la $S.C._{grupos}$ es mayor que esta *S.C.* crítica.. Es de interés investigar por qué la $S.C._{grupos}$ es tan grande y probar la significación de las diversas aportaciones hechas a esta *S.C.* por las diferencias entre las medias de muestreo. Esto se ha discutido en la sección anterior, donde se han calculado diferentes sumas de cuadrados basados en comparaciones entre medias planeadas antes de haberse examinado los datos. Una comparación se ha denominado significativa si su razón F_s era mayor que $F_{\alpha[k-1,\alpha(n-1)]}$, donde k es el número de medias que se comparan. Ahora podemos expresar también esto en términos de sumas de cuadrados. Una *S.C.* es significativa si es mayor que $(k-1)\,M.C._{intra}\ F_{\alpha[k-1,\alpha(n-1)]}$.

Las pruebas anteriores eran comparaciones a priori. Un procedimiento para probar comparaciones a posteriori sería fijar $k = a$ en la última fórmula, sin importar cuántas medias comparemos. De este modo el valor crítico de la *S.C.* será mayor que en el método anterior, resultando más difícil demostrar la significación de una *S.C.* de muestreo. Esto tiene en cuenta el hecho de que escogemos para probar aquellas diferencias entre medias de grupo que parece que contribuyen sustancialmente a la significación del análisis de la varianza global.

Por ejemplo, vamos a volver a los efectos de los azúcares en el crecimiento de las secciones de guisante (cuadro 8.1). Anotamos las medias en orden ascendente de magnitud: 58,0 (glucosa + fructosa), 58,2 (fructosa), 59,3 (glucosa), 64,1 (sacarosa), 70,1 (control). Observamos que los tres primeros tratamientos tienen medias bastante similares y sospechamos que no difieren significativamente entre sí y por tanto no contribuyen sustancialmente a la significación de la $S.C._{grupos}$.

Para probar esto calculamos la *S.C.* entre estas tres medias por la fórmula habitual:

$$S.C. = \frac{593^2 + 582^2 + 580^2}{10} - \frac{(593 + 582 + 580)^2}{3(10)}$$

$$= 102\,677{,}3 - 102\,667{,}5 = 9{,}8$$

Las diferencias entre estas medias no son significativas porque esta *S.C.* es menor que la *S.C.* crítica (56,35) calculada más arriba.

La media de la sacarosa parece sospechosamente diferente de las medias de los otros azúcares. Para probar esto calculamos

$$S.C. = \frac{641^2}{10} + \frac{(593 + 582 + 580)^2}{30} - \frac{(641 + 593 + 582 + 580)^2}{10 + 30}$$

$$= 41\,088{,}1 + 102\,667{,}5 - 143\,520{,}4 = 235{,}2$$

que es muy superior al valor crítico de *S.C.* Por lo tanto, concluimos que la sacarosa retarda el crecimiento significativamente menos que los otros azúcares ensayados. Podemos continuar de este modo, probando todas las diferencias que parezcan sospechosas o incluso probando todos los grupos de medias posibles, considerándolas 2, 3, 4 y 5 a la vez. Este último enfoque puede requerir un computador si hay más de 5 medias para comparar, ya que hay muchas pruebas posibles que podrían realizarse. Este procedimiento fue propuesto por Gabriel (1964), quien lo denominó *suma de cuadrados, procedimiento de pruebas simultáneas (S.C.-P.T.S.)*.

En el (*S.C.-P.T.S.*) y en el análisis de la varianza original, la probabilidad de cometer cualquier error de tipo I en absoluto es α, la probabilidad seleccionada para el valor crítico F de la tabla V. Por "cometer cualquier error de tipo I en absoluto" queremos decir que se comete tal error en la prueba de significación global del análisis de la varianza y en cualquiera de las comparaciones secundarias entre medias o grupos de medias necesarias para completar el análisis del experimento. Esta probabilidad α se denomina grado de error "del experimento". Debería advertirse que aunque la probabilidad de cualquier error absoluto sea α, la probabilidad de error para cualquier prueba particular de algún subgrupo, tal como una prueba de la diferencia entre 3 o entre dos medias, es necesariamente menor que α. Así para la prueba de cada subgrupo realmente se está utilizando un nivel de significación α', que puede ser mucho menor que α; y si hay muchas medias en el análisis de la varianza este grado de error efectivo α' puede ser 1/10, 1/100, o incluso 1/1000 del valor de α (Gabriel, 1964). Por esta razón las pruebas a posteriori discutidas hasta ahora y el análisis de la varianza global, no son muy sensibles a diferencias de medias individuales ni a diferencias dentro de pequeños subgrupos. Evidentemente, si α' es pequeño, no muchas diferencias se van a considerar significativas. Este es el precio que se paga por no planear las comparaciones antes de examinar los datos. Si se hacen pruebas a priori, el grado de error de cada uno seguiría siendo α.

En la literatura estadística se han descrito otras muchas técnicas de pruebas de comparaciones múltiples a posteriori, pero el procedimiento de pruebas simultáneas dado más arriba, bastaría como una moderada introducción al tema.

Ejercicios 8

8.1 El siguiente es un ejemplo con números sencillos para ayudar a la familiarización con el análisis de la varianza. Un ecólogo vegetal desea probar la hipótesis de que la altura de la especie vegetal X depende del tipo de suelo en que crece. Midió la altura de tres plantas en cada una de cuatro parcelas que representaban diferentes tipos de suelo, estando incluidas las cuatro parcelas en una superficie de dos millas cuadradas. Sus resultados están tabulados más abajo. (La altura se da en centímetros.) ¿Confirma el análisis esta hipótesis? SOLUCION. $F_s = 6{,}951$, $F_{0,05[3,9]} = 4{,}07$.

Observación número	*Localidades* 1	2	3	4
1	15	25	17	10
2	9	21	23	13
3	14	19	20	16

8.2 Las siguientes son medidas (en unidades del micrómetro codificado) de la longitud del tórax del áfido *Pemphigus populi-transversus*. Los áfidos se recogieron de 28 agallas del chopo de la Carolina, *Populas deltoides*. Se seleccionaron al azar cuatro áfidos alados de cada agalla y se midieron. Los áfidos alados de cada agalla son isogénicos (gemelos idénticos), habiendo descendido parteno-genéticamente de una hembra apomíctica. Así, cualquier varianza intraagalla sólo puede deberse al medio ambiente. La varianza entre diferentes agallas puede deberse a diferencias

en genotipo y también a diferencias ambientales entre agallas. Si este carácter, longitud del tórax, es afectado por variación genética, la varianza entre agallas debe ser significativa. Lo contrario no es necesariamente cierto; varianza entre agallas significativa no tiene por qué indicar variación genética; podría deberse también a diferencias ambientales entre agallas (datos de Sokal 1952). Analícese la varianza de la longitud del tórax. ¿Hay varianza interagallas significativa? Dense estimaciones del componente aditivo de la varianza interagallas si lo hay. ¿Qué porcentaje de la varianza está controlado por factores intragallas y qué porcentaje por factores interagallas? Discútanse los resultados. (Recuérdese comprobar los cálculos paso por paso; de lo contrario un error cometido al principio del cálculo puede arruinar el esfuerzo total. Para fines de cálculo ignórese la coma decimal en las variantes.)

Agalla N.º					*Agalla N.º*				
1.	6.1,	6.0,	5.7,	6.0	15.	6.3,	6.5,	6.1,	6.3
2.	6.2,	5.1,	6.1,	5.3	16.	5.9,	6.1,	6.1,	6.0
3.	6.2,	6.2,	5.3,	6.3	17.	5.8,	6.0,	5.9,	5.7
4.	5.1,	6.0,	5.8,	5.9	18.	6.5,	6.3,	6.5,	7.0
5.	4.4,	4.9,	4.7,	4.8	19.	5.9,	5.2,	5.7,	5.7
6.	5.7,	5.1,	5.8,	5.5	20.	5.2,	5.3,	5.4,	5.3
7.	6.3,	6.6,	6.4,	6.3	21.	5.4,	5.5,	5.2,	6.3
8.	4.5,	4.5,	4.0,	3.7	22.	4.3,	4.7,	4.5,	4.4
9.	6.3,	6.2,	5.9,	6.2	23.	6.0,	5.8,	5.7,	5.9
10.	5.4,	5.3,	5.0,	5.3	24.	5.5,	6.1,	5.5,	6.1
11.	5.9,	5.8,	6.3,	5.7	25.	4.0,	4.2,	4.3,	4.4
12.	5.9,	5.9,	5.5,	5.5	26.	5.8,	5.6,	5.6,	6.1
13.	5.8,	5.9,	5.4,	5.5	27.	4.3,	4.0,	4.4,	4.6
14.	5.6,	6.4,	6.4,	6.1	28.	6.1,	6.0,	5.6,	6.5

8.3 Millis y Seng (1954) publicaron un estudio de la relación del orden de nacimiento con los pesos de nacimiento de niños. Los datos que siguen de primero y octavo nacimientos se han extraído de una tabla de pesos de nacimiento de niños varones, de pacientes chinos de tercera clase de la Maternidad del Hospital Kandang Kerbau de Singapur en 1950 y 1951.

Peso al nacer libras : onzas	*Orden de nacimiento* 1	8
3:0–3:7	—	—
3:8–3:15	2	—
4:0–4:7	3	—
4:8–4:15	7	4
5:0–5:7	111	5
5:8–5:15	267	19
6:0–6:7	457	52
6:8–6:15	485	55
7:0–7:7	363	61
7:8–7:15	162	48
8:0–8:7	64	39
8:8–8:15	6	19
9:0–9:7	5	4
9:8–9:15	—	—
10:0–10:7	—	1
10:8–10:15	—	—
	1932	307

¿Qué orden de nacimiento parece ir acompañado por niños más pesados? ¿Es significativa esta diferencia? ¿Puede concluirse que el orden de nacimiento produce diferencias en el peso de nacimiento? (Nota: La variable debería codificarse lo más sencillamente posible). Volver a analizar utilizando la prueba t y comprobar que $t_2 = F_s$. SOLUCION. $t_s^2 = 11{,}016$.

8.4 De un amplio estudio de Brown y Brown (1956) se han tomado los siguientes valores de citocromo oxidasa de cucarachas macho *Periplaneta* en milímetros cúbicos por miligramo por diez minutos.

	n	$\bar{Y}$	$s_{\bar{Y}}$
24 horas después de la inyección de clorometoxi	5	24,8	0,9
Control	3	19,7	1,4

¿Son significativamente diferentes las dos medias?

8.5 Los datos que siguen son medidas de cinco muestras al azar de palomas domésticas recogidas durante los meses de enero, febrero y marzo en Chicago en 1955. La variable es la longitud desde el extremo anterior de la abertura de los orificios nasales hasta el extremo del pico óseo y se registra en milímetros. Datos de Olson y Miller (1958).

Muestras 1	2	3	4	5
5,4	5,2	5,5	5,1	5,1
5,3	5,1	4,7	4,6	5,5
5,2	4,7	4,8	5,4	5,9
4,5	5,0	4,9	5,5	6,1
5,0	5,9	5,9	5,2	5,2
5,4	5,3	5,2	5,0	5,0
3,8	6,0	4,8	4,8	5,9
5,9	5,2	4,9	5,1	5,0
5,4	6,6	6,4	4,4	4,9
5,1	5,6	5,1	6,5	5,3
5,4	5,1	5,1	4,8	5,3
4,1	5,7	4,5	4,9	5,1
5,2	5,1	5,3	6,0	4,9
4,8	4,7	4,8	4,8	5,8
4,6	6,5	5,3	5,7	5,0
5,7	5,1	5,4	5,5	5,6
5,9	5,4	4,9	5,8	6,1
5,8	5,8	4,7	5,6	5,1
5,0	5,8	4,8	5,5	4,8
5,0	5,9	5,0	5,0	4,9

¿Son homogéneas las cinco muestras?

8.6 P. E. Hunter (1959, datos detallados no publicados) seleccionó dos cepas de *Drosophila melanogaster*, una de período larvario corto (L.C.) y otra de período larvario largo (L.L.). Además mantuvo una cepa control no seleccionada (C.C.). A la generación 42 se obtuvieron los datos siguientes para el período larvario (medidos en horas). Analícese e interprétese.

	Cepas			
	LC	*CC*	LL	
n_i	80	69	33	
$\sum^{n_i} Y$	8070	7291	3640	$\sum^{3}\sum^{n_i} Y^2 = 1\,994\,650$

Obsérvese que parte del cálculo ya se ha realizado. Háganse pruebas a priori entre las tres medias (períodos larvarios cortos respecto a largos y cada uno respecto al control). Fíjense límites de confianza del 95 % para las diferencias de medias observadas, para las que se han hecho estas comparaciones. SOLUCION. $M.C._{\text{(L.C. respecto a L.L.)}} = 2\,076{,}6697$.

8.7 Los datos que siguen se han recogido de un estudio de las variaciones de proteínas sanguíneas en el ciervo (Cowan and Johnston, 1962). La variable es la movilidad de la fracción II de las proteínas del suero, expresada como 10^{-5} cm^2/volt. segundos.

	$\overline{Y}$	$s_{\overline{Y}}$
Sitka	2,8	0,07
Ciervo californiano de cola negra	2,5	0,05
Ciervo de Vancouver de cola negra	2,9	0,05
Ciervo norteamericano de orejas largas	2,5	0,05
Ciervo de cola blanca	2,8	0,07

$n = 12$ para cada media. Hágase un análisis de la varianza y una prueba de comparaciones múltiples, utilizando el procedimiento sumas de cuadrados *P.T.S.* SOLUCION. $M.C._{\text{intra}} = 0{,}0416$, grupos máximos no significativos (para $P = 0{,}05$) son las muestras 1, 3, 5 y 2, 4 (numeradas por orden de magnitud de las medias).

Capítulo 9

Análisis de la varianza de clasificación doble

Del análisis de la varianza de clasificación simple del capítulo 8, pasamos al análisis de la varianza de clasificación doble por un simple paso lógico. Los ítems individuales pueden agruparse en clases que representan las diferentes combinaciones posibles de dos tratamientos o factores. Así las longitudes del ala de moscas domésticas estudiadas en capítulos anteriores, las cuales daban muestras representantes de diferentes formulaciones del medio, podrían dividirse también en machos y hembras. No solamente nos gustaría conocer si el medio 1 determinaba una longitud del ala diferente que el medio 2, sino también si las moscas macho diferían en longitud del ala respecto de las moscas hembras. Evidentemente cada combinación de factores estaría representada por una muestra de moscas. Así para siete medias y dos sexos necesitamos al menos $7 \times 2 = 14$ muestras. Igualmente, el experimento que contrasta cinco tratamientos de azúcar en secciones de guisantes (cuadro 9.4) podría haberse realizado a tres temperaturas diferentes. Esto habría conducido a un *análisis de la varianza de clasificación doble* tanto de los efectos de los azúcares como de las temperaturas.

Este método de análisis de varianza supone que una determinada temperatura y un determinado azúcar contribuyen en cierta medida al crecimiento de una sección de guisante y que estas dos aportaciones suman sus efectos sin influenciarse entre sí. En la sección 9.1 veremos cómo se miden las desviaciones de este supuesto; consideraremos también la expresión para descomponer variantes en un análisis de la varianza de clasificación doble.

En el presente diseño los dos factores pueden representar efectos modelo I o modelo II o uno de cada, en cuyo caso hablamos de un *modelo mixto*.

El cálculo de un análisis de la varianza de clasificación doble para subclases múltiples (más de una variante por subclase o factor de combinación), se muestra en la sección 9.1, que contiene además una discusión del significado de la interacción tal como se utiliza en estadística. La prueba de significación en un análisis de la varianza de clasificación doble es materia de la sección 9.2. Esta va seguida por una sección 9.3 de análisis de la varianza

de clasificación doble sin réplica o con una sola variante por subclase. El bien conocido método de comparaciones apareadas es un caso especial de análisis de la varianza de clasificación doble sin réplica.

Pasaremos ahora a exponer el cálculo de un análisis de la varianza de clasificación doble. Conforme expliquemos los cálculos nos podremos formar una idea más precisa de este modelo.

9.1 Análisis de la varianza de clasificación doble con réplica

Representamos el cálculo de un análisis de la varianza de clasificación doble, en un estudio de consumo de oxígeno por dos especies de lapas en tres concentraciones salinas de agua del mar. Para cada combinación de especie y salinidad del agua del mar se han obtenido ocho lecturas repetidas. Hemos continuado llamando a al número de columnas y llamamos b al número de filas. El tamaño de muestreo para cada casilla (combinación de fila y columna) de la tabla es n. Las casillas se denominan también subgrupos o subclases.

Los datos figuran en el cuadro 9.1. Los pasos a seguir en el cálculo designados como *Cálculos preliminares* proporcionan un procedimiento eficiente para el análisis de la varianza, pero haremos algunas digresiones para asegurar que los conceptos fundamentales de este modelo son apreciados por el lector. Comenzaremos por considerar las seis subclases como si fueran seis grupos en un análisis de la varianza de clasificación simple. Cada subgrupo o subclase representa ocho lecturas de consumo de oxígeno. Este análisis de la varianza probaría si hay alguna variación entre los seis subgrupos. Si no existiera esta variación aditiva sería improbable que ni la especie ni la salinidad afectaran significativamente la toma de oxígeno. Los pasos **1** al **3** en el cuadro 9.1 corresponden a los mismos pasos en el cuadro 8.1, aunque el simbolismo ha cambiado ligeramente ya que en lugar de a grupos ahora tenemos ab subgrupos. Para completar el análisis de la varianza necesitamos un término de corrección que se designa como paso **6** en el cuadro 9.1. A partir de estas cantidades obtenemos $S.C._{\text{total}}$ y $S.C._{\text{intragrupos}}$ en los pasos **7**, **8** y **12**, que corresponden a los **5**, **6** y **7** en el esquema del cuadro 8.1. Los resultados de este análisis de varianza preliminar se presentan en la tabla 9.1. Ellos indican claramente que hay considerable variación aditiva entre subgrupos, lo que hace probable que encontremos efectos significativos para al menos uno de los factores.

A continuación del cálculo se hallan las sumas de cuadrados para las filas y columnas de la tabla. Esto se hace por la fórmula general expuesta al final de la sección 8.1. Así para las columnas, elevamos al cuadrado las sumas de columna, sumamos estos cuadrados y los dividimos por 24, el número de ítems por fila. Este es el paso **4** en el cuadro 9.1. Una cantidad similar se calcula para las filas (paso **5**). De estos cocientes restamos el término de corrección, calculado como cantidad **6**. Estas sustracciones se efectúan como pasos **9** y **10**, respectivamente. Como las filas y columnas están basadas en tamaños de muestra iguales, no tenemos que obtener cocientes diferentes para el cuadrado de cada suma de fila o columna sino efectuar una sola división después de acumular los cuadrados de las sumas.

CUADRO 9.1

Análisis de la varianza de clasificación doble con réplica.

Velocidades de consumo de oxígeno de dos especies de lapas, *Acmaea scabra* y *A. digitalis*, a tres concentraciones salinas del agua de mar. La variable que se mide es μl O_2/mg peso corporal seco/min a 22° C. Hay ocho réplicas por combinación de especie y salinidad ($n = 8$). Este es un análisis de varianza modelo I.

Factor B: concentraciones salinas del agua del mar ($b = 3$)	*Factor A: Especie* ($a = 2$) *Acmaea scabra*	*Acmaea digitalis*	Σ
100%	7,16 8,26	6,14 6,14	
	6,78 14,00	3,86 10,00	
	13,60 16,10	10,40 11,60	
	8,93 9,66	5,49 5,80	
	Σ = 84,49	Σ = 59,43	143,92
75%	5,20 13,20	4,47 4,95	
	5,20 8,39	9,90 6,49	
	7,18 10,40	5,75 5,44	
	6,37 7,18	11,80 9,90	
	Σ = 63,12	Σ = 58,70	121,82
50%	11,11 10,50	9,63 14,50	
	9,74 14,60	6,38 10,20	
	18,80 11,10	13,40 17,70	
	9,74 11,80	14,50 12,30	
	Σ = 97,39	Σ = 98,61	196,00
Σ	245,00	216,74	461,74

Fuente: Estudio no publicado de F. J. Rohlf.

CUADRO 9.1 (continuación)

Cálculos preliminares

1. Suma total $= \sum^{a}\sum^{b}\sum^{n} Y = 461.74$

2. Suma de los cuadrados de las observaciones $= \sum^{a}\sum^{b}\sum^{n} Y^2 = (7,16)^2 + \cdots + (12,30)^2 = 5065,1530$

3. Suma de los cuadrados de las sumas de subgrupo (casilla), dividida por el tamaño de muestreo de los subgrupos

$$= \frac{\sum^{a}\sum^{b}\left(\sum^{n} Y\right)^2}{n} = \frac{[(84,49)^2 + \cdots + (98,61)^2]}{8} = 4663,6317$$

4. Suma de los cuadrados de las sumas de fila dividida por el tamaño de muestreo de una fila $= \dfrac{\sum^{a}\left(\sum^{b}\sum^{n} Y\right)^2}{bn}$

$$= \frac{[(245,00)^2 + (216,74)^2]}{(3 \times 8)} = 4458,3844$$

5. ,Suma de los cuadrados de las sumas de fila dividida por el tamaño de muestreo de una fila $= \dfrac{\sum^{b}\left(\sum^{a}\sum^{n} Y\right)^2}{an}$

$$= \frac{[(143,92)^2 + (121,82)^2 + (196,00)^2]}{(2 \times 8)} = 4623,0674$$

6. Suma total al cuadrado y dividida por el tamaño de muestreo total = término de corrección *T.C.*

$$= \frac{\left(\sum^{a}\sum^{b}\sum^{n} Y\right)^2}{abn} = \frac{(\text{quantity } 1)^2}{abn} = \frac{(461,74)^2}{(2 \times 3 \times 8)} = 4441,7464$$

7. $S.C._{\text{total}} = \sum^{a}\sum^{b}\sum^{n} Y^2 - T.C.$ = cantidad **2** − cantidad **6** = 5 065,1530 − 4 441,7464 = 623,4066

8. $S.C._{\text{subgrupos}} = \dfrac{\sum^{a}\sum^{b}\left(\sum^{n} Y\right)^2}{n} - T.C.$ = cantidad **3** − cantidad **6** = 4 663,6317 − 4 441,7464 = 221,8853.

9. $S.C._A$ (*S.C.* de las columnas) $= \dfrac{\sum^{a}\left(\sum^{b}\sum^{n} Y\right)^2}{bn} - T.C.$ = cantidad **4** − cantidad **6** = 4 458,3844 − 4 441,7464 = 16,6380.

10. $S.C._B$ (*S.C.* de las filas) $= \dfrac{\sum^{b}\left(\sum^{a}\sum^{n} Y\right)^2}{an} - T.C.$ = cantidad **5** − cantidad **6** = 4 623,0674 − 4 441,7464 = 181,3210.

11. $S.C._{A\times B}$ (*S.C.* de la Interacción) $= S.C._{\text{subgrupos}} - SC_A - S.C._B$ = cantidad **8** − cantidad **9** − cantidad **10**
$= 221,8853 - 16,6380 - 181,3210 = 23,9263$

12. $S.C._{\text{intra}}$ (Intrasubgrupos; *S.C.* de los errores) $= S.C._{\text{total}} - S.C._{\text{subgrupos}}$ = cantidad 7 − cantidad 8 = 623,4066
$- 221,8853 = 401,5213$

Como comprobación de los cálculos, cerciórese de que son aplicables las siguientes relaciones para algunas de las cantidades anteriores: $2 \geqslant 3 \geqslant 4 \geqslant 6$; $3 \geqslant 5 \geqslant 6$.

Ahora rellénese la tabla de análisis de la varianza.

Origen de la variación		*g.l.*	*S.C.*	*M.C.*	*M.C. esperada (Modelo I)*
$\bar{Y} - \bar{\bar{Y}}$	Subgrupos	$ab - 1$	**8**	$\dfrac{\mathbf{8}}{(ab-1)}$	
$\bar{Y}_A - \bar{\bar{Y}}$	A (columnas)	$a - 1$	**9**	$\dfrac{\mathbf{9}}{(a-1)}$	$\sigma^2 + \dfrac{nb}{a-1}\sum^{a}\alpha^2$
$\bar{Y}_B - \bar{\bar{Y}}$	B (filas)	$b - 1$	**10**	$\dfrac{\mathbf{10}}{(b-1)}$	$\sigma^2 + \dfrac{na}{b-1}\sum^{b}\beta^2$
$\bar{Y} - \bar{Y}_A - \bar{Y}_B + \bar{\bar{Y}}$	$A \times B$ (interacción)	$(a-1)(b-1)$	**11**	$\dfrac{\mathbf{11}}{(a-1)(b-1)}$	$\sigma^2 + \dfrac{n}{(a-1)(b-1)}\sum^{ab}(\alpha\beta)^2$
$Y - \bar{Y}$	Intrasubgrupos	$ab(n-1)$	**12**	$\dfrac{\mathbf{12}}{ab(n-1)}$	σ^2
$Y - \bar{\bar{Y}}$	Total	$abn - 1$	**7**		

Ya que el presente ejemplo es un análisis de la varianza modelo I para ambos factores, las *M.C.* esperadas anteriores son correctas. Más abajo se dan las expresiones correspondientes para otros modelos.

CUADRO 9.1 (continuación)

Origen de la variación	*Modelo II*	*Modelo mixto (A fijos, B aleatorios)*
A	$\sigma^2 + n\sigma^2_{A\times B} + nb\sigma^2_A$	$\sigma^2 + n\sigma^2_{A\times B} + \frac{nb}{a-1}\sum^{a}\alpha^2$
B	$\sigma^2 + n\sigma^2_{A\times B} + na\sigma^2_B$	$\sigma^2 + na\sigma^2_B$
$A \times B$	$\sigma^2 + n\sigma^2_{A\times B}$	$\sigma^2 + n\sigma^2_{A\times B}$
Intrasubgrupos	σ^2	σ^2

Tabla de análisis de la varianza

Origen de la variación	*g.l.*	*S.C.*	*M.C.*	F_s
Subgrupos	5	221,8853	44,377	
A (columnas; especies)	1	16,6380	16,638	1,740 *ns*
B (filas; salinidades)	2	181,3210	90,660	9,483**
$A \times B$ (interacción)	2	23,9263	11,963	1,251 *ns*
Intrasubgrupos (error)	42	401,5213	9,560	
Total	47	623,4066		

$F_{0,05[1,42]} = 4{,}07 \qquad F_{0,05[2,42]} = 3{,}22 \qquad F_{0,01[2,42]} = 5{,}15$

Como este es un análisis de la varianza modelo I, todas las medias cuadráticas se contrastan con la *M.C.* del error. Para una discusión de pruebas de significación, véase sección 9.2.

Conclusiones. – El consumo de oxígeno no difiere significativamente entre las dos especies de lapas pero varía con la salinidad. El consumo de O_2 aumenta a una salinidad del 50 %. La salinidad parece que afecta por igual a las dos especies, puesto que hay insuficiente evidencia de una interacción especie × salinidad.

TABLA 9.1

Análisis preliminar de subgrupos en el análisis de varianza de clasificación doble. Datos del cuadro 9.1

Origen de la variación	*g.l.*	*S.C.*	*M.C.*
$\bar{Y} - \bar{\bar{Y}}$ Entre subgrupos	5 $ab-1$	221,8853	44,377**
$Y - \bar{Y}$ Intrasubgrupos	42 $ab(n-1)$	401,5213	9,560
$Y - \bar{\bar{Y}}$ Total	47 $abn-1$	623,4066	

Vamos a volver por un momento al análisis de la varianza preliminar de la tabla 9.1 que dividía la suma de cuadrados totales en dos partes, la suma de cuadrados entre los seis subgrupos y la intrasubgrupos, la suma de cuadrados de los errores. Las nuevas sumas de cuadrados pertenecientes a efectos de fila y de columna, sin duda no son parte del error, pero deben contribuir a la diferencia que comprende la suma de cuadrados entre los cuatro subgrupos. Por lo tanto restamos la *S.C.* entre filas y columnas de la *S.C.* entre subgrupos. Esta última es 221,8853. La *S.C.* entre filas es 181,3210, y la *S.C.* entre columnas es 16,6380. Juntas suman 197,9590, casi pero no totalmente el valor de la suma de cuadrados entre subgrupos. La diferencia representa una tercera suma de cuadrados denominada *suma de cuadrados de la interacción*, cuyo valor en este caso es 23,9263. Luego discutiremos el significado de esta nueva suma de cuadrados. De momento solamente vamos a decir que está casi siempre presente (pero no es necesariamente significativa) y generalmente no tiene que calcularse independientemente sino que puede obtenerse, como se ha representado anteriormente, por sustracción de la *S.C.* entre filas y la *S.C.* entre columnas de la *S.C.* entre subgrupos. Este procedimiento se muestra gráficamente en la figura 9.1 que representa la *S.C.* entre subgrupos y la *S.C.* intrasubgrupos (error), subdividiéndose la primera en la *S.C.* entre filas, *S.C.* entre columnas y *S.C.* de la interacción. Las magnitudes relativas de estas sumas de cuadrados variarán de un experimento a otro. En la figura 9.1 no se representan proporcionales a sus valores reales en el experimento de lapas; de lo contrario el área que representa la *S.C.* entre filas tendría que ser aproximadamente once veces la asignada a la *S.C.* entre columnas. Antes de que podamos probar inteligentemente la significación en este análisis de la varianza, debemos conocer el significado de la interacción.

Podemos explicar mejor la interacción en un análisis de la varianza de clasificación doble, valiéndonos de un ejemplo artificial basado en los datos de lapas que acabamos de estudiar. Si intercambiamos las lecturas para el 75 % y 50 % de *A. digitalis* solamente, obtenemos la tabla de datos que aparece en la tabla 9.2. Solamente se indican las sumas de los subgrupos, filas y columnas. Completamos el análisis de la varianza del modo descrito anteriormente y observamos los resultados al pie de la tabla 9.2. La *S.C.* total, la *S.C.* entre subgrupos y la *S.C.* de los errores son las mismas que antes (tabla 9.1). Esto no debería ser sorprendente puesto que utilizamos los mismos datos. Todo lo que hemos hecho es intercambiar los contenidos de las dos casillas inferiores en la columna de la derecha de la tabla. Cuando descomponemos la *S.C.* entre subgrupos hallamos algunas diferencias. Observamos que la *S.C.* entre especies (entre columnas) no ha cambiado.

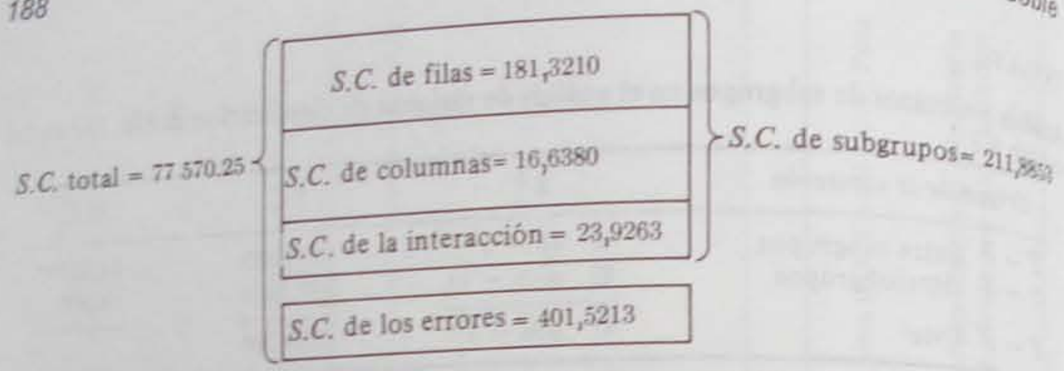


Fig. 9.1. Representación esquemática de la descomposición de las sumas de cuadrados totales en un análisis de la varianza de clasificación doble ortogonal. Las áreas de las subdivisiones no se representan proporcionales a las magnitudes de las sumas de cuadrados.

Como el cambio que hemos hecho ha sido dentro de una columna, el total para esa columna no se ha alterado y en consecuencia la *S.C.* entre columnas no ha variado. Sin embargo, las sumas de la segunda y tercera filas se han modificado apreciablemente como resultado del intercambio de las lecturas para el 75 % y 50 % de salinidad en *A. digitalis*.

TABLA 9.2

Ejemplo artificial para explicar el significado de la interacción. Se han intercambiado las lecturas del cuadro 9.1 para las concentraciones salinas del 75 % y 50 % de *Acmaea digitalis*. Sólo se indican las sumas de subgrupo y las marginales.

	Especies		
Salinidad del agua del mar	*A. scabra*	*A. digitalis*	Σ
100%	84,49	59,43	143,92
75%	63,12	98,61	161,73
50%	97,39	58,70	156,09
Σ	245,00	216,74	461,74

Análisis de la varianza completado

Origen de la variación	*g.l.*	*S.C.*	*M.C.*
Subgrupos	5	221,8853	44,377**
Especies	1	16,6380	16,638 *ns*
Salinidad	2	10,3566	5,178 *ns*
Sp X Sal	2	194,8907	97,445**
Error	42	401,5213	9,560
Total	47	623,4066	

La suma para el 75 % de salinidad es ahora muy próxima a la que se halla para el 50 % y la diferencia entre las salinidades, previamente muy notable, ahora ya no es tan grande. Por el contrario, la *S.C.* de la interacción, obtenida restando las sumas de cuadrados entre filas y columnas de la *S.C.* entre subgrupos es ahora una cantidad grande. Recuérdese que la *S.C.* entre subgrupos es la misma en los dos ejemplos. En el primero hemos restado las sumas de cuadrados debidas a los efectos de los dos factores, especies y salinidades, dejando solamente un pequeño residual que representa la interacción. En el segundo ejemplo estos dos *efectos principales* (especies y salinidades) solamente responden de una pequeña parte de la suma de cuadrados entre subgrupos, quedando la suma de cuadrados de la interacción como un residuo considerable. ¿Cuál es la diferencia esencial entre estos dos ejemplos?

En la tabla 9.3 hemos mostrado las medias de subgrupos y marginales para los datos originales de la tabla 9.1 y para los datos "tratados" de la tabla 9.2. Los resultados originales están muy claros: al 75 % de salinidad el consumo de oxígeno es inferior que a las otras dos salinidades y esto es cierto para las dos especies. Observamos además que *A. scabra* consume más oxígeno que *A. digitalis* a dos de las salinidades. Así nuestras afirmaciones respecto a las diferencias debidas a las especies o a la salinidad pueden hacerse con toda independencia una de otra. Sin embargo, si tuviésemos que interpretar los datos artificiales (mitad inferior de la tabla 9.3), observaríamos que aunque *A. scabra* sigue consumiendo más oxígeno que *A. digitalis* (puesto que las sumas de columnas no han cambiado), esta diferencia depende en gran parte de la salinidad. Al 100 % y 50 % *A. scabra* consume considerablemente más oxígeno que *A. digitalis*, pero al 75 % esta rela-

TABLA 9.3

Comparación de medias de los datos de las tablas 9.1 y 9.2

	Especies		
Salinidad de agua de mar	*A. scabra*	*A. digitalis*	*Media*
	Datos originales de la tabla 9.1		
100%	10,56	7,43	9,00
75%	7,89	7,34	7,61
50%	12,17	12,33	12,25
Media	10,21	9,03	9,62
	Datos artificiales de la tabla 9.2		
100%	10,56	7,43	9,00
75%	7,89	12,33	10,11
50%	12,17	7,34	9,76
Media	10,21	9,03	9,62

ción se invierte. De este modo, ya no podemos hacer una afirmación inequívoca acerca de la cantidad de oxígeno tomado por las dos especies. Tenemos que calificar nuestra afirmación con la concentración salina a la que se mantienen. Al 100 % y 50 % $\bar{Y}_{scabra} > \bar{Y}_{digitalis}$, pero al 75 % $\bar{Y}_{scabra} < \bar{Y}_{digitalis}$. Si examinamos los efectos de la salinidad en el ejemplo artificial, observamos un ligero incremento en el consumo de oxígeno al 75 %. Sin embargo, nuevamente tenemos que calificar esta afirmación con la especie de la lapa consumidora; *scabra* tiene el menor consumo al 75 %, mientras *digitalis* consume la mayor cantidad a esta concentración.

A esta dependencia del efecto de un factor respecto del nivel de otro factor, se le denomina *interacción*. Es una idea científica común y fundamental. Indica que los efectos de los dos factores no son simplemente aditivos sino que cualquier combinación dada de niveles de factores tales como salinidad combinado con especie, aporta un incremento positivo o negativo al nivel de expresión de la variable. En la terminología biológica ordinaria un gran incremento positivo de este tipo se denomina *sinergismo*. Cuando las drogas actúan sinérgicamente, el resultado de la interacción de las dos drogas puede ser superior a los efectos aislados de cada droga. Cuando una combinación de niveles de dos factores inhiben mutuamente sus efectos lo llamamos *interferencia*. Sinergismo e interferencia tenderán ambos a aumentar la *S.C.* de la interacción.

El contraste de la interacción es un procedimiento importante en análisis de la varianza. Si los datos artificiales de la tabla 9.2 fuesen reales, sería de poco valor afirmar que el 75 % de salinidad lleva a un consumo de oxígeno ligeramente mayor. Esta afirmación encubre completamente las diferencias importantes en los datos, las cuales son que *scabra* tiene el menor consumo a esta concentración, mientras *digitalis* consume la mayor cantidad.

Ya podemos escribir una expresión que simbolice la descomposición de una variante individual en un análisis de la varianza de clasificación doble, a semejanza de la expresión (7.3), para el análisis de la varianza de clasificación simple. La expresión que sigue supone que ambos factores representan efectos de tratamiento fijos, modelo I. Esto parece razonable, ya que tanto las especies como la salinidad son tratamientos fijos. La variante Y_{ijk} es el ítem k en el subgrupo que representa el grupo i de tratamiento A y el grupo j de tratamiento B. Se descompone como sigue:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \tag{9.1}$$

donde μ es la media paramétrica de la población, α_i es el efecto de tratamiento fijo para el grupo i de tratamiento A, β_j es el efecto de tratamiento fijo para el grupo j de tratamiento B, $(\alpha\beta)_{ij}$ es el efecto de interacción en el subgrupo que representa el grupo i de factor A y el grupo j de factor B, y ϵ_{ijk} es el término error del ítem k en el subgrupo ij. Suponemos, como es habitual, que ϵ_{ijk} está normalmente distribuido con una media de 0 y una varianza de σ^2. Si uno o los dos factores son modelo II, sustituimos la α_i y β_j de la fórmula por A_i y/o B_j.

En capítulos anteriores hemos visto que cada suma de cuadrados representa una suma de desviaciones al cuadrado. ¿Qué desviaciones reales representa una *S.C.* de la interacción? Esto podemos verlo fácilmente volviendo a los análisis de la varianza de la tabla 9.1. La variación entre subgrupos se representa por $\bar{Y} - \bar{\bar{Y}}$, donde $\bar{Y}$ simboliza la media de

subgrupo e $\bar{\bar{Y}}$ simboliza la media total. Cuando restamos las desviaciones debidas a las filas $\bar{F} - \bar{\bar{Y}}$ y a las columnas $\bar{C} - \bar{\bar{Y}}$ de las de los subgrupos obtenemos

$$(\bar{Y} - \bar{\bar{Y}}) - (\bar{F} - \bar{\bar{Y}}) - (\bar{C} - \bar{\bar{Y}}) = \bar{Y} - \bar{\bar{Y}} - \bar{F} + \bar{\bar{Y}} - \bar{C} + \bar{\bar{Y}}$$
$$= \bar{Y} - \bar{F} - \bar{C} + \bar{\bar{Y}}$$

Esta expresión un tanto complicada es la desviación debida a la interacción. Cuando calculamos una expresión de este tipo para cada subgrupo, la elevamos al cuadrado, sumamos estos cuadrados y multiplicamos la suma por n, obtenemos la *S.C.* de la interacción. Esta separación de las desviaciones se conserva también para sus cuadrados. Esto se debe a que las sumas de los productos de los diferentes términos se anulan.

Un método sencillo para revelar la naturaleza de la interacción presente en los datos, es examinar las medias de la tabla de datos originales. Esto podemos hacerlo en la tabla 9.3. Los datos originales, que no presentan interacción, darían el siguiente patrón de magnitudes relativas:

	scabra	*digitalis*
100%		
	V	V
75%		
	Λ	Λ
50%		

Las magnitudes relativas de las medias de la tabla inferior que da la interacción pueden resumirse como sigue:

	scabra	*digitalis*
100%		
	V	Λ
75%		
	Λ	V
50%		

Cuando el patrón de signos que expresa las magnitudes relativas no es uniforme, como en la tabla inferior, está indicada la interacción. Mientras el patrón de medias sea uniforme, como en la tabla superior, la interacción puede no presentarse. Sin embargo, la interacción se presenta frecuentemente sin cambio en la *dirección* de las diferencias; solamente pueden verse afectadas las magnitudes relativas. Así pues, el examen visual no debería sustituir a la prueba estadística.

En resumen, cuando el efecto de dos tratamientos aplicados juntos no puede predecirse del promedio de respuestas de los factores por separado, los estadísticos llaman a este fenómeno interacción y prueban su significación por medio de una media cuadrática de la interacción. Esta es una relación científica muy común. Si decimos que el efecto de la densidad en la fecundidad o en el peso de un coleóptero depende de su genotipo, implicamos que está presente una interacción genotipo X densidad. Si la variación geográfica de

un parásito depende de la naturaleza de la especie huésped a la que ataca, hablamos de una interacción huésped X localidad. Si el efecto de la temperatura en un proceso metabólico es independiente del efecto de la concentración de oxígeno, decimos que no hay interacción temperatura X oxígeno.

La prueba de significación en un análisis de la varianza de clasificación doble, se pospondrá hasta la próxima sección. No obstante, deberíamos hacer notar que los pasos 4 y 9 del cuadro 9.1 podrían haberse acortado utilizando la fórmula simplificada para una suma de cuadrados entre *dos* grupos ilustrada en la sección 8.4. En un análisis con dos filas y dos columnas solamente, la *S.C.* de la interacción puede calcularse directamente como

$$(\text{suma de una diagonal} - \text{suma de otra diagonal})^2 / abn$$

9.2 Análisis de la varianza de clasificación doble: prueba de significación

Antes de que podamos contrastar hipótesis sobre las fuentes de variación aisladas en el cuadro 9.1, debemos familiarizarnos con las medias cuadráticas esperadas para este modelo. En la tabla de análisis de la varianza del cuadro 9.1 mostramos en primer lugar las medias cuadráticas esperadas para el modelo I, siendo las diferencias de especies y las salinidades del agua del mar efectos de tratamiento fijo. Obsérvese que la *M.C.* intrasubgrupos o *M.C.* del error de nuevo estima la varianza paramétrica de los ítems. El más importante dato a tener en cuenta sobre un análisis de la varianza modelo I es que la media cuadrática a cada nivel de variación incluye solamente el efecto aditivo debido a ese nivel de tratamiento; con excepción de la varianza paramétrica de los ítems, no contienen ningún término de una línea inferior. Así, la *M.C.* esperada del factor *A* contiene solamente la varianza paramétrica de los ítems más el término aditivo debido al factor *A*, pero no incluye además efectos de interacción. Por lo tanto, en el modelo I, la prueba de significación es sencilla y directa. Cualquier fuente de variación se contrasta con la razón de varianzas de la media cuadrática apropiada a la *M.C.* de los errores. De este modo, para las pruebas apropiadas utilizamos las razones de varianza A/**Error**, B/**Error** y $(A \times B)$/Error, donde cada término de trazo grueso significa una media cuadrática. Así $A = M.C._A$, Error $= M.C._{intra}$.

Cuando hacemos esto en el ejemplo del cuadro 9.1 solamente encontramos significativo el factor *B*, la salinidad. Ni el factor *A* ni la interacción son significativos. Concluimos que las diferencias en consumo de oxígeno son inducidas por variación de las salinidades (las dos variables aparecen inversamente relacionadas), y no parece haber evidencia suficiente de que las especies se diferencien en consumo de oxígeno. La tabulación de las magnitudes relativas de las medias en la sección anterior revela que el patrón de signos en las dos líneas es idéntico. Sin embargo, esto puede inducir a error, ya que la media de *A. scabra* es muy superior al 100 % de salinidad que al 75 %, pero la de *A. digitalis* es sólo muy ligeramente superior. Aunque las curvas de consumo de oxígeno de las dos especies cuando se representan gráficamente aparecen lejos del paralelismo (véase figura 9.2), esta indicación de una interacción especie X salinidad no puede mostrarse significativa cuando se compara con la varianza intrasubgrupos. El hallar una diferencia significativa entre las

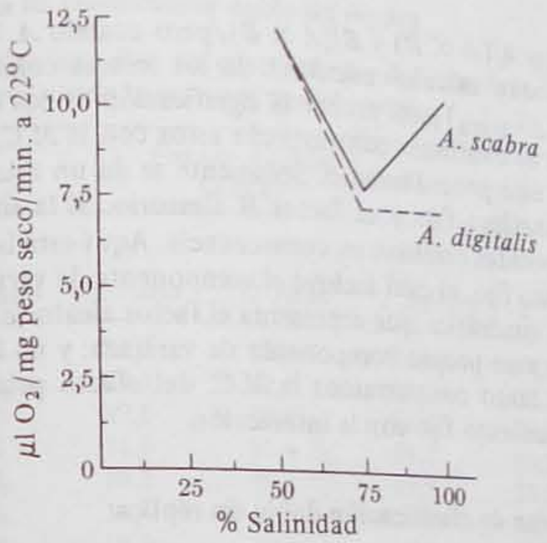


Fig. 9.2. Consumo de oxígeno por dos especies de lapas a tres salinidades. Datos del cuadro 9.1.

salinidades no concluye el análisis. Los datos sugieren que al 75 % de salinidad hay una reducción real del consumo de oxígeno. Si esto es realmente cierto podría demostrarse por los métodos de la sección 8.6.

Cuando analizamos los resultados del ejemplo artificial en la tabla 9.2, solamente hallamos significativa la *M.C.* de la interacción. Así concluiríamos que la respuesta a la salinidad difiere en las dos especies. Esto se demuestra por el examen de los datos, el cual revela que al 75 % de salinidad *A. scabra* consume la menor cantidad de oxígeno y *A. digitalis* la mayor.

En el último ejemplo (artificial) las medias cuadráticas de los dos factores (efectos principales) no son significativas en ningún caso. Sin embargo, muchos estadísticos ni siquiera las contrastarían una vez que encontrasen que la media cuadrática de la interacción era significativa, ya que en este caso un enunciado completo para cada factor tendría poco sentido. Una simple manifestación de respuesta a la salinidad sería incompleta. La presencia de interacción nos hace calificar nuestras afirmaciones: el patrón de respuesta a los cambios en la salinidad es diferente en las dos especies. En consecuencia tendríamos que describir curvas de respuesta diferentes, no paralelas, para las dos especies. En ocasiones resulta importante la prueba de significación global en un análisis de la varianza modelo I, a pesar de la presencia de interacción. Podemos querer demostrar la significación del efecto de una droga independientemente de su interacción significativa con la edad del paciente. Para justificar este argumento podríamos querer contrastar la media cuadrática entre concentraciones de droga (sobre la *M.C.* del error) independientemente de que la *M.C.* de la interacción sea significativa.

El cuadro 9.1 señala además las medias cuadráticas esperadas para un modelo mixto y un modelo II. En el modelo II observamos que los dos efectos principales contienen el componente de varianza de la interacción así como su propio componente de varianza. En un modelo II contrastamos primero $(A \times B)$/**Error**. Si la interacción es significativa

continuamos probando $A/(A \times B)$ y $B/(A \times B)$, pero cuando $A \times B$ no es significativa algunos autores proponen calcular una *M.C.* de los errores combinada $= (S.C._{A\times B} + S.C._{\text{intra}})/(g.l._{A\times B} + g.l._{\text{intra}})$ para probar la significación de los efectos principales. La posición conservadora es continuar contrastando éstos con la *M.C.* de la interacción y en este libro seguiremos este procedimiento. Solamente se da un tipo de modelo mixto en que el factor *A* se considera fijo y el factor *B* aleatorio. Si la situación se invierte, las medias cuadráticas esperadas cambian en consecuencia. Aquí está la media cuadrática que representa el tratamiento fijo, el cual incluye el componente de varianza de la interacción, mientras que la media cuadrática que representa el factor aleatorio contiene solamente la varianza de los errores y su propio componente de varianza; y no incluye el componente de la interacción. Por tanto contrastamos la *M.C.* del efecto principal aleatorio con el error y la *M.C.* del tratamiento fijo con la interacción.

9.3 Análisis de la varianza de clasificación doble sin réplica

En muchos experimentos no habrá réplica para cada combinación de factores representada por una casilla en la tabla de datos. En tales casos no podemos hablar sencillamente de "subgrupos", ya que cada casilla contiene solamente una lectura. Con frecuencia puede resultar demasiado difícil o demasiado caro obtener más de una lectura por casilla, o puede saberse que las medidas son tan repetibles que sea de poca importancia estimar su error. Como veremos más abajo, un análisis de la varianza de clasificación doble sin réplica solamente puede aplicarse correctamente con ciertas consideraciones. Para algunos modelos y pruebas del análisis de la varianza debemos suponer que no existe interacción.

Nuestro ejemplo para este modelo está tomado de limnología. En el cuadro 9.2 mostramos las temperaturas de un lago tomadas aproximadamente a la misma hora, en cuatro tardes de verano sucesivas. Estas medidas de temperatura se han hecho a diez profundidades diferentes y se ha visto que son muy repetibles. Por lo tanto, solamente se han tomado lecturas individuales a cada profundidad en cualquier día. ¿Cuál es el modelo apropiado para este análisis de la varianza? Sin duda las profundidades son modelo I. Los cuatro días, sin embargo, probablemente no son de peculiar interés. Es improbable que un investigador se pregunte si el agua estaba más fría el 30 de julio que el 31 de julio. Una forma más significativa de enfocar este problema sería considerar estos cuatro días de verano como muestras al azar que nos permitan estimar la variación día a día de la estratificación de temperatura en el lago durante el período estival estable.

En el cuadro 9.2 se presentan los cálculos. Son los mismos que los del cuadro 9.1 excepto que las expresiones a calcular son considerablemente más simples. Como $n = 1$, muchas de las sumas pueden omitirse. En este ejemplo la suma de cuadrados entre subgrupos es la misma que la suma de cuadrados total. Si esto no es intuitivamente claro, consúltese la figura 9.3, la cual, cuando se compara con la figura 9,1, demuestra que la suma de cuadrados de los errores basada en la variación intrasubgrupos ha desaparecido en este ejemplo. Así, una vez restada la suma de cuadrados entre columnas (factor *A*) y entre filas (factor *B*) de la *S.C.* total, nos quedamos solamente con una suma de cuadrados que es el equivalente de la anterior *S.C.* de la interacción, pero que ahora es la única fuente de error en el análisis de la varianza. Esta *S.C.* es conocida como la *S.C. residual* o la *discrepancia.*

CUADRO 9.2

Análisis de la varianza de clasificación doble sin réplica.

Temperaturas (° C) del lago Rot en cuatro tardes del verano de 1952, a 10 profundidades. Este es un análisis de la varianza modelo mixto.

Factor B: Profundidades en metros ($b = 10$)	*Factor A: Días* ($a = 4$) *29 Julio*	*20 Julio*	*31 Julio*	*1 Agosto*	Σ
0	23,8	24,0	24,6	24,8	97,2
1	22,6	22,4	22,9	23,2	91,1
2	22,2	22,1	22,1	22,2	88,6
3	21,2	21,8	21,0	21,2	85,2
4	18,4	19,3	19,0	18,8	75,5
5	13,5	14,4	14,2	13,8	55,9
6	9,8	9,9	10,4	9,6	39,7
9	6,0	6,0	6,3	6,3	24,6
12	5,8	5,9	6,0	5,8	23,5
15,5	5,6	5,6	5,5	5,6	22,3
Σ	148,9	151,4	152,0	151,3	603,6

Fuente: Datos de Vollenweider y Frei (1953).

Los cuatro grupos de lecturas son tratados como réplicas (lotes) en este análisis. La profundidad es un efecto de tratamiento fijo, los días se consideran como efectos aleatorios, por lo tanto éste es un análisis de varianza modelo mixto.

Cálculos preliminares

1. Suma total $= \sum^{a}\sum^{b} Y = 603,6$
2. Suma de los cuadrados de las observaciones $= \sum^{a}\sum^{b} Y^2$
$= (23,8)^2 + \cdots + (5,6)^2 = 11\,230,78$
3. Suma de los cuadrados de las sumas de columna dividida por el tamaño de muestreo de una columna
$$= \frac{\sum\left(\sum^{b} Y\right)^2}{b} = \frac{[(148,9)^2 + \cdots + (151,3)^2]}{10} = 9108,89$$
4. Suma de los cuadrados de las sumas de fila dividida por el tamaño de muestreo de una fila
$$= \frac{\sum^{b}\left(\sum^{a} Y\right)^2}{a} = \frac{[(97,2)^2 + \cdots + (22,3)^2]}{4} = 11\,227,98$$
5. Suma total al cuadrado y dividida por el tamaño de muestreo total = término de corrección
$$T.C. = \frac{\left(\sum^{a}\sum^{b} Y\right)^2}{ab} = \frac{(\text{cantidad } 1)}{ab} = \frac{(603,6)^2}{40} = 9108,32$$

CUADRO 9.2 (continuación)

6. $S.C._{\text{total}} = \sum^{a}\sum^{b} Y^2 - T.C.$
= cantidad **2** − cantidad **5** = 11 230,78 − 9108,32 = 2122,46

7. $S.C._A$ (*S.C.* de columnas) $= \dfrac{\sum^{a}\left(\sum^{b} Y\right)^2}{b} - T.C.$
= cantidad **3** − cantidad **5** = 9 108,89 − 9 108,32 = 0,57

8. $S.C._B$ (*S.C.* de filas) $= \dfrac{\sum^{b}\left(\sum^{a} Y\right)^2}{a} - T.C.$
= cantidad **4** − cantidad **5** = 11 227,98 − 9 108,32 = 2 119,66

9. $S.C._{\text{error}}$ (residuo; discrepancia) $= S.C._{\text{total}} - S.C._A - S.C._B$
= cantidad **6** − cantidad **7** − cantidad **8** = 2 122,46 − 0,57 − 2 119,66
= 2,23

Tabla de análisis de la varianza

Fuente de la variación		*g.l.*	*S.C.*	*M.C.*	F_s	*M.C. esperada*
$\bar{Y}_A - \bar{\bar{Y}}$	A (columna; días)	3	0,57	0,190	2,30 *ns*	$\sigma^2 + b\sigma^2_A$
$\bar{Y}_B - \bar{\bar{Y}}$	B (filas; profundidades)	9	2119,66	235,5	2851,1**	$\sigma^2+\sigma^2_{AB}+\frac{a}{b-1}\Sigma\beta^2$
$Y - \bar{Y}_A - \bar{Y}_B + \bar{\bar{Y}}$	Error (residuo; discrepancia)	27	2,23	0,0826		$\sigma^2 + \sigma^2_{AB}$
$Y - \bar{\bar{Y}}$	Total	39	2122,46			

Conclusiones. – Al aumentar la profundidad se encuentra una disminución en la temperatura altamente significativa. Para contrastar "días" debemos suponer que la interacción entre los días y las profundidades es cero. No puede demostrarse ningún componente aditivo de la varianza entre días.

Si se refiriese a las medias cuadráticas esperadas para el análisis de la varianza de clasificación doble del cuadro 9.1, se descubriría por qué hemos afirmado anteriormente que para algunos modelos y pruebas de un análisis de la varianza de clasificación doble, debemos suponer que la interacción no es significativa. Si hay interacción, solamente puede contrastarse un modelo II, mientras que en un modelo mixto solamente puede contrastarse el nivel fijo con la media cuadrática residual. Pero en un modelo puro o para el factor aleatorio en un modelo mixto, sería incorrecto contrastar los efectos principales con el residual a no ser que pudiésemos considerar con seguridad que no está presente efecto aditivo alguno debido a la interacción. El examen general de los datos del cuadro

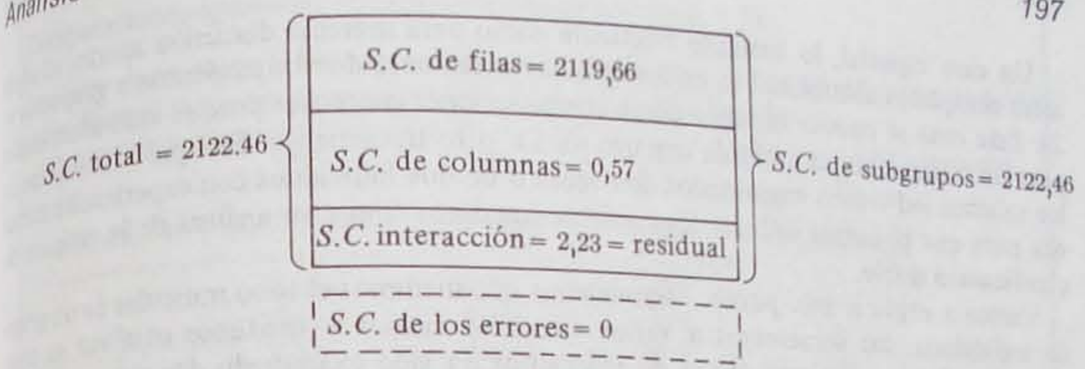


Fig. 9.3. Representación esquemática de la descomposición de la suma de cuadrados total en un análisis de la varianza ortogonal de clasificación doble sin réplica. Las áreas de las subdivisiones no aparecen proporcionales a las magnitudes de las sumas de cuadrados.

9.2 nos convence de que las tendencias de temperatura presentes en un día cualquiera se reproducen fielmente en los otros días. Así, es improbable que exista interacción. Si, por ejemplo, una fuerte tormenta hubiese agitado el lago en un día, cambiando las relaciones de temperatura a diversas profundidades, la interacción habría sido evidente, y la prueba de la media cuadrática entre días realizada en el cuadro 9.2 no habría sido razonable.

Como suponemos que no hay interacción, las medias cuadráticas de las filas y columnas se contrastan con la *M.C.* de los errores. Los resultados no son sorprendentes; el examen casual de los datos habría predicho nuestros hallazgos. La varianza aditiva no es significativa entre días, pero las diferencias de temperaturas debidas a las profundidades son altamente significativas, dando un valor de $F_s = 2\,851{,}1$. Sería sumamente improbable que tales diferencias se produjesen sólo por azar.

Una aplicación normal del análisis de la varianza de clasificación doble sin réplica es el *contraste repetido de los mismos individuos*. Con esto queremos decir que el mismo grupo de individuos se prueba repetidamente durante un período de tiempo. Los individuos son un factor (ordinariamente considerado como aleatorio y que sirve como réplica) y la dimensión tiempo es el segundo factor, un efecto de tratamiento fijo. Por ejemplo, podríamos medir el crecimiento de una estructura en diez individuos a intervalos regulares. Cuando se comprueba la presencia de un componente aditivo de la varianza (debido al factor aleatorio), este modelo supone nuevamente que no hay interacción entre el tiempo y los individuos; es decir, las respuestas de los distintos individuos son paralelas a lo largo del tiempo. Otro uso de este modelo se encuentra en diversos experimentos fisiológicos y psicológicos en los cuales examinamos el mismo grupo de individuos con respecto a la aparición de alguna respuesta después del tratamiento. Los ejemplos incluyen inmunidad creciente tras las inoculaciones de antígeno, respuestas alteradas tras el condicionamiento, y medidas de aprendizaje después de un número de pruebas. De este modo podemos estudiar la velocidad con que diez ratas, repetidamente probadas en el mismo laberinto, alcanzan el punto final. El efecto de tratamiento fijo serían las pruebas sucesivas a las que se han sometido las ratas. El segundo factor es aleatorio, representando probablemente una muestra al azar de ratas de la población del laboratorio.

Un caso especial, lo bastante frecuente como para merecer discusión aparte, es el de lotes completos aleatorizados, en los que hay solamente dos tratamientos o grupos ($a = 2$). Este caso se conoce también como *comparaciones apareadas* porque cada observación para un tratamiento va asociada con una para el otro tratamiento. Este par se compone de los mismos individuos examinados dos veces o de dos individuos con experiencias comunes para que podamos ordenar lógicamente los datos como un análisis de la varianza de clasificación doble.

Vamos a explicar este punto. Supongamos que medimos el tono muscular de un grupo de individuos, los sometemos a varios ejercicios físicos y medimos otra vez su tono muscular. Como el mismo grupo de individuos ha sido examinado dos veces, podemos ordenar por parejas nuestras lecturas de tono muscular, representando cada pareja las de un individuo (antes y después del ejercicio). Estos datos se tratan apropiadamente por un análisis de la varianza de clasificación doble sin replicación, que en este caso sería una prueba de comparaciones apareadas, porque hay solamente dos clases de tratamiento. Esta comparación "antes y después del tratamiento" es un modelo muy frecuente que conduce a comparaciones apareadas. Otro modelo mide simplemente dos estadios del desarrollo de un grupo de organismos, siendo el tiempo el tratamiento que interviene entre los dos estadios. El ejemplo del cuadro 9.3 es de este tipo. Mide el perímetro cefálico en un grupo de niñas de cinco años y en el mismo grupo de niñas cuando tienen seis años. La comparación apareada es para cada niña, entre su perímetro cefálico cuando tiene cinco años y su perímetro cefálico a los seis años.

Con frecuencia las comparaciones apareadas resultan de dividir un organismo u otra unidad individual de modo que la mitad reciba el tratamiento 1 y la otra mitad el tratamiento 2, que puede ser el control. Por tanto, si queremos probar la potencia de dos antígenos o alérgenos podríamos inyectar uno en cada brazo de un solo individuo y medir el diámetro del área enrojecida que se produce. No sería prudente desde el punto de vista del diseño experimental probar el antígeno 1 en el individuo 1 y el antígeno 2 en el individuo 2. Estos individuos pueden tener diferente susceptibilidad a estos antígenos y podemos averiguar poco acerca de la potencia relativa de los antígenos, ya que ésta se confundiría por las respuestas diferenciales de los sujetos. Un diseño mucho mejor sería inyectar el antígeno 1 en el brazo izquierdo y el antígeno 2 en el brazo derecho de un grupo de n individuos y analizar los datos como un análisis de varianza de clasificación doble sin réplica con n filas (individuos) y dos columnas (tratamientos). Probablemente sea indiferente que un antígeno se inyecte en el brazo derecho o en el izquierdo, pero si tuviésemos que diseñar un experimento de este tipo y conociésemos poco sobre la reacción de los humanos a los antígenos podríamos, como precaución, asignar al azar el antígeno 1 al brazo izquierdo o derecho para los diferentes sujetos, inyectando el antígeno 2 en el brazo contrario. Un ejemplo similar es la titulación de ciertos virus de plantas frotando una cierta concentración del virus sobre la superficie de una hoja y contando las lesiones resultantes. Puesto que las diferentes hojas tienen diferente susceptibilidad, una manera convencional de medir la potencia del virus es frotarlo sobre la mitad de la hoja a un lado del nervio central, frotando la otra mitad de la hoja con una disolución control o standard.

Otro diseño que conduce a comparaciones apareadas es cuando el tratamiento se da a dos individuos que comparten una experiencia común, sea ésta genética o ambiental. Así,

CUADRO 9.3

Comparaciones apareadas (bloques aleatorizados con $a = 2$).

Anchura de la parte inferior de la cara (diámetro esquelético bigonial en cm) de 15 niñas blancas norteamericanas medida a los 5 y 6 años de edad.

Individuos	(1) *De 5 años*	(2) *De 6 años*	(3) Σ	(4) $D = Y_{(2)} - Y_{(1)}$ *(diferencia)*
1	7,33	7,53	14,86	0,20
2	7,49	7,70	15,19	,21
3	7,27	7,46	14,73	,19
4	7,93	8,21	16,14	,28
5	7,56	7,81	15,37	,25
6	7,81	8,01	15,82	,20
7	7,46	7,72	15,18	,26
8	6,94	7,13	14,07	,19
9	7,49	7,68	15,17	,19
10	7,44	7,66	15,10	,22
11	7,95	8,11	16,06	,16
12	7,47	7,66	15,13	,19
13	7,04	7,20	14,24	,16
14	7,10	7,25	14,35	,15
15	7,64	7,79	15,43	,15
ΣY	111,92	114,92	226,84	3,00
ΣY^2	836,3300	881,8304	3435,6992	0,6216

Fuente: Datos de un amplio estudio de Newman y Meredith (1956).

Análisis de la varianza de clasificación doble sin réplica.

Tabla de análisis de la varianza

Fuente de la variación	*g.l.*	*S.C.*	*M.C.*	F_s	*M.C. esperada*
Edades (columnas factor A)	1	0,3000	0,3000	389,11**	$\sigma^2 + \sigma^2_{AB} + \frac{b}{a-1}\Sigma\alpha^2$
Individuos (filas, factor B)	14	2,6367	0,1883	242,02**	$\sigma^2 + a\sigma^2_B$
Residuo	14	0,0108	0,000771		$\sigma^2 + \sigma^2_{AB}$
Total	29	2,9475			

$F_{0,01[1,14]} = 8,86$ $\quad$ $F_{0,01[12,14]} = 3,80$

Conclusiones. – La razón de varianza para edades es altamente significativa. Concluimos que los perímetros cefálicos de las niñas de 6 años son mayores que los correspondientes de las de 5 años. Si estamos dispuestos a suponer que la interacción σ^2_{AB} es cero, podemos buscar un componente aditivo de la varianza entre individuos y lo encontraríamos significativo.

CUADRO 9.3 (continuación)

Prueba t *para comparaciones apareadas*

$$t_s = \frac{\bar{D} - (\mu_1 - \mu_2)}{s_{\bar{D}}}$$

donde $\bar{D}$ es la diferencia media entre las observaciones apareadas

$(\bar{D} = \sum D/b = 3{,}00/15 = 0{,}20)$

y $s_{\bar{D}} = s_D/\sqrt{b}$ es el error standard de $\bar{D}$ calculado a partir de las diferencias observadas en la columna (4).

$$s_D = \sqrt{\frac{\sum D^2 - (\sum D)^2/b}{b-1}} = \sqrt{\frac{0{,}6216 - (3{,}00^2/15)}{14}} = \sqrt{0{,}0216/14}$$

$$= \sqrt{0{,}001543} = 0{,}0392810$$

$$s_{\bar{D}} = \frac{s}{\sqrt{b}} = \frac{0{,}0392810}{\sqrt{15}} = 0{,}0101423$$

Suponemos que la verdadera diferencia entre las medias de los dos grupos, $\mu_1 - \mu_2$, es cero.

$$t_s = \frac{\bar{D} - 0}{s_{\bar{D}}} = \frac{0{,}20 - 0}{0{,}0101423} = 19{,}7194 \qquad \text{por } b - 1 = 14 \text{ } g.l., \text{ y } P \ll 0{,}001.$$

$$t_s^2 = 388{,}85$$

lo que concuerda con el F_s anterior dentro de un aceptable error de redondeo.

podría darse a grupos de gemelos o parientes una droga o una prueba psicológica, recibiendo el tratamiento uno de cada pareja y el otro no.

Finalmente, la técnica de comparaciones apareadas puede utilizarse cuando los individuos a comparar forman parte de una sola unidad experimental y están por tanto sometidos a experiencias ambientales comunes. Si tenemos un grupo de jaulas de ratas, cada una de las cuales contiene dos ratas, y tratamos de comparar el efecto de una inyección hormonal con un control, podríamos inyectar una de cada pareja de ratas con la hormona y utilizar su compañera de jaula como control. Esto daría lugar a un análisis de la varianza $2 \times n$ para n jaulas.

Una razón para destacar por separado la prueba de comparaciones apareadas es que solamente entre los análisis de la varianza de clasificación doble sin réplica, tiene un método de análisis alternativo y equivalente, la prueba t para comparaciones apareadas, que es el método tradicional de resolverlo.

El caso de comparaciones apareadas presentado en el cuadro 9.3 analiza los perímetros cefálicos de niñas de cinco y seis años como ya se ha dicho. La cuestión a responder es si las frentes de las niñas de seis años son significativamente más anchas que las de niñas de cinco años. Los datos se exponen en las columnas (1) y (2) del cuadro 9.3 para 15 niñas.

La columna (3) destaca las sumas de fila que son necesarias para el análisis de la varianza. Los cálculos para el análisis de la varianza de clasificación doble sin réplica del cuadro 9.3, son los mismos que los ya presentados en el cuadro 9.2 y no se exponen con detalle. La tabla de análisis de la varianza muestra que hay una diferencia altamente significativa en los perímetros cefálicos entre los dos grupos de edad. Si se supone que la interacción es cero, hay un considerable componente aditivo de la varianza entre las niñas, que representa indudablemente diferencias tanto genéticas como ambientales.

El otro método para analizar modelos de comparaciones apareadas es la bien conocida *prueba t para comparaciones apareadas*. Es muy sencilla de aplicar y se ilustra en la segunda mitad del cuadro 9.3. Comprueba si la media de las diferencias de muestreo entre pares de lecturas en las dos columnas es significativamente diferente de una media hipotética, que la hipótesis nula sitúa en cero. El error estándar con el cual se contrasta ésta es el error estándar de la diferencia media. Tiene que calcularse la columna de diferencias, y se presenta en la columna (4) de la tabla de datos del cuadro 9.3. Los cálculos son muy sencillos y las conclusiones son las mismas que para el análisis de la varianza de clasificación doble. Este es otro ejemplo en el cual obtenemos el valor de F_s dentro del error de redondeo cuando elevamos al cuadrado el valor de t_s.

Aunque la prueba t de comparaciones apareadas es el método tradicional para resolver este tipo de problema, preferimos el análisis de la varianza de clasificación doble. Su cálculo no es más pesado, evita hallar una raíz cuadrada, y tiene la ventaja de proporcionar una media del componente de la varianza entre filas (bloques). Este conocimiento es útil porque si no hay componente aditivo significativo de la varianza entre bloques, se podría simplificar el análisis y diseño de estudios posteriores similares, utilizando un análisis de la varianza completamente aleatorizado.

Ejercicios 9

9.1 Swanson, Latshaw, y Tague (1921) determinaron electrométricamente el pH del suelo para varias muestras de suelo de Kansas. Más abajo se presenta un extracto de sus datos (suelos ácidos). ¿Difieren en pH los subsuelos de los suelos de la superficie? SOLUCION. $F_s = 0{,}894$.

Condado	*Tipo de suelo*	*pH en la superficie*	*pH en el subsuelo*
Finney	Rico en sedimentos de arcilla	6,57	8,34
Montgomery	Sedimento arcilloso en la cumbre	6,77	6,13
Doniphan	Sedimento de arcilla morena	6,53	6,32
Jewell	Sedimento arcilloso con piedras preciosas	6,71	8,30
Jewell	Sedimento arcilloso de Colby	6,72	8,44
Shawnee	Sedimento arcilloso de Crawford	6,01	6,80
Cherokee	Sedimento arcilloso de Oswego	4,99	4,42
Greenwood	Cumbre de sedimento arcilloso	5,49	7,90
Montgomery	Sedimento arcilloso Cherokee	5,56	5,20
Montgomery	Sedimento arcilloso Oswego	5,32	5,32
Cherokee	Sedimento arcilloso Bates	5,92	5,21
Cherokee	Sedimento arcilloso Cherokee	6,55	5,66
Cherokee	Sedimento arcilloso Neosho	6,53	5,66

9.2 Los datos siguientes fueron extraídos de un libro de registro de ganado vacuno de pura raza canadiense. De cada una de cinco razas (lista de honor, clase 305 días) se tomaron muestras al azar de 10 vacas maduras (cinco años y mayores) y diez vacas de dos años. Se registraron los porcentajes medios de grasa de la mantequilla de estas vacas. Esto nos dio un total de 100 porcentajes, descompuestos en cinco razas y en dos clases de edad. Los 100 porcentajes de grasa se dan a continuación. Analícese y discútanse los resultados. Se observará que la parte más pesada del cálculo se da hecha.

	Razas									
	Ayshire		*Canadian*		*Guernsey*		*Holstein-Friesian*		*Jersey*	
	Madura	*2 años*	*Madura*	*2 años*	*Madura*	*2 años*	*Madura*	*2 años*	*Madura*	*2 años*
	3,74	4,44	3,92	4,29	4,54	5,30	3,40	3,79	4,80	5,75
	4,01	4,37	4,95	5,24	5,18	4,50	3,55	3,66	6,45	5,14
	3,77	4,25	4,47	4,43	5,75	4,59	3,83	3,58	5,18	5,25
	3,78	3,71	4,28	4,00	5,04	5,04	3,95	3,38	4,49	4,76
	4,10	4,08	4,07	4,62	4,64	4,83	4,43	3,71	5,24	5,18
	4,06	3,90	4,10	4,29	4,79	4,55	3,70	3,94	5,70	4,22
	4,27	4,41	4,38	4,85	4,72	4,97	3,30	3,59	5,41	5,98
	3,94	4,11	3,98	4,66	3,88	5,38	3,93	3,55	4,77	4,85
	4,11	4,37	4,46	4,40	5,28	5,39	3,58	3,55	5,18	6,55
	4,25	3,53	5,05	4,33	4,66	5,97	3,54	3,43	5,23	5,72
$\sum Y$	40,03	41,17	43,66	45,11	48,48	50,52	37,21	36,18	52,45	53,40
$\bar{Y}$	4,003	4,117	4,366	4,511	4,848	5,052	3,721	3,618	5,245	5,340

$$\sum_{abn} Y^2 = 2059{,}6109$$

9.3 Blakeslee (1921) estudió las razones longitud/anchura de las hojas de plantas de semillero de segunda clase de dos tipos de hierbas llamadas globo (G) y nominal (N). Tres semillas de cada tipo se plantaron en 16 macetas. ¿Hay evidencia suficiente para concluir que globo y nominal difieren en razón longitud/anchura?

Número de identificación de maceta	*Tipos*					
	G			*N*		
16533	1,67	1,53	1,61	2,18	2,23	2,32
16534	1,68	1,70	1,49	2,00	2,12	2,18
16550	1,38	1,76	1,52	2,41	2,11	2,60
16668	1,66	1,48	1,69	1,93	2,00	2,00
16767	1,38	1,61	1,64	2,32	2,23	1,90
16768	1,70	1,71	1,71	2,48	2,11	2,00
16770	1,58	1,59	1,38	2,00	2,18	2,16
16771	1,49	1,52	1,68	1,94	2,13	2,29
16773	1,48	1,44	1,58	1,93	1,95	2,10
16775	1,28	1,45	1,50	1,77	2,03	2,08
16776	1,55	1,45	1,44	2,06	1,85	1,92
16777	1,29	1,57	1,44	2,00	1,94	1,80
16780	1,36	1,22	1,41	1,87	1,87	2,26
16781	1,47	1,43	1,61	2,24	2,00	2,23
16787	1,52	1,56	1,56	1,79	2,08	1,89
16789	1,37	1,38	1,40	1,85	2,10	2,00

9.4 Los datos siguientes se han extraído de un estudio más amplio de Sokal y Harten (1964). Los datos representan pesos secos medios (en mg) de tres genotipos de escarabajos, *Tribolium castaneum*, criados a una densidad de 20 escarabajos por gramo de harina. Las cuatro series de experimentos representan réplicas.

	Genotipos		
Series	$++$	$+b$	bb
1	0,958	0,986	0,925
2	0,971	1,051	0,952
3	0,927	0,891	0,829
4	0,971	1,010	0,955

Comprobar si los genotipos difieren en peso seco medio.

Capítulo 10

Supuestos teóricos del análisis de la varianza

Examinaremos ahora los supuestos básicos del análisis de la varianza, los métodos para comprobar si éstos son válidos, las consecuencias para un análisis de la varianza en caso de violación de estos supuestos y los pasos a seguir si no pueden reunirse los requisitos. Deberíamos hacer hincapié en que antes de realizar un análisis de la varianza en un problema de investigación real, deberíamos asegurarnos de que los supuestos enumerados en este capítulo parecen razonables y si no lo son, deberíamos tomar una de las diversas medidas alternativas posibles para remediar la situación.

La primera sección (10.1) enumera brevemente las diversas suposiciones del análisis de la varianza, describe procedimientos para comprobar algunas de ellas, expone brevemente las consecuencias en caso de que no se cumplan estos requisitos y (en este caso) da instrucciones acerca de cómo proceder. Los supuestos incluyen muestreo al azar, independencia, homogeneidad de varianzas, normalidad y aditividad.

En muchos casos la desviación de los supuestos del análisis de la varianza puede rectificarse por transformación de los datos originales utilizando una nueva escala. En la sección 10.2 se da el fundamento de esto y algunas de las transformaciones comunes.

Cuando las transformaciones son incapaces de conformar los datos a los supuestos del análisis de la varianza, deben utilizarse otras técnicas analíticas, análogas al propuesto análisis de la varianza. Estas son técnicas no paramétricas mencionadas en la sección 2.6. Como se expuso en ella, estas técnicas se utilizan a veces por preferencia aun cuando el método paramétrico (análisis de la varianza en este caso) pueda emplearse legítimamente. La rapidez del cálculo y una preferencia de los supuestos generalmente sencillos de los análisis no paramétricos, hace que muchos investigadores recurran a ellas. No obstante, cuando se reúnen los requisitos del análisis de la varianza, estos métodos son menos eficientes que él. La sección 10.3 examina tres métodos no paramétricos capaces de sustituir el análisis de la varianza para casos de dos muestras solamente.

10.1 Los supuestos teóricos del análisis de la varianza

Aleatoriedad. Todos los análisis de la varianza requieren que el muestreo de individuos sea al azar. Así, en un estudio de los efectos de tres dosis de una droga (más un control) en cinco ratas cada una, las cinco ratas asignadas a cada tratamiento deben seleccionarse al azar. Si las cinco ratas utilizadas como controles son las más jóvenes o las más pequeñas o las más pesadas, mientras las asignadas a algún otro tratamiento se seleccionan de alguna otra manera, está claro que los resultados no son aptos para dar una estimación no sesgada de los verdaderos efectos del tratamiento. La no aleatoriedad de la selección de la muestra puede reflejarse en la falta de independencia de los ítems, en la heterogeneidad de las varianzas, o en la no normalidad de la distribución, discutido todo en esta sección. Son esenciales las precauciones suficientes para asegurar el muestreo al azar durante el diseño de un experimento o al muestrear poblaciones naturales.

Independencia. Un supuesto establecido en cada expresión explícita para el valor esperado de una variante [por ejemplo, la expresión (7.3) era $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$] es que el término error ϵ_{ij} es una variable aleatoria normal. Además, para completarse debería afirmar también que se supone que los valores de ϵ se distribuyen de forma independiente e idéntica (véase más adelante).

Así, si se ordenan las variantes de un grupo cualquiera en algún orden lógico independiente de su magnitud (tal como el orden en que se han obtenido las medidas) se podría esperar que los valores de ϵ_{ij} se sucediesen uno a otro en una secuencia al azar. Por consiguiente, consideramos muy improbable una larga secuencia de valores positivos seguida por una secuencia igualmente larga de valores negativos. Tampoco esperaríamos que alternasen con regularidad valores positivos y negativos.

¿Cómo podrían producirse las desviaciones de esta independencia? Un ejemplo sencillo sería un experimento en el que las unidades experimentales fuesen parcelas de terreno trazadas en un campo. En este caso se encuentra con frecuencia que las parcelas adyacentes dan rendimientos muy similares. Así pues, sería importante no agrupar todas las parcelas que contienen el mismo tratamiento en una serie de parcelas adyacentes, sino más bien aleatorizar la asignación de los tratamientos entre las parcelas experimentales. El proceso físico de asignación de los tratamientos al azar a las parcelas experimentales asegura que los valores de ϵ serán independientes.

La falta de independencia de los ϵ puede resultar de correlación en tiempo más que en espacio. En un experimento podríamos medir el efecto de un tratamiento registrando los pesos de diez individuos. Nuestra estimación puede quedar perjudicada por efecto de un ajuste defectuoso que resultaría al dar subestimaciones sucesivas compensadas por diversas sobreestimaciones. A la inversa, la compensación del equilibrio por el operador puede conducir a sobre y subestimaciones del verdadero peso, regularmente alternantes. De nuevo en este caso la aleatorización puede superar el problema de la no independencia de los errores. Por ejemplo, podemos determinar la secuencia en que se pesan individuos de los diversos grupos según algún procedimiento aleatorio.

No hay ninguna simple adaptación ni transformación para superar la falta de independencia de los errores. Debe cambiarse el diseño básico del experimento o la forma en que se ha realizado. Si los ϵ no son independientes, la validez de la prueba F de significación habitual puede resultar gravemente perjudicada.

Homogeneidad de varianzas. En la sección 8.4 y en el cuadro 8.2, en los que hemos descrito la prueba t para la diferencia entre dos medias, se ha dicho que las pruebas estadísticas solamente son válidas si podemos suponer que las varianzas de las dos muestras son iguales. Aunque hasta ahora no hemos hecho hincapié en ello, esta suposición de que los ϵ_{ij} tienen idénticas varianzas también es el fundamento de la prueba equivalente de análisis de la varianza para dos muestras, y en realidad de cualquier tipo de análisis de la varianza. La *igualdad de varianzas* en un grupo de muestras es un prerrequisito importante para varias pruebas estadísticas. Sinónimos de esta condición son *homogeneidad de varianzas* u *homoscedasticidad*. Este término, etimológicamente griego, significa igual dispersión; la condición contraria (desigualdad de varianzas entre muestras) se denomina *heteroscedasticidad*. Como suponemos que cada varianza de muestreo es una estimación de la misma varianza paramétrica, el supuesto de homogeneidad de las varianzas tiene sentido intuitivo.

Ya hemos visto cómo comprobar si dos muestras son homoscedásticas antes de realizar una prueba t de las diferencias entre dos medias o un análisis de la varianza para dos muestras: utilizamos una prueba F para las hipótesis $H_0 : \sigma_1^2$ y $H_1 : \sigma_1^2 \neq \sigma_2^2$, como se ha demostrado en la sección 7.3 y en el cuadro 7.1. Para más de dos muestras hay un método "rápido y malo", preferido por muchos debido a su simplicidad. Se trata de la prueba $F_{máx}$. Esta prueba cuenta con la distribución de probabilidad acumulativa de un estadístico, que es la razón de varianzas de la mayor a la menor de diferentes varianzas de muestreo. Esta distribución se presenta en la tabla **VI**. Vamos a suponer que tenemos seis muestras antropológicas de 10 longitudes de hueso cada una, para las cuales queremos hacer un análisis de la varianza. Las varianzas de las seis muestras varían de 1,2 a 10,8. Calculamos la máxima razón de varianzas $s^2_{máx}/s^2_{mín} = 10{,}8/1{,}2 = 9{,}0$ y la comparamos con $F_{máx\,\alpha[a,\nu]}$, cuyos valores críticos se encuentran en la tabla **VI**. Para $a = 6$ y $\nu = n - 1$, $F_{máx}$ es 7,80 y 12,1 a los niveles de significación del 5 % y 1 %, respectivamente. Concluimos que las varianzas de las seis muestras son significativamente heterogéneas.

¿Qué puede causar tal heterogeneidad? En este caso, sospechamos que algunas de las poblaciones son inherentemente más variables que otras. Ciertas razas o especies son relativamente uniformes para un carácter mientras que otras son muy variables para el mismo carácter. En un análisis de la varianza que represente los resultados de un experimento, es muy posible que una muestra se haya obtenido bajo condiciones menos estandarizadas que las otras y por consiguiente tenga una varianza superior. Hay también muchos casos en que la heterogeneidad de varianzas es función de una elección incorrecta de la escala de medida, las varianzas varían como funciones de las medias. Así pues las diferencias entre las medias originan varianzas heterogéneas. Por ejemplo, en las variables que siguen la distribución de Poisson, la varianza es en realidad igual a la media, y las poblaciones con medias mayores tendrán por lo tanto varianzas mayores. Estas excepciones del supuesto de homoscedasticidad pueden a veces corregirse fácilmente por una transformación adecuada, como se discute posteriormente en este capítulo. Un primer examen rápido con respecto a la heteroscedasticidad es comprobar la correlación entre las medias y las varianzas o entre las medias y los rangos de las muestras. Si las varianzas aumentan con las medias (como en una distribución de Poisson), las razones $s^2/\bar{Y}$ ó $s/\bar{Y} = C.V.$ serán aproximadamente constantes para las muestras. Si las medias y las varianzas son independientes, estas razones variarán ampliamente.

Las consecuencias de una moderada heterogeneidad de varianzas no son demasiado graves para la prueba de significación global, pero las simples comparaciones de grados de libertad pueden estar lejos de la exactitud.

Si la transformación no puede hacer frente a la heteroscedasticidad, puede ser necesario recurrir a los métodos no paramétricos (sección 10.3).

Normalidad. Hemos supuesto que los errores ϵ_{ij} de las variantes en cada muestra serán independientes, que las varianzas de los errores de las diversas muestras serán iguales, y finalmente, que los errores siguen la distribución normal. Si hay verdadera duda acerca de la normalidad de los datos, podría aplicarse a cada muestra por separado una prueba gráfica como se ha demostrado en la sección 5.5.

Las consecuencias de la no normalidad del error no son demasiado graves. Solamente una distribución muy sesgada tendría un marcado efecto sobre el nivel de significación de la prueba F o sobre la eficacia del diseño. El mejor modo de corregir la falta de normalidad es hacer una transformación que conduzca a una distribución normal de los datos, como se explica en la próxima sección. Si ninguna transformación sencilla es satisfactoria, el análisis de la varianza debería sustituirse por una prueba no paramétrica, como la realizada en la sección 10.3.

Aditividad. En un análisis de la varianza de clasificación doble sin replicación, es necesario suponer que no existe interacción si se van a hacer pruebas de los efectos principales, en un análisis de la varianza modelo I. Este supuesto de no interacción en un análisis de la varianza de clasificación doble, también se menciona a veces como el supuesto de aditividad de los efectos principales. Con esto queremos decir que cualquier variante individual observada puede descomponerse en componentes aditivos representando tanto los efectos de tratamiento de una fila y columna particular como un término aleatorio especial para ella. Si la interacción está realmente presente, entonces la prueba F será muy ineficaz y posiblemente induzca a error si el efecto de la interacción es muy grande. Una verificación de este supuesto requiere o bien más de una sola observación por casilla (para que pueda calcularse una media cuadrática del error) o bien una estimación independiente de la media cuadrática del error a partir de experimentos *comparables* previos.

Las interacciones pueden deberse a varias causas. La mayor parte de las veces significa que una determinada combinación de tratamientos, tal como nivel 2 de factor A al combinarlo con nivel 3 de factor B, hace que una variable se desvíe del valor esperado. Esta desviación se considera como una propiedad inherente del sistema natural en estudio, como en los ejemplos de sinergismo o interferencia. Se encuentran efectos similares cuando una réplica determinada es muy aberrante, como puede ocurrir si se incluye una parcela excepcional en un experimento agrícola, si un individuo enfermo se incluye en un experimento fisiológico, o si por error se incluye un individuo de diferente especie en un estudio biométrico. Finalmente, resultará una interacción si el efecto de los dos factores A y B en la variable respuesta Y son multiplicativos en lugar de aditivos. Esto se clarificará por medio de un ejemplo.

En la tabla 10.1 representamos los efectos de tratamiento aditivos y multiplicativos en un hipotético análisis de la varianza de clasificación doble. Supongamos que la media paramétrica esperada μ es cero. En tal caso la media de la muestra sometida al tratamiento 1 de factor A y al tratamiento 1 de factor B debería ser 2 según el modelo aditivo convencional. Esto es así porque cada factor al nivel 1 aporta una unidad a la media.

TABLA 10.1

Ilustración de efectos aditivos y multiplicativos.

Factor B	*Factor A* $\alpha_1 = 1$	$\alpha_2 = 2$	$\alpha_3 = 3$	
	2	3	4	Efectos aditivos
	1	2	3	Efectos multiplicativos
$\beta_1 = 1$	0	0,30	0,48	Log. de los efectos multipl.
	6	7	8	Efectos aditivos
	5	10	15	Efectos multiplicativos
$\beta_2 = 5$	0,70	1,00	1,18	Log. de los efectos multipl.

Igualmente, la media de subgrupo esperada sometida al nivel 3 para el factor A y al nivel 2 para el factor B es 8, siendo las respectivas aportaciones a la media 3 y 5. Sin embargo, si el proceso es multiplicativo en vez de aditivo, como ocurre en una variedad de fenómenos fisicoquímicos y biológicos, los valores esperados serían muy diferentes. Para el tratamiento A_1B_1, el valor esperado es igual a 1, que es el producto de 1 y 1. Para el tratamiento A_3B_2, el valor esperado es 15, el producto de 3 y 5. Si analizásemos datos multiplicativos de este tipo por un análisis de la varianza convencional, hallaríamos que la suma de cuadrados de la interacción estaría enormemente aumentada debido a la no aditividad de los efectos del tratamiento. En este caso, hay un remedio sencillo. Transformando la variable en logaritmos (tabla 10.1), podemos restablecer la aditividad de los datos. El tercer ítem de cada casilla da el logaritmo del valor esperado, suponiendo relaciones multiplicativas. Nótese que los incrementos son de nuevo estrictamente aditivos ($S.C._{A\times B} = 0$). En realidad, en una escala logarítmica podríamos escribir simplemente $\alpha_1 = 0, \alpha_2 = 0{,}30, \alpha_3 = 0{,}48, \beta_1 = 0, \beta_2 = 0{,}70$. Esta es una buena demostración de cómo la transformación de escala, discutida con detalle en la sección 10.2, nos ayuda a satisfacer los requisitos del análisis de la varianza.

10.2 Transformaciones

Si la evidencia indica que no pueden mantenerse los requisitos para un análisis de la varianza o una prueba t, se nos presentan dos posibilidades. Podemos realizar una prueba diferente que no requiera los supuestos rechazados, tal como las pruebas de distribución libre, en lugar del análisis de la varianza, discutidas en la próxima sección. Un segundo enfoque sería transformar la variable a analizar de tal manera que las variantes transformadas que resultan reúnan los requisitos del análisis. Vamos a considerar un ejemplo sencillo de lo que hará la transformación. Una variante del tipo más sencillo de análisis de la varianza (completamente aleatorizado, de clasificación simple, modelo I) se descompone de la forma siguiente: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$. En este modelo los componentes son aditivos con el error ϵ_{ij} normalmente distribuido. Sin embargo, podríamos encontrar una situación

en la cual los componentes fuesen multiplicativos de modo que $Y_{ij} = \mu\alpha_i\epsilon_{ij}$, el producto de estos tres términos. En este caso fallan los supuestos de normalidad y de homoscedasticidad. La media paramétrica general μ es constante en cualquier análisis de la varianza, pero el efecto de tratamiento α_i difiere de un grupo a otro. Evidentemente, la dispersión entre las variantes Y_{ij} se duplicaría en un grupo en que α_i fuese el doble que en otro. Supongamos que $\mu = 1$, el menor $\epsilon_{ij} = 1$, y el mayor 3; entonces si $\alpha_i = 1$, el rango de los valores de Y será $3 - 1 = 2$. Sin embargo, cuando $\alpha_i = 4$, el rango correspondiente será cuatro veces más amplio, de $4 \times 1 = 4$ a $4 \times 3 = 12$, un rango de 8. Estos datos serán heteroscedásticos. Podemos corregir esta situación fácilmente transformando nuestro modelo en logaritmos. Obtendríamos pues $\log Y_{ij} = \log \mu + \log \alpha_i + \log \epsilon_{ij}$, que es aditivo y homoscedástico. El análisis de la varianza completo se realizaría entonces con las variantes transformadas.

En este punto muchos se sentirán más o menos incómodos por lo que hemos hecho. La transformación se parece mucho a una "pulidora de datos". Cuando se descubre que a veces una prueba estadística puede hacerse significativa tras la transformación de una serie de datos, aun cuando no lo habría sido sin esta transformación, nos podemos sentir aún más recelosos. ¿Qué es lo que justifica la transformación de los datos? Es necesario acostumbrarse a la idea, pero realmente no hay necesidad científica de emplear la escala lineal o aritmética común a la que estamos acostumbrados. La enseñanza de la "matemática moderna" en escuelas elementales ha hecho mucho por disipar la creencia ingenua de que el sistema de numeración decimal es el único "natural". Se necesita gran experiencia en ciencias y en el tratamiento de datos estadísticos para apreciar el hecho de que la escala lineal, tan familiar a todos nosotros desde nuestra más remota experiencia, ocupa con relación a otras escalas de medida una posición similar a la del sistema decimal de numeración con respecto a los sistemas de numeración binario u octal y otros. Si un sistema es multiplicativo en una escala lineal, puede resultar mucho más conveniente considerarlo como un sistema aditivo en una escala logarítmica. La raíz cuadrada de una variable es otra transformación frecuente. La raíz cuadrada del área superficial de un organismo es a veces una medida más apropiada de la variable biológica fundamental sometida a fuerzas fisiológicas y evolutivas que el área. Esto se refleja en la distribución normal de la raíz cuadrada de la variable cuando se compara con la distribución sesgada de las áreas. En muchos casos la experiencia nos ha enseñado a expresar variables experimentales, no en escala lineal, sino más bien como logaritmos, raíces cuadradas, recíprocos o ángulos. Así, los valores de pH son logaritmos y las series de dilución en titulaciones microbiológicas se expresan como recíprocos. Tan pronto como se esté dispuesto a aceptar la idea de que la escala de medida es arbitraria, simplemente se tendrá que mirar la distribución de las variantes transformadas para decidir qué transformación satisface más rigurosamente los requisitos del análisis de la varianza antes de realizarlo.

Un dato afortunado acerca de las transformaciones es que muy frecuentemente varias excepciones de los supuestos del análisis de la varianza se reparan simultáneamente por la misma transformación a una nueva escala. Así, frecuentemente basta hacer los datos homoscedásticos para aproximarlos también a la normalidad y asegurar la aditividad de los efectos del tratamiento.

Cuando se aplica una transformación, las pruebas de significación se realizan con los datos transformados, pero las estimaciones de las medias se dan ordinariamente en la

escala familiar no transformada. Como las transformaciones discutidas en este capítulo son no lineales, los límites de confianza calculados en la escala transformada y vueltos a cambiar, a la escala original serían asimétricos. Sería pues erróneo expresar el error standard en la escala original. El presentar los resultados de la investigación con variables que requieren transformación, proporciona medias en la escala nuevamente transformada seguidas por sus límites de confianza (asimétricos) en vez de por sus errores standard.

Una manera sencilla de hallar si una determinada transformación dará una distribución que cumpla los requisitos del análisis de la varianza es representar las distribuciones acumulativas de las diversas muestras en papel probabilístico. Transformando la escala del segundo eje de coordenadas de lineal a logarítmica, raíz cuadrada u otra cualquiera, podemos ver si una línea previamente curva, indicando oblicuidad, se hace recta para indicar normalidad (es conveniente refrescar la memoria sobre estas técnicas gráficas estudiadas en la sección 5.5). Podemos buscar límites de clase superiores en escalas transformadas o emplear una variedad de papeles cuyo segundo eje está en escala logarítmica, angular u otra. De este modo, no solamente comprobamos si los datos se normalizan por medio de la transformación, sino que además podemos dar una estimación de la desviación típica transformada que se mide por la pendiente de la línea ajustada. El supuesto de homoscedasticidad implica que las pendientes deberían ser idénticas para las diversas muestras. Si las pendientes son muy heterogéneas, no se ha conseguido la homoscedasticidad.

La transformación logarítmica. La más común de las transformaciones aplicadas es la conversión de todas las variantes en logaritmos, ordinariamente logaritmos vulgares. Siempre que la media esté positivamente correlacionada con la varianza (medias mayores acompañadas de varianzas mayores), es muy probable que la transformación logarítmica remedie la situación y haga que la varianza sea independiente de la media. Las distribuciones de frecuencias inclinadas hacia la derecha se hacen a veces más simétricas por transformación a la escala logarítmica. En la sección anterior y en la tabla 10.1 hemos visto que la transformación logarítmica es necesaria también cuando los efectos son multiplicativos.

La transformación raíz cuadrada. Utilizaremos una transformación raíz cuadrada como ejemplo detallado de transformación de escala. Cuando los datos son recuentos, como los de insectos en una hoja o de células sanguíneas en un hemacitómetro, frecuentemente encontramos de utilidad la transformación raíz cuadrada. Se recordará que probablemente estas distribuciones sean de Poisson en vez de normales y que en una distribución de Poisson la varianza es igual a la media. Por consiguiente, la media y la varianza no pueden ser independientes sino que variarán idénticamente. Al transformar las variantes en raíces cuadradas, las varianzas se harán generalmente independientes de las medias. Cuando los recuentos incluyen valores de cero, se ha visto conveniente codificar todas las variantes sumándoles 0,5. La transformación en este caso es $\sqrt{Y + \frac{1}{2}}$.

La tabla 10.2 muestra una aplicación de la transformación raíz cuadrada. La muestra de media superior tiene una varianza significativamente mayor antes de la transformación. Después de la transformación las varianzas no son significativamente diferentes. Para notificación de medias, las medias transformadas se elevan al cuadrado nuevamente y se dan límites de confianza en lugar de errores standard.

La transformación arco seno. Esta transformación, conocida también como la *transformación angular*, es especialmente apropiada para porcentajes y proporciones. De la sec-

TABLA 10.2

Aplicación de la transformación raíz cuadrada. Los datos representan el número de adultos de *Drosophila* que salen de cultivos de una sola pareja para dos medios de diferente formulación (el medio A contiene DDT).

(1)	(2)	(3)	(4)
Número de moscas que salen Y	*Raíz cuadrada del número de moscas* $\sqrt{Y}$	*Medio A* f	*Medio B* f
0	0,00	1	—
1	1,00	5	—
2	1,41	6	—
3	1,73	—	—
4	2,00	3	—
5	2,24	—	—
6	2,45	—	—
7	2,65	—	2
8	2,83	—	1
9	3,00	—	2
10	3,16	—	3
11	3,32	—	1
12	3,46	—	1
13	3,61	—	1
14	3,74	—	1
15	3,87	—	1
16	4,00	—	2
		15	15
Variable sin transformar			
$\overline{Y}$		1,933	11,133
s^2		1,495	9,410
Transformación raíz cuadrada			
$\overline{\sqrt{Y}}$		1,297	3,307
$s^2_{\sqrt{Y}}$		0,2630	0,2089
Pruebas de igualdad de varianzas			
$F_s = \frac{s^2_2}{s^2_1} = \frac{9,410}{1,495} = 6,294*$		$F_{.025[14,14]} = 2,98$	$F_s = \frac{s^2_{\sqrt{Y_2}}}{s^2_{\sqrt{Y_1}}} = \frac{0,2630}{0,2089} = 1,259\ ns$
Medias transformadas de nuevo (al cuadrado)			
$(\overline{\sqrt{Y}})^2$		1,682	10,936
límites de confianza del 95 %			
$L_1 = \overline{\sqrt{Y}} - t_{.05\,\nu\,\overline{\sqrt{Y}}}$		$1,297 - 2,145\sqrt{\frac{0,2630}{15}}$ $= 1,013$	$3,307 - 2,145\sqrt{\frac{0,2089}{15}}$ $= 3,054$
$L_2 = \overline{\sqrt{Y}} + t_{.05\,\nu\,\overline{\sqrt{Y}}}$		1,581	3,560
Límites de confianza (al cuadrado) transformados nuevamente			
L_1^2		1,026	9,327
L_2^2		2,500	12,674

Fuente: Datos de un estudio no publicado de R.R. Sokal.

ción 4.2 se puede recordar que la desviación típica de una distribución binominal es $\sigma = \sqrt{pq/k}$. Como $\mu = p$, $q = 1 - p$, y k es constante para cualquier problema, está claro que en una distribución binominal la varianza sería una función de la media. La transformación arco seno evita esto.

La transformación halla $\theta = \text{arcsen } \sqrt{p}$, en donde p es una proporción. El término arco seno es sinónimo de seno inverso o seno^{-1}, que simboliza el ángulo cuyo seno es la cantidad dada. Así, si buscamos arcsen 0,431 encontramos 41,03°, el ángulo cuyo seno es 0,431. La transformación arco seno extiende las dos colas de una distribución de porcentajes y proporciones y reduce el centro. Cuando los porcentajes en los datos originales caen entre 30 % y 70 % generalmente no es necesario aplicar la transformación arco seno.

10.3 Métodos no paramétricos en lugar del análisis de la varianza

Si ninguna de las transformaciones anteriores consigue hacer que nuestros datos reúnan los requisitos del análisis de la varianza, podemos recurrir a un método no paramétrico análogo. Estas técnicas se denominan también *métodos de distribución-libre*, ya que no dependen de una distribución determinada (como la normal en el análisis de la varianza), sino que ordinariamente son eficaces para un amplio rango de distribuciones diferentes. Se denominan métodos no paramétricos porque su hipótesis nula no tiene nada que ver con parámetros específicos (tales como la media en el análisis de la varianza) sino solamente con la distribución de las variantes. En los últimos años, el análisis de la varianza no paramétrico se ha hecho muy popular porque es sencillo de calcular y libera de la preocupación por los supuestos de un análisis de la varianza. No obstante, deberíamos señalar que en los casos en que los requisitos se cumplen completa o incluso aproximadamente, el análisis de la varianza es generalmente la prueba estadística más eficiente para detectar desviaciones de la hipótesis nula.

En esta sección solamente discutiremos las pruebas no paramétricas para dos muestras. Para un modelo que origine una prueba t o análisis de la varianza con dos clases, empleamos la *prueba U* no paramétrica de *Mann-Whitney* (cuadro 10.1). La hipótesis nula es que las dos muestras proceden de poblaciones que tienen la misma distribución. Los datos del cuadro 10.1 son medidas morfológicas de dos muestras de ninfas de mariposas. Como se ha visto en el cuadro 10.1, la prueba U de Mann-Whitney es semigráfica y muy sencilla de aplicar. Será especialmente conveniente cuando los datos estén ya representados gráficamente y no haya demasiados ítems en cada muestra.

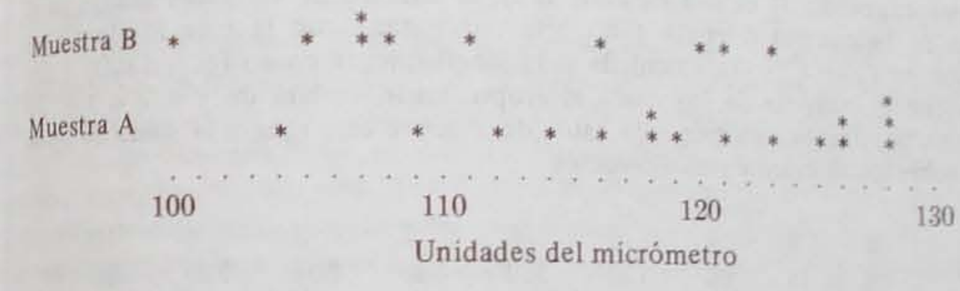
Nótese que este método realmente no requiere que cada observación individual represente una medida precisa. Siempre que se puedan ordenar los datos se pueden realizar estas pruebas. Así, por ejemplo, supóngase que se pone al descubierto un poco de carne y se estudian los tiempos de llegada de individuos de dos especies de moscardas. Se podría registrar exactamente el tiempo de llegada de cada moscarda, partiendo de un punto cero en tiempo cuando se ha expuesto la carne. Por otra parte, simplemente se podrían clasificar los tiempos de llegada de las dos especies, señalando que 1 individuo de la especie *B* llega primero, a continuación 2 individuos de la especie *A*, luego 3 individuos de *B*, seguidos por la llegada simultánea de uno de cada especie (un ligamiento), y así

CUADRO 10.1

Test *U* de Mann-Whitney para dos muestras, observaciones ordenadas, no apareadas.

Dos muestras de ninfas de la mariposa *Trombicula lipovskyi* (datos no publicados de D.A. Crossley). La variante medida es la longitud de la base del quelícero indicada en unidades del micrómetro. Los datos reales no se presentan. Como se demuestra en el paso 1 solamente se necesita una gráfica de las dos muestras. El tamaño de muestreo de la muestra más grande se designa por n_1 y el de la más pequeña por n_2. En este caso, $n_1 = 16$, $n_2 = 10$. Si las dos muestras son de igual tamaño, no importa cuál se designe por 1.

1. Representar gráficamente las dos muestras como se expone más abajo. Indicar los ligamientos poniendo asteriscos o cruces uno sobre el otro.



2. Para cada observación de una muestra (es conveniente utilizar la muestra más pequeña) contar el número de observaciones de la otra muestra que son de menor valor (a la izquierda). Contar ½ para cada observación ligada. Por ejemplo, hay cero observaciones en la muestra A inferiores a la primera observación de la muestra B; 1 observación menor que la segunda, tercera, cuarta y quinta observaciones de la muestra B; 2 observaciones de A inferiores a la sexta de B; 4 observaciones de A menores que la séptima de B, pero una es igual (ligada) a ella, en este caso contamos 4 ½. Continuando de manera similar, obtenemos recuentos de 8, 8 ½ y 9 ½. La suma de estos recuentos es $C = 36$ ½. El estadístico U_s de Mann-Whitney es la mayor de las dos cantidades C y $(n_1 n_2 - C)$, en este caso 36 ½ y $[(16 \times 10) - 36\ ½] = 123$ ½.

3. Comparamos el valor de $U_s = 123$ ½ con los valores críticos para $U_{\alpha[n_1,n_2]}$ de la tabla **XIII**. Se rechaza la hipótesis nula si el valor observado es demasiado grande. Como $U_{0,025[16,10]} = 118$ y $U_{0,01[16,10]} = 124$, las dos muestras son significativamente diferentes a $0,05 > P > 0,02$ (duplicamos las probabilidades porque esta prueba es de dos colas).

En los casos en que $n_1 > 20$, calculamos la siguiente cantidad

$$t_s = \frac{\left(U_s - \frac{n_1 n_2}{2}\right)}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

CUADRO 10.1 (continuación)

que sigue aproximadamente la distribución normal. El denominador 12 es una constante. Se busca la significación de t_s en la tabla **III** respecto a los valores críticos de $t_{\alpha[\infty]}$ para una prueba de una cola o de dos colas según requiera la hipótesis. Cuando aparecen valores ligados, la fórmula anterior se modifica como sigue:

$$t_s = \frac{\left(U_s - \frac{n_1 n_2}{2}\right)}{\sqrt{\left(\frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-1)}\right)\left(\frac{(n_1+n_2)^3 - (n_1+n_2) - \sum^m T_j}{12}\right)}}$$

En esta expresión T_j es una función de t_j, el número de variantes ligadas en el grupo de ligamiento j. (Esta t no está relacionada con la t de Student.) La función es $T_j = t_j^3 - t_j$, calculada más sencillamente como $(t_j - 1)t_j(t_j + 1)$. Como en la mayoría de los casos el grupo ligado variará de $t = 2$ a $t = 10$ ligamientos, damos una pequeña tabla de T sobre este rango; la suma de T_j se hace sobre los m ligamientos diferentes.

t_j	2	3	4	5	6	7	8	9	10
T_j	6	24	60	120	210	336	504	720	990

Por ejemplo, si tuviésemos que calcular $\sum^m T_j$ para el problema anterior (no necesario puesto que $n_1 < 20$), calcularíamos $t_1 = 2$ para las dos primeras variantes ligadas (107 unidades del micrómetro). Igualmente, construiríamos una tabla para todos los valores de t_j y T_j de este problema.

t_j	2	2	2	2	2	2	3
T_j	6	6	6	6	6	6	24

$\sum^m T_j = 6 + 6 + \ldots + 24 = 60$

sucesivamente. Si bien estos datos clasificados u ordenados no podrían ser analizados por los métodos paramétricos ya estudiados, las técnicas del cuadro 10.1 son completamente aplicables.

El método de cálculo del estadístico de muestreo U_s para las pruebas de Mann-Whitney y Wilcoxon es directo, como se demuestra en el cuadro 10.1. Los valores críticos para $U_{\alpha[n_1,n_2]}$ se exponen en la tabla **XIII**, que es adecuada para casos en los que el tamaño de muestreo mayor $n_1 \leqslant 20$. Las probabilidades de la tabla **XIII** suponen una prueba de una cola. Para una prueba de dos colas se debería multiplicar por dos el valor de la probabilidad representado en esa tabla. Cuando $n_1 > 20$, se calcula la expresión que aparece casi al final del cuadro 10.1. Puesto que esta expresión se distribuye como una desviante normal, consúltese la tabla de t (tabla **III**), para $t_{\alpha[\infty]}$ utilizando probabilidades

de una o dos colas dependiendo de la hipótesis. Una complicación adicional surge de observaciones ligadas, las cuales requieren la fórmula más detallada que se muestra al final del cuadro 10.1. No obstante, se necesita un número considerable de ligamientos para que afecten al resultado de la prueba apreciablemente. Las correcciones para ligamientos aumentan ligeramente el valor de t_s; por lo tanto la fórmula sin corregir es más conservativa.

Es conveniente adquirir una comprensión intuitiva del fundamento de esta prueba. En la prueba de Mann-Whitney podemos considerar dos situaciones extremas: en un caso las dos muestras se solapan y coinciden completamente; en el otro están bastante separadas. En el segundo caso, si tomamos la muestra con las variantes de menor valor, no habrá puntos de la muestra que se contrasta a su izquierda; es decir, podemos pasar cada observación de la muestra de valores más bajos sin que tenga ningún ítem de la de valores más altos a su izquierda. Por el contrario, si hubiésemos empezado por la segunda, todos los puntos de la muestra inferior estarían a la izquierda de cada punto de la superior. Nuestro recuento total sería por tanto el recuento total de una muestra multiplicado por el total de observaciones de la segunda muestra, lo cual da $n_1 n_2$. De este modo, puesto que se nos ha dicho que tomemos el mayor de los dos valores, la suma de los recuentos C o $n_1 n_2 - C$, nuestro resultado sería en este caso $n_1 n_2$. Por otra parte, si las dos muestras coinciden completamente, entonces para cada punto de una muestra tendríamos los puntos inferiores a él más medio punto para el valor ligado que representa esa observación en la segunda muestra,el cual está exactamente al mismo nivel que la observación que se considera. Una corta experimentación demostrará que este valor es $[n(n-1)/2] + (n/2) = n^2/2$. Naturalmente el rango de posibles valores de U debe estar entre éste y $n_1 n_2$, y el valor crítico debe estar en algún punto dentro de este rango.

Como resultado de las pruebas del cuadro 10.1 nuestra conclusión es que las dos muestras difieren significativamente en la distribución de la longitud de la base del quelícero. Está claro que las mariposas de la muestra A tienen las bases del quelícero más largas que las de la muestra B.

Finalmente presentaremos un método no paramétrico para el modelo de comparaciones apareadas, discutido en la sección 9.3 e ilustrado en el cuadro 9.3. El método más ampliamente utilizado es la *prueba de rangos con signo de Wilcoxon*, ilustrado en el cuadro 10.2. El ejemplo al que se aplica aún no se ha encontrado. Registra el tamaño medio de camada en dos razas de cobayas mantenidas en grandes colonias durante los años 1916 hasta 1924. Cada uno de estos valores es el promedio de un gran número de camadas. Obsérvese el paralelismo en los cambios de la variable en las dos razas. Durante 1917 y 1918 (años de guerra para los EE.UU.), una escasez de cuidados y alimentos condujo a un descenso en el número de crías por camada. En cuanto volvieron mejores condiciones, el tamaño medio de camada aumentó nuevamente. Obsérvese que en 1922 otra vez se refleja en ambas líneas una caída subsiguiente, sugiriendo que estas fluctuaciones son de causa ambiental. Es pues muy conveniente que los datos sean tratados como comparaciones apareadas, considerando los años como replicaciones y las diferencias de razas como los tratamientos fijos que se contrastan. La columna (3) del cuadro 10.2 presenta las diferencias con las cuales podría realizarse una prueba t de comparaciones apareadas convencional. Para la prueba de Wilcoxon, estas diferencias se ordenan *sin considerar el signo* de modo que la diferencia absoluta mínima se clasifica como 1, y la

CUADRO 10.2

Prueba del rango con signo de Wilcoxon para dos grupos, dispuestos como observaciones apareadas.

Tamaño medio de camada de dos razas de cobayas, comparadas durante $n = 9$ años.

	(1)	*(2)*	*(3)*	*(4)*
Año	*Raza B*	*Raza 13*	D	*Rango* (R)
1916	2,68	2,36	+0,32	+9
1917	2,60	2,41	+0,19	+8
1918	2,43	2,39	+0,04	+2
1919	2,90	2,85	+0,05	+3
1920	2,94	2,82	+0,12	+7
1921	2,70	2,73	−0,03	−1
1922	2,68	2,58	+0,10	+6
1923	2,98	2,89	+0,09	+5
1924	2,85	2,78	+0,07	+4
		Suma absoluta de rangos negativos		1
		Suma de rangos positivos		44

Fuente: Datos de S. Wright.

Método

1. Calcular las diferencias entre los n pares de observaciones. Estas se introducen en la columna (3), señalada como D.
2. Ordenar estas diferencias de menor a mayor *sin considerar el signo.*
3. Asignar a los rangos los signos originales de las diferencias.
4. Sumar los rangos positivos y negativos por separado. La suma que sea menor en valor absoluto, T_s, se compara con los valores de la tabla **XIV** para $n = 9$.

Como $T_s = 1$, que es igual o menor que la entrada en la tabla para una cola $\alpha = 0{,}005$, nuestra diferencia observada es significativamente diferente de la de la cepa 13.

Para muestras grandes ($n > 50$) se calcula

$$t_s = \frac{T_s - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+\frac{1}{2})(n+1)}{12}}}$$

donde T_s es como se definió en el punto 4 arriba. Comparar el valor calculado con $t_{\alpha[\infty]}$ en la tabla **III**.

diferencia absoluta máxima (de las nueve diferencias) se clasifica como 9. Las filas ligadas se calculan como promedios de las filas; así pues, si la cuarta y quinta diferencia tienen la misma magnitud absoluta se les asignaría a ambas la fila 4,5. Una vez calculadas las filas, el signo original de cada diferencia se asigna a la fila correspondiente. A continuación se calcula la suma de las filas positivas o de las negativas, la que sea menor en valor absoluto (se denomina T_s) y se compara con el valor crítico de T en la tabla **XIV** para el tamaño muestral correspondiente. A la vista de la significación de la suma de filas, es evidente que la raza *B* tiene un tamaño de camada diferente del de la raza 13. Esta es una prueba muy fácil de realizar, pero naturalmente no es tan eficiente como la correspondiente prueba *t* paramétrica, la cual sería preferible si se cumpliesen los requisitos necesarios. Hay que hacer notar que se necesitan seis diferencias como mínimo para realizar la prueba de rangos con signo de Wilcoxon. En seis comparaciones apareadas, todas las diferencias deben ser del mismo signo para que la prueba sea significativa al 5 %.

Para una muestra grande se puede utilizar una aproximación a la curva normal, que se presenta en el cuadro 10.2. Obsérvese que las magnitudes absolutas de estas diferencias desempeñan un papel, solamente en tanto en cuanto afecten a las posiciones de las diferencias.

Una prueba aún más sencilla es la *prueba de signos* la cual cuenta el número de signos positivos y negativos entre las diferencias (omitiendo todas las diferencias de cero). Entonces contrastamos la hipótesis de que los n signos más y menos se muestrean de una población en la que los dos tipos de signos se presentan en las mismas proporciones, como pudiera esperarse si no hubiese verdadera diferencia entre las dos muestras apareadas. Este muestreo debería seguir la distribución binomial, y el contraste de la hipótesis de que la frecuencia paramétrica de los signos más es $\hat{p} = 0{,}5$, puede hacerse de varias maneras. Vamos a estudiar esto aplicando la prueba de signos a los datos de cobayas del cuadro 10.2. Hay nueve diferencias, de las cuales ocho son positivas y una negativa. Podríamos seguir los métodos de la sección 4.2 (ilustrados en la tabla 4.3) en los cuales calculamos la probabilidad esperada de muestrear un signo menos en una muestra de nueve en el supuesto de que $\hat{p} = \hat{q} = 0{,}5$. La probabilidad de este suceso y todos los "peores" es igual a 0,0195. Como no tenemos nociones a priori de que una raza tuviese un tamaño de camada mayor que la otra, ésta es una prueba de dos colas y duplicamos la probabilidad a 0,0390. Sin duda éste es un suceso improbable y rechazamos la hipótesis nula de que $\hat{p} = \hat{q} = 0{,}5$.

Como el cálculo de las probabilidades exactas puede resultar muy pesado si no se tiene a mano una tabla de probabilidades acumulativas binomiales, otra posibilidad es utilizar la tabla **IX**, que da límites de confianza de p para varios tamaños y resultados de muestreo. Buscando tamaño de muestreo 9 e $Y = 1$ (número que representa la propiedad) encontramos que los límites de confianza para el 95 % son 0,0028 y 0,4751 por interpolación, excluyendo así el valor $\hat{p} = \hat{q} = 0{,}5$ postulado por la hipótesis nula. Por lo menos al nivel de significación del 5 % podemos concluir que es improbable que el número de signos más y menos sea el mismo. Los límites de confianza implican una distribución de dos colas; si hacemos una prueba de una cola, podemos deducir un nivel de significación de 0,025 de los límites de confianza del 95 % y un nivel 0,005 de los límites del 99 %. Naturalmente, esta prueba de una cola solamente se realizaría si los resultados estuviesen en la dirección de la hipótesis alternativa. Así, si la hipótesis alternativa fuese que la raza 13 en el cuadro

10.2 tenía mayor tamaño de camada que la raza B, no nos hubiésemos molestado en absoluto en probar este ejemplo, ya que la proporción observada de años que exhiben esta relación era menor de la mitad. Para muestras mayores, podemos utilizar la aproximación normal a la distribución binomial del siguiente modo: $t_s = (Y - \mu)/\sigma_Y = (Y - kp)/\sqrt{kpq}$, sustituyendo la media y la desviación típica de la distribución binomial mostradas en la sección 4.2. En nuestro caso, admitimos que n representa k y suponemos que $p = q = 0{,}5$. Por lo tanto, $t_s = (Y - \frac{1}{2}n)/\sqrt{\frac{1}{4}n} = (Y - \frac{1}{2}n)/\frac{1}{2}\sqrt{n}$. El valor de t_s se compara entonces con $t_{\alpha[\infty]}$ en la tabla **III**, utilizando una o dos colas de la distribución según se justifique. Cuando el tamaño de muestreo $n \geq 12$, ésta es una aproximación satisfactoria.

Una tercera vía sería probar la desviación de lo esperado $\hat{p} = \hat{q} = 0{,}5$ por uno de los métodos del capítulo 13.

Ejercicios 10

10.1 En un estudio del color de las flores en *Asclepias tuberosa*, Woodson (1964) obtuvo los siguientes resultados:

Región geográfica	$\bar{Y}$	n	s
C1	29,3	226	4,59
SW2	15,8	94	10,15
SW3	6,3	23	1,22

La variable considerada fue una valoración de color (variando desde 1 para amarillo puro hasta 40 para rojo anaranjado fuerte) obtenido comparando pétalos de flor con colores de muestra del *Diccionario de Colores* de Maerz y Paul. Probar si las muestras son homoscedásticas.

10.2 Allee y Bowen (1932) estudiaron el tiempo de supervivencia de la carpa dorada (en minutos) cuando se coloca en suspensiones coloidales de plata. El experimento N.º 9 incluye 5 replicaciones y el experimento N.º 10 incluye 10. ¿Difieren los resultados de los dos experimentos? La adición de urea, ClNa y SNa_2 a una tercera serie de suspensiones prolonga aparentemente la vida del pez.

Suspensión coloidal de plata		
Experimento N.º 9	*Experimento N.º 10*	*Urea y sales añadidas*
210	150	330
180	180	300
240	210	300
210	240	420
210	240	360
	120	270
	180	360
	240	360
	120	300
	150	120

Analizar e interpretar. Probar la igualdad de varianzas. Comparar los resultados del análisis de la varianza con los obtenidos utilizando la prueba U de Mann-Whitney para las dos comparaciones en estudio. Para probar el efecto de la urea puede que fuese mejor combinar los experimentos 9 y 10, si demuestran no diferir significativamente. SOLUCION. $U_s = 33$ entre los experimentos 9 y 10.

10.3 Número de bacterias en 1 cm^3 de leche de tres vacas contadas en tres períodos (datos de Park, Williams y Krumwiede, 1924).

	En el momento de ordeñar	*A las 24 horas*	*A las 48 horas*
Vaca N.º 1	12 000	14 000	57 000
2	13 000	20 000	65 000
3	21 500	31 000	106 000

(a) Calcular las medias y varianzas para los tres períodos y examinar la relación entre estos dos estadísticos. Transformar las variantes en logaritmos y comparar las medias y varianzas basadas en los datos transformados. Discutir.

(b) Hallar el análisis de la varianza basado en los datos transformados y no transformados. Discutir los resultados.

10.4 Probar la diferencia de pH entre la superficie y el subsuelo en los datos del ejercicio 9.1, utilizando la prueba del rango con signo de Wilcoxon. SOLUCION. $T_s = 38$; $P > 0{,}10$.

Capítulo 11

Regresión

Hasta ahora nuestros estudios han tratado de una variable solamente. Sin embargo, con frecuencia medimos dos o más variables en cada individuo y, por consiguiente, nos gustaría poder expresar con más precisión la naturaleza de las relaciones entre estas variables. Esto nos lleva a las materias de *regresión y correlación.* En la regresión estimamos la relación de una variable con otra, expresando la primera en términos de una función lineal (o más compleja) de la otra. En el análisis de correlación, que a veces se confunde con regresión, estimamos el grado en que dos variables varían simultáneamente. El capítulo 12 trata de correlación, y aplazaremos hasta entonces nuestro esfuerzo para clarificar la relación y distinción entre regresión y correlación. Las variables implicadas en regresión y correlación son continuas, o si son merísticas, se tratan como si fuesen continuas. Si las variables son cualitativas (es decir, atributos), no pueden aplicarse los métodos de regresión y correlación.

En la sección 11.1, revisamos la noción de funciones matemáticas e introducimos la nueva terminología requerida para el análisis de regresión. Esto va seguido en la sección 11.2 por una discusión de los modelos estadísticos apropiados para el análisis de regresión. En la sección 11.3 se presentan los cálculos básicos de la regresión lineal simple para el caso de una variante dependiente por cada variante independiente. El caso de varias variantes dependientes por cada variante independiente se trata en la sección 11.4. En la sección 11.5 se discuten pruebas de significación y cálculo de intervalos de confianza para problemas de regresión.

La sección 11.6 sirve como un resumen de la regresión y trata de las diversas aplicaciones del análisis de regresión en biología. Finalmente, la sección 11.7 demuestra como la transformación de escala puede rectificar las relaciones curvilíneas para facilitar el análisis.

11.1 Introducción a la regresión

Gran parte del pensamiento científico atañe a las relaciones entre pares de variables que se supone están en una relación causa-efecto. Nos contentaremos con establecer la forma y significado de relaciones funcionales entre dos variables, dejando la demostración de relaciones causa-efecto para los procedimientos establecidos del método científico. Una *función* es una relación matemática que nos permite predecir qué valores de una variable Y corresponden a determinados valores de una variable X. Esta relación, escrita generalmente como $Y = f(X)$, nos resulta familiar a todos.

Una regresión lineal típica es de la forma representada en la figura 11.1, la cual ilustra el efecto de dos drogas en las presiones sanguíneas de dos especies de animales. Las relaciones representadas en esta gráfica pueden expresarse por la fórmula $Y = a + bX$. Claramente, Y es una función de X. A la variable Y la denominamos *variable dependiente,* mientras que X se denomina *variable independiente.* La magnitud de la presión sanguínea Y depende de la cantidad de la droga X y, por lo tanto, puede predecirse a partir de la variable independiente, que se supone es libre de variar. Aunque una causa siempre se considerará una variable independiente y un efecto una variable dependiente, una relación funcional observada en la naturaleza puede no ser en realidad una relación causa-efecto. La línea más alta es de la relación $Y = 20 + 15X$, que representa el efecto de la droga A en el animal P. La cantidad de droga se mide en microgramos, la presión sanguínea en milímetros de mercurio. Así, después de haberse administrado 4 μg de la droga, la presión sanguínea sería $Y = 20 + (15)(4) = 80$ mm Hg. La variable independiente X se multiplica por un coeficiente b, la pendiente. En el ejemplo elegido, $b = 15$; esto es, para un incremento en la droga de un microgramo la presión sanguínea se eleva en 15 mm.

En biología, esta relación puede claramente ser apropiada sobre un rango limitado de valores de X solamente. Los valores negativos de X no tienen sentido en este caso y es improbable que la presión sanguínea continúe aumentando a una velocidad uniforme. Muy probablemente la pendiente de la relación funcional se aplanará al aumentar el nivel de droga. Pero para una porción limitada del rango de la variable X (microgramos de la droga), la relación lineal $Y = a + bX$ puede ser una descripción adecuada de la dependencia funcional de Y sobre X.

Según esta fórmula, cuando la variable independiente es igual a cero, la variable dependiente es igual a a. Este punto es la intersección de la línea función con el eje Y. Se denomina *ordenada en el origen.* En la figura 11.1, cuando $X = 0$, la función recién estudiada dará una presión sanguínea de 20 mm, que es la presión sanguínea normal del animal P en ausencia de la droga.

Las otras dos funciones de la figura 11.1 muestran los efectos de variar ambos a, la ordenada en el origen, y b, la pendiente. En la línea inferior $Y = 20 + 7{,}5X$, la ordenada en el origen no cambia, pero la pendiente se ha reducido a la mitad. Concebimos esto como el efecto de una droga diferente B en el mismo organismo P. Evidentemente, cuando no se administra droga, la presión sanguínea estaría en la misma ordenada en el origen, puesto que se está estudiando el mismo organismo. Sin embargo, una droga diferente es probable que ejerza un efecto hipertensor diferente, como se refleja por la diferente pendiente. La tercera relación describe también el efecto de la droga B, que se supone permanece igual, pero el experimento se realiza en una especie diferente Q, cuya

presión sanguínea normal se supone que es 40 mm de Hg. Así, la ecuación para el efecto de la droga B en la especie Q se escribe como $Y = 40 + 7{,}5X$. Esta línea es paralela a la estudiada previamente.

Por conocimientos básicos de geometría analítica se habrá reconocido el factor pendiente b como la *pendiente* de la función $Y = a + bX$, simbolizada generalmente por m. En cálculo diferencial, b es la *derivada* de esa misma función ($dY/dX = b$). En bioestadística, b se denomina el *coeficiente de regresión*. Cuando queremos subrayar que el coeficiente de regresión es de la variable Y en función de la variable X, escribimos $b_{Y \cdot X}$.

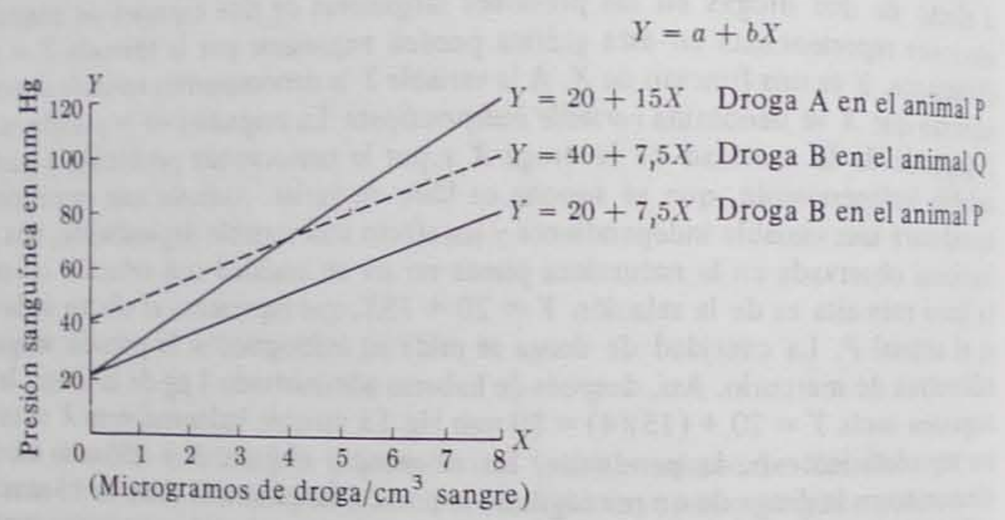


Fig. 11.1. Presión sanguínea de un animal en mm de Hg en función de la concentración de droga en μg por cm^3 de sangre.

11.2 Modelos en regresión

En cualquier ejemplo real las observaciones no se hallan situadas completamente a lo largo de una línea de regresión, a causa del error aleatorio de medida y de los efectos de factores ambientales impredecibles. Así, en regresión una relación funcional no significa que dado un valor de X, el valor de Y debe ser $a + bX$, sino más bien que la media (o valor esperado) de Y es $a + bX$.

Las pruebas de significación en regresión se basan en los dos modelos siguientes. El más común de éstos, *modelo I de regresión*, es especialmente adecuado en situaciones experimentales. Está basado en cuatro supuestos.

1. La variable independiente X se mide sin error. Por esto decimos que los valores de X son "fijos". Esto significa que solamente Y, la variable dependiente, es una variable aleatoria; X no varía al azar sino que está bajo el control del investigador. Así en el ejemplo de la figura 11.1 hemos variado a voluntad la dosis de droga y estudiado la respuesta de la variable aleatoria presión sanguínea. Podemos manipular X de la misma manera que podíamos manipular el efecto de tratamiento en un análisis de la varianza

modelo I. En realidad, como se verá más adelante, hay una relación muy íntima entre análisis de la varianza modelo I y regresión modelo I.

2. El valor esperado de la variable Y para un determinado valor de X está descrito por la función lineal $\mu_Y = \alpha + \beta X$. Esta es la misma relación que hemos encontrado antes, pero utilizamos letras griegas para a y b, puesto que estamos describiendo una relación paramétrica. Otra forma de establecer este supuesto es que las medias paramétricas μ_Y de los valores de Y son una función de X y se hallan en una línea recta descrita por esta ecuación.

3. Para cualquier valor dado X_i, los valores de Y_i se distribuyen independiente y normalmente. Esto puede representarse por la ecuación $Y_{ij} = \alpha + \beta X_i + \epsilon_{ij}$, en la que ϵ_{ij} se supone que es un error normalmente distribuido con una media de cero. La figura 11.2 ilustra este concepto con una línea de regresión similar a las de la figura 11.1. Un determinado experimento puede repetirse varias veces. Así, por ejemplo, podríamos administrar 2, 4, 6, 8 y 10 μg de droga a cada uno de 20 individuos de una especie animal y obtener una distribución de frecuencias de las respuestas de presión sanguínea Y para las variantes independientes $X = 2, 4, 6, 8$ y 10 μg. En vista de la inherente variabilidad del material biológico, es evidente que las respuestas a cada dosis no serían las mismas en todos los individuos; se obtendría una distribución de frecuencias de valores de Y (presión sanguínea) en torno al valor esperado. El supuesto 3 establece que estos valores de muestreo estarían independiente y normalmente distribuidos. Esto se indica en la figura 11-2 por las curvas normales, que se superponen en varios puntos en la línea de regresión.

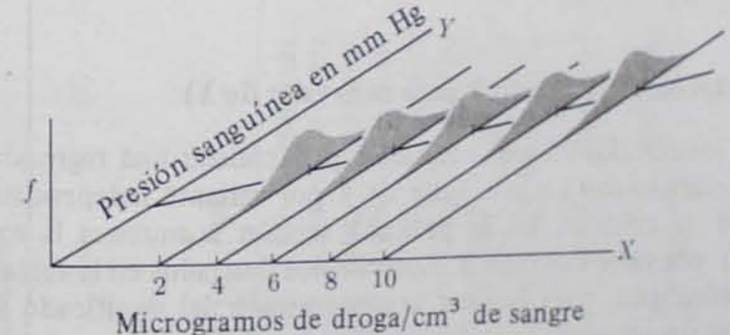


Fig. 11.2. Presión sanguínea de un animal en mm de Hg como función de la concentración de droga en μg por cm^3 de sangre. Muestreo repetido para una determinada concentración de droga.

En realidad hay claramente una dispersión continua, como si estas distribuciones normales diferentes estuviesen apiladas muy próximas unas a otras, habiendo, después de todo, una infinidad de valores intermedios de X entre dos dosis cualesquiera. En los casos raros en que la variable independiente es discontinua, las distribuciones de Y estarían físicamente separadas entre sí y se hallarían solamente a lo largo de aquellos puntos de la abscisa que corresponden a las variantes independientes. Un ejemplo de este caso sería el peso de las crías (Y) como una función del número de crías (X) en camadas de ratones.

Puede haber tres o cuatro crías por camada, pero no existirían valores intermedios de X representando 3,25 ratones por camada.

No todo experimento tendrá más de una réplica de Y para cada valor de X. De hecho, los cálculos básicos que aprenderemos en la próxima sección son para un solo valor de Y por valor de X, siendo éste el caso más común. No obstante, deberías advertir que incluso en estos casos el supuesto básico del modelo I de regresión es que la variante única de Y correspondiente al valor dado de X es una muestra de una población de variantes independiente y normalmente distribuidas.

4. El supuesto final es ya familiar. Suponemos que estas muestras a lo largo de la línea de regresión son homoscedásticas; es decir, tienen una varianza común, σ^2, que es la varianza de los ϵ en la expresión anterior. Así pues, suponemos que la varianza en torno a la línea de regresión es constante e independiente de la magnitud de X ó Y.

En la *regresión modelo II* la variable independiente también se mide con error. No consideramos que X es fija y bajo control del investigador. Supongamos que se muestrea una población de moscas hembras y se mide un número de ovariolas y el peso total de cada individuo. Probablemente las distribuciones de estas variables serán diferentes. Podría interesar estudiar el número de ovariolas como función del peso. En este caso el peso, que se supone es la variable independiente, no es fijo e indudablemente no es la "causa" del desarrollo ovárico. Aunque, como se discutirá en el próximo capítulo, los casos de este tipo se analizan mucho mejor por los métodos del análisis de correlación, a veces queremos describir la relación funcional entre tales variables. Para hacer esto, necesitamos recurrir a las técnicas especiales de regresión modelo II. En este libro nos limitaremos a un tratamiento de la regresión modelo I.

11.3 Los cálculos básicos (un solo Y para cada valor de X)

Para aprender los cálculos básicos necesarios para realizar una regresión lineal modelo I, elegiremos un ejemplo con un solo valor de Y por variante independiente X, ya que éste es más sencillo de calcular. En la próxima sección se muestra la extensión a valores replicados de Y por valor único de X. Los cálculos ilustrados en la tabla 11.1 se presentan por razones pedagógicas para facilitar la comprensión del significado de la regresión. Al final de esta sección aparecen fórmulas sencillas de cálculo.

Los datos en los que aprenderemos la regresión proceden de un estudio de pérdida de agua en el *Tribolium confusum*, el coleóptero de la harina. Se pesaron nueve lotes de 25 coleópteros (los coleópteros individuales no podrían pesarse con el equipo disponible), se guardaron a diferentes humedades relativas, y se pesaron de nuevo después de seis días de inanición. Se calculó la pérdida de peso en miligramos para cada lote. Esto es sin duda una regresión modelo I en la que la pérdida de peso es la variable dependiente Y y la humedad relativa es la variable independiente X, un efecto de tratamiento fijo bajo el control del investigador. El objeto del análisis es establecer si la relación entre humedad relativa y pérdida de peso puede ser descrita adecuadamente por una regresión lineal de la forma general $Y = a + bX$. Los datos originales se exponen en las columnas (1) y (2) de la tabla 11.1. Se representan gráficamente en la figura 11.3, en la cual se muestra que existe una relación negativa entre pérdida de peso y humedad; al aumentar la humedad, disminuye la

TABLA 11.1

Cálculos básicos en regresión. Pérdida de peso (en mg) de nueve lotes de 25 coleópteros *Tribolium* tras seis días de inanición a nueve humedades diferentes.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Porcentaje de humedad relativa X	*Pérdida de peso en mg* Y	$x = (X - \overline{X})$	$y = (Y - \overline{Y})$	x^2	xy	y^2	$\hat{Y}$	$d_{Y \cdot X} = Y - \hat{Y}$	$d^2_{Y \cdot X}$	$\hat{y} = \hat{Y} - \overline{Y}$	$\hat{y}^2$
0	8.98	−50.39	2.958	2539.1521	−149.0536	8.7498	8.7038	0.2762	0.0763	2.6818	7.1921
12	8.14	−38.39	2.118	1473.7921	− 81.3100	4.4859	8.0652	0.0748	0.0056	2.0432	4.1747
29.5	6.67	−20.89	0.648	436.3921	− 13.5367	0.4199	7.1338	−0.4638	0.2151	1.1118	1.2361
43	6.08	− 7.39	0.058	54.6121	− 0.4286	0.0034	6.4153	−0.3353	0.1124	0.3933	0.1547
53	5.90	2.61	−0.122	6.8121	− 0.3184	0.0149	5.8831	0.0169	0.0003	−0.1389	0.0193
62.5	5.83	12.11	−0.192	146.6521	− 2.3251	0.0369	5.3776	0.4524	0.2047	−0.6444	0.4153
75.5	4.68	25.11	−1.342	630.5121	− 33.6976	1.8010	4.6857	−0.0057	0.0000	−1.3363	1.7857
85	4.20	34.61	−1.822	1197.8521	− 63.0594	3.3197	4.1801	0.0199	0.0004	−1.8419	3.3926
93	3.72	42.61	−2.302	1815.6121	− 98.0882	5.2992	3.7543	−0.0343	0.0012	−2.2677	5.1425
Suma 453.5	54.20	− 0.01	0.002	8301.3889	−441.8176	24.1307	54.1989	0.0011	0.6160	0.0009	23.5130
Media 50.39	6.022						6.022				
Suma/$(n-1)$				1037.6736	− 55.2272	3.0163			0.0880[a]		

Fuente: Nelson (1964).
[a] Suma dividida por $n - 2$.

pérdida de peso. Las medias de pérdida de peso y humedad relativa $\bar{Y}$ y $\bar{X}$, respectivamente, se marcan a lo largo de los ejes de coordenadas. La humedad media es 50,39 % y la pérdida de peso media es 6,022 mg. ¿Cómo podemos ajustar una línea de regresión a estos datos, que nos permita estimar un valor de Y para un valor dado de X? A no ser que las observaciones reales se hallen exactamente en una línea recta, necesitaremos un criterio para determinar la mejor colocación posible de la línea de regresión. Generalmente los estadísticos han seguido el principio de mínimos cuadrados, que vimos por primera vez en el capítulo 3 al instruirnos sobre la media aritmética y la varianza. Si trazásemos una línea horizontal a través de $\bar{X}$, $\bar{Y}$ (es decir, una línea paralela al eje X a la altura de $\bar{Y}$), las desviaciones de esa línea trazadas paralelas al eje Y representarían las desviaciones de la media para estas observaciones con respecto a la variable Y (véase figura 11.4). En el

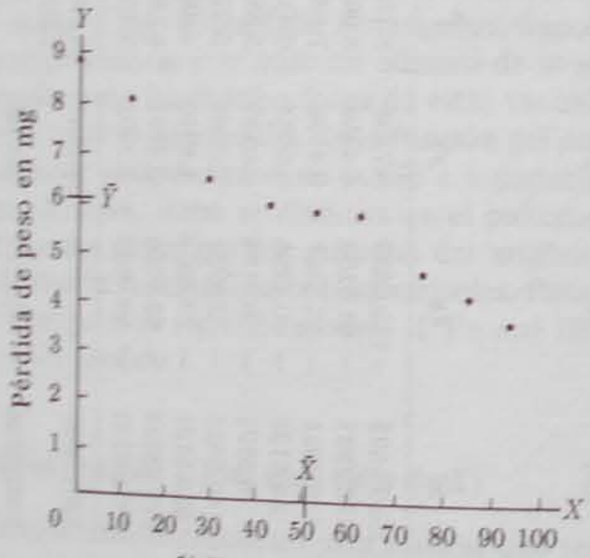


Fig. 11.3. Pérdida de peso (en mg) de nueve lotes de 25 coleópteros *Tribolium* después de seis días de inanición a nueve humedades relativas diferentes. Datos de la tabla 11.1 de Nelson (1964).

capítulo 3 vimos que la suma de estas desviaciones, $\sum(Y - \bar{Y}) = \sum y = 0$. La suma de cuadrados de estas desviaciones, $\sum(Y - \bar{Y})^2 = \sum y^2$, es menor que la de cualquier otra línea horizontal. Otra forma de expresar esto es que la media aritmética de Y representa la línea horizontal de mínimos cuadrados. Cualquier línea horizontal trazada entre los datos en un punto distinto de $\bar{Y}$ daría una suma de desviaciones distinta de cero y una suma de desviaciones cuadráticas mayor que $\sum y^2$. Por consiguiente, un método matemáticamente correcto pero no práctico para hallar la media de Y sería trazar una serie de líneas horizontales sobre una gráfica, calcular la suma de cuadrados de sus desviaciones y escoger la línea que diese la mínima suma de cuadrados.

En regresión lineal, seguimos trazando una línea recta entre nuestras observaciones, pero ya no es necesariamente horizontal. Una línea de regresión inclinada indicará

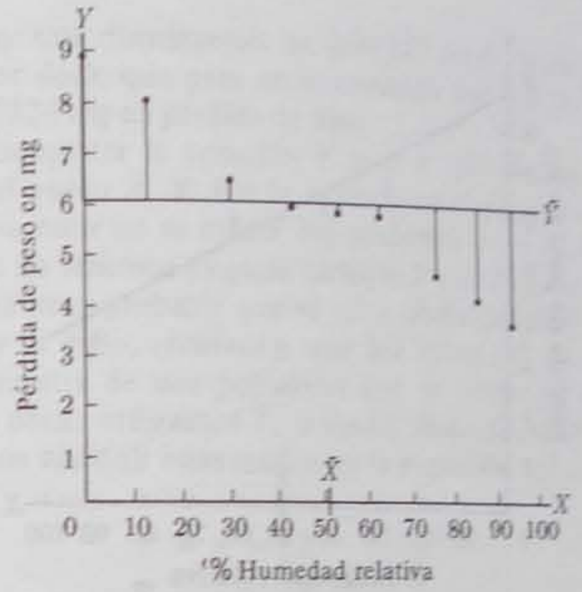


Fig. 11.4. Desviaciones de la media (de Y) para los datos de la figura 11.3.

para cada valor de la variable independiente X_i un valor estimado de la variable dependiente. Deberíamos distinguir el valor estimado de Y_i, que de aquí en adelante designaremos como $\hat{Y}_i$ (léase: Y con signo de intercalación), y los valores observados, designados convencionalmente como Y_i. Por lo tanto, la ecuación de regresión se expresaría

$$\hat{Y} = a + bX \tag{11.1}$$

la cual indica que para valores dados de X, esta ecuación calcula los valores estimados de $\hat{Y}$ (distintos de los valores observados de Y en cualquier caso real). La desviación de una observación Y_i respecto de la línea de regresión es $(Y_i - \hat{Y}_i)$ y se simboliza generalmente por $d_{Y \cdot X}$. Estas desviaciones aún pueden trazarse paralelas al eje Y, pero como la línea de regresión es oblicua la cruzan en un ángulo (véase figura 11.5). La suma de estas desviaciones sigue siendo cero ($\sum d_{Y \cdot X} = 0$), y la suma de sus cuadrados da una cantidad $\sum(Y - \hat{Y})^2 = \sum d^2_{Y \cdot X}$ análoga a la suma de cuadrados $\sum y^2$. Por razones que se aclararán más adelante, $\sum d^2_{Y \cdot X}$ se denomina suma de cuadrados inexplicable. La recta de regresión lineal mínima cuadrática entre una serie de puntos se define como la línea recta que conduce a que $\sum d^2_{Y \cdot X}$ sea mínima. Geométricamente, la idea básica es que sería preferible utilizar una línea que fuese en cierto sentido próxima al mayor número posible de puntos. Para los fines del análisis de regresión, es más conveniente definir proximidad en términos de las distancias verticales de los puntos a una línea, y utilizar la línea que haga mínima la suma de estas desviaciones cuadráticas. Una consecuencia a propósito de este criterio es que la línea debe pasar por el punto $\bar{X}$, $\bar{Y}$. Una vez más sería factible pero no práctico calcular la pendiente de regresión correcta girando una regla alrededor del punto $\bar{X}$, $\bar{Y}$, y calcular la suma de cuadrados inexplicable, $\sum d^2_{Y \cdot X}$, para cada una de las innumerables posiciones posibles. La posición que diese el valor mínimo de $\sum d^2_{Y \cdot X}$ sería la línea de regresión de mínimos cuadrados.

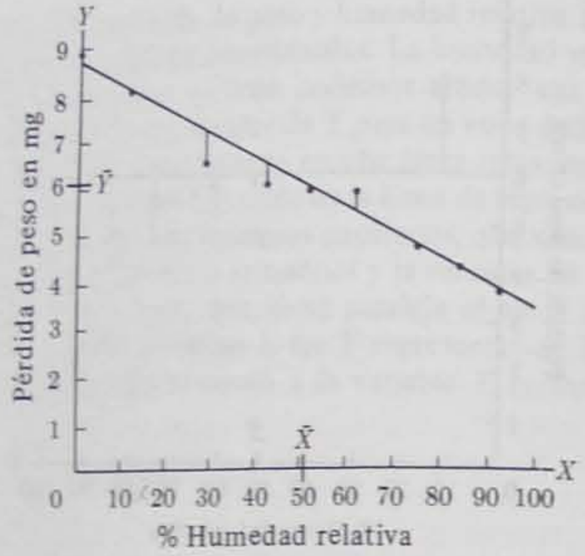


Fig. 11.5. Desviaciones de la línea de regresión para los datos de la figura 11.3.

La fórmula de la pendiente de una línea basada en un valor mínimo de $\sum d^2_{Y \cdot X}$ se obtiene por medio del cálculo diferencial. Esta es

$$b_{Y \cdot X} = \frac{\sum xy}{\sum x^2} \tag{11.2}$$

Vamos a calcular $b = \sum xy / \sum x^2$ para nuestros datos de pérdida de peso.

Primero calculamos las desviaciones de las respectivas medias de X e Y como se muestra en las columnas (3) y (4) de la tabla 11.1. Las sumas de estas desviaciones, $\sum x$ y $\sum y$, son ligeramente diferentes de su valor esperado de cero por errores de redondeo. Los cuadrados de estas desviaciones dan las sumas de cuadrados y varianzas en las columnas (5) y (7). En la columna (6) hemos calculado los productos xy, que en este ejemplo son todos negativos porque las desviaciones son de distinto signo. Un aumento en la humedad conduce a una disminución en la pérdida de peso. La suma de estos productos $\sum xy$ es una nueva cantidad, denominada *suma de productos*. Este es un término inadecuado pero bien establecido, que hace referencia a $\sum xy$, la suma de los productos de las desviaciones en lugar de $\sum XY$, la suma de los productos de las variantes. Se recordará que $\sum y^2$ se denomina suma de cuadrados, mientras que $\sum Y^2$ es la suma de los cuadrados de las variantes. La suma de productos es análoga a la suma de cuadrados. Al dividirla por los grados de libertad, da la *covarianza* por analogía con la varianza que resulta de una división similar de la suma de cuadrados. Se puede recordar que ya se han encontrado covarianzas anteriormente en la sección 7.4. Obsérvese que la suma de productos puede ser tanto positiva como negativa. Si es negativa, esto indica una pendiente negativa de la línea de regresión: al aumentar X, disminuye Y. Con respecto a esto difiere de una suma de cuadrados, la cual solamente puede ser positiva. En la tabla 11.1 hallamos que $\sum xy = -441{,}8176$, $\sum x^2 = 8301{,}3889$, y $b = \sum xy / \sum x^2 = 0{,}05322$. Así, para un incremento en

X de una unidad, hay una disminución de 0,05322 en Y. Relacionándolo con nuestro ejemplo real, podemos decir que para un incremento del 1 % en humedad relativa, hay una reducción de 0,05322 mg en pérdida de peso.

¿Cómo podemos completar la ecuación $Y = a + bX$? Hemos dicho que la línea de regresión pasará por el punto $\bar{X}$, $\bar{Y}$. Por lo tanto, cuando X está en su media, estimamos que Y también debería estar en su media. No podemos decir si esto es realmente así en nuestros datos, ya que no tenemos ninguna variante X exactamente en la media. Incluso si la tuviésemos, no sería muy probable que el valor observado de Y estuviese exactamente en la media. Al fin y al cabo, recuérdese que los valores Y de nuestras observaciones solamente son una muestra de una población que se centra en torno a $\mu_{\hat{Y}}$. Para $\bar{X} = 50{,}39$, $\bar{Y} = 6{,}022$; es decir, utilizamos $\bar{Y}$, la media observada de Y, como una estimación $\hat{Y}$ de la media. Podemos sustituir estas medias en la expresión (11.1):

$$\begin{aligned} \hat{Y} &= a + bX \\ \bar{Y} &= a + b\bar{X} \\ a &= \bar{Y} - b\bar{X} \\ a &= 6{,}022 - (-0{,}05322)(50{,}39) \\ &= 8{,}7038 \end{aligned}$$

Por lo tanto

$$\hat{Y} = 8{,}7038 - 0{,}05322X.$$

Esta es la ecuación que relaciona la pérdida de peso con la humedad relativa. Obsérvese que cuando X es cero (humedad cero), la pérdida de peso estimada es mayor. Es entonces igual a $a = 8{,}7038$ mg. Pero al aumentar X hasta un máximo de 100, la pérdida de peso disminuiría hasta 3,3818 mg.

Podemos utilizar la fórmula de regresión para trazar la línea de regresión: simplemente estimamos $\hat{Y}$ para dos puntos convenientes de X tales como $X = 0$ y $X = 100$ y trazamos una línea recta entre ellos. Esta línea se ha adicionado a los datos observados y se muestra en la figura 11.6. Obsérvese que pasa por el punto $\bar{X}$, $\bar{Y}$. De hecho, para trazar la línea de regresión, frecuentemente utilizamos la intersección de las dos medias y otro punto.

Como

$$a = \bar{Y} - b\bar{X}$$

podemos escribir la expresión (11.1), $\hat{Y} = a + bX$, como

$$\begin{aligned} \hat{Y} &= (\bar{Y} - b\bar{X}) + bX \\ &= \bar{Y} + b(X - \bar{X}) \end{aligned}$$

Por lo tanto

$$\hat{Y} = \bar{Y} + bx$$

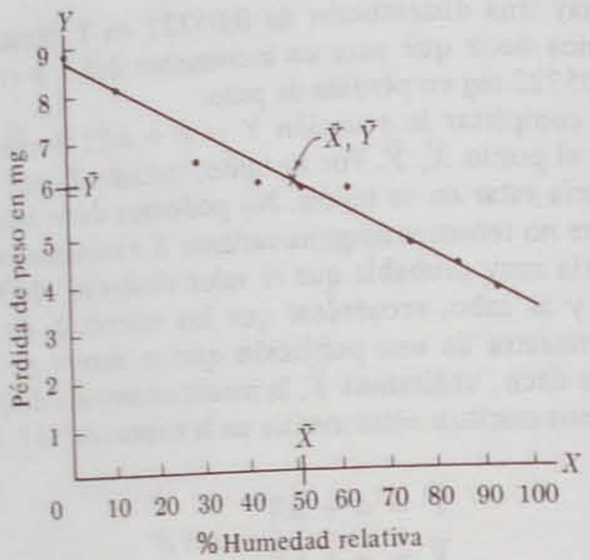


Fig. 11.6. Regresión lineal ajustada a los datos de la figura 11.3.

Asimismo

$$\hat{Y} - \bar{Y} = bx$$
$$\hat{y} = bx \tag{11.3}$$

donde $\hat{y}$ se define como la desviación $\hat{Y} - \bar{Y}$. A continuación, utilizando la expresión (11.1), estimamos $\hat{Y}$ para cada uno de nuestros valores dados de X. Los valores estimados $\hat{Y}$ se exponen en la columna (8) de la tabla 11.1. Se pueden comparar con los valores observados de Y en la columna (2). La concordancia global entre las dos columnas es buena. Se pone de manifiesto que, salvo errores de redondeo, $\sum \hat{Y} = \sum Y$ y por consiguiente $\bar{\hat{Y}} = \bar{Y}$. Sin embargo, nuestros valores reales de Y ordinariamente son diferentes de los valores estimados $\hat{Y}$. Esto es debido a la variación individual en torno a la línea de regresión. No obstante, la línea de regresión es una base sobre la que se pueden calcular desviaciones, mejor que a partir de la media aritmética $\bar{Y}$, ya que se ha tenido en cuenta el valor de X al construirla.

Cuando calculamos desviaciones de cada valor Y observado respecto de su valor estimado $(Y - \hat{Y}) = d_{Y \cdot X}$ y las ponemos en una lista en la columna (9), observamos que estas desviaciones exhiben una de las propiedades de las desviaciones de una media: suman cero excepto por errores de redondeo. Así $\sum d_{Y \cdot X} = 0$, del mismo modo que $\sum y = 0$. A continuación, en la columna (10) calculamos los cuadrados de estas desviaciones de valores observados de Y respecto de valores estimados por regresión y los sumamos para dar una nueva suma de cuadrados, $\sum d^2_{Y \cdot X} = 0{,}6160$. Cuando comparamos $\sum (Y - \bar{Y})^2 = \sum y^2 = 24{,}1307$ con $\sum (Y - \hat{Y})^2 = \sum d^2_{Y \cdot X} = 0{,}6160$, observamos que la nueva suma de cuadrados es mucho menor que la anterior. ¿Qué ha causado esta reducción? La línea de regresión es una serie de medias (una para cada valor de X) cuando se compara con la única media aritmética de Y. Al tener en cuenta diferentes magnitudes de X se ha

eliminado la mayor parte de la varianza de Y respecto de la muestra. Lo que falta es la *suma de cuadrados inexplicable* $\sum d^2_{Y \cdot X}$, que expresa la porción de la *S.C.* total de Y que no está justificada por diferencias en X. Es inexplicable con respecto a X. La diferencia entre la *S.C.* total, $\sum y^2$, y la *S.C.* inexplicable, $\sum d^2_{Y \cdot X}$, se denomina la *suma de cuadrados explicable*, $\sum \hat{y}^2$ y está basada en las desviaciones $\hat{y} = \hat{Y} - \bar{Y}$. En las columnas (11) y (12) se muestra el cálculo de esta desviación y su cuadrado. Obsérvese que $\sum \hat{y}$ se aproxima a cero y que $\sum \hat{y}^2 = 23{,}5130$. Al sumar a ésta la *S.C.* inexplicable = 0,6160 se obtiene $\sum y^2 = \sum \hat{y}^2 + \sum d^2_{Y \cdot X} = 24{,}1290$, que es igual (salvo errores de redondeo) al valor calculado independientemente de 24,1307 en la columna (7). En las secciones que siguen volveremos al significado de las sumas de cuadrados inexplicable y explicable.

Ahora pasamos a demostrar un método eficiente de cálculo para una ecuación de regresión en datos con valores únicos de Y para cada valor de X.

El coeficiente de regresión $\sum xy / \sum x^2$ puede volver a escribirse como

$$b_{Y \cdot X} = \frac{\sum^n (X - \bar{X})(Y - \bar{Y})}{\sum^n (X - \bar{X})^2} \tag{11.4}$$

El denominador de esta expresión es la suma de cuadrados de X. Su fórmula, encontrada por primera vez en la sección 3.9, es $\sum x^2 = \sum X^2 - (\sum X)^2 / n$. Ahora aprenderemos una fórmula análoga para el numerador de la expresión (11.4), la suma de productos. La fórmula habitual es

$$\sum^n xy = \sum^n XY - \frac{(\sum^n X)(\sum^n Y)}{n} \tag{11.5}$$

La cantidad $\sum XY$ es simplemente el producto acumulado de las dos variables. La expresión (11.5) se demuestra en el apéndice A1.6. Los cálculos propiamente dichos para una ecuación de regresión (un solo valor de Y por valor de X) se ilustran en el cuadro 11.1, utilizando los datos de pérdida de peso de la tabla 11.1.

El cuadro 11.1 muestra también cómo calcular la suma de cuadrados explicable $\sum \hat{y}^2 = \sum (\hat{Y} - \bar{Y})^2$ y la suma de cuadrados inexplicable $\sum d^2_{Y \cdot X} = \sum (Y - \hat{Y})^2$. Esta

$$\sum d^2_{Y \cdot x} = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2} \tag{11.6}$$

se demuestra en el apéndice A1.7. En esta demostración también se ve claro que la suma de cuadrados explicable es

$$\sum \hat{y}^2 = \sum b^2 x^2 = b^2 \sum x^2 = \frac{(\sum xy)^2}{(\sum x^2)^2} \sum x^2 \tag{11.7}$$

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

CUADRO 11.1

Cálculo de estadísticos de regresión. Un solo valor de Y para cada valor de X.

Datos de la tabla 11.1.

Pérdida de peso en mg (Y)	8,98	8,14	6,67	6,08	5,90	5,83	4,68	4,20	3,72
Porcentaje de humedad relativa (X)	0	12,0	29,5	43,0	53,0	62,5	75,5	85,0	93,0

Cálculos básicos

1. Calcular el tamaño de muestreo, las sumas, las sumas de los cuadrados de las observaciones y la suma de los valores de XY.

$$n = 9 \qquad \sum X = 453{,}5 \qquad \sum Y = 54{,}20$$

$$\sum X^2 = 31\,152{,}75 \qquad \sum Y^2 = 350{,}5350 \qquad \sum XY = 2289{,}260$$

2. Las medias, sumas de cuadrados, y sumas de productos son

$$\bar{X} = 50{,}389 \qquad \bar{Y} = 6{,}022$$

$$\sum x^2 = 8301{,}3889 \qquad \sum y^2 = 24{,}1306$$

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

$$= 2289{,}260 - \frac{(453{,}5)(54{,}20)}{9} = -441{,}8178$$

3. El coeficiente de regresión es

$$b_{Y \cdot X} = \frac{\sum xy}{\sum x^2} = \frac{-441{,}8178}{8301{,}3889} = -0{,}05322$$

4. La ordenada en el origen es

$$a = \bar{Y} - b_{Y \cdot X}\bar{X} = 6{,}022 - (-0{,}05322)(50{,}389) = 8{,}7037$$

5. La suma de cuadrados explicable es

$$\sum \hat{y}^2 = \frac{(\sum xy)^2}{\sum x^2} = \frac{(-441{,}8178)^2}{8301{,}3889} = 23{,}5145$$

6. La suma de cuadrados inexplicable es

$$\sum d^2_{Y \cdot X} = \sum y^2 - \sum \hat{y}^2 = 24{,}1306 - 23{,}5145 = 0{,}6161$$

11.4 Más de un valor de Y para cada valor de X

Ahora nos ocupamos de la regresión modelo I tal como se ha definido originalmente en la sección 11.2 e ilustrado por la figura 11.2. Para cada valor del tratamiento X muestreamos repetidamente Y, obteniendo una distribución de muestreo de valores de Y para cada uno de los puntos de X elegidos. Hemos seleccionado un experimento del laboratorio de uno de nosotros (Sokal) en el cual se han desarrollado coleópteros *Tribolium* desde huevos hasta adultos, a cuatro densidades diferentes. Se ha calculado el porcentaje de supervivencia hasta el estado adulto para números de réplicas variables a estas densidades. Siguiendo la sección 10.2, a estos porcentajes se les han aplicado las transformaciones arcseno, que se indican en el cuadro 11.2. Es más probable que estos valores transformados sean normales y homoscedásticos que los porcentajes. La disposición de estos datos es muy parecida a la de un análisis de varianza modelo I de clasificación simple. Hay cuatro densidades diferentes y múltiples valores de supervivencia para cada densidad. Ahora nos gustaría determinar si hay diferencias en supervivencia entre los cuatro grupos, y también si podemos establecer una regresión de supervivencia en función de la densidad.

Por lo tanto, un primer paso es hacer un análisis de la varianza, utilizando los métodos de la sección 8.3 y de la tabla 8.1. El objeto de que hagamos esto se representa gráficamente en la figura 11.7. Si el análisis de la varianza no fuese significativo, esto indicaría que las medias no son significativamente diferentes unas de otras como se muestra en la figura 11.7A, y sería improbable que una línea de regresión ajustada a estos datos tuviese una pendiente significativamente diferente de cero. A veces, cuando las medias aumentan o disminuyen ligeramente al aumentar X, pueden no ser suficientemente diferentes para que la media cuadrática entre grupos sea significativa por análisis de la varianza; no obstante, podría encontrarse una regresión significativa. Ordinariamente éstos son casos límite de significación estadística. Cuando encontramos una marcada regresión de las medias sobre X, como se muestra en la figura 11.7B, ordinariamente encontraremos una diferencia significativa entre las medias por un análisis de la varianza. Sin embargo, no podemos utilizar este argumento en el sentido de decir que una diferencia significativa entre medias, probada por un análisis de la varianza, indica necesariamente que puede ajustarse a estos datos una regresión lineal significativa. En la figura 11.7C, las medias siguen una función en forma de U (una parábola). Aunque probablemente las medias serían significativamente diferentes entre sí, sin duda una línea recta ajustada a estos datos sería una línea horizontal situada a mitad de la distancia entre los puntos superior e inferior. En esta serie de datos, la regresión *lineal* solamente puede explicar una pequeña parte de la variación de la variable dependiente. Sin embargo, una regresión parabólica curvilínea se adaptaría a estos datos y eliminaría la mayor parte de la varianza de Y. Un caso similar se señala en la figura 11.7D, en el cual las medias describen un fenómeno periódicamente variable, aumentando y disminuyendo alternativamente. De nuevo la línea de regresión para estos datos tiene una pendiente cero. Una regresión curvilínea (cíclica) también podría ajustarse a estos datos, pero nuestro fin primordial al mostrar ese ejemplo es indicar que podría haber heterogeneidad entre las medias de Y aparentemente no relacionada con la magnitud de X. Recuérdese que en ejemplos reales casi nunca se obtendrá una regresión tan bien definida como el caso lineal de la fig. 11.7B, o el curvilíneo de la fig. 11.7C, ni necesariamente se tendrá una heterogeneidad del tipo

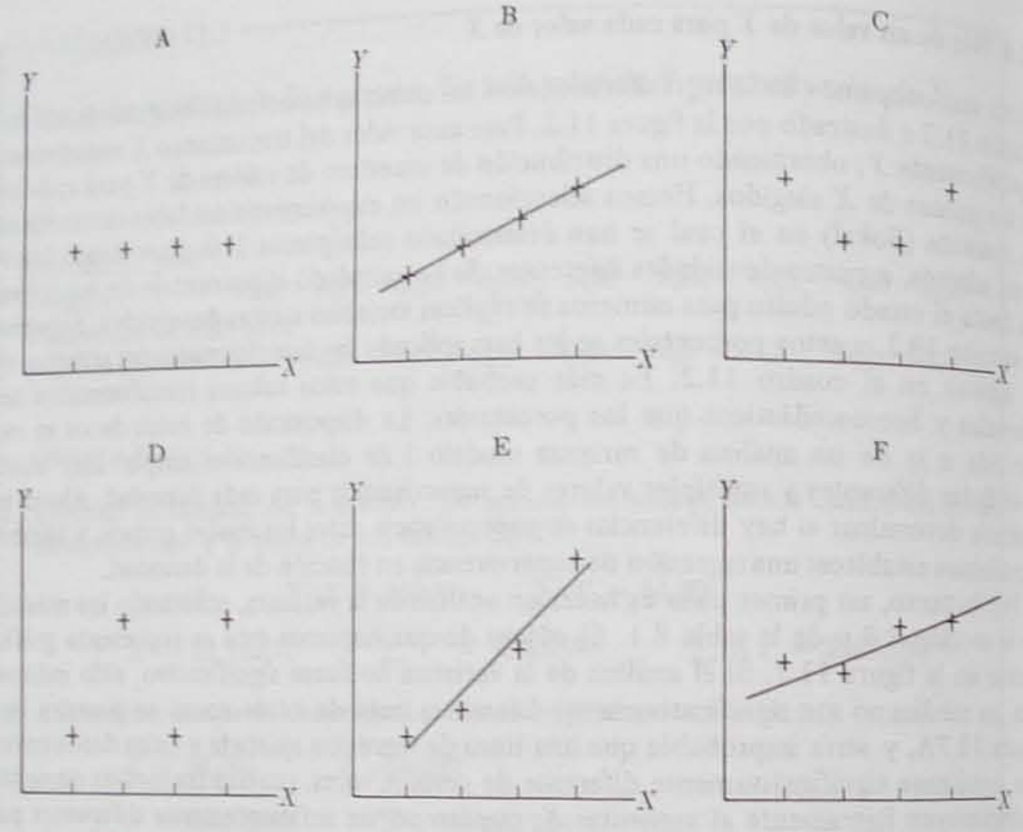


Fig. 11.7. Diferencias entre medias y regresión lineal. En estas figuras solamente se indican las inclinaciones generales. La significación de cualquiera de éstas dependería de los resultados de las pruebas apropiadas. (Para más explicación véase el texto.)

mostrado en la fig. 11.7D, en el que cualquier línea recta ajustada a los datos sería horizontal y a la mitad de distancia entre los puntos superior e inferior. Es más probable que tengamos datos en los que pueda demostrarse regresión lineal, pero que no se ajusten perfectamente a una línea recta. Las desviaciones residuales de las medias en torno a la regresión lineal podrían ser eliminadas cambiando de regresión lineal a curvilínea (como se sugiere por el patrón de puntos de la figura 11.7E) o quedar como heterogeneidad residual inexplicable en torno a la línea de regresión, como se indica en la figura 11.7F.

Realizamos los cálculos siguiendo el por ahora familiar esquema del análisis de la varianza y obtenemos la tabla presentada en el cuadro 11.2. Los tres grados de libertad entre los cuatro grupos dan una media cuadrática que es altamente significativa cuando se contrasta con la media cuadrática intragrupos. Por lo tanto, está claro que vale la pena continuar el análisis para ver si existe una regresión significativa de supervivencia en función de la densidad. Los pasos adicionales para el análisis de regresión se siguen en el cuadro 11.2. Calculamos la suma de cuadrados de X, la suma de productos de X e Y, la suma de cuadrados explicable de Y, y la suma de cuadrados inexplicable de Y. Las fórmulas no nos resultarán familiares debido a la complicación de los distintos valores de

Y para cada valor de X. Los cálculos de las sumas de cuadrados de X incluyen la multiplicación de X por el número de ítems del estudio. Así, aunque pueda parecer que hay solamente cuatro densidades, en realidad hay tantas densidades (aunque de cuatro magnitudes solamente) como valores de Y en el estudio. Una vez completados los cálculos, presentamos nuevamente los resultados en forma de análisis de la varianza, como se expone en el cuadro 11.2. Obsérvese que las cantidades principales de esta tabla son las mismas que en un análisis de la varianza de clasificación simple, pero ahora tenemos además una suma de cuadrados que representa la regresión lineal, basada siempre en un grado de libertad. Esta suma de cuadrados se resta de la *S.C.* entre grupos, dejando una suma de cuadrados residual (de dos grados de libertad en este caso), que representa las desviaciones de la regresión lineal.

Deberíamos estar completamente seguros de lo que representan estas fuentes de variación. La *S.C.* debida a regresión lineal representa la porción de la *S.C.* entre grupos que se explica por regresión lineal en X. La *S.C.* debida a desviaciones de la regresión representa la variación residual o dispersión en torno a la línea de regresión, como se hace patente en los diversos ejemplos de la figura 11.7. La *S.C.* intragrupo es una medida de la variación de los ítems en torno a cada media de grupo.

CUADRO 11.2

Cálculo de regresión con más de un valor de Y por valor de X.

Las variantes Y son transformaciones arcseno del porcentaje de supervivencia del coleóptero *Tribolium castaneum* a 4 densidades (X = número de huevos por gramo de harina).

	Densidad = X ($a = 4$)			
	5/g	*20/g*	*50/g*	*100/g*
Supervivencia; en grados	61,68 58,37 69,30 61,68 69,30	68,21 66,72 63,44 60,84	58,69 58,37 58,37	53,13 49,89 49,82
$\sum^{n_i} Y$	320,33	259,21	175,43	152,84
n_i	5	4	3	3
$\bar{Y}_i$	64,07	64,80	58,48	50,95

$\sum^{a} n_i = 15$

$\sum\sum^{n_i} Y = 907{,}81$

Fuente: Datos de Sokal (1967).

Los cálculos del análisis de la varianza se realizan como en la tabla 8.1.

CUADRO 11.2 (continuación)

Tabla de análisis de la varianza

Origen de la variación	*g.l.*	*S.C.*	*M.C.*	F_s
$\bar{Y} - \bar{\bar{Y}}$ Entre grupos	3	423,7016	141,2339	11,20**
$Y - \bar{Y}$ Intragrupos	11	138,6867	12,6079	
$Y - \bar{\bar{Y}}$ Total	14	562,3883		

Los grupos difieren significativamente.

Pasamos a probar si las diferencias entre los valores de supervivencia pueden justificarse por regresión lineal en función de la densidad. Si se han encontrado diferencias no significativas entre grupos, sería posible, aunque improbable, que la regresión lineal fuese significativa. Si $S.C._{\text{grupos}} < (M.C._{\text{intra}} \times F_{[1,\,\sum^{a} n_i - a]})$, es imposible que la regresión sea significativa.

Cálculos para el análisis de la regresión

1. Suma de X multiplicado por el tamaño de muestra $= \sum^{a} n_i X$
$$= 5(5) + 4(20) + 3(50) + 3(100) = 555$$

2. Suma de X^2 multiplicado por el tamaño de muestra $= \sum^{a} n_i X^2$
$$= 5(5)^2 + 4(20)^2 + 3(50)^2 + 3(100)^2 = 39\,225$$

3. Suma de productos de X e Y multiplicada por el tamaño de muestra $= \sum^{a} n_i X \bar{Y}$
$$= \sum^{a} X\left(\sum^{n_i} Y\right) = 5(320{,}33) + \cdots + 100(152{,}84) = 30\,841{,}35$$

4. Término de corrección para $X = TC_X = \dfrac{\left(\sum^{a} n_i X\right)^2}{\sum^{a} n_i}$
$$= \frac{(\text{cantidad } 1)^2}{\sum^{a} n_i} = \frac{(555)^2}{15} = 20\,535{,}00$$

5. Suma de cuadrados de $X = \sum x^2 = \sum^{a} n_i X^2 - CT_X$
$$= \text{cantidad } 2 - \text{cantidad } 4 = 39\,225 - 20\,535 = 18\,690$$

6. Suma de productos $= \sum xy$
$$= \sum^{n} X\left(\sum^{n_i} Y\right) - \frac{\left(\sum^{a} n_i X\right)\left(\sum^{a}\sum^{n_i} Y\right)}{\sum^{a} n_i}$$
$$= \text{cantidad } 3 - \frac{\text{cantidad } 1 \times \sum^{a}\sum^{n_i} Y}{\sum^{a} n_i}$$
$$= 30\,841{,}35 - \frac{(555)(907{,}81)}{15} = -2747{,}62$$

CUADRO 11.2 (continuación)

7. Suma de cuadrados explicable $= \sum \hat{y}^2 = \dfrac{(\sum xy)^2}{\sum x^2}$
$$= \frac{(\text{cantidad } 6)^2}{\text{cantidad } 5} = \frac{(-2747{,}62)^2}{18\,690} = 403{,}9281$$

8. Suma de cuadrados inexplicables $= \sum d^2_{Y \cdot X} = S.C._{\text{grupos}} - \sum \hat{y}^2$
$$= S.C._{\text{grupos}} - \text{cantidad } 7 = 423{,}7016 - 403{,}9281 = 19{,}7735$$

Tabla de análisis de la varianza, completada con regresión

Origen de la variación	*g.l.*	*S.C.*	*M.C.*	F_s
$\bar{Y} - \bar{\bar{Y}}$ Entre densidades (grupos)	3	423,7016	141,2339	11,20**
$\hat{Y} - \bar{\bar{Y}}$ Regresión lineal	1	403,9281	403,9281	40,86*
$\bar{Y} - \hat{Y}$ Desviaciones de la regresión	2	19,7735	9,8868	<1 *ns*
$Y - \bar{Y}$ Intragrupos	11	138,6867	12,6079	
$Y - \bar{\bar{Y}}$ Total	14	562,3883		

Además de las medias cuadráticas familiares, $M.C._{\text{grupos}}$ y $M.C._{\text{intra}}$, ahora tenemos la media cuadrática debida a la regresión lineal, $M.C.\hat{y}$, y la media cuadrática para desviaciones de la regresión $M.C._{Y \cdot X}$ $(= s^2_{Y \cdot X})$. Para probar si las desviaciones de la regresión lineal son significativas, se compara la razón $F_s = M.C._{Y \cdot X}/M.C._{\text{intra}}$ con $F_{\alpha[a-2,\,\sum^{a} n_i - a]}$. Como encontramos $F_s < 1$, aceptamos la hipótesis nula de que las desviaciones de la regresión lineal son cero.

Para demostrar la presencia de regresión lineal, contrastamos pues la $M.C.\hat{y}$ con la media cuadrática de las desviaciones de la regresión $s^2_{Y \cdot X}$, y como $F_s = 403{,}9281/9{,}8868 = 40{,}86$ es mayor que $F_{0{,}05[1,2]} = 18{,}5$, rechazamos sin duda la hipótesis nula de que no hay regresión, o que $\beta = 0$.

9. Coeficiente de regresión (pendiente de la línea de regresión) $= b_{Y \cdot X} = \dfrac{\sum xy}{\sum x^2}$
$$= \frac{\text{cantidad } 6}{\text{cantidad } 5} = \frac{-2747{,}62}{18\,690} = -0{,}14701$$

10. Ordenada en el origen $Y = a = \bar{\bar{Y}} - b_{Y \cdot X}\bar{X}$
$$= \frac{\sum^{a}\sum^{n_i} Y}{\sum^{a} n_i} - \frac{\text{cantidad } 9 \times \text{cantidad } 1}{\sum^{a} n_i}$$
$$= \frac{907{,}81}{15} - \frac{(-0{,}14701)555}{15} = 60{,}5207 + 5{,}4394 = 65{,}9601$$

Por lo tanto, la ecuación de regresión es $\hat{Y} = 65{,}9601 - 0{,}14701X$

En primer lugar probamos si la media cuadrática para las desviaciones de la regresión ($M.C._{Y\cdot X} = s^2_{Y\cdot X}$) es significativa calculando la razón de varianzas de $M.C._{Y\cdot X}$ sobre la *M.C.* intragrupos. En nuestro caso las desviaciones de la regresión son sin duda no significativas, ya que la media cuadrática de las desviaciones es menor que la intragrupos. Ahora probamos la media cuadrática de la regresión, $M.C._{\hat{Y}}$, sobre la media cuadrática de las desviaciones de la regresión y encontramos que es significativa. Así pues la regresión lineal sobre densidad, sin duda ha eliminado una parte significativa de la variación de los valores de supervivencia. La significación de la media cuadrática de las desviaciones de la regresión podría significar, o bien que Y es una función curvilínea de X o que hay una gran cantidad de heterogeneidad aleatoria en torno a la línea de regresión (como ya se ha discutido en relación con la figura 11.7; realmente puede predominar una mezcla de ambas condiciones).

Completamos el cálculo del coeficiente de regresión y la ecuación de regresión como se muestra al final del cuadro 11.2. Nuestras conclusiones son que al aumentar la densidad disminuye la supervivencia, y que esta relación puede expresarse por una regresión lineal significativa de la forma $\hat{Y} = 65{,}9601 - 0{,}14701\,X$, en donde X es la densidad por gramo e Y es la transformación arcseno del porcentaje de supervivencia. Esta relación se representa gráficamente en la figura 11.8. Las sumas de productos y pendientes de regresión de los dos ejemplos discutidos hasta ahora han sido negativas y se pudiera creer que esto siempre sucede así. Sin embargo, es solamente un accidente de elección de estos dos ejemplos. En los ejercicios de práctica se encontrará un coeficiente de regresión positivo.

Cuando tenemos iguales tamaños de muestra de valores de Y para cada valor de X, los cálculos resultan más sencillos. Primero hacemos el análisis de la varianza del mismo modo que en el cuadro 8.1. Desde el paso 1 hasta el 8 del cuadro 11.2 se simplifican debido a que los distintos tamaños de muestreo n_i están reemplazados por un tamaño de muestreo constante n, que generalmente puede sacarse como factor común de las diversas expresiones. Además, $\sum_a^a n_i = an$. Las pruebas de significación aplicadas a estos casos también se simplifican.

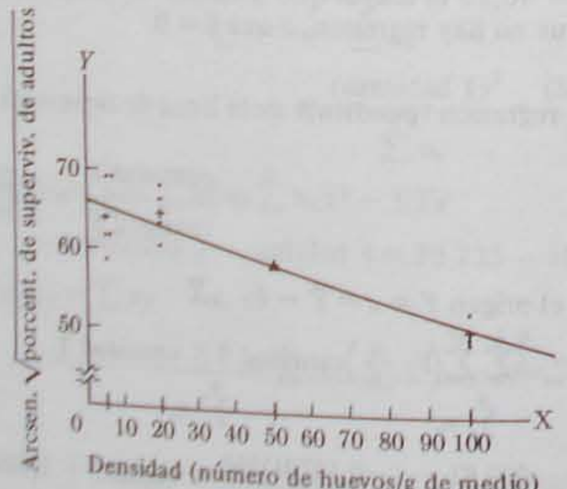


Fig. 11.8. Regresión lineal ajustada a los datos del cuadro 11.2. Las medias de muestreo se identifican por signos +.

11.5 Pruebas de significación en regresión

Hasta ahora hemos interpretado la regresión como un método para proporcionar una estimación $\hat{Y}_1$, dado un valor de X_1. Otra interpretación es como un método para explicar parte de la variación de la variable dependiente Y en términos de la variación de la variable independiente X. La *S.C.* de una muestra de valores de Y, $\sum y^2$, se calcula sumando y elevando al cuadrado las desviaciones $y = Y - \bar{Y}$. En la figura 11.9 podemos ver que la desviación y puede descomponerse en dos partes, $\hat{y}$ y $d_{Y\cdot X}$. También se ve claro en esta figura que la desviación $\hat{y} = \hat{Y} - \bar{Y}$ representa la desviación del valor estimado $\hat{Y}$ respecto de la media de Y La altura de $\hat{y}$ es sin duda una función de x. Ya hemos visto que $\hat{y} = bx$ [expresión (11.3)]. En geometría analítica ésta se denomina la forma punto-pendiente de la ecuación. Si la pendiente de la línea de regresión, b, fuese más elevada, $\hat{y}$ sería relativamente mayor para un valor determinado de x. La parte restante de la desviación y es $d_{Y\cdot X}$. Representa la variación residual de la variable Y después de haberse restado la variación explicable. Podemos ver que $y = \hat{y} + d_{Y\cdot X}$ escribiendo estas desviaciones explícitamente como $Y - \bar{Y} = (\hat{Y} - \bar{Y}) + (Y - \hat{Y})$.

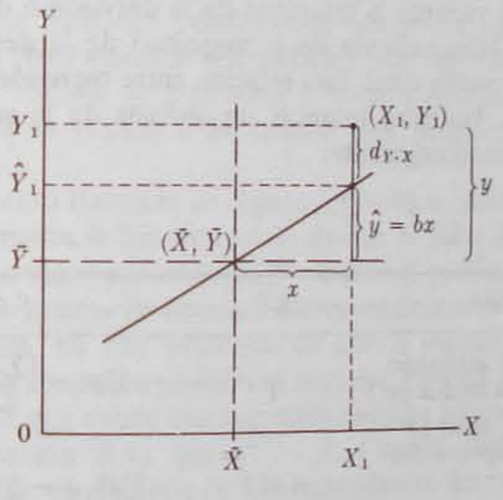


Fig. 11.9. Diagrama esquemático para mostrar las relaciones implicadas en la descomposición de las sumas de cuadrados de la variable dependiente.

Para cada una de estas desviaciones podemos calcular una suma de cuadrados correspondiente. El apéndice A1.7 da la fórmula para calcular la suma de cuadrados inexplicable,

$$\sum d^2_{Y\cdot X} = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2}$$

Trasponiendo términos, ésta da

$$\sum y^2 = \frac{(\sum xy)^2}{\sum x^2} + \sum d^2_{Y \cdot X}$$

Naturalmente, $\sum y^2$ corresponde a y, $\sum d^2_{Y \cdot X}$ a $d_{Y \cdot X}$, y

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

a $\hat{y}$ (como se ha visto en la sección anterior). De este modo podemos descomponer la suma de cuadrados de la variable dependiente en la regresión de un modo análogo a la descomposición de la *S.C.* total en el análisis de la varianza. Nos podemos preguntar cómo la relación aditiva de las desviaciones puede corresponder a una relación aditiva de sus cuadrados, sin la presencia de ningún doble producto. En el apéndice A1.8 una parte de álgebra sencilla mostrará que los dobles productos se anulan. La magnitud de la desviación inexplicable $d_{Y \cdot X}$ es independiente de la magnitud de la desviación explicable $\hat{y}$, así como en el análisis de la varianza la magnitud de la desviación de un ítem respecto de la media de muestreo es independiente de la magnitud de la desviación de la media de muestreo respecto de la media total. Esta relación entre regresión y análisis de la varianza puede llevarse más allá. Podemos intentar un análisis de la varianza de las sumas de cuadrados separadas del modo siguiente:

Fuente de variación	*g.l.*	*S.C.*	*M.C.*
$\hat{Y} - \bar{Y}$ Explicable (Y estimado respecto de la media de Y)	1	$\sum \hat{y}^2 = \frac{(\sum xy)^2}{\sum x^2}$	$s^2_{\hat{Y}}$
$Y - \hat{Y}$ Inexplicable, error (Y observado respecto del Y estimado)	$n - 2$	$\sum d^2_{Y \cdot X} = \sum y^2 - \sum \hat{y}^2$	$s^2_{Y \cdot X}$
$Y - \bar{Y}$ Total (Y observado respecto de la media de Y)	$n - 1$	$\sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$	s^2_Y

La *media cuadrática explicable*, o *media cuadrática debida a la regresión lineal*, mide la cantidad de variación en Y explicada por variación de X. Se contrasta con la *media cuadrática inexplicable*, que mide la variación residual, y se utiliza como una *M.C.* de los errores. La media cuadrática debida a regresión lineal, $s^2_{\hat{Y}}$, está basada en un grado de libertad, y por consiguiente quedan $n - 2$ *g.l.* para la *M.C.* de los errores, ya que la suma de cuadrados total posee $n - 1$ grados de libertad. La prueba es de la hipótesis nula $H_0 : \beta = 0$. Cuando realizamos este análisis de la varianza con los datos de pérdida de peso del cuadro 11.1, obtenemos los resultados siguientes:

Fuente de variación	*g.l.*	*S.C.*	*M.C.*	F_s
Explicable –debida a regresión lineal–	1	23,5145	23,5145	267,18**
Inexplicable –error respecto de la línea de regresión	7	0,6161	0,08801	
Total	8	24,1306		

La prueba de significación es $F_s = s^2_{\hat{Y}} / s^2_{\hat{Y} \cdot X}$. Según el valor observado de F_s está claro que una porción grande y significativa de la varianza de Y se explica por regresión sobre X.

Pasamos ahora a los errores estándar para diversos estadísticos de regresión, su utilización en pruebas de hipótesis y el cálculo de límites de confianza. El cuadro 11.3 registra estos errores estándar en dos columnas. La columna de la derecha es para el caso de un solo valor de Y para cada valor de X. La primera fila de la tabla considera el *error estándar del coeficiente de regresión*, que es simplemente la raíz cuadrada del cociente de la varianza inexplicable, dividida por la suma de cuadrados de X. Obsérvese que la varianza inexplicable $s^2_{Y \cdot X}$ es una cantidad fundamental que forma parte de todos los errores estándar en regresión. El error estándar del coeficiente de regresión nos permite contrastar diversas hipótesis y dar límites de confianza a nuestro estimador de muestreo de b. El cálculo de s_b se presenta en el paso 1 del cuadro 11.4, utilizando el ejemplo de pérdida de peso del cuadro 11.1.

La prueba de significación ilustrada en el paso 2, prueba la *"significación" del coeficiente de regresión;* es decir, prueba la hipótesis nula de que el valor de muestreo de b procede de una población con un valor paramétrico $\beta = 0$ para el coeficiente de regresión. Esta es una prueba t, siendo los grados de libertad correspondientes $n - 2 = 7$. Si no podemos rechazar la hipótesis nula, no hay evidencia de que la regresión sea significativamente desviante de cero ni en la dirección positiva ni negativa. Para los datos de pérdida de peso, nuestras conclusiones son que existe una regresión negativa altamente significativa. Hemos visto anteriormente (sección 8.4) que $t^2 = F$. Cuando elevamos al cuadrado $t_s = -16{,}345$ del cuadro 11.4, obtenemos 267,16, que (dentro del error de redondeo) corresponde al valor de F_s encontrado en el análisis de la varianza anterior en esta sección. La prueba de significación del paso 2 del cuadro 11.4 podría naturalmente haberse utilizado también para probar si b es significativamente diferente de un valor paramétrico β distinto de cero.

La fijación de límites de confianza para el coeficiente de regresión presenta característicamente conocidas. El cálculo se expone en el paso 3 del cuadro 11.4. En vista de la pequeña magnitud de s_b, el intervalo de confianza es bastante reducido. Los límites de confianza se muestran en la figura 11.10 como líneas de puntos que representan los límites de la pendiente al 95 %. Obsérvese que tanto la línea de regresión como sus límites de confianza pasan por las medias de X e Y. Por tanto, la variaciónen b hace girar la línea de regresión sobre el punto $\bar{X}$, $\bar{Y}$.

A continuación, calculamos un *error estándar para la media de muestreo observada* $\bar{Y}$.

CUADRO 11.3

Errores estándar de estadísticos de regresión y sus grados de libertad.

Para la explicación de este cuadro, véase la sección 11.5: ν identifica grados de libertad; a = número de valores de X cuando hay n_i valores de Y para cada X; n = tamaño de muestreo cuando hay un solo valor de Y para cada valor de X.

Estadístico	s	*Más de un valor de Y para cada valor de X*		*Un solo valor de Y para cada valor de X*	
$b_{Y \cdot X}$ (coeficiente de regresión)	s_b	$\sqrt{\frac{s^2_{Y \cdot X}}{\Sigma x^2}}$	$\nu = a - 2$	$\sqrt{\frac{s^2_{Y \cdot X}}{\Sigma x^2}}$	$\nu = n - 2$
		Para cualquier valor X_i		Para $\overline{X}$	
$\overline{Y}_i$ (media de muestreo)	$s_{\overline{Y}}$	$\sqrt{\frac{M.C._{\text{intra}}}{n_i}}$	$\nu = \Sigma n_i - a$	$\sqrt{\frac{s^2_{Y \cdot X}}{n}}$	$\nu = n - 2$
$\hat{Y}_i$ (Y estimado para un valor dado X_i)	$s_{\hat{Y}}$	$\sqrt{s^2_{Y \cdot X}\left[\frac{1}{\Sigma n_i} + \frac{(X_i - \overline{X})^2}{\Sigma x^2}\right]}$	$\nu = a - 2$ ($\nu = \Sigma n_i - 2$ si se utiliza $s^2_{Y \cdot X}$ ponderada)	$\sqrt{s^2_{Y \cdot X}\left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\Sigma x^2}\right]}$	$\nu = n - 2$

CUADRO 11.4

Prueba de significación y cálculo de límites de confianza de estadísticos de regresión. Un solo valor de Y para cada valor de X.

Basado en los errores estándar y grados de libertad del cuadro 11.3; utilizando el ejemplo del cuadro 11.1.

$n = 9 \qquad \overline{X} = 50{,}389 \qquad \overline{Y} = 6{,}022$

$b_{Y \cdot X} = -0{,}05322 \qquad \sum x^2 = 8301{,}3889$

$$s^2_{Y \cdot X} = \frac{\sum d^2_{Y \cdot X}}{(n-2)} = \frac{0{,}6161}{7} = 0{,}08801$$

1. Error estándar del coeficiente de regresión

$$s_b = \sqrt{\frac{s^2_{Y \cdot X}}{\sum x^2}} = \sqrt{\frac{0{,}08801}{8301{,}3889}} = \sqrt{0{,}000\,010\,602} = 0{,}003\,2561$$

2. Prueba de significación del coeficiente de regresión

$$t_s = \frac{(b - 0)}{s_b} = \frac{-0{,}05322}{0{,}003\,2561} = -16{,}345$$

$$t_{0{,}001[7]} = 5{,}408 \qquad P < 0{,}001$$

3. Límites de confianza al 95 % para el coeficiente de regresión

$$t_{0{,}05[7]} s_b = 2{,}365(0{,}003\,2561) = 0{,}00770$$

$$L_1 = b - t_{0{,}05[7]} s_b = -0{,}05322 - 0{,}00770 = -0{,}06092$$

$$L_2 = b + t_{0{,}05[7]} s_b = -0{,}05322 + 0{,}00770 = -0{,}04552$$

4. Error estándar de la media muestreada $\overline{Y}$ (para $\overline{X}$)

$$s_{\overline{Y}} = \sqrt{\frac{s^2_{Y \cdot X}}{n}} = \sqrt{\frac{0{,}08801}{9}} = 0{,}098\,883$$

5. Límites de confianza al 95 % para la media μ_Y correspondientes a X (Y = 6,022)

$$t_{0{,}05[7]} s_{\overline{Y}} = 2{,}365(0{,}098\,8883) = 0{,}233871$$

$$L_1 = \overline{Y} - t_{0{,}05[7]} s_{\overline{Y}} = 6{,}022 - 0{,}2339 = 5{,}7881$$

$$L_2 = \overline{Y} + t_{0{,}05[7]} s_{\overline{Y}} = 6{,}022 + 0{,}2339 = 6{,}2559$$

CUADRO 11.4 (continuación)

6. Error estándar de $\hat{Y}$, un Y estimado para un valor dado de X_i

$$s_{\hat{Y}} = \sqrt{s^2_{Y \cdot X}\left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum x^2}\right]}$$

por ejemplo, para $X_i = 100$ % de humedad relativa,

$$s_{\hat{Y}} = \sqrt{0{,}08801\left[\frac{1}{9} + \frac{(100 - 50{,}389)^2}{8301{,}3889}\right]}$$

$$= \sqrt{0{,}08801[0{,}40760]} = \sqrt{0{,}035\,873} = 0{,}18940$$

7. Límites de confianza del 95 % para μ_{Y_i} correspondientes a la estimación $\hat{Y}_i = 3{,}3817$ para $X_i = 100$ % de humedad relativa.

$$t_{0{,}05[7]}s_{\hat{Y}} = 2{,}365(0{,}18940) = 0{,}44793$$

$$L_1 = \hat{Y}_i - t_{0{,}05[7]}s_{\hat{Y}} = 3{,}3817 - 0{,}4479 = 2{,}9338$$

$$L_2 = \hat{Y}_i + t_{0{,}05[7]}s_{\hat{Y}} = 3{,}3817 + 0{,}4479 = 3{,}8296$$

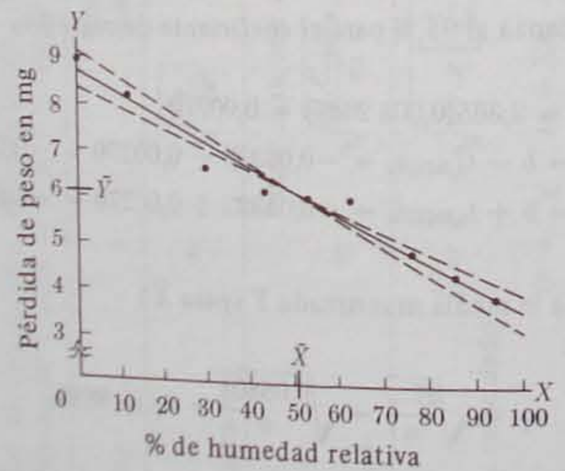


Fig. 11.10. Límites de confianza del 95 % para la línea de regresión de la figura 11.6.

De la sección 6.1 se recordará que $s^2_{\bar{Y}} = s^2_Y/n$. Sin embargo, ahora que hemos hecho la regresión de Y sobre X, podemos explicar una parte de la variación de Y en términos de la

variación de X. La varianza de Y en torno del punto $\bar{X}$, $\bar{Y}$ en la línea de regresión es menor que $s^2_{\bar{Y}}$; es $s^2_{Y \cdot X}$. Por lo tanto, para $\bar{X}$ podemos calcular los límites de confianza de $\bar{Y}$, utilizando como error estándar de la media $s_{\bar{Y}} = \sqrt{s^2_{Y \cdot X}/n}$ con $n - 2$ grados de libertad. Este error estándar se calcula en el paso 4 del cuadro 11.4, y los límites de confianza al 95 % de la media muestreada $\bar{Y}$ para $\bar{X}$ se calculan en el paso 5. Estos límites (5,7881 – 6,2559) son considerablemente más reducidos que los límites de confianza de la media basada en el error estándar convencional $s_{\bar{Y}}$, que serían de 4,687 a 7,357. Sin duda, las diferencias en la humedad relativa explican gran parte de la variación en pérdida de peso.

El error estándar de $\bar{Y}$ es solamente un caso particular del *error estándar para cualquier valor $\hat{Y}$ estimado a lo largo de la línea de regresión.* Un nuevo factor introduce ahora la varianza del error, cuya magnitud es en parte una función de la distancia de un determinado valor X_i respecto de su media $\bar{X}$. Así, cuanto más alejado esté X_i de su media, mayor será el error de la estimación. Este factor se observa en la tercera fila del cuadro 11.3 como la desviación $X_i - \bar{X}$, elevada al cuadrado y dividida por la suma de cuadrados de X. El error estándar de una estimación $\hat{Y}_i$ para una humedad relativa $X_i = 100$ %, se presenta en el paso 6 del cuadro 11.4. Los límites de confianza del 95 % para $\mu_{\hat{Y}_i}$, el valor paramétrico correspondiente a la estimación $\hat{Y}_i$, se exponen en el paso 7 de dicho cuadro. Obsérvese que la amplitud del intervalo de confianza es 3,8296 – 2,9338 = 0,8958, considerablemente mayor que el intervalo de confianza para $\bar{X}$ calculado en el paso 5, el cual era 6,2559 – 5,7881 = 0,4678. Esto demuestra el hecho de que los límites de confianza están más separados fuera de la media que en la media. Si calculamos una serie de límites de confianza para diferentes valores de X_i, obtenemos una franja de confianza bicóncava como se muestra en la figura 11.11. Cuanto más lejos estemos de la media, menos fiables son nuestras estimaciones de Y debido a la inseguridad acerca de la verdadera pendiente, β, de la línea de regresión.

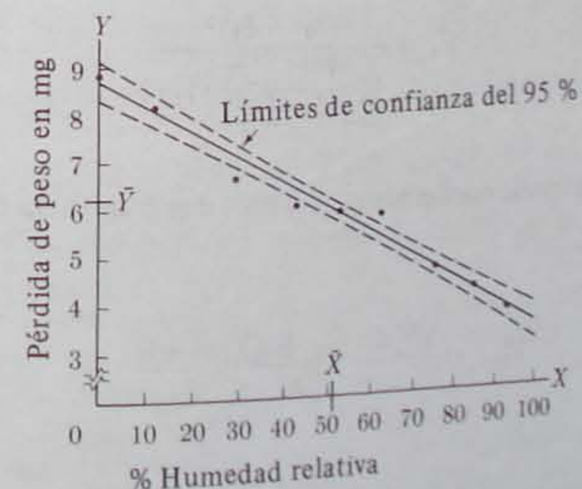


Fig. 11.11. Límites de confianza del 95 % para estimaciones de regresión en los datos de la figura 11.6.

Además, las regresiones lineales que ajustamos, con frecuencia son solamente aproximaciones groseras a las relaciones funcionales más complicadas entre variables biológicas. Es muy frecuente que haya una relación aproximadamente lineal a lo largo de un cierto rango de la variable independiente, después de cuyo rango la pendiente cambia rápidamente. Por ejemplo, el latido cardíaco de un animal poiquilotérmico será directamente proporcional a la temperatura sobre un rango de temperaturas tolerables, pero por debajo y por encima de este rango el latido disminuirá con el tiempo al enfriarse el animal o sufrir depresión térmica. Por tanto, el sentido común indica que se debería ser muy prudente al extrapolar en una ecuación de regresión si se tiene alguna duda sobre la linealidad de la relación.

Los límites de confianza para α, el valor paramétrico de a, son un caso especial de los de $\mu_{\hat{Y}_i}$ para $X_i = 0$.

Las pruebas de significación en los análisis de regresión en que haya más de una variante Y por valor de X se realizan de una manera similar a los del cuadro 11.4, salvo que se utilizan los errores estándar de la columna de la izquierda del cuadro 11.3.

Otra prueba de significación en regresión es una prueba de las diferencias entre dos líneas de regresión. ¿Por qué nos interesaría contrastar diferencias entre pendientes de regresión? Podríamos encontrar que diferentes tóxicos dan diferentes curvas dosis-mortalidad o que diferentes drogas dan diferentes relaciones entre dosis y respuesta (véase, por ejemplo, figura 11.1). O cultivos que difieren genéticamente podrían dar diferentes respuestas al aumentar la densidad, un hecho importante para comprender el efecto de la selección natural en estos cultivos. La pendiente de regresión de una variable sobre otra es tan fundamental como estadístico de una muestra, como la media o la desviación típica, y al comparar muestras puede ser tan importante comparar coeficientes de regresión como lo es comparar estos otros estadísticos.

La prueba de la diferencia entre dos coeficientes de regresión puede realizarse como una prueba F. Calculamos

$$F_s = \frac{(b_1 - b_2)^2}{\dfrac{\sum x_1^2 + \sum x_2^2}{(\sum x_1^2)(\sum x_2^2)}\,\bar{s}^2_{Y \cdot X}}$$

donde $\bar{s}^2_{Y \cdot X}$ es la media ponderada $s^2_{Y \cdot X}$ de los dos grupos. Su fórmula es

$$\bar{s}^2_{Y \cdot X} = \frac{(\sum d^2_{Y \cdot X})_1 + (\sum d^2_{Y \cdot X})_2}{\nu_2}$$

Para un solo Y por valor de X, $\nu_2 = n_1 + n_2 - 4$, pero cuando hay más de una variante Y por valor de X, $\nu_2 = a_1 + a_2 - 4$. Comparar F_s con $F_{\alpha[1,\nu_2]}$. Como hay sólo un grado de libertad en el numerador, $t_s = \sqrt{F_s}$.

11.6 Las aplicaciones de la regresión

Hemos estado tan ocupados aprendiendo la mecánica del análisis de regresión, que no hemos tenido tiempo para pensar en las diversas aplicaciones de la regresión. En esta sección consideraremos cuatro aplicaciones más o menos distintas. Se tratan todas en términos de regresión modelo I.

En primer lugar podríamos mencionar el *estudio de causalidad*. Si queremos saber si la variación en una variable Y es causada por cambios en otra variable X, manipulamos X en un experimento y vemos si podemos obtener una regresión significativa de Y sobre X. La idea de causalidad es una idea compleja, filosófica, que no trataremos aquí. Indudablemente se ha llamado la atención desde la más temprana experiencia científica en no confundir variación concomitante con causalidad. Las variables pueden variar simultáneamente, pero esta covariación puede ser accidental o ambas pueden ser funciones de una causa común que las afecta. Los últimos casos son ordinariamente regresión modelo II con las dos variables variando libremente. Cuando manipulamos una variable y descubrimos que estas manipulaciones afectan a una segunda variable, generalmente estamos seguros de que la variación de la variable independiente X es la causa de la variación de la variable dependiente Y (no la causa de la variable). No obstante, incluso en este punto es mejor ser precavido. Cuando hallamos que la velocidad del latido cardíaco en un animal de sangre fría es una función de la temperatura ambiente, podemos concluir que la temperatura es una de las causas de las diferencias en frecuencia cardíaca. Puede haber también otros factores que afecten a la frecuencia cardíaca. Un posible error es invertir la relación causa-efecto. Es improbable que alguien suponga que la frecuencia cardíaca afecta a la temperatura del ambiente general, pero podríamos estar equivocados sobre las relaciones causa-efecto entre dos sustancias químicas en la sangre, por ejemplo. A pesar de estas precauciones, el análisis de regresión es un mecanismo habitualmente utilizado para exponer relaciones causales. Mientras que una regresión significativa de Y en función de X no demuestra que los cambios en X sean la causa de las variaciones de Y, la afirmación contraria es cierta. Cuando encontramos regresión no significativa de Y en función de X, podemos, en todos los casos excepto los más complejos, inferir con bastante seguridad (teniendo en cuenta la posibilidad de error de tipo II) que las desviaciones de X no afectan a Y.

La *descripción de leyes* y *predicciones científicas* es una segunda área general de aplicación del análisis de regresión. Las ciencias naturales apuntan hacia una descripción matemática de relaciones entre variables en la naturaleza, y el análisis de regresión modelo I nos permite estimar las relaciones funcionales entre variables, una de las cuales está sujeta a error. Estas relaciones funcionales no siempre tienen sentido biológico claramente interpretable. Así, en muchos casos puede resultar difícil asignar una interpretación biológica a los estadísticos a y b, o a sus parámetros correspondientes α y β. Cuando podemos hacer esto, hablamos de un modelo *matemático estructural*, una de cuyas partes componentes tiene claro significado científico. Sin embargo, las curvas matemáticas que no son modelos estructurales también son de importancia en la ciencia. La mayoría de las líneas de regresión son *curvas empíricamente ajustadas*, en las que la función representa simplemente el mejor ajuste matemático (por un criterio tal como el de mínimos cuadrados) a una serie de datos observados.

La *comparación de variantes dependientes* es otra aplicación de la regresión. Tan pronto como se establezca que una determinada variable es función de otra, como en el cuadro 11.2 donde encontramos que la supervivencia de los coleópteros es una función de la densidad, uno está obligado a preguntarse hasta qué punto cualquier diferencia en supervivencia observada entre dos muestras de coleópteros es función de la densidad en la que se han desarrollado. Sería injusto comparar coleópteros desarrollados en densidad muyalta (y se espera que tengan baja supervivencia) con los criados bajo condiciones óptimas de baja densidad. Este es el mismo punto de vista que nos quita la idea de comparar los conocimientos matemáticos de un alumno de quinto grado con los de un estudiante universitario. Puesto que sin duda podríamos obtener una regresión de conocimientos matemáticos sobre años de enseñanza en matemáticas, deberíamos estar comparando hasta qué punto un determinado individuo se desvía de su valor esperado basado en esta regresión. Así, con respecto a sus condiscípulos y grupo de edad, el de quinto grado puede estar mucho mejor que el estudiante universitario con respecto a su grupo de compañeros. Esto sugiere que calculemos *valores de Y ajustados* que tengan en cuenta la magnitud de la variable independiente X. Una manera convencional de calcular estos valores de Y ajustados es como la media total de la población $\bar{Y}$ más la desviación $d_{Y \cdot X} = (Y_i - \hat{Y}_i)$ del valor dado de Y, Y_i, respecto de la línea de regresión. Como ya se sabe, estas desviaciones pueden ser positivas o negativas; por lo tanto, los valores de Y ajustados pueden estar por encima o por debajo de la media. Definiremos formalmente un valor de Y ajustado como

$$Y_{aj} = \bar{Y} + d_{Y \cdot X} = Y - bx$$

El *control estadístico* es una aplicación de la regresión que no es muy conocida entre los biólogos y representa una filosofía científica que no está bien establecida en biología excepto en los círculos agrícolas. Frecuentemente los biólogos categorizan el trabajo como descriptivo o como experimental, con la implicación de que solamente el segundo puede ser analítico. No obstante, los métodos estadísticos aplicados al trabajo descriptivo pueden, en muchos casos, sustituir a las técnicas experimentales muy adecuadamente; a veces son incluso preferibles. Estos métodos tratan de sustituir la manipulación estadística de una variable concomitante por el control de la variable por medios experimentales. Un ejemplo clarificará esta técnica.

Vamos a suponer que estamos estudiando los efectos de varias dietas en la presión sanguínea en ratas. Encontramos que la variabilidad de la presión sanguínea en nuestra población de ratas es considerable, incluso antes de que introduzcamos diferencias en la dieta. Otro estudio revela que la variabilidad se debe en gran parte a diferencias de edad entre las ratas de la población experimental. Esto puede demostrarse por una regresión lineal significativa de la presión sanguínea sobre la edad. Para reducir la variabilidad de la presión sanguínea en la población, deberíamos mantener constante la edad de las ratas. En este punto la reacción de la mayoría de los biólogos será repetir el experimento utilizando solamente ratas de la misma edad; éste es un enfoque válido, de sentido común, que forma parte del método experimental. En algunos casos, cuando no es práctico o es demasiado costoso mantener constante la variable, es mejor un planteamiento alternativo.

Podemos continuar utilizando ratas de edades variables y registrar simplemente la edad de cada rata así como su presión sanguínea. Entonces hacemos la regresión de presión sanguínea sobre edad y utilizamos una media ajustada como presión sanguínea básica para cada individuo. Ahora podemos valorar el efecto de diferencias en la dieta sobre estas medias ajustadas. O bien podemos analizar los efectos de la dieta en las desviaciones inexplicables, $d_{Y \cdot X}$, después de haber hecho la regresión de las presiones sanguíneas experimentales sobre la edad (que equivale a lo mismo).

¿Cuáles son las ventajas de este procedimiento? A menudo será imposible obtener un número suficiente de individuos todos de la misma edad. Haciendo uso de la regresión podemos utilizar todos los individuos de la población. El empleo del control estadístico supone que es relativamente fácil medir la variable independiente X y por supuesto, que esta variable puede medirse sin error, lo que generalmente sería cierto en una variable tal como la edad de un animal de laboratorio. El control estadístico puede ser también preferible porque obtenemos conocimientos adicionales sobre las relaciones entre estas dos variables, lo que no ocurriría si nos limitásemos a un solo grupo de edad.

11.7 Transformaciones en regresión

Al transformar una o las dos variables en la regresión, nos proponemos simplificar una relación curvilínea a lineal. Como un subproducto de este procedimiento,la proporción de la varianza de la variable dependiente explicada por la variable independiente es generalmente incrementada,y la distribución de las desviaciones de los puntos en torno a la línea de regresión tiende a hacerse normal y homoescedástica. En lugar de ajustar una regresión curvilínea complicada a los puntos representados en una escala aritmética, es mucho más conveniente calcular una regresión lineal simple para variantes representadas en una escala transformada. Una prueba general de si la transformación mejora la regresión lineal es representar en papel milimetrado corriente los puntos que van a ser ajustados, así como en otro papel en una escala que se sospeche que mejora la relación. Si la función se rectifica y se reduce la desviación de los puntos en torno a una línea ajustada visualmente, vale la pena la transformación.

Discutiremos brevemente algunas de las transformaciones ordinariamente aplicadas en el análisis de regresión. Las transformaciones raíz cuadrada y arcseno (sección 10.2) no se mencionan, pero también son efectivas en casos de regresión que implican datos apropiados para estas transformaciones.

La *transformación logarítmica* es la más frecuentemente utilizada. Por consiguiente, es aconsejable que cualquiera que esté haciendo trabajo estadístico tenga a mano repuesto de papel semilogarítmico. La mayoría de las veces transformamos la variable dependiente Y. Esta transformación está indicada cuando los cambios de porcentaje en la variable dependiente varían directamente con los cambios en la variable independiente. Esta relación se indica por la ecuación $\hat{Y} = ae^{bX}$, donde a y b son constantes y e es la base del logaritmo natural. Después de la transformación obtenemos $\log \hat{Y} = \log a + b(\log e)X$. En esta expresión $\log e$ es una constante que al multiplicarla por b da un nuevo factor constante b', que es equivalente a un coeficiente de regresión. Igualmente, $\log a$, es una

nueva ordenada en el origen, a'. Podemos, por lo tanto, hacer simplemente la regresión de log Y sobre X para obtener la función $\log \hat{Y} = a' + b'X$, y de esta forma conseguir todas nuestras ecuaciones de predicción e intervalos de confianza. La figura 11.2 muestra un ejemplo de transformación de la variante dependiente en la forma logarítmica, lo que conduce a una considerable rectificación de la curva respuesta.

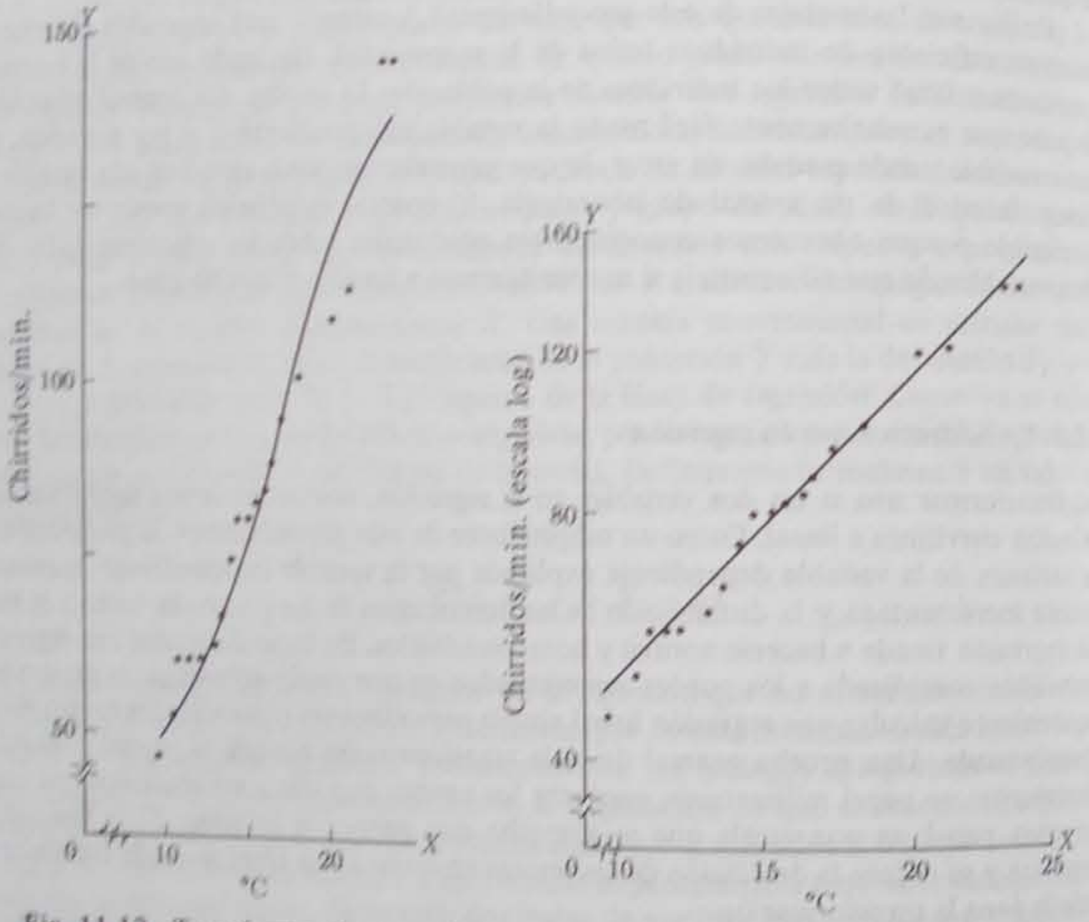


Fig. 11.12. Transformación logarítmica de una variable dependiente en regresión. Chirridos en función de la temperatura en machos del grillo arbóreo *Oecanthus fultoni*. Cada punto representa el número medio de chirridos por minuto para todo el grupo de observaciones a una determinada temperatura en °C. Datos originales en la parte izquierda, Y representada en escala logarítmica en la derecha. (Datos de Block, 1966).

Una transformación logarítmica de la variable independiente en regresión es efectiva cuando cambios proporcionales en la variable independiente producen respuestas lineales en la variable dependiente. Un ejemplo podría ser la caída de peso de un organismo al aumentar la densidad, donde los sucesivos incrementos en densidad tienen que estar en una proporción constante para producir idénticas disminuciones en peso. Éste pertenece a un tipo de fenómenos biológicos bien conocidos, otro ejemplo de los cuales es la ley de

Weber-Fechner en fisiología y sicología, que afirma que un estímulo tiene que ser incrementado en una proporción constante para que produzca un incremento constante en la respuesta. La figura 11.3 ilustra cómo la transformación logarítmica de la variable independiente da por resultado la rectificación de la línea de regresión. Para los cálculos se transformaría X en logaritmos.

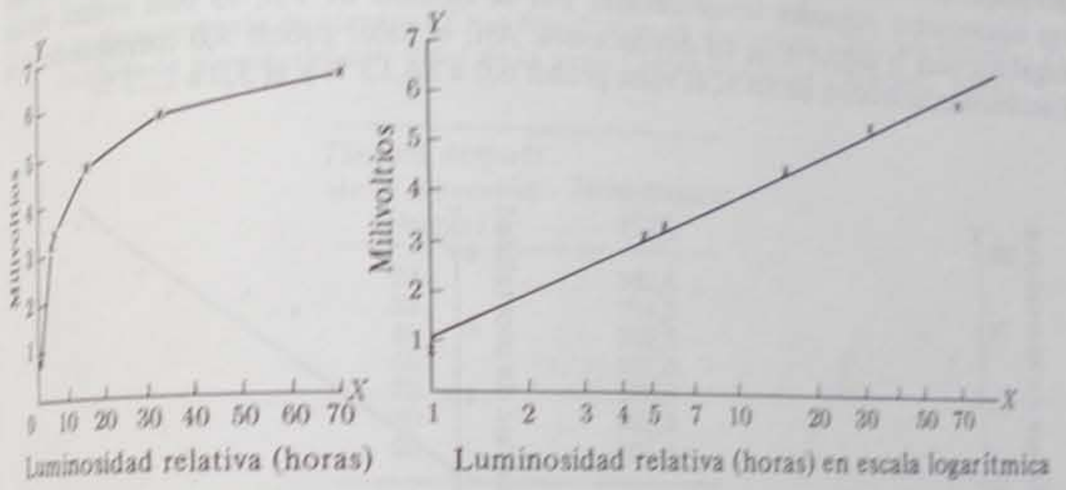


Fig. 11.13. Transformación logarítmica de la variable independiente en regresión. Esto representa magnitud de respuesta eléctrica a la iluminación en el ojo de cefalópodos. En la ordenada milivoltios; en la abscisa luminosidad relativa. Un incremento proporcional en X (luminosidad relativa) produce una respuesta eléctrica lineal Y. (Datos de Fröhlich, 1921).

La transformación logarítmica para las dos variables es aplicable en situaciones en las cuales la relación correcta puede describirse por la fórmula $\hat{Y} = aX^b$. Se vuelve a escribir la ecuación de regresión como $\log \hat{Y} = \log a + b \log X$ y se hace el cálculo de manera convencional. Como ejemplos de esto están el crecimiento enormemente desproporcionado de diversos órganos en algunos organismos, como los tamaños de las astas del ciervo o de las antenas del ciervo volante con respecto a sus tamaños corporales en general. Una transformación logarítmica doble está indicada cuando la representación en papel log-log conduce a una línea recta.

Transformación recíproca. Muchos fenómenos (una determinada acción por unidad de tiempo o por unidad de población) como los batidos de alas por segundo o el número de huevos puestos por hembra, darán curvas hiperbólicas al representarlos en la escala de medida original. De este modo, forman curvas descritas por las ecuaciones matemáticas generales $bXY = 1$ o $(a + bX)\,Y = 1$. De éstas podemos deducir $1/Y = bX$ o $1/Y = a + bX$. Transformando la variable dependiente en su recíproco, frecuentemente podemos obtener regresiones rectilíneas.

Finalmente, ciertas curvas acumulativas pueden rectificarse por medio de la *transformación probit*. Recuérdese la curva acumulativa normal mostrada en la figura 5.5. Téngase en cuenta que cambiando la ordenada de la escala normal acumulativa a la escala probabi-

lística se podía rectificar esta curva. Aquí hacemos lo mismo excepto que graduamos la escala probabilística en unidades de desviación típica. Así, el punto 50 % corresponde a 0 desviaciones típicas, el punto 84'13 % a + 1 desviación típica, y el punto 2'27 % a − 2 desviaciones típicas. Estas desviaciones, correspondientes a un porcentaje acumulativo, se denominan *desviaciones equivalentes normales (D.E.N.)*. Si utilizásemos papel milimetrado ordinario y marcásemos la ordenada en unidades *D.E.N.*, al representar la curva normal acumulativa frente a ella obtendríamos una línea recta. *Probits* son simplemente desviantes equivalentes normales transformadas por la adición de 5,0, lo cual evitará valores negativos para la mayoría de las desviaciones. Así, el valor probit 5,0 corresponde a una frecuencia acumulativa de 50 %, el valor probit 6,0 a 84,13 % y el 3,0 a 2,27 %.

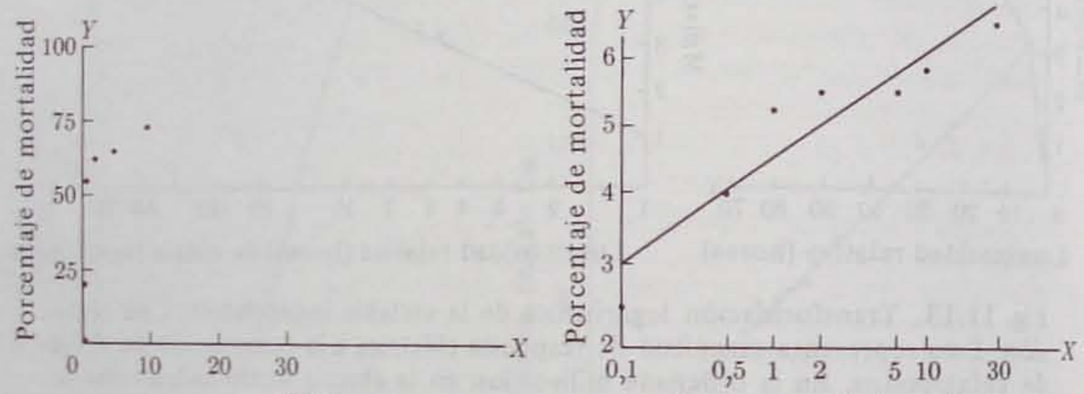


Fig. 11.14. Datos de dosificación de mortalidad que ilustran una aplicación de la transformación probit. Los datos son mortalidades medias para dos réplicas. Se sometieron veinte *Drosophila melanogaster* por réplica a siete dosis de un insecticida "desconocido" en un experimento de clase. Al punto que dio un 0 % de mortalidad a la dosis 0,1 se le ha asignado un valor probit de 2,5 en lugar de $-\infty$, que no puede ser representado.

La figura 11.14 muestra un ejemplo de porcentajes de mortalidad para dosis crecientes de un insecticida. Estos representan diferentes puntos de una distribución de frecuencias acumulativas. Con dosificaciones crecientes muere una proporción de la muestra cada vez mayor hasta que a una dosis suficientemente alta se muere la muestra entera. Con frecuencia se descubre que si las dosis de tóxicos se transforman en logaritmos, las tolerancias de muchos organismos a estos tóxicos siguen aproximadamente la distribución normal. Estas dosis transformadas se denominan a veces *dosificaciones*. El aumentar las dosificaciones conduce a una distribución normal acumulativa de mortalidades, a veces denominadas *curvas de dosificación-mortalidad*. Estas curvas constituyen la materia de un área completa del análisis biométrico, el *bioensayo*, al que aquí sólo podemos referirnos de pasada. La técnica más común en este campo es el *análisis probit*. Los resultados de este análisis se representan en *papel probit*, que es papel milimetrado probabilístico en el cual la abscisa se ha transformado en escala logarítmica. Se ajusta una línea de regresión a

los datos de *dosificación-mortalidad* representados en papel probit (ver figura 11.14). A partir de la línea de regresión se estima la dosis letal 50 % por un proceso de predicción inversa, es decir, estimamos el valor de X correspondiente a un valor probit de mortalidad 5,0, que es equivalente al 50 %.

Ejercicios 11

11.1 Las temperaturas siguientes (Y) se registraron en un conejo a diversos tiempos (X), después de inocularlo con virus de la peste (datos de Carter y Mitchell, 1958).

Tiempo después de la inyección (horas)	*Temperatura (°F)*
24	102,8
32	104,5
48	106,5
56	107,0
72	103,9
80	103,2
96	103,1

Hacer la gráfica de los datos. Sin duda los tres últimos puntos representan un fenómeno diferente de los cuatro primeros pares. *Para los cuatro primeros puntos:* a) Calcular b. b) Calcular la ecuación de regresión y trazar la línea de regresión. c) Probar la hipótesis de que $\beta = 0$ y fijar límites de confianza del 95 % para la estimación de la temperatura del conejo 50 horas después de la inyección. SOLUCION. $b = 0{,}1300$, $\hat{Y}_{50} = 106{,}5$.

11.2 La siguiente tabla se ha obtenido de los datos de Sokoloff (1955). Los pesos de hembras adultas *Drosophila persimilis* desarrolladas a 24°C están afectados por su densidad larvaria. Hacer un análisis de la varianza entre densidades. Luego, calcular la regresión de peso sobre densidad y descomponer las sumas de cuadrados entre grupos en la explicada y la inexplicada por regresión lineal. Hacer la gráfica de los datos ajustando a las medias la línea de regresión. Interpretar los resultados.

Densidad larvaria	*Peso medio de las adultas (en mg)*	*s de los pesos (no $s_{\bar{Y}}$)*	*n*
1	1,356	0,180	9
3	1,356	0,133	34
5	1,284	0,130	50
6	1,252	0,105	63
10	0,989	0,130	83
20	0,664	0,141	144
40	0,475	0,083	24

11.3 Utilizando todos los datos del ejercicio 11.1 calcular la ecuación de regresión y compararla con la obtenida anteriormente. Discutir el efecto de la inclusión en el análisis de los tres últimos puntos.

11.4 Davis (1955) dio los siguientes resultados de un estudio de la cantidad de energía metabolizada por el gorrión *Passer dometicus*, bajo diversas condiciones de temperatura constante y un fotoperíodo de diez horas.

Temperatura (^oC)	*Calorías* $\bar{Y}$	n	s
0	24,9	6	1,77
4	23,4	4	1,99
10	24,2	4	2,07
18	18,7	5	1,43
26	15,2	7	1,52
34	13,7	7	2,70

Analizar e interpretar.

11.5 En un estudio del consumo de oxígeno (microlitros/mg peso seco/hora) en *Heliothis zea* realizado por Phillips y Newsom (1966) bajo temperaturas y fotoperíodos controlados, se obtuvieron los siguientes resultados:

Temperatura (^oC)	*Fotoperíodo (horas)*	
	10	14
18	0,51	1,61
21	0,53	1,64
24	0,89	1,73

Calcular la regresión para cada fotoperíodo por separado y probar la homogeneidad de las pendientes. SOLUCION. $s^2_{Y.X} = 0{,}019267$ y $0{,}00060$ para fotoperíodos de 10 y 14 horas.

11.6 Longitud del período embrionario (en días) del saltamontes de la patata *Empoasca fabae*, desde huevo hasta adulto a diversas temperaturas constantes (Kouskolekas y Decker, 1966). Los datos originales eran medias ponderadas, pero para los fines de este análisis las consideraremos como si fuesen valores observados únicos.

Temperatura ^oF	*Longitud media del período embrionario en días* $\bar{Y}$
59,8	58,1
67,6	27,3
70,0	26,8
70,4	26,3
74,0	19,1
75,3	19,0
78,0	16,5
80,4	15,9
81,4	14,8
83,2	14,2
88,4	14,4
91,4	14,6
92,5	15,3

Analizar e interpretar. Calcular las desviaciones de la línea de regresión $(\bar{Y}_i - \hat{Y}_i)$ y representar frente a temperatura.

11.7 El experimento citado en el ejercicio 11.4 se repitió utilizando un fotoperíodo de 15 horas y se obtuvieron los siguientes resultados.

Temperatura (^oC)	*Calorías* $\bar{Y}$	n	s
0	24,3	6	1,93
10	25,1	7	1,98
18	22,2	8	3,67
26	13,8	10	4,01
34	16,4	6	2,92

Hacer una prueba de la igualdad de pendientes de las líneas de regresión para los fotoperíodos de 10 y de 15 horas. SOLUCION. $F_s = 0{,}003$.

Capítulo 12

Correlación

En este capítulo continuamos nuestra discusión de estadísticos bivariantes. El capítulo de regresión trata de la relación funcional de una variable con otra; éste trata de la medida del grado de asociación entre dos variables. Este tema, en general, se denomina análisis de correlación.

No siempre está claro qué tipo de análisis, regresión o correlación, debería emplearse en un determinado problema. Sobre esta materia ha habido considerable confusión en la mente de los investigadores y también en la literatura. Al principio de la sección 12.1 trataremos de clarificar la distinción entre estos dos métodos. En la sección que sigue (12.2) se hará la introducción del coeficiente de correlación producto-momento, el coeficiente de correlación ordinario de la literatura. Deduciremos una fórmula para este coeficiente y expondremos algo de su fundamento teórico. En esta sección se examinará la relación matemática íntima entre análisis de regresión y correlación; además calcularemos un coeficiente de correlación producto-momento. A continuación, sección (12.3) se tratan diversas pruebas de significación que incluyen coeficientes de correlación. Una vez que se ha aprendido algo sobre coeficientes de correlación, es hora de discutir sus aplicaciones en la sección 12.4.

La sección 12.5 contiene un método no paramétrico para probar la asociación. Se utiliza en aquellos casos en que no se cumplen los requisitos necesarios para las pruebas que incluyen coeficientes de correlación, o en los que son preferibles pruebas rápidas pero de inferior rendimiento, por razones de rapidez en el cálculo o por conveniencia.

12.1 Correlación y regresión

Ha habido mucha confusión acerca de la materia de correlación y regresión. Es muy frecuente que problemas de correlación sean tratados como regresión en la literatura científica, e igualmente ocurre lo contrario. Hay varias razones para esta confusión. En

primer lugar, las relaciones matemáticas entre los dos métodos de análisis son muy estrechas y matemáticamente se puede ir de uno al otro con facilidad. Por tanto, la tentación a hacer esto es grande. En segundo lugar, los textos anteriores no han clarificado suficientemente la distinción entre los dos métodos, y este riesgo aún no ha sido totalmente superado. Por lo menos un texto general sinonimiza las dos, una medida que sentimos, pues solamente puede sembrar la confusión. Finalmente, aun cuando el método elegido por un investigador pueda ser correcto por lo que se refiere a sus intenciones, los datos disponibles para el análisis pueden ser tales como para hacer inadecuada una u otra de las técnicas.

Vamos a examinar estos puntos con cierta extensión. Las muchas y estrechas relaciones matemáticas entre regresión y correlación se detallarán en la sección siguiente. Por ahora basta establecer que para cualquier problema dado la mayoría de los pasos del cálculo son los mismos si se realiza un análisis de regresión o de correlación. Se recordará que la cantidad fundamental requerida para el análisis de regresión es la suma de productos. Esta es la misma cantidad que sirve de base para calcular el coeficiente de correlación. Hay ciertas relaciones matemáticas simples entre coeficientes de regresión y sus correspondientes coeficientes de correlación. Así pues, existe la tentación de calcular un coeficiente de correlación correspondiente a un determinado coeficiente de regresión. Sin embargo, como veremos más abajo, esto sería un error a no ser que nuestra intención al principio hubiese sido estudiar la asociación y los datos fuesen apropiados para este cálculo.

Así, un coeficiente de correlación calculado a partir de datos que se han analizado correctamente por regresión modelo I, carece de sentido como estimación de cualquier coeficiente de correlación de la población. Por el contrario, podemos evaluar un coeficiente de regresión de una variable sobre otra con datos que han sido debidamente estimados como correlaciones. La explicación de tal dependencia funcional para estas variables no solamente no satisfacería nuestras intenciones, sino que deberíamos señalar que un coeficiente de regresión convencional calculado con datos en los que las dos variables se miden sin error, proporciona estimaciones sesgadas de la relación funcional.

Consideremos entonces las intenciones o fines que hay tras los dos tipos de análisis. En regresión nos proponemos describir la dependencia de una variable Y sobre una variable independiente X. Como hemos visto, utilizamos ecuaciones de regresión para prestar apoyo a hipótesis que consideran la posible causalidad de cambios en Y por cambios en X a efectos de predicción, de Y en función de X; y a efectos de explicar parte de la variación de Y por X, utilizando la segunda variable como un control estadístico. Los estudios de los efectos de la temperatura sobre la frecuencia cardíaca, contenido de nitrógeno del suelo sobre la velocidad de crecimiento en una planta, edad de un animal sobre presión sanguínea, o dosis de un insecticida sobre mortalidad de la población de insectos, son todos ejemplos típicos de regresión a los efectos señalados anteriormente.

En correlación, en cambio, estamos en gran parte interesados en si dos variables son interdependientes o *covarían*, es decir, varían juntas. No expresamos una como una función de la otra. No hay distinción entre variables dependientes e independientes. Puede ser que del par de variables cuya correlación se estudia, una sea la causa de la otra, pero ni sabemos ni suponemos esto. Un supuesto más típico (pero no esencial) es que las dos variables son efectos de una causa común. Lo que queremos estimar es el grado en que estas variables varían juntas. Así, pudiéramos estar interesados en la correlación entre

longitud del brazo y longitud de la pierna en una población de mamíferos, o entre peso corporal y producción de huevos en hembras de moscarda, o entre días hasta la madurez y número de semillas en una hierba. No es necesario ocuparnos todavía de las razones por las que desearíamos demostrar y medir la asociación entre pares de variables. Nos ocuparemos de esto en la sección 12.5. Por ahora basta afirmar que cuando queremos establecer el grado de asociación entre pares de variables en una muestra de población, el análisis de correlación es el método apropiado.

Aunque procuremos el método correcto de acuerdo con nuestros propósitos, podemos chocar con la naturaleza de los datos. Así, podemos querer establecer el contenido de colesterol en la sangre como función del peso, y para esto podemos coger una muestra al azar de hombres de la misma edad, obtener el contenido de colesterol y el peso de cada individuo, y hacer la regresión del primero en función del segundo. No obstante, estas dos variables se han medido sin error. Las variantes individuales de la supuestamente independiente variable X no han sido deliberadamente elegidas ni controladas por el investigador. Los requisitos fundamentales del modelo I de regresión no se cumplen, y no es lógico ajustar a los datos una regresión modelo I, aunque no es difícil que se encuentren ejemplos de tales prácticas incorrectas en la literatura de investigación publicada. Si lo que buscamos es realmente una ecuación que describa la dependencia de Y sobre X, deberíamos realizar una regresión modelo II. En cambio, si lo que interesa es el grado de asociación entre las variables (interdependencia), deberíamos hacer un análisis de correlación para el que sean adecuados estos datos. La dificultad contraria es tratar de obtener un coeficiente de correlación de datos que sean propiamente computados como regresión, es decir, cuando X es fija. Un ejemplo sería los latidos del corazón de un animal poiquilotermo como función de la temperatura, habiendo aplicado varias temperaturas en un experimento. Este coeficiente de correlación se obtiene matemáticamente con facilidad

TABLA 12.1

Las relaciones entre correlación y regresión. Esta tabla indica el cálculo correcto para cualquier combinación de intenciones y variables, como se expone.

Propósito del investigador	*Naturaleza de las dos variables*	
	Y al azar, X fija	*Y_1, Y_2 ambas al azar*
Establecer y estimar la dependencia de una variable respecto de otra	Regresión modelo I.	Regresión modelo II. (No tratado en este libro.)
Establecer y estimar la asociación (interdependencia) entre dos variables	Sin sentido para este caso. Si se desea, puede obtenerse una estimación de la proporción de la variación de Y explicable por X, como el cuadrado del coeficiente de correlación entre X e Y.	Coeficiente de correlación. (Las pruebas de significación sólo son totalmente adecuadas si Y_1, Y_2 se distribuyen como variables normales bivariantes.)

pero sería simplemente un valor numérico, no un estimador de una medida paramétrica de correlación. Existe una interpretación que puede darse al cuadrado del coeficiente de correlación de cierta aplicabilidad a un problema de regresión. No obstante, en ningún caso es una estimación de una correlación paramétrica.

Esta discusión se resume en la tabla 12.1, que muestra las relaciones entre correlación y regresión. Las dos columnas de la tabla indican las dos condiciones del par de variables: en un caso una variable aleatoria y medida con error, la otra fija; en el otro caso, las dos variables al azar. Hemos dejado el convencionalismo usual de designar el par de variables Y y X o X_1, X_2 para ambos análisis, correlación y regresión. En regresión seguimos utilizando Y para la variable dependiente y X para la variable independiente, pero en correlación las dos variables son en realidad variables aleatorias, que hemos designado como Y en todo el texto. Por lo tanto mencionamos las dos variables como Y_1, Y_2. Las filas de la tabla indican la intención del investigador al hacer su análisis, y los cuatro cuadrantes de la tabla indican los procedimientos adecuados para una determinada combinación de intención del investigador y naturaleza del par de variables.

12.2 El coeficiente de correlación producto-momento

Hay numerosos coeficientes de correlación en estadística. El más común de éstos se denomina *coeficiente de correlación producto-momento*, que en su formulación general se debe a Karl Pearson. Deduciremos su fórmula por medio de un procedimiento intuitivo.

Se ha visto que la suma de productos es una medida de covariación y, por consiguiente, es probable que ésta sea la cantidad básica de la que obtener una fórmula para el coeficiente de correlación. Las variables cuya correlación se va a estimar las designaremos por Y_1 e Y_2. Su suma de productos será por tanto $\sum y_1 y_2$ y su covarianza $[1/(n-1)]$ $\sum y_1 y_2 = s_{12}$. La segunda cantidad es análoga a una varianza, es decir, una suma de cuadrados dividida por sus grados de libertad.

Una desviación típica se expresa en unidades de medida originales tales como pulgadas, gramos, o centímetros cúbicos. Igualmente, un coeficiente de regresión se expresa como tantas unidades de Y por unidad de X, tal como 5,2 gramos/día. Sin embargo, una medida de asociación debería ser independiente de la escala original de medida a fin de poder comparar el grado de asociación en un par de variables con el de otro. Una forma de hacer esto es dividir la covarianza por las desviaciones típicas de las variables Y_1 e Y_2. Esto conduce a dividir cada desviación y_1 e y_2 por su propia desviación típica y convertirla en una desviación tipificada. La expresión se convierte ahora en la suma de los productos de las desviaciones tipificadas dividida por $n-1$.

$$r_{Y_1Y_2} = \frac{\sum y_1 y_2}{(n-1)s_{Y_1}s_{Y_2}} \tag{12.1}$$

Esta es la fórmula para el coeficiente de correlación producto-momento $r_{Y_1Y_2}$ entre las variables Y_1 e Y_2. Simplificaremos el simbolismo a

$$r_{12} = \frac{\sum y_1 y_2}{(n-1)s_1 s_2} = \frac{s_{12}}{s_1 s_2} \tag{12.2}$$

La expresión (12.2) puede volverse a escribir de otra forma común. Como la expresión

$$s\sqrt{n-1} = \sqrt{s^2(n-1)} = \sqrt{\frac{\sum y^2}{n-1}}(n-1) = \sqrt{\sum y^2}$$

(12.2) puede escribirse como

$$r_{12} = \frac{\sum y_1 y_2}{\sqrt{\sum y_1^2 \sum y_2^2}} \tag{12.3}$$

que a veces es preferible para el cálculo. Para establecer la expresión (12.2) de un caso más general, para las variables Y_j e Y_k, podemos escribirla como

$$r_{jk} = \frac{\sum y_j y_k}{(n-1)s_j s_k} \tag{12.4}$$

El coeficiente de correlación r_{jk} puede variar desde + 1 para asociación positiva perfecta hasta − 1 para asociación negativa perfecta. Esto es intuitivamente evidente cuando consideramos la correlación de una variable Y_j consigo misma. La expresión (12.4) daría en este caso $r_{jj} = \sum y_j y_j / \sqrt{\sum y_j^2 \sum y_j^2} = \sum y_j^2 / \sum y_j^2 = 1$, lo cual da una correlación perfecta de + 1. Si las desviaciones de una variable se apareasen con desviaciones idénticas pero opuestas representantes de otra variable, esto daría una correlación de − 1 porque la suma de productos del numerador sería negativa. La prueba de que el coeficiente de correlación está limitado por + 1 y − 1 se dará en breve.

Si las variantes siguen una distribución particular, la *distribución normal bivariante*, el coeficiente de correlación r_{jk} estimará un parámetro de la distribución simbolizado por ρ_{jk}.

Vamos a aproximar empíricamente la distribución. Supongamos que se han muestreado cien ítems y medido dos variables en cada ítem, obteniendo de esta manera dos muestras de 100 variantes. Si se representan estos 100 ítems en una gráfica en la cual las variables Y_1 e Y_2 sean las coordenadas, se obtendrá un diagrama de esparcimiento de puntos como en la figura 12.3A. Supongamos que ambas variables, Y_1 e Y_2, se distribuyen normalmente y son completamente independientes una de otra, de modo que el hecho de que un individuo resulte ser mayor que la media en el carácter Y_1 no afecta en absoluto a su valor para la variable Y_2. Así, este mismo individuo puede ser mayor o menor que la media para la variable Y_2. Si no hay absolutamente ninguna relación entre Y_1 e Y_2 y si las dos variables se tipifican para hacer comparables sus escalas, se encontraría que el perfil del diagrama de esparcimiento es aproximadamente circular. Naturalmente, para una muestra de 100 ítems, el círculo estaría sólo imperfectamente delimitado; pero cuanto más grande sea la muestra más claramente podría distinguirse un círculo con el área central alrededor de la intersección $\bar{Y}_1$, $\bar{Y}_2$ intensamente oscurecida debido a la

agregación en ella de muchos puntos. Si se continúa muestreando, habrá que superponer nuevos puntos sobre los anteriores, y si se hacen visibles estos puntos en sentido físico, tales como granos de arena, se acumularían gradualmente en un montículo puntiagudo acampanado. Esta es una percepción tridimensional de una distribución normal mostrada en perspectiva en la figura 12.1. Observado desde cualquiera de los ejes de coordenadas, el montículo presentaría un aspecto bidimensional, y su contorno sería el de una curva de distribución normal, dando las dos perspectivas las distribuciones de Y_1 e Y_2, respectivamente.

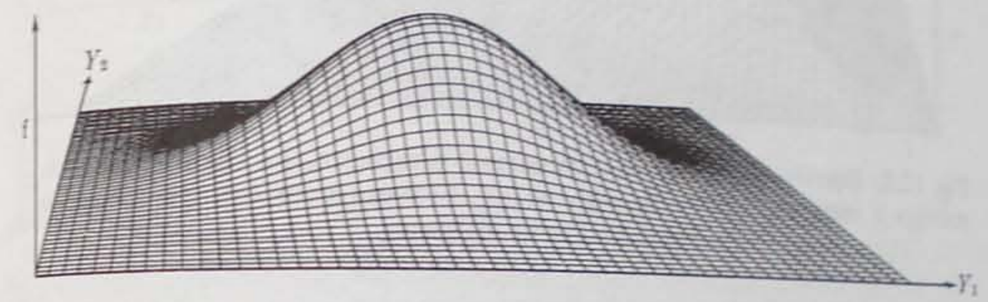


Fig. 12.1. Distribución de frecuencias normal bivariante. La correlación paramétrica ρ entre las variables Y_1 e Y_2 es igual a cero. La distribución de frecuencias puede visualizarse como un montículo acampanado.

Si suponemos que las dos variables Y_1 e Y_2 no son independientes sino que están positivamente correlacionadas en cierto grado, entonces si un determinado individuo tiene un valor grande de Y_1, es más probable que tenga también un valor grande de Y_2, que no lo contrario. Igualmente, un pequeño valor de Y_1 probablemente estará asociado con un pequeño valor de Y_2. Si se muestreasen ítems de esta población, el diagrama de esparcimiento resultante (representado en la figura 12.3 D) se alargaría en forma de elipse. Esto ocurre porque aquellas partes del círculo que primeramente incluían individuos altos para una variable y bajos para la otra (y viceversa), están ahora escasamente representadas. El muestreo continuado (con el modelo del grano de arena) da un montículo elíptico tridimensional presentado en la figura 12.2. Si la correlación es perfecta, todos los datos se hallarían a lo largo de una sola línea de regresión (la misma línea describiría la regresión de Y_1 sobre Y_2 y de Y_2 sobre Y_1), y si las dejamos acumularse en un modelo físico, darían por resultado una curva normal plana, esencialmente bidimensional, hallándose sobre su línea de regresión.

La forma elíptica o circular del contorno del diagrama de esparcimiento y del montículo resultante es sin duda una función del grado de correlación entre las dos variables, y éste es el parámetro ρ_{jk} de la distribución normal bivariante. Por analogía con la expresión (12.2), el parámetro ρ_{jk} puede definirse como

$$\rho_{jk} = \sigma_{jk}/\sigma_j\sigma_k \tag{12.5}$$

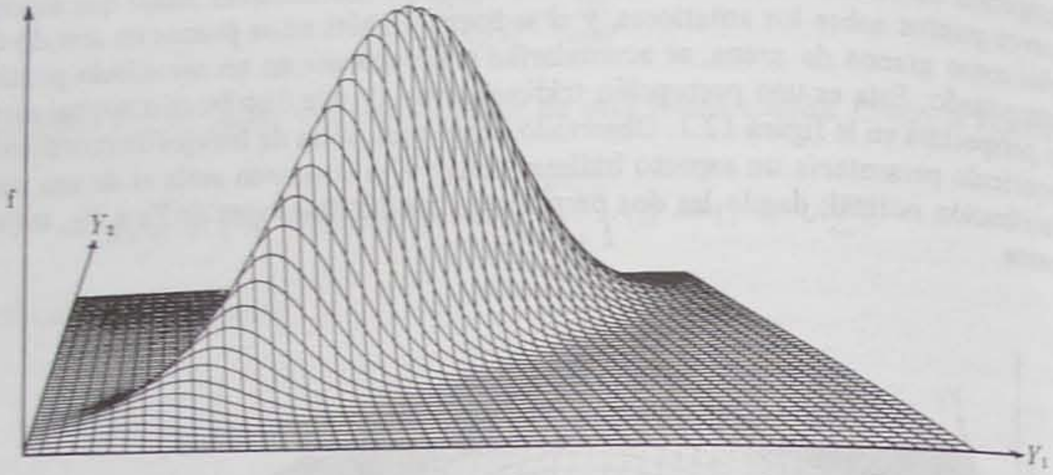
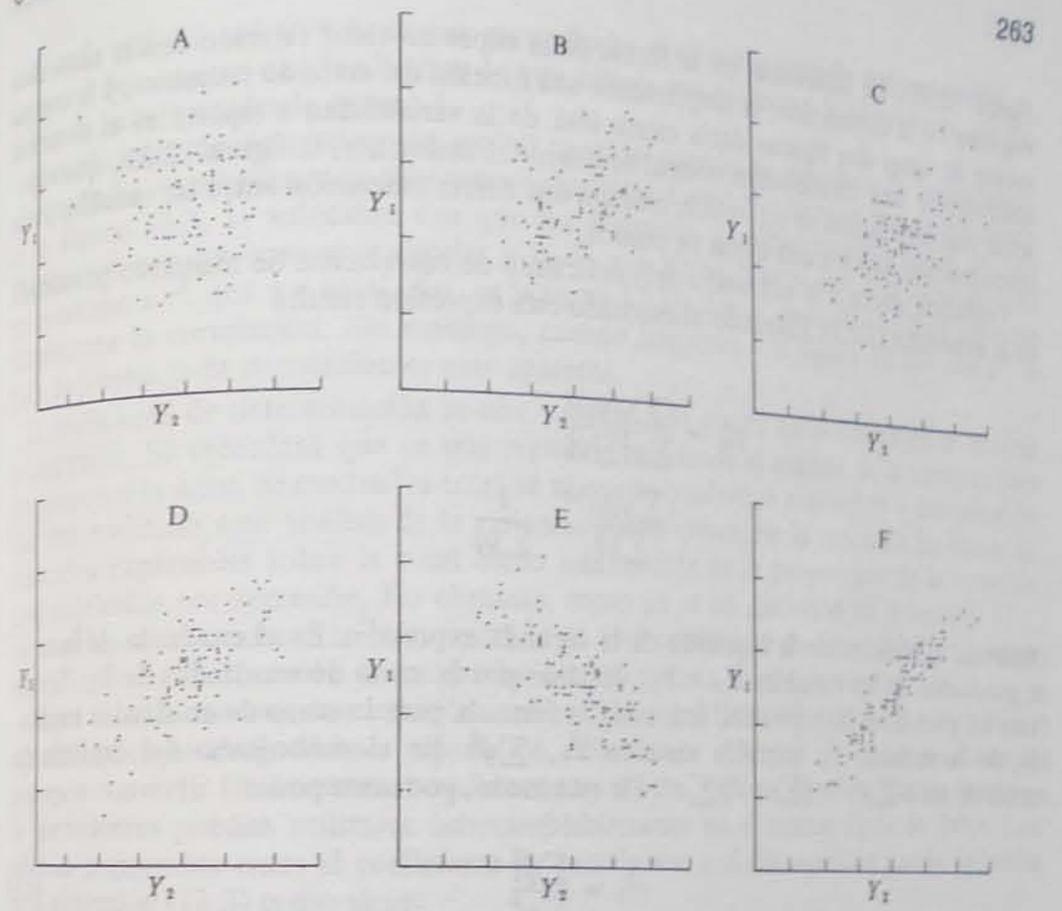


Fig. 12.2. Distribución de frecuencias normal bivariante. La correlación paramétrica ρ entre las variables Y_1 e Y_2 es 0,9. El montículo acampanado de la figura 12.1 se ha alargado.

donde σ_{jk} es la covarianza paramétrica de las variables Y_j e Y_k, y σ_j y σ_k son las desviaciones típicas paramétricas de las variables Y_j e Y_k, como antes. Cuando dos variables se distribuyen según la normal bivariante, un coeficiente de correlación de muestreo r_{jk}, estima el coeficiente de correlación paramétrico ρ_{jk}. Podemos hacer ciertas afirmaciones sobre la distribución de muestreo de ρ_{jk} y darles límites de confianza.

Lamentablemente, la forma elíptica de los diagramas de esparcimiento de variables correlacionadas, ordinariamente no es muy clara a menos que se hayan cogido muestras muy grandes o que la correlación paramétrica ρ_{jk} sea muy alta. Para ilustrar este punto, en la figura 12.3 presentamos varias gráficas que representan diagramas de esparcimiento resultantes de muestras de 100 ítems de poblaciones normales bivariantes, con diferentes valores de ρ_{jk}. Obsérvese que en la primera gráfica (figura 12.3A), con $\rho_{jk} = 0$, la distribución circular sólo está muy vagamente esbozada. Para demostrar más claramente la forma circular de la distribución se necesita una muestra mucho mayor. En la figura 12.3B, basada en $\rho_{jk} = 0{,}3$ no se observa ninguna diferencia sustancial. Sabiendo que ésta representa una correlación positiva, puede visualizarse una pendiente positiva en el diagrama de esparcimiento; pero sin conocimiento previo ésta sería difícil de detectar visualmente. La próxima gráfica (figura 12.3C, basada en $\rho_{jk} = 0{,}5$) es un poco más clara, pero a pesar de esto no muestra una tendencia inequívoca. En general, la correlación no puede deducirse de la inspección de los diagramas de esparcimiento basados en muestras de poblaciones con ρ_{jk} entre $-0{,}5$ y $+0{,}5$, a no ser que la muestra sea muy numerosa. Este punto se demuestra en la última gráfica (figura 12.3G), extraída también de una población con $\rho_{jk} = 0{,}5$ pero basada en una muestra de 500. En ella, la pendiente positiva y el contorno elíptico del diagrama de esparcimiento son completamente evidentes. La figura 12.3D, basada en $\rho_{jk} = 0{,}7$ y $n = 100$, exhibe la tendencia con más claridad. Obsérvese que la próxima gráfica (figura 12.3 E), basada en la misma magnitud de ρ_{jk} pero representando correlación negativa, también muestra la inclinación pero es más extendida que la



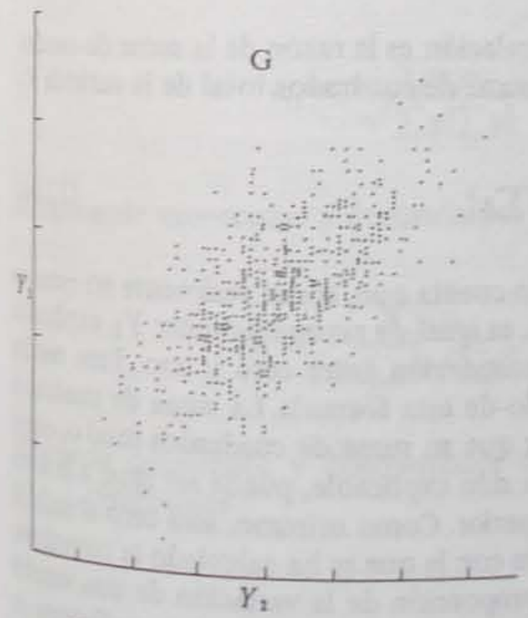


Fig. 12.3. Muestras al azar de distribuciones normales bivariantes, con valores variables del coeficiente de correlación paramétrico ρ. Tamaños de muestreo n = 100 en todas las gráficas excepto G, que tiene $n = 500$. A. $\rho = 0{,}4$. B. $\rho =$ 0,3. C. $\rho = 0{,}5$. D. $\rho = 0{,}7$. E. $\rho = -0{,}7$. F. $\rho = 0{,}9$. G. $\rho = 0{,}5$.

figura 12.3D. La diferencia en la forma de la elipse no tiene relación con la naturaleza negativa de la correlación; es simplemente una función del error de muestreo, y la comparación de estas dos figuras daría cierta idea de la variabilidad a esperar en el muestreo aleatorio de una distribución normal bivariante. Finalmente, la figura 12.3F, que representa una correlación de $\rho_{jk} = 0{,}9$, muestra una fuerte asociación entre las variables y una aproximación lógica a una elipse de puntos.

Volvamos ahora a la expresión del coeficiente de correlación de muestreo presentado en la expresión (12.3). Elevando al cuadrado esta expresión resulta

$$r_{12}^2 = \frac{(\sum y_1 y_2)^2}{\sum y_1^2 \sum y_2^2}$$

$$= \frac{(\sum y_1 y_2)^2}{\sum y_1^2} \cdot \frac{1}{\sum y_2^2}$$

Obsérvese el término de la izquierda de la segunda expresión. Es el cuadrado de la suma de productos de las variables Y_1 e Y_2, dividido por la suma de cuadrados de Y_1. Si éste fuera un problema de regresión, ésta sería la fórmula para la suma de cuadrados explicable de la variable Y_2 sobre la variable Y_1, $\sum \hat{y}_2^2$. En el simbolismo del capítulo de regresión, sería $\sum \hat{y}^2 = (\sum xy)^2 / \sum x^2$. De este modo, podemos poner

$$r_{12}^2 = \frac{\sum \hat{y}_2^2}{\sum y_2^2} \tag{12.6}$$

Por lo tanto, el cuadrado del coeficiente de correlación es la razón de la suma de cuadrados explicable de la variable Y_2, dividida por la suma de cuadrados total de la variable Y_2 o, equivalentemente,

$$r_{12}^2 = \sum \hat{y}_1^2 / \sum y_1^2 \tag{12.6a}$$

la cual podría deducirse con facilitad. Téngase en cuenta que, como realmente no estamos haciendo regresión de una variable sobre la otra, es igual de razonable tener Y_1 explicable por Y_2 como lo contrario. Esta razón es una proporción entre cero y uno. Esto resulta obvio tras una ligera consideración del significado de esta fórmula. La suma de cuadrados explicable de cualquier variable debe ser menor que su suma de cuadrados total o, como máximo, si toda la variación de una variable ha sido explicable, puede ser igual a la suma de cuadrados total, pero indudablemente no superior. Como mínimo, será cero si nada de la variable pudiera justificarse por la otra variable con la que se ha calculado la covarianza. Así, obtenemos una medida importante de la proporción de la variación de una variable determinada por la variación de la otra. Esta cantidad, el cuadrado del coeficiente de correlación, r_{12}^2, se denomina *coeficiente de determinación*. Varía desde cero hasta 1 y debe ser positivo independientemente de que el coeficiente de correlación sea negativo o positivo. Casualmente, ésta es una prueba de que el coeficiente de correlación no puede

variar más allá de -1 y $+1$. Como su cuadrado es el coeficiente de determinación, y acabamos de exponer que los límites de este último son de cero a 1, es evidente que los límites de su raíz cuadrada serán ± 1.

El coeficiente de determinación es útil también cuando se considera la importancia relativa de las correlaciones de diferentes magnitudes. Como puede verse por un reexamen de la figura 12.3, la velocidad con que los diagramas de esparcimiento pasan de una distribución con un contorno circular a la de una elipse, parece ser más directamente proporcional a r^2 que a r en sí. Así, en la figura 12.3B, con $\rho^2 = 0{,}09$, es difícil detectar visualmente la correlación. Sin embargo, cuando llegamos a la figura 12.3D, con $\rho^2 = 0{,}49$, la presencia de correlación es muy aparente.

El coeficiente de determinación es una cantidad que puede ser útil también en análisis de regresión. Se recordará que en una regresión usábamos el análisis de la varianza para descomponer la suma de cuadrados total en sumas de cuadrados explicable e inexplicable. Una vez realizado este análisis de la varianza, puede obtenerse la razón de las sumas de cuadrados explicables sobre la total como una medida de la proporción de la variación total explicable por regresión. No obstante, como ya se ha discutido en la sección 12.1, no tendría sentido extraer la raíz cuadrada de este coeficiente de determinación y considerarla como una estimación de la correlación paramétrica de estas variables.

Ahora consideraremos una relación matemática entre los coeficientes de correlación y regresión. Con el riesgo de ser repetitivos, deberíamos subrayar de nuevo que aunque podemos convertir fácilmente un coeficiente en el otro, esto no significa que los dos tipos de coeficientes puedan utilizarse intercambiablemente en el mismo tipos de datos. Una relación importante entre el coeficiente de correlación y el de regresión puede deducirse de la expresión (12.3) como sigue:

$$r_{12} = \frac{\sum y_1 y_2}{\sqrt{\sum y_1^2 \sum y_2^2}} = \frac{\sum y_1 y_2}{\sqrt{\sum y_1^2}} \cdot \frac{1}{\sqrt{\sum y_2^2}}$$

Multiplicando numerador y denominador de esta expresión por $\sqrt{\sum y_1^2}$, obtenemos

$$r_{12} = \frac{\sum y_1 y_2}{\sqrt{\sum y_1^2 \sum y_1^2}} \cdot \frac{\sqrt{\sum y_1^2}}{\sqrt{\sum y_2^2}} = \frac{\sum y_1 y_2}{\sum y_1^2} \cdot \frac{\sqrt{\sum y_1^2}}{\sqrt{\sum y_2^2}}$$

Dividiendo numerador y denominador del término de la derecha de esta expresión por $\sqrt{n-1}$, obtenemos

$$r_{12} = \frac{\sum y_1 y_2}{\sum y_1^2} \cdot \frac{\sqrt{\frac{\sum y_1^2}{n-1}}}{\sqrt{\frac{\sum y_2^2}{n-1}}} = b_{2 \cdot 1} \frac{s_1}{s_2} \tag{12.7}$$

Igualmente, podríamos demostrar que

$$r_{12} = b_{1\cdot 2}\frac{s_2}{s_1} \tag{12.7a}$$

y por tanto

$$b_{2\cdot 1} = r_{12}\frac{s_2}{s_1} \qquad b_{1\cdot 2} = r_{12}\frac{s_1}{s_2} \tag{12.7b}$$

En estas expresiones $b_{2\cdot 1}$ es el coeficiente de regresión para la variable Y_2 sobre Y_1. Vemos pues que el coeficiente de correlación es la pendiente de regresión multiplicada por la razón de las desviaciones típicas de las variables. El coeficiente de correlación puede considerarse por tanto como un coeficiente de regresión tipificado. Si las dos desviaciones típicas son idénticas, los dos coeficientes de regresión y el coeficiente de correlación tendrán idéntico valor.

Ahora que tenemos conocimiento del coeficiente de correlación, puede ponerse en perspectiva correcta parte de la labor anterior sobre comparaciones apareadas (véase sección 9.3). En el apéndice A1.9 demostramos, para las expresiones paramétricas correspondientes, que la varianza de una suma de dos variables es

$$s^2_{(Y_1+Y_2)} = s_1^2 + s_2^2 + 2r_{12}s_1s_2 \tag{12.8}$$

donde s_1 y s_2 son las desviaciones típicas de Y_1 e Y_2, respectivamente, y r_{12} es el coeficiente de correlación entre estas variables. Igualmente, para una diferencia entre dos variables, obtenemos

$$s^2_{(Y_1-Y_2)} = s_1^2 + s_2^2 - 2r_{12}s_1s_2 \tag{12.9}$$

Lo que indica la expresión (12.8) es que si componemos una nueva variable que sea la suma de otras dos variables, la varianza de esta nueva variable será la suma de las varianzas de las variables de que se compone, más un término aditivo que es una función de las desviaciones típicas de estas dos variables y de la correlación entre ellas. En el apéndice A1.9 se indica que este término aditivo es dos veces la covarianza de Y_1 e Y_2. Cuando las dos variables que se suman no están en correlación, esta covarianza aditiva será también cero, y la varianza de la suma será simplemente la suma de las varianzas de las dos variables. Esta es la razón por la cual en un análisis de la varianza o en una prueba t de la diferencia entre las dos medias, teníamos que suponer la independencia de las dos variables para permitirnos sumar sus varianzas. De lo contrario, deberíamos haber tenido en cuenta un término de covarianza. En cambio, en la técnica de comparaciones apareadas esperamos correlación entre las variables, ya que cada par comparte una experiencia común. La prueba de comparaciones apareadas sustrae automáticamente un término de covarianza, conduciendo a un error estándar menor y en consecuencia a un valor de t mayor, ya que el numerador de la razón sigue siendo el mismo. Así, siempre que la correlación entre dos variables sea positiva, la varianza de sus diferencias será considerablemente menor que la suma de sus varianzas; ésta es la razón por la cual tiene que

utilizarse la prueba de comparaciones apareadas en lugar de la prueba t para diferencia de medias. Estas consideraciones son igualmente ciertas para los correspondientes análisis de la varianza, de clasificación simple y doble.

El cálculo de un coeficiente de correlación producto-momento es muy sencillo. Las cantidades básicas necesarias son las seis mismas requeridas para calcular el coeficiente de regresión (sección 11.3). El cuadro 12.1 ilustra cómo debería calcularse el coeficiente. El ejemplo está basado en una muestra de doce cangrejos de mar en los cuales se ha medido el peso de las branquias Y_1 y el peso corporal Y_2. Deseamos saber si hay una correlación entre el peso de las branquias y el del cuerpo, representando este último una medida de tipo global. La existencia de una correlación positiva podría llevar a concluir que un cangrejo de cuerpo más grande, con su mayor cantidad de metabolismo resultante, requeriría branquias más grandes a fin de suministrar el oxígeno necesario. Los cálculos se presentan en el cuadro 12.1. El coeficiente de correlación de 0,87 concuerda con la pendiente definida y estricto contorno elíptico del diagrama de esparcimiento para estos datos en la figura 12.4.

CUADRO 12.1

Cálculo del coeficiente de correlación producto-momento.

Relación entre peso branquial y peso corporal en el cangrejo de mar *Pachygrapsus crassipes*. $n = 12$.

(1) Y_1 *Peso branquial en miligramos*	(2) Y_2 *Peso corporal en gramos*
159	14,40
179	15,20
100	11,30
45	2,50
384	22,70
230	14,90
100	1,41
320	15,81
80	4,19
220	15,39
320	17,25
210	9,52

Fuente: Datos no publicados de L. Miller.

CUADRO 12.1 (continuación)

Cálculo

1. $\sum Y_1 = 159 + \cdots + 210 = 2347$
2. $\sum Y_1^2 = 159^2 + \cdots + 210^2 = 583\,403$
3. $\sum Y_2 = 14{,}40 + \cdots + 9{,}52 = 144{,}57$
4. $\sum Y_2^2 = (14{,}40)^2 + \cdots + (9{,}52)^2 = 2204{,}1853$
5. $\sum Y_1 Y_2 = 14{,}40(159) + \cdots + 9{,}52(210) = 34\,837{,}10$
6. Suma de cuadrados de $Y_1 = \sum y_1^2 = \sum Y_1^2 - \dfrac{(\sum Y_1)^2}{n}$

$$= \text{cantidad } 2 - \frac{(\text{cantidad } 1)^2}{n} = 583\,403 - \frac{(2347)^2}{12}$$

$$= 124\,368{,}9167$$

7. Suma de cuadrados de $Y_2 = \sum y_2^2 = \sum Y_2^2 - \dfrac{(\sum Y_2)^2}{n}$

$$= \text{cantidad } 4 - \frac{(\text{cantidad } 3)^2}{n} = 2204{,}1853 - \frac{(144{,}57)^2}{12}$$

$$= 462{,}4782$$

8. Suma de productos $= \sum y_1 y_2 = \sum Y_1 Y_2 - \dfrac{(\sum Y_1)(\sum Y_2)}{n}$

$$= \text{cantidad } 5 - \frac{\text{cantidad } 1 \times \text{cantidad } 3}{n}$$

$$= 34\,837{,}10 - \frac{(2347)(144{,}57)}{12} = 6561{,}6175$$

9. Coeficiente de correlación producto-momento [según expresión (12.3)] =

$$r_{12} = \frac{\sum y_1 y_2}{\sqrt{\sum y_1^2 \sum y_2^2}} = \frac{\text{cantidad } 8}{\sqrt{\text{cantidad } 6 \times \text{cantidad } 7}}$$

$$= \frac{6561{,}6175}{\sqrt{(124\,368{,}9167)(462{,}4782)}} = \frac{6561{,}6175}{\sqrt{57\,517\,912{,}7314}}$$

$$= \frac{6561{,}6175}{7584{,}0565} = 0{,}8652 \approx 0.87$$

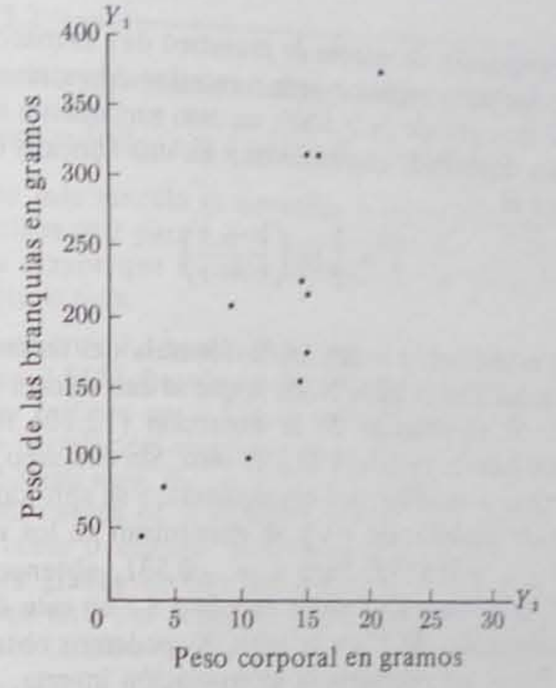


Fig. 12.4. Diagrama de esparcimiento para los datos del cangrejo de mar del cuadro 12.1.

12.3 Prueba de significación en correlación

La prueba de significación más común es si un coeficiente de correlación de muestreo podría proceder de una población con un coeficiente de correlación paramétrico de cero. La hipótesis nula es, por tanto, $H_0 : \rho = 0$. Esto implica que las dos variables no están correlacionadas. Si la muestra procede de una distribución normal bivariante y $\rho = 0$, el error estándar del coeficiente de correlación es $s_r = \sqrt{(1 - r^2)/(n - 2)}$. La hipótesis se contrasta como una prueba t con $n - 2$ grados de libertad, $t_s = (r - 0)/\sqrt{(1 - r^2)/(n - 2)}$ $= r\sqrt{(n - 2)/(1 - r^2)}$. Deberíamos subrayar que este error estándar solamente es apropiado cuando $\rho = 0$, de modo que no puede aplicarse para contrastar una hipótesis de que ρ es un valor específico distinto de cero. La prueba t para la significación de r, es matemáticamente equivalente a la prueba t para la significación de b, midiendo en ambos casos la fuerza de la asociación entre las dos variables que se están contrastando. Esta situación es un tanto análoga a la del modelo I y modelo II en el análisis de la varianza de clasificación simple, donde la misma prueba F establece la significación, independientemente del modelo.

Se han realizado sistemáticamente pruebas de significación que siguen esta fórmula y están tabuladas en la tabla **VIII**, la cual permite el reconocimiento directo de un coeficiente de correlación de muestreo con respecto a significación sin cálculo adicional. El cuadro 12.3 ilustra pruebas de la hipótesis $H_0 : \rho = 0$, utilizando la tabla **VIII** así como la prueba t discutida al principio.

Cuando $\rho \neq 0$, la distribución de valores de muestreo de r es marcadamente asimétrica, y aunque se ha hallado un error estándar para r en estos casos, no debería aplicarse a no ser que la muestra fuese muy grande ($n > 500$), un caso sumamente infrecuente y de poco interés. Para superar esta dificultad, convertimos r en una función de z, desarrollada por Fisher. La fórmula para z es

$$z = \frac{1}{2} \ln \left(\frac{1 + r}{1 - r} \right) \tag{12.10}$$

Se podría reconocer esto como $z = \text{tgh}^{-1} r$, la fórmula del arctangente hiperbólico de r. Esta función se ha tabulado en la tabla X, en la que se dan valores de z correspondientes a valores absolutos de r. El examen de la expresión (12.10) revelará que cuando $r = 0$, z también será igual a cero, ya que ½ ln 1 es cero. Sin embargo, al aproximarse r a 1, $(1 + r)/(1 - r)$ se aproxima a infinito; por consiguiente, z se aproxima a infinito. Por esta razón, las diferencias sustanciales entre r y z se encuentran en los valores de r más altos. Así, cuando r es 0,115, $z = 0{,}1155$. Para $r = -0{,}531$, obtenemos $z = -0{,}5915$; $r = 0{,}972$ da $z = 2{,}1273$. Obsérvese en cuánto excede z a r en este último par de valores. Buscando un valor determinado de z en la tabla X, podemos obtener también el valor correspondiente de r. Puede ser necesaria la interpolación inversa. Así, $z = 0{,}70$ corresponde a $r = 0{,}604$, y un valor de $z = 2{,}76$ corresponde a $r = 0{,}992$.

La ventaja de la transformación z es que mientras los coeficientes de correlación siguen una distribución sesgada para valores de $\rho \neq 0$, los valores de z siguen aproximadamente la distribución normal para cualquier valor de su parámetro, que denominamos ζ (zeta), siguiendo el convencionalismo ordinario. La varianza esperada de z es

$$\sigma_z^2 = \frac{1}{n - 3} \tag{12.11}$$

Esta es una aproximación adecuada para tamaños de muestreo $n \geq 50$ y una aproximación tolerable incluso cuando $n \geq 25$. Un aspecto interesante de la varianza de z, manifiesto en la expresión (12.11), es que es independiente de la magnitud de r, pero es simplemente una función del tamaño de muestreo n.

Como se expone en el cuadro 12.2, para tamaños de muestreo mayores que 50 podemos utilizar también la transformación-z para probar la significación de un r de muestreo empleando la hipótesis $H_0 : \rho = 0$. En la segunda sección del cuadro 12.2 explicamos la prueba de la hipótesis nula de que $\rho \neq 0$. Es posible que tengamos la hipótesis de que la verdadera correlación entre dos variables es un valor determinado ρ diferente de cero. Tales hipótesis sobre la correlación esperada entre dos variables son frecuentes en trabajos de genética, y es posible que queramos probar los datos observados en contraste con esta hipótesis. Aunque no hay razón a priori para suponer que la verdadera correlación entre las longitudes de los nervios del ala de la abeja de la parte derecha e izquierda es 0,5, presentamos la prueba de esta hipótesis para ilustrar el método. Correspondiente a $\rho = 0{,}5$, existe ζ, el valor paramétrico de z. Es la transformación-z de ρ. Observamos que la probabilidad de que el r de muestreo de 0,837 pudiera haberse extraído de una población con $\rho = 0{,}5$ es despreciablemente pequeña.

CUADRO 12.2

Pruebas de significación y límites de confianza para coeficientes de correlación.

Prueba de la hipótesis nula $H_0 : \rho = 0$ contra $H_1 : \rho \neq 0$.

El procedimiento más sencillo es consultar la tabla **VIII**, en la que están tabulados los valores críticos de r para $g.l. = n - 2$ desde 1 a 1.000. Si el valor absoluto del r observado es mayor que el valor tabulado en la columna para dos variables, rechazamos la hipótesis nula.

Ejemplos. — En el cuadro 12.1 hemos hallado que la correlación entre peso corporal y peso branquial es 0,8652, basada en una muestra de $n = 12$. Para 10 grados de libertad, los valores críticos son 0,576 al nivel de significación del 5 % y 0,708 al 1 %. Puesto que la correlación observada es superior que cualquiera de éstos, podemos rechazar la hipótesis nula, $H_0 : \rho = 0$, a $P < 0{,}01$.

La tabla **VIII** está basada en la siguiente prueba, que puede realizarse cuando no se dispone de la tabla o cuando se necesita una prueba exacta para niveles de significación o para grados de libertad distintos de los provistos en la tabla. Se contrasta la hipótesis nula por medio de la distribución t (con $n - 2$ $g.l.$) utilizando el error estándar de r. Cuando $\rho = 0$,

$$s_r = \sqrt{(1 - r^2)/(n - 2)}$$

Por lo tanto,

$$t_s = \frac{(r - 0)}{\sqrt{(1 - r^2)/(n - 2)}} = r\sqrt{(n - 2)/(1 - r^2)}$$

Para los datos del cuadro 12.1, éste sería

$$t_s = 0{,}8652\sqrt{(12 - 2)/(1 - 0{,}8652^2)} = 0{,}8652\sqrt{10/0{,}25143}$$

$$= 0{,}8652\sqrt{39{,}7725} = 0{,}8652\ (6{,}3065) = 5{,}4564 > t_{0{,}001[10]}$$

Para una prueba de una cola deberían utilizarse los valores de t 0,10 y 0,02 para pruebas de significación del 5 % y del 1 % respectivamente. Estas pruebas se aplicarían si la hipótesis alternativa fuese $H_1 : \rho > 0$ o $H_1 : \rho < 0$, en lugar de $H_1 : \rho \neq 0$.

Cuando n es mayor que 50, podemos utilizar también la transformación z descrita en el texto. Como $\sigma_z = 1/\sqrt{n - 3}$, probamos

$$t_s = \frac{z - 0}{1/\sqrt{n - 3}} = z\sqrt{n - 3}$$

Puesto que z se distribuye normalmente y estamos utilizando una desviación típica paramétrica, comparamos t_s con $t_{\alpha[\infty]}$ o empleamos la tabla **II**, áreas de la curva normal. Si tuviésemos un coeficiente de correlación de $r = 0{,}837$ entre la longitud de los nervios del ala derecha e izquierda de las abejas basado en $n = 500$, encontraríamos $z = 1{,}2111$ en la tabla **X**.

$$t_s = 1{,}2111\sqrt{497} = 26{,}997$$

Este valor, cuando se busca en la tabla de áreas de una curva normal (tabla **II**), da una probabilidad muy pequeña ($< 10^{-6}$).

CUADRO 12.2 (continuación)

Prueba de la hipótesis nula $H_0 : \rho = \rho_1$, donde $\rho_1 \neq 0$.

Para probar esta hipótesis no podemos utilizar la tabla **VIII** ni la prueba t dada más arriba, sino que debemos valernos de la transformación z.

Supongamos que queremos probar la hipótesis nula $H_0 : \rho = +0{,}50$ contra $H_1 : \rho \neq +0{,}50$ para el caso recién considerado. Utilizaríamos la siguiente expresión:

$$t_s = \frac{z - \zeta}{1/\sqrt{n-3}} = (z - \zeta)\sqrt{n-3}$$

en la que z y ζ son las transformaciones z de r y ρ, respectivamente. De nuevo comparamos t_s con $t_{s(\infty)}$ o lo buscamos en la tabla **II**. En la tabla **VIII** encontramos

Para $r = 0{,}837$ $\quad z = 1{,}2111$

Para $\rho = 0{,}500$ $\quad z = 0{,}5493$

Por consiguiente

$$t_s = (1{,}2111 - 0{,}5493)(\sqrt{497}) = 14{,}7538$$

La probabilidad de obtener este valor de t_s por muestreo al azar es $P < 10^{-9}$ (ver tabla **II**). Es sumamente improbable que la correlación paramétrica entre los nervios del ala derecha e izquierda sea 0,5.

Límites de confianza

Si $n > 50$, podemos establecer límites de confianza para r utilizando la transformación z. Primero convertimos el r de muestreo en z, fijamos límites de confianza para este z, y luego volvemos a cambiar estos límites a la escala r. Hallaremos límites de confianza del 95 % para los anteriores datos de longitud de los nervios del ala.

Para $r = 0{,}837, z = 1{,}2111; \alpha = 0{,}05$

$$L_1 = z - t_{s(\infty)\alpha_s} = z - \frac{t_{s(\infty)}}{\sqrt{n-3}} = 1.2111 - \frac{1{,}960}{22{,}2053}$$

$$= 1{,}2111 - 0{,}0879 = 1{,}1232$$

$$L_2 = z + \frac{t_{s(\infty)}}{\sqrt{n-3}} = 1{,}2111 + 0{,}0879 = 1{,}2990$$

Volvemos a cambiar estos valores z a la escala r hallando los correspondientes argumentos para la función z en la tabla **X**.

$L_1 \approx 0{,}808$ $\quad$ y $\quad$ $L_2 \approx 0{,}862$

son los límites de confianza del 95 % en torno a $r = 0{,}837$.

Prueba de la diferencia entre dos coeficientes de correlación

Para dos coeficientes de correlación podemos probar $H_0 : \rho_1 = \rho_2$ en contraste con $H_1 : \rho_1 \neq \rho_2$ del modo siguiente:

$$t_s = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Como $z_1 - z_2$ se distribuye normalmente y estamos utilizando una desviación típica paramétrica, comparamos t_s con $t_{s(\infty)}$ o empleamos la tabla **II**, áreas de la curva normal.

CUADRO 12.2 (continuación)

Por ejemplo, Sokoloff (1966) encontró que la correlación entre peso corporal y longitud del ala en *Drosophila pseudoobscura* era 0,552 en una muestra de $n_1 = 39$ en el Gran Cañón y 0,665 en una muestra de $n_2 = 20$ en Flagstaff, Arizona.

Gran Cañón: $z_1 = 0{,}6213$ $\quad$ Flagstaff: $z_2 = 0{,}8017$

$$t_s = \frac{0{,}6213 - 0{,}8017}{\sqrt{s_1^2 + s_2^2}} = \frac{-0{,}1804}{\sqrt{0{,}096601}} = \frac{-0{,}1804}{0{,}29426} = -0{,}6130$$

Por interpolación lineal en la tabla **II**, hallamos que la probabilidad de que un t_s esté entre $\pm 0{,}6130$ es alrededor de $2(0{,}22941) = 0{,}45882$, de este modo no tenemos prueba con la que rechazar la hipótesis nula.

A continuación, en el cuadro 12.2 vemos como fijar límites de confianza para el coeficiente de correlación de muestreo r. Esto se hace por medio de la transformación z; dará por resultado límites de confianza asimétricos cuando éstos se vuelvan a cambiar a la escala-r, como al establecer límites de confianza con variables sometidas a transformaciones raíz cuadrada o logarítmicas.

El ejemplo final ilustrado en el cuadro 12.2 es una prueba de significación de la diferencia entre dos coeficientes de correlación de muestreo. Se calcula un error estándar para la diferencia y se contrasta con una tabla de áreas de la curva normal. En el ejemplo se ha examinado la correlación entre peso corporal y longitud del ala en dos poblaciones de *Drosophila*, y se ha encontrado no significativa la diferencia entre coeficientes de correlación de las dos poblaciones. La fórmula dada es una aproximación aceptable cuando la menor de las dos muestras es mayor que 25. Se utiliza frecuentemente con tamaño de muestreo aún menor, como se indica en nuestro ejemplo del cuadro 12.2.

12.4 Aplicaciones de la correlación

El objeto del análisis de correlación es medir la intensidad de asociación observada entre cualquier par de variables y probar si es mayor de lo que podría esperarse sólo por azar. Una vez establecida esta asociación, es probable que conduzca a razonar sobre las relaciones causales entre las variables. A los estudiantes de estadística se les dijo desde el principio que no confundiesen correlación significativa con causalidad. También se nos ha advertido acerca de las llamadas correlaciones sin sentido, siendo un caso bien conocido la correlación positiva entre el número de ministros baptistas y el consumo de bebidas alcohólicas per cápita en ciudades con poblaciones de más de 10.000 en los Estados Unidos. Los casos individuales de correlación deben analizarse detenidamente antes de sacar conclusiones de ellos. Es conveniente distinguir las correlaciones en las que una variable es la causa total, o más probablemente, la causa parcial de la otra, de aquellas en

que las dos variables correlacionadas tienen una causa común y de situaciones más complicadas que incluyen ambas, influencia directa y causas comunes. El establecimiento de una correlación significativa no nos dice cuál de los muchos modelos estructurales posibles es adecuado. Se requiere otro análisis para discriminar entre los diversos modelos.

La distinción tradicional de correlación real frente a ilusoria o sin sentido es de poca utilidad. En las correlaciones supuestamente lógicas, se sabe o al menos se cree que las conexiones causales se comprenden claramente. En las llamadas correlaciones ilusorias, no puede encontrarse ninguna conexión razonable entre las variables, o si se demuestra una, no es de verdadero interés o puede demostrarse que es un artefacto del procedimiento de muestreo. Así la correlación entre ministros baptistas y consumo de alcohol es simplemente una consecuencia del tamaño de la ciudad. Cuanto más grande sea la ciudad, más ministros baptistas contendrá por término medio y mayor será el consumo de bebidas alcohólicas. La correlación es de poco interés para cualquiera que estudie la distribución de ministros o el consumo de alcohol. Ciertas correlaciones tienen como factor común el tiempo, y los procesos que cambian con el tiempo es muy probable que estén frecuentemente correlacionados, no por causa de ningunas razones biológicas funcionales sino simplemente porque el cambio con el tiempo en las dos variables que se consideran resulta ser en la misma dirección. Así, el tamaño de una población de insectos formada a lo largo del verano puede estar correlacionado con la altura de ciertas hierbas, pero esto puede ser simplemente una función del paso del tiempo. Puede no haber relación ecológica entre la planta y los insectos.

Quizás las únicas correlaciones propiamente llamadas sin sentido o ilusorias, son las adoptadas por convicción popular o intuición científica, las cuales, al probarlas por la metodología estadística apropiada utilizando tamaños de muestreo adecuados, se encuentra que son no significativas. Así, si podemos demostrar que no hay correlación significativa entre la cantidad de ácidos grasos saturados ingeridos y el grado de arterioesclerosis, podemos considerar esto como una correlación ilusoria. Recuérdese además que al probar la significación de correlaciones a niveles convencionales de significación, se debe tener en cuenta el error tipo I, que conducirá a que un cierto porcentaje de correlaciones se juzguen significativas cuando en realidad el valor paramétrico de $\rho = 0$.

Los coeficientes de correlación tienen una historia de amplia utilidad y aplicación que data de la escuela de biometría inglesa de finales de siglo. En recientes años se ha visto una aplicación algo menor de esta técnica conforme se han hecho experimentales campos crecientes de la investigación biológica. En experimentos en los cuales se varía un factor y se examina la respuesta de otra variable a la variación deliberada del primero, el método de regresión es más apropiado, como ya se ha discutido. No obstante, grandes áreas de la biología y de otras ciencias se quedan en donde el método experimental no es satisfactorio porque las variables no pueden someterse al control del investigador. Hay muchas áreas de la ecología, sistemática, evolución, y otros campos en los cuales los métodos experimentales son difíciles de aplicar. Hasta ahora no puede controlarse el estado de la atmósfera ni alterarse los factores evolutivos históricos. Sin embargo, necesitamos comprender los mecanismos científicos en que se basan estos fenómenos tanto como los de bioquímica o embriología experimental. En estos casos, el análisis de correlación sirve como una primera tónica descriptiva que estima los grados de asociación entre las variables incluidas.

12.5 Coeficiente de correlación por rangos de Kendall

A veces, aunque se sabe que los datos no son bivariantes normalmente distribuidos, no obstante, queremos probar la significación de la asociación entre las dos variables. Un método de analizar estos datos es alineando las variables y calculando un coeficiente de correlación por rangos. Este método pertenece a la familia general de métodos no paramétricos que vimos en el capítulo 10, en el que descubrimos métodos para análisis de variantes clasificables paralelamente al análisis de la varianza. En otros casos especialmente a propósito para métodos de graduación, no podemos medir la variable en una escala absoluta, sino solamente en una escala ordinal. Esto es típico de datos en los cuales estimamos acción relativa, como al asignar posiciones en una clase. Podemos decir que A es el mejor estudiante, B el mejor después del primero, C y D son los dos iguales entre sí y los mejores que siguen, y así sucesivamente. Dos instructores pueden ordenar independientemente un grupo de estudiantes y luego podemos comprobar si estas dos series de graduaciones están correlacionadas, como debería ocurrir si los juicios de los instructores estuviesen basados en la evidencia objetiva. Los ejemplos que siguen son de mayor interés biológico. Pudiéramos querer correlacionar el orden de eclosión en una muestra de insectos con una graduación en tamaño, o el orden de germinación en una muestra de plantas con el orden lineal de floración. Un genetista podría predecir el rango de acción de una serie de n genotipos que sintetiza y desearía demostrar la correlación de su predicción con los rangos de la acción realizada por estos genotipos. Un taxonomista podría querer ordenar n organismos desde los más parecidos a la forma X a los menos parecidos. Una disposición similar preparada por un segundo taxonomista ¿estará significativamente correlacionada con la primera? es decir, ¿están correlacionados los juicios taxonómicos de los dos observadores?

En el cuadro 12.3 presentamos el *coeficiente de correlación por rangos de Kendall*, simbolizado generalmente por τ (tau), aunque es un estadístico de muestreo, no un parámetro. La fórmula para el coeficiente de correlación lineal de Kendall es $\tau = N/n(n-1)$, donde n es el tamaño de muestreo convencional y N es un número de filas, que puede obtenerse de varias maneras. Una segunda variable Y_2, perfectamente correlacionada con la primera variable Y_1, debería estar en el mismo orden que las variantes Y_1. Sin embargo, si la correlación no es completa, el orden de las variantes Y_2 no corresponderá totalmente al de las Y_1. La cantidad N mide hasta qué punto la segunda variable se corresponde con el rango de la primera. Tiene un valor máximo de $n(n-1)$ y un valor mínimo de $-n(n-1)$. El siguiente ejemplo clarificará esto. Supongamos que tenemos una muestra de cinco individuos que se han ordenado por el rango de la variable Y_1:

Y_1	1	2	3	4	5
Y_2	1	3	2	5	4

Obsérvese que la graduación para la variable Y_2 no es totalmente concordante con la de Y_1. La técnica empleada en el cuadro 12.3 es contar el número de rangos superiores que siguen a un rango dado, sumar esta cantidad para todos los rangos, multiplicar la suma $\sum_{i}^{n} C_i$ por cuatro, y restarle un factor de corrección $n(n-1)$ para obtener un estadístico N. Para la variable Y_1 hallamos $\sum C_i = 4+3+2+1+0 = 10$, luego calculamos $N =$

CUADRO 12.3

Coeficiente de correlación por rangos de Kendall, τ.

Cálculo de un coeficiente de correlación por rangos entre la longitud total (Y_1) de 15 hembras apomícticas de áfidos y la longitud media del tórax (Y_2) de su descendencia partenogenética (basada en la medida de cuatro formas aladas): $n = 15$ pares de observaciones.

(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Hembra apomíctica	Y_1	R_1	Y_2	R_2	*Hembra apomíctica*	Y_1	R_1	Y_2	R_2
1	8,7	8	5,95	9	8	6,5	2	4,18	1
2	8,5	6	5,65	4	9	6,6	3	6,15	13
3	9,4	9	6,00	10	10	10,6	12	5,93	8
4	10,0	10	5,70	6,5	11	10,2	11	5,70	6,5
5	6,3	1	4,70	2	12	7,2	4	5,68	5
6	7,8	5	5,53	3	13	8,6	7	6,13	12
7	11,9	15	6,40	15	14	11,1	13	6,30	14
					15	11,6	14	6,03	11

Fuente: Datos de un estudio más amplio de R. R. Sokal.

Etapas del cálculo

1. Ordenar las variables Y_1 e Y_2 por separado y luego reemplazar las variantes originales por los rangos (asignar rangos ligados si es necesario). Estos rangos se alistan en las columnas (3) y (5). En la variable Y_2 había un ligamiento, así a las variantes 4.ª y 11.ª se les asignó un rango medio de 6,5.
2. Anotar por orden los n rangos de una de las dos variables, apareados con los valores del rango asignados a la otra variable (como se muestra más abajo). Si solamente una variable tiene ligamientos, ordenar las parejas con la variable sin ligamientos (como en el ejemplo presente). Si las dos variables tienen ligamientos, no importa cuál de las variables se ordene.
3. Obtener una suma de los recuentos C_i, como sigue. Examinar el primer valor de la columna de rangos apareada con la columna ordenada. En nuestro caso, éste es el rango 2. Contar todos los rangos siguientes a éste que son superiores al rango que se está considerando. Así, en este caso, contar todos los rangos mayores que 2. Hay catorce rangos que siguen al 2 y todos ellos excepto el rango 1 son mayores que 2. Por consiguiente, $C_1 = 13$. Ahora consideramos el rango siguiente (rango 1) y encontramos que los trece rangos subsiguientes son superiores a él; por lo tanto, C_2 es también igual a 13. Sin embargo, C_3 es solamente 2 ya que sólo los rangos 14 y 15 son mayores que el rango 13. Continuar de esta manera, considerando sucesivamente cada rango de la variable y contar el número de rangos superiores subsiguientes. Este cálculo puede hacerse mentalmente, pero lo mostramos explícitamente a continuación para que el método esté completamente claro. En los casos de ligamiento, contar ½. Así, para C_{10}, hay 4 ½ rangos superiores al primer rango de 6,5.

CUADRO 12.3 (continuación)

R_1	R_2	*Rangos subsiguientes mayores que el rango fundamental* R_2	*Recuentos* C_i
1	2	13, 5, 3, 4, 12, 9, 10, 6,5, 6,5, 8, 14, 11, 15	13
2	1	13, 5, 3, 4, 12, 9, 10, 6,5, 6,5, 8, 14, 11, 15	13
3	13	14, 15	2
4	5	12, 9, 10, 6,5, 6,5, 8, 14, 11, 15	9
5	3	4, 12, 9, 10, 6,5, 6,5, 8, 14, 11, 15	10
6	4	12, 9, 10, 6,5, 6,5, 8, 14, 11, 15	9
7	12	14, 15	2
8	9	10, 14, 11, 15	4
9	10	14, 11, 15	3
10	6,5	(6,5), 8, 14, 11, 15	4½
11	6,5	8, 14, 11, 15	4
12	8	14, 11, 15	3
13	14	15	1
14	11	15	1
15	15		0
			$78\frac{1}{2} = \sum^{n} C_i$

Necesitamos pues la siguiente cantidad:

$$N = 4\sum^{n} C_i - n(n-1) = 4(78\tfrac{1}{2}) - 15(14) = 314 - 210 = 104$$

4. El coeficiente Kendall de correlación por rangos, τ, puede hallarse del modo siguiente:

$$\tau = \frac{N}{\sqrt{\left[n(n-1) - \sum^{m} T_1\right]\left[n(n-1) - \sum^{m} T_2\right]}}$$

donde $\sum^{m} T_1$ y $\sum^{m} T_2$ son las sumas de los términos de corrección para ligamientos en los rangos de la variable Y_1 e Y_2, respectivamente, definidas a continuación. Se calcula un valor T igual a $t(t-1)$ para cada grupo de t variantes ligadas y se suman estos m grupos. En nuestro ejemplo $\sum^{m} T_1 = 0$, puesto que no había ligamientos en los rangos R_1, $\sum^{m} T_2 = 2$ porque hay $m = 1$ grupo de $t = 2$ rangos ligados R_2. $T = t(t-1) = 2(2-1) = 2$. Si hubiese habido más grupos de ligamientos, hubiéramos sumado los valores de T.

$$\tau = \frac{104}{\sqrt{[15(14)][15(14)-2]}} = \frac{104}{\sqrt{210(208)}} = \frac{104}{\sqrt{43\,680}}$$

$$= \frac{104}{208{,}9976} = 0{,}4976$$

Si no hay ligamientos, la ecuación puede simplificarse a

$$\tau = \frac{N}{n(n-1)}$$

CUADRO 12.3 (continuación)

5. Para la prueba de significación con tamaños de muestreo > 10, podemos valernos de una aproximación normal a la prueba de la hipótesis nula, de que el verdadero valor de $\tau = 0$.

$$t_s = \frac{\tau}{\sqrt{2(2n+5)/9n(n-1)}} = \frac{0{,}4976}{\sqrt{2[2(15)+5]/9(15)(14)}}$$

$$= \frac{0{,}4976}{\sqrt{70/1890}} = \frac{0{,}4976}{0{,}19253} = 2{,}59 \quad \text{comparado con } t_{\alpha[\infty]}$$

Cuando se busca este valor en la tabla II (áreas de la curva normal) hallamos que la probabilidad de que este t_s aparezca por azar es 0,0096 (ambas colas).

Cuando n es $\leqslant 10$, la aproximación dada más arriba no es adecuada y debe utilizarse la tabla especial que se da a continuación. La tabla da los valores críticos de N (el numerador de τ) para $n = 5$ a 10, al 5 y 1 % (dos colas). Estos solamente son exactos si no hay ligamientos. Si hay ligamientos, debe consultarse una tabla especial (véase Burr, 1960).

Valor crítico de N tal que $P(N_s > N_\alpha) \leqslant \alpha$ donde N_s y N_α se refieren a estimaciones de muestreo y valores críticos de N, respectivamente. Esta tabla está basada en la tabla Q de Siegel (1956).

	N_α	
n	$\alpha = 0{,}05$	$\alpha = 0{,}01$
5	20	—
6	26	30
7	30	38
8	36	44
9	40	52
10	46	58

$= 4\sum C_i - n(n-1) = 40 - 5(4) = 20$, para obtener el máximo recuento posible $N = n(n-1) = 20$. Evidentemente, una vez ordenada Y_1 siempre es perfectamente concordante consigo misma. Sin embargo, para Y_2 solamente obtenemos $\sum C_i = 4+2+1+1+0+0 = 8$, y $N = 4(8) - 5(4) = 12$. Como el máximo valor de N es $n(n-1) = 20$ y el valor observado 12, se sugiere un coeficiente obvio tal como $N/n(n-1) = [4\sum C_i - n(n-1)]/n(n-1) = 12/20 = 0{,}6$. Los ligamentos presentan complicaciones secundarias que se tratan en el cuadro 12.3. En este cuadro se trata de la correlación entre el tamaño corporal total de las hembras apomícticas de áfidos y la longitud media del tibrax de su descendencia. En este caso, no había especial necesidad de recurrir a la correlación por rangos a menos que haya cierta evidencia de que estos datos sean bimodales y no normalmente distribuidos. La significación de τ para tamaños de muestreo superiores a 10 puede

probarse fácilmente por el medio de un error estándar que aparece en el cuadro 12.3. En los casos en que el tamaño de muestreo es menor que 10, se buscan los valores críticos de N al final del cuadro 12.3.

Ejercicios 12

12.1 Representar gráficamente los datos siguientes en forma de un diagrama de esparcimiento bivariante. Calcular el coeficiente de correlación y fijar intervalos de confianza al 95 % para ρ. Los datos se recogieron de un estudio de variación geográfica en el áfido *Pemphigus populi-transversus*. Los valores de la tabla representan medias de localidad basadas en tamaños de muestreo idénticos para 23 localidades del Este de Norteamérica. Las variables, obtenidas de Sokal y Thomas (1965) se expresan en milímetros. Y_1 = longitud de la tibia, Y_2 = longitud del tarso. El coeficiente de correlación estimará la correlación de estas dos variables sobre localidades.

Número de código de localidad	Y_1	Y_2
1	0.631	0.140
2	.644	.139
3	.612	.140
4	.632	.141
5	.675	.155
6	.653	.148
7	.655	.146
8	.615	.136
9	.712	.159
10	.626	.140
11	.597	.133
12	.625	.144
13	.657	.147
14	.586	.134
15	.574	.134
16	.551	.127
17	.556	.130
18	.665	.147
19	.585	.138
20	.629	.150
21	.671	.148
22	.703	.151
23	.662	.142

SOLUCION. $r = 0{,}910$.

12.2 Los datos que siguen se han tomado de un amplio estudio de Brower (1959) sobre especiación en un grupo de mariposas de cola ahorquillada. Las medidas morfológicas se dan en milímetros X 8.

Especie	*Número de ejemplar*	Y_1 *Longitud del 8.º tergila*	Y_2 *Longitud del superuncus*
Papilo multicaudatus	1	24.0	14.0
	2	21.0	15.0
	3	20.0	17.5
	4	21.5	16.5
	5	21.5	16.0
	6	25.5	16.0
	7	25.5	17.5
	8	28.5	16.5
	9	23.5	15.0
	10	22.0	15.5
	11	22.5	17.5
	12	20.5	19.0
	13	21.0	13.5
	14	19.5	19.0
	15	26.0	18.0
	16	23.0	17.0
	17	21.0	18.0
	18	21.0	17.0
	19	20.5	16.0
	20	22.5	15.5
Papilo rutulus	21	20.0	11.5
	22	21.5	11.0
	23	18.5	10.0
	24	20.0	11.0
	25	19.0	11.0
	26	02.5	11.0
	27	19.5	11.0
	28	19.0	10.5
	29	21.5	11.0
	30	20.0	11.5
	31	21.5	10.0
	32	20.5	12.0
	33	20.0	10.5
	34	21.5	12.5
	35	17.5	12.0
	36	21.0	12.5
	37	21.0	11.5
	38	21.0	12.0
	39	19.5	10.5
	40	19.0	11.0
	41	18.0	11.5
	42	21.5	10.5
	43	23.0	11.0
	44	22.5	11.5
	45	19.0	13.0
	46	22.5	14.0
	47	21.0	12.5

Calcular el coeficiente de correlación para cada especie por separado y hacer la prueba de significación de cada uno. Probar si los dos coeficientes de correlación difieren significativamente.

12.3 Probar la presencia de asociación entre longitud de la tibia y longitud del tarso en los datos del ejercicio 12.1 utilizando el coeficiente de correlación por rangos de Kendall.

12.4 La siguiente tabla de datos es de un estudio morfométrico no publicado del chopo de La Carolina *Populus deltoides*, realizado por T.J. Crovello. Se midieron veintiséis hojas de un árbol cuando estaban frescas y después de secarse. Las variables representadas son: anchura de la hoja fresca (Y_1) y anchura de la hoja seca (Y_2), ambas en milímetros. Calcular r y probar su significación.

Y_1	Y_2
90	88
88	87
55	52
100	95
86	83
90	88
82	77
78	75
115	109
100	95
110	105
84	78
76	71
100	97
110	105
95	90
99	98
104	100
92	92
80	82
110	106
105	97
101	98
95	91
80	76
103	97

Capítulo 13

Análisis de frecuencias

Hasta aquí casi todo nuestro trabajo ha tratado de estimación de parámetros y prueba de hipótesis en variables continuas. Este capítulo trata de un tipo importante de casos, las pruebas de hipótesis sobre frecuencias. Las variables biológicas pueden distribuirse en dos o más clases, dependiendo de algún criterio tal como límites de clase arbitrarios en una variable continua o una serie de atributos mutuamente exclusivos. Un ejemplo del primero sería una distribución de frecuencias de pesos de nacimiento (una variable continua dividida arbitrariamente en un número de clases contiguas); uno del segundo sería una distribución de frecuencias cualitativa tal como la frecuencia de individuos de diez especies diferentes obtenidos de una muestra de suelo. Para cualquier distribución de este tipo podemos hipotetizar que se ha muestreado de una población en que las frecuencias de las diversas clases representan ciertas proporciones paramétricas de la frecuencia total. Necesitamos una prueba de bondad de ajuste de nuestra distribución de frecuencias observada, a la distribución esperada que representa nuestra hipótesis. Recuérdese que anteriormente vimos la necesidad de esta prueba en los capítulos 4 y 5, donde calculábamos distribuciones de frecuencias esperadas binomiales, de Poisson y normales pero éramos incapaces de decidir si una distribución de muestreo observada se alejaba significativamente de la teórica.

En la sección 13.1 introducimos la idea de bondad de ajuste, discutimos los tipos de prueba de significación que son apropiados, el fundamento básico de dichas pruebas y desarrollamos fórmulas de cálculo generales para ellas.

La sección 13.2 ilustra los cálculos reales para bondad de ajuste cuando los datos están ordenados por un solo criterio de clasificación, como en una distribución de frecuencias cuantitativa o cualitativa. Esto se aplica a los casos que se espera sigan una de las bien conocidas distribuciones de frecuencias tales como la distribución binomial, de Poisson o normal, así como a las distribuciones esperadas que sigan alguna otra ley sugerida por la materia científica que se investiga, tal como, por ejemplo, prueba de bondad de ajuste de frecuencias mendelianas esperadas.

En la sección 13.3 pasamos a discutir pruebas de significación de frecuencias en clasificaciones de doble entrada, llamadas pruebas de independencia. Discutiremos las pruebas ordinarias de tablas 2×2, en las cuales cada uno de los dos criterios de clasificación divide las frecuencias en dos clases, dando una tabla de cuatro casillas, así como tablas $F \times C$ con más filas y columnas.

A lo largo de este capítulo realizamos pruebas de bondad de ajuste por medio del estadístico G. Aludimos brevemente a las pruebas ji-cuadrado, pero tal como se explica en diversas partes a lo largo del texto, G tiene ventajas teóricas sobre X^2, siendo además más sencillo de calcular para pruebas de independencia.

13.1 Pruebas de bondad de ajuste: introducción

La idea básica de una prueba de bondad de ajuste es fácil de comprender, dada la gran experiencia que se tiene ya con el contraste de hipótesis estadísticas. Vamos a suponer que un genetista ha realizado un experimento de cruzamiento entre dos híbridos F_1 y obtiene una progenie F_2 de 90 crías, de las cuales 80 son de tipo salvaje y 10 son mutantes. El genetista supone dominancia y espera una proporción de fenotipos 3 : 1. Sin embargo, cuando calculamos las proporciones reales observamos que los datos están en una proporción $80/10 = 8 : 1$. Los valores esperados de p y q son $\hat{p} = 0{,}75$ y $\hat{q} = 0{,}25$ para el tipo salvaje y mutante, respectivamente. Obsérvese que utilizamos el signo de intercalación (llamado generalmente "sombrerito" en estadística) para indicar valores hipotéticos o esperados de las proporciones binomiales. Sin embargo, las proporciones observadas de estos dos tipos son $p = 0{,}89$ y $q = 0{,}11$ respectivamente. Otra forma más de observar el contraste entre lo observado y lo esperado es exponerlo en frecuencias: las frecuencias observadas son 80 y 10 para los dos fenotipos. Las frecuencias esperadas serían $\hat{f}_1 = \hat{p}n = 0{,}75(90) = 67{,}5$ y $\hat{f}_2 = \hat{q}n = 0{,}25(90) = 22{,}5$, respectivamente, en donde n se refiere al tamaño de muestreo de la descendencia del cruzamiento. Nótese que cuando sumamos las frecuencias esperadas dan $67{,}5 + 22{,}5 = n = 90$, tal como debería ser.

La cuestión obvia que acude a la mente es si la desviación de la hipótesis 3 : 1 observada en nuestra muestra es de tal magnitud como para resultar improbable. En otras palabras, ¿difieren los datos observados de los esperados lo suficiente como para hacer que rechacemos la hipótesis nula? Para el caso que se acaba de considerar, ya se conocen dos métodos para llegar a una decisión sobre la hipótesis nula. Naturalmente, ésta es una distribución binomial en la que p es la probabilidad de que sea un tipo salvaje y q es la probabilidad de que sea un mutante. Es posible hallar la probabilidad de obtener un resultado de 80 tipo salvaje y 10 mutantes así como todos los casos "peores" para $\hat{p} = 0{,}75$ y $\hat{q} = 0{,}25$, y una muestra de $n = 90$ descendientes. Utilizamos aquí la expresión binomial convencional $(\hat{p} + \hat{q})^n$ con la salvedad de que p y q son hipotéticos, y reemplazamos el símbolo k por n, el cual aceptábamos en el capítulo 4 como el símbolo correspondiente a la suma de todas las frecuencias en una distribución de frecuencias. En este ejemplo tenemos una muestra solamente, por tanto lo que ordinariamente sería k en la binomial es, al mismo tiempo, n. Este problema fue expuesto en la tabla 4.3 y sección 4.2, y podemos calcular la probabilidad acumulativa de la cola de la distribución

binomial. Cuando se hace esto, obtenemos una probabilidad de 0,00085 para todos los resultados tan alejados o más de la hipótesis. Nótese que ésta es una prueba de una cola, siendo la hipótesis alternativa que, de hecho, hay más descendientes tipo salvaje de los que postularía la hipótesis mendeliana. Suponiendo $\hat{p} = 0{,}75$ y $\hat{q} = 0{,}25$, la muestra observada es en consecuencia un suceso muy raro, y concluimos que hay una desviación significativa de lo esperado.

Un camino más rápido basado en el mismo principio es buscar límites de confianza para las proporciones binomiales como se hizo para la prueba del signo en la sección 10.3. La interpolación en la tabla IX indica que para una muestra de $n = 90$, un porcentaje observado de 89 % daría límites de confianza aproximados al 99 % de 78 y 96 para el porcentaje real de individuos tipo salvaje. Sin duda el valor hipotetizado de $\hat{p} = 0{,}75$ está fuera de los límites de confianza del 99 %.

Ahora vamos a desarrollar un tercer método mediante una prueba de bondad de ajuste. La tabla 13.1 muestra cómo podríamos proceder. En la primera columna se dan las frecuencias observadas f que representan el resultado del experimento. La columna (2) muestra las frecuencias esperadas $\hat{f}$ basadas en la hipótesis particular que se contrasta. En este caso, la hipótesis es una proporción 3 : 1 y ya hemos calculado las frecuencias esperadas bajo estas condiciones como $\hat{f}_1 = \hat{p}n = 0{,}75(90) = 67{,}5$ y $\hat{f}_2 = \hat{q}n = 0{,}25(90) = 22{,}5$.

TABLA 13.1

Desarrollo de la prueba ji-cuadrado para bondad de ajuste. Frecuencias observadas y esperadas del resultado de un cruzamiento genético, suponiendo una proporción 3:1 de fenotipos entre la descendencia.

	(1)	(2)	(3)	(4)	(5)
Fenotipos	*Frecuencias observadas* f	*Frecuencias esperadas* $\hat{f}$	*Desviaciones de lo esperado* $f - \hat{f}$	*Cuadrado de las desviaciones* $(f - \hat{f})^2$	$\frac{(f - \hat{f})^2}{\hat{f}}$
Tipo salvaje	80	$\hat{p}n = 67{,}5$	12,5	156,25	2,315
Mutante	10	$\hat{q}n = 22{,}5$	−12,5	156,25	6,944
Suma	90	90,0	0		$X^2 = 9{,}259$

¿Cómo podemos desarrollar un estadístico para probar en qué medida difieren las frecuencias observadas en la columna (1) de las frecuencias esperadas en la columna (2)? La siguiente prueba estadística es fácil de comprender y su estructura tiene un sentido lógico. Primero medimos $f - \hat{f}$, la desviación de las frecuencias observadas respecto a las esperadas. Obsérvese que la suma de estas desviaciones es igual a cero, por razones muy similares a las que hacen que la suma de las desviaciones de una media sumen cero. De aquí que en un ejemplo con dos clases, las desviaciones sean siempre iguales y de signo opuesto.

Siguiendo nuestro método previo de elevar al cuadrado todas las desviaciones para hacerlas positivas, elevamos al cuadrado $(f - \hat{f})$ en la columna (4) para dar una medida de la magnitud de la desviación de lo que se espera. Esta cantidad debe expresarse como una proporción de la frecuencia esperada. Después de todo, si la frecuencia esperada fuese 13,0, una desviación de 12,5 sería extremadamente grande, comprendiendo casi el 100 % de $\hat{f}$, pero esta desviación representaría solamente el 10 % de una frecuencia esperada de 125,0. Así, obtenemos la columna (5) como el cociente de la cantidad de la columna (4) dividida por la de la columna (2). Es de notar que la magnitud del cociente es mayor para la segunda línea, en la cual $\hat{f}$ es menor. El próximo paso en el desarrollo de nuestra prueba estadística es sumar estos cocientes, lo cual se hace al pie de la columna (5), dando un valor de 9,259.

¿Cómo denominaremos a este nuevo estadístico? Aquí tenemos ciertos problemas de nomenclatura. Ya se habrá reconocido ésta como la llamada *prueba ji-cuadrado*, normalmente enseñada al comenzar las clases de genética. El nombre de la prueba está demasiado bien establecido como para que un cambio resulte práctico, pero de hecho esta cantidad que acabamos de calcular, la suma de la columna (5), posiblemente podría no ser un χ^2. Esta es una distribución de frecuencias continua y teórica, mientras nuestra cantidad 9,259 es un estadístico de muestreo basado en frecuencias discretas. Este último punto se ve fácilmente si se conciben otros posibles resultados. Por ejemplo, podríamos haber obtenido tan pocos mutantes como cero equilibrados por 90 individuos tipo salvaje (suponiendo que el número total de descendientes $n = 90$ permanezca constante), o podríamos haber obtenido 1, 2, 3 o más mutantes, compensados en cada caso por el número exacto de descendientes tipo salvaje para dar un total de 90. Las frecuencias observadas cambian en incrementos unitarios, y como las frecuencias esperadas permanecen constantes, está claro que las desviaciones, sus cuadrados y los cocientes, no son variables continuas sino que sólo pueden adoptar ciertos valores.

La razón por la que esta prueba se ha denominado prueba ji-cuadrado y por la que muchas personas llaman ji-cuadrado al estadístico obtenido como la suma de la columna (5), es que la distribución de muestreo de esta suma se aproxima a la de una distribución ji-cuadrado con un grado de libertad. Podemos comprender la razón para el grado de libertad único cuando consideramos las frecuencias en las dos clases de la tabla y su suma, $80 + 10 = 90$. En este ejemplo la frecuencia total es fija. Por lo tanto, si variásemos la frecuencia de una cualquiera de las clases, la otra tendría que compensar los cambios en la primera clase para conservar un total correcto. Si la primera clase tuviese una frecuencia de 75, la segunda clase debería contener 15 ítems para hacer el total 90. Así, sólo puede variar libremente una clase, estando la otra clase limitada por la suma constante. Aquí el significado de *un grado de libertad* resulta muy claro. Una de las clases es libre de variar; la otra no. No obstante, como el estadístico de muestreo no es un ji-cuadrado, hemos seguido la costumbre,crecientemente en boga,de designar al estadístico de muestreo como X^2 en lugar de χ^2. El valor de $X^2 = 9{,}259$ de la tabla 13.1, cuando se compara con el valor crítico de χ^2 (tabla IV), es altamente significativo ($P < 0{,}005$). La prueba ji-cuadrado es siempre de una cola. Como las desviaciones están elevadas al cuadrado, tanto las positivas como las negativas conducen a valores positivos de X^2. Naturalmente, rechazamos la hipótesis 3 : 1 y concluimos que la proporción de tipo salvaje es mayor que

0,75. En consecuencia, el genetista debe buscar un mecanismo que explique esta desviación de lo esperado.

La prueba de bondad de ajuste puede aplicarse a una distribución con más de dos clases. El ejemplo de la parte 1 del cuadro 13.1 es un cruzamiento genético complicado en el que se espera una proporción 18:6:6:2:12:4:12:4. Se establece la tabla como antes, con las frecuencias esperadas calculadas como $\hat{f}_i = \hat{p}_i n$, en donde los valores de $\hat{p}_i$ son las probabilidades esperadas para las $a = 8$ clases. Naturalmente $\sum^{a} \hat{p}_i = 1$, puesto que las probabilidades de todos los resultados posibles suman uno. Los valores de $\hat{p}_i$ para este ejemplo son $\hat{p}_1 = \frac{18}{64}$, $\hat{p}_2 = \frac{6}{64}$, y así sucesivamente. Calculamos nuevamente las desviaciones de lo esperado, las elevamos al cuadrado, y las dividimos por las frecuencias esperadas. La operación puede describirse por la fórmula

$$X^2 = \sum^{a} \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i} \tag{13.1}$$

Sin embargo, su práctica es pesada y requiere el cálculo de las desviaciones, sus cuadrados, y su división por las frecuencias esperadas para el conjunto de las clases. Una fórmula de cálculo generalmente aplicable para X^2 puede deducirse fácilmente de la expresión (13.1), como se demuestra en el apéndice A1.10. Esta es

$$X^2 = \sum^{a} \frac{f_i^2}{\hat{f}_i} - n \tag{13.2}$$

que puede obtenerse fácilmente como la suma de los cocientes de los cuadrados de las frecuencias observadas divididos por sus frecuencias esperadas. De esta suma de cocientes se resta n, la suma de todas las frecuencias.

¿Qué podemos concluir acerca de nuestra prueba de bondad de ajuste? Si por el momento suponemos que X^2 en este caso también se distribuye aproximadamente como χ^2, necesitamos saber cuántos grados de libertad hay en este ejemplo para poder compararlo con la distribución χ^2 apropiada. En este caso son ocho clases, siete cualesquiera de ellas pueden variar libremente; pero las ocho clases deben constituir la diferencia entre la suma total y la suma de las siete primeras. Así, en un caso de ocho clases tenemos siete grados de libertad y, en general, cuando tenemos a clases, tenemos $a - 1$ grados de libertad. En el ejemplo de la parte 1 del cuadro 13.1, la aplicación de la expresión (13.2) lleva a $X^2 = 9{,}4909$, que es menor que $\chi^2_{0,05[7]} = 14{,}067$. Así pues las frecuencias observadas son compatibles con las proporciones postuladas.

Recientemente se ha introducido mucho una nueva prueba de bondad de ajuste. Se trata de la prueba G, basada en el estadístico de razón de verosimilitudes. Tiene varias ventajas sobre la más antigua ji-cuadrado y, por lo tanto, lo hemos destacado exclusivamente en este capítulo en lugar del método más antiguo.

Vamos a reconsiderar el ejemplo de la tabla 13.1. Utilizando la expresión (4.1) para las frecuencias relativas esperadas en una distribución binomial, podemos calcular dos cantidades que en este momento nos interesan:

CUADRO 13.1

Pruebas de bondad de ajuste. Clasificación simple, frecuencias esperadas basadas en hipótesis extrínsecas a los datos muestreados.

1. Frecuencias divididas en $a \geqslant 2$ clases.

En un experimento genético que incluye un cruzamiento entre dos variedades de la judía *Phaseolus vulgaris*, Smith (1933) obtuvo los siguientes resultados:

Fenotipos ($a = 8$)	*Frecuencias observadas* f	*Frecuencias esperadas* $\hat{f}$
Púrpura/ante	63	67,8
Púrpura/testáceo	31	22,6
Rojo/ante	28	22,6
Rojo/testáceo	12	7,5
Púrpura	39	45,2
Rojo, sangre de toro	16	15,1
Ante	40	45,2
Testáceo	12	15,1
Total	241	241,1

Las frecuencias esperadas $\hat{f}_i$ se calcularon basándose en la proporción esperada de 18:6:6:2:12:4:12:4. Calcular $\hat{f}_i$ como $\hat{p}_i n$. Así, $\hat{f}_1 = \hat{p}_1 n = (\frac{18}{64}) \times 241 = 67{,}8$.

Utilícese la expresión (13.4a) cuando los valores de $\hat{f}$ ya se hayan calculado o (13.5) cuando se den los valores de $\hat{p}$. Como en el ejemplo de judías dado más arriba hemos presentado los valores de $\hat{f}$, utilizaremos la expresión (13.4a):

$$G = 2[\sum^{a} f_i \ln f_i - 2{,}30259 \sum^{a} f_i \log \hat{f}_i]$$

Utilizando la tabla **XI** hallamos

$$\sum^{a} f_i \ln f_i = 63 \ln 63 + \cdots + 12 \ln 12 = 261{,}017 + \cdots + 29{,}819$$
$$= 855{,}206$$

Utilizando una tabla de logaritmos, podemos calcular

$$\sum^{a} f_i \log \hat{f}_i = 63 \log 67{,}8 + \cdots + 12 \log 15{,}1$$
$$= 63(1{,}83123) + \cdots + 12(1{,}17898) = 369{,}5282$$

Entonces G puede calcularse como

$$G = 2[\sum^{a} f_i \ln f_i - 2{,}30259 \sum^{a} f_i \log \hat{f}_i]$$
$$= 2[855{,}206 - 2{,}30259(369{,}5282)] = 2[4{,}334] = 8{,}668$$

Puesto que nuestro valor observado de $G = 8{,}668 < \chi^2_{0.05[7]} = 14{,}067$, podemos considerar los datos compatibles con la hipótesis nula.

CUADRO 13.1 (continuación)

2. Caso particular de frecuencias divididas en $a = 2$ clases.

En un cruzamiento F_2 en *Drosophila* se obtuvieron 176 descendientes, de los cuales 130 fueron moscas tipo salvaje y 46 mutantes ebony. Considerando que el mutante es un autosómico recesivo, se esperaría una razón de 3 moscas tipo salvaje por cada mutante.

Una complicación adicional la constituye el hecho de que cuando n es menor que 200 aproximadamente debería aplicarse una corrección de continuidad. Los cálculos se realizan tal como se describe más arriba en la parte **1**, pero utilizando valores de f ajustados en vez de los f originales, si hay que aplicar esta corrección (ver texto). Ahora aplicamos la prueba G a los datos de *Drosophila*.

Moscas	f	f *Ajustados*	*Hipótesis*	$\hat{f}$
Tipo salvaje	130	130,5	$\hat{p} = 0{,}75$	$\hat{p}n = 132{,}0$
Mutante ebony	46	45,5	$\hat{q} = 0{,}25$	$\hat{q}n = 44{,}0$
	176	176,0		176,0

Utilizando la tabla **XII**, y frecuencias f_i ajustadas,

$$\sum^{a} f_i \ln f_i = 130{,}5 \ln 130{,}5 + 45{,}5 \ln 45{,}5 = 635{,}714 + 173{,}706$$
$$= 809{,}420$$

Utilizando una tabla de logaritmos,

$$\sum^{a} f_i \log \hat{f}_i = 130{,}5 \log 132{,}0 + 45{,}5 \log 44{,}0$$
$$= 130{,}5(2{,}12057) + 45{,}5(1{,}64345) = 351{,}5114$$
$$G_{adj} = 2[809{,}420 - 2.30259(351{,}5114)] = 0{,}0668$$

que es evidentemente no significativa cuando se compara con $\chi^2_{0,05[1]}$.

$$C(90, 80)\left(\frac{80}{90}\right)^{80}\left(\frac{10}{90}\right)^{10} = 0{,}1327$$

$$C(90, 80)\left(\frac{3}{4}\right)^{80}\left(\frac{1}{4}\right)^{10} = 0{,}0005514$$

La primera cantidad es la probabilidad de observar los resultados muestreados (80 tipo salvaje y 10 mutantes) según la hipótesis de que $\hat{p} = p$, es decir, que el parámetro de población es igual a la proporción muestral observada. El segundo es la probabilidad de

observar los resultados muestreados suponiendo que $\hat{p} = \frac{3}{4}$, de acuerdo con la hipótesis nula mendeliana. Obsérvese que estas expresiones solamente dan las probabilidades para los resultados observados, *no para los resultados observados y todos los peores*. Así, $P = 0{,}0005514$ es menor que el valor calculado anteriormente de $P = 0{,}00085$, que es la probabilidad de 10 *y menos* mutantes, suponiendo $\hat{p} = \frac{3}{4}$, $\hat{q} = \frac{1}{4}$.

La primera probabilidad (0,1327) es mayor que la segunda (0,0005514), ya que el parámetro se basa en los datos observados. Si la proporción observada p es en realidad igual a la proporción $\hat{p}$ postulada según la hipótesis nula, las dos probabilidades calculadas serán iguales y su razón, L, será igual a 1,0. Cuando mayor sea la diferencia entre p y $\hat{p}$ (la proporción esperada bajo la hipótesis nula), mayor será la razón (la probabilidad basada en p se divide por la probabilidad basada en $\hat{p}$ o definida por la hipótesis nula). Esto indica que la razón de estas dos probabilidades o *verosimilitudes* puede utilizarse como un estadístico para medir el grado de ajuste entre frecuencias observadas y esperadas. Una prueba basada en esta razón se denomina *prueba de razón de verosimilitudes*. En nuestro caso, $L = 0{,}1327/0{,}0005514 = 240{,}66$. La distribución teórica de esta razón es en general compleja y escasamente conocida. No obstante, se ha demostrado que la distribución de

$$G = 2 \ln L = 2 (\ln 10) \log L \tag{13.3}$$

puede ser aproximada para la distribución χ^2 cuando los tamaños de muestreo son grandes (para una definición de "grande" en este caso, véase más adelante). Los grados de libertad apropiados para una determinada prueba son los mismos que para las pruebas ji-cuadrado discutidas anteriormente. En nuestro caso,

$$G = 2 \ln L = 2(\ln 10) \log L = 2(2{,}302585)(2{,}38140) = 2(5{,}48356) = 10{,}967$$

Si comparamos este valor observado con el de una distribución χ^2 con un grado de libertad, hallamos que el resultado es significativo ($P < 0{,}005$), como se ha visto anteriormente para la prueba ji-cuadrado. En general, G será numéricamente muy similar a X^2. La notación para la *prueba del logaritmo de la razón de verosimilitudes* está tan poco establecida como la de la prueba ji-cuadrado. A veces se utiliza para G el símbolo $2I$.

Para desarrollar una fórmula de cálculo, la expresión (13.3) para el estadístico G puede volver a escribirse de varias maneras, dependiendo de la aplicación particular. En el apéndice A1.11 se demuestra que para una prueba de bondad de ajuste general, puede escribirse como sigue:

$$G = 2\sum^{a} f_i \ln\left(\frac{f_i}{\hat{f}_i}\right) \tag{13.4}$$

Si se dispone de una calculadora electrónica con logaritmos, ésta es con mucho la fórmula más sencilla. En otros casos para simplificar el cálculo puede expresarse como sigue

$$G = 2[\sum^{a} f_i \ln f_i - \sum^{a} f_i \ln \hat{f}_i]$$
$$= 2[\sum^{a} f_i \ln f_i - (2{,}30259)\sum^{a} f_i \log \hat{f}_i] \tag{13.4a}$$

que es conveniente utilizar con la tabla **XI**. La tabla **XI** da $f \ln f$ para valores enteros de f entre 0 y 10 000. Así, las cantidades $f \ln f$ ó $n \ln n$ pueden buscarse directamente y acumularse en una máquina sumadora. Como existe la posibilidad de transcripción errónea de las tablas, es más conveniente utilizar una calculadora impresora para que solamente baste confrontar la inscripción impresa con la tabla para verificar los resultados. Si se encuentran errores es fácil corregir los resultados finales sin necesidad de repetir todos los cálculos. Otra fórmula de cálculo frecuentemente utilizada para G es

$$G = 2[\sum^{a} f_i \ln f_i - \sum^{a} f_i \ln \hat{p}_i - n \ln n] \qquad (13.5)$$

demostrada también en el apéndice A1.11.

13.2 Prueba de bondad de ajuste de clasificación simple

En el cuadro 13.1 ilustramos pruebas G de bondad de ajuste en los casos en que las frecuencias esperadas están basadas en una hipótesis extrínseca a los datos, hay a clases y las proporciones esperadas en cada clase se suponen basadas en el conocimiento exterior, no siendo funciones de parámetros estimados de la muestra. Comenzamos con el caso general para un número cualquiera de clases, donde el número de clases se simboliza por a, para resaltar la analogía con el análisis de varianza. Los datos son los resultados de un cruzamiento genético complicado que se espera conduzca a una proporción 18:6:6:2:12:4:12:4. Se obtuvo una progenie de 241 asignada a las ocho clases fenotípicas. Las frecuencias esperadas pueden calcularse sencillamente multiplicando el tamaño de muestreo total n por las probabilidades de ocurrencia esperadas. Observamos que el ajuste en conjunto es bastante bueno. Este es un ejemplo de una prueba de bondad de ajuste con una hipótesis extrínseca, porque la proporción genética probada está basada en consideraciones anteriores y externas a la muestra de observaciones examinada en el ejemplo. El cálculo es muy sencillo, como se indica en el cuadro.

En la sección anterior vimos que el valor de G ha de ser comparado con un valor crítico de χ^2 para $a - 1$ grados de libertad. Cuando comparamos nuestro resultado con la distribución χ^2, encontramos que el valor de G obtenido de nuestra muestra no es significativo. No tenemos suficiente evidencia para rechazar la hipótesis nula y concluimos que la muestra es concordante con la proporción genética especificada. No obstante, especialmente en ejemplos tales como el que se acaba de analizar, debe recordarse que no hemos especificado una hipótesis alternativa; hay una variedad de hipótesis alternativas que también podrían no ser excluidas si realizásemos una prueba de significación para ellas. Realmente no hemos comprobado que los datos se distribuyan como se ha especificado, y hay otras varias hipótesis de proporción genética que también podrían ser aceptables.

No se encuentran direcciones generales en la literatura con respecto a lo pequeña que puede ser una muestra y ser, no obstante, adecuada para pruebas G o X^2 de bondad de ajuste. Sin embargo, deberían tomarse precauciones al interpretar los resultados con tamaños de muestreo de n menor que 50.

El ejemplo de la parte 2 del cuadro 13.1 es un cruzamiento monohíbrido con una

proporción esperada de 3 tipo salvaje a 1 mutante. En las pruebas de bondad de ajuste que incluyen solamente dos clases, el valor de G tal como se ha calculado en las expresiones (13.4) ó (13.5) mostrará un sesgo que puede modificarse aplicando una *corrección de continuidad*, aproximando más estrictamente el valor de G a la distribución χ^2. Esta corrección consiste en sumar o restar 0,5 de las frecuencias observadas, de tal manera que se haga mínimo el valor de G. Simplemente se ajustan las f_i transformándolas para reducir la diferencia entre éstas y las frecuencias esperadas correspondientes en un medio. Entonces se utilizan las expresiones (13.4a) ó (13.5) como antes para obtener G_{aj}. Los valores de $(f + \frac{1}{2}) \ln (f + \frac{1}{2})$ necesarios para los cálculos de G_{aj} pueden encontrarse en la tabla **XII**. La corrección de continuidad se aplica siempre que $n < 200$. Cuando $n < 25$ incluso esta corrección es insuficiente para rectificar el sesgo. En tal caso es conveniente un cálculo exacto de las probabilidades binomiales como en la tabla 4.3. El bajo valor de G_{aj} encontrado en el cuadro 13.1 muestra que los datos observados se ajustan estrechamente a las proporciones esperadas.

En algunas pruebas de bondad de ajuste restamos más de un grado de libertad del número de clases, a. Estos son ejemplos en que los parámetros para la hipótesis nula han sido extraídos de los mismos datos de muestreo, en contraste con las hipótesis nulas encontradas hasta ahora (en la tabla 13.1 y el cuadro 13.1). En estos casos, la hipótesis a contrastar se ha originado basándose en el conocimiento general del problema específico y de la genética mendeliana por parte del investigador. Por esta razón, las frecuencias esperadas están basadas en una *hipótesis extrínseca* o hipótesis externa a los datos. En contraposición, consideremos las frecuencias de Poisson esperadas de células de levadura en un hemocitómetro (cuadro 4.1). Se recordará que para calcular estas frecuencias necesitábamos valores de μ, que se estimaban a partir de la media de muestreo $\bar{Y}$. Por consiguiente, el parámetro de la distribución de Poisson calculada procede de las mismas observaciones muestreadas. Las frecuencias de Poisson esperadas representan una *hipótesis intrínseca*. En este caso, para obtener el número correcto de grados de libertad para la prueba de bondad de ajuste (ji-cuadrado ó G) restaríamos de a, el número de clases en que se han agrupado los datos, no solamente un grado de libertad para n, la suma de las frecuencias, sino también un grado de libertad más para la estimación de la media. Así, en este caso, un estadístico de muestreo G se compararía con ji-cuadrado para $a - 2$ grados de libertad.

Cuando aplicamos la prueba G a las frecuencias observadas y esperadas del cuadro 4.1, al utilizar cualquier fórmula del cuadro 13.1, que es lo más conveniente desde el punto de vista del cálculo, obtenemos $G = 7{,}529$. Dos aspectos distinguen este cálculo del de la parte 1 del cuadro 13.1. Como regla general, eludimos las frecuencias esperadas menores que 5. Por lo tanto, las clases de $\hat{f}_i$ en la cola superior de la distribución son demasiado pequeñas. Las aumentamos sumando sus frecuencias a las de clases contiguas como se muestra en el cuadro 4.1. Naturalmente, las frecuencias observadas deben agruparse para que se equilibren. El número de clases a es el número *después* de la agrupación. En nuestro caso, $a = 6$.

La otra nueva característica ya se ha discutido. Es el número de grados de libertad considerado para la prueba de significación. Restamos siempre un grado de libertad para la suma prefijada (en este caso $n = 400$). No obstante, restamos un grado de libertad adicional para cada parámetro de la distribución de frecuencias esperada que se ha

estimado de la distribución muestreada. En este caso, estimamos μ de la muestra y, por consiguiente, se resta de a un segundo grado de libertad, haciendo el número final de grados de libertad $a - 2 = 6 - 2 = 4$. Comparando el valor muestral de $G = 7{,}529$ con el valor crítico de χ^2 para 4 grados de libertad lo hallamos no significativo. Por lo tanto, aceptamos la hipótesis nula y concluimos que las células de levadura se distribuyen al azar.

La prueba G para probar la bondad del ajuste de una serie de datos a una distribución de frecuencias esperada, puede aplicarse no solamente a la de Poisson sino también a la normal, binomial y otras distribuciones. Para una distribución normal, ordinariamente estimamos dos parámetros de los datos muestreados μ y σ. De aquí que los grados de libertad correspondientes sean $a - 3$. En la binomial solamente debe estimarse un parámetro, $\hat{p}$; los grados de libertad correspondientes son $a - 2$.

13.3 Pruebas de independencia: tablas de doble entrada

La noción de independencia estadística o probabilística se ha introducido primeramente en la sección 4.1, en donde se ha demostrado que si dos sucesos eran independientes, la probabilidad de que ocurriesen juntos podría calcularse como el producto de sus probabilidades por separado. Así, si entre la progenie de un cierto cruzamiento genético la probabilidad de que un grano de cereal sea rojo es $\frac{1}{2}$ y la probabilidad de que el grano sea dentado es $\frac{1}{3}$, la probabilidad de obtener un grano dentado y rojo sería $\frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$, si los sucesos unidos de estas dos características fuesen estadísticamente independientes.

La prueba estadística apropiada para este problema genético sería probar las frecuencias para bondad del ajuste a las proporciones esperadas de 2 (rojo, no dentado):2 (no rojo, no dentado):1 (rojo, dentado):1 (no rojo, dentado). Esta sería una prueba simultánea de dos hipótesis nulas: que las proporciones esperadas sean $\frac{1}{2}$ y $\frac{1}{3}$ para rojo y dentado, respectivamente, y que estas dos propiedades sean independientes. La primera hipótesis nula prueba el modelo mendeliano en general. La segunda prueba si estos caracteres se clasifican independientemente, es decir, si están determinados por genes situados en diferentes grupos de ligamiento. Si la segunda hipótesis debe ser rechazada, esto se considera como evidencia de que los caracteres están ligados, es decir, situados en el mismo cromosoma.

Hay numerosos ejemplos en biología en los cuales la segunda hipótesis acerca de la independencia de dos propiedades es de gran interés y la primera hipótesis sobre la verdadera proporción de una o ambas propiedades es de poco interés. En realidad, a veces no puede ser formulada por el investigador ninguna hipótesis respecto a los valores paramétricos $\hat{p}_i$. Citaremos varios ejemplos de estas situaciones, que conducen a la prueba de independencia que se estudia en esta sección. Utilizaremos esta prueba siempre que queramos probar si dos propiedades diferentes, cada una apareciendo en dos estados, son dependientes entre sí. Por ejemplo, los individuos de una cierta polilla pueden presentarse en dos fases de color, claro y oscuro. Cincuenta especímenes de cada fase pueden dejarse al descubierto expuestas a depredación por los pájaros. Después de un intervalo de tiempo prefijado se cuentan el número de polillas que sobreviven. La proporción atrapada puede diferir en las dos fases de color. En este ejemplo las dos propiedades son color y supervivencia. Podemos dividir nuestra muestra en cuatro clases: supervivientes de color claro, presas de color claro, supervivientes oscuros, y presas oscuras. Si la probabilidad de ser capturado es independiente del color de la polilla, las frecuencias esperadas de estas cuatro clases pueden calcularse sencillamente como productos independientes de la proporción de cada color (en nuestro experimento $\frac{1}{2}$) y la proporción total capturada en la muestra total. Si la prueba de independencia explicada más abajo demuestra que las dos propiedades no son independientes, debemos concluir que una de las fases de color es más susceptible de depredación que la otra. Este es un fenómeno biológico importante; las proporciones exactas de las dos propiedades son de poco interés en este caso. La proporción de las fases de color es arbitraria, y la proporción de supervivencia interesa solamente hasta donde difiera para las dos fases.

Un segundo ejemplo podría referirse a un experimento de muestreo realizado por un ecólogo vegetal. Obtiene una muestra al azar de 100 individuos de una especie de árbol bastante rara distribuida sobre un área de 400 millas cuadradas. Para cada árbol observa si está plantado en un suelo de serpentina o no, y si las hojas son pubescentes ó lisas. Así la muestra de $n = 100$ árboles puede dividirse en cuatro grupos: pubescente-serpentina, serpentina-lisa, no serpentina-pubescente, y no serpentina-lisa. Si la probabilidad de que un árbol sea pubescente o no es independiente de su situación, se sostendrá nuestra hipótesis nula de la independencia de estas propiedades. Si por el contrario, la proporción de pubescentes difiere para los dos tipos de suelo, nuestra prueba estadística conducirá probablemente al rechazo de la hipótesis nula de independencia. De nuevo las frecuencias esperadas serán simplemente los productos de las proporciones independientes de las dos propiedades, serpentina contra no serpentina y pubescente contra lisa. En este ejemplo las proporciones en sí pueden ser de interés para el investigador.

El ejemplo que resolveremos con detalle es de inmunología. Una muestra de 111 ratones se dividió en dos grupos, 57 que recibieron una dosis estándar de bacterias patógenas, seguida por un antisuero, y un grupo control de 54 que recibieron las bacterias pero no el antisuero. Una vez que había transcurrido suficiente tiempo para un período de incubación y para que la enfermedad siguiera su curso, se contaron 38 ratones muertos y 73 supervivientes. De los que habían muerto, 13 habían recibido bacterias y antisuero mientras que 25 habían recibido solamente bacterias. Una cuestión de interés es si el antisuero había protegido de algún modo a los ratones de modo que hubiese proporcionalmente más supervivientes en ese grupo. De nuevo aquí las proporciones de estas propiedades son de no más interés que en el primer ejemplo (depredación en polillas).

Estos datos se exponen convenientemente en forma de *tabla de doble entrada* como se muestra más abajo. Las tablas de doble entrada y de múltiple entrada (más de dos criterios) se conocen a veces como *tablas de contingencia*. Este tipo de tabla de doble entrada, en la que cada uno de los dos criterios se divide en dos clases, se conoce como *tabla* 2×2.

	Muertos	*Vivos*	Σ
Bacterias y antisuero	13	44	57
Bacterias solamente	25	29	54
Σ	38	73	111

Así, 13 ratones recibieron bacterias y antisuero pero murieron, como se ve en la tabla. Los totales marginales dan el número de ratones que manifiestan cualquier propiedad: 57 ratones recibieron bacterias y antisuero; 73 ratones sobrevivieron al experimento. En conjunto se incluyeron en el experimento 111 ratones y constituyen la muestra total.

Al discutir esta tabla es conveniente marcar las casillas de la tabla y las sumas de filas y columnas como sigue:

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	n

A partir de una tabla de doble entrada se pueden calcular sistemáticamente las frecuencias esperadas (basadas en la hipótesis nula de independencia) y compararlas con las frecuencias observadas. Por ejemplo, la frecuencia esperada para d (bacterias, vivos) sería

$$\hat{f}_{\text{bact,vivos}} = n\hat{p}_{\text{bact,vivos}} = n\hat{p}_{\text{bact}} \times \hat{p}_{\text{vivos}} = n\left(\frac{c+d}{n}\right)\left(\frac{b+d}{n}\right) = (c+d)(b+d)/n$$

que en nuestro caso sería (54)(73)/111 = 35,5, un valor más alto que la frecuencia observada de 29. Podemos proceder de un modo similar *para calcular las frecuencias esperadas para cada casilla de la tabla multiplicando un total de fila por un total de columna, y dividiendo el producto por la suma total.* Las frecuencias esperadas pueden exponerse convenientemente en forma de tabla de doble entrada:

		Muertos	*Vivos*	Σ
Bacterias y antisuero		19,5	37,5	57,0
Solamente bacterias		18,5	35,5	54,0
	Σ	38,0	73,0	111,0

Se notará que las sumas de fila y columna de esta tabla son idénticas a las de la tabla de frecuencias observadas, lo cual no debería sorprender ya que las frecuencias esperadas se calcularon partiendo de estas sumas de fila y columna. Por lo tanto, debería quedar claro que una prueba de independencia no probará si una propiedad se presenta en una determinada proporción sino que solamente puede probar si las dos propiedades se manifiestan independientemente o no.

La prueba estadística apropiada para una determinada tabla 2 × 2 depende del modelo fundamental que éste represente. En la literatura estadística ha existido considerable confusión sobre este asunto. Para nuestros fines actuales no es necesario distinguir entre los tres modelos de tablas de contingencia. La prueba G ilustrada más adelante, dará al menos aproximadamente resultados correctos con muestras de tamaño moderado o grande independientemente del modelo fundamental. También se podría realizar una

CUADRO 13.2

Prueba de independencia 2 × 2

Un ecólogo vegetal muestrea 100 árboles de una especie rara de un área de 400 millas cuadradas. Para cada árbol registra si está plantado en suelos de serpentina o no, y si sus hojas son pubescentes ó lisas.

Suelo	*Pubescente*	*Lisa*	*Totales*
Serpentina	12	22	34
No serpentina	16	50	66
Totales	28	72	100 = n

La representación algebraica convencional de esta tabla es como sigue:

		Σ	
	a	b	$a+b$
	c	d	$c+d$
Σ	$a+c$	$b+d$	$a+b+c+d=n$

Si $ad - bc$ es positivo, como ocurre en nuestro ejemplo, puesto que (12 × 50) − (16 × 22) = 248, se resta $\frac{1}{2}$ de a y d y se suma $\frac{1}{2}$ a b y c. Si $ad - bc$ es negativo, se suma $\frac{1}{2}$ a a y d y se resta $\frac{1}{2}$ de b y c. Esta es la corrección de Yate y puede ignorarse cuando $n > 200$. La nueva tabla 2 × 2 se presenta como sigue.

Suelo	*Pubescente*	*Lisa*	*Totales*
Serpentina	$11\frac{1}{2}$	$22\frac{1}{2}$	34
No serpentina	$16\frac{1}{2}$	$49\frac{1}{2}$	66
Totals	28	72	100

Calcular las siguientes cantidades valiéndose de las tablas **XI** y **XII**.

1. $\Sigma f \ln f$ para las frecuencias de casilla
$$= 11\tfrac{1}{2} \ln 11\tfrac{1}{2} + 22\tfrac{1}{2} \ln 22\tfrac{1}{2} + 16\tfrac{1}{2} \ln 16\tfrac{1}{2} + 49\tfrac{1}{2} \ln 49\tfrac{1}{2}$$
$$= 28{,}087 + 70{,}054 + 46{,}255 + 193{,}148 = 337{,}544$$

2. $\Sigma f \ln f$ para las sumas de fila y columna
$$= 34 \ln 34 + 66 \ln 66 + 28 \ln 28 + 72 \ln 72$$
$$= 119{,}896 + 276{,}517 + 93{,}302 + 307{,}920 = 797{,}635$$

3. Buscar $n \ln n = 100 \ln 100 = 460{,}517$

CUADRO 13.2 (continuación)

4. $G_{aj} = 2[\text{cantidad } 1 - \text{cantidad } 2 + \text{cantidad } 3]$
$= 2[337{,}544 - 797{,}635 + 460{{,}}517] = 2[0{,}426] = 0{,}852$

Comparar G_{aj} con el valor crítico de χ^2 para un grado de libertad. Como nuestro G_{aj} observado es mucho menor que $\chi^2_{0{,}05[1]} = 3{,}841$, aceptamos la hipótesis nula de que el tipo de hoja es independiente del tipo de suelo en que está plantado el árbol.

prueba ji-cuadrado sobre las desviaciones de las frecuencias observadas respecto de las esperadas utilizando la expresión (13.2). Este daría $X^2 = 6{,}767$, utilizando frecuencias esperadas redondeadas hasta un decimal. Vamos a establecer sin explicación que la X^2 observada debería compararse con χ^2 para un grado de libertad. Al final de esta sección examinaremos las razones para esto. La probabilidad de encontrar un ajuste tan malo, o peor, a estos datos es $0{,}005 < P < 0{,}01$. Por tanto concluimos que la mortalidad en estos ratones no es independiente de la presencia de antisuero. Observamos que el porcentaje de mortalidad entre los que se les ha dado bacterias y antisuero es $(13)(100)/57 = 22{,}8$ %, considerablemente más bajo que la mortalidad de $(25)(100)/54 = 46{,}3$ % entre los ratones a los que se les ha administrado solamente bacterias. Sin duda el antisuero ha sido efectivo reduciendo la mortalidad.

En el cuadro 13.2 ilustramos la prueba G aplicada al experimento de muestreo de ecología vegetal, tratándose de árboles plantados en dos suelos diferentes y que poseen dos tipos de hojas. Con tamaños de muestreo pequeños ($n < 200$) aplicamos nuevamente una *corrección de continuidad* (*corrección de Yates*), cuya aplicación se muestra en el cuadro. El resultado del análisis demuestra claramente que no podemos rechazar la hipótesis nula de independencia entre tipo de suelo y tipo de hoja. La presencia de hojas pubescentes es independiente de que el árbol esté plantado en suelos de serpentina o no.

Las pruebas de independencia no tienen por qué restringirse a tablas 2×2. En los casos de doble entrada considerados en esta sección solamente nos ocupamos de dos propiedades, pero cada una de estas propiedades puede dividirse en un número cualquiera de clases. Así los organismos pueden presentarse en cuatro clases de color y muestrearse de tres localidades, dando una prueba de independencia 4×3. Esta prueba examinaría si las proporciones de color manifestadas por los totales marginales son independientes de las localidades en las que se han muestreado los individuos. Estas pruebas se denominan frecuentemente pruebas de independencia $F \times C$, representando F y C el número de filas y columnas en la tabla de frecuencias. Otro caso, examinado con detalle en el cuadro 13.3, trata de patrones de color rojo claro hallados en muestras de una especie de cicindela, en cuatro ocasiones durante la primavera y verano. Era de interés conocer si el porcentaje de individuos rojo claro cambiaba significativamente durante el tiempo de observación. Examinamos si la proporción de coleópteros de color rojo claro (55,7 % para el estudio completo) es independiente del tiempo de recolección.

Como se muestra en el cuadro 13.3, la que sigue es una simple regla general para el cálculo de la prueba G de independencia:

$G = 2[(\sum f \ln f \text{ para las frecuencias de casilla})$
$- (\sum f \ln f \text{ para las sumas de fila y columna}) + n \ln n]$

Las transformaciones pueden buscarse en la tabla **XI**. En las fórmulas del cuadro 13.3 empleamos un subíndice doble para referir las entradas en una tabla de doble entrada, como en el caso estructuralmente similar de análisis de varianza bidireccional. La cantidad f_{ij} del cuadro 13.3 hace referencia a la frecuencia observada en la fila i y la columna j de la tabla.

Los resultados del cuadro 13.3 muestran claramente que la frecuencia de patrones de color rojo claro en estas cicindelas es dependiente de la estación. Observamos una disminución de rojos claros a finales de primavera y principio de verano, seguido por un nuevo incremento a final del verano.

Los grados de libertad para pruebas de independencia son siempre los mismos y pueden calcularse utilizando las reglas dadas anteriormente (sección 13.2). Hay k casillas en la tabla pero debemos restar un grado de libertad para cada parámetro independiente que hayamos estimado de los datos. Naturalmente debemos restar un grado de libertad para el tamaño de muestreo total observado, n. Hemos estimado además $a - 1$ probabilidades de fila y $b - 1$ probabilidades de columna, donde a y b son el número de filas y columnas de la tabla, respectivamente. Así, hay $k - (a - 1) - (b - 1) - 1 = k - a - b + 1$ grados de libertad para la prueba. Pero ya que $k = a \times b$, esta expresión se convierte en $(a \times b) - a - b + 1 = (a - 1) \times (b - 1)$, la expresión convencional para los grados de libertad en una prueba de independencia de doble entrada. Así, los grados de libertad en el ejemplo del cuadro 13.3, un caso 4×2, eran $(4 - 1) \times (2 - 1) = 3$. En todos los casos 2×2 naturalmente hay sólo $(2 - 1) \times (2 - 1) = 1$ grado de libertad.

Otro nombre para prueba de independencia es *prueba de asociación*. Si dos propiedades no son independientes entre sí, *están asociadas*. Así, en el ejemplo que examina la frecuencia relativa de dos tipos de hoja en dos suelos diferentes podemos hablar de una asociación entre tipos de hoja y suelos. En el experimento de inmunología hay una asociación negativa entre presencia de antisuero y mortalidad. *Asociación* es pues similar a correlación, pero es un término más general que se aplica tanto a los atributos como a las variables continuas. En las pruebas de independencia 2×2 de esta sección, una forma de buscar la falta de independencia sospechada era examinar el porcentaje de manifestación de una de las propiedades en las dos clases, basado en la otra propiedad. Así, comparábamos el porcentaje de hojas lisas en los dos tipos de suelos, o estudiábamos el porcentaje de mortalidad con ó sin antisuero. Esta forma de considerar una prueba de independencia sugiere otra interpretación de éstas como pruebas de significación de las diferencias entre dos porcentajes.

Ejercicios 13

13.1 En un experimento para determinar el modo de herencia de un mutante *verde*, se obtuvieron 146 descendientes tipo salvaje y 30 mutantes cuando se cruzaron las moscas de la generación F_1. Probar si los datos concuerdan con la hipótesis de que la razón de tipo salvaje a mutantes es 3:1. SOLUCION. $G = 6{,}4624$.

CUADRO 13.3

Prueba de independencia $F \times C$ utilizando la prueba G.

Frecuencias de patrones de color de una especie de cicindela (*Cicindela fulgida*) halladas en varias estaciones

Estación	*Patrón de color* ($a = 2$)		*Sumas*	*% Rojo claro*
($b = 4$)	*Rojo claro*	*No rojo claro*		
Primavera temprana	29	11	40	72,5
Primavera tardía	273	191	464	58,8
Verano temprano	8	31	39	20,5
Verano tardío	64	64	128	50,0
Sumas	374	297	671 = n	55,7

Fuente: Datos no publicados de H. L. Willis.

Calcular las sumas siguientes, utilizando la tabla **XI** para $f \ln f$.

1. Suma de transformaciones de las frecuencias en el cuerpo de la tabla de contingencia
$$= \sum^{b}\sum^{a} f_{ij} \ln f_{ij} = 29 \ln 29 + 11 \ln 11 + \cdots + 64 \ln 64$$
$$= 97{,}652 + 26{,}377 + \cdots + 266{,}169 = 3314{,}027$$
2. Suma de transformaciones de los totales de fila
$$= \sum^{b}\left(\sum^{a} f_{ij}\right) \ln \left(\sum^{a} f_{ij}\right)$$
$$= 40 \ln 40 + \cdots + 128 \ln 128 = 147{,}555 + \cdots + 621{,}060$$
$$= 3760{,}400$$
3. Suma de las transformaciones de los totales de columna
$$= \sum^{a}\left(\sum^{b} f_{ij}\right) \ln \left(\sum^{b} f_{ij}\right)$$
$$= 374 \ln 374 + 297 \ln 297 = 2215{,}672 + 1691{,}038 = 3906{,}710$$
4. Transformación de la suma total
$$= n \ln n = 671 \ln 671 = 4367{,}384$$
5. $G = 2[\text{cantidad } 1 - \text{cantidad } 2 - \text{cantidad } 3 + \text{cantidad } 4]$
$$= 2[3314{,}027 - 3760{,}400 - 3906{,}710 + 4367{,}384 = 2[14{,}301] = 28{,}602$$

Este valor es para compararlo con una distribución χ^2 con $(a - 1)(b - 1)$ grados de libertad, donde a es el número de columnas y b el número de filas de la tabla. En nuestro caso, $gl = (2 - 1)(4 - 1) = 3$.

Como $\chi^2_{0,005(3)} = 12{,}838$, nuestro valor G es significativo a $P < 0{,}005$, y debemos rechazar nuestra hipótesis nula de que la frecuencia de patrón de color es independiente de la estación.

13.2 En la localidad A se ha hecho una recolecta exhaustiva de la especie S. Un examen de los 167 machos adultos que se han recogido revela que 35 de ellos tienen bandas de color pálido alrededor de su cuello. De la localidad B, situada a 90 millas, obtenemos una muestra de 27 machos adultos de la misma especie, de los que 6 muestran las bandas. ¿Cuál es la probabilidad de que ambas muestras sean de la misma población estadística con respecto a la frecuencia de bandas?

13.3 De 445 ejemplares de la mariposa *Erebia epipsodea* en áreas montañosas, 2,5 % tienen en sus alas manchas de color claro. De 65 especímenes de la pradera 70,8 % tienen estas manchas (datos no publicados de P. R. Ehrlich). ¿Es significativa esta diferencia? SOLUCION. $G_{adj} = 170{,}998$.

13.4 En un estudio de polimorfismo de inversiones cromosómicas en la langosta *Moraba Scurra*, Lewontin y White (1960) dieron los siguientes resultados para la composición de una población en Royalla "B" en 1958.

		Cromosoma CD		
		St/St	St/Bl	Bl/Bl
Cromosoma EF	Td/Td	22	96	75
	St/Td	8	56	64
	St/St	0	6	6

¿Son las frecuencias de las tres combinaciones diferentes del cromosoma EF independientes de las frecuencias de las tres combinaciones del cromosoma CD? SOLUCION. $G = 7{,}396$.

13.5 Comprobar si el porcentaje de ninfas del áfido *Myzus persicae* que se ha desarrollado en formas aladas depende del tipo de dieta suministrada. A las hembras apomícticas se les han puesto las dietas un día antes del nacimiento de las ninfas (dato de Mittler y Dadd, 1966).

Tipo de dieta	*% formas aladas*	n
Dieta sintética	100	216
"Sandwich" cotiledón	92	230
Cotiledón libre	36	75

13.6 Probar el ajuste de las frecuencias observadas respecto de las esperadas, basándose en una distribución binomial de los datos dados en las tablas 4.1 y 4.2.

13.7 Probar el ajuste de las frecuencias observadas respecto de las esperadas basándose en una distribución de Poisson para los datos de la tabla 4.5 y tabla 4.6.

Apéndice 1

Apéndice matemático

A1.1 Demostración de que la suma de las desviaciones de la media es igual a cero.

Hemos de aprender dos reglas comunes de álgebra estadística. Podemos abrir un par de paréntesis con un signo $\sum$ delante tratando el signo $\sum$ como si fuese un factor común.

Tenemos

$$\sum_{i=1}^{n}(A_i + B_i) = (A_1 + B_1) + (A_2 + B_2) + \cdots + (A_n + B_n)$$
$$= (A_1 + A_2 + \cdots + A_n) + (B_1 + B_2 + \cdots + B_n)$$

Por consiguiente

$$\sum_{i=1}^{n}(A_i + B_i) = \sum_{i=1}^{n} A_i + \sum_{i=1}^{n} B_i$$

Además, cuando se desarrolla $\sum_{i=1}^{n} C$ en una operación algebraica, en que C es constante, ésta puede calcularse como sigue:

$$\sum_{i=1}^{n} C = C + C + \cdots + C \qquad (n \text{ términos})$$
$$= nC$$

Puesto que en un problema determinado una media es un valor constante, $\sum_{}^{n}\overline{Y} = n\overline{Y}$. Si se desea, se pueden comprobar estas reglas, utilizando números sencillos. En la demostración

que sigue y otras, siempre que todas las sumas sean de n ítems hemos simplificado la notación omitiendo los subíndices de las variables y los sobrescritos sobre los signos sumatorios.

Queremos demostrar que $\sum y = 0$. Por definición,

$$\sum y = \sum(Y - \overline{Y})$$
$$= \sum Y - n\overline{Y}$$
$$= \sum Y - \frac{n\sum Y}{n} \qquad \left(\text{ya que } \overline{Y} = \frac{\sum Y}{n}\right)$$
$$= \sum Y - \sum Y$$

Por lo tanto $\sum y = 0$.

A1.2 Demostración de los efectos de codificación aditiva, multiplicativa, y de combinación sobre las medias, varianzas y desviaciones típicas.

Para esta prueba tenemos que aprender otro convenio de álgebra estadística. En el apéndice A1.1 hemos visto que $\sum C = nC$. Sin embargo, cuando el $\sum$ precede a una constante y a una variable, como en $\sum CY$, la constante puede ponerse delante del $\sum$, ya que

$$\sum_{i=1}^{n} CY_i = CY_1 + CY_2 + \cdots + CY_n$$
$$= C(Y_1 + Y_2 + \cdots + Y_n)$$
$$= C\left(\sum_{i=1}^{n} Y_i\right)$$

Por lo tanto

$$\sum_{i=1}^{n} CY_i = C\sum_{i=1}^{n} Y_i$$

Así $\sum CY = C\sum Y$, $\sum C^2y^2 = C^2\sum y^2$, y $\sum 2\overline{Y}Y = 2\overline{Y}\sum Y$, porque ambos 2 e $\overline{Y}$ son constantes.

Medias de datos transformados

Codificación aditiva. – La variable se codifica $Y_c = Y + C$, en donde C es una constante, el código aditivo. Por lo tanto

$$\sum Y_c = \sum(Y + C) = \sum Y + nC$$

y

$$\overline{Y}_c = \frac{\sum Y_c}{n} = \frac{\sum Y}{n} + C = \overline{Y} + C$$

Para descodificar $\overline{Y}_c$, se le resta C y se obtiene $\overline{Y}$; es decir, $\overline{Y} = \overline{Y}_c - C$.

Codificación multiplicativa. – La variable se codifica $Y_c = DY$, en donde D es una constante, el código multiplicativo. Por lo tanto

$$\sum Y_c = D\sum Y$$

y

$$\overline{Y}_c = \frac{\sum Y_c}{n} = D\frac{\sum Y}{n} = D\overline{Y}$$

Para descodificar $\overline{Y}_c$ se divide por D y se obtiene $\overline{Y}$; es decir $\overline{Y} = \overline{Y}_c/D$.

Codificación de combinación. – La variable se codifica $Y_c = D(Y + C)$, en donde C y D son constantes, los códigos aditivos y multiplicativo, respectivamente. Por consiguiente

$$\sum Y_c = D\sum(Y + C) = D\sum Y + nDC$$

y

$$\overline{Y}_c = \frac{\sum Y_c}{n} = D\frac{\sum Y}{n} + DC = D\overline{Y} + DC$$

Para descodificar $\overline{Y}_c$, se divide por D, a continuación se resta C y se obtiene $\overline{Y}$; es decir,

$$\overline{Y} = \frac{\overline{Y}_c}{D} - C.$$

Varianzas y desviaciones típicas de datos codificados

Codificación aditiva. – La variable se codifica $Y_c = Y + C$, en donde C es una constante, el código aditivo. Por definición, $y = Y - \overline{Y}$, y

$$\begin{aligned} y_c &= Y_c - \overline{Y}_c \\ &= [(Y + C) - (\overline{Y} + C)] \quad \text{(como se ha visto más arriba para las medias)} \\ &= [Y + C - \overline{Y} - C] \\ &= [Y - \overline{Y}] \\ &= y \end{aligned}$$

Por consiguiente $\sum y_c^2 = \sum y^2$, y

$$\frac{\sum y_c^2}{n-1} = \frac{\sum y^2}{n-1}$$

Así pues, la codificación aditiva no tiene efecto sobre las sumas de cuadrados, varianzas, ni desviaciones típicas.

Codificación multiplicativa. – La variable se codifica $Y_c = DY$, en donde D es una constante, el código multiplicativo. Por definición, $y = Y - \overline{Y}$, y

$$\begin{aligned} y_c &= Y_c - \overline{Y}_c \\ &= DY - D\overline{Y} \quad \text{(como se ha visto para las medias)} \\ &= D(Y - \overline{Y}) \\ &= Dy \end{aligned}$$

Por lo tanto $y_c^2 = D^2y^2$, $\sum y_c^2 = D^2\sum y^2$, y

$$\frac{\sum y_c^2}{n-1} = D^2\frac{\sum y^2}{n-1}$$

$$s_c^2 = D^2s^2$$

Así, cuando los datos se han sometido al código multiplicativo, una suma de cuadrados o varianza puede descodificarse dividiéndola por el *cuadrado* del código multiplicativo; una desviación típica puede descodificarse dividiéndola por el código como tal, es decir, $s^2 = s_c^2/D^2$ y $s = s_c/D$.

Codificación de combinación. – La variable se codifica $Y_c = D(Y + C)$, en donde C y D son constantes, los códigos aditivo y multiplicativo, respectivamente. Por definición $y = Y - \overline{Y}$, y

$$\begin{aligned} y_c &= Y_c - \overline{Y}_c \\ &= [D(Y + C) - (D\overline{Y} + DC)] \quad \text{(como se ha visto para las medias)} \\ &= [DY + DC - D\overline{Y} - DC] \\ &= D[Y - \overline{Y}] \end{aligned}$$

Por consiguiente $y_c = Dy$, como antes.

Así, en la codificación de combinación, al descodificar sumas de cuadrados, varianzas, o desviaciones típicas solamente tiene que considerarse el código multiplicativo.

A1.3 Demostración de que la expresión (3.7), la fórmula para calcular la suma de cuadrados, es igual a la expresión (3.6), expresión desarrollada originalmente para este estadístico.

Queremos demostrar que $\sum(Y - \overline{Y})^2 = \sum Y^2 - ((\sum Y)^2/n)$. Tenemos

$$\begin{aligned} \sum(Y - \overline{Y})^2 &= \sum(Y^2 - 2Y\overline{Y} + \overline{Y}^2) \\ &= \sum Y^2 - 2\overline{Y}\sum Y + n\overline{Y}^2 \\ &= \sum Y^2 - \frac{2(\sum Y)^2}{n} + \frac{n(\sum Y)^2}{n^2} \qquad \left(\text{ya que } \overline{Y} = \frac{\sum Y}{n}\right) \\ &= \sum Y^2 - \frac{2(\sum Y)^2}{n} + \frac{(\sum Y)^2}{n} \end{aligned}$$

Por lo tanto

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

A1.4 Fórmulas simplificadas para el error estándar de la diferencia entre dos medias. El error estándar al cuadrado, de la expresión (8.2) es

$$\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right]\left(\frac{n_1 + n_2}{n_1 n_2}\right).$$

Cuando $n_1 = n_2 = n$, éste se simplifica hasta

$$\left[\frac{(n - 1)s_1^2 + (n - 1)s_2^2}{2n - 2}\right]\left(\frac{2n}{n^2}\right) = \left[\frac{(n - 1)(s_1^2 + s_2^2)(2)}{2(n - 1)(n)}\right] = \frac{1}{n}(s_1^2 + s_2^2)$$

que es el error estándar al cuadrado, de la expresión (8.3).

Cuando $n_1 \neq n_2$, pero cada uno es grande de modo que $(n_1 - 1) \approx n_1$ y $(n_2 - 1) \approx n_2$, el error estándar al cuadrado, de la expresión (8.2) se simplifica a

$$\left[\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}\right]\left(\frac{n_1 + n_2}{n_1 n_2}\right) = \left[\frac{n_1 s_1^2}{n_1 n_2} + \frac{n_2 s_2^2}{n_1 n_2}\right] = \frac{s_1^2}{n_2} + \frac{s_2^2}{n_1}$$

que es el error estándar al cuadrado, de la expresión (8.4).

A1.5 Demostración de que t_s^2 obtenido a partir de una prueba de significación de la diferencia entre dos medias (como en el cuadro 8.2), es idéntico al valor F_s obtenido en un análisis de la varianza de clasificación simple de dos grupos de igual tamaño de muestreo (en el mismo cuadro).

$$t_s \text{ (del cuadro 8.2)} = = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{1}{n(n-1)}\left(\sum^n y_1^2 + \sum^n y_2^2\right)}}$$

$$t_s^2 = \frac{(\bar{Y}_1 - \bar{Y}_2)^2}{\frac{1}{n(n-1)}\left(\sum^n y_1^2 + \sum^n y_2^2\right)} = \frac{n(n-1)(\bar{Y}_1 - \bar{Y}_2)^2}{\sum^n y_1^2 + \sum^n y_2^2}$$

En el análisis de varianza de dos muestras,

$$M.C._{\text{medias}} = \frac{1}{2-1}\sum^2 (\bar{Y}_i - \bar{\bar{Y}})^2$$
$$= (\bar{Y}_1 - \bar{\bar{Y}})^2 + (\bar{Y}_2 - \bar{\bar{Y}})^2$$
$$= \left(\bar{Y}_1 - \frac{\bar{Y}_1 + \bar{Y}_2}{2}\right)^2 + \left(\bar{Y}_2 - \frac{\bar{Y}_1 + \bar{Y}_2}{2}\right)^2 \quad (\text{puesto que } \bar{\bar{Y}} = (\bar{Y}_1 + \bar{Y}_2)/2)$$
$$= \left(\frac{\bar{Y}_1 - \bar{Y}_2}{2}\right)^2 + \left(\frac{\bar{Y}_2 - \bar{Y}_1}{2}\right)^2$$
$$= \tfrac{1}{2}(\bar{Y}_1 - \bar{Y}_2)^2$$

ya que los cuadrados de los numeradores son idénticos. En este caso

$$M.C._{\text{grupos}} = n \times M.C._{\text{medias}} = n[\tfrac{1}{2}(\bar{Y}_1 - \bar{Y}_2)^2]$$
$$= \frac{n}{2}(\bar{Y}_1 - \bar{Y}_2)^2$$

$$M.C._{\text{intra}} = \frac{\sum^n y_1^2 + \sum^n y_2^2}{2(n-1)}$$

$$F_s = \frac{M.C._{\text{grupos}}}{M.C._{\text{intra}}}$$
$$= \frac{\frac{n}{2}(\bar{Y}_1 - \bar{Y}_2)^2}{\left(\sum^n y_1^2 + \sum^n y_2^2\right) \Big/ [2(n-1)]}$$
$$= \frac{n(n-1)(\bar{Y}_1 - \bar{Y}_2)^2}{\sum^n y_1^2 + \sum^n y_2^2}$$
$$= t_s^2$$

A1.6 Demostración de que la expresión (11.5), fórmula para calcular la suma de productos, es igual a $\sum(X - \bar{X})(Y - \bar{Y})$, expresión desarrollada originalmente para esta cantidad.

Todas las sumas son de n ítems. Tenemos

$$\begin{aligned}\sum xy &= \sum(X - \bar{X})(Y - \bar{Y}) \\ &= \sum XY - \bar{X}\sum Y - \bar{Y}\sum X + n\bar{X}\bar{Y} \quad (\text{ya que } \textstyle\sum \bar{X}\bar{Y} = n\bar{X}\bar{Y}) \\ &= \sum XY - \bar{X}n\bar{Y} - \bar{Y}n\bar{X} + n\bar{X}\bar{Y} \quad (\text{ya que } \textstyle\sum Y/n = \bar{Y}, \\ &\qquad\qquad\qquad\qquad\qquad\qquad \textstyle\sum Y = n\bar{Y}; \text{ igualmente, } \sum X = n\bar{X})\end{aligned}$$

$$= \sum XY - n\bar{X}\bar{Y}$$
$$= \sum XY - n\bar{X}\sum Y/n$$
$$= \sum XY - \bar{X}\sum Y$$

Igualmente

$$\sum xy = \sum XY - \bar{Y}\sum X$$

y

$$\sum xy = \sum XY - \frac{(\sum X)(\sum Y)}{n} \tag{11.5}$$

A1.7 Deducción de la fórmula para calcular $\sum d_{Y\cdot X}^2 = \sum y^2 - ((\sum xy)^2/\sum x^2)$.

Por definición, $d_{Y\cdot X} = Y - \hat{Y}$. Como $\bar{Y} = \bar{\hat{Y}}$, podemos restar $\bar{Y}$ de ambos Y e $\hat{Y}$ para obtener

$$d_{Y\cdot X} = y - \hat{y} = y - bx \qquad \text{(ya que } \hat{y} = bx\text{)}$$

Por lo tanto

$$\sum d_{Y\cdot X}^2 = \sum (y - bx)^2 = \sum y^2 - 2b\sum xy + b^2\sum x^2$$
$$= \sum y^2 - 2\frac{\sum xy}{\sum x^2}\sum xy + \frac{(\sum xy)^2}{(\sum x^2)^2}\sum x^2 = \sum y^2 - 2\frac{(\sum xy)^2}{\sum x^2} + \frac{(\sum xy)^2}{\sum x^2}$$

o

$$\sum d_{Y\cdot X}^2 = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2} \tag{11.7}$$

A1.8 Demostración de que la suma de cuadrados de la variable dependiente en regresión puede descomponerse exactamente en sumas de cuadrados explicable e inexplicable, anulándose los dobles productos.

Por definición (sección 11.5)

$$y = \hat{y} + d_{Y\cdot X}$$
$$\sum y^2 = \sum (\hat{y} + d_{Y\cdot X})^2 = \sum \hat{y}^2 + \sum d_{Y\cdot X}^2 + 2\sum \hat{y} d_{Y\cdot X}$$

Si podemos demostrar que $\sum \hat{y} d_{Y\cdot X} = 0$, hemos demostrado la identidad requerida. Tenemos

$$\sum \hat{y} d_{Y\cdot X} = \sum bx(y - bx) \qquad \text{[puesto que } \hat{y} = bx \text{ según la expresión (11.3) y}$$
$$= b\sum xy - b^2\sum x^2 \qquad d_{Y\cdot X} = y - bx \text{ según el apéndice A1.7]}$$
$$= b\sum xy - b\frac{\sum xy}{\sum x^2}\sum x^2 \qquad \text{(puesto que } b = \sum xy/\sum x^2\text{)}$$
$$= b\sum xy - b\sum xy$$
$$= 0$$

Por consiguiente $\sum y^2 = \sum \hat{y}^2 + \sum d_{Y\cdot X}^2$

o, escrito en términos de variantes,

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

A1.9 Probar que la varianza de la suma de dos variables es

$$\sigma_{(Y_1+Y_2)}^2 = \sigma_1^2 + \sigma_2^2 + 2\rho_{12}\sigma_1\sigma_2$$

donde σ_1 y σ_2 son desviaciones típicas de Y_1 e Y_2 respectivamente, y ρ_{12} es el coeficiente de correlación paramétrico entre Y_1 e Y_2.

Si $Z = Y_1 + Y_2$, en este caso

$$\sigma_Z^2 = \frac{1}{n}\sum (Z - \bar{Z})^2 = \frac{1}{n}\sum \left[(Y_1 + Y_2) - \frac{1}{n}\sum (Y_1 + Y_2) \right]^2$$
$$= \frac{1}{n}\sum \left[(Y_1 + Y_2) - \frac{1}{n}\sum Y_1 - \frac{1}{n}\sum Y_2 \right]^2 = \frac{1}{n}\sum \left[(Y_1 + Y_2) - \bar{Y}_1 - \bar{Y}_2 \right]^2$$
$$= \frac{1}{n}\sum \left[(Y_1 - \bar{Y}_1) + (Y_2 - \bar{Y}_2) \right]^2 = \frac{1}{n}\sum \left[y_1 + y_2 \right]^2$$
$$= \frac{1}{n}\sum \left[y_1^2 + y_2^2 + 2y_1y_2 \right] = \frac{1}{n}\sum y_1^2 + \frac{1}{n}\sum y_2^2 + \frac{2}{n}\sum y_1y_2$$
$$= \sigma_1^2 + \sigma_2^2 + 2\sigma_{12}$$

Pero, como $\rho_{12} = \sigma_{12}/\sigma_1\sigma_2$, tenemos

$$\sigma_{12} = \rho_{12}\sigma_1\sigma_2$$

Por lo tanto

$$\sigma_Z^2 = \sigma_{(Y_1+Y_2)}^2 = \sigma_1^2 + \sigma_2^2 + 2\rho_{12}\sigma_1\sigma_2$$

Igualmente,

$$\sigma_D^2 = \sigma_{(Y_1-Y_2)}^2 = \sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2$$

La expresión análoga se aplica a estadísticos de muestreo. Así

$$s_{(Y_1+Y_2)}^2 = s_1^2 + s_2^2 + 2r_{12}s_1s_2 \tag{12.8}$$
$$s_{(Y_1-Y_2)}^2 = s_1^2 + s_2^2 - 2r_{12}s_1s_2 \tag{12.9}$$

A1.10 Deducción de la fórmula para calcular X^2 [expresión (13.2)] según la expresión (13.1).

Desarrollando la expresión (13.1),

$$X^2 = \sum^a \frac{(f_i - \hat{f}_i)^2}{\hat{f}_i}$$

$$= \sum^a \frac{f_i^2}{\hat{f}_i} + \sum^a \frac{\hat{f}_i^2}{\hat{f}_i} - 2\sum^a \frac{f_i \hat{f}_i}{\hat{f}_i}$$

$$= \sum^a \frac{f_i^2}{\hat{f}_i} + \sum^a \hat{f}_i - 2\sum^a f_i$$

Pero, como $\sum^a f_i = \sum^a \hat{f}_i = n$

$$X^2 = \sum^a \frac{f_i^2}{\hat{f}_i} - n \tag{13.2}$$

A1.11 Probar que la expresión general para la prueba G puede simplificarse hasta las expresiones (13.4) y (13.5).

En general, G es dos veces el logaritmo natural de la razón de la probabilidad de la muestra con todos los parámetros estimados de los datos y la probabilidad de la muestra suponiendo que la hipótesis nula sea cierta. Suponiendo una distribución multinomial, esta razón es

$$L = \frac{\dfrac{n!}{f_1!f_2! \cdots f_a!} p_1^{f_1} p_2^{f_2} \cdots p_a^{f_a}}{\dfrac{n!}{f_1!f_2! \cdots f_a!} \hat{p}_1^{f_1} \hat{p}_2^{f_2} \cdots \hat{p}_a^{f_a}}$$

$$= \prod_{i=1}^{a} \left(\frac{p_i}{\hat{p}_i}\right)^{f_i}$$

$$G = 2 \ln L$$

$$= 2\sum^a f_i \ln \left(\frac{p_i}{\hat{p}_i}\right)$$

Como $f_i = np_i$ y $\hat{f}_i = n\hat{p}_i$,

$$G = 2\sum^a f_i \ln \left(\frac{f_i}{\hat{f}_i}\right) \tag{13.4}$$

Si ahora sustituimos $\hat{f}_i$ por $n\hat{p}_i$,

$$G = 2\sum^a f_i \ln \left(\frac{f_i}{n\hat{p}_i}\right) = 2[\sum^a f_i \ln f_i - \sum^a f_i \ln \hat{p}_i - \sum^a f_i \ln n]$$

$$= 2[\sum^a f_i \ln f_i - \sum^a f_i \ln \hat{p}_i - n \ln n] \tag{13.5}$$

Apéndice 2

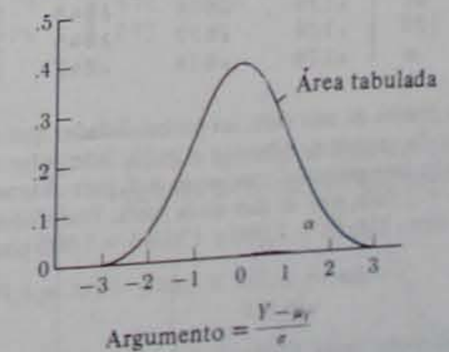
Tablas estadísticas

TABLA I. Dos mil quinientos dígitos aleatorios.

	1	2	3	4	5	6	7	8	9	10	
1	48461	14952	72619	73689	52059	37086	60050	86192	67049	64739	1
2	76534	38149	49692	31366	52093	15422	20498	33901	10319	43397	2
3	70437	25861	38504	14752	23757	59660	67844	78815	23758	86814	3
4	59584	03370	42806	11393	71722	93804	09095	07856	55589	46020	4
5	04285	58554	16085	51555	27501	73883	33427	33343	45507	50063	5
6	77340	10412	69189	85171	29082	44785	83638	02583	96483	76553	6
7	59183	62687	91778	80354	23512	97219	65921	02035	59847	91403	7
8	91800	04281	39979	03927	82564	28777	59049	97532	54540	79472	8
9	12066	24817	81099	48940	69554	55925	48379	12866	51232	21580	9
10	69907	91751	53512	23748	65906	91385	84983	27915	48491	91068	10
11	80467	04873	54053	25955	48518	13815	37707	68687	15570	08890	11
12	78057	67835	28302	45048	56761	97725	58438	91528	24645	18544	12
13	05648	39387	78191	88415	60269	94880	58812	42931	71898	61534	13
14	22304	39246	01350	99451	61862	78688	30339	60222	74052	25740	14
15	61346	50269	67005	40442	33100	16742	61640	21046	31909	72641	15
16	66793	37696	27965	30459	91011	51426	31006	77468	61029	57108	16
17	86411	48809	36698	42453	83061	43769	39948	87031	30767	13953	17
18	62098	12825	81744	28882	27369	88183	65846	92545	09065	22655	18
19	68775	06261	54265	16203	23340	84750	16317	88686	86842	00879	19
20	52679	19595	13687	74872	89181	01939	18447	10787	76246	80072	20
21	84096	87152	20719	25215	04349	54434	72344	93008	83282	31670	21
22	63964	55937	21417	49944	38356	98404	14850	17994	17161	98981	22
23	31191	75131	72386	11689	95727	05414	88727	45583	22568	77700	23
24	30545	68523	29850	67833	05622	89975	79042	27142	99257	32349	24
25	52573	91001	52315	26430	54175	30122	31796	98842	37600	26025	25
26	16586	81842	01076	99414	31574	94719	34656	80018	86988	79234	26
27	81841	88481	61191	25013	30272	23388	22463	65774	10029	58376	27
28	43563	66829	72838	08074	57080	15446	11034	98143	74989	26885	28
29	19945	84193	57581	77252	85604	45412	43556	27518	90572	00563	29
30	79374	23796	16919	99691	80276	32818	62953	78831	54395	30705	30
31	48503	26615	43980	09810	38289	66679	73799	48418	12647	40044	31
32	32049	65541	37937	41105	70106	89706	40829	40789	59547	00783	32
33	18547	71562	95493	34112	76895	46766	96395	31718	48302	45893	33
34	03180	96742	61486	43305	34183	99605	67803	13491	09243	29557	34
35	94822	24738	67749	83748	59799	25210	31093	62925	72061	69991	35
36	34330	60599	85828	19152	68499	27977	35611	96240	62747	89529	36
37	43770	81537	59527	95674	76692	86420	69930	10020	72881	12532	37
38	56908	77192	50623	41215	14311	42834	80651	93750	59957	31211	38
39	32787	07189	80539	75927	75475	73965	11796	72140	48944	74156	39
40	52441	78392	11733	57703	29133	71164	55355	31006	25526	55790	40
41	22377	54723	18227	28449	04570	18882	00023	67101	06895	08915	41
42	18376	73460	88841	39602	34049	20589	05701	08249	74213	25220	42
43	53201	28610	87957	21497	64729	64983	71551	99016	87903	63875	43
44	34919	78901	59710	27396	02593	05665	11964	44134	00273	76358	44
45	33617	92159	21971	16901	57383	34262	41744	60891	57624	06962	45
46	70010	40964	98780	72418	52571	18415	64362	90636	38034	04909	46
47	19282	68447	35665	31530	59832	49181	21914	65742	89815	39231	47
48	91429	73328	13266	54898	68795	40948	80808	63887	89939	47938	48
49	97637	78393	33021	05867	86520	45363	43066	00988	64040	09803	49
50	95150	07625	05255	83254	93943	52325	93230	62668	79529	65964	50

TABLA II. Áreas de la curva normal.

Unidades de desviación típica	0	1	2	3	4	5	6	7	8	9	Unidades de desviación típica
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359	0.0
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636-	.0675	.0714	.0753	0.1
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141	0.2
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517	0.3
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879	0.4
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224	0.5
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549	0.6
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852	0.7
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133	0.8
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389	0.9
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621	1.0
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830	1.1
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015	1.2
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177	1.3
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319	1.4
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441	1.5
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545	1.6
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633	1.7
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706	1.8
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767	1.9
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817	2.0
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857	2.1
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890	2.2
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916	2.3
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936	2.4
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952	2.5
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964	2.6
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974	2.7
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981	2.8
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986	2.9
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990	3.0
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993	3.1
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995	3.2
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997	3.3
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998	3.4
3.5	.499767										
3.6	.499841										
3.7	.499892										
3.8	.499928										
3.9	.499952										
4.0	.499968										
4.1	.499979										
4.2	.499987										
4.3	.499991										
4.4	.499995										
4.5	.499997										
4.6	.499998										
4.7	.499999										
4.8	.499999										
4.9	.500000										



Nota: La cantidad dada es el área bajo la función de densidad normal estándar entre la media y el punto crítico. El área se designa generalmente como $\frac{1}{2} - \alpha$ (como se muestra en la figura). Por interpolación inversa se puede hallar el número de desviaciones típicas correspondientes a un área determinada.

TABLA III. Valores críticos de la distribución t de Student.

ν \ α	0.9	0.5	0.4	0.2	0.1
1	.158	1.000	1.376	3.078	6.314
2	.142	.816	1.061	1.886	2.920
3	.137	.765	.978	1.638	2.353
4	.134	.741	.941	1.533	2.132
5	.132	.727	.920	1.476	2.015
6	.131	.718	.906	1.440	1.943
7	.130	.711	.896	1.415	1.895
8	.130	.706	.889	1.397	1.860
9	.129	.703	.883	1.383	1.833
10	.129	.700	.879	1.372	1.812
11	.129	.697	.876	1.363	1.796
12	.128	.695	.873	1.356	1.782
13	.128	.694	.870	1.350	1.771
14	.128	.692	.868	1.345	1.761
15	.128	.691	.866	1.341	1.753
16	.128	.690	.865	1.337	1.746
17	.128	.689	.863	1.333	1.740
18	.127	.688	.862	1.330	1.734
19	.127	.688	.861	1.328	1.729
20	.127	.687	.860	1.325	1.725
21	.127	.686	.859	1.323	1.721
22	.127	.686	.858	1.321	1.717
23	.127	.685	.858	1.319	1.714
24	.127	.685	.857	1.318	1.711
25	.127	.684	.856	1.316	1.708
26	.127	.684	.856	1.315	1.706
27	.127	.684	.855	1.314	1.703
28	.127	.683	.855	1.313	1.701
29	.127	.683	.854	1.311	1.699
30	.127	.683	.854	1.310	1.697
40	.126	.681	.851	1.303	1.684
60	.126	.679	.848	1.296	1.671
120	.126	.677	.845	1.289	1.658
∞	.126	.674	.842	1.282	1.645

Nota: Si se desea una prueba de una cola, las probabilidades que encabezan la tabla deben dividirse por dos. Para un número de grados de libertad $\nu > 30$, interpolar entre los valores del argumento ν. La tabla está diseñada para interpolación armónica. Así, para obtener $t_{.05[43]}$ interpolar entre $t_{.05[40]} = 2{,}021$ y $t_{.05[60]} = 2{,}000$, que se dan en la tabla. Transformar los argumentos en $120/\nu = 120/43 = 2{,}791$ e interpolar entre $120/60 = 2{,}000$ y $120/40 = 3{,}000$ por interpolación lineal ordinaria:

$$t_{.05[43]} = (0{,}791 \times 2{,}021) + [(1 - 0{,}791) \times 2{,}000] = 2{,}017$$

Cuando $\nu > 120$, interpolar entre $120/\infty = 0$ y $120/120 = 1$. Los valores de esta tabla se han tomado de otra más extensa (tabla III) de R. A. Fisher y F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, 5th ed. (Oliver & Boyd, Edinburgh, 1958) con permiso de los autores y editores.

0.05	0.02	0.01	0.001	α / ν
12.706	31.821	63.657	636.619	1
4.303	6.965	9.925	31.598	2
3.182	4.541	5.841	12.924	3
2.776	3.747	4.604	8.610	4
2.571	3.365	4.032	6.869	5
2.447	3.143	3.707	5.959	6
2.365	2.998	3.499	5.408	7
2.306	2.896	3.355	5.041	8
2.262	2.821	3.250	4.781	9
2.228	2.764	3.169	4.587	10
2.201	2.718	3.106	4.437	11
2.179	2.681	3.055	4.318	12
2.160	2.650	3.012	4.221	13
2.145	2.624	2.977	4.140	14
2.131	2.602	2.947	4.073	15
2.120	2.583	2.921	4.015	16
2.110	2.567	2.898	3.965	17
2.101	2.552	2.878	3.922	18
2.093	2.539	2.861	3.883	19
2.086	2.528	2.845	3.850	20
2.080	2.518	2.831	3.819	21
2.074	2.508	2.819	3.792	22
2.069	2.500	2.807	3.767	23
2.064	2.492	2.797	3.745	24
2.060	2.485	2.787	3.725	25
2.056	2.479	2.779	3.707	26
2.052	2.473	2.771	3.690	27
2.048	2.467	2.763	3.674	28
2.045	2.462	2.756	3.659	29
2.042	2.457	2.750	3.646	30
2.021	2.423	2.704	3.551	40
2.000	2.390	2.660	3.460	60
1.980	2.358	2.617	3.373	120
1.960	2.326	2.576	3.291	∞

Área α correspondiente al porcentaje de puntos incluidos en dos colas de $\alpha/2$ cada una

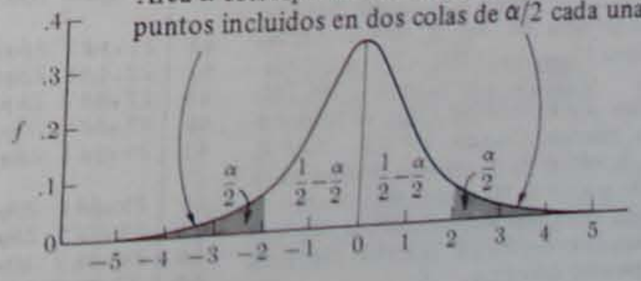


TABLA IV. Valores críticos de la distribución ji-cuadrado.

ν \ α	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	ν
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	5
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	7
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	9
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	10
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757	11
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300	12
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819	13
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319	14
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801	15
16	5.142	6.908	9.312	15.338	23.542	26.296	28.845	32.000	34.267	16
17	5.697	7.564	10.085	16.338	24.769	27.587	30.191	33.409	35.718	17
18	6.265	8.231	10.865	17.338	25.989	28.869	31.526	34.805	37.156	18
19	6.844	8.907	11.651	18.338	27.204	30.144	32.852	36.191	38.582	19
20	7.434	9.591	12.443	19.337	28.412	31.410	34.170	37.566	39.997	20
21	8.034	10.283	13.240	20.337	29.615	32.670	35.479	38.932	41.401	21
22	8.643	10.982	14.042	21.337	30.813	33.924	36.781	40.289	42.796	22
23	9.260	11.688	14.848	22.337	32.007	35.172	38.076	41.638	44.181	23
24	9.886	12.401	15.659	23.337	33.196	36.415	39.364	42.980	45.558	24
25	10.520	13.120	16.473	24.337	34.382	37.652	40.646	44.314	46.928	25
26	11.160	13.844	17.292	25.336	35.563	38.885	41.923	45.642	48.290	26
27	11.808	14.573	18.114	26.336	36.741	40.113	43.194	46.963	49.645	27
28	12.461	15.308	18.939	27.336	37.916	41.337	44.461	48.278	50.993	28
29	13.121	16.047	19.768	28.336	39.088	42.557	45.722	49.588	52.336	29
30	13.787	16.791	20.599	29.336	40.256	43.773	46.979	50.892	53.672	30
31	14.458	17.539	21.434	30.336	41.422	44.985	48.232	52.191	55.003	31
32	15.134	18.291	22.271	31.336	42.585	46.194	49.480	53.486	56.329	32
33	15.815	19.047	23.110	32.336	43.745	47.400	50.725	54.776	57.649	33
34	16.501	19.806	23.952	33.336	44.903	48.602	51.966	56.061	58.964	34
35	17.192	20.569	24.797	34.336	46.059	49.802	53.203	57.342	60.275	35
36	17.887	21.336	25.643	35.336	47.212	50.998	54.437	58.619	61.582	36
37	18.586	22.106	26.492	36.335	48.363	52.192	55.668	59.892	62.884	37
38	19.289	22.878	27.343	37.335	49.513	53.384	56.896	61.162	64.182	38
39	19.996	23.654	28.196	38.335	50.660	54.572	58.120	62.428	65.476	39
40	20.707	24.433	29.051	39.335	51.805	55.758	59.342	63.691	66.766	40
41	21.421	25.215	29.907	40.335	52.949	56.942	60.561	64.950	68.053	41
42	22.138	25.999	30.765	41.335	54.090	58.124	61.777	66.206	69.336	42
43	22.859	26.785	31.625	42.335	55.230	59.304	62.990	67.459	70.616	43
44	23.584	27.575	32.487	43.335	56.369	60.481	64.202	68.710	71.893	44
45	24.311	28.366	33.350	44.335	57.505	61.656	65.410	69.957	73.166	45
46	25.042	29.160	34.215	45.335	58.641	62.830	66.617	71.201	74.437	46
47	25.775	29.956	35.081	46.335	59.774	64.001	67.821	72.443	75.704	47
48	26.511	30.755	35.949	47.335	60.907	65.171	69.023	73.683	76.969	48
49	27.249	31.555	36.818	48.335	62.038	66.339	70.222	74.919	78.231	49
50	27.991	32.357	37.689	49.335	63.167	67.505	71.420	76.154	79.490	50

Nota: Para valores de $\nu > 100$, calcular valores críticos aproximados de χ^2 por la fórmula siguiente: $\chi^2_{\alpha[\nu]} = \frac{1}{2}(t_{2\alpha[\infty]} + \sqrt{2\nu - 1})^2$, donde $t_{2\alpha[\infty]}$ puede buscarse en la tabla III. Así $\chi^2_{.05[120]}$ se calcula como $\frac{1}{2}(t_{.10[\infty]} + \sqrt{240 - 1})^2 = \frac{1}{2}(1.645 + \sqrt{239})^2 = \frac{1}{2}(17.10462)^2 = 146.284$. Para $\alpha > 0{,}5$ utilizar $t_{1-2\alpha[\infty]}$ en la fórmula anterior. Cuando $\alpha = 0{,}5$, $t_{1\infty} = 0$. Los valores de ji-cuadrado desde 1 hasta 30 grados de libertad se han tomado de una tabla más amplia de C. M. Thompson (*Biometrika* 32: 188-189, 1941) con permiso del editor.

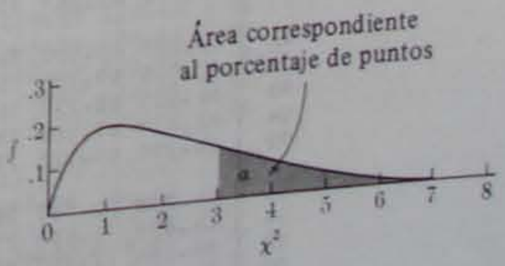


TABLA IV (continuación)

ν \ α	0.995	0.975	0.9	0.5	0.1
51	28.735	33.162	38.560	50.335	64.295
52	29.481	33.968	39.433	51.335	65.422
53	30.230	34.776	40.308	52.335	66.548
54	30.981	35.586	41.183	53.335	67.673
55	31.735	36.398	42.060	54.335	68.796
56	32.490	37.212	42.937	55.335	69.918
57	33.248	38.027	43.816	56.335	71.040
58	34.008	38.844	44.696	57.335	72.160
59	34.770	39.662	45.577	58.335	73.279
60	35.534	40.482	46.459	59.335	74.397
61	36.300	41.303	47.342	60.335	75.514
62	37.068	42.126	48.226	61.335	76.630
63	37.838	42.950	49.111	62.335	77.745
64	38.610	43.776	49.996	63.335	78.860
65	39.383	44.603	50.883	64.335	79.973
66	40.158	45.431	51.770	65.335	81.085
67	40.935	46.261	52.659	66.335	82.197
68	41.713	47.092	53.548	67.334	83.308
69	42.494	47.924	54.438	68.334	84.418
70	43.275	48.758	55.329	69.334	85.527
71	44.058	49.592	56.221	70.334	86.635
72	44.843	50.428	57.113	71.334	87.743
73	45.629	51.265	58.006	72.334	88.850
74	46.417	52.103	58.900	73.334	89.956
75	47.206	52.942	59.795	74.334	91.061
76	47.997	53.782	60.690	75.334	92.166
77	48.788	54.623	61.586	76.334	93.270
78	49.582	55.466	62.483	77.334	94.373
79	50.376	56.309	63.380	78.334	95.476
80	51.172	57.153	64.278	79.334	96.578
81	51.969	57.998	65.176	80.334	97.680
82	52.767	58.845	66.076	81.334	98.780
83	53.567	59.692	66.976	82.334	99.880
84	54.368	60.540	67.876	83.334	100.98
85	55.170	61.389	68.777	84.334	102.08
86	55.973	62.239	69.679	85.334	103.18
87	56.777	63.089	70.581	86.334	104.28
88	57.582	63.941	71.484	87.334	105.37
89	58.389	64.793	72.387	88.334	106.47
90	59.196	65.647	73.291	89.334	107.56
91	60.005	66.501	74.196	90.334	108.66
92	60.815	67.356	75.101	91.334	109.76
93	61.625	68.211	76.006	92.334	110.85
94	62.437	69.068	76.912	93.334	111.94
95	63.250	69.925	77.818	94.334	113.04
96	64.063	70.783	78.725	95.334	114.13
97	64.878	71.642	79.633	96.334	115.22
98	65.694	72.501	80.541	97.334	116.32
99	66.510	73.361	81.449	98.334	117.41
100	67.328	74.222	82.358	99.334	118.50

0.05	0.025	0.01	0.005	ν
68.669	72.616	77.386	80.747	51
69.832	73.810	78.616	82.001	52
70.993	75.002	79.843	83.253	53
72.153	76.192	81.069	84.502	54
73.311	77.380	82.292	85.749	55
74.468	78.567	83.513	86.994	56
75.624	79.752	84.733	88.237	57
76.778	80.936	85.950	89.477	58
77.931	82.117	87.166	90.715	59
79.082	83.298	88.379	91.952	60
80.232	84.476	89.591	93.186	61
81.381	85.654	90.802	94.419	62
82.529	86.830	92.010	95.649	63
83.675	88.004	93.217	96.878	64
84.821	89.177	94.422	98.105	65
85.965	90.349	95.626	99.331	66
87.108	91.519	96.828	100.55	67
88.250	92.689	98.028	101.78	68
89.391	93.856	99.228	103.00	69
90.531	95.023	100.43	104.21	70
91.670	96.189	101.62	105.43	71
92.808	97.353	102.82	106.65	72
93.945	98.516	104.01	107.86	73
95.081	99.678	105.20	109.07	74
96.217	100.84	106.39	110.29	75
97.351	102.00	107.58	111.50	76
98.484	103.16	108.77	112.70	77
99.617	104.32	109.96	113.91	78
100.75	105.47	111.14	115.12	79
101.88	106.63	112.33	116.32	80
103.01	107.78	113.51	117.52	81
104.14	108.94	114.69	118.73	82
105.27	110.09	115.88	119.93	83
106.39	111.24	117.06	121.13	84
107.52	112.39	118.24	122.32	85
108.65	113.54	119.41	123.52	86
109.77	114.69	120.59	124.72	87
110.90	115.84	121.77	125.91	88
112.02	116.99	122.94	127.11	89
113.15	118.14	124.12	128.30	90
114.27	119.28	125.29	129.49	91
115.39	120.43	126.46	130.68	92
116.51	121.57	127.63	131.87	93
117.63	122.72	128.80	133.06	94
118.75	123.86	129.97	134.25	95
119.87	125.00	131.14	135.43	96
120.99	126.14	132.31	136.62	97
122.11	127.28	133.48	137.80	98
123.23	128.42	134.64	138.99	99
124.34	129.56	135.81	140.17	100

TABLA V. Valores críticos de la distribución F.

ν_2 (grados de libertad de la media cuadrática del denominador)	α	ν_1 (grados de libertad de la media cuadrática del numerador) 1	2	3	4	5	6
1	.05	161	199	216	225	230	234
	.025	648	800	864	900	922	937
	.01	4050	5000	5400	5620	5760	5860
2	.05	18.5	19.0	19.2	19.2	19.3	19.3
	.025	38.5	39.0	39.2	39.2	39.3	39.3
	.01	98.5	99.0	99.2	99.2	99.3	99.3
3	.05	10.1	9.55	9.28	9.12	9.01	8.94
	.025	17.4	16.0	15.4	15.1	14.9	14.7
	.01	34.1	30.8	29.5	28.7	28.2	27.9
4	.05	7.71	6.94	6.59	6.39	6.26	6.16
	.025	12.2	10.6	9.98	9.60	9.36	9.20
	.01	21.2	18.0	16.7	16.0	15.5	15.2
5	.05	6.61	5.79	5.41	5.19	5.05	4.95
	.025	10.0	8.43	7.76	7.39	7.15	6.98
	.01	16.3	13.3	12.1	11.4	11.0	10.7
6	.05	5.99	5.14	4.76	4.53	4.39	4.28
	.025	8.81	7.26	6.60	6.23	5.99	5.82
	.01	13.7	10.9	9.78	9.15	8.75	8.47
7	.05	5.59	4.74	4.35	4.12	3.97	3.87
	.025	8.07	6.54	5.89	5.52	5.29	5.12
	.01	12.2	9.55	8.45	7.85	7.46	7.19
8	.05	5.32	4.46	4.07	3.84	3.69	3.58
	.025	7.57	6.06	5.42	5.05	4.82	4.65
	.01	11.3	8.65	7.59	7.01	6.63	6.37
9	.05	5.12	4.26	3.86	3.63	3.48	3.37
	.025	7.21	5.71	5.08	4.72	4.48	4.32
	.01	10.6	8.02	6.99	6.42	6.06	5.80
10	.05	4.96	4.10	3.71	3.48	3.33	3.22
	.025	6.94	5.46	4.83	4.47	4.24	4.07
	.01	10.0	7.56	6.55	5.99	5.64	5.39

Nota: La interpolación para un número de grados de libertad no dados en los argumentos se hace por medio de interpolación armónica (véase nota al pie de la tabla III). Si los dos ν_1 y ν_2 requieren interpolación, es necesario interpolar para cada uno de estos argumentos sucesivamente. Así, para obtener $F_{.05[32,60]}$ se interpola primero entre $F_{.05[30,40]}$ y $F_{.05[60,60]}$ y entre $F_{.05[30,120]}$ y $F_{.05[60,120]}$, para estimar $F_{.05[32,60]}$ y $F_{.05[32,120]}$, respectivamente. Luego se interpola entre estos dos valores para obtener la cantidad deseada. Las entradas para $\alpha = 0{,}05$, 0,025, 0,01 y 0,005, y para ν_1 y $\nu_2 = 1$ a 10, 12, 15, 20, 24, 30, 40, 60, 120 e ∞ se han copiado de una tabla de M. Merrington y C. M. Thompson (*Biometrika* 33: 73-88, 1943) con permiso del editor.

ν_2 (grados de libertad de la media cuadrática del denominador)	α	ν_1 (grados de libertad de la media cuadrática del numerador) 7	8	9	10	11
1	.05	237	239	241	241	243
	.025	948	957	963	969	973
	.01	5930	5980	6020	6060	6080
2	.05	19.4	19.4	19.4	19.4	19.4
	.025	39.4	39.4	39.4	39.4	39.4
	.01	99.4	99.4	99.4	99.4	99.4
3	.05	8.89	8.85	8.81	8.79	8.76
	.025	14.6	14.5	14.5	14.4	14.3
	.01	27.7	27.5	27.3	27.2	27.1
4	.05	6.09	6.04	6.00	5.96	5.93
	.025	9.07	8.98	8.90	8.84	8.79
	.01	15.0	14.8	14.7	14.5	14.4
5	.05	4.88	4.82	4.77	4.74	4.71
	.025	6.85	6.76	6.68	6.62	6.57
	.01	10.5	10.3	10.2	10.1	9.99
6	.05	4.21	4.15	4.10	4.06	4.03
	.025	5.70	5.60	5.52	5.46	5.41
	.01	8.26	8.10	7.98	7.87	7.79
7	.05	3.77	3.73	3.68	3.64	3.60
	.025	4.99	4.89	4.82	4.76	4.71
	.01	6.99	6.84	6.72	6.62	6.54
8	.05	3.50	3.44	3.39	3.35	3.31
	.025	4.53	4.43	4.36	4.30	4.25
	.01	6.18	6.03	5.91	5.81	5.73
9	.05	3.29	3.23	3.18	3.14	3.10
	.025	4.20	4.10	4.03	3.96	3.91
	.01	5.61	5.47	5.35	5.26	5.18
10	.05	3.14	3.07	3.02	2.98	2.94
	.025	3.95	3.85	3.78	3.72	3.67
	.01	5.20	5.06	4.94	4.85	4.77

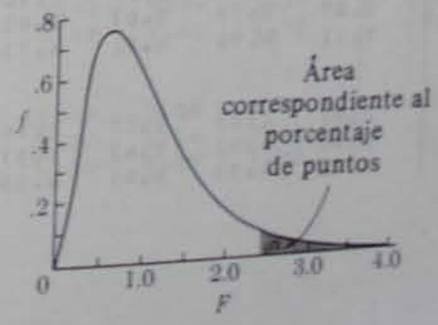


TABLA V (continuación)

ν_2	α	ν_1 = 12	15	20	24	30
1	.05	244	246	248	249	250
	.025	977	985	993	997	1000
	.01	6110	6160	6210	6230	6260
2	.05	19.4	19.4	19.4	19.5	19.5
	.025	39.4	39.4	39.4	39.5	39.5
	.01	99.4	99.4	99.4	99.5	99.5
3	.05	8.74	8.70	8.66	8.64	8.62
	.025	14.3	14.3	14.2	14.1	14.1
	.01	27.1	26.9	26.7	26.6	26.5
4	.05	5.91	5.86	5.80	5.77	5.75
	.025	8.75	8.66	8.56	8.51	8.46
	.01	14.4	14.2	14.0	13.9	13.8
5	.05	4.68	4.62	4.56	4.53	4.50
	.025	6.52	6.43	6.33	6.28	6.23
	.01	9.89	9.72	9.55	9.47	9.38
6	.05	4.00	3.94	3.87	3.84	3.81
	.025	5.37	5.27	5.17	5.12	5.07
	.01	7.72	7.56	7.40	7.31	7.23
7	.05	3.57	3.51	3.44	3.41	3.38
	.025	4.67	4.57	4.47	4.42	4.36
	.01	6.47	6.31	6.16	6.07	5.99
8	.05	3.28	3.22	3.15	3.12	3.08
	.025	4.20	4.10	4.00	3.95	3.89
	.01	5.67	5.52	5.36	5.28	5.20
9	.05	3.07	3.01	2.94	2.90	2.86
	.025	3.87	3.77	3.67	3.61	3.56
	.01	5.11	4.96	4.81	4.73	4.65
10	.05	2.91	2.85	2.77	2.74	2.70
	.025	3.62	3.52	3.42	3.37	3.31
	.01	4.71	4.56	4.41	4.33	4.25

ν_2 (grados de libertad de la media cuadrática del denominador)

ν_2	α	ν_1 = 40	60	120	∞	α	ν_2
1	.05	251	252	253	254	.05	1
	.025	1010	1010	1010	1020	.025	
	.01	6290	6310	6340	6370	.01	
2	.05	19.5	19.5	19.5	19.5	.05	2
	.025	39.5	39.5	39.5	39.5	.025	
	.01	99.5	99.5	99.5	99.5	.01	
3	.05	8.59	8.57	8.55	8.53	.05	3
	.025	14.0	14.0	13.9	13.9	.025	
	.01	26.4	26.3	26.2	26.1	.01	
4	.05	5.72	5.69	5.66	5.63	.05	4
	.025	8.41	8.36	8.31	8.26	.025	
	.01	13.7	13.7	13.6	13.5	.01	
5	.05	4.46	4.43	4.40	4.36	.05	5
	.025	6.18	6.12	6.07	6.02	.025	
	.01	9.29	9.20	9.11	9.02	.01	
6	.05	3.77	3.74	3.70	3.67	.05	6
	.025	5.01	4.96	4.90	4.85	.025	
	.01	7.14	7.06	6.97	6.88	.01	
7	.05	3.34	3.30	3.27	3.23	.05	7
	.025	4.31	4.25	4.20	4.14	.025	
	.01	5.91	5.82	5.74	5.65	.01	
8	.05	3.04	3.01	2.97	2.93	.05	8
	.025	3.84	3.78	3.73	3.67	.025	
	.01	5.12	5.03	4.95	4.86	.01	
9	.05	2.83	2.79	2.75	2.71	.05	9
	.025	3.51	3.45	3.39	3.33	.025	
	.01	4.57	4.48	4.40	4.31	.01	
10	.05	2.66	2.62	2.58	2.54	.05	10
	.025	3.26	3.20	3.14	3.08	.025	
	.01	4.17	4.08	4.00	3.91	.01	

ν_1 (grados de libertad de la media cuadrática del numerador)

TABLA V (continuación)

ν_2 (grados de libertad de la media cuadrática del denominador)

ν_2	α	ν_1 = 1	2	3	4	5	6
11	.05	4.84	3.98	3.59	3.36	3.20	3.09
	.025	6.72	5.26	4.63	4.28	4.04	3.88
	.01	9.65	7.21	6.22	5.67	5.32	5.07
12	.05	4.75	3.89	3.49	3.26	3.11	3.00
	.025	6.55	5.10	4.47	4.12	3.89	3.73
	.01	9.33	6.93	5.95	5.41	5.06	4.82
15	.05	4.54	3.68	3.29	3.06	2.90	2.79
	.025	6.20	4.77	4.15	3.80	3.58	3.41
	.01	8.68	6.36	5.42	4.89	4.56	4.32
20	.05	4.35	3.49	3.10	2.87	2.71	2.60
	.025	5.87	4.46	3.86	3.51	3.29	3.13
	.01	8.10	5.85	4.94	4.43	4.10	3.87
24	.05	4.26	3.40	3.01	2.78	2.62	2.51
	.025	5.72	4.32	3.72	3.38	3.15	2.99
	.01	7.82	5.61	4.72	4.22	3.90	3.67
30	.05	4.17	3.32	2.92	2.69	2.53	2.42
	.025	5.57	4.18	3.59	3.25	3.03	2.87
	.01	7.56	5.39	4.51	4.02	3.70	3.47
40	.05	4.08	3.23	2.84	2.61	2.45	2.34
	.025	5.42	4.05	3.46	3.13	2.90	2.74
	.01	7.31	5.18	4.31	3.83	3.51	3.29
60	.05	4.00	3.15	2.76	2.53	2.37	2.25
	.025	5.29	3.93	3.34	3.01	2.79	2.63
	.01	7.08	4.98	4.13	3.65	3.34	3.12
120	.05	3.92	3.07	2.68	2.45	2.29	2.17
	.025	5.15	3.80	3.23	2.89	2.67	2.52
	.01	6.85	4.79	3.95	3.48	3.17	2.96
∞	.05	3.84	3.00	2.60	2.37	2.21	2.10
	.025	5.02	3.69	3.11	2.79	2.57	2.41
	.01	6.63	4.61	3.78	3.32	3.02	2.80

ν_1 (grados de libertad de la media cuadrática del numerador)

7	8	9	10	11	α	ν_2
3.01	2.95	2.90	2.85	2.82	.05	11
3.76	3.66	3.59	3.53	3.48	.025	
4.89	4.74	4.63	4.54	4.46	.01	
2.91	2.85	2.80	2.75	2.72	.05	12
3.61	3.51	3.44	3.37	3.32	.025	
4.64	4.50	4.39	4.30	4.22	.01	
2.71	2.64	2.59	2.54	2.51	.05	15
3.29	3.20	3.12	3.06	3.01	.025	
4.14	4.00	3.89	3.80	3.73	.01	
2.51	2.45	2.39	2.35	2.31	.05	20
3.01	2.91	2.84	2.77	2.72	.025	
3.70	3.56	3.46	3.37	3.29	.01	
2.42	2.36	2.30	2.25	2.22	.05	24
2.87	2.78	2.70	2.64	2.59	.025	
3.50	3.36	3.26	3.17	3.09	.01	
2.33	2.27	2.21	2.16	2.13	.05	30
2.75	2.65	2.57	2.51	2.46	.025	
3.30	3.17	3.07	2.98	2.90	.01	
2.25	2.18	2.12	2.08	2.04	.05	40
2.62	2.53	2.45	2.39	2.33	.025	
3.12	2.99	2.89	2.80	2.73	.01	
2.17	2.10	2.04	1.99	1.95	.05	60
2.51	2.41	2.33	2.27	2.22	.025	
2.95	2.82	2.72	2.63	2.56	.01	
2.09	2.02	1.96	1.91	1.87	.05	120
2.39	2.30	2.22	2.16	2.10	.025	
2.79	2.66	2.56	2.47	2.40	.01	
2.01	1.94	1.88	1.83	1.79	.05	∞
2.29	2.19	2.11	2.05	1.99	.025	
2.64	2.51	2.41	2.32	2.25	.01	

ν_2 (grados de libertad de la media cuadrática del denominador)

TABLA V (continuación)

ν_2 (grados de libertad de la media cuadrática del denominador)	α	ν_1 (grados de libertad de la media cuadrática del numerador) 12	15	20	24	30
11	.05	2.79	2.72	2.65	2.61	2.57
	.025	3.43	3.33	3.23	3.17	3.12
	.01	4.40	4.25	4.10	4.02	3.94
12	.05	2.69	2.62	2.54	2.51	2.47
	.025	3.28	3.18	3.07	3.02	2.96
	.01	4.16	4.01	3.86	3.78	3.70
15	.05	2.48	2.40	2.33	2.39	2.25
	.025	2.96	2.86	2.76	2.70	2.64
	.01	3.67	3.52	3.37	3.29	3.21
20	.05	2.28	2.20	2.12	2.08	2.04
	.025	2.68	2.57	2.46	2.41	2.35
	.01	3.23	3.09	2.94	2.86	2.78
24	.05	2.18	2.11	2.03	1.98	1.94
	.025	2.54	2.44	2.33	2.27	2.21
	.01	3.03	2.89	2.74	2.66	2.58
30	.05	2.09	2.01	1.93	1.89	1.84
	.025	2.41	2.31	2.20	2.14	2.07
	.01	2.84	2.70	2.55	2.47	2.39
40	.05	2.04	1.92	1.84	1.79	1.74
	.025	2.29	2.18	2.07	2.01	1.94
	.01	2.66	2.52	2.37	2.29	2.20
60	.05	1.92	1.84	1.75	1.70	1.65
	.025	2.17	2.06	1.94	1.88	1.82
	.01	2.50	2.35	2.20	2.12	2.03
120	.05	1.83	1.75	1.66	1.61	1.55
	.025	2.05	1.95	1.82	1.76	1.69
	.01	2.34	2.19	2.03	1.95	1.86
∞	.05	1.75	1.67	1.57	1.52	1.46
	.025	1.94	1.83	1.71	1.64	1.57
	.01	2.18	2.04	1.88	1.79	1.70

ν_1 (grados de libertad de la media cuadrática del numerador) 40	60	120	∞	α	ν_2 (grados de libertad de la media cuadrática del denominador)
2.53	2.49	2.45	2.40	.05	11
3.06	3.00	2.94	2.88	.025	
3.86	3.78	3.69	3.60	.01	
2.43	2.38	2.34	2.30	.05	12
2.91	2.85	2.79	2.72	.025	
3.62	3.54	3.45	3.36	.01	
2.20	2.16	2.11	2.07	.05	15
2.59	2.52	2.46	2.40	.025	
3.13	3.05	2.96	2.87	.01	
1.99	1.95	1.90	1.84	.05	20
2.29	2.22	2.16	2.09	.025	
2.69	2.61	2.52	2.42	.01	
1.89	1.84	1.79	1.73	.05	24
2.15	2.08	2.01	1.94	.025	
2.49	2.40	2.31	2.21	.01	
1.79	1.74	1.68	1.62	.05	30
2.01	1.94	1.87	1.79	.025	
2.30	2.21	2.11	2.01	.01	
1.69	1.64	1.58	1.51	.05	40
1.88	1.80	1.72	1.64	.025	
2.11	2.02	1.92	1.80	.01	
1.59	1.53	1.47	1.39	.05	60
1.74	1.67	1.58	1.48	.025	
1.94	1.84	1.73	1.60	.01	
1.50	1.43	1.35	1.25	.05	120
1.61	1.53	1.43	1.31	.025	
1.76	1.66	1.53	1.38	.01	
1.39	1.32	1.22	1.00	.05	∞
1.48	1.39	1.27	1.00	.025	
1.59	1.47	1.32	1.00	.01	

TABLA VI. Valores críticos de $F_{máx}$

ν \ a	2	3	4	5	6
2	39.0	87.5	142.	202.	266.
	199.	448.	729.	1036	1362
3	15.4	27.8	39.2	50.7	62.0
	47.5	85.	120.	151.	184.
4	9.60	15.5	20.6	25.2	29.5
	23.2	37.	49.	59.	69.
5	7.15	10.8	13.7	16.3	18.7
	14.9	22.	28.	33.	38.
6	5.82	8.38	10.4	12.1	13.7
	11.1	15.5	19.1	22.	25.
7	4.99	6.94	8.44	9.70	10.8
	8.89	12.1	14.5	16.5	18.4
8	4.43	6.00	7.18	8.12	9.03
	7.50	9.9	11.7	13.2	14.5
9	4.03	5.34	6.31	7.11	7.80
	6.54	8.5	9.9	11.1	12.1
10	3.72	4.85	5.67	6.34	6.92
	5.85	7.4	8.6	9.6	10.4
12	3.28	4.16	4.79	5.30	5.72
	4.91	6.1	6.9	7.6	8.2
15	2.86	3.54	4.01	4.37	4.68
	4.07	4.9	5.5	6.0	6.4
20	2.46	2.95	3.29	3.54	3.76
	3.32	3.8	4.3	4.6	4.9
30	2.07	2.40	2.61	2.78	2.91
	2.63	3.0	3.3	3.4	3.6
60	1.67	1.85	1.96	2.04	2.11
	1.96	2.2	2.3	2.4	2.4
∞	1.00	1.00	1.00	1.00	1.00
	1.00	1.0	1.0	1.0	1.0

Nota: Para cada valor de a (número de muestras) y ν (grados de libertad) hay dos valores críticos de F_{max} que representan los puntos 5 % y 1 %. Las probabilidades correspondientes a $\alpha = 0{,}05$ y $0{,}01$ representan *una cola* de la distribución F_{max}. Esta tabla se ha copiado de H. A. David (*Biometrika* 39: 422-424, 1952) con permiso del editor y autor.

7	8	9	10	11	12
333.	403.	475.	550.	626.	704.
1705	2063	2432	2813	3204	3605
72.9	83.5	93.9	104.	114.	124.
21(6)	24(9)	28(1)	31(0)	33(7)	36(1)
33.6	37.5	41.1	44.6	48.0	51.4
79.	89.	97.	106.	113.	120.
20.8	22.9	24.7	26.5	28.2	29.9
42.	46.	50.	54.	57.	60.
15.0	16.3	17.5	18.6	19.7	20.7
27.	30.	32.	34.	36.	37.
11.8	12.7	13.5	14.3	15.1	15.8
20.	22.	23.	24.	26.	27.
9.78	10.5	11.1	11.7	12.2	12.7
15.8	16.9	17.9	18.9	19.8	21.
8.41	8.95	9.45	9.91	10.3	10.7
13.1	13.9	14.7	15.3	16.0	16.6
7.42	7.87	8.28	8.66	9.01	9.34
11.1	11.8	12.4	12.9	13.4	13.9
6.09	6.42	6.72	7.00	7.25	7.48
8.7	9.1	9.5	9.9	10.2	10.6
4.95	5.19	5.40	5.59	5.77	5.93
6.7	7.1	7.3	7.5	7.8	8.0
3.94	4.10	4.24	4.37	4.49	4.59
5.1	5.3	5.5	5.6	5.8	5.9
3.02	3.12	3.21	3.29	3.36	3.39
3.7	3.8	3.9	4.0	4.1	4.2
2.17	2.22	2.26	2.30	2.33	2.36
2.5	2.5	2.6	2.6	2.7	2.7
1.00	1.00	1.00	1.00	1.00	1.00
1.0	1.0	1.0	1.0	1.0	1.0

TABLA VII. Límites de confianza, más cortos, no sesgados, de la varianza.

ν	Coeficientes de confianza 0.95	0.99	ν	Coeficientes de confianza 0.95	0.99	ν	Coeficientes de confianza 0.95	0.99
2	.2099 23.605	.1505 114.489	14	.5135 2.354	.4289 3.244	26	.6057 1.825	.5261 2.262
3	.2681 10.127	.1983 29.689	15	.5242 2.276	.4399 3.091	27	.6110 1.802	.5319 2.223
4	.3125 6.590	.2367 15.154	16	.5341 2.208	.4502 2.961	28	.6160 1.782	.5374 2.187
5	.3480 5.054	.2685 10.076	17	.5433 2.149	.4598 2.848	29	.6209 1.762	.5427 2.153
6	.3774 4.211	.2956 7.637	18	.5520 2.097	.4689 2.750	30	.6255 1.744	.5478 2.122
7	.4025 3.679	.3192 6.238	19	.5601 2.050	.4774 2.664	40	.6636 1.608	.5900 1.896
8	.4242 3.314	.3400 5.341	20	.5677 2.008	.4855 2.588	50	.6913 1.523	.6213 1.760
9	.4432 3.048	.3585 4.720	21	.5749 1.971	.4931 2.519	60	.7128 1.464	.6458 1.668
10	.4602 2.844	.3752 4.265	22	.5817 1.936	.5004 2.458	70	.7300 1.421	.6657 1.607
11	.4755 2.683	.3904 3.919	23	.5882 1.905	.5073 2.402	80	.7443 1.387	.6824 1.549
12	.4893 2.553	.4043 3.646	24	.5943 1.876	.5139 2.351	90	.7564 1.360	.6966 1.508
13	.5019 2.445	.4171 3.426	25	.6001 1.850	.5201 2.305	100	.7669 1.338	.7090 1.475

Nota: Los factores de esta tabla se han obtenido dividiendo la cantidad $n - 1$ por los valores encontrados en una tabla preparada por D. V. Lindley, D. A. East, y P. A. Hamilton (*Biometrika* 47: 433-437, 1960).

TABLA VIII. Valores críticos para coeficientes de correlación.

ν	α	r	ν	α	r	ν	α	r
1	.05	.997	16	.05	.468	35	.05	.325
	.01	1.000		.01	.590		.01	.418
2	.05	.950	17	.05	.456	40	.05	.304
	.01	.990		.01	.575		.01	.393
3	.05	.878	18	.05	.444	45	.05	.288
	.01	.959		.01	.561		.01	.372
4	.05	.811	19	.05	.433	50	.05	.273
	.01	.917		.01	.549		.01	.354
5	.05	.754	20	.05	.423	60	.05	.250
	.01	.874		.01	.537		.01	.325
6	.05	.707	21	.05	.413	70	.05	.232
	.01	.834		.01	.526		.01	.302
7	.05	.666	22	.05	.404	80	.05	.217
	.01	.798		.01	.515		.01	.283
8	.05	.632	23	.05	.396	90	.05	.205
	.01	.765		.01	.505		.01	.267
9	.05	.602	24	.05	.388	100	.05	.195
	.01	.735		.01	.496		.01	.254
10	.05	.576	25	.05	.381	125	.05	.174
	.01	.708		.01	.487		.01	.228
11	.05	.553	26	.05	.374	150	.05	.159
	.01	.684		.01	.478		.01	.208
12	.05	.532	27	.05	.367	200	.05	.138
	.01	.661		.01	.470		.01	.181
13	.05	.514	28	.05	.361	300	.05	.113
	.01	.641		.01	.463		.01	.148
14	.05	.497	29	.05	.355	400	.05	.098
	.01	.623		.01	.456		.01	.128
15	.05	.482	30	.05	.349	500	.05	.088
	.01	.606		.01	.449		.01	.115
						1.000	.05	.062
							.01	.081

Nota: El valor de arriba es el valor crítico 5 %, el de abajo el 1 %. Esta tabla se ha reproducido con permiso de *Statistical Methods*, 5th edition, de George W. Snedecor, © 1956, por The Iowa State University Press.

TABLA IX. Límites de confianza para porcentajes

Esta tabla da los límites de confianza para porcentajes basados en la distribución binomial.

La primera parte de la tabla da los límites para muestras hasta un tamaño $n = 30$. Los argumentos son Y, el número de ítems de la muestra que presenta una determinada propiedad, y n, el tamaño de muestreo. El argumento Y está tabulado para valores enteros entre 0 y 15, lo que da porcentajes de hasta el 50 %. Para cada tamaño de muestreo n y número de ítems con la propiedad dada, se muestran tres líneas de valores numéricos. La primera línea de valores da los límites de confianza del 95 % para el porcentaje, la segunda línea presenta el porcentaje de incidencia observado de esa propiedad, y la tercera línea de valores da los límites de confianza del 99 % para el porcentaje. Así, por ejemplo, para $Y = 8$ individuos que presentan la propiedad entre una muestra de $n = 20$, la segunda línea indica que esto representa una incidencia de la propiedad de 40,00 %, la primera línea da los límites de confianza al 95 % de este porcentaje que son de 19,10 % a 63,95 %, y la tercera línea da los límites de confianza al 99 % que son de 14,60 % a 70,10 %.

Interpolar en esta tabla (hasta $n = 49$) dividiendo $L_1{}^-$ y $L_2{}^-$, los límites de confianza inferior y superior al tamaño de muestreo inmediato inferior tabulado n^- por el tamaño de muestreo deseado n, y multiplicarlos por el tamaño de muestreo inmediato inferior tabulado n^-. Así, por ejemplo, para obtener los límites de confianza del porcentaje correspondiente a 8 individuos que exhiben la propiedad en una muestra de 22 individuos (lo que corresponde al 36,36 % de los individuos), calcular el límite de confianza inferior $L_1 = L_1{}^- n^-/n = (19{,}10)20/22 = 17{,}36$ % y el límite de confianza superior $L_2 = L_2{}^- n^-/n = (63{,}95)20/22 = 58{,}14$ %.

La segunda mitad de la tabla es para tamaños de muestreo más grandes ($n = 50$, 100, 200, 500 y 1 000). Los argumentos que aparecen a lo largo del margen izquierdo de la tabla son ahora porcentajes desde 0 % hasta 50 % en incrementos del 1 %, en vez de números. Las funciones que se dan en la tabla en dos líneas son los límites de confianza al 95 % y 99 % correspondientes a un porcentaje de incidencia p y un tamaño muestral n determinado. Por ejemplo, los límites de confianza al 99 % de una incidencia observada del 12 % en una muestra de 500 se encuentra que son 8,56-16,19 %, en la segunda de las dos líneas. En esta tabla, la interpolación entre los tamaños de muestreo provistos puede realizarse por medio de la siguiente fórmula para el límite inferior

$$L_1 = \frac{[L_1{}^- n^-(n^+ - n) + L_1{}^+ n^+(n - n^-)]}{n(n^+ - n^-)}$$

En la expresión anterior, n es el tamaño de la muestra observada, n^- y n^+ son los tamaños de muestreo inmediatos inferior y superior tabulados, respectivamente, $L_1{}^-$ y $L_1{}^+$ son los límites de confianza tabulados correspondientes a estos tamaños de muestreo, y L_1 es el límite de confianza inferior a calcular por interpolación. El límite de confianza superior, L_2, puede obtenerse por una fórmula análoga sustituyendo el subíndice 2 por 1. A modo

de ejemplo ilustraremos el establecimiento de límites de confianza del 95 % para un porcentaje observado del 25 % en un tamaño muestral de 80. Los límites del 95 % tabulados para $n = 50$ son 13,84-39,27 %. Para $n = 100$ los límites tabulados correspondientes son 16,88-34,66 %. Cuando sustituimos los valores para los límites inferiores en la fórmula anterior obtenemos

$$L_1 = [(13{,}84)(50)(100 - 80) + (16{,}88)(100)(80 - 50)]/80(100 - 50) = 16{,}12\%$$

para el límite de confianza inferior. Igualmente, para el límite de confianza superior calculamos

$$L_2 = [(39{,}27)(50)(100 - 80) + (34{,}66)(100)(80 - 50)]/80(100 - 50) = 35{,}81\%$$

Los valores tabulados en paréntesis son límites para porcentajes que no podrían obtenerse en ningún problema de muestreo real (por ejemplo, 25 % en 50 ítems), pero son necesarios para interpolación. Para porcentajes superiores al 50 % buscar el porcentaje complementario como el argumento. Los complementos de los límites de confianza binomiales tabulados son los límites deseados.

Estas tablas se han extraído de otras más extensas de D. Mainland, L. Herrera and M. I. Sutcliffe, *Tables for Use with binomial Samples* (1956) con permiso de los autores. Las fórmulas de interpolación citadas se deben también a estos autores. Los límites de confianza de porcentajes impares hasta 13 % para $n = 50$ se han calculado por interpolación.

TABLA IX. Límites de confianza para porcentajes.

Y	Coeficientes de confianza	n = 5	n = 10	n = 15	n = 20	n = 25	n = 30	Coeficientes de confianza	Y
0	95	0.00 - 52.18	0.00 - 30.85	0.00 - 21.80	0.00 - 16.85	0.00 - 13.72	0.00 - 11.57	95	0
		0.00	0.00	0.00	0.00	0.00	0.00		
	99	0.00 - 65.34	0.00 - 41.13	0.00 - 29.76	0.00 - 23.27	0.00 - 19.10	0.00 - 16.19	99	
1	95	0.51 - 71.60	0.25 - 44.50	0.17 - 32.00	0.13 - 24.85	0.10 - 20.36	0.08 - 17.23	95	1
		20.00	10.00	6.67	5.00	4.00	3.33		
	99	0.10 - 81.40	0.05 - 54.40	0.03 - 40.27	0.02 - 31.70	0.02 - 26.24	0.02 - 22.33	99	
2	95	5.28 - 85.34	2.52 - 55.60	1.66 - 40.49	1.24 - 31.70	0.98 - 26.05	0.82 - 22.09	95	2
		40.00	20.00	13.33	10.00	8.00	6.67		
	99	2.28 - 91.72	1.08 - 64.80	0.71 - 48.71	0.53 - 38.70	0.42 - 32.08	0.35 - 27.35	99	
3	95		6.67 - 65.20	4.33 - 48.07	3.21 - 37.93	2.55 - 31.24	2.11 - 26.53	95	3
			30.00	20.00	15.00	12.00	10.00		
	99		3.70 - 73.50	2.39 - 56.07	1.77 - 45.05	1.40 - 37.48	1.16 - 32.03	99	
4	95		12.20 - 73.80	7.80 - 55.14	5.75 - 43.65	4.55 - 36.10	3.77 - 30.74	95	4
			40.00	26.67	20.00	16.00	13.33		
	99		7.68 - 80.91	4.88 - 62.78	3.58 - 50.65	2.83 - 42.41	2.34 - 36.39	99	
5	95		18.70 - 81.30	11.85 - 61.62	8.68 - 49.13	6.84 - 40.72	5.64 - 34.74	95	5
			50.00	33.33	25.00	20.00	16.67		
	99		12.80 - 87.20	8.03 - 68.89	5.85 - 56.05	4.60 - 47.00	3.79 - 40.44	99	
6	95			16.33 - 67.74	11.90 - 54.30	9.35 - 45.14	7.70 - 38.56	95	6
				40.00	30.00	24.00	20.00		
	99			11.67 - 74.40	8.45 - 60.95	6.62 - 51.38	5.43 - 44.26	99	
7	95			21.29 - 73.38	15.38 - 59.20	12.06 - 49.38	9.92 - 42.29	95	7
				46.67	35.00	28.00	23.33		
	99			15.87 - 79.54	11.40 - 65.70	8.90 - 55.56	7.29 - 48.01	99	
8	95				19.10 - 63.95	14.96 - 53.50	12.29 - 45.89	95	8
					40.00	32.00	26.67		
	99				14.60 - 70.10	11.36 - 59.54	9.30 - 51.58	99	
9	95				23.05 - 68.48	17.97 - 57.48	14.73 - 49.40	95	9
					45.00	36.00	30.00		
	99				18.08 - 74.30	14.01 - 63.36	11.43 - 55.00	99	
10	95				27.20 - 72.80	21.12 - 61.32	17.29 - 52.80	95	10
					50.00	40.00	33.33		
	99				21.75 - 78.25	16.80 - 67.04	13.69 - 58.35	99	
11	95					24.41 - 65.06	19.93 - 56.13	95	11
						44.00	36.67		
	99					19.75 - 70.55	16.06 - 61.57	99	
12	95					27.81 - 68.69	22.66 - 59.39	95	12
						48.00	40.00		
	99					22.84 - 73.93	18.50 - 64.69	99	
13	95						25.46 - 62.56	95	13
							43.33		
	99						21.07 - 67.72	99	
14	95						28.35 - 65.66	95	14
							46.67		
	99						23.73 - 70.66	99	
15	95						31.30 - 68.70	95	15
							50.00		
	99						26.47 - 73.53	99	

TABLA IX (continuación)

%	Coeficientes de confianza	n = 50	n = 100	n = 200	n = 500	n = 1000
0	95	.00- 7.11	.00- 3.62	.00- 1.83	.00- 0.74	.00- 0.37
	99	.00-10.05	.00- 5.16	.00- 2.62	.00- 1.05	.00- 0.53
1	95	(.02- 8.88)	.02- 5.45	.12- 3.57	.32- 2.32	.48- 1.83
	99	(.00-12.02)	.00- 7.21	.05- 4.55	.22- 2.80	.37- 2.13
2	95	.05-10.66	.24- 7.04	.55- 5.04	1.06- 3.56	1.29- 3.01
	99	.01-13.98	.10- 8.94	.34- 6.17	.87- 4.12	1.13- 3.36
3	95	(.27-12.19)	.62- 8.53	1.11- 6.42	1.79- 4.81	2.11- 4.19
	99	(.16-15.60)	.34-10.57	.78- 7.65	1.52- 5.44	1.88- 4.59
4	95	.49-13.72	1.10- 9.93	1.74- 7.73	2.53- 6.05	2.92- 5.36
	99	.21-17.21	.68-12.08	1.31- 9.05	2.17- 6.75	2.64- 5.82
5	95	(.88-15.14)	1.64-11.29	2.43- 9.00	3.26- 7.29	3.73- 6.54
	99	(.45-18.76)	1.10-13.53	1.89-10.40	2.83- 8.07	3.39- 7.05
6	95	1.26-16.57	2.24-12.60	3.18-10.21	4.11- 8.43	4.63- 7.64
	99	.69-20.32	1.56-14.93	2.57-11.66	3.63- 9.24	4.25- 8.18
7	95	(1.74-17.91)	2.86-13.90	3.88-11.47	4.96- 9.56	5.52- 8.73
	99	(1.04-21.72)	2.08-16.28	3.17-12.99	4.43-10.42	5.12- 9.31
8	95	2.23-19.25	3.51-15.16	4.70-12.61	5.81-10.70	6.42- 9.83
	99	1.38-23.13	2.63-17.61	3.93-14.18	5.23-11.60	5.98-10.43
9	95	(2.78-20.54)	4.20-16.40	5.46-13.82	6.66-11.83	7.32-10.93
	99	(1.80-24.46)	3.21-18.92	4.61-15.44	6.04-12.77	6.84-11.56
10	95	3.32-21.82	4.90-17.62	6.22-15.02	7.51-12.97	8.21-12.03
	99	2.22-25.80	3.82-20.20	5.29-16.70	6.84-13.95	7.70-12.69
11	95	(3.93-23.06)	5.65-18.80	7.05-16.16	8.41-14.06	9.14-13.10
	99	(2.70-27.11)	4.48-21.42	6.06-17.87	7.70-15.07	8.60-13.78
12	95	4.54-24.31	6.40-19.98	7.87-17.30	9.30-15.16	10.06-14.16
	99	3.18-28.42	5.15-22.65	6.83-19.05	8.56-16.19	9.51-14.86
13	95	(5.18-27.03)	7.11-21.20	8.70-18.44	10.20-16.25	10.99-15.23
	99	(3.72-29.67)	5.77-23.92	7.60-20.23	9.42-17.31	10.41-15.95
14	95	5.82-26.75	7.87-22.37	9.53-19.58	11.09-17.34	11.92-16.30
	99	4.25-30.92	6.46-25.13	8.38-21.40	10.28-18.43	11.31-17.04
15	95	(6.50-27.94)	8.64-23.53	10.36-20.72	11.98-18.44	12.84-17.37
	99	(4.82-32.14)	7.15-26.33	9.15-22.58	11.14-19.55	12.21-18.13

TABLA IX (continuación)

%	Coeficientes de confianza	n = 50	n = 100	n = 200	n = 500	n = 1000
16	95	7.17-29.12	9.45-24.66	11.22-21.82	12.90-19.50	13.79-18.42
	99	5.40-33.36	7.89-27.49	9.97-23.71	12.03-20.63	13.14-19.19
17	95	(7.88-30.28)	10.25-25.79	12.09-22.92	13.82-20.57	14.73-19.47
	99	(6.00-34.54)	8.63-28.65	10.79-24.84	12.92-21.72	14.07-20.25
18	95	8.58-31.44	11.06-26.92	12.96-24.02	14.74-21.64	15.67-20.52
	99	6.60-35.73	9.37-29.80	11.61-25.96	13.81-22.81	14.99-21.32
19	95	(9.31-32.58)	11.86-28.06	13.82-25.12	15.66-22.71	16.62-21.57
	99	(7.23-36.88)	10.10-30.96	12.43-27.09	14.71-23.90	15.92-22.38
20	95	10.04-33.72	12.66-29.19	14.69-26.22	16.58-23.78	17.56-22.62
	99	7.86-38.04	10.84-32.12	13.26-28.22	15.60-24.99	16.84-23.45
21	95	(10.79-34.84)	13.51-30.28	15.58-27.30	17.52-24.83	18.52-23.65
	99	(8.53-39.18)	11.63-33.24	14.11-29.31	16.51-26.05	17.78-24.50
22	95	11.54-35.95	14.35-31.37	16.48-28.37	18.45-25.88	19.47-24.69
	99	9.20-40.32	12.41-34.35	14.97-30.40	17.43-27.12	18.72-25.55
23	95	(12.30-37.06)	15.19-32.47	17.37-29.45	19.39-26.93	20.43-25.73
	99	(9.88-41.44)	13.60-34.82	15.83-31.50	18.34-28.18	19.67-26.59
24	95	13.07-38.17	16.03-33.56	18.27-30.52	20.33-27.99	21.39-26.77
	99	10.56-42.56	13.98-36.57	16.68-32.59	19.26-29.25	20.61-27.64
25	95	(13.84-39.27)	16.88-34.66	19.16-31.60	21.26-29.04	22.34-27.81
	99	(11.25-43.65)	14.77-37.69	17.54-33.68	20.17-30.31	21.55-28.69
26	95	14.63-40.34	17.75-35.72	20.08-32.65	22.21-30.08	23.31-28.83
	99	11.98-44.73	15.59-38.76	18.43-34.75	21.10-31.36	22.50-29.73
27	95	(15.45-41.40)	18.62-36.79	20.99-33.70	23.16-31.11	24.27-29.86
	99	(12.71-45.79)	16.42-39.84	19.31-35.81	22.04-32.41	23.46-30.76
28	95	16.23-42.48	19.50-37.85	21.91-34.76	24.11-32.15	25.24-30.89
	99	13.42-46.88	17.25-40.91	20.20-36.88	22.97-33.46	24.41-31.80
29	95	(17.06-43.54)	20.37-38.92	22.82-35.81	25.06-33.19	26.21-31.92
	99	(14.18-47.92)	18.07-41.99	21.08-37.94	23.90-34.51	25.37-32.84
30	95	17.87-44.61	21.24-39.98	23.74-36.87	26.01-34.23	27.17-32.95
	99	14.91-48.99	18.90-43.06	21.97-39.01	24.83-35.55	26.32-33.87

TABLA IX (continuación)

%	Coeficientes de confianza	n = 50	n = 100	n = 200	n = 500	n = 1000
31	95	(18.71-45.65)	22.14-41.02	24.67-37.90	26.97-35.25	28.15-33.97
	99	(15.68-50.02)	19.76-44.11	22.88-40.05	25.78-36.59	27.29-34.90
32	95	19.55-46.68	23.04-42.06	25.61-38.94	27.93-36.28	29.12-34.99
	99	16.46-51.05	20.61-45.15	23.79-41.09	26.73-37.62	28.25-35.92
33	95	(20.38-47.72)	23.93-43.10	26.54-39.97	28.90-37.31	30.09-36.01
	99	(17.23-52.08)	21.47-46.19	24.69-42.13	27.68-38.65	29.22-36.95
34	95	21.22-48.76	24.83-44.15	27.47-41.01	29.86-38.33	31.07-37.03
	99	18.01-53.11	22.33-47.24	25.60-43.18	28.62-39.69	30.18-37.97
35	95	(22.06-49.80)	25.73-45.19	28.41-42.04	30.82-39.36	32.04-38.05
	99	(18.78-54.14)	23.19-48.28	26.51-44.22	29.57-40.72	31.14-39.00
36	95	22.93-50.80	26.65-46.20	29.36-43.06	31.79-40.38	33.02-39.06
	99	19.60-55.13	24.08-49.30	27.44-45.24	30.53-41.74	32.12-40.02
37	95	(23.80-51.81)	27.57-47.22	30.31-44.08	32.76-41.39	34.00-40.07
	99	(20.42-56.12)	24.96-50.31	28.37-46.26	31.49-42.76	33.09-41.03
38	95	24.67-52.81	28.49-48.24	31.25-45.10	33.73-42.41	34.98-41.09
	99	21.23-57.10	25.85-51.32	29.30-47.29	32.45-43.78	34.07-42.05
39	95	(25.54-53.82)	29.41-49.26	32.20-46.12	34.70-43.43	35.97-42.10
	99	(22.05-58.09)	26.74-52.34	30.23-48.31	33.42-44.80	35.04-43.06
40	95	26.41-54.82	30.33-50.28	33.15-47.14	35.68-44.44	36.95-43.11
	99	22.87-59.08	27.63-53.35	31.16-49.33	34.38-45.82	36.02-44.08
41	95	(27.31-55.80)	31.27-51.28	34.12-48.15	36.66-45.45	37.93-44.12
	99	(23.72-60.04)	28.54-54.34	32.11-50.33	35.35-46.83	37.00-45.09
42	95	28.21-56.78	32.21-52.28	35.08-49.16	37.64-46.46	38.92-45.12
	99	24.57-60.99	29.45-55.33	33.06-51.33	36.32-47.83	37.98-46.10
43	95	(29.10-57.76)	33.15-53.27	36.05-50.16	38.62-47.46	39.91-46.13
	99	(25.42-61.95)	30.37-56.32	34.01-52.34	37.29-48.84	38.96-47.10
44	95	30.00-58.74	34.09-54.27	37.01-51.17	39.60-48.47	40.90-47.14
	99	26.27-62.90	31.28-57.31	34.95-53.34	38.27-49.85	39.95-48.11
45	95	(30.90-59.71)	35.03-55.27	37.97-52.17	40.58-49.48	41.89-48.14
	99	(27.12-63.86)	32.19-58.30	35.90-54.34	39.24-50.86	40.93-49.12
46	95	31.83-60.67	35.99-56.25	38.95-53.17	41.57-50.48	42.88-49.14
	99	28.00-64.78	33.13-59.26	36.87-55.33	40.22-51.85	41.92-50.12
47	95	(32.75-61.62)	36.95-57.23	39.93-54.16	42.56-51.48	43.87-50.14
	99	(28.89-65.69)	34.07-60.22	37.84-56.31	41.21-52.85	42.91-51.12
48	95	33.68-62.57	37.91-58.21	40.91-55.15	43.55-52.47	44.87-51.14
	99	29.78-66.61	35.01-61.19	38.80-57.30	42.19-53.85	43.90-52.12
49	95	(34.61-63.52)	38.87-59.19	41.89-56.14	44.54-53.47	45.86-52.14
	99	(30.67-67.53)	35.95-62.15	39.77-58.28	43.18-54.84	44.89-53.12
50	95	35.53-64.47	39.83-60.17	42.86-57.14	45.53-54.47	46.85-53.15
	99	31.55-68.45	36.89-63.11	40.74-59.26	44.16-55.84	45.89-54.11

TABLA X. La transformación z del coeficiente de correlación r.

r	z	r	z
.00	0.0000	.50	0.5493
.01	0.0100	.51	0.5627
.02	0.0200	.52	0.5763
.03	0.0300	.53	0.5901
.04	0.0400	.54	0.6042
.05	0.0500	.55	0.6184
.06	0.0601	.56	0.6328
.07	0.0701	.57	0.6475
.08	0.0802	.58	0.6625
.09	0.0902	.59	0.6777
.10	0.1003	.60	0.6931
.11	0.1104	.61	0.7089
.12	0.1206	.62	0.7250
.13	0.1307	.63	0.7414
.14	0.1409	.64	0.7582
.15	0.1511	.65	0.7753
.16	0.1614	.66	0.7928
.17	0.1717	.67	0.8107
.18	0.1820	.68	0.8291
.19	0.1923	.69	0.8480
.20	0.2027	.70	0.8673
.21	0.2132	.71	0.8872
.22	0.2237	.72	0.9076
.23	0.2342	.73	0.9287
.24	0.2448	.74	0.9505
.25	0.2554	.75	0.9730
.26	0.2661	.76	0.9962
.27	0.2769	.77	1.0203
.28	0.2877	.78	1.0454
.29	0.2986	.79	1.0714
.30	0.3095	.80	1.0986
.31	0.3205	.81	1.1270
.32	0.3316	.82	1.1568
.33	0.3428	.83	1.1881
.34	0.3541	.84	1.2212
.35	0.3654	.85	1.2562
.36	0.3769	.86	1.2933
.37	0.3884	.87	1.3331
.38	0.4001	.88	1.3758
.39	0.4118	.89	1.4219
.40	0.4236	.90	1.4722
.41	0.4356	.91	1.5275
.42	0.4477	.92	1.5890
.43	0.4599	.93	1.6584
.44	0.4722	.94	1.7380
.45	0.4847	.95	1.8318
.46	0.4973	.96	1.9459
.47	0.5101	.97	2.0923
.48	0.5230	.98	2.2976
.49	0.5361	.99	2.6467

TABLA XI. $f \ln f$ como función de f.

f	0	1	2	3	4	5	6	7	8	9	f
0	0.000	0.000	1.386	3.296	5.545	8.047	10.751	13.621	16.636	19.775	0
10	23.026	26.377	29.819	33.344	36.947	40.621	44.361	48.165	52.027	55.944	10
20	59.915	63.935	68.003	72.116	76.273	80.472	84.711	88.988	93.302	97.652	20
30	102.036	106.454	110.904	115.385	119.896	124.437	129.007	133.604	138.228	142.879	30
40	147.555	152.256	156.982	161.732	166.504	171.300	176.118	180.957	185.818	190.699	40
50	195.601	200.523	205.465	210.425	215.405	220.403	225.420	230.454	235.506	240.575	50
60	245.661	250.763	255.882	261.017	266.169	271.335	276.517	281.714	286.927	292.153	60
70	297.395	302.650	307.920	313.204	318.501	323.812	329.136	334.473	339.823	345.186	70
80	350.562	355.950	361.351	366.764	372.189	377.625	383.074	388.534	394.006	399.489	80
90	404.983	410.488	416.005	421.532	427.070	432.618	438.177	443.747	449.327	454.917	90
100	460.517	466.127	471.747	477.377	483.017	488.666	494.325	499.993	505.670	511.357	100
110	517.053	522.758	528.472	534.195	539.927	545.667	551.416	557.174	562.941	568.716	110
120	574.499	580.291	586.091	591.899	597.715	603.539	609.372	615.212	621.060	626.916	120
130	632.779	638.651	644.530	650.416	656.311	662.212	668.121	674.037	679.961	685.892	130
140	691.830	697.775	703.727	709.687	715.653	721.626	727.607	733.594	739.587	745.588	140
150	751.595	757.609	763.630	769.657	775.691	781.731	787.778	793.831	799.890	805.956	150
160	812.028	818.106	824.191	830.281	836.378	842.481	848.590	854.705	860.826	866.953	160
170	873.086	879.224	885.369	891.519	897.676	903.838	910.005	916.179	922.357	928.542	170
180	934.732	940.928	947.129	953.336	959.548	965.766	971.989	978.217	984.451	990.690	180
190	996.935	1003.184	1009.439	1015.699	1021.964	1028.235	1034.510	1040.791	1047.077	1053.368	190
200	1059.663	1065.964	1072.270	1078.581	1084.896	1091.217	1097.542	1103.873	1110.208	1116.548	200
210	1122.893	1129.242	1135.596	1141.955	1148.319	1154.687	1161.060	1167.438	1173.820	1180.207	210
220	1186.598	1192.994	1199.394	1205.799	1212.209	1218.623	1225.041	1231.464	1237.891	1244.322	220
230	1250.758	1257.198	1263.643	1270.092	1276.545	1283.003	1289.464	1295.930	1302.400	1308.875	230
240	1315.353	1321.636	1328.323	1334.814	1341.309	1347.808	1354.312	1360.819	1367.330	1373.846	240
250	1380.365	1386.889	1393.416	1399.948	1406.483	1413.022	1419.565	1426.113	1432.664	1439.218	250
260	1445.777	1452.340	1458.906	1465.477	1472.051	1478.628	1485.210	1491.795	1498.385	1504.977	260
270	1511.574	1518.174	1524.778	1531.386	1537.997	1544.612	1551.231	1557.853	1564.479	1571.108	270
280	1577.741	1584.378	1591.018	1597.661	1604.309	1610.959	1617.614	1624.271	1630.933	1637.597	280
290	1644.265	1650.937	1657.612	1664.291	1670.972	1677.658	1684.346	1691.038	1697.734	1704.433	290
300	1711.135	1717.840	1724.549	1731.261	1737.976	1744.695	1751.417	1758.142	1764.871	1771.602	300
310	1778.337	1785.076	1791.817	1798.562	1805.309	1812.060	1818.815	1825.572	1832.332	1839.096	310
320	1845.863	1852.633	1859.406	1866.182	1872.961	1879.743	1886.529	1893.317	1900.108	1906.903	320
330	1913.701	1920.501	1927.305	1934.111	1940.921	1947.734	1954.549	1961.368	1968.190	1975.014	330
340	1981.842	1988.672	1995.505	2002.342	2009.181	2016.023	2022.868	2029.716	2036.566	2043.420	340
350	2050.277	2057.136	2063.998	2070.863	2077.731	2084.602	2091.475	2098.352	2105.231	2112.113	350
360	2118.997	2125.885	2132.775	2139.668	2146.564	2153.463	2160.364	2167.268	2174.175	2181.084	360
370	2187.996	2194.911	2201.829	2208.749	2215.672	2222.597	2229.526	2236.456	2243.390	2250.326	370
380	2257.265	2264.207	2271.151	2278.097	2285.047	2291.999	2298.953	2305.910	2312.870	2319.832	380
390	2326.797	2333.765	2340.735	2347.707	2354.682	2361.660	2368.640	2375.623	2382.608	2389.596	390
400	2396.586	2403.579	2410.574	2417.571	2424.572	2431.574	2438.579	2445.587	2452.597	2459.609	400
410	2466.624	2473.642	2480.662	2487.684	2494.709	2501.736	2508.765	2515.797	2522.831	2529.868	410
420	2536.907	2543.948	2550.992	2558.038	2565.087	2572.138	2579.191	2586.247	2593.305	2600.365	420
430	2607.428	2614.493	2621.560	2628.629	2635.701	2642.776	2649.852	2656.931	2664.012	2671.095	430
440	2678.181	2685.269	2692.359	2699.451	2706.546	2713.643	2720.742	2727.844	2734.947	2742.053	440
450	2749.161	2756.272	2763.384	2770.499	2777.616	2784.735	2791.857	2798.980	2806.106	2813.234	450
460	2820.364	2827.496	2834.631	2841.768	2848.906	2856.047	2863.191	2870.336	2877.483	2884.633	460
470	2891.784	2898.935	2906.094	2913.252	2920.412	2927.575	2934.739	2941.905	2949.074	2956.245	470
480	2963.417	2970.592	2977.769	2984.948	2992.129	2999.312	3006.497	3013.685	3020.874	3028.065	480
490	3035.259	3042.454	3049.652	3056.851	3064.053	3071.256	3078.462	3085.669	3092.879	3100.090	490

Nota: Cuando $f > 999$, y puede descomponerse en factores $f = ab$, en que a y b son inferiores a 999, es útil la siguiente cantidad para calcular $f \ln f$: $ab \ln ab = a(b \ln b) + b(a \ln a)$.

TABLA XI (continuación)

	f	0	1	2	3	4
500-990	500	3107.304	3114.520	3121.737	3128.957	3136.178
	510	3179.549	3186.785	3194.022	3201.262	3208.503
	520	3251.991	3259.246	3266.502	3273.761	3281.022
	530	3324.625	3331.899	3339.174	3346.452	3353.731
	540	3397.447	3404.740	3412.034	3419.330	3426.628
	550	3470.455	3477.766	3485.079	3492.393	3499.709
	560	3543.645	3550.973	3558.304	3565.636	3572.971
	570	3617.013	3624.359	3631.708	3639.058	3646.409
	580	3690.556	3697.920	3705.286	3712.653	3720.022
	590	3764.272	3771.653	3779.036	3786.420	3793.806
	600	3838.158	3845.556	3852.955	3860.356	3867.759
	610	3912.210	3919.624	3927.040	3934.458	3941.877
	620	3986.426	3993.857	4001.289	4008.722	4016.158
	630	4060.803	4068.250	4075.698	4083.148	4090.599
	640	4135.340	4142.802	4150.266	4157.731	4165.198
	650	4210.032	4217.510	4224.989	4232.470	4239.952
	660	4284.878	4292.371	4299.866	4307.362	4314.859
	670	4359.876	4367.384	4374.894	4382.405	4389.917
	680	4435.023	4442.546	4450.070	4457.596	4465.123
	690	4510.317	4517.855	4525.393	4532.934	4540.476
	700	4585.756	4593.308	4600.861	4608.416	4615.972
	710	4661.328	4668.904	4676.471	4684.040	4691.610
	720	4737.061	4744.641	4752.222	4759.805	4767.389
	730	4812.923	4820.516	4828.111	4835.708	4843.306
	740	4888.921	4896.528	4904.137	4911.747	4919.359
	750	4965.055	4972.676	4980.298	4987.921	4995.546
	760	5041.322	5048.956	5056.591	5064.228	5071.866
	770	5117.721	5125.368	5133.016	5140.666	5148.317
	780	5194.249	5201.909	5209.570	5217.233	5224.897
	790	5270.906	5278.579	5286.253	5293.928	5301.604
	800	5347.689	5355.375	5363.061	5370.749	5378.438
	810	5424.598	5432.295	5439.994	5447.694	5455.396
	820	5501.630	5509.339	5517.051	5524.763	5532.477
	830	5578.783	5586.505	5594.229	5601.953	5609.679
	840	5656.058	5663.792	5671.527	5679.263	5687.001
	850	5733.451	5741.197	5748.944	5756.692	5764.441
	860	5810.962	5818.719	5826.478	5834.238	5841.999
	870	5888.589	5896.358	5904.128	5911.900	5919.672
	880	5966.331	5974.112	5981.893	5989.676	5997.460
	890	6044.187	6051.979	6059.772	6067.566	6075.361
	900	6122.155	6129.958	6137.762	6145.567	6153.374
	910	6200.235	6208.049	6215.864	6223.680	6231.497
	920	6278.424	6286.249	6294.075	6301.902	6309.730
	930	6356.722	6364.557	6372.394	6380.232	6388.071
	940	6435.127	6442.973	6450.821	6458.670	6466.519
	950	6513.639	6521.496	6529.354	6537.213	6545.073
	960	6592.256	6600.123	6607.992	6615.861	6623.732
	970	6670.977	6678.855	6686.734	6694.614	6702.495
	980	6749.802	6757.690	6765.579	6773.469	6781.360
	990	6828.728	6835.626	6844.525	6852.426	6860.327

5	6	7	8	9	f
3143.402	3150.628	3157.855	3165.085	3172.316	500
3215.746	3222.991	3230.238	3237.487	3244.738	510
3288.284	3295.548	3302.815	3310.083	3317.353	520
3361.013	3368.296	3375.581	3382.868	3390.157	530
3433.928	3441.230	3448.533	3455.839	3463.146	540
3507.027	3514.347	3521.669	3528.992	3536.318	550
3580.307	3587.644	3594.984	3602.325	3609.668	560
3653.763	3661.118	3668.475	3675.834	3683.194	570
3727.393	3734.765	3742.140	3749.515	3756.893	580
3801.194	3808.583	3815.975	3823.367	3830.762	590
3875.163	3882.569	3889.977	3897.386	3904.797	600
3949.298	3956.720	3964.144	3971.570	3978.997	610
4023.595	4031.033	4038.473	4045.915	4053.359	620
4098.052	4105.506	4112.962	4120.420	4127.879	630
4172.666	4180.136	4187.608	4195.081	4202.556	640
4247.436	4254.921	4262.408	4269.897	4277.387	650
4322.358	4329.859	4337.361	4344.864	4352.370	660
4397.431	4404.947	4412.463	4419.982	4427.502	670
4472.652	4480.182	4487.714	4495.247	4502.781	680
4548.019	4555.563	4563.109	4570.657	4578.206	690
4623.529	4631.088	4638.649	4646.210	4653.774	700
4699.182	4706.755	4714.329	4721.905	4729.482	710
4774.974	4782.561	4790.150	4797.739	4805.330	720
4850.905	4858.505	4866.107	4873.711	4881.315	730
4926.971	4934.585	4942.201	4949.817	4957.435	740
5003.172	5010.799	5018.428	5026.058	5033.689	750
5079.505	5087.146	5094.787	5102.431	5110.075	760
5155.969	5163.622	5171.277	5178.933	5186.591	770
5232.562	5240.228	5247.896	5255.564	5263.235	780
5309.282	5316.961	5324.641	5332.323	5340.005	790
5386.128	5393.819	5401.512	5409.206	5416.901	800
5463.098	5470.802	5478.507	5486.213	5493.921	810
5540.191	5547.907	5555.624	5563.343	5571.063	820
5617.405	5625.134	5632.863	5640.593	5648.325	830
5694.739	5702.479	5710.220	5717.963	5725.706	840
5772.192	5779.943	5787.696	5795.450	5803.206	850
5849.761	5857.524	5865.289	5873.054	5880.821	860
5927.446	5935.221	5942.997	5950.774	5958.552	870
6005.245	6013.031	6020.818	6028.607	6036.396	880
6083.157	6090.955	6098.753	6106.553	6114.353	890
6161.181	6168.990	6176.799	6184.610	6192.422	900
6239.316	6247.135	6254.956	6262.777	6270.600	910
6317.559	6325.390	6333.221	6341.053	6348.887	920
6395.911	6403.752	6411.594	6419.437	6427.282	930
6474.370	6482.221	6490.074	6497.928	6505.783	940
6552.934	6560.797	6568.660	6576.524	6584.390	950
6631.604	6639.476	6647.350	6655.225	6663.100	960
6710.377	6718.259	6726.143	6734.028	6741.914	970
6789.252	6797.145	6805.039	6812.935	6820.831	980
6868.229	6876.132	6884.037	6891.942	6899.848	990

TABLA XII. $(f+\frac{1}{2})\ln(f+\frac{1}{2})$ como función de f.

f	0	1	2	3	4
0	-0.347	0.608	2.291	4.385	6.768
10	24.689	28.087	31.572	35.136	38.775
20	61.919	65.963	70.054	74.190	78.367
30	104.241	108.675	113.140	117.637	122.163
40	149.903	154.616	159.354	164.115	168.899
50	198.060	202.991	207.943	212.913	217.902
60	248.210	253.321	258.448	263.591	268.750
70	300.021	305.283	310.560	315.850	321.155
80	353.255	358.649	364.056	369.475	374.906
90	407.734	413.245	418.767	424.299	429.843
100	463.321	468.936	474.561	480.196	485.840
110	519.904	525.614	531.332	537.060	542.796
120	577.394	583.190	588.994	594.806	600.626
130	635.714	641.589	647.472	653.363	659.260
140	694.802	700.750	706.706	712.669	718.639
150	754.601	760.619	766.643	772.673	778.710
160	815.066	821.148	827.235	833.329	839.429
170	876.154	882.296	888.444	894.597	900.756
180	937.829	944.028	950.232	956.441	962.656
190	1000.059	1006.311	1012.569	1018.831	1025.099
200	1062.813	1069.117	1075.425	1081.738	1088.056
210	1126.067	1132.419	1138.775	1145.136	1151.502
220	1189.795	1196.194	1202.596	1209.003	1215.415
230	1253.978	1260.420	1266.867	1273.318	1279.773
240	1318.594	1325.079	1331.568	1338.061	1344.558
250	1383.626	1390.152	1396.681	1403.215	1409.752
260	1449.058	1455.623	1462.191	1468.763	1475.339
270	1514.874	1521.476	1528.082	1534.691	1541.304
280	1581.059	1587.697	1594.339	1600.985	1607.634
290	1647.601	1654.274	1660.951	1667.631	1674.315
300	1714.487	1721.194	1727.905	1734.618	1741.335
310	1781.706	1788.446	1795.189	1801.935	1808.684
320	1849.247	1856.019	1862.793	1869.571	1876.352
330	1917.101	1923.903	1930.708	1937.516	1944.327
340	1985.256	1992.088	1998.923	2005.761	2012.601
350	2053.706	2060.567	2067.430	2074.297	2081.166
360	2122.441	2129.330	2136.221	2143.116	2150.013
370	2191.453	2198.369	2205.288	2212.210	2219.134
380	2260.735	2267.678	2274.624	2281.572	2288.522
390	2330.281	2337.249	2344.221	2351.194	2358.171
400	2400.082	2407.076	2414.072	2421.071	2428.073
410	2470.133	2477.151	2484.172	2491.196	2498.222
420	2540.427	2547.470	2554.515	2561.562	2568.612
430	2610.960	2618.026	2625.094	2632.165	2639.238
440	2681.725	2688.814	2695.905	2702.998	2710.094
450	2752.716	2759.828	2766.941	2774.057	2781.175
460	2823.930	2831.063	2838.199	2845.337	2852.477
470	2895.361	2902.516	2909.673	2916.832	2923.993
480	2967.004	2974.180	2981.358	2988.538	2995.720
490	3038.856	3046.053	3053.251	3060.452	3067.654

5	6	7	8	9	f
9.376	12.167	15.112	18.191	21.387	0
42.483	46.255	50.089	53.979	57.923	10
82.586	86.844	91.140	95.472	99.840	20
126.718	131.302	135.913	140.550	145.214	30
173.706	178.535	183.385	188.256	193.148	40
222.909	227.935	232.978	238.038	243.116	50
273.924	279.114	284.319	289.538	294.772	60
326.472	331.803	337.147	342.503	347.873	70
380.348	385.802	391.268	396.746	402.234	80
435.397	440.961	446.536	452.121	457.716	90
491.494	497.157	502.830	508.512	514.204	100
548.541	554.294	560.057	565.827	571.606	110
606.454	612.291	618.135	623.987	629.847	120
665.166	671.078	676.998	682.926	688.860	130
724.616	730.599	736.590	742.587	748.591	140
784.753	790.803	796.860	802.922	808.991	150
845.535	851.647	857.765	863.889	870.019	160
906.921	913.091	919.267	925.449	931.636	170
968.877	975.102	981.334	987.570	993.812	180
1031.372	1037.650	1043.933	1050.222	1056.515	190
1094.379	1100.707	1107.040	1113.377	1119.720	200
1157.873	1164.248	1170.628	1177.013	1183.402	210
1221.831	1228.252	1234.677	1241.106	1247.540	220
1286.233	1292.697	1299.165	1305.637	1312.114	230
1351.059	1357.565	1364.074	1370.588	1377.105	240
1416.293	1422.839	1429.388	1435.941	1442.497	250
1481.919	1488.502	1495.089	1501.680	1508.275	260
1547.921	1554.541	1561.165	1567.793	1574.424	270
1614.286	1620.942	1627.602	1634.265	1640.931	280
1681.002	1687.692	1694.386	1701.083	1707.783	290
1748.056	1754.779	1761.506	1768.236	1774.970	300
1815.437	1822.193	1828.952	1835.714	1842.479	310
1883.135	1889.922	1896.712	1903.505	1910.301	320
1951.141	1957.958	1964.778	1971.601	1978.427	330
2019.445	2026.291	2033.141	2039.993	2046.848	340
2088.038	2094.913	2101.791	2108.671	2115.555	350
2156.913	2163.815	2170.721	2177.629	2184.540	360
2226.061	2232.991	2239.923	2246.858	2253.795	370
2295.476	2302.431	2309.390	2316.351	2323.314	380
2365.150	2372.131	2379.115	2386.101	2393.090	390
2435.077	2442.083	2449.092	2456.103	2463.117	400
2505.250	2512.281	2519.314	2526.349	2533.387	410
2575.664	2582.719	2589.775	2596.835	2603.896	420
2646.313	2653.391	2660.471	2667.553	2674.638	430
2717.192	2724.293	2731.395	2738.500	2745.607	440
2788.296	2795.418	2802.543	2809.670	2816.799	450
2859.619	2866.763	2873.909	2881.058	2888.208	460
2931.156	2938.322	2945.489	2952.659	2959.831	470
3002.905	3010.091	3017.279	3024.469	3031.662	480
3074.859	3082.065	3089.274	3096.484	3103.697	490

TABLA XIII. Valores críticos de U, el estadístico de Mann-Whitney.

n_1	n_2 \ α	0.10	0.05	0.025	0.01	0.005	0.001
3	2	6					
	3	8	9				
4	2	8					
	3	11	12				
	4	13	15	16			
5	2	9	10				
	3	13	14	15			
	4	16	18	19	20		
	5	20	21	23	24	25	
6	2	11	12				
	3	15	16	17			
	4	19	21	22	23	24	
	5	23	25	27	28	29	
	6	27	29	31	33	34	
7	2	13	14				
	3	17	19	20	21		
	4	22	24	25	27	28	
	5	27	29	30	32	34	
	6	31	34	36	38	39	42
	7	36	38	41	43	45	48
8	2	14	15	16			
	3	19	21	22	24		
	4	25	27	28	30	31	
	5	30	32	34	36	38	40
	6	35	38	40	42	44	47
	7	40	43	46	49	50	54
	8	45	49	51	55	57	60
9	1	9					
	2	16	17	18			
	3	22	23	25	26	27	
	4	27	30	32	33	35	
	5	33	36	38	40	42	44
	6	39	42	44	47	49	52
	7	45	48	51	54	56	60
	8	50	54	57	61	63	67
	9	56	60	64	67	70	74
10	1	10					
	2	17	19	20			
	3	24	26	27	29	30	
	4	30	33	35	37	38	40
	5	37	39	42	44	46	49
	6	43	46	49	52	54	57
	7	49	53	56	59	61	65
	8	56	60	63	67	69	74
	9	62	66	70	74	77	82
	10	68	73	77	81	84	90

Nota: Los valores críticos están tabulados para dos muestras de tamaños n_1 y n_2, donde $n_1 \geqslant n_2$ hasta $n_1 = n_2 = 20$. Los limites superiores de los valores críticos se dan de modo que el estadístico de muestreo U_s tenga que ser mayor que un valor crítico determinado. Las probabilidades que encabezan las columnas están basadas en una prueba de una cola y representan la proporción del área de la distribución de U en una cola más allá del valor crítico. Para una prueba de dos colas se utilizan los mismos valores críticos pero se multiplica por dos la probabilidad que encabeza las columnas. Esta tabla se ha extraído de otra más extensa (tabla 11.4) de D. B. Owen, *Handbook of Statistical Tables* (Addison-Wesley Publishing Co., Reading, Mass., 1962). Cortesía de U.S. Atomic Energy Comission, con permiso de los autores.

TABLA XIII (continuación)

n_1	n_2 \ α	0.10	0.05	0.025	0.01	0.005	0.001
11	1	11					
	2	19	21	22			
	3	26	28	30	32	33	
	4	33	36	38	40	42	44
	5	40	43	46	48	50	53
	6	47	50	53	57	59	62
	7	54	58	61	65	67	71
	8	61	65	69	73	75	80
	9	68	72	76	81	83	89
	10	74	79	84	88	92	98
	11	81	87	91	96	100	106
12	1	12					
	2	20	22	23			
	3	28	31	32	34	35	
	4	36	39	41	42	45	48
	5	43	47	49	52	54	58
	6	51	55	58	61	63	68
	7	58	63	66	70	72	77
	8	66	70	74	79	81	87
	9	73	78	82	87	90	96
	10	81	86	91	96	99	106
	11	88	94	99	104	108	115
	12	95	102	107	113	117	124
13	1	13					
	2	22	24	25	26		
	3	30	33	35	37	38	
	4	39	42	44	47	49	51
	5	47	50	53	56	58	62
	6	55	59	62	66	68	73
	7	63	67	71	75	78	83
	8	71	76	80	84	87	93
	9	79	84	89	94	97	103
	10	87	93	97	103	106	113
	11	95	101	106	112	116	123
	12	103	109	115	121	125	133
	13	111	118	124	130	135	143
14	1	14					
	2	24	25	27	28		
	3	32	35	37	40	41	
	4	41	45	47	50	52	55
	5	50	54	57	60	63	67
	6	59	63	67	71	73	78
	7	67	72	76	81	83	89
	8	76	81	86	90	94	100
	9	85	90	95	100	104	111
	10	93	99	104	110	114	121
	11	102	108	114	120	124	132
	12	110	117	123	130	134	143
	13	119	126	132	139	144	153
	14	127	135	141	149	154	164

TABLA XIII (continuación)

n_1	n_2 \ α	0.10	0.05	0.025	0.01	0.005	0.001
15	1	15					
	2	25	27	29	30		
	3	35	38	40	42	43	
	4	44	48	50	53	55	59
	5	53	57	61	64	67	71
	6	63	67	71	75	78	83
	7	72	77	81	86	89	95
	8	81	87	91	96	100	106
	9	90	96	101	107	111	118
	10	99	106	111	117	121	129
	11	108	115	121	128	132	141
	12	117	125	131	138	143	152
	13	127	134	141	148	153	163
	14	136	144	151	159	164	174
	15	145	153	161	169	174	185
16	1	16					
	2	27	29	31	32		
	3	37	40	42	45	46	
	4	47	50	53	57	59	62
	5	57	61	65	68	71	75
	6	67	71	75	80	83	88
	7	76	82	86	91	94	101
	8	86	92	97	102	106	113
	9	96	102	107	113	117	125
	10	106	112	118	124	129	137
	11	115	122	129	135	140	149
	12	125	132	139	146	151	161
	13	134	143	149	157	163	173
	14	144	153	160	168	174	185
	15	154	163	170	179	185	197
	16	163	173	181	190	196	208
17	1	17					
	2	28	31	32	34		
	3	39	42	45	47	49	51
	4	50	53	57	60	62	66
	5	60	65	68	72	75	80
	6	71	76	80	84	87	93
	7	81	86	91	96	100	106
	8	91	97	102	108	112	119
	9	101	108	114	120	124	132
	10	112	119	125	132	136	145
	11	122	130	136	143	148	158
	12	132	140	147	155	160	170
	13	142	151	158	166	172	183
	14	153	161	169	178	184	195
	15	163	172	180	189	195	208
	16	173	183	191	201	207	220
	17	183	193	202	212	219	232

TABLA XIII (continuación)

n_1	n_2 \ α	0.10	0.05	0.025	0.01	0.005	0.001
18	1	18					
	2	30	32	34	36		
	3	41	45	47	50	52	54
	4	52	56	60	63	66	69
	5	63	68	72	76	79	84
	6	74	80	84	89	92	98
	7	85	91	96	102	105	112
	8	96	103	108	114	118	126
	9	107	114	120	126	131	139
	10	118	125	132	139	143	153
	11	129	137	143	151	156	166
	12	139	148	155	163	169	179
	13	150	159	167	175	181	192
	14	161	170	178	187	194	206
	15	172	182	190	200	206	219
	16	182	193	202	212	218	232
	17	193	204	213	224	231	245
	18	204	215	225	236	243	258
19	1	18	19				
	2	31	34	36	37	38	
	3	43	47	50	53	54	57
	4	55	59	63	67	69	73
	5	67	72	76	80	83	88
	6	78	84	89	94	97	103
	7	90	96	101	107	111	118
	8	101	108	114	120	124	132
	9	113	120	126	133	138	146
	10	124	132	138	146	151	161
	11	136	144	151	159	164	175
	12	147	156	163	172	177	188
	13	158	167	175	184	190	202
	14	169	179	188	197	203	216
	15	181	191	200	210	216	230
	16	192	203	212	222	230	244
	17	203	214	224	235	242	257
	18	214	226	236	248	255	271
	19	226	238	248	260	268	284
20	1	19	20				
	2	33	36	38	39	40	
	3	45	49	52	55	57	60
	4	58	62	66	70	72	77
	5	70	75	80	84	87	93
	6	82	88	93	98	102	108
	7	94	101	106	112	116	124
	8	106	113	119	126	130	139
	9	118	126	132	140	144	154
	10	130	138	145	153	158	168
	11	142	151	158	167	172	183
	12	154	163	171	180	186	198
	13	166	176	184	193	200	212
	14	178	188	197	207	213	226
	15	190	200	210	220	227	241
	16	201	213	222	233	241	255
	17	213	225	235	247	254	270
	18	225	237	248	260	268	284
	19	237	250	261	273	281	298
	20	249	262	273	286	295	312

TABLA XIV. Valores críticos de la suma del rango de Wilcoxon.

n \ nominal α	0.05		0.025		0.01		0.005	
	T	α	T	α	T	α	T	α
5	0	.0312						
	1	.0625						
6	2	.0469	0	.0156				
	3	.0781	1	.0312				
7	3	.0391	2	.0234	0	.0078		
	4	.0547	3	.0391	1	.0156		
8	5	.0391	3	.0195	1	.0078	0	.0039
	6	.0547	4	.0273	2	.0117	1	.0078
9	8	.0488	5	.0195	3	.0098	1	.0039
	9	.0645	6	.0273	4	.0137	2	.0059
10	10	.0420	8	.0244	5	.0098	3	.0049
	11	.0527	9	.0322	6	.0137	4	.0068
11	13	.0415	10	.0210	7	.0093	5	.0049
	14	.0508	11	.0269	8	.0122	6	.0068
12	17	.0461	13	.0212	9	.0081	7	.0046
	18	.0549	14	.0261	10	.0105	8	.0061
13	21	.0471	17	.0239	12	.0085	9	.0040
	22	.0549	18	.0287	13	.0107	10	.0052
14	25	.0453	21	.0247	15	.0083	12	.0043
	26	.0520	22	.0290	16	.0101	13	.0054
15	30	.0473	25	.0240	19	.0090	15	.0042
	31	.0535	26	.0277	20	.0108	16	.0051
16	35	.0467	29	.0222	23	.0091	19	.0046
	36	.0523	30	.0253	24	.0107	20	.0055
17	41	.0492	34	.0224	27	.0087	23	.0047
	42	.0544	35	.0253	28	.0101	24	.0055
18	47	.0494	40	.0241	32	.0091	27	.0045
	48	.0542	41	.0269	33	.0104	28	.0052
19	53	.0478	46	.0247	37	.0090	32	.0047
	54	.0521	47	.0273	38	.0102	33	.0054
20	60	.0487	52	.0242	43	.0096	37	.0047
	61	.0527	53	.0266	44	.0107	38	.0053

Nota: Esta tabla da los valores críticos para la prueba de significación (de una cola) de la suma del rango T, obtenido a partir de la prueba de los rangos con signo apareados de Wilcoxon. Como el nivel exacto de probabilidad deseado no puede obtenerse con valores críticos enteros de T, se dan dos de estos valores y sus consiguientes probabilidades poniendo entre corchetes el nivel de significación deseado. Así, para hallar los valores significativos al 1 % para $n = 19$ observamos los dos valores

TABLA XIV continuación

n \ nominal α	0.05		0.025		0.01		0.005	
	T	α	T	α	T	α	T	α
21	67	.0479	58	.0230	49	.0097	42	.0045
	68	.0516	59	.0251	50	.0108	43	.0051
22	75	.0492	65	.0231	55	.0095	48	.0046
	76	.0527	66	.0250	56	.0104	49	.0052
23	83	.0490	73	.0242	62	.0098	54	.0046
	84	.0523	74	.0261	63	.0107	55	.0051
24	91	.0475	81	.0245	69	.0097	61	.0048
	92	.0505	82	.0263	70	.0106	62	.0053
25	100	.0479	89	.0241	76	.0094	68	.0048
	101	.0507	90	.0258	77	.0101	69	.0053
26	110	.0497	98	.0247	84	.0095	75	.0047
	111	.0524	99	.0263	85	.0102	76	.0051
27	119	.0477	107	.0246	92	.0093	83	.0048
	120	.0502	108	.0260	93	.0100	84	.0052
28	130	.0496	116	.0239	101	.0096	91	.0048
	131	.0521	117	.0252	102	.0102	92	.0051
29	140	.0482	126	.0240	110	.0095	100	.0049
	141	.0504	127	.0253	111	.0101	101	.0053
30	151	.0481	137	.0249	120	.0098	109	.0050
	152	.0502	138	.0261	121	.0104	110	.0053
31	163	.0491	147	.0239	130	.0099	118	.0049
	164	.0512	148	.0251	131	.0105	119	.0052
32	175	.0492	159	.0249	140	.0097	128	.0050
	176	.0512	160	.0260	141	.0103	129	.0053
33	187	.0485	170	.0242	151	.0099	138	.0049
	188	.0503	171	.0253	152	.0104	139	.0052
34	200	.0488	182	.0242	162	.0098	148	.0048
	201	.0506	183	.0252	163	.0103	149	.0051
35	213	.0484	195	.0247	173	.0096	159	.0048
	214	.0501	196	.0257	174	.0100	160	.0051

críticos de T, 37 y 38, en la tabla. Las probabilidades correspondientes a estos dos valores de T son 0,0090 y 0,0102. Sin duda una suma de rangos de $T_+ = 37$ tendría una probabilidad inferior a 0,01 y se consideraría significativa por el criterio establecido. Para pruebas de dos colas en que la hipótesis alternativa es que las parejas podrían diferir en una u otra dirección, se doblan las probabilidades que encabezan la tabla. Para tamaños de muestreo $n > 50$ se calcula

$$t_{\alpha[\infty]} = \left[T_+ - \frac{n(n+1)}{4}\right] \Big/ \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

TABLA XIV continuación

n \ nominal α	0.05		0.025		0.01		0.005	
	T	α	T	α	T	α	T	α
36	227	.0489	208	.0248	185	.0096	171	.0050
	228	.0505	209	.0258	186	.0100	172	.0052
37	241	.0487	221	.0245	198	.0099	182	.0048
	242	.0503	222	.0254	199	.0103	183	.0050
38	256	.0493	235	.0247	211	.0099	194	.0048
	257	.0509	236	.0256	212	.0104	195	.0050
39	271	.0493	249	.0246	224	.0099	207	.0049
	272	.0507	250	.0254	225	.0103	208	.0051
40	286	.0486	264	.0249	238	.0100	220	.0049
	287	.0500	265	.0257	239	.0104	221	.0051
41	302	.0488	279	.0248	252	.0100	233	.0048
	303	.0501	280	.0256	253	.0103	234	.0050
42	319	.0496	294	.0245	266	.0098	247	.0049
	320	.0509	295	.0252	267	.0102	248	.0051
43	336	.0498	310	.0245	281	.0098	261	.0048
	337	.0511	311	.0252	282	.0102	262	.0050
44	353	.0495	327	.0250	296	.0097	276	.0049
	354	.0507	328	.0257	297	.0101	277	.0051
45	371	.0498	343	.0244	312	.0098	291	.0049
	372	.0510	344	.0251	313	.0101	292	.0051
46	389	.0497	361	.0249	328	.0098	307	.0050
	390	.0508	362	.0256	329	.0101	308	.0052
47	407	.0490	378	.0245	345	.0099	322	.0048
	408	.0501	379	.0251	346	.0102	323	.0050
48	426	.0490	396	.0244	362	.0099	339	.0050
	427	.0500	397	.0251	363	.0102	340	.0051
49	446	.0495	415	.0247	379	.0098	355	.0049
	447	.0505	416	.0253	380	.0100	356	.0050
50	466	.0495	434	.0247	397	.0098	373	.0050
	467	.0506	435	.0253	398	.0101	374	.0051

Bibliografía

Allee, W. C., y E. Bowen. 1932. Studies in animal aggregations: Mass protection against colloidal silver among goldfishes. *J. Exp. Zool.*, **61**:185–207.

Archibald, E. E. A. 1950. Plant populations. II. The estimation of the number of individuals per unit area of species in heterogeneous plant populations. *Ann. Bot. N.S.*, **14**:7–21.

Banta, A. M. 1939. Studies on the physiology, genetics, and evolution of some Cladocera. Carnegie Institution of Washington, Dept. Genetics, Paper 39. 285 pp.

Blakeslee, A. F. 1921. The globe mutant in the Jimson Weed (*Datura stramonium*). *Genetics*, **6**:241–264.

Block, B. C. 1966. The relation of temperature to the chirp-rate of male snowy tree crickets, *Oecanthus fultoni* (Orthoptera: Gryllidae). *Ann. Entomol. Soc. Amer.*, **59**:56–59.

Brower, L. P. 1959. Speciation in butterflies of the *Papilio glaucus* group. I. Morphological relationships and hybridization. *Evolution*, **13**:40–63.

Brown, B. E., y A. W. A. Brown. 1956. The effects of insecticidal poisoning on the level of cytochrome oxidase in the American cockroach. *J. Econ. Entomol.*, **49**:675–679.

Burr, E. J. 1960. The distribution of Kendall's score S for a pair of tied rankings. *Biometrika*, **47**:151–171.

Carter, G. R., y C. A. Mitchell. 1958. Methods for adapting the virus of Rinderpest to rabbits. *Science*, **128**:252–253.

Cowan, I. M., y P. A. Johnston. 1962. Blood serum protein variations at the species and subspecies level in deer of the genus *Odocoilcus*. *Syst. Zool.*, **11**:131–138.

Davis, E. A., Jr. 1955. Seasonal changes in the energy balance of the English sparrow. *Auk*, **72**:385–411.

Fröhlich, F. W. 1921. *Grundzüge einer Lehre vom Licht- und Farbensinn. Ein Beitrag zur allgemeinen Physiologie der Sinne.* Fischer, Jena. 86 pp.

Gabriel, K. R. 1964. A procedure for testing the homogeneity of all sets of means in analysis of variance. *Biometrics*, **20**:459–477.

Gartler, S. M., I. L. Firschein, and T. Dobzhansky. 1956. Chromatographic investigation of urinary amino-acids in the great apes. *Am. J. Phys. Anthropol.*, **14**:41–57.

Geissler, A. 1889. Beiträge zur Frage des Geschlechtsverhältnisses der Geborenen. *Z. K. Sächs. Stat. Bur.*, **35**:1–24.

Hunter, P. E. 1959. Selection of *Drosophila melanogaster* for length of larval period. *Z. Vererbungsl.*, **90**:7–28.

Johnson, N. K. 1966. Bill size and the question of competition in allopatric and sympatic populations of Dusky and Gray Flycatchers. *Syst. Zool.*, **15**:70–87.

Kouskolekas, C. A., y G. C. Decker. 1966. The effect of temperature on the rate of development of the potato leafhopper, *Empoasca fabae* (Homoptera: Cicadellidae). *Ann. Entomol. Soc. Amer.*, **59**:292–298.

Lewontin, R. C., y M. J. D. White. 1960. Interaction between inversion polymorphisms of two chromosome pairs in the grasshopper, *Moraba scurra*. *Evolution*, **14**:116–129.

Littlejohn, M. J. 1965. Premating isolation in the *Hyla ewingi* complex. *Evolution*, **19**:234–243.

Millis, J., y Y. P. Seng. 1954. The effect of age and parity of the mother on birth weight of the offspring. *Ann. Human Genetics*, **19**:58–73.

Mittler, T. E., y R. H. Dadd. 1966. Food and wing determination in *Myzus persicae* (Homoptera: Aphidae). *Ann. Entomol. Soc. Amer.*, **59**:1162–1166.

Mosimann, J. E. 1968. *Elementary Probability for the Biological Sciences*. Appleton-Century-Crofts, New York. 255 pp.

Nelson, V. E. 1964. The effects of starvati on and humidity on water content in *Tribolium confusum* Duval (Coleoptera). Unpublished Ph.D. thesis, University of Colorado. 111 pp.

Newman, K. J., y H. V. Meredith. 1956. Individual growth in skeletal bigonial diameter during the childhood period from 5 to 11 years of age. *Am. J. Anatomy*, **99**:157–187.

Olson, E. C., y R. L. Miller. 1958. *Morphological Integration*. University of Chicago Press, Chicago. 317 pp.

Park, W. H., A. W. Williams, y C. Krumwiede. 1924. *Pathogenic Microöganisms*. Lea & Febiger, Philadelphia and New York. 811 pp.

Phillips, J. R., y L. D. Newsom. 1966. Diapause in *Heliothis zea* and *Heliothis virescens* (Lepidoptera: Noctuidae). *Ann. Entomol. Soc. Amer.*, **59**:154–159.

Siegel, S. 1956. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, Toronto, y London. 312 pp.

Sinnott, E. W. y D. Hammond. 1935. Factorial balance in the determination of fruit shape in *Cucurbita*. *Amer. Nat.*, **64**:509–524.

Smith, F. L. 1939. A genetic analysis of red seed-coat color in *Phaseolus vulgaris*. *Hilgardia*, **12**:553–621.

Sokal, R. R. 1952. Variation in a local population of *Pemphigus*. *Evolution*, **6**:296–315.

Sokal, R. R. 1967. A comparison of fitness characters and their responses to density in stock and selected cultures of wild type and black *Tribolium castaneum*. *Tribolium Inf. Bull.*, **10**:142–147.

Sokal, R. R., y P. E. Hunter. 1955. A morphometric analysis of DDT-resistant and non-resistant housefly strains. *Ann. Entomol. Soc. Amer.* **48**:499–507.

Sokal, R. R., y I. Karten. 1964. Competition among genotypes in *Tribolium castaneum* at varying densities and gene frequencies (the black locus). *Genetics*, **49**:195–211.

Sokal, R. R., y F. J. Rohlf. 1969. *Biometry*. W. H. Freeman and Company, San Francisco. 776 pp.

Sokal, R. R., y P. A. Thomas. 1965. Geographic variation of *Pemphigus populitransversus* in Eastern North America: Stem mothers and new data on alates. *Univ. Kansas Sci. Bull.*, **46**:201–252.

Sokoloff, A. 1955. Competition between sibling species of the *Pseudoobscura* subgroup of *Drosophila*. *Ecol. Monogr.*, **25**:387–409.

Sokoloff, A. 1966. Morphological variation in natural and experimental populations of *Drosophila pseudoobscura* and *Drosophila persimilis*. *Evolution*, **20**:49–71.

Student (W. S. Gossett). 1907. On the error of counting with a haemacytometer. *Biometrika*, **5**:351–360.

Swanson, C. O., W. L. Latshaw, y E. L. Tague. 1921. Relation of the calcium content of some Kansas soils to soil reaction by the electrometric titration. *J. Agr. Res.*, **20**:855–868.

Tate, R. F., y G. W. Klett. 1959. Optimal confidence intervals for the variance of a normal distribution. *J. Am. Stat. Assoc.*, **54**:674–682.

Utida, S. 1943. Studies on experimental population of the Azuki bean weevil, *Callosobruchus chinensis* (L.). VIII. Statistical analysis of the frequency distribution of the emerging weevils on beans. *Mem. Coll. Agr. Kyoto Imp. Univ.*, **54**:1–22.

Vollenweider, R. A., y M. Frei. 1953. Vertikale und zeitliche Verteilung der Leitfähigkeit in einem eutrophen Gewässer während der Sommerstagnation. *Schweiz. Z. Hydrol.* **15**:158–167.

Whittaker, R. H. 1952. A study of summerfoliage insect communities in the Great Smoky Mountains. *Ecol. Monogr.*, **22**:1–44.

Willis, E. R., y N. Lewis. 1957. The longevity of starved cockroaches. *J. Econ. Entomol.*, **50**:438–440.

Woodson, R. E., Jr. 1964. The geography of flower color in butterflyweed. *Evolution*, **18**:143–163.

Índice alfabético

Programas Educativos, S.A. de C.V.
Calz. de Chabacano No. 65, Local A
Col. Asturias, C.P. 06850, México, D.F.
Fecha: Octubre de 1999
Empresa Certificada por el
Instituto Mexicano de Normalización
y Certificación A.C., bajo la Norma
ISO-9002: 1994/NMX-CC-004: 1995
con el No. de Registro RSC-048