un nivel de $\alpha = 0.05$.
b) Compárense los resultados de la parte *a* con los que se obtienen en la parte *f* del ejercicio 13.6. Formúlese un comentario sobre la relación entre el valor de la estadística F, calculado aquí, con el de la estadística T determinado en la parte *f* del ejercicio 13.6.
c) Calcúlense los coeficientes de determinación y explíquese su significado. ¿Puede concluirse que las verdaderas ecuaciones de regresión entre la temperatura y el consumo de energía, o entre el ingreso anual y el monto del seguro de vida, son estrictamente lineales?

13.11. Los siguientes datos son las alturas X y los pesos Y de una muestra aleatoria de 10 empleados del sexo femenino de una gran empresa.

Altura (pulgadas)	68	67	65	68	64	67	66	65	64	66
Peso (libras)	119	118	129	135	123	140	125	132	118	130

a) Grafíquense los datos.

b) Calcúlese el coeficiente de correlación de la muestra y formúlese un comentario sobre cualquier linealidad aparente entre la altura y el peso.

13.12. Los siguientes datos* representan la potencia diaria, en megawatts, generada por una central eléctrica de servicio regional, durante el mes de agosto de 1980, y la temperatura atmosférica en grados Fahrenheit registrada a las 11 a.m. en una localidad central.

Temperatura	99	99	99	99	99	96	96	97	97
Potencia	153.4	141.0	143.1	156.8	158.7	158.5	158.7	159.6	148.3
Temperatura	97	99	94	91	97	96	85	79	76
Potencia	137.8	160.0	154.0	142.2	149.4	147.9	114.2	94.7	112.5
Temperatura	84	90	76	78	81	90	93	90	96
Potencia	123.6	131.1	119.4	111.9	103.5	103.7	125.4	129.0	135.6
Temperatura	98	95	95	95					
Potencia	142.3	142.5	128.9	124.3					

a) Grafíquense los datos.

b) Calcúlese el coeficiente de correlación de la muestra y formúlese un comentario sobre cualquier linealidad aparente entre la temperatura y la cantidad de potencia generada.

13.13. Supóngase que se sabe que la curva de regresión entre una respuesta Y y una variable de predicción x es lineal. Para estimar la ecuación de regresión, se toman $n/2$ observaciones de Y en el extremo inferior del intervalo de valores de x y $n/2$ observaciones en el extremo superior de x. Por conveniencia, los valores extremos de x se han escalado a -1 y $+1$.

a) Empléese la ecuación (13.18) para obtener el intervalo para la varianza de la respuesta media para cualquier punto x_p de x que se encuentre dentro del intervalo $(-1, 1)$.

b) Úsese la ecuación (13.20) para obtener una expresión similar para la varianza de una respuesta en particular.

c) Supóngase que se registran las siguientes observaciones:

x	-1	-1	-1	-1	-1	1	1	1	1	1
Y	10	12	9	13	8	20	17	23	24	19

* Cortesía de K. L. Fugett.

Úsese álgebra de matrices para obtener las estimaciones de mínimos cuadrados para la pendiente y la intersección.

13.14. Supóngase que la siguiente información sobre el ingreso anual bruto x y el porcentaje de impuestos pagados Y, proviene de una muestra aleatoria de 14 declaraciones de impuestos.

Ingreso bruto (miles de dólares)	25.6	42.2	57.6	98.8	10.4	30.1	40.0
Porcentaje pagado en impuesto	15.4	16.8	19.7	21.7	10.8	15.2	18.9
Ingreso bruto (miles de dólares)	29.3	16.1	18.0	88.2	34.0	22.1	70.0
Porcentaje pagado en impuesto	15.9	12.0	14.1	21.1	17.6	14.8	21.6

a) Grafíquense los datos. ¿Sugiere esta gráfica una asociación lineal?
b) Mediante la suposición de un ajuste lineal, estímese la ecuación de regresión y dibújese la recta sobre la gráfica de la parte *a*.
c) Realícese un análisis de varianza, obténganse los coeficientes de determinación y coméntese si se puede pensar que la ecuación de regresión estimada proporciona una predicción apropiada. Úsese $\alpha = 0.05$.
d) Predígase el porcentaje promedio de impuestos que pagará la gente con ingresos brutos de 15 y 85 mil dólares y obténganse las estimaciones de sus desviaciones estándar.

13.15. El gerente de una industria desea determinar si existe una relación lineal entre el número de unidades Y, armadas por los operadores de una línea de ensamble, y el lapso x que transcurre antes de que se presente una falla. Con base en una muestra aleatoria de operadores de la línea de ensamble, se observa la siguiente información:

Tiempo (en horas)	1	2	3	4
Unidades ensambladas	25, 29, 23, 31	55, 65, 63, 59	73, 75, 74, 71	90, 88, 91, 87

a) Grafíquense los datos y coméntese el resultado.
b) Estímese una ecuación de regresión lineal mediante el uso del álgebra de matrices.
c) Determínese si la relación lineal es estadísticamente discernible para un nivel $\alpha = 0.01$
d) Obténgase un intervalo de confianza del 95% para la pendiente.

13.16. Los siguientes datos muestran el porcentaje de la población con cuatro o más años de educación superior x, y la tasa de mortalidad infantil por cada 1 000 nacimientos Y para una muestra de 15 estados.*

a) Grafíquense los datos y calcúlese el coeficiente de correlación de la muestra.
b) Ajústese una función de regresión lineal con la tasa de mortalidad como la respuesta y el porcentaje de la población con cuatro o más años de educación superior como la variable de predicción. Interprétese el coeficiente de regresión estimado para la pendiente.

* *Hammond almanac, 1981.*

x	19.4	12.3	13.7	11.0	11.5	16.8	11.8	12.8
Y	12.0	15.4	16.0	14.2	17.9	11.9	14.2	12.7
x	15.3	11.8	11.7	10.4	17.5	15.6	16.1	
Y	13.8	15.8	13.7	17.6	10.1	10.1	12.1	

c) La regresión lineal, ¿es estadísticamente discernible para un nivel $\alpha = 0.05$ ¿Cómo podría explicarse cualquier asociación lineal que existiese entre estas dos cantidades?

13.17. Los datos* que figuran en la tabla 13.8 consisten en información anual sobre los precios relativos del alcohol x y el consumo *per cápita* en litros de alcohol absoluto Y para el periodo 1948-1967 en Ontario.

a) Grafíquense los datos y calcúlese el coeficiente de correlación de la muestra.
b) Mediante el empleo del análisis de varianza, determínese si la regresión lineal entre el precio relativo y el consumo *per cápita* es estadísticamente discernible para un nivel $\alpha = 0.01$

13.18. Se llevó a cabo un estudio para determinar la relación entre el número de años de experiencia a x y el salario anual Y para una profesión en particular en una región geográfica dada. Se seleccionó una muestra aleatoria de 17 personas, las cuales ejercen esta profesión, y se obtuvo la siguiente información:

TABLA 13.8 Datos de la muestra para el ejercicio 13.17

Año	*Precio relativo*	*Consumo per cápita*
1948	0.057	7.09
1949	0.058	7.18
1950	0.055	7.23
1951	0.052	7.23
1952	0.051	7.32
1953	0.055	7.64
1954	0.056	7.73
1955	0.047	7.55
1956	0.045	7.91
1957	0.044	7.86
1958	0.043	7.96
1959	0.043	7.77
1960	0.043	8.14
1961	0.043	8.14
1962	0.041	8.23
1963	0.040	8.46
1964	0.039	8.73
1965	0.038	8.77
1966	0.039	9.18
1967	0.035	8.91

* R.E. Popham, W. Schmidt, y J. de Lint, *The prevention of alcoholism: Epidemiological studies of the effects of government control measures,* Brit. J. of Addiction 70 (1975), 125—144.

Años de experiencia	13	16	30	2	8	31	19	20	1
Salario anual actual (miles de dólares)	26.1	33.2	36.1	16.5	26.4	36.4	33.8	36.5	16.9
Años de experiencia	4	27	25	7	15	13	6	10	
Salario anual actual (miles de dólares)	19.8	36.0	36.5	21.4	31.0	31.4	19.1	24.6	

a) Grafíquense los datos y, con base en esta gráfica, determínese si un ajuste lineal es suficiente.
b) Ajústese un modelo lineal e interprétense los coeficientes de regresión estimados.
c) ¿Puede rechazarse la hipótesis nula de pendiente cero para un nivel $\alpha = 0.01$?
d) Estímese el salario promedio para una persona que ejerce esta profesión, la cual tiene 12 años de experiencia; además, calcúlese un intervalo de confianza del 99% para este valor.
e) Obténganse los residuos y grafíquense contra los correspondientes años de experiencia, ¿se observa algo fuera de lo común? Explíquese.

13.19. Los siguientes datos representan el producto nacional bruto x y los gastos de consumo Y en miles de millones de dólares en 1972, para los años 1960-1980.*

Año	1960	1961	1962	1963	1964	1965	1966
x	737.2	756.6	800.3	832.5	876.4	929.3	984.8
Y	452.0	461.4	482.0	500.5	528.0	557.5	585.7

Año	1967	1968	1969	1970	1971	1972	1973
x	1 011.4	1 058.1	1 087.6	1 085.6	1 122.4	1 185.9	1 255.0
Y	602.7	634.4	657.9	672.1	696.8	737.1	768.5

Año	1974	1975	1976	1977	1978	1979	1980
x	1 248.0	1 233.9	1 300.4	1 371.7	1 436.9	1 483.0	1 480.7
Y	763.6	780.2	823.7	863.9	904.8	930.9	935.1

a) Ajústese un modelo lineal e interprétense los coeficientes de regresión estimados.
b) Hágase una gráfica de los residuos estandarizados contra el tiempo. ¿Se puede detectar algún patrón?
c) Calcúlese la estadística de Durbin-Watson y determínese si los errores se encuentran positivamente autocorrelacionados. Úsese $\alpha = 0.05$.
d) Si la autocorrelación positiva es estadísticamente discernible, ajústese la ecuación de regresión estimada mediante la transformación de los datos.

13.20. Los siguientes datos representan las ganancias de las empresas por inventario y ajustes al capital x y los impuestos sobre estas ganancias Y en miles de millones de dólares para los años 1960-1980.* Repítanse todas las partes del ejercicio 13.19.

Año	1960	1961	1962	1963	1964	1965	1966
x	47.6	48.6	56.6	62.1	69.2	80.0	85.1
Y	22.7	22.8	24.0	26.2	28.0	30.9	33.7

* *Economic report of the president* february 1982.

Año	1967	1968	1969	1970	1971	1972	1973
x	82.4	89.1	85.1	71.4	83.2	96.6	108.3
Y	32.5	39.2	39.5	34.2	37.5	41.6	49.0
Año	**1974**	**1975**	**1976**	**1977**	**1978**	**1979**	**1980**
x	94.9	110.5	138.1	164.7	185.5	196.8	182.7
Y	51.6	50.6	63.8	72.6	83.0	87.6	82.3

APÉNDICE

Breve revisión del álgebra de matrices

Una matriz es un arreglo rectangular de elementos en renglones y columnas. Por ejemplo,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mj} & \cdots & x_{mn} \end{bmatrix}$$

es una matriz que contiene **m** renglones y n columnas. Las entradas x_{ij}, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, son los elementos de la matriz **X.** El primer índice (i) identifica el renglón en el que se encuentra el elemento, y el segundo (j) la columna a la que pertenece. La matriz **X** de **m** renglones y n columnas se conoce como una matriz de orden (o dimensión) m por n. En general, una matriz se denota por una letra mayúscula en negritas, mientras que la correspondiente letra minúscula designa a un elemento de ésta. Es una práctica común utilizar la siguiente notación abreviada:

$$\mathbf{X} = [x_{ij}], \qquad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n$$

para designar a la matriz **X** de dimensión $m \times n$.

Una matriz que contiene sólo una columna recibe el nombre *vector columna,* y una matriz formada por un renglón *vector renglón.* Las matrices

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{Z}' = [z_1 \quad z_2 \quad \cdots \quad z_n]$$

son ejemplos de vectores columna y renglón, respectivamente; **Y** es un vector columna de $n \times 1$, y $\mathbf{Z}'$ es un vector renglón de $1 \times n$. La razón para emplear el símbolo

de virgulilla en el vector renglón $\mathbf{Z}'$ se explicará en forma breve. Ya que un vector columna o renglón tienen sólo un renglón o una sola columna, únicamente es necesario usar una notación para identificar la posición de los elementos.

Una matriz que tiene el mismo número de renglones que de columnas recibe el nombre de *matriz cuadrada*.

$$\mathbf{A} = \begin{bmatrix} 3 & -2 & 1 \\ 0 & 2 & 4 \\ -3 & 1 & 5 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 10 & -3 \\ 2 & 1 \end{bmatrix}$$

son ejemplo de matrices cuadradas. $\mathbf{A}$ es una matriz cuadrada de 3×3, y $\mathbf{B}$ es una matriz cuadrada de 2×2.

El intercambio de los renglones y las columnas de una matriz $\mathbf{X}$ de $m \times n$ da origen a una nueva matriz denotada por $\mathbf{X}'$, de dimensión $n \times m$ que recibe el nombre de *transpuesta* de $\mathbf{X}$. Por ejemplo, dada la matriz

$$\mathbf{X} = \begin{bmatrix} 2 & -3 & 0 \\ 1 & 5 & -12 \end{bmatrix},$$

la transpuesta de $\mathbf{X}$ es la matriz cuya primera columna es igual al primer renglón de $\mathbf{X}$ y cuya segunda columna es igual al segundo renglón de $\mathbf{X}$,

$$\mathbf{X}' = \begin{bmatrix} 2 & 1 \\ -3 & 5 \\ 0 & -12 \end{bmatrix}.$$

En general, dada

$$\mathbf{X} = [x_{ij}], \qquad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n,$$

se tiene

$$\mathbf{X}' = [x_{ji}], \qquad j = 1, 2, \ldots, n, \quad i = 1, 2, \ldots, m.$$

En otras palabras, el elemento en el i-ésimo renglón y la j-ésima columna de $\mathbf{X}$ se encuentra en el j-ésimo renglón y la i-ésima columna de la matriz transpuesta $\mathbf{X}'$. La transpuesta de un vector columna es un vector renglón y viceversa. Por esta razón se acostumbra emplear el símbolo de virgulilla para denotar un vector renglón.

Se dice que dos matrices son iguales sólo si sus correspondientes elementos son iguales. De esta forma, una condición necesaria para que dos matrices sean iguales es que tengan la misma dimensión. Por ejemplo, las dos matrices

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} -2 & 0 \\ 5 & 6 \\ 12 & -5 \end{bmatrix}$$

son iguales si

$$\begin{aligned} a_{11} &= -2 & a_{12} &= 0 \\ a_{21} &= 5 & a_{22} &= 6 \\ a_{31} &= 12 & a_{32} &= -5. \end{aligned}$$

La suma o diferencia entre dos matrices sólo es posible cuando sus dimensiones son las mismas. La suma (diferencia) de dos matrices es una matriz cuyos elementos son las sumas (diferencias) de los correspondientes elementos de las dos matrices. Por ejemplo, dadas

$$\mathbf{A} = \begin{bmatrix} -2 & 5 \\ 3 & 8 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 4 & -3 \\ 2 & -6 \end{bmatrix},$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} -2+4 & 5-3 \\ 3+2 & 8-6 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 5 & 2 \end{bmatrix},$$

$$\mathbf{A} - \mathbf{B} = \begin{bmatrix} -2-4 & 5-(-3) \\ 3-2 & 8-(-6) \end{bmatrix} = \begin{bmatrix} -6 & 8 \\ 1 & 14 \end{bmatrix}.$$

Dadas dos matrices **A** y **B**, la matriz producto **AB** se define sólo si el número de columnas de **A** es igual al número de renglones de **B**. Entonces, si **A** es de $m \times n$ y **B** es de $n \times p$, el producto **AB** es una matriz de dimensión $m \times p$ para la que el elemento que se encuentra en el i-ésimo renglón y la j-ésima columna es igual a la suma de los productos de los elementos que se encuentran en el i-ésimo renglón de **A** y la j-ésima columna de **B**. Si

$$\mathbf{A} = \begin{bmatrix} 1 & -2 \\ -3 & 4 \\ 0 & -1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} -2 & 1 \\ 4 & 3 \end{bmatrix},$$

$$\mathbf{AB} = \begin{bmatrix} 1 & -2 \\ -3 & 4 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} -2 & 1 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} (1)(-2)+(-2)(4) & (1)(1)+(-2)(3) \\ (-3)(-2)+(4)(4) & (-3)(1)+(4)(3) \\ (0)(-2)+(-1)(4) & (0)(1)+(-1)(3) \end{bmatrix}$$

$$= \begin{bmatrix} -10 & -5 \\ 22 & 9 \\ -4 & -3 \end{bmatrix}.$$

Nótese que para este par de matrices el producto **BA** no está definido; en general, la multiplicación de matrices no es conmutativa. También es interesante notar que si **Y** es un vector columna de $n \times 1$ y **Y**′ es un vector renglón de $1 \times n$, entonces **YY**′ es una matriz cuadrada de dimensión n y **Y**′**Y** es un escalar. Un *escalar* es cualquier número de la recta real $(-\infty, \infty)$. La multiplicación de una matriz por un escalar da origen a una matriz cuyos elementos son los productos de los correspondientes elementos originales y la cantidad escalar. Por ejemplo, dada

$$\mathbf{A} = \begin{bmatrix} -2 & 1 & -2 \\ 3 & 4 & 1 \end{bmatrix},$$

$$-5\mathbf{A} = \begin{bmatrix} (-5)(-2) & (-5)(1) & (-5)(-2) \\ (-5)(3) & (-5)(4) & (-5)(1) \end{bmatrix} = \begin{bmatrix} 10 & -5 & 10 \\ -15 & -20 & -5 \end{bmatrix}.$$

Existen ciertas matrices especiales que vale la pena mencionar. Una matriz cuadrada de dimensión n cuyos elementos son cero excepto los que se encuentran sobre la diagonal principal,* elementos iguales a uno, recibe el nombre de *matriz identidad* de orden n. Por ejemplo,

$$\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

son matrices identidad de orden 3 y 2, respectivamente. En la multiplicación de matrices cuadradas, la matriz identidad juega el mismo papel que el número 1 tiene en la multiplicación entre escalares. Esto es, dada cualquier matriz $\mathbf{A}$, el producto de la correspondiente matriz identidad y $\mathbf{A}$ da como resultado la matriz $\mathbf{A}$, siempre que exista compatibilidad para llevar a cabo la multiplicación. De esta forma,

$$\mathbf{IA} = \mathbf{AI} = \mathbf{A}.$$

Se dice que una matriz cuadrada es *simétrica,* si es igual a su transpuesta. Dada cualquier matriz cuadrada $\mathbf{A}$, si $\mathbf{A} = \mathbf{A}'$, entonces $\mathbf{A}$ es simétrica. Por ejemplo,

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & -2 \\ 1 & 4 & 3 \\ -2 & 3 & 1 \end{bmatrix}$$

es una matriz simétrica. Nótese que los elementos que se encuentran formando un triángulo por debajo de la diagonal principal son idénticos a los correspondientes en el triángulo que se encuentra por encima de la diagonal principal. Si una matriz $\mathbf{A}$ de $m \times n$ se premultiplica por su transpuesta, la matriz producto será simétrica. De esta forma, $\mathbf{A}'\mathbf{A}$ es una matriz simétrica de orden n. Por ejemplo, dadas

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 4 \\ 3 & 2 \end{bmatrix}, \qquad \mathbf{A}' = \begin{bmatrix} 2 & 1 & 3 \\ 1 & 4 & 2 \end{bmatrix},$$

$$\mathbf{A}'\mathbf{A} = \begin{bmatrix} 2 & 1 & 3 \\ 1 & 4 & 2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 4 \\ 3 & 2 \end{bmatrix} = \begin{bmatrix} 14 & 12 \\ 12 & 21 \end{bmatrix}.$$

Una *matriz diagonal* es cualquier matriz cuadrada para la que todos los elementos que se encuentran fuera de la diagonal principal son cero.

$$\mathbf{A} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 7 \end{bmatrix}$$

es una matriz diagonal. Debe ser evidente que la matriz identidad es un caso especial de una matriz diagonal.

* La diagonal principal contiene los elementos cuyas posiciones en el renglón y la columna son las mismas.

Un vector cero es cualquier vector columna para el que todos sus elementos son cero.

$$\mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

es un vector cero de 4×1.

A continuación se define un concepto importante en el álgebra matricial, que se conoce como la inversa de una matriz cuadrada. Sea **A** una matriz cuadrada de orden n. Si existe una matriz denotada por $\mathbf{A}^{-1}$, tal que

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

donde **I** es la correspondiente matriz identidad, entonces $\mathbf{A}^{-1}$ es la ***única matriz inversa de*** **A**. Si una matriz cuadrada tiene inversa, se dice que es ***no singular***; de otra forma, recibe el nombre de matriz ***singular***.

Para cada matriz cuadrada, es posible definir y calcular una cantidad escalar que se conoce como el *determinante* de la matriz. El valor del determinante de una matriz cuadrada es el factor para decidir si ésta tiene o no inversa. Sea **A** cualquier matriz cuadrada. Si el determinante de **A**, denotado por det(**A**) no es igual a cero, existe la matriz inversa de **A**. Si det(**A**) = 0, entonces **A** es singular. La noción de una matriz inversa es el análogo del inverso multiplicativo en el álgebra de escalares.

Como ilustración, se encontrará la inversa de matrices sólo para el caso de 2×2. En general, sea

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

cualquier matriz de 2×2. El determinante de **A** se define como

$$\det(\mathbf{A}) = a_{11}\,a_{22} - a_{12}\,a_{21}$$

y puede demostrarse que la matriz inversa de **A** está dada por

$$\mathbf{A}^{-1} = \begin{bmatrix} \dfrac{a_{22}}{\det(\mathbf{A})} & -\dfrac{a_{12}}{\det(\mathbf{A})} \\ -\dfrac{a_{21}}{\det(\mathbf{A})} & \dfrac{a_{11}}{\det(\mathbf{A})} \end{bmatrix}$$

Por ejemplo, dada

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ -1 & 1 \end{bmatrix},$$

$$\det(\mathbf{A}) = (2)(1) - (-1)(3) = 5,$$

y

$$\mathbf{A}^{-1} = \begin{bmatrix} 1/5 & -3/5 \\ 1/5 & 2/5 \end{bmatrix}$$

es la matriz inversa de **A**. Este resultado puede verificarse en forma sencilla, ya que

$$\mathbf{A}^{-1}\mathbf{A} = \begin{bmatrix} 1/5 & -3/5 \\ 1/5 & 2/5 \end{bmatrix}\begin{bmatrix} 2 & 3 \\ -1 & 1 \end{bmatrix} = \mathbf{A}\mathbf{A}^{-1}$$

$$= \begin{bmatrix} 2 & 3 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 1/5 & -3/5 \\ 1/5 & 2/5 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Para finalizar, debe notarse que la inversa de cualquier matriz diagonal también es una matriz diagonal, cuyos elementos sobre la diagonal principal son los recíprocos de los elementos que se encuentran en la diagonal principal de la matriz original. Por ejemplo. dado que

$$\mathbf{A} = \begin{bmatrix} 9 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 10 \end{bmatrix},$$

$$\mathbf{A}^{-1} = \begin{bmatrix} 1/9 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & 1/10 \end{bmatrix}.$$

CAPÍTULO CATORCE

Análisis de regresión: el modelo lineal general

14.1 Introducción

En el capítulo 13 se examinaron los fundamentos del análisis de regresión para el modelo lineal simple. En este capítulo se extenderán los conceptos ya presentados al modelo lineal general para el cual una respuesta dada se considera como una función de varias variables de predicción. Al examinar este modelo se estudiarán algunas formas para determinar el mejor conjunto de variables de predicción por incluir en la ecuación de regresión. También se proporcionará un estudio detallado del análisis de residuos (también conocidos como residuales), mínimos cuadrados con factores de peso (ponderados) y variables indicadoras, así como ejemplos resueltos con gran detalle. Para este capítulo se emplearán los paquetes estadísticos para computadoras Minitab y SAS (véase [6]). Se supone que este tipo de paquetes o algunos similares se encuentran disponibles para el lector. Para un estudio más teórico de los temas presentados en este capítulo, se invita al lector a que consulte [4].

14.2 El modelo lineal general

Sean x_1, x_2, ..., x_k k variables de predicción, las cuales pueden tener alguna influencia sobre una respuesta Y, y supóngase que el modelo tiene la forma donde Y_i es la

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \qquad (14.1)$$

i-ésima observación de la respuesta para un conjunto de valores fijos x_{i1}, x_{i2}, ..., x_{ik} de las variables de predicción, ε_i es el error aleatorio no observable asociado con Y_i, y β_0, β_1, ..., β_k son $m = k + 1$ parámetros lineales desconocidos. La ecuación (14.1) recibe el nombre de *modelo lineal general* y da origen a lo que se conoce como una *regresión lineal múltiple.*

Si se supone el caso de la teoría basada en el modelo normal, las observaciones Y_i son variables aleatorias independientes, normalmente distribuidas con

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik},$$

$$Var(Y_i) = \sigma^2, \qquad i = 1, 2, \ldots, n.$$

De esta forma, los errores aleatorios ε_i son $N(0, \sigma^2)$ independientes. El modelo lineal general define una ecuación de regresión la cual representa un hiperplano, para la que el parámetro β_0 es el valor de la respuesta media cuando todas las variables de predicción tienen un valor igual a cero. El parámetro β_j, $j = 1, 2, \ldots, k$, representa el cambio en la respuesta promedio para un cambio igual a una unidad de la correspondiente variable de predicción x_j, cuando todas las demás variables de predicción se mantienen constantes. En este sentido, β_j representa el efecto parcial de x_j sobre la respuesta.

La única restricción funcional que se impone al modelo lineal general es que sea lineal en los parámetros desconocidos; el modelo no tiene ninguna restricción con respecto a la naturaleza de las variables de predicción; por lo tanto surgen muchos casos especiales e interesantes, algunos de los cuales cabe mencionar. El modelo dado por (14.1) implica que los efectos que las variables de predición $x_1, x_2, \ldots, x_k$ tienen sobre la respuesta son aditivos, de tal manera que la ecuación de regresión propuesta es una función lineal de las variables de predicción. Una ecuación de este tipo se denomina *modelo de primer orden*. Sin embargo, es posible que dos o más variables de predicción interactúen, es decir, el efecto de una de las variables de predicción sobre la variable de respuesta depende del valor de otra variable de predicción. Cuando esto ocurre, los efectos no son aditivos debido a la presencia en el modelo de un término que contiene un producto cruzado el cual representa el efecto de interacción. Por ejemplo, considérese un modelo que contiene dos variables de predicción que interactúan. El modelo es

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i, \tag{14.2}$$

donde el sumando $\beta_3 x_{i1} x_{i2}$ refleja la interacción entre x_1 y x_2. Si se define

$$x_{i3} = x_{i1} x_{i2}, \qquad i = 1, 2, \ldots, n,$$

entonces (14.2) puede escribirse en la forma del modelo lineal general (14.1), y de esta manera se advierte que es un caso especial de éste. Nótese que para este caso especial el significado de β_1 y β_2 no es el mismo dado con anterioridad. La derivada parcial de la respuesta media con respecto a x_1 (o con respecto a x_2) representa el efecto sobre la respuesta media por unidad de cambio en x_1 (x_2) cuando x_2 (x_1) se mantiene fija. Las derivadas parciales son

$$\frac{\partial E(Y)}{\partial x_1} = \beta_1 + \beta_3 x_2$$

$$\frac{\partial E(Y)}{\partial x_2} = \beta_2 + \beta_3 x_1.$$

Otro caso interesante surge cuando en (14.1) se tiene

$$x_{ij} = x_i^j, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, k.$$

Entonces el modelo lineal general toma la forma

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_i, \tag{14.3}$$

la cual se conoce como *modelo curvilineal o polinomial.* En este caso se supone que la respuesta promedio es una función polinómica de grado k de una sola variable de predicción. Por lo tanto, la ecuación de regresión propuesta para la respuesta promedio es una función no lineal de la variable de predicción, pero sigue siendo lineal en los parámetros. Es importante notar que lo que se busca en este caso es el grado k que mejor se ajusta a una muestra aleatoria de la variable respuesta.

Para describir en forma adecuada una variable respuesta dada, muchas veces es necesario incluir términos lineales, cuadráticos y de interacción en el modelo propuesto. Por ejemplo, un modelo para dos variables de predicción podría ser

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1} x_{i2} + \varepsilon_i. \tag{14.4}$$

Al definir nuevas variables de predicción, así como se hizo anteriormente para los términos cuadráticos y de interacción, se observa que (14.4) también es un caso especial del modelo general. Este tipo de modelo se denomina *ecuación completa de segundo orden* y define varias superficies para la respuesta promedio como una función no lineal de las variables de predicción x_1 y x_2. Para $k \geq 2$ variables de predicción distintas, una ecuación de regresión completa de segundo orden consiste en un término constante, k términos lineales, k términos cuadráticos y $k(k-1)/2$ términos de interacción.

A continuación se regresará al modelo lineal general dado en (14.1) para obtener los estimadores de mínimos cuadrados de los parámetros y para desarrollar técnicas de regresión para este modelo. Todos los casos especiales mencionados con anterioridad así como muchos otros que no se citaron de manera específica, se encuentran incluidos en el siguiente análisis. Se empleará el álgebra de matrices, ya que ésta simplifica en gran medida la presentación.

Dada una muestra aleatoria de observaciones $Y_1, Y_2, \ldots, Y_n$ en los puntos de observación $x_{11}, x_1 x_{12}, \ldots, x_{1k}, x_{21}, x_{22}, \ldots, x_{2k}, \ldots, x_{n1}, x_{n2}, \ldots, x_{nk}$, respectivamente, con base en el modelo lineal general, se tienen las n ecuaciones siguientes:

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\
Y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n.
\end{aligned}$$

Como resultado, el modelo lineal general también puede expresarse en forma matricial como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{14.5}$$

donde

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

El lector no tendrá ninguna dificultad para reconocer que (14.5) tiene la misma forma matricial que el modelo lineal simple (13.37), excepto que ahora $\mathbf{X}$ es una matriz de $n \times m$ para las variables de predicción, y $\boldsymbol{\beta}$ es un vector de parámetros desconocidos de $m \times 1$, mientras que $\mathbf{Y}$ y $\boldsymbol{\varepsilon}$ siguen siendo vectores de $n \times 1$, los que contienen las observaciones de la variable de respuesta y los errores aleatorios asociados con éstas, respectivamente.

Bajo el caso de la teoría normal

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}),$$
$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}),$$

donde

$$Var(\mathbf{Y}) = Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}.$$

De esta manera $\mathbf{Y}$ y $\boldsymbol{\varepsilon}$ son vectores de variables aleatorias independientes normalmente distribuidas.

Para la estimación de los parámetros por mínimos cuadrados las ecuaciones normales toman la misma forma dada por (13.40), o

$$(\mathbf{X}'\mathbf{X})\mathbf{B} = \mathbf{X}'\mathbf{Y}$$

donde, ahora, $(\mathbf{X}'\mathbf{X})$ es un matriz de $m \times n$ y $\mathbf{B}$ es un vector de $m \times 1$ el cual contiene los estimadores de mínimos cuadrados $B_0, B_1, \ldots, B_k$. Si $(\mathbf{X}'\mathbf{X})$ tiene inversa, la solución para el vector $\mathbf{B}$ está dada por

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Por lo tanto, la ecuación estimada de regresión es

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{B}, \tag{14.6}$$

donde el vector $\hat{\mathbf{Y}}$ de $n \times 1$ contiene los valores estimados para la respuesta promedio correspondientes a los n puntos de observación de las variables de predicción. La diferencia entre los vectores $\mathbf{Y}$ y $\hat{\mathbf{Y}}$ proporciona el vector de residuos.

Puede demostrarse que las propiedades de los estimadores de mínimos cuadrados $B_0, B_1, \ldots, B_k$ son extensiones de las propiedades de los estimadores para el modelo lineal simple, es decir, de acuerdo con el caso de la teoría normal, los estimadores también son de máxima verosimilitud, de tal manera que lo siguiente se verifica:

1. Cada B_j tiene una distribución normal con media $E(B_j) = \beta_j$, $j = 0, 1, 2, \ldots, k$, y varianza $Var(B_j) = c_{(j+1)}\sigma^2$, $j = 0, 1, 2, \ldots, k$, donde $c_{(j+1)}$ es el elemento de la diagonal $(j + 1)$ de $(\mathbf{X'X})^{-1}$.
2. $Cov(B_i, B_j) = c_{(i+1),(j+1)}\sigma^2$, $i \neq j = 0, 1, 2, \ldots, k$, donde $c_{(i+1),(j+1)}$ es el elemento de $(\mathbf{X'X})^{-1}$ que se encuentra en el renglón $(i + 1)$ y la columna $(j + 1)$ para $i \neq j$.

Un estimador no sesgado de la varianza del error es

$$S^2 = \frac{\mathbf{Y'Y} - \mathbf{B'X'Y}}{n - m}, \tag{14.7}$$

donde el numerador de (14.7) no es más que la suma de los cuadrados de los residuos. Nótese que el denominador de (14.7) es igual al número de observaciones, menos el número de parámetros que figuran en el modelo, el que para el modelo lineal general es $m = k + 1$. Por lo tanto, una estimación de $Var(B_j)$ es

$$s^2(B_j) = c_{(j+1)}s^2, \qquad j = 0, 1, 2, \ldots, k,$$

donde $c_{(j+1)}$ tiene un valor igual al ya definido con anterioridad.

De los resultados anteriores puede deducirse que la cantidad

$$(B_j - \beta_j)/s(B_j), \qquad j = 0, 1, 2, \ldots, k,$$

es una variable aleatoria t de Student con $n - m$ grados de libertad. Entonces, un intervalo de confianza del $100(1 - \alpha)\%$ para el parámetro β_j es

$$b_j \pm t_{1-\alpha/2,\, n-m}s(B_j), \qquad j = 0, 1, 2, \ldots, k, \tag{14.8}$$

y una estadística apropiada para probar la hipótesis nula

$$H_0\colon \beta_j = 0$$

contra cualquier alternativa, ya sea ésta uni o bilateral, es la ya familiar t de Student

$$T = B_j/s(B_j), \qquad j = 0, 1, 2, \ldots, k,$$

con $n - m$ grados de libertad.

Considérese la técnica del análisis de varianza para probar la hipótesis nula

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

contra la alternativa

$$H_1\colon \beta_j \neq 0 \text{ para algún } j = 1, 2, \ldots, k.$$

Dado que H_0 establece que todos los parámetros de regresión son iguales a cero, excepto el término constante, esto implica que no existe ninguna relación igual a la especificada por el modelo propuesto entre la respuesta y el conjunto de variables de predicción. No obstante, se advierte al lector que el hecho de rechazar a H_0 no nece-

sariamente implica que la ecuación estimada de regresión sea útil para efectuar predicciones. Se necesita profundizar el análisis antes de que se pueda dar un juicio definitivo sobre la utilidad de la ecuación de regresión.

Al seguir el mismo argumento para el modelo lineal general que el dado para el modelo lineal simple en la sección 13.7, puede demostrarse que la suma total de cuadrados se encuentra dividida en la suma de cuadrados de la regresión y en la suma de cuadrados de los errores. Mediante el empleo de la notación matricial, *STC, SCR* y *SCE* se encuentran definidos en la tabla 14.1.

El número total de grados de libertad sigue siendo $n - 1$, pero el número de grados de libertad para el error ahora es de $n - m$. Los grados de libertad para la regresión son $(n - 1) - (n - m) = m - 1 = k$, dado que $m = k + 1$. La varianza residual o $SCE/(n - m)$ es el cuadrado medio del error y $SCR/(m - 1)$ es el cuadrado medio de la regresión. Bajo la hipótesis nula, la estadística de prueba apropiada es

$$F = \text{CMR/CME},$$

la cual tiene una distribución F con $m - 1$ y $n - m$ grados de libertad. Al igual que en los casos anteriores, puede argumentarse que si un valor de esta estadística es lo suficientemente grande, entonces una porción considerable de la variación en las observaciones puede atribuirse a la regresión de Y sobre las variables de predicción como se encuentran definidas por el modelo. De esta forma se rechaza la hipótesis nula siempre que el valor calculado se encuentre en el interior de una región crítica de tamaño α en el extremo superior de la distribución. En la tabla 14.1 se da la tabla de análisis de varianza para el modelo lineal general.

Para el modelo lineal general la noción del coeficiente de determinación se extiende para dar origen a lo que se conoce como *coeficiente de correlación múltiple* o *coeficiente de determinación múltiple.* El coeficiente de correlación múltiple se define como

$$R^2 = \frac{\text{SCR}}{\text{STC}} = 1 - \frac{\text{SCE}}{\text{STC}}, \tag{14.9}$$

y al igual que r^2, mide la proporción de la variación total de las observaciones con respecto a su media, atribuible a la ecuación de regresión estimada. En otras pala-

TABLA 14.1 Tabla ANOVA para el modelo lineal general

Fuente de variación	*Número de grados de libertad*	*Sumas de los cuadrados*	*Cuadrados medios*	*Estadística F*
Regresión	$k = m - 1$	$\mathbf{B'X'Y} - \frac{(\Sigma Y_i)^2}{n}$	$\text{SCR}/(m - 1)$	$\frac{\text{SCR}/(m - 1)}{\text{SCE}/(n - m)}$
Error	$n - m$	$\mathbf{Y'Y} - \mathbf{B'X'Y}$	$\text{SCE}/(n - m)$	
Total	$n - 1$	$\mathbf{Y'Y} - \frac{(\Sigma Y_i)^2}{n}$		

bras, R^2 es una medida relativa de qué tanto las variables de predicción incluidas en el modelo explican la variación de las observaciones. Al igual que para el modelo lineal simple, $0 \leq R^2 \leq 1$, y entre más cercano a uno es el valor de R^2 mayor es la cantidad de la variación total que puede explicarse por medio de los términos que aparecen en el modelo. Por sí mismo, R^2 no puede validar el modelo propuesto, ni tener un valor de R^2 cercano a uno necesariamente implica que la ecuación de regresión estimada sea apropiada para predicción.

Supóngase que se desea predecir la respuesta promedio cuando las k variables de predicción toman los valores específicos $x_1, x_2, \ldots, x_k$, respectivamente. En notación matricial, sea

$$\mathbf{X}_p' = [1\ x_1\ x_2 \cdots x_k]$$

un vector renglón el cual identifica las coordenadas para las cuales se va a formular la predicción. Entonces la respuesta promedio estimada es

$$\begin{aligned} \hat{Y}_p &= \mathbf{X}_p'\mathbf{B} \\ &= B_0 + B_1x_1 + B_2x_2 + \cdots + B_kx_k. \end{aligned} \tag{14.10}$$

Dada (14.10), puede demostrarse que

$$Var(\hat{Y}_p) = \sigma^2\mathbf{X}_p'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_p.$$

De esta forma, una estimación de $Var(\hat{Y}_p)$ es

$$s^2(\hat{Y}_p) = s^2\mathbf{X}_p'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_p, \tag{14.11}$$

donde s^2 es la varianza residual y $\mathbf{X}$ es la matriz original de valores x, los cuales dieron origen a la ecuación de regresión estimada. De acuerdo con el caso de la teoría normal, un intervalo de confianza del $100(1 - \alpha)\%$ para la respuesta promedio en $x_1, x_2, \ldots, x_k$, es

$$\hat{y}_p \pm t_{1-\alpha/2,\ n-m}s(\hat{Y}_p). \tag{14.12}$$

Si se desea estimar una respuesta particular para $x_1, x_2, \ldots, x_k$, la predicción estará dada por (14.10), pero la varianza será

$$Var(\hat{Y}_{\text{part}}) = \sigma^2[1 + \mathbf{X}_p'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_p].$$

Por lo tanto, un intervalo de predicción del $100(1 - \alpha)\%$ para el valor real de la respuesta en $x_1, x_2, \ldots, x_k$, es

$$\hat{y}_{\text{part}} \pm t_{1-\alpha/2,\ n-m}s(\hat{Y}_{\text{part}}), \tag{14.13}$$

donde

$$s^2(\hat{Y}_{\text{part}}) = s^2[1 + \mathbf{X}_p'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_p].$$

Ejemplo 14.1 N.H.* Prater desarrolló una ecuación de regresión para estimar la producción de gasolina como una función de las propiedades de destilación de cierto tipo de petróleo crudo. Se identificaron cuatro variables de predicción: la gravedad del petróleo crudo, °API (x_1); la presión de vapor del petróleo crudo, psi (x_2); el punto de 10% *ASTM* para el petróleo crudo, °F (x_3) y el punto final *ASTM* para la gasolina, °F (x_4). Los primeros dos miden la gravedad y la presión de vapor del petróleo crudo. El punto de 10% *ASTM* es la temperatura para la cual se ha evaporado cierta cantidad de líquido, y el punto final para la gasolina es la temperatura para la cual se ha evaporado todo el líquido. La variable respuesta fue la cantidad de gasolina producida expresada como un porcentaje respecto al total de petróleo crudo. El objetivo radicó en determinar una ecuación de regresión para la producción de gasolina como una función lineal de las propiedades de destilación de cierto tipo de petróleo crudo x_1, x_2, x_3 y el punto final deseado para la gasolina x_4. Los datos de laboratorio obtenidos por Prater se muestran en la tabla 14.2.

Se emplearán los datos que aparecen en la tabla 14.2 para ilustrar las técnicas que hasta este momento se han presentado para regresión lineal múltiple mediante el empleo del paquete *SAS*. Este problema también se considerará como perteneciente a un problema particular que puede encontrarse en la regresión lineal múltiple y que se conoce como *multicolinealidad*. Debe notarse que desde la publicación de los datos de Prater, en 1956, varios autores los han empleado con el propósito de ilustrar diferentes aspectos del modelo lineal general. Entre ellos, Daniel y Wood [2] desarrollaron una ecuación de regresión muy diferente a la dada por Prater.

Mediante el empleo de una opción de *SAS*, denominada *GLM*, se ajusta el modelo lineal

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon.$$

En la figura 14.1 se proporciona el listado de computadora. Nótese que en la parte inferior de éste se encuentran cinco columnas de información. La primera columna de la izquierda identifica a las variables de predicción en el modelo que incluyen al término constante. La segunda columna proporciona las estimaciones por mínimos cuadrados; en la tercera se encuentran los valores t de Student para probar la hipótesis nula de que el valor del parámetro es cero; la cuarta columna da la probabilidad (valor p) de observar un valor t de Student, al menos tan grande en magnitud, como el valor observado (ignorando su signo) y la quinta columna proporciona las desviaciones estándar (errores) para las estimaciones por mínimos cuadrados. De esta forma, la ecuación estimada de regresión (tomando en cuenta sólo dos cifras decimales) es

$$\hat{y} = -6.82 + 0.23x_1 + 0.55x_2 - 0.15x_3 + 0.15x_4.$$

En la parte superior de la figura se encuentra la tabla ANOVA con $gl(CMR) = 4$, $SCR = 3\,429.27$, $CMR = 857.32$, $gl(SCE) = 27$, $SCE = 134.80$, $ECM = 4.99$,

* N.H. Prater, *Estimate gasoline yields from crudes*, Petroleum Refiner **35** (1956), 236-238. La reproducción de la tabla se hizo con el permiso de Petroleum Refiner (posteriormente Hydrocarbon Processing), Mayo 1956.

TABLA 14.2 Datos de la muestra para el ejemplo 14.1

Observación	Y	x_1	x_2	x_3	x_4
1	6.9	38.4	6.1	220	235
2	14.4	40.3	4.8	231	307
3	7.4	40.0	6.1	217	212
4	8.5	31.8	0.2	316	365
5	8.0	40.8	3.5	210	218
6	2.8	41.3	1.8	267	235
7	5.0	38.1	1.2	274	285
8	12.2	50.8	8.6	190	205
9	10.0	32.2	5.2	236	267
10	15.2	38.4	6.1	220	300
11	26.8	40.3	4.8	231	367
12	14.0	32.2	2.4	284	351
13	14.7	31.8	0.2	316	379
14	6.4	41.3	1.8	267	275
15	17.6	38.1	1.2	274	365
16	22.3	50.8	8.6	190	275
17	24.8	32.2	5.2	236	360
18	26.0	38.4	6.1	220	365
19	34.9	40.3	4.8	231	395
20	18.2	40.0	6.1	217	272
21	23.2	32.2	2.4	284	424
22	18.0	31.8	0.2	316	428
23	13.1	40.8	3.5	210	273
24	16.1	41.3	1.8	267	358
25	32.1	38.1	1.2	274	444
26	34.7	50.8	8.6	190	345
27	31.7	32.2	5.2	236	402
28	33.6	38.4	6.1	220	410
29	30.4	40.0	6.1	217	340
30	26.6	40.8	3.5	210	347
31	27.8	41.3	1.8	267	416
32	45.7	50.8	8.6	190	407

$gl\,(STC) = 31$ y $STC = 3\,564.08$. El valor F calculado para probar la hipótesis nula

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

es de 171.71, y la probabilidad de observar un valor mayor se encuentra inmediatamente a la derecha de éste. Debajo del valor p está la desviación estándar residual, $s = 2.23$. El coeficiente de correlación múltiple es 0.9622, lo cual significa que alrededor de un 96% de la variación total de las observaciones con respecto a su media puede explicarse por las cuatro variables de predicción incluidos en la ecuación de regresión.

En el extremo superior derecho, está el coeficiente de variación, el cual se definió en el capítulo 3. En este caso, el valor de CV es el cociente de la desviación estándar residual entre la media de las observaciones. Ya que en este caso $s = 2.23$ y $\bar{y} =$

VARIABLE DEPENDIENTE: Y

FUENTE	DF	SUMA DE CUADRADO	CUADRADO MEDIO	VALOR F	PR > F	R-CUADRADA	C.V.
MODELO	4	3429.27322460	857.31830615	171.71	0.0001	0.962177	11.3658
ERROR	27	134.80396290	4.99273937		STD DEV		Y MEDIA
TOTAL CORREGIDO	31	3564.07718750			2.23444386		19.65937500

FUENTE	DF	SC TIPO I	VALOR F	PR > F	DF	SC TIPO IV	VALOR F	PR > F
X1	1	216.25576661	43.31	0.0001	1	25.81557060	5.17	0.0311
X2	1	309.85082754	62.06	0.0001	1	11.19716648	2.24	0.1458
X3	1	29.21431690	5.85	0.0226	1	130.67556799	26.17	0.0001
X4	1	2873.95231355	575.63	0.0001	1	2873.95231355	575.63	0.0001

PARAMETRO	ESTIMACION	T PARA HO: PARAMETRO = 0	PR > |T|	DEST ERROR DE LA ESTIMACION
INTERSECCION	-6.82077407	-0.67	0.5062	10.12315182
X1	0.22724595	2.27	0.0311	0.09993664
X2	0.55372621	1.50	0.1458	0.36975194
X3	-0.14953562	-5.12	0.0001	0.02922920
X4	0.15465009	23.99	0.0001	0.00644584

FIGURA 14.1 Listado de computadora para la regresión lineal de Y sobre x_1, x_2, x_3 y x_4 para los datos de Prater

19.66, $CV = 11.37\%$. En el análisis de regresión es deseable que la desviación estándar residual sea una pequeña fracción de la media de las observaciones, ya que lo anterior, en general, implica que gran parte de la variación en la respuesta se explica mediante las variables de predicción en la ecuación de regresión. En la siguiente sección se darán más explicaciones con respecto a la información que se encuentra en la parte media de la figura.

Con base en el análisis anterior, existe una pequeña duda de que la regresión entre la producción de gasolina y las cuatro variables de predicción sea estadísticamente significativa. Debido a que se rechaza la hipótesis nula de que todos los coeficientes de regresión (excepto el término constante) son iguales a cero y el valor del coeficiente de correlación múltiple es relativamente alto al 0.9622. Sin embargo, existe una razón para preocuparse con respecto a la utilidad de la ecuación de regresión dada. Por ejemplo, las desviaciones estándar de los estimadores de mínimos cuadrados para β_0 y β_2 son grandes, lo que sugiere que x_2 y posiblemente otras variables de predicción, puedan no tener un gran efecto sobre la producción de gasolina. En las siguientes secciones se examinarán los procedimientos adecuados para obtener la mejor ecuación de regresión para un conjunto dado de variables de predicción. Los datos del ejemplo 14.1 se utilizarán de vez en cuando para otros ejemplos en este capítulo.

14.3 Principio de la suma de cuadrados extra

La inclusión de una variable de predicción en un modelo de regresión no implica, en forma necesaria, que tenga un efecto substancial sobre la respuesta dada; es decir, cuando un investigador identifica un conjunto de variables de predicción, esto indica el *potencial* de las variables para explicar la variación en la respuesta. Queda por comprobarse si algunas realmente lo hacen.

El procedimiento apropiado para encontrar los efectos individuales de las variables de predicción se basa en el *principio de la suma de cuadrados extra.* Este principio permite determinar la reducción en la suma de los cuadrados de los errores cuando se introduce un coeficiente adicional de regresión para alguna función de una variable de predicción en la ecuación de regresión. Cabe recordar dos cosas importantes: 1) la suma total de cuadrados sigue siendo la misma sin importar el número de términos que se introduzcan en el modelo de regresión. 2) La suma de los cuadrados de los errores siempre disminuye (cuando menos un poco) conforme se añaden más términos al modelo.

Dado que la suma de los cuadrados de regresión es la diferencia entre *STC* y *SCE*, el incremento en *SCR* tiene un límite conforme se suman más términos al modelo. Una estrategia lógica en la regresión lineal múltiple es la de añadir no cualesquiera términos, al modelo, sino sólo aquéllos que incrementen en forma significativa la suma de los cuadrados de regresión y de esta manera disminuyan significativamente la suma de los cuadrados de los errores. Como ejemplo, en el modelo lineal simple, *SCR* es la suma extra de los cuadrados debida a la inclusión del término lineal $\beta_1 x$ en el modelo. En otras palabras, *SCR* representa la reducción en la suma de los cuadrados de

los errores cuando se añade un efecto lineal de la variable de predicción al modelo original.

$$Y_i = \beta_0 + \varepsilon_i.$$

Para ilustrar el principio de la suma de cuadrados extra, se emplearán, de los datos de Prater como variables de predicción potenciales, sólo a x_2 y x_3 y se ajustarán todas las posibles regresiones de la producción de gasolina para esas dos variables. Existen tres ecuaciones de regresión; dos que toman en cuenta a x_2 y x_3 en forma individual y la tercera que contiene a ambas variables x_2 y x_3. En la tabla 14.3 se proporcionan las ecuaciones de regresión estimadas y sus correspondientes tablas de análisis de varianza. Nótese que se ha empleado la notación $SCR(x_2)$, $SCR(x_2, x_3)$ y $SCE(x_2, x_3)$, para denotar que estas sumas de cuadrados son funciones de las variables de predicción ya indicadas en la ecuación de regresión y de los correspondientes coeficientes de mínimos cuadrados.

A continuación se examinarán los resultados que se encuentran en la tabla 14.3. Como ya se ha mencionado, para las 32 observaciones dadas de la respuesta, la

TABLA 14.3 Ecuaciones estimadas de regresión y tablas ANOVA para la producción de gasolina, tomando en cuenta a x_2 y/o x_3

a) $\hat{y} = 13.09 + 1.57x_2$

Fuente de variación	*gl*	*SC*	*CM*	*Valor F*
Regresión	1	$SCR(x_2) = 525.74$	$CMR(x_2) = 525.74$	5.19
Error	30	$SCE(x_2) = 3038.34$	$CME(x_2) = 101.28$	
Total	31	$STC = 3564.08$	$f_{0.95,\,1,\,30} = 4.17$	

b) $\hat{y} = 41.39 - 0.09x_3$

Fuente de variación	*gl*	*SC*	*CM*	*Valor F*
Regresión	1	$SCR(x_3) = 353.70$	$CMR(x_3) = 353.70$	3.31
Error	30	$SCE(x_3) = 3210.38$	$CME(x_3) = 107.01$	
Total	31	$STC = 3564.08$	$f_{0.95,\,1,\,30} = 4.17$	

c) $\hat{y} = -2.52 + 2.26x_2 + 0.05x_3$

Fuente de variación	*gl*	*SC*	*CM*	*Valor F*
Regresión	2	$SCR(x_2, x_3) = 547.49$	$CMR(x_2, x_3) = 273.74$	2.63
Error	29	$SCE(x_2, x_3) = 3016.59$	$CME(x_2, x_3) = 104.02$	
Total	31	$STC = 3564.08$	$f_{0.95,\,2,\,29} = 3.33$	

suma total de cuadrados es $STC = 3\ 564.08$ sin importar cuántas variables de predicción se incluyan en el modelo. Para la regresión de Y sobre x_2, $SCR(x_2) = 525.74$ es la reducción en la suma de los cuadrados de los errores cuando se añade el término $\beta_2 x_2$ al modelo $Y_i = \beta_0 + \varepsilon_i$. En otras palabras, si se ajusta el modelo $Y_i = \beta_0 + \varepsilon_i$, se supone que la única fuente de variación en Y_i es el error aleatorio; la recta de regresión estimada es simplemente $\hat{Y}_i = \bar{Y}$. Cuando se agrega el término $\beta_2 x_2$ al modelo, entonces parte de la variación total puede explicarse por la presencia de x_2. Esto es lo que precisamente representa $SCR(x_2) = 525.74$, $SCR(x_2)$ es la suma extra de cuadrados en la que disminuye SCE cuando se añade el término $\beta_2 x_2$ al modelo. Al emplear el mismo argumento para la regresión de Y sobre x_3, $SCR(x_3) = 353.70$ es la suma extra de cuadrados en los que disminuye el error cuando se añade el término $\beta_3 x_3$ al modelo $Y_i = \beta_0 + \varepsilon_i$. Para cualquier otro caso, si la reducción en la suma de los cuadrados de los errores es substancial, se rechaza la hipótesis nula de valor cero para el correspondiente coeficiente de regresión. Nótese que se rechaza x_2, H_0: $\beta_2 = 0$ para la regresión de Y sobre x_2 (valor $f = 5.19 > f_{0.95,1,30} = 4.17$), pero para la regresión de Y sobre x_3, H_0: $\beta_3 = 0$ no puede rechazarse.

Considérese la regresión de Y sobre x_2 y x_3. Lo que se desea determinar es la reducción en la suma de los cuadrados de los errores cuando se añade el término $\beta_3 x_3$ al modelo, el cual ya contiene el término constante β_0 y el término $\beta_2 x_2$, o la reducción en SCE cuando se introduce el término $\beta_2 x_2$ al modelo, el cual ya contiene a β_0 y $\beta_3 x_3$. Nótese que para el modelo c de la tabla 14.3 la suma de los cuadrados de los errores cuando se incluye en el modelo de regresión, tanto a x_2 como a x_3 es $SCE(x_2, x_3) = 3\ 016.59$. Pero cuando sólo se tiene a x_2 en el modelo, $SCE(x_2) = 3\ 038.34$. Por lo tanto, la diferencia entre $SCE(x_2)$ y $SCE(x_2, x_3)$ debe ser la suma de cuadrados extra debida a la inclusión del término $\beta_3 x_3$ en el modelo que ya contiene a los términos β_0 y $\beta_2 x_2$. Se denotará esta diferencia por $\text{SCR}(x_3 \mid x_2)$. De esta forma

$$\begin{aligned} \text{SCR}(x_3 \mid x_2) &= \text{SCE}(x_2) - \text{SCE}(x_2, x_3) \\ &= 3038.34 - 3016.59 \\ &= 21.75 \end{aligned} \tag{14.14}$$

es la reducción adicional en la suma de los cuadrados de los errores cuando se introduce x_3 en el modelo que ya contiene a x_2.

Dado que una reducción en la suma de los cuadrados de los errores significa un aumento correspondiente a la suma de los cuadrados de la regresión,

$$\begin{aligned} \text{SCR}(x_2, x_3) &= \text{SCR}(x_3 \mid x_2) + \text{SCR}(x_2) \\ &= 21.75 + 525.74 \\ &= 547.49. \end{aligned} \tag{14.15}$$

La suma de los cuadrados de la regresión, cuando figuran en el modelo, tanto x_2 como x_3, se separa en dos componentes, cada uno de éstos con un grado de libertad. $SCR(x_3 \mid x_2)$, el cual refleja la contribución de x_3 cuando ésta se añade al modelo $Y = \beta_0 + \beta_2 x_2 + \varepsilon$, y $SCR(x_2)$ la cual mide la contribución de x_2 cuando ésta se añade al modelo $Y = \beta_0 + \varepsilon$.

TABLA 14.4 Tabla ANOVA aumentada para la regresión de Y sobre x_2 y x_3

Fuente de variación	*gl*	*SC*	*CM*	*Valor F*
Regresión	2	SCR $(x_2, x_3) = 547.49$	CMR $(x_2, x_3) = 273.74$	2.63
x_2	1	SCR $(x_2) = 525.74$	CMR $(x_2) = 525.74$	5.05
$x_3 \mid x_2$	1	SCR$(x_3 \mid x_2) = 21.75$	CMR$(x_3 \mid x_2) = 21.75$	0.2
Error	29	SCE $(x_2, x_3) = 3016.59$	CME $(x_2, x_3) = 104.02$	
Total	31	STC $= 3564.08$	$f_{0.95,\,2,\,29} = 3.33$; $f_{0.95,\,1,\,29} = 4.18$	

Se puede demostrar que $SCR(x_3 \mid x_2)$ y $SCR(x_2)$ son variables aleatorias independientes chi-cuadrada, cada una con un grado de libertad; entonces puede hacerse una comparación entre el cuadrado medio correspondiente a $SCR(x_3 \mid x_2)$, o el de $SCR(x_2)$, y el cuadrado medio del error, $CME(x_2, x_3)$ por medio de la estadística F. Esta prueba se conoce como *prueba F parcial* sobre una variable de predicción. En realidad, la *prueba F parcial* determina si la contribución de un coeficiente de regresión es lo suficientemente grande para garantizar su inclusión en el modelo, dado que otros términos no toman en cuenta al coeficiente que ya se encuentra en el mismo. Por lo tanto, en cierto sentido, se intenta enjuiciar el efecto individual de la correspondiente variable de predicción sobre una respuesta dada. La tabla ANOVA aumentada para la regresión de Y sobre x_2 y x_3, la cual incluye las pruebas F parciales, se muestra en la tabla 14.4. Nótese que la inclusión del término $\beta_2 x_2$ en el modelo $Y = \beta_0 + \varepsilon$ tiene un efecto benéfico, mientras que la inclusión de $\beta_3 x_3$ en $Y = \beta_0 + \beta_2 X_2 + \varepsilon$ no.

A lo largo de toda la presentación anterior se supuso que el término $\beta_3 x_3$ era el último en sumarse al modelo que incluye a x_2 y x_3. Sin embargo, es posible realizar pruebas parciales F para cada coeficiente de regresión, como si la correspondiente variable de predicción fuese la última en haberse añadido al modelo. De esta forma, los efectos individuales de cada variable de predicción, en presencia de las otras, pueden comprobarse. Para el ejemplo, lo que se desea es determinar la contribución del término $\beta_2 x_2$ cuando el modelo ya contiene a β_0 y a $\beta_3 x_3$.

Al seguir el mismo procedimiento dado con anterioridad, la suma de los cuadrados de los errores cuando, tanto x_2 como x_3 se encuentran en el modelo, es $SCE(x_2, x_3) = 3\ 016.59$. Pero cuando sólo se encuentra x_3 en el modelo, $SCE(x_3) = 3\ 210.38$. De esta forma, la reducción en el valor de la suma de los cuadrados de los errores cuando se añade el término $\beta_2 x_2$ al modelo que ya contiene a β_0 y $\beta_3 x_3$ es

$$\begin{aligned} \text{SCR}(x_2 \mid x_3) &= \text{SCE}(x_3) - \text{SCE}(x_2, x_3) \\ &= 3210.38 - 3016.59 \\ &= 193.79. \end{aligned} \tag{14.16}$$

Entonces la suma de los cuadrados de regresión, cuando x_2 y x_3 se encuentran en el

modelo, la suma de los dos componentes es

$$\begin{aligned} \text{SCR}(x_2, x_3) &= \text{SCR}(x_2 \mid x_3) + \text{SCR}(x_3) \\ &= 193.79 + 353.70 \\ &= 547.49, \end{aligned} \tag{14.17}$$

cada componente con un grado de libertad. Una consecuencia importante de todo lo anterior, es que tanto $SCR(x_2 \mid x_3) = 193.79$ y $SCR(x_3 \mid x_2) = 21.75$ son más pequeños que $SCR(x_2) = 525.74$ y $SCR(x_3) = 353.70$, respectivamente. El porqué de lo anterior constituye el tema de la siguiente sección.

Para determinar las pruebas F parciales para la regresión debida a x_3, o a x_2 dada x_3, ahora es posible tener otra versión de la tabla 14.4; ésta se muestra en la tabla 14.5. Nótese que una comparación entre los resultados de las tablas 14.4 y 14.5 muestra un desacuerdo con respecto al efecto de x_2 sobre la producción de gasolina. Mientras que la regresión lineal simple de Y sobre x_2 es estadísticamente significativa ($f = 5.19$), la regresión de Y sobre x_2 dada la presencia de x_3, no lo es ($f = 1.86$). Se dará más información con respecto a esta ocurrencia en la siguiente sección.

El principio de la suma de cuadrados extra se extiende de manera directa para aplicar la idea básica a cualquier número de variables de predicción. Por ejemplo, supóngase que se tienen tres variables de predicción x_1, x_2 y x_3. Se puede definir la reducción en la suma de los cuadrados de los errores, cuando una de éstas se añade al modelo que ya contiene a las otras dos, de la siguiente manera:

$$\text{SCR}(x_3 \mid x_1, x_2) = \text{SCE}(x_1, x_2) - \text{SCE}(x_1, x_2, x_3), \tag{14.18}$$

$$\text{SCR}(x_2 \mid x_1, x_3) = \text{SCE}(x_1, x_3) - \text{SCE}(x_1, x_2, x_3), \tag{14.19}$$

$$\text{SCR}(x_1 \mid x_2, x_3) = \text{SCE}(x_2, x_3) - \text{SCE}(x_1, x_2, x_3). \tag{14.20}$$

Para desarrollar expresiones similares a (14.15) o (14.17), de (14.14) se deduce que

$$\text{SCR}(x_2 \mid x_1) = \text{SCE}(x_1) - \text{SCE}(x_1, x_2),$$

TABLA 14.5 Tabla ANOVA aumentada para la regresión de Y sobre x_2 y x_3

Fuente de variación	*gl*	*SC*	*CM*	*Valor F*
Regresión	2	$\text{SCR}(x_2, x_3) = 547.49$	$\text{CMR}(x_2, x_3) = 273.74$	2.63
x_3	1	$\text{SCR}(x_3) = 353.70$	$\text{CMR}(x_3) = 353.70$	3.40
$x_2 \mid x_3$	1	$\text{SCR}(x_2 \mid x_3) = 193.79$	$\text{CMR}(x_2 \mid x_3) = 193.79$	1.86
Error	29	$\text{SCE}(x_2, x_3) = 3016.59$	$\text{CME}(x_2, x_3) = 104.02$	
Total	31	$\text{STC} = 3564.08$	$f_{0.95,\,2,\,29} = 3.33$; $f_{0.95,\,1,\,29} = 4.18$	

o

$$\text{SCE}(x_1, x_2) = \text{SCE}(x_1) - \text{SCR}(x_2 \mid x_1). \tag{14.21}$$

Ahora, cuando sólo se tiene a x_1 en el modelo, por definición

$$\text{SCE}(x_1) = \text{STC} - \text{SCR}(x_1);$$

pero cuando todas las tres variables se encuentran en el modelo,

$$\text{STC} = \text{SCR}(x_1, x_2, x_3) + \text{SCE}(x_1, x_2, x_3).$$

Entonces

$$\text{SCE}(x_1) = \text{SCR}(x_1, x_2, x_3) + \text{SCE}(x_1, x_2, x_3) - \text{SCR}(x_1),$$

y al sustituir $SCE(x_1)$ en (14.21), se obtiene

$$\text{SCE}(x_1, x_2) = \text{SCR}(x_1, x_2, x_3) + \text{SCE}(x_1, x_2, x_3) - \text{SCR}(x_1) - \text{SCR}(x_2 \mid x_1). \tag{14.22}$$

Al sustituir (14.22) por $SCE(x_1, x_2)$ en (14.18) se obtiene el resultado deseado

$$\text{SCR}(x_3 \mid x_1, x_2) = \text{SCR}(x_1, x_2, x_3) - \text{SCR}(x_1) - \text{SCR}(x_2 \mid x_1), \tag{14.23}$$

o

$$\text{SCR}(x_1, x_2, x_3) = \text{SCR}(x_1) + \text{SCR}(x_2 \mid x_1) + \text{SCR}(x_3 \mid x_1, x_2). \tag{14.24}$$

La suma de los cuadrados de regresión, cuando las tres variables se encuentran en el modelo, tiene tres componentes, cada uno con un grado de libertad. $SCR(x_1)$ mide la contribución (reducción en la suma de los cuadrados de los errores) de x_1 cuando se añade x_1 al modelo $Y = \beta_0 + \varepsilon$; $SCR(x_2 \mid x_1)$ representa la contribución de x_2 cuando ésta se introduce al modelo $Y = \beta_0 + \beta_1 x_1 + \varepsilon$; y $\text{SCR}(x_3 \mid x_1, x_2)$ es la contribucion de x_3 cuando ésta se agrega al modelo $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

Al emplear (14.19) o (14.20) y si se sigue el mismo procedimiento, pueden establecerse resultados similares a (14.24) de la siguiente manera:

$$\text{SCR}(x_1, x_2, x_3) = \text{SCR}(x_1) + \text{SCR}(x_3 \mid x_1) + \text{SCR}(x_2 \mid x_1, x_3), \tag{14.25}$$

$$\text{SCR}(x_1, x_2, x_3) = \text{SCR}(x_2) + \text{SCR}(x_3 \mid x_2) + \text{SCR}(x_1 \mid x_2, x_3). \tag{14.26}$$

Estos resultados permiten que se lleven a cabo pruebas F parciales sobre cada coeficiente de regresión como si la variable de predicción asociada con éste hubiese sido la última en incluirse en el modelo. En otras palabras, con las pruebas parciales F puede determinarse si el efecto individual de una variable de predicción en presencia de las demás es estadísticamente discernible. Debe notarse que al intercambiar el orden de entrada al modelo para las variables de predicción, entonces es posible identificar otras relaciones similares a (14.24)-(14.26). Por ejemplo,

$$\text{SCR}(x_1, x_2, x_3) = \text{SCR}(x_2) + \text{SCR}(x_1 \mid x_2) + \text{SCR}(x_3 \mid x_2, x_1)$$

es otra separación de $SCR(x_1, x_2, x_3)$. Conforme crece el número de variables de predicción, el número posible de separaciones se vuelve más grande.

Con base en lo anterior, puede explicarse ahora, para los datos de Prater dados en la sección anterior, la información que aparece en la parte media de la figura 14.1. El lector notará dos columnas identificadas como "*SC* tipo I" y "*SC* tipo IV". La de tipo I contiene las cuatro componentes de $SCR(x_1, x_2, x_3, x_4)$, tales que

$$\begin{aligned}\text{SCR}(x_1, x_2, x_3, x_4) = \text{SCR}(x_1) + \text{SCR}(x_2 \mid x_1) \\ + \text{SCR}(x_3 \mid x_1, x_2) + \text{SCR}(x_4 \mid x_1, x_2, x_3).\end{aligned}$$

Cada componente tiene un grado de libertad y representa la reducción en la suma de los cuadrados de los errores cuando se añade al modelo la variable indentificada. El orden de entrada de variables al modelo es el mismo para el cual fueron identificadas las variables de predicción por el usuario, así que

$$\begin{aligned}\text{SCR}(x_1) &= 216.26, \\ \text{SCR}(x_2 \mid x_1) &= 309.85, \\ \text{SCR}(x_3 \mid x_1, x_2) &= 29.21, \\ \text{SCR}(x_4 \mid x_1, x_2, x_3) &= 2873.95.\end{aligned}$$

Las dos columnas que se encuentran inmediatamente a la derecha de la columna que corresponde a "*SC* tipo I", dan los valores de las pruebas *F* parciales y los valores correspondientes *p* para cada una de las cuatro componentes. A partir de esta información, es claro que el efecto individual de cada coeficiente de regresión en presencia de otros términos en el modelo es estadísticamente apreciable.

La *SC* tipo IV representa la reducción en la suma de los cuadrados de los errores debida a la edición, en el modelo, de la variable de predicción correspondiente, dado que las otras tres ya se encuentran en el mismo. Para el ejemplo, las componentes son

$$\begin{aligned}\text{SCR}(x_1 \mid x_2, x_3, x_4) &= 25.82, \\ \text{SCR}(x_2 \mid x_1, x_3, x_4) &= 11.20, \\ \text{SCR}(x_3 \mid x_1, x_2, x_4) &= 130.68, \\ \text{SCR}(x_4 \mid x_1, x_2, x_3) &= 2873.95.\end{aligned}$$

Nótese que no existe ninguna razón teórica para que la suma de estas cuatro componentes sea igual a $SCR(x_1, x_2, x_3, x_4)$.

Con base en los valores de las pruebas *F* parciales para estas componentes, es clara la existencia de cierta discrepancia entre estos resultados y los que se tienen para *SC* tipo I. Por ejemplo, la contribución de x_2 en presencia sólo de x_1, es estadísticamente discernible, pero no puede decirse lo mismo de la contribución de x_2 en presencia de x_1, x_3 y x_4.

14.4 El problema de la multicolinealidad

Es muy común obtener conclusiones equivocadas con un punto de vista casual para la aplicación de análisis de regresión, cuando no se tiene una completa apreciación de los problemas que pueden encontrarse. En la sección anterior se notaron varias de las discrepancias que pueden presentarse en la regresión lineal múltiple. Éstas proporcionan información valiosa para identificar problemas que necesitan una atencion adicional. El enfoque para el análisis de regresión no debe ser simplemente maximizar el coeficiente de correlación múltiple sin tomar en cuenta la debida consideración de los coeficientes de regresión estimados y sus desviaciones estándar, o la de comprobar las suposiciones fundamentales del análisis de regresión.

Un problema frecuente en regresión lineal múltiple es el que algunas de la variables de predicción están correlacionadas. Si la correlación es pequeña, las consecuencias serán de índole menor. Sin embargo, si existe una correlación muy fuerte entre dos o más variables de predicción, los resultados de la regresión serán ambiguos, especialmente con respecto a los valores de los coeficientes de regresión estimados. Un coeficiente de correlación muy alto entre dos o más variables de predicción constituye lo que se conoce como *multicolinealidad.* Este problema muchas veces es difícil de detectar ya que surge como consecuencia de datos deficientes. Éste es el precio que se paga cuando no es posible diseñar los experimentos en forma estadística y recabar los datos en arreglos balanceados, tal como se analizó en el capítulo 12.

Recuérdese que la ecuación de predicción, a pesar de que no es precisa en un sentido físico, debe ser un medio, empírico, viable para predecir la respuesta promedio dada una condición de las variables de predicción. La multicolinealidad no impide tener un buen ajuste ni evita que la respuesta sea, en forma adecuada, predicha dentro del intervalo de las observaciones; lo que sucede es que ésta afecta en forma severa las estimaciones de mínimos cuadrados, ya que bajo los efectos de la multicolinealidad éstas tienden a ser menos precisas para los efectos individuales de las variables de predicción, es decir, cuando dos o más variables de predicción son colineales los coeficientes de regresión estimados no miden los efectos individuales sobre la respuesta, sino que reflejan un efecto parcial sobre la misma, sujeto a todo lo que pase con las demás variables de predicción en la ecuación de regresión.

Para apreciar la naturaleza de la multicolinealidad, primero se estudiará una situación en la que ésta no existe. Considérese un modelo de regresión con dos variables de predicción. Si el coeficiente de correlación simple entre las dos variables es cero, entonces se dice que las variables son *ortogonales.** Al tener variables de predicción ortogonales el efecto que una de éstas tiene sobre la respuesta dada se mide en forma totalmente independiente del efecto individual que la otra variable tiene sobre la misma respuesta. Si una o ambas variables de predicción se encuentran en la ecuación de regresión, las estimaciones de mínimos cuadrados no cambiarán su valor

* Una de las principales razones para diseñar experimentos en forma estadística es la de adquirir factores o variables que sean ortogonales. Para muchos de los experimentos que emplean el análisis de varianza, los factores son ortogonales.

TABLA 14.6 Datos de la muestra para el ejemplo 14.2

Y(°F)	x_1(°F)	x_2(%)
66	70	20
72	75	20
77	80	20
67	70	30
73	75	30
78	80	30
68	70	40
74	75	40
79	80	40

Fuente: Servicio Climatológico Nacional.

Ejemplo 14.2 Para ilustrar los efectos ortogonales se examinarán los datos (limitados) que aparecen en la tabla 14.6 que consisten en la temperatura aparente Y (qué tan caliente se siente) como una función de la temperatura del aire x_1 y de la humedad relativa x_2.

El lector no tendrá ningún problema pra verificar que el coeficiente de correlación entre x_1 y x_2 tiene un valor de cero. Se procederá a ajustar los modelos $Y = \beta_0 + \beta_1 x_1 + \varepsilon$, $Y = \beta_0 + \beta_2 x_2 + \varepsilon$, y $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. La información relevante se encuentra en la tabla 14.7.

Nótese que los coeficientes de regresión estimados para x_1 y x_2 son 1.10 y 0.10, respectivamente, sin importar que una o ambas variables de predicción se encuentren en la ecuación de regresión. De esta forma, por cada grado que aumenta la temperatura del aire, la temperatura aparente aumenta en 1.10 grados, y por cada incremento en porcentaje de la humedad relativa, la temperatura aparente aumenta 0.10 grados.* Además, nótese que

$$\text{SCR}(x_2 \mid x_1) = \text{SCR}(x_2),$$

$$\text{SCR}(x_1, x_2) = \text{SCR}(x_1) + \text{SCR}(x_2).$$

Los resultados anteriores son los que se esperan cuando las variables de predicción son ortogonales y no existe multicolinealidad.

Si se consideran de nuevo los datos de Prater y las regresiones que incluyen a x_2 o x_3, dadas en la tabla 14.3, se mostrará que existe una razón para sospechar la existencia de multicolinealidad entre x_2 y x_3. Primero, nótese que el coeficiente de regresión estimado para x_2 es 1.57 cuando sólo se encuentra presente en la ecuación de regresión x_2, pero su valor es de 2.26 cuando se añade x_3. De manera similar, el coeficiente de x_3 es -0.09 para el modelo de línea recta, pero éste cambia tanto en signo como en valor para ser igual a 0.05 cuando también se incluye a x_2 en la ecuación de regresión. Segundo, es claro que la reducción en el valor de la suma de los cuadrados

* El lector no debe generalizar de estos resultados por lo limitado de los datos.

TABLA 14.7 Ecuaciones de regresión estimadas y tablas ANOVA para la temperatura aparente, tomando a x_1 y/o x_2.

$$\hat{y} = -9.83 + 1.10x_1$$

Fuente de variación	*gl*	*SC*	*CM*	*Valor F*
Regresión	1	$SCR(x_1) = 181.5$	$CMR(x_1) = 181.5$	195.46
Error	7	$SCE(x_1) = 6.5$	$CME(x_1) = 0.9286$	
Total	8	$STC = 188.0$	$f_{0.95,1,7} = 5.59$	

$$\hat{y} = 69.67 + 0.10x_2$$

Fuente de variación	*gl*	*SC*	*CM*	*Valor F*
Regresión	1	$SCR(x_2) = 6.0$	$CMR(x_2) = 6.0$	0.23
Error	7	$SCE(x_2) = 182.0$	$CME(x_2) = 26.0$	
Total	8	$STC = 188.0$	$f_{0.95,1,7} = 5.59$	

$$\hat{y} = 12.83 + 1.10x_1 + 0.10x_2$$

Fuente de variación	*gl*	*SC*	*CM*	*Valor F*
Regresión	2	$SCR(x_1, x_2) = 187.5$	$CMR(x_1, x_2) = 93.75$	1125.0
x_1	1	$SCR(x_1) = 181.5$	$CMR(x_1) = 181.5$	2178.0
$x_2 \mid x_1$	1	$SCR(x_2 \mid x_1) = 6.0$	$CMR(x_2 \mid x_1) = 6.0$	72.0
Error	6	$SSE(x_1, x_2) = 0.5$	$CME(x_1, x_2) = 0.0833$	
Total	8	$STC = 188.0$	$f_{0.95,2,6} = 5.14$, $f_{0.95,1,6} = 5.99$	

de los errores debida a x_3 cuando x_2 se encuentra en el modelo, $SCR(x_3 \mid x_2) = 21.75$ es mucho menor que cuando sólo se encuentra x_3 en el modelo, $SCR(x_3) = 353.70$. La fuerte correlación que en forma aparente existe entre x_2 y x_3 ha disminuido de manera drástica el efecto individual que sobre la respuesta tiene x_3 en presencia de x_2. Puede hacerse el mismo comentario con respecto al efecto de x_2, ya que éste es estadísticamente apreciable en ausencia de x_3 ($SCR(x_2) = 525.74$, $f = 5.19$), pero se encuentra sustancialmente disminuido cuando x_3 se encuentra presente ($SCR(x_2 \mid x_3) = 193.79$).

Para mostrar que existe una fuerte correlación entre x_2 y x_3, se determinará la matriz de correlación para las cuatro variables de predicción de los datos de Prater. Esta matriz contiene todos los pares posibles de coeficientes de correlación y puede

determinarse para un conjunto dado de variables en forma muy fácil mediante el empleo de un paquete para computadora.* La matriz de correlación para x_1, x_2, x_3, y x_4 es la siguiente:

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \begin{bmatrix} 1.00 & 0.62 & -0.70 & -0.32 \\ 0.62 & 1.00 & -0.91 & -0.30 \\ -0.70 & -0.91 & 1.00 & 0.41 \\ -0.32 & -0.30 & 0.41 & 1.00 \end{bmatrix} \end{array}$$

Nótese que el valor de cada uno de los elementos que se encuentran en la diagonal es uno, ya que cada variable se encuentra correlacionada de manera perfecta consigo misma. Los elementos que se encuentran fuera de la diagonal son los valores de los coeficientes de correlación simple. Por ejemplo, $r_{12} = 0.62$ es el coficiente de correlación entre x_1 y x_2; por lo tanto, el valor $r_{23} = -0.91$ al encontrarse muy cercano a -1 sugiere una fuerte asociación lineal entre x_2 y x_3. Este resultado es predecible si se inspeccionan en forma visual los datos dados en el ejemplo 14.1. Nótese que conforme aumenta la presión de vapor del petróleo crudo x_2, el punto x_3 *ASTM* 10% disminuye y viceversa. Estos resultados proporcionan la causa para sospechar la presencia de multicolinealidad en este ejemplo.

¿Qué es lo que se puede hacer cuando se descubre la presencia de multicolinealidad? Una alternativa es la de añadir puntos de observación para las variables colineales, los cuales tiendan a disminuir la severidad de la correlación. Pero puede ocurrir que estos puntos de observación no se encuentren disponibles fácilmente. Por ejemplo, para los datos de la gasolina podrían no existir los tipos de petróleo crudo que pueden disminuir la fuerte linealidad que existe entre x_2 y x_3. Una segunda alternativa es la de omitir una o más de las variables que son colineales, lo que reduce la variabilidad de los coeficientes de regresión de las restantes variables. Se han desarrollado enfoques más sofisticados para resolver los problemas que plantea la multicolinealidad, incluyendo la regresión por componentes principales y la regresión ridge. Estos temas se encuentran más allá del objetivo de este libro; se invita al lector a que consulte las referencias [1] y [3].

Para ilustrar la segunda alternativa y resolver el problema de la multicolinealidad, se examinarán las regresiones para las cuales se omiten x_2 o x_3. Como comparación, también se considerará la regresión de la producción de gasolina con respecto sólo al punto (x_3) *ASTM* 10% y al punto final (x_4). Sin proporcionar argumentación adicional se piensa que estas tres regresiones son las candidatas para la "mejor" ecuación de regresión lineal para los datos de Prater. La información más importante se encuentra en la tabla 14.8.

Al comparar parece que la regresión de Y sólo sobre x_3 y x_4 es la mejor con respecto a las proporcionadas por los otros dos modelos. Para el modelo *b*, la desviación estándar del estimador por mínimos cuadrados para el término constante es muy gran-

* Para *SAS* puede ser apropiado utilizar *PROC CORR*.

TABLA 14.8 Candidatos para la mejor ecuación de regresión para los datos de Prater

a) Regresión de Y sobre x_1, x_2, x_4

Variable	*Coeficiente de regresión estimado*	*Desviación estándar estimada*	*Valor t*
Constante	−53.899	5.8135	−9.27
x_1	0.422	0.1273	3.32
x_2	2.154	0.2716	7.93
x_4	0.144	0.0084	17.10

$R^2 = 0.9255$ $\quad t_{0.975,\,28} = 2.048$

ANOVA

Fuente	*gl*	*SC*	*CM*	*Valor F*
Regresión	3	3298.60	1099.53	115.97
x_1	1	216.26	216.26	22.81
$x_2 \mid x_1$	1	309.85	309.85	32.68
$x_4 \mid x_1, x_2$	1	2772.49	2772.49	292.41
Error	28	265.48	9.48	
Total	31	3564.08	$f_{0.95,\,3,\,28} = 2.95$; $f_{0.95,\,1,\,28} \doteq 4.20$	

b) Regresión de Y sobre x_1, x_3, x_4

Variable	*Coeficiente de regresión estimado*	*Desviación estándar estimada*	*Valor t*
Constante	4.032	7.2233	0.56
x_1	0.222	0.1021	2.17
x_3	−0.187	0.0159	−11.72
x_4	0.157	0.0065	24.22

$R^2 = 0.959$ $\quad t_{0.975,\,28} = 2.048$

ANOVA

Fuente	*gl*	*SC*	*CM*	*Valor F*
Regresión	3	3418.08	1139.38	218.51
x_1	1	216.26	216.26	41.47
$x_3 \mid x_1$	1	142.08	142.08	27.25
$x_4 \mid x_1, x_3$	1	3059.74	3059.74	586.79
Error	28	146.00	5.21	
Total	31	3564.08	$f_{0.95,\,3,\,28} = 2.95$; $f_{0.95,\,1,\,28} = 4.20$	

(continúa)

TABLA 14.8 *(continuación)*

c) Regresión de Y sobre x_3, x_4

Variable	*Coeficiente de regresión estimado*	*Desviación estándar estimada*	*Valor t*
Constante	18.468	3.0090	6.14
x_3	−0.209	0.0127	−16.43
x_4	0.156	0.0069	22.73

$R^2 = 0.9521$ $\quad t_{0.975,\,29} = 2.045$

ANOVA

Fuente	*gl*	*SC*	*CM*	*Valor F*
Regresión	2	3393.47	1696.73	228.41
x_3	1	353.70	353.70	60.12
$x_4 \mid x_3$	1	3039.77	3039.77	516.69
Error	29	170.61	5.88	
Total	31	3564.08	$f_{0.95,\,2,\,29} = 3.33$; $f_{0.95,\,1,\,29} = 4.18$	

de, y la desviación estándar del coeficiente x_1 es casi igual a la mitad del valor de éste. Para el modelo *a*, $R^2 = 0.9255$, mientras que para el modelo *c*, $R^2 = 0.9521$, el cual es un valor mucho más cercano al valor de R^2 cuando todas las variables de predicción figuran en la ecuación de regresión. Además, las desviaciones estándar de los coeficientes de regresión estimados son, en forma relativa, mejores para el modelo *c* que para el *b*. Para finalizar, los factores físicos claves como la consistencia lógica de los coeficientes de regresión estimados, son los que por lo general definen la elección final.

14.5 Determinación del mejor conjunto de variables de predicción

Un problema muy importante en el análisis de regresión es determinar cuáles de las variables de predicción en la lista inicial deberán incluirse en el modelo de regresión. En casi todas las ocasiones, un investigador decidirá, de una lista inicial de variables de predicción, a aquéllas que tienen la mayor probabilidad de contener los factores más importantes para la respuesta dada. Por lo tanto, es necesario tener una manera de determinar, de la lista inicial de variables de predicción, a aquéllas que parecen ser las mejores para describir el cambio en la respuesta promedio, y de esta forma proporcionarán una ecuación de predicción representativa de las condiciones bajo las cuales se recabaron los datos. La palabra "mejores" no debe interpretarse como poseedora de la connotación teórica de óptimo; ésta debe considerarse como representativa de los medios por los cuales se aíslan las características más sobresalientes, de tal manera que puede llevarse a cabo un análisis significativo.

Sea k el número inicial de potenciales variables de predicción; el número de términos en el modelo lineal completo, incluyendo al término constante, es $m = k + 1$. Un procedimiento que es muy recomendable para determinar el mejor conjunto de variables de predicción por incluir en la ecuación de regresión es calcular y comparar todas las posibles 2^k ecuaciones de regresión. Con este proceso se tendrá una ecuación, la cual no contiene ninguna variable de predicción ($\hat{Y} = \overline{Y}$), k ecuaciones cada una con una variable de predicción, $k(k-1)/2$ ecuaciones con dos variables de predicción y así sucesivamente. El procedimiento proporciona al investigador la oportunidad de evaluar y comparar todas las ecuaciones de regresión y, con base en la investigación de todas las discrepancias aparentes, debe surgir la mejor ecuación. Dado que hoy en día la capacidad de cómputo es muy extensa, la determinación de todas las posibles ecuaciones de regresión es el mejor método, aun si k tiene un valor tan grande como 9 o 10.

Cuando k es grande, puede no ser práctico determinar y evaluar todas las posibles ecuaciones de regresión. Para estos casos, se han desarrollado *técnicas para la selección de las variables* que pueden proporcionar al usuario información muy útil, sin tener que evaluar todas las posibles ecuaciones de regresión. Sin embargo, estas técnicas tienen algunos inconvenientes y no deben considerarse como iguales con respecto a la evaluación de todas las posibles regresiones. Mientras que los procedimientos para la selección de variables dan resultados confiables, cuando la muticolinealidad no es problema, éstos producirán resultados contradictorios para datos colineales. De esta forma, si se sospecha la presencia de multicolinealidad, no deberán emplearse métodos para la selección de variables. La técnica más usual de selección de variables emplea un procedimiento de *regresión por pasos* para obtener la mejor ecuación de regresión. Existen dos versiones principales de esta técnica: la selección hacia adelante y la eliminación hacia atrás.

El procedimiento de selección hacia adelante comienza con una ecuación que no contiene variables de predicción. La primera variable incluida en la ecuación es aquella que produce la mayor reducción en el valor de la suma de los cuadrados de los errores; ésta es la variable de predicción con el coeficiente de correlación simple más alto para la respuesta dada. Con base en una prueba de hipótesis, si el coeficiente de regresión es significativamente diferente de cero, la variable permanece en la ecuación y se comienza la búsqueda de una segunda variable. La segunda variable por incluir en la ecuación es aquella que produce la mayor reducción en la suma de los cuadrados de los errores, dada la presencia de la primera variable.* Ésta es la variable que posee el coeficiente de correlación más alto con la respuesta, después de que ésta se ha ajustado para tomar en cuenta el efecto de la primera variable. Si la significancia estadística es discernible para el coeficiente de regresión de la segunda variable, ésta se mantiene en la ecuación y se comienza la búsqueda de una tercera variable de predicción. El proceso se continúa de esta forma hasta que la significancia estadística no sea discernible para el coeficiente de la última variable que ha entrado a la ecuación.

El procedimiento de eliminación hacia atrás comienza con la ecuación de regresión que contiene a todas las variables de predicción. Entonces se eliminan, una a la

* En este momento pueden surgir dificultades cuando los datos son colineales.

vez, las variables menos importantes con base en su contribución a la reducción en el valor de la suma de los cuadrados de los errores. Por ejemplo, la primera variable por omitir será aquella cuyo efecto sobre la reducción en el valor de la suma de los cuadrados de los errores, dada la presencia de las demás variables, sea el más pequeño. El procedimiento concluye cuando los coeficientes de todas las variables que aún permanecen en la ecuación tienen una significancia estadísticamente discernible.

El procedimiento de selección hacia adelante se ha modificado de tal manera que se considere la posibilidad de eliminar una variable en cada etapa. Esta modificación da origen a lo que en forma usual se conoce en los paquetes de computación como procedimiento de regresión por pasos *(stepwise)*. Con este método puede eliminarse, en una etapa posterior, una variable de predicción cuya inclusión se llevó a cabo en una etapa anterior. De nuevo, el proceso de decisión se basa en la reducción en el valor de la suma de los cuadrados de los errores y de las pruebas F parciales y depende de la combinación particular de las variables que se tienen en la ecuación de regresión.

Con el desarrollo de paquetes para computadora cada vez más elaborados se tienen disponibles otras técnicas, pero la característica común sigue siendo el valor de la suma de los cuadrados de los errores cuando una variable entra a (o es removida de) la regresión, dada la presencia de las demás variables de predicción. Para datos "con buen comportamiento", los procedimientos de regresión por pasos y de eliminación hacia atrás en general proporcionan los mismos resultados. Si existe alguna diferencia entre éstos, este hecho muchas veces constituye una buena indicación para considerar el problema con mayor cuidado, así como la realización de análisis adicionales.

Para evaluar y comparar las ecuaciones de regresión, de manera especial dentro del contexto de todas las posibles regresiones, es necesario tener criterios efectivos. Dos de los criterios más útiles son el del cuadrado medio del error (*CME*) y el criterio C_p. Con el propósito de tener un panorama más completo, también se estudiará el coeficiente de correlación múltiple R^2.

1. *El criterio del cuadrado medio del error.* Recuérdese que el cuadrado medio del error es igual a la varianza residual. Dado que *CME* es la suma de los cuadrados de los residuos dividida entre el número de grados de libertad de *SCE, CME* toma en cuenta el número de parámetros en el modelo a través del número de grados de libertad. Mientras que la suma de los cuadrados de los errores no puede aumentar si se permiten más variables en el modelo, no ocurre lo mismo con el cuadrado medio del error si la reducción en el valor de *SCE* es tan pequeña que no pueda compensar la pérdida del número de grados de libertad adicionales. Por ejemplo, recuérdese la tabla 14.3 y en particular los modelos *a* y *c*. Nótese que $SCE\,(x_2) = 3038.34$ es mayor que $SCE\,(x_2, x_3) = 3016.59$, pero $\text{CME}\,(x_2) = 101.28$ es menor que $(x_2, x_3) = 104.02$. Con el criterio *CME* puede determinarse el conjunto de variables de predicción que minimice a *CME* o casi lo haga en el momento para el que la introducción de más variables de predicción en la ecuación de regresión ya no se encuentre garantizada.

2. *El criterio* C_p. Recuérdese que la varianza residual S^2 es un estimador no sesgado de la varianza del error σ^2 sólo cuando se ha escogido la forma correcta para el

modelo de regresión. De otra forma, puede demostrarse que

$$E(S^2) = \sigma^2 + \frac{\sum_{i=1}^{n} A_i^2}{(n - p)}, \tag{14.27}$$

donde p es el número de términos que aparecen en el modelo, incluido el término constante y

$$A_i = E(Y_i) - E(\hat{Y}_i)$$

es el sesgo.

Supóngase que la ecuación de regresión, la cual contiene k variables de predicción, se ha escogido en forma cuidadosa, de tal manera que $CME \equiv S^2$ es un estimador no sesgado de σ^2. Pero para cualquier ecuación de regresión que sólo contenga a un subconjunto de las k variables de predicción, es posible que $A_i \neq 0$, y las predicciones de la respuesta con base en la ecuación de regresión estimada pueden encontrarse sesgadas. Para evaluar la efectividad de esta ecuación de regresión, como un medio para formular predicciones, debe considerarse el cuadrado medio del error de un valor predicho, más que la varianza de éste. El cuadrado medio del error total estandarizado que se define como

$$\begin{aligned} \Gamma_p &= \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{CME}(\hat{Y}_i) \\ &= \frac{1}{\sigma^2} \left[\sum_{i=1}^{n} A_i^2 + \sum_{i=1}^{n} Var(\hat{Y}_i) \right], \end{aligned} \tag{14.28}$$

se ha propuesto como un criterio apropiado de la bondad del ajuste para una ecuación de regresión estimada la cual contiene p términos. La cantidad Γ_p considera tanto a la componente del sesgo en $\hat{Y}_i$, ya que algunas de las variables de predicción no se encuentran incluidas, así como a la varianza en $\hat{Y}_i$ para todas las n observaciones de la respuesta. A continuación se obtendrá un estimador para Γ_p.

Puede demostrarse que

$$\sum_{i=1}^{n} Var(\hat{Y}_i) = p\sigma^2,$$

lo cual implica que la varianza total de la predicción aumenta conforme el número de términos en la ecuación de regresión también aumenta. Al sustituir en (14.28), se tiene

$$\Gamma_p = \frac{1}{\sigma^2} \sum_{i=1}^{n} A_i^2 + p. \tag{14.29}$$

Dado que para una ecuación de regresión que contiene p términos

$$\text{SCE}_p = (n - p)S_p^2,$$

se tiene

$$\begin{aligned} E(\text{SCE}_p) &= (n - p)E(S_p^2) \\ &= (n - p)\left[\sigma^2 + \frac{\sum A_i^2}{(n - p)}\right] \\ &= (n - p)\sigma^2 + \sum A_i^2, \end{aligned}$$

o

$$\sum A_i^2 = E(\text{SCE}_p) - (n - p)\sigma^2.$$

Al sustituir en (14.29), se obtiene

$$\begin{aligned} \Gamma_p &= \frac{E(\text{SCE}_p) - (n - p)\sigma^2}{\sigma^2} + p \\ &= \frac{E(\text{SCE}_p)}{\sigma^2} - (n - p) + p \\ &= \frac{E(\text{SCE}_p)}{\sigma^2} - (n - 2p). \end{aligned}$$

Dado que SCE_p es un estimador de $E(\text{SCE}_p)$ y S_k^2 lo es a su vez de σ^2, un estimador de Γ_p es la estadística

$$C_p = \frac{\text{CSE}_p}{S_k^2} - (n - 2p). \tag{14.30}$$

Nótese que SCE_p es la suma de los cuadrados de los errores para la ecuación de regresión, la cual contiene p términos y que $S_k^2 = \text{CME}(x_1, x_2, \ldots, x_k)$ es el estimador de σ^2 basado en todas las k variables de predicción.

Los valores deseables para C_p para la bondad del ajuste de una ecuación de regresión que contiene p términos son aquellos que se encuentran muy cercanos a p. Lo anterior surge del hecho de que si el sesgo de una ecuación de regresión de p términos es despreciable $\Sigma\ A_i^2 \approx 0$ y $E(\text{SCE}_p) = (n - p)\sigma^2$. Bajo esta condición, el valor esperado de la estadística C_p es

$$\begin{aligned} E(C_p \mid A_i = 0) &= \frac{(n - p)\sigma^2}{\sigma^2} - (n - 2p) \\ &= p. \end{aligned}$$

De esta forma, cuando se obtienen todas las posibles regresiones, se calcula un valor de C_p para cada caso. Las regresiones que tienen valores de C_p cercanos a p se consideran como deseables.

Puede ser benéfico aceptar un pequeño sesgo en la predicción, mediante la eliminación de algunas variables de predicción, aun si sus coeficientes de regresión son es-

tadísticamente significativos, con excepción de los que tienen un valor igual a cero. Lo anterior es especialmente cierto si los coeficientes de regresión del nuevo modelo se estiman con varianzas pequeñas; además, dado que la varianza total de la predicción aumenta conforme se añaden más variables al modelo de regresión, puede ser ventajoso eliminar algunas variables con el propósito de disminuir el error promedio de la predicción.

Además de considerar a *CME* y a C_p, también se debe considerar el coeficiente de correlación múltiple R^2 para evaluar las ecuaciones de regresión. Dado que R^2 varía en forma inversa a como lo hace la suma de los cuadrados de los errores, R^2 aumentará conforme se añadan más variables al modelo de regresión y R^2 alcanzará su valor máximo cuando todas las variables de predicción se encuentren en la ecuación de regresión. Por lo tanto, la razón para emplear a R^2 como un criterio, no es la de encontrar el conjunto de variables que maximiza R^2, sino más bien determinar el punto más allá del cual sumar más variables no es deseable, ya que el incremento que se tiene en R^2 es mínimo.

Para proporcionar una ilustración de todas las posibles regresiones y sus comparaciones, tomando en cuenta los criterios anteriores, de nuevo considérense los datos de Prater. La tabla 14.9 contiene las estimaciones por mínimos cuadrados para los coeficientes de cada regresión (distintas de la trivial $\hat{y}_i = \bar{y} = 19.66$), y la tabla 14.10 identifica los correspondientes valores de SCE, CME, C_p y R^2.

El cuadrado medio del error cuando las cuatro variables de predicción se encuentran en el modelo de regresión es $\text{CME}(x_1, x_2, x_3, x_4) = 4.99$. De esta forma, por ejemplo, para obtener el valor de C_p para la regresión de Y sobre x_1, x_3, y x_4, se tiene que $SCE(x_1, x_3, x_4) = 146.00$, $p = 4$, $n = 32$ y

$$C_p = \frac{146}{4.99} - (32 - 8) = 5.26.$$

TABLA 14.9 Todas las regresiones posibles para los datos de Prater

Variables de predicción en el modelo	b_0	b_1	b_2	b_3	b_4
x_1	1.264	0.469			
x_2	13.087		1.572		
x_3	41.389			−0.090	
x_4	−16.662				0.019
x_1, x_2	12.256	0.025	1.539		
x_1, x_3	35.174	0.096		−0.080	
x_1, x_4	−64.951	1.009			0.136
x_2, x_3	− 2.524		2.257	0.053	
x_2, x_4	−37.808		2.677		0.139
x_3, x_4	18.468			−0.209	0.156
x_1, x_2, x_3	−11.013	0.125	2.278	0.067	
x_1, x_2, x_4	−53.899	0.422	2.154		0.144
x_1, x_3, x_4	4.032	0.222		−0.187	0.157
x_2, x_3, x_4	8.562		0.523	−0.175	0.154
x_1, x_2, x_3, x_4	− 6.821	0.227	0.554	−0.150	0.155

TABLA 14.10 Criterios de bondad de ajuste para todas las posibles regresiones para los datos de Prater

Variables de predicción	R^2	SCE	CME	C_p
x_1	0.0607	3347.82	111.59	642.91
x_2	0.1475	3038.34	101.28	580.89
x_3	0.0992	3210.38	107.01	615.36
x_4	0.5063	1759.69	58.66	324.64
x_1, x_2	0.1476	3037.97	104.76	582.81
x_1, x_3	0.1005	3205.74	110.54	616.43
x_1, x_4	0.7582	861.95	29.72	146.74
x_2, x_3	0.1536	3016.59	104.02	578.53
x_2, x_4	0.8962	369.87	12.75	48.12
x_3, x_4	0.9521	170.61	5.88	8.19
x_1, x_2, x_3	0.1558	3008.76	107.46	578.96
x_1, x_2, x_4	0.9255	265.48	9.48	29.20
x_1, x_3, x_4	0.9590	146.00	5.21	5.26
x_2, x_3, x_4	0 9549	160.62	5.74	8.19
x_1, x_2, x_3, x_4	0.9622	134.80	4.99	5.00

Al tomar en cuenta, tanto a *CME* como C_p, la mejor ecuación de predicción para la producción de gasolina debe seleccionarse de las regiones que incluyen (x_3, x_4), (x_1, x_3, x_4), (x_2, x_3, x_4), y (x_1, x_2, x_3, x_4). Esta última no es en particular atractiva, ya que las estimaciones de los coeficientes de regresión para el término constante y para x_2 tienen desviaciones estándar muy grandes. A pesar de que la ecuación de regresión que contenga x_2, x_3, y x_4 tiene valores de CME y C_p muy cercanos a los óptimos, ésta carece de una precisión satisfactoria para la estimación del coeficiente de x_2, dado que $b_2 = 0.523$, con $s(B_2) = 0.396$. Puede decirse lo mismo de la regresión que comprenda a x_1, x_3, y x_4 para las estimaciones de β_0 y del coeficiente de x_1 (véase el modelo *b* en la tabla 14.8). De acuerdo con lo anterior, se acepta un pequeño sesgo en la predicción y se concluye que la ecuación de regresión que contiene a x_3 y a x_4 es la mejor para predecir la producción de gasolina en el intervalo de valores de las observaciones.

A continuación se dan las etapas por seguir en un procedimiento de regresión paso a paso:

1. El procedimiento comienza mediante la obtención de k ecuaciones de regresión lineal simples.
   La estadística F

$$F = \text{CMR}(x_i)/\text{CME}(x_i)$$

   se calcula para cada $i = 1, 2, \ldots, k$ variables. Si el mayor valor F excede un nivel de significancia estadística, previamente determinado, la variable correspondiente es la primera que se incluye en la regresión. De otro modo, la mejor ecuación es $\hat{Y} = \bar{Y}$. Este proceso es idéntico al que se sigue para determinar la variable de predicción que tiene la mayor correlación con la respuesta.

2. Supóngase que la variable x_3 entra a la ecuación de regresión en el paso 1. En este momento, el procedimiento de regresión paso a paso calcula todas las ecuaciones que contienen dos variables, incluyendo a x_3. Para cada caso, el valor de la estadística F parcial

$$F = \text{CMR}(x_i \mid x_3)/\text{CME}(x_i, x_3)$$

se calcula para determinar si puede rechazarse H_0: $\beta_i = 0$ en presencia de x_3. Si el mayor valor de F es suficiente para la significancia estadística, la segunda variable correspondiente se añade a la ecuación.

3. Supóngase que se añade x_1 a la ecuación en el paso 2. El procedimiento continúa mediante un examen para determinar si alguna de las otras variables que ya se encuentran en la ecuación debe eliminarse ahora; en este caso, ésta podría ser x_3. Se calcula el valor de la estadística F parcial

$$F = \text{CMR}(x_3 \mid x_1)/\text{CME}(x_1, x_3)$$

y se compara con un nivel predeterminado de significancia. Si el efecto de x_3 dado x_1 no es estadísticamente discernible, se elimina a x_3 de la ecuación; de otro modo se retiene. Para etapas posteriores existirá un cierto número de las pruebas F parciales para todas las variables que se añadieron en etapas anteriores. La variable que puede eliminarse es aquella para la que el valor de F es el más pequeño.

4. Supóngase que se retiene a x_3; en este momento la ecuación de regresión incluye a x_1 y a x_3. El proceso se continúa mediante un examen para determinar cuál de las variables restantes es candidata para incluirse en el modelo. Entonces, se examina si alguna de las variables que ya se encuentran incluidas debe eliminarse ahora. El proceso termina cuando ninguna de las demás variables de predicción puede añadirse o eliminarse del modelo de regresión.

Se deja como un ejercicio para el lector emplear los datos de producción de gasolina con todas las opciones de selección posibles de variables y se compararán los resultados.

14.6 Análisis de residuos o residuales

En la sección anterior se examinaron algunas formas para determinar la "mejor" ecuación de regresión, bajo las circunstancias impuestas por el conjunto de datos. Una manera muy efectiva de detectar las posibles deficiencias de un modelo radica en llevar a cabo un análisis de residuos. Ningún otro aspecto es tan importante en el análisis de regresión como el análisis de los residuos. El conocido economista Paul A. Samuelson comentaba: "al científico que hace predicciones le recomiendo que siempre estudie sus residuales".

Como se hizo notar en el capítulo 12, el análisis de los residuos puede descubrir las violaciones de las suposiciones o las deficiencias del modelo. Se examinarán tres deficiencias muy comunes: la ecuación de regresión puede no ser lineal en las variables de predicción; la varianza del error σ^2 puede no ser constante y una o más de las variables de predicción que ejercen una influencia importante pueden no estar in-

cluidas en el modelo. También se considerará el problema de las observaciones *discrepantes o aberrantes,* que son aquellas cuyos valores se encuentran alejados del comportamiento general del resto de los datos.

Recuérdese que el i-ésimo residuo e_i es la diferencia numérica que existe entre el valor observado y_i y el correspondiente valor estimado $\hat{y}_i$, para toda $i = 1, 2, \ldots, n$. El residuo e_i se considera como una estimación del verdadero error no observable ε_i. El error cuadrático medio es la varianza de los residuos, la que a su vez es una estimación de σ^2.

En esencia, el análisis de residuos significa realizar un análisis de sus gráficas de los residuos. Si se ha definido la ecuación de regresión en forma correcta y no existe ninguna deficiencia, entonces una gráfica de los residuos contra cualesquiera de los valores estimados $\hat{y}_i$ a los correspondientes valores de cada variable de predicción en la ecuación no mostrará ningún patrón, es decir, no existirá ninguna relación entre los residuos y los valores ajustados o entre los residuos y ios valores de las variables de predicción. Si existe alguna relación, ésta sugerirá el hecho de que hay una deficiencia en la ecuación de regresión. Para detectar las áreas de problemas a través del análisis de los residuos, es preferible, de nuevo, emplear los residuos estandarizados. Dado que la media de los residuos es igual a cero,

$$e_{i_s} = e_i/s$$

define al i-ésimo residuo estandarizado donde s es la desviación estándar residual ($\sqrt{\text{CME}}$). Debe notarse que si el tamaño de la muestra n es muy grande, la distribución de los residuos estandarizados deberá encontrarse aproximada en forma adecuada por una distribución normal estándar. De hecho, muchos investigadores han sugerido que cualquier alejamiento notable de la normalidad en la distribución de los residuos puede indicar una deficiencia en el modelo.

Para determinar si un modelo de regresión es lineal o no en las variables de predicción, se grafican los residuos contra los correspondientes valores de cada una de las variables de predicción que figuran en la ecuación de regresión. Para determinar si la varianza del error es o no constante, se grafican los residuos estandarizados contra los correspondientes valores estimados de la respuesta. Finalmente, para determinar si una variable de predicción, potencialmente importante, debe incluirse o no en el modelo de regresión, se grafican los residuos contra los valores de esta variable. Si la ecuación de regresión estimada está prácticamente libre de cualquier deficiencia o violación de suposiciones, entonces los residuos estandarizados tenderán a encontrarse dentro de una banda horizontal centrada alrededor del valor cero, sin ninguna tendencia sistemática a ser positivos o negativos, y en forma muy rara se encontrarán fuera del intervalo ± 3. Cualquier desviación significativa con respecto a este comportamiento indicará la existencia de un problema.

La figura 14.2 representa algunas gráficas usuales de residuos: *a)* cuando se encuentra presente un efecto cuadrático causado por una variable de predicción y que debe incluirse en el modelo; *b)* cuando la varianza del error no es constante y deben emplearse mínimos cuadrados con factores de peso (ponderados) para estimar los coeficientes de regresión y *c)* cuando una variable que se ha eliminado muestra una fuerte asociación (lineal) con los residuos y por lo tanto debe incluirse en el modelo

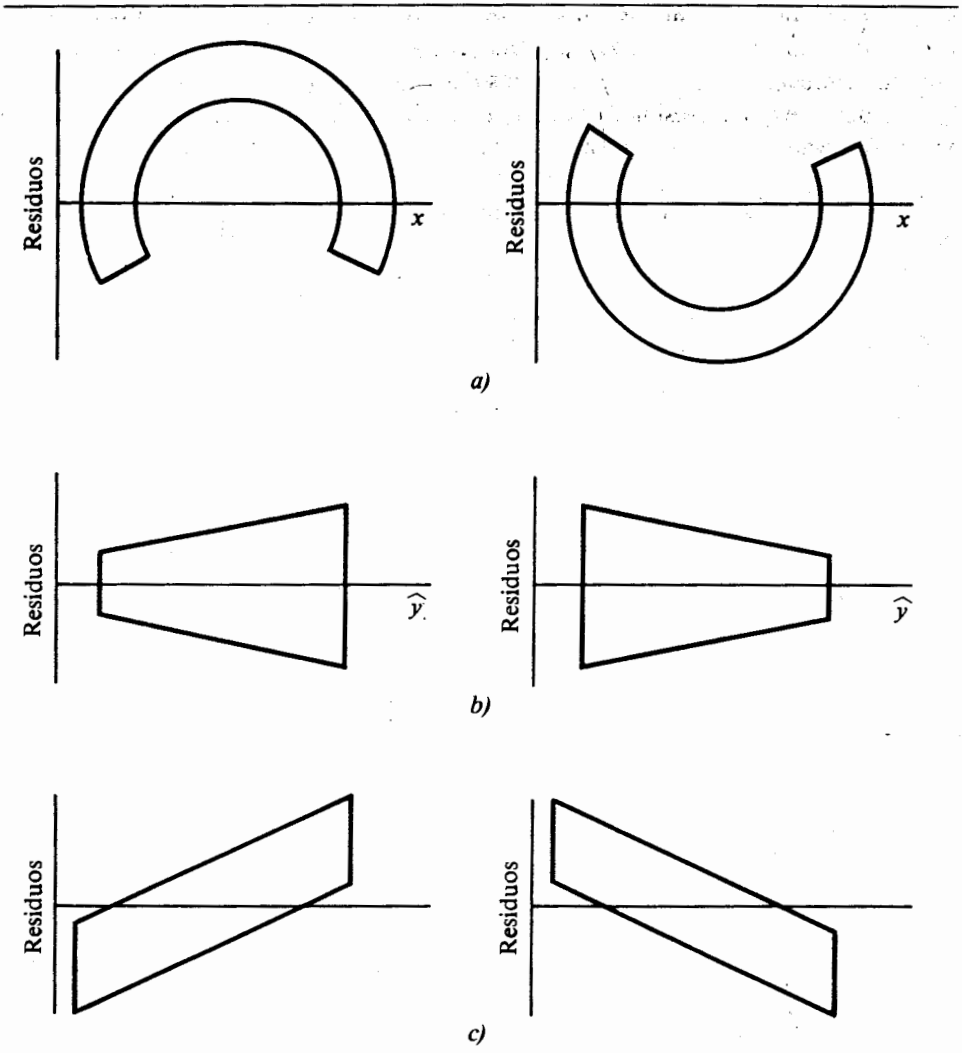


FIGURA 14.2 Gráficas comunes de residuos para: *a)* la presencia de un efecto cuadrático; *b)* la varianza no constante del error, y *c)* el efecto lineal de una variable omitida

de regresión. Puede decirse más con respecto a estos tres casos. Si la ecuación de regresión contiene sólo un efecto lineal causado por una variable de predicción x, cuando en realidad existe un efecto cuadrático estadísticamente apreciable, entonces la gráfica de los residuos estandarizados contra x será una curva en forma de U (o de U invertida). Bajo un efecto cuadrático, los residuos correspondientes a los valores extremos de x tenderán a ser grandes y positivos (negativos), y los residuos que se encuentran en la parte media del intervalo de valores de x tenderán a ser pequeños pero

negativos (positivos). En general, mediante la inclusión de un término cuadrático de x en el modelo, mejora considerablemente el valor predictivo de la ecuación de regresión resultante con respecto a la ecuación previa. Los efectos de orden superior también pueden detectarse de la misma manera.

Si la gráfica de los residuos da como resultado una figura en forma de cuña, entonces es posible que la suposición de que la varianza del error es constante no se cumpla. En otras palabras, si existe una tendencia a aumentar o disminuir los residuos estandarizados al aumentar los valores estimados de la respuesta, la varianza del error puede no ser constante. Esto da origen a lo que se conoce como *modelo heterocedástico*. Para remediar esta situación se emplea el método de *mínimos cuadrados con factores de peso,* en donde los pesos son inversamente proporcionales a la varianza de los errores. De esta forma, en lugar de intentar determinar las estimaciones de los coeficientes de regresión mediante la minimización de la suma de los cuadrados de los errores, se determina el conjunto de valores para los cuales la suma de pesos de los cuadrados de los errores es un mínimo. El motivo para emplear mínimos cuadrados con factores de peso en una situación heterocedástica es estimar los coeficientes de regresión con pequeñas desviaciones lo que a su vez produce un mejor ajuste.

Si cuando los residuos estandarizados se grafican contra una variable que no forma parte de la ecuación de regresión, pero bajo la cual se pudo observar la respuesta, se observa una tendencia lineal (o de orden superior); entonces, como se mencionó en el capítulo 13, los errores no pueden considerarse más como independientes de esta variable. En general, este tipo de variable resulta ser un efecto demográfico o relacionado con el tiempo. Por ejemplo, para muchos experimentos en los que los datos se observan durante un periodo significativo, el investigador podría inicialmente decidir no incluir al tiempo como una variable de predicción potencial. Pero si los residuos revelan un patrón sistemático cuando se grafican contra el tiempo, la variable tiempo deberá introducirse en la ecuación de regresión.

Las gráficas de residuos también son una ayuda al tratar con observaciones extremas o discrepantes. En general, las observaciones extremas tienen residuos que son, en forma relativa, grandes, comparados con los de las demás observaciones. En general, el valor del residuo estandarizado de una observación discrepante se encontrará más allá del intervalo ± 3. Las observaciones discrepantes pueden crear situaciones difíciles en una ecuación de regresión, debido a que tienen un efecto desproporcionado sobre los valores estimados de los coeficientes de regresión. Recuérde que una de las suposiciones de la estimación por mínimos cuadrados es que el conjunto de datos es típico de la situación para la cual se intenta identificar una buena ecuación de predicción. Por lo tanto, la remoción de cualquier observación del conjunto de datos no tendrá, en forma virtual, ningún efecto sobre la ecuación de regresión. Lo anterior constituye precisamente el porqué puede removerse, sólo con extremo cuidado, una observación discrepante. Un método lógico que se ha sugerido es remover una observación discrepante sólo si existe evidencia comprobada de que ésta es el producto de un error causado, por ejemplo, por un mal funcionamiento del instrumento de medición. En ausencia de clara evidencia de error, la observación discrepante puede ser información única con respecto a la respuesta y ser vital para el entendimiento del fenómeno.

Los siguientes dos ejemplos ilustrarán los casos *a)* y *c)* que se muestran en la figura 14.2. El caso en el cual se tiene una varianza no constante se analizará en la sección 14.8.

Ejemplo 14.3 Una compañía manufacturera desea predecir el costo unitario de fabricación Y de uno de sus productos como una función de la tasa de producción (que fluctúa en el tiempo) x_1 y de los costos de material y mano de obra x_2. Los datos se recabaron durante un periodo de 20 meses durante el cual la tasa de producción y los costos del material y la mano de obra experimentaron una fluctuación muy amplia. La tasa de producción se midió como un porcentaje de la capacidad total de producción, y se utilizó un índice apropiado para reflejar los costos del material y mano de obra. Las observaciones se encuentran en la tabla 14.11. Obténgase la mejor ecuación de regresión para predecir el costo por unidad.

Primero se supondrá un modelo de regresión lineal que sólo tome en cuenta a x_1 y a x_2. En la tabla 14.12 se proporcionan las estimaciones y otra información importante. Hasta aquí parece que todo marcha muy bien. Las estimaciones tienen sentido (valor negativo para el coeficiente x_1 y positivo para el de x_2), las desviaciones estándar son pequeñas, el valor de R^2 es relativamente alto y todos los efectos son estadísticamente discernibles. Por lo tanto, se podría concluir que se ha obtenido una buena ecuación de predicción, pero una gráfica de los residuos estandarizados contra x_1 revela un patrón cuadrático en la mitad superior de la figura 14.3. Ningún patrón es evidente para x_2 .

TABLA 14.11 Datos de la muestra para el ejemplo 14.3

Y	x_1	x_2
13.59	87	80
15.71	78	95
15.97	81	106
20.21	65	115
24.64	51	128
21.25	62	128
18.94	70	115
14.85	91	92
15.18	94	93
16.30	100	111
15.93	102	116
16.45	82	117
19.02	74	127
18.16	85	133
18.57	86	135
17.01	90	136
18.03	93	140
19.22	81	142
21.12	72	148
23.32	60	150

TABLA 14.12 Análisis de regresión para el ejemplo 14.3

Regresión de Y sobre x_1 y x_2

Variable	*Coeficiente de regresión estimado*	*Desviación estándar estimada*	*Valor t*
Constante	20.2800	2.1300	9.54
x_1	−0.1377	0.0159	−8.69
x_2	0.0742	0.0110	6.77

$R^2 = 0.914$ $\quad t_{0.975,\,17} = 2.11$

ANOVA

Fuente	gl	SC	CM	*Valor F*
Regresión	2	144.39	72.19	90.24
x_1	1	107.72	107.72	134.65
$x_2 \mid x_1$	1	36.67	36.67	45.84
Error	17	13.59	0.80	
Total	19	157.98	$f_{0.95,\,2,\,17} = 3.59$; $f_{0.95,\,1,\,17} = 4.45$	

La gráfica de los residuos para x_1 implica que debe incluirse un término cuadrático en x_1 en el modelo de regresión. De esta forma, se ajustará el modelo

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \varepsilon$$

obteniéndose los resultados que se muestran en la tabla 14.13.

Una comparación con los resultados anteriores revela que la inclusión de un efecto cuadrático en x_1 mejora en forma considerable la ecuación de regresión estimada. Por ejemplo, los coeficientes de regresión, tanto de x_1 como de x_2, se estiman con una mejor precisión comparada con la anterior y el valor de R^2 se incrementa hasta 0.981. Además, la nueva gráfica de residuos contra x_1 (véase la Fig. 14.4) no muestra ningún patrón apreciable.

Ejemplo 14.4 Recuérdese el ejemplo de los salarios iniciales contra la calificación promedio, empleado a través de todo el capítulo 13. Quizá el lector se pregunte si existiesen otras variables de predicción potenciales. Supóngase que también se ha observado la edad de cada estudiante en la muestra. Ya que algunas compañías tienen como requisito poseer alguna experiencia en el campo y un recién egresado de mayor edad podría tenerla, es posible que la edad de éste pueda influenciar en el salario inicial que percibirá. Los datos, tomando en cuenta la edad, se encuentran en la tabla 14.14.

Cuando se hace una gráfica de los residuos estandarizados de la ecuación de regresión estimada $\hat{y} = -6.63 + 8.12x_1$ contra los correspondientes valores de x_2

TABLA 14.13 Análisis de regresión revisado para el ejemplo 14.3

Regresión de Y sobre x_1 y x_2 y x_1^2

Variable	*Coeficiente de regresión estimado*	*Desviación estándar estimada*	*Valor t*
Constante	41.550000	3.050000	13.64
x_1	−0.700300	0.076200	−9.20
x_2	0.073400	0.005400	13.68
x_1^2	0.003624	0.000488	7.43

$R^2 = 0.981$ $\quad t_{0.975,\,16} = 2.12$

ANOVA

Fuente	gl	SC	CM	*Valor F*
Regresión	3	154.92	51.640	270.37
x_1	1	107.72	107.72	563.98
$x_2 \mid x_1$	1	36.66	36.66	191.94
$x_1^2 \mid x_1, x_2$	1	10.54	10.54	55.18
Error	16	3.06	0.191	
Total	19	157.98	$f_{0.95,\,3,\,16} = 3.24$; $f_{0.95,\,1,\,16} = 4.49$	

(véase la Fig. 14.5), se observa una tendencia lineal ascendente. Por lo tanto, se incluye el efecto lineal de x_2 en el modelo de regresión y se ajusta

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

En la tabla 14.15 se muestran los nuevos resultados. Dado que ahora se estiman con mejor precisión el término constante, el coeficiente de x_1 y el valor de R^2 ha aumentado en forma apreciable, la inclusión de x_2 da como resultado una mejor ecuación de predicción.

14.7 Regresión polinomial

En la sección 14.2 se mencionó que el modelo polinomial dado por (14.3), o alguno que contenga términos de interacción como (14.4), es un caso especial del modelo lineal general. De hecho, en el ejemplo 14.3 se mostró cómo el efecto cuadrático de una variable de predicción puede mejorar la capacidad predictiva de la ecuación de regresión. En esta sección se ahondará más sobre este tipo de modelos.

Si se ha identificado sólo una variable de predicción x y la gráfica de las respuestas observadas contra los valores de x revela una curvatura, entonces debe usarse un polinomio en x, de cierto grado, para aproximar la verdadera curva de regresión.

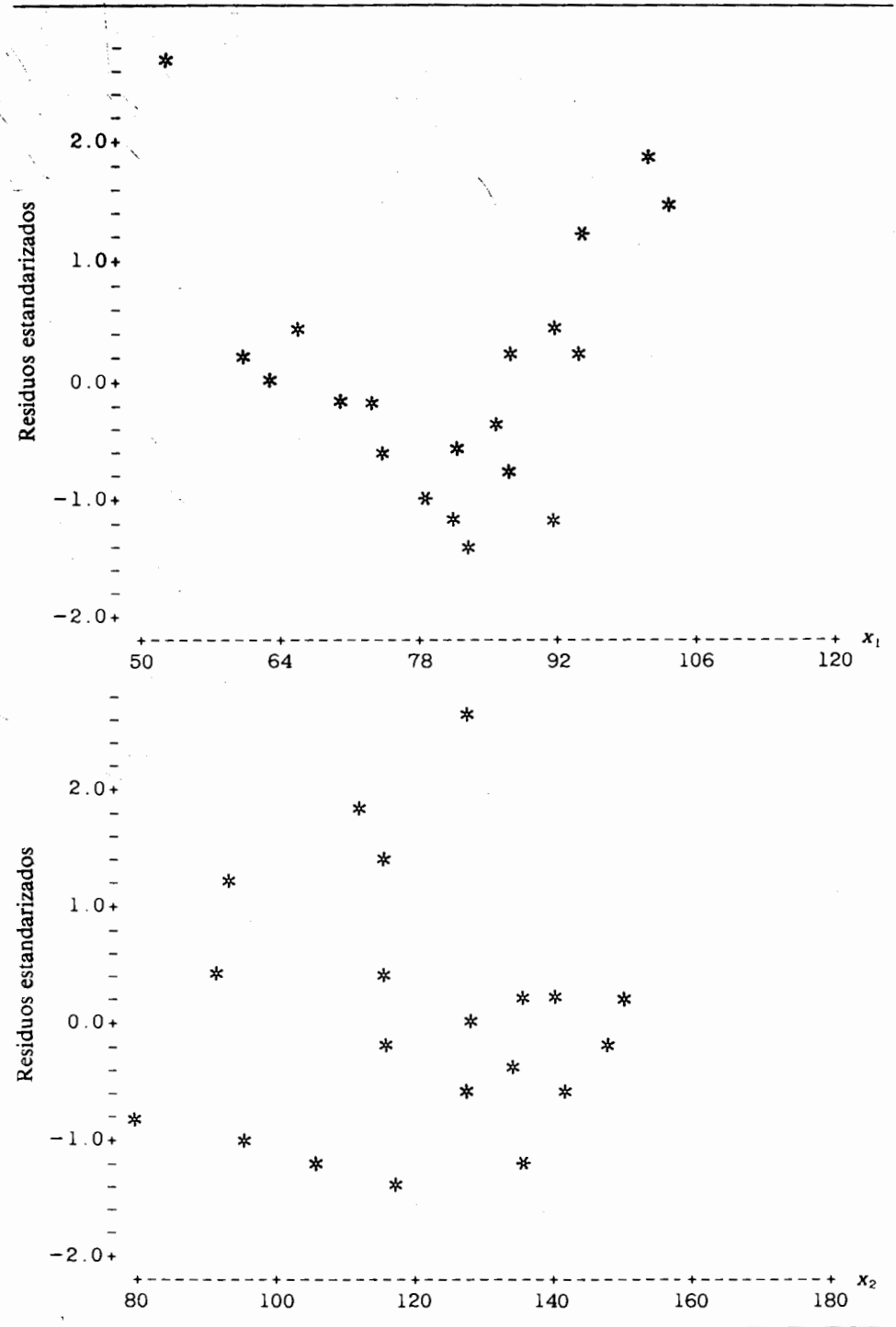


FIGURA 14.3 Gráficas de los residuos estandarizados para el ejemplo 14.3

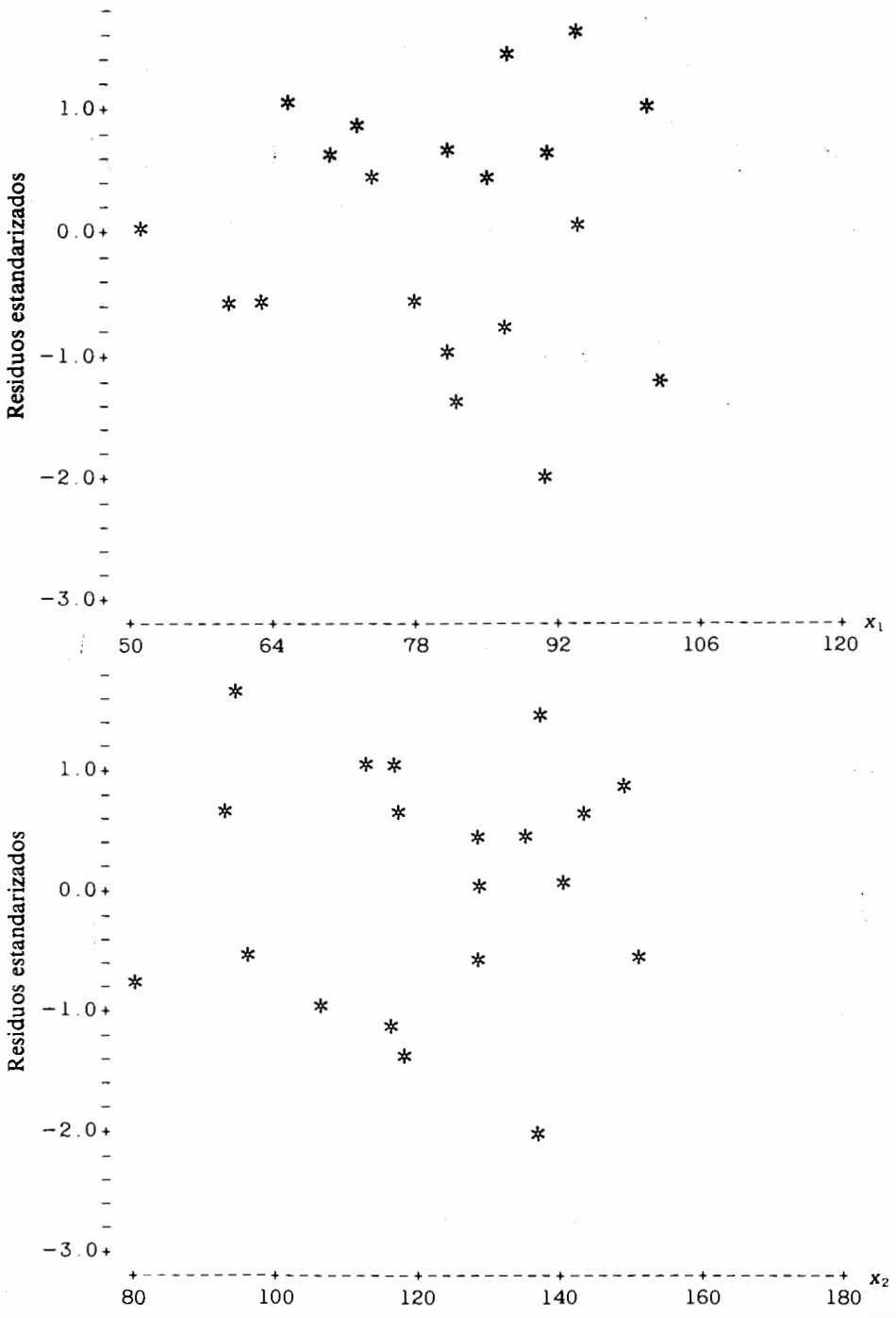


FIGURA 14.4 Gráficas de los residuos estandarizados para la ecuación de regresión revisada en el ejemplo 14.3

TABLA 14.14 Datos aumentados para el ejemplo de los salarios iniciales

Y	x_1 (*CP*)	x_2 *Edad*
18.5	2.95	22
20.0	3.20	23
21.1	3.40	23
22.4	3.60	23
21.2	3.20	27
15.0	2.85	22
18.0	3.10	25
18.8	2.85	28
15.7	3.05	23
14.4	2.70	22
15.5	2.75	28
17.2	3.10	22
19.0	3.15	26
17.2	2.95	23
16.8	2.75	26

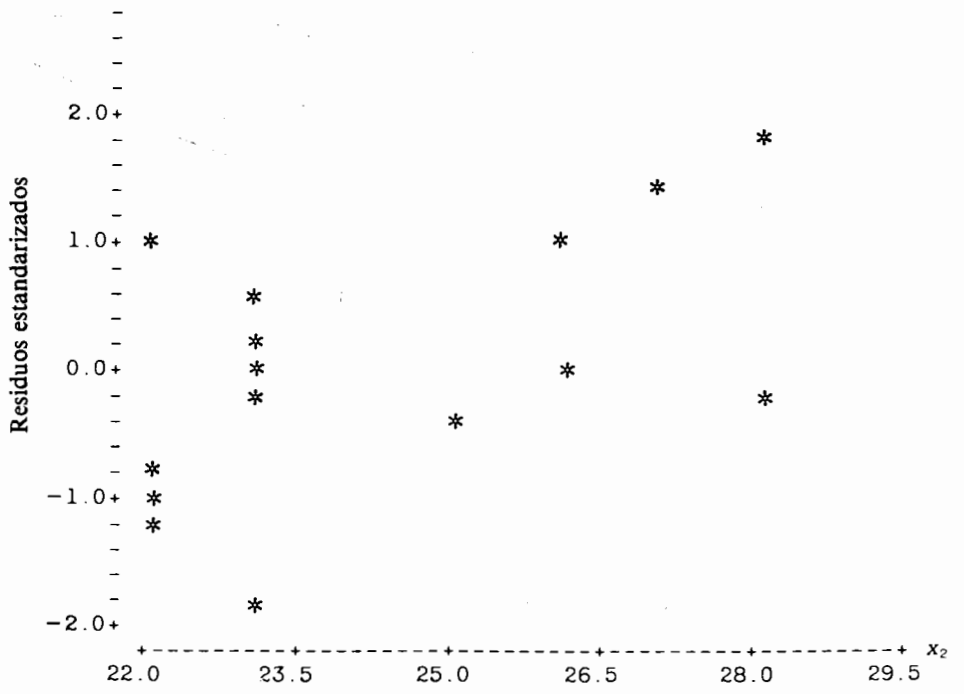


FIGURA 14.5 Residuos estandarizados contra la edad para el ejemplo 14.4

TABLA 14.15 Análisis de regresión para el ejemplo 14.4

Variable	*Coeficiente de regresión estimado*	*Desviación estándar estimada*	*Valor t*
Constante	−16.880	5.470	−3.05
x_1	8.740	1.220	7.16
x_2	0.338	0.137	2.47

$R^2 = 0.813$ $\quad t_{0.975,\,12} = 2.179$

ANOVA

Fuente	gl	SC	CM	*Valor F*
Regresión	2	66.10	33.05	26.23
x_1	1	58.40	58.40	46.35
$x_2 \mid x_1$	1	7.70	7.70	6.11
Error	12	15.17	1.26	
Total	14	81.27	$f_{0.95,\,2,\,12} = 3.89$; $f_{0.95,\,1,\,12} = 4.75$	

Por ejemplo, un modelo cúbico en x está dado por

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{111} x_i^3 + \varepsilon_i,$$

donde β_1 recibe el nombre de *coeficiente lineal,* β_{11} es el *coeficiente cuadrático* y β_{111} es el *coeficiente cúbico*. Para mantener la costumbre se ha alterado en forma ligera la notación para estos coeficientes de regresión para reflejar el patrón de la correspondiente potencia de x.

Como se mencionó con anterioridad, lo que se busca con un polinomio es el grado que mejor ajuste los datos dados. De acuerdo con lo anterior, el interés recae en probar hipótesis, como por ejemplo, H_0: $\beta_{11} = 0$ o H_0: $\beta_{111} = 0$. Mediante el empleo de este enfoque se tiene la capacidad para determinar el polinomio más apropiado para estimar la respuesta promedio. Sin embargo, se advierte al lector que lo que se busca y se prefiere en forma general es un polinomio de un orden relativamente bajo. Se deberá evitar el empleo de potencias muy grandes de la variable de predicción, debido a que lo que ocurre la mayor parte de las veces es un ajuste que explica incluso las variaciones aleatorias que se encuentran en los datos; en otras palabras, siempre se puede encontrar un modelo polinomial de un grado, lo suficientemente alto para ajustar los datos de manera perfecta, ya que un polinomio de grado $n - 1$ pasará a través de todos los n valores de la respuesta.

Muchas veces un modelo completo de segundo orden que contiene términos lineales, cuadráticos y de interacción, proporciona una aproximación funcional excelente en comparación con una función de respuesta desconocida y, en forma general, compleja. Por ejemplo, un modelo de segundo orden en dos variables es

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i,$$

donde β_1 y β_2 son los coeficientes lineales, β_{11} y β_{22} son los coeficientes cuadráticos y β_{12} es el coeficiente de interacción. Para este modelo, la matriz $\mathbf{X}$ y el vector de parámetros $\boldsymbol{\beta}$ que figuran en la ecuación matricial

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

son

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{11}^2 & x_{12}^2 & x_{11}x_{12} \\ 1 & x_{21} & x_{22} & x_{21}^2 & x_{22}^2 & x_{21}x_{22} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n1}^2 & x_{n2}^2 & x_{n1}x_{n2} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{12} \end{bmatrix}.$$

Con los dos siguientes ejemplos se ilustrarán tanto un modelo polinómico en una variable, así como un modelo completo de segundo orden.

Ejemplo 14.5 La demanda de cierto producto cambió debido a una variación rápida de su precio por unidad. Supóngase que la demanda Y del producto se observa en una región geográfica en particular sobre un intervalo bastante amplio de precios x. Dados los datos que se encuentran en la tabla 14.16, determínese el grado de un polinomio que mejor ajuste estos datos.

Dado que sólo se tiene una variable de predicción, lo primero que se tiene que hacer es una gráfica de la demanda contra el precio por unidad. La figura 14.6 revela una curvatura, lo cual indica que debe intentarse el ajuste con un modelo cuadrático.

Para ilustrar cómo se detecta la curvatura, supóngase que se propone un modelo lineal sencillo. En la figura 14.7 se muestra un listado de computadora generado por Minitab y los residuos estandarizados resultantes contra el precio en la figura 14.8. La necesidad de incluir un efecto cuadrático en x es evidente.

TABLA 14.16 Datos de la muestra para el ejemplo 14.5

Y unidades	*x dólares*
360	8.8
305	9.7
230	9.9
242	10.3
180	11.0
172	12.5
121	13.2
83	14.8
122	15.8
91	17.4
105	18.2

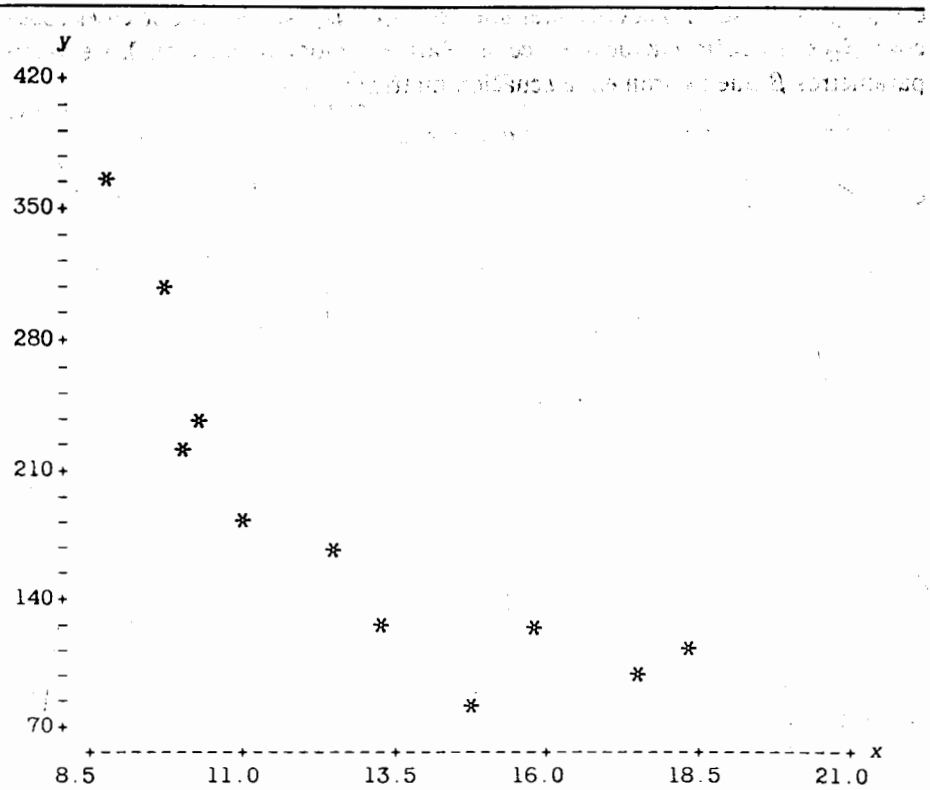


FIGURA 14.6 Gráfica de la demanda contra el precio por unidad para el ejemplo 14.5

LA ECUACION DE REGRESION ES
Y = 497. − 24.4 X1

	COLUMNA	COEFICIENTE	DEV. EST. DEL COEF.	COEFICIENTE-T = COEF/D.E.
	—	497.15	60.85	8.17
X1	C2	−24.419	4.594	−5.32

LA DEV. EST. DE Y CON RESPECTO A LA RECTA DE REGRESION ES
S = 47.53
CON (11− 2) = 9 GRADOS DE LIBERTAD

R-CUADRADA = 75.8 POR CIENTO

ANALISIS DE VARIANZA

DEBIDO A	GL	SC	CM = SC/GL
REGRESION	1	63815	63815
RESIDUO	9	20330	2259
TOTAL	10	84145	

FIGURA 14.7 Listado de computadora para el ejemplo 14.5

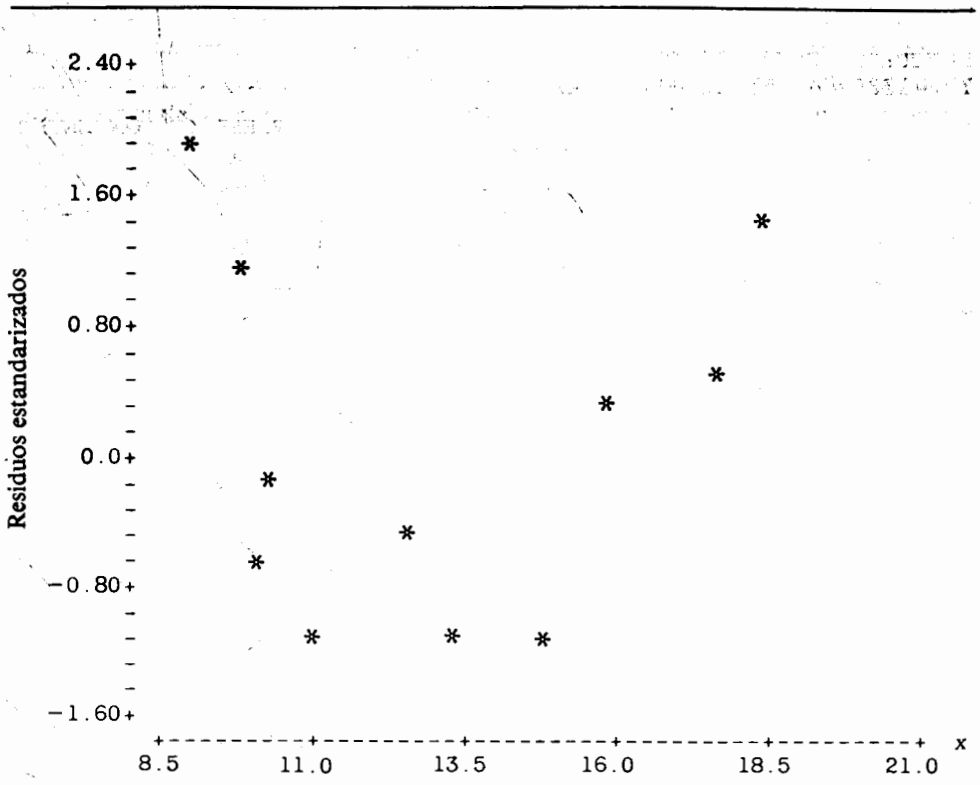


FIGURA 14.8 Residuos estandarizados contra el precio por unidad para el modelo lineal del ejemplo 14.5

El listado de salida para un modelo cuadrático se muestra en la figura 14.9. Como se esperaba, existe una considerable mejoría en la predicción proporcionada por la ecuación de regresión estimada, que la que se tenía con un modelo lineal simple. Nótese que Minitab también proporciona la "SC tipo *I*", es decir, a través de las entradas identificadas por *"C2"* y *"C3"* se tiene que SCR(x) = 63 814.5 y SCR($x^2 \mid x$) = 14 961.4, respectivamente.

Aunque no se proporciona una gráfica de los residuos estandarizados contra el precio para el modelo cuadrático, no mostrará ningún patrón evidente; además, no se obtiene ninguna mejoría apreciable si se añaden al modelo términos de orden superior. Una ecuación de regresión estimada de orden cuadrático es lo más adecuado para predecir la demanda de este producto como una función del precio por unidad.

Ejemplo 14.6 En el ejemplo 14.2 se consideró la regresión lineal de la temperatura aparente Y sobre la temperatura del aire x_1 y la humanidad relativa x_2 para un intervalo limitado de x_1 y x_2. Para el conjunto aumentado de datos dado en la tabla 14.17 se desea ajustar y analizar una ecuación de regresión completa de segundo orden.

LA ECUACION DE REGRESION ES
Y = 1330. − 155. X1 + 4.87 X2

	COLUMNA	COEFICIENTE	DEV. EST. DEL COEF.	COCIENTE-T = COEF/D.E.
	—	1330.4	179.6	7.41
X1	C2	−155.47	27.87	−5.58
X2	C3	4.866	1.031	4.72

LA DEV. EST. DE Y CON RESPECTO A LA RECTA DE REGRESION ES
S = 25.91
CON (11− 3) = 8 GRADOS DE LIBERTAD

R-CUADRADA = 93.6 POR CIENTO

ANALISIS DE VARIANZA

DEBIDO A	GL	SC	CM = SC/GL
REGRESION	2	78775.8	39387.9
RESIDUO	8	5368.8	671.1
TOTAL	10	84144.7	

ANALISIS DE VARIANZA ADICIONAL
SC EXPLICADA POR CADA VARIABLE QUE ENTRE EN EL ORDEN DADO

DEBIDO A	GL	SC
REGRESION	2	78775.8
C2	1	63814.5
C3	1	14961.4

FIGURA 14.9 Listado revisado para el ejemplo 14.5

TABLA 14.17 Datos de la muestra para el ejemplo 14.6

x_2 \ x_1	70°	75	80	85	90	95
0%	64	69	73	78	83	87
10	65	70	75	80	85	90
20	66	72	77	82	87	93
30	67	73	78	84	90	96
40	68	74	79	86	93	101
50	69	75	81	88	96	107
60	70	76	82	90	100	114
70	70	77	85	93	106	124
80	71	78	86	97	113	136

Con base en la experiencia cotidiana de cualquier persona con respecto al clima, debe ser evidente que la temperatura del aire y la humedad relativa tienen una interacción con la temperatura aparente. Por ejemplo, el calor que se siente cuando la temperatura del aire es de 90° y la humedad relativa es del 10%, es muy diferente a la que se percibe cuando la humedad relativa es del 70%. Los resultados que se muestran en la tabla 14.18 son los que se obtienen cuando se supone un modelo completo de segundo orden.

El efecto de cada término en el modelo sobre la temperatura aparente es estadísticamente discernible; los coeficientes de regresión se encuentran estimados con una exactitud razonablemente buena y el valor de R^2 es muy alto. De esta forma, la ecuación de regresión estimada completa de segundo orden es adecuada para la predicción.

14.8 Mínimos cuadrados con factores de peso

Una suposición clave en la estimación por mínimos cuadrados es que la varianza de cada error aleatorio es la misma. De la sección 14.6 recuérdese que si los residuos es-

TABLA 14.18 Análisis de regresión para el ejemplo 14.6

Regresión de Y sobre x_1, x_2, x_1^2, x_2^2, y x_1x_2

Variable	*Coeficiente de regresión estimado*	*Desviación estándar estimada*	*Valor t*
Constante	175.3300	36.11000	4.86
x_1	−3.1689	0.87580	−3.62
x_2	−1.4351	0.13210	−10.87
x_1^2	0.0236	0.00530	4.46
x_2^2	0.0017	0.00056	3.07
x_1x_2	0.0188	0.00150	12.56

$R^2 = 0.977$ $\qquad t_{0.975,\,48} = 2.01$

ANOVA

Fuente	gl	SC	CM	*Valor F*
Regresión	5	11,966.71	2393.34	407.20
Efecto lineal de x_1	1	8536.13	8536.13	1452.32
Efecto lineal de x_2	1	2330.71	2330.71	396.54
$x_1^2 \mid x_1, x_2$	1	116.68	116.68	19.85
$x_2^2 \mid x_1, x_2, x_1^2$	1	55.41	55.41	9.43
Interacción de x_1, x_2	1	927.78	927.78	157.85
Error	48	282.12	5.88	
Total	53	12,248.83	$f_{0.95,\,5,\,48} = 2.42$; $f_{0.95,\,1,\,48} = 4.04$	

tandarizados tienden a disminuir o a aumentar conforme se incrementan los valores estimados de la respuesta, la varianza del error no puede considerarse como constante. El remedio apropiado para esta situación es aplicar mínimos cuadrados con *factores de peso*, en los cuales las estimaciones para los coeficientes de regresión se obtienen mediante la minimización de la suma con pesos de los cuadrados de los errores. Si se empleara la estimación por mínimos cuadrados ordinarios en una situación para la cual la varianza del error no es constante, los coeficientes de regresión no serían estimados con la misma precisión.

Antes de resolver algunos ejemplos, se examinarán en forma breve los aspectos teóricos clave de la estimación por mínimos cuadrados con factores de peso. Al igual que en los mínimos cuadrados ordinarios se supone que para el modelo lineal general

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$\boldsymbol{\varepsilon}$ es un vector de errores aleatorios no observable, tal que

$$E(\boldsymbol{\varepsilon}) = \mathbf{0},$$

y la matriz de varianza-covarianza está dada por

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \mathbf{Q}.$$

La matriz $\mathbf{Q}$ es de tal naturaleza que el elemento que se encuentra sobre la diagonal q_{ii} es la varianza de ε_i, y q_{ij} es la covarianza entre ε_i y ε_j para toda $i \neq j$. $\mathbf{Q}$ debe ser no singular; de hecho, $\mathbf{Q}^{-1}$ recibe el nombre de matriz de ponderación y la debe especificar el investigador, es decir, los pesos se asignan a cada observación de la respuesta de acuerdo con alguna información respecto a la correspondiente varianza del error. Existen algunos procedimientos disponibles para los usuarios para determinar los pesos; lo anterior se ilustrará más adelante.

Las estimaciones de los coeficientes de regresión se obtienen mediante la minimización de la suma con pesos de los cuadrados de los errores dada por

$$\boldsymbol{\varepsilon}'\mathbf{Q}^{-1}\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{Q}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Puede demostrarse que las ecuaciones normales en forma matricial son

$$\mathbf{X}'\mathbf{Q}^{-1}\mathbf{X}\mathbf{B} = \mathbf{X}'\mathbf{Q}^{-1}\mathbf{Y}.$$

Si existe la matriz inversa $(\mathbf{X}'\mathbf{Q}^{-1}\mathbf{X})^{-1}$, los estimadores por mínimos cuadrados con factores de peso se obtienen mediante

$$\mathbf{B} = (\mathbf{X}'\mathbf{Q}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}^{-1}\mathbf{Y}. \tag{14.31}$$

Es importante notar que los mínimos cuadrados ordinarios son un caso especial de los mínimos cuadrados con factores de peso, es decir, si $\mathbf{Q} = \sigma^2\mathbf{I}$, entonces es relativamente fácil demostrar que (14.31) se reduce a la expresión usual

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

La definición de la matriz **Q** implica una estructura de covarianza entre los errores aleatorios. En la práctica, esta estructura resulta difícil de identificar. La aplicación más sencilla de la estimación por mínimos cuadrados con factores de peso es la de suponer que **Q** es una matriz diagonal de la forma

$$\mathbf{Q} = \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \mathbf{0} \\ & & \ddots & \\ & \mathbf{0} & & \sigma_n^2 \end{bmatrix},$$

donde σ_i^2 es la varianza de ε_i. Entonces

$$\mathbf{Q}^{-1} = \begin{bmatrix} 1/\sigma_1^2 & & & \\ & 1/\sigma_2^2 & & \mathbf{0} \\ & & \ddots & \\ & \mathbf{0} & & 1/\sigma_n^2 \end{bmatrix}.$$

Por lo tanto, los errores aleatorios se suponen independientes, pero algunas de sus varianzas (si no es que todas) pueden ser diferentes.

A continuación se examinarán algunas situaciones para las cuales es probable que se viole la suposición de que la varianza del error es constante si se emplean mínimos cuadrados ordinarios. Una práctica muy frecuente en la adquisición de datos experimentales es tomar varias mediciones de la respuesta para cada uno de los puntos de observación y después calcular el promedio de las mediciones para cada uno. La principal razón para llevar a cabo este procedimiento es estabilizar la variabilidad de las observaciones individuales. Bajo este procedimiento la respuesta se convierte en un promedio. Dado que la desviación estándar de un promedio es proporcional a la raíz cuadrada del tamaño de la muestra sobre la cual se basa este promedio, la variación de $\overline{Y}_i$, y de esta forma de ε_i, es σ^2/n_i, donde σ^2 es la varianza común del error y n_i es el tamaño de la muestra en relación con $\overline{Y}_i$. Esto conduce a un procedimiento de estimación por mínimos cuadrados con factores de peso para el cual la inversa de la matriz **Q** está dada por

$$\mathbf{Q}^{-1} = \frac{1}{\sigma^2} \begin{bmatrix} n_1 & & & \\ & n_2 & & \mathbf{0} \\ & & \ddots & \\ & \mathbf{0} & & n_n \end{bmatrix}.$$

Los pesos son los tamaños individuales de cada muestra $n_1, n_2, \ldots, n_n$ para los n puntos de observación. La lógica que se encuentra detrás de lo anterior es muy senci-

lla; los promedio basados en un gran número de observaciones deben tener un mayor peso en la determinación de las estimaciones que aquellas que se basan en pocas observaciones.

Ejemplo 14.7 Una compañía implanta un programa de inspección en el cual las unidades de cierto producto se revisan en forma visual en la línea de producción, con el objetivo de detectar las que se encuentran defectuosas. El gerente sabe que la velocidad de la línea afectará el número de unidades defectuosas encontradas. Se selecciona un lote de unidades de un tamaño suficiente y se envía a un departamento que se encargará de revisar el 100% de los mismos minuciosamente con el propósito de encontrar el número total de unidades defectuosas. Entonces el lote se coloca sobre la línea un número variable de veces para cada una de las ocho velocidades que posee la misma. Para cada velocidad de la línea x se calcula el número promedio $\overline{Y}$ de unidades defectuosas que no se descubrieron. Los datos que aparecen en la tabla 14.19 son los resultados de este experimento y la última columna representa los tamaños individuales de cada muestra. Obténganse y compárense las regresiones simples de $\overline{Y}$ sobre x, con base en mínimos cuadrados ordinarios y con factores de peso.

Para mínimos cuadrados ordinarios se descartan los tamaños variables de cada muestra y se procede de la manera usual. Los resultados se muestran en la tabla 14.20. Para mínimos cuadrados con factores de peso se ilustrará el cálculo de las estimaciones. Dado que los pesos son los tamaños de cada una de las muestras,

$$\mathbf{Q}^{-1} = \frac{1}{\sigma^2}\begin{bmatrix} 14 & & & \\ & 3 & & \mathbf{0} \\ & & \cdot & \\ & & & \cdot \\ & & & \cdot \\ \mathbf{0} & & & 2 \end{bmatrix},$$

TABLA 14.19 Datos de la muestra para el ejemplo 14.7

$\overline{Y}$	x(pie/min)	n
0.50	10	14
4.67	20	3
6.25	30	25
10.00	40	2
13.50	50	3
13.70	60	22
17.50	70	5
23.00	80	2

TABLA 14.20 Estimaciones por mínimos cuadrados ordinarios y tabla ANOVA para el ejemplo 14.7

Variable	*Coeficiente de regresión estimado*	*Desviación estándar estimada*	*Valor t*
Constante	−2.1190	0.9490	−2.23
x	0.2946	0.0188	15.68

$r^2 = 0.976$ $\quad t_{0.975,\,6} = 2.447$

ANOVA

Fuente	gl	SC	CM	*Valor F*
Regresión	1	364.62	364.62	246.36
Error	6	8.89	1.48	
Total	7	373.51	$f_{0.95,\,1,\,6} = 5.99$	

y

$$\mathbf{X'Q^{-1}X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 10 & 20 & \cdots & 80 \end{bmatrix} \frac{1}{\sigma^2} \begin{bmatrix} 14 & & & \\ & 3 & & \\ & & \ddots & \\ & & & 2 \end{bmatrix} \begin{bmatrix} 1 & 10 \\ 1 & 20 \\ \vdots & \vdots \\ 1 & 80 \end{bmatrix}$$

$$= \frac{1}{\sigma^2} \begin{bmatrix} 76 & 3010 \\ 3010 & 152\,300 \end{bmatrix}$$

Además,

$$(\mathbf{X'Q^{-1}X})^{-1} = \sigma^2 \begin{bmatrix} 0.06056388 & -0.00119696 \\ -0.00119696 & 0.00003022 \end{bmatrix}$$

Entonces, mediante el empleo de (14.31), las estimaciones de mínimos cuadrados con factores de peso son

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \sigma^2 \begin{bmatrix} 0.06056388 & -0.00119696 \\ -0.00119696 & 0.00003022 \end{bmatrix} \begin{bmatrix} 1 & 1 \cdots 1 \\ 10 & 20 \cdots 80 \end{bmatrix}$$

$$\frac{1}{\sigma^2}\begin{bmatrix} 14 & & & 0 \\ & 3 & & \\ & & \ddots & \\ 0 & & & 2 \end{bmatrix}\begin{bmatrix} 0.5 \\ 4.67 \\ \cdot \\ \cdot \\ \cdot \\ 23.0 \end{bmatrix} = \begin{bmatrix} -2.0540 \\ 0.2753 \end{bmatrix}.$$

Los resultados del análisis de regresión con factores de peso se dan en la tabla 14.21.

Una comparación de los resultados basados en mínimos cuadrados con factores de peso y ordinarios muestran una ligera mejoría en la precisión de las estimaciones de mínimos cuadrados con factores de peso, así como un pequeño aumento en el valor de r^2. Debe notarse que si los tamaños de cada muestra individual no difieren mucho entre sí, es muy probable que los resultados que se obtengan mediante el empleo de los dos métodos sean muy similares.

En el ejemplo 14.7 se supuso de antemano una varianza del error no constante, debido a que el registro de las observaciones de la respuesta son promedios basados en tamaños variables de las muestras. Sin embargo, la mayoría de las veces la falta de una varianza constante para el error debe determinarse en forma empírica. Ya se ha indicado que una gráfica de los residuos estandarizados contra las respuestas estimadas resulta ser una ayuda considerable en la detección de la heterocedasticidad, pero para aquellos problemas para los cuales se tienen disponibles observaciones repetidas de la respuesta para el mismo punto de observación, es preferible registrar las observaciones reales más que sus promedios y emplearlas para detectar una varianza

TABLA 14.21 Estimaciones por mínimos cuadrados con factores de peso y tabla ANOVA para el ejemplo 14.7

Variable	*Coeficiente de regresión estimado*	*Desviación estándar estimada*	*Valor t*
Constante	−2.0540	0.6990	−2.94
x	0.2753	0.0156	17.63

$r^2 = 0.981$ $\qquad t_{0.975,\,6} = 2.447$

ANOVA

Fuente	gl	SC	CM	*Valor F*
Regresión	1	2508.66	2508.66	310.86
Error	6	48.43	8.07	
Total	7	2557.09*	$f_{0.95,\,1,\,6} = 5.99$	

* Las sumas de los cuadrados son diferentes de las anteriores debido a que ahora son funciones de los pesos impuestos.

no constante del error. Para estos tipos de problemas la gráfica de las observaciones contra los valores de la variable de predicción revelará una varianza no constante, si es que ésta existe. El siguiente ejemplo ilustra este problema.

Ejemplo 14.8 Recientemente, la variabilidad del ozono en la estratósfera ha recibido gran atención, especialmente en el impacto que el hombre tiene sobre el clima. El ozono es una forma de oxígeno que se encuentra en diversas cantidades en la estratósfera y constituye un componente muy importante de la atmósfera, ya que tiene la propiedad de bloquear la radiación ultravioleta que provienen del sol. Los datos que se encuentran en la tabla 14.22 muestran la cantidad de ozono registrada Y y su presión parcial x para cada capa de altitud, donde cada capa tiene aproximadamente un kilómetro de altura. Por conveniencia, las capas se han escalado a un intervalo de -7 a $+7$. Determínese si la varianza del error puede considerarse como constante.

TABLA 14.22 Datos de la muestra para el ejemplo 14.8

Capa	*Ozono*	*Capa*	*Ozono*
−7.00	53.8	−1.00	102.8
−7.00	53.3	−1.00	96.9
−7.00	54.8	−1.00	98.2
−7.00	54.6	0.0	98.9
−7.00	53.7	0.0	96.1
−7.00	55.2	0.0	99.6
−7.00	55.7	0.0	91.4
−7.00	54.1	1.00	101.1
−6.00	63.8	1.00	94.6
−6.00	64.2	1.00	95.9
−6.00	66.9	2.00	92.3
−6.00	67.2	2.00	96.6
−6.00	65.4	2.00	98.5
−6.00	67.3	3.00	93.6
−5.00	71.8	3.00	86.2
−5.00	73.2	3.00	87.9
−5.00	75.6	3.00	89.5
−5.00	76.2	4.00	74.8
−5.00	72.7	4.00	82.3
−4.00	79.4	4.00	76.9
−4.00	81.1	4.00	81.2
−4.00	85.2	5.00	73.6
−4.00	83.0	5.00	65.4
−4.00	84.1	5.00	67.1
−4.00	82.8	6.00	60.2
−3.00	90.3	6.00	54.9
−3.00	84.2	6.00	50.8
−3.00	88.3	7.00	44.7
−3.00	86.0	7.00	38.5
−2.00	93.2		
−2.00	97.4		
−2.00	98.3		

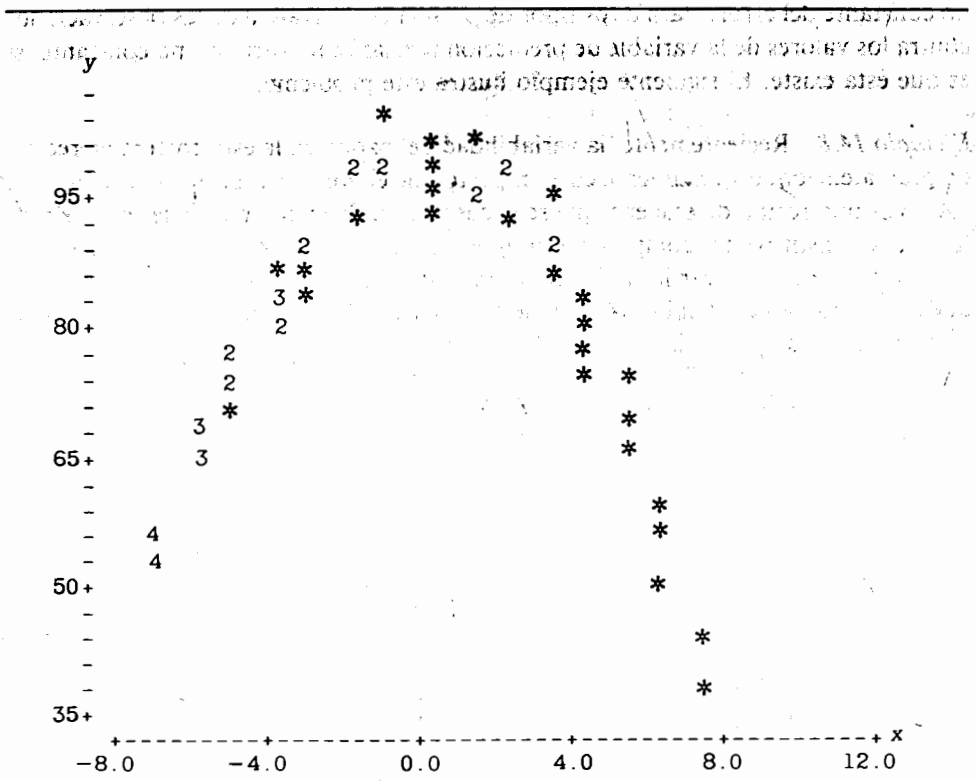


FIGURA 14.10 Gráfica del ozono contra la altitud de la capa para el ejemplo 14.8

Una gráfica de la cantidad de ozono contra la capa, figura 14.10, revela que la varianza del error no puede considerarse como constante debido a que la variabilidad en la cantidad de ozono aumenta conforme la capa crece. La figura 14.10 también sugiere que el modelo apropiado por utilizar es una ecuación cuadrática.

En una situación como ésta, en la que existen repeticiones para varios puntos de observación, los pesos se determinan mediante el cálculo de la varianza de las mediciones de la respuesta para cada punto de observación. De esta forma, cada peso es el recíproco de la correspondiente varianza. Por ejemplo, si y_{ij} denota la i-ésima medición de ozono en la j-ésima capa, la varianza de la muestra de la j-ésima capa es

$$s_j^2 = \sum_{j=1}^{n_j} (y_{ij} - \bar{y}_{\cdot j})^2/(n_j - 1),$$

y el correspondiente peso es $w_j = 1/s_j^2$. Como ilustración, considérense las observaciones para $x = 0$. Éstas son 98.9, 99.6 y 91.4. Entonces, $n_j = 4$, $\bar{y}_{\cdot j} = 96.5$, $s_j^2 = 13.8467$, y $w_j = 1/13.8467 = 0.0722$. Al seguir este procedimiento, los pesos correspondientes para cada capa son los que se muestran en la tabla 14.23.

TABLA 14.23 Pesos para el ejemplo 14.8

Capa	*Peso*	*Capa*	*Peso*
−7	1.4956	1	0.0845
−6	0.4119	2	0.0991
−5	0.2755	3	0.0997
−4	0.2304	4	0.0797
−3	0.1411	5	0.0534
−2	0.1349	6	0.0450
−1	0.1041	7	0.0520
0	0.0722		

Mediante el uso de estos pesos y al ajustar un modelo cuadrático, se obtienen, para el ozono, los resultados que se encuentran en la tabla 14.24. Es evidente que una ecuación cuadrática de regresión basada en mínimos cuadrados con factores de peso es muy adecuada para describir la variabilidad de la cantidad promedio de ozono como una función de la altitud.

La mayoría de las veces no existen observaciones repetidas, pero los datos se recaban en agrupaciones naturales las que pueden, *a priori*, sugerir varianzas diferentes para el error en cada grupo. Lo que en general se hace es suponer que la varianza del j-ésimo grupo es $c_j^2\sigma^2$, donde c_j es única para el j-ésimo grupo, pero σ^2 es común para todos los grupos. En general, los valores de las c_j no son conocidos, pero pueden estimarse primero al determinar la varianza residual para cada grupo, s_j^2 basada en

TABLA 14.24 Estimaciones por mínimos cuadrados con factores de peso y tabla ANOVA para el ejemplo 14.8

Variable	*Coeficiente de regresión estimado*	*Desviación estándar estimada*	*Valor t*
Constante	96.7590	0.6367	151.98
x	−0.5585	0.1266	−4.41
x^2	−0.9495	0.0238	−39.83

$R^2 = 0.9817$ $\quad t_{0.975,\,58} = 2.00$

ANOVA

Fuente	gl	SC	CM	*Valor F*
Regresión	2	4082.33	2041.17	1556.30
Efecto lineal	1	2001.11	2001.11	1525.78
Efecto cuadrático	1	2081.22	2081.22	1586.82
Error	58	76.07	1.31	
Total	60	4158.40	$f_{0.95,\,2,\,58} = 3.15$; $f_{0.95,\,1,\,58} = 4.00$	

los residuos de éstos. Los residuos se obtienen mediante el ajuste de un modelo lineal general empleando mínimos cuadrados ordinarios. Entonces, una estimación de c_j es s_j/s, donde s es la desviación estándar residual global basada en los mínimos cuadrados ordinarios y s_j es la desviación estándar residual para el j-ésimo grupo. Entonces el peso para el j-ésimo grupo es $w_j = 1/c_j^2 = s^2/s_j^2$.

14.9 Variables indicadoras

En casi todos los problemas que se han considerado hasta este momento, las variables de predicción han sido cuantitativas en el sentido en que toman valores de una escala numérica bien definida. Sin embargo, para muchas variables como la localización geográfica, el estado civil, las poblaciones urbanas o rurales o alguna otra, no es evidente tener una escala bien definida. Dado que estas variables cualitativas son factores importantes en muchas situaciones, a continuación se examinará una manera de cuantificar los niveles de una variable de predicción cualitativa para su empleo en el análisis de regresión. Se considerarán las que comúnmente se conocen como variables indicadores o mudas. A cada una de estas variables se le asignan los valores 0 y 1.

Como ilustración, considérese la tasa de crímenes para dos estados adyacentes, para los que los datos figuran en el ejercicio 14.16 que se encuentra al final de este capítulo. En particular, supóngase que se desea hacer una regresión de la tasa de crímenes sobre el porcentaje de la población urbana en un estado para aquellos que se encuentren sólo en las regiones 1 y 5. El modelo de regresión será una función de la variable cuantitativa x_1 (porcentaje de población urbana) y una variable de predicción cualitativa que representa las dos regiones de interés.

Dado que sólo se tienen dos regiones, es conveniente definir dos variables indicadoras x_2 y x_3 tales, que

$$x_2 = \begin{cases} 1 & \text{si un estado es encuentra en la región 1,} \\ 0 & \text{de otro modo,} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{si un estado se encuentra en la región 5,} \\ 0 & \text{de otro modo.} \end{cases}$$

Entonces, para obtener una sola ecuación de regresión para ambas regiones, se deberá ajustar el modelo

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.$$

Pero si se hace esto, entonces la matriz $\mathbf{X}'\mathbf{X}$ no tendría inversa. Una manera fácil de salir de este problema es eliminar una de las dos variables indicadoras y emplear solamente una, por ejemplo x_2, en donde al igual que antes,

$$x_2 = \begin{cases} 1 & \text{si un estado se encuentra en la región 1,} \\ 0 & \text{de otro modo,} \end{cases}$$

En otra palabras, para cualquier estado que se encuentre en la región 1 ($x_2 = 1$) o si se encuentra en la región 5 ($x_2 = 0$). En general, si una variable cualitativa tiene m niveles, puede representarse por medio de $m - 1$ variables indicadoras, asignando a cada una los valores de 0 y 1.

Considérese de nuevo el problema de la tasa de crímenes para las regiones 1 y 5. Existen varias maneras de abordar el desarrollo de un modelo de regresión. Se puede reunir la información proveniente de ambas regiones y entonces ajustar el modelo lineal simple

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon,$$

ignorando las diferencias regionales pueden obtenerse ecuaciones de regresión separadas para las regiones, cada una con diferentes estimaciones para los coeficientes de regresión. La elección entre estas dos opciones debe hacerse con mucho cuidado. En realidad debe decidirse si cada una de estas dos regiones es distinta con respecto a la tasa de crímenes, o si existe alguna relación en común. Si lo primero es cierto y se ajusta el modelo dado con anterioridad, entonces es probable que la tasa de crímenes en una región se encuentre sobreestimada mientras que para la otra ocurre lo contrario. Si existe una relación en común, entonces no es necesario tener dos ecuaciones de regresión separadas.

Al comparar los resultados que se obtienen con base en las ecuaciones de regresión separadas y la única para las regiones 1 y 5 mediante el empleo del porcentaje de población urbana como la única variable de predicción, se obtienen los datos que se encuentran en la tabla 14.25.

La comparación revela que las estimaciones para cada una de las pendientes son, en esencia, las mismas, pero las estimaciones de las intersecciones son significativamente diferentes. Nótese también que la ecuación de regresión única exhibe las propiedades menos deseables. De hecho, con esta última ecuación las tasas para los estados que se encuentran en las región 1 se sobreestiman, mientras que para los estados que se encuentran en la región 5 se subestiman con una sola excepción. Por lo tanto, en forma aparente existen diferencias regionales para la respuesta y no deben ignorarse.

Para incorporar las diferencias regionales dentro del modelo, sólo se utilizará la variable indicadora x_2 definida con anterioridad; el modelo se convierte en

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon. \tag{14.32}$$

Para interpretar los coeficientes de regresión, considérense los estados de la región 5. Dado que para éstos $x_2 = 0$, se supone una curva de regresión dada por

$$E(Y) = \beta_0 + \beta_1 x_1,$$

que es la ecuación de una línea recta con pendiente β_1 e intersección β_0. Para los estados que se encuentran en la región 1, $x_2 = 1$, y la respuesta media toma la forma

$$\begin{aligned} E(Y) &= \beta_0 + \beta_1 x_1 + \beta_2 \\ &= (\beta_0 + \beta_2) + \beta_1 x_1, \end{aligned}$$

TABLA 14.25 Modelos de regresión combinado y separado para el ejemplo de la tasa de crímenes

Modelo de regresión estimado

Variable	*Coeficiente de regresión estimado*	*Desviación estándar estimada*	*Valor t*
Constante	7.0350	4.2300	1.66
x_1	−0.0094	0.0673	−0.14
$n = 12$	$r^2 = 0.002$	$t_{0.975,\,10} = 2.228$	

Modelo de regresión para la región 1

Variable	*Coeficiente de regresión estimado*	*Desviación estándar estimada*	*Valor t*
Constante	0.4170	0.8020	0.52
x_1	0.0404	0.0118	3.41
$n = 6$	$r^2 = 0.745$	$t_{0.975,\,4} = 2.776$	

Modelo de regresión para la región 5

Variable	*Coeficiente de regresión estimado*	*Desviación estándar estimada*	*Valor t*
Constante	7.4400	3.9500	1.88
x_1	0.0439	0.0686	0.64
$n = 6$	$r^2 = 0.093$	$t_{0.975,\,4} = 2.776$	

la que también es la ecuación de una línea recta con la misma pendiente, β_1, pero con una intersección diferente $\beta_0 + \beta_2$. Entonces el modelo dado por (14.32) proporciona la respuesta promedio como una función lineal de x_1 con la misma pendiente para ambas regiones, pero con diferentes intersecciones. El parámetro β_2 representa el efecto diferencial que existe entre las intersecciones de las dos regiones. Para ajustar el modelo (14.32) el vector $\boldsymbol{Y}$ y la matriz $\boldsymbol{X}$ son

$$\mathbf{Y} = \begin{bmatrix} 4.2 \\ 2.4 \\ 3.1 \\ 3.2 \\ 3.9 \\ 1.4 \\ 10.2 \\ 11.7 \\ 10.6 \\ 11.9 \\ 9.0 \\ 6.0 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 77.6 & 1 \\ 1 & 50.8 & 1 \\ 1 & 84.6 & 1 \\ 1 & 56.4 & 1 \\ 1 & 87.1 & 1 \\ 1 & 32.2 & 1 \\ 1 & 80.5 & 0 \\ 1 & 60.3 & 0 \\ 1 & 45.0 & 0 \\ 1 & 47.6 & 0 \\ 1 & 63.1 & 0 \\ 1 & 39.0 & 0 \end{bmatrix}$$

Los resultados de la regresión se muestran en la tabla 14.26

TABLA 14.26 Análisis de regresión para el ejemplo de la tasa de crímenes

Variable	*Coeficiente de regresión estimado*	*Desviación estándar estimada*	*Valor t*
Constante	7.5800	1.6400	4.62
x_1	0.0416	0.0269	1.54
x_2	−7.2340	0.9520	−7.60
$n = 12$	$R^2 = 0.865$	$t_{0.975,\,9} = 2.262$	

Con base en estos resultados, las diferencias regionales son estadísticamente significativas. De esta forma, la última ecuación de regresión es superior con respecto al modelo único en el cual no se consideraban las diferencias regionales. En particular, las dos regiones tienen la misma estimación para la pendiente (0.0416), pero las intersecciones son iguales a 7.58 para la región 5 y 7.58 − 7.23 = 0.35 para la región 1. En general puede suponerse que la pendiente es la misma y, por lo tanto, es mejor emplear un modelo con una variable indicadora que un modelo único. Además, también es mejor un modelo con una variable indicadora que emplear dos modelos de regresión separados debido a que para el primero se tiene un mayor número de grados de libertad disponible para el error que para el segundo. De acuerdo con lo anterior, β_0 y β_2 son las estimaciones con la mejor precisión como es el caso para este ejemplo.

¿Qué ocurre si la pendiente no es la misma? Esta situación puede manejarse mediante la introducción en el modelo de un término de interacción para la variable cuantitativa x_1 y para la variable indicadora x_2. El modelo propuesto se convierte en

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon. \tag{14.33}$$

Para los estados que se encuentran en la región 5, $x_2 = 0$. Entonces, $x_1 x_2 = 0$, y la respuesta promedio para esta región es

$$E(Y) = \beta_0 + \beta_1 x_1.$$

Para los estados que se encuentran en la región 1, $x_2 = 1$, y $x_1 x_2 = x_1$, la respuesta media para esta región es

$$\begin{aligned} E(Y) &= \beta_0 + \beta_1 x_1 + \beta_2 + \beta_{12} x_1 \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) x_1. \end{aligned}$$

Nótese que el coeficiente de regresión de x_2 es el efecto diferencial que existe entre las intersecciones de las dos regiones y el coeficiente de regresión del producto cruzado $x_1 x_2$ es el efecto diferencial entre las pendientes de las dos regiones. Por lo tanto, suponiendo que existe una interacción estadísticamente apreciable entre x_1 y x_2, pueden obtenerse las ecuaciones estimadas de regresión para cada región mediante el empleo del modelo dado por (14.33).

Para finalizar, se examinará el problema en el cual una variable cualitativa tiene más de dos niveles. Este caso requiere del uso de más de una variable indicadora en

el modelo de regresión. Como ilustración, se continuará con el problema de la tasa de crímenes al llevar a cabo una comparación entre los estados de las regiones 1, 5 y 7. Dado que se han identificado tres niveles de una variable cualitativa, se definirán dos variables indicadoras de la siguiente manera:

$$x_2 = \begin{cases} 1 & \text{si un estado se encuentra en la región 1,} \\ 0 & \text{de otro modo,} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{si un estado se encuentra en la región 5,} \\ 0 & \text{de otro modo.} \end{cases}$$

Este arreglo proporciona el mismo número de combinaciones posible de los valores de x_2 y x_3 de acuerdo con el número de niveles de la variable cualitativa. Estos son $x_2 = 1, x_3 = 0$; $x_2 = 0, x_3 = 1$; y $x_2 = x_3 = 0$. Representan los estados en las regiones 1, 5 y 7 respectivamente.

Si se supone que las pendientes son iguales para las tres regiones, el modelo es

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.$$

Para los estados que se encuentran en la región 7, $x_2 = 0$ y $x_3 = 0$, de tal manera que la respuesta se reduce a

$$E(Y) = \beta_0 + \beta_1 x_1,$$

que es la ecuación de una línea recta con pendiente β_1 e intersección β_0. Para los estados que se encuentran en la región 5, $x_2 = 0$ y $x_3 = 1$. De acuerdo con lo anterior, la curva de regresión es

$$\begin{aligned} E(Y) &= \beta_0 + \beta_1 x_1 + \beta_3 \\ &= (\beta_0 + \beta_3) + \beta_1 x_1, \end{aligned}$$

donde β_3 representa el cambio en la intersección de la región 5 con respecto al de la región 7. De manera similar, cuando $x_2 = 1$ y $x_3 = 0$, la respuesta media es

$$\begin{aligned} E(Y) &= \beta_0 + \beta_1 x_1 + \beta_2 \\ &= (\beta_0 + \beta_2) + \beta_1 x_1, \end{aligned}$$

donde ahora β_2 es el cambio en la intersección de la región 1 con respecto al de la región 7. Se deduce que tanto β_2 como β_3 representan los efectos diferenciales para las intersecciones de las regiones 1 y 5, respectivamente, en relación con la intersección de la región 7.

El caso para el cual no es posible suponer que las pendientes son iguales, en este momento debe ser ya evidente, es decir, si se asume un modelo de la forma

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \varepsilon, \qquad (14.34)$$

donde β_{12} y β_{13} son los coeficientes de regresión para las interacciones que

comprenden a la variable cuantitativa x_1 y a cada una de las dos variables indicadoras x_2 y x_3.

Ejemplo 14.9 Se seleccionan al azar cinco casas recientemente vendidas para tres distintas zonas residenciales (A, B y C) en cierta ciudad, y el precio de venta Y se compara con el valor catastral de la propiedad x_1 determinado por la oficina estatal local correspondiente. Los datos se encuentran en la tabla 14.27 donde el precio de venta y el valor catastral de la propiedad se dan en miles de dólares. Mediante el empleo de variables indicadoras, ajústese una ecuación de regresión lineal y determínese si las pendientes para las tres zonas residenciales son las mismas.

Dado que se tienen tres zonas residenciales, se definen dos variables indicadoras x_2 y x_3 tales, que

$$x_2 = \begin{cases} 1 & \text{si una casa se encuentra en la zona B,} \\ 0 & \text{de otro modo,} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{si una casa se encuentra en la zona C,} \\ 0 & \text{de otro modo.} \end{cases}$$

Para el modelo (14.34) el vector Y y la matriz X son iguales a

$$\mathbf{Y} = \begin{bmatrix} 42.5 \\ 36.8 \\ 42.6 \\ 41.2 \\ 48.6 \\ 75.2 \\ 83.4 \\ 83.3 \\ 116.8 \\ 114.3 \\ 122.8 \\ 125.6 \\ 132.5 \\ 127.4 \\ 147.8 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 33.1 & 0 & 0 & 0 & 0 \\ 1 & 42.0 & 0 & 0 & 0 & 0 \\ 1 & 47.8 & 0 & 0 & 0 & 0 \\ 1 & 53.4 & 0 & 0 & 0 & 0 \\ 1 & 59.6 & 0 & 0 & 0 & 0 \\ 1 & 63.9 & 1 & 0 & 63.9 & 0 \\ 1 & 68.4 & 1 & 0 & 68.4 & 0 \\ 1 & 72.3 & 1 & 0 & 72.3 & 0 \\ 1 & 77.8 & 1 & 0 & 77.8 & 0 \\ 1 & 80.8 & 1 & 0 & 80.8 & 0 \\ 1 & 96.5 & 0 & 1 & 0 & 96.5 \\ 1 & 101.8 & 0 & 1 & 0 & 101.8 \\ 1 & 106.2 & 0 & 1 & 0 & 106.2 \\ 1 & 112.6 & 0 & 1 & 0 & 112.6 \\ 1 & 120.5 & 0 & 1 & 0 & 120.5 \end{bmatrix}$$

TABLA 14.27 Datos de la muestra para el ejemplo 14.9

Zona A		*Zona* B		*Zona* C	
x	Y	x	Y	x	Y
33.1	42.5	63.9	75.2	96.5	122.8
42.0	36.8	68.4	83.4	101.8	125.6
47.8	42.6	72.3	83.3	106.2	132.5
53.4	41.2	77.8	116.8	112.6	127.4
59.6	48.6	80.8	114.3	120.5	147.8

El listado de computadora que produce Minitab se encuentra en la figura 14.11, donde C2-C6 se refieren a x_1, x_2, x_3, x_1x_2, y x_1x_3, respectivamente.

Nótese que la hipótesis nula

$$H_0: \beta_{12} = 0$$

puede rechazarse, pero la hipótesis

$$H_0: \beta_{13} = 0$$

no; por lo tanto, existe una razón para creer que las pendientes para las zonas residenciales A y B no son las mismas. Del listado se determina que las ecuaciones esti-

---

LA ECUACION DE REGRESION ES

Y = 31.4 + 0.232 X1 − 129.X2
+ 1.89 X3 + 2.41 X4 + 0.679 X5

	COLUMNA	COEFICIENTE	DEV. EST. DEL COEF.	COCIENTE-T = COEF/D.E.
	—	31.37	14.66	2.14
X1	C2	0.2325	0.3050	0.76
X2	C3	−128.81	36.29	−3.55
X3	C4	1.89	38.82	0.05
X4	C5	2.4112	0.5481	4.40
X5	C6	0.6786	0.4518	1.50

LA DEV. EST. DE Y CON RESPECTO A LA RECTA DE REGRESION ES
S = 6.238
CON (15 − 6) = 9 GRADOS DE LIBERTAD

R-CUADRADA = 98.4 POR CIENTO

ANALISIS DE VARIANZA DE

DEBIDO A	GL	SC	CM = SC/GL
REGRESION N	5	21577.96	4315.59
RESIDUO	9	350.26	38.92
TOTAL	14	21928.22	

ANALISIS DE VARIANZA ADICIONAL
SC EXPLICADA POR CADA VARIABLE QUE ENTRA EN EL ORDEN DADO

DEBIDO A	GL	SC
REGRESION	5	21577.96
C2	1	19892.61
C3	1	698.16
C4	1	232.89
C5	1	666.52
C6	1	87.79

FIGURA 14.11 Listado de computadora para el ejemplo 14.9

madas de regresión para cada zona residencial son las siguientes:

$$\text{Zona A:}\quad \hat{y} = 31.37 + 0.2325x_1, \quad (x_2 = x_3 = 0)$$

$$\text{Zona B:}\quad \hat{y} = -97.44 + 2.6437x_1, \quad (x_2 = 1, x_3 = 0)$$

$$\text{Zona C:}\quad \hat{y} = 33.26 + 0.9111x_1. \quad (x_2 = 0, x_3 = 1)$$

Referencias

1. S. Chatterjee and B. Price, *Regression analysis by example*, Wiley, New York, 1977.
2. C. Daniel and F. S. Wood, *Fitting equations to data*, Wiley, New York, 1971.
3. N. R. Draper and H. Smith, *Applied regression analysis*, 2nd ed., Wiley, New York, 1981.
4. F. A. Graybill, *Theory and application of the linear model*, Duxbury, North Scituate, Mass., 1976.
5. J. Neter, W. Wasserman, and M. H. Kutner, *Applied linear regression models*, Richard D. Irwin, Homewood, Ill., 1983.
6. *SAS users guide*, SAS Institute, Raleigh, N. C., 1982.

Ejercicios

14.1. De los siguientes modelos, ¿cuáles no son casos del modelo lineal general y por que?

a) $Y = \beta_0 + \beta_1 \exp(\beta_2 x_1) + \beta_3 x_2 + \varepsilon$

b) $Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1^2 x_2 + \varepsilon$

c) $Y = \beta_0 + \beta_1 / x_1 + \beta_2 x_2^{1/2} + \varepsilon$

14.2. Una agencia de alquiler de automóviles obtiene la siguiente ecuación de regresión:

$$\hat{y} = 0.75 + 1.2x_1 + 0.15x_2$$

para predecir el costo promedio anual Y en miles de dólares como una función del número de automóviles alquilados x_1 y del número promedio de millas que cada automóvil recorre x_2 en miles de millas. Explíquese el significado de cada uno de los coeficientes estimados de la regresión.

14.3. Supóngase que la ecuación estimada de regresión que describe una respuesta media como una función de dos variables de predicción está dada por

$$\hat{y} = 15 + 6x_1 - 2x_2 - 1.5x_1x_2.$$

a) ¿Cúal es el efecto sobre la respuesta media por unidad de cambio en x_1 cuando $x_2 = 2$?

b) ¿Cuál es el efecto sobre la respuesta media por unidad de cambio en x_2 cuando $x_1 = 1$?

14.4. Mediante el empleo de los datos de Prater, ajústense todos los modelos lineales que incluyan sólo a x_1 y a x_3, e ilústrese el principio de la suma de cuadrados extra mediante el cálculo de lo siguiente:

a) Las tablas de análisis de varianza correspondientes.
b) $SCR(x_3 \mid x_1)$ y $SCR(x_1 \mid x_3)$.
c) Las pruebas F parciales apropiadas.

14.5. Una agencia desea estimar los gastos en alimentación de una familia con base en el ingreso y su tamaño. Los datos que se encuentran en la tabla 14.28 representan los gastos de alimentación por mes Y en miles de dólares, contra el ingreso mensual x_1 y el tamaño de la familia x_2 para 15 familias que se seleccionaron al azar en cierta localidad geográfica.

TABLA 14.28 Datos de la muestra para el ejercicio 14.5

Y	x_1	x_2
0.43	2.1	3
0.31	1.1	4
0.32	0.9	5
0.46	1.6	4
1.25	6.2	4
0.44	2.3	3
0.52	1.8	6
0.29	1.0	5
1.29	8.9	3
0.35	2.4	2
0.35	1.2	4
0.78	4.7	3
0.43	3.5	2
0.47	2.9	3
0.38	1.4	4

a) Ajústense todos los modelos lineales que abarcan a x_1 y/o x_2, e interprétense los coeficientes de regresión estimados.
b) Pruébese la hipótesis nula H_0: $\beta_1 = \beta_2 = 0$.
c) Calcúlese $SCR(x_2 \mid x_1)$ y $SCR(x_1 \mid x_2)$ y llévense a cabo las pruebas F parciales apropiadas.
d) Calcúlese e interprétese el coeficiente de correlación múltiple R^2.
e) Con base en los resultados anteriores, decídase cuál es la mejor ecuación para predecir el gasto de alimentación y empléese para estimar el gasto promedio mensual en alimentación para una familia de cuatro personas con un ingreso mensual de $2 500. Determínese un intervalo de confianza del 98% para esta cantidad.

14.6. Con respecto al ejercicio 14.5 hágase lo siguiente:

a) Para la regresión que comprende, tanto a x_1 como a x_2, efectúense las pruebas individuales t para los coeficientes de regresión β_1 y β_2. Úsese $\alpha = 0.05$.
b) Determínense intervalos de confianza de 95% para β_1 y β_2 y formúlense las conclusiones apropiadas.

14.7. Mediante el uso de los datos del ejercicio 14.5, constrúyase un modelo lineal general que abarque tanto a x_1 como a x_2 en forma matricial; identifíquense todas las matrices y obténganse las ecuaciones normales.

14.8. En muchas agencias gubernamentales y compañías privadas el problema de identificar aquellos factores que son importantes para predecir la aptitud para el trabajo de los aspirantes a obtener un ejemplo constituye un proceso continuo. El procedimiento usual es el de aplicar al solicitante un conjunto de pruebas apropiadas y tomar las decisión de contratarlo o no con base en los resultados de éstas. El asunto clave es conocer *a priori* qué pruebas pueden predecir la aptitud para el trabajo de una persona. Supóngase que el personal de una compañía muy grande ha desarrollado cuatro pruebas para una determinada clasificación con respecto al trabajo. Estas pruebas se aplicaron a 20 individuos que fueron contratados por la compañía. Después de un periodo de dos años, cada uno de estos empleados se clasifica de acuerdo con su aptitud para el trabajo. La puntuación para la aptitud hacia el trabajo Y y la correspondiente a cada una de las cuatro pruebas x_1, x_2, x_3, x_4 se dan en la tabla 14.29.

TABLA 14.29 Datos de la muestra para el ejercicio 14.8

Empleado	Y	x_1	x_2	x_3	x_4
1	94	122	121	96	89
2	71	108	115	98	78
3	82	120	115	95	90
4	76	118	117	93	95
5	111	113	102	109	109
6	64	112	96	90	88
7	109	109	129	102	108
8	104	112	119	106	105
9	80	115	101	95	88
10	73	111	95	95	84
11	127	119	118	107	110
12	88	112	110	100	87
13	99	120	89	105	97
14	80	117	108	99	100
15	99	109	125	108	95
16	116	116	122	116	102
17	100	104	83	100	102
18	96	110	101	103	103
19	126	117	120	113	108
20	58	120	77	80	74

a) Utilícese la rutina *PROC GLM* de *SAS* (o algún otro paquete comparable) para ajustar la regresión lineal de Y sobre x_1, x_2, x_3 y x_4.
b) Con base en el listado de la computadora que se obtiene en la parte *a*, prepárese una tabla de análisis de varianza mostrando todas las posibles pruebas F parciales.
c) Interprétense los coeficientes de regresión estimados y el coeficiente de correlación múltiple.

14.9. Empléense los datos del ejercicio 14.8 para hacer lo siguiente:

a) Obténganse todas las posibles ecuaciones de regresión, y para cada una cálculese la

suma de los cuadrados de los errores, el cuadrado medio del error, el valor de C_p y el valor R^2 (véase el ejercicio 14.12).

b) Demuéstrese que la regresión por pasos y el procedimiento de eliminación hacia atrás proporcionan los mismos resultados para la mejor ecuación de predicción.

c) Con base en los resultados anteriores, dedúzcase la mejor ecuación de predicción y empléese para estimar la aptitud para el trabajo de un individuo que tiene las siguientes puntuaciones, en las pruebas: $x_1 = 105$, $x_2 = 110$, $x_3 = 99$, y $x_4 = 107$. Obténgase un intervalo de predicción del 95% para esta cantidad.

14.10. De manera reciente, se ha dirigido el interés hacia el desarrollo de métodos más rápidos y económicos para vigilar la concentración de sedimentos y contaminantes en los recursos acuíferos de cierta nación. Para los encargados de vigilar el medio ambiente, el interés principal recae en la necesidad de cuantificar los valores de concentración en el agua con base en datos de percepción remota. El uso de las técnicas de percepción remota para vigilar distintos parámetros que miden la calidad del agua parece ser prometedor. Un tipo de sistema de percepción remota es la variedad pasiva el cual depende, en forma única, de la radiación de sol como fuente de energía y mide el flujo total de radiación emitido por el sistema agua-atmósfera. Una componente muy grande de este flujo de radiación es el flujo de luz emitido por el agua, el cual, bajo condiciones normales, es una función de los constituyentes que se encuentran presentes en el agua. Para medir el espectro de esta radiación se encuentran disponibles un gran número de sistemas de rastreo multiespectral. Sin embargo, cada sistema tiene diferentes localizaciones de las bandas y anchos diferentes.

Se piensa que un cambio en la concentración de un contaminante causará un cambio en el valor del flujo de radiaciones, es decir, si se conocen los valores de la radiación para diferentes bandas espectrales, entonces es posibles predecir la concentración de un contaminante en una fuente de agua dada. El problema reside en el hecho de identificar, de entre todas las bandas, cuál es la que puede predecir la concentración del contaminante. En una tesis doctoral reciente, Whitlock* obtuvo datos reales de percepción remota proporcionados por un laboratorio, bajo condiciones controladas, que empleó cinco bandas y varios constituyentes, entre ellos el sedimento del feldespato. Los datos de la muestra se proporcionan en la tabla 14.30.

a) Empléense las cinco bandas como variables de predicción y las concentraciones de feldespato como la respuesta, para ajustar un modelo de regresión lineal.

b) Calcúlese la matriz de correlación para las cinco bandas de radiación. Interprétese el resultado.

c) Úsese la regresión por pasos y el procedimiento de eliminación hacia atrás, para determinar el mejor conjunto de variables de predicción. ¿Son los resultados idénticos?

d) Con base en los resultados anteriores, analícese cualquier aspecto que se considera evidente con respecto a este problema y que sirva para decidir por una ecuación de predicción adecuada.

14.11. En la sección 14.5 se mencionó que el valor de R^2 aumenta conforme se añaden más términos a la ecuación de regresión debido a que *SCE* siempre disminuye al sumar más términos y *STC* siempre permanece constante. Es por esta razón que se sugiere una medida alternativa que tome en cuenta el número de términos que figuran en el modelo.

* Charles H. Whitlock, tesis doctoral, Universidad Old Dominion, mayo, 1977.

TABLA 14.30 Datos de la muestra para el ejercicio 14.10

Concentración Y de feldespato	*Bandas de radiación*				
	x_1	x_2	x_3	x_4	x_5
17	0.297	0.310	0.290	0.220	0.156
17	0.360	0.390	0.369	0.297	0.205
35	0.075	0.058	0.047	0.034	0.023
69	0.114	0.100	0.081	0.058	0.042
69	0.229	0.213	0.198	0.142	0.102
173	0.315	0.304	0.267	0.202	0.147
173	0.477	0.518	0.496	0.395	0.285
17	0.072	0.063	0.047	0.036	0.024
17	0.099	0.092	0.074	0.056	0.038
73	0.420	0.452	0.425	0.332	0.235
17	0.189	0.178	0.153	0.107	0.076
35	0.369	0.391	0.364	0.286	0.200
69	0.142	0.124	0.105	0.077	0.056
35	0.094	0.087	0.072	0.049	0.032
35	0.171	0.161	0.145	0.094	0.068
52	0.378	0.420	0.380	0.281	0.200

Esta medida recibe el nombre de *coeficiente de correlación múltiple ajustado* y se define por

$$R_A^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{\text{SCE}}{\text{STC}}$$

donde p es el número de términos que contiene el modelo, incluyendo al término constante. Use la información que se encuentra en la tabla 14.10 para demostrar que el coeficiente de correlación múltiple ajustado para una regresión lineal que contiene a x_1, x_2, y x_3 es más pequeña que la de la ecuación que sólo contiene a x_2 y a x_3.

14.12. Empléese la ecuación de regresión estimada en el ejercicio 14.9, parte *a*, que incluye sólo a x_1, x_2, y x_4 para calcular los residuos estandarizados. Entonces, grafíquense contra los valores correspondientes de x_3. Explíquese el resultado.

14.13. Úsese la ecuación de reducción simple estimada del eje. 13.14; calcúlense y grafíquense los residuos estandarizados frente al producto anual bruto (PAB) X. ¿Qué se puede concluir? Ajústese una nueva ecuación de reducción, como lo sugiere la gráfica residual, para demostrar el riesgo de la extrapolación al estimar el pago de los impuestos porcentuales si el PAB es $ 250 000.

14.14. Empléese la ecuación estimada de regresión lineal simple obtenida en el ejercicio 13.13 para calcular y graficar los residuos estandarizados contra los años de experiencia. Obténgase una nueva ecuación de regresión, calcúlense los nuevos residuos estandarizados y de nuevo grafíquense contra x. ¿Qué se puede concluir?

14.15. Los datos que se encuentran en la tabla 14.31 representan la temperatura atmosférica promedio Y en enero para 50 estaciones climatológicas situadas en el estado de Virginia, donde cada estación se identifica por medio de su latitud x_1, longitud x_2 y altitud x_3.

TABLA 14.31 Datos de la muestra para el ejercicio 14.13*

Estación número	Y	x_1	x_2	x_3
1	37.9	37.35	79.52	975
2	28.7	38.52	78.43	3535
3	38.3	37.08	77.95	440
4	37.3	37.53	79.68	870
5	31.5	37.08	81.33	3300
6	35.0	37.38	80.08	1890
7	36.0	38.03	78.52	870
8	37.4	36.83	79.37	700
9	40.4	37.28	75.97	11
10	35.8	37.77	78.15	300
11	35.3	38.47	78.00	420
12	33.2	38.45	78.93	1400
13	39.3	36.58	79.38	410
14	41.3	36.90	76.20	25
15	34.7	38.45	77.67	300
16	38.0	37.33	78.38	450
17	34.2	36.93	80.30	2600
18	35.4	38.30	77.47	100
19	35.7	37.37	80.87	1524
20	39.7	36.68	76.78	80
21	40.5	37.30	77.30	40
22	31.6	38.00	79.83	2238
23	40.0	37.08	76.35	10
24	36.1	37.78	79.43	1060
25	34.1	39.12	77.72	500
26	36.1	38.03	78.00	420
27	33.9	38.67	78.38	1200
28	36.6	37.33	79.20	916
29	37.1	36.70	79.88	760
30	28.6	38.42	79.58	2910
31	29.3	39.07	77.88	1720
32	37.4	37.70	78.30	300
33	40.5	36.90	76.20	22
34	38.9	37.58	75.82	300
35	34.4	36.75	83.03	1510
36	35.3	38.50	77.32	12
37	37.5	37.50	77.33	164
38	36.4	37.32	79.97	1149
39	35.0	36.88	81.77	1735
40	34.0	38.15	79.03	1385
41	33.3	38.65	78.72	1000
42	38.6	37.65	76.57	25
43	37.5	37.75	77.05	50
44	36.2	37.85	75.48	9
45	32.1	38.95	77.45	291
46	35.6	38.85	77.03	10
47	39.3	37.30	76.70	70
48	33.7	39.20	78.17	760
49	34.4	38.88	78.52	887
50	34.4	36.93	81.08	2450

* *Fuente: Monthly normals of temperature, precipitation and heating and cooling degree days 1941–70*, No. 81, NOAA, U. S. Department of Commerce.

a) Ajústese un modelo de regresión de segundo orden completo y llévense a cabo los análisis apropiados sobre sus resultados.

b) Úsense los medios apropiados para evaluar si todos los términos que aparecen en la ecuación estimada de regresión deben retenerse. Si no es así, proporciónense argumentos suficientes para la elección de una ecuación de predicción adecuada.

14.16. Los datos de la tabla 14.32 representan la tasa de crímenes Y por cada 100 000 habitantes para los 48 estados de Estados Unidos y algunas variables de predicción potenciales como el porcentaje de población urbana x_1, el porcentaje de la población minoritaria x_2, la tasa de desempleo x_3, el porcentaje de la población que tiene cuatro o más años de educación x_4 y la región geográfica x_5.

a) Empléese un procedimiento de regresión por pasos para obtener el mejor conjunto de variables de predicción por incluir en un modelo lineal.

b) Para la mejor ecuación de predicción, calcúlense los residuos estandarizados y grafíquense contra las regiones. ¿La dispersión de estos residuos es esencialmente igual para todas las regiones? Si no es así obténganse los pesos para cada región mediante el empleo del procedimiento sugerido en la sección 14.8 y después utilícese el método de mínimos cuadrados con factores de peso para obtener las estimaciones de los coeficientes de regresión del mejor conjunto de variables de predicción. Compárense los resultados con los que se obtienen al emplear el método ordinario de mínimos cuadrados.

14.17. Emplénse los datos del ejercicio 14.16 para obtener la matriz de correlación para todas las variables potenciales de predicción y la respuesta. ¿Cuáles variables de predicción son las que tienen mayor correlación con la respuesta? ¿Está este resultado de acuerdo con la parte *a* del ejercicio 14.16? ¿Existen otras variables de predicción que se encuentren muy correlacionadas? Coméntese en términos del problema de la multicolinealidad.

14.18. Se cree que los salarios Y, en miles de dólares, para los profesores de cierta universidad por año académico están influenciados por tres variables: los años de experiencia en la enseñanza x_1; el rango académico x_2, y la disciplina x_3. Los datos que figuran en la tabla 14.33 provienen de una muestra aleatoria de 18 profesores de esta universidad. Los rangos académicos se identifican por un 1 para profesor asistente, 2 para profesor asociado y 3 para profesor titular. Las disciplinas se identifican mediante un 1 para ciencias, 2 para humanidades, 3 para artes y 4 para finanzas.

a) Defínanse las variables indicadoras para el rango y la disciplina. Entonces, ajústese un modelo lineal con el salario como la respuesta, los años de experiencia en la enseñanza, como la variable cuantitativa y las variables indicadoras representan el rango académico y la disciplina.

b) Interprétense los coeficientes de la regresión estimada.

c) Ajústese un modelo lineal que incluya todos los términos que contienen productos cruzados entre las variables indicadoras y x_1. Llévese a cabo un análisis completo sobre esta ecuación de regresión y obténganse las conclusiones adecuadas.

d) Para cada disciplina y rango académico, grafíquese la ecuación de regresión estimada obtenida en la parte *c* como una función de x_1.

TABLA 14.32 Datos de la muestra para el ejercicio 14.16

Estado	Y	x_1	x_2	x_3	x_4	x_5
1	14.2	58.4	25.8	7.4	10.2	6
2	9.5	79.6	9.2	9.8	15.7	8
3	8.8	50.0	18.4	6.6	9.1	6
4	11.5	90.9	12.0	8.2	16.8	8
5	6.3	78.5	4.7	5.6	19.4	7
6	4.2	77.4	6.6	7.1	18.3	1
7	6.0	72.2	15.2	8.9	15.5	2
8	10.2	80.5	14.9	9.0	13.7	5
9	11.7	60.3	26.5	6.9	12.3	5
10	5.5	54.1	1.8	6.3	13.5	7
11	9.9	83.0	14.7	6.5	13.7	3
12	7.4	64.9	7.6	5.7	11.0	3
13	2.3	57.2	1.6	4.0	12.8	4
14	6.6	66.1	5.6	4.0	14.6	4
15	10.1	52.3	7.5	4.6	10.0	6
16	15.5	66.1	30.2	7.0	11.5	6
17	2.4	50.8	0.7	8.9	13.6	1
18	8.0	76.6	21.1	6.8	18.6	2
19	3.1	84.6	4.3	9.5	16.8	1
20	9.3	73.8	12.5	8.2	12.6	3
21	2.7	66.4	2.0	5.9	13.2	4
22	14.3	44.5	36.4	7.4	11.5	6
23	9.6	70.1	11.2	6.2	11.8	4
24	5.4	53.4	4.8	6.2	14.2	7
25	3.9	61.5	3.8	5.0	12.8	4
26	15.8	80.9	8.3	9.0	13.1	8
27	3.2	56.4	0.7	6.4	15.3	1
28	5.6	88.9	12.8	9.4	14.9	2
29	8.8	69.8	9.8	7.8	15.3	8
30	10.7	88.9	14.6	9.1	16.0	2
31	10.6	45.0	23.1	6.2	11.8	5
32	0.9	44.3	3.3	5.5	12.2	4
33	7.8	75.3	10.1	7.8	11.5	3
34	8.6	68.0	11.3	5.0	11.7	6
35	4.9	67.1	3.0	9.5	15.4	7
36	5.6	71.5	9.4	7.9	11.9	2
37	3.9	87.1	3.7	8.6	14.9	1
38	11.9	47.6	31.2	5.0	10.4	5
39	2.0	44.6	6.1	3.6	11.4	4
40	10.1	58.7	15.9	6.0	10.5	6
41	13.3	79.7	13.1	5.7	13.7	6
42	3.5	80.4	2.5	5.3	17.5	7
43	1.4	32.2	0.8	8.0	15.6	1
44	9.0	63.1	19.5	5.6	16.4	5
45	4.3	72.6	5.1	8.8	16.1	7
46	6.0	39.0	3.9	7.5	9.2	5
47	2.8	65.9	3.9	4.5	12.7	3
48	5.4	60.5	3.1	3.6	14.5	7

Fuente: World almanac, 1979

TABLA 14.33 Datos de la muestra para el ejercicio 14.18

Y	x_1	x_2	x_3
25.7	10	1	2
18.8	4	1	1
18.6	5	1	3
21.8	13	2	3
26.3	4	1	4
29.4	24	3	3
28.6	7	2	2
34.5	12	3	4
24.3	11	2	1
21.2	6	1	3
28.8	6	1	4
24.7	4	1	2
32.4	12	3	2
33.4	20	3	1
27.4	11	2	1
29.8	6	2	4
31.4	11	2	4
27.7	8	2	3

CAPÍTULO QUINCE

Métodos no paramétricos

15.1 Introducción

Los procedimientos inferenciales que hasta este momento se han estudiado, con excepción de los límites de tolerancia independientes de distribución analizados en el capítulo 8, y de la estadística de Kolmogorov-Smirnov, presentada en el capítulo 10, necesitan de la especificación de una distribución para la población de interés. Por ejemplo, el procedimiento del análisis de varianza se hace posible al asumir que las observaciones provienen de distribuciones normales. De esta forma, la mayor parte de los procesos inferenciales que se han presentado representan estimaciones con respecto a los parámetros de la población de interés. Por esta razón, este tipo de inferencias reciben el nombre de *métodos paramétricos.*

Para muchos de los métodos inferenciales que se han examinado se ha hecho un intento por determinar su robustez, y en muchas ocasiones se ha encontrado que los métodos son razonablemente robustos con respecto a las distribuciones supuestas. No obstante, en general los métodos paramétricos son más sensibles a las suposiciones para muestras de tamaño pequeño y, para muchos de ellos, su aplicación se encuentra limitada a aquellas observaciones que tienen un carácter cuantitativo, es decir, se supone que lo que se observa es una cantidad numérica continua como el volumen de ventas semanal, la cantidad de cierta sustancia que se vacía en un recipiente, la resistencia de una muestra de metal y otros más.

Las observaciones de tipo cuantitativo se definen, en forma general, sobre un *intervalo* o sobre una escala de *proporciones.* Las mediciones que se definen en una escala de intervalo se pueden distinguir y ordenar en forma numérica, y sus diferencias son significativas. Un ejemplo clásico de una escala de intervalo es aquel que incluye la medición de la temperatura. Puede escogerse entre registrar la temperatura en grados Celsius (para los cuales el punto de congelación del agua es de 0°C) o en grados Fahrenheit (para los que el punto de congelación es de 32°F). De esta forma el origen de las escalas es diferente, pero el significado de la diferencia entre 10°C y 15°C es el mismo que tiene la diferencia entre 20°C y 25°C.

Si una medición reúne los requisitos de una escala de intervalo y además tiene un verdadero punto de origen, entonces la medición se define sobre una escala de pro-

porciones. Por ejemplo, las alturas, los pesos, las resistencias y otros se encuentran definidos sobre una escala de proporciones ya que tienen verdaderos puntos cero, sin importar la unidad de medición. Las escalas de intervalo y de proporción son verdaderamente cuantitativas. Para la mayor parte de los métodos paramétricos que se han presentado, como son la construcción de intervalos de confianza, la prueba de hipótesis estadísticas y el ajuste de ecuaciones son aplicables a todas aquellas observaciones que se encuentran definidas, por lo menos, sobre una escala de intervalo.

Sin embargo, en muchas situaciones lo que se observa tiene un carácter cualitativo (no cuantitativo) y, por lo tanto, no puede definirse sobre una escala de intervalo o de proporciones. Tales situaciones se encuentran con frecuencia en las ciencias sociales y en las encuestas de mercado. Por ejemplo, no es probable que al evaluar las preferencias del consumidor con respecto a una bebida, se adhieran a una escala numérica significativa, incluso si se le pidiese al consumidor su opinión con respecto a la bebida en una escala de cinco puntos, donde 1 y 5 pueden representar reacciones muy negativas o muy positivas, respectivamente, la escala es arbitraria. En otras palabras, los números no tienen ningún significado físico más allá que el de representar con un número más grande la respuesta más favorable para la bebida.

Las observaciones de este tipo pueden definirse sobre una escala *ordinal,* dado que la distancia entre dos puntos no es de consecuencia y sólo es importante el *orden* o *rango* de los números. En algunas ocasiones, las observaciones sólo pueden definirse sobre una escala *nominal* debido a que se emplea, ya sea un nombre (símbolo) o un número para clasificar una característica de interés, pero el principio de orden no es de consecuencia. Por ejemplo, las personas pueden clasificarse de acuerdo con su sexo. Pueden emplearse los símbolos M y H o utilizar números como 122 y 48 para denotar mujer u hombre. Las observaciones que se definen sobre escalas nominales son mediciones con pocas propiedades.

Se han desarrollado procedimientos inferenciales que no se encuentran sujetos a la forma de la distribución de la población de interés y no requieren, en forma necesaria, que las observaciones se definan por lo menos en una escala de intervalo. Estos procedimientos inferenciales se conocen como *métodos no paramétricos.* Dado que estos métodos no necesitan que se especifique la forma de la distribución de la población de interés, también se conocen como métodos *independientes de la distribución* (recuérdese, por ejemplo, los límites de tolerancia independientes de la distribución estudiados en el capítulo 8). En un sentido relativo, los métodos paramétricos requieren de pocas suposiciones y, la mayor parte de las veces, son más fáciles de aplicar que los procedimientos paramétricos que se han presentado en los capítulos anteriores; además, los métodos no paramétricos pueden aplicarse en aquellas situaciones para las que las observaciones se definen, por lo menos, en una escala de intervalo y, en ocasiones, sobre escalas nominales. Pero si las observaciones se definen por lo menos en una escala de intervalo y la distribución de la población de interés es normal, los métodos no paramétricos son menos eficientes comparados con los procedimientos paramétricos que se basan en la suposición de normalidad.

Se han desarrollado muchos métodos paramétricos en los que se han incluido procedimientos de análisis de varianza y de regresión. Las referencias citadas al final del capítulo proporcionan un panorama completo de todos los métodos no para-

métricos. El propósito de este capítulo radica sólo en introducir los conceptos básicos y presentar algunos métodos que son, en forma especial, útiles. Estos procedimientos no paramétricos son comparables con los métodos paramétricos para la prueba de hipótesis con respecto a las medias de dos distribuciones normales independientes (sección 9.6.2), la prueba de hipótesis con respecto a las medias para observaciones igualadas (sección 9.6.4), experimentos unifactoriales en diseños y en bloque completamente aleatorios (sección 12.4 y 12.5) y correlación lineal (sección 13.8).

15.2 Pruebas no paramétricas para comparar dos poblaciones con base en muestras aleatorias independientes

En la sección 9.6.2 se consideró el problema de comparar las medias de dos distribuciones cuando se supone que son normales. En esta sección se analizarán dos procedimientos no paramétricos para comparar las distribuciones de dos poblaciones: la prueba U de Mann-Whitney y la prueba de tendencias de Wald-Wolfowitz. La única suposición necesaria para su aplicación es que las distribuciones de interés sean continuas. De acuerdo con lo anterior, se supondrá que $X_1, X_2, \ldots, X_{n_1}$ y $Y_1, Y_2, \ldots, Y_{n_2}$ son muestras aleatorias independientes de dos poblaciones con distribuciones continuas.

15.2.1 Prueba de Mann-Whitney

Dadas muestras aleatorias independientes de dos poblaciones, considérese la prueba de la hipótesis nula de que las poblaciones tienen la misma distribución. La hipótesis puede establecerse como

$$H_0: f_1(x) \equiv f_2(y), \tag{15.1}$$

donde $f_1(x)$ y $f_2(y)$ son las correspondientes funciones de densidad de probabilidad. La hipótesis alternativa puede ser uni o bilateral. La hipótesis alternativa bilateral establece en forma sencilla que las distribuciones no son las mismas. Pero la hipótesis alternativa sólo implica un desplazamiento en la tendencia central de una distribución con respecto a la otra y no sugiere una diferencia en la forma o en la dispersión. En otras palabras, al igual que para el procedimiento t de Student, se supone que las distribuciones tienen la misma forma y dispersión.

Un procedimiento popular no paramétrico para probar la hipótesis nula dada por (15.1) es la *prueba U de Mann-Whitney*.* Esta prueba es el equivalente no paramétrico de la prueba t de student para dos muestras estudiada en la sección 9.6.2. La prueba de Mann-Whitney se basa en una combinación de las n_1 y n_2 observaciones para formar un solo conjunto de $n_1 + n_2$ observaciones arregladas en orden creciente de magnitud. Entonces se asigna un *rango* a cada observación en la secuencia ordenada que comienza con un rango 1 y termina con un rango $n_1 + n_2$. Si las muestras aleatorias provienen de poblaciones que tienen la misma distribución, se espera que los rangos se encuentren lo suficientemente dispersos cuando se observa

* Este procedimiento es, en forma esencial, igual a la prueba de Wilcoxon de la suma del rango.

en qué muestra se encuentran las observaciones. De otra forma, debe esperarse que los rangos de las observaciones en cada muestra se encuentren muy agrupados en los extremos. En esencia, la estadística de Mann-Whitney determina cuándo un agregado de rangos observados es suficiente para concluir que las dos muestras aleatorias provienen de poblaciones cuyas distribuciones difieren en la tendencia central.

Para implementar el procedimiento se obtiene la suma de los rangos asociados con las observaciones de una de las dos muestras, por ejemplo la muestra 1, la cual se escoge en forma arbitraria. Denótese esta suma por R_1. Entonces la estadística U de Mann-Whitney está dada por

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1. \tag{15.2}$$

La estadística U es una función de la variable aleatoria R_1 y de los tamaños de las muestras n_1 y n_2. Si H_0 es cierta, la ocurrencia de cualquier orden particular para las observaciones en el conjunto combinado es equiprobable. Por lo tanto, bajo H_0, R_1 es la suma de n_1 enteros positivos seleccionados en forma aleatoria de entre los primeros $n_1 + n_2$. De acuerdo con lo anterior, puede determinarse que

$$E(R_1) = n_1(n_1 + n_2 + 1)/2, \tag{15.3}$$

$$Var(R_1) = n_1 n_2(n_1 + n_2 + 1)/12. \tag{15.4}$$

De (15.2) sigue que

$$E(U) = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - E(R_1) = n_1 n_2/2, \tag{15.5}$$

y

$$Var(U) = Var(R_1) = n_1 n_2(n_1 + n_2 + 1)/12. \tag{15.6}$$

Se ha determinado y tabulado la distribución exacta de U. Se invita al lector a que consulte [1] y [2] para conocer los detalles. Para una hipótesis alternativa bilateral, es probable que se rechace H_0 si se obtiene un valor muy grande o muy pequeño de U. Lo anterior ocurrirá cuando el valor de R_1 es muy grande o muy pequeño, respectivamente. Sin embargo, cuando tanto n_1 y n_2 son mayores de 10, la distribución de U se encuentra, en forma adecuada, aproximada por una distribución normal con media y varianza dadas por (15.5) y (15.6), respectivamente, es decir, bajo H_0 la variable aleatoria

$$Z = \frac{U - E(U)}{\sqrt{Var(U)}}$$

es aproximadamente $N(0, 1)$ para valores grandes de n_1 y n_2.

Debe notarse que a pesar de que no pueden ocurrir empates en la práctica desde un punto de vista teórico, esto ocurre en muchas ocasiones. Cuando ocurre un empate en la secuencia ordenada, se sugiere asignar el promedio de los rangos a las observaciones para las cuales ocurre el empate. Por ejemplo, supóngase que las obser-

vaciones octava y novena en la secuencia ordenada son las mismas. Entonces a cada una de estas observaciones se les asigna un rango de 8.5.

Ejemplo 15.1 Se sospecha que una compañía lleva a cabo una política de discriminación, con respecto al sexo, en los salarios de sus empleados. Se seleccionaron 12 empleados masculinos y 12 femeninos de entre los que tienen responsabilidades y experiencia similares en el trabajo; sus salarios anuales en miles de dólares son los siguientes:

Mujeres	22.5	19.8	20.6	24.7	23.2	19.2	18.7	20.9	21.6	23.5	20.7	21.6
Hombres	21.9	21.6	22.4	24.0	24.1	23.4	21.2	23.9	20.5	24.5	22.3	23.6

¿Existe alguna razón para creer que estas muestras aleatorias provienen de poblaciones con diferentes distribuciones? Úsese $\alpha = 0.05$.

Se combinan los salarios de las dos muestras para formar un solo conjunto de 24 salarios anuales. Entonces se ordenan los salarios y se les asigna un rango de la siguiente manera:

Sexo	*M*	*M*	*M*	*H*	*M*	*M*	*M*	*H*	*H*	*M*	*M*	*H*
Rango del salario	18.7 1	19.2 2	19.8 3	20.5 4	20.6 5	20.7 6	20.9 7	21.2 8	21.6 10	21.6 10	21.6 10	21.9 12

Sexo	*H*	*H*	*F*	*M*	*H*	*M*	*H*	*H*	*H*	*H*	*H*	*M*
Rango del salario	22.3 13	22.4 14	22.5 15	23.2 16	23.4 17	23.5 18	23.6 19	23.9 20	24.0 21	24.1 22	24.5 23	24.7 24

Para obtener la suma de los rangos se seleccionará la muestra de mujeres. De esta forma la suma de los rangos es

$$1 + 2 + 3 + 5 + 6 + 7 + 10 + 10 + 15 + 16 + 18 + 24 = 117,$$

y el valor de la estadística U de Mann-Whitney es

$$u = (12)(12) + \frac{12(13)}{2} - 117 = 105.$$

Dado que $E(U) = (12)(12)/2 = 72$ y $Var(U) = (12)(12)(25)/12 = 300$, mediante el empleo de la aproximación normal,

$$z = (105 - 72)/\sqrt{300} = 1.91$$

es un valor de una variable aleatoria normal estándar. Para $\alpha = 0.05$, los valores críticos son ± 1.96. Por lo tanto, no puede rechazarse la hipótesis nula de que las muestras aleatorias provienen de poblaciones con distribuciones idénticas.

15.2.2 Prueba de tendencias de Wald-Wolfowitz

Otro método no paramétrico que compara las distribuciones de dos poblaciones con base en muestras aleatorias independientes es la *prueba de tendencias de Wald-Wolfowitz*. Para esta prueba, la hipótesis nula es que las dos muestras aleatorias provienen de poblaciones que tienen distribuciones idénticas, pero a diferenca de la prueba U de Mann-Whitney, no sugiere una diferencia sólo en la tendencia central; es decir, la hipótesis alternativa en la prueba de Wald-Wolfowitz es mucho más amplia. Ésta establece simplemente que las distribuciones difieren en algún aspecto como en la tendencia central, en la dispersión o la asimetría.

Al igual que en la prueba de Mann-Whitney, las observaciones en las dos muestras aleatorias se combinan y ordenan de acuerdo con sus magnitudes. Pero en lugar de considerar los rangos, el procedimiento de Mann-Wolfowitz busca el número de tendencias en la secuencia ordenada.

Definición 15.1 Se define una tendencia de longitud j como una secuencia de j observaciones, todas pertenecientes al mismo grupo, que se encuentran, ya sea precedidas o seguidas por observaciones que pertenecen a un grupo diferente.

Como ilustración, recuérdese la secuencia ordenada del ejemplo 15.1. La secuencia ordenada de acuerdo con el sexo de los empleados es la siguiente:

F	F	F	M	F	F	F	M	M	F	F	M
M	M	F	F	M	F	M	M	M	M	M	F

Para el sexo del empleado, la secuencia ordenada exhibe tendencias de M y H. En particular, la secuencia comienza con una tendencia de longitud tres, seguida por una tendencia de longitud uno, seguida por otra de longitud tres, y así consecutivamente. El número total de tendencias en esta secuencia es de 11.

Si la hipótesis nula de que las distribuciones son idénticas es cierta, las observaciones de las dos muestras en la secuencia ordenada deben encontrarse bien mezcladas, produciendo de esta forma un número grande de tendencias. Pero si las distribuciones de interés difieren en algún aspecto, es probable que la secuencia ordenada contenga tendencias de corta longitud obteniéndose de esta forma un número total de tendencias pequeño.

Sea R el número total de tendencias observadas en una secuencia ordenada de $n_1 + n_2$ observaciones, donde n_1 y n_2 son los respectivos tamaños de las muestras. Los posibles valores de R son $2, 3, \ldots, (n_1 + n_2)$. Puede demostrarse que la función de probabilidad de R está dada por

$$p(r) = \begin{cases} \dfrac{2\binom{n_1 - 1}{r/2 - 1}\binom{n_2 - 1}{r/2 - 1}}{\binom{n_1 + n_2}{n_1}} & r \text{ par}, \\ \dfrac{\binom{n_1 - 1}{r/2 - 1/2}\binom{n_2 - 1}{r/2 - 3/2} + \binom{n_1 - 1}{r/2 - 3/2}\binom{n_2 - 1}{r/2 - 1/2}}{\binom{n_1 + n_2}{n_1}} & r \text{ impar}. \end{cases} \tag{15.7}$$

La media y la varianza de R son

$$E(R) = \frac{2n_1n_2}{n_1 + n_2} + 1, \tag{15.8}$$

$$Var(R) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}. \tag{15.9}$$

Para probar H_0 con una probabilidad α, para el error de tipo I, debe encontrarse un entero r_0 tal, que en la medida de lo posible

$$\sum_{r=2}^{r_0} p(r) = \alpha.$$

Se rechaza la hipótesis nula cuando el número observado de tendencias es menor o igual a r_0. Nótese que la región crítica es una región unilateral inferior dado que se rechaza H_0 cuando el número de tendencias es bastante pequeño.

La distribución acumulativa de R se encuentra tabulada en forma extensa; pero si tanto n_1 como n_2 son mayores que 10, la distribución de R se encuentra, en forma adecuada, aproximada por una distribución normal con media y varianza dadas por (15.8) y (15.9), respectivamente. Como ilustración, recuérdese el ejemplo 15.1. El número observado de tendencias es 11, y para $n_1 = n_2 = 12$ los valores de la media y la varianza de R son 13 y 5.7391, respectivamente. Entonces, mediante el empleo de la aproximación normal,

$$z = (11 - 13)/\sqrt{5.7391} = -0.83$$

es un valor de una variable aleatoria normal estándar. Para $\alpha = 0.05$, se observa que la hipótesis nula no puede ser rechazada.

En la aplicación de la prueba de tendencias de Wald-Wolfowitz surge un problema muy serio cuando ocurren empates entre las observaciones que se encuentran en grupos diferentes. Este problema se debe a que el número de tendencias depende de cómo se manejen los empates en la secuencia ordenada. El procesamiento que se sugiere en estos casos es el de ordenar las observaciones empatadas en forma tal, que sea lo menos favorable para el rechazo de H_0. Pero si se tienen muchos empates, la validez de la prueba de Wald-Wolfowitz es cuestionable.

Por causa de la naturaleza extensa de la hipótesis alternativa en la prueba de Wald-Wolfowitz, ésta y la prueba de Mann-Whitney no son comparables. Si un investigador desea comparar las tendencias centrales de las distribuciones de dos poblaciones y sólo se tienen observaciones ordinales disponibles, la estadística de Mann-Whitney es el procedimiento no paramétrico más poderoso para detectar diferencias entre las tendencias centrales. Si se va a hacer una comparación más amplia, la prueba de Wald-Wolfowitz es un procedimiento viable pero menos poderoso.

15.3 Pruebas no paramétricas para observaciones por pares

En la sección 9.6.4 se consideró la comparación entre las medias de dos tratamientos cuando las observaciones se encuentran igualadas con el propósito de eliminar los

efectos causados por factores externos. En esta sección se discutirán dos pruebas no paramétricas que son equivalentes al procedimiento t de Student de la sección 9.6.4. Éstas son la *prueba del signo* y la *prueba de rangos y signos de Wilcoxon.*

15.3.1 La prueba del signo

La *prueba del signo* se basa en los signos de las diferencias entre las observaciones por pares de dos variables aleatorias X y Y. Sean (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) pares de n observaciones muestrales de las distribuciones de X y Y, donde se supone que éstas son continuas. En muchas ocasiones existe una relación natural entre X y Y, por lo que X y Y no necesitan ser independientes. Por ejemplo, X y Y pueden representar las respuestas de parejas de matrimonios.

Para cada par en el que X es mayor que Y se registra un signo (+) de otra forma se registra un signo (–). Dado que se supone que las distribuciones de X y Y son continuas, en forma teórica, no pueden ocurrir empates. Sea p la probabilidad de que X sea mayor que Y. Entonces si la hipótesis nula es que X y Y tienen la misma distribución, el valor de p debe ser igual a 0.5. Sin embargo, debe notarse que p puede ser igual a 0.5, aun cuando las distribuciones de X y Y no sean idénticas. Por lo tanto, y en esencia, la hipótesis nula para la prueba del signo es

$$H_0: \quad p = 0.5,$$

La cual puede probarse contra hipótesis alternativas, ya sean uni o bilaterales, lo cual depende de lo que el investigador desee. Nótese que si H_0 es cierta, debe esperarse que, en forma aproximada, la mitad de los n pares tengan signos +.

La estadística para la prueba del signo, denotada por S, es el número de signos + entre los n pares. Dado que bajo H_0 cada par constituye un ensayo independiente con una probabilidad para el signo + de 0.5, la estadística S tiene una distribución binomial con $p = 0.5$. De acuerdo con lo anterior, para n dado y $p = 0.5$, se emplea la distribución binomial para obtener las regiones críticas de tamaño α para el error de tipo I. Para valores grandes de n puede utilizarse la aproximación normal de la distribución binomial, estudiada en la sección 5.2.

Cuando ocurren empates al aplicar la prueba del signo, el procedimiento que se recomienda seguir es el de ignorarlos y emplear la prueba sólo para aquellos pares en los que no ocurren empates. Este procedimiento puede representar un problema si se tienen numerosos empates y el número original de pares es relativamente pequeño.

Ejemplo 15.2 Se seleccionaron al azar 10 parejas de recién casados, y se les preguntó por separado, tanto al marido como a la esposa, cuántos hijos deseaban tener. Se obtuvo la siguiente información.

Pareja	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Esposa X	3	2	1	0	0	1	2	2	2	0
Marido Y	2	3	2	2	0	2	1	3	1	2

Mediante el empleo de la prueba del signo, ¿existe alguna razón para creer que las

esposas desean menos hijos que sus esposos? Supóngase un tamaño máximo del error del tipo I de 0.05.

Considérese la prueba de la hipótesis nula

$$H_0: \ p = 0.5$$

contra la alternativa

$$H_1: \ p < 0.5.$$

Nótese que deberá rechazarse H_0 si el número de signos + es muy pequeño. Al restar las respuestas de cada esposo de la de su esposa, y notando que las respuestas de cinco de las parejas son las mismas, se obtiene el siguiente arreglo de signos + y –:

Pareja	*1*	*2*	*3*	*4*	*6*	*7*	*8*	*9*	*10*
Signo	+	–	–	–	–	+	–	+	–

Existen tres signos + de manera tal, que el valor de la estadística S es 3. Dado que bajo H_0, S es binomial con $n = 10$ y $p = 0.5$, el valor p, o la probabilidad de observar tres o menos signos +, se obtiene de la tabla A del apéndice y es

$$P(S \leq 3) = 0.2539.$$

Dado que 0.2539 es mayor que $\alpha = 0.05$ la hipótesis nula no puede rechazarse. Nótese que para este ejemplo el valor crítico de S debe ser igual a uno si el tamaño máximo del error de tipo I es de 0.05.

15.3.2 Prueba de rangos de signos de Wilcoxon

La prueba del signo considera sólo las diferencias en el signo entre cada par de observaciones e ignora sus magnitudes. Si las observaciones se definen sobre una escala ordinal, las magnitudes de las diferencias tienen poco valor. Pero si las observaciones son magnitudes físicas, la prueba del signo puede ignorar mucha información debido a que no se toman en cuenta las magnitudes de las diferencias. La *prueba de rangos y de signos de Wilcoxon* toma en cuenta tanto el signo como la magnitud de las diferencias entre cada par de observaciones. Por lo tanto, para tener un buen balance, éste es el mejor método no paramétrico por utilizar para observaciones en parejas.

Para implementar la prueba de Wilcoxon, se obtienen las diferencias para los n pares de observaciones. Entonces, se ordenan sin importar el signo y de acuerdo con este orden se les asigna un rango, es decir, la diferencia más pequeña recibe un rango uno y a la diferencia absoluta más grande se le asigna un rango igual a n. Entonces, el signo de cada diferencia se une al rango de ésta. Los empates entre las diferencias se manejan de la misma manera que en la prueba de Mann-Whitney, pero si una diferencia es igual a cero, el procedimiento que se sugiere es omitir el par y ajustar n.

La estadística de la prueba de Wilcoxon es la suma de los rangos positivos y se denota por T_+. Nótese que T_+ contiene no sólo información proporcionada por la estadística de la prueba del signo sino también información con respecto a la magni-

tud relativa de las diferencias. Si la hipótesis nula de que las observaciones en cada par provienen de distribuciones idénticas es cierta, la ocurrencia de cualquier secuencia, en particular de los rangos y signos, es equiprobable de entre las 2_n secuencias posibles de signos + y −. Bajo la hipótesis nula, se espera que T_+ tenga el mismo valor, aproximadamente, que la suma de las magnitudes de los rangos negativos. Por lo tanto, dependiendo de la naturaleza de la hipótesis alternativa, se rechaza H_0 cuando se observa un valor de T_+ suficientemente grande o pequeño.

Se ha determinado y tabulado la distribución exacta de T_+. Sin embargo, al igual que para algunas otras estadísticas, la distribución de muestre de T_+ se encuentra aproximada, en forma adecuada, por una distribución normal para $n > 10$, donde

$$E(T_+) = n(n + 1)/4, \tag{15.10}$$

$$Var(T_+) = n(n + 1)(2n + 1)/24. \tag{15.11}$$

En otras palabras, la variable aleatoria

$$Z = \frac{T_+ - E(T_+)}{\sqrt{Var(T_+)}}$$

es aproximadamente $N(0, 1)$ para valores grandes de n.

Ejemplo 15.3 De una clase de estadística se seleccionan al azar 11 estudiantes y se observan sus calificaciones en dos exámenes sucesivos. Para las calificaciones dadas en la tabla 15.1, utilícese la prueba de rangos y de signos de Wilcoxon para determinar si el segundo examen fue más difícil que el primero. Úsese $\alpha = 0.1$.

En la tabla se encuentran las diferencias (examen 1 − examen 2), rangos, y rangos con signos para los 11 estudiantes. Dado que se desea determinar si el segundo examen fue más difícil que el primero, la hipótesis alternativa es unilateral, y la región crítica se encuentra en el extremo superior de la distribución de muestreo de T_+ es decir, si el valor observado de la suma de los rangos positivos es grande, lo anterior

TABLA 15.1 Datos de la muestra para el ejemplo 15.3

Estudiante	*Prueba 1*	*Prueba 2*	*Diferencia*	*Rango*	*Rango con signo*
1	94	85	9	8	8
2	78	65	13	10	10
3	89	92	−3	4	−4
4	62	56	6	7	7
5	49	52	−3	4	−4
6	78	74	4	6	6
7	80	79	1	1	1
8	82	84	−2	2	−2
9	62	48	14	11	11
10	83	71	12	9	9
11	79	82	−3	4	−4

implicaría tener calificaciones bajas, en forma suficiente, para el examen 2, y debe rechazarse la hipótesis nula de no diferencia.

La suma de los rangos positivos es $8 + 10 + 7 + 6 + 1 + 11 + 9 = 52$. Para $n = 11$, los valores de la media y la varianza de T_+ son $E(T_+) = 33$ y $Var(T_+) = 126.5$. Entonces, mediante el empleo de la aproximación normal,

$$z = \frac{52 - 33}{\sqrt{126.5}} = 1.69.$$

Para $\alpha = 0.1$, $z_{0.9} = 1.28$, y por lo tanto se rechaza la hipótesis nula.

15.4 Prueba de Kruskal-Wallis para k muestras aleatorias independientes

Recuérdese el procedimiento paramétrico del análisis de varianza de la sección 12.4, en el que el interés radica en probar la hipótesis nula

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k,$$

con base en k muestras aleatorias mutuamente independientes provenientes de poblaciones cuyas distribuciones se suponen como normales. Se han desarrollado métodos no paramétricos para, de manera esencial, el mismo propósito siempre que por lo menos se encuentren disponibles mediciones ordinales y las distribuciones de las poblaciones de interés sean continuas. Uno de estos métodos es el procedimiento de *Kruskal-Wallis,* el cual prueba las hipótesis nulas de que los efectos de los tratamientos son los mismos, o que las k muestras aleatorias provienen de poblaciones con distribuciones idénticas.

Sean las observaciones de las k muestras aleatorias las dadas en la tabla 15.2, donde n_j es el tamaño de la j-ésima muestra y $N = \sum_{j=1}^{k} n_j$ es el número total de observaciones para todas las muestras.

La hipótesis nula puede establecerse como

$$H_0: \quad f_1(y) \equiv f_2(y) \equiv \cdots \equiv f_k(y) \tag{15.12}$$

donde $f_1(y), f_2(y), \cdots, f_k(y)$ son las correspondientes funciones de densidad de probabilidad. La hipótesis alternativa puede ser general y establecer sólo que las k

TABLA 15.2 Observaciones de k muestras aleatorias para la prueba de Kruskal-Wallis

Muestra					
1	2	⋯	j	⋯	k
Y_{11}	Y_{12}	⋯	Y_{1j}	⋯	Y_{1k}
Y_{21}	Y_{22}	⋯	Y_{2j}	⋯	Y_{2k}
⋮	⋮		⋮		⋮
$Y_{n_1 1}$	$Y_{n_2 2}$	⋯	$Y_{n_j j}$	⋯	$Y_{n_k k}$

distribuciones no son idénticas. Sin embargo, la prueba de Kruskal-Wallis es sensible a las diferencias en tendencia central y es muy útil cuando se sospecha que las distribuciones de interés difieren sólo en ese aspecto. De acuerdo con lo anterior, el procedimiento de Kruskal-Wallis se considera, en general, como una extensión de la prueba U, de Mann-Whitney.

Al igual que en la prueba de Mann-Whitney, el procedimiento de Kruskal-Wallis se basa en la combinación de todas las observaciones en las muestras aleatorias para formar un solo conjunto de N observaciones; entonces, éstas se arreglan en orden creciente de magnitud y se asigna un rango a cada observación comenzando con un rango 1 y terminando con un rango N. Cuando el rango de todas las observaciones está completo, se determina la suma de los rangos para cada muestra. Sea R_j la suma de los rangos de la j-ésima muestra. En esencia, la prueba de Kruskal-Wallis determina si la disparidad entre las R_j con respecto a los tamaños n_j de las muestras es suficiente para garantizar el rechazo de la hipótesis nula.

Bajo la suposición de que las k muestras provienen de poblaciones con distribuciones idénticas, la estadística de la prueba de Kruskal-Wallis es

$$H = \frac{12}{N(N+1)}\left[\sum_{j=1}^{k}\frac{R_j^2}{n_j}\right] - 3(N+1), \tag{15.13}$$

la que para tamaños n_j relativamente grandes de las muestras se encuentra aproximada, en forma adecuada, por una distribución chi-cuadrada con $k-1$ grados de libertad. Para un tamaño específico del error de tipo I, la región crítica es la porción superior de la distribución chi-cuadrada. De acuerdo con lo anterior, se rechaza la hipótesis nula para valores grandes de la estadística de la prueba de Kruskal-Wallis. Debe notarse que la aproximación chi-cuadrada es, por lo general, satisfactoria, excepto cuando $k = 3$ y ninguno de los tamaños de las muestras n_j sea mayor que cinco.

El procedimiento que se recomienda para manejar los empates es igual al de la prueba de Mann-Whitney. Si el número de empates es grande, se ha propuesto un factor de corrección para la estadística de pruebas dada por (15.3); véanse cualesquiera de las referencias que se encuentran al final de este capítulo. A pesar de que esta correción siempre incrementa el valor de la estadística de prueba, en muchos casos este efecto es despreciable, aun si existen numerosos empates.

Ejemplo 15.4 Se tomaron muestras aleatorias independientes de casas recientemente vendidas en cuatro zonas residenciales de una gran ciudad. El problema era determinar si existían diferencias en las zonas con respecto al valor de la propiedad y el precio de venta. Los datos que figuran en la tabla 15.3 son los cocientes entre los precios de venta y el valor catastral de la propiedad. Para $\alpha = 0.05$, empléese la estadística de Kruskal-Wallis para probar si estas muestras provienen de poblaciones con distribuciones idénticas.

Los valores que se encuentran entre paréntesis en la tabla son los rangos de las observaciones después de haberlas combinado y ordenado. Nótese que $n_1 = n_4 = 5$, $n_2 = n_3 = 6$ y $N = 22$. Las sumas de los rangos de cada muestra

TABLA 15.3 Datos de la muestra para el ejemplo 15.4

Zona residencial			
1	*2*	*3*	*4*
1.19 (15)	1.08 (4.5)	0.98 (2)	1.12 (7.5)
1.05 (3)	1.23 (17.5)	1.19 (15)	1.14 (10)
1.14 (10)	1:26 (20)	1.08 (4.5)	1.31 (22)
1.25 (19)	1.10 (6)	0.93 (1)	1.12 (7.5)
1.29 (21)	1.18 (12.5)	1.23 (17.5)	1.19 (15)
	1.14 (10)	1.18 (12.5)	

son $R_1 = 68$, $R_2 = 70.5$, $R_3 = 52.5$, y $R_4 = 62$. Entonces el valor de la estadística de Kruskal-Wallis es

$$h = \frac{12}{(22)(23)}\left[\frac{(68)^2}{5} + \frac{(70.5)^2}{6} + \frac{(52.5)^2}{6} + \frac{(62)^2}{5}\right] - 3(23) = 1.70.$$

De la tabla *E* del apéndice, para $\alpha = 0.05$ y $k - 1 = 3$ grados de libertad, el valor crítico es 7.82. Dado que $h = 1.70 < 7.82$, no puede rechazarse la hipótesis nula. Por lo tanto, no existe alguna razón para creer que existen diferencias entre las zonas cuando se comparan el precio de venta y el valor real de la propiedad.

15.5 Prueba de Friedman para *k* muestras igualadas

La prueba de rango de signos de Wilcoxon se considera como el equivalente no paramétrico del método *t* de Student para observaciones por pares o del procedimiento de análisis de varianza para experimentos con dos tratamientos en un diseño en bloque completamente aleatorio. Cuando es necesario investigar $k \geq 3$ tratamientos de un solo factor en presencia de un factor externo y por lo menos se encuentran disponibles mediciones ordinales, un método no paramétrico útil para determinar si los efectos debidos a los tratamientos son los mismos, es la *prueba de Friedman.*

De manera similar al procedimiento paramétrico, se crea un bloque para cada una de las *n* condiciones de los factores externos de tal manera que cada bloque contiene una observación proveniente de cada uno de los *k* tratamientos. Además, se supone que los tratamientos se asignan en forma aleatoria y que no existe ninguna interacción entre los bloques y los tratamientos. Las *nk* observaciones se arreglan como se ilustra en la tabla 15.4, donde los bloques son los renglones y los tratamientos las columnas.

La hipótesis nula para el procedimiento de Friedman es que los efectos atribuidos a los tratamientos son los mismos (es decir, las problaciones de interés tienen distribuciones idénticas) y la hipótesis alternativa es que existe una diferencia entre los tratamientos. Al igual que para la estadística de Kruskal-Wallis, las diferencias en los tratamientos descubiertas a través de la estadística de Friedman implican diferencias en la tendencia central.

TABLA 15.4 Arreglo de las observaciones para la prueba de Friedman

	Tratamiento					
Bloque	1	2	⋯	*j*	⋯	*k*
1	Y_{11}	Y_{12}	⋯	Y_{1j}	⋯	Y_{1k}
2	Y_{21}	Y_{22}	⋯	Y_{2j}	⋯	Y_{2k}
⋮	⋮	⋮		⋮		⋮
n	Y_{n1}	Y_{n2}	⋯	Y_{nj}	⋯	Y_{nk}

Al igual que en los otros procedimientos no paramétricos, la prueba de Friedman se basa en los rangos. Para cada bloque (renglón) se asigna un rango a las observaciones comenzando con un rango 1 y terminando con un rango k; entonces se suman los rangos para cada tratamiento. Sea R_j la suma de los rangos del j-ésimo tratamiento (columna). Si dentro de cada bloque los efectos del tratamiento son los mismos, entonces para cualquier bloque los rangos deben ser una permutación aleatoria de los enteros del 1 al k, donde cada permutación tiene la misma probabilidad de ocurrencia. De esta forma, se espera que para cada tratamiento los rangos del 1 al k aparezcan, en forma aproximada, con la misma frecuencia. Si los efectos de los tratamientos son idénticos, R_j deberá tener prácticamente el mismo valor para toda j. Por lo tanto, el procedimiento de Friedman determina cuándo una disparidad observada entre los R_j es suficiente para rechazar la hipótesis nula.

La estadística de Friedman está dada por

$$S = \frac{12}{nk(k+1)}\left[\sum_{j=1}^{k} R_j^2\right] - 3n(k+1). \tag{15.14}$$

Las probabilidades para los valores de S se encuentran tabuladas para valores pequeños de n y k (véase [3]). Pero si el número de bloques n y el de tratamientos k no es muy pequeño (por ejemplo $n \geq 10$ y $k \geq 4$), la estadística S es, en forma aproximada, una variable aleatoria chi-cuadrada con $k-1$ grados de libertad. Al igual que para la prueba de Kruskal-Wallis, la región crítica de tamaño α es la porción superior de la distribución chi-cuadrada con $k-1$ grados de libertad. Se rechaza la hipótesis nula cuando el valor de S es mayor que el valor crítico. Al igual que en los casos anteriores, los empates se manejan mediante el uso de rangos promedio.

Ejemplo 15.5 Cuatro jueces se encargan de calificar en una competencia de salto que incluye a 10 finalistas. Los datos que figuran en la tabla 15.5 son las calificaciones, donde un 10 indica un salto perfecto. Para $\alpha = 0.01$, empléese la estadística de Friedman para determinar si existen diferencias discernibles en las calificaciones que otorgan cada uno de los cuatro jueces.

Los valores que figuran entre paréntesis en la tabla 15.5 son los rangos de las observaciones para cada competidor (bloque). Entonces, para cada juez, la suma de los

TABLA 15.5 Datos de la muestra para el ejemplo 15.5

Competidor	*Juez* 1	2	3	4
1	8.5 (3)	8.6 (4)	8.2 (1)	8.4 (2)
2	9.8 (4)	9.7 (3)	9.4 (1)	9.6 (2)
3	7.9 (2)	8.1 (3)	7.5 (1)	8.2 (4)
4	9.7 (3)	9.8 (4)	9.6 (1.5)	9.6 (1.5)
5	6.2 (1)	6.8 (3)	6.9 (4)	6.5 (2)
6	8.9 (3)	9.2 (4)	8.1 (1)	8.7 (2)
7	9.2 (3.5)	9.2 (3.5)	8.7 (1)	8.9 (2)
8	8.4 (1.5)	8.5 (3)	8.4 (1.5)	8.6 (4)
9	9.2 (2)	9.6 (4)	8.9 (1)	9.5 (3)
10	8.8 (2)	9.2 (3)	8.6 (1)	9.3 (4)

rangos es la siguiente: $R_1 = 25$, $R_2 = 34.5$, $R_3 = 14$, $R_4 = 26.5$. El valor de la estadística de Fiedman es

$$s = \frac{12}{(10)(4)(5)}[25^2 + 34.5^2 + 14^2 + 26.5^2] - (3)(10)(5) = 12.81.$$

Para $\alpha = 0.01$ y $k - 1 = 3$ grados de libertad, el valor crítico se obtiene de la tabla E del apéndice y es igual a 11.32. Dado que $s = 12.81 > 11.32$, se rechaza la hipótesis nula de que los efectos de los tratamientos son los mismos; las diferencias entre las calificaciones que otorgan los cuatro jueces son estadísticamente discernibles.

15.6 Coeficiente de correlación de rangos de Spearman

En la sección 13.8 se definió el coeficiente de correlación de la muestra como una medida de la asociación lineal que existe entre dos variables X y Y. El enfoque empleado en esa sección fue paramétrico, ya que se supuso una distribución normal bivariada para X y Y. En esta sección se define una popular medida no paramétrica de asociación cuando se emplean los rangos, que se conoce como coeficiente de correlación de rangos de Spearman, denotado por r_S.

Sean X y Y dos características de interés y supóngase que existe una muestra aleatoria de n pares que consiste sólo en los rangos de X y Y. El coeficiente de correlación del rango de Spearman es el coeficiente ordinario de correlación de la muestra que puede determinarse mediante el empleo, ya sea de (13.27) o (13.28), excepto que para este caso se emplean los rangos en lugar de las observaciones de X y Y. Al igual que el coeficiente de correlación de la muestra r, el coeficiente de correlación del rango r_S se define en el intervalo $-1 \leq r_S \leq 1$; y mide el grado de asociación lineal entre los rangos de X y Y. Para las características X y Y, la interpretación de r_S no es completamente idéntica a la de r. Si se tienen disponibles observaciones de X y Y, entonces el coeficiente de correlación de la muestra r es una medida del grado de asociación lineal que existe entre X y Y. Pero si se emplean los rangos, r_S mide la ten-

dencia de X y Y al relacionarse en forma monótona, es decir, r_S se encuentra cercano a 1 o a -1, se sugiere una asociación monótona creciente o decreciente para X y Y. En cierto sentido, r_S tiene un significado mayor que el de r debido a que al medir el grado de asociación monótona entre X y Y, r_S no se encuentra restringido a descubrir sólo una asociación lineal entre éstas.

Sea (X_i', Y_i'), $i = 1, 2, \ldots, n$ la representación de una muestra de rangos de X y Y. Entonces, de (13.28) el coeficiente de correlación de rangos de Spearman es

$$r_S = \frac{\sum_{i=1}^{n} X_i' Y_i' - \frac{\left(\sum_{i=1}^{n} X_i'\right)\left(\sum_{i=1}^{n} Y_i'\right)}{n}}{\left[\sum_{i=1}^{n} X_i'^2 - \frac{\left(\sum_{i=1}^{n} X_i'\right)^2}{n}\right]^{1/2}\left[\sum_{i=1}^{n} Y_i'^2 - \frac{\left(\sum_{i=1}^{n} Y_i'\right)^2}{n}\right]^{1/2}} \tag{15.15}$$

Si no existen empates puede desarrollarse una relación alternativa más simple para (15.15) al tomar ventajas de la naturaleza del rango. Los rangos (X_i', Y_i') son arreglos de los primeros n enteros positivos. Dado que la suma de los primeros n enteros positivos es $n(n + 1)/2$, y la suma de sus cuadrados es $n(n + 1)(2n + 1)/6$,

$$\sum X_i' = \sum Y_i' = n(n + 1)/2 \tag{15.16}$$

y

$$\sum X_i'^2 = \sum Y_i'^2 = n(n + 1)(2n + 1)/6. \tag{15.17}$$

Además, dado que la relación

$$\sum X_i' Y_i' = \left[\sum X_i'^2 + \sum Y_i'^2 - \sum (X_i' - Y_i')^2\right]/2 \tag{15.18}$$

es válida para cualquier valor, al sustituir (15.16) a (15.18) en (15.15) y después de algunos manejos algebráicos, se obtiene la expresión alternativa

$$r_S = 1 - \frac{6\sum (X_i' - Y_i')^2}{n(n^2 - 1)}. \tag{15.19}$$

Ejemplo 15.6 Se pide a dos catadores de vinos que clasifiquen 10 vinos tintos ligeros en una escala del 1 (pobre) al 10 (excelente). Se obtienen los resultados que se muestran en la tabla 15.6. Calcúlese el coeficiente de correlación de rangos de Spearman.

Dado que no existen empates, puede usarse (15.19) para calcular r_S.

$$r_S = 1 - \frac{6[(5 - 3)^2 + (2 - 4)^2 + \cdots + (3 - 1)^2]}{10(100 - 1)} = 0.73,$$

lo cual sugiere una fuerte concordancia entre los dos catadores.

TABLA 15.6 Datos de la muestra para el ejemplo 15.6

Vino	*Catador 1* X'	*Catador 2* Y'
1	5	3
2	2	4
3	8	7
4	9	6
5	10	9
6	7	9
7	1	3
8	4	6
9	4	7
10	3	1

15.7 Comentarios finales

Para los métodos presentados en este capítulo, se tienen tres ventajas:

1. Las suposiciones para su empleo son menos estrictas que las de los correspondientes métodos paramétricos.
2. Los métodos no paramétricos pueden aplicarse en forma muy fácil a todas aquellas observaciones que se definen sobre una escala ordinal.
3. Los cálculos por efectuar son más fáciles cuando se comparan con los de los correspondientes métodos paramétricos.

A causa de la primera ventaja, los métodos paramétricos son particularmente útiles cuando se tienen muestras de tamaño pequeño y existe interés en adherirse a las suposiciones de distribución para los métodos paramétricos. En particular, las pruebas de Mann-Whitney, Wilcoxon, Kruskal-Wallis y de Friedman se comparan, en potencia, a las de los correspondientes métodos paramétricos, lo que incluye a la distribución t de Student o a la estadística F en el análisis de varianza, pero como ya se indicó en el capítulo 9, para muestras de tamaño mayor de 15, la distribución t de Student es bastante más robusta con respecto a la suposición de normalidad. Además, la estadística T es robusta con respecto a la suposición de varianzas iguales para muestras de gran tamaño y con el mismo número de observaciones, cuando se comparan dos medias, de la misma manera en que la estadística F lo es en el análisis de varianza, siempre y cuando los tamaños de la muestra de los tratamientos sean los mismos. De esta forma, cuando se tienen muestras de gran tamaño y las observaciones contenidas en éstas se definen por lo menos sobre una escala ordinal, puede perderse información muy importante al convertir las observaciones en rangos y signos y utilizar métodos no paramétricos. Para tales casos, la eficiencia en potencia de los métodos no paramétricos es menor que la de los procedimientos paramétricos. Por lo tanto, la ventaja más clara que tienen los métodos no paramétricos sobre los de tipo paramétrico es la segunda que se encuentra en la lista mencionada con anterioridad. La aplicación de los métodos paramétricos a observaciones que se en-

cuentran definidas sólo sobre una escala ordinal es muy difícil, ya que la interpretación de un intervalo en este caso tiene poco significado.

Referencias

1. J. D. Gibbons, *Nonparametric statistical inference*, McGraw-Hill, New York, 1971.
2. M. Hollander and D. A. Wolfe, *Nonparametric statistical methods*, Wiley, New York, 1973.
3. S. Siegel, *Nonparametric statistics for the behavioral sciences*, McGraw-Hill, New York, 1956.

Ejercicios

15.1. Para los datos del ejemplo 15.1, pruébese la hipótesis nula de que no existe diferencia entre las medias mediante el empleo del procedimiento t de Student de la sección 9.6.2. Para el mismo tamaño del error de tipo I dado en este ejemplo, ¿es diferente la conclusión?

15.2. Durante cinco años se llevó a cabo un estudio para determinar si existe alguna diferencia en el número de resfriados que sufren los fumadores y los no fumadores. Con base en muestras aleatorias de 14 no fumadores y 12 fumadores se observaron, a lo largo de los cinco años, los siguientes datos.

No fumadores	1	0	2	7	3	1	2	2	4	3	5	0	2	1
Fumadores	4	2	6	5	8	10	8	7	6	4	9	3		

Úsese la estadística U de Mann-Whitney para determinar si existe alguna razón para creer que estas muestras aleatorias provienen de poblaciones con diferentes distribuciones. Supóngase que $\alpha = 0.05$. ¿Existen algunas suposiciones que se estén violando?

15.3. Una compañía de mercadotecnia se interesa en comparar la aceptación por parte del consumidor de dos nuevos productos, A y B. Se seleccionaron, en forma aleatoria, 12 consumidores y se les pidió que dieran su opinión, con respecto al producto A, sobre una escala de 1 (Poca aceptación) a 5 (mucha aceptación). Se hizo lo mismo para el producto B, empleando para ello el mismo número de consumidores. Se obtuvieron los siguientes datos:

Producto A	1	2	5	5	4	3	5	4	4	3	5	2
Producto B	2	2	1	1	3	1	2	2	4	3	1	3

Mediante el empleo de la estadística U de Mann-Whitney, determínese si puede rechazarse, con $\alpha = 0.05$ la hipótesis nula de que estas muestras aleatorias provienen de poblaciones con distribuciones idénticas.

15.4. La siguiente información representa el número de unidades terminadas para dos trabajadores, A y B, en un periodo de cinco días.

A	49	52	53	47	50
B	56	48	58	46	55

a) Mediante el uso de la expresión (15.7), obténgase la función de probabilidad para el número de tendencias posible.

b) Para $\alpha = 0.05$, empléese el procedimiento de tendencias de Wald-Wolfowitz para probar la hipótesis nula de que estas muestras provienen de distribuciones idénticas.

15.5. El procedimiento de tendencias de Wald-Wolfowitz se emplea muchas veces para probar la aleatoriedad de una secuencia dada de observaciones. Si la aleatoriedad existe, entonces el número de tendencias para dos grupos distintos no deberá ser ni muy grande ni muy pequeño. Supóngase que los siguientes datos constituyen la secuencia de residuos para una ecuación de regresión estimada:

−2.98	−4.19	−0.51	5.19	2.38	6.73	0.93	1.29	−3.18
−1.14	−0.54	−2.76	−1.89	−4.28	−0.18	0.32	0.48	1.48
−2.43	−4.69	3.18	0.64	0.89	2.08	0.98	−3.28	

¿Existe alguna razón para creer que esta secuencia de residuos no es aleatoria? Úsese $\alpha = 0.05$.

15.6. Una compañía de mercadotecnia se interesa en la preferencia del consumidor con respecto a dos marcas de refresco que compiten entre sí. Se seleccionan, en forma aleatoria, 14 personas y se les pide que clasifiquen las bebidas mediante una escala del 1 (poca aceptación) al 10 (mucha aceptación). El orden en la selección de la bebida fue aleatorio. Se obtiene la siguiente información:

Persona	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*
Marca A	7	5	9	4	8	10	4	3	7	2	8	6	6	9
Marca B	3	2	7	6	9	3	5	1	4	2	4	7	5	4

Mediante el uso de la prueba del signo, ¿se tiene alguna razón para creer que existe una diferencia en la preferencia para estos dos refrescos? Supóngase $\alpha = 0.1$.

15.7. Para los datos que figuran en el ejercicio 15.6, empléese la prueba de rangos de signos de Wilcoxon. ¿Se obtienen las mismas conclusiones?

15.8. Para los datos del ejemplo 9.10, supóngase que no puede formularse la suposición de normalidad. Mediante el empleo de la prueba de rangos de signos de Wilcoxon, determínese si puede rechazarse la correspondiente hipótesis nula no paramétrica para un nivel $\alpha = 0.01$

15.9. Durante 12 días seleccionados al azar, dos tiendas, A y B vendieron el siguiente número de unidades del mismo producto:

Día	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
A	42	58	47	39	41	56	59	37	38	46	43	51
B	64	57	48	59	64	52	65	59	37	65	68	49

Mediante el empleo de la prueba del signo, ¿puede rechazarse la hipótesis nula de que las muestras provienen de distribuciones idénticas para un nivel $\alpha = 0.05$?

15.10. Para los datos que figuran en el ejercicio 15.9, úsese la prueba de rangos de signos de Wilcoxon y compárense los resultados.

15.11. Se desea determinar si el campo de especialización del estudiante no graduado tiene algún efecto sobre su desempeño en una escuela de leyes. Se toma una muestra aleatoria

TABLA 15.7 Datos de la muestra para el ejemplo 15.11

Finanzas	*Ciencia o ingeniería*	*Artes liberales*	*Otros*
9	3	2	14
22	7	4	34
24	10	15	48
31	18	26	52
47	23	38	59
65	25	43	63
		45	67
		49	72
		55	79

de 30 estudiantes de una clase de graduados de cierta escuela de leyes, la cual clasifica a los estudiantes y anota su campo de especialización; los datos que se encuentran en la tabla 15.7 son los resultados de este procedimiento. Mediante el empleo de la prueba de Kruskal-Wallis, determínese si el campo de especialización tiene algún efecto sobre el desempeño en la escuela de leyes, con $\alpha = 0.05$.

15.12. Con referencia a los datos que se encuentran en el ejercicio 12.7, empléese el procedimiento de Kruskal-Wallis para probar la hipótesis nula de que no existe ninguna diferencia con respecto a la durabilidad entre las dos marcas con $\alpha = 0.05$. La conclusión, ¿es la misma que la que se obtuvo para el ejercicio 12.7?

15.13. Se seleccionaron 12 estudiantes al azar, de una clase muy grande; sus calificaciones, en los cuatro exámenes que se llevaron a cabo durante el trimestre, se encuentran en la tabla 15.8. Mediante el uso de la prueba de Friedman, determínese si las diferencias entre los cuatro exámenes son estadísticamente discernibles para un nivel $\alpha = 0.01$. ¿Se estaría de acuerdo con la hipótesis de que no existe interacción alguna entre los estudiantes? Coméntese.

TABLA 15.8 Datos de la muestra para el ejercicio 15.13

Estudiante	*1*	*2*	*3*	*4*
1	72	68	80	75
2	89	87	78	92
3	48	56	64	58
4	65	76	70	62
5	86	94	93	85
6	56	73	78	87
7	75	84	65	69
8	39	45	48	56
9	78	67	69	59
10	98	87	86	95
11	64	87	92	48
12	82	76	85	79

15.14. Con referencia a los datos del ejercicio 12.6, úsese el procedimiento de Friedman para determinar si las diferencias que existen entre los cuatro supermercados son estadísticamente discernibles para un nivel $\alpha = 0.01$.

15.15. Para el ejercicio 13.12, conviértanse los datos en rangos y calcúlese el coeficiente de correlación de rangos de Spearman.

15.16. Dos jueces se encargan de calificar a ocho patinadores que patinan sobre hielo, mediante el empleo de una escala del 1 (muy malo) al 10 (el mejor). Se obtienen los siguiente resultados.

Patinador	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Juez 1	3	4	8	8	4	6	4	7
Juez 2	2	4	9	7	2	8	7	9

Calcúlese el coeficiente de correlación de rangos de Spearman y formúlese un comentario con respecto a si existe una relación que sea evidente.

15.17. Mediante el empleo de la misma escala que se menciona en el ejercicio 15.16, dos jueces califican el talento de las 10 semifinalistas del concurso señorita América. Se tienen los siguientes resultados:

Semifinalista	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Juez 1	2	6	5	9	3	7	9	2	6	2
Juez 2	7	1	4	4	8	9	3	9	10	8

Calcúlese el coeficiente de correlación de rangos de Spearman y formúlese un comentario con respecto a si existe alguna relación que sea evidente.

15.18. Un grupo de analistas de inversión clasificaron 10 compañías de acuerdo con su crecimiento potencial y el valor de sus acciones de la siguiente manera:

Compañía	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Valores en libros	8	3	10	1	6	2	5	7	4	9
Crecimiento	4	8	6	5	9	3	7	1	10	2

Calcúlese el coeficiente de correlación de rangos de Spearman y formúlese un comentario con respecto a si existe una relación, que sea evidente, entre el valor de las acciones de la compañía y su crecimiento potencial.

Apéndice

TABLA A Valores de la función de distribución acumulativa binomial
TABLA B Valores de la función de distribución acumulativa de Poisson
TABLA C Valores de las funciones de probabilidad y distribución acumulativa para la distribución hipergeométrica
TABLA D Valores de la función de distribución acumulativa normal estándar
TABLA E Valores de cuantiles de la distribución chi-cuadrada
TABLA F Valores de cuantiles de la distribución t de Student
TABLA G Valores de cuantiles de la distribución F
TABLA H k-valores para los límites de tolerancia bilaterales cuando se muestrean distribuciones normales
TABLA I k-valores para los límites de tolerancia unilaterales cuando se muestrean distribuciones normales
TABLA J Valores de cuantiles superiores de la distribución de la estadística D_n de Kolmogorov-Smirnov
TABLA K Límites de la estadística de Durbin-Watson

TABLA A Valores de la función de distribución acumulativa binomial

$$P(X \leq x) = F(x; n, p) = \sum_{i=0}^{x} \binom{n}{i} p^i (1 - p)^{n-i}$$

		p										
n	x	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
2	0	0.9801	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
	1	0.9999	0.9975	0.9900	0.9775	0.9600	0.9375	0.9100	0.8775	0.8400	0.7975	0.7500
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	0	0.9703	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
	1	0.9997	0.9928	0.9720	0.9392	0.8960	0.8438	0.7840	0.7183	0.6480	0.5748	0.5000
	2	1.0000	0.9999	0.9990	0.9966	0.9920	0.9844	0.9730	0.9571	0.9360	0.9089	0.8750
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	0	0.9606	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
	1	0.9994	0.9860	0.9477	0.8905	0.8192	0.7383	0.6517	0.5630	0.4752	0.3910	0.3125
	2	1.0000	0.9995	0.9963	0.9880	0.9728	0.9492	0.9163	0.8735	0.8208	0.7585	0.6875
	3	1.0000	1.0000	0.9999	0.9995	0.9984	0.9961	0.9919	0.9850	0.9744	0.9590	0.9375
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	0	0.9510	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0313
	1	0.9990	0.9774	0.9185	0.8352	0.7373	0.6328	0.5282	0.4284	0.3370	0.2562	0.1875
	2	1.0000	0.9988	0.9914	0.9734	0.9421	0.8965	0.8369	0.7648	0.6826	0.5931	0.5000
	3	1.0000	1.0000	0.9995	0.9978	0.9933	0.9844	0.9692	0.9460	0.9130	0.8688	0.8125
	4	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990	0.9976	0.9947	0.9898	0.9815	0.9688
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	0	0.9415	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
	1	0.9985	0.9672	0.8857	0.7765	0.6554	0.5339	0.4202	0.3191	0.2333	0.1636	0.1094
	2	1.0000	0.9978	0.9842	0.9527	0.9011	0.8306	0.7443	0.6471	0.5443	0.4415	0.3438
	3	1.0000	0.9999	0.9987	0.9941	0.9830	0.9624	0.9295	0.8826	0.8208	0.7447	0.6563
	4	1.0000	1.0000	0.9999	0.9996	0.9984	0.9954	0.9891	0.9777	0.9590	0.9308	0.8906
	5	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9982	0.9959	0.9917	0.9844
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

7	0	0.9321	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
	1	0.9980	0.9556	0.8503	0.7166	0.5767	0.4449	0.3294	0.2338	0.1586	0.1024	0.0625
	2	1.0000	0.9962	0.9743	0.9262	0.8520	0.7564	0.6471	0.5323	0.4199	0.3164	0.2266
	3	1.0000	0.9998	0.9973	0.9879	0.9667	0.9294	0.8740	0.8002	0.7102	0.6083	0.5000
	4	1.0000	1.0000	0.9998	0.9988	0.9953	0.9871	0.9712	0.9444	0.9037	0.8471	0.7734
	5	1.0000	1.0000	1.0000	0.9999	0.9996	0.9987	0.9962	0.9910	0.9812	0.9643	0.9375
	6	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9984	0.9963	0.9922
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
8	0	0.9227	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
	1	0.9973	0.9428	0.8131	0.6572	0.5033	0.3671	0.2553	0.1691	0.1064	0.0632	0.0352
	2	0.9999	0.9942	0.9619	0.8948	0.7969	0.6785	0.5518	0.4278	0.3154	0.2201	0.1445
	3	1.0000	0.9996	0.9950	0.9786	0.9437	0.8862	0.8059	0.7064	0.5941	0.4770	0.3633
	4	1.0000	1.0000	0.9996	0.9971	0.9896	0.9727	0.9420	0.8939	0.8263	0.7396	0.6367
	5	1.0000	1.0000	1.0000	0.9998	0.9988	0.9958	0.9887	0.9747	0.9502	0.9115	0.8555
	6	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9987	0.9964	0.9915	0.9819	0.9648
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9983	0.9961
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	0	0.9135	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
	1	0.9966	0.9288	0.7748	0.5995	0.4362	0.3003	0.1960	0.1211	0.0705	0.0385	0.0195
	2	0.9999	0.9916	0.9470	0.8591	0.7382	0.6007	0.4628	0.3373	0.2318	0.1495	0.0898
	3	1.0000	0.9994	0.9917	0.9661	0.9144	0.8343	0.7297	0.6089	0.4826	0.3614	0.2539
	4	1.0000	1.0000	0.9991	0.9944	0.9804	0.9511	0.9012	0.8283	0.7334	0.6214	0.5000
	5	1.0000	1.0000	0.9999	0.9994	0.9969	0.9900	0.9747	0.9464	0.9006	0.8342	0.7461
	6	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9957	0.9888	0.9750	0.9502	0.9102
	7	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9986	0.9962	0.9909	0.9805
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9980
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	0	0.9044	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
	1	0.9957	0.9139	0.7361	0.5443	0.3758	0.2440	0.1493	0.0860	0.0464	0.0233	0.0107
	2	0.9999	0.9885	0.9298	0.8202	0.6778	0.5256	0.3828	0.2616	0.1673	0.0996	0.0547
	3	1.0000	0.9990	0.9872	0.9500	0.8791	0.7759	0.6496	0.5138	0.3823	0.2660	0.1719
	4	1.0000	0.9999	0.9984	0.9901	0.9672	0.9219	0.8497	0.7515	0.6331	0.5044	0.3770
	5	1.0000	1.0000	0.9999	0.9986	0.9936	0.9803	0.9527	0.9051	0.8338	0.7384	0.6230
	6	1.0000	1.0000	1.0000	0.9999	0.9991	0.9965	0.9894	0.9740	0.9452	0.8980	0.8281
	7	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9984	0.9952	0.9877	0.9726	0.9453
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983	0.9955	0.9893
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

TABLA A (continuación) Valores de la función de distribución acumulativa binomial

		p										
n	x	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
11	0	0.8953	0.5688	0.3138	0.1673	0.0859	0.0422	0.0198	0.0088	0.0036	0.0014	0.0005
	1	0.9948	0.8981	0.6974	0.4922	0.3221	0.1971	0.1130	0.0606	0.0302	0.0139	0.0059
	2	0.9998	0.9848	0.9104	0.7788	0.6174	0.4552	0.3127	0.2001	0.1189	0.0652	0.0327
	3	1.0000	0.9984	0.9815	0.9306	0.8389	0.7133	0.5696	0.4256	0.2963	0.1911	0.1133
	4	1.0000	0.9999	0.9972	0.9841	0.9496	0.8854	0.7897	0.6683	0.5328	0.3971	0.2744
	5	1.0000	1.0000	0.9997	0.9973	0.9883	0.9657	0.9218	0.8513	0.7535	0.6331	0.5000
	6	1.0000	1.0000	1.0000	0.9997	0.9980	0.9924	0.9784	0.9499	0.9006	0.8262	0.7256
	7	1.0000	1.0000	1.0000	1.0000	0.9998	0.9988	0.9957	0.9878	0.9707	0.9390	0.8867
	8	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9980	0.9941	0.9852	0.9673
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9993	0.9978	0.9941
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9995
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
12	0	0.8864	0.5404	0.2824	0.1422	0.0687	0.0317	0.0138	0.0057	0.0022	0.0008	0.0002
	1	0.9938	0.8816	0.6590	0.4435	0.2749	0.1584	0.0850	0.0424	0.0196	0.0083	0.0032
	2	0.9998	0.9804	0.8891	0.7358	0.5583	0.3907	0.2528	0.1513	0.0834	0.0421	0.0193
	3	1.0000	0.9978	0.9744	0.9078	0.7946	0.6488	0.4925	0.3467	0.2253	0.1345	0.0730
	4	1.0000	0.9998	0.9957	0.9761	0.9274	0.8424	0.7237	0.5833	0.4382	0.3044	0.1938
	5	1.0000	1.0000	0.9995	0.9954	0.9806	0.9456	0.8822	0.7873	0.6652	0.5269	0.3872
	6	1.0000	1.0000	0.9999	0.9993	0.9961	0.9857	0.9614	0.9154	0.8418	0.7393	0.6128
	7	1.0000	1.0000	1.0000	0.9999	0.9994	0.9972	0.9905	0.9745	0.9427	0.8883	0.8062
	8	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9983	0.9944	0.9847	0.9644	0.9270
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9992	0.9972	0.9921	0.9807
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9968
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

13	0	0.8775	0.5133	0.2542	0.1209	0.0550	0.0238	0.0097	0.0037	0.0013	0.0004	0.0001
	1	0.9928	0.8646	0.6213	0.3983	0.2336	0.1267	0.0637	0.0296	0.0126	0.0049	0.0017
	2	0.9997	0.9755	0.8661	0.6920	0.5017	0.3326	0.2025	0.1132	0.0579	0.0269	0.0112
	3	1.0000	0.9969	0.9658	0.8820	0.7473	0.5843	0.4206	0.2783	0.1686	0.0929	0.0461
	4	1.0000	0.9997	0.9935	0.9658	0.9009	0.7940	0.6543	0.5005	0.3530	0.2279	0.1334
	5	1.0000	1.0000	0.9991	0.9925	0.9700	0.9198	0.8346	0.7159	0.5744	0.4268	0.2905
	6	1.0000	1.0000	0.9999	0.9987	0.9930	0.9757	0.9376	0.8705	0.7712	0.6437	0.5000
	7	1.0000	1.0000	1.0000	0.9998	0.9988	0.9944	0.9818	0.9538	0.9023	0.8212	0.7095
	8	1.0000	1.0000	1.0000	1.0000	0.9998	0.9990	0.9960	0.9874	0.9679	0.9302	0.8666
	9	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9975	0.9922	0.9797	0.9539
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9987	0.9959	0.9888
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
14	0	0.8687	0.4877	0.2288	0.1028	0.0440	0.0178	0.0068	0.0024	0.0008	0.0002	0.0001
	1	0.9916	0.8470	0.5846	0.3567	0.1979	0.1010	0.0475	0.0205	0.0081	0.0029	0.0009
	2	0.9997	0.9699	0.8416	0.6479	0.4481	0.2811	0.1608	0.0839	0.0398	0.0170	0.0065
	3	1.0000	0.9958	0.9559	0.8535	0.6982	0.5213	0.3552	0.2205	0.1243	0.0632	0.0287
	4	1.0000	0.9996	0.9908	0.9533	0.8702	0.7415	0.5842	0.4227	0.2793	0.1672	0.0898
	5	1.0000	1.0000	0.9985	0.9885	0.9561	0.8883	0.7805	0.6405	0.4859	0.3373	0.2120
	6	1.0000	1.0000	0.9998	0.9978	0.9884	0.9617	0.9067	0.8164	0.6925	0.5461	0.3953
	7	1.0000	1.0000	1.0000	0.9997	0.9976	0.9897	0.9685	0.9247	0.8499	0.7414	0.6047
	8	1.0000	1.0000	1.0000	1.0000	0.9996	0.9978	0.9917	0.9757	0.9417	0.8811	0.7880
	9	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9983	0.9940	0.9825	0.9574	0.9102
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9989	0.9961	0.9886	0.9713
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9978	0.9935
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9991
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

TABLA A (continuación) Valores de la función de distribución acumulativa binomial

		p										
n	x	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
15	0	0.8601	0.4633	0.2059	0.0874	0.0352	0.0134	0.0047	0.0016	0.0005	0.0001	0.0000
	1	0.9904	0.8290	0.5490	0.3186	0.1671	0.0802	0.0353	0.0142	0.0052	0.0017	0.0005
	2	0.9996	0.9638	0.8159	0.6042	0.3980	0.2361	0.1268	0.0617	0.0271	0.0107	0.0037
	3	1.0000	0.9945	0.9444	0.8227	0.6482	0.4613	0.2969	0.1727	0.0905	0.0424	0.0176
	4	1.0000	0.9994	0.9873	0.9383	0.8358	0.6865	0.5155	0.3519	0.2173	0.1204	0.0592
	5	1.0000	0.9999	0.9978	0.9832	0.9389	0.8516	0.7216	0.5643	0.4032	0.2608	0.1509
	6	1.0000	1.0000	0.9997	0.9964	0.9819	0.9434	0.8689	0.7548	0.6098	0.4522	0.3036
	7	1.0000	1.0000	1.0000	0.9994	0.9958	0.9827	0.9500	0.8868	0.7869	0.6535	0.5000
	8	1.0000	1.0000	1.0000	0.9999	0.9992	0.9958	0.9848	0.9578	0.9050	0.8182	0.6964
	9	1.0000	1.0000	1.0000	1.0000	0.9999	0.9992	0.9963	0.9876	0.9662	0.9231	0.8491
	10	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9972	0.9907	0.9745	0.9408
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9981	0.9937	0.9824
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9989	0.9963
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
16	0	0.8515	0.4401	0.1853	0.0743	0.0281	0.0100	0.0033	0.0010	0.0003	0.0001	0.0000
	1	0.9891	0.8108	0.5147	0.2839	0.1407	0.0635	0.0261	0.0098	0.0033	0.0010	0.0003
	2	0.9995	0.9571	0.7893	0.5614	0.3518	0.1971	0.0994	0.0451	0.0183	0.0066	0.0021
	3	1.0000	0.9930	0.9316	0.7899	0.5981	0.4050	0.2459	0.1339	0.0651	0.0281	0.0106
	4	1.0000	0.9991	0.9830	0.9209	0.7982	0.6302	0.4499	0.2892	0.1666	0.0853	0.0384
	5	1.0000	0.9999	0.9967	0.9765	0.9183	0.8103	0.6598	0.4900	0.3288	0.1976	0.1051
	6	1.0000	1.0000	0.9995	0.9944	0.9733	0.9204	0.8247	0.6881	0.5272	0.3660	0.2272
	7	1.0000	1.0000	0.9999	0.9989	0.9930	0.9729	0.9256	0.8406	0.7161	0.5629	0.4018
	8	1.0000	1.0000	1.0000	0.9998	0.9985	0.9925	0.9743	0.9329	0.8577	0.7441	0.5982
	9	1.0000	1.0000	1.0000	1.0000	0.9998	0.9984	0.9929	0.9771	0.9417	0.8759	0.7728
	10	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9984	0.9938	0.9809	0.9514	0.8949
	11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9951	0.9851	0.9616
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9991	0.9965	0.9894
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9979
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

17	0	0.8429	0.4181	0.1668	0.0631	0.0225	0.0075	0.0023	0.0007	0.0002	0.0000	0.0000
	1	0.9877	0.7922	0.4818	0.2525	0.1182	0.0501	0.0193	0.0067	0.0021	0.0006	0.0001
	2	0.9994	0.9497	0.7618	0.5198	0.3096	0.1637	0.0774	0.0327	0.0123	0.0041	0.0012
	3	1.0000	0.9912	0.9174	0.7556	0.5489	0.3530	0.2019	0.1028	0.0464	0.0184	0.0064
	4	1.0000	0.9988	0.9779	0.9013	0.7582	0.5739	0.3887	0.2348	0.1260	0.0596	0.0245
	5	1.0000	0.9999	0.9953	0.9681	0.8943	0.7653	0.5968	0.4197	0.2639	0.1471	0.0717
	6	1.0000	1.0000	0.9992	0.9917	0.9623	0.8929	0.7752	0.6188	0.4478	0.2902	0.1662
	7	1.0000	1.0000	0.9999	0.9983	0.9891	0.9598	0.8954	0.7872	0.6405	0.4743	0.3145
	8	1.0000	1.0000	1.0000	0.9997	0.9974	0.9876	0.9597	0.9006	0.8011	0.6626	0.5000
	9	1.0000	1.0000	1.0000	1.0000	0.9995	0.9969	0.9873	0.9617	0.9081	0.8166	0.6855
	10	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9968	0.9880	0.9652	0.9174	0.8338
	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9970	0.9894	0.9699	0.9283
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9975	0.9914	0.9755
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9981	0.9936
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9988
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
18	0	0.8345	0.3972	0.1501	0.0536	0.0180	0.0056	0.0016	0.0004	0.0001	0.0000	0.0000
	1	0.9862	0.7735	0.4503	0.2241	0.0991	0.0395	0.0142	0.0046	0.0013	0.0003	0.0001
	2	0.9993	0.9419	0.7338	0.4797	0.2713	0.1353	0.0600	0.0236	0.0082	0.0025	0.0007
	3	1.0000	0.9891	0.9018	0.7202	0.5010	0.3057	0.1646	0.0783	0.0328	0.0120	0.0038
	4	1.0000	0.9985	0.9718	0.8794	0.7164	0.5187	0.3327	0.1886	0.0942	0.0411	0.0154
	5	1.0000	0.9998	0.9936	0.9581	0.8671	0.7175	0.5344	0.3550	0.2088	0.1077	0.0481
	6	1.0000	1.0000	0.9988	0.9882	0.9487	0.8610	0.7217	0.5491	0.3743	0.2258	0.1189
	7	1.0000	1.0000	0.9998	0.9973	0.9837	0.9431	0.8593	0.7283	0.5634	0.3915	0.2403
	8	1.0000	1.0000	1.0000	0.9995	0.9957	0.9807	0.9404	0.8609	0.7368	0.5778	0.4073
	9	1.0000	1.0000	1.0000	0.9999	0.9991	0.9946	0.9790	0.9403	0.8653	0.7473	0.5927
	10	1.0000	1.0000	1.0000	1.0000	0.9998	0.9988	0.9939	0.9788	0.9424	0.8720	0.7597
	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9986	0.9938	0.9797	0.9463	0.8811
	12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9986	0.9942	0.9817	0.9519
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9951	0.9846
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9990	0.9962
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

TABLA A (continuación) Valores de la función de distribución acumulativa binomial

		p										
n	x	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
19	0	0.8262	0.3774	0.1351	0.0456	0.0144	0.0042	0.0011	0.0003	0.0001	0.0000	0.0000
	1	0.9847	0.7547	0.4203	0.1985	0.0829	0.0310	0.0104	0.0031	0.0008	0.0002	0.0000
	2	0.9991	0.9335	0.7054	0.4413	0.2369	0.1113	0.0462	0.0170	0.0055	0.0015	0.0004
	3	1.0000	0.9868	0.8850	0.6841	0.4551	0.2631	0.1332	0.0591	0.0230	0.0077	0.0022
	4	1.0000	0.9980	0.9648	0.8556	0.6733	0.4654	0.2822	0.1500	0.0696	0.0280	0.0096
	5	1.0000	0.9998	0.9914	0.9463	0.8369	0.6678	0.4739	0.2968	0.1629	0.0777	0.0318
	6	1.0000	1.0000	0.9983	0.9837	0.9324	0.8251	0.6655	0.4812	0.3081	0.1727	0.0835
	7	1.0000	1.0000	0.9997	0.9959	0.9767	0.9225	0.8180	0.6656	0.4878	0.3169	0.1796
	8	1.0000	1.0000	1.0000	0.9992	0.9933	0.9713	0.9161	0.8145	0.6675	0.4940	0.3238
	9	1.0000	1.0000	1.0000	0.9999	0.9984	0.9911	0.9674	0.9125	0.8139	0.6710	0.5000
	10	1.0000	1.0000	1.0000	1.0000	0.9997	0.9977	0.9895	0.9653	0.9115	0.8159	0.6762
	11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9972	0.9886	0.9648	0.9129	0.8204
	12	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9969	0.9884	0.9658	0.9165
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9969	0.9891	0.9682
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9972	0.9904
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9978
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
20	0	0.8179	0.3585	0.1216	0.0388	0.0115	0.0032	0.0008	0.0002	0.0000	0.0000	0.0000
	1	0.9831	0.7358	0.3917	0.1756	0.0692	0.0243	0.0076	0.0021	0.0005	0.0001	0.0000
	2	0.9990	0.9245	0.6769	0.4049	0.2061	0.0913	0.0355	0.0121	0.0036	0.0009	0.0002
	3	1.0000	0.9841	0.8670	0.6477	0.4114	0.2252	0.1071	0.0444	0.0160	0.0049	0.0013
	4	1.0000	0.9974	0.9568	0.8298	0.6296	0.4148	0.2375	0.1182	0.0510	0.0189	0.0059
	5	1.0000	0.9997	0.9887	0.9327	0.8042	0.6172	0.4164	0.2454	0.1256	0.0553	0.0207
	6	1.0000	1.0000	0.9976	0.9781	0.9133	0.7858	0.6080	0.4166	0.2500	0.1299	0.0577
	7	1.0000	1.0000	0.9996	0.9941	0.9679	0.8982	0.7723	0.6010	0.4159	0.2520	0.1316
	8	1.0000	1.0000	0.9999	0.9987	0.9900	0.9591	0.8867	0.7624	0.5956	0.4143	0.2517
	9	1.0000	1.0000	1.0000	0.9998	0.9974	0.9861	0.9520	0.8782	0.7553	0.5914	0.4119
	10	1.0000	1.0000	1.0000	1.0000	0.9994	0.9961	0.9829	0.9468	0.8725	0.7507	0.5881

	11	1.0000	1.0000	1.0000	1.0000	0.9999	0.9991	0.9949	0.9804	0.9435	0.8692	0.7483
	12	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9987	0.9940	0.9790	0.9420	0.8684
	13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9985	0.9935	0.9786	0.9423
	14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9984	0.9936	0.9793
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9985	0.9941
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
25	0	0.7778	0.2774	0.0718	0.0172	0.0038	0.0008	0.0001	0.0000	0.0000	0.0000	0.0000
	1	0.9742	0.6424	0.2712	0.0931	0.0274	0.0070	0.0016	0.0003	0.0001	0.0000	0.0000
	2	0.9980	0.8729	0.5371	0.2537	0.0982	0.0321	0.0090	0.0021	0.0004	0.0001	0.0000
	3	0.9999	0.9659	0.7636	0.4711	0.2340	0.0962	0.0332	0.0097	0.0024	0.0005	0.0001
	4	1.0000	0.9928	0.9020	0.6821	0.4207	0.2137	0.0905	0.0320	0.0095	0.0023	0.0005
	5	1.0000	0.9988	0.9666	0.8385	0.6167	0.3783	0.1935	0.0826	0.0294	0.0086	0.0020
	6	1.0000	0.9998	0.9905	0.9305	0.7800	0.5611	0.3407	0.1734	0.0736	0.0258	0.0073
	7	1.0000	1.0000	0.9977	0.9745	0.8909	0.7265	0.5118	0.3061	0.1536	0.0639	0.0216
	8	1.0000	1.0000	0.9995	0.9920	0.9532	0.8506	0.6769	0.4668	0.2735	0.1340	0.0539
	9	1.0000	1.0000	0.9999	0.9979	0.9827	0.9287	0.8106	0.6303	0.4246	0.2424	0.1148
	10	1.0000	1.0000	1.0000	0.9995	0.9944	0.9703	0.9022	0.7712	0.5858	0.3843	0.2122
	11	1.0000	1.0000	1.0000	0.9999	0.9985	0.9893	0.9558	0.8746	0.7323	0.5426	0.3450
	12	1.0000	1.0000	1.0000	1.0000	0.9996	0.9966	0.9825	0.9396	0.8462	0.6937	0.5000
	13	1.0000	1.0000	1.0000	1.0000	0.9999	0.9991	0.9940	0.9745	0.9222	0.8173	0.6550
	14	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9982	0.9907	0.9656	0.9040	0.7878
	15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9971	0.9868	0.9560	0.8852
	16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9992	0.9957	0.9826	0.9461
	17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9988	0.9942	0.9784
	18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9984	0.9927
	19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9980
	20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995
	21	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
	22	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	23	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	24	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	25	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

TABLA B Valores de la función de distribución acumulativa de Poisson

$$P(X \leq x) = F(x; \lambda) = \sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^{i}}{i!}$$

	λ									
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725	0.7358
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371	0.9197
3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865	0.9810
4	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977	0.9963
5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9994
6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

	λ									
x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	0.3329	0.3012	0.2725	0.2466	0.2231	0.2019	0.1827	0.1653	0.1496	0.1353
1	0.6990	0.6626	0.6268	0.5918	0.5578	0.5249	0.4932	0.4628	0.4338	0.4060
2	0.9004	0.8795	0.8571	0.8335	0.8088	0.7834	0.7572	0.7306	0.7037	0.6767
3	0.9743	0.9662	0.9569	0.9463	0.9344	0.9212	0.9068	0.8913	0.8747	0.8571
4	0.9946	0.9923	0.9893	0.9857	0.9814	0.9763	0.9704	0.9636	0.9559	0.9473
5	0.9990	0.9985	0.9978	0.9968	0.9955	0.9940	0.9920	0.9896	0.9868	0.9834
6	0.9999	0.9997	0.9996	0.9994	0.9991	0.9987	0.9981	0.9974	0.9966	0.9955
7	1.0000	1.0000	0.9999	0.9999	0.9998	0.9997	0.9996	0.9994	0.9992	0.9989
8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9998
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

	λ									
x	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	0.1225	0.1108	0.1003	0.0907	0.0821	0.0743	0.0672	0.0608	0.0550	0.0498
1	0.3796	0.3546	0.3309	0.3084	0.2873	0.2674	0.2487	0.2311	0.2146	0.1991
2	0.6496	0.6227	0.5960	0.5697	0.5438	0.5184	0.4936	0.4695	0.4460	0.4232
3	0.8386	0.8194	0.7993	0.7787	0.7576	0.7360	0.7141	0.6919	0.6696	0.6472
4	0.9379	0.9275	0.9163	0.9041	0.8912	0.8774	0.8629	0.8477	0.8318	0.8153
5	0.9796	0.9751	0.9700	0.9643	0.9580	0.9510	0.9433	0.9349	0.9258	0.9161
6	0.9941	0.9925	0.9906	0.9884	0.9858	0.9828	0.9794	0.9756	0.9713	0.9665
7	0.9985	0.9980	0.9974	0.9967	0.9958	0.9947	0.9934	0.9919	0.9901	0.9881
8	0.9997	0.9995	0.9994	0.9991	0.9989	0.9985	0.9981	0.9976	0.9969	0.9962
9	0.9999	0.9999	0.9999	0.9998	0.9997	0.9996	0.9995	0.9993	0.9991	0.9989
10	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9998	0.9998	0.9997
11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999

	λ									
x	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	0.0450	0.0408	0.0369	0.0334	0.0302	0.0273	0.0247	0.0224	0.0202	0.0183
1	0.1847	0.1712	0.1586	0.1468	0.1359	0.1257	0.1162	0.1074	0.0992	0.0916
2	0.4012	0.3799	0.3594	0.3397	0.3208	0.3027	0.2854	0.2689	0.2531	0.2381
3	0.6248	0.6025	0.5803	0.5584	0.5366	0.5152	0.4942	0.4735	0.4533	0.4335
4	0.7982	0.7806	0.7626	0.7442	0.7254	0.7064	0.6872	0.6678	0.6484	0.6288
5	0.9057	0.8946	0.8829	0.8705	0.8576	0.8441	0.8301	0.8156	0.8006	0.7851
6	0.9612	0.9554	0.9490	0.9421	0.9347	0.9267	0.9182	0.9091	0.8995	0.8893
7	0.9858	0.9832	0.9802	0.9769	0.9733	0.9692	0.9648	0.9599	0.9546	0.9489
8	0.9953	0.9943	0.9931	0.9917	0.9901	0.9883	0.9863	0.9840	0.9815	0.9786
9	0.9986	0.9982	0.9978	0.9973	0.9967	0.9960	0.9952	0.9942	0.9931	0.9919
10	0.9996	0.9995	0.9994	0.9992	0.9990	0.9987	0.9984	0.9981	0.9977	0.9972
11	0.9999	0.9999	0.9998	0.9998	0.9997	0.9996	0.9995	0.9994	0.9993	0.9991
12	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998	0.9997
13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999

TABLA B (continuación) Valores de la función de distribución acumulativa de Poisson

	λ									
x	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	0.0166	0.0150	0.0136	0.0123	0.0111	0.0101	0.0091	0.0082	0.0074	0.0067
1	0.0845	0.0780	0.0719	0.0663	0.0611	0.0563	0.0518	0.0477	0.0439	0.0404
2	0.2238	0.2102	0.1974	0.1851	0.1736	0.1626	0.1523	0.1425	0.1333	0.1247
3	0.4142	0.3954	0.3772	0.3595	0.3423	0.3257	0.3097	0.2942	0.2793	0.2650
4	0.6093	0.5898	0.5704	0.5512	0.5321	0.5132	0.4946	0.4763	0.4582	0.4405
5	0.7693	0.7531	0.7367	0.7199	0.7029	0.6858	0.6684	0.6510	0.6335	0.6160
6	0.8787	0.8675	0.8558	0.8436	0.8311	0.8180	0.8046	0.7908	0.7767	0.7622
7	0.9427	0.9361	0.9290	0.9214	0.9134	0.9050	0.8960	0.8867	0.8769	0.8666
8	0.9755	0.9721	0.9683	0.9642	0.9597	0.9549	0.9497	0.9442	0.9382	0.9319
9	0.9905	0.9889	0.9871	0.9851	0.9829	0.9805	0.9778	0.9749	0.9717	0.9682
10	0.9966	0.9959	0.9952	0.9943	0.9933	0.9922	0.9910	0.9896	0.9880	0.9863
11	0.9989	0.9986	0.9983	0.9980	0.9976	0.9971	0.9966	0.9960	0.9953	0.9945
12	0.9997	0.9996	0.9995	0.9993	0.9992	0.9990	0.9988	0.9986	0.9983	0.9980
13	0.9999	0.9999	0.9998	0.9998	0.9997	0.9997	0.9996	0.9995	0.9994	0.9993
14	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998
15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999

	λ									
x	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
0	0.0061	0.0055	0.0050	0.0045	0.0041	0.0037	0.0033	0.0030	0.0027	0.0025
1	0.0372	0.0342	0.0314	0.0289	0.0266	0.0244	0.0224	0.0206	0.0189	0.0174
2	0.1165	0.1088	0.1016	0.0948	0.0884	0.0824	0.0768	0.0715	0.0666	0.0620
3	0.2513	0.2381	0.2254	0.2133	0.2017	0.1906	0.1801	0.1700	0.1604	0.1512
4	0.4231	0.4061	0.3895	0.3733	0.3575	0.3422	0.3272	0.3127	0.2987	0.2851
5	0.5984	0.5809	0.5635	0.5461	0.5289	0.5119	0.4950	0.4783	0.4619	0.4457
6	0.7474	0.7324	0.7171	0.7017	0.6860	0.6703	0.6544	0.6384	0.6224	0.6063
7	0.8560	0.8449	0.8335	0.8217	0.8095	0.7970	0.7842	0.7710	0.7576	0.7440
8	0.9252	0.9181	0.9106	0.9027	0.8944	0.8857	0.8766	0.8672	0.8574	0.8472
9	0.9644	0.9603	0.9559	0.9512	0.9462	0.9409	0.9352	0.9292	0.9228	0.9161
10	0.9844	0.9823	0.9800	0.9775	0.9747	0.9718	0.9686	0.9651	0.9614	0.9574
11	0.9937	0.9927	0.9916	0.9904	0.9890	0.9875	0.9859	0.9841	0.9821	0.9799
12	0.9976	0.9972	0.9967	0.9962	0.9955	0.9949	0.9941	0.9932	0.9922	0.9912
13	0.9992	0.9990	0.9988	0.9986	0.9983	0.9980	0.9977	0.9973	0.9969	0.9964
14	0.9997	0.9997	0.9996	0.9995	0.9994	0.9993	0.9991	0.9990	0.9988	0.9986
15	0.9999	0.9999	0.9999	0.9998	0.9998	0.9998	0.9997	0.9996	0.9996	0.9995
16	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9998
17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999

	λ									
x	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
0	0.0022	0.0020	0.0018	0.0017	0.0015	0.0014	0.0012	0.0011	0.0010	0.0009
1	0.0159	0.0146	0.0134	0.0123	0.0113	0.0103	0.0095	0.0087	0.0080	0.0073
2	0.0577	0.0536	0.0498	0.0463	0.0430	0.0400	0.0371	0.0344	0.0320	0.0296
3	0.1425	0.1342	0.1264	0.1189	0.1119	0.1052	0.0988	0.0928	0.0871	0.0818
4	0.2719	0.2592	0.2469	0.2351	0.2237	0.2127	0.2022	0.1920	0.1823	0.1730
5	0.4298	0.4141	0.3988	0.3837	0.3690	0.3547	0.3407	0.3270	0.3137	0.3007
6	0.5902	0.5742	0.5582	0.5423	0.5265	0.5108	0.4953	0.4799	0.4647	0.4497
7	0.7301	0.7160	0.7018	0.6873	0.6728	0.6581	0.6433	0.6285	0.6136	0.5987
8	0.8367	0.8259	0.8148	0.8033	0.7916	0.7796	0.7673	0.7548	0.7420	0.7291
9	0.9090	0.9016	0.8939	0.8858	0.8774	0.8686	0.8596	0.8502	0.8405	0.8305
10	0.9531	0.9486	0.9437	0.9386	0.9332	0.9274	0.9214	0.9151	0.9084	0.9015
11	0.9776	0.9750	0.9723	0.9693	0.9661	0.9627	0.9591	0.9552	0.9510	0.9467
12	0.9900	0.9887	0.9873	0.9857	0.9840	0.9821	0.9801	0.9779	0.9755	0.9730
13	0.9958	0.9952	0.9945	0.9937	0.9929	0.9920	0.9909	0.9898	0.9885	0.9872
14	0.9984	0.9981	0.9978	0.9974	0.9970	0.9966	0.9961	0.9956	0.9950	0.9943
15	0.9994	0.9993	0.9992	0.9990	0.9988	0.9986	0.9984	0.9982	0.9979	0.9976
16	0.9998	0.9997	0.9997	0.9996	0.9996	0.9995	0.9994	0.9993	0.9992	0.9990
17	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998	0.9998	0.9997	0.9997	0.9996
18	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

TABLA B (continuación) Valores de la función de distribución acumulativa de Poisson

	λ									
x	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0
0	0.0008	0.0007	0.0007	0.0006	0.0006	0.0005	0.0005	0.0004	0.0004	0.0003
1	0.0067	0.0061	0.0056	0.0051	0.0047	0.0043	0.0039	0.0036	0.0033	0.0030
2	0.0275	0.0255	0.0236	0.0219	0.0203	0.0188	0.0174	0.0161	0.0149	0.0138
3	0.0767	0.0719	0.0674	0.0632	0.0591	0.0554	0.0518	0.0485	0.0453	0.0424
4	0.1641	0.1555	0.1473	0.1395	0.1321	0.1249	0.1181	0.1117	0.1055	0.0996
5	0.2881	0.2759	0.2640	0.2526	0.2414	0.2307	0.2203	0.2103	0.2006	0.1912
6	0.4349	0.4204	0.4060	0.3920	0.3782	0.3646	0.3514	0.3384	0.3257	0.3134
7	0.5838	0.5689	0.5541	0.5393	0.5246	0.5100	0.4956	0.4812	0.4670	0.4530
8	0.7160	0.7027	0.6892	0.6757	0.6620	0.6482	0.6343	0.6204	0.6065	0.5926
9	0.8202	0.8097	0.7988	0.7877	0.7764	0.7649	0.7531	0.7411	0.7290	0.7166
10	0.8942	0.8867	0.8788	0.8707	0.8622	0.8535	0.8445	0.8352	0.8257	0.8159
11	0.9420	0.9371	0.9319	0.9265	0.9208	0.9148	0.9085	0.9020	0.8952	0.8881
12	0.9703	0.9673	0.9642	0.9609	0.9573	0.9536	0.9496	0.9454	0.9409	0.9362
13	0.9857	0.9841	0.9824	0.9805	0.9784	0.9762	0.9739	0.9714	0.9687	0.9658
14	0.9935	0.9927	0.9918	0.9908	0.9897	0.9886	0.9873	0.9859	0.9844	0.9827
15	0.9972	0.9969	0.9964	0.9959	0.9954	0.9948	0.9941	0.9934	0.9926	0.9918
16	0.9989	0.9987	0.9985	0.9983	0.9980	0.9978	0.9974	0.9971	0.9967	0.9963
17	0.9996	0.9995	0.9994	0.9993	0.9992	0.9991	0.9989	0.9988	0.9986	0.9984
18	0.9998	0.9998	0.9998	0.9997	0.9997	0.9996	0.9996	0.9995	0.9994	0.9993
19	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998	0.9998	0.9997
20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9999
21	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

	λ									
x	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
0	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001
1	0.0028	0.0025	0.0023	0.0021	0.0019	0.0018	0.0016	0.0015	0.0014	0.0012
2	0.0127	0.0118	0.0109	0.0100	0.0093	0.0086	0.0079	0.0073	0.0068	0.0062
3	0.0396	0.0370	0.0346	0.0323	0.0301	0.0281	0.0262	0.0244	0.0228	0.0212
4	0.0941	0.0887	0.0837	0.0789	0.0744	0.0701	0.0660	0.0621	0.0584	0.0550
5	0.1823	0.1736	0.1653	0.1573	0.1496	0.1422	0 1352	0.1284	0.1219	0.1157
6	0.3013	0.2896	0.2781	0.2670	0.2562	0.2457	0.2355	0.2256	0.2160	0.2068
7	0.4391	0.4254	0.4119	0.3987	0.3856	0.3728	0.3602	0.3478	0.3357	0.3239
8	0.5786	0.5647	0.5508	0.5369	0.5231	0.5094	0.4958	0.4823	0.4689	0.4557
9	0.7041	0.6915	0.6788	0.6659	0.6530	0.6400	0.6269	0.6137	0.6006	0.5874
10	0.8058	0.7956	0.7850	0.7743	0.7634	0.7522	0.7409	0.7294	0.7178	0.7060
11	0.8807	0.8731	0.8652	0.8571	0.8487	0.8400	0.8311	0.8220	0.8126	0.8030
12	0.9313	0.9261	0.9207	0.9150	0.9091	0.9029	0.8965	0.8898	0.8829	0.8758
13	0.9628	0.9595	0.9561	0.9524	0.9486	0.9445	0.9403	0.9358	0.9311	0.9262
14	0.9810	0.9791	0.9771	0.9749	0.9726	0.9701	0.9675	0.9647	0.9617	0.9585
15	0.9908	0.9898	0.9887	0.9875	0.9862	0.9848	0.9832	0.9816	0.9798	0.9780
16	0.9958	0.9953	0.9947	0.9941	0.9934	0.9926	0.9913	0.9909	0.9899	0.9889
17	0.9982	0.9979	0.9977	0.9973	0.9970	0.9966	0.9962	0.9957	0.9952	0.9947
18	0.9992	0.9991	0.9990	0.9989	0.9987	0.9985	0.9983	0.9981	0.9978	0.9976
19	0.9997	0.9997	0.9996	0.9995	0.9995	0.9994	0.9993	0.9992	0.9991	0.9989
20	0.9999	0.9999	0.9998	0.9998	0.9998	0.9998	0.9997	0.9997	0.9996	0.9996
21	1.0000	1.0000	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998
22	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999
23	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

TABLA B (continuación) Valores de la función de distribución acumulativa de Poisson

	λ									
x	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10.0
0	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000
1	0.0011	0.0010	0.0009	0.0009	0.0008	0.0007	0.0007	0.0006	0.0005	0.0005
2	0.0058	0.0053	0.0049	0.0045	0.0042	0.0038	0.0035	0.0033	0.0030	0.0028
3	0.0198	0.0184	0.0172	0.0160	0.0149	0.0138	0.0129	0.0120	0.0111	0.0103
4	0.0517	0.0486	0.0456	0.0429	0.0403	0.0378	0.0355	0.0333	0.0312	0.0293
5	0.1098	0.1041	0.0987	0.0935	0.0885	0.0838	0.0793	0.0750	0.0710	0.0671
6	0.1978	0.1892	0.1808	0.1727	0.1650	0.1575	0.1502	0.1433	0.1366	0.1301
7	0.3123	0.3010	0.2900	0.2792	0.2687	0.2584	0.2485	0.2388	0.2294	0.2202
8	0.4426	0.4296	0.4168	0.4042	0.3918	0.3796	0.3676	0.3558	0.3442	0.3328
9	0.5742	0.5611	0.5480	0.5349	0.5218	0.5089	0.4960	0.4832	0.4705	0.4579
10	0.6941	0.6820	0.6699	0.6576	0.6453	0.6330	0.6205	0.6081	0.5956	0.5830
11	0.7932	0.7832	0.7730	0.7626	0.7520	0.7412	0.7303	0.7193	0.7081	0.6968
12	0.8684	0.8607	0.8529	0.8448	0.8364	0.8279	0.8191	0.8101	0.8009	0.7916
13	0.9210	0.9156	0.9100	0.9042	0.8981	0.8919	0.8853	0.8786	0.8716	0.8645
14	0.9552	0.9517	0.9480	0.9441	0.9400	0.9357	0.9312	0.9265	0.9216	0.9165
15	0.9760	0.9738	0.9715	0.9691	0.9665	0.9638	0.9609	0.9579	0.9546	0.9513
16	0.9878	0.9865	0.9852	0.9838	0.9823	0.9806	0.9789	0.9770	0.9751	0.9730
17	0.9941	0.9934	0.9927	0.9919	0.9911	0.9902	0.9892	0.9881	0.9870	0.9857
18	0.9973	0.9969	0.9966	0.9962	0.9957	0.9952	0.9947	0.9941	0.9935	0.9928
19	0.9988	0.9986	0.9985	0.9983	0.9980	0.9978	0.9975	0.9972	0.9969	0.9965
20	0.9995	0.9994	0.9993	0.9992	0.9991	0.9990	0.9989	0.9987	0.9986	0.9984
21	0.9998	0.9998	0.9997	0.9997	0.9996	0.9996	0.9995	0.9995	0.9994	0.9993
22	0.9999	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998	0.9998	0.9997	0.9997
23	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
24	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
25	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

	λ									
x	11.0	12.0	13.0	14.0	15.0	16.0	17.0	18.0	19.0	20.0
0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0012	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0049	0.0023	0.0011	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000
4	0.0151	0.0076	0.0037	0.0018	0.0009	0.0004	0.0002	0.0001	0.0000	0.0000
5	0.0375	0.0203	0.0107	0.0055	0.0028	0.0014	0.0007	0.0003	0.0002	0.0001
6	0.0786	0.0458	0.0259	0.0142	0.0076	0.0040	0.0021	0.0010	0.0005	0.0003
7	0.1432	0.0895	0.0540	0.0316	0.0180	0.0100	0.0054	0.0029	0.0015	0.0008
8	0.2320	0.1550	0.0998	0.0621	0.0374	0.0220	0.0126	0.0071	0.0039	0.0021
9	0.3405	0.2424	0.1658	0.1094	0.0699	0.0433	0.0261	0.0154	0.0089	0.0050
10	0.4599	0.3472	0.2517	0.1757	0.1185	0.0774	0.0491	0.0304	0.0183	0.0108
11	0.5793	0.4616	0.3532	0.2600	0.1848	0.1270	0.0847	0.0549	0.0347	0.0214
12	0.6887	0.5760	0.4631	0.3585	0.2676	0.1931	0.1350	0.0917	0.0606	0.0390
13	0.7813	0.6815	0.5730	0.4644	0.3632	0.2745	0.2009	0.1426	0.0984	0.0661
14	0.8540	0.7720	0.6751	0.5704	0.4657	0.3675	0.2808	0.2081	0.1497	0.1049
15	0.9074	0.8444	0.7636	0.6694	0.5681	0.4667	0.3715	0.2867	0.2148	0.1565
16	0.9441	0.8987	0.8355	0.7559	0.6641	0.5660	0.4677	0.3751	0.2920	0.2211
17	0.9678	0.9370	0.8905	0.8272	0.7489	0.6593	0.5640	0.4686	0.3784	0.2970
18	0.9823	0.9626	0.9302	0.8826	0.8195	0.7423	0.6550	0.5622	0.4695	0.3814
19	0.9907	0.9787	0.9573	0.9235	0.8752	0.8122	0.7363	0.6509	0.5606	0.4703
20	0.9953	0.9884	0.9750	0.9521	0.9170	0.8682	0.8055	0.7307	0.6472	0.5591
21	0.9977	0.9939	0.9859	0.9712	0.9469	0.9108	0.8615	0.7991	0.7255	0.6437
22	0.9990	0.9970	0.9924	0.9833	0.9673	0.9418	0.9047	0.8551	0.7931	0.7206
23	0.9995	0.9985	0.9960	0.9907	0.9805	0.9633	0.9367	0.8989	0.8490	0.7875
24	0.9998	0.9993	0.9980	0.9950	0.9888	0.9777	0.9594	0.9317	0.8933	0.8432
25	0.9999	0.9997	0.9990	0.9974	0.9938	0.9869	0.9748	0.9554	0.9269	0.8878
26	1.0000	0.9999	0.9995	0.9987	0.9967	0.9925	0.9848	0.9718	0.9514	0.9221
27	1.0000	0.9999	0.9998	0.9994	0.9983	0.9959	0.9912	0.9827	0.9687	0.9475
28	1.0000	1.0000	0.9999	0.9997	0.9991	0.9978	0.9950	0.9897	0.9805	0.9657
29	1.0000	1.0000	1.0000	0.9999	0.9996	0.9989	0.9973	0.9941	0.9882	0.9782
30	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9986	0.9967	0.9930	0.9865
31	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9993	0.9982	0.9960	0.9919
32	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9990	0.9978	0.9953
33	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9995	0.9988	0.9973
34	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9985
35	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992
36	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996
37	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
38	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999

TABLA C Valores de las funciones de probabilidad y distribución acumulativa para la distribución hipergeométrica

$$P(X \leq x) = F(x; N, n, k) = \sum_{i=0}^{x} \frac{\binom{k}{i}\binom{N-k}{n-i}}{\binom{N}{n}} \qquad p(x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

N	n	k	x	$F(x)$	$p(x)$	N	n	k	x	$F(x)$	$p(x)$
2	1	1	0	0.500000	0.500000	6	2	2	2	1.000000	0.066667
2	1	1	1	1.000000	0.500000	6	3	1	0	0.500000	0.500000
3	1	1	0	0.666667	0.666667	6	3	1	1	1.000000	0.500000
3	1	1	1	1.000000	0.333333	6	3	2	0	0.200000	0.200000
3	2	1	0	0.333333	0.333333	6	3	2	1	0.800000	0.600000
3	2	1	1	1.000000	0.666667	6	3	2	2	1.000000	0.200000
3	2	2	1	0.666667	0.666667	6	3	3	0	0.050000	0.050000
3	2	2	2	1.000000	0.333333	6	3	3	1	0.500000	0.450000
4	1	1	0	0.750000	0.750000	6	3	3	2	0.950000	0.450000
4	1	1	1	1.000000	0.250000	6	3	3	3	1.000000	0.050000
4	2	1	0	0.500000	0.500000	6	4	1	0	0.333333	0.333333
4	2	1	1	1.000000	0.500000	6	4	1	1	1.000000	0.666667
4	2	2	0	0.166667	0.166667	6	4	2	0	0.066667	0.066667
4	2	2	1	0.833333	0.666667	6	4	2	1	0.600000	0.533333
4	2	2	2	1.000000	0.166667	6	4	2	2	1.000000	0.400000
4	3	1	0	0.250000	0.250000	6	4	3	1	0.200000	0.200000
4	3	1	1	1.000000	0.750000	6	4	3	2	0.800000	0.600000
4	3	2	1	0.500000	0.500000	6	4	3	3	1.000000	0.200000
4	3	2	2	1.000000	0.500000	6	4	4	2	0.400000	0.400000
4	3	3	2	0.750000	0.750000	6	4	4	3	0.933333	0.533333
4	3	3	3	1.000000	0.250000	6	4	4	4	1.000000	0.066667
5	1	1	0	0.800000	0.800000	6	5	1	0	0.166667	0.166667
5	1	1	1	1.000000	0.200000	6	5	1	1	1.000000	0.833333
5	2	1	0	0.600000	0.600000	6	5	2	1	0.333333	0.333333
5	2	1	1	1.000000	0.400000	6	5	2	2	1.000000	0.666667
5	2	2	0	0.300000	0.300000	6	5	3	2	0.500000	0.500000
5	2	2	1	0.900000	0.600000	6	5	3	3	1.000000	0.500000
5	2	2	2	1.000000	0.100000	6	5	4	3	0.666667	0.666667
5	3	1	0	0.400000	0.400000	6	5	4	4	1.000000	0.333333
5	3	1	1	1.000000	0.600000	6	5	5	4	0.833333	0.833333

5	3	2	0	0.100000	0.100000	6	5	5	5	1.000000	0.166667
5	3	2	1	0.700000	0.600000	7	1	1	0	0.857143	0.857143
5	3	2	2	1.000000	0.300000	7	1	1	1	1.000000	0.142857
5	3	3	1	0.300000	0.300000	7	2	1	0	0.714286	0.714286
5	3	3	2	0.900000	0.600000	7	2	1	1	1.000000	0.285714
5	3	3	3	1.000000	0.100000	7	2	2	0	0.476190	0.476190
5	4	1	0	0.200000	0.200000	7	2	2	1	0.952381	0.476190
5	4	1	1	1.000000	0.800000	7	2	2	2	1.000000	0.047619
5	4	2	1	0.400000	0.400000	7	3	1	0	0.571429	0.571429
5	4	2	2	0.000000	0.600000	7	3	1	1	1.000000	0.428571
5	4	3	2	0.600000	0.600000	7	3	2	0	0.285714	0.285714
5	4	3	3	1.000000	0.400000	7	3	2	1	0.857143	0.571429
5	4	4	3	0.800000	0.800000	7	3	2	2	1.000000	0.142857
5	4	4	4	1.000000	0.200000	7	3	3	0	0.114286	0.114286
6	1	1	0	0.833333	0.833333	7	3	3	1	0.628571	0.514286
6	1	1	1	1.000000	0.166667	7	3	3	2	0.971428	0.342857
6	2	1	0	0.666667	0.666667	7	3	3	3	1.000000	0.028571
6	2	1	1	1.000000	0.333333	7	4	1	0	0.428571	0.428571
6	2	2	0	0.400000	0.400000	7	4	1	1	1.000000	0.571429
6	2	2	1	0.933333	0.533333	7	4	2	0	0.142857	0.142857
7	4	2	1	0.714286	0.571429	8	3	3	2	0.982143	0.267857
7	4	2	2	1.000000	0.285714	8	3	3	3	1.000000	0.017857
7	4	3	0	0.028571	0.028571	8	4	1	0	0.500000	0.500000
7	4	3	1	0.371429	0.342857	8	4	1	1	1.000000	0.500000
7	4	3	2	0.885714	0.514286	8	4	2	0	0.214286	0.214286
7	4	3	3	1.000000	0.114286	8	4	2	1	0.785714	0.571429
7	4	4	1	0.114286	0.114286	8	4	2	2	1.000000	0.214286
7	4	4	2	0.628571	0.514286	8	4	3	0	0.071429	0.071429
7	4	4	3	0.971428	0.342857	8	4	3	1	0.500000	0.428571
7	4	4	4	1.000000	0.028571	8	4	3	2	0.928571	0.428571
7	5	1	0	0.285714	0.285714	8	4	3	3	1.000000	0.071429
7	5	1	1	1.000000	0.714286	8	4	4	0	0.014286	0.014286
7	5	2	0	0.047619	0.047619	8	4	4	1	0.242857	0.228571
7	5	2	1	0.523809	0.476190	8	4	4	2	0.757143	0.514286
7	5	2	2	1.000000	0.476190	8	4	4	3	0.985714	0.228571
7	5	3	1	0.142857	0.142857	8	4	4	4	1.000000	0.014286
7	5	3	2	0.714286	0.571429	8	5	1	0	0.375000	0.375000
7	5	3	3	1.000000	0.285714	8	5	1	1	1.000000	0.625000
7	5	4	2	0.285714	0.285714	8	5	2	0	0.107143	0.107143
7	5	4	3	0.857143	0.571429	8	5	2	1	0.642857	0.535714

TABLA C (continuación) Valores de las funciones de probabilidad y distribución acumulativa para la distribución hipergeométrica

N	n	k	x	$F(x)$	$p(x)$	N	n	k	x	$F(x)$	$p(x)$
7	5	4	4	1.000000	0.142857	8	5	2	2	1.000000	0.357143
7	5	5	3	0.476190	0.476190	8	5	3	0	0.017857	0.017857
7	5	5	4	0.952381	0.476190	8	5	3	1	0.285714	0.267857
7	5	5	5	1.000000	0.047619	8	5	3	2	0.821429	0.535714
7	6	1	0	0.142857	0.142857	8	5	3	3	1.000000	0.178571
7	6	1	1	1.000000	0.857143	8	5	4	1	0.071429	0.071429
7	6	2	1	0.285714	0.285714	8	5	4	2	0.500000	0.428571
7	6	2	2	1.000000	0.714286	8	5	4	3	0.928571	0.428571
7	6	3	2	0.428571	0.428571	8	5	4	4	1.000000	0.071429
7	6	3	3	1.000000	0.571429	8	5	5	2	0.178571	0.178571
7	6	4	3	0.571429	0.571429	8	5	5	3	0.714286	0.535714
7	6	4	4	1.000000	0.428571	8	5	5	4	0.982143	0.267857
7	6	5	4	0.714286	0.714286	8	5	5	5	1.000000	0.017857
7	6	5	5	1.000000	0.285714	8	6	1	0	0.250000	0.250000
7	6	6	5	0.857143	0.857143	8	6	1	1	1.000000	0.750000
7	6	6	6	1.000000	0.142857	8	6	2	0	0.035714	0.035714
8	1	1	0	0.875000	0.875000	8	6	2	1	0.464286	0.428571
8	1	1	1	1.000000	0.125000	8	6	2	2	1.000000	0.535714
8	2	1	0	0.750000	0.750000	8	6	3	1	0.107143	0.107143
8	2	1	1	1.000000	0.250000	8	6	3	2	0.642857	0.535714
8	2	2	0	0.535714	0.535714	8	6	3	3	1.000000	0.357143
8	2	2	1	0.964286	0.428571	8	6	4	2	0.214286	0.214286
8	2	2	2	1.000000	0.035714	8	6	4	3	0.785714	0.571429
8	3	1	0	0.625000	0.625000	8	6	4	4	1.000000	0.214286
8	3	1	1	1.000000	0.375000	8	6	5	3	0.357143	0.357143
8	3	2	0	0.357143	0.357143	8	6	5	4	0.892857	0.535714
8	3	2	1	0.892857	0.535714	8	6	5	5	1.000000	0.107143
8	3	2	2	1.000000	0.107143	8	6	6	4	0.535714	0.535714
8	3	3	0	0.178571	0.178571	8	6	6	5	0.964286	0.428571
8	3	3	1	0.714286	0.535714	8	6	6	6	1.000000	0.035714

8	7	1	0	0.125000	0.125000	9	5	3	1	0.404762	0.357143
8	7	1	1	1.000000	0.875000	9	5	3	2	0.880952	0.476190
8	7	2	1	0.250000	0.250000	9	5	3	3	1.000000	0.119048
8	7	2	2	1.000000	0.750000	9	5	4	0	0.007936	0.007936
8	7	3	2	0.375000	0.375000	9	5	4	1	0.166667	0.158730
8	7	3	3	1.000000	0.625000	9	5	4	2	0.642857	0.476190
8	7	4	3	0.500000	0.500000	9	5	4	3	0.960317	0.317460
8	7	4	4	1.000000	0.500000	9	5	4	4	1.000000	0.039683
8	7	5	4	0.625000	0.625000	9	5	5	1	0.039683	0.039683
8	7	5	5	1.000000	0.375000	9	5	5	2	0.357143	0.317460
8	7	6	5	0.750000	0.750000	9	5	5	3	0.833333	0.476190
8	7	6	6	1.000000	0.250000	9	5	5	4	0.992063	0.158730
8	7	7	6	0.875000	0.875000	9	5	5	5	1.000000	0.007936
8	7	7	7	1.000000	0.125000	9	6	1	0	0.333333	0.333333
9	1	1	0	0.888889	0.888889	9	6	1	1	1.000000	0.666667
9	1	1	1	1.000000	0.111111	9	6	2	0	0.083333	0.083333
9	2	1	0	0.777778	0.777778	9	6	2	1	0.583333	0.500000
9	2	1	1	1.000000	0.222222	9	6	2	2	1.000000	0.416667
9	2	2	0	0.583333	0.583333	9	6	3	0	0.011905	0.011905
9	2	2	1	0.972222	0.388889	9	6	3	1	0.226190	0.214286
9	2	2	2	1.000000	0.027778	9	6	3	2	0.761905	0.535714
9	3	1	0	0.666667	0.666667	9	6	3	3	1.000000	0.238095
9	3	1	1	1.000000	0.333333	9	6	4	1	0.047619	0.047619
9	3	2	0	0.416667	0.416667	9	6	4	2	0.404762	0.357143
9	3	2	1	0.916667	0.500000	9	6	4	3	0.880952	0.476190
9	3	2	2	1.000000	0.083333	9	6	4	4	1.000000	0.119048
9	3	3	0	0.238095	0.238095	9	6	5	2	0.119048	0.119048
9	3	3	1	0.773809	0.535714	9	6	5	3	0.595238	0.476190
9	3	3	2	0.988095	0.214286	9	6	5	4	0.952381	0.357143
9	3	3	3	1.000000	0.011905	9	6	5	5	1.000000	0.047619
9	4	1	0	0.555556	0.555556	9	6	6	3	0.238095	0.238095
9	4	1	1	1.000000	0.444444	9	6	6	4	0.773809	0.535714
9	4	2	0	0.277778	0.277778	9	6	6	5	0.988095	0.214286
9	4	2	1	0.833333	0.555556	9	6	6	6	1.000000	0.011905
9	4	2	2	1.000000	0.166667	9	7	1	0	0.222222	0.222222
9	4	3	0	0.119048	0.119048	9	7	1	1	1.000000	0.777778
9	4	3	1	0.595238	0.476190	9	7	2	0	0.027778	0.027778
9	4	3	2	0.952381	0.357143	9	7	2	1	0.416667	0.388889
9	4	3	3	1.000000	0.047619	9	7	2	2	1.000000	0.583333
9	4	4	0	0.039683	0.039683	9	7	3	1	0.083333	0.083333

TABLA C (continuación) Valores de las funciones de probabilidad y distribución acumulativa para la distribución hipergeométrica

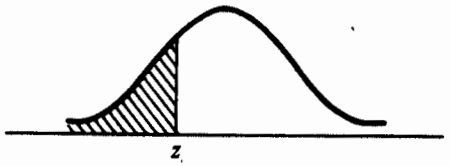
N	n	k	x	$F(x)$	$p(x)$	N	n	k	x	$F(x)$	$p(x)$
9	4	4	1	0.357143	0.317460	9	7	3	2	0.583333	0.500000
9	4	4	2	0.833333	0.476190	9	7	3	3	1.000000	0.416667
9	4	4	3	0.992063	0.158730	9	7	4	2	0.166667	0.166667
9	4	4	4	1.000000	0.007936	9	7	4	3	0.722222	0.555556
9	5	1	0	0.444444	0.444444	9	7	4	4	1.000000	0.277778
9	5	1	1	1.000000	0.555556	9	7	5	3	0.277778	0.277778
9	5	2	0	0.166667	0.166667	9	7	5	4	0.833333	0.555556
9	5	2	1	0.722222	0.555556	9	7	5	5	1.000000	0.166667
9	5	2	2	1.000000	0.277778	9	7	6	4	0.416667	0.416667
9	5	3	0	0.047619	0.047619	9	7	6	5	0.916667	0.500000
9	7	6	6	1.000000	0.833333	10	5	1	0	0.500000	0.500000
9	7	7	5	0.583333	0.583333	10	5	1	1	1.000000	0.500000
9	7	7	6	0.972222	0.388889	10	5	2	0	0.222222	0.222222
9	7	7	7	1.000000	0.027778	10	5	2	1	0.777778	0.555556
9	8	1	0	0.111111	0.111111	10	5	2	2	1.000000	0.222222
9	8	1	1	1.000000	0.888889	10	5	3	0	0.083333	0.083333
9	8	2	1	0.222222	0.222222	10	5	3	1	0.500000	0.416667
9	8	2	2	1.000000	0.777778	10	5	3	2	0.916667	0.416667
9	8	3	2	0.333333	0.333333	10	5	3	3	1.000000	0.083333
9	8	3	3	1.000000	0.666667	10	5	4	0	0.023810	0.023810
9	8	4	3	0.444444	0.444444	10	5	4	1	0.261905	0.238095
9	8	4	4	1.000000	0.555556	10	5	4	2	0.738095	0.476190
9	8	5	4	0.555556	0.555556	10	5	4	3	0.976190	0.238095
9	8	5	5	1.000000	0.444444	10	5	4	4	1.000000	0.023810
9	8	6	5	0.666667	0.666667	10	5	5	0	0.003968	0.003968
9	8	6	6	1.000000	0.333333	10	5	5	1	0.103175	0.099206
9	8	7	6	0.777778	0.777778	10	5	5	2	0.500000	0.396825
9	8	7	7	1.000000	0.222222	10	5	5	3	0.896825	0.396825
9	8	8	7	0.888889	0.888889	10	5	5	4	0.996032	0.099206
9	8	8	8	1.000000	0.111111	10	5	5	5	1.000000	0.003968

10	1	1	0	0.900000	0.900000	10	6	1	0	0.400000	0.400000
10	1	1	1	1.000000	0.100000	10	6	1	1	1.000000	0.600000
10	2	1	0	0.800000	0.800000	10	6	2	0	0.133333	0.133333
10	2	1	1	1.000000	0.200000	10	6	2	1	0.666667	0.533333
10	2	2	0	0.622222	0.622222	10	6	2	2	1.000000	0.333333
10	2	2	1	0.977778	0.355556	10	6	3	0	0.033333	0.033333
10	2	2	2	1.000000	0.022222	10	6	3	1	0.333333	0.300000
10	3	1	0	0.700000	0.700000	10	6	3	2	0.833333	0.500000
10	3	1	1	1.000000	0.300000	10	6	3	3	1.000000	0.166667
10	3	2	0	0.466667	0.466667	10	6	4	0	0.004762	0.004762
10	3	2	1	0.933333	0.466667	10	6	4	1	0.119048	0.114286
10	3	2	2	1.000000	0.066667	10	6	4	2	0.547619	0.428571
10	3	3	0	0.291667	0.291667	10	6	4	3	0.928571	0.380952
10	3	3	1	0.816667	0.525000	10	6	4	4	1.000000	0.071429
10	3	3	2	0.991667	0.175000	10	6	5	1	0.023810	0.023810
10	3	3	3	1.000000	0.008333	10	6	5	2	0.261905	0.238095
10	4	1	0	0.600000	0.600000	10	6	5	3	0.738095	0.476190
10	4	1	1	1.000000	0.400000	10	6	5	4	0.976190	0.238095
10	4	2	0	0.333333	0.333333	10	6	5	5	1.000000	0.023810
10	4	2	1	0.866667	0.533333	10	6	6	2	0.071429	0.071429
10	4	2	2	1.000000	0.133333	10	6	6	3	0.452381	0.380952
10	4	3	0	0.166667	0.166667	10	6	6	4	0.880952	0.428571
10	4	3	1	0.666667	0.500000	10	6	6	5	0.995238	0.114286
10	4	3	2	0.966667	0.300000	10	6	6	6	1.000000	0.004762
10	4	3	3	1.000000	0.033333	10	7	1	0	0.300000	0.300000
10	4	4	0	0.071429	0.071429	10	7	1	1	1.000000	0.700000
10	4	4	1	0.452381	0.380952	10	7	2	0	0.066667	0.066667
10	4	4	2	0.880952	0.428571	10	7	2	1	0.533333	0.466667
10	4	4	3	0.995238	0.114286	10	7	2	2	1.000000	0.466667
10	4	4	4	1.000000	0.004762	10	7	3	0	0.008333	0.008333

Fuente: Tomado de *Tables of the hypergeometric probability distribution,* por Gerald J. Lieberman y Donald B. Owen, con permiso de los editores, Stanford University Press. Copyright © 1961 by the Board of Trustees of the Leland Stanford Junior University.

TABLA D Valores de la función de distribución acumulativa normal estándar

$$P(Z \leq z) = F(z; 0, 1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-t^2/2)dt$$



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183

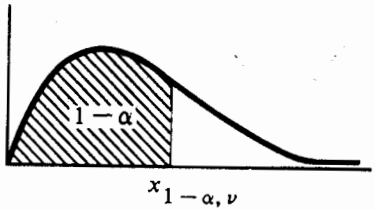
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389

TABLA D (continuación) Valores de la función de distribución acumulativa normal estándar

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

TABLA E Valores de cuantiles de la distribución chi-cuadrada

$$P(X \leq x_{1-\alpha,\nu}) = F(x_{1-\alpha};\nu) = \frac{1}{\Gamma(\nu/2)\,2^{\nu/2}} \int_0^{x_{1-\alpha,\nu}} t^{\nu/2-1}\exp(-t/2)dt = 1-\alpha$$



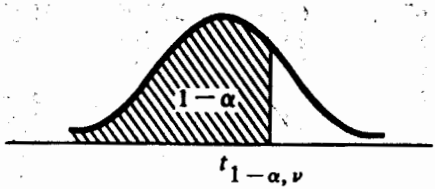
ν	$x_{0.005}$	$x_{0.010}$	$x_{0.025}$	$x_{0.050}$	$x_{0.100}$	$x_{0.900}$	$x_{0.950}$	$x_{0.975}$	$x_{0.990}$	$x_{0.995}$
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.64	7.90
2	0.01	0.02	0.05	0.10	0.21	4.60	5.99	7.38	9.22	10.59
3	0.07	0.11	0.22	0.35	0.58	6.25	7.82	9.36	11.32	12.82
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.15	13.28	14.82
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.84	15.09	16.76
6	0.67	0.87	1.24	1.63	2.20	10.65	12.60	14.46	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.02	18.47	20.27
8	1.34	1.64	2.18	2.73	3.49	13.36	15.51	17.55	20.08	21.94
9	1.73	2.09	2.70	3.32	4.17	14.69	16.93	19.03	21.65	23.56
10	2.15	2.55	3.24	3.94	4.86	15.99	18.31	20.50	23.19	25.15
11	2.60	3.05	3.81	4.57	5.58	17.28	19.68	21.93	24.75	26.71
12	3.06	3.57	4.40	5.22	6.30	18.55	21.03	23.35	26.25	28.25
13	3.56	4.10	5.01	5.89	7.04	19.81	22.37	24.75	27.72	29.88
14	4.07	4.65	5.62	6.57	7.79	21.07	23.69	26.13	29.17	31.38
15	4.59	5.23	6.26	7.26	8.55	22.31	25.00	27.50	30.61	32.86
16	5.14	5.81	6.90	7.96	9.31	23.55	26.30	28.86	32.03	34.32
17	5.69	6.40	7.56	8.67	10.08	24.77	27.59	30.20	33.43	35.77
18	6.25	7.00	8.23	9.39	10.86	25.99	28.88	31.54	34.83	37.21
19	6.82	7.63	8.90	10.11	11.65	27.21	30.15	32.87	36.22	38.63
20	7.42	8.25	9.59	10.85	12.44	28.42	31.42	34.18	37.59	40.05

TABLA E (continuación) Valores de cuantiles de la distribución chi-cuadrada

ν	$x_{0.005}$	$x_{0.010}$	$x_{0.025}$	$x_{0.050}$	$x_{0.100}$	$x_{0.900}$	$x_{0.950}$	$x_{0.975}$	$x_{0.990}$	$x_{0.995}$
21	8.02	8.89	10.28	11.59	13.24	29.62	32.68	35.49	38.96	41.45
22	8.62	9.53	10.98	12.34	14.04	30.82	33.93	36.79	40.31	42.84
23	9.25	10.19	11.69	13.09	14.85	32.01	35.18	38.09	41.66	44.23
24	9.87	10.85	12.40	13.84	15.66	33.20	36.42	39.38	43.00	45.60
25	10.50	11.51	13.11	14.61	16.47	34.38	37.66	40.66	44.34	46.97
26	11.13	12.19	13.84	15.38	17.29	35.57	38.89	41.94	45.66	48.33
27	11.79	12.87	14.57	16.15	18.11	36.74	40.12	43.21	46.99	49.69
28	12.44	13.55	15.30	16.92	18.94	37.92	41.34	44.47	48.30	51.04
29	13.09	14.24	16.04	17.70	19.77	39.09	42.56	45.74	49.61	52.38
30	13.77	14.94	16.78	18.49	20.60	40.26	43.78	46.99	50.91	53.71
35	17.16	18.49	20.56	22.46	24.79	46.06	49.81	53.22	57.36	60.31
40	20.67	22.14	24.42	26.51	29.06	51.80	55.75	59.34	63.71	66.80
45	24.28	25.88	28.36	30.61	33.36	57.50	61.65	65.41	69.98	73.20
50	27.96	29.68	32.35	34.76	37.69	63.16	67.50	71.42	76.17	79.52
60	35.50	37.46	40.47	43.19	46.46	74.39	79.08	83.30	88.40	91.98
70	43.25	45.42	48.75	51.74	55.33	85.52	90.53	95.03	100.44	104.24
80	51.14	53.52	57.15	60.39	64.28	96.57	101.88	106.63	112.34	116.35
90	59.17	61.74	65.64	69.13	73.29	107.56	113.14	118.14	124.13	128.32
100	67.30	70.05	74.22	77.93	82.36	118.49	124.34	129.56	135.82	140.19

TABLA F Valores de cuantiles de la distribución t de Student

$$P(T \leq t_{1-\alpha,\nu}) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\pi\nu}\,\Gamma(\nu/2)} \int_{-\infty}^{t_{1-\alpha,\nu}} [1 + (t^2/\nu)]^{-(\nu+1)/2} dt = 1 - \alpha$$



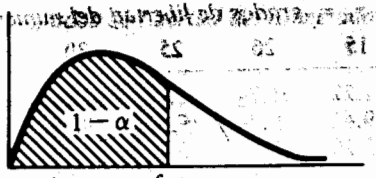
ν	$t_{0.001}$	$t_{0.005}$	$t_{0.010}$	$t_{0.025}$	$t_{0.050}$	$t_{0.100}$	$t_{0.200}$
1	−318.309	−63.657	−31.821	−12.706	−6.314	−3.078	−1.376
2	−22.327	−9.925	−6.965	−4.303	−2.920	−1.886	−1.061
3	−10.215	−5.841	−4.541	−3.182	−2.353	−1.638	−0.978
4	−7.173	−4.604	−3.747	−2.776	−2.132	−1.533	−0.941
5	−5.893	−4.032	−3.365	−2.571	−2.015	−1.476	−0.920
6	−5.208	−3.707	−3.143	−2.447	−1.943	−1.440	−0.906
7	−4.785	−3.499	−2.998	−2.365	−1.895	−1.415	−0.896
8	−4.501	−3.355	−2.896	−2.306	−1.860	−1.397	−0.889
9	−4.297	−3.250	−2.821	−2.262	−1.833	−1.383	−0.883
10	−4.144	−3.169	−2.764	−2.228	−1.812	−1.372	−0.879
11	−4.025	−3.106	−2.718	−2.201	−1.796	−1.363	−0.876
12	−3.930	−3.055	−2.681	−2.179	−1.782	−1.356	−0.873
13	−3.852	−3.012	−2.650	−2.160	−1.771	−1.350	−0.870
14	−3.787	−2.977	−2.624	−2.145	−1.761	−1.345	−0.868
15	−3.733	−2.947	−2.602	−2.131	−1.753	−1.341	−0.866
16	−3.686	−2.921	−2.583	−2.120	−1.746	−1.337	−0.865
17	−3.646	−2.898	−2.567	−2.110	−1.740	−1.333	−0.863
18	−3.610	−2.878	−2.552	−2.101	−1.734	−1.330	−0.862
19	−3.579	−2.861	−2.539	−2.093	−1.729	−1.328	−0.861
20	−3.552	−2.845	−2.528	−2.086	−1.725	−1.325	−0.860
21	−3.527	−2.831	−2.518	−2.080	−1.721	−1.323	−0.859
22	−3.505	−2.819	−2.508	−2.074	−1.717	−1.321	−0.858
23	−3.485	−2.807	−2.500	−2.069	−1.714	−1.319	−0.858
24	−3.467	−2.797	−2.492	−2.064	−1.711	−1.318	−0.857
25	−3.450	−2.787	−2.485	−2.060	−1.708	−1.316	−0.856
26	−3.435	−2.779	−2.479	−2.056	−1.706	−1.315	−0.856
27	−3.421	−2.771	−2.473	−2.052	−1.703	−1.314	−0.855
28	−3.408	−2.763	−2.467	−2.048	−1.701	−1.313	−0.855
29	−3.396	−2.756	−2.462	−2.045	−1.699	−1.311	−0.854
30	−3.385	−2.750	−2.457	−2.042	−1.697	−1.310	−0.854
35	−3.340	−2.724	−2.438	−2.030	−1.690	−1.306	−0.852
40	−3.307	−2.704	−2.423	−2.021	−1.684	−1.303	−0.851
45	−3.281	−2.690	−2.412	−2.014	−1.679	−1.301	−0.850
50	−3.261	−2.678	−2.403	−2.009	−1.676	−1.299	−0.849
60	−3.232	−2.660	−2.390	−2.000	−1.671	−1.296	−0.848
70	−3.211	−2.648	−2.381	−1.994	−1.667	−1.294	−0.847
80	−3.195	−2.639	−2.374	−1.990	−1.664	−1.292	−0.846
90	−3.183	−2.632	−2.369	−1.987	−1.662	−1.291	−0.846
100	−3.174	−2.626	−2.364	−1.984	−1.660	−1.290	−0.845
200	−3.131	−2.601	−2.345	−1.972	−1.652	−1.286	−0.843
500	−3.107	−2.586	−2.334	−1.965	−1.648	−1.283	−0.842
1000	−3.098	−2.581	−2.330	−1.962	−1.646	−1.282	−0.842

TABLA F (continuación) Valores de cuantiles de la distribución t de Student

ν	$t_{0.800}$	$t_{0.900}$	$t_{0.950}$	$t_{0.975}$	$t_{0.990}$	$t_{0.995}$	$t_{0.999}$
1	1.376	3.078	6.314	12.706	31.820	63.656	318.294
2	1.061	1.886	2.920	4.303	6.965	9.925	22.327
3	0.978	1.638	2.353	3.182	4.541	5.841	10.214
4	0.941	1.533	2.132	2.776	3.747	4.604	7.173
5	0.920	1.476	2.015	2.571	3.365	4.032	5.893
6	0.906	1.440	1.943	2.447	3.143	3.707	5.208
7	0.896	1.415	1.895	2.365	2.998	3.499	4.785
8	0.889	1.397	1.860	2.306	2.896	3.355	4.501
9	0.883	1.383	1.833	2.262	2.821	3.250	4.297
10	0.879	1.372	1.812	2.228	2.764	3.169	4.144
11	0.876	1.363	1.796	2.201	2.718	3.106	4.025
12	0.873	1.356	1.782	2.179	2.681	3.055	3.930
13	0.870	1.350	1.771	2.160	2.650	3.012	3.852
14	0.868	1.345	1.761	2.145	2.624	2.977	3.787
15	0.866	1.341	1.753	2.131	2.602	2.947	3.733
16	0.865	1.337	1.746	2.120	2.583	2.921	3.686
17	0.863	1.333	1.740	2.110	2.567	2.898	3.646
18	0.862	1.330	1.734	2.101	2.552	2.878	3.610
19	0.861	1.328	1.729	2.093	2.539	2.861	3.579
20	0.860	1.325	1.725	2.086	2.528	2.845	3.552
21	0.859	1.323	1.721	2.080	2.518	2.831	3.527
22	0.858	1.321	1.717	2.074	2.508	2.819	3.505
23	0.858	1.319	1.714	2.069	2.500	2.807	3.485
24	0.857	1.318	1.711	2.064	2.492	2.797	3.467
25	0.856	1.316	1.708	2.060	2.485	2.787	3.450
26	0.856	1.315	1.706	2.056	2.479	2.779	3.435
27	0.855	1.314	1.703	2.052	2.473	2.771	3.421
28	0.855	1.313	1.701	2.048	2.467	2.763	3.408
29	0.854	1.311	1.699	2.045	2.462	2.756	3.396
30	0.854	1.310	1.697	2.042	2.457	2.750	3.385
35	0.852	1.306	1.690	2.030	2.438	2.724	3.340
40	0.851	1.303	1.684	2.021	2.423	2.704	3.307
45	0.850	1.301	1.679	2.014	2.412	2.690	3.281
50	0.849	1.299	1.676	2.009	2.403	2.678	3.261
60	0.848	1.296	1.671	2.000	2.390	2.660	3.232
70	0.847	1.294	1.667	1.994	2.381	2.648	3.211
80	0.846	1.292	1.664	1.990	2.374	2.639	3.195
90	0.846	1.291	1.662	1.987	2.368	2.632	3.183
100	0.845	1.290	1.660	1.984	2.364	2.626	3.174
200	0.843	1.286	1.652	1.972	2.345	2.601	3.131
500	0.842	1.283	1.648	1.965	2.334	2.586	3.107
1000	0.842	1.282	1.646	1.962	2.330	2.581	3.098

TABLA G Valores de cuantiles de la distribución F

$$P(F \leq f_{1-\alpha,\nu_1,\nu_2}) = \frac{\Gamma[(\nu_1+\nu_2)/2]\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \int_0^{f_{1-\alpha,\nu_1,\nu_2}} t^{(\nu_1-2)/2}(\nu_2+\nu_1 t)^{(\nu_1+\nu_2)/2}\,dt = 1-\alpha$$



$f_{1-\alpha,\nu_1,\nu_2}$

$1-\alpha = 0.9$

	ν_1 = *grados de libertad del numerador*									
ν_2	1	2	3	4	5	6	7	8	9	10
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
7	3.59	3.26	3.07	2.96	2.88	2.83	2.79	2.75	2.72	2.70
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
35	2.85	2.46	2.25	2.11	2.02	1.95	1.90	1.85	1.82	1.79
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
80	2.77	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.71	1.68
100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66
200	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63
500	2.72	2.31	2.09	1.96	1.86	1.79	1.73	1.68	1.64	1.61
1000	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64	1.61

TABLA G (continuación) Valores de cuantiles de la distribución F

$1 - \alpha = 0.9$

	ν_1 = *grados de libertad del numerador*									
ν_2	11	12	15	20	25	30	40	50	100	1000
1	60.47	60.71	61.22	61.74	62.06	62.26	62.53	62.69	63.00	63.29
2	9.40	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.22	5.22	5.20	5.19	5.17	5.17	5.16	5.15	5.14	5.13
4	3.91	3.90	3.87	3.84	3.83	3.82	3.80	3.80	3.78	3.76
5	3.28	3.27	3.24	3.21	3.19	3.17	3.16	3.15	3.13	3.11
6	2.92	2.90	2.87	2.84	2.81	2.80	2.78	2.77	2.75	2.72
7	2.68	2.67	2.63	2.59	2.57	2.56	2.54	2.52	2.50	2.47
8	2.52	2.50	2.46	2.42	2.40	2.38	2.36	2.35	2.32	2.30
9	2.40	2.38	2.34	2.30	2.27	2.25	2.23	2.22	2.19	2.16
10	2.30	2.28	2.24	2.20	2.17	2.16	2.13	2.12	2.09	2.06
11	2.23	2.21	2.17	2.12	2.10	2.08	2.05	2.04	2.00	1.98
12	2.17	2.15	2.10	2.06	2.03	2.01	1.99	1.97	1.94	1.91
13	2.12	2.10	2.05	2.01	1.98	1.96	1.93	1.92	1.88	1.85
14	2.07	2.05	2.01	1.96	1.93	1.91	1.89	1.87	1.83	1.80
15	2.04	2.02	1.97	1.92	1.89	1.87	1.85	1.83	1.79	1.76
16	2.01	1.99	1.94	1.89	1.86	1.84	1.81	1.79	1.76	1.72
17	1.98	1.96	1.91	1.86	1.83	1.81	1.78	1.76	1.73	1.69
18	1.95	1.93	1.89	1.84	1.80	1.78	1.75	1.74	1.70	1.66
19	1.93	1.91	1.86	1.81	1.78	1.76	1.73	1.71	1.67	1.64
20	1.91	1.89	1.84	1.79	1.76	1.74	1.71	1.69	1.65	1.61
21	1.90	1.87	1.83	1.78	1.74	1.72	1.69	1.67	1.63	1.59
22	1.88	1.86	1.81	1.76	1.73	1.70	1.67	1.65	1.61	1.57
23	1.87	1.84	1.80	1.74	1.71	1.69	1.66	1.64	1.59	1.55
24	1.85	1.83	1.78	1.73	1.70	1.67	1.64	1.62	1.58	1.54
25	1.84	1.82	1.77	1.72	1.68	1.66	1.63	1.61	1.56	1.52
26	1.83	1.81	1.76	1.71	1.67	1.65	1.61	1.59	1.55	1.51
27	1.82	1.80	1.75	1.70	1.66	1.64	1.60	1.58	1.54	1.50
28	1.81	1.79	1.74	1.69	1.65	1.63	1.59	1.57	1.53	1.48
29	1.80	1.78	1.73	1.68	1.64	1.62	1.58	1.56	1.52	1.47
30	1.79	1.77	1.72	1.67	1.63	1.61	1.57	1.55	1.51	1.46
35	1.76	1.74	1.69	1.63	1.60	1.57	1.53	1.51	1.47	1.42
40	1.74	1.71	1.66	1.61	1.57	1.54	1.51	1.48	1.43	1.38
50	1.70	1.68	1.63	1.57	1.53	1.50	1.46	1.44	1.39	1.33
60	1.68	1.66	1.60	1.54	1.50	1.48	1.44	1.41	1.36	1.30
80	1.65	1.63	1.57	1.51	1.47	1.44	1.40	1.38	1.32	1.25
100	1.64	1.61	1.56	1.49	1.45	1.42	1.38	1.35	1.29	1.22
200	1.60	1.58	1.52	1.46	1.41	1.38	1.34	1.31	1.24	1.16
500	1.58	1.56	1.50	1.44	1.39	1.36	1.31	1.28	1.21	1.11
1000	1.58	1.55	1.49	1.43	1.38	1.35	1.30	1.27	1.20	1.08

TABLA G (continuación) Valores de cuantiles de la distribución F

$1 - \alpha = 0.95$

	ν_1 = *grados de libertad del numerador*									
ν_2	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.97
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.73
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85
1000	3.85	3.01	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84

TABLA G (continuación) Valores de cuantiles de la distribución F

$1 - \alpha = 0.95$

	ν_1 = *grados de libertad del numerador*									
ν_2	11	12	15	20	25	30	40	50	100	1000
1	242.98	243.91	245.96	248.01	249.26	250.08	251.15	251.77	253.01	254.17
2	19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48	19.49	19.50
3	8.76	8.74	8.70	8.66	8.63	8.62	8.59	8.58	8.55	8.53
4	5.94	5.91	5.86	5.80	5.77	5.74	5.72	5.70	5.66	5.63
5	4.70	4.68	4.62	4.56	4.52	4.50	4.46	4.44	4.41	4.37
6	4.03	4.00	3.94	3.87	3.84	3.81	3.77	3.75	3.71	3.67
7	3.60	3.57	3.51	3.44	3.40	3.38	3.34	3.32	3.27	3.23
8	3.31	3.28	3.22	3.15	3.11	3.08	3.04	3.02	2.97	2.93
9	3.10	3.07	3.01	2.94	2.89	2.86	2.83	2.80	2.76	2.71
10	2.94	2.91	2.85	2.77	2.73	2.70	2.66	2.64	2.59	2.54
11	2.82	2.79	2.72	2.65	2.60	2.57	2.53	2.51	2.46	2.41
12	2.72	2.69	2.62	2.54	2.50	2.47	2.43	2.40	2.35	2.30
13	2.63	2.60	2.53	2.46	2.41	2.38	2.34	2.31	2.26	2.21
14	2.57	2.53	2.46	2.39	2.34	2.31	2.27	2.24	2.19	2.14
15	2.51	2.48	2.40	2.33	2.28	2.25	2.20	2.18	2.12	2.07
16	2.46	2.42	2.35	2.28	2.23	2.19	2.15	2.12	2.07	2.02
17	2.41	2.38	2.31	2.23	2.18	2.15	2.10	2.08	2.02	1.97
18	2.37	2.34	2.27	2.19	2.14	2.11	2.06	2.04	1.98	1.92
19	2.34	2.31	2.23	2.16	2.11	2.07	2.03	2.00	1.94	1.88
20	2.31	2.28	2.20	2.12	2.07	2.04	1.99	1.97	1.91	1.85
21	2.28	2.25	2.18	2.10	2.05	2.01	1.96	1.94	1.88	1.82
22	2.26	2.23	2.15	2.07	2.02	1.98	1.94	1.91	1.85	1.79
23	2.24	2.20	2.13	2.05	2.00	1.96	1.91	1.88	1.82	1.76
24	2.22	2.18	2.11	2.03	1.97	1.94	1.89	1.86	1.80	1.74
25	2.20	2.16	2.09	2.01	1.96	1.92	1.87	1.84	1.78	1.72
26	2.18	2.15	2.07	1.99	1.94	1.90	1.85	1.82	1.76	1.70
27	2.17	2.13	2.06	1.97	1.92	1.88	1.84	1.81	1.74	1.68
28	2.15	2.12	2.04	1.96	1.91	1.87	1.82	1.79	1.73	1.66
29	2.14	2.10	2.03	1.94	1.89	1.85	1.81	1.77	1.71	1.65
30	2.13	2.09	2.01	1.93	1.88	1.84	1.79	1.76	1.70	1.63
35	2.07	2.04	1.96	1.88	1.82	1.79	1.74	1.70	1.63	1.57
40	2.04	2.00	1.92	1.84	1.78	1.74	1.69	1.66	1.59	1.52
50	1.99	1.95	1.87	1.78	1.73	1.69	1.63	1.60	1.52	1.45
60	1.95	1.92	1.84	1.75	1.69	1.65	1.59	1.56	1.48	1.40
80	1.91	1.88	1.79	1.70	1.64	1.60	1.54	1.51	1.43	1.34
100	1.89	1.85	1.77	1.68	1.62	1.57	1.52	1.48	1.39	1.30
200	1.84	1.80	1.72	1.62	1.56	1.52	1.46	1.41	1.32	1.21
500	1.81	1.77	1.69	1.59	1.53	1.48	1.42	1.38	1.28	1.14
1000	1.80	1.76	1.68	1.58	1.52	1.47	1.41	1.36	1.26	1.11

TABLA G (continuación) Valores de cuantiles de la distribución F

$1 - \alpha = 0.99$

	ν_1 = *grados de libertad del numerador*									
ν_2	1	2	3	4	5	6	7	8	9	10
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.50	27.34	27.22
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41
500	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34

TABLA G (continuación) Valores de cuantiles de la distribución F

$1 - \alpha = 0.99$

	ν_1 = *grados de libertad del numerador*									
ν_2	11	12	15	20	25	30	40	50	100	1000
2	99.41	99.42	99.43	99.45	99.46	99.46	99.47	99.48	99.49	99.51
3	27.12	27.03	26.85	26.67	26.58	26.50	26.41	26.35	26.24	26.14
4	14.45	14.37	14.19	14.02	13.91	13.84	13.75	13.69	13.58	13.48
5	9.96	9.89	9.72	9.55	9.45	9.38	9.30	9.24	9.13	9.03
6	7.79	7.72	7.56	7.40	7.29	7.23	7.15	7.09	6.99	6.89
7	6.54	6.47	6.31	6.16	6.06	5.99	5.91	5.86	5.75	5.66
8	5.73	5.67	5.52	5.36	5.26	5.20	5.12	5.07	4.96	4.87
9	5.18	5.11	4.96	4.81	4.71	4.65	4.57	4.52	4.41	4.32
10	4.77	4.71	4.56	4.41	4.31	4.25	4.17	4.12	4.01	3.92
11	4.46	4.40	4.25	4.10	4.00	3.94	3.86	3.81	3.71	3.61
12	4.22	4.16	4.01	3.86	3.76	3.70	3.62	3.57	3.47	3.37
13	4.02	3.96	3.82	3.66	3.57	3.51	3.43	3.38	3.27	3.18
14	3.86	3.80	3.66	3.51	3.41	3.35	3.27	3.22	3.11	3.02
15	3.73	3.67	3.52	3.37	3.28	3.21	3.13	3.08	2.98	2.88
16	3.62	3.55	3.41	3.26	3.16	3.10	3.02	2.97	2.86	2.76
17	3.52	3.46	3.31	3.16	3.07	3.00	2.92	2.87	2.76	2.66
18	3.43	3.37	3.23	3.08	2.98	2.92	2.84	2.78	2.68	2.58
19	3.36	3.30	3.15	3.00	2.91	2.84	2.76	2.71	2.60	2.50
20	3.29	3.23	3.09	2.94	2.84	2.78	2.69	2.64	2.54	2.43
21	3.24	3.17	3.03	2.88	2.78	2.72	2.64	2.58	2.48	2.37
22	3.18	3.12	2.98	2.83	2.73	2.67	2.58	2.53	2.42	2.32
23	3.14	3.07	2.93	2.78	2.69	2.62	2.54	2.48	2.37	2.27
24	3.09	3.03	2.89	2.74	2.64	2.58	2.49	2.44	2.33	2.22
25	3.06	2.99	2.85	2.70	2.60	2.54	2.45	2.40	2.29	2.18
26	3.02	2.96	2.81	2.66	2.57	2.50	2.42	2.36	2.25	2.14
27	2.99	2.93	2.78	2.63	2.54	2.47	2.38	2.33	2.22	2.11
28	2.96	2.90	2.75	2.60	2.51	2.44	2.35	2.30	2.19	2.08
29	2.93	2.87	2.73	2.57	2.48	2.41	2.33	2.27	2.16	2.05
30	2.91	2.84	2.70	2.55	2.45	2.39	2.30	2.24	2.13	2.02
35	2.80	2.74	2.60	2.44	2.35	2.28	2.19	2.14	2.02	1.90
40	2.73	2.66	2.52	2.37	2.27	2.20	2.11	2.06	1.94	1.82
50	2.62	2.56	2.42	2.27	2.17	2.10	2.01	1.95	1.82	1.70
60	2.56	2.50	2.35	2.20	2.10	2.03	1.94	1.88	1.75	1.62
80	2.48	2.42	2.27	2.12	2.01	1.94	1.85	1.79	1.65	1.51
100	2.43	2.37	2.22	2.07	1.97	1.89	1.80	1.74	1.60	1.45
200	2.34	2.27	2.13	1.97	1.87	1.79	1.69	1.63	1.48	1.30
500	2.28	2.22	2.07	1.92	1.81	1.74	1.63	1.57	1.41	1.20
1000	2.27	2.20	2.06	1.90	1.79	1.72	1.61	1.54	1.38	1.16

TABLA H *k*-valores para los límites de tolerancia bilaterales cuando se muestrean distribuciones normales

n \ d	$\gamma = 0.75$					$\gamma = 0.90$				
	0.75	0.90	0.95	0.99	0.999	0.75	0.90	0.95	0.99	0.999
6	1.704	2.429	2.889	3.779	4.802	2.196	3.131	3.723	4.870	6.188
7	1.624	2.318	2.757	3.611	4.593	2.034	2.902	3.452	4.521	5.750
8	1.568	2.238	2.663	3.491	4.444	1.921	2.743	3.264	4.278	5.446
9	1.525	2.178	2.593	3.400	4.330	1.839	2.626	3.125	4.098	5.220
10	1.492	2.131	2.537	3.328	4.241	1.775	2.535	3.018	3.959	5.046
11	1.465	2.093	2.493	3.271	4.169	1.724	2.463	2.933	3.849	4.906
12	1.443	2.062	2.456	3.223	4.110	1.683	2.404	2.863	3.758	4.792
13	1.425	2.036	2.424	3.183	4.059	1.648	2.355	2.805	3.682	4.697
14	1.409	2.013	2.398	3.148	4.016	1.619	2.314	2.756	3.618	4.615
15	1.395	1.994	2.375	3.118	3.979	1.594	2.278	2.713	3.562	4.545
16	1.383	1.977	2.355	3.092	3.946	1.572	2.246	2.676	3.514	4.484
17	1.372	1.962	2.337	3.069	3.917	1.552	2.219	2.643	3.471	4.430
18	1.363	1.948	2.321	3.048	3.891	1.535	2.194	2.614	3.433	4.382
19	1.355	1.936	2.307	3.030	3.867	1.520	2.172	2.588	3.399	4.339
20	1.347	1.925	2.294	3.013	3.846	1.506	2.152	2.564	3.368	4.300
21	1.340	1.915	2.282	2.998	3.827	1.493	2.135	2.543	3.340	4.264
22	1.334	1.906	2.271	2.984	3.809	1.482	2.118	2.524	3.315	4.232
23	1.328	1.898	2.261	2.971	3.793	1.471	2.103	2.506	3.292	4.203
24	1.322	1.891	2.252	2.959	3.778	1.462	2.089	2.489	3.270	4.176
25	1.317	1.883	2.244	2.948	3.764	1.453	2.077	2.474	3.251	4.151
26	1.313	1.877	2.236	2.938	3.751	1.444	2.065	2.460	3.232	4.127
27	1.309	1.871	2.229	2.929	3.740	1.437	2.054	2.447	3.215	4.106
28	1.305	1.865	2.222	2.920	3.728	1.430	2.044	2.435	3.199	4.085
29	1.301	1.860	2.216	2.911	3.718	1.423	2.034	2.424	3.184	4.066
30	1.297	1.855	2.210	2.904	3.708	1.417	2.025	2.413	3.170	4.049
31	1.294	1.850	2.204	2.896	3.699	1.411	2.017	2.403	3.157	4.032
32	1.291	1.846	2.199	2.890	3.690	1.405	2.009	2.393	3.145	4.016
33	1.288	1.842	2.194	2.883	3.682	1.400	2.001	2.385	3.133	4.001
34	1.285	1.838	2.189	2.877	3.674	1.395	1.994	2.376	3.122	3.987
35	1.283	1.834	2.185	2.871	3.667	1.390	1.988	2.368	3.112	3.974
36	1.280	1.830	2.181	2.866	3.660	1.386	1.981	2.361	3.102	3.961
37	1.278	1.827	2.177	2.860	3.653	1.381	1.975	2.353	3.092	3.949
38	1.275	1.824	2.173	2.855	3.647	1.377	1.969	2.346	3.083	3.938
39	1.273	1.821	2.169	2.850	3.641	1.374	1.964	2.340	3.075	3.927
40	1.271	1.818	2.166	2.846	3.635	1.370	1.959	2.334	3.066	3.917
41	1.269	1.815	2.162	2.841	3.629	1.366	1.954	2.328	3.059	3.907
42	1.267	1.812	2.159	2.837	3.624	1.363	1.949	2.322	3.051	3.897
43	1.266	1.810	2.156	2.833	3.619	1.360	1.944	2.316	3.044	3.888
44	1.264	1.807	2.153	2.829	3.614	1.357	1.940	2.311	3.037	3.879
45	1.262	1.805	2.150	2.826	3.609	1.354	1.935	2.306	3.030	3.871
46	1.261	1.802	2.148	2.822	3.605	1.351	1.931	2.301	3.024	3.863
47	1.259	1.800	2.145	2.819	3.600	1.348	1.927	2.297	3.018	3.855
48	1.258	1.798	2.143	2.815	3.596	1.345	1.924	2.292	3.012	3.847
49	1.256	1.796	2.140	2.812	3.592	1.343	1.920	2.288	3.006	3.840
50	1.255	1.794	2.138	2.809	3.588	1.340	1.916	2.284	3.001	3.833

TABLA H (continuación) *k*-valores para los límites de tolerancia bilaterales cuando se muestrean distribuciones normales

n \ d	$\gamma = 0.95$					$\gamma = 0.99$				
	0.75	0.90	0.95	0.99	0.999	0.75	0.90	0.95	0.99	0.999
6	2.604	3.712	4.414	5.775	7.337	3.743	5.337	6.345	8.301	10.548
7	2.361	3.369	4.007	5.248	6.676	3.233	4.613	5.488	7.187	9.142
8	2.197	3.136	3.732	4.891	6.226	2.905	4.147	4.936	6.468	8.234
9	2.078	2.967	3.532	4.631	5.899	2.677	3.822	4.550	5.966	7.600
10	1.987	2.839	3.379	4.433	5.649	2.508	3.582	4.265	5.594	7.129
11	1.916	2.737	3.259	4.277	5.452	2.378	3.397	4.045	5.308	6.766
12	1.858	2.655	3.162	4.150	5.291	2.274	3.250	3.870	5.079	6.477
13	1.810	2.587	3.081	4.044	5.158	2.190	3.130	3.727	4.893	6.240
14	1.770	2.529	3.012	3.955	5.045	2.120	3.029	3.608	4.737	6.043
15	1.735	2.480	2.954	3.878	4.949	2.060	2.945	3.507	4.605	5.876
16	1.705	2.437	2.903	3.812	4.865	2.009	2.872	3.421	4.492	5.732
17	1.679	2.400	2.858	3.754	4.791	1.965	2.808	3.345	4.393	5.607
18	1.655	2.366	2.819	3.702	4.725	1.926	2.753	3.279	4.307	5.497
19	1.635	2.337	2.784	3.656	4.667	1.891	2.703	3.221	4.230	5.399
20	1.616	2.310	2.752	3.615	4.614	1.860	2.659	3.168	4.161	5.312
21	1.599	2.286	2.723	3.577	4.567	1.833	2.620	3.121	4.100	5.234
22	1.584	2.264	2.697	3.543	4.523	1.808	2.584	3.078	4.044	5.163
23	1.570	2.244	2.673	3.512	4.484	1.785	2.551	3.040	3.993	5.098
24	1.557	2.225	2.651	3.483	4.447	1.764	2.522	3.004	3.947	5.039
25	1.545	2.208	2.631	3.457	4.413	1.745	2.494	2.972	3.904	4.985
26	1.534	2.193	2.612	3.432	4.382	1.727	2.469	2.941	3.865	4.935
27	1.523	2.178	2.595	3.409	4.353	1.711	2.446	2.914	3.828	4.888
28	1.514	2.164	2.579	3.388	4.326	1.695	2.424	2.888	3.794	4.845
29	1.505	2.152	2.554	3.368	4.301	1.681	2.404	2.864	3.763	4.805
30	1.497	2.140	2.549	3.350	4.278	1.668	2.385	2.841	3.733	4.768
31	1.489	2.129	2.536	3.332	4.256	1.656	2.367	2.820	3.706	4.732
32	1.481	2.118	2.524	3.316	4.235	1.644	2.351	2.801	3.680	4.699
33	1.475	2.108	2.512	3.300	4.215	1.633	2.335	2.782	3.655	4.668
34	1.468	2.099	2.501	3.286	4.197	1.623	2.320	2.764	3.632	4.639
35	1.462	2.090	2.490	3.272	4.179	1.613	2.306	2.748	3.611	4.611
36	1.455	2.081	2.479	3.258	4.161	1.604	2.293	2.732	3.590	4.585
37	1.450	2.073	2.470	3.246	4.146	1.595	2.281	2.717	3.571	4.560
38	1.446	2.068	2.464	3.237	4.134	1.587	2.269	2.703	3.552	4.537
39	1.441	2.060	2.455	3.226	4.120	1.579	2.257	2.690	3.534	4.514
40	1.435	2.052	2.445	3.213	4.104	1.571	2.247	2.677	3.518	4.493
41	1.430	2.045	2.437	3.202	4.090	1.564	2.236	2.665	3.502	4.472
42	1.426	2.039	2.429	3.192	4.077	1.557	2.227	2.653	3.486	4.453
43	1.422	2.033	2.422	3.183	4.065	1.551	2.217	2.642	3.472	4.434
44	1.418	2.027	2.415	3.173	4.053	1.545	2.208	2.631	3.458	4.416
45	1.414	2.021	2.408	3.165	4.042	1.539	2.200	2.621	3.444	4.399
46	1.410	2.016	2.402	3.156	4.031	1.533	2.192	2.611	3.431	4.383
47	1.406	2.011	2.396	3.148	4.021	1.527	2.184	2.602	3.419	4.367
48	1.403	2.006	2.390	3.140	4.011	1.522	2.176	2.593	3.407	4.352
49	1.399	2.001	2.384	3.133	4.002	1.517	2.169	2.584	3.396	4.337
50	1.396	1.969	2.379	3.126	3.993	1.512	2.162	2.576	3.385	4.323

Source: C. Eisenhart, M. W. Hastay, and W. A. Wallis, *Techniques of statistical analysis*, McGraw-Hill, New York, 1947. Publicado con permiso.

TABLA I *k*-valores para los límites de tolerancia unilaterales cuando se muestrean distribuciones normales

n \ d	$\gamma = 0.75$					$\gamma = 0.90$				
	0.75	0.90	0.95	0.99	0.999	0.75	0.90	0.95	0.99	0.999
6	1.087	1.860	2.336	3.243	4.273	1.540	2.494	3.091	4.242	5.556
7	1.043	1.791	2.250	3.126	4.118	1.435	2.333	2.894	3.972	5.201
8	1.010	1.740	2.190	3.042	4.008	1.360	2.219	2.755	3.783	4.955
9	0.984	1.702	2.141	2.977	3.924	1.302	2.133	2.649	3.641	4.772
10	0.964	1.671	2.103	2.927	3.858	1.257	2.065	2.568	3.532	4.629
11	0.947	1.646	2.073	2.885	3.804	1.219	2.012	2.503	3.444	4.515
12	0.933	1.624	2.048	2.851	3.760	1.188	1.966	2.448	3.371	4.420
13	0.919	1.606	2.026	2.822	3.722	1.162	1.928	2.403	3.310	4.341
14	0.909	1.591	2.007	2.796	3.690	1.139	1.895	2.363	3.257	4.274
15	0.899	1.577	1.991	2.776	3.661	1.119	1.866	2.329	3.212	4.215
16	0.891	1.566	1.977	2.756	3.637	1.101	1.842	2.299	3.172	4.164
17	0.883	1.554	1.964	2.739	3.615	1.085	1.820	2.272	3.136	4.118
18	0.876	1.544	1.951	2.723	3.595	1.071	1.800	2.249	3.106	4.078
19	0.870	1.536	1.942	2.710	3.577	1.058	1.781	2.228	3.078	4.041
20	0.865	1.528	1.933	2.697	3.561	1.046	1.765	2.208	3.052	4.009
21	0.859	1.520	1.923	2.686	3.545	1.035	1.750	2.190	3.028	3.979
22	0.854	1.514	1.916	2.675	3.532	1.025	1.736	2.174	3.007	3.952
23	0.849	1.508	1.907	2.665	3.520	1.016	1.724	2.159	2.987	3.927
24	0.845	1.502	1.901	2.656	3.509	1.007	1.712	2.145	2.969	3.904
25	0.842	1.496	1.895	2.647	3.497	0.999	1.702	2.132	2.952	3.882
30	0.825	1.475	1.869	2.613	3.454	0.966	1.657	2.080	2.884	3.794
35	0.812	1.458	1.849	2.588	3.421	0.942	1.623	2.041	2.833	3.730
40	0.803	1.445	1.834	2.568	3.395	0.923	1.598	2.010	2.793	3.679
45	0.795	1.435	1.821	2.552	3.375	0.908	1.577	1.986	2.762	3.638
50	0.788	1.426	1.811	2.538	3.358	0.894	1.560	1.965	2.735	3.604

TABLA I (continuación) *k*-valores para los límites de tolerancia unilaterales cuando se muestrean distribuciones normales

n \ d	$\gamma = 0.95$					$\gamma = 0.99$				
	0.75	0.90	0.95	0.99	0.999	0.75	0.90	0.95	0.99	0.999
6	1.895	3.006	3.707	5.062	6.612	2.849	4.408	5.409	7.334	9.540
7	1.732	2.755	3.399	4.641	6.061	2.490	3.856	4.730	6.411	8.348
8	1.617	2.582	3.188	4.353	5.686	2.252	3.496	4.287	5.811	7.566
9	1.532	2.454	3.031	4.143	5.414	2.085	3.242	3.971	5.389	7.014
10	1.465	2.355	2.911	3.981	5.203	1.954	3.048	3.739	5.075	6.603
11	1.411	2.275	2.815	3.852	5.036	1.854	2.897	3.557	4.828	6.284
12	1.366	2.210	2.736	3.747	4.900	1.771	2.773	3.410	4.633	6.032
13	1.329	2.155	2.670	3.659	4.787	1.702	2.677	3.290	4.472	5.826
14	1.296	2.108	2.614	3.585	4.690	1.645	2.592	3.189	4.336	5.651
15	1.268	2.068	2.566	3.520	4.607	1.596	2.521	3.102	4.224	5.507
16	1.242	2.032	2.523	3.463	4.534	1.553	2.458	3.028	4.124	5.374
17	1.220	2.001	2.486	3.415	4.471	1.514	2.405	2.962	4.038	5.268
18	1.200	1.974	2.453	3.370	4.415	1.481	2.357	2.906	3.961	5.167
19	1.183	1.949	2.423	3.331	4.364	1.450	2.315	2.855	3.893	5.078
20	1.167	1.926	2.396	3.295	4.319	1.424	2.275	2.807	3.832	5.003
21	1.152	1.905	2.371	3.262	4.276	1.397	2.241	2.768	3.776	4.932
22	1.138	1.887	2.350	3.233	4.238	1.376	2.208	2.729	3.727	4.866
23	1.126	1.869	2.329	3.206	4.204	1.355	2.179	2.693	3.680	4.806
24	1.114	1.853	2.309	3.181	4.171	1.336	2.154	2.663	3.638	4.755
25	1.103	1.838	2.292	3.158	4.143	1.319	2.129	2.632	3.601	4.706
30	1.059	1.778	2.220	3.064	4.022	1.249	2.029	2.516	3.446	4.508
35	1.025	1.732	2.166	2.994	3.934	1.195	1.957	2.431	3.334	4.364
40	0.999	1.697	2.126	2.941	3.866	1.154	1.902	2.365	3.250	4.255
45	0.978	1.669	2.092	2.897	3.811	1.122	1.857	2.313	3.181	4.168
50	0.961	1.646	2.065	2.863	3.766	1.096	1.821	2.296	3.124	4.096

Source: G. J. Lieberman, *Table for one-sided statistical tolerance limits,* Industrial Quality Control XIV, 1958, 7–9. Reprinted with permission.

TABLA J Valores de cuantiles superiores de la distribución de la estadística D_n de Kolgomorov-Smirnov

	$1 - \alpha$				
n	0.80	0.85	0.90	0.95	0.99
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.828
4	.494	.525	.564	.624	.733
5	.446	.474	.510	.565	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.410	.490
11	.307	.326	.352	.391	.468
12	.295	.313	.338	.375	.450
13	.284	.302	.325	.361	.433
14	.274	.292	.314	.349	.418
15	.266	.283	.304	.338	.404
16	.258	.274	.295	.328	.392
17	.250	.266	.286	.318	.381
18	.244	.259	.278	.309	.371
19	.237	.252	.272	.301	.363
20	.231	.246	.264	.294	.356
25	.21	.22	.24	.27	.32
30	.19	.20	.22	.24	.29
35	.18	.19	.21	.23	.27
Fórmula para una n mayor	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Fuente: F. J. Massey, Jr., *The Kolmogorov-Smornov test for goodness of fit,* J. Amer Statistical Assoc. 46 (1951), 68-78. Publicado con permiso.

TABLA K Límites de la estadística de Durbin-Watson

$1 - \alpha = 0.95$

	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

TABLA K (continuación) Límites de la estadística de Durbin-Watson

$1 - \alpha = 0.99$

	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0.81	1.07	0.70	1.25	0.59	1.46	0.49	1.70	0.39	1.96
16	0.84	1.09	0.74	1.25	0.63	1.44	0.53	1.66	0.44	1.90
17	0.87	1.10	0.77	1.25	0.67	1.43	0.57	1.63	0.48	1.85
18	0.90	1.12	0.80	1.26	0.71	1.42	0.61	1.60	0.52	1.80
19	0.93	1.13	0.83	1.26	0.74	1.41	0.65	1.58	0.56	1.77
20	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
21	0.97	1.16	0.89	1.27	0.80	1.41	0.72	1.55	0.63	1.71
22	1.00	1.17	0.91	1.28	0.83	1.40	0.75	1.54	0.66	1.69
23	1.02	1.19	0.94	1.29	0.86	1.40	0.77	1.53	0.70	1.67
24	1.04	1.20	0.96	1.30	0.88	1.41	0.80	1.53	0.72	1.66
25	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
26	1.07	1.22	1.00	1.31	0.93	1.41	0.85	1.52	0.78	1.64
27	1.09	1.23	1.02	1.32	0.95	1.41	0.88	1.51	0.81	1.63
28	1.10	1.24	1.04	1.32	0.97	1.41	0.90	1.51	0.83	1.62
29	1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	0.96	1.51	0.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	0.98	1.51	0.92	1.60
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	0.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	0.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	0.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	0.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

Fuente: J. Durbin and G. S. Watson, *Testing for serial correlation in least squares regression, II,* Biometrika **38** (1951), 159-178. Publicado con permiso de Biometrika Trustees.

Respuestas a los ejercicios seleccionados de número impar

Capítulo 1

1.1. *a), b)*

Límites verdaderos	*Frecuencia de la clase*	*Frecuencia relativa*	*Frecuencia relativa acumulativa*
(0.15, 1.55)	17	0.34	0.34
(1.55, 2.95)	11	0.22	0.56
(2.95, 4.35)	7	0.14	0.70
(4.35, 5.75)	6	0.12	0.82
(5.75, 7.15)	4	0.08	0.90
(7.15, 8.55)	3	0.06	0.96
(8.55, 9.95)	2	0.04	1.00
Totales	50	1.00	

c) Los intervalos intercuantil e interdecil son, en forma aproximada, iguales a 3.8 min y 6.7 min, respectivamente.

d) $\bar{x} = 3.258$; *Mediana* = 2.6182; *Moda* = 0.85; $s = 2.4986$; *M.D.* = 2.081; y *Md.D.* = 2.0042

e) $\bar{x} = 3.26$; *Median* = 2.75; *Moda* = 0.4; $s = 2.4819$; *M.D.* = 2.0056; y *Md.D.* = 1.948

1.3. $\bar{x} = 3.5$; las varianzas son $s_1^2 = 3.5$, $s_2^2 = 7.5$, y $s_3^2 = 109.9$

1.5. *a)*

Límites verdaderos	*Frecuencia*	*Frecuencia relativa*
(−1.875, −1.125)	5	0.1667
(−1.125, −0.375)	5	0.1667
(−0.375, 0.375)	8	0.2667
(0.375, 1.125)	8	0.2667
(1.125, 1.875)	4	0.1333
Totales	30	1.0001

Ningún cambio en la distribución de frecuencia relativa

1.7. *b)* $\bar{x} = 18.82$; *Moda* $= 14.0$
c) $s^2 = 123.4196$; $s = 11.11$; *M.D.* $= 9.27$

Capítulo 2

2.1. *a)* Los eventos no son mutuamente excluyentes
b) 1. 180/400; 2. 150/400; 3. 30/400; 4. 60/180; 5. 60/200
c) $P(S \mid M) = 50/220$; $P(S) = 150/400$; no son estadísticamente independientes
d) No; $P(A \mid F) = 0.1111$, $P(A) = 0.125$
e) 1. 240/400; 2. 210/400; 3. 60/400; 4. 30/50

2.3. Cuando alguno o ambos eventos son vacíos

2.5. Las permutaciones son GGG, GGB, GBG, BGG, BBG, BGB, y BBB. La probabilidad de tener dos niños del mismo sexo es 6/8; la probabilidad de un niño y dos niños es 3/8; la probabilidad de que todos sean del mismo sexo es 2/8.

2.7. $(1/2)^{10}$; $1/2$

2.9. $13/30$

2.11. *a)* Cuatro resultados posibles: ambos componentes trabajan; ambos no y uno trabaja y el otro no (en dos formas posibles)
b) 0.99

2.13. $n = 4$

2.15. 0.6571

2.17. 0.41

Capítulo 3

3.1. *a), c)*

x	$p(x)$	$F(x)$
0	0.0498	0.0498
1	0.1494	0.1992
2	0.2240	0.4232
3	0.2240	0.6472
4	0.1680	0.8152
5	0.1008	0.9160
6	0.0504	0.9664
7	0.0216	0.9880

3.3. *a)* $3/2$; *b)* $(x^3 + 1)/2$; *c)* $7/16$, $1/8$

3.5. *a)* $(1/100)\exp(-x/100)$; *b)* 0.8187

3.7. $E(X) = 4$; $Var(X) = 4.1$

3.9. *a)* $1/3$; *b)* $1/18$

3.11. *a)* 4; *b)* 16; *c)* 2; *d)* 9
e) La distribución del ejercicio 3.10 es simétrica y se encuentra centrada alrededor del valor 5, tiene varianza igual a 8.33 y desviación estándar de 2.8868. Esta distribución tiene un sesgo positivo y un pico relativamente grande; la dispersión relativa también es grande.

3.13. *a)* $\sigma^2 + (\mu - c)^2$; *b)* $c = \mu$

3.15. *M.D.*$(X) = 0.19753$, *d.e.* $(X) = 0.2357$

3.17. *a)* Media = 800, Mediana = 554.52; *b)* 878.89 *c)* 1757.78; *d)* 0.3679

3.19. *a)* $(1 - 4t)^{-2}$ *b)* $E(X) = 8$, $Var(X) = 32$

Capítulo 4

4.5. *a)* 0.6562, 0.9346; *b)* 0.5696, 0.9391

4.7. $P(X = 4) = 0.0049$, $P(X \geq 4) = 0.0055$; existe una inclinación a concluir que la afirmación es incorrecta

4.9. 0.2122

4.11. Seis o más

4.15. 0.7601, 0.9718

4.17. 0.0488

4.19. 0.0803, 0.9862

4.21. 0.6767

4.23. 0.0293, sí

4.25. 0.1837, no

4.27. *a)* 0.5973; *b)* 5987; *c)* 0.6065

4.31.

x	*Frecuencia relativa*	*Probabilidad teórica*
0	0.715	0.7201
1	0.179	0.1689
2	0.063	0.0630
3	0.019	0.0263
4	0.010	0.0116
5	0.010	0.0053
6	0.002	0.0025
7	0.000	0.0012
8	0.002	0.0006

4.33. *a)* 0.0189; *b)* 0.0180; la ocurrencia es poco probable

Capítulo 5

5.3. *a)* 0.4649; *b)* 0.2204; *c)* 0.0228; *d)* 0.8643

5.5. *a)* 1.775; *b)* 18.225; *c)* 21.65; *d)* −1.65; *e)* 0.2; *f)* 19.8

5.7. 1018

5.9. 0.00069

5.11. 0.000008; la ocurrencia es muy poco probable

5.13. $228 000

5.15. *a)* 0.0256; *b)* ≈0; *c)* ≈0

5.17. Sí, la probabilidad de ocurrencia es virtualmente cero.

5.19. *a)* 0.5774; *b)* no

5.21. *a)* $= 4$, $b = 16$

5.23. *b)* $E(X) = 0.75$, $Var(X) = 0.0375$, $D.M.(X) = 0.1582$, $\alpha_3(X) = -0.8607$, $\alpha_4(X) = 3.0952$
c) 0.6679, 0.9523; *d)* 0.63, 0.7937, 0.9086

5.25. 0.64, 0.9728

5.27. 0.1314, 0.0582

5.29. *a)* 0.594; *b)* 0.0466; *c)* 0.2642

5.31. $\alpha = 3.75$, $\theta = 8$

5.35. 0.9409

5.37. *a)* $E(X) = 44.3113$; 16.23, 23.62, 29.86, 35.74, 41.63, 47.86, 54.86, 63.43, 75.87
b) 0.1054

5.39. *a)* 0.3679; *b)* 0.8647, 0.9502

5.41. *a)* 0.1353; *b)* 433

5.45. Exponencial con parámetro de escala θ^α

Capítulo 6

6.1. 0.0022; la ocurrencia es poco probable

6.3. *a)* $F(x, y) = (3x^2y - xy^2 - 3x^2 + x - 3y + y^2 + 2)/10$
b) 0.225
c) $F_X(x) = (3x - 1)(x - 1)/5$, $F_Y(y) = (9y - y^2 - 8)/10$
d) $f_X(x) = (6x - 4)/5$, $f_Y(y) = (9 - 2y)/10$

6.5. *a)* $p_X(x) = p_Y(y) = 5/16, 6/16, 5/16$, $x = y = -1, 0, 1$, respectivamente
b) no; *c)* 0

6.7. *a)* 0.69; *b)* $f_{T_1}(t_1) = (1/5)\exp(-t_1/5)$, $f_{T_2}(t_2) = 10\exp(-10t_2)$

6.9. 1029.2152

6.13. Sí $Cov(X, Y) > 0$, $Var(X + Y) > Var(X) + Var(Y)$ y $Var(X - Y) < Var(X) + Var(Y)$; si $Cov(X, Y) < 0$, $Var(X + Y) < Var(X) + Var(Y)$, y $Var(X - Y) > Var(X) + Var(Y)$

6.15. 11/27

6.17. *a)* $\mu = 0.04$, $\sigma^2 = 0.0014769$; *b)* $f(p \mid x) = 1260p(1 - p)^{34}$
c) $\mu = 0.054$, $\sigma^2 = 0.0013456$; *d)* 0.5432

6.19. *a)* 1/2; *b)* 50
c) $f(x \mid y) = \exp\{-(1/150)[x - 50 - (y - 25)/2]^2\}/\sqrt{150\pi}$
d) $f(x \mid Y = 30) = \exp[-(1/150)(x - 52.5)^2]/\sqrt{150\pi}$, 0.9251

Capítulo 7

7.3. *a)* $\lambda^{\Sigma x_i}\exp(-n\lambda)/\Pi x_i!$; *b)* $p^n(1 - p)^{\Sigma x_i}$
c) $1/(b - a)^n$; *d)* $(1/\sigma\sqrt{2\pi})^n \exp[-\Sigma(x_i - \mu)^2/2\sigma^2]$

7.5. Partes *c)*, *b)*, y *f)*

7.11. 0.0075

7.15. *a)* 0.7698; *b)* 0.9986; *c)* 0.0548; *d)* 0.0228

7.17. ≈0; el inspector debe emprender la acción apropiada

7.19. 255.82

7.23. 0.99

7.25. Muy poco probable $P(T > 3.429) < 0.005$

7.27. Sí; $P(T < -3.516) < 0.001$

7.29. Dudoso; $P(F_{15,20} > 1.999) < 0.10$

Capítulo 8

8.1. *a)* $\text{ECM}(T_1) = p(1 - p)/n$; $\text{ECM}(T_2) = [np(1 - p) + (2p - 1)^2]/(n + 2)^2$
b) No. Para $n = 10$, sí $0.138 < p < 0.862$, $\text{ECM}(T_2) < \text{ECM}(T_1)$; de otro modo $\text{ECM}(T_U) < \text{ECM}(T_2)$. Para $n = 25$, sí $0.142 < p < 0.858$, $\text{ECM}(T_2) < \text{ECM}(T_1)$; de otro modo, $\text{ECM}(T_1) < \text{ECM}(T_2)$

8.5. T_3; $Var(T_3)/Var(T_1) = 0.9$

8.9. $\hat{\lambda} = \overline{X}$

8.11. $\hat{\sigma}^2 = \Sigma X_i^2/2n$; sí

8.13. Los factores de la muestra de forma son −0.0028 y 2.21, respectivamente; la distribución es, es forma aparente, simétrica y ligeramente plana en su parte superior.

8.15. *a)* $\hat{\theta} = 100.0696$; *b)* sí, $\tilde{\theta} = 103.575$; *c)* 0.1057

8.17. *a)* 2532.7; *b)* 0.2061

8.19. *a)* 214.9289; *b)* 0.8410, 0.5340

8.21. (20.1191, 20.6434)

8.27. (151.31, 165.69), (149.75, 167.25), (147.82, 169.18)

8.31. (−3.89, −1.51), (−4.12, −1.28), (−4.58, −0.82), sí

8.33. (4.84, 21.16), (2.07, 23.93), sí

8.35. (146.98, 645.69)

8.39. (0.2048, 4.0744), sí

8.41. (0.0172, 0.0628), (0.0128, 0.0672), (0.0043, 0.0757); existe una razón para dudar det la afirmación

8.43. 663 8.45. *a)* 88; *b)* (66.40, 109.60)

8.47. (2.98514, 3.01486) 8.49. 0.8609, 299 8.51. 152

Capítulo 9

9.1. Prueba *b*

9.3. *a)* H_0: $p = 0.05$ contra H_1: $p > 0.05$; *b)* 0.2642
c) 0.3396, 0.4831, 0.6083, 0.8244, 0.9308, y 0.9757, respectivamente
d) $\alpha = 0.0755$, la potencia es 0.1150, 0.2120, 0.3231, 0.5951, 0.7939 y 0.9087, respectivamente.

9.5. Los valores críticos para la prueba 1 son 19.8333 y 20.1667; para la prueba 2 éstos son 19.7917 y 20.2083.

	Potencia										
μ	19.5	19.6	19.7	19.8	19.9	20.0	20.1	20.2	20.3	20.4	20.5
Prueba 1	≈ 1	0.9974	0.9452	0.6554	0.2126	0.0456	0.2126	0.6554	0.9452	0.9974	≈ 1
Test 2	0.9998	0.9893	0.8643	0.4602	0.0968	0.0124	0.0968	0.4602	0.8643	0.9893	0.9998

9.7. El extremo izquierdo de la distribución de muestreo de $\overline{X}$

9.9. *a)* H_0: $\lambda = 2.5$ contra H_1: $\lambda < 2.5$
b) Para las cuatro semanas, el valor crítico es 5
c) 0.8088

9.11. No puede rechazarse H_0, ya que $\bar{x} = 145 < \bar{x}_0 = 233.8$

9.13. 30

9.15. 7

9.17. 100.62; H_0 no puede rechazarse si el valor propuesto es $\geq$ 100.62

9.19. *a)* Valores relativamente grandes de manera que es fácil rechazar H_0
b) $t = 0.667$; H_0 no puede ser rechazada con $\alpha = 0.1$; el valor p es mayor que 0.2
c) Sí; los valores extremos pueden ser críticos

9.21. $t = 0.54$; H_0 no puede rechazarse

9.23. *a)* Sí; el valor crítico es tres, el valor p es 0.0755
b) $z = 2.05$ y H_0 se rechaza, el valor p es 0.0202

9.25. $z = -3.54$, H_0 se rechaza

9.27. $z = 1.62$, H_0 no puede rechazarse, el valor p es 0.1052

9.29. $[(z_{1-\beta} + z_{1-\alpha})^2(\sigma_X^2 + \sigma_Y^2)]/(\delta_0 - \delta_1)^2$

9.31. $t = -1.36$, H_0 no puede rechazarse; el valor p es 0.19

9.33. $t = -1.729$, H_0 no puede rechazarse

9.37. *a)* $t = 2.11$, H_0 se rechaza; *b)* el valor p es 0.039; *c)* (1.66, 7.84)

9.39. $\chi^2 = 17.28$, H_0 no puede rechazarse

9.41. Aproximadamente 0.1

9.43. *a)* $\chi^2 = 47.04$, H_0 se rechaza
b) los valores en el intervalo (1.8083, 7.1666); no son equidistantes debido a que la distribución de muestreo no es simétrica

9.45. *a)* $f = 3.24$, H_0 se rechaza; *b)* 0.1374

9.47. $z = 3.03$, H_0 se rechaza; el valor p es 0.0012

9.49. $z = 1.33$, H_0 no puede rechazarse; el valor p es 0.1836

Capítulo 10

10.1. $\chi^2 = 12$, H_0 se rechaza; el valor p es aproximadamente 0.008

10.3. *a)* $\chi^2 = 400$, H_0 se rechaza; la conclusión es diferente a la del ejercicio 10.2

10.5. $\chi^2 = 40$, H_0 se rechaza

10.7. $\chi^2 = 4.25$, H_0 no puede rechazarse

10.9. *a)* $\chi^2 = 5.8501$, H_0 no puede rechazarse
b) $\hat{\lambda} = 2.673$, $\chi^2 = 5.8969$, y H_0 a pesar de lo anterior, no puede rechazarse

10.11. La desviación máxima es 0.1263, H_0 no puede rechazarse

10.13. $\chi^2 = 1.0097$ (para $k = 5$ clases), H_0 no puede rechazarse

10.15. $\chi^2 = 7.8628$, H_0 no puede rechazarse

10.17.

2.16	3.78	2.70	1.36
2.60	4.54	3.24	1.62
3.24	5.68	4.06	2.02

10.19. $\chi^2 = 22.04$, H_0 se rechaza

10.21. $\chi^2 = 2.69$, H_0 no puede rechazarse

Capítulo 11

11.1. *a)* (4.3292, 5.6708), el promedio de la muestra de la decimoprimera semana es mayor que el valor del límite de control superior
b) probabilidad ≈ 0; c. 0.0301

11.3. *a)* (475.5051, 524.4949), no; *b)* 0.9884; *c)* (0.574, 37.486)

11.5. *a)* (378.36, 422.04), (0, 31.96); *b)* 16.28

11.7. *a)* (0, 0.0797); *b)* (0, 0.0758); *c)* 0.0013

11.9. *a)* 0.6767; *b)* 0.5438

11.11. 0.5526

11.13. Aproximadamente 0.175

11.15. $n = 99$, $c = 4$; $n = 131$, $c = 5$; $n = 100$, $c = 4$; $n = 116$, $c = 5$

11.17. *a)* 0.3679; *b)* 0.019; *c)* 0.3971; *d)* 0.216

11.19. $n = 65$, $\bar{x}_a = 71.53$

Capítulo 12

12.5.

Fuente	gl	SC	CM	*Valor F*
Tratamientos	2	0.492	0.246000	43.41
Error	12	0.068	0.005667	
Total	14	0.560	$f_{0.95,2,12} = 3.89$	

12.7. *a)*

Fuente	gl	SC	CM	*Valor F*
Tratamientos	3	2305.5	768.50	2.75
Error	28	7838.0	279.93	
Total	31	10143.5	$f_{0.95,3,28} = 2.95$	

b) Los residuos estandarizados no sugieren varianzas desiguales.

12.9. *a)*

Fuente	gl	SC	CM	*Valor F*
Tratamientos	4	522,744	130,686	66.41
Error	20	39,360	1,968	
Total	24	562,104	$f_{0.99,4,20} = 4.43$	

b) Algunos contrastes y sus intervalos de confianza son: $L_1 = \mu_5 - \mu_4$, (41.87, 278); $L_2 = 3\mu_5 - \mu_2 - \mu_3 - \mu_4$, (406.65, 985.35); $L_3 = \mu_2 - \mu_1$, (33.87, 270.13); $L_4 = 2\mu_5 - \mu_3 - \mu_4$, (205.39, 614.61); $L_5 = \mu_3 - \mu_2$, (−82.13, 154.13)

12.11. El uso del análisis de varianza es cuestionable debido a que la variación en el interior de la región es demasiado grande para ser atribuida sólo a un error aleatorio.

12.13. *b)* $Y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}$, $i = 1, 2, \ldots, 5$, $j = 1, 2, 3, 4$; H_0: $\tau_j = 0$ para toda j;

Fuente	gl	SC	CM	*Valor F*
Bloques	4	1026.2875		
Tratamientos	3	17.6260	5.8753	41.09
Error	12	1.7165	0.1430	
Total	19	1045.6300	$f_{0.95,3,12} = 3.49$	

c) Algunos contrastes y sus intervalos de confianza son $L_1 = 3\mu_4 - \mu_1 - \mu_2 - \mu_3$, (4.48, 8.28); $L_2 = \mu_4 - \mu_1$, (1.57, 3.11); $L_3 = \mu_2 + \mu_3 - 2\mu_1$, (−0.70, 1.98); $L_4 = \mu_3 - \mu_2$, (−0.41, 1.13)

12.15. *a)* $Y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}$, $i = 1, 2, \ldots, 7$, $j = 1, 2, 3, 4$

b)

Fuente	gl	SC	CM	*Valor F*
Bloques	6	1,471,772.429		
Tratamientos	3	44,826.572	14,942.19	16.48
Error	18	16,316.428	906.47	
Total	27	1,532,915.429	$f_{0.95,3,18} = 3.16$	

c) $f = 16.48 > f_{0.95,1,6} = 5.99$; H_0 a pesar de esto se rechaza

d) Dos contrastes y sus intervalos de confianza son: $L_1 = \mu_1 - \mu_3$, (12.01, 111.13); $L_2 = \mu_2 - \mu_4$, (−29.13, 69.99)

12.17. *a)* $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, $i = 1, 2$, $j = 1, 2$, $k = 1, 2, 3$

b) H_0: $(\alpha\beta)_{ij} = 0$ para toda i y j; H_0: $\alpha_i = 0$ para toda i y j; H_0: $\beta_j = 0$ para toda j

c)

Fuente	gl	SC	CM	*Valor F*
Horno	1	0.022534	0.022534	3.92
Temperatura	1	0.005634	0.005634	0.98
Horno × Temp	1	0.554699	0.554699	96.47
Error	8	0.046000	0.005750	
Total	11	0.628867	$f_{0.95,1,8} = 5.32$	

12.19. *a)* $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, $i = 1, 2, 3, 4$, $j = 1, 2, 3$, $k = 1, 2, 3, 4$

b) H_0: $\sigma^2_{\alpha\beta} = 0$; H_0: $\alpha_i = 0$ para toda i; H_0: $\sigma^2_\beta = 0$

c)

Fuente	gl	SC	CM	*Valor F*
Variedades	3	331.750	110.58	0.64
Fertilizantes	2	22,764.875	11,382.44	230.65
Variedades × Fert.	6	1,052.125	173.35	3.51
Error	36	1,776.500	49.35	
Total	47	25,925.250	$f_{0.95,3,6} = 4.76$	

$f_{0.95,2,36} = 3.27$ $f_{0.95,6,36} = 2.37$

Capítulo 13

13.3. *a)* $\Sigma Y_i x_i / \Sigma x_i^2$; *b)* $E(B) = \beta$

13.5. *a)* En algún grado; *b)* $\hat{y} = 2.50 + 1.7774x$

c) Para cualquier aumento de $1 000 en el ingreso familiar, la cobertura del seguro de vida también aumenta.

d) Debe ajustarse una ecuación cuadrática

13.7. *a)*

x	45	20	40	40	47	30	25	20	15
residuos	−12.48	11.95	−13.60	−23.60	3.96	−0.82	8.06	−3.05	10.84
x	35	40	55	50	60	15	30	35	45
residuos	0.29	1.40	4.74	18.63	10.85	0.84	−15.82	0.29	−2.48

c) 124.7; *d)* 7.727, 0.2021; *e)* (1.3489, 2.2059); *f)* Sí, $t = 8.79$

g)

x_p	*Intervalo de confianza*	x_p	*Intervalo de confianza*
45	(75.67, 89.29)	35	(59.11, 70.31)
20	(29.23, 46.87)	40	(67.75, 79.45)
40	(67.75, 79.45)	55	(90.38, 110.14)
40	(67.75, 79.45)	50	(83.17, 99.57)
47	(78.73, 93.35)	60	(97.43, 120.87)
30	(49.69, 61.95)	15	(18.60, 39.72)
25	(39.65, 54.23)	30	(49.69, 61.95)
20	(29.23, 46.87)	35	(59.11, 70.31)
15	(18.60, 39.72)	45	(75.67, 89.29)

13.9.

x_p	$\hat{y}_{part}$	*intervalo de predicción del 95%*
18	34.49	(8.98, 60.00)
28	52.27	(27.71, 76.83)
38	70.04	(45.70, 94.38)
48	87.82	(62.95, 112.69)
58	105.59	(79.50, 131.68)

13.11. *b)* 0.2262; la asociación lineal es vaga

13.13. *a)* $(\sigma^2/n) \leq Var(\hat{Y}_p) \leq 2(\sigma^2/n)$;

b) $[(n + 1)\sigma^2/n] \leq Var(\hat{Y}_{part}) \leq [(n + 2)\sigma^2/n]$;

c) $b_0 = 15.5$, $b_1 = 5.1$

13.15. *b)* $\hat{y} = 12.75 + 19.875x$; *c)* sí, $t = 14.24$; *d)* (16.881, 22.869)

13.17. *a)* $r = -0.8829$

b)

Fuente	gl	SC	CM	*Valor F*
Regresión	1	5.64305	5.64305	63.65
Error	18	1.59595	0.08866	
Total	19	7.23900	$f_{0.99,1,18} = 8.29$	

13.19. *a)* $\hat{y} = -53.119 + 0.6639x$; *b)* sí

c) se detecta una autocorrelación positiva $d = 0.7075$

d) $\hat{y} = -45.116 + 0.6704x$

Capítulo 14

14.1. *a)* β_2 es no lineal

14.3. *a)* 3; b. -3.5

14.5. *a)*

Variable en el modelo	b_0	b_1	b_2
x_1	0.1619	0.1342	
x_2	0.6713		−0.0363
x_1, x_2	−0.1605	0.1487	0.0769

b) $f = 113.14$, rechazar H_0

c) $\text{SCR}(x_2 \mid x_1) = 0.0879$, $f = 14.63$; $\text{SCR}(x_1 \mid x_2) = 1.3366$, $f = 222.47$

d) $R^2 = 0.9496$

e) $\hat{y} = -0.1605 + 0.1487x_1 + 0.0769x_2$, $\hat{y}_p = \$518.85$, (\$462.7, \$575.0)

14.7. $\begin{bmatrix} 15 & 42.00 & 55.0 \\ 42 & 188.08 & 140.8 \\ 55 & 140.80 & 219.0 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 8.070 \\ 32.063 \\ 28.960 \end{bmatrix}$

14.9. *a)*

Variable en el modelo	R^2	SCE	CME	C_p
x_1	0.000	7326	407.02	114.07
x_2	0.229	5648	313.77	84.27
x_3	0.784	1581	87.82	12.06
x_4	0.748	1846	102.53	16.76
x_1, x_2	0.230	5641	332.00	86.20
x_1, x_3	0.802	1451	85.34	11.75
x_1, x_4	0.754	1800	106.00	17.97
x_2, x_3	0.785	1576	92.73	13.99
x_2, x_4	0.774	1653	97.20	15.36
x_3, x_4	0.869	958	56.34	3.00
x_1, x_2, x_3	0.802	1451	90.70	13.77
x_1, x_2, x_4	0.778	1624	102.00	16.85
x_1, x_3, x_4	0.885	846	52.88	3.02
x_2, x_3, x_4	0.870	950	59.38	4.87
x_1, x_2, x_3, x_4	0.885	845	56.33	5.00

c) $\hat{y} = -114.988 + 1.2657x_3 + 0.8414x_4$, $\hat{y}_{part} = 100.35$, (82.05, 118.64)

14.11. 0.0654, 0.0952

14.13. La gráfica revela una tendencia cuadrática $\hat{y} = 8.238 + 0.3126x - 0.001823x^2$; $\hat{y} = -27.55\%$, lo cual es absurdo

14.15. *a)* $\hat{y} = 5284.28 - 114.85x_1 - 78.67x_2 + 0.129x_3 + 0.189x_1^2 + 0.201x_2^2 + 2.63 \times 10^{-7}x_3^2 + 1.268x_1x_2 - 0.0017x_1x_3 - 0.0008x_2x_3$

b) La elección para la mejor ecuación se encuentra entre las dos siguientes:

$\hat{y} = 2163.98 - 56.47x_1 - 26.17x_2 + 0.0162x_3 + 0.6952x_1x_2 - 0.0005x_1x_3$

$\hat{y} = 1676.19 - 43.57x_1 - 19.77x_2 + 0.526x_1x_2 - 5.91 \times 10^{-5}x_1x_3$

14.17.

	Y	x_1	x_2	x_3	x_4	x_5
Y	1.00	0.18	0 78	0.15	−0.29	0.45
x_1	0.18	1.00	−0.05	0.41	0.55	−0.04
x_2	0.78	−0.05	1.00	0.08	−0.30	0.16
x_3	0.15	0.41	0.08	1.00	0.27	−0.14
x_4	−0.29	0.55	−0.30	0.27	1.00	−0.11
x_5	0.45	−0.04	0.16	−0.14	−0.11	1.00

Capítulo 15

15.1. Sí, $t = -2.12$ y se rechaza H_0

15.3. $z = -2.51$, H_0 se rechaza

15.5. $z = -2.80$ y, por lo tanto, existe una razón para creer que la secuencia no es aleatoria.

15.7. $z = 2.24$ y la H_0 de que no existe diferencia entre la preferencia se rechaza. La conclusión es diferente a la del ejercicio 15.6.

15.9. $s = 4$, los valores críticos son 2 y 10, no puede rechazarse H_0

15.11. $h = 11.40$ y se rechaza la hipótesis H_0 de que las distribuciones son idénticas para ambas disciplinas.

15.13. $s = 1.1$ y las diferencias entre las cuatro pruebas no son estadísticamente significativas.

15.15. $r_S = 0.7915$

15.17. $r_S = -0.4667$; existe alguna tendencia en uno de los jueces para dar un puntaje alto cuando los demás lo dan bajo.

Índice analítico

Análisis:
  residual:
    de regresión, 532
    de varianza, 415
  de varianza, 405
    para análisis de regresión, 469, 508
    para experimentos factoriales, 428
    para un solo factor, 405, 420
    para modelos de efectos aleatorios, 418, 434
Asimetría, coeficiente de, 70
Autocorrelación, 479
Axiomas de probabilidad, 34, 35

Bondad del ajuste, 363

Coeficiente:
  de asimetría, 70
  de confianza, 273
  de correlación, 192
    de la muestra, 477
    múltiple, 508
      corregido, 567
  de determinación, 475
  de regresión, inferencia para, 508
    en el modelo lineal general, 508
    en el modelo lineal sencillo, 465
  de variación, 69
Combinaciones, 47
Contraste, 413
Correlación lineal, 477
Corridas, 577
Cota inferior de Cramer-Roo, 260
Covarianza, 191
Criterio del error medio cuadrático, 527
Criterio C_p, 527
Cuartil, 7, 8, 77
Curtosis relativa, 71

Desviación estándar, 15, 68
Diferencia de dos medias, 7, 8
  distribución de la, 278
  límites de confianza para la, 278
  prueba de hipótesis de la, 333
Diseño:
  en bloque completamente aleatorizado, 403
  completamente aleatorizado, 403
    para experimentos factoriales, 428
    para experimentos con un solo factor, 404
Diseños experimentales, 401
Dispersión, 15, 75
Distribución:
  Beta, 147
    momentos de la, 149
  binominal, 89
    función generadora de momentos de la, 98
    momentos de la, 93
  binominal negativa, 115
    función generadora de momentos de la, 120
    momentos de la, 118
    relación con la binomial, 117
  condicionales, 197
  Chi-cuadrada, 158
    función generadora de momentos de la, 159
  de Erlang, 157
  experimental, 158,163
    función de confiabilidad de la, 166
    momentos de la, 164
  expotencial negativa, 158, 163
    F, 240
    en el análisis de varianza, 409
    para inferencia acerca de las varianzas, 242
  de frecuencia acumulativa, 7
  de una función de variable aleatoria, 167
  gamma, 152
    función generadora de momentos de la, 155
    momentos de la, 153
  geométrica, 117

hipergeométrica, 108
fórmula de recursión para, 110
momentos de la, 113
marginal, 189
de muestreo, 221
de media muestral, 221
de varianza muestral, 232
multinominal, 186
normal, 130
aproximación a la binominal, 142
bivariada, 207
estandarizada, 135
función generadora de momentos de la, 133
logarítmica, 170
momentos de la, 131
propiedad aditiva, 223
Pascal, 116
de poisson, 100
función generadora de momentos de la, 107
momentos de la, 105
relación con la binominal, 104
a posteriori, 200, 201
a priori o distribución inicial, 201, 202
de probabilidad, 53,56
de Rayleigh, 163
t de Student, 234, 249
para la diferencia entre dos medias, 240
para observaciones pareadas, 340
trinominal, 186, 187
uniforme, 143
momentos de la, 144
de Weibull, 159
función de confiabilidad de la, 167
momentos de la, 161
duplicación, 403

Ecuación de segundo orden, 505
Ecuaciones normales, 451, 506
Enfoque matricial para:
el modelo lineal general, 505
el modelo general sencillo, 488
Error:
cuadrático medio, 253
experimental, 402, 406
tipo I, 305
tipo II, 305
varianza, 410, 448

Escala:
de intervalo, 572
nominal, 573
Ordinal, 573
de proporción 572
Espacio muestral, 32
Estadística, 220
distribución de muestreo, 221
de Kolgomorov-Smirnof 368
Estadísticas:
de Durbin-Watson, 480
suficientes, 261
Estimación:
de máxima verosimilitud, 264
para el modelo lineal sencillo, 455
para muestras truncadas, 269
por mínimos cuadrados, 446
para el modelo lineal general, 506, 507
para el modelo lineal sencillo, 448
propiedades de los estimadores MC, 457
puntual, 251
Bayesiana, 285, 288
intervalo, 271
método de, 264
Estimador, 220
Bayes, 286, 287
consistente, 256
eficiente, 260
de máxima verosimilitud, 264
mínimos cuadrados, 446
no sesgado, 255
no sesgado de varianza mínima, 259
Eventos, 33
estadísticamente independientes, 41
mutuamente excluyentes, 33
Experimentos factoriales, 426

Factor, 401
Factores de forma, 72
Frecuencia:
de falla, 164
relativa, 3
distribución de, 3

Función:
beta, 147
beta incompleta, 148
característica de operación, 312
confiabilidad, 164
de densidad conjunta, 187, 188
de densidad de probabilidad, 56, 59
distribución acumulativa, 54, 60
gamma, 65
gamma incompleta, 154
generadora de momentos, 80
de pérdida, 286
potencia, 311, 312
probabilidad, 53-55
verosimilitud, 200, 216, 217

Grados de libertad, 158, 235

Hipótesis nula, 304
alternativa, 307
bilateral, 311
compuesta, 304
sencilla, 304
unilateral, 311
Histograma, 3
Hoja de control:
para desviación estándar, 384, 385
para medias, 381, 384
para proporción, 388

Independencia estadística de eventos, 41
de variables aleatorias. 194
Inferencia:
estadística, 214
inductiva, 2
Interacción, 426, 427
Intervalo de predicción, 469, 509, 510
Intervalos de confianza:
para medias, 274, 278, 279
para proporciones, 282, 350
para varianzas, 280, 281

Jacobiano, 168

Lema de Neyman-Pearson, 315
Ley de los grandes números, 258
Límites:
de clase, 3, 6
de tolerancia, 150
para distribuciones normales, 293
independientes de la distribución 290

Media:
cálculo de la, 11, 12
definición teórica de la, 75
distribución de muestreo para la, 224, 225
intervalo de confianza para la, 267, 274
prueba de hipótesis para la, 327, 333
Mejor conjunto de variables de predicción, 525
Mejor estimador lineal no sesgado, 455
Mejor prueba, 314
Método:
de los momentos, 268
de Scheffé, 413
Métodos:
independientes de la distribución, 249, 290
no paramétricos, 573, 574
Mínimos cuadrados con factores de peso, 535, 548
Moda, 12
Modelo:
autorregresivo, 514
curvilíneal, 504, 538
de efectos aleatorios, 406
de efectos fijos, 406
heterosedástico, 535
lineal general, 503
de primer orden, 504
Momentos:
factoriales, 95
de una variable aleatoria, 67-69
de distribuciones bivariadas, 191
de muestras, 268
Muestra aleatoria, 217
Muestreo para aceptación:
por atributos, 392
por variables, 393
Multicolinealidad, 510, 520

Nivel de calidad aceptable, 392
Notación de suma, 25

Observaciones discordantes (aberrantes o anormales) 533, 535

Parámetro, 1
definición de, 218
tipos de, 88

Permutaciones, 45
Población, 1
Principio:
de aleatorización, 341
de la suma cuadrada extra, 513
Probabilidad:
condicional, 37
conjunta, 36
definición clásica de la, 29
definición de frecuencia relativa de la, 30
interpretación subjetiva de la, 31
marginal, 37
a posteriori o posterior, 44
a priori o inicial, 44
Proporción:
defectuosa tolerable en un lote, 392
diferencia de dos proporciones, 350
intervalo de confianza para 390
Prueba:
para corridas de Wald-Wolfowitz, 577
estadística, 306
F conservadora, 102
F parcial, 516
de hipótesis estadística, 303
de Kruskal-Wallis, 582
de Mann-Whitney, 574
para observaciones pareadas, 340
del rango signado de Wilcoxon para la suma, 580
del signo, 579
uniforme más poderosa, 317
de Wilcoxon para el rango de la suma, 574

Rango, 573
de correlación de Spearman, 586
Recorrido, 20
intercuartil, 20, 77
interdecil, 20, 77
Región crítica, 306
Regla:
de adición de probabilidades, 35
de multiplicación para probabilidades, 38
Regresión:
curva de, 445
curvilineal, 538
lineal múltiple, 503
para el modelo lineal general, 503
para el modelo lineal sencillo, 465
paso a paso, 526, 527, 532
significado de la, 444
suposiciones para la, 446
Repaso de álgebra matricial, 497
Residual, 415, 452
estandarizado, 416
varianza, 455
Riesgo:
del consumidor, 391
del fabricante, 391
Robusto, 338

Sensibilidad, 338
Series, de tiempo, 479
Sesgo, 253
Suma de cuadrados, 409, 472

Tablas de contendencia, 371
Técnicas de relación de variables, 525
Tendencia central, 12, 75
Teorema:
de Bayes, 43, 44, 200
de DeMoivre-Lapace, 141
del límite central, 230, 247, 248,
de Tchebusheff, 257, 258
Tratamiento, 402

Valor:
esperado, 197
condicional, 198
propiedades del, 65-67
P, 326
Valores aleatorios, generación de, 171
para la distribución binominal, 174
para la distribución normal, 174
para la distribución de Poisson, 175
para la distribución de Weibull, 173
Variable
aleatoria, 52
continua, 53
discreta, 53
distribución de una función de, 167, 168
estandarizada, 73
de predicción, 444
de respuesta, 444

Variables:
de indicación, 556
ortogonales, 520
Variación, 15, 76
coeficiente de, 69
Varianza:
cálculo de la, 15
definición teórica de la, 68
intervalos de confianza para, 280, 281
prueba de hipótesis para, 346, 347
Violación de suposiciones, 338