y de ahí deducimos $b_0 = -3.38$ y $b_1 = 1.036$, y por tanto, $X = -3.38 + 1.036Y$. El gráfico de estas ecuaciones se ve en la Figura 14.2.

*Otro método*

$$b_0 = \frac{(\sum X)(\sum Y^2) - (\sum Y)(\sum XY)}{N\sum Y^2 - (\sum Y)^2} = -3.38 \qquad b_1 = \frac{N\sum XY - (\sum Y)(\sum X)}{N\sum Y^2 - (\sum Y)^2} = 1.036$$

**14.2.** Rehacer los Problemas 14.1(*b*) y 14.1(*c*) usando las rectas de regresión

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \quad \text{y} \quad x = \left(\frac{\sum xy}{\sum y^2}\right)y$$

donde $x = X - \bar{X}$ e $y = Y - \bar{Y}$.

**Solución**

*Primer método*

La Tabla 14.3 resume la tarea. La recta de regresión de $Y$ sobre $X$ es

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x = \left(\frac{40.34}{84.68}\right)x = 0.476x \qquad \text{o sea} \qquad Y - 67.6 = 0.476(X - 66.7)$$

**Tabla 14.3**

| $X$ | $Y$ | $x = X - \bar{X}$ | $y = Y - \bar{Y}$ | $x^2$ | $xy$ | $y^2$ |
|---|---|---|---|---|---|---|
| 65 | 68 | −1.7 | 0.4 | 2.89 | −0.68 | 0.16 |
| 63 | 66 | −3.7 | −1.6 | 13.69 | 5.92 | 2.56 |
| 67 | 68 | 0.3 | 0.4 | 0.09 | 0.12 | 0.16 |
| 64 | 65 | −2.7 | −2.6 | 7.29 | 7.02 | 6.76 |
| 68 | 69 | 1.3 | 1.4 | 1.69 | 1.82 | 1.96 |
| 62 | 66 | −4.7 | −1.6 | 22.09 | 7.52 | 2.56 |
| 70 | 68 | 3.3 | 0.4 | 10.89 | 1.32 | 0.16 |
| 66 | 65 | −0.7 | −2.6 | 0.49 | 1.82 | 6.76 |
| 68 | 71 | 1.3 | 3.4 | 1.69 | 4.42 | 11.56 |
| 67 | 67 | 0.3 | −0.6 | 0.09 | −0.18 | 0.36 |
| 69 | 68 | 2.3 | 0.4 | 5.29 | 0.92 | 0.16 |
| 71 | 70 | 4.3 | 2.4 | 18.49 | 10.32 | 5.76 |
| $\sum X = 800$ | $\sum Y = 811$ | | | | | |
| $\bar{X} = 800/12$ | $\bar{Y} = 811/12$ | | | $\sum x^2 = 84.68$ | $\sum xy = 40.34$ | $\sum y^2 = 38.92$ |
| $= 66.7$ | $= 67.6$ | | | | | |

La recta de regresión de $X$ sobre $Y$ es

$$x = \left(\frac{\sum xy}{\sum y^2}\right)y = \left(\frac{40.34}{38.92}\right)y = 1.036y \qquad \text{o sea} \qquad X - 66.7 = 1.036(Y - 67.6)$$

Coinciden con los resultados del Problema 14.1.

*Segundo método*

Restar una constante adecuada, 60, por ejemplo, de cada valor de $X$ e $Y$. Los resultados se pueden ordenar como en la Tabla 14.4. Procedamos con el segundo método del Problema 13.17. Así pues.

$$a_1 = \frac{N\sum X'Y' - (\sum X')(\sum Y')}{N\sum X'^2 - (\sum X')^2} = 0.476 \qquad b_1 = \frac{N\sum X'Y' - (\sum Y')(\sum X')}{N\sum Y'^2 - (\sum Y')^2} = 1.036$$

Como $\bar{X} = 60 + 80/12 = 66.7$ e $\bar{Y} = 60 + 91/12 = 67.6$, las requeridas ecuaciones de regresión son las de antes.

Nótese que si $a_0$ y $b_0$ se hallasen por este método, *no* obtendríamos los mismos resultados que antes, ya que $a_0$ y $b_0$ dependen de la elección del origen. De manera que este método se usa *sólo* para hallar $a_1$ y $b_1$, que son independientes de la elección del origen.

**Tabla 14.4**

| $X'$ | $Y'$ | $X'^2$ | $X'Y'$ | $Y'^2$ |
|---|---|---|---|---|
| 5 | 8 | 25 | 40 | 64 |
| 3 | 6 | 9 | 18 | 36 |
| 7 | 8 | 49 | 56 | 64 |
| 4 | 5 | 16 | 20 | 25 |
| 8 | 9 | 64 | 72 | 81 |
| 2 | 6 | 4 | 12 | 36 |
| 10 | 8 | 100 | 80 | 64 |
| 6 | 5 | 36 | 30 | 25 |
| 8 | 11 | 64 | 88 | 121 |
| 7 | 7 | 49 | 49 | 49 |
| 9 | 8 | 81 | 72 | 64 |
| 11 | 10 | 121 | 110 | 100 |
| $\sum X' = 80$ | $\sum Y' = 91$ | $\sum X'^2 = 618$ | $\sum X'Y' = 647$ | $\sum Y'^2 = 729$ |

## ERROR TIPICO DE ESTIMACION

**14.3.** Si la recta de regresión de $Y$ sobre $X$ viene dada por $Y = a_0 + a_1X$, probar que el error típico de estimación $s_{Y.X}$ viene dado por

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0\sum Y - a_1\sum XY}{N}$$

**Solución**

Los valores de $Y$ estimados por la recta de regresión están dados por $Y_{est} = a_0 + a_1X$. Luego

$$s_{Y.X}^2 = \frac{\sum(Y - Y_{est})^2}{N} = \frac{\sum(Y - a_0 - a_1X)^2}{N}$$
$$= \frac{\sum Y(Y - a_0 - a_1X) - a_0\sum(Y - a_0 - a_1X) - a_1\sum X(Y - a_0 - a_1X)}{N}$$

Ahora bien $$\sum(Y - a_0 - a_1X) = \sum Y - a_0N - a_1\sum X = 0$$

y $$\sum X(Y - a_0 - a_1X) = \sum XY - a_0\sum X - a_1\sum X^2 = 0$$

ya que de las ecuaciones normales

$$\sum Y = a_0N + a_1\sum X$$
$$\sum XY = a_0\sum X + a_1\sum X^2$$

Por tanto $$s_{Y.X}^2 = \frac{\sum Y(Y - a_0 - a_1X)}{N} = \frac{\sum Y^2 - a_0\sum Y - a_1\sum XY}{N}$$

Este resultado puede ser extendido a ecuaciones de regresión no lineales.

**14.4.** Si $x = X - \bar{X}$ e $y = Y - \bar{Y}$, probar que el resultado del Problema 14.3 puede expresarse

$$s_{Y.X}^2 = \frac{\sum y^2 - a_1\sum xy}{N}$$

**Solución**

Del Problema 14.3, con $X = x + \bar{X}$ e $Y = y + \bar{Y}$, tenemos

$$\begin{aligned} Ns_{Y.X}^2 &= \sum Y^2 - a_0\sum Y - a_1\sum XY = \sum(y + \bar{Y})^2 - a_0\sum(y + \bar{Y}) - a_1\sum(x + \bar{X})(y + \bar{Y}) \\ &= \sum(y^2 + 2y\bar{Y} + \bar{Y}^2) - a_0(\sum y + N\bar{Y}) - a_1\sum(xy + \bar{X}y + x\bar{Y} + \bar{X}\bar{Y}) \\ &= \sum y^2 + 2\bar{Y}\sum y + N\bar{Y}^2 - a_0N\bar{Y} - a_1\sum xy - a_1\bar{X}\sum y - a_1\bar{Y}\sum x - a_1N\bar{X}\bar{Y} \\ &= \sum y^2 + N\bar{Y}^2 - a_0N\bar{Y} - a_1\sum xy - a_1N\bar{X}\bar{Y} \\ &= \sum y^2 - a_1\sum xy + N\bar{Y}(\bar{Y} - a_0 - a_1\bar{X}) \\ &= \sum y^2 - a_1\sum xy \end{aligned}$$

donde hemos usado los resultados $\sum x = 0$, $\sum y = 0$ e $\bar{Y} = a_0 + a_1\bar{X}$ (que se siguen al dividir ambos lados de la ecuación normal $\sum Y = a_0N + a_1\sum X$ por $N$).

**14.5.** Calcular el error típico de estimación, $s_{Y.X}$, para los datos del Problema 14.1, usando (*a*) la definición y (*b*) el resultado del Problema 14.4.

**Solución**

(*a*) Según el Problema 14.1(*b*) la recta de regresión de $Y$ sobre $X$ es $Y = 35.82 + 0.476X$. La Tabla 14.5 da los valores reales de $Y$(de la Tabla 14.1) y los valores estimados de $Y$, denotados por $Y_{est}$, que se obtienen de la recta de regresión; por ejemplo, correspondiente a $X = 65$ tenemos $Y_{est} = 35.82 + 0.476(65) = 66.76$. También se recogen los valores $Y - Y_{est}$, que se necesitan al calcular $s_{Y.X}$:

$$s_{Y.X}^2 = \frac{\sum(Y - Y_{est})}{N} = \frac{(1.24)^2 + (0.19)^2 + \cdots + (0.38)^2}{12} = 1.642$$

y $s_{Y.X} = \sqrt{1.642} = 1.28$ in.

**Tabla 14.5**

| $X$ | 65 | 63 | 67 | 64 | 68 | 62 | 70 | 66 | 68 | 67 | 69 | 71 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 68 | 66 | 68 | 65 | 69 | 66 | 68 | 65 | 71 | 67 | 68 | 70 |
| $Y_{est}$ | 66.76 | 65.81 | 67.71 | 66.28 | 68.19 | 65.33 | 69.14 | 67.24 | 68.19 | 67.71 | 68.66 | 69.62 |
| $Y - Y_{est}$ | 1.24 | 0.19 | 0.29 | −1.28 | 0.81 | 0.67 | −1.14 | −2.24 | 2.81 | −0.71 | −0.66 | 0.38 |

(*b*) De los Problemas 14.1, 14.2 y 14.4

$$s_{Y.X}^2 = \frac{\sum y^2 - a_1 \sum xy}{N} = \frac{38.92 - 0.476(40.34)}{12} = 1.643$$

y $s_{Y.X} = \sqrt{1.643} = 1.28$ in.

**14.6.** (*a*) Construir dos rectas paralelas a la recta de regresión del Problema 14.1 y que estén a una distancia vertical $s_{Y.X}$ de ella.

(*b*) Determinar el porcentaje de puntos dato que caen entre esas dos rectas.

**Solución**

(*a*) La recta de regresión $Y = 35.82 + 0.476X$, obtenida en el Problema 14.1, es la de trazo grueso en la Figura 14.3. Las paralelas a distancia vertical $s_{Y.X} = 1.28$ de ella (véase Prob. 14.5), son las de trazo discontinuo en esa figura.

(*b*) De la Figura 14.3 se ve que mientras 7 de los 12 puntos dato caen entre esas rectas, 3 aparecen sobre ellas. Un examen más detallado (usando la fila inferior de la Tabla 14.5, por ejemplo) revela que dos de ellos están entre esas dos rectas. Luego el porcentaje requerido es 9/12 = 75%.

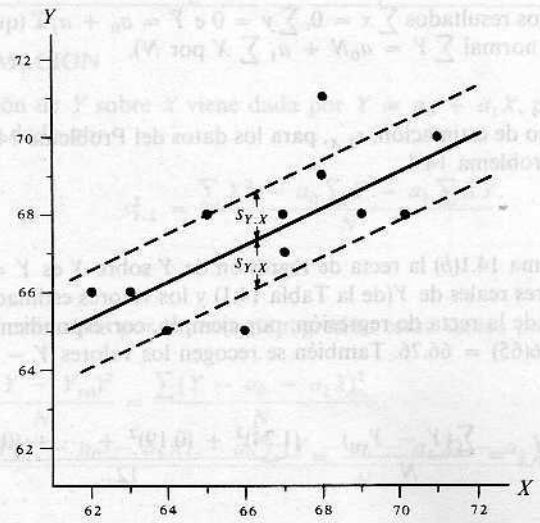



**Figura 14.3.**

*Otro método*

Según la fila de abajo en la Tabla 14.5, $Y - Y_{est}$ está entre $-1.28$ y $1.28$ (por ejemplo, $+s_{Y.X}$) para nueve puntos $(X, Y)$. Luego el porcentaje pedido es $9/12 = 75\%$.

Si los puntos están normalmente distribuidos respecto de la recta de regresión, la teoría predice que alrededor del 68% de los puntos están entre las dos rectas. Ello sería más preciso si el tamaño de la muestra fuese grande.

*Nota:* Una estimación mejor del error típico de estimación de la población de la que procede la muestra viene dada por $\hat{s}_{Y.X} = \sqrt{N/(N-2)}s_{Y.X} = \sqrt{12/10}(1.28) = 1.40$ in.

## VARIACION EXPLICADA Y VARIACION INEXPLICADA

**14.7.** Probar que $\sum(Y - \bar{Y})^2 = \sum(Y - Y_{est})^2 + \sum(Y_{est} - \bar{Y})^2$.

**Solución**

Elevando al cuadrado ambos mienbros de $Y - \bar{Y} = (Y - Y_{est}) + (Y_{est} - \bar{Y})$ y sumando, tenemos

$$\sum(Y - \bar{Y})^2 = \sum(Y - Y_{est})^2 + \sum(Y_{est} - \bar{Y})^2 + 2\sum(Y - Y_{est})(Y_{est} - \bar{Y})$$

El resultado buscado se sigue inmediatamente si conseguimos ver que la última suma es cero; en el caso de regresión lineal, eso es cierto, porque

$$\begin{aligned}\sum(Y - Y_{est})(Y_{est} - \bar{Y}) &= \sum(Y - a_0 - a_1X)(a_0 + a_1X - \bar{Y}) \\ &= a_0\sum(Y - a_0 - a_1X) + a_1\sum X(Y - a_0 - a_1X) - \bar{Y}\sum(Y - a_0 - a_1X) = 0\end{aligned}$$

a causa de las ecuaciones normales $\sum(Y - a_0 - a_1X) = 0$ y $\sum X(Y - a_0 - a_1X) = 0$.

Análogamente se ve que el resultado es válido para regresión no lineal usando una curva de mínimos cuadrados dada por $Y_{est} = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n$.

**14.8.** Calcular (*a*) la variación total, (*b*) la variación inexplicada y (*c*) la variación explicada para los datos del Problema 14.1.

**Solución**

(*a*) La variación total (Prob. 14.2) es $\sum(Y - \bar{Y})^2 = \sum y^2 = 38.92$.

(*b*) La variación inexplicada (Prob. 14.5) es $\sum(Y - Y_{est})^2 = Ns_{Y.X}^2 = 19.70$.

(*c*) La variación explicada (Prob. 14.7) es $\sum(Y_{est} - \bar{Y})^2 = 38.92 - 19.70 = 19.22$.

*Otro método*

Como $\bar{Y} = 811/12 = 67.58$, podemos construir la Tabla 14.6 usando los valores $Y_{est}$ de la Tabla 14.5; entonces $\sum(Y_{est} - \bar{Y})^2 = (-0.82)^2 + (-1.77)^2 + \cdots + (2.04)^2 = 19.21$. Los resultados de las partes (*a*) y (*b*) se pueden deducir también directamente.

**Tabla 14.6**

| $Y_{est} - \bar{Y} =$ $Y_{est} - 67.58$ | $-0.82$ | $-1.77$ | $0.13$ | $-1.30$ | $0.61$ | $-2.25$ | $1.56$ | $-0.34$ | $0.61$ | $0.13$ | $1.08$ | $2.04$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|

## COEFICIENTE DE CORRELACION

**14.9.** Hallar (*a*) el coeficiente de determinación y (*b*) el coeficiente de correlación para los datos del Problema 14.1. Usar los resultados del Problema 14.8.

**Solución**

(*a*) Coeficiente de determinación $= r^2 = \dfrac{\text{variación explicada}}{\text{variación total}} = \dfrac{19.22}{38.92} = 0.4938.$

(*b*) Coeficiente de correlación $= r = \pm\sqrt{0.4938} = \pm 0.7027.$

Como la variable $Y$ crece al crecer $X$, la correlación es positiva y por tanto escribimos $r = 0.7027$, o sea 0.70 con dos cifras significativas.

**14.10.** Probar que para regresión lineal el coeficiente de correlación entre las variables $X$ e $Y$ se puede escribir

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

donde $x = X - \bar{X}$ e $y = Y - \bar{Y}$.

**Solución**

La recta de regresión de mínimos cuadrados de $Y$ sobre $X$ es $Y_{est} = a_0 + a_1 X$ ó $y_{est} = a_1 x$, donde [véase Prob. 13.15(*a*)]

$$a_1 = \frac{\sum xy}{\sum x^2} \quad \text{e} \quad y_{est} = Y_{est} - \bar{Y}$$

Entonces

$$r^2 = \frac{\text{variación explicada}}{\text{variación total}} = \frac{\sum (Y_{est} - \bar{Y})^2}{\sum (Y - \bar{Y})^2} = \frac{\sum y_{est}^2}{\sum y^2}$$

$$= \frac{\sum a_1^2 x^2}{\sum y^2} = \frac{a_1^2 \sum x^2}{\sum y^2} = \left(\frac{\sum xy}{\sum x^2}\right)^2 \frac{\sum x^2}{\sum y^2} = \frac{(\sum xy)^2}{(\sum x^2)(\sum y^2)}$$

y

$$r = \pm \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

Sin embargo, como la cantidad

$$\frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

es positiva cuando $y_{est}$ crece al crecer $x$ (o sea, correlación lineal positiva) y negativa cuando $y$ decrece al crecer $x$ (o sea, correlación lineal negativa), *automáticamente* tiene el signo correcto. Por tanto, definimos el coeficiente de correlación lineal como

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

Esto se suele llamar la *fórmula momento-producto* para el coeficiente de correlación lineal.

## FORMULA MOMENTO-PRODUCTO PARA EL COEFICIENTE DE CORRELACION LINEAL

**14.11.** Hallar el coeficiente de correlación lineal entre las variables $X$ e $Y$ presentadas en la Tabla 14.7.

**Tabla 14.7**

| $X$ | 1 | 3 | 4 | 6 | 8 | 9 | 11 | 14 |
|---|---|---|---|---|---|---|---|---|
| $Y$ | 1 | 2 | 4 | 4 | 5 | 7 | 8 | 9 |

**Solución**

Los cálculos se resumen en la Tabla 14.8.

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{84}{\sqrt{(132)(56)}} = 0.977$$

De ahí observamos que hay una correlación lineal muy alta entre las variables, como ya se comprobó en los Problemas 13.8 y 13.12.

**Tabla 14.8**

| $X$ | $Y$ | $x = X - \bar{X}$ | $y = Y - \bar{Y}$ | $x^2$ | $xy$ | $y^2$ |
|---|---|---|---|---|---|---|
| 1 | 1 | −6 | −4 | 36 | 24 | 16 |
| 3 | 2 | −4 | −3 | 16 | 12 | 9 |
| 4 | 4 | −3 | −1 | 9 | 3 | 1 |
| 6 | 4 | −1 | −1 | 1 | 1 | 1 |
| 8 | 5 | 1 | 0 | 1 | 0 | 0 |
| 9 | 7 | 2 | 2 | 4 | 4 | 4 |
| 11 | 8 | 4 | 3 | 16 | 12 | 9 |
| 14 | 9 | 7 | 4 | 49 | 28 | 16 |
| $\sum X = 56$ | $\sum Y = 40$ | | | $\sum x^2 = 132$ | $\sum xy = 84$ | $\sum y^2 = 56$ |
| $\bar{X} = 56/8 = 7$ | $\bar{Y} = 40/8 = 5$ | | | | | |

**14.12.** Para los datos del Problema 14.11, hallar (*a*) la desviación típica de $X$, (*b*) la desviación típica de $Y$, (*c*) la varianza de $X$, (*d*) la varianza de $Y$ y (*e*) la covarianza de $X$ e $Y$.

**Solución**

(*a*) Desviación típica de $X = xs_X = \sqrt{\dfrac{\sum (X - \bar{X})^2}{N}} = \sqrt{\dfrac{\sum x^2}{N}} = \sqrt{\dfrac{132}{8}} = 4.06$

(*b*) Desviación típica de $Y = s_Y = \sqrt{\dfrac{\sum (Y - \bar{Y})^2}{N}} = \sqrt{\dfrac{\sum y^2}{N}} = \sqrt{\dfrac{56}{8}} = 2.65$

(*c*) Varianza de $X = s_X^2 = 16.50$

(*d*) Varianza de $Y = s_Y^2 = 7.00$

(*e*) Covarianza de $X$ e $Y = s_{XY} = \dfrac{\sum xy}{N} = \dfrac{84}{8} = 10.50.$

**14.13.** Para los datos del Problema 14.11, verificar la fórmula

$$r = \frac{s_{XY}}{s_X s_Y}$$

**Solución**

Del Problema 14.12

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{10.50}{(4.06)(2.65)} = 0.976$$

que, salvo por errores de redondeo, coincide con el resultado del Problema 14.11.

**14.14.** Obtener, mediante la fórmula momento-producto, el coeficiente de correlación lineal para los datos del Problema 14.1.

**Solución**

Se puede organizar el trabajo como en la Tabla 14.3 del Problema 14.2. Entonces

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{40.34}{\sqrt{(84.68)(38.92)}} = 0.7027$$

que está de acuerdo con el método más largo del Problema 14.9.

**14.15.** Demostrar que el coeficiente de correlación lineal viene dado por

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

**Solución**

Haciendo $x = X - \bar{X}$ e $y = Y - \bar{Y}$ en el resultado del Problema 14.10, tenemos

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{[\sum (X - \bar{X})^2][\sum (Y - \bar{Y}]}} \tag{34}$$

Pero $\sum (X - \bar{X})(Y - \bar{Y}) = \sum (XY - \bar{X}Y - X\bar{Y} + \bar{X}\bar{Y}) = \sum XY - \bar{X}\sum Y - \bar{Y}\sum X + N\bar{X}\bar{Y}$

$$= \sum XY - N\bar{X}\bar{Y} - N\bar{Y}\bar{X} + N\bar{X}\bar{Y} = \sum XY - N\bar{X}\bar{Y}$$

$$= \sum XY - \frac{(\sum X)(\sum Y)}{N}$$

ya que $\bar{X} = (\sum X)/N$ e $\bar{Y} = (\sum Y)/N$. Análogamente,

$$\sum (X - \bar{X})^2 = \sum (X^2 - 2X\bar{X} + \bar{X}^2) = \sum X^2 - 2\bar{X}\sum X + N\bar{X}^2$$

$$= \sum X^2 - \frac{2(\sum X)^2}{N} + \frac{(\sum X)^2}{N} = \sum X^2 - \frac{(\sum X)^2}{N}$$

y

$$\sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{N}$$

Así pues, la ecuación (34) se convierte en

$$r = \frac{\sum XY - (\sum X)(\sum Y)/N}{\sqrt{[\sum X^2 - (\sum X)^2/N][\sum Y^2 - (\sum Y)^2/N]}} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

**14.16.** Mediante la fórmula del Problema 14.15, hallar el coeficiente de correlación lineal para los datos del Problema 14.1.

**Solución**

Según la Tabla 14.2 del Problema 14.1 se tiene

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

$$= \frac{(12)(54{,}107) - (800)(811)}{\sqrt{[(12)(53{,}418) - (800)^2][(12)(54{,}849) - (811)^2]}} = 0.7027$$

como en los Problemas 14.9 y 14.14.

*Otro método*

El valor de $r$ es independiente de la elección del origen de $X$ e $Y$. Así pues, podemos usar los resultados del segundo método del Problema 14.2, con lo que se obtiene

$$r = \frac{N\sum X'Y' - (\sum X')(\sum Y')}{\sqrt{[N\sum X'^2 - (\sum X')^2][N\sum Y'^2 - (\sum Y')^2]}} = \frac{(12)(647) - (80)(91)}{\sqrt{[(12)(618) - (80)^2][(12)(729) - (91)^2]}} = 0.7027$$

## COEFICIENTE DE CORRELACION PARA DATOS AGRUPADOS

**14.17.** La Tabla 14.9 da las distribuciones de frecuencias de las notas finales de 100 estudiantes en Matemáticas y Física. Con referencia a esa tabla, determinar:

(*a*) El número de estudiantes que sacó notas entre 70-79 en Matemáticas y entre 80-89 en Física.
(*b*) El porcentaje de estudiantes con nota de Matemáticas menor que 70.
(*c*) El número de estudiantes que obtuvo 70 o más en Física y menos de 80 en Matemáticas.
(*d*) El porcentaje de estudiantes que aprobó al menos una de las dos materias, si se exigían 60 puntos para aprobar.

**Tabla 14.9**

| | | Calificación en Matemáticas | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 40-49 | 50-59 | 60-69 | 70-79 | 80-89 | 90-99 | Total |
| Calificación en Física | 90-99 | | | | 2 | 4 | 4 | 10 |
| | 80-89 | | | 1 | 4 | 6 | 5 | 16 |
| | 70-79 | | | 5 | 10 | 8 | 1 | 24 |
| | 60-69 | 1 | 4 | 9 | 5 | 2 | | 21 |
| | 50-59 | 3 | 6 | 6 | 2 | | | 17 |
| | 40-49 | 3 | 5 | 4 | | | | 12 |
| | Total | 7 | 15 | 25 | 23 | 20 | 10 | 100 |

**Solución**

(*a*) En la Tabla 14.9, miramos hacia abajo en la columna encabezada con 70-79 (nota de Matemáticas) a la fila con rótulo 80-89 (nota de Física), donde la entrada es 4, que es el número de estudiantes pedido.

(*b*) El número total de estudiantes con nota de Matemáticas inferior a 70 es la suma de los que tienen 40-49, 50-59 y 60-69 = 7 + 15 + 25 = 47. Luego el porcentaje pedido es 47/100 = 47%.

(*c*) El número pedido es el total de las entradas de la Tabla 14.10 (que representa parte de la Tabla 14.9). Por tanto, el número de estudiantes requerido es 1 + 5 + 2 + 4 + 10 = 22.

(*d*) La Tabla 14.11 (sacada de la Tabla 14.9), dice que el número de estudiantes con notas menores que 60 en ambas asignaturas es 3 + 3 + 6 + 5 = 17. Luego el número de los que tienen al menos una nota de 60 o más es 100 − 17 = 83, y el porcentaje requerido es 83/100 = 83%.

**Tabla 14.10**

| | | Calificación en Matemáticas | |
|---|---|---|---|
| | | 60-69 | 70-79 |
| Calificación en Física | 90-99 | | 2 |
| | 80-89 | 1 | 4 |
| | 70-79 | 5 | 10 |

**Tabla 14.11**

| | | Calificación en Matemáticas | |
|---|---|---|---|
| | | 40-49 | 50-59 |
| Calificación en Física | 50-59 | 3 | 6 |
| | 40-49 | 3 | 5 |

La Tabla 14.9 se llama a veces una *tabla de frecuencias de dos variables*. Cada cuadrado de ese tabla se llama una celda y corresponde a un par de clases o intervalos de confianza. El número indicado en la celda se llama *frecuencia de celda*. Así, en la parte (*a*) el número 4 es la frecuencia de la celda correspondiente al par de intervalos de confianza 70-79 en Matemáticas y 80-89 en Física.

Los totales indicados en la última fila y en la última columna se llaman *totales marginales* o *frecuencias marginales*. Corresponden, respectivamente, a las frecuencias de clase de las distribuciones de frecuencias separadas de las notas de Matemáticas y Física.

**14.18.** Mostrar cómo modificar la fórmula del Problema 14.15 para el caso de datos agrupados como en la tabla de frecuencias de dos variables (Tabla 14.9).

**Solución**

Para datos agrupados, podemos considerar los diversos valores de las variables $X$ e $Y$ como coincidentes con las marcas de clase, mientras $f_X$ y $f_Y$ son las correspondientes frecuencias de clase, o frecuencias marginales, que se recogen en la última fila y columna de la tabla de frecuencias de dos variables. Si denotamos por $f$ las diversas frecuencias de celda asociadas a los pares de marcas de clase $(X, Y)$, podemos sustituir la fórmula del Problema 14.15 por

$$r = \frac{N\sum fXY - (\sum f_X X)(\sum f_Y Y)}{\sqrt{[N\sum f_X X^2 - (\sum f_X X)^2][N\sum f_Y Y^2 - (\sum f_Y Y)^2]}} \tag{35}$$

Si hacemos $X = A + c_X u_X$ e $Y = B + c_Y u_Y$, donde $c_X$ y $c_Y$ son las anchuras de intervalos de clase

(supuestas constantes) y $A$ y $B$ son marcas de clase arbitrarias correspondientes a las variables, la fórmula (35) se convierte en la (21):

$$r = \frac{N\sum f u_X u_Y - (\sum f_X u_X)(\sum f_Y u_Y)}{\sqrt{[N\sum f_X u_X^2 - (\sum f_X u_X)^2][N\sum f_Y u_Y^2 - (\sum f_Y u_Y)^2]}} \tag{21}$$

Este es el *método de compilación* empleado en capítulos precedentes como método abreviado para calcular medias, desviaciones típicas y momentos superiores.

**Tabla 14.12**

| Calificación en Física | | Calificación en Matemáticas $X$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $X$ | 44.5 | 54.5 | 64.5 | 74.5 | 84.5 | 94.5 | | | | |
| | $Y$ \ $u_X$ / $u_Y$ | −2 | −1 | 0 | 1 | 2 | 3 | $f_Y$ | $f_Y u_Y$ | $f_Y u_Y^2$ | Suma de los números de las esquinas en cada fila |
| | 94.5 2 | | | | 2 [4] | 4 [16] | 4 [24] | 10 | 20 | 40 | 44 |
| | 84.5 1 | | | 1 [0] | 4 [4] | 6 [12] | 5 [15] | 16 | 16 | 16 | 31 |
| | 74.5 0 | | | 5 [0] | 10 [0] | 8 [0] | 1 [0] | 24 | 0 | 0 | 0 |
| | 64.5 −1 | 1 [2] | 4 [4] | 9 [0] | 5 [−5] | 2 [−4] | | 21 | −21 | 21 | −3 |
| | 54.5 −2 | 3 [12] | 6 [12] | 6 [0] | 2 [−4] | | | 17 | −34 | 68 | 20 |
| | 44.5 −3 | 3 [18] | 5 [15] | 4 [0] | | | | 12 | −36 | 108 | 33 |
| | $f_X$ | 7 | 15 | 25 | 23 | 20 | 10 | $\sum f_X = \sum f_Y = N = 100$ | $\sum f_Y u_Y = -55$ | $\sum f_Y u_Y^2 = 253$ | $\sum f u_X u_Y = 125$ |
| | $f_X u_X$ | −14 | −15 | 0 | 23 | 40 | 30 | $\sum f_X u_X = 64$ | | | |
| | $f_X u_X^2$ | 28 | 15 | 0 | 23 | 80 | 90 | $\sum f_X u_X^2 = 236$ | | | |
| | Suma de los números de las esquinas en cada columna | 32 | 31 | 0 | −1 | 24 | 39 | $\sum f u_X u_Y = 125$ | | | |

Comprobación

**14.19.** Hallar el coeficiente de correlación lineal de las notas del Problema 14.17.

**Solución**

Usamos la fórmula (21). El proceso se resume en la Tabla 14.12, que se llama una tabla de correlación. Las sumas $\sum f_X$, $\sum f_X u_X$, $\sum f_X u_X^2$, $\sum f_Y$, $\sum f_Y u_Y$ y $\sum f_Y u_Y^2$ se obtienen mediante el método de compilación, como en capítulos anteriores.

El número en la esquina de cada celda en la Tabla 14.12 representa el producto $fu_X u_Y$, donde $f$ es la frecuencia de celda. Su suma en cada fila se indica en la fila correspondiente de la última columna. Y su suma en cada columna se indica en la correspondiente columna de la última fila. Los totales finales de la última fila y columna son iguales y representan

$$r = \frac{N\sum fu_X u_Y - (\sum f_X u_X)(\sum f_Y u_Y)}{\sqrt{[N\sum f_X u_X^2 - (\sum f_X u_X)^2][N\sum f_Y u_Y^2 - (\sum f_Y u_Y)^2]}}$$

$$= \frac{(100)(125) - (64)(-55)}{\sqrt{[(100)(236) - (64)^2][(100)(253) - (-55)^2]}} = \frac{16{,}020}{\sqrt{(19{,}504)(22{,}275)}} = 0.7686$$

**14.20.** Usar la Tabla 14.12 para calcular (*a*) $s_X$, (*b*) $s_Y$ y (*c*) $s_{XY}$ y así verificar la fórmula $r = s_{XY}/(s_X s_Y)$.

**Solución**

(*a*) $$s_X = c_X\sqrt{\frac{\sum f_X u_X^2}{N} - \left(\frac{\sum f_X u_X}{N}\right)^2} = 10\sqrt{\frac{236}{100} - \left(\frac{64}{100}\right)^2} = 13.966$$

(*b*) $$s_Y = c_Y\sqrt{\frac{\sum f_Y u_Y^2}{N} - \left(\frac{\sum f_Y u_Y}{N}\right)^2} = 10\sqrt{\frac{253}{100} - \left(\frac{-55}{100}\right)^2} = 14.925$$

(*c*) $$s_{XY} = c_X c_Y\left[\frac{\sum fu_X u_Y}{N} - \left(\frac{\sum f_X u_X}{N}\right)\left(\frac{\sum f_Y u_Y}{N}\right)\right] = (10)(10)\left[\frac{125}{100} - \left(\frac{64}{100}\right)\left(\frac{-55}{100}\right)\right] = 160.20$$

Luego las desviaciones típicas de las notas de Matemáticas y Física son 14.0 y 14.9 respectivamente, mientras que su covarianza es 160.2. El coeficiente de correlación $r$ es, por tanto,

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{160.20}{(13.966)(14.925)} = 0.7686$$

en coincidencia con el Problema 14.19.

## RECTAS DE REGRESION Y EL COEFICIENTE DE CORRELACION

**14.21.** Probar que las rectas de regresión de $Y$ sobre $X$ y de $X$ sobre $Y$ tienen ecuaciones respectivas (*a*) $Y - \bar{Y} = (rs_Y/s_X)(X - \bar{X})$ y (*b*) $X - \bar{X} = (rs_X/s_Y)(Y - \bar{Y})$.

**Solución**

(*a*) Del Problema 13.15(*a*) sabemos que la recta de regresión de $Y$ sobre $X$ es

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \qquad \text{o sea} \qquad Y - \bar{Y} = \left(\frac{\sum xy}{\sum x^2}\right)(X - \bar{X})$$

Entonces, como $r = \dfrac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$ (véase Prob. 14.10)

tenemos $$\frac{\sum xy}{\sum x^2} = \frac{r\sqrt{(\sum x^2)(\sum y^2)}}{\sum x^2} = \frac{r\sqrt{\sum y^2}}{\sqrt{\sum x^2}} = \frac{rs_Y}{s_X}$$

y el resultado es el deseado.

(*b*) Esto se deduce intercambiando $X$ e $Y$ en la parte (*a*).

**14.22.** Si las rectas de regresión de $Y$ sobre $X$ y de $X$ sobre $Y$ son, respectivamente, $Y = a_0 + a_1 X$ y $X = b_0 + b_1 Y$, probar que $a_1 b_1 = r^2$.

**Solución**

Del Problema 14.21, partes (*a*) y (*b*),

$$a_1 = \frac{rs_Y}{s_X} \quad \text{y} \quad b_1 = \frac{rs_X}{s_Y}$$

Luego $$a_1 b_1 = \left(\frac{rs_Y}{s_X}\right)\left(\frac{rs_X}{s_Y}\right) = r^2$$

Cabe tomar este resultando como punto de partida para la definición del coeficiente de correlación lineal.

**14.23.** Usar el resultado del Problema 14.22 para hallar el coeficiente de correlación lineal para los datos del Problema 14.1.

**Solución**

Del Problema 14.1 [partes (*b*) y (*c*), respectivamente] $a_1 = 484/1016 = 0.476$ y $b_1 = 484/467 = 1.036$. Así que $r^2 = a_1 b_1 = (484/1016)(484/467)$ y $r = 0.7027$, de acuerdo con los Problemas 14.9, 14.14 y 14.16.

**14.24.** Para los datos del Problema 14.19, escribir las ecuaciones de las rectas de regresión de (*a*) $Y$ sobre $X$ y (*b*) $X$ sobre $Y$.

**Solución**

De la tabla de correlación (Tabla 14.12) del Problema 14.19, tenemos

$$\bar{X} = A + c_X \frac{\sum f_X u_X}{N} = 64.5 + \frac{(10)(64)}{100} = 70.9$$

$$\bar{Y} = B + c_Y \frac{\sum f_Y u_Y}{N} = 74.5 + \frac{(10)(-55)}{100} = 69.0$$

Por el Problema 14.20, $s_X = 13.966$, $s_Y = 14.925$ y $r = 0.7686$. Ahora usamos el Problema 14.21, partes (*a*) y (*b*), para obtener las ecuaciones de las rectas de regresión.

(*a*) $Y - \bar{Y} = \dfrac{rs_Y}{s_X}(X - \bar{X})$ $\quad Y - 69.0 = \dfrac{(0.7686)(14.925)}{13.966}(X - 70.9) = 0.821(X - 70.9)$

(*b*) $X - \bar{X} = \dfrac{rs_X}{s_Y}(Y - \bar{Y})$ $\quad X - 70.9 = \dfrac{(0.7686)(13.966)}{14.925}(Y - 69.0) = 0.719(Y - 69.0)$

**14.25.** Calcular, para los datos del Problema 14.19, los errores típicos de estimación (*a*) $s_{Y.X}$ y (*b*) $s_{X.Y}$. Usar los resultados del Problema 14.20.

**Solución**

(*a*) $s_{Y.X} = s_Y\sqrt{1 - r^2} = 14.925\sqrt{1 - (0.7686)^2} = 9.548$

(*b*) $s_{X.Y} = s_X\sqrt{1 - r^2} = 13.966\sqrt{1 - (0.7686)^2} = 8.934$

**14.26.** La Tabla 14.13 muestra los índices de precios al consumo de alimentación y de asistencia sanitaria durante los años 1975-1983 comparados con los precios en un año base, 1967 (tomados como 100). Calcular el coeficiente de correlación entre esos dos índices.

**Tabla 14.13**

| Año | 1975 | 1976 | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 |
|---|---|---|---|---|---|---|---|---|---|
| Alimentación | 175 | 181 | 192 | 211 | 235 | 255 | 275 | 286 | 292 |
| Asistencia sanitaria | 169 | 185 | 202 | 219 | 240 | 266 | 295 | 329 | 357 |

*Fuente:* Survey of Current Business.

**Solución**

(*a*) Denotando por $X$ e $Y$ los índices de alimentación y de asistencia sanitaria, respectivamente, el cálculo del coeficiente de correlación procede como sugiere la Tabla 14.14. (Nótese que el año se emplea sólo para especificar los valores correspondientes de $X$ e $Y$). Entonces, por la fórmula momento-producto,

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{23.442}{\sqrt{(16,774)(34,107)}} = 0.98$$

Luego existe una correlación lineal muy buena entre ambos índices de costo. Hay que hacer constar, no obstante, que eso no quiere decir que los costes hayan aumentado *lo mismo* a lo largo de los años: así, por ejemplo, de 1975 a 1983 los alimentos han subido un 67% mientras que la asistencia sanitaria lo ha hecho en un 111%.

**Tabla 14.14**

| $X$ | $Y$ | $x = X - \bar{X}$ | $y = Y - \bar{Y}$ | $x^2$ | $xy$ | $y^2$ |
|---|---|---|---|---|---|---|
| 175 | 169 | −59 | −82 | 3,481 | 4,838 | 6,724 |
| 181 | 185 | −53 | −66 | 2,809 | 3,498 | 4,356 |
| 192 | 202 | −42 | −49 | 1,764 | 2,058 | 2,401 |
| 211 | 219 | −23 | −32 | 529 | 736 | 1,024 |
| 235 | 240 | 1 | −11 | 1 | −11 | 121 |
| 255 | 266 | 21 | 15 | 441 | 315 | 225 |
| 275 | 295 | 41 | 44 | 1,681 | 1,804 | 1,936 |
| 286 | 329 | 52 | 78 | 2,704 | 4,056 | 6,084 |
| 292 | 357 | 58 | 106 | 3,364 | 6,148 | 11,236 |
| $\sum X = 2,102$<br>$\bar{X} = 234$ | $\sum Y = 2,262$<br>$\bar{Y} = 251$ | | | $\sum x^2 = 16,774$ | $\sum xy = 23,442$ | $\sum y^2 = 34,107$ |

## CORRELACION NO LINEAL

**14.27.** Ajustar una parábola de mínimos cuadrados de la forma $Y = a_0 + a_1X + a_2X^2$ al conjunto de datos de la Tabla 14.15.

**Tabla 14.15**

| $X$ | 1.2 | 1.8 | 3.1 | 4.9 | 5.7 | 7.1 | 8.6 | 9.8 |
|---|---|---|---|---|---|---|---|---|
| $Y$ | 4.5 | 5.9 | 7.0 | 7.8 | 7.2 | 6.8 | 4.5 | 2.7 |

**Solución**

Las ecuaciones normales (23) del Capítulo 13 son

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X + a_2 \sum X^2 \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum X^3 \\ \sum X^2 Y &= a_0 \sum X^2 + a_1 \sum X^3 + a_2 \sum X^4 \end{aligned} \tag{36}$$

El proceso de cálculo de las sumas se presenta en la Tabla 14.16. Como $N = 8$, las ecuaciones normales (36) pasan a ser

$$\begin{aligned} 8a_0 + 42.2a_1 + 291.20a_2 &= 46.4 \\ 42.2a_0 + 291.20a_1 + 2275.35a_2 &= 230.42 \\ 291.20a_0 + 2275.35a_1 + 18971.92a_2 &= 1449.00 \end{aligned} \tag{37}$$

Resolviendo, $a_0 = 2.588$, $a_1 = 2.065$, y $a_2 = -0.2110$; por tanto, la parábola de mínimos cuadrados buscada es

$$Y = 2.588 + 2.065X - 0.2110X^2$$

**14.28.** Estimar, mediante la parábola de mínimos cuadrados del Problema 14.27, los valores de $Y$ a partir de los valores de $X$ dados.

**Solución**

Para $X = 1.2$, $Y_{est} = 2.588 + 2.065(1.2) - 0.2110(1.2)^2 = 4.762$. Otros valores estimados se obtienen análogamente. Los resultados, junto con los valores reales de $Y$, se muestran en la Tabla 14.17.

**Tabla 14.16**

| $X$ | $Y$ | $X^2$ | $X^3$ | $X^4$ | $XY$ | $X^2Y$ |
|---|---|---|---|---|---|---|
| 1.2 | 4.5 | 1.44 | 1.73 | 2.08 | 5.40 | 6.48 |
| 1.8 | 5.9 | 3.24 | 5.83 | 10.49 | 10.62 | 19.12 |
| 3.1 | 7.0 | 9.61 | 29.79 | 92.35 | 21.70 | 67.27 |
| 4.9 | 7.8 | 24.01 | 117.65 | 576.48 | 38.22 | 187.28 |
| 5.7 | 7.2 | 32.49 | 185.19 | 1055.58 | 41.04 | 233.93 |
| 7.1 | 6.8 | 50.41 | 357.91 | 2541.16 | 48.28 | 342.79 |
| 8.6 | 4.5 | 73.96 | 636.06 | 5470.12 | 38.70 | 332.82 |
| 9.8 | 2.7 | 96.04 | 941.19 | 9223.66 | 26.46 | 259.31 |
| $\sum X$ = 42.2 | $\sum Y$ = 46.4 | $\sum X^2$ = 291.20 | $\sum X^3$ = 2275.35 | $\sum X^4$ = 18,971.92 | $\sum XY$ = 230.42 | $\sum X^2Y$ = 1449.00 |

**Tabla 14.17**

| $Y_{est}$ | 4.762 | 5.621 | 6.962 | 7.640 | 7.503 | 6.613 | 4.741 | 2.561 |
|---|---|---|---|---|---|---|---|---|
| $Y$ | 4.5 | 5.9 | 7.0 | 7.8 | 7.2 | 6.8 | 4.5 | 2.7 |

**14.29.** (*a*) Hallar el coeficiente de correlación lineal entre las variables $X$ e $Y$ del Problema 14.27.
(*b*) Hallar el coeficiente de correlación no lineal entre estas variables, suponiendo la relación parabólica obtenida en el Problema 14.27.
(*c*) Explicar la diferencia entre los coeficientes de correlación obtenidos en las partes (*a*) y (*b*).
(*d*) ¿Qué porcentaje de la variación total queda inexplicada al suponer una relación parabólica entre $X$ e $Y$?

**Solución**

(*a*) Haciendo uso de los cálculos ya realizados en la Tabla 14.16 y el hecho añadido de que $\sum Y^2 = 290.52$, vemos que

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}} = \frac{(8)(230.42) - (42.2)(46.4)}{\sqrt{[(8)(291.20) - (42.2)^2][(8)(290.52) - (46.4)^2]}} = -0.3743$$

(*b*) De la Tabla 14.16, $\bar{Y} = (\sum Y)/N = 46.4/8 = 5.80$; luego la variación total es $\sum (Y - \bar{Y})^2 = 21.40$. De la Tabla 14.17 vemos que la variación explicada es $\sum (Y_{est} - \bar{Y})^2 = 21.02$. Luego

$$r^2 = \frac{\text{variación explicada}}{\text{variación total}} = \frac{21.02}{21.40} = 0.9822 \quad \text{y} \quad r = 0.9911 \quad \text{o sea} \quad 0.99$$

(*c*) El que (*a*) haya dado un coeficiente de correlación lineal de sólo $-0.3743$ indica que *no hay* prácticamente *relación lineal* entre $X$ e $Y$. Sin embargo, hay una *relación no lineal* muy fuerte dada por la parábola del Problema 14.27, como ratifica el hecho de que el coeficiente de correlación en (*b*) es 0.9.

(*d*) $$\frac{\text{Variación inexplicada}}{\text{Variación total}} = 1 - r^2 = 1 - 0.9822 = 0.0178$$

Luego el 1.78% de la variación total queda inexplicada. Ello podría ser debido a fluctuaciones aleatorias o a una variable adicional que no se ha tenido en cuenta.

**14.30.** Hallar (*a*) $s_Y$ y (*b*) $s_{Y.X}$ para los datos del Problema 14.27.

**Solución**

(*a*) Del Problema 14.29(*a*), $\sum (Y - \bar{Y})^2 = 21.40$. Así pues, la desviación típica de $Y$ es

$$s_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}} = \sqrt{\frac{21.40}{8}} = 1.636 \quad \text{o sea} \quad 1.64$$

(*b*) *Primer método*

Usando la parte (*a*) y el Problema 14.29(*b*), el error típico de estimación de $Y$ sobre $X$ es

$$s_{Y.X} = s_Y\sqrt{1 - r^2} = 1.636\sqrt{1 - (0.9911)^2} = 0.218 \quad \text{o sea} \quad 0.22$$

*Segundo método*

Usando el Problema 14.29,

$$s_{Y.X} = \sqrt{\frac{\sum(Y - Y_{est})^2}{N}} = \sqrt{\frac{\text{variación inexplicada}}{N}} = \sqrt{\frac{21.40 - 21.02}{8}} = 0.218 \quad \text{o sea} \quad 0.22$$

*Tercer método*

Por el Problema 14.27 y el cálculo adicional $\sum Y^2 = 290.52$, tenemos

$$s_{Y.X} = \sqrt{\frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY - a_2 \sum X^2 Y}{N}} = 0.218 \quad \text{o sea} \quad 0.22$$

## TEORIA MUESTRAL DE LA CORRELACION

**14.31.** Al calcular el coeficiente de correlación de una muestra de tamaño 18, ha dado el valor 0.32. ¿Podemos concluir al nivel de significación (*a*) 0.05 y (*b*) 0.01 que el coeficiente de correlación de la población correspondiente difiere de cero?

**Solución**

Queremos decidir entre las hipótesis $H_0$: $\rho = 0$ y $H_1$: $\rho > 0$.

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{0.32\sqrt{18-2}}{\sqrt{1-(0.32)^2}} = 1.35$$

(*a*) Usando un contraste unilateral con la distribución de Student en el nivel 0.05, rechazaríamos la hipótesis $H_0$ si $t > t_{.95} = 1.75$ para $(18 - 2) = 16$ grados de libertad. Luego no podemos rechazar $H$ al nivel 0.05.

(*b*) Puesto que no podemos rechazar $H$ al nivel 0.05, ciertamente, tampoco al 0.01.

**14.32.** ¿Cuál es el mínimo tamaño de muestra necesario para poder concluir que un coeficiente de correlación de 0.32 difiere significativamente de cero al nivel 0.05?

**Solución**

Con un contraste de una cola de la distribución de Student en el nivel 0.05, el mínimo valor de $N$ debe ser tal que

$$\frac{0.32\sqrt{N-2}}{\sqrt{1-(0.32)^2}} = t_{.95}$$

para $N - 2$ grados de libertad. Para un número infinito de grados de libertad, $t_{95} = 1.64$ y por tanto, $N = 25.6$.

Para $N = 26$: $\nu = 24$ $t_{.95} = 1.71$ $t = 0.32\sqrt{24}/\sqrt{1-(0.32)^2} = 1.65$

Para $N = 27$: $\nu = 25$ $t_{.95} = 1.71$ $t = 0.32\sqrt{25}/\sqrt{1-(0.32)^2} = 1.69$

Para $N = 28$: $\nu = 26$ $t_{.96} = 1.71$ $t = 0.32\sqrt{26}/\sqrt{1-(0.32)^2} = 1.72$

Así que el tamaño mínimo de la muestra es $N = 28$.

**14.33.** Un coeficiente de correlación de una muestra de tamaño 24 resulta ser $r = 0.75$. Al nivel de significación 0.05, ¿podemos rechazar la hipótesis de que el coeficiente de correlación de la población es tan pequeño como (*a*) $\rho = 0.60$ y (*b*) $\rho = 0.50$?

**Solución**

(*a*) $$Z = 1.1513 \log\left(\frac{1 + 0.75}{1 - 0.75}\right) = 0.9730 \qquad \mu_Z = 1.1513 \log\left(\frac{1 + 0.60}{1 - 0.60}\right) = 0.6932$$

y $$\sigma_Z = \frac{1}{\sqrt{N - 3}} = \frac{1}{\sqrt{21}} = 0.2182$$

Por tanto $$z = \frac{Z - \mu_Z}{\sigma_Z} = \frac{0.9730 - 0.6932}{0.2182} = 1.28$$

Usando un contraste de una cola con la distribución normal al nivel 0.05, rechazaríamos la hipótesis sólo si $z$ fuera mayor que 1.64. Luego no podemos rechazar la hipótesis de que el coeficiente de correlación de la población es tan pequeño como 0.60.

(*b*) Si $\rho = 0.50$, entonces $\mu_Z = 1.1513 \log 3 = 0.5493$ y $z = (0.9730 - 0.5493)/0.2182 = 1.94$. Luego podemos rechazar la hipótesis de que el coeficiente de correlación de la población sea tan pequeño como $\rho = 0.50$, al nivel 0.05.

**14.34.** El coeficiente de correlación entre las notas en Física y Matemáticas para un grupo de 21 estudiantes resulta ser 0.80. Hallar los límites de confianza 95% para este coeficiente.

**Solución**

Como $r = 0.80$ y $N = 21$, los límites de confianza 95% para $\mu_Z$ vienen dados por

$$Z \pm 1.96\sigma_Z = 1.1513 \log\left(\frac{1 + r}{1 - r}\right) \pm 1.96\left(\frac{1}{\sqrt{N - 3}}\right) = 1.0986 \pm 0.4620$$

Así pues, $\mu_Z$ tiene el intervalo de confianza 95% desde 0.5366 a 1.5606. Ahora bien, si

$$\mu_Z = 1.1513 \log\left(\frac{1 + \rho}{1 - \rho}\right) = 0.5366 \qquad \text{entonces} \qquad \rho = 0.4904$$

y si $$\mu_Z = 1.1513 \log\left(\frac{1 + \rho}{1 - \rho}\right) = 1.5606 \qquad \text{entonces} \qquad \rho = 0.9155$$

Luego los límites de confianza 95% para $\rho$ son 0.49 y 0.92.

**14.35.** Dos coeficientes de correlación obtenidos de muestras de tamaños $N_1 = 28$ y $N_3 = 35$ han resultado ser $r_1 = 0.50$ y $r_2 = 0.30$, respectivamente. ¿Hay diferencia significativa entre los dos coeficientes al nivel 0.05?

**Solución**

$$Z_1 = 1.1513 \log\left(\frac{1 + r_1}{1 - r_1}\right) = 0.5493 \qquad Z_2 = 1.1513 \log\left(\frac{1 + r_2}{1 - r_2}\right) = 0.3095$$

y $$\sigma_{Z1\ Z2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} = 0.2669$$

Queremos decidir entre dos hipótesis $H_0$: $\mu_{Z1} = \mu_{Z2}$ y $H_1$: $\mu_{Z1} \neq \mu_{Z2}$. Bajo la hipótesis $H_0$,

$$z = \frac{Z_1 - Z_2 - (\mu_{Z1} - \mu_{Z2})}{\sigma_{Z1-Z2}} = \frac{0.5493 - 0.3095 - 0}{0.2669} = 0.8985$$

Con un contraste bilateral mediante la distribución normal, rechazaríamos $H$ sólo si $z > 1.96$ o si $z < -1.96$. Por tanto, no podemos rechazar $H$, y concluimos que los resultados no son significativamente diferentes al nivel 0.05.

## TEORIA MUESTRAL DE LA REGRESION

**14.36.** En el Problema 14.1 hallamos como ecuación de regresión de $Y$ sobre $X$ la que sigue: $Y = 35.82 + 0.476X$. Contrastar la hipótesis, al nivel de significación 0.05, de que el coeficiente de correlación de la ecuación de regresión de la población es 0.180.

**Solución**

$$t = \frac{a_1 - A_1}{s_{Y.X}/s_X}\sqrt{N-2} = \frac{0.476 - 0.180}{1.28/2.66}\sqrt{12-2} = 1.95$$

como $s_{Y.X} = 1.28$ (calculado en el Problema 14.5) y $s_X = \sqrt{(\sum x^2)/N} = \sqrt{84.68/12} = 2.66$ (del Problema 14.2). Usando un contraste de una cola con la distribución de Student al nivel 0.05, rechazaríamos la hipótesis de que el coeficiente de regresión es tan bajo como 0.180 si $t > t_{.95} = 1.81$ para $(12 - 2) = 10$ grados de libertad. Luego no podemos rechazar la hipótesis.

**14.37.** Hallar los límites de confianza 95% para el coeficiente de regresión del Problema 14.36.

**Solución**

$$A_1 = a_1 - \frac{t}{\sqrt{N-2}}\left(\frac{s_{Y.X}}{s_X}\right)$$

Luego los límites de confianza para $A$ (obtenidos haciendo $t = \pm t_{.975} = \pm 2.23$ para $12 - 2 = 10$ grados de libertad) vienen dados por

$$a_1 \pm \frac{2.23}{\sqrt{12-2}}\left(\frac{s_{Y.X}}{s_X}\right) = 0.476 \pm \frac{2.23}{\sqrt{10}}\left(\frac{1.28}{2.66}\right) = 0.476 \pm 0.340$$

Es decir, tenemos 95% de confianza de que $A$ está entre 0.136 y 0.816.

**14.38.** En el Problema 14.1, hallar los límites de confianza 9% para las alturas de los hijos cuyos padres miden (*a*) 65.0 y (*b*) 70.0 in.

**Solución**

Como $t_{.975} = 2.23$ para $(12 - 2) = 10$ grados de libertad, los límites de confianza 95% para $Y_p$ (véase pág. 330) vienen dados por

$$Y_0 \pm \frac{2.23}{\sqrt{N-2}} s_{Y.X}\sqrt{N + 1 + \frac{(X_0 - \bar{X})^2}{s_X^2}}$$

donde $Y_0 = 35.82 + 0.476X_0$ (Problema 14.1), $s_{Y.X} = 1.28$, $s_X = 2.66$ (Problema 14.36) y $N = 12$.

(*a*) Si $X_0 = 65.0$, entonces $Y_0 = 66.76$ in. Además $(X_0 - \bar{X})^2 = (65.0 - 800/12)^2 = 2.78$. Así pues los límites de confianza al 95% son

$$66.76 \pm \frac{2.23}{\sqrt{10}}(1.28)\sqrt{12 + 1 + \frac{2.78}{(2.66)^2}} = 66.76 \pm 3.31 \text{ in}$$

Esto es, podemos tener un 95% de confianza de que las alturas de los hijos están entre 63.4 y 70.1.

(*b*) Si $X_0 = 70.0$, entonces $Y_0 = 69.14$ in. Además, $(X_0 - \bar{X})^2 = (70.0 - 800/12)^2 = 11.11$. Luego los límites de confianza 95% resultan ser $69.14 \pm 3.45$ in; es decir, con un 95% de confianza las alturas de los hijos están entre 65.7 y 72.6 in.

Nótese que para los valores grandes de $N$, los límites de confianza 95% vienen dados aproximadamente por $Y_0 \pm 1.96s_{Y.X}$ o sea $Y_0 \pm 2s_{Y.X}$, supuesto que $(X_0 - \bar{X})$ no sea demasiado grande. Eso coincide con los resultados aproximados mencionados en la página 210. Los métodos de este problema son válidos con independencia del valor de $N$ o de $(X_0 - \bar{X})$; esto es, los métodos de muestreo son exactos.

**14.39.** En el Problema 14.1 hallar los límites de confianza 95% para las alturas medias de los hijos cuyos padres miden (*a*) 65.0 in y (*b*) 70.0 in.

**Solución**

Ya que $t_{.975} = 2.23$ para 10 grados de libertad, los límites de confianza 95% para $\bar{Y}_p$ (véase página 330) vienen dados por

$$Y_0 \pm \frac{2.23}{\sqrt{10}} s_{Y.X}\sqrt{1 + \frac{(X_0 - \bar{X})^2}{s_X^2}}$$

donde $Y_0 = 35.82 + 0.476X_0$ (Problema 14.1), $s_{Y.X} = 1.28$ y $s_X = 2.66$ (Problema 14.36).

(*a*) Si $X_0 = 65.0$, vemos que los límites de confianza 95% son $66.76 \pm 1.07$ in [comparar con el Problema 14.38(*a*)]. Es decir, podemos tener 95% de confianza de que la altura media de todos los hijos cuyos padres miden 65.0 in está entre 65.7 y 67.8 in.

(*b*) Si $X_0 = 70.0$, vemos que los límites de confianza 95% son $69.14 \pm 1.45$ in [comparar con el Problema 14.38(*b*)]. Es decir, podemos tener 95% de confianza de que la *altura media* de todos los hijos cuyos padres miden 70.0 in estará entre 67.7 y 70.6 in.

## PROBLEMAS SUPLEMENTARIOS

### REGRESION LINEAL Y CORRELACION LINEAL

**14.40.** La Tabla 14.18 presenta las notas de dos exámenes de Biología, $X$ e $Y$, de 10 estudiantes.

(*a*) Construir un diagrama de dispersión.

(*b*) Hallar la recta de regresión de mínimos cuadrados de $Y$ sobre $X$.

(*c*) Hallar la recta de regresión de mínimos cuadrados de $X$ sobre $Y$.

(*d*) Representar las dos rectas de las partes (*b*) y (*c*) en el diagrama de dispersión de la parte (*a*).

**Tabla 14.18**

| Calificaciones en el primer examen ($X$) | Calificaciones en el segundo examen ($Y$) |
|---|---|
| 6 | 8 |
| 5 | 7 |
| 8 | 7 |
| 8 | 10 |
| 7 | 5 |
| 6 | 8 |
| 10 | 0 |
| 4 | 6 |
| 9 | 8 |
| 7 | 6 |

**14.41.** Hallar (*a*) $s_{Y.X}$ y (*b*) $s_{X.Y}$ para los datos de la Tabla 14.18.

**14.42.** Calcular (*a*) la variación total en $Y$, (*b*) la variación inexplicada en $Y$ y (*c*) la variación explicada en $Y$, para los datos del Problema 14.40.

**14.43.** Usar los resultados del Problema 14.42 para hallar el coeficiente de correlación entre los dos conjuntos de notas del Problema 14.40.

**14.44.** (*a*) Hallar el coeficiente de correlación entre los dos conjuntos de notas del Problema 14.40 usando la fórmula momento-producto, y comparar con el resultado del Problema 14.45.

(*b*) Obtener el coeficiente de correlación directamente a partir de las pendientes de las rectas de regresión del Problema 14.42, partes (*b*) y (*c*).

**14.45.** Hallar la covarianza para los datos del Problema 14.40 (*a*) directamente y (*b*) usando la fórmula $s_{XY} = rs_X s_Y$ y el resultado del Problema 14.43 ó 14.44.

**14.46.** La Tabla 14.19 da las edades $X$ y las presiones sanguíneas (en sístole) $Y$ de 12 mujeres.

(*a*) Hallar el coeficiente de correlación entre $X$ e $Y$.

(*b*) Determinar la ecuación de regresión de mínimos cuadrados de $Y$ sobre $X$

(*c*) Estimar la presión sanguínea de una mujer de 45 años.

**Tabla 14.19**

| Edad ($X$) | Presión sanguínea |
|---|---|
| 56 | 147 |
| 42 | 125 |
| 72 | 160 |
| 36 | 118 |
| 63 | 149 |
| 47 | 128 |
| 55 | 150 |
| 49 | 145 |
| 38 | 115 |
| 42 | 140 |
| 68 | 152 |
| 60 | 155 |

**14.47.** Hallar el coeficiente de correlación para los datos del (*a*) Problema 13.32 y (*b*) Problema 13.35.

**14.48.** El coeficiente de correlación entre las variables $X$ e $Y$ es $r = 0.60$. Si $s_X = 1.50$, $s_Y = 2.00$, $\bar{X} = 10$ e $\bar{Y} = 20$, hallar la ecuación de la recta de regresión de (*a*) $Y$ sobre $X$ y (*b*) $X$ sobre $Y$.

**14.49.** Calcular (*a*) $s_{Y.X}$ y (*b*) $s_{X.Y}$ para los datos del Problema 14.48.

**14.50.** Si $s_{Y.X} = 3$ y $s_Y = 5$, calcular $r$

**14.51.** Si el coeficiente de correlación entre $X$ e $Y$ es 0.50, ¿qué porcentaje de la variación total queda inexplicado por la ecuación de regresión?

**14.52.** (*a*) Probar que la ecuación de la recta de regresión de $Y$ sobre $X$ puede escribirse

$$Y - \bar{Y} = \frac{s_{XY}}{s_X^2}(X - \bar{X})$$

(*b*) Escribir una ecuación análoga para la recta de regresión de $X$ sobre $Y$

**14.53.** (*a*) Calcular el coeficiente de correlación entre los valores correspondientes de $X$ e $Y$ dados en la Tabla 14.20.

(*b*) Multiplicar cada valor de $X$ en la tabla por 2 y sumar 6. Multiplicar cada valor de $Y$ en la tabla por 3 y restar 15. Hallar el coeficiente de correlación entre los dos nuevos conjuntos de valores, explicando por qué se obtiene o por qué no se obtiene el mismo resultado que en (*a*).

**Tabla 14.20**

| $X$ | $Y$ |
|---|---|
| 2 | 18 |
| 4 | 12 |
| 5 | 10 |
| 6 | 8 |
| 11 | 5 |

**14.54.** (*a*) Hallar las ecuaciones de regresión de $Y$ sobre $X$ para los datos considerados en el Problema 14.53, partes (*a*) y (*b*).

(*b*) Discutir la relación entre estas ecuaciones de regresión.

**14.55.** (*a*) Probar que el coeficiente de correlación entre $X$ e $Y$ puede expresarse

$$r = \frac{\overline{XY} - \bar{X}\bar{Y}}{\sqrt{[\overline{X^2} - \bar{X}^2][\overline{Y^2} - \bar{Y}^2]}}$$

(*b*) Usando ese método, resolver el Problema 14.1.

**14.56.** Probar que un coeficiente de correlación es independiente de la elección de origen de las variables o de las unidades en que se expresan. (*Ayuda:* Supóngase que $X' = c_1X + A$ e $Y' = c_2 Y + B$, donde $c_1$, $c_2$, $A$ y $B$ son constantes arbitrarias, y pruébese que el coeficiente de correlación entre $X'$ e $Y'$ es el mismo que entre $X$ e $Y$).

**14.57.** (*a*) Probar que, para regresión lineal,

$$\frac{s_{Y.X}^2}{s_Y^2} = \frac{s_{X.Y}^2}{s_X^2}$$

(*b*) ¿Es válido el resultado para regresión no lineal?

## COEFICIENTE DE CORRELACION PARA DATOS AGRUPADOS

**14.58.** Hallar el coeficiente de correlación entre las alturas y pesos de los 300 hombres adultos de EE.UU recogidos en la tabla de frecuencias dada en la Tabla 14.21.

**Tabla 14.21**

| Pesos | Alturas $X$ (in) | | | | |
|---|---|---|---|---|---|
| $Y$ (lb) | 59-62 | 63-66 | 67-70 | 71-74 | 75-78 |
| 90-109 | 2 | 1 | | | |
| 110-129 | 7 | 8 | 4 | 2 | |
| 130-149 | 5 | 15 | 22 | 7 | 1 |
| 150-169 | 2 | 12 | 63 | 19 | 5 |
| 170-189 | | 7 | 28 | 32 | 12 |
| 190-209 | | 2 | 10 | 20 | 7 |
| 210-229 | | | 1 | 4 | 2 |

**14.59.** (*a*) Hallar la recta de regresión de mínimos cuadrados de $Y$ sobre $X$ para los datos del Problema 14.58.

(*b*) Estimar los pesos de dos hombres cuyas alturas son 64 y 72 in.

**14.60.** Hallar (*a*) $s_{Y.X}$ y (*b*) $s_{X.Y}$ para los datos del Problema 14.58.

**14.61.** Establecer la fórmula (21) de este capítulo para el coeficiente de correlación de datos agrupados.

## CORRELACION DE SERIES EN EL TIEMPO

**14.62.** La Tabla 14.22 muestra los precios al por menor del cinc en EE.UU. y los correspondientes índices de precios al consumo en los

años 1978-1985. Hallar el coeficiente de correlación.

**14.63.** La Tabla 14.23 da la temperatura media y la precipitación en una ciudad durante el mes de julio de los años 1975-1984. Hallar el coeficiente de correlación.

## TEORIA MUESTRAL DE LA CORRELACION

**14.64.** Un coeficiente de correlación basado en una muestra de tamaño 27 resultó ser 0.40. ¿Se puede concluir que el coeficiente de correlación de la población correspondiente, al nivel de significación (*a*) 0.05 y (*b*) 0.01, difiere de cero?

**Tabla 14.22**

| Año | Precio de cinc (centavos por libra) | Indice de precios al consumo (1967 = 100) |
|---|---|---|
| 1978 | 31.0 | 195.4 |
| 1979 | 37.3 | 217.4 |
| 1980 | 37.4 | 246.8 |
| 1981 | 44.6 | 272.4 |
| 1982 | 38.5 | 289.1 |
| 1983 | 41.4 | 298.4 |
| 1984 | 48.6 | 311.1 |
| 1985 | 40.3 | 322.2 |

Fuente: U.S. Bureau of Labor Statistics and Bureau of Mines.

**Tabla 14.23**

| Año | Temperatura (°F) | Precipitación (in) |
|---|---|---|
| 1975 | 78.1 | 6.23 |
| 1976 | 71.8 | 3.64 |

**Tabla 14.23.** (*Continuación*)

| Año | Temperatura (°F) | Precipitación (in) |
|---|---|---|
| 1977 | 75.6 | 3.42 |
| 1978 | 72.7 | 2.84 |
| 1979 | 75.3 | 1.83 |
| 1980 | 73.6 | 2.82 |
| 1981 | 75.1 | 4.04 |
| 1982 | 75.3 | 2.56 |
| 1983 | 73.8 | 1.18 |
| 1984 | 70.4 | 4.19 |

**14.65.** Un coeficiente de correlación basado en una muestra de tamaño 35 ha dado 0.50. Al nivel de significación 0.05, ¿podemos rechazar la hipótesis de que el coeficiente de correlación de la población es (*a*) tan pequeño como 0.30 y (*b*) tan grande como 0.70?

**14.66.** Hallar los límites de confianza (*a*) 95% y (*b*) 99% para un coeficiente de correlación que se ha calculado como 0.60 a partir de una muestra de tamaño 28.

**14.67.** Resolver el Problema 14.66 con una muestra de tamaño 52.

**14.68.** Hallar los límites de confianza 95% para el coeficiente de correlación calculado en (*a*) el Problema 14.46 y (*b*) el Problema 14.58.

**14.69.** Dos coeficientes de correlación obtenidos de muestras de tamaños 23 y 28 resultan ser 0.80 y 0.95 respectivamente. ¿Podemos concluir a nivel de significación (*a*) 0.05 y (*b*) 0.01 que hay una diferencia significativa entre ellos?

## TEORIA MUESTRAL DE LA REGRESION

**14.70.** Con una muestra de tamaño 27 se ha encontrado una ecuación de regresión de $Y$

sobre $X$ dada por $Y = 25.0 + 2.00X$. Si $s_{Y.X} = 1{,}50$, $s_X = 3.00$ y $\bar{X} = 7.50$, hallar los límites de confianza (*a*) 95% y (*b*) 99% para el coeficiente de regresión.

**14.71.** En el Problema 14.70, contrastar la hipótesis de que el coeficiente de regresión de la población al nivel de significación 0.01 es (*a*) tan bajo como 1.70 y (*b*) tan alto como 2.20.

**14.72.** En el Problema 14.70, hallar los límites de confianza (*a*) 95% y (*b*) 99% para $Y$ cuando $X = 6.00$.

**14.73.** En el Problema 14.70, hallar los límites de confianza (*a*) 95% y (*b*) 99% para la media de todos los valores de $Y$ correspondientes a $X = 6.00$.

**14.74.** Con referencia al Problema 14.46, hallar los límites de confianza del 95% para (*a*) el coeficiente de regresión de $Y$ sobre $X$, (*b*) las presiones sanguíneas de las mujeres de 45 años y (*c*) la media de las presiones sanguíneas de las mujeres de 45 años.

# CAPITULO 15

# Correlación múltiple y parcial

## CORRELACION MULTIPLE

El grado de correlación existente entre tres o más variables se llama *correlación múltiple*. Los principios fundamentales implicados en los problemas de correlación múltiple son análogos a los de la correlación simple, tratados en el Capítulo 14.

## NOTACION DE SUBINDICES

Para permitir generalizaciones a números grandes de variables, conviene adoptar una notación de subíndices.

Denotaremos por $X_1$, $X_2$, $X_3$, ... las variables bajo consideración. Entonces denotaremos por $X_{11}, X_{12}, X_{13}, \ldots$ los valores que toma la variable $X_1$, y $X_{21}, X_{22}, X_{23}, \ldots$ los que toma la variable $X_2$, etcétera. Con esta notación, una suma tal como $X_{21} + X_{22} + X_{23} + \cdots + X_{2N}$ se escribirá $\sum_{j=1}^{N} X_{2j}$, $\sum_j X_{2j}$, o simplemente $\sum X_2$. Cuando no haya ambigüedad, usaremos la última notación. En tal caso, la media de $X_2$ se escribe $\bar{X}_2 = \sum X_2/N$.

## ECUACIONES DE REGRESION Y PLANOS DE REGRESION

Una *ecuación de regresión* es una ecuación para estimar una variable dependiente, digamos $X_1$, a partir de las variables independientes $X_2$, $X_3$, ... y se llama una *ecuación de regresión de* $X_1$ *sobre* $X_2$, $X_3$, ... En notación funcional eso se escribe a veces brevemente como $X_1 = F(X_2, X_3, \ldots)$ (léase «$X_1$ es una función de $X_2$, $X_3$, etc.»).

Para el caso de tres variables, la ecuación de regresión más simple de $X_1$ sobre $X_2$ y $X_3$ tiene la forma

$$X_1 = b_{1.23} + b_{12.3}X_2 + b_{13.2}X_3 \tag{1}$$

donde $b_{1.23}$ $b_{12.3}$, y $b_{13.2}$ son constantes. Si mantenemos $X_3$ constante en la ecuación (1), el gráfico de $X_1$ versus $X_2$ es una recta con pendiente $b_{12.3}$. Si mantenemos constante $X_2$, el gráfico de $X_1$ versus $X_3$ es una recta con pendiente $b_{13.2}$. Es claro que los subíndices tras el punto indican las variables que se mantienen constantes en cada caso.

Debido al hecho de que $X_1$ varía parcialmente a causa de la variación en $X_2$ y parcialmente a

causa de la de $X_3$, se llama a $b_{12.3}$ y $b_{13.2}$ los *coeficientes de regresión parcial* de $X_1$ sobre $X_2$ dejando $X_3$ constante, y de $X_1$ sobre $X_3$ dejando $X_2$ constante, respectivamente.

La ecuación (1) se llama una *ecuación de regresión lineal* de $X_1$ sobre $X_2$ y $X_3$. En un sistema rectangular tridimensional de coordenadas representa un plano llamado *plano de regresión* y es generalización de la recta de regresión en dos variables, tal como se consideró en el Capítulo 13.

## ECUACIONES NORMALES PARA EL PLANO DE REGRESION DE MINIMOS CUADRADOS

Así como existen rectas de regresión de mínimos cuadrados que aproximan un conjunto de $N$ puntos dato $(X, Y)$ en un diagrama de dispersión, existen también *planos de regresión de mínimos cuadrados* que ajustan un conjunto de $N$ puntos dato $(X_1, X_2, X_3)$ en un diagrama de dispersión tridimensional.

El plano de regresión de mínimos cuadrados de $X_1$ sobre $X_2$ y $X_3$ tiene ecuación (1) donde $b_{1.23}$, $b_{12.3}$ y $b_{13.2}$ se determinan resolviendo simultáneamente las *ecuaciones normales*

$$\begin{aligned} \sum X_1 &= b_{1.23}N + b_{12.3}\sum X_2 + b_{13.2}\sum X_3 \\ \sum X_1X_2 &= b_{1.23}\sum X_2 + b_{12.3}\sum X_2^2 + b_{13.2}\sum X_2X_3 \\ \sum X_1X_3 &= b_{1.23}\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum X_3^2 \end{aligned} \tag{2}$$

Estas pueden obtenerse formalmente multiplicando ambos lados de la ecuación (1) por 1, $X_2$ y $X_3$ sucesivamente y sumando en ambos lados.

A menos que se especifique lo contrario, siempre que nos refiramos a una ecuación de regresión se supondrá que se habla de la ecuación de regresión de mínimos cuadrados.

Si $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$ y $x_3 = X_3 - \bar{X}_3$, la ecuación de regresión de $X_1$ sobre $X_2$ y $X_3$ pueden escribirse más sencillamente como

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \tag{3}$$

donde $b_{12.3}$ y $b_{13.2}$ se obtienen resolviendo simultáneamente las ecuaciones

$$\begin{aligned} \sum x_1x_2 &= b_{12.3}\sum x_2^2 + b_{13.2}\sum x_2x_3 \\ \sum x_1x_3 &= b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2 \end{aligned} \tag{4}$$

Estas ecuaciones que son equivalentes a las ecuaciones normales (2) se pueden obtener formalmente multiplicando (3) por $x_2$ y $x_3$ sucesivamente y sumando (véase Prob. 15.8).

## PLANOS DE REGRESION Y COEFICIENTES DE CORRELACION

Si los coeficientes de correlación entre variables $X_1$ y $X_2$, $X_1$ y $X_3$ y $X_2$ y $X_3$, tal como se calculaban en el Capítulo 14, se denotan respectivamente por $r_{12}$, $r_{13}$ y $r_{23}$ (llamados a veces *coeficientes de correlación de orden cero*), entonces el plano de regresión de mínimos cuadrados tiene la ecuación

$$\frac{x_1}{s_1} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\frac{x_2}{s_2} + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\frac{x_3}{s_3} \tag{5}$$

donde $x_1 = X - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$ y $x_3 = X_3 - \bar{X}_3$ y donde $s_1$, $s_2$ y $s_3$ son la desviación típica de $X_1$, $X_2$ y $X_3$, respectivamente (véase Prob. 15.9).

Nótese que si la variable $X_3$ no existiese y si $X_1 = Y$ y $X_2 = X$, entonces la ecuación (5) se reduce a la ecuación (25) del Capítulo 14.

## ERROR TIPICO DE ESTIMACION

Por una generalización obvia de la ecuación 8 del Capítulo 14, podemos definir el *error típico de estimación de* $X_1$ sobre $X_2$ y $X_3$ como

$$s_{1.23} = \sqrt{\frac{\sum (X_1 - X_{1,\,\mathrm{est}})^2}{N}} \tag{6}$$

donde $X_{1,\,\mathrm{est}}$ indica los valores estimados de $X_1$ tal como se calculan mediante las ecuaciones de regresión (1) o (5).

En términos de los coeficientes de correlación $r_{12}$, $r_{13}$ y $r_{23}$, el error típico de estimación se puede calcular también a partir del resultado

$$s_{1.23} = s_1 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \tag{7}$$

La interpretación muestral del error típico de estimación para dos variables, vista en la página 324 para el caso en que $N$ es grande, puede extenderse a tres dimensiones sustituyendo las rectas paralelas a la de regresión por planos paralelos al plano de regresión. Una estimación mejor del error típico de estimación de la población viene dada por $\hat{s}_{1.23} = \sqrt{N/(N-3)}\, s_{1.23}$.

## COEFICIENTE DE CORRELACION MULTIPLE

El *coeficiente de correlación múltiple* se define por extensión de la ecuación (12) o (14) del Capítulo 14. En el caso de dos variables independientes, por ejemplo, el coeficiente de correlación múltiple viene dado por

$$R_{1.23} = \sqrt{1 - \frac{s_{1.23}^2}{s_1^2}} \tag{8}$$

donde $s_1$ es la desviación típica de $X_1$ y $s_{1.23}$ viene dado por la ecuación (6) o (7). La cantidad $R_{1.23}^2$ se llama *coeficiente de determinación múltiple*.

Cuando se usa una ecuación de regresión lineal, el coeficiente de correlación múltiple se llama *coeficiente de correlación múltiple lineal*. Salvo que se especifique lo contrario, siempre que nos refiramos a correlación múltiple querremos decir correlación múltiple lineal.

En términos de $r_{12}$, $r_{13}$ y $r_{23}$, la ecuación (8) se puede expresar

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \tag{9}$$

Un coeficiente de correlación múltiple, tal como $R_{1.23}$, está entre 0 y 1. Cuanto más cerca de 1, más precisa es la relación lineal entre las variables. Cuanto más cerca de 0, peor es la relación lineal. Si el coeficiente de correlación múltiple es 1, la correlación se dice *perfecta*. Aunque un coeficiente de correlación igual a 0 indica que no hay relación lineal entre las variables, puede haber una *relación no lineal*.

## CAMBIO DE VARIABLE DEPENDIENTE

Los resultados anteriores son válidos cuando se considera a $X_1$ como variable dependiente. Sin embargo, si queremos considerar a $X_3$ (por ejemplo) como la variable dependiente en vez de $X_1$, sólo tendríamos que reemplazar los subíndices 1 por 3 y 3 por 1 en las fórmulas ya obtenidas. Por ejemplo, la ecuación de regresión de $X_3$ sobre $X_1$ y $X_2$ sería

$$\frac{x_3}{s_3} = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2}\right)\frac{x_2}{s_2} + \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right)\frac{x_1}{s_1} \tag{10}$$

que se deduce de (5) haciendo uso de $r_{32} = r_{23}$, $r_{31} = r_{13}$ y $r_{21} = r_{12}$.

## GENERALIZACIONES A MAS DE TRES VARIABLES

Estas se obtienen por analogía con los resultados precedentes. Así, las ecuaciones de regresión lineales de $X_1$ sobre $X_2$, $X_3$ y $X_4$ pueden escribirse

$$X_1 = b_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 \tag{11}$$

y representan un *hiperplano en el espacio de cuatro dimensiones*. Multiplicando ambos miembros de (11) por 1, $X_2$, $X_3$ y $X_4$ sucesivamente y sumando, se llega a las ecuaciones normales para determinar $b_{1.234}$, $b_{12.34}$, $b_{13.24}$ y $b_{14.23}$; sustituyendo estas en la ecuación (11) nos da la *ecuación de regresión de mínimos cuadrados* de $X_1$ sobre $X_2$, $X_3$ y $X_4$. Esta ecuación de regresión de mínimos cuadrados se puede escribir de modo similar a la (5). (Véase Prob. 15.41.)

## CORRELACION PARCIAL

A menudo es importante medir la correlación entre una variable dependiente y una variable independiente particular, cuando todas las demás variables se suprimen (indicado con frecuencia con la frase «quedando iguales las restantes»). Esto se consigue definiendo un *coeficiente de correlación parcial*, como en la ecuación (12) del Capítulo 14, excepto que hemos de considerar la variación explicada y la variación inexplicada que aparecen tanto con como sin la variable independiente particular.

Si denotamos por $r_{12.3}$ el coeficiente de correlación parcial entre $X_1$ y $X_2$ manteniendo $X_3$ constante encontramos que

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \tag{12}$$

De la misma manera, si $r_{12.34}$ es el coeficiente de correlación parcial entre $X_1$ y $X_2$ manteniendo $X_3$ y $X_4$ constante, entonces

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{(1 - r_{13.4}^2)(1 - r_{23.4}^2)}} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}} \tag{13}$$

Estos resultados son útiles porque por su mediación cualquier coeficiente de correlación parcial se puede hacer depender en última instancia de los coeficientes de correlación $r_{12}$, $r_{23}$, etc. (o sea, los *coeficientes de correlación de orden cero*).

En el caso de dos variables $X$ e $Y$, si las dos rectas de regresión tienen ecuaciones $Y = a_0 + a_1X$ y $X = b_0 + b_1Y$, hemos visto que $r^2 = a_1b_1$ (véase Prob. 14.22). Este resultado admite generalización. Así, si

$$X_1 = b_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 \tag{14}$$

y

$$X_4 = b_{4.123} + b_{41.23}X_1 + b_{42.13}X_2 + b_{43.12}X_3 \tag{15}$$

son ecuaciones de regresión lineales de $X_1$ sobre $X_2$, $X_3$ y $X_4$ y de $X_4$ sobre $X_1$, $X_2$ y $X_3$, respectivamente, entonces

$$r_{14.23}^2 = b_{14.23}b_{41.23} \tag{16}$$

(véase Prob. 15.18). Esto se puede adoptar como punto de partida para una definición de los coeficientes de correlación parcial lineales.

## RELACIONES ENTRE COEFICIENTES DE CORRELACION PARCIAL Y MULTIPLE

Hay interesantes resultados que conectan los coeficientes de correlación múltiple. Como ejemplo,

$$1 - R_{1.23}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2) \tag{17}$$

$$1 - R_{1.234}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2) \tag{18}$$

Es fácil generalizar estos resultados.

## REGRESION MULTIPLE NO LINEAL

Los resultados anteriores para regresión múltiple lineal se pueden extender a la regresión múltiple no lineal. Se pueden definir coeficientes de correlación parcial y múltiple por métodos similares a los ya vistos.

# PROBLEMAS RESUELTOS

## ECUACION DE REGRESION EN TRES VARIABLES

**15.1.** Usando notación de subíndices adecuada, escribir la ecuación de regresión de (*a*) $X_2$ sobre $X_1$ y $X_3$; (*b*) $X_3$ sobre $X_1$, $X_2$ y $X_4$, y (*c*) $X_5$ sobre $X_1$, $X_2$, $X_3$ y $X_4$.

**Solución**

(*a*) $X_2 = b_{2.13} + b_{21.3}X_1 + b_{23.1}X_3$

(*b*) $X_3 = b_{3.124} + b_{31.24}X_1 + b_{32.14}X_2 + b_{34.12}X_4$

(*c*) $X_5 = b_{5.1234} + b_{51.234}X_1 + b_{52.134}X_2 + b_{53.124}X_3 + b_{54.123}X_4$

**15.2.** Escribir las ecuaciones normales correspondientes a la ecuación de regresión (*a*) $X_3 = b_{3.12} + b_{31.2}X_1 + b_{32.1}X_2$ y (*b*) $X_1 = b_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$.

**Solución**

(*a*) Multiplicar la ecuación sucesivamente por 1, $X_1$ y $X_2$, y sumar en ambos lados. Las ecuaciones normales son

$$\sum X_3 = b_{3.12}N + b_{31.2}\sum X_1 + b_{32.1}\sum X_2$$

$$\sum X_1X_3 = b_{3.12}\sum X_1 + b_{31.2}\sum X_1^2 + b_{32.1}\sum X_1X_2$$

$$\sum X_2X_3 = b_{3.12}\sum X_2 + b_{31.2}\sum X_1X_2 + b_{32.1}\sum X_2^2$$

(*b*) Multiplicar la ecuación sucesivamente por 1, $X_2$, $X_3$ y $X_4$, y sumar en ambos lados. Las ecuaciones normales son

$$\sum X_1 = b_{1.234}N + b_{12.34}\sum X_2 + b_{13.24}\sum X_3 + b_{14.23}\sum X_4$$

$$\sum X_1X_2 = b_{1.234}\sum X_2 + b_{12.34}\sum X_2^2 + b_{13.24}\sum X_2X_3 + b_{14.23}\sum X_2X_4$$

$$\sum X_1X_3 = b_{1.234}\sum X_3 + b_{12.34}\sum X_2X_3 + b_{13.24}\sum X_3^2 + b_{14.23}\sum X_3X_4$$

$$\sum X_1X_4 = b_{1.234}\sum X_4 + b_{12.34}\sum X_2X_4 + b_{13.24}\sum X_3X_4 + b_{14.23}\sum X_4^2$$

Nótese que esto no es una demostración de las ecuaciones normales, sino sólo un medio de acordarse de ellas.

El número de ecuaciones normales es igual al número de constantes desconocidas.

**15.3.** La Tabla 15.1 da los pesos $X_1$ redondeados en libras (lb), las alturas $X_2$ redondeadas en pulgadas (in), y las edades $X_3$ redondeadas en años, de niños.

(*a*) Hallar la ecuación de regresión de mínimos cuadrados de $X_1$ sobre $X_2$ y $X_3$.

(*b*) Determinar los valores estimados de $X_1$ a partir de los valores dados de $X_2$ y $X_3$.

(*c*) Estimar el peso de un niño de 9 años que mide 54 in.

**Tabla 15.1**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Peso ($X_1$) | 64 | 71 | 53 | 67 | 55 | 58 | 77 | 57 | 56 | 51 | 76 | 68 |
| Altura ($X_2$) | 57 | 59 | 49 | 62 | 51 | 50 | 55 | 48 | 52 | 42 | 61 | 57 |
| Edad ($X_3$) | 8 | 10 | 6 | 11 | 8 | 7 | 10 | 9 | 10 | 6 | 12 | 9 |

**Solución**

(*a*) La ecuación de regresión lineal de $X_1$ sobre $X_2$ y $X_3$ puede expresarse

$$X_1 = b_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

Las ecuaciones normales de la ecuación de regresión de mínimos cuadrados son

$$\begin{aligned} \sum X_1 &= b_{1.23}N + b_{12.3}\sum X_2 + b_{13.2}\sum X_3 \\ \sum X_1X_2 &= b_{1.23}\sum X_2 + b_{12.3}\sum X_2^2 + b_{13.2}\sum X_2X_3 \\ \sum X_1X_3 &= b_{1.23}\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum X_3^2 \end{aligned} \tag{19}$$

El camino a seguir se indica en la Tabla 15.2. (Aunque la columna encabezada por $X_1^2$ no se necesita ahora, se ha añadido para referencia posterior.)

**Tabla 15.2**

| $X_1$ | $X_2$ | $X_3$ | $X_1^2$ | $X_2^2$ | $X_3^2$ | $X_1X_2$ | $X_1X_3$ | $X_2X_3$ |
|---|---|---|---|---|---|---|---|---|
| 64 | 57 | 8 | 4096 | 3249 | 64 | 3648 | 512 | 456 |
| 71 | 59 | 10 | 5041 | 3481 | 100 | 4189 | 710 | 590 |
| 53 | 49 | 6 | 2809 | 2401 | 36 | 2597 | 318 | 294 |
| 67 | 62 | 11 | 4489 | 3844 | 121 | 4154 | 737 | 682 |
| 55 | 51 | 8 | 3025 | 2601 | 64 | 2805 | 440 | 408 |
| 58 | 50 | 7 | 3364 | 2500 | 49 | 2900 | 406 | 350 |
| 77 | 55 | 10 | 5929 | 3025 | 100 | 4235 | 770 | 550 |
| 57 | 48 | 9 | 3249 | 2304 | 81 | 2736 | 513 | 432 |
| 56 | 52 | 10 | 3136 | 2704 | 100 | 2912 | 560 | 520 |
| 51 | 42 | 6 | 2601 | 1764 | 36 | 2142 | 306 | 252 |
| 76 | 61 | 12 | 5776 | 3721 | 144 | 4636 | 912 | 732 |
| 68 | 57 | 9 | 4624 | 3249 | 81 | 3876 | 612 | 513 |
| $\sum X_1 = 753$ | $\sum X_2 = 643$ | $\sum X_3 = 106$ | $\sum X_1^2 = 48{,}139$ | $\sum X_2^2 = 34{,}843$ | $\sum X_3^2 = 976$ | $\sum X_1X_2 = 40{,}830$ | $\sum X_1X_3 = 6796$ | $\sum X_2X_3 = 5779$ |

Usando la Tabla 15.2, las ecuaciones normales (19) pasan a ser

$$\begin{aligned} 12b_{1.23} + 643b_{12.3} + 106b_{13.2} &= 753 \\ 643b_{1.23} + 34{,}843b_{12.3} + 5.779b_{13.2} &= 40.830 \\ 106b_{1.23} + 5{,}779b_{12.3} + 976b_{13.2} &= 6{,}796 \end{aligned} \tag{20}$$

Resolviendo, $b_{1.23} = 3.6512$, $b_{12.3} = 0.8546$ y $b_{13.2} = 1.5063$, y la ecuación de regresión pedida será

$$X_1 = 3.6512 + 0.8546X_2 + 1.5063X_3 \quad \text{o sea} \quad X_1 = 3.65 + 0.855X_2 + 1.506\ X_3 \tag{21}$$

Para otro método, que evita resolver ecuaciones simultáneas, véase el Problema 15.6.

(*b*) Usando la ecuación de regresión (21), obtenemos los valores estimados de $X_1$, denotados por

$X_{1,\text{est}}$, sustituyendo los valores correspondientes de $X_2$ y $X_3$. Por ejemplo, sustituyendo $X_2 = 57$ y $X_3 = 8$ en (21), vemos que $X_{1,\text{est}} = 64.414$.

Los otros valores estimados de $X_1$ se obtienen del mismo modo. Se recogen en la Tabla 15.3 junto con los valores muestrales de $X_1$.

(*c*) Poniendo $X_2 = 54$ y $X_3 = 9$ en la ecuación (21), el peso estimado es $X_{1,\text{est}} = 63.356$, es decir, unas 63 lb.

**Tabla 15.3**

| $X_{1,\text{est}}$ | 64.414 | 69.136 | 54.564 | 73.206 | 59.286 | 56.925 | 65.717 | 58.229 | 63.153 | 48.582 | 73.857 | 65.920 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | 64 | 71 | 53 | 67 | 55 | 58 | 77 | 57 | 56 | 51 | 76 | 68 |

**15.4.** Calcular las derivaciones estándar (*a*) $s_1$, (*b*) $s_2$ y (*c*) $s_3$ para los datos del Problema 15.3.

**Solución**

(*a*) La cantidad $s_1$ es la desviación típica de la variable $X_1$. Entonces, usando la Tabla 15.2 del Problema 15.3(*a*) y los métodos del Capítulo 4, se ve que

$$s_1 = \sqrt{\frac{\sum X_1^2}{N} - \left(\frac{\sum X_1}{N}\right)^2} = \sqrt{\frac{48{,}139}{12} - \left(\frac{753}{12}\right)^2} = 8.6035 \quad \text{o sea} \quad 8.6 \text{ lb}$$

(*b*)
$$s_2 = \sqrt{\frac{\sum X_2^2}{N} - \left(\frac{\sum X_2}{N}\right)^2} = \sqrt{\frac{34{,}843}{12} - \left(\frac{643}{12}\right)^2} = 5.6930 \quad \text{o sea} \quad 5.7 \text{ in}$$

(*c*)
$$s_3 = \sqrt{\frac{\sum X_3^2}{N} - \left(\frac{\sum X_3}{N}\right)^2} = \sqrt{\frac{976}{12} - \left(\frac{106}{12}\right)^2} = 1.8181 \quad \text{o sea} \quad 1.8 \text{ años}$$

**15.5.** Calcular (*a*) $r_{12}$, (*b*) $r_{13}$ y (*c*) $r_{23}$ para los datos del Problema 15.3.

**Solución**

(*a*) La cantidad $r_{12}$ es el coeficiente de correlación lineal entre las variables $X_1$ y $X_2$, ignorando la variable $X_3$. Entonces, usando los métodos del Capítulo 14, se tiene

$$r_{12} = \frac{N\sum X_1X_2 - (\sum X_1)(\sum X_2)}{\sqrt{[N\sum X_1^2 - (\sum X_1)^2][N\sum X_2^2 - (\sum X_2)^2]}} =$$

$$= \frac{(12)(40{,}830) - (753)(643)}{\sqrt{[(12)(48.139) - (753)^2][(12)(34.843) - (643)^2]}} = 0.8196 \quad \text{o sea} \quad 0.82$$

(*b*) y (*c*) Usando las fórmulas correspondientes, se obtiene $r_{12} = 0.7698$, o sea 0.77 y $r_{23} = 0.7984$, ó 0.80.

**15.6.** Resolver el Problema 15.3(*a*) usando la ecuación (5) y los resultados de los Problemas 15.4 y 15.5.

**Solución**

La ecuación de regresión de $X_1$ sobre $X_2$ y $X_3$ es, multiplicando cada miembro de la ecuación (5) por $s_1$,

$$x_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_2}\right)x_2 + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_3}\right)x_3 \tag{22}$$

donde $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$ y $x_3 = X_3 - \bar{X}_3$. Usando los resultados de los Problemas 15.4 y 15.5, la (22) se convierte en

$$x_1 = 0.8546x_2 + 1.5063x_3$$

Como $\bar{X}_1 = \dfrac{\sum X_1}{N} = \dfrac{753}{12} = 62.750$ $\bar{X}_2 = \dfrac{\sum X_2}{N} = 53.583$ y $\bar{X}_3 = 8.833$

(por la Tabla 15.2 del Prob. 15.3), la requerida ecuación se puede expresar

$$X_1 - 62.750 = 0.8546(X_2 - 53.583) + 1.506(X_3 - 8.833)$$

que coincide con el resultado del Problema 15.3(*a*).

**15.7.** Para los datos del Problema 15.3, determinar (*a*) el crecimiento promedio en peso por pulgada de crecimiento en altura, para niños de la misma edad y (*b*) el crecimiento promedio en peso por año, para niños de la misma altura.

**Solución**

De la ecuación de regresión obtenida en el Problema 15.3(*a*) o en el 15.6 vemos que la respuesta a (*a*) es 0.8546, o sea unas 0.9 lb, y la de (*b*) es 1.5063 lb, o sea unas 1.5 lb.

**15.8.** Probar que las ecuaciones (3) y (4) de este capítulo se siguen de las ecuaciones (1) y (2).

**Solución**

De la primera de las ecuaciones (2), dividiendo ambos lados por $N$, se tiene

$$\bar{X}_1 = b_{1.23} + b_{12.3}\bar{X}_2 + b_{13.2}\bar{X}_3 \tag{23}$$

Restando (23) de (1) vemos que

$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

o

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \tag{24}$$

que no es sino la ecuación (3).

Sean $X_1 = x_1 + \bar{X}_1$, $X_2 = x_2 + \bar{X}_2$ y $X_3 = x_3 + \bar{X}_3$ en la segunda y tercera ecuaciones (2). Entonces, tras algunas manipulaciones algebraicas, usando los resultados $\sum x_1 = \sum x_2 = \sum x_3 = 0$, pasan a ser

$$\sum x_1x_2 = b_{12.3}\sum x_2^2 + b_{13.2}\sum x_2x_3 + N\bar{X}_2[b_{1.23} + b_{12.3}\bar{X}_2 + b_{13.2}\bar{X}_3 - \bar{X}_1] \tag{25}$$

$$\sum x_1x_3 = b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2 + N\bar{X}_3[b_{1.23} + b_{12.3}\bar{X}_2 + b_{13.2}\bar{X}_3 - \bar{X}_1] \tag{26}$$

que se reducen a (4) pues las cantidades entre corchetes de la derecha en las ecuaciones (25) y (26) son cero debido a la ecuación (1).

*Otro método*

Véase Problema 15.30.

**15.9.** Establecer la ecuación (5), que copiamos aquí:

$$\frac{x_1}{s_1} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\frac{x_2}{s_2} + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\frac{x_3}{s_3} \tag{5}$$

**Solución**

De las ecuaciones (25) y (26)

$$\begin{aligned} b_{12.3}\sum x_2^2 \quad + b_{13.2}\sum x_2x_3 &= \sum x_1x_2 \\ b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2 \quad &= \sum x_1x_3 \end{aligned} \tag{27}$$

Como $s_2^2 = \dfrac{\sum x_2^2}{N}$ y $s_3^2 = \dfrac{\sum x_3^2}{N}$

$\sum x_2^2 = Ns_2^2$ y $\sum x_3^2 = Ns_3^2$. Puesto que

$$r_{23} = \frac{\sum x_2x_3}{\sqrt{(\sum x_2^2)(\sum x_3^2)}} = \frac{\sum x_2x_3}{Ns_2s_3}$$

$\sum x_2x_3 = Ns_2s_3r_{23}$. Análogamente, $\sum x_1x_2 = Ns_1s_2r_{12}$ y $\sum x_1x_3 = Ns_1s_3r_{13}$. Sustituyendo en (27) y simplificando, hallamos

$$\begin{aligned} b_{12.3}s_2 \quad + b_{13.2}s_3r_{23} &= s_1r_{12} \\ b_{12.3}s_2r_{23} + b_{13.2}s_3 \quad &= s_1r_{13} \end{aligned} \tag{28}$$

Resolviendo simultáneamente, tenemos

$$b_{12.3} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_2}\right) \quad \text{y} \quad b_{13.2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_3}\right)$$

que sustituidas en la ecuación $x_1 = b_{12.3}x_2 + b_{13.2}x_3$ [ecuación (24)] y dividiendo por $s_1$, dan el resultado anunciado.

## ERROR TIPICO DE ESTIMACION

**15.10.** Calcular el error típico de estimación de $X_1$ sobre $X_2$ y $X_3$ para los datos del Problema 15.3.

**Solución**

De la Tabla 15.3 del Problema 15.3(*b*) vemos que

$$s_{1.23} = \sqrt{\frac{\sum (X_1 - X_{1.\text{est}})^2}{N}} = \sqrt{\frac{(64 - 64.414)^2 + (71 - 69.136)^2 + \cdots + (68 - 65.920)^2}{12}} = 4.6447 \text{ o sea } 4.6 \text{ lb}$$

El error típico de estimación de la población se estima como $\hat{s}_{1.23} = \sqrt{N/(N-3)}s_{1.23} = 5.3$ lb en este caso.

**15.11.** Deducir el resultado del Problema 15.10, usando

$$s_{1.23} = s_1\sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

**Solución**

Por los Problemas 15.4(*a*) y 15.5 tenemos

$$s_{1.23} = 8.6035\sqrt{\frac{1 - (0.8196)^2 - (0.7698)^2 - (0.7984)^2 + 2(0.8196)(0.7698)(0.7984)}{1 - (0.7984)^2}} = 4.6 \text{ lb}$$

Nótese que con el método de este problema el error típico de estimación se puede encontrar sin recurrir a la ecuación de regresión.

## COEFICIENTE DE CORRELACION MULTIPLE

**15.12.** Calcular el coeficiente de correlación múltiple lineal de $X_1$ sobre $X_2$ y $X_3$ para los datos del Problema 15.3.

**Solución**

*Primer método*

De los resultados de los Problemas 15.4(*a*) y 15.10 tenemos

$$R_{1.23} = \sqrt{1 - \frac{s_{1.23}^2}{s_1^2}} = \sqrt{1 - \frac{(4.6447)^2}{(8.6035)^2}} = 0.8418$$

*Segundo método*

De los resultados del Problema 15.5 tenemos

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} = \sqrt{\frac{(0.8169)^2 + (0.7698)^2 - 2(0.8196)(0.7698)(0.7984)}{1 - (0.7984)^2}} = 0.8418$$

Obsérvese que el coeficiente de correlación múltiple, $R_{1.23}$, es mayor que cualquiera de los coeficientes $r_{12}$ o $r_{13}$ (véase Prob. 15.5). Esto ocurre siempre y era de esperar, de hecho, ya que teniendo en cuenta variables independientes relevantes adicionales llegaríamos a una relación más exacta entre las variables.

**15.13.** Calcular el coeficiente de determinación múltiple de $X_1$ sobre $X_2$ y $X_3$ para los datos del Problema 15.3.

**Solución**

El coeficiente de determinación múltiple de $X_1$ sobre $X_2$ y $X_3$ es

$$R_{1.23}^2 = (0.8418)^2 = 0.7086$$

usando el Problema 15.12. Así pues, alrededor del 71% de la variación total de $X$ es explicada por la ecuación de regresión.

**15.14.** Para los datos del Problema 15.3, calcular (*a*) $R_{2.13}$ y (*b*) $R_{3.12}$ y comparar sus valores con el valor de $R_{1.23}$.

**Solución**

(*a*) $$R_{2.13}=\sqrt{\frac{r_{12}^2+r_{23}^2-2r_{12}r_{13}r_{23}}{1-r_{13}^2}}=\sqrt{\frac{(0.8196)^2+(0.7984)^2-2(0.8196)(0.7698)(0.7984)}{1-(0.7698)^2}}=0.8606$$

(*b*) $$R_{3.12}=\sqrt{\frac{r_{13}^2+r_{23}^2-2r_{12}r_{13}r_{23}}{1-r_{12}^2}}=\sqrt{\frac{(0.7698)^2+(0.7984)^2-2(0.8196)(0.7698)(0.7984)}{1-(0.8196)^2}}=0.8234$$

Este problema ilustra el hecho de que, en general, $R_{2.13}$, $R_{3.12}$ y $R_{1.23}$ no son necesariamente iguales, como se ve comparando con el Problema 15.12.

**15.15.** Si $R_{1.23} = 1$, probar que (*a*) $R_{2.13} = 1$ y (*b*) $R_{3.12} = 1$.

**Solución**

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \tag{29}$$

y

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}} \tag{30}$$

(*a*) En la ecuación (29), poniendo $R_{1.23} = 1$ y elevando al cuadrado ambos lados, $r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23} = 1 - r_{23}^2$. Entonces

$$r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23} = 1 - r_{13}^2 \quad \text{o sea} \quad \frac{r_{12}^2 + r_{23}^2 - 2r_1r_{13}r_{23}}{1 - r_{13}^2} = 1$$

Esto es, $R_{2.13}^2 = 1$ o sea $R_{2.13} = 1$, ya que el coeficiente de correlación múltiple se considera no negativo.

(*b*) $R_{3.12} = 1$ se sigue de la parte (*a*) intercambiando los subíndices 2 y 3 en el resultado $R_{2.13} = 1$.

**15.16.** Si $R_{1.23} = 0$, ¿se deduce necesariamente que $R_{2.13} = 0$?

**Solución**

De la ecuación (29), $R_{1.23} = 0$ si y sólo si

$$r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23} = 0 \quad \text{o sea} \quad 2r_{12}r_{13}r_{23} = r_{12}^2 + r_{13}^2$$

Entonces, de la ecuación (30) tenemos

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - (r_{12}^2 + r_{13}^2)}{1 - r_{13}^2}} = \sqrt{\frac{r_{23}^2 - r_{13}^2}{1 - r_{13}^2}}$$

que no es necesariamente cero.

## CORRELACION PARCIAL

**15.17.** Para los datos del Problema 15.3, calcular los coeficientes de correlación parcial lineal (*a*) $r_{12.3}$, (*b*) $r_{13.2}$ y (*c*) $r_{23.1}$.

**Solución**

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \qquad r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}} \qquad r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}}$$

Para los resultados del Problema 15.5 sabemos que $r_{12.3} = 0.5334$, $r_{13.2} = 0.3346$ y $r_{23.1} = 0.4580$. Se sigue que para niños de la misma edad, el coeficiente de correlación entre peso y edad es 0.53; para niños del mismo peso, el coeficiente de correlación entre peso y edad es sólo 0.33. Como estos resultados se basan en una muestra pequeña de sólo 12 niños, no son, claro está, tan fiables como los que se obtendrían con una muestra grande.

**15.18.** Si $X_1 = b_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$ y $X_3 = b_{3.12} + b_{32.1}X_2 + b_{31.2}X_1$ son la ecuación de regresión de $X_1$ sobre $X_2$ y $X_3$ y de $X_3$ sobre $X_2$ y $X_1$, respectivamente, probar que $r_{13.2}^2 = b_{13.2}b_{31.2}$.

**Solución**

La ecuación de regresión de $X_1$ sobre $X_2$ y $X_3$ se puede escribir [véase ecuación (5) de este capítulo]

$$X_1 - \bar{X}_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_3}\right)(X_3 - \bar{X}_3) \tag{31}$$

La ecuación de regresión de $X_3$ sobre $X_2$ y $X_1$ se puede escribir [véase ecuación (10)]

$$X_3 - \bar{X}_3 = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_3}{s_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_3}{s_1}\right)(X_1 - \bar{X}_1) \tag{32}$$

De (31) y (32) los coeficientes de $X_3$ y $X_1$ son, respectivamente,

$$b_{13.2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_3}\right) \quad \text{y} \quad b_{31.2} = \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_3}{s_1}\right)$$

Luego
$$b_{13.2}b_{31.2} = \frac{(r_{13} - r_{12}r_{23})^2}{(1 - r_{23}^2)(1 - r_{12}^2)} = r_{13.2}^2$$

**15.19.** Si $r_{12.3} = 0$, demostrar que

$$(a)\ r_{13.2} = r_{13}\sqrt{\frac{1 - r_{23}^2}{1 - r_{12}^2}} \qquad (b)\ r_{23.1} = r_{23}\sqrt{\frac{1 - r_{13}^2}{1 - r_{12}^2}}$$

**Solución**

Si
$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = 0$$

tenemos $r_{12} = r_{13}r_{23}$

$$(a)\quad r_{13.2}=\frac{r_{13}-r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}=\frac{r_{13}-(r_{13}r_{23})r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}=\frac{r_{13}(1-r_{23}^2)}{\sqrt{(1-r_{12}^2)(1-r_{23}^3)}}=r_{13}\sqrt{\frac{1-r_{23}^2}{1-r_{12}^2}}$$

(*b*) Intercambiar los subíndices 1 y 2 en el resultado de la parte (*a*).

## CORRELACION MULTIPLE Y PARCIAL EN CUATRO O MAS VARIABLES

**15.20.** Un examen de ingreso en cierta universidad consistía de tres partes; matemáticas, inglés y cultura general. Para analizar la capacidad del examen a la hora de predecir el rendimiento en un curso de estadística, se estudiaron los datos de 200 estudiantes. Llamando

$X_1$ = nota en estadística  $X_3$ = nota en inglés
$X_2$ = nota en matemáticas  $X_4$ = nota en cultura general

se han obtenido los siguientes resultados:

$$\bar{X}_1 = 75 \quad s_1 = 10 \quad \bar{X}_2 = 24 \quad s_2 = 5$$
$$\bar{X}_3 = 15 \quad s_3 = 3 \quad \bar{X}_4 = 36 \quad s_4 = 6$$
$$r_{12} = 0.90 \quad r_{13} = 0.75 \quad r_{14} = 0.80 \quad r_{23} = 0.70 \quad r_{24} = 0.70 \quad r_{34} = 0.85$$

Hallar la ecuación de regresión de mínimos cuadrados de $X_1$ sobre $X_2$, $X_3$ y $X_4$.

**Solución**

Generalizando el resultado del Problema 15.8, podemos escribir la ecuación de regresión de mínimos cuadrados de $X_1$ sobre $X_2$, $X_3$ y $X_4$ en la forma

$$x_1 = b_{12.34}x_2 + b_{13.24}x_3 + b_{14.23}x_4 \tag{33}$$

donde $b_{12.34}$, $b_{13.24}$ y $b_{14.23}$ pueden obtenerse de las ecuaciones normales

$$\begin{aligned}
\sum x_1x_2 &= b_{12.34}\sum x_2^2 + b_{13.24}\sum x_2x_3 + b_{14.23}\sum x_2x_4 \\
\sum x_1x_3 &= b_{12.34}\sum x_2x_3 + b_{13.24}\sum x_3^2 + b_{14.23}\sum x_3x_4 \\
\sum x_1x_4 &= b_{12.34}\sum x_2x_4 + b_{13.24}\sum x_3x_4 + b_{14.23}\sum x_4^2
\end{aligned} \tag{34}$$

y donde $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$, $x_3 = X_3 - \bar{X}_3$ y $x_4 = X_4 - \bar{X}_4$.

De los datos, deducimos

$$\sum x_2^2 = Ns_2^2 = 5000 \quad \sum x_1x_2 = Ns_1s_2r_{12} = 9000 \quad \sum x_2x_3 = Ns_1s_3r_{23} = 2100$$
$$\sum x_3^2 = Ns_3^2 = 1800 \quad \sum x_1x_3 = Ns_1s_3r_{13} = 4500 \quad \sum x_2x_4 = Ns_2s_4r_{24} = 4200$$
$$\sum x_4^2 = Ns_4^2 = 7200 \quad \sum x_1x_4 = Ns_1s_4r_{14} = 9600 \quad \sum x_3x_4 = Ns_3s_4r_{34} = 3060$$

Poniendo esos resultados en las ecuaciones (34), obtenemos

$$b_{12.34} = 1.3333 \quad b_{13.24} = 0.0000 \quad b_{14.23} = 0.5556 \tag{35}$$

que, al ser sustituidos en (33), dan la ecuación de regresión pedida

$$x_1 = 1.3333x_2 + 0.0000x_3 + 0.5556x_4$$

o sea $$X_1 - 75 = 1.3333(X_2 - 24) + 0.5556(X_4 - 27) \tag{36}$$

es decir $$X_1 = 22.9999 + 1.3333X_2 + 0.5556X_4$$

Una solución exacta de las ecuaciones (34) da $b_{12.34} = \frac{4}{3}$, $b_{13.24} = 0$ y $b_{14.23} = \frac{5}{9}$, así que la ecuación de regresión se puede también escribir como

$$X_1 = 23 + \tfrac{4}{3}X_2 + \tfrac{5}{9}X_4 \tag{37}$$

Es interesante observar que la ecuación de regresión no involucra la nota de inglés $X_3$. Ello no quiere decir que el conocimiento del inglés no tenga peso en el rendimiento en estadística. Más bien, significa que la necesidad del inglés, en lo que concierne a la predicción del rendimiento en estadística, queda ampliamente reflejada en las notas de las restantes materias.

**15.21.** Dos estudiantes obtuvieron en el examen del Problema 15.20 notas respectivas de (*a*) 30 en matemáticas, 18 en inglés y 32 en cultura general y (*b*) 18 en matemáticas, 20 en inglés y 36 en cultura general. ¿Cuál sería la predicción para sus notas en estadística?

**Solución**

(*a*) Sustituyendo $X_2 = 30$, $X_3 = 18$ y $X_4 = 32$ en (37), la predicción de la nota en estadística es $X_1 = 81$.

(*b*) Procediendo como en la parte (*a*) con $X_2 = 18$, $X_3 = 20$ y $X_4 = 36$, vemos que $X_1 = 67$.

**15.22.** Para los datos del Problema 15.20, hallar los coeficientes de correlación parcial (*a*) $r_{12.34}$, (*b*) $r_{13.24}$ y (*c*) $r_{14.23}$.

**Solución**

(*a*) y (*b*)

$$r_{12.4} = \frac{r_{12} - r_{14}r_{24}}{\sqrt{(1-r_{14}^2)(1-r_{24}^2)}} \qquad r_{13.4} = \frac{r_{13} - r_{14}r_{34}}{\sqrt{(1-r_{14}^2)(1-r_{34}^2)}} \qquad r_{23.4} = \frac{r_{23} - r_{24}r_{34}}{\sqrt{(1-r_{24}^2)(1-r_{34}^2)}}$$

Sustituyendo los valores del Problema 15.20, obtenemos $r_{12.4} = 0.7935$, $r_{13.4} = 0.2215$ y $r_{23.4} = 0.2791$. Luego

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{(1-r_{13.4}^2)(1-r_{23.4}^2)}} = 0.7814 \qquad \text{y} \qquad r_{13.24} = \frac{r_{13.4} - r_{12.4}r_{23.4}}{\sqrt{(1-r_{12.4}^2)(1-r_{23.4}^2)}} = 0.0000$$

(*c*)

$$r_{14.3} = \frac{r_{14} - r_{13}r_{34}}{\sqrt{(1-r_{13}^2)(1-r_{34}^2)}} \qquad r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \qquad r_{24.3} = \frac{r_{24} - r_{23}r_{34}}{\sqrt{(1-r_{23}^2)(1-r_{34}^2)}}$$

Sustituyendo los valores del Problema 15.20, obtenemos $r_{14.3} = 0.4664$, $r_{12.3} = 0.7939$ y $r_{24.3} = 0.2791$. Por tanto

$$r_{14.23} = \frac{r_{14.3} - r_{12.3}r_{24.3}}{\sqrt{1-r_{12.3}^2)(1-r_{24.3}^2)}} = 0.4193$$

**15.23.** Interpretar los coeficientes de correlación parcial (*a*) $r_{12.4}$, (*b*) $r_{13.4}$, (*c*) $r_{12.34}$, (*d*) $r_{14.3}$ y (*e*) $r_{14.23}$.

**Solución**

(*a*) $r_{12.4} = 0.7935$ representa el coeficiente de correlación (lineal) entre las notas de estadística y matemáticas para estudiantes con iguales notas en cultura general. Al obtener este coeficiente, las notas en inglés (así como otros factores que no se han tenido en cuenta) no se consideran, como lo evidencia el hecho de que el subíndice 3 se ha omitido.

(*b*) $r_{13.4} = 0.2215$ representa el coeficiente de correlación entre las notas de estadística e inglés para estudiantes con la misma nota en cultura general. Ahora, las notas en matemáticas no se han considerado.

(*c*) $r_{12.34} = 0.7814$ representada el coeficiente de correlación entre las notas de estadística y matemáticas para estudiantes con la misma nota en inglés y en cultura general.

(*d*) $r_{14.3} = 0.4664$ representa el coeficiente de correlación entre las notas de estadística y cultura general para estudiantes con la misma nota en inglés.

(*e*) $r_{14.23} = 0.4193$ representa el coeficiente de correlación entre las notas de estadística y cultura general para estudiantes con iguales notas en matemáticas e inglés.

**15.24.** (*a*) Para los datos del Problema 15.20, mostrar que

$$\frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{(1 - r_{13.4}^2)(1 - r_{23.4}^2)}} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}} \tag{38}$$

(*b*) Explicar el significado de la igualdad en la parte (*a*).

**Solución**

(*a*) El lado izquierdo de (38) se calcula en el Problema 15.22(*a*), con el resultado 0.7814. Para calcular el lado derecho, usamos el Problema 15.22(*c*); de nuevo, resulta 0.7814. Luego la igualdad es válida en este caso especial. Se puede demostrar, por métodos algebraicos directos, que la igualdad es válida en general.

(*b*) El lado izquierdo de (38) es $r_{12.34}$, y el lado derecho es $r_{12.43}$. Como $r_{12.34}$ es la correlación entre $X_1$ y $X_2$ dejando $X_3$ y $X_4$ constantes, mientras que $r_{12.43}$ es la correlación entre $X_1$ y $X_2$ dejando $X_4$ y $X_3$ constantes, salta a la vista por qué es cierta la igualdad.

**15.25.** Para los datos del Problema 15.20, hallar (*a*) el coeficiente de correlación múltiple $R_{1.234}$ y (*b*) el error típico de estimación $S_{1.234}$.

**Solución**

(*a*) $$1 - R_{1.234}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2) \quad \text{o sea} \quad R_{1.234} = 0.9310$$

como $r_{12} = 0.90$ por el Problema 15.20, $r_{14.23} = 0.4193$ por el Problema 15.22(*c*), y

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}} = \frac{0.75 - (0.90)(0.70)}{\sqrt{[1 - (0.90)^2][1 - (0.70)^2]}} = 0.3855$$

*Otro método*

Intercambiando los subíndices 2 y 4 en la primera ecuación se deduce

$$1 - R_{1.234}^2 = (1 - r_{14}^2)(1 - r_{13.4}^2)(1 - r_{12.34}^2) \quad \text{o sea} \quad R_{1.234} = 0.9319$$

donde se ha hecho uso directo de los resultados del Problema 15.22(*a*).

(*b*) $R_{1.234}=\sqrt{\dfrac{1-s_{1.234}^2}{s_1^2}}$ o sea $s_{1.234}=s_1\sqrt{1-R_{1.234}^2}=10\sqrt{1-(0.9310)^2}=3.659$

Comparar con la ecuación (8) de este capítulo.

## PROBLEMAS SUPLEMENTARIOS

### ECUACION DE REGRESION EN TRES VARIABLES

**15.26.** Usando notación de subíndices adecuada, escribir las ecuaciones de regresión (*a*) $X_3$ sobre $X_1$ y $X_2$ y (*b*) $X_4$ sobre $X_1$, $X_2$, $X_3$ y $X_5$.

**15.27.** Escribir las ecuaciones normales correspondientes a la ecuación de regresión de (*a*) $X_2$ sobre $X_1$ y $X_3$ y (*b*) $X_5$ sobre $X_1$, $X_2$, $X_3$ y $X_4$.

**15.28.** La Tabla 15.4 muestra los valores correspondientes de tres variables: $X_1$, $X_2$ y $X_3$.

(*a*) Hallar la ecuación de regresión de mínimos cuadrados de $X_3$ sobre $X_1$ y $X_2$.
(*b*) Estimar $X_3$ cuando $X_1 = 10$ y $X_2 = 6$.

**Tabla 15.4**

| | | | | | | |
|---|---|---|---|---|---|---|
| $X_1$ | 3 | 5 | 6 | 8 | 12 | 14 |
| $X_2$ | 16 | 10 | 7 | 4 | 3 | 2 |
| $X_3$ | 90 | 72 | 54 | 42 | 30 | 12 |

**15.29.** Un profesor de matemáticas desea determinar la relación de las notas del examen final con las de dos parciales anteriores. Llamando $X_1$, $X_2$ y $X_3$ a las notas en el primer parcial, segundo parcial y examen final, efectuó los siguientes cálculos para un total de 120 estudiantes:

$$\bar{X}_1 = 6.8 \qquad \bar{X}_2 = 7.0 \qquad \bar{X}_3 = 74$$
$$s_1 = 1.0 \qquad s_2 = 0.80 \qquad s_2 = 9.0$$
$$r_{12} = 0.60 \qquad r_{13} = 0.70 \qquad r_{23} = 0.65$$

(*a*) Hallar la ecuación de regresión de mínimos cuadrados de $X_3$ sobre $X_1$ y $X_2$.
(*b*) Estimar las notas finales de dos estudiantes cuyas respectivas notas en los parciales fueron (1) 9 y 7 y (2) 4 y 8.

**15.30.** Resolver el Problema 15.8, enunciado anteriormente, escogiendo las variables $X_2$ y $X_3$ tales que $\sum X_2 = \sum X_3 = 0$.

### ERROR TIPICO DE ESTIMACION

**15.31.** Para los datos del Problema 15.28, hallar el error típico de estimación de $X_3$ sobre $X_1$ y $X_2$.

**15.32.** Para los datos del Problema 15.29, hallar el error típico de estimación de (*a*) $X_3$ sobre $X_1$ y $X_2$ y (*b*) $X_1$ sobre $X_2$ y $X_3$.

### COEFICIENTE DE CORRELACION MULTIPLE

**15.33.** Para los datos del Problema 15.28, calcular el coeficiente de correlación múltiple de $X_3$ sobre $X_1$ y $X_2$.

**15.34.** Para los datos del Problema 15.29, calcular (*a*) $R_{3.12}$, (*b*) $R_{1.23}$ y (*c*) $R_{2.13}$.

**15.35.** (*a*) Si $r_{12} = r_{13} = r_{23} = r \neq 1$, mostrar que

$$R_{1.23} = R_{2.31} = R_{3.12} = \frac{r\sqrt{2}}{\sqrt{1+r}}$$

(*b*) Discutir el caso $r = 1$.

**15.36.** Si $R_{1.23} = 0$, probar que $|r_{23}| \geq |r_{12}|$ y $|r_{23}| \geq |r_{13}|$ e interpretar

## CORRELACION PARCIAL

**15.37.** Calcular los coeficientes de correlación parcial lineal (*a*) $r_{12.3}$, (*b*) $r_{13.2}$ y (*c*) $r_{23.1}$ para los datos del Problema 15.28 e interpretar la respuesta.

**15.38.** Rehacer el Problema 15.37 para los datos del Problema 15.29.

**15.39.** Si $r_{12} = r_{13} = r_{23} = r \neq 1$, probar que $r_{12.3} = r_{13.2} = r_{23.1} = r/(1 + r)$. Discutir el caso $r = 1$.

**15.40.** Si $r_{12.3} = 1$, probar que (*a*) $|r_{13.2}| = 1$, (*b*) $|r_{23.1}| = 1$, (*c*) $R_{1.23} = 1$ y (*d*) $s_{1.23} = 0$.

## CORRELACION MULTIPLE Y PARCIAL EN CUATRO O MAS VARIABLES

**15.41.** Probar que la ecuación de regresión de $X_4$ sobre $X_1$, $X_2$ y $X_3$ puede escribirse

$$\frac{x_4}{s_4} = a_1\left(\frac{x_1}{s_1}\right) + a_2\left(\frac{x_2}{s_2}\right) + a_3\left(\frac{x_3}{s_3}\right)$$

donde $a_1$, $a_2$ y $a_3$ vienen determinados al resolver simultáneamente las ecuaciones

$$a_1 r_{11} + a_2 r_{12} + a_3 r_{13} = r_{14}$$
$$a_1 r_{21} + a_2 r_{22} + a_3 r_{23} = r_{24}$$
$$a_1 r_{31} + a_2 r_{32} + a_3 r_{33} = r_{34}$$

y donde $x_j = X_j - \bar{X}_j$, $r_{jj} = 1$ y $j = 1, 2, 3$ y 4. Generalizar al caso de más de cuatro variables.

**15.42.** Dados $\bar{X}_1 = 20$, $\bar{X}_2 = 36$, $\bar{X}_3 = 12$, $\bar{X}_4 = 80$, $s_1 = 1.0$, $s_2 = 2.0$, $s_3 = 1.5$, $s_4 = 6.0$, $r_{12} = -0.20$, $r_{13} = 0.40$, $r_{23} = 0.50$, $r_{14} = 0.40$, $r_{24} = 0.30$ y $r_{34} = -0.10$, (*a*) hallar la ecuación de regresión de $X_4$ sobre $X_1$, $X_2$ y $X_3$ y (*b*) estimar $X_4$ cuando $X_1 = 15$, $X_2 = 40$ y $X_3 = 14$.

**15.43.** Hallar (*a*) $r_{41.23}$, (*b*) $r_{42.13}$ y (*c*) $r_{43.12}$ para los datos del Problema 15.42 e interpretar el resultado.

**15.44.** Para los datos del Problema 15.42, hallar (*a*) $R_{4.123}$ y (*b*) $s_{4.123}$.

**15.45.** Un científico ha coleccionado datos relativos a cuatro variables $T$, $U$, $V$ y $W$. Piensa que una ecuación de la forma $W = aT^bU^cV^d$, donde $a$, $b$, $c$ y $d$ son constantes desconocidas, podría ser válida para determinar $W$ a partir del conocimiento de $T$, $U$ y $V$. Describir un procedimiento por el cual se pueda lograr ese objetivo. [*Ayuda:* Tomar logaritmos en ambos lados de esa ecuación.]

# CAPITULO 16

# Análisis de varianza

## OBJETIVO DEL ANALISIS DE VARIANZA

En el Capítulo 8 hemos usado la teoría del muestreo para contrastar la significación de diferencias entre dos medias muestrales, en el supuesto de que las dos poblaciones de las que se tomaban las muestras tenían la misma varianza. En muchas situaciones es necesario hacer eso mismo con tres o más medias muestrales, o sea, equivalentemente, contrastar la hipótesis de que todas las medias son iguales.

**EJEMPLO 1.** Supongamos que en un experimento agrario, cuatro tratamientos químicos con abonos distintos han producido cosechas medias de trigo de 28, 22, 18 y 24 bushels por acre. ¿Hay diferencia significativa en esas medias o la dispersión se debe simplemente al azar?

Problemas como éste se pueden resolver usando una importante técnica conocida como *análisis de varianza*, desarrollada por Fisher. Hace uso de la distribución $F$ ya considerada en el Capítulo 11.

## EXPERIMENTOS DE FACTOR UNICO

En un *experimento de un factor*, las medidas (u observaciones) se obtienen para $a$ grupos independientes de muestras, donde el número de medidas en cada grupo es $b$. Hablamos de *a tratamientos*, cada uno de los cuales tiene $b$ *repeticiones* o *réplicas*. En el Ejemplo 1, $a = 4$.

Los resultados de un experimento de un factor se pueden presentar en una tabla con $a$ filas y $b$ columnas, como indica la Tabla 16.1. Aquí $X_{jk}$ denota la medida en la $j$-ésima fila y en la $k$-ésima columna, donde $j = 1, 2, ..., a$ y donde $k = 1, 2, ..., b$. Por ejemplo, $X_{35}$ se refiere a la quinta medida para el tercer tratamiento.

**Tabla 16.1**

| | | |
|---|---|---|
| Tratamiento 1 | $X_{11}, X_{12}, ..., X_{1b}$ | $\bar{X}_1$ |
| Tratamiento 2 | $X_{21}, X_{22}, ..., X_{2b}$ | $\bar{X}_2$ |
| ⋮ | ⋮ | ⋮ |
| Tratamiento $a$ | $X_{a1}, X_{a2}, ..., X_{ab}$ | $\bar{X}_a$ |

Denotaremos por $\bar{X}_{j.}$ la media de las medidas en la fila $j$-ésima. Tenemos

$$\bar{X}_{j.} = \frac{1}{b}\sum_{k=1}^{b} X_{jk} \qquad j = 1, 2, ..., a \tag{1}$$

El punto en $\bar{X}_{j.}$ se usa para anunciar que el índice $k$ se ha sumado. Los valores $\bar{X}_{j.}$ se llaman *medias de grupo, medias de tratamiento* o *medias de fila*. La *media global* es la media de todas las medidas en todos los grupos y se denota por $\bar{X}$:

$$\bar{X} = \frac{1}{ab}\sum_{j=1}^{a}\sum_{k=1}^{b} X_{jk} \tag{2}$$

## VARIACION TOTAL, VARIACION DENTRO DE LOS TRATAMIENTOS Y VARIACION ENTRE TRATAMIENTOS

Definimos la *variación total*, denotada por $V$, como la suma de los cuadrados de las desviaciones de cada medida respecto de la media global $\bar{X}$:

$$\text{Variación total} = V = \sum_{j,k} (X_{jk} - \bar{X})^2 \tag{3}$$

Escribiendo la identidad

$$X_{jk} - \bar{X} = (X_{jk} - \bar{X}_{j.}) + (\bar{X}_{j.} - \bar{X}) \tag{4}$$

elevando al cuadrado y sumando en $j$ y $k$, se tiene (Prob. 16.1)

$$\sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 + \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 \tag{5}$$

o sea

$$\sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 + b\sum_{j} (\bar{X}_{j.} - \bar{X})^2 \tag{6}$$

Llamamos a la primera suma de la derecha de (5) y (6) la *variación dentro de los tratamientos* (puesto que implica a los cuadrados de las desviaciones de $X_{jk}$ respecto de las medias de tratamientos $\bar{X}_{j.}$) y la denotamos por $V_W$. Luego

$$V_W = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 \tag{7}$$

La segunda suma del lado derecho de (5) y (6) se llama la *variación entre tratamientos* (ya que involucra a los cuadrados de las desviaciones de las diversas medias de tratamientos $\bar{X}_{j.}$ respecto de la media global $\bar{X}$) y se denota por $V_B$. Así pues,

$$V_B = \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 = b\sum_{j} (\bar{X}_{j} - \bar{X})^2 \tag{8}$$

Las ecuaciones (5) y (6) se pueden expresar, por tanto, como

$$V = V_W + V_B \tag{9}$$

## METODOS ABREVIADOS PARA CALCULAR VARIACIONES

Para minimizar la tarea de calcular las variaciones precedentes, son convenientes las formas siguientes:

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} \tag{10}$$

$$V_B = \frac{1}{b}\sum_j T_{j\cdot}^2 - \frac{T^2}{ab} \tag{11}$$

$$V_W = V - V_B \tag{12}$$

donde $T$ es el total de los valores $X_{jk}$ y $T_{j\cdot}$ es el total de los valores en el tratamiento $j$-ésimo:

$$T = \sum_{j,k} X_{jk} \qquad T_{j\cdot} = \sum_k X_{jk} \tag{13}$$

En la práctica es conveniente restar alguna cantidad fija de todos los datos de la tabla para simplificar los cálculos; tal operación no tiene efecto alguno sobre el resultado final.

## MODELOS MATEMATICOS PARA EL ANALISIS DE VARIANZA

Podemos considerar cada fila de la Tabla 16.1 como una muestra aleatoria de tamaño $b$ de la población para un tratamiento particular. Los $X_{jk}$ diferirán de la media poblacional $\mu_j$ para el tratamiento $j$-ésimo por un *error de azar* o *error aleatorio*, que denotamos por $\varepsilon_{jk}$; así pues

$$X_{jk} = \mu_j + \varepsilon_{jk} \tag{14}$$

Estos errores se suponen normalmente distribuidos con media 0 y varianza $\sigma^2$. Si $\mu$ es la media de la población para todos los tratamientos y hacemos $\alpha_j = \mu_j - \mu$, de manera que $\mu_j = \mu + \alpha_j$, entonces la ecuación (14) se convierte en

$$X_{jk} = \mu + \alpha_j + \varepsilon_{jk} \tag{15}$$

donde $\sum_j \alpha_j = 0$ (véase Prob. 16.9). De la ecuación (15) y de la hipótesis de que los $\varepsilon_{jk}$ están normalmente distribuidos con media 0 y varianza $\sigma^2$, concluimos que los $X_{jk}$ se pueden considerar como variables aleatorias normalmente distribuidas con media $\mu$ y varianza $\sigma^2$.

La hipótesis nula de que todas las medias de los tratamientos son iguales viene dada por ($H_0$: $\alpha_j = 0$; $j = 1, 2, \ldots, a$), o lo que es equivalente, por ($H_0$: $\mu_j = \mu$; $j = 1, 2, \ldots, a$). Si $H_0$ es verdadera, las poblaciones de los tratamientos tendrán todas la misma distribución normal (o sea,

con la misma media y varianza). En tales casos hay sólo una población de tratamiento (o sea, todos los tratamientos son estadísticamente idénticos); en otras palabras, no hay diferencia significativa entre los tratamientos.

## VALORES ESPERADOS DE LAS VARIACIONES

Se puede demostrar (véase Prob. 16.10) que los valores esperados de $V_W$, $V_B$ y $V$ vienen dados por

$$E(V_W) = a(b - 1)\sigma^2 \tag{16}$$

$$E(V_B) = (a - 1)\sigma^2 + b\sum_j \alpha_j^2 \tag{17}$$

$$E(V) = (ab - 1)\sigma^2 + b\sum_j \alpha_j^2 \tag{18}$$

De la ecuación (16) se deduce que

$$E\left[\frac{V_W}{a(b - 1)}\right] = \sigma^2 \tag{19}$$

luego

$$\hat{S}_W^2 = \frac{V_W}{a(b - 1)} \tag{20}$$

es siempre una estimación óptima (no sesgada) de $\sigma^2$ independientemente de que $H_0$ sea verdadera o no. Por otro lado, vemos de (16) y (18) que sólo si $H_0$ es verdadera (o sea, $\alpha_j = 0$) tendremos

$$E\left(\frac{V_B}{a - 1}\right) = \sigma^2 \quad \text{y} \quad E\left(\frac{V}{ab - 1}\right) = \sigma^2 \tag{21}$$

así que sólo en tal circunstancia proporcionan

$$\hat{S}_B^2 = \frac{V_B}{a - 1} \quad \text{y} \quad \hat{S}^2 = \frac{V}{ab - 1} \tag{22}$$

estimaciones sin sesgo de $\sigma^2$. Si $H_0$ es falsa, sin embargo, tenemos de la ecuación (16) que

$$E(\hat{S}_B^2) = \sigma^2 + \frac{b}{a - 1}\sum_j \alpha_j^2 \tag{23}$$

## DISTRIBUCIONES DE LAS VARIACIONES

Usando la propiedad aditiva de *ji*-cuadrado (página 272, podemos probar los siguientes teoremas fundamentales sobre las distribuciones de las variaciones $V_W$, $V_B$ y $V$:

**TEOREMA 1.** $V_W/\sigma^2$ tiene distribución *ji*-cuadrado con $a(b-1)$ grados de libertad.

**TEOREMA 2.** Bajo la hipótesis nula $H_0$, $V_B/\sigma^2$ y $V/\sigma^2$ tiene distribución *ji*-cuadrado con $a-1$ y $ab-1$ grados de libertad, respectivamente.

Es importante recalcar que el Teorema 1 es válido independientemente de que se suponga $H_0$ o no, mientras que el Teorema 2 es válido sólo cuando se supone $H_0$.

## EL CONTRASTE $F$ PARA LA HIPOTESIS NULA DE IGUALDAD DE MEDIAS

Si la hipótesis nula $H_0$ es falsa (o sea, si las medias de los tratamientos no son iguales), vemos de (23) que cabe esperar que $\hat{S}_B^2$ sea mayor que $\sigma^2$, con el efecto tanto más pronunciado cuanto mayor sea la discrepancia entre las medias. Por otra parte, de (19) y (20) cabe esperar que $\hat{S}_W^2$ sea igual a $\sigma^2$ independientemente de que las medias sean o no iguales. Deducimos que un buen estadístico para contrastar $H_0$ viene dado por $\hat{S}_B^2/\hat{S}_W^2$. Si este estadístico es significativamente grande, podemos concluir que hay una diferencia significativa entre las medias de los tratamientos y podemos, por tanto, rechazar $H_0$; en caso contrario, podemos ya sea aceptar $H_0$ o reservar la decisión, pendiente de posteriores análisis adicionales.

Para usar el estadístico $\hat{S}_B^2/\hat{S}_W^2$, debemos conocer su distribución muestral. Esto lo proporciona el Teorema 3.

**TEOREMA 3.** El estadístico $F = \hat{S}_B^2/\hat{S}_W^2$ tiene distribución $F$ con $a-1$ y $a(b-1)$ grados de libertad.

El Teorema 3 nos capacita para contrastar la hipótesis nula a algún nivel de significación especificado mediante un contraste unilateral con la distribución $F$ (Cap. 11).

## TABLAS DE ANALISIS DE VARIANZA

Los cálculos que requiere el contraste anterior se resumen en la Tabla 16.2, que se llama una *tabla de análisis de varianza*. En la práctica, calcularíamos $V$ y $V_B$ por el método largo [ecuaciones (3) y (8)] o por el método corto [ecuaciones (10) y (11)], calculando después $V_W = V - V_B$. Hagamos notar que los grados de libertad para la variación total (o sea, $ab-1$) son igual a la suma de los grados de libertad para las variaciones dentro de los tratamientos y las variaciones entre tratamientos.

**Tabla 16.2**

| Variación | Grados de libertad | Cuadrado medio | $F$ |
|---|---|---|---|
| Entre tratamientos, $V_B = b\sum_j (\bar{X}_{j\cdot} - \bar{X})^2$ | $a-1$ | $\hat{S}_B^2 = \dfrac{V_B}{a-1}$ | $\dfrac{\hat{S}_B^2}{\hat{S}_W^2}$ con $a-1$ y $a(b-1)$ grados de libertad |
| Dentro de los tratamientos, $V_W = V - V_B$ | $a(b-1)$ | $\hat{S}_W^2 = \dfrac{V_W}{a(b-1)}$ | |
| Total, $V = V_B + V_W = \sum_{j,k} (X_{jk} - \bar{X})^2$ | $ab-1$ | | |

## MODIFICACIONES PARA NUMEROS DISTINTOS DE OBSERVACIONES

Si los tratamientos 1, ..., $a$ tienen diferentes números de observaciones, iguales a $N_1$, ..., $N_a$, respectivamente, los resultados anteriores se modifican sin dificultad y se obtiene

$$V = \sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} X_{jk}^2 - \frac{T^2}{N} \tag{24}$$

$$V_B = \sum_{j,k} (\bar{X}_{j\cdot} - \bar{X})^2 = \sum_{j} N_j(\bar{X}_{j\cdot} - \bar{X})^2 = \sum_{j} \frac{T_{j\cdot}^2}{N_j} - \frac{T^2}{N} \tag{25}$$

$$V_W = V - V_B \tag{26}$$

donde $\sum_{j,k}$ denota la suma sobre $k$ desde 1 hasta $N_j$ y después la suma sobre $j$ desde 1 hasta $a$. La Tabla 16.3 es la tabla del análisis de varianza para este caso.

**Tabla 16.3**

| Variación | Grados de libertad | Cuadrado medio | $F$ |
|---|---|---|---|
| Entre tratamientos, $V_B = \sum_j N_j(\bar{X}_{j\cdot} - \bar{X})^2$ | $a - 1$ | $\hat{S}_B^2 = \dfrac{V_B}{a-1}$ | $\dfrac{\hat{S}_B^2}{\hat{S}_W^2}$ con $a - 1$ y $N - a$ grados de libertad |
| Dentro de los tratamientos, $V_W = V - V_B$ | $N - a$ | $\hat{S}_W^2 = \dfrac{V_W}{N-a}$ | |
| Total, $V = V_B + V_W = \sum_{j,k} (X_{jk} - \bar{X})^2$ | $N - 1$ | | |

## EXPERIMENTOS DE DOS FACTORES

Las ideas del análisis de varianza para un solo factor, pueden generalizarse a *experimentos de dos factores*, tal como ilustra el Ejemplo 2.

**EJEMPLO 2.** Supongamos que en un experimento agrario se examina la producción por acre de 4 variedades de trigo, cada una sembrada en 5 parcelas de terreno. Se necesitan en total 20 parcelas. Conviene, en tal caso, combinarlas en bloques, digamos 4 por bloque, con una variedad distinta de trigo en cada una de ellas dentro de un bloque. Eso requiere 5 bloques.

En este caso hay dos factores, ya que puede haber diferencias en la producción por acre debidas a (1) la variedad de trigo elegida y (2) el bloque particular usado (por distinta fertilidad del terreno, etc.).

Por analogía con el Ejemplo 2, nos referimos con frecuencia a los dos factores de un experimento como *tratamientos* y *bloques*, pero naturalmente podíamos llamarlos simplemente factor 1 y factor 2.

## NOTACION PARA EXPERIMENTOS DE DOS FACTORES

Si hay $a$ tratamientos y $b$ bloques, construimos la Tabla 16.4, donde se supone que hay un valor experimental (tal como producción por acre) correspondiente a cada tratamiento y bloque. Para el tratamiento $j$ y el bloque $k$, lo denotamos por $X_{jk}$. La media de las entradas de la fija $j$-ésima se denota por $\bar{X}_{j.}$, donde $j = 1, ..., a$, mientras la media de las entradas de la columna $k$-ésima se denota $\bar{X}_{.k}$, donde $k = 1, ..., b$. La media global se denota por $\bar{X}$. En símbolos,

$$\bar{X}_{j.} = \frac{1}{b}\sum_{k=1}^{b} X_{jk} \qquad \bar{X}_{.k} = \frac{1}{a}\sum_{j=1}^{a} X_{jk} \qquad \bar{X} = \frac{1}{ab}\sum_{j,k} X_{jk} \tag{27}$$

**Tabla 16.4**

| | Bloque | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | ... | $b$ | |
| Tratamiento 1 | $X_{11}$ | $X_{12}$ | ... | $X_{1b}$ | $\bar{X}_{1.}$ |
| Tratamiento 2 | $X_{21}$ | $X_{22}$ | ... | $X_{2b}$ | $\bar{X}_{2.}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Tratamiento $a$ | $X_{a1}$ | $X_{a2}$ | ... | $X_{ab}$ | $\bar{X}_{a.}$ |
| | $\bar{X}_{.1}$ | $\bar{X}_{.2}$ | | $\bar{X}_{.b}$ | |

## VARIACIONES PARA EXPERIMENTOS DE DOS FACTORES

Como en el caso de experimentos de un factor, podemos definir variaciones para experimentos de dos factores. Definimos primero la *variación total*, como en la ecuación (3), a saber

$$V = \sum_{j,k} (X_{jk} - \bar{X})^2 \tag{28}$$

Escribiendo la identidad

$$X_{jk} - \bar{X} = (X_{jk} - \bar{X}_{j.} - \bar{X}_{.k} + \bar{X}) + (\bar{X}_{j.} - \bar{X}) + (\bar{X}_{.k} - \bar{X}) \tag{29}$$

elevando ahora al cuadrado y sumando sobre $j$ y $k$, se ve que

$$V = V_E + V_R + V_C \tag{30}$$

donde

$V_E$ = variación debida a error o azar $= \sum_{j,k} (X_{jk} - \bar{X}_{j.} - \bar{X}_{.k} + \bar{X})^2$

$V_R$ = variación entre filas (tratamientos) $= b\sum_{j=1}^{a} (\bar{X}_{j.} - \bar{X})^2$

$V_C$ = variación entre columnas (bloques) $= a\sum_{k=1}^{b} (\bar{X}_{.k} - \bar{X})^2$

La variación debida al error aleatorio se conoce como *variación residual* o *aleatoria*.

Las que siguen, análogas a las ecuaciones (10), (11) y (12), son fórmulas abreviadas para el cálculo:

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} \tag{31}$$

$$V_R = \frac{1}{b} \sum_{j=1}^{a} T_{j.}^2 - \frac{T^2}{ab} \tag{32}$$

$$V_C = \frac{1}{a} \sum_{k=1}^{b} T_{.k}^2 - \frac{T^2}{ab} \tag{33}$$

$$V_E = V - V_R - V_C \tag{34}$$

donde $T_{j.}$ es el total de las entradas en la fija $j$-ésima, $T_{.k}$ es el total de entradas en la columna $k$-ésima, y $T$ el total de las entradas.

## ANALISIS DE VARIANZA PARA EXPERIMENTOS DE DOS FACTORES

La generalización del modelo matemático para experimentos de un factor dado por (15) nos lleva a suponer para experimentos de dos factores que

$$X_{jk} = \mu + \alpha_j + \beta_k + \varepsilon_{jk} \tag{35}$$

donde $\sum \alpha_j = 0$ y $\sum \beta_k = 0$. Aquí $\mu$ es la media global de la población, $\alpha_j$ es la parte de $X_{jk}$ debida a los diferentes tratamientos (llamados *efectos de los tratamientos*), $\beta_k$ la parte de $X_{jk}$ debida a los diferentes bloques (*efectos de los bloques*) y $\varepsilon_{jk}$ es la parte debida a error o azar. Como antes, suponemos que los $\varepsilon_{jk}$ están normalmente distribuidos con media 0 y varianza $\sigma^2$, así que los $X_{jk}$ también están normalmente distribuidos con media $\mu$ y varianza $\sigma^2$.

Correspondientes a los resultados (16), (17) y (18), podemos probar que las esperanzas de las variaciones vienen dadas por

$$E(V_E) = (a - 1)(b - 1)\sigma^2 \tag{36}$$

$$E(V_R) = (a - 1)\sigma^2 + b \sum_j \alpha_j^2 \tag{37}$$

$$E(V_C) = (b - 1)\sigma^2 + a \sum_k \beta_k^2 \tag{38}$$

$$E(V) = (ab - 1)\sigma^2 + b \sum_j \alpha_j^2 + a \sum_k \beta_k^2 \tag{39}$$

Hay dos hipótesis nulas que querríamos contrastar:

$H_0^{(1)}$: Todos los tratamientos (fila) tienen la misma media; o sea, $\alpha_j = 0$ y $j = 1, \ldots, a$.

$H_0^{(2)}$: Todos los bloques (columna) tienen la misma media; es decir, $\beta_k = 0$ y $k = 1, \ldots, b$.

Vemos de (38) que, independientemente de $H_0^{(1)}$ o $H_0^{(2)}$, una estimación óptima (sin sesgo) de $\sigma^2$ la da

$$\hat{S}_E^2 = \frac{V_E}{(a-1)(b-1)} \quad \text{es decir,} \quad E(\hat{S}_E^2) = \sigma^2 \tag{40}$$

Además, si las hipótesis $H_0^{(1)}$ y $H_0^{(2)}$ son verdaderas, entonces

$$\hat{S}_R^2 = \frac{V_R}{a-1} \qquad \hat{S}_C^2 = \frac{V_C}{b-1} \qquad \hat{S}^2 = \frac{V}{ab-1} \tag{41}$$

serán estimaciones sin sesgo de $\sigma^2$. Si $H_0^{(1)}$ y $H_0^{(2)}$ son falsas, no obstante, de las ecuaciones (36) y (37), respectivamente, tendremos

$$E(\hat{S}_R^2) = \sigma^2 + \frac{b}{a-1}\sum_j \alpha_j^2 \tag{42}$$

$$E(\hat{S}_C^2) = \sigma^2 + \frac{a}{b-1}\sum_k \beta_k^2 \tag{43}$$

Los siguientes teoremas son similares a los Teoremas 1 y 2:

**TEOREMA 4.** $V_E/\sigma^2$ tiene una distribución *ji*-cuadrado con $(a-1)(b-1)$ grados de libertad, independientemente de $H_0^{(1)}$ o $H_0^{(2)}$.

**TEOREMA 5.** Bajo la hipótesis $H_0^{(1)}$, $V_R/\sigma^2$ tiene una distribución *ji*-cuadrado con $a-1$ grados de libertad. Bajo $H_0^{(2)}$, $V_C/\sigma^2$ tiene una distribución *ji*-cuadrado con $b-1$ grados de libertad. Bajo ambas hipótesis, $H_0^{(1)}$ y $H_0^{(2)}$, $V/\sigma^2$ tiene una distribución *ji*-cuadrado con $ab-1$ grados de libertad.

Para contrastar la hipótesis $H_0^{(1)}$, es natural considerar el estadístico $\hat{S}_R^2/\hat{S}_E^2$ ya que podemos ver de la ecuación (42) que $\hat{S}_R^2$ se espera que difiera significativamente de $\sigma^2$ si las medias de fila (tratamiento) son significativamente diferentes. Análogamente, para contrastar $H_0^{(2)}$, consideramos el estadístico $\hat{S}_C^2/\hat{S}_E^2$. Las distribuciones de $\hat{S}_R^2/\hat{S}_E^2$ y $\hat{S}_C^2/\hat{S}_E^2$ vienen dadas por el Teorema 6. que es análogo al Teorema 3.

**TEOREMA 6.** Bajo la hipótesis $H_0^{(1)}$, el estadístico $\hat{S}_R^2/\hat{S}_E^2$ tiene una distribución $F$ con $a-1$ y $(a-1)(b-1)$ grados de libertad. Bajo la hipótesis $H_0^{(2)}$, el estadístico $\hat{S}_C^2/\hat{S}_E^2$ tiene una distribución $F$ con $b-1$ y $(a-1)(b-1)$ grados de libertad.

El Teorema 6 nos capacita para aceptar o rechazar $H_0^{(1)}$ o $H_0^{(2)}$ a niveles de significación específicos. Por conveniencia, como en el caso de experimentos de un factor, se puede construir una tabla de análisis de varianza, como indica la Tabla 16.5.

## EXPERIMENTOS DE DOS FACTORES CON REPETICION

En la Tabla 16.4 hay sólo una entrada correspondiente a un tratamiento y un bloque dados. Se puede obtener más información acerca de los factores repitiendo el experimento, un proceso

**Tabla 16.5**

| Variación | Grados de libertad | Cuadrado medio | $F$ |
|---|---|---|---|
| Entre tratamientos, $V_R = b\sum_j (\bar{X}_{j.} - \bar{X})^2$ | $a - 1$ | $\hat{S}_R^2 = \dfrac{V_R}{a - 1}$ | $\hat{S}_R^2/\hat{S}_E^2$ con $a - 1$ y $(a - 1)(b - 1)$ grados de libertad |
| Entre bloques, $V_C = a\sum_k (\bar{X}_{.k} - \bar{X})^2$ | $b - 1$ | $\hat{S}_C^2 = \dfrac{V_C}{b - 1}$ | $\hat{S}_C^2/\hat{S}_E^2$ con $b - 1$ y $(a - 1)(b - 1)$ grados de libertad |
| Residual o aleatoría, $V_E = V - V_R - V_C$ | $(a - 1)(b - 1)$ | $\hat{S}_E^2 = \dfrac{V_E}{(a - 1)(b - 1)}$ | |
| Total, $V = V_R + V_C + V_E = \sum_{j,k} (X_{jk} - \bar{X})^2$ | $ab - 1$ | | |

llamado *repetición*. En tal caso habrá más de una entrada correspondiente a un tratamiento y a un bloque dados. Supondremos que hay $c$ entradas para toda posición; cuando los números de repeticiones no son iguales han de hacerse las modificaciones pertinentes.

A causa de la repetición, se debe usar un modelo apropiado para sustituir el dado por la ecuación (35). Usaremos

$$X_{jkl} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{jkl} \tag{44}$$

donde los subíndices $j$, $k$ y $l$ de $X_{jkl}$ corresponden a la fila $j$-ésima (o tratamiento), la $k$-ésima columna (o bloque) y la $l$-ésima repetición, respectivamente. En la ecuación (44) los $\mu$, $\alpha_j$ y $\beta_k$ se definen como antes; $\varepsilon_{jkl}$ es un término de azar o error, mientras que los $\gamma_{jk}$ denotan los *efectos de interacción* fila-columna (o sea, tratamiento-bloque), llamados a menudo *interacciones*. Tenemos las restricciones

$$\sum_j \alpha_j = 0 \qquad \sum_k \beta_k = 0 \qquad \sum_j \gamma_{jk} = 0 \qquad \sum_k \gamma_{jk} = 0 \tag{45}$$

y los $X_{jkl}$ se suponen normalmente distribuidos con media $\mu$ y varianza $\sigma^2$.

Como antes, la variación total $V$ de todos los datos se puede romper en variaciones debidas a filas $V_R$, columnas $V_C$, interacción $V_I$ y error residual o aleatorio $V_E$:

$$V = V_R + V_C + V_I + V_E \tag{46}$$

donde

$$V = \sum_{j,k,l} (X_{jkl} - \bar{X})^2 \tag{47}$$

$$V_R = bc \sum_{j=1}^{a} (\bar{X}_{j..} - \bar{X})^2 \tag{48}$$

$$V_C = ac \sum_{k=1}^{b} (\bar{X}_{.k.} - \bar{X})^2 \tag{49}$$

$$V_I = c \sum_{j,k} (\bar{X}_{jk.} - \bar{X}_{j..} - \bar{X}_{.k.} + \bar{X})^2 \tag{50}$$

$$V_E = \sum_{j,k,l} (X_{jkl} - \bar{X}_{jk.})^2 \tag{51}$$

En estos resultados los puntos en los subíndices tienen significados análogos a los antes citados (página 375); así, por ejemplo,

$$\bar{X}_{j..} = \frac{1}{bc} \sum_{k,l} X_{jkl} = \frac{1}{b} \sum_{k} \bar{X}_{jk.} \tag{52}$$

Los valores esperados de las variaciones se hallan como antes. Usando el número apropiado de grados de libertad para cada fuente de variación, podemos establecer la tabla del análisis de varianza como indica la Tabla 16.6. Los $F$-cocientes en la última columna de esa tabla se pueden utilizar para contrastar las hipótesis nula:

$H_0^{(1)}$: Todas las medias de tratamiento (fila) son iguales; esto es, $\alpha_j = 0$.

$H_0^{(2)}$: Todas las medias de bloque (columna) son iguales; o sea, $\beta_k = 0$.

$H_0^{(3)}$: No hay interacciones entre tratamientos y bloques, es decir, $\gamma_{jk} = 0$.

**Tabla 16.6**

| Variación | Grados de libertad | Cuadrado medio | $F$ |
|---|---|---|---|
| Entre tratamientos, $V_R$ | $a - 1$ | $\hat{S}_R^2 = \dfrac{V_R}{a - 1}$ | $\hat{S}_R^2/\hat{S}_E^2$ con $a - 1$ y $ab(c - 1)$ grados de libertad |
| Entre bloques, $V_C$ | $b - 1$ | $\hat{S}_C^2 = \dfrac{V_C}{b - 1}$ | $\hat{S}_C^2/\hat{S}_E^2$ con $b - 1$ y $ab(c - 1)$ grados de libertad |
| Interacción, $V_I$ | $(a - 1)(b - 1)$ | $\hat{S}_I^2 = \dfrac{V_I}{(a - 1)(b - 1)}$ | $\hat{S}_I^2/\hat{S}_E^2$ con $(a - 1)(b - 1)$ y $ab(c - 1)$ grados de libertad |
| Residual o aleatoria, $V_E$ | $ab(c - 1)$ | $\hat{S}_E^2 = \dfrac{V_E}{ab(c - 1)}$ | |
| Total, $V$ | $abc - 1$ | | |

Desde un punto de vista práctico debemos decidir primero si $H_0^{(3)}$ puede ser rechazada o no a un nivel de significación apropiado, usando el $F$-cociente $\hat{S}_I^2/\hat{S}_E^2$ de la Tabla 16.6. Dos casos son posibles:

1. $H_0^{(3)}$ *no se puede rechazar*. En este caso podemos concluir que las interacciones no son demasiado grandes. Podemos entonces contrastar $H_0^{(1)}$ y $H_0^{(2)}$ usando los $F$-cocientes $\hat{S}_R^2/\hat{S}_E^2$ y $\hat{S}_C^2/\hat{S}_E^2$, respectivamente, como se muestra en la Tabla 16.6. Algunos estadísticos recomiendan tomar el total de $V_I + V_E$ y dividirlo por el total correspondiente de grados de libertad $(a-1)(b-1) + ab(c-1)$ y usar este valor como sustituto del denominador $\hat{S}_E^2$ en $F$ test.
2. $H_0^{(3)}$ *puede ser rechazada*. En este caso podemos concluir que las interacciones son significativamente grandes. Diferencias en los factores serían entonces importantes sólo si fueran grandes comparadas con tales interacciones. Por esta razón muchos estadísticos recomiendan contrastar $H_0^{(1)}$ y $H_0^{(2)}$ mediante los $F$-cocientes $\hat{S}_R^2/\hat{S}_I^2$ y $\hat{S}_C^2/\hat{S}_I^2$ más bien que con los de la Tabla 16.6. Nosotros usaremos también aquí este procedimiento alternativo.

El análisis de varianza con repetición se realiza de forma sencilla totalizando primero los valores de repetición que corresponden a tratamientos (filas) y bloques (columnas) particulares. Esto produce una tabla de dos factores con entradas únicas, que puede analizarse como en la Tabla 16.5. Este procedimiento se ilustra en el Problema 16.16.

## DISEÑO EXPERIMENTAL

Las técnicas del análisis de varianza discutidas hasta ahora se emplean una vez que se han obtenido los resultados de un experimento. Sin embargo, con el fin de adquirir cuanta información sea posible, el diseño de un experimento debe planificarse cuidadosamente; eso se conoce como el *diseño del experimento*. He aquí varios ejemplos importantes de diseño experimental:

1. **Aleatorización completa.** Supongamos que tenemos un experimento agrario como el del Ejemplo 1. Para su diseño, debemos dividir el campo en $4 \times 4 = 16$ parcelas (indicadas en la Figura 16.1 por cuadrados, aunque se puede usar cualquier forma) y asignar cada tratamiento (indicado por $A$, $B$, $C$ y $D$) a cuatro bloques elegidos completamente al azar. El objetivo de la aleatorización completa es eliminar varias fuentes de error, tales como la fertilidad del suelo.

| | | | |
|---|---|---|---|
| $D$ | $A$ | $C$ | $C$ |
| $B$ | $D$ | $B$ | $A$ |
| $D$ | $C$ | $B$ | $D$ |
| $A$ | $B$ | $C$ | $A$ |

Aleatorización completa

**Figura 16.1.**

| | | | | |
|---|---|---|---|---|
| $I$ | $C$ | $B$ | $A$ | $D$ |
| $II$ | $A$ | $B$ | $D$ | $C$ |
| $III$ | $B$ | $C$ | $D$ | $A$ |
| $IV$ | $A$ | $D$ | $C$ | $B$ |

Bloques aleatorizados

**Figura 16.2.**

| | | | |
|---|---|---|---|
| $D$ | $B$ | $C$ | $A$ |
| $B$ | $D$ | $A$ | $C$ |
| $C$ | $A$ | $D$ | $B$ |
| $A$ | $C$ | $B$ | $D$ |

Cuadrado latino

**Figura 16.3.**

| | | | |
|---|---|---|---|
| $B_\gamma$ | $A_\beta$ | $D_\delta$ | $C_\alpha$ |
| $A_\delta$ | $B_\alpha$ | $C_\gamma$ | $D_\beta$ |
| $D_\alpha$ | $C_\delta$ | $B_\beta$ | $A_\gamma$ |
| $C_\beta$ | $D_\gamma$ | $A_\alpha$ | $B_\delta$ |

Cuadrado greco-latino

**Figura 16.4.**

2. **Bloques aleatorios.** Cuando, como en el Ejemplo 2, es necesario tener un conjunto completo de tratamientos para cada bloque, los tratamientos $A$, $B$, $C$ y $D$ se introducen en orden aleatorio dentro de cada bloque: I, II, III y IV (o sea, las filas en la Fig. 16.2), y por esa razón se habla de los bloques como *bloques aleatorios*. Este tipo de diseño se usa cuando se desea controlar *una fuente de error o variabilidad*: a saber, la diferencia en bloques.
3. **Cuadrados latinos.** Para algunos propósitos es preciso controlar *dos fuentes de error o variabilidad* al mismo tiempo, tales como la diferencia en filas y la diferencia en columnas. Así, en el experimento del Ejemplo 1, errores en diferentes filas y columnas podrían ser debidos a cambios en la fertilidad en diferentes partes del campo. En tal caso es deseable que cada tratamiento ocurra una vez en cada fila y una vez en cada columna, como en la Figura 16.3. Esa disposición se llama un *cuadrado latino* por cuanto se usan las letras latinas $A$, $B$, $C$ y $D$.
4. **Cuadrados greco-latinos.** Si es necesario controlar *tres fuentes de error o variabilidad*, se usa un *cuadrado greco-latino* como el que muestra la Figura 16.4. Tal cuadrado es esencialmente como un par de cuadrados latinos unidos, con letras unidas $A$, $B$, $C$ y $D$ para uno y griegas $\beta$, $\gamma$ y $\delta$ para el otro. El requisito adicional que deben satisfacer es que cada letra latina ha de usarse una y sólo una vez con cada letra griega; cuando ese requisito se cumple, el cuadrado se dice *ortogonal*.

## PROBLEMAS RESUELTOS

### EXPERIMENTOS DE UN FACTOR

**16.1.** Probar que $V = V_W + V_B$; esto es

$$\sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 + \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2$$

**Solución**

Tenemos $$X_{jk} - \bar{X} = (X_{jk} - \bar{X}_{j.}) + (\bar{X}_{j.} - \bar{X})$$

Entonces, elevando al cuadrado y sumando en $j$ y $k$, obtenemos

$$\sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 + \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 + 2\sum_{j,k} (X_{jk} - \bar{X}_{j.})(\bar{X}_{j.} - \bar{X})$$

Para probar el resultado pedido, debemos mostrar que la última suma es cero. Para ello, procedemos como sigue:

$$\sum_{j,k} (X_{jk} - \bar{X}_{j.})(\bar{X}_{j.} - \bar{X}) = \sum_{j=1}^{a} (\bar{X}_{j.} - \bar{X}) \left[ \sum_{k=1}^{b} (X_{jk} - \bar{X}_{j.}) \right]$$

$$= \sum_{j=1}^{a} = (\bar{X}_{j.} - \bar{X}) \left[ \left( \sum_{k=1}^{b} X_{jk} \right) - b\bar{X}_{j.} \right] = 0$$

ya que $$\bar{X}_{j.} = \frac{1}{b} \sum_{k=1}^{b} X_{jk}$$

**16.2.** Comprobar que (*a*) $T = ab\bar{X}$, (*b*) $T_{j.} = b\bar{X}_{j.}$ y (*c*) $\sum_j T_{j.} = ab\bar{X}$, usando la notación de la página 376.

**Solución**

(*a*)
$$T = \sum_{j,k} X_{jk} = ab\left(\frac{1}{b}\sum_{j,k} X_{jk}\right) = ab\bar{X}$$

(*b*)
$$T_{j.} = \sum_{k} X_{jk} = b\left(\frac{1}{b}\sum_{k} X_{jk}\right) = b\bar{X}_{j.}$$

(*c*) Como $T_{j.} = \sum_k X_{jk}$, por la parte (*a*) se tiene

$$\sum_j T_{j.} = \sum_j \sum_k X_{jk} = T = ab\bar{X}$$

**16.3.** Verificar las fórmulas (10), (11) y (12) de este capítulo.

**Solución**

Tenemos

$$\begin{aligned} V &= \sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} (X_{jk}^2 - 2\bar{X}X_{jk} + \bar{X}^2) \\ &= \sum_{j,k} X_{jk}^2 - 2\bar{X}\sum_{j,k} X_{jk} + ab\bar{X}^2 \\ &= \sum_{j,k} X_{jk}^2 - 2\bar{X}(ab\bar{X}) + ab\bar{X}^2 \\ &= \sum_{j,k} X_{jk}^2 - ab\bar{X}^2 \\ &= \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} \end{aligned}$$

usando el Problema 16.2(*a*) en la tercera y en la última línea. De igual modo,

$$\begin{aligned} V_B &= \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 = \sum_{j,k} (\bar{X}_{j.}^2 - 2\bar{X}\bar{X}_{j.} + \bar{X}^2) \\ &= \sum_{j,k} \bar{X}_{j.}^2 - 2\bar{X}\sum_{j,k} \bar{X}_{j.} + ab\bar{X}^2 \\ &= \sum_{j,k} \left(\frac{T_{j.}}{b}\right)^2 - 2\bar{X}\sum_{j,k} \frac{T_{j.}}{b} + ab\bar{X}^2 \\ &= \frac{1}{b^2}\sum_{j=1}^{a}\sum_{k=1}^{b} T_{j.}^2 - 2\bar{X}(ab\bar{X}) + ab\bar{X}^2 \\ &= \frac{1}{b}\sum_{j=1}^{a} T_{j.}^2 - ab\bar{X}^2 \\ &= \frac{1}{b}\sum_{j=1}^{a} T_{j.}^2 - \frac{T^2}{ab} \end{aligned}$$

usando el Problema 16.2(*b*) en la tercera línea y el Problema 16.2(*a*) en la última. Finalmente, la ecuación (12) se sigue de que $V = V_W + V_B$, o sea $V_W = V - V_B$.

**16.4.** La Tabla 16.7 da las producciones por acre de una cierta variedad de trigo que crece en terrenos tratados con fertilizantes $A$, $B$ y $C$. Hallar (*a*) las producciones medias para los diferentes tratamientos, (*b*) la media global para todos los tratamientos, (*c*) la variación total, (*d*) la variación entre tratamientos y (*e*) la variación dentro de los tratamientos. Usar el método largo.

**Tabla 16.7**

| | | | | |
|---|---|---|---|---|
| $A$ | 48 | 49 | 50 | 49 |
| $B$ | 47 | 49 | 48 | 48 |
| $C$ | 49 | 51 | 50 | 50 |

**Tabla 16.8**

| | | | |
|---|---|---|---|
| 3 | 4 | 5 | 4 |
| 2 | 4 | 3 | 3 |
| 4 | 6 | 5 | 5 |

**Solución**

Para simplificar la aritmética, podemos restar 45 a todos los datos sin que ello afecte a los valores de las variaciones. Entonces obtenemos los datos de la Tabla 16.8.

(*a*) Las medias de tratamiento (fila) para la Tabla 16.8 vienen dadas por

$$\bar{X}_{1.} = \tfrac{1}{4}(3+4+5+4) = 4 \qquad \bar{X}_{2.} = \tfrac{1}{4}(2+4+3+3) = 3 \qquad \bar{X}_{3.} = \tfrac{1}{4}(4+6+5+5) = 5$$

Luego las producciones medias, obtenidas añadiendo 45 a éstas, son de 49, 48 y 50 bushels por acre para $A$, $B$ y $C$, respectivamente.

(*b*) La media global para todos los tratamientos es

$$\bar{X} = \tfrac{1}{12}(3+4+5+4+2+4+3+3+4+6+5+5) = 4$$

Así que la media global para los datos originales es 45 + 4 = 49 bushels por acre.

(*c*) La variación es

$$V = \sum_{j,k} (X_{jk} - X)^2 = (3-4)^2 + (4-4)^2 + (5-4)^2 + (4-4)^2 + (2-4)^2 + (4-4)^2 +$$
$$+ (3-4)^2 + (3-4)^2 + (4-4)^2 + (6-4)^2 + (5-4)^2 + (5-4)^2 = 14$$

(*d*) La variación entre tratamientos es

$$V_B = b \sum_{j} (\bar{X}_{j.} - \bar{X})^2 = 4[(4-4)^2 + (3-4)^2 + (5-4)^2] = 8$$

(*e*) La variación dentro de los tratamientos es

$$V_W = V - V_B = 14 - 8 = 6$$

*Otro método*

$$V_W = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 = (3-4)^2 + (4-4)^2 + (5-4)^2 + (4-4)^2 + (2-3)^2 + (4-3)^2 +$$
$$+ (3-3)^2 + (3-3)^2 + (4-5)^2 + (6-5)^2 + (5-5)^2 + (5-5)^2 = 6$$

*Nota:* La Tabla 16.9 es la tabla de análisis de varianza para los Problemas 16.4, 16.5 y 16.6.

**Tabla 16.9**

| Variación | Grados de libertad | Cuadrado medio | $F$ |
|---|---|---|---|
| Entre tratamientos, $V_B = 8$ | $a - 1 = 2$ | $\hat{S}_B^2 = \frac{8}{2} = 4$ | $\frac{\hat{S}_B^2}{\hat{S}_W^2} = \frac{4}{2/3} = 6$ con 2 y 9 grados de libertad |
| Dentro de los tratamientos, $V_W = V - V_B = 14 - 8 = 6$ | $a(b-1) = (3)(3) = 9$ | $\hat{S}_W^2 = \frac{6}{9} = \frac{2}{3}$ | |
| Total, $V = 14$ | $ab - 1 = (3)(4) - 1 = 11$ | | |

**16.5.** Con referencia al Problema 16.4, hallar una estimación sin sesgo de la varianza de la población $\sigma^2$ de (*a*) la variación entre tratamientos bajo la hipótesis nula de medias de tratamiento iguales y (*b*) la variación entre tratamientos.

**Solución**

(*a*) $$\hat{S}_B^2 = \frac{V_B}{a-1} = \frac{8}{3-1} = 4$$

(*b*) $$\hat{S}_W^2 = \frac{V_W}{a(b-1)} = \frac{6}{3(4-1)} = \frac{2}{3}$$

**16.6.** En el Problema 16.4, ¿podemos rechazar la hipótesis nula de medias iguales al nivel de significación (*a*) 0.05 y (*b*) 0.01?

**Solución**

Se tiene $$F = \frac{\hat{S}_B^2}{\hat{S}_W^2} = \frac{4}{2/3} = 6$$

con $a - 1 = 3 - 1 = 2$ grados de libertad y $a(b-1) = 3(4-1) = 9$ grados de libertad.

(*a*) En el Apéndice V, con $\nu_1 = 2$ y $\nu_2 = 9$, vemos que $F_{.95} = 4.26$. Como $F = 6 > F_{.95}$, podemos rechazar la hipótesis nula de medias iguales al nivel 0.05.

(*b*) En el Apéndice VI, con $\nu_1 = 2$ y $\nu_2 = 9$, vemos que $F_{.99} = 8.02$. Puesto que $F = 6 < F_{.99}$, no podemos rechazar la hipótesis nula de medias iguales al nivel 0.01.

**16.7.** Usar las fórmulas abreviadas (10), (11) y (12) para llegar a los resultados del Problema 16.4.

**Solución**

Conviene disponer los datos como en la Tabla 16.10.

**Tabla 16.10**

| | | $T_{j.}$ | $T_{j.}^2$ |
|---|---|---|---|
| $A$ | 3 4 5 4 | 16 | 256 |
| $B$ | 2 4 3 3 | 12 | 144 |
| $C$ | 4 6 5 5 | 20 | 400 |
| | $\sum_{j,k} X_{jk}^2 = 206$ | $T = \sum_j T_{j.} = 48$ | $\sum_j T_{j.}^2 = 800$ |

(*a*) Usando la fórmula (10), vemos que

$$\sum_{j,k} X_{jk}^2 = 9 + 16 + 25 + 16 + 4 + 16 + 9 + 9 + 16 + 36 + 25 + 25 = 206$$

y

$$T = 3 + 4 + 5 + 4 + 2 + 4 + 3 + 3 + 4 + 6 + 5 + 5 = 48$$

Luego

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} = 206 - \frac{(48)^2}{(3)(4)} = 206 - 192 = 14$$

(*b*) Los totales de las filas son

$$T_{1.} = 3+4+5+4 = 16 \qquad T_{2.} = 2+4+3+3 = 12 \qquad T_{3.} = 4+6+5+5 = 20$$

y

$$T = 16 + 12 + 20 = 48$$

Así que, por la fórmula (11), se deduce

$$V_B = \frac{1}{b}\sum_j T_{j.}^2 - \frac{T^2}{ab} = \frac{1}{4}(16^2 + 12^2 + 20^2) - \frac{(48)^2}{(3)(4)} = 200 - 192 = 8$$

(*c*) Mediante la fórmula (12), se obtiene

$$V_W = V - V_B = 14 - 8 = 6$$

Los resultados coinciden con los obtenidos en el Problema 16.4, y desde este punto en adelante el análisis es como antes.

**16.8.** Una empresa quiere comprar una de entre cinco máquinas diferentes: $A$, $B$, $C$, $D$ o $E$. En un experimento diseñado para comprobar si hay diferencia entre ellas, cada máquina fue manejada por un operario experto distinto en cada una, durante tiempos iguales. La Tabla 16.11 muestra los números de unidades producidas por las máquinas. Contrastar la hipótesis de que no hay diferencia entre las máquinas al nivel de significación (*a*) 0.05 y (*b*) 0.01.

**Solución**

Restar un número adecuado, 60 por ejemplo, a todos los datos de la Tabla 16.12. Entonces

$$V = 2658 - \frac{(54)^2}{(5)(4)} = 2658 - 145.8 = 2512.2$$

y

$$V_B = \frac{1}{5}(3874) - \frac{(54)^2}{(5)(4)} = 774.8 - 145.8 = 629.0$$

Ahora formamos la Tabla 16.13. Para 4 y 20 grados de libertad tenemos $F_{.95} = 2.87$. Luego no podemos rechazar la hipótesis nula al nivel 0.05 y por tanto con menos motivo al 0.01.

**Tabla 16.11**

| | | | | | |
|---|---|---|---|---|---|
| $A$ | 68 | 72 | 77 | 42 | 53 |
| $B$ | 72 | 53 | 63 | 53 | 48 |
| $C$ | 60 | 82 | 64 | 75 | 72 |
| $D$ | 48 | 61 | 57 | 64 | 50 |
| $E$ | 64 | 65 | 70 | 68 | 53 |

**Tabla 16.12**

| | | | | | | $T_{j.}$ | $T_{j.}^2$ |
|---|---|---|---|---|---|---|---|
| $A$ | 8 | 12 | 17 | −18 | −7 | 12 | 144 |
| $B$ | 12 | −7 | 3 | −7 | −12 | −11 | 121 |
| $C$ | 0 | 22 | 4 | 15 | 12 | 53 | 2809 |
| $D$ | −12 | 1 | −3 | 4 | −10 | −20 | 400 |
| $E$ | 4 | 5 | 10 | 8 | −7 | 20 | 400 |
| | $\sum X_{jk}^2 = 2658$ | | | | | 54 | 3874 |

**Tabla 16.13**

| Variación | Grados de libertad | Cuadrado medio | $F$ |
|---|---|---|---|
| Entre tratamientos, $V_B = 629.0$ | $a-1=4$ | $\hat{S}_B^2 = \frac{629.0}{4} = 157.25$ | $\frac{\hat{S}_B^2}{\hat{S}_W^2} = 1.67$ |
| Dentro de los tratamientos, $V_W = 1883.8$ | $a(b-1)=(5)(4)=20$ | $\hat{S}_W^2 = \frac{1883.2}{(5)(4)} = 94.16$ | |
| Total, $V = 2512.2$ | $ab-1=24$ | | |

## MODIFICACIONES PARA NUMEROS DISTINTOS DE OBSERVACIONES

**16.9.** La Tabla 16.4 da las vidas medias, en horas, de muestras de tres tipos distintos de tubos de televisión producidos por cierta empresa. Usando el método largo, determinar si hay diferencia entre ellos al nivel de significación (*a*) 0.05 y (*b*) 0.01.

**Tabla 16.14**

| | | | | | |
|---|---|---|---|---|---|
| Muestra 1 | 407 | 411 | 409 | | |
| Muestra 2 | 404 | 406 | 408 | 405 | 402 |
| Muestra 3 | 410 | 408 | 406 | 408 | |

**Solución**

Conviene restar a los datos un número apropiado, digamos 400, con lo que se obtiene la Tabla 16.15. Esta muestra los totales de fila, las medias muestrales (o de grupo) y la media global. Así pues, se tiene

$$V = \sum_{j,k} (X_{jk} - \bar{X})^2 = (7 - 7)^2 + (11 - 7)^2 + \cdots + (8 - 7)^2 = 72$$

$$V_B = \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 = \sum_{j} N_j(\bar{X}_{j.} - \bar{X})^2 = 3(9 - 7)^2 + 5(7 - 5)^2 + 4(8 - 7)^2 = 36$$

$$V_W = V - V_B = 72 - 36 = 36$$

**Tabla 16.15**

| | | | | | | Total | Media |
|---|---|---|---|---|---|---|---|
| Muestra 1 | 7 | 11 | 9 | | | 27 | 9 |
| Muestra 2 | 4 | 6 | 8 | 5 | 2 | 25 | 5 |
| Muestra 3 | 10 | 8 | 6 | 8 | | 32 | 8 |
| | $\bar{X}$ = media final = $\frac{84}{12} = 7$ | | | | | | |

Podemos también obtener $V_W$ directamente observando que es igual a

$$(7 - 9)^2 + (11 - 9)^2 + (9 - 9)^2 + (4 - 5)^2 + (6 - 5)^2 + (8 - 5)^2 + (5 - 5)^2 +$$
$$+ (2 - 5)^2 + (10 - 8)^2 + (8 - 8)^2 + (6 - 8)^2 + (8 - 8)^2$$

Los datos se resumen en la Tabla 16.16, la tabla del análisis de varianza. Para 2 y 9 grados de libertad, vemos en el Apéndice V que $F_{.95} = 4.26$ y en el Apéndice VI vemos que $F_{.99} = 8.02$. Luego podemos rechazar la hipótesis de medias iguales (o sea, no hay diferencia entre los tres tipos de tubos) al nivel 0.05, pero no al 0.01.

**Tabla 16.16**

| Variación | Grados de libertad | Cuadrado medio | $F$ |
|---|---|---|---|
| $V_B = 36$ | $a - 1 = 2$ | $\hat{S}_B^2 = \frac{36}{2} = 18$ | $\frac{\hat{S}_B^2}{\hat{S}_W^2} = \frac{18}{4}$ |
| $V_W = 36$ | $N - a = 9$ | $\hat{S}_W^2 = \frac{36}{9} = 4$ | $= 4.5$ |

**16.10.** Resolver el Problema 16.9 usando las fórmulas abreviadas incluidas en las ecuaciones (24), (25) y (26).

**Solución**

De la Tabla 16.15 se sigue $N_1 = 3$, $N_2 = 5$, $N_3 = 4$, $N = 12$, $T_1 = 27$, $T_2 = 25$, $T_3 = 32$ y $T = 84$. En consecuencia,

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{N} = 7^2 + 11^2 + \cdots + 6^2 + 8^2 - \frac{(84)^2}{12} = 72$$

$$V_B = \sum_j \frac{T_{j.}^2}{N_j} - \frac{T^2}{N} = \frac{(27)^2}{3} + \frac{(25)^2}{5} + \frac{(32)^2}{4} - \frac{(84)^2}{12} = 36$$

$$V_W = V - V_B = 36$$

Usando esto, el análisis de varianza se hace ya como en el Problema 16.9.

## EXPERIMENTOS DE DOS FACTORES

**16.11.** La Tabla 16.17 muestra las producciones por acre de cuatro semillas sembradas en campos tratados con tres fertilizantes distintos. Por el método largo, determinar el nivel de significación 0.01 si hay diferencia en producción por acre (*a*) debida a los fertilizantes y (*b*) debida a las semillas.

**Tabla 16.17**

| | Semilla I | Semilla II | Semilla III | Semilla IV |
|---|---|---|---|---|
| Fertilizante *A* | 4.5 | 6.4 | 7.2 | 6.7 |
| Fertilizante *B* | 8.8 | 7.8 | 9.6 | 7.0 |
| Fertilizante *C* | 5.9 | 6.8 | 5.7 | 5.2 |

**Solución**

Calcular los totales de fila, de columna, las medias de columna, el total global y la media global, como indica la Tabla 16.18. De esa tabla se obtiene:

**Tabla 16.18**

| | Cosecha I | Cosecha II | Cosecha III | Cosecha IV | Total de fila | Media de fila |
|---|---|---|---|---|---|---|
| Fertilizante *A* | 4.5 | 6.4 | 7.2 | 6.7 | 24.8 | 6.2 |
| Fertilizante *B* | 8.8 | 7.8 | 9.6 | 7.0 | 33.2 | 8.3 |
| Fertilizante *C* | 5.9 | 6.8 | 5.7 | 5.2 | 23.6 | 5.9 |
| Total de columna | 19.2 | 21.0 | 22.5 | 18.9 | Total final = 81.6 | |
| Media de columna | 6.4 | 7.0 | 7.5 | 6.3 | Media final = 6.8 | |

La variación de las medias de fila respecto de la media global es

$$V_R = 4[(6.2 - 6.8)^2 + (8.3 - 6.8)^2 + (5.9 - 6.8)^2] = 13.68$$

La variación de las medias de columna respecto de la media global es

$$V_C = 3[(6.4 - 6.8)^2 + (7.0 - 6.8)^2 + (7.5 - 6.8)^2 + (6.3 - 6.8)^2] = 2.82$$

La variación total es

$$\begin{aligned} V = {} & (4.5 - 6.8)^2 + (6.4 - 6.8)^2 + (7.2 - 6.8)^2 + (6.7 - 6.8)^2 + \\ & + (8.8 - 6.8)^2 + (7.8 - 6.8)^2 + (9.6 - 6.8)^2 + (7.0 - 6.8)^2 + \\ & + (5.9 - 6.8)^2 + (6.8 - 6.8)^2 + (5.7 - 6.8)^2 + (5.2 - 6.8)^2 = 23.08 \end{aligned}$$

La variación aleatoria es

$$V_E = V - V_R - V_C = 6.58$$

Eso conduce al análisis de varianza de la Tabla 16.19.

Al nivel de significación 0.05 con 2 y 6 grados de libertad, $F_{.95} = 5.14$. Por tanto, desde $6.24 > 5.14$, podemos rechazar la hipótesis de que las medias de fila son iguales y concluir que hay diferencia significativa en producción debida a los fertilizantes.

Como el valor $F$ correspondiente a la diferencia en medias de columna es menor que 1, concluimos que no hay diferencia significativa debida a las semillas en la producción

**Tabla 16.19**

| Variación | Grados de libertad | Cuadrado medio | $F$ |
|---|---|---|---|
| $V_R = 13.68$ | 2 | $\hat{S}_R^2 = 6.84$ | $\hat{S}_R^2/\hat{S}_E^2 = 6.24$ con 2 y 6 grados de libertad |
| $V_C = 2.82$ | 3 | $\hat{S}_C^2 = 0.94$ | $\hat{S}_C^2/\hat{S}_E^2 = 0.86$ con 3 y 6 grados de libertad |
| $V_E = 6.58$ | 6 | $\hat{S}_E^2 = 1.097$ | |
| $V = 23.08$ | 11 | | |

**16.12.** Usar las fórmulas abreviadas para llegar a los resultados del Problema 16.11.

**Solución**

De la Tabla 16.18 tenemos

$$\sum_{j,k} X_{jk}^2 = (4.5)^2 + (6.4)^2 + \cdots + (5.2)^2 = 577.96$$

$$T = 24.8 + 33.2 + 23.6 = 81.6$$

$$\sum T_{j.}^2 = (24.8)^2 + (33.2)^2 + (23.6)^2 = 2274.24$$

$$\sum T_{.k}^2 = (19.2)^2 + (21.0)^2 + (22.5)^2 + (18.9)^2 = 1673.10$$

Entonces

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} = 577.96 - 554.88 = 23.08$$

$$V_R = \frac{1}{b}\sum T_{j.}^2 - \frac{T^2}{ab} = \frac{1}{4}(2274.24) - 554.88 = 13.68$$

$$V_C = \frac{1}{a}\sum T_{.k}^2 - \frac{T^2}{ab} = \frac{1}{3}(1673.10) - 554.88 = 2.82$$

$$V_E = V - V_R - V_C = 23.08 - 13.68 - 2.82 = 6.58$$

de acuerdo con el Problema 16.11.

## EXPERIMENTOS DE DOS FACTORES CON REPETICION

**16.13.** Un empresario desea determinar la eficacia de cuatro tipos distintos de máquinas ($A$, $B$, $C$ y $D$) en la producción de tornillos. Para ello, anota el número de tornillos defectuosos cada día de una semana en dos turnos de trabajo, con los resultados que recoge la Tabla 16.20. Hacer un análisis de varianza para determinar al nivel de significación 0.05 si hay diferencia ($a$) entre las máquinas y ($b$) entre los turnos.

**Solución**

Los datos se organizan de modo equivalente en la Tabla 16.21, en la que los dos factores, máquinas y turnos, quedan indicados. Hay dos turnos para cada máquina. Los días de la semana pueden considerarse como repeticiones del trabajo de cada máquina. La variación total para todos los datos de la Tabla 16.21 es

$$V = 6^2 + 4^2 + 5^2 + \cdots + 7^2 + 10^2 - \frac{(268)^2}{40} = 1946 - 1795.6 = 150.4$$

**Tabla 16.20**

| Máquina | Primer turno | | | | | Segundo turno | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | L. | Mar. | Miér. | J. | V. | L. | Mar. | Miér. | J. | V. |
| $A$ | 6 | 4 | 5 | 5 | 4 | 5 | 7 | 4 | 6 | 8 |
| $B$ | 10 | 8 | 7 | 7 | 9 | 7 | 9 | 12 | 8 | 8 |
| $C$ | 7 | 5 | 6 | 5 | 9 | 9 | 7 | 5 | 4 | 6 |
| $D$ | 8 | 4 | 6 | 5 | 5 | 5 | 7 | 9 | 7 | 10 |

**Tabla 16.21**

| Factor I: Máquina | Factor II: Ensayo | Réplicas | | | | | Total |
|---|---|---|---|---|---|---|---|
| | | L. | Mar. | Miér. | J. | V. | |
| *A* | 1 | 6 | 4 | 5 | 5 | 4 | 24 |
| | 2 | 5 | 7 | 4 | 6 | 8 | 30 |
| *B* | 1 | 10 | 8 | 7 | 7 | 9 | 41 |
| | 2 | 7 | 9 | 12 | 8 | 8 | 44 |
| *C* | 1 | 7 | 5 | 6 | 5 | 9 | 32 |
| | 2 | 9 | 7 | 5 | 4 | 6 | 31 |
| *D* | 1 | 8 | 4 | 6 | 5 | 5 | 28 |
| | 2 | 5 | 7 | 9 | 7 | 10 | 38 |
| | Total | 57 | 51 | 54 | 47 | 59 | 268 |

Con el fin de considerar los dos factores, limitamos nuestra atención al total de valores de repetición correspondientes a cada combinación de factores. Recogidos en la Tabla 16.22 hacen de ésta una tabla de dos factores con entrada única. La variación total para la Tabla 16.22, que llamaremos *variación subtotal* $V_S$, viene dada por

$$V_S = \frac{(24)^2}{5} + \frac{(41)^2}{5} + \frac{(32)^2}{5} + \frac{(28)^2}{5} + \frac{(30)^2}{5} + \frac{(44)^2}{5} + \frac{(31)^2}{5} + \frac{(38)^2}{5} - \frac{(268)^2}{40} =$$

$$= 1861.2 - 1795.6 = 65.6$$

La variación entre filas es

$$V_R = \frac{(54)^2}{10} + \frac{(85)^2}{10} + \frac{(63)^2}{10} + \frac{(66)^2}{10} - \frac{(268)^2}{40} = 1846.6 - 1795.6 = 51.0$$

La variación entre columnas viene dada por

$$V_C = \frac{(125)^2}{20} + \frac{(143)^2}{20} - \frac{(268)^2}{40} = 1803.7 - 1795.6 = 8.1$$

**Tabla 16.22**

| Máquina | Primer ensayo | Segundo ensayo | Total |
|---|---|---|---|
| *A* | 24 | 30 | 54 |
| *B* | 41 | 44 | 85 |
| *C* | 32 | 31 | 63 |
| *D* | 28 | 38 | 66 |
| Total | 125 | 143 | 268 |

**Tabla 16.21**

| Factor I: Máquina | Factor II: Ensayo | Réplicas L. | Mar. | Miér. | J. | V. | Total |
|---|---|---|---|---|---|---|---|
| *A* | 1 | 6 | 4 | 5 | 5 | 4 | 24 |
| | 2 | 5 | 7 | 4 | 6 | 8 | 30 |
| *B* | 1 | 10 | 8 | 7 | 7 | 9 | 41 |
| | 2 | 7 | 9 | 12 | 8 | 8 | 44 |
| *C* | 1 | 7 | 5 | 6 | 5 | 9 | 32 |
| | 2 | 9 | 7 | 5 | 4 | 6 | 31 |
| *D* | 1 | 8 | 4 | 6 | 5 | 5 | 28 |
| | 2 | 5 | 7 | 9 | 7 | 10 | 38 |
| | Total | 57 | 51 | 54 | 47 | 59 | 268 |

Con el fin de considerar los dos factores, limitamos nuestra atención al total de valores de repetición correspondientes a cada combinación de factores. Recogidos en la Tabla 16.22 hacen de ésta una tabla de dos factores con entrada única. La variación total para la Tabla 16.22, que llamaremos *variación subtotal* $V_S$, viene dada por

$$V_S = \frac{(24)^2}{5} + \frac{(41)^2}{5} + \frac{(32)^2}{5} + \frac{(28)^2}{5} + \frac{(30)^2}{5} + \frac{(44)^2}{5} + \frac{(31)^2}{5} + \frac{(38)^2}{5} - \frac{(268)^2}{40} =$$

$$= 1861.2 - 1795.6 = 65.6$$

La variación entre filas es

$$V_R = \frac{(54)^2}{10} + \frac{(85)^2}{10} + \frac{(63)^2}{10} + \frac{(66)^2}{10} - \frac{(268)^2}{40} = 1846.6 - 1795.6 = 51.0$$

La variación entre columnas viene dada por

$$V_C = \frac{(125)^2}{20} + \frac{(143)^2}{20} - \frac{(268)^2}{40} = 1803.7 - 1795.6 = 8.1$$

**Tabla 16.22**

| Máquina | Primer ensayo | Segundo ensayo | Total |
|---|---|---|---|
| *A* | 24 | 30 | 54 |
| *B* | 41 | 44 | 85 |
| *C* | 32 | 31 | 63 |
| *D* | 28 | 38 | 66 |
| Total | 125 | 143 | 268 |

Si restamos ahora de $V_S$ la suma de las variaciones entre filas y columnas ($V_R + V_C$), obtenemos la variación debida a la *interacción* entre filas y columnas, que está dada por

$$V_I = V_S - V_R - V_C = 65.6 - 51.0 - 8.1 = 6.5$$

Finalmente, la variación residual, que se puede ver como la variación de error o azar $V_E$ (supuesto que creemos que los diversos días de la semana no producen diferencias relevantes), se halla restando la variación subtotal (o sea, la suma de las variaciones de fila, columna e interacción) de la variación total $V$. Eso da

$$V_E = V - (V_R + V_C + V_I) = V - V_S = 150.4 - 65.6 = 84.8$$

Estas variaciones se recogen en la Tabla 16.23, el análisis de varianza. La tabla da también el número de grados de libertad correspondiente a cada tipo de variación. Así pues, como hay cuatro filas en la Tabla 16.22, la variación debida a filas tiene $4 - 1 = 3$ grados de libertad, mientras que la variación debida a las dos columnas tiene $2 - 1 = 1$ grados de libertad. Para hallar los grados de libertad debidos a la interacción, notemos que hay ocho entradas en la Tabla 16.22; luego los grados de libertad totales son $8 - 1 = 7$. Puesto que 3 de ellos se deben a las filas y 1 a las columnas, los restantes $[7 - (3 + 1) = 3]$ se deben a la interacción. Puesto que hay 40 entradas en la tabla original 16.21, el total de grados de libertad es $40 - 1 = 39$. De modo que los grados de libertad debidos a la variación residual o de azar son $39 - 7 = 32$.

**Tabla 16.23**

| Variación | Grados de libertad | Cuadrado medio | $F$ |
|---|---|---|---|
| Filas (máquinas), $V_R = 51.0$ | 3 | $\hat{S}_R^2 = 17.0$ | $\frac{17.0}{2.65} = 6.42$ |
| Columnas (turnos), $V_C = 8.1$ | 1 | $\hat{S}_C^2 = 8.1$ | $\frac{8.1}{2.65} = 3.06$ |
| Interacción, $V_I = 6.5$ | 3 | $\hat{S}_I^2 = 2.167$ | $\frac{2.167}{2.65} = 0.817$ |
| Subtotal, $V_S = 65.6$ | 7 | | |
| Aleatorio o residual, $V_E = 84.8$ | 32 | $\hat{S}_E^2 = 2.65$ | |
| Total, $V = 150.4$ | 39 | | |

Para continuar, hemos de determinar primero si hay interacción significativa entre los factores básicos (o sea, las filas y columnas de la Tabla 16.22). De la Tabla 16.23 vemos que para la interacción es $F = 0.817$, lo cual nos dice que la interacción no es significativa; esto es, no podemos rechazar la hipótesis $H_0^{(3)}$ de la página 385. Siguiendo las reglas de la misma página, vemos que la $F$ calculada para filas es 6.42. Como $F_{.95} = 2.90$ para 3 y 32 grados de libertad, podemos rechazar la hipótesis $H_0^{(1)}$

de que las filas tienen medias iguales. Ello equivale a decir que al nivel 0.05 podemos concluir que las máquinas no son igualmente eficaces.

Para 1 y 32 grados de libertad, $F_{.95} = 4.15$. Entonces, ya que la $F$ calculada para columnas es 3.06, no podemos rechazar la $H_0^{(2)}$ de que las columnas tienen medias iguales. Lo que equivale a decir que al nivel 0.05 no hay diferencia significativa entre los turnos.

Si podemos optar por analizar los resultados uniendo las variaciones residual y de interacción, como propugnan algunos estadísticos, encontramos que $V_I + V_E = 6.5 + 84.8 = 91.3$ para la variación conjunta y $V_I + V_E = 3 + 32 = 35$ para los grados de libertad conjuntos, que nos da una varianza conjunta de $91.3/35 = 2.61$. Usar este valor en lugar de 2.65 para el denominador de $F$ en la Tabla 16.23 no afecta a las conclusiones antes alcanzadas.

**16.14.** Rehacer el Problema 16.13 al nivel de significación 0.01.

**Solución**

A este nivel no hay todavía interacción apreciable, así que podemos continuar.

Como $F_{.99} = 4.47$ para 3 y 32 grados de libertad y el $F$ calculado para filas es 6.42, podemos concluir que incluso al nivel 0.01 las máquinas no son igualmente efectivas.

Como $F_{.99} = 7.51$ para 1 y 32 grados de libertad y la $F$ para columnas es 3.06, podemos concluir que al nivel de significación 0.01 no hay diferencia significativa entre turnos.

## CUADRADOS LATINOS

**16.15.** Un labrador quiere contrastar los efectos de cuatro fertilizantes ($A$, $B$, $C$ y $D$) en la producción de trigo. Con el fin de eliminar fuentes de error debidas a la variabilidad en la fertilidad del suelo, los utiliza en una disposición de cuadrado latino, tal como indica la Tabla 16.24, donde los números están en bushels por unidad de área. Hacer un análisis de varianza para determinar si hay diferencia entre los fertilizantes al nivel de significación (*a*) 0.05 y (*b*) 0.01.

**Solución**

Primero obtenemos totales de filas y columnas (véase Tabla 16.25). También obtenemos las producciones totales de cada uno de los fertilizantes (véase Tabla 16.26). La variación total y las variaciones para filas, columnas y tratamientos se deducen de ahí del modo usual. Encontramos:

La variación total es

$$V = (18)^2 + (21)^2 + (25)^5 + \cdots + (10)^2 + (17)^2 - \frac{(295)^2}{16} =$$

$$= 5769 - 5439.06 = 329.94$$

**Tabla 16.24**

| | | | |
|---|---|---|---|
| $A$ 18 | $C$ 21 | $D$ 25 | $B$ 11 |
| $D$ 22 | $B$ 12 | $A$ 15 | $C$ 19 |
| $B$ 15 | $A$ 20 | $C$ 23 | $D$ 24 |
| $C$ 22 | $D$ 21 | $B$ 10 | $A$ 17 |

**Tabla 16.25**

| | | | | | Total |
|---|---|---|---|---|---|
| | $A$ 18 | $C$ 21 | $D$ 25 | $B$ 11 | 75 |
| | $D$ 22 | $B$ 12 | $A$ 15 | $C$ 19 | 68 |
| | $B$ 15 | $A$ 20 | $C$ 23 | $D$ 24 | 82 |
| | $C$ 22 | $D$ 21 | $B$ 10 | $A$ 17 | 70 |
| Total | 77 | 74 | 73 | 71 | 295 |

**Tabla 16.26**

| | *A* | *B* | *C* | *D* | |
|---|---|---|---|---|---|
| Total | 70 | 48 | 85 | 92 | 295 |

La variación entre filas es

$$V_R = \frac{(75)^2}{4} + \frac{(68)^2}{4} + \frac{(82)^2}{4} + \frac{(70)^2}{4} - \frac{(295)^2}{16} =$$

$$= 5468.25 - 5439.06 = 29.19$$

La variación entre columnas es

$$V_C = \frac{(77)^2}{4} + \frac{(74)^2}{4} + \frac{(73)^2}{4} + \frac{(71)^2}{4} - \frac{(295)^2}{16} =$$

$$= 5443.75 - 5439.06 = 4.69$$

La variación entre tratamientos es

$$V_B = \frac{(70)^2}{4} + \frac{(48)^2}{4} + \frac{(85)^2}{4} + \frac{(92)^2}{4} - \frac{(295)^2}{16} =$$

$$= 5723.25 - 5439.06 = 284.19$$

La Tabla 16.27 muestra el análisis de la varianza.

**Tabla 16.27**

| Variación | Grados de libertad | Cuadrado medio | *F* |
|---|---|---|---|
| Filas, 29.19 | 3 | 9.73 | 4.92 |
| Columnas, 4.69 | 3 | 1.563 | 0.79 |
| Tratamientos, 284.19 | 3 | 94.73 | 47.9 |
| Residuales, 11.87 | 6 | 1.978 | |
| Total, 329.94 | 15 | | |

(*a*) Como $F_{.95,\,3,\,6} = 4.76$, podemos rechazar al nivel 0.05 la hipótesis de medias de fila iguales. Se sigue que al nivel 0.05 hay diferencia en fertilidad del terreno de una fila a otra.

Como el valor $F$ para columnas es menor que 1, no hay diferencia en fertilidad en las columnas.

Ya que el valor $F$ para tratamientos es $47.9 > 4.76$, concluimos que hay diferencia entre los fertilizantes.

(*b*) Puesto que $F_{99,3,6} = 9.78$, podemos aceptar la hipótesis de que no hay diferencia en fertilidad en las filas (o en las columnas) al nivel 0.01. Sin embargo, debemos concluir todavía que hay diferencia entre los fertilizantes al nivel 0.01.

## CUADRADOS GRECO-LATINOS

**16.16.** Interesa saber si hay diferencia en millas recorridas por galón entre las gasolinas *A*, *B*, *C* y *D*. Diseñar un experimento con cuatro conductores distintos, cuatro coches distintos y cuatro carreteras distintas.

**Solución**

Como se usa el mismo número de cada uno de los factores, podemos recurrir a un cuadrado greco-latino. Supongamos que los diferentes coches se representan por filas y los diferentes conductores por columnas, como en la Tabla 16.28. Ahora asignamos las diferentes gasolinas (*A*, *B*, *C* y *D*) a las filas y columnas al azar, con el único requisito de que cada letra aparezca una vez en cada fila y en cada columna. Así pues, cada conductor conducirá una vez cada coche y usará una vez cada gasolina, y ningún coche será conducido dos veces con la misma gasolina.

Ahora asignamos al azar las cuatro carreteras, denotadas por $\alpha$, $\beta$, $\gamma$ y $\delta$, con el mismo requisito impuesto sobre los cuadrados latinos. Así que cada conductor tendrá oportunidad de conducir por cada una de ellas. La Tabla 16.28 muestra una de las posibles disposiciones.

**Tabla 16.28**

| | Conductor | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Coche 1 | $B_\gamma$ | $A_\beta$ | $D_\delta$ | $C_\alpha$ |
| Coche 2 | $A_\delta$ | $B_\alpha$ | $C_\gamma$ | $D_\beta$ |
| Coche 3 | $D_\alpha$ | $C_\delta$ | $B_\beta$ | $A_\gamma$ |
| Coche 4 | $C_\beta$ | $D_\gamma$ | $A_\alpha$ | $B_\delta$ |

**16.17.** Supongamos que al realizar el experimento del Problema 16.16, el número de millas por galón resulta ser el que indica la Tabla 16.29. Determinar por análisis de varianza si hay diferencias al nivel de significación 0.05.

**Tabla 16.29**

| | Conductor | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Coche 1 | $B_\gamma$ 19 | $A_\beta$ 16 | $D_\delta$ 16 | $C_\alpha$ 14 |
| Coche 2 | $A_\delta$ 15 | $B_\alpha$ 18 | $C_\gamma$ 11 | $D_\beta$ 15 |
| Coche 3 | $D_\alpha$ 14 | $C_\delta$ 11 | $B_\beta$ 21 | $A_\gamma$ 16 |
| Coche 4 | $C_\beta$ 16 | $D_\gamma$ 16 | $A_\alpha$ 15 | $B_\delta$ 23 |

**Solución**

Primero obtenemos los totales de filas y columnas (véase Tabla 16.30) y a continuación los totales para cada letra latina y para cada letra griega, como sigue:

$$\begin{aligned}
A \text{ total:} &\quad 15 + 16 + 15 + 16 = 62 \\
B \text{ total:} &\quad 19 + 18 + 21 + 23 = 81 \\
C \text{ total:} &\quad 16 + 11 + 11 + 14 = 52 \\
D \text{ total:} &\quad 14 + 16 + 16 + 15 = 61 \\
\alpha \text{ total:} &\quad 14 + 18 + 15 + 14 = 61 \\
\beta \text{ total:} &\quad 16 + 16 + 21 + 15 = 68 \\
\gamma \text{ total:} &\quad 19 + 16 + 11 + 16 = 62 \\
\delta \text{ total:} &\quad 15 + 11 + 16 + 23 = 65
\end{aligned}$$

**Tabla 16.30**

| | | | | | Total |
|---|---|---|---|---|---|
| | $B_\gamma$ 19 | $A_\beta$ 16 | $D_\delta$ 16 | $C_\alpha$ 14 | 65 |
| | $A_\delta$ 15 | $B_\alpha$ 18 | $C_\gamma$ 11 | $D_\beta$ 15 | 59 |
| | $D_\alpha$ 14 | $C_\delta$ 11 | $B_\beta$ 21 | $A_\gamma$ 16 | 62 |
| | $C_\beta$ 16 | $D_\gamma$ 16 | $A_\alpha$ 15 | $B_\delta$ 23 | 70 |
| Total | 64 | 61 | 63 | 68 | 256 |

Ahora calculamos las variaciones correspondientes a todas éstas, mediante el método abreviado:

$$\text{Filas:} \quad \frac{(65)^2}{4} + \frac{(59)^2}{4} + \frac{(62)^2}{4} + \frac{(70)^2}{4} - \frac{(256)^2}{16} = 4112.50 - 4096 = 16.50$$

$$\text{Columnas:} \quad \frac{(64)^2}{4} + \frac{(61)^2}{4} + \frac{(63)^2}{4} + \frac{(68)^2}{4} - \frac{(256)^2}{16} = 4102.50 - 4096 = 6.50$$

$$\text{Gasolinas } (A, B, C, D): \quad \frac{(62)^2}{4} + \frac{(81)^2}{4} + \frac{(52)^2}{4} + \frac{(61)^2}{4} - \frac{(256)^2}{16} = 4207.50 - 4096 = 111.50$$

$$\text{Carreteras } (\alpha, \beta, \gamma, \delta): \quad \frac{(61)^2}{4} + \frac{(68)^2}{4} + \frac{(62)^2}{4} + \frac{(65)^2}{4} - \frac{(256)^2}{16} = 4103.50 - 4096 = 7.50$$

La variación total es

$$(19)^2 + (16)^2 + (16)^2 + \cdots + (15)^2 + (23)^2 - \frac{(256)^2}{16} = 4244 - 4096 = 148.00$$

de manera que la variación debida a error es

$$148.00 - 16.50 - 6.50 - 111.50 - 7.50 = 6.00$$

Los resultados del análisis de varianza se recogen en la Tabla 16.31. El número total de grados de libertad es $N^2 - 1$ para un cuadrado $N \times N$. Cada fila, columna, letra latina y letra griega tiene $N - 1$ grados de libertad. Así pues, los grados de libertad para el error son $N^2 - 1 - 4(N - 1) = (N - 1)(N - 3)$. En nuestro caso, $N = 4$.

Tenemos $F_{.95,3,3} = 9.28$ y $F_{.99,3,3} = 29.5$. Luego podemos rechazar la hipótesis de que las gasolinas son iguales al nivel 0.05 pero no al 0.01.

## PROBLEMAS DIVERSOS

**16.18.** Probar [como en la ecuación (15) de este capítulo] que $\sum_j \alpha_j = 0$.

**Solución**

Las medias de tratamiento de la población $\mu_j$ y la media de la población están relacionadas por

$$\mu = \frac{1}{a}\sum_j \mu_j \tag{53}$$

Entonces, como $\alpha_j = \mu_j - \mu$, tenemos, usando la ecuación (53),

$$\sum_j \alpha_j = \sum_j (\mu_j - \mu) = \sum_j \mu_j - a\mu = 0 \tag{54}$$

**Tabla 16.31**

| Variación | Grados de libertad | Cuadrado medio | $F$ |
|---|---|---|---|
| Filas (coches), 16.50 | 3 | 5.500 | $\frac{5.500}{2.000} = 2.75$ |
| Columnas (conductores), 6.50 | 3 | 2.167 | $\frac{2.167}{2.000} = 1.08$ |
| Gasolinas ($A$, $B$, $C$, $D$), 111.50 | 3 | 37.167 | $\frac{37.167}{2.000} = 18.6$ |
| Carreteras ($\alpha$, $\beta$, $\gamma$, $\delta$), 7.50 | 3 | 2.500 | $\frac{2.500}{2.000} = 1.25$ |
| Error, 6.00 | 3 | 2.000 | |
| Total, 148.00 | 15 | | |

**16.19.** Deducir (*a*) la ecuación (16) y (*b*) la ecuación (17) de este capítulo.

**Solución**

(*a*) Por definición se tiene

$$V_W = \sum_{j,k} (X_{jk} - \bar{X}_{j.})^2 = b \sum_{j=1}^{a} \left[ \frac{1}{b} \sum_{k=1}^{b} (X_{jk} - \bar{X}_{j.})^2 \right] = b \sum_{j=1}^{a} S_j^2$$

donde $S_j^2$ es la varianza de la muestra para el $j$-ésimo tratamiento. Entonces, como el tamaño de la muestra es $b$,

$$E(V_W) = b \sum_{j=1}^{a} E(S_j^2) = b \sum_{j=1}^{a} \left( \frac{b-1}{b} \sigma^2 \right) = a(b-1)\sigma^2$$

(*b*) Por definición,

$$V_B = b \sum_{j=1}^{a} (\bar{X}_{j.} - \bar{X})^2 = b \sum_{j=1}^{a} \bar{X}_{j.}^2 - 2b\bar{X} \sum_{j=1}^{a} \bar{X}_{j.} + ab\bar{X}^2 = b \sum_{j=1}^{a} \bar{X}_{j.}^2 - ab\bar{X}^2$$

ya que $\bar{X} = (\sum_j X_{jk})/a$. Omitiendo el índice de suma, se tiene

$$E(V_B) = b \sum E(\bar{X}_{j.}^2) - abE(\bar{X}^2) \tag{55}$$

Ahora bien, para cualquier variable aleatoria $U$, $E(U^2) = \text{var}\,(U) + [E(U)]^2$, donde var $(U)$ denota la varianza de $U$. Así pues,

$$E(\bar{X}_{j.}^2) = \text{var}\,(\bar{X}_{j.}) + [E(\bar{X}_{j.})]^2 \tag{56}$$

$$E(\bar{X}^2) = \text{var}\,(\bar{X}) + [E(\bar{X})]^2 \tag{57}$$

Pero como las poblaciones de los tratamientos son normales con medias $\mu_j = \mu + \alpha$, tenemos que

$$\text{var}\,(\bar{X}_{j.}) = \frac{\sigma^2}{b} \tag{58}$$

$$\text{var}\,(\bar{X}) = \frac{\sigma^2}{ab} \tag{59}$$

$$E(\bar{X}_{j.}) = \mu_j = \mu + \alpha_j \tag{60}$$

$$E(\bar{X}) = \mu \tag{61}$$

Los resultados (56) a (61) junto con (53) nos dan

$$\begin{aligned} E(V_B) &= b \sum \left[ \frac{\sigma^2}{b} + (\mu + \alpha_j)^2 \right] - ab \left[ \frac{\sigma^2}{ab} + \mu^2 \right] = \\ &= a\sigma^2 + b \sum (\mu + \alpha_j)^2 - \sigma^2 - ab\mu^2 = \\ &= (a-1)\sigma^2 + ab\mu^2 + 2b\mu \sum \alpha_j + b \sum \alpha_j^2 + ab\mu^2 = \\ &= (a-1)\sigma^2 + b \sum \alpha_j^2 \end{aligned}$$

**16.20.** Demostrar el Teorema 1 de este capítulo.

**Solución**

Como muestra el Problema 16.19,

$$V_W = b\sum_{j=1}^{a} S_j^2 \qquad \text{o sea} \qquad \frac{V_W}{\sigma^2} = \sum_{j=1}^{a} \frac{bS_j^2}{\sigma^2}$$

donde $S_j^2$ es la varianza de la muestra para muestras de tamaño $b$ tomadas en la población del tratamiento $j$. De la página 254 vemos que $bS_j^2/\sigma^2$ tiene una distribución $ji$-cuadrado con $b - 1$ grados de libertad. Luego, como las varianzas $S_j^2$ son independientes, concluimos de la página 272 que $V_W/\sigma^2$ tiene una distribución $ji$-cuadrado con $a(b - 1)$ grados de libertad.

## PROBLEMAS SUPLEMENTARIOS

### EXPERIMENTOS DE UN FACTOR

**16.21.** Se realiza un experimento para determinar las producciones de 5 variedades de trigo: $A$, $B$, $C$, $D$ y $E$. Se asignan 4 parcelas a cada variedad. Las producciones (en bushels por acre) se dan en la Tabla 16.32. Supuesto que las parcelas son de la misma fertilidad y que las variedades se asignan al azar a las parcelas, determinar si hay diferencia entre las producciones al nivel de significación ($a$) 0.05 y ($b$) 0.01.

**Tabla 16.32**

| | | | | |
|---|---|---|---|---|
| $A$ | 20 | 12 | 15 | 19 |
| $B$ | 17 | 14 | 12 | 15 |
| $C$ | 23 | 16 | 18 | 14 |
| $D$ | 15 | 17 | 20 | 12 |
| $E$ | 21 | 14 | 17 | 18 |

**16.22.** Una empresa quiere comparar cuatro tipos de llantas: $A$, $B$, $C$ y $D$. Sus vidas medias en rodaje (en miles de millas) se dan en la Tabla 16.33, donde cada tipo ha sido probado en seis coches similares asignados al azar a las llantas. Determinar si hay diferencia significativa al nivel de significación ($a$) 0.05 y ($b$) 0.01 entre las llantas.

**Tabla 16.33**

| | | | | | | |
|---|---|---|---|---|---|---|
| $A$ | 33 | 38 | 36 | 40 | 31 | 35 |
| $B$ | 32 | 40 | 42 | 38 | 30 | 34 |
| $C$ | 31 | 37 | 35 | 33 | 34 | 30 |
| $D$ | 29 | 34 | 32 | 30 | 33 | 31 |

**16.23.** Un profesor quiere contrastar tres tipos distintos de enseñanza: I, II y III. Para ello, escoge al azar tres grupos de 5 estudiantes cada uno, y aplica a cada uno un método distinto. Tras proponer, al final del curso, el mismo examen a todos ellos, se obtienen las notas que indica la Tabla 16.34. Determinar si hay diferencia significativa entre los tres métodos al nivel de significación ($a$) 0.05 y ($b$) 0.01.

**Tabla 16.34**

| | | | | | |
|---|---|---|---|---|---|
| Método I | 75 | 62 | 71 | 58 | 73 |
| Método II | 81 | 85 | 68 | 92 | 90 |
| Método III | 73 | 79 | 60 | 75 | 81 |

## MODIFICACIONES PARA NUMEROS DISTINTOS DE OBSERVACIONES

**16.24.** La Tabla 16.35 da el número de millas por galón recorridas por coches similares usando cinco tipos distintos de gasolina. Determinar si hay diferencia significativa entre las gasolinas al nivel de significación (*a*) 0.05 y (*b*) 0.01.

**Tabla 16.35**

| | | | | | |
|---|---|---|---|---|---|
| Tipo *A* | 12 | 15 | 14 | 11 | 15 |
| Tipo *B* | 14 | 12 | 15 | | |
| Tipo *C* | 11 | 12 | 10 | 14 | |
| Tipo *D* | 15 | 18 | 16 | 17 | 14 |
| Tipo *E* | 10 | 12 | 14 | 12 | |

**16.25.** Durante un curso, un estudiante obtuvo las calificaciones que figuran en la Tabla 16.36. Determinar si hay diferencia significativa entre esas calificaciones al nivel de significación.

**Tabla 16.36**

| | | | | | |
|---|---|---|---|---|---|
| Matemáticas | 72 | 80 | 83 | 75 | |
| Ciencias | 81 | 74 | 77 | | |
| Inglés | 88 | 82 | 90 | 87 | 80 |
| Economía | 74 | 71 | 77 | 70 | |

## EXPERIMENTOS DE DOS FACTORES

**16.26.** Los artículos manufacturados por una compañía se producen en 3 máquinas distintas manejadas por 3 operarios diferentes. El dueño desea saber si hay diferencia (*a*) entre los operarios y (*b*) entre las máquinas. Se realiza un experimento para conocer el número de artículos producidos al día, con los resultados que recoge la Tabla 16.37. Establecer la deseada información al nivel de significación 0.05.

**Tabla 16.37**

| | Operador | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Máquina *A* | 23 | 27 | 24 |
| Máquina *B* | 34 | 30 | 28 |
| Máquina *C* | 28 | 25 | 27 |

**16.27.** Rehacer el Problema 16.26 al nivel de significación 0.01.

**16.28.** Se siembran semillas de maíz de 4 tipos distintos en 5 bloques, cada bloque dividido en 4 parcelas que se asignan al azar a dichos 4 tipos de semillas. Determinar el nivel de significación 0.05 si las producciones en bushels por acre, dadas en la Tabla 16.38, varían significativamente con diferentes (*a*) terrenos (o sea, los 5 bloques) y (*b*) tipos de maíz.

**Tabla 16.38**

| | Tipo de maíz | | | |
|---|---|---|---|---|
| | I | II | III | IV |
| Bloque *A* | 12 | 15 | 10 | 14 |
| Bloque *B* | 15 | 19 | 12 | 11 |
| Bloque *C* | 14 | 18 | 15 | 12 |
| Bloque *D* | 11 | 16 | 12 | 16 |
| Bloque *E* | 16 | 17 | 11 | 14 |

**16.29.** Resolver el Problema 16.28 al nivel de significación 0.01.

**16.30.** Supongamos que en el Problema 16.22 se hace la primera observación para cada tipo de llanta usando un tipo particular de coche, la segunda con otro tipo de coche, etc. Determinar si hay diferencia significativa al nivel de significación 0.05 entre (*a*) los tipos de llantas y (*b*) las clases de coches usados.

**16.31.** Rehacer el Problema 16.30 al nivel de significación 0.01.

**16.32.** Supongamos que en el Problema 16.23 la primera entrada para cada método de enseñanza corresponde a un estudiante de un colegio concreto, la segunda a uno de otro colegio, etc. Contrastar la hipótesis, al nivel de significación 0.05, de que hay diferencia entre (*a*) los métodos de enseñanza y (*b*) los colegios.

**16.33.** Se realiza un experimento para saber si el color del cabello y la altura de mujeres adultas en EE.UU. tienen alguna influencia sobre el rendimiento escolar. Los resultados figuran en la Tabla 16.39, donde los números indican individuos en el 10% más alto de entre los que se gradúan. Analizar el experimento al nivel de significación 0.05.

**Tabla 16.39**

| | Pelirroja | Rubia | Castaña |
|---|---|---|---|
| Alta | 75 | 78 | 80 |
| Media | 81 | 76 | 79 |
| Baja | 73 | 75 | 77 |

**16.34.** Repetir el Problema 16.33 al nivel de significación 0.01.

## EXPERIMENTOS DE DOS FACTORES CON REPETICION

**16.35.** Supongamos que el experimento del Problema 16.21 se realizó en el sur de EE.UU. y que las columnas de la Tabla 16.32 indican ahora 4 tipos de fertilizantes, mientras que un experimento similar se llevó a cabo en el Oeste con los resultados de la Tabla 16.40. Determinar al nivel de significación 0.05 si hay diferencia en producción debida a (*a*) los fertilizantes y (*b*) la localización.

**Tabla 16.40**

| | | | | |
|---|---|---|---|---|
| *A* | 16 | 18 | 20 | 23 |
| *B* | 15 | 17 | 16 | 19 |
| *C* | 21 | 19 | 18 | 21 |
| *D* | 18 | 22 | 21 | 23 |
| *E* | 17 | 18 | 24 | 20 |

**16.36.** Rehacer el Problema 16.35 al nivel de significación 0.01.

**16.37.** La Tabla 16.41 da el número de artículos producidos por 4 trabajadores en dos máquinas distintas, I y II, en diferentes días de la semana. Determinar si hay diferencia significativa al nivel 0.05 entre (*a*) los trabajadores y (*b*) las máquinas.

**Tabla 16.41**

| | Máquina I | | | | |
|---|---|---|---|---|---|
| | L. | Mar. | Miér. | J. | V. |
| Operador *A* | 15 | 18 | 17 | 20 | 12 |
| Operador *B* | 12 | 16 | 14 | 18 | 11 |
| Operador *C* | 14 | 17 | 18 | 16 | 13 |
| Operador *D* | 19 | 16 | 21 | 23 | 18 |

| | Máquina II | | | | |
|---|---|---|---|---|---|
| | L. | Mar. | Miér. | J. | V. |
| Operador *A* | 14 | 16 | 18 | 17 | 15 |
| Operador *B* | 11 | 15 | 12 | 16 | 12 |
| Operador *C* | 12 | 14 | 16 | 14 | 11 |
| Operador *D* | 17 | 15 | 18 | 20 | 17 |

## CUADRADOS LATINOS

**16.38.** Se lleva a cabo un experimento para comprobar los efectos en la producción de maíz de 4 fertilizantes (*A*, *B*, *C* y *D*) y de las variaciones del terreno en dos direcciones perpendiculares. El cuadrado latino de la Tabla 16.42 da los resultados obtenidos, donde los números muestran la producción de maíz por unidad de área. Contrastar al nivel de significación 0.01 la hipótesis de que no hay diferencia entre (*a*) los fertilizantes y (*b*) las variaciones del terreno.

**Tabla 16.42**

| | | | |
|---|---|---|---|
| $C$ 8 | $A$ 10 | $D$ 12 | $B$ 11 |
| $A$ 14 | $C$ 12 | $B$ 11 | $D$ 15 |
| $D$ 10 | $B$ 14 | $C$ 16 | $A$ 10 |
| $B$ 7 | $D$ 16 | $A$ 14 | $C$ 12 |

**16.39.** Resolver el Problema 16.38 al nivel de significación 0.05.

**16.40.** Con referencia al Problema 16.33, supongamos que introducimos un factor adicional, dando la parte $E$, $M$ o $W$ de los EE.UU. en que nació un estudiante, como muestra la Tabla 16.43. Determinar si hay diferencia significativa al nivel 0.05 en los rendimientos escolares debidas a diferencias en (*a*) altura, (*b*) color del cabello y (*c*) lugar de nacimiento.

**Tabla 16.43**

| | | |
|---|---|---|
| $E$ 75 | $W$ 78 | $M$ 80 |
| $M$ 81 | $E$ 76 | $W$ 79 |
| $W$ 73 | $M$ 75 | $E$ 77 |

## CUADRADOS GRECO-LATINOS

**16.41.** Con objeto de lograr mejorar la calidad de un pienso para gallinas, se han añadido dos productos químicos a sus ingredientes básicos. Las distintas cantidades del primero se indican por $A$, $B$, $C$ y $D$, y las del segundo por $\alpha$, $\beta$, $\gamma$ y $\delta$. Se da el pienso a animales ordenados en grupos de acuerdo con cuatro pesos iniciales diferentes ($W_1$, $W_2$, $W_3$ y $W_4$) y cuatro especies diferentes ($S_1$, $S_2$, $S_3$ y $S_4$). Los aumentos de peso por unidad de tiempo vienen dados en el cuadrado greco-latino de la Tabla 16.44. Hacer un análisis de varianza del experimento al nivel de significación 0.05, sacando las conclusiones pertinentes.

**Tabla 16.44**

| | $W_1$ | $W_2$ | $W_3$ | $W_4$ |
|---|---|---|---|---|
| $S_1$ | $C_\gamma$ 8 | $B_\beta$ 6 | $A_\alpha$ 5 | $D_\delta$ 6 |
| $S_2$ | $A_\delta$ 4 | $D_\alpha$ 3 | $C_\beta$ 7 | $B_\gamma$ 3 |
| $S_3$ | $D_\beta$ 5 | $A_\gamma$ 6 | $B_\delta$ 5 | $C_\alpha$ 6 |
| $S_4$ | $B_\alpha$ 6 | $C_\delta$ 10 | $D_\gamma$ 10 | $A_\beta$ 8 |

**16.42.** Cuatro tipos de cables ($T_1$, $T_2$, $T_3$ y $T_4$) se fabrican en cada una de las empresas ($C_1$, $C_2$, $C_3$ y $C_4$). Cuatro operarios ($A$, $B$, $C$ y $D$) usando cuatro máquinas distintas ($\alpha$, $\beta$, $\gamma$ y $\delta$) miden las tensiones de ruptura de esos cables, obteniendo los valores promedio que indica el cuadrado greco-latino de la Tabla 16.45. Hacer un análisis de varianza al nivel de significación 0.05 para llegar a las conclusiones pertinentes.

**Tabla 16.45**

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ |
|---|---|---|---|---|
| $T_1$ | $A_\beta$ 164 | $B_\gamma$ 181 | $C_\alpha$ 193 | $D_\delta$ 160 |
| $T_2$ | $C_\delta$ 171 | $D_\alpha$ 162 | $A_\gamma$ 183 | $B_\beta$ 145 |
| $T_3$ | $D_\gamma$ 198 | $C_\beta$ 212 | $B_\delta$ 207 | $A_\alpha$ 188 |
| $T_4$ | $B_\alpha$ 157 | $A_\delta$ 172 | $D_\beta$ 166 | $C_\gamma$ 136 |

## PROBLEMAS DIVERSOS

**16.43.** La Tabla 16.46 proporciona datos sobre la herrumbre acumulada sobre el hierro tra-

**Tabla 16.46**

| | | | | |
|---|---|---|---|---|
| $A$ | 3 | 5 | 4 | 4 |
| $B$ | 4 | 2 | 3 | 3 |
| $C$ | 6 | 4 | 5 | 5 |

tado con productos químicos $A$, $B$ o $C$, respectivamente. Determinar al nivel de significación (*a*) 0.05 y (*b*) 0.01 si hay diferencia significativa entre esos tratamientos.

**16.44.** Un experimento mide los coeficientes de inteligencia (IQ) de estudiantes varones adultos de estatura alta, media y baja, con los resultados que figuran en la Tabla 16.47. Determinar si hay diferencia significativa al nivel de significación (*a*) 0.05 y (*b*) 0.01 en los IQ por efecto de las diferencias en altura.

**Tabla 16.47**

| | | | | | | |
|---|---|---|---|---|---|---|
| Alto | 110 | 105 | 118 | 112 | 90 | |
| Bajo | 95 | 103 | 115 | 107 | | |
| Medio | 108 | 112 | 93 | 104 | 96 | 102 |

**16.45.** Probar los resultados (10), (11) y (12) de este capítulo.

**16.46.** Se hace una prueba para saber si responden mejor los veteranos o los no veteranos de diversos IQ. Las calificaciones obtenidas son las de la Tabla 16.48. Determinar si hay diferencia significativa al nivel de significación 0.05, debida a diferencias en (*a*) ser o no veterano y (*b*) IQ.

**Tabla 16.48**

| | Resultado del test | | |
|---|---|---|---|
| | Alto IQ | Medio IQ | Bajo IQ |
| Veterano | 90 | 81 | 74 |
| No veterano | 85 | 78 | 70 |

**16.47.** Repetir el Problema 16.46 al nivel de significación 0.01.

**16.48.** La Tabla 16.49 muestra las notas de una muestra de estudiantes procedentes de diferentes partes del país y con diferentes IQ. Analizar los datos de la tabla al nivel de significación 0.05 y establecer conclusiones.

**Tabla 16.49**

| | Resultado del test | | |
|---|---|---|---|
| | Alto IQ | Medio IQ | Bajo IQ |
| Este | 88 | 80 | 72 |
| Oeste | 84 | 78 | 75 |
| Sur | 86 | 82 | 70 |
| Norte y central | 80 | 75 | 79 |

**16.49.** Resolver el Problema 16.48 al nivel de significación 0.01.

**16.50.** En el Problema 16.37, ¿puede determinar si hay diferencia significativa en el número de artículos producidos en distintos días de la semana? Explíquese.

**16.51.** En cálculos de análisis de varianza se sabe que puede añadirse o restarse una constante adecuada a cada entrada sin que ello afecte a las conclusiones. ¿Es eso cierto también si cada entrada se multiplica por una constante? Justificar la respuesta.

**16.52.** Deducir los resultados (24), (25) y (26) para números distintos de observaciones.

**16.53.** Supongamos que los resultados de la Tabla 16.46 del Problema 16.43 son válidos para la parte nordeste de los EE.UU., mientras que los de la Tabla 16.50 lo son para la parte oeste. Determinar al nivel de significación 0.05 si hay diferencias debidas a (*a*) los productos químicos y (*b*) la localización.

**Tabla 16.50**

| | | | | |
|---|---|---|---|---|
| $A$ | 5 | 4 | 6 | 3 |
| $B$ | 3 | 4 | 2 | 3 |
| $C$ | 5 | 7 | 4 | 6 |

**16.54.** Refiriéndonos a los Problemas 16.21 y 16.35, supongamos que se realiza un experimento adicional en la parte nordeste de EE.UU. y produce los resultados de la Tabla 16.51. Determinar al nivel 0.05 si hay diferencia en la producción debida (*a*) a los fertilizantes, y (*b*) a las tres localizaciones.

**Tabla 16.51**

| | | | | |
|---|---|---|---|---|
| *A* | 17 | 14 | 18 | 12 |
| *B* | 20 | 10 | 20 | 15 |
| *C* | 18 | 15 | 16 | 17 |
| *D* | 12 | 11 | 14 | 11 |
| *E* | 15 | 12 | 19 | 14 |

**16.55.** Repetir el Problema 16.54 al nivel de significación 0.01.

**16.56.** Hacer un análisis de varianza del cuadrado latino de la Tabla 16.52 al nivel de significación 0.05 y establecer las conclusiones pertinentes.

**16.57.** Describir un experimento que conduzca al cuadrado latino de la Tabla 16.52.

**Tabla 16.52**

Factor 1

| Factor 2 | | | |
|---|---|---|---|
| | *B* 16 | *C* 21 | *A* 15 |
| | *A* 18 | *B* 23 | *C* 14 |
| | *C* 15 | *A* 18 | *B* 12 |

**16.58.** Hacer un análisis de varianza del cuadrado greco-latino de la Tabla 16.53 al nivel de significación 0.05 y sacar las conclusiones.

**Tabla 16.53**

Factor 1

| Factor 2 | | | | |
|---|---|---|---|---|
| | $A_\gamma$ 6 | $B_\beta$ 12 | $C_\delta$ 4 | $D_\alpha$ 18 |
| | $B_\delta$ 3 | $A_\alpha$ 8 | $D_\gamma$ 15 | $C_\beta$ 14 |
| | $D_\beta$ 15 | $C_\gamma$ 20 | $B_\alpha$ 9 | $A_\delta$ 5 |
| | $C_\alpha$ 16 | $D_\delta$ 6 | $A_\beta$ 17 | $B_\gamma$ 7 |

**16.59.** Describir un experimento que conduzca al cuadrado greco-latino de la Tabla 16.53.

**16.60.** Describir cómo usar el análisis de varianza para experimentos de tres factores con repetición.

**16.61.** Enunciar y resolver un problema que ilustre el procedimiento del Problema 16.60.

**16.62.** Probar (*a*) la ecuación (30) y (*b*) los resultados (31) a (34) de este capítulo.

**16.63.** En la práctica, ¿cabe esperar hallar (*a*) un cuadrado latino $2 \times 2$ y (*b*) un cuadrado greco-latino $3 \times 3$? Explicar la razón.

# CAPITULO 17

# Contrastes no paramétricos

## INTRODUCCION

La mayor parte de los contrastes de hipótesis y significación (o reglas de decisión) considerados en los capítulos precedentes requieren varias suposiciones acerca de la distribución de la población cuyas muestras se analizan. Por ejemplo, en la página 187 las distribuciones de la población se exigían normales o casi normales.

En la práctica aparecen situaciones en las que tales requisitos no están justificados, como es el caso de una población fuertemente asimétrica. A causa de ello, los estadísticos han creado varios contrastes y métodos que son independientes de las distribuciones de la población y de los parámetros asociados. Estos se llaman *contrastes* o *tests no paramétricos*.

Los tests no paramétricos se pueden usar como abreviaciones de contrastes más complicados. Son especialmente útiles cuando se trata con datos no numéricos, por ejemplo, cuando los consumidores colocan productos por orden de preferencia.

## EL TEST DE LOS SIGNOS

Consideremos la Tabla 17.1, que indica los números de tuercas defectuosas producidas por dos tipos de máquinas, I y II, en 12 días consecutivos y que supone que ambas máquinas tienen la misma producción diaria. Deseamos contrastar la hipótesis $H_0$ de que no hay diferencia entre las máquinas: que las diferencias observadas se deben simplemente al azar, lo que equivale a decir que las muestras proceden de la misma población.

Un sencillo test no paramétrico en este caso de muestras emparejadas la proporciona el *test de los signos*, que consiste en tomar la diferencia entre los números de tuercas defectuosas cada día y escribir sólo el *signo* de esa diferencia; por ejemplo, para el primer día se tiene 47-71, que es negativo. De este modo se obtiene de la Tabla 17.1 la secuencia de signos

$$- \quad - \quad + \quad - \quad - \quad - \quad + \quad - \quad + \quad - \quad - \quad - \tag{1}$$

(o sea, tres + y nueve −). Ahora bien, si fuese tan probable obtener + como −, esperaríamos seis + + y seis −. El contraste de $H$ equivale al de si una moneda es buena sabiendo que en 12 tiradas han salido 3 caras (+) y 9 cruces (−). Ello involucra a la distribución binomial del Capítulo 7. El Problema 17.1 muestra que mediante un contraste de dos colas con la distribución binomial al nivel de significación 0.05, no podemos rechazar $H_0$; esto es, no hay diferencia entre las máquinas a ese nivel.

**Tabla 17.1**

| Día | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Máquina I | 47 | 56 | 54 | 49 | 36 | 48 | 51 | 38 | 61 | 49 | 56 | 52 |
| Máquina II | 71 | 63 | 45 | 64 | 50 | 55 | 42 | 46 | 53 | 57 | 75 | 60 |

**Nota 1:** Si un día las máquinas producen el mismo número de tuercas defectuosas, aparecerá una diferencia *cero* en la secuencia (1). En tal caso podemos omitir ese par de valores muestrales y utilizar 11 en vez de 12 observaciones.

**Nota 2:** Se puede usar también una aproximación normal a la distribución binomial, mediante corrección por continuidad (véase Prob. 17.2).

Aunque el test de los signos es particularmente útil para muestras emparejadas, como en la Tabla 17.1, se puede usar también en problemas con una sola muestra (véase Probs. 17.3 y 17.4).

## EL *U*-TEST DE MANN-WHITNEY

Consideremos la Tabla 17.2, que da las resistencias de cables fabricados con dos aleaciones distintas, I y II. En esa tabla tenemos dos muestras: 8 cables de la aleación I y 10 de la II. Queremos decidir si hay o no diferencia entre las muestras, o sea, si proceden o no de una misma población. Si bien este problema se puede atacar con el contraste $t$ del Capítulo 11, es conveniente un test no paramétrico llamado el *U-test de Mann-Whitney*, o abreviadamente, *U-test*. Consiste en los siguientes pasos:

**Tabla 17.2**

| Aleación I | | | | Aleación II | | | | |
|---|---|---|---|---|---|---|---|---|
| 18.3 | 16.4 | 22.7 | 17.8 | 12.6 | 14.1 | 20.5 | 10.7 | 15.9 |
| 18.9 | 25.3 | 16.1 | 24.2 | 19.6 | 12.9 | 15.2 | 11.8 | 14.7 |

*Paso 1.* Combinar todos los valores muestrales en una ordenación del menor al mayor, y asignar rangos (en este caso de 1 a 8) a todos esos valores. Si dos o más valores muestrales son idénticos (o sea, son coincidencias), se les asigna a cada uno un rango que es la media de los rangos que les hubieran correspondido sin tal coincidencia. Si la entrada 18.9 en la Tabla 17.2 fuese 18.3, dos valores idénticos 18.3 ocuparían los rangos 12 y 13 en la ordenación, de modo que se asignaría a cada uno el rango $\frac{1}{2}(12 + 13) = 12.5$.

*Paso 2.* Hallar la suma de los rangos para cada muestra. Las denotamos $R_1$ y $R_2$, donde $N_1$ y $N_2$ son los respectivos tamaños muestrales. Por conveniencia elegimos $N_1$ que es el menor si son desiguales tales que $N_1 \leqslant N_2$. Una diferencia significativa entre las sumas de rangos $R_1$ y $R_2$ implica una diferencia significativa entre las muestras.

*Paso 3.* Para contrastar la diferencia entre las sumas de rangos, usamos el estadístico

$$U = N_1N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 \tag{2}$$

correspondiente a la muestra 1. La distribución muestral $U$ es simétrica y tiene una media y una varianza dadas por

$$\mu_U = \frac{N_1N_2}{2} \qquad \sigma_U^2 = \frac{N_1N_2(N_1 + N_2 + 1)}{12} \tag{3}$$

Si $N_1$ y $N_2$ son ambos al menos iguales a 8, resulta que la distribución de $U$ es aproximadamente normal, de manera que

$$z = \frac{U - \mu_U}{\sigma_U} \tag{4}$$

está normalmente distribuido con media 0 y varianza 1. Usando el Apéndice II, podemos entonces decidir si las muestras son significativamente diferentes. El Problema 17.5 enseña que hay diferencia significativa entre los cables al nivel 0.05.

**Nota 3:** Un valor correspondiente a la muestra 2 viene dado por el estadístico

$$U = N_1N_2 + \frac{N_2(N_2 + 1)}{2} - R_2 \tag{5}$$

y tiene la misma distribución muestral que el (2), con la media y la varianza de las fórmulas (3). El estadístico (5) está relacionado con el (2), porque si $U_1$ y $U_2$ son los valores correspondientes a los estadísticos (2) y (5), respectivamente, se tiene

$$U_1 + U_2 = N_1N_2 \tag{6}$$

Se tiene además

$$R_1 + R_2 = \frac{N(N + 1)}{2} \tag{7}$$

donde $N = N_1 + N_2$. El resultado (7) proporciona una comprobación para los cálculos.

**Nota 4:** El estadístico $U$ en (2) es el número total de veces que los valores de la muestra 1 preceden a los de la muestra 2 cuando todos los valores se ordenan de modo creciente. Ello proporciona un *método* alternativo *de recuento* para hallar $U$.

## EL *H*-TEST DE KRUSKAL-WALLIS

El $U$-test es un test no paramétrico para decidir si dos muestras provienen o no de la misma población. Una generalización para $k$ muestras la da el *H-test de Kruskal-Wallis*, o simplemente *H-test*.

El *H-test* puede describirse como sigue: Sean $k$ muestras de tamaños $N_1, N_2, ..., N_k$, con tamaño suma total $N = N_1 + N_2 + \cdots + N_k$. Supongamos que los datos de todas las muestras se ordenan y que las sumas de rangos para las $k$ muestras son $R_1, R_2, ..., R_k$, respectivamente. Si definimos el estadístico

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{N_j} - 3(N+1) \tag{8}$$

se puede demostrar que su distribución de muestreo es muy próxima a una *distribución ji-cuadrado* con $k - 1$ grados de libertad, supuesto que $N_1, N_2, ..., N_k$ son al menos 5 todos ellos.

El *H*-test nos da un test no paramétrico en el *análisis de varianza* para experimentos de un factor, y admite generalización.

## EL *H*-TEST CORREGIDO POR COINCIDENCIAS

En caso de haber demasiadas coincidencias entre las observaciones en los datos muestrales, el valor de $H$ dado por (8) es menor de lo que debiera. El valor corregido de $H$, denotado $H$, se obtiene dividiendo el valor dado en (8) por el factor de corrección

$$1 - \frac{\sum(T^3 - T)}{N^3 - N} \tag{9}$$

donde $T$ es el número de coincidencias correspondientes a cada observación y donde la suma se toma sobre todas las observaciones. Si no hay coincidencias, $T = 0$ y el factor (9) se reduce a 1, así que no se precisa corrección. En la práctica, la corrección suele ser despreciable (o sea, no suficiente para cambiar la decisión).

## EL TEST DE LAS RACHAS PARA EL CARACTER ALEATORIO

Aunque la palabra «aleatorio» ha sido utilizada con frecuencia en este libro (por ejemplo en «muestreo aleatorio»), no hemos visto ningún criterio de aleatoriedad. Un test no paramétrico a tal fin lo proporciona la *teoría de rachas*.

Para entender qué son las rachas (o escalones) consideremos una secuencia con dos símbolos, $a$ y $b$ tal como

$$a \;\; a \mid b \;\; b \;\; b \mid a \mid b \;\; b \mid a \;\; a \;\; a \;\; a \;\; a \mid b \;\; b \;\; b \mid a \;\; a \;\; a \;\; a \mid \tag{10}$$

Al tirar una moneda, por ejemplo, $a$ sería «cara» y $b$ «cruz»; en el muestreo de tuercas defectuosas, $a$ sería «defectuosa» y $b$ «no defectuosa».

Una *racha* se define como un conjunto de símbolos idénticos (o relacionados) contenido entre dos símbolos diferentes o uno sólo si estamos al comienzo o al final de la secuencia. Leyendo de izquierda a derecha en la secuencia (10) la primera racha, indicada por una barra vertical, consiste de dos aes, la segunda de tres bes, la tercera de una $a$, etc. Hay siete rachas en total.

Parece claro que existe relación entre aleatoriedad y el número de rachas. Así, para la secuencia

$$a \mid b \mid a \mid b \mid a \mid b \mid a \mid b \mid a \mid b \mid a \mid b \mid \tag{11}$$

hay un *esquema cíclico*, en el que vamos de $a$ a $b$, vuelta al $a$, etc, que difícilmente puede ser aleatorio. En ese caso tenemos *demasiadas* rachas (de hecho, hay el máximo posible con ese número de letras $a$ y $b$).

Por otra parte, para la secuencia

$$a\ a\ a\ a\ a\ a \mid b\ b\ b\ b \mid a\ a\ a\ a\ a \mid b\ b\ b \mid \tag{12}$$

parece haber un *esquema de tendencia* o de inercia, en el que las *aes* y las *bes* están agrupadas. En este caso hay *demasiado pocas* rachas, y no consideraríamos tampoco aleatoria a esa secuencia.

Así pues, una secuencia se considera no aleatoria si hay demasiadas o demasiado pocas rachas, y aleatoria en los demás casos. Para cuantificar esa idea, supongamos que formamos todas las posibles secuencias con $N_1$ *aes* y $N_2$ *bes*, para un total de $N$ símbolos, $(N_1 + N_2 = N)$. La colección de todas esas secuencias nos da una distribución muestral. Cada secuencia tiene asociado un número de rachas, denotado por $V$. De este modo nos vemos conducidos a la distribución muestral del estadístico $V$. Se demuestra que esta distribución tiene media y varianza dadas por

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1 \qquad \sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)} \tag{13}$$

Mediante las fórmulas (13), podemos contrastar la hipótesis de aleatoriedad a niveles de significación apropiados. Resulta que si $N_1$ y $N_2$ son ambos al menos iguales a 8, entonces la distribución muestral de $V$ es muy próxima una distribución normal. Luego

$$z = \frac{V - \mu_V}{\sigma_V} \tag{14}$$

está normalmente distribuido con media 0 y varianza 1, y se puede utilizar el Apéndice II.

## OTRAS APLICACIONES DEL TEST DE LAS RACHAS

He aquí otras aplicaciones del test de las rachas en problemas de estadística:

1. **Test sobre- y bajo-mediana para la aleatoriedad de datos numéricos.** Para determinar si unos datos numéricos (como los tomados en una muestra) son aleatorios, los colocamos primero en el *mismo orden* en que fueron tomados, hallamos la mediana y sustituimos cada entrada por la letra $a$ o $b$ según que ese valor esté *sobre* o *bajo* la mediana. Si un valor coincide con la mediana, lo suprimimos. La muestra es aleatoria o no según lo sea la secuencia de *aes* y *bes* así obtenida. (Véase Prob. 17.20).
2. **Diferencias en poblaciones de las que se toman muestras.** Sean dos muestras de tamaños $m$ y $n$, denotadas por $a_1, a_2, \ldots, a_m$ y $b_1\ b_2, \ldots, b_n$. Para decidir si las muestras proceden o no de una misma población, colocamos los $m + n$ valores en orden creciente. Si varios valores coinciden, se ordenan por algún procedimiento de azar (usando números aleatorios, por

ejemplo). Si la secuencia resultante es aleatoria, concluimos que las dos muestras no son realmente diferentes y provienen, por tanto, de una misma población; si no es aleatoria, no podemos sacar esa conclusión. Este test proporciona una alternativa al $U$-test de Mann-Whitney (véase Prob. 17.21).

## CORRELACION DE RANGO DE SPEARMAN

Se pueden usar también métodos no paramétricos para medir la correlación de dos variables $X$ e $Y$. En lugar de usar valores precisos de las variables, o cuando tal precisión no es alcanzable, a los datos se les pueden asignar un rango de 1 a $N$ ordenándolos por su tamaño, importancia, etc. Si $X$ e $Y$ tienen asignado un rango así, el *coeficiente de correlación de rango*, o *fórmula de Spearman para la correlación de rango* (como se suele llamar), viene dado por

$$r_S = 1 - \frac{6\sum D^2}{N(N^2 - 1)} \tag{15}$$

donde $D$ denota la diferencia entre los rangos de valores correspondientes de $X$ e $Y$, y donde $N$ es el número de pares de valores $(X, Y)$ en los datos.

## PROBLEMAS RESUELTOS

### EL TEST DE LOS SIGNOS

**17.1.** Con referencia a la Tabla 17.1, contrastar la hipótesis $H_0$ de que no hay diferencia entre las máquinas I y II frente a la hipótesis alternativa $H_1$ de que sí la hay, al nivel de significación 0.05

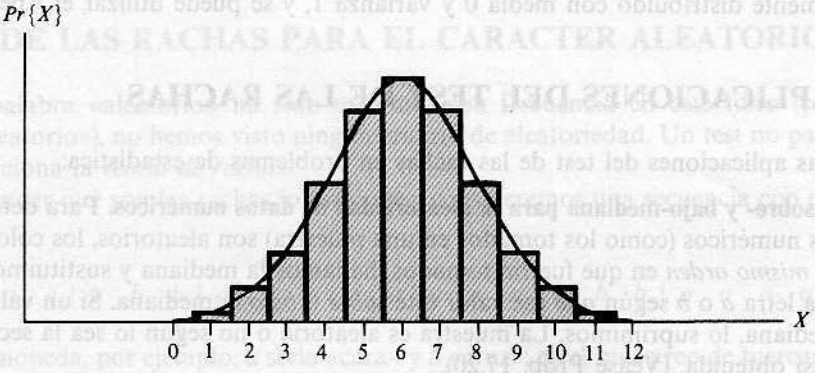



**Figura 17.1.**

**Solución**

La Figura 17.1 es un gráfico de la distribución binomial (y una aproximación normal a ella) que da

las probabilidades de $X$ caras en 12 tiradas de una moneda buena, donde $X = 0, 1, 2, \ldots, 12$. Del Capítulo 7 sabemos que la probabilidad de $X$ caras es

$$\Pr\{X\} = \binom{12}{X}\left(\frac{1}{2}\right)^X\left(\frac{1}{2}\right)^{12-X} = \binom{12}{X}\left(\frac{1}{2}\right)^{12}$$

de donde $\Pr\{0\} = 0.00024$, $\Pr\{1\} = 0.00293$, $\Pr\{2\} = 0.01611$ y $\Pr\{3\} = 0.05371$.

Como $H_1$ es la hipótesis de que hay *diferencia* entre las máquinas, no que la $I$ sea *mejor* que la II, usamos un contraste de dos colas. Al nivel de significación 0.05 cada cola tiene asociada la probabilidad $\frac{1}{2}(0.05) = 0.025$. Ahora sumamos las probabilidades de la cola izquierda hasta que la suma sobrepase 0.025. Luego

$$\Pr\{0, 1 \text{ ó } 2 \text{ caras}\} = 0.00024 + 0.00293 + 0.01611 = \boldsymbol{0.01928}$$

$$\boldsymbol{Pr}\{0, 1, 2 \text{ ó } 3 \text{ caras}\} = 0.00024 + 0.00293 + 0.01611 + 0.05371 = 0.07299$$

Como 0.025 es mayor que 0.01928 pero menor que 0.07299, podemos rechazar $H_0$ si el número de caras es 2 o menor (o por simetría, si es 10 o mayor); no obstante, el número de caras [los signos + en la secuencia (1)] es 3. Luego no podemos rechazar $H_0$ al nivel de significación 0.05 y debemos concluir que no hay diferencia entre las máquinas a ese nivel.

**17.2.** Rehacer el Problema 17.1 usando una aproximación normal a la distribución binomial.

**Solución**

Para lograr una aproximación normal a la distribución binomial, usaremos el hecho de que el recuento $z$ correspondiente al número de caras es

$$z = \frac{X - \mu}{\sigma} = \frac{X - Np}{\sqrt{Npq}}$$

(véase pág. 161). Como la variable $X$ para la distribución binomial es discreta mientras que para una distribución normal es continua, hacemos una *corrección por continuidad* (por ejemplo, 3 caras es realmente un valor entre 2,5 y 3,5 caras). Eso equivale a disminuir $X$ en 0,5 si $X > Np$ y a aumentar $X$ en 0,5 si $X < Np$. Ahora bien, $N = 12$, $\mu = Np = (12)(0.5) = 6$ y $\sigma = \sqrt{Npq} = \sqrt{(12)(0.5)(0.5)} = 1.73$, de modo que

$$z = \frac{(3 + 0.5) - 6}{1.73} = -1.45$$

Como esto es mayor que $-1.96$ (el valor de $z$ para el cual el área en la cola izquierda es 0.025), llegamos a la misma conclusión que en el Problema 17.1.

Nótese que $\Pr\{z \leq -1.45\} = 0.0735$, que está en buen acuerdo con la $\Pr\{X \leq 3 \text{ caras}\} = 0.07299$ del Problema 17.1.

**17.3.** La empresa PQR afirma que la vida media de un tipo de baterías que fabrica es superior a 250 horas(h). Un defensor de los consumidores desea saber si tal afirmación está justificada, y para ello mide las vidas medias de 24 baterías, con los resultados que figuran en la Tabla 17.3. Supuesto que la muestra era aleatoria, determinar si la empresa tiene razón al nivel de significación 0.05.

**Solución**

Sea $H_0$ la hipótesis de que las baterías de esa empresa tienen vida media igual a 250 h, y sea $H_1$ la hipótesis de que la vida media es mayor que 250 h. Para contrastar $H_0$, podemos usar el test de los signos. Para ello, restamos 250 a cada entrada de la Tabla 17.3 y anotamos los signos de las diferencias, tal como indica la Tabla 17.4. Vemos que hay 15 signos + y 9 signos −.

**Tabla 17.3**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 271 | 230 | 198 | 275 | 282 | 225 | 284 | 219 |
| 253 | 216 | 262 | 288 | 236 | 291 | 253 | 224 |
| 264 | 295 | 211 | 252 | 294 | 243 | 272 | 268 |

**Tabla 17.4**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| + | − | − | + | + | − | + | − |
| + | − | + | + | − | + | + | − |
| + | + | − | + | + | − | + | + |

Usando un contraste unilateral al nivel de significación 0.05, rechazaríamos $H_0$ si el recuento $z$ fuese mayor que 1.645 (Fig. 17.2). Como el $z$, usando corrección por continuidad, es

$$z = \frac{(15 - 0.5) - (24)(0.5)}{\sqrt{(24)(0.5)(0.5)}} = 1.02$$

la afirmación de la empresa no estaba justificada al nivel 0.05.

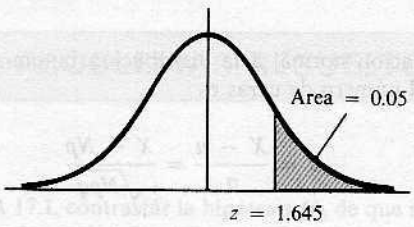



**Figura 17.2.**

**17.4.** La Tabla 17.5 recoge una muestra de 40 notas en un examen de ámbito nacional. Contrastar al nivel de significación 0.05 la hipótesis de que la nota mediana de todos los participantes es (*a*) 66 y (*b*) 75.

**Solución**

(*a*) Restando 66 de las entradas de la Tabla 17.5 y reteniendo sólo los signos de las diferencias, se obtiene la Tabla 17.6, en la que hay 23 signos +, 15 signos − y 2 ceros. Descartados los ceros, quedan 23 + y 15 −. Usando un contraste bilateral con la distribución normal con probabilidades $\frac{1}{2}(0.05) = 0.025$ en cada cola (Fig. 17.3), adoptamos la siguiente regla de decisión:

**Tabla 17.5**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 71 | 67 | 55 | 64 | 82 | 66 | 74 | 58 | 79 | 61 |
| 78 | 46 | 84 | 93 | 72 | 54 | 78 | 86 | 48 | 52 |
| 67 | 95 | 70 | 43 | 70 | 73 | 57 | 64 | 60 | 83 |
| 73 | 40 | 78 | 70 | 64 | 86 | 76 | 62 | 95 | 66 |

**Tabla 17.6**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| + | + | − | − | + | 0 | + | − | + | − |
| + | − | + | + | + | − | + | + | − | − |
| + | + | + | − | + | + | − | − | − | + |
| + | − | + | + | − | + | + | − | + | 0 |

**Tabla 17.7**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| − | − | − | − | + | − | − | − | + | − |
| + | − | + | + | − | − | + | + | − | − |
| − | + | − | − | − | − | − | − | − | + |
| − | − | + | − | − | + | + | − | + | − |

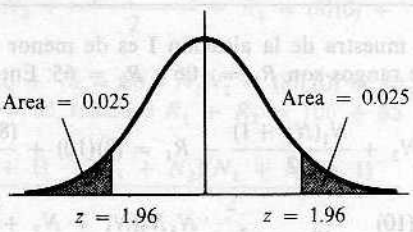



**Figura 17.3.**

Aceptar la hipótesis si $-1.96 \leqslant z \leqslant 1.96$.
Rechazarla en caso contrario.

Como
$$z = \frac{X - Np}{\sqrt{Npq}} = \frac{(23 - 0.5) - (38)(0.5)}{\sqrt{(38)(0.5)(0.5)}} = 1.14$$

aceptamos la hipótesis de que la mediana es 66, al nivel 0.05.

Nótese que podríamos haber usado 15, el número de signos −. En ese caso,

$$z = \frac{(15 + 0.5) - (38)(0.5)}{\sqrt{(38)(0.5)(0.5)}} = -1.14$$

con la misma conclusión.

(*b*) Restando 75 de las entradas de la Tabla 17.5 se llega a la Tabla 17.7, con 13 + y 27 −. Como

$$z = \frac{(13 + 0.5) - (40)(0.5)}{\sqrt{(40)(0.5)(0.5)}} = -2.06$$

rechazamos la hipótesis de que la mediana es 75, al nivel 0.05.

Por este método, podemos llegar al intervalo de confianza del 95% para la nota mediana del examen. (Véase Prob. 17.30.)

## EL U-TEST DE MANN-WHITNEY

**17.5.** Con referencia a la Tabla 17.2, determinar si hay diferencia entre los cables de aleaciones I y II, al nivel de significación 0.05.

**Solución**

Seguimos los pasos 1, 2 y 3 descritos antes en este capítulo.

*Paso 1.* Combinando los 18 valores de la muestra en una ordenación de menor a mayor tenemos la primera fila de la Tabla 17.8. La segunda fila les asigna rango de 1 a 18.

*Paso 2.* Para hallar la suma de los rangos de cada muestra, reescribimos la Tabla 17.2 usando los rangos asociados de la Tabla 17.8, lo que nos da la Tabla 17.9. La suma de los rangos es 106 para la aleación I y 65 para la aleación II.

**Tabla 17.8**

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10.7 | 11.8 | 12.6 | 12.9 | 14.1 | 14.7 | 15.2 | 15.9 | 16.1 | 16.4 | 17.8 | 18.3 | 18.9 | 19.6 | 20.5 | 22.7 | 24.2 | 25.3 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |

*Paso 3.* Puesto que la muestra de la aleación I es de menor tamaño, $N_1 = 8$ y $N_2 = 10$. Las correspondientes sumas de rangos son $R_1 = 106$ y $R_2 = 65$. Entonces

$$U = N_1N_2 + \frac{N_1(N_1+1)}{2} - R_1 = (8)(10) + \frac{(8)(9)}{2} - 106 = 10$$

$$\mu_U = \frac{N_1N_2}{2} = \frac{(8)(10)}{2} = 40 \qquad \sigma_U^2 = \frac{N_1N_2(N_1+N_2+1)}{12} = \frac{(8)(10)(19)}{12} = 126.67$$

**Tabla 17.9**

| Aleación I | | Aleación II | |
|---|---|---|---|
| Resistencia del cable | Rango | Resistencia del cable | Rango |
| 18.3 | 12 | 12.6 | 3 |
| 16.4 | 10 | 14.1 | 5 |
| 22.7 | 16 | 20.5 | 15 |
| 17.8 | 11 | 10.7 | 1 |
| 18.9 | 13 | 15.9 | 8 |
| 25.3 | 18 | 19.6 | 14 |
| 16.1 | 9 | 12.9 | 4 |
| 24.2 | 17 | 15.2 | 7 |
| | Suma 106 | 11.8 | 2 |
| | | 14.7 | 6 |
| | | | Suma 65 |

Así pues $\sigma_U = 11.25$ y

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{10 - 40}{11.25} = -2.67$$

Como la hipótesis $H_0$ que estamos estudiando es que *no* hay diferencia entre las aleaciones, se requiere un contraste de dos colas. Al nivel de significación 0.05, tenemos como regla de decisión:

Aceptar $H_0$ si $-1.96 \leqslant z \leqslant 1.96$.
Rechazarla en caso contrario.

Como $z = -2.67$, rechazamos $H_0$ y concluimos que hay diferencia entre las dos aleaciones al nivel 0.05.

**17.6.** Comprobar los resultados (6) y (7) de este capítulo para los datos del Problema 17.5.

**Solución**

(*a*) Dado que las muestras 1 y 2 resultan valores para $U$ dados por

$$U_1 = N_1N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 = (8)(10) + \frac{(8)(9)}{2} - 106 = 10$$

$$U_2 = N_1N_2 + \frac{N_2(N_2 + 1)}{2} - R_2 = (8)(0) + \frac{(10)(11)}{2} - 65 = 70$$

tenemos $U_1 + U_2 = 10 + 70 = 80$ y $N_1N_2 = (8)(10) = 80$.

(*b*) Como $R_1 = 106$ y $R_2 = 65$, tenemos $R_1 + R_2 = 106 + 65 = 171$ y

$$\frac{N(N + 1)}{2} = \frac{(N_1 + N_2)(N_1 + N_2 + 1)}{2} = \frac{(18)(19)}{2} = 171$$

**17.7.** Resolver el Problema 17.5 usando el estadístico $U$ para la muestra de la aleación II.

**Solución**

Para la muestra de la aleación II,

$$U = N_1N_2 + \frac{N_2(N_2 + 1)}{2} - R_2 = (8)(10) + \frac{(10)(11)}{2} - 65 = 70$$

así que

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{70 - 40}{11.25} = 2.67$$

Este valor de $z$ es el *negativo* del $z$ del Problema 17.5, y se usa la cola derecha de la distribución normal en vez de la izquierda. Ya que este valor de $z$ también cae fuera de $-1.96 \leqslant z \leqslant 1.96$, la conclusión es la misma que en el Problema 17.5.

**17.8.** Un profesor de psicología tiene dos clases, una matinal de 9 estudiantes y otra vespertina de 12. En un examen común a todos ellos, las notas fueron las que recoge la Tabla 17.10. ¿Podemos concluir al nivel de significación 0.05 que la clase de la mañana es peor que la de la tarde?

**Tabla 17.10**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Clase matinal | 73 | 87 | 79 | 75 | 82 | 66 | 95 | 75 | 70 | | | |
| Clase vespertina | 86 | 81 | 84 | 88 | 90 | 85 | 84 | 92 | 83 | 91 | 53 | 84 |

**Solución**

*Paso 1.* La Tabla 17.11 muestra la ordenación de notas y rangos. Nótese que el rango para las dos notas de 75 es $\frac{1}{2}(5 + 6) = 5.5$, mientras que para las tres de 84 es $\frac{1}{3}(11 + 12 + 13) = 12$.

*Paso 2.* Reescribiendo la Tabla 17.10 en términos de rangos obtenemos la Tabla 17.12.

*Comprobación:* $R_1 = 73$, $R_2 = 158$ y $N = N_1 + N_2 = 9 + 12 = 21$; luego $R_1 + R_2 = 73 + 158 = 231$ y

$$\frac{N(N+1)}{2} = \frac{(21)(22)}{2} = 231 = R_1 + R_2$$

**Tabla 17.11**

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 66 | 70 | 73 | 75 | 75 | 79 | 81 | 82 | 83 | 84 | 84 | 84 | 85 | 86 | 87 | 88 | 90 | 91 | 92 | 95 |
| 1 | 2 | 3 | 5.5 | 7 | | 8 | 9 | 10 | 12 | | 14 | | 15 | 16 | 17 | 18 | 19 | 20 | 21 | |

**Tabla 17.12**

| | | | | | | | | | | | | | Suma de rangos |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Clase matinal | 4 | 16 | 7 | 5.5 | 9 | 2 | 21 | 5.5 | 3 | | | | 73 |
| Clase vespertina | 15 | 8 | 12 | 17 | 18 | 14 | 12 | 20 | 10 | 19 | 1 | 12 | 158 |

*Paso 3.*

$$U = N_1N_2 + \frac{N_1(N_1+1)}{2} - R_1 = (9)(12) + \frac{(9)(10)}{2} - 73 = 80$$

$$\mu_U = \frac{N_1N_2}{2} = \frac{(9)(12)}{2} = 54 \qquad \sigma_U^2 = \frac{N_1N_2(N_1+N_2+1)}{12} = \frac{(9)(12)(22)}{12} = 198$$

Por tanto,

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{80 - 54}{14.07} = 1.85$$

Puesto que deseamos contrastar la hipótesis $H_0$ de que la clase de la mañana es peor que la otra frente a la $H_0$ de que no hay diferencia al nivel 0.05, necesitamos un contraste unilateral. Con referencia a la Figura 17.2, que se aplica aquí, tenemos la regla de decisión:

Aceptar $H_0$ si $z \leq 1.645$.
Rechazar $H_0$ si $z > 1.645$.

Como el valor real es $z = 1.85 > 1.645$, rechazamos $H_0$ y concluimos que la clase matinal es peor al nivel 0.05. Esa conclusión no se mantiene, sin embargo, al nivel de significación 0.01 (véase Problema 17.33).

**17.9.** Hallar $U$ para los datos de la Tabla 17.13, usando (*a*) la fórmula (2) de este capítulo y (*b*) el método de recuentos descrito en la Nota 4 de este capítulo.

**Solución**

(*a*) Ordenando los datos de ambas muestras en orden de magnitud creciente y asignándoles rangos

de 1 a 5, se llega a la Tabla 17.14. Sustituyendo los datos de la Tabla 17.13 por los rangos correspondientes se obtiene la Tabla 17.15, en la cual las sumas de rangos son $R_1 = 5$ y $R_2 = 10$. Como $N_1 = 2$ y $N_2 = 3$, el valor de $U$ para la muestra 1 es

$$U = N_1N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 = (2)(3) + \frac{(2)(3)}{2} - 5 = 4$$

El valor de $U$ para la muestra 2 se halla de forma similar y es $U = 2$.

**Tabla 17.13**

| | | | |
|---|---|---|---|
| Muestra 1 | 22 | 10 | |
| Muestra 2 | 17 | 25 | 14 |

**Tabla 17.14**

| | | | | | |
|---|---|---|---|---|---|
| Datos | 10 | 14 | 17 | 22 | 25 |
| Rango | 1 | 2 | 3 | 4 | 5 |

**Tabla 17.15**

| | | | | Suma de rangos |
|---|---|---|---|---|
| Muestra 1 | 4 | 1 | | 5 |
| Muestra 2 | 3 | 5 | 2 | 10 |

(*b*) Sustituyamos los valores muestrales en la Tabla 17.14 por I o II, según la muestra a la que el valor pertenezca. Entonces la primera línea de la Tabla 17.14 pasa a ser

| | | | | | | |
|---|---|---|---|---|---|---|
| Datos | I | II | II | I | I | II |

De ahí vemos que

Número de valores de la muestra 1 que preceden al primero de la muestra 2 = 1
Número de valores de la muestra 1 que preceden al segundo de la muestra 2 = 1
Número de valores de la muestra 1 que preceden al tercero de la muestra 2 = 2
Total = 4

Luego el valor de $U$ correspondiente a la muestra 1 es 4.

Análogamente se tiene

Número de valores de la muestra 2 que preceden al primero de la muestra 1 = 0
Número de valores de la muestra 2 que preceden al segundo de la muestra 1 = 2
Total = 2

Luego el valor de $U$ para la muestra 2 es 2.

Nótese que como $N_1 = 2$ y $N_2 = 3$, estos valores satisfacen $U_1 + U_2 = N_1N_2$; es decir, $4 + 2 = (2)(3) = 6$.

**17.10.** Se toman dos muestras sin reposición de una población que consiste en los valores 7, 12 y 15: la primera muestra consta de un solo valor y la segunda de dos valores. [Entre ambas muestras cubren toda la población.]

(*a*) Hallar la distribución de muestreo de $U$ y su gráfico.
(*b*) Hallar la media y la varianza de esa distribución.
(*c*) Comprobar los resultados de la parte (*b*) mediante las fórmulas (3) de este capítulo.

**Solución**

(*a*) Escogemos muestreo sin reposición para evitar coincidencias, que ocurrirían si, por ejemplo, el valor 12 apareciese en ambas muestras.

Hay $3 \cdot 2 = 6$ posibilidades para escoger las muestras, como indica la Tabla 17.16. Debemos observar que podríamos usar los rangos 1, 2 y 3 en vez de 7, 12 y 15. El valor de $U$ en la Tabla 17.16, es el hallado para la muestra 1, pero si se usara el $U$ para la muestra 2, la distribución sería la misma.

**Tabla 17.16**

| Muestra 1 | Muestra 2 | $U$ |
|---|---|---|
| 7 | 12 15 | 2 |
| 7 | 15 12 | 2 |
| 12 | 7 15 | 1 |
| 12 | 15 7 | 1 |
| 15 | 7 12 | 0 |
| 15 | 12 7 | 0 |

El gráfico de esta distribución aparece en la Figura 17.4, donde $f$ es la frecuencia. La distribución de probabilidad de $U$ también puede representarse; en este caso $\Pr\{U = 0\} = \Pr\{U = 1\} = \Pr\{U = 2\} = \frac{1}{3}$. El gráfico pedido es el mismo que el de la Figura 17.4, pero con las ordenadas 1 y 2 sustituidas por $\frac{1}{6}$ y $\frac{1}{3}$, respectivamente.

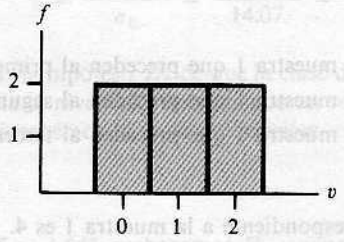



**Figura 17.4.**

(*b*) La media y la varianza halladas a partir de la Tabla 17.15 vienen dadas por

$$\mu_U = \frac{2 + 2 + 1 + 1 + 0 + 0}{6} = 1$$

$$\sigma_U^2 = \frac{(2 - 1)^2 + (2 - 1)^2 + (1 - 1)^2 + (1 - 1)^2 + (0 - 1)^2 + (0 - 1)^2}{6} = \frac{2}{3}$$

(*c*) Por las fórmulas (3),

$$\mu_U = \frac{N_1 N_2}{2} = \frac{(1)(12)}{2} = 1$$

$$\sigma_U^2 = \frac{N_1 N_2 (N_1 + N_2 + 1)}{12} = \frac{(1)(2)(1 + 2 + 1)}{12} = \frac{2}{3}$$

en buen acuerdo con la parte (*a*).

**17.11.** (*a*) Hallar la distribución muestral de $U$ en el Problema 17.9 y representarla gráficamente.
(*b*) Representar la correspondiente distribución de probabilidad de $U$.
(*c*) Obtener la media y la varianza de $U$ directamente de los resultados de la parte (*a*).
(*d*) Verificar la parte (*c*) usando las fórmulas (3) de este capítulo.

**Solución**

(*a*) En este caso hay $5 \cdot 4 \cdot 3 \cdot 2 = 120$ posibilidades para escoger valores en las dos muestras y el método del Problema 17.9 es demasiado laborioso. Para simplificar el proceso, vamos a concentrarnos en la muestra menor (de tamaño $N_1 = 2$) y las posibles sumas de rangos, $R$. La suma de los rangos para la muestra 1 es *mínima* cuando la muestra consiste en los dos números de rango más bajo (1, 2); entonces $R_1 = 1 + 2 = 3$. Análogamente, es *máxima* cuando la muestra 1 consta de los números de rango más alto (4, 5); entonces $R_1 = 4 + 5 = 9$. Luego $R_1$ varía de 3 a 9.

La columna 1 de la Tabla 17.17 da esos valores de $R_1$ (desde 3 hasta 9), y la columna 2 da los correspondientes valores en la muestra 1 cuya suma es $R_1$. La columna 3 da la frecuencia (o número) de muestras con suma $R_1$; por ejemplo, hay $f = 2$ muestras con $R_1 = 5$. Como $N_1 = 2$ $N_2 = 3$, tenemos

$$U = N_1 N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 = (2)(3) + \frac{(2)(3)}{2} - R_1 = 9 - R_1$$

**Tabla 17.17**

| $R_1$ | Valores de la muestra 1 | $f$ | $U$ | $\Pr\{U = R_1\}$ |
|---|---|---|---|---|
| 3 | (1, 2) | 1 | 6 | 0.1 |
| 4 | (1, 3) | 1 | 5 | 0.1 |
| 5 | (1, 4), (2, 3) | 2 | 4 | 0.2 |
| 6 | (1, 5), (2, 4) | 2 | 3 | 0.2 |
| 7 | (2, 5), (3, 4) | 2 | 2 | 0.2 |
| 8 | (3, 5) | 1 | 1 | 0.1 |
| 9 | (4, 5) | 1 | 0 | 0.1 |

Hallamos los correspondientes valores de $U$ en la columna 4; nótese que cuando $R_1$ varía de 3 a 9, $U$ varía de 6 a 0. La distribución muestral viene dada por las columnas 3 y 4, y su gráfico por la Figura 17.5

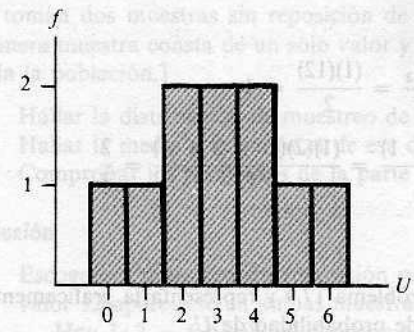



Figura 17.5.

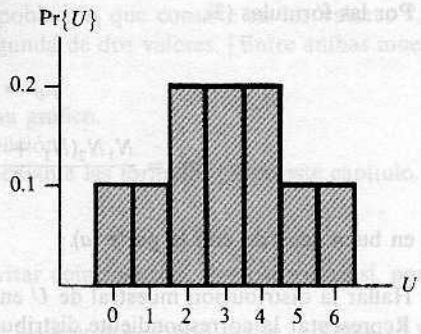



Figura 17.6.

(*b*) La probabilidad de que $U = R_1$ (es decir, $\Pr\{U = R_1\}$) aparece en la columna 5 de la Tabla 17.17 y se obtiene hallando la frecuencia relativa, cociente de cada frecuencia $f$ por la suma de todas las frecuencias, o sea 10; así, $\Pr\{U = 5\} = \frac{2}{10} = 0.2$. El gráfico de la distribución de probabilidad se muestra en la Figura 17.6.

(*c*) De las columnas 3 y 4 de la Tabla 17.17, se deduce

$$\mu_U = \bar{U} = \frac{\sum fU}{\sum f} = \frac{(1)(6)+(1)(5)+(2)(4)+(2)(3)+(2)(2)+(1)(1)+(1)(0)}{1+1+2+2+2+1+1} = 3$$

$$\sigma_U^2 = \frac{\sum f(U-\bar{U})^2}{\sum f}$$

$$= \frac{(1)(6-3)^2+(1)(5-3)^2+(2)(4-3)^2+(2)(3-3)^2+(2)(2-3)^2+(1)(1-3)^2+(1)(0-3)^2}{10} = 3$$

*Otro método*

$$\sigma_U^2 = \overline{U^2} - \bar{U}^2 = \frac{(1)(6)^2+(1)(5)^2+(2)(4)^2+(2)(3)^2+(2)(2)^2+(1)(1)^2+(1)(0)^2}{10} - (3)^2 = 3$$

(*d*) Por las fórmulas (3), usando $N_1 = 2$ y $N_2 = 3$, tenemos

$$\mu_U = \frac{N_1N_2}{2} = \frac{(2)(3)}{2} = 3 \qquad \sigma_U^2 = \frac{N_1N_2(N_1+N_2+1)}{12} = \frac{(2)(3)(6)}{12} = 3$$

**17.12.** Si $N$ números en un conjunto se enumeran con rangos de 1 a $N$, probar que la suma de rangos es $[N(N+1)]/2$.

**Solución**

Si llamamos $R$ a la suma de rangos, tenemos

$$R = 1 + 2 + 3 + \cdots + (N-1) + N \tag{16}$$

$$R = N + (N-1) + (N-2) + \cdots + 2 + 1 \tag{17}$$

donde la suma en (17) se obtiene escribiendo la de (16) hacia atrás. Sumando las ecuaciones (16) y (17) resulta

$$2R = (N + 1) + (N + 1) + (N + 1) + \cdots + (N + 1) + (N + 1) = N(N + 1)$$

ya que $(N + 1)$ aparece $N$ veces en la suma; así pues, $R = [N(N + 1)]/2$. Esto se puede obtener también recurriendo al álgebra elemental de progresiones aritméticas.

**17.13.** Si $R_1$ y $R_2$ son las respectivas sumas de rangos para las muestras 1 y 2 en el $U$-test, probar que $R_1 + R_2 = [N(N + 1)]/2$.

**Solución**

Suponemos que no hay coincidencias en los datos muestrales. Entonces $R_1$ ha de ser la suma de los rangos (números) del conjunto 1, 2, 3, .., $N$ y $R_2$, la suma de los restantes rangos. Así que la suma $R_1 + R_2$ debe ser la suma de todos los rangos del conjunto; es decir, $R_1 + R_2 = 1 + 2 + 3 + \cdots + N = [N(N + 1)]/2$ por el Problema 17.12.

## EL $H$-TEST DE KRUSKAL-WALLIS

**17.14.** Una empresa desea comprar una de las cinco máquinas distintas $A$, $B$, $C$, $D$ y $E$. En un experimento diseñado para saber si hay diferencia en la eficacia de tales máquinas, cinco operarios expertos trabajaron cada uno con las máquinas un mismo tiempo en cada una. Los resultados se recogen en la Tabla 17.18, en número de unidades producidas. Contrastar la hipótesis de que no hay diferencia entre ellas al nivel de significación ($a$) 0.05 y ($b$) 0.01.

**Tabla 17.18**

| | | | | | |
|---|---|---|---|---|---|
| $A$ | 68 | 72 | 77 | 42 | 53 |
| $B$ | 72 | 53 | 63 | 53 | 48 |
| $C$ | 60 | 82 | 64 | 75 | 72 |
| $D$ | 48 | 61 | 57 | 64 | 50 |
| E | 64 | 65 | 70 | 68 | 53 |

**Tabla 17.19**

| | | | | | | Suma de rangos |
|---|---|---|---|---|---|---|
| $A$ | 17.5 | 21 | 24 | 1 | 6.5 | 70 |
| $B$ | 21 | 6.5 | 12 | 6.5 | 2.5 | 48.5 |
| $C$ | 10 | 25 | 14 | 23 | 21 | 93 |
| $D$ | 2.5 | 11 | 9 | 14 | 4 | 40.5 |
| E | 14 | 16 | 19 | 17.5 | 6.5 | 73 |

**Solución**

Como hay 5 muestras ($A$, $B$, $C$, $D$ y $E$), $k = 5$. Y como cada muestra consta de 5 valores, tenemos $N_1 = N_2 = N_3 = N_4 = N_5 = 5$ y $N = N_1 + N_2 + N_3 + N_4 + N_5 = 25$. Ordenando todos los valores en orden creciente de magnitud y asignando rangos apropiados a las coincidencias, cambiamos la Tabla 17.18 por la 17.19, cuya columna de la derecha da la suma de rangos. Vemos de la Tabla 17.19 que $R_1 = 70$, $R_2 = 48.5$, $R_3 = 93$, $R_4 = 40.5$ y $R_5 = 73$. Luego

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{N_j} - 3(N + 1)$$

$$= \frac{12}{(25)(26)} \left[ \frac{(70)^2}{5} + + \frac{(48.5)^2}{5} + \frac{(93)^2}{5} + \frac{(40.5)^2}{5} + \frac{(73)^2}{5} \right] - 3(26) = 6.44$$

Para $k - 1 = 4$ grados de libertad al nivel de significación 0.05, por el Apéndice IV sabemos que $\chi^2_{.95} = 9.49$. Puesto que $6.44 < 9.49$, no podemos rechazar la hipótesis de igualdad entre la máquinas al nivel 0.05 y, por tanto, tampoco al 0.01. En otras palabras, podemos aceptar la hipótesis de que no hay diferencia entre las máquinas a ambos niveles (o reservar nuestra opinión).

Nótese que ya hemos resuelto este problema mediante análisis de varianza (véase Prob. 16.8) y llegamos a la misma conclusión.

**17.15.** Repetir el Problema 17.14 haciendo corrección por coincidencias.

**Solución**

La Tabla 17.20 da el número de coincidencias correspondientes a cada una de las observaciones con coincidencias. Por ejemplo, 48 aparece dos veces, de donde $T = 2$, y 53 aparece cuatro veces, luego $T = 4$. Calculando $T^3 - T$ para cada valor de $T$ y sumando, encontramos que $\sum(T^3 - T) = 6 + 60 + 24 + 6 + 24 = 120$, como indica la Tabla 17.20. Entonces, como $N = 25$, el factor de corrección es

$$1 - \frac{\sum(T^3 - T)}{N^3 - N} = 1 - \frac{120}{(25)^3 - 25} = 0.9923$$

y el valor corregido de $H$ es

$$H_c = \frac{6.44}{0.9923} = 6.49$$

Esta corrección no es suficiente para cambiar la decisión del Problema 17.14.

**Tabla 17.20**

| Observación | 48 | 53 | 64 | 68 | 72 | |
|---|---|---|---|---|---|---|
| Número de coincidencias ($T$) | 2 | 4 | 3 | 2 | 3 | |
| $T^3 - T$ | 6 | 60 | 24 | 6 | 24 | $\sum(T^3 - T) = 120$ |

**17.16.** Se toman al azar tres muestras de una población. Al ordenar los datos de acuerdo con el rango se obtiene la Tabla 17.21. Determinar si hay diferencia entre las muestras al nivel de significación (*a*) 0.05 y (*b*) 0.01.

**Solución**

Aquí $k = 3$, $N_1 = 4$, $N_2 = 3$, $N_3 = 5$, $N = N_1 + N_2 + N_3 = 12$, $R_1 = 7 + 4 + 6 + 10 = 27$, $R_2 = 11 + 9 + 12 = 32$ y $R_3 = 5 + 1 + 3 + 8 + 19$. Por tanto,

$$H = \frac{12}{N(N+1)}\sum_{j=1}^{k}\frac{R_j^2}{N_j} - 3(N+1) = \frac{12}{(12)(13)}\left[\frac{(27)^2}{4} + \frac{(32)^2}{3} + \frac{(19)^2}{5}\right] - 3(13) = 6.83$$

(*a*) Para $k - 1 = 3 - 1 = 2$ grados de libertad, $\chi^2_{.95} = 5.99$. Luego, como $6.83 > 5.99$, concluimos que hay diferencia significativa entre las muestras al nivel 0.05.

(*b*) Para 2 grados de libertad, $\chi^2_{.95} = 9.21$. Luego, como $6.83 < 9.21$ no podemos concluir que haya diferencia al nivel 0.01.

**Tabla 17.21**

| | |
|---|---|
| Muestra 1 | 7 4 6 10 |
| Muestra 2 | 11 9 12 |
| Muestra 3 | 5 1 3 8 2 |

## EL TEST DE LAS RACHAS PARA EL CARACTER ALEATORIO

**17.17.** En 30 tiradas de una moneda se ha obtenido la siguiente secuencia de caras ($H$) y cruces ($T$):

H T T H T H H H T H H T T H T
H T H H T H T T H T H H T H T

(*a*) Hallar el número de rachas, $V$.
(*b*) Decidir al nivel de significación 0.05 si la secuencia es aleatoria.

**Solución**

(*a*) Separando las rachas con barras verticales, vemos en

H | T T | H | T | H H H | T | H H | T T | H | T |
H | T | H H | T | H | T T | H | T | H H | T | H | T |

que el número de rachas es $V = 22$.

(*b*) Hay $N_1 = 16$ caras y $N_2 = 14$ cruces en la muestra dada, y por la parte (*a*) sabemos que el número de rachas es $V = 22$. Luego de (13) se deduce

$$\mu_V = \frac{2(16)(14)}{16 + 14} + 1 = 15.93 \qquad \sigma_V^2 = \frac{2(16)(14)[2(16)(14) - 16 - 14]}{(16 + 14)^2(16 + 14 - 1)} = 7.175$$

o sea $\sigma_V = 2.679$. El $z$ correspondiente a $V = 22$ es, en consecuencia,

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{22 - 15.93}{2.679} = 2.27$$

Ahora bien, para un contraste bilateral al nivel de significación 0.05, aceptaríamos la hipótesis $H_0$ de aleatoriedad si $-1.96 \leq z \leq 1.96$ y la rechazaríamos en caso contrario (véase Fig. 17.7). Como el valor calculado de $z$ es $2.27 > 1.96$, concluimos que los lanzamientos no son aleatorios al nivel 0.05. El test nos ha hecho ver que hay *demasiadas* rachas, sugiriendo un *esquema cíclico*.

Si se hace corrección por continuidad, el $z$ anterior pasa a ser

$$z = \frac{(22 - 0.5) - 15.93}{2.679} = 2.08$$

y se obtiene la misma conclusión.

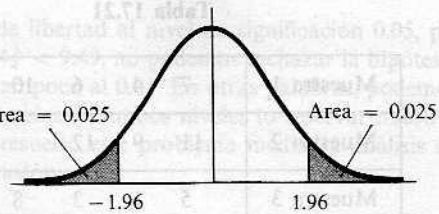



**Figura 17.7.**

**17.18.** Una muestra de 48 piezas producidas por una máquina ha dado la siguiente secuencia de piezas correctas (G) y defectuosas (D):

G G G G G G D D G G G G G G G G
G G D D D D G G G G G G D G G G
G G G G G G D D G G G G G D G G

Contrastar la aleatoriedad de esa secuencia al nivel de significación 0.05.

**Solución**

Los números de Des y Ges son $N_1 = 10$ y $N_2 = 38$, respectivamente, y el número de rachas es $V = 11$. Luego la media y la varianza vienen dadas por

$$\mu_V = \frac{2(10)(38)}{10 + 38} + 1 = 16.83 \qquad \sigma_V^2 = \frac{2(10)(38)[2(10)(38) - 10 - 38]}{(10 + 38)^2(10 + 38 - 1)} = 4.997$$

así que $\sigma_V = 2.235$.

Para un contraste bilateral al nivel de significación 0.05, aceptaríamos la hipótesis $H_0$ de aleatoriedad si $-1.96 \leqslant z \leqslant 1.96$ (véase Fig. 17.7) y la rechazaríamos en caso contrario. Como el $z$ correspondiente a $V = 11$ es

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{11 - 16.83}{2.235} = -2.16$$

y $-2.61 < -1.96$, podemos rechazar $H_0$ al nivel 0.05.

El test pone de manifiesto que hay *demasiado pocas* rachas, indicando un hacinamiento de piezas defectuosas. En otras palabras, parece haber un *esquema de tendencia* en la producción de piezas defectuosas. Debe examinarse con más profundidad el proceso de fabricación.

**17.19.** (*a*) Formar todas las posibles secuencias consistentes en tres *aes* y dos *bes*, y dar los números de rachas $V$ para cada una de ellas.
(*b*) Obtener la distribución muestral de $V$ y su gráfico.
(*c*) Obtener la distribución de probabilidad de $V$ y su gráfico.

**Solución**

(*a*) El número de posibles secuencias de ese tipo es

$$\binom{5}{2} = \frac{5!}{2!3!} = 10$$

Estas secuencias se recogen en la Tabla 17.22, junto con el número de rachas de cada una.

(*b*) La distribución muestral de $V$ viene dada en la Tabla 17.23 (deducida de la Tabla 17.21), donde $V$ denota el número de rachas y $f$ la frecuencia. Por ejemplo, la Tabla 17.23 dice que hay 1 cinco, 4 cuatros, etc. El gráfico correspondiente se puede ver en la Figura 17.8.

**Tabla 17.22**

| Secuencia | | | | | Rachas ($V$) |
|---|---|---|---|---|---|
| *a* | *a* | *a* | *b* | *b* | 2 |
| *a* | *a* | *b* | *a* | *b* | 4 |
| *a* | *a* | *b* | *b* | *a* | 3 |
| *a* | *b* | *a* | *b* | *a* | 5 |
| *a* | *b* | *b* | *a* | *a* | 3 |
| *a* | *b* | *a* | *a* | *b* | 4 |
| *b* | *b* | *a* | *a* | *a* | 2 |
| *b* | *a* | *b* | *a* | *a* | 4 |
| *b* | *a* | *a* | *a* | *b* | 3 |
| *b* | *a* | *a* | *b* | *a* | 4 |

**Tabla 17.23**

| $V$ | $f$ |
|---|---|
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 5 | 1 |

(*c*) La distribución de probabilidad de $V$, dibujada en la Figura 17.9, se obtiene de la Tabla 17.23 dividiendo cada frecuencia por la frecuencia total $2 + 33 + 4 + 1 = 10$. Por ejemplo, $\Pr\{V = 5\} = \frac{1}{10} = 0.1$.

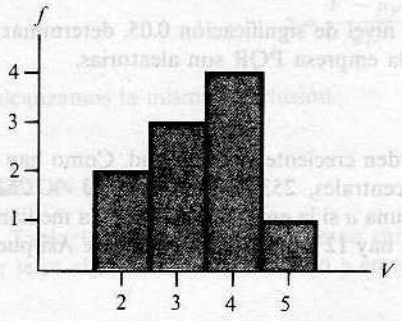



**Figura 17.8.**

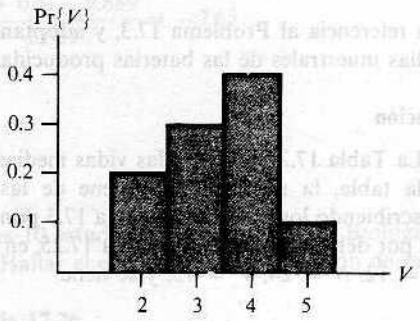



**Fig. 17.9.**

**17.20.** Hallar (*a*) la media y (*b*) la varianza del número de rachas en el Problema 17.19 directamente de los resultados allí obtenidos.

**Solución**

(*a*) De la Tabla 17.22 tenemos

$$\mu_V = \frac{2 + 4 + 3 + 5 + 3 + 4 + 2 + 4 + 3 + 4}{10} = \frac{17}{5}$$

*Otro método*

De la Tabla 17.22 el método de datos agrupados da

$$\mu_V = \frac{\sum fV}{\sum f} = \frac{(2)(2) + (3)(3) + (4)(4) + (1)(5)}{2 + 3 + 4 + 1} = \frac{17}{5}$$

(*b*) Usando el método de datos agrupados para calcular la varianza, se sigue de la Tabla 17.23 que

$$\sigma_V^2 = \frac{\sum f(V-\bar{V})^2}{\sum f} = \frac{1}{10}\left[(2)\left(2-\frac{17}{5}\right)^2 + (3)\left(3-\frac{17}{5}\right)^2 + (4)\left(4-\frac{17}{5}\right)^2 + (1)\left(5-\frac{17}{5}\right)^2\right] = \frac{21}{25}$$

*Otro método*

Como en el Capítulo 3, la varianza viene dada por

$$\sigma_V^2 = \overline{V^2} - \bar{V}^2 = \frac{(2)(2)^2 + (3)(3)^2 + (4)(4)^2 + (1)(5)^2}{10} - \left(\frac{17}{5}\right)^2 = \frac{21}{25}$$

**17.21.** Resolver el Problema 17.20 con las fórmulas (13) de este capítulo.

**Solución**

Puesto que hay tres *aes* y dos *bes*, se tiene $N_1 = 3$ y $N_2 = 2$. Así pues

(*a*) $$\mu_V = \frac{2N_1N_2}{N_1+N_2} + 1 = \frac{2(3)(2)}{3+2} + 1 = \frac{17}{5}$$

(*b*) $$\sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1+N_2)^2(N_1+N_2-1)} = \frac{2(3)(2)[2(3)(2)-3-2]}{(3+2)^2(3+2-1)} = \frac{21}{25}$$

## OTRAS APLICACIONES DEL TEST DE LAS RACHAS

**17.22.** Con referencia al Problema 17.3, y adoptando un nivel de significación 0.05, determinar si las vidas medias muestrales de las baterías producidas por la empresa PQR son aleatorias.

**Solución**

La Tabla 17.24 presenta las vidas medias en orden creciente de magnitud. Como hay 24 entradas en la tabla, la mediana se obtiene de las dos centrales, 253 y 262, es $\frac{1}{2}(253 + 262) = 257.5$. Reescribiendo los datos de la Tabla 17.3 poniendo una *a* si la entrada está sobre la mediana y una *b* si está por debajo, se llega a la Tabla 17.25, en la que hay 12 *aes*, 12 *bes* y 15 rachas. Así pues, $N_1 = 12$, $N_2 = 12$, $N = 24$, $V = 15$, y se tiene

$$\mu_V = \frac{2N_1N_2}{N_1+N_2} + 1 = \frac{2(12)(12)}{12+12} + 1 = 13 \qquad \sigma_V^2 = \frac{2(12)(12)(264)}{(24)^2(23)} = 5.739$$

luego $$z = \frac{V-\mu_V}{\sigma_V} = \frac{15-13}{2.396} = 0.835$$

Con un contraste de dos colas al nivel de significación 0.05, aceptaríamos la hipótesis de aleatoriedad si $-1.96 \leq z \leq 1.96$. Como 0.835 cae dentro de ese intervalo, concluimos que la muestra es aleatoria.

**Tabla 17.24**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 198 | 211 | 216 | 219 | 224 | 225 | 230 | 236 |
| 243 | 252 | 253 | 253 | 262 | 264 | 268 | 271 |
| 272 | 275 | 282 | 284 | 288 | 291 | 294 | 295 |

**Tabla 17.25**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *a* | *b* | *b* | *a* | *a* | *b* | *a* | *b* |
| *b* | *b* | *a* | *a* | *b* | *a* | *b* | *b* |
| *a* | *a* | *b* | *b* | *a* | *b* | *a* | *a* |

**17.23.** Resolver el Problema 17.5 aplicando el test de las rachas para decidir sobre la aleatoriedad.

**Solución**

La ordenación de todos los valores de ambas muestras aparece en la línea 1 de la Tabla 17.8. Usando los símbolos respectivos $a$ y $b$ para los datos de las muestras I y II, se convierte en

$$b \quad b \quad b \quad b \quad b \quad b \quad b \quad b \quad a \quad a \quad a \quad a \quad a \quad b \quad b \quad a \quad a \quad a$$

Como hay 4 rachas, tenemos $V = 4$, $N_1 = 8$ y $N_2 = 10$. Entonces

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1 = \frac{2(8)(10)}{18} + 1 = 9.889$$

$$\sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)} = \frac{2(8)(10)(142)}{(18)^2(17)} = 4.125$$

así que

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{4 - 9.889}{2.031} = -2.90$$

Si $H_0$ es la hipótesis de que no hay diferencia entre las aleaciones, esa es también la hipótesis de que la secuencia anterior es aleatoria. La aceptaríamos si $-1.96 \leqslant z \leqslant 1.96$ y la rechazaríamos en caso contrario. Puesto que $z = -2.90$ está fuera de ese intervalo, rechazamos $H_0$ y llegamos a la misma conclusión que en el Problema 17.5.

Nótese que si se hace corrección por continuidad,

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{(4 + 0.5) - 9.889}{2.031} = -2.65$$

y alcanzamos la misma conclusión.

## CORRELACION DE RANGO

**17.24.** La Tabla 17.26 muestra cómo fueron calificados 10 estudiantes de un curso de Biología, ordenados por letra alfabética, en laboratorio y en teoría. Hallar el coeficiente de correlación de rango.

**Tabla 17.26**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Laboratorio | 8 | 3 | 9 | 2 | 7 | 10 | 4 | 6 | 1 | 5 |
| Teoría | 9 | 5 | 10 | 1 | 8 | 7 | 3 | 4 | 2 | 6 |

**Solución**

La diferencia en rangos, $D$, en laboratorio y en teoría, para cada estudiante se da en la Tabla 17.27, que da también $D^2$ y $\sum D^2$. Luego

$$r_S = 1 - \frac{6\sum D^2}{N(N^2 - 1)} = 1 - \frac{6(24)}{10(10^2 - 1)} = 0.8545$$

indicando que hay una marcada relación entre las calificaciones de laboratorio y de teoría.

**Tabla 17.27**

| Diferencia de rangos ($D$) | $-1$ | $-2$ | $-1$ | 1 | $-1$ | 3 | 1 | 2 | $-1$ | $-1$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $D^2$ | 1 | 4 | 1 | 1 | 1 | 9 | 1 | 4 | 1 | 1 | $\sum D^2 = 24$ |

**17.25.** En la Tabla 17.28 aparecen las alturas de una muestra de 12 padres y sus hijos mayores. Hallar el coeficiente de correlación de rango.

**Tabla 17.28**

| Altura del padre (pulgadas) | 65 | 63 | 67 | 64 | 68 | 62 | 70 | 66 | 68 | 67 | 69 | 71 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Altura del hijo (pulgadas) | 68 | 66 | 68 | 65 | 69 | 66 | 68 | 65 | 71 | 67 | 68 | 70 |

**Solución**

Ordenados de menor a mayor, las alturas de los padres son

$$62 \quad 63 \quad 64 \quad 65 \quad 66 \quad 67 \quad 67 \quad 68 \quad 68 \quad 69 \quad 71 \tag{18}$$

Como el sexto y el séptimo lugares representan la misma altura (67 in), asignamos a esos lugares un rango medio $\frac{1}{2}(6 + 7) = 6.5$. Análogamente, al octavo y noveno lugar se les asigna rango $\frac{1}{2}(8 + 9) = 8.5$. Así que las alturas de los padres tienen asignados los rangos

$$1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6.5 \quad 6.5 \quad 8.5 \quad 8.5 \quad 10 \quad 11 \quad 12 \tag{19}$$

De la misma manera, ordenadas de menor a mayor, las alturas de los hijos son

$$65 \quad 65 \quad 66 \quad 66 \quad 67 \quad 68 \quad 68 \quad 68 \quad 68 \quad 69 \quad 70 \quad 71 \tag{20}$$

y como los lugares del sexto al noveno tienen la misma altura anotada (68 in), les asignamos el rango medio $\frac{1}{4}(6 + 7 + 8 + 9) = 7.5$. Así pues, a las alturas de los hijos se les asignan los rangos

$$1.5 \quad 1.5 \quad 3.5 \quad 3.5 \quad 5 \quad 7.5 \quad 7.5 \quad 7.5 \quad 7.5 \quad 10 \quad 11 \quad 12 \tag{21}$$

Usando las correspondencia(18) y (19), y (20) y (21), podemos sustituir la Tabla 17.28 por la Tabla 17.29. La Tabla 17.30 da las diferencias en rangos, $D$, y los cálculos de $D^2$ y $\sum D^2$, de donde

$$r_S = 1 - \frac{6\sum D^2}{N(N^2 - 1)} = 1 - \frac{6(72.50)}{12(12^2 - 1)} = 0.7465$$

Este resultado está en buen acuerdo con el coeficiente de correlación obtenido por otros métodos (véanse Probs. 14.9, 14.14, 14.16 y 14.23).

**Tabla 17.29**

| Rango del padre | 4 | 2 | 6.5 | 3 | 8.5 | 1 | 11 | 5 | 8.5 | 6.5 | 10 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rango del hijo | 7.5 | 3.5 | 7.5 | 1.5 | 10 | 3.5 | 7.5 | 1.5 | 12 | 5 | 7.5 | 11 |

**Tabla 17.30**

| $D$ | −3.5 | −1.5 | −1.0 | 1.5 | −1.5 | −2.5 | 3.5 | 3.5 | −3.5 | 1.5 | 2.5 | 1.0 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $D^2$ | 12.25 | 2.25 | 1.00 | 2.25 | 2.25 | 6.25 | 12.25 | 12.25 | 12.25 | 2.25 | 6.25 | 1.00 | $\sum D^2 = 72.50$ |

## PROBLEMAS SUPLEMENTARIOS

### EL TEST DE LOS SIGNOS

**17.26.** Una empresa afirma que si se añade su producto en el depósito de gasolina de un automóvil, las millas recorridas por galón aumentan. Para contrastar tal afirmación, se toman 15 automóviles distintos y se miden las millas por galón recorridas con y sin ese producto, con los resultados de la Tabla 17.31. Suponiendo que las condiciones de conducción son las mismas, determinar si hay diferencia debida a ese producto, al nivel de significación (*a*) 0.05 y (*b*) 0.01.

**Tabla 17.31**

| Con aditivo | Sin aditivo |
|---|---|
| 34.7 | 31.4 |
| 28.3 | 27.2 |
| 19.6 | 20.4 |
| 25.1 | 24.6 |
| 15.7 | 14.9 |
| 24.5 | 22.3 |
| 28.7 | 26.8 |
| 23.5 | 24.1 |
| 27.7 | 26.2 |
| 32.1 | 31.4 |
| 29.6 | 28.8 |
| 22.4 | 23.1 |
| 25.7 | 24.0 |
| 28.1 | 27.3 |
| 24.3 | 22.9 |

**17.27.** ¿Se puede concluir al nivel de significación 0.05 que las millas recorridas por galón en el Problema 17.26 aumentan al añadir ese producto?

**17.28.** Un club de adelgazamiento anuncia que ha preparado un programa especial que producirá pérdidas de peso de al menos un 6% en un mes, si se sigue rigurosamente. Para comprobar esa afirmación, 36 adultos siguen el programa. De ellos, 25 perdieron lo anunciado, 6 engordaron y el resto no sufrió cambio esencialmente. Determinar al nive de significación 0.05 si el programa era eficaz.

**17.29.** Un director de personal sostiene que con un curso especial para el personal de la sección de ventas, una empresa aumentará sus ventas. Para comprobarlo, se impartió el curso a 24 personas, de las que 16 vieron las ventas aumentadas, 6 las vieron decrecer y las de 2 quedaron sin cambio. Contrastar al nivel de significación 0.05 la hipótesis de que el curso hizo crecer las ventas de la empresa.

**17.30.** Una empresa fabricante de refrescos hizo degustaciones en 27 localidades del país para saber hacia qué refresco de cola, *A* o *B*, se inclinaban la preferencias del público. En 8 localidades se prefirió el *A*, en 17 el *B* y en las restantes no hubo preferencia por ninguno sobre el otro. ¿Se puede concluir que, al nivel de significación 0.05, el *B* es el preferido?

**17.31.** Las tensiones de ruptura de una muestra aleatoria de 25 sogas de un cierto fabricante se dan en la Tabla 17.32. Contrastar con esa muestra, al nivel de significación 0.05, la

afirmación del fabricante de que tal tensión es (*a*) 25, (*b*) 30, (*c*) 35 y (*d*) 40.

**Tabla 17.32**

| | | | | |
|---|---|---|---|---|
| 41 | 28 | 35 | 38 | 23 |
| 37 | 32 | 24 | 46 | 30 |
| 25 | 36 | 22 | 41 | 37 |
| 43 | 27 | 34 | 27 | 36 |
| 42 | 33 | 28 | 31 | 24 |

**17.32.** Indicar cómo se pueden obtener los límites de confianza 95% para los datos del Problema 17.4.

**17.33.** Plantear y resolver un problema que utilice el test de los signos.

## EL *U*-TEST DE MANN-WHITNEY

**17.34.** Los profesores *A* y *B* dan cursos de química en la Universidad XYZ. En un examen común, sus estudiantes recibieron las calificaciones que aparecen en la Tabla 17.33. Contrastar al nivel de significación 0.05 la hipótesis de que no hay diferencia entre las calificaciones de ambos profesores.

**Tabla 17.33**

| *A* | *B* |
|---|---|
| 88 | 72 |
| 75 | 65 |
| 92 | 84 |
| 71 | 53 |
| 63 | 76 |
| 84 | 80 |
| 55 | 51 |
| 64 | 60 |
| 82 | 57 |
| 96 | 85 |
| | 94 |
| | 87 |
| | 73 |
| | 61 |

**17.35.** Refiriéndonos al Problema 17.34, ¿puede concluirse al nivel de significación 0.01, que las notas de la clase matinal son peores que las de la vespertina?

**17.36.** Un agricultor quiere saber si hay diferencia entre las producciones de dos variedades de trigo, I y II. La Tabla 17.34 indica las producciones de trigo por unidad de área con ambas variedades. ¿Puede concluirse que existe diferencia al nivel de significación (*a*) 0.05 y (*b*) 0.01?

**Tabla 17.34**

| Trigo I | Trigo II |
|---|---|
| 15.9 | 16.4 |
| 15.3 | 16.8 |
| 16.4 | 17.1 |
| 14.9 | 16.9 |
| 15.3 | 18.0 |
| 16.0 | 15.6 |
| 14.6 | 18.1 |
| 15.3 | 17.2 |
| 14.5 | 15.4 |
| 16.6 | |
| 16.0 | |

**17.37.** Puede el agricultor del Problema 17.36 concluir al nivel de significación 0.05 que la variedad II da mayor producción que la I?

**17.38.** Se desea averiguar si hay diferencia entre dos clases de gasolina, *A* y *B*. La Tabla 17.35 da las distancias recorridas por galón para cada clase. ¿Se puede concluir al nivel de significación 0.05 que (*a*) hay diferencia entre ambas y (*b*) la *B* es mejor que la *A*?

**Tabla 17.35**

| *A* | *B* |
|---|---|
| 30.4 | 33.5 |
| 28.7 | 29.8 |
| 29.2 | 30.1 |
| 32.5 | 31.4 |
| 31.7 | 33.8 |
| 29.5 | 30.9 |
| 30.8 | 31.3 |
| 31.1 | 29.6 |
| 30.7 | 32.8 |
| 31.8 | 33.0 |

**17.39.** ¿Puede usarse el $U$-test para determinar si hay diferencia entre las máquinas I y II de la Tabla 17.1? Explicar la respuesta.

**17.40.** Proponer y resolver un problema que utilice el $U$-test.

**17.41.** Hallar $U$ para los datos de la Tabla 17.36, usando (*a*) el método de la fórmula y (*b*) el método de recuento.

**Tabla 17.36**

| | | |
|---|---|---|
| Muestra 1 | 15 | 25 |
| Muestra 2 | 20 | 32 |

**17.42.** Resolver el Problema 17.41 para los datos de la Tabla 17.37.

**Tabla 17.37**

| | | | | |
|---|---|---|---|---|
| Muestra 1 | 40 | 27 | 30 | 56 |
| Muestra 2 | 10 | 35 | | |

**17.43.** Una población consta de los valores 2, 5, 9 y 12. Se toman dos muestras, la primera de uno de esos valores y la segunda de los tres restantes.

(*a*) Obtener la distribución muestral de $U$ y su gráfico.

(*b*) Hallar la media y la varianza de esa distribución, directamente y por la fórmula.

**17.44.** Probar que $U_1 + U_2 = N_1N_2$.

**17.45.** Probar que $R_1 + R_2 = [N(N + 1)]/2$ para el caso en que el número de coincidencias es (*a*), 1, (*b*) 2 y (*c*) cualquier número.

**17.46.** Si $N_1 = 14$, $N_2 = 12$ y $R_1 = 105$, hallar (*a*) $R_2$, (*b*) $U_1$ y (*c*) $U_2$.

**17.47.** Si $N_1 = 10$, $N_2 = 16$, y $U_2 = 60$, hallar (*a*) $R_1$, (*b*) $R_2$ y (*c*) $U_1$.

**17.48.** ¿Cuál es el mayor número de valores $N_1$, $N_2$, $R_1$, $R_2$, $U_1$ y $U_2$ que puede calcularse a partir de los restantes? Demostrar la respuesta.

## EL $H$-TEST DE KRUSKAL-WALLIS

**17.49.** Se realiza un experimento para determinar las producciones de cinco variedades de trigo: $A$, $B$, $C$, $D$ y $E$. Se asignan a cada variedad cuatro parcelas. La producción (en bushels por acre) se indica en la Tabla 17.38. Suponiendo que las parcelas tienen igual fertilidad y que las variedades se asignan a las parcelas de modo aleatorio, determinar si hay diferencia significativa entre las producciones al nivel de significación (*a*) 0.05 y (*b*) 0.01.

**Tabla 17.38**

| | | | | |
|---|---|---|---|---|
| $A$ | 20 | 12 | 15 | 19 |
| $B$ | 17 | 14 | 12 | 15 |
| $C$ | 23 | 16 | 18 | 14 |
| $D$ | 15 | 17 | 20 | 12 |
| $E$ | 21 | 14 | 17 | 18 |

**17.50.** Las vidas medias de cuatro tipos de llantas $A$, $B$, $C$ y $D$, vienen dadas en la Tabla 17.39 (en miles de millas de rodaje); cada tipo se ha probado con seis automóviles similares asignados a las llantas al azar. Determinar si hay diferencia significativa entre las llantas al nivel de significación (*a*) 0.05 y (*b*) 0.01.

**Tabla 17.39**

| | | | | | | |
|---|---|---|---|---|---|---|
| $A$ | 33 | 38 | 36 | 40 | 31 | 35 |
| $B$ | 32 | 40 | 42 | 38 | 30 | 34 |
| $C$ | 31 | 37 | 35 | 33 | 34 | 30 |
| $D$ | 27 | 33 | 32 | 29 | 31 | 28 |

**17.51.** Un pedagogo quiere probar tres métodos de enseñanza: I, II y III. Para ello, escoge al azar tres grupos de 5 estudiantes cada uno y les aplica métodos diferentes. Se da el mismo examen a todos ellos y se producen las notas que figuran en la Tabla 17.40. Determinar si hay diferencia entre esos métodos de enseñanza al nivel de significación (*a*) 0.05 y (*b*) 0.01.

**Tabla 17.40**

| | | | | | |
|---|---|---|---|---|---|
| Método I | 78 | 62 | 71 | 58 | 73 |
| Método II | 76 | 85 | 77 | 90 | 87 |
| Método III | 74 | 79 | 60 | 75 | 80 |

**17.52.** En la Tabla 17.41 se ven las notas de un alumno. Al nivel de significación (*a*) 0.05 y (*b*) 0.01 decidir si hay diferencia entre las notas en las diversas materias.

**Tabla 17.41**

| | | | | | |
|---|---|---|---|---|---|
| Matemáticas | 72 | 80 | 83 | 75 | |
| Ciencias | 81 | 74 | 77 | | |
| Inglés | 88 | 82 | 90 | 87 | 80 |
| Economía | 74 | 71 | 77 | 70 | |

**17.53.** Usando el *H* test, resolver (*a*) Problema 16.9, (*b*) Problema 16.21 y (*c*) Problema 16.25.

**17.54.** Usando el *H* test, resolver (*a*) Problema 16.23, (*b*) Problema 16.24 y (*c*) Problema 16.25.

## EL TEST DE LAS RACHAS PARA EL CARACTER ALEATORIO

**17.55.** Determinar el número de rachas, *V*, para cada una de estas secuencias:

(*a*) A B A B B A A A B B A B

(*b*) H H T H H H T T T T H H T H H T H T

**17.56.** Se ha preguntado a 25 individuos si les gusta (Y) o no (N) un cierto producto, y se ha obtenido la secuencia de respuestas siguiente:

Y Y N N N N Y Y Y N Y N N
Y N N N N N Y Y Y Y N N

(*a*) Hallar el número de rachas.

(*b*) Decidir al nivel de significación 0.05 si las respuestas son aleatorias.

**17.57.** Usar el test de las rachas en las secuencias (10) y (11) de este capítulo, y establecer las conclusiones acerca de su aleatoriedad.

**17.58.** (*a*) Formar todas las posibles secuencias con dos *aes* y una *b*, y dar el número *V*, de rachas en cada una.

(*b*) Hallar la distribución muestral de *V* y su gráfico.

(c) Obtener la distribución de probabilidad de *V* y su gráfico.

**17.59.** En el Problema 17.58, hallar la media y la varianza de *V* (*a*) directamente de la distribución muestral y (*b*) por la fórmula.

**17.60.** Resolver los Problemas 17.58 y 17.59 para los casos en que hay (*a*) dos *aes* y dos *bes* (*b*) una *a* y tres *bes*, y (*c*) una *a* y cuatro *bes*.

**17.61.** Resolver los Problemas 17.58 y 17.59 con (*a*) dos *aes* y cuatro *bes* y (*b*) tres *aes* y tres *bes*.

## OTRAS APLICACIONES DEL TEST DE LAS RACHAS

**17.62.** Determinar, al nivel de significación 0.05, si la muestra de 40 calificaciones de la Tabla 17.5 es aleatoria.

**17.63.** Las cotizaciones de ciertas acciones en 25 días sucesivos vienen dadas en la Tabla 17.42. Determinar al nivel de significación 0.05 si son aleatorias.

**Tabla 17.42**

| | | | | |
|---|---|---|---|---|
| 10.375 | 11.125 | 10.875 | 10.625 | 11.500 |
| 11.625 | 11.250 | 11.375 | 10.750 | 11.000 |
| 10.875 | 10.750 | 11.500 | 11.250 | 12.125 |
| 11.875 | 11.375 | 11.875 | 11.125 | 11.750 |
| 11.375 | 12.125 | 11.750 | 11.500 | 12.250 |

**17.64.** Los primeros dígitos de $\sqrt{2}$ son 1.41421 35623 73095 0488 ··· . ¿Qué conclusiones se pueden sacar sobre su aleatoriedad?

**17.65.** ¿Qué conclusiones se pueden sacar sobre el carácter aleatorio de los siguientes dígitos?

(*a*) $\sqrt{3} = 1.73205\ 08075\ 68877\ 2935...$

(*b*) $\pi = 3.14159\ 26535\ 89793\ 2643...$

**17.66** En el Problema 17.30, aplicar el test de las rachas para decidir sobre su aleatoriedad.

**17.67.** En el Problema 17.32, aplicar el test de las rachas para decidir sobre su aleatoriedad.

**17.68.** En el Problema 17.34, aplicar el test de las rachas para decidir sobre su aleatoriedad.

## CORRELACION DE RANGO

**17.69.** En un concurso, dos jueces hubieron de colocar a ocho candidatos (numerados de 1 a 8) por orden de preferencia, con el resultado que recoge la Tabla 17.43.

(*a*) Hallar el coeficiente de correlación de rango.

(*b*) Decidir cuántos coincidentes fueron las elecciones de ambos jueces.

**Tabla 17.43**

| Primer juez | Segundo juez |
|---|---|
| 5 | 4 |
| 2 | 5 |
| 8 | 7 |
| 1 | 3 |
| 4 | 2 |
| 6 | 8 |
| 3 | 1 |
| 7 | 6 |

**17.70.** Aplicar correlación de rango al (*a*) Problema 14.26, (*b*) Problema 14.42, (*c*) Problema 14.46 y (*d*) Problema 14.63.

**17.71.** El coeficiente de correlación de rango se deduce usando los datos con rango en la fórmula momento-producto del Capítulo 14. Ilustrar esto resolviendo algún problema por ambos métodos.

**17.72.** ¿Puede hallarse el coeficiente de correlación de rango para datos agrupados? Explicar la respuesta e ilustrarla con un ejemplo.

# CAPITULO 18

# Análisis de series en el tiempo

## SERIES EN EL TIEMPO

Una *serie en el tiempo* es un conjunto de observaciones tomadas en instantes específicos, generalmente a intervalos iguales. Ejemplos de tales series en el tiempo son la producción anual total de acero en EE.UU. durante un cierto número de años, la cotización diaria al cierre de la sesión bursátil de ciertas acciones, las temperaturas anunciadas cada hora por el instituto meteorológico para una ciudad o el total de ventas mensuales en una empresa.

Matemáticamente, una serie en el tiempo se define por los valores $Y_1$, $Y_2$, ... de una variable $Y$ (temperatura, cotización, etc.) en tiempos $t_1$, $t_2$, ... Así pues, $Y$ es una función de $t$; esto se denota por $Y = F(t)$.

## GRAFICOS DE SERIES EN EL TIEMPO

Una serie en el tiempo que involucra a una variable $Y$ se representa por un gráfico de $Y$ respecto de $t$, como se ha hecho ya muchas veces en capítulos anteriores. Por ejemplo, la Figura 18.1 es el gráfico de una serie en el tiempo que muestra el número de cabezas de ganado en EE.UU. durante los años 1870-1980.

## MOVIMIENTOS CARACTERISTICOS DE SERIES EN EL TIEMPO

Es interesante pensar en el gráfico de una serie en el tiempo (tal como el de la Fig. 18.1) como un gráfico que describe un punto moviéndose con el paso del tiempo, análogo en muchos aspectos a la trayectoria de una partícula física que se mueve bajo la influencia de fuerzas físicas. Claro está que, en lugar de fuerzas físicas, aquí cabe pensar en el resultado de una combinación de fuerzas económicas, sociológicas, psicológicas o de otros tipos.

La experiencia con muchos ejemplos de series en el tiempo ha revelado ciertos *movimientos* o *variaciones características* que aparecen a menudo, y cuyo análisis es de gran interés por muchas razones, una de ellas el problema de *predicción* de futuros movimientos. No puede sorprendernos, en consecuencia, que muchas empresas y gobiernos estén preocupados por este importante tema.

## CLASIFICACION DE MOVIMIENTOS DE SERIES EN EL TIEMPO

Los movimientos característicos de series en el tiempo se pueden clasificar en cuatro tipos principales, a menudo llamados *componentes* de una serie en el tiempo:

1. **Movimientos a largo plazo o seculares.** Se refieren a la dirección general en la que el gráfico de una serie en el tiempo parece progresar en un largo período de tiempo. En la Figura 18.1, este movimiento secular (o *variación secular* o *tendencia secular*, como se llama a veces) se indica por una *curva de tendencia*, en trazo discontinuo. Para algunas series en el tiempo puede ser apropiada una *recta de tendencia*. La determinación de tales curvas o rectas de tendencia por mínimos cuadrados se ha considerado en el Capítulo 13. Otros métodos se discutirán más adelante en este capítulo.
2. **Movimientos característicos o variaciones cíclicas.** Estas se refieren a las oscilaciones a largo término en torno a una recta o curva de tendencia. Estos *ciclos*, como se les llama, pueden ser *periódicos* o no; es decir, pueden seguir o no esquemas repetidos en intervalos iguales de tiempo. En actividades de negocios o financieras, los movimientos se consideran cíclicos sólo si son recurrentes en un período de tiempo de al menos un año. Un importante ejemplo de movimientos característicos lo constituyen los llamados *ciclos económicos*, que representan intervalos de prosperidad, recesión, depresión y recuperación. Los movimientos característicos en torno a las curvas de tendencia son muy nítidos en la Figura 18.1

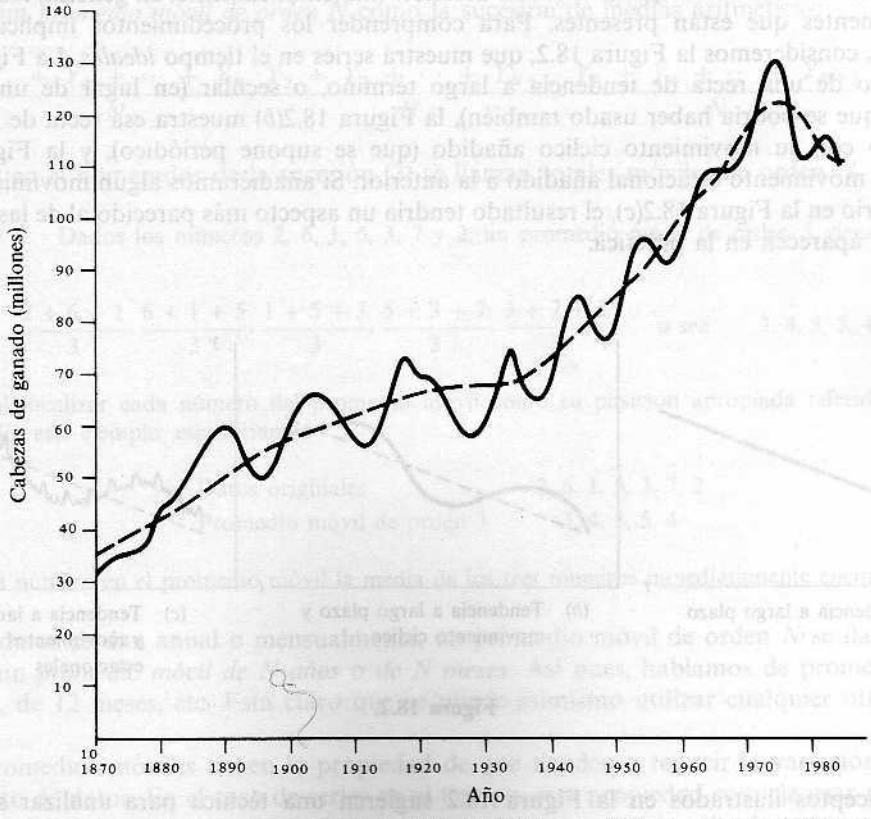


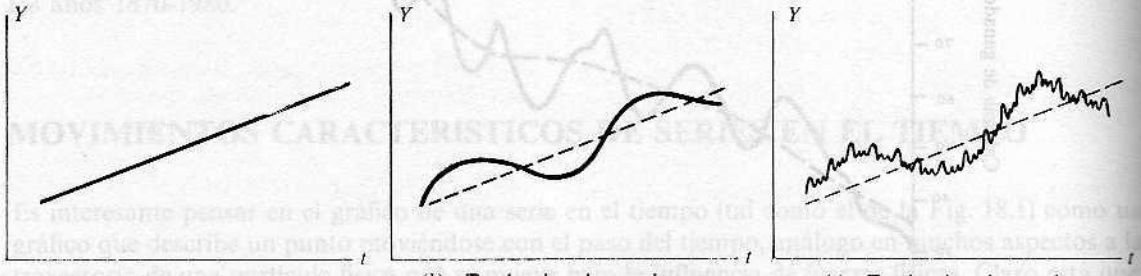

**Figura 18.1.** Censo de ganado en EE.UU., 1870-1980 (*Fuente:* U.S. Department of Agriculture).

3. **Movimientos estacionales o variaciones estacionales.** Estos se refieren a los esquemas idénticos o casi idénticos que una serie en el tiempo parece seguir durante meses correspondientes en años sucesivos. Tales movimientos se deben a sucesos recurrentes que tienen lugar anualmente, tales como el brusco aumento de precios al consumo antes de la Navidad. En la Figura 18.1 no se aprecian movimientos estacionales, pues el gráfico fue obtenido mediante datos anuales.

   Aunque los movimientos estacionales se refieren generalmente en teoría económica a periodicidad *anual*, las ideas en juego admiten extensión a intervalos cualesquiera de periodicidad (días, horas o semanas), según el tipo de datos de que disponemos.
4. **Movimientos irregulares o aleatorios.** Estos se refieren a los movimientos esporádicos de las series en el tiempo debidos a sucesos de azar, tales como inundaciones, huelgas o elecciones. Si bien se suele suponer que tales sucesos producen variaciones que pierden su influencia tras poco tiempo, cabe la posibilidad de que sean tan intensos que den lugar a nuevos movimientos cíclicos o de otro tipo.

## ANALISIS DE SERIES EN EL TIEMPO

El análisis de series en el tiempo consiste en describir (matemáticamente, en general) los movimientos componentes que están presentes. Para comprender los procedimientos implicados en tal descripción, consideremos la Figura 18.2, que muestra series en el tiempo *ideales*. La Figura 18.2(*a*) es el gráfico de una recta de tendencia a largo término, o secular (en lugar de una curva de tendencia, que se podría haber usado también), la Figura 18.2(*b*) muestra esa recta de tendencia a largo plazo con su movimiento cíclico añadido (que se supone periódico), y la Figura 18.2(*c*) muestra un movimiento estacional añadido a la anterior. Si añadiéramos algún movimiento irregular o aleatorio en la Figura 18.2(*c*), el resultado tendría un aspecto más parecido al de las series en el tiempo que aparecen en la práctica.




(*a*) Tendencia a largo plazo

(*b*) Tendencia a largo plazo y movimiento cíclico

(*c*) Tendencia a largo plazo y movimientos cíclicos y estacionales

**Figura 18.2.**

Los conceptos ilustrados en la Figura 18.2 sugieren una técnica para analizar series en el tiempo. Supongamos que la serie en el tiempo tiene por variable $Y$ el producto de varias variables

$T$, $C$, $S$ e $I$ que producen los movimientos de tendencia, cíclicos, estacionales e irregulares, respectivamente. En símbolos,

$$Y = T \times C \times S \times I = TCSI \tag{1}$$

El análisis de series en el tiempo requiere investigar los factores $T$, $C$, $S$ e $I$, y se conoce a menudo como una *descomposición* de una serie en el tiempo en movimientos componentes básicos.

Hay que hacer constar que algunos estadísticos prefieren considerar $Y$ como la suma $T + C + S + I$ de las variables básicas involucradas. Aunque supondremos la descomposición dada por la ecuación (1) cuando examinemos los métodos discutidos en este capítulo, procedimientos análogos entran en juego cuando se trata con una suma. En la práctica, la decisión sobre cuál de los métodos de descomposición se adopta depende del grado de éxito a que conduce la aplicación de cada uno.

## PROMEDIOS MOVILES; SUAVIZACION DE SERIES EN EL TIEMPO

Dado un conjunto de números

$$Y_1, Y_2, Y_3, \ldots \tag{2}$$

definimos un *promedio móvil de orden N* como la sucesión de medias aritméticas:

$$\frac{Y_1 + Y_2 + \cdots + Y_N}{N}, \frac{Y_2 + Y_3 + \cdots + Y_{N+1}}{N}, \frac{Y_3 + Y_4 + \cdots + Y_{N+2}}{N}, \ldots \tag{3}$$

Las sumas en el numerador de la sucesión (3) se llaman totales móviles de orden $N$.

**EJEMPLO 1.** Dados los números 2, 6, 1, 5, 3, 7 y 2, un promedio móvil de orden 3 viene dado por la sucesión

$$\frac{2+6+1}{3}, \frac{6+1+5}{3}, \frac{1+5+3}{3}, \frac{5+3+7}{3}, \frac{3+7+2}{3} \quad \text{o sea} \quad 3, 4, 3, 5, 4$$

Es usual localizar cada número del promedio móvil como su posición apropiada referida a los datos originales. En este ejemplo escribiríamos

| | |
|---|---|
| Datos originales | 2, 6, 1, 5, 3, 7, 2 |
| Promedio móvil de orden 3 | 3, 4, 3, 5, 4 |

siendo cada número en el promedio móvil la media de los tres números inmediatamente encima de él.

Si los datos se dan anual o mensualmente, un promedio móvil de orden $N$ se llama, respectivamente, un *promedio móvil de N años* o *de N meses*. Así pues, hablamos de promedios móviles de 5 años, de 12 meses, etc. Está claro que se puede asimismo utilizar cualquier otra unidad de tiempo.

Los promedios móviles tienen la propiedad de que tienden a reducir la variación presente en un conjunto de datos. En el caso de series en el tiempo, esta propiedad se suele usar para eliminar fluctuaciones indeseables, en un proceso que se conoce como *suavización de series en el tiempo*.

Si se usan medias aritméticas ponderadas en la sucesión (3), con pesos especificados de antemano, la sucesión resultante se llama un *promedio móvil ponderado de orden N.*

**EJEMPLO 2.** Si se usan pesos 1, 4 y 1 en el Ejemplo 1, un promedio móvil ponderado de orden 3 viene dado por la sucesión

$$\frac{1(2)+4(6)+1(1)}{1+4+1}, \frac{1(6)+4(1)+1(5)}{1+4+1}, \frac{1(1)+4(5)+1(3)}{1+4+1}, \frac{1(5)+4(3)+1(7)}{1+4+1}, \frac{1(3)+4(7)+1(2)}{1+4+1}$$

o sea, 4.5, 2.5, 4.0, 4.0, 5.5

## ESTIMACION DE LA TENDENCIA

1. **Método de los mínimos cuadrados.** Este método, descrito en el Capítulo 13, se puede utilizar para hallar la ecuación de la recta o curva de tendencia adecuada. De esta ecuación se podrán calcular los valores de tendencia $T$.
2. **Método «a mano».** Este método, que consiste en ajustar una curva o recta de tendencia por simple inspección del gráfico, también se puede usar para estimar $T$. No obstante, tiene la desventaja evidente de depender muy fuertemente del criterio personal de cada cual.
3. **Método del promedio móvil.** Usando promedios móviles de órdenes apropiados, podemos eliminar esquemas cíclicos, estacionales e irregulares, dejando así tan sólo el movimiento de tendencia.

   Una desventaja de este método es que los datos al comienzo y al final de una serie se pierden: así, en el Ejemplo 1 comenzamos con siete números, y con un promedio móvil de orden 3 llegamos a cinco números. Otra desventaja es que los promedios móviles pueden generar ciclos u otros movimientos que no estaban presentes en los datos originales. Una tercera desventaja es que los promedios móviles se ven muy afectados por los valores extremos. Para obviar esto último en cierta medida, se usa a veces un promedio móvil ponderado con pesos adecuados; en tal caso, al valor o valores centrales se les asigna peso máximo, y a los valores extremos, pesos pequeños.
4. **Método de semipromedios.** Consiste en separar los datos en dos partes (preferible que sean iguales) y promediar los datos de cada parte, obteniendo con ello dos puntos en el gráfico de la serie en el tiempo. Entonces se traza una recta de tendencia entre esos dos puntos, y los valores de tendencia se determinan de esa recta de tendencia. Los valores de tendencia se pueden determinar también directamente, sin gráfico (véase Prob. 18.6).

   Aunque el método es sencillo de aplicar, puede conducir a resultados pobres si se usa indiscriminadamente. Además es sólo aplicable cuando la tendencia es lineal o aproximadamente lineal, si bien puede extenderse a casos en que los datos pueden agruparse en varias partes, en cada una de las cuales la tendencia es lineal.

## ESTIMACION DE LAS VARIACIONES ESTACIONALES; EL INDICE ESTACIONAL

Para determinar el factor estacional $S$ en la ecuación (1), debemos estimar cómo varían los datos de la serie en el tiempo de mes a mes en un año típico. Un conjunto de números que muestra los

valores relativos de una variable durante los meses del año se llama un *índice estacional* para la variable. Por ejemplo, si sabemos que las ventas durante enero, febrero, marzo, etc., son el 50, 120, 90, ... % del promedio de ventas mensual en el total del año, entonces los números 50, 120, 90, ... dan el índice estacional de ese año, y se llaman *números índice estacionales*. El índice estacional medio del año ha de ser 100%; esto es, la suma de los números índice de los 12 meses ha de ser 1200%.

Se dispone de varios métodos para calcular un índice estacional:

1. **Método de porcentaje medio.** En este método expresamos los datos de cada mes como porcentajes del promedio anual. Los porcentajes para meses correspondientes en distintos años se promedian entonces, usando una media o una mediana; si se usa la media, es mejor evitar valores extremos que puedan aparecer. Los 12 porcentajes resultantes dan el índice estacional. Si su media no es el 100% (o sea, si su suma no es 1200%), deben ser ajustados, lo que se logra multiplicándolos por un factor adecuado.
2. **Método del porcentaje de tendencia.** En este método expresamos los datos para cada mes como porcentajes de valores de tendencia mensuales. Un promedio apropiado de los porcentajes para meses correspondientes da entonces el índice requerido. Como en el método 1, los ajustamos si no tienen promedio 100%.

   Nótese que al dividir cada valor mensual $Y$ por el correspondiente valor de tendencia $T$ resulta $Y/T = CSI$, de la ecuación (1), y que el subsiguiente promedio de $Y/T$ produce los índices estacionales. En tanto en cuanto estos índices incluyen variaciones cíclicas e irregulares, puede ser una desventaja del método, especialmente si las variaciones son grandes.
3. **Método del promedio móvil en porcentaje.** En este método calculamos un promedio móvil de 12 meses. Como los resultados obtenidos así caen entre meses sucesivos en lugar de en el centro del mes (que es donde caen los datos originales), calculamos un promedio móvil de 2 meses de ese promedio móvil de 12 meses. El resultado se llama a veces un *promedio móvil de 12 meses centrado*.

   Tras hacer eso, expresamos los datos originales de cada mes como un porcentaje del promedio móvil centrado de 12 meses que corresponde a los datos originales. Los porcentajes de los meses correspondientes se promedian a continuación, dando el índice buscado. Como antes, los ajustamos si no promedian 100%.

   Obsérvese que el razonamiento lógico que subyace a este método se sigue de la ecuación (1). Un promedio móvil centrado de 12 meses de $Y$ sirve para eliminar los movimientos estacionales e irregulares $S$ e $I$, y es por tanto equivalente a los valores dados por $TC$. Al dividir los datos originales por $TC$ nos da $SI$. Los promedios subsiguientes sobre meses correspondientes sirven para eliminar la irregularidad $I$ y en consecuencia producen un índice $S$ adecuado.
4. **Método de la relación de enlace.** En este método expresamos los datos para cada mes como un porcentaje de los datos para los meses previos; estos porcentajes mensuales se llaman *relaciones de enlace* porque relacionan cada mes con el precedente. Entonces tomamos un promedio adecuado de los enlaces relativos para los meses correspondientes. De estas 12 relaciones de enlace promedio obtenemos los porcentajes relativos de cada mes respecto a enero, que se adopta como el 100%.

   Tras hacer eso, encontraremos que el siguiente enero tiene un porcentaje asociado que es mayor o menor que 100%, según haya habido un crecimiento o decrecimiento en la tendencia. Usando este porcentaje del próximo enero, ajustamos los diversos porcentajes relativos mensuales (antes obtenidos) para esta tendencia. Estos porcentajes finales, ajustados de modo que promedien 100%, dan el índice estacional requerido.

## DATOS AJUSTADOS A LA VARIACION ESTACIONAL

Si los datos mensuales originales se dividen por los correspondientes números índice estacionales, los datos resultantes se llaman *desestacionalizados* o *ajustados a la variación estacional*. Tales datos incluyen todavía movimientos de tendencia, cíclicos e irregulares.

## ESTIMACION DE LAS VARIACIONES CICLICAS

Una vez ajustados los datos a la variación estacional, pueden ser ajustados también a la tendencia sin más que dividirlos por los correspondientes valores de tendencia. De acuerdo con la ecuación (1), el proceso de ajustar a la variación estacional y a la tendencia corresponde a dividir $Y$ por $ST$, lo que da $CI$ (las variaciones cíclicas e irregulares). Un promedio móvil apropiado de unos pocos meses de duración (digamos 3, 5 ó 7 meses, de manera que el centrado subsiguiente no sea necesario) sirve entonces para suavizar las variaciones irregulares $I$ y para dejar sólo las variaciones cíclicas $C$. Una vez que estas variaciones cíclicas han sido aisladas de esa forma, se pueden estudiar en detalle. Si ocurre una periodicidad, exacta o aproximada, de ciclos, se pueden construir *índices cíclicos* de manera parecida a como se ha hecho para los índices estacionales.

## ESTIMACION DE LAS VARIACIONES IRREGULARES

Las variaciones irregulares (o aleatorias) se pueden estimar ajustando los datos a las variaciones de tendencia, estacionales y cíclicas. Eso significa tener que dividir los datos originales $Y$ por $T$, $S$ y $C$, que [por la ecuación (1)] da $I$. En la práctica se encuentra que las variaciones irregulares tienden a tener pequeña magnitud y con frecuencia tienden a seguir el esquema de una distribución normal; es decir, las pequeñas desviaciones ocurren con gran frecuencia y grandes desviaciones ocurren con pequeña frecuencia.

## COMPARACION DE DATOS

Al comparar datos, hay que tener siempre mucho cuidado de que tal comparación esté justificada. Por ejemplo, al comparar datos de marzo con datos de febrero, debemos tener bien presente que febrero tiene 28 ó 29 días y marzo tiene 31; y al comparar datos de febrero de años diferentes, hay que recordar que en un año bisiesto febrero tiene 29 días en lugar de 28. Para poner otro ejemplo, el número de días laborables durante varios meses del mismo año o de años diferentes, pueden ser distintos a causa de las vacaciones, huelgas, etc.

En la práctica, no se sigue una regla definida para ajustar tales variaciones. La necesidad de tales ajustes queda a voluntad del investigador.

## PREDICCION

Los métodos y principios anteriores se usan en la importante tarea de predecir series en el tiempo. Hay que ser conscientes de que, naturalmente, el tratamiento matemático de los datos no resuelve por sí mismo todos los problemas. No obstante, acoplado al sentido común del investigador, a su