

---

# BIOESTADISTICA: PRINCIPIOS Y PROCEDIMIENTOS

---

segunda edición  
(primera en español)

Robert G. D. Steel  
Profesor de Estadística  
*North Carolina State University*

James H. Torrie  
Profesor Emérito de Agronomía  
*University of Wisconsin*

Traducción  
Ricardo Martínez B.  
Profesor Asociado  
*Universidad Nacional de Colombia*

Revisión  
Jesús María Castaño  
Profesor de Matemáticas

<http://arveja.awardspace.com>

McGRAW-HILL

Bogotá, Buenos Aires, Guatemala, Lisboa, Madrid, México, Nueva York,  
Panamá, San Juan, Santiago, São Paulo

Auckland, Hamburgo, Johannesburgo, Londres, Montreal, Nueva Delhi, París,  
San Francisco, San Luis, Sidney, Singapur, Tokio, Toronto

**RESERVADOS TODOS LOS DERECHOS (D.R.)**  
Copyright © 1985, por **EDITORIAL McGRAW-HILL**  
**LATINOAMERICANA, S. A.**  
**Apartado 81078**  
**Bogotá - Colombia**

Ni este libro ni parte de él pueden ser reproducidos o transmitidos de alguna forma o por algún medio electrónico o mecánico, incluyendo fotocopia o grabación o por cualquier otro sistema de memoria o archivo, sin el permiso escrito del editor.

Traducido de la segunda edición de  
**PRINCIPLES AND PROCEDURES OF STATISTICS**  
**A Biometrical Approach**  
Copyright © 1980 por McGraw-Hill Inc.  
ISBN 0-07-060926-8

0987654321

8976432105

ISBN 968-451-495-6

*Impreso en Colombia*

*Printed in Colombia*

Esta obra se terminó en septiembre de 1985  
en Editorial Presencia Ltda  
Calle 23 No. 24-20

A la memoria de James H. Torrie, quien no pudo participar en la redacción de esta segunda edición. Los dos años anteriores a su muerte, el 30 de mayo de 1976, fueron para él de precaria salud. Era entonces Profesor Emérito de Agronomía de la Universidad de Wisconsin en Madison. Jim tuvo una larga y productiva carrera, fue un profesor amistoso y paciente y una persona tranquila, nada presuntuoso.

---

# CONTENIDO

---

## Prefacio

## Símbolos escogidos

<b>Capítulo 1</b>	<b>Introducción</b>	<b>1</b>
1.1	Definición de la estadística	1
1.2	Breve historia de la estadística	2
1.3	La estadística y el método científico	4
1.4	El estudio de la estadística	5
<b>Capítulo 2</b>	<b>Observaciones</b>	<b>7</b>
2.1	Introducción	7
2.2	Variables	1
2.3	Distribuciones	8
2.4	Poblaciones y muestras	9
2.5	Muestras aleatorias: Recolección de datos	10
2.6	Presentación, resumen y caracterización de la información	12
2.7	Medidas de tendencia central	15
2.8	Medidas de dispersión	19
2.9	Desviación estándar de las medias	24
2.10	Coeficiente de variabilidad o de variación	25
2.11	Ejemplo	26
2.12	Modelo lineal aditivo	27
2.13	Ejemplo	28
2.14	El uso de codificación en el cálculo de estadígrafos	30
2.15	La tabla de frecuencia	32
2.16	Ejemplo	33

2.17	Cálculo de la media y la desviación estándar con una tabla de frecuencia	34
2.18	Presentación gráfica de la tabla de frecuencia	35
2.19	Dígitos significativos	35
<b>Capítulo 3</b>	<b>Probabilidad</b>	<b>37</b>
3.1	Introducción	37
3.2	Algunos elementos de probabilidad	37
3.3	La distribución binomial	40
3.4	Funciones de probabilidad para variables continuas	44
3.5	La distribución normal	46
3.6	Probabilidades de una distribución normal. Uso de una tabla de probabilidades	48
3.7	La distribución normal con media $\mu$ y varianza $\sigma^2$	53
3.8	Distribución de medias	55
3.9	Distribución $\chi^2$	56
3.10	Distribución $t$ de Student	58
3.11	Estimación e inferencia	59
3.12	Predicción de resultados de muestras	63
<b>Capítulo 4</b>	<b>Muestreo de una población normal</b>	<b>65</b>
4.1	Introducción	65
4.2	Una población con distribución normal	65
4.3	Muestras aleatorias de una distribución normal	68
4.4	Distribución de medias muestrales	68
4.5	Distribución de varianzas y muestrales y desviaciones estándar	71
4.6	Insesgamiento de $s^2$	72
4.7	Desviación estándar de la media o error estándar	72
4.8	La distribución $t$ de Student	73
4.9	El enunciado de confianza	75
4.10	Muestreo de diferencias	76
4.11	Resumen sobre muestreo	80
<b>Capítulo 5</b>	<b>Comparaciones entre dos medias muestrales</b>	<b>83</b>
5.1	Introducción	83
5.2	Pruebas de significancia	83
5.3	Prueba de hipótesis de que una media poblacional es un valor dado	88
5.4	Pruebas de dos o más medias	91
5.5	Comparación de dos medias muestrales, muestras independientes y varianzas iguales	93
5.6	Modelo lineal aditivo	97
5.7	Comparación de medias muestrales; observaciones pareadas de importancia	98
5.8	El modelo lineal aditivo para comparaciones pareadas	101
5.9	Muestras independientes y varianzas desiguales	102
5.10	La media y la varianza de una función lineal	103

5.11	Prueba de hipótesis de igualdad de varianzas	108
5.12	Poder, tamaño de la muestra y determinación de diferencias	109
5.13	Muestras bietápicas de Stein	116
<b>Capítulo 6</b>	<b>Principios de diseño experimental</b>	<b>118</b>
6.1	Introducción	118
6.2	¿Qué es un experimento?	118
6.3	Objetivos de un experimento	119
6.4	Unidad experimental y tratamiento	120
6.5	Error experimental	121
6.6	Repeticiones y sus funciones	122
6.7	Factores que afectan el número de repeticiones	123
6.8	Precisión relativa de diseños con pocos tratamientos	125
6.9	Control del error	125
6.10	Elección de los tratamientos	128
6.11	Refinamiento de la técnica	128
6.12	Aleatorización	129
6.13	Inferencia estadística	130
<b>Capítulo 7</b>	<b>Análisis de la varianza I: Clasificación de una vía</b>	<b>132</b>
7.1	Introducción	132
7.2	El diseño completamente aleatorio	132
7.3	Datos con un solo criterio de clasificación: El análisis de la varianza para cualquier número de grupos con igual número de repeticiones	134
7.4	Datos con un solo criterio de clasificación: El análisis de la varianza para cualquier número de grupos con número desigual de repeticiones	140
7.5	El modelo lineal aditivo	144
7.6	Análisis de la varianza con submuestras: Número igual de submuestras	148
7.7	Modelo lineal para submuestreo	154
7.8	Análisis de la varianza con submuestras: Desigual número de submuestras	156
7.9	Componentes de la varianza en experimentos planeados con submuestras	159
7.10	Supuestos en que se fundamenta el análisis de la varianza	162
<b>Capítulo 8</b>	<b>Comparaciones múltiples</b>	<b>166</b>
8.1	Introducción	166
8.2	La diferencia mínima significante	167
8.3	Comparaciones	171
8.4	Prueba de efectos sugeridos por los datos	175
8.5	Prueba de Scheffé	177
8.6	Procedimiento <i>w</i> de Tukey	179

## **x CONTENIDO**

8.7	Prueba de Student-Newman-Keuls o S-N-K	180
8.8	Nueva prueba de amplitud múltiple de Duncan	181
8.9	Comparación de todas las medias con un control	182
8.10	Prueba de $r$ de razón $k$ bayesiana de Waller-Duncan	184
8.11	Pruebas de medias con número desigual de repeticiones	185
<b>Capítulo 9</b>	<b>Análisis de la varianza II: clasificaciones múltiples.</b>	<b>188</b>
9.1	Introducción	188
9.2	El diseño de bloque completo al azar	188
9.3	Análisis de la varianza para cualquier número de tratamientos; diseño de bloque completo al azar	190
9.4	La naturaleza del término de error	195
9.5	Partición del error experimental	198
9.6	Datos faltantes	202
9.7	Estimación de la ganancia en eficiencia	207
9.8	El diseño de bloques completos al azar: más de una observación por tratamiento por bloque	208
9.9	Modelos lineales y el análisis de la varianza	211
9.10	Agrupamiento doble: cuadrados latinos	213
9.11	Análisis de la varianza del cuadrado latino	215
9.12	Parcelas faltantes en el cuadrado latino	219
9.13	Estimación de la ganancia en eficiencia	221
9.14	El modelo lineal para el cuadrado latino	223
9.15	El tamaño de un experimento	224
9.16	Transformaciones	226
<b>Capítulo 10</b>	<b>Regresión lineal</b>	<b>231</b>
10.1	Introducción	231
10.2	La regresión lineal de $Y$ con respecto a $X$	231
10.3	El modelo y la ecuación de regresión lineal	236
10.4	Fuentes de variación en la línea de regresión lineal	240
10.5	Valores de regresión y valores ajustados	242
10.6	Desviaciones estándar, intervalos de confianza y pruebas de hipótesis	244
10.7	Control de la variación por observaciones concomitantes	248
10.8	Diferencia entre dos regresiones independientes	250
10.9	Una predicción y su varianza	253
10.10	Predicción de $X$ , modelo I	256
10.11	Distribuciones bivariantes, modelo II	256
10.12	Regresión a través del origen	258
10.13	Ánálisis de regresión ponderada	261
<b>Capítulo 11</b>	<b>Correlación lineal</b>	<b>263</b>
11.1	Introducción	263
11.2	La correlación y el coeficiente de correlación	263
11.3	Correlación y regresión	268

11.4	Distribuciones muestrales, intervalos de confianza y pruebas de hipótesis	269
11.5	Homogeneidad de los coeficientes de correlación	271
11.6	Correlación intraclasses	273
<b>Capítulo 12</b>	<b>Notación matricial</b>	<b>276</b>
12.1	Introducción	276
12.2	Matrices	277
12.3	Operaciones con matrices	278
12.4	Inversas, dependencia lineal, y rango	283
<b>Capítulo 13</b>	<b>Regresión lineal en notación matricial</b>	<b>288</b>
13.1	Introducción	288
13.2	El modelo y la estimación de mínimos cuadrados	288
13.3	El análisis de la varianza	292
13.4	Desviaciones estándar, intervalos de confianza y pruebas de hipótesis	294
13.5	Estimación y predicción	296
13.6	Variables indicadoras o binarias	298
<b>Capítulo 14</b>	<b>Regresión y correlación múltiple y parcial</b>	<b>303</b>
14.1	Introducción	303
14.2	La ecuación lineal y su interpretación en más de dos dimensiones	304
14.3	Regresión lineal parcial, total y múltiple	306
14.4	La ecuación muestral de regresión lineal múltiple	308
14.5	Regresión lineal múltiple; dos variables independientes	309
14.6	Correlación parcial y múltiple	316
14.7	Regresión lineal múltiple; resultados impresos para $k$ variables independientes	320
14.8	Miscelánea	324
14.9	Coeficientes de regresión parcial estándar	325
<b>Capítulo 15</b>	<b>Análisis de la varianza III: experimentos factoriales</b>	<b>328</b>
15.1	Introducción	328
15.2	Experimentos factoriales	328
15.3	El experimento factorial $2 \times 2$ : un ejemplo	334
15.4	Factorial $3 \times 3 \times 2$ ó $3^2 \times 2$ : un ejemplo	340
15.5	Modelos lineales para experimentos factoriales	346
15.6	Clasificaciones de $n$ vías y experimentos factoriales; superficies de respuesta	352
15.7	Grados de libertad individuales; tratamientos igualmente espaciados	354
15.8	Un solo grado de libertad para no aditividad	363

<b>Capítulo 16</b>	<b>Análisis de la varianza IV: diseño y análisis de parcelas divididas</b>	<b>368</b>
16.1	Introducción	368
16.2	Diseños de parcelas divididas	368
16.3	Un ejemplo de parcelas divididas	374
16.4	Datos faltantes en diseños de parcelas divididas	379
16.5	Diseño de bloques divididos	381
16.6	Modelos de parcelas y de bloques divididos	384
16.7	Parcelas divididas en espacio y tiempo	384
16.8	Serie de experimentos semejantes	386
<b>Capítulo 17</b>	<b>Análisis de la covarianza</b>	<b>392</b>
17.1	Introducción	392
17.2	Usos del análisis de la covarianza	392
17.3	El modelo y los supuestos para la covarianza	396
17.4	Prueba de medias de tratamientos ajustadas	398
17.5	La covarianza en el diseño de bloques completos al azar	401
17.6	Ajuste de las medias de tratamiento	406
17.7	Aumento de precisión debido a la covarianza	408
17.8	Partición de la covarianza	409
17.9	Homogeneidad de coeficientes de regresión	412
17.10	La varianza cuando se partitiona la suma de cuadrados de tratamiento	413
17.11	Estimación de observaciones faltantes mediante la covarianza	417
17.12	Covarianza con dos variables independientes	418
17.13	Cálculos de alta velocidad y salidas de computador	424
<b>Capítulo 18</b>	<b>Análisis de la varianza V: número desigual de subclases</b>	<b>428</b>
18.1	Introducción	428
18.2	Observaciones múltiples dentro de subclases	428
18.3	Análisis de un número proporcionado de subclases	429
18.4	Análisis de un número no proporcionado de subclases	432
18.5	Otras técnicas analíticas	440
<b>Capítulo 19</b>	<b>Ajuste de curvas</b>	<b>442</b>
19.1	Introducción	442
19.2	Regresión no lineal	442
19.3	Curvas logarítmicas o exponenciales	444
19.4	El polinomio de segundo grado	450
19.5	Polinomios ortogonales	451
<b>Capítulo 20</b>	<b>Algunos usos del Ji-cuadrado</b>	<b>458</b>
20.1	Introducción	458
20.2	Intervalos de confianza para $\sigma^2$	458

20.3	Homogeneidad de la varianza	460
20.4	Bondad de ajuste para distribuciones continuas	461
20.5	Combinaciones de probabilidades de pruebas de significancia	464-
<b>Capítulo 21</b>	<b>Datos enumerativos I: clasificaciones de una vía</b>	<b>466</b>
21.1	Introducción	466
21.2	El criterio de prueba $\chi^2$	466
21.3	Tablas de dos celdas, límites de confianza para una proporción o porcentaje	467
21.4	Tablas de dos celdas, pruebas de hipótesis	471
21.5	Pruebas de hipótesis para un conjunto limitado de alternativas	474
21.6	Tamaño de la muestra	478
21.7	Tablas de una vía con $n$ celdas	480
<b>Capítulo 22</b>	<b>Datos enumerativos II: tablas de contingencia</b>	<b>482</b>
22.1	Introducción	482
22.2	El modelo de muestreo aleatorio	482
22.3	El modelo de muestreo aleatorio estratificado	486
22.4	Tabla cuádruple o de $2 \times 2$	489
22.5	“Prueba exacta” de Fisher	491
22.6	Muestras no independientes en tablas $2 \times 2$	493
22.7	Homogeneidad de muestras de dos celdas	495
22.8	Aditividad de $\chi^2$	497
22.9	Más sobre la aditividad de $\chi^2$	498
22.10	Regresión lineal, tablas $r \times 2$	501
22.11	Tamaño de la muestra en tablas $2 \times 2$	503
22.12	Clasificación de $n$ vías	504
<b>Capítulo 23</b>	<b>Algunas distribuciones discretas</b>	<b>508</b>
23.1	Introducción	508
23.2	La distribución hipergeométrica	508
23.3	La distribución binomial	510
23.4	Ajuste de una distribución binomial	510
23.5	Transformación para la distribución binomial	514
23.6	La distribución de Poisson	515
23.7	Otras pruebas con distribuciones de Poisson	517
<b>Capítulo 24</b>	<b>Estadística no paramétrica</b>	<b>520</b>
24.1	Introducción	520
24.2	Prueba $\chi^2$ de bondad de ajuste	521
24.3	Prueba de Kolmogorov-Smirnov con una muestra	522
24.4	La prueba de signos	524
24.5	Prueba de rangos signados de Wilcoxon	526
24.6	Prueba de Kolmogorov-Smirnov de dos muestras	527

24.7	Prueba de Wilcoxon-Mann-Whitney con dos muestras	528
24.8	Prueba de la mediana	530
24.9	Prueba de Kruskal-Wallis con $k$ muestras	530
24.10	Prueba de la mediana para $k$ muestras.	532
24.11	Prueba de Friedman para la clasificación de dos vías	532
24.12	Una prueba de la mediana para la clasificación de dos vías	534
24.13	Desigualdad de Chebyshew	534
24.14	Coeficiente de correlación de rangos de Spearman	536
24.15	Prueba de asociación del cuadrante de Olmstead-Tukey	537
24.16	Prueba de aleatorización para regresión	539
Capítulo 25		
Muestreo de poblaciones finitas		541
25.1	Introducción	541
25.2	Organización del estudio	542
25.3	Muestreo probabilístico	543
25.4	Muestreo aleatorio simple	544
25.5	Muestreo estratificado	547
25.6	Asignación óptima	550
25.7	Muestreo multietápico o por conglomerados	553
Apéndice		559
Tablas		559
Índice		613

---

## PREFACIO

---

Esta segunda edición de *Bioestadística (Principles and Procedures of Statistics: A Biometrical Approach)* reconoce el hecho de que la estadística es necesaria, y ya es usada por un creciente número de disciplinas. Los principios estadísticos son independientes de la materia en la cual se aplican, y los procedimientos aplicados en la agricultura y las ciencias biológicas pueden llevarse a otras áreas como la industria, el gobierno, la ingeniería, la medicina, y dar allí tan buenos resultados como en aquellas; podría decirse, a cualquier área donde se adelante la investigación. Las universidades y colegios superiores aceptan generalmente este hecho y exigen uno o varios cursos de estadística como requisito para otorgar títulos superiores. El rápido crecimiento de la enseñanza de estadística en los cursos de pregrado también está asociado con los requisitos exigidos en este nivel. Este extraordinario crecimiento en el uso de la estadística es paralelo en cierta medida con un rápido desarrollo de los procedimientos estadísticos, algunos de los cuales se cubren en este libro. Otros escapan a su nivel docente. Tendencias como la expansión en el uso de la estadística y el crecimiento de los métodos explican por qué después de una vida vigorosa y de éxito en estos 20 años, la edición original *Bioestadística (Principles and Procedures of Statistics: A Biometrical Approach)* ha debido ceder el paso a esta nueva edición actualizada, reorganizada y ampliada.

Los supuestos básicos de ambas ediciones siguen siendo iguales: un enfoque esencialmente no matemático porque los desarrollos algebraicos parecen crear temores en algunos estudiantes; presentación y análisis tempranos del diseño experimental de modo que los estudiantes y los profesionales en centros de investigación puedan aplicar los métodos estadísticos aunque todavía se hallen en el proceso de aprenderlos, e incorporación de suficientes técnicas que satisfagan las necesidades de la mayoría de investigadores.

Esta edición tiene unas 200 páginas más que la primera. Obviamente se cubre más material. Entre otras cosas, tiene en cuenta los comentarios y las sugerencias que durante estos años se han hecho a la primera edición. Una de ellas y muy frecuente se refería al nivel de la lengua, argumentando que era muy difícil debido a su concisión. El autor ha

revisado el libro palabra por palabra y párrafo por párrafo en un esfuerzo por ampliar las explicaciones y, por consiguiente, simplificar el contenido.

Se han incluido muchas técnicas nuevas, algunas de las cuales se estudian brevemente y otras, con mayor detenimiento. Entre los múltiples y nuevos procedimientos de comparación se incluyen la dms protegida de Fisher, la prueba de Scheffé y la prueba  $t$  para la razón  $k$  bayesiana de Waller-Duncan, las útiles variables de indicadores se tratan breve, pero apropiadamente; el procedimiento de Satterthwaite para calcular pruebas cuasi razones  $F$  se explica cuidadosamente en una parte y se demuestra en otras; los polinomios ortogonales se usan para producir ecuaciones de superficies de respuesta en experimentos factoriales, y el capítulo 24, sobre estadística no paramétrica, incluye ahora las pruebas de una y dos muestras para la bondad de ajuste de Kolmogorov-Smirnov.

No se ha descuidado la modernización de las técnicas estándar. Los diseños de parcelas divididas en el tiempo se consideran como ejemplos de diseño de bloques divididos. El capítulo 14, sobre regresión múltiple, se presenta en notación matricial, el enfoque moderno que necesita de la interpretación de resultados impresos que resultan de la tecnología de la computación. La notación matricial se ha presentado en los dos capítulos anteriores: el capítulo 12, sobre definiciones y procedimientos de operación, pretende estimular al usuario de los paquetes de cálculos estadísticos a que adquiera un mayor dominio de los resultados impresos del computador, que los acompañan; el capítulo 13, sobre regresión lineal en presentación de matricial, presenta en un desarrollo paralelo el enfoque usual, tal como se desarrolló en el capítulo anterior. Este capítulo ayuda al lector a hacer la transición a la notación matricial; y el análisis de datos numérico de subclases desproporcionadas se relaciona con la regresión múltiple y con los resultados impresos del computador.

Hay una mejora en la organización del material respecto de la primera edición. Por ejemplo, el análisis de la distribución binomial aparece ya desde el capítulo 2, en donde es muy útil para la presentación de la distribución normal; la discusión de una función lineal, su media y su varianza se han adelantado, de suerte que pueda relacionarse con la comparación de dos medias, sea de muestras dependientes o independientes; los temas sobre contrastes y comparaciones múltiples se han reunido en un nuevo capítulo 8, en donde las tasas de error se han definido cuidadosamente y se presentan guías y advertencias sobre la confusión que puede generar un exceso de pruebas; el capítulo 9 se beneficia de una discusión mejorada del método para determinar el tamaño de un experimento; y el capítulo 24 ofrece una presentación más ordenada de la estadística no paramétrica.

Se ha prestado atención a presentaciones alternativas, por ejemplo, un tratamiento más adecuado del uso de los componentes de la varianza en el planeamiento de experimentos con atención a los costos involucrados; un manejo más apropiado de correlaciones entre clases; un mejor enfoque —olvidado en la primera edición— de las tablas de contingencia, usando modelos.

Por último, en atención a muchas solicitudes y sugerencias, la selección de ejercicios se ha aumentado considerablemente para incluir datos tomados de una gama más amplia dentro de las ciencias biológicas. Además, se han incluido datos obtenidos de las ciencias sociales. Para un conjunto de datos de pre-prueba y post-prueba, los análisis propuestos incluyen tratamientos de dos conjuntos de datos, como problema de regresión y como diseño de bloques divididos.

Los autores quedan muy reconocidos con el profesor Sir Ronald A. Fisher, Cambridge, con el doctor Frank Yates, Rothamsted, y con Oliver and Boyd, Ltd., Edimburgo, por haber autorizado la reproducción de la tabla III de su libro *Statistical Tables for Biological, Agricultural and Medical Research*.

Los autores también expresan sus agradecimientos a Fred Gruenberger y al Numerical Analysis Laboratory de la Universidad de Wisconsin por su preparación de la tabla A.1; a E. S. Pearson y H. O. Hartley, editores de *Biometrika Tables for Statisticians*, Vol I, y a *Biometrika* por su permiso para reproducir las tablas A.2, A.6, A.8 y A.15; a C.M. Thompson y a *Biometrika* por su permiso para reproducir la tabla A.5; a D. B. Duncan y al editor de *Biometrika* por su permiso para reproducir la tabla A.7; a C. W. Dunnett y al editor del *Journal of the American Statistical Association* por su permiso para reproducir la tabla A.9; a C. I. Bliss por su permiso para reproducir la tabla A.10 y a F. N. David y *Biometrika* por su permiso para reproducir la tabla A.11; a L. M. Milne-Thomson y L. J. Comrie, autores de *Standard Four-figure Mathematical Tables* y a MacMillan Co. Ltd., Londres, por su permiso para reproducir la tabla A.12; a G. W. Snedecor, autor de *Statistical Methods*, 4a. ed. y a la Iowa State College Press por su permiso para reproducir la tabla A.13; a D. Mainland, L. Herrera y M. I. Sutcliffe por su permiso para reproducir la tabla A.14; a F. Mosteller y J. W. Tukey, editor del *Journal of the American Statistical Association*, y a Codex Book Company Inc., por su permiso para reproducir la tabla A.16; a Prasert Na Nagara, por su permiso para reproducir la tabla A.17; a Frank Wilcoxon y a la American Cyanamid Company por su permiso para reproducir la tabla A.18; a Colin White y al editor de *Biometrics* por su permiso para reproducir la tabla A.19; a P. S. Olmstead, J. W. Tukey, Bell Telephone Laboratories y al editor de *Annals of Mathematical Statistics* por su permiso para reproducir la tabla A.20; a D. B. Duncan por su permiso para reproducir la tabla A.21; a L. H. Miller y al editor de *Journal of the American Statistical Association* por su permiso para reproducir la tabla A.22; a Z. W. Birnbaum, R. A. Hall y al editor de *Annals of Mathematical Statistics* por su permiso para reproducir la tabla A.23;

En particular, deseo agradecer a Wyman Nyquist por su valiosa crítica de la primera edición y del manuscrito de la revisión.

Además, tengo deuda de gratitud con muchos de mis colegas por sus sugerencias acerca de varios temas, con otros por sus generosos permisos para usar datos, y con aquellas personas que me ayudaron en la preparación del manuscrito. Me hubiera extraviado sin las destrezas de Dorothy Green, quien mecanografió, cortó y pegó el manuscrito final. Por último deseo agradecer a mi esposa, Jennie, por su lectura cuidadosa de las pruebas y ayuda editorial.

## SIMBOLOS ESCOGIDOS

$\neq$	no es igual a; por ejemplo, $3 \neq 4$
$>$	mayor que; por ejemplo, $5 > 2$
$\geq$	mayor que o igual a
$<$	menor que; por ejemplo, $3 < 7$
$\leq$	menor que o igual a
$  \quad  $	valor absoluto; por ejemplo $  -7   = 7$
$\sum$	suma de
$\dots$	indica un conjunto de cantidades faltantes; por ejemplo $1, 2, \dots, 10$
$n!$	$n(n - 1) \dots 1$ llamado $n$ factorial; por ejemplo, $3! = 3(2) 1 = 6$
$\bar{}$	se usa para indicar el promedio aritmético de una media
$\hat{\quad}$	sombrero; se usa para indicar una estimación, no tanto un valor verdadero; por lo general aparece sobre letras griegas.

<b>Letras griegas</b>	con pocas excepciones se refieren a parámetros de una población
$\mu$	media poblacional
$\sigma^2, \sigma$	varianza poblacional y desviación estándar
$\tau, \beta, \text{etc.}$	componentes de las medias poblacionales; se usan comúnmente junto con modelos lineales
$\varepsilon$	error experimental verdadero
$\delta$	error verdadero en la muestra; diferencia real
$\beta$	coeficiente de regresión de la población, efecto de bloque
$\rho$	- coeficiente de correlación de la población

## xx SIMBOLOS ESCOGIDOS

$N, S^2$  estas letras latinas se usan como símbolos que indican poblaciones finitas, en especial en el cap. 25.

Las anteriores letras griegas se usan también con subíndices para mayor claridad. Por ejemplo:

- $\mu_Y$  media poblacional de las  $\bar{Y}$
- $\beta_{Y \cdot X \cdot Z}$  regresión de  $Y$  sobre  $X$  con  $Z$  fijo
- $\tau_i$  contribución de la media poblacional que recibe el  $i$ -ésimo tratamiento

Algunas excepciones en el uso de letras griegas para indicar parámetros son:

- $\alpha$  probabilidad de un error de Tipo I
- $1 - \alpha$  coeficiente de confianza
- $\beta$  probabilidad de un error de Tipo II
- $1 - \beta$  poder de una prueba estadística
- $\chi^2$  criterio común de prueba

Letras latinas	se usan como símbolos generales, incluyendo los de estadística muestral
$Y$	variable
$Y_i, Y_{ij}$	observaciones individuales
$Y_i, Y_{..}$	totales de observaciones
$D_j$	diferencia entre observaciones pareadas, $Y_{1j} - Y_{2j}$
$n, n_{..}$	total de tamaño de la muestra
$n_{ij}$	número de observaciones en $i, j$ -ésima celda
$\bar{Y}, \bar{Y}_{..}, \bar{Y}_i$	medias muestrales, total o parte de una muestra
$\bar{\bar{Y}}$	media de las medias muestrales
$s^2, s_y^2, s_D^2$	varianzas muestrales, estimaciones no sesgadas de $\sigma^2, \sigma_y^2$ , y $\sigma_D^2$
$s, s_y, s_D$	desviaciones estándar de la muestra
$s_{Y \cdot X}^2, s_{Y \cdot 1 \dots k}^2$	varianzas muestrales ajustadas por regresión
CL, CI	límites de confianza, o intervalos
$l_1, l_2$	puntos extremos de los límites de confianza
$b$	coeficiente de regresión de la muestra
$b_{Y \cdot 1 \dots k}$	coeficiente de regresión parcial de la muestra
$b'$	coeficiente de regresión estándar
$r$	total de la muestra o coeficiente de correlación simple
$r_{1 \cdot 2 \cdot 3 \dots k}$	coeficiente de correlación parcial de la muestra en $X_1$ y $X_2$
$R_{1 \cdot 2 \dots k}$	coeficiente de correlación múltiple entre $X_1$ y otras variables

$gl$ , $f$	grados de libertad
$FC, C, TC$	factor de corrección, valor de corrección, término de corrección
$SC$	$\sum (Y_i - \bar{Y})^2$ , suma de cuadrados
$CM$	cuadrado medio
$E_a, E_b$	cuadrados de las medias del error en un diseño de parcelas divididas
$E_{YY}, E_{XY}, E_{XX}$	sumas del error de los productos en la covarianza (se usan otras letras para indicar otras fuentes de variación)
*	significante, por ejemplo, 2.3*
**	altamente significante, por ejemplo, 14.37**
$ns$	no significante
$dms$	diferencia mínima significante
$RE$	eficiencia relativa
$CV$	coeficiente de variabilidad ( $s/Y$ )100
$L = \sum c_i Y_i$	una función lineal de observaciones, $c_i$ es constante
$Q = \sum c_i Y_i$	comparación en la cual $c_i$ es constante, $Y_i$ es a menudo un tratamiento total y $\sum c_i = 0$
$cpf$	corrección por población finita
$upm$	unidad primaria de muestreo
$est$	estratificado, se usa como subíndice
$P$	probabilidad
$p, 1 - p$	probabilidades en una distribución binomial
$H_0$	hipótesis nula
$H_1$	hipótesis alternativa, usualmente conjunto de alternativas
$\infty$	infinito

---

CAPITULO  
UNO

---

INTRODUCCION

### 1.1 Definición de la estadística

La estadística moderna proporciona conocimientos a los investigadores. Es un tema nuevo y estimulante, producto del siglo XX. Para el científico, particularmente para el científico en Biología, la estadística comenzó aproximadamente en 1925 cuando apareció el libro de Fisher, *Statistical Methods for Research Workers*.

La estadística es un tema de rápido crecimiento con mucho material original que todavía no se encuentra en textos; crece a medida que los estadísticos encuentran respuestas a más y más problemas propuestos por los investigadores. Algunos de los primeros investigadores que contribuyeron al desarrollo inicial de la estadística todavía laboran activamente, y los nuevos encuentran diversas oportunidades para sus talentos investigativos. En la aplicación de la estadística, los principios son generales aun cuando las técnicas puedan diferir, y la necesidad de formación estadística crece a medida que se incrementa la aplicación a las ciencias biológicas y sociales, la ingeniería y la industria.

Este tema nuevo y vigoroso afecta a todos los aspectos de la vida moderna. Por ejemplo, el planeamiento estadístico y la evaluación de la investigación contribuyen a los avances tecnológicos en el cultivo y procesamiento de alimentos; el control estadístico de calidad de los productos manufacturados hace confiables los equipos automotores y eléctricos. La estadística ayuda a los encuestadores a recolectar datos para determinar las preferencias de esparcimiento del público; proporciona información para los estudios de impacto ambiental y ayuda en la evaluación de las exigencias gubernamentales para que la industria farmacéutica demuestre que un producto es benéfico y no sólo inofensivo. Cada vez son más los grupos de investigación en los cuales se encuentra un estadístico.

La extensión de la estadística hace difícil su definición. Su desarrollo obedeció a la necesidad de tratar problemas en los cuales, para observaciones individuales, las leyes de causa y efecto no aparecen claramente al observador y donde es necesario un enfoque objetivo. En tales problemas siempre existe un cierto grado de incertidumbre en toda inferencia basada en un número limitado de observaciones. Por lo tanto, para nuestro propó-

sito, una definición razonable y satisfactoria sería: *La estadística es la ciencia, pura y aplicada, que crea, desarrolla y aplica técnicas de modo que pueda evaluarse la incertidumbre de inferencias inductivas.*

Para la mayoría de los científicos, la estadística es lógica o sentido común con un fuerte ingrediente de procedimientos aritméticos. La lógica proporciona el método mediante el cual se deben recolectar los datos y determinar cuánto deben abarcar; la aritmética, junto con ciertas tablas numéricas, produce el material sobre el cual se basa la inferencia y se mide la incertidumbre. La parte aritmética es a menudo rutinaria, y el estudiante necesita de formación matemática especial. No vamos a ocuparnos directamente con las matemáticas, ya que es difícil encontrar un campo de esta materia que no haya dado al estadístico alguna teoría útil.

## 1.2 Breve historia de la estadística

La historia de la estadística aclara en gran medida la naturaleza de la misma en el siglo XX. La perspectiva histórica también es importante para ver las necesidades y las presiones que la crearon.

El término estadística no es nuevo. La estadística debió comenzar como una aritmética estatal para asistir al gobernante que necesitaba conocer la riqueza y el número de sus súbditos con el objeto de recaudar impuestos o presupuestar la guerra. Es de presumir que todas las culturas que intencionalmente registraron su historia también registraron sus estadísticas. Sabemos que César Augusto decretó que todos los súbditos tenían que tributar y por lo tanto exigió a todas las personas que se presentaran al estadístico más cercano, que entonces era el recaudador de impuestos. Debido a lo anterior, Jesús nació en Belén, no en Nazareth. Guillermo el Conquistador ordenó un censo de las tierras de Inglaterra para fines de tributación y de servicio militar. Este se llamó "Domesday Book". Tales estadísticas son historia.

Varios siglos después del "Domesday Book", encontramos una aplicación de la probabilidad empírica al seguro de embarque, del cual parece haber dispuesto la navegación flamenca del siglo XIV. Esto pudo haber sido poco más que pura especulación o juego de azar, pero llegó a ser la forma muy respetable de la estadística llamada seguros.

El juego, en forma de juegos al azar, originó la teoría de las probabilidades, desarrollada por Pascal y Fermat, a mediados del siglo XVII, debido a su interés en las experiencias de juego del Caballero de Meré. Para el estadístico y el científico experimental, tal teoría tiene mucho uso práctico en la informática.

La curva normal o la curva normal de errores ha sido muy importante en el desarrollo de la estadística. La ecuación de esta curva fue originalmente publicada en 1733 por de Moivre, quien no supo cómo aplicar sus resultados a observaciones experimentales y su escrito permaneció desconocido hasta que Karl Pearson lo encontró en una biblioteca en 1924. Sin embargo, al mismo resultado llegaron luego dos astrónomos matemáticos, Laplace, 1749-1827, y Gauss, 1777-1855, independientemente el uno del otro.

Un razonamiento esencialmente estadístico fue aplicado en el siglo XIX por Charles Lyell a un problema geológico. En el período comprendido entre 1830 y 1833 aparecieron tres volúmenes de *Principles of Geology* de Lyell, quien estableció el orden de las rocas terciarias y les asignó nombres. Con M. Deshayes, un conquiliólogo francés, identificó y enumeró especies fósiles que se presentaban en uno o más estratos, y también logra-

ron dar las proporciones de las que aún vivían en ciertas partes de los mares. Basados en estas proporciones asignaron los nombres de: Pleistoceno (novísimo), Plioceno (más reciente), Mioceno (menos reciente) y Eoceno (reciente). El razonamiento de Lyell fue esencialmente estadístico. Una vez establecidos y aceptados los nombres, el método fue casi inmediatamente olvidado. No había geólogos evolucionistas que se preguntaran si se trataba de etapas discretas, como lo implican los nombres, o bien si era un proceso continuo y se podía utilizar para hacer predicciones.

Otros descubrimientos científicos del siglo XX también se hicieron sobre una base estadística sin que se advirtiera apenas la naturaleza estadística de la técnica, y desafortunadamente el método cayó pronto en el olvido. Esta afirmación es válida para las ciencias biológicas y las físicas.

Charles Darwin, 1809-1882, biólogo, recibió en el *Beagle* el segundo volumen del libro de Lyell. Posteriormente Darwin formuló sus teorías y bien pudo haber influido en él la lectura de ese libro. La obra de Darwin fue, en gran parte, la naturaleza biométrica o estadística, y ciertamente renovó el entusiasmo por la Biología. Mendel, con sus estudios sobre híbridos vegetales publicados en 1866, también tuvo un problema biométrico o estadístico.

En el siglo XIX, la necesidad de una base más sólida para la estadística se hizo manifiesta. Karl Pearson, 1857-1936, inicialmente físico matemático aplicó sus matemáticas a la evolución, como resultado del entusiasmo que generó Darwin en la Biología. Pearson dedicó casi medio siglo a la investigación estadística rigurosa. Además, fundó la revista *Biometrika* y una escuela de estadística; con ello tomó impulso el estudio de la estadística.

Si bien Pearson se ocupaba de muestras grandes, la teoría correspondiente resultaba inadecuada para los experimentadores que trabajan con muestras necesariamente pequeñas. Entre estos estaba W. S. Gosset, 1876-1937, quien estudiaba con Karl Pearson y era técnico de la firma de cerveceros Guinness. Parece que la matemática de Gosset era insuficiente para encontrar distribuciones exactas de la desviación estándar de la muestra, la relación entre la media de la muestra y la desviación estándar de la muestra, del coeficiente de correlación, estadígrafo al que dedicó especial interés. Por lo tanto, recurrió a sacar cartas calculando y compilando distribuciones de frecuencia empírica. Sus escritos sobre los resultados aparecieron en *Biometrika* en 1908 bajo el nombre de Student, seudónimo de Gosset mientras trabajaba con Guinness. Hoy, la *t* de Student es instrumento fundamental para estadísticos y experimentadores, y "estudiantizar" es expresión corriente en estadística. Ahora que el uso de la distribución *t* de Student está tan generalizado, es interesante anotar que el astrónomo alemán, Helmert, ya la había obtenido matemáticamente en 1875.

R. A. Fisher, 1890-1962, recibió influencias de Karl Pearson y de Student, e hizo numerosas e importantes contribuciones a la estadística. El y sus estudiantes dieron considerable impulso al uso de los procedimientos estadísticos en muchos campos, particularmente en agricultura, biología y genética.

J. Neyman, 1894, y E. S. Pearson, 1895, presentaron una teoría sobre la verificación o prueba de hipótesis estadísticas en 1936 y 1938. La teoría fomentó en forma considerable la investigación y muchos de los resultados son de uso práctico.

En esta breve historia, mencionaremos sólo otro estadístico. Abraham Wald, 1902-1950. Sus dos libros, *Sequential Analysis* y *Statistical Decision Functions*, se ocupan de

grandes conquistas estadísticas no tratadas en este texto, no obstante, una aplicación, la solución mínima de un problema de genética, se ilustra en el capítulo 21.

En este siglo entonces se han desarrollado la mayoría de los métodos que actualmente se utilizan. La estadística de este texto es parte de esos métodos.

### 1.3 La estadística y el método científico

Se dice que los científicos usan el método científico. Sería difícil definir la expresión método científico, dado que los científicos usan cuantos métodos o medios puedan concebir. Sin embargo, la mayoría de estos métodos tienen puntos esenciales en común. Sin intentar promover una controversia, consideramos que éstos son:

1. Una revisión de hechos, teorías y propuestas.
2. Formulación de una hipótesis lógica sujeta a prueba mediante métodos experimentales.
3. Evaluación objetiva de las hipótesis con base en los resultados experimentales.

Mucho podría escribirse respecto a estos puntos esenciales: ¿Cómo se llega a una hipótesis? ¿Cómo se diseña un experimento? ¿Cómo se evalúa objetivamente una hipótesis?

La ciencia es un estudio que se ocupa de la observación y clasificación de los hechos. Los científicos deben, entonces, ser capaces de observar un suceso o conjunto de eventos como resultado de un plan o diseño. Esto es el experimento, la sustancia del método científico. El diseño experimental es un campo de la estadística.

La evaluación objetiva de una hipótesis presenta problemas, puesto que no es posible observar todos los eventos concebibles, y como las leyes exactas de causa-efecto generalmente se desconocen, existirá variación entre los que son observados. El científico debe entonces razonar partiendo de casos particulares a casos más generales. Este proceso es de inferencia incierta. Es un proceso que nos capacita para desaprobar hipótesis incorrectas, pero no nos permite aprobar hipótesis correctas. Lo único que podemos dar como demostración es una *comprobación fuera de duda razonable*. Los procedimientos estadísticos son métodos que nos conducen a esta suerte de pruebas.

Una parte de la información posible, necesariamente conduce sólo a inferencia incierta. El azar entra en juego en la obtención de información y es la causa de la incertidumbre. Al aplicar las leyes del azar, el estadístico de hoy puede realizar una medición objetiva y precisa de la incertidumbre de las inferencias. Ciertamente, esto se hace para la totalidad de las inferencias y no para cada inferencia individual. O sea que se sigue un procedimiento que asegure que 9 de 10 inferencias serán correctas, o 99 de 100, o algo por el estilo. ¿Por qué no estar siempre en lo correcto o muy cerca a lo correcto? El inconveniente es el costo. El costo puede subir debido al incremento del tamaño de la muestra, a consecuencia de una decisión incorrecta, o a la vaguedad de la inferencia necesaria para incluir la respuesta correcta.

El método científico no es una sucesión dispersa de secuencias de hipótesis experimento-inferencia que se ajusten perfectamente en compartimientos. Mas bien, si un científico no logra demostrar la falsedad de una hipótesis, quizá la teoría abarque hechos fuera del alcance de la inferencia del experimento o acaso modificándola, pueda abarcar tales hechos. El ciclo se repite entonces. Por otra parte todos los supuestos que entran en la hipótesis

pueden no ser necesarios; entonces se formula una nueva hipótesis con nuevos supuestos y se repite el ciclo.

En resumen, la estadística es un instrumento aplicable en el método científico, para el cual fue desarrollada. Su aplicación particular está en los muchos aspectos del diseño de un experimento, desde el plan inicial para la recolección de los datos, y en el análisis de los resultados a partir de los datos resumidos, hasta la evaluación de la incertidumbre de toda la inferencia extraída de ellos.

#### 1.4 El estudio de la estadística

No se intenta convertir en estadísticos profesionales a aquellos que lean y estudien este libro. Nuestro propósito es promover una forma de pensar clara y disciplinada, especialmente cuando se trata de recolectar e interpretar información numérica, y presentar un considerable número de técnicas estadísticas de aplicabilidad y utilidad generales en la investigación. Se requiere hacer cálculos en estadística, pero es cosa de aritmética, no de matemática ni estadística.

La estadística implica, para la mayoría de los estudiantes, una forma nueva de pensar en términos de incertidumbre o de improbabilidades. Acá como en otros casos, los estudiantes difieren en habilidad, y cuando se enfrentan a la estadística por primera vez, para algunos puede parecer una tortura mental que puede ser emocionalmente perturbadora. Creemos haber hecho todo el esfuerzo compatible con nuestros objetivos para minimizar los problemas del aprendizaje de la estadística.

Muchos estudiantes encontrarán que se aprende mejor la estadística mediante la aplicación directa a sus propios problemas; pocos encontrarán, en el transcurso de uno o dos períodos, la utilidad del material presentado. Por consiguiente, muchos estudiantes necesitarán considerable reflexión y discusión para obtener el máximo provecho de un curso basado en este texto. Se dan preguntas y ejercicios para estimular la reflexión y ofrecer alguna oportunidad de aplicar las técnicas y familiarizarse con ellas.

Finalmente, es necesario tener en cuenta que la estadística se ha propuesto como instrumento de investigación. La investigación puede ser en genética, mercadeo, nutrición, agronomía, etc. Es el campo de investigación, no el instrumento, el que debe proporcionar los "por qué" del problema de investigación. A veces, este hecho se pasa por alto y los usuarios olvidan que tienen que pensar, que la estadística no puede pensar por ellos. La estadística, sin embargo, ayuda a los investigadores a diseñar experimentos y a evaluar objetivamente los datos numéricos resultantes. Es nuestra intención proporcionar a los investigadores instrumentos estadísticos útiles para este fin.

#### Referencias

- 1.1. Box, Joan Fisher: *R. A. Fisher, The life of a scientist*, Wiley, Nueva York, 1978.
- 1.2. Committee of Presidents of Statistical Societies: *Careers in statistics*, current edition, American Statistical Association, Washington, D.C.
- 1.3. Eisenhart, Churchill: "Anniversaries in 1965 of interest to statisticians," *Amer. Statist.*, 19: 21-29 (1965)
- 1.4. Eisenhart, Churchill, y Allan Birnbaum: "Anniversaries in 1966-67 of interest to statisticians", *Amer. Statist.*, 21:22-29 (1967).
- 1.5. Fisher, R. A.: "Biometry." *Biom.*, 4:217-219 (1948).

## 6 BIOESTADISTICA: PRINCIPIOS Y PROCEDIMIENTOS

- 1.6. Fisher, R. A.: "The expansion of statistics," *J. Roy. Statist. Soc., Ser. A.*, 116:1-6 (1953).
- 1.7. Fisher, R. A.: "The expansion of statistics," *Amer. Sci.*, 42:275-282 y 293 (1954).
- 1.8. Freeman, Linton C., y Douglas M. More: "Teaching introductory statistics in the liberal arts curriculum," *Amer. Statist.*, 10:20-21 (1956).
- 1.9. Hotelling, Harold: "The teaching of statistics," *Ann. Math. Statist.*, 11:1-14 (1940).
- 1.10. Hotelling, Harold: "The impact of R. A. Fisher on statistics," *J. Amer. Statist. Ass.*, 46:35-46 (1951).
- 1.11. Hotelling, Harold: "Abraham Wald," *Amer. Statist.*, 5:18-19 (1951).
- 1.12. Hotelling, Harold: "The statistical method and the philosophy of science," *Amer. Statist.*, 12: 9-14 (1958).
- 1.13. McMullen, Launce: Foreword, en E. S. Pearson y John Wishart (eds.), "Student's" collected papers, *Biometrika* Office, University College, London, 1947.
- 1.14. Mahalanobis, P. C.: "Professor Ronald Aylmer Fisher," *Sankhya*, 4:265-272 (1938).
- 1.15. Mainland, Donald: "Statistics in clinical research; some general principles," *Ann. N.Y. Acad. Sci.*, 52:922-930 (1950).
- 1.16. Mather, Kenneth: "R. A. Fisher's *Statistical Methods for Research Workers*, an appreciation," *J. Amer. Statist. Ass.*, 46:51-54 (1951).
- 1.17. Menger, Karl: "The formative years of Abraham Wald and his work in geometry," *Ann. Math. Statist.*, 23:13-20 (1952).
- 1.18. Pearson, E. S.: "Karl Pearson, an appreciation of some aspects of his life and work, part I: 1857-1906," *Biometrika*, 28:193-257 (1936).
- 1.19. Pearson, E. S.: "Karl Pearson, an appreciation of some aspects of his life and work, part II: 1906-1936," *Biometrika*, 29:161-248 (1938).
- 1.20. Reid, R. D.: "Statistics in clinical research," *Ann. N.Y. Acad. Sci.*, 52:931-934 (1950).
- 1.21. Tintner, G.: "Abraham Wald's contributions to econometrics," *Ann. Math. Statist.*, 23:21-28 (1952).
- 1.22. Walker, Helen M.: "Bicentenary of the normal curve," *J. Amer. Statist. Ass.*, 29:72-75 (1934).
- 1.23. Walker, Helen M.: "Statistical literacy in the social sciences," *Amer. Statist.*, 5:6-12 (1951).
- 1.24. Walker, Helen M.: "The contributions of Karl Pearson," *J. Amer. Statist. Ass.*, 53:11-27 (1958).
- 1.25. Wolfowitz, J.: "Abraham Wald, 1902-1950," *Ann. Math. Statist.*, 23:1-13 (1952).
- 1.26. Yates, F.: "The influence of *Statistical Methods for Research Workers* on the development of the science of statistics," *J. Amer. Statist. Ass.*, 46:19-34 (1951).
- 1.27. Youden, W. J.: "The Fisherian revolution in methods of experimentation," *J. Amer. Statist. Ass.*, 46:47-50 (1951).

## OBSERVACIONES

### 2.1 Introducción

→ Las observaciones constituyen la materia prima con la cual trabajan los investigadores. Para que se pueda aplicar la estadística a esas observaciones éstas deben estar en forma numérica. En el mejoramiento de cultivos, los números bien pueden ser rendimientos por parcela; en la investigación médica, pueden ser tiempos de recuperación bajo varios tratamientos; en la industria, pueden ser cantidad de defectos en varios lotes de un artículo producido en una línea de montaje. Tales números constituyen *datos* y su característica común es la *variabilidad* o *variación*.

Este capítulo se refiere a la recolección, presentación, resumen y caracterización de la información. Se discutirán los conceptos de poblaciones, muestras, modelo lineal e inferencia estadística.

### 2.2 Variables

Proposiciones tales como "María es rubia", o "El pesa más de 20 libras" son comunes e informativas. Se refieren a características que no son constantes, sino que varían de una persona a otra y que sirven para distinguir o describir.

Las características que presentan variabilidad o variación se denominan *variables*, *variables aleatorias* o *variables de azar*.

Como gran parte de nuestro estudio debe ser general, empleamos algunos símbolos. En vez de escribir variable a cada oportunidad, sean  $Y$  la variable  $Y$  e  $Y_i$  (léase  $Y$  sub- $i$ ) la observación  $i$ -ésima. Aquí no tenemos en mente ninguna observación en particular. Cuando tengamos que referirnos a una observación específica, remplazaremos  $i$  por un número. Por ejemplo, si en una familia tres niños pesan 52, 29 y 28 libras, y  $Y$  denota peso,  $Y_1 = 52$  libras,  $Y_2 = 29$  libras y  $Y_3 = 28$  libras. En términos más generales y abstractos, denotamos un conjunto de observaciones mediante  $Y_1, Y_2, \dots, Y_n$ . Aquí  $Y_n$  se refiere al último término, el subíndice nos dice el número total, y los tres puntos entre  $Y_2$  e  $Y_n$  se

refieren al resto de observaciones, si las hay. En nuestro ejemplo,  $n = 3$ . Los símbolos se consideran una taquigrafía.

Las variables pueden ser cuantitativas o cualitativas.

Una variable cuantitativa es aquella para la cual las observaciones resultantes pueden medirse porque poseen un orden o rango natural; por ejemplo, estaturas, pesos y número de caras que se presentan cuando se lanzan 10 monedas.

Las observaciones sobre variables cuantitativas pueden clasificarse además en continuas o discretas.

Una variable continua es aquella que puede presentar cualquier valor dentro de cierto intervalo. Estatura y peso son ejemplos obvios. La estatura puede medirse con una aproximación de  $1/4$  de pulgada, pero esto no quiere decir que las estaturas existen solamente para valores de  $1/4$  de pulgada. Nuestra limitación depende del tipo de aparato de medida usado para obtener los valores de una variable continua.

Una variable discreta o discontinua es aquella para la cual los valores posibles no se pueden observar en una escala continua debido a la existencia de espacios entre estos posibles valores. A menudo las observaciones discretas son enteros porque provienen de conteos. Son ejemplos, el número de pétalos de una flor, el número de familias residentes en una manzana o el número de insectos atrapados en una red. Pero el espacio entre posibilidades sucesivas puede ser diferente de la unidad. El promedio de puntos que se presentan al lanzar dos dados también es una variable discreta, y sus valores van de uno a seis por incrementos de un medio.

Una variable cualitativa es aquella para la cual no es posible hacer mediciones numéricas. Se hace una observación cuando se asigna un individuo a una o varias categorías mutuamente excluyentes (no se puede asignar a más de una). Las observaciones no se pueden ordenar o medir en forma significativa, sólo se pueden clasificar y enumerar.

**Ejercicio 2.2.1** Clasificar las siguientes variables como cuantitativas, cualitativas, continuas o discretas según el caso: color de los ojos, recuento de insectos, número de errores por estudiante en un examen de deletreo, kilómetros recorridos por llanta hasta el primer pinchazo, tiempo para recargar de tinta un estilógrafo que se usa normalmente, posibles rendimientos de maíz en un campo determinado, número de niños nacidos en el hospital más cercano en el día de año nuevo, posibles resultados al lanzar 50 monedas, número de peces en un estanque.

**Ejercicio 2.2.2** Al lanzar 10 veces 7 monedas, el número de caras fue de 2, 6, 2, 2, 5, 3, 5, 3, 3, 4. Si se denotan las observaciones  $Y_1, Y_2, \dots, Y_n$ , ¿cuál es el valor de  $n$ ? ¿Cuál es el valor de  $Y_2$ ? ¿de  $Y_5$ ? ¿Para qué valores de  $i$  es  $Y_i$  igual a 2, 3 y 4? Distinguir entre  $Y_{i-1}$  e  $Y_i - 1$ . Cuando  $i = 2$ , ¿qué valores toman  $Y_{i-1}$  e  $Y_i - 1$ ?

### 2.3 Distribuciones

Los valores de una variable sirven para describir o clasificar individuos o distinguir entre ellos. La mayoría de nosotros hacemos algo más que simplemente describir, clasificar o distinguir, porque tenemos ideas respecto a las *frecuencias relativas* de los valores de una variable. Así, en Minnesota, el ser rubio no es valor raro del cutis de una persona; la mayoría de la gente no creería mucho en historias sobre un gato doméstico de 65 libras; un bebé de peso de  $7\frac{1}{2}$  libras al nacer se considera común y corriente con excepción de la familia. La mente asocia una medida con el valor de una variable, una medida de lo corriente que es o no, a la probabilidad de ocurrencia de un valor semejante. En estadística decimos que la variable tiene una *función de probabilidad*, una *función de densidad de* ↗

*probabilidad* o simplemente *una función de densidad*. Así, para una moneda equilibrada o justa, la probabilidad de que caiga cara es la misma de que caiga sello, es decir,  $1/2$ ; ésta es una afirmación con respecto a la función de densidad de una variable discreta. La afirmación de que cierto porcentaje del peso de adultos es menor que un valor dado, corresponde a una *distribución de probabilidad acumulada* o, simplemente, a una *función de distribución* de una variable continua cuando tenemos los porcentajes de todos y cada uno de los pesos. Las expresiones "al azar" y "variable aleatoria" se usan más particularmente para variables que poseen funciones de densidad de probabilidad.

Tan pronto los términos muy relacionados de función de densidad y función de distribución estén bien definidos, comenzaremos a usar simplemente el término *distribución* para significar uno y otro, según convenga.

La noción de azar o aleatoriedad no ha sido definida. Sólo hemos dado a entender que las leyes del azar son aplicables. La aleatoriedad se estudia más extensamente en la siguiente sección.

**Ejercicio 2.3.1** Con base en su experiencia, clasifique la ocurrencia de los siguientes sucesos con frecuencias relativas, alta, media o baja, según el caso: Un bebé de 2 libras; un jugador de baloncesto de 6 pies y 8 pulgadas; una temperatura nocturna baja de  $0^{\circ}\text{C}$  por lo menos una vez en octubre en su localidad; una estudiante de primer año de universidad de 112 libras; una serie de 5 caras al lanzar 5 veces una moneda; un caballo de 3400 libras; 27 caras al lanzar una moneda 50 veces.

**Ejercicio 2.3.2** Si se lanza un clavo, ¿caerá de cabeza tantas veces como de punta? De acuerdo con su experiencia, ¿es probable que la señal de ocupado en un teléfono ocurra tan frecuentemente como 1 vez en 10 llamadas a 10 personas diferentes? ¿Qué porcentaje de estudiantes de primer año de universidad espera usted que pasen al curso siguiente el próximo año? ¿Qué porcentaje completará todo un programa en la misma universidad?

(Luego de reflexionar con respecto al material de los ejercicios 2.3.1 y 2.3.2, se caerá en cuenta de que la idea de las distribuciones, de sus promedios y su variabilidad, no es enteramente nueva).

## 2.4 Poblaciones y muestras

La primera preocupación respecto a un conjunto de datos es si se puede considerar como todos los datos posibles o sólo una parte de un conjunto más grande. Esto es de gran importancia, y el no hacer una distinción clara ha producido errores en la forma de pensar y una explicación ambigua en algunos escritos.

Una *población* o *universo* consiste en todos los posibles valores de una variable. Estos valores no tienen que ser todos diferentes ni en número finito. Son ejemplos los pesos de una camada de cerditos al nacer, el número de caras al lanzar 500 veces 10 monedas, todos los posibles valores de rendimiento de maíz por acre en el estado de Iowa. La variable puede ser continua o discreta, observable o no observable. Cuando se conocen todos los valores de una población, es posible describirla sin ambigüedad.

Una *muestra* es una parte de una población. (En algunos casos, una muestra puede incluir la población entera). Por lo general, se trata de usar la información de muestra para hacer inferencias acerca de una población. Por esta razón es particularmente importante definir la población que se estudia y obtener una muestra representativa de la población definida, lo cual no es cosa trivial.

Los pesos de los cerditos de una camada al nacer y el número de caras en cada uno de los lanzamientos de 10 monedas, bien pueden ser muestras de poblaciones indefinidamente numerosas y usualmente tiene más valor considerarlas como muestras que como poblaciones.

Una muestra debe ser representativa de la población si tiene como fin obtener inferencias válidas. Para obtener una muestra representativa, el principio de aleatoriedad se incorpora a las reglas para obtener la muestra. La *aleatoriedad* es el resultado de un proceso mecánico para asegurar que los sesgos individuales, conocidos o desconocidos en su naturaleza, no influyan en la selección de las observaciones de la muestra. En consecuencia, se aplican las leyes de la probabilidad y se usan para extraer inferencias. Al efectuar una encuesta de opinión pública, las conclusiones que se intenta aplicar a la población adulta de los Estados Unidos, rara vez serían válidas si la muestra fuera de tal modo no aleatoria que sólo incluyesen mujeres o solamente ciudadanos de Nueva Inglaterra. En este texto, la palabra muestra implicará muestra aleatoria. A medida que avancemos se darán ilustraciones del procedimiento de aleatorización.

**Ejercicio 2.4.1** ¿Sería objetable considerar las siguientes muestras como provenientes de poblaciones posiblemente infinitas: una muestra de los pesos de 100 arenques de los Grandes Bancos? Una muestra de 200 familias de Madison, Wisconsin? Una muestra de las longitudes de los cuerpos de 20 orcas? Una muestra de 10 cc de sangre de una persona adulta? Explique sus respuestas.

**Ejercicios 2.4.2** ¿Serían muestras aleatorias las siguientes: las truchas pescadas en un día en un lago de tamaño moderado? Las ardillas capturadas en un día en una trampa? Las respuestas escritas a un pronunciamiento político solicitadas en un anuncio de televisión? Una muestra autorizada de un botánico de la vegetación de un campo? Explique sus respuestas.

## 2.5 Muestras aleatorias: Recolección de datos

Ha sido ampliamente demostrado que no se puede tomar una muestra aleatoria sin emplear un proceso mecánico. En el proceso usado para obtener una muestra aleatoria o para introducir la aleatoriedad en un experimento o encuesta, por lo general interviene una tabla de números aleatorios, como la tabla A.1. Esta tabla está formada por los dígitos 0, 1, 2, 3, 4, 5, 6, 7, 8 y 9 distribuidos en una tabla de 100 por 100, dando lugar a 10,000 dígitos aleatorios. Estos números se obtuvieron en una máquina y no hay razón para pensar que algún número apareciera con más frecuencia que otro, ni que alguna sucesión de números fuese más frecuente que otra, excepto por el azar. Hay 1,015 ceros, 1,026 unos, 1,013 doces, 975 tréces, 976 cuatros, 932 cinco, 1,067 seis, 1,013 sietes, 1,023 ochos, 960 nueves; 5,094 son pares y 4,906 son impares. Ilustremos el uso de la tabla tomando una muestra aleatoria de 10 observaciones de la tabla 4.1. Los datos de la tabla 4.1 se han clasificado de acuerdo con la magnitud asignándoles números de orden. La organización por orden no es necesaria para extraer muestras al azar; el orden de los números aleatorios pudo haberse asignado en forma arbitraria.

Para obtener una muestra aleatoria de 10 pesos, tómense 20 dígitos consecutivos de la tabla A.1 y regístrense como 10 pares. Estos serán los números de orden de los pesos correspondientes. Se puede comenzar en cualquier parte de la tabla, pero una forma más satisfactoria es señalar con el dedo en una de las páginas, leer los cuatro números opuestos más cercanos a la punta del dedo y utilizar éstos para localizar el punto de partida. Así:

1. En la primera página de la tabla A.1, el dedo encuentra el número 1188 (frente a 10 y son los primeros cuatro dígitos en la columna 20-24).
2. Se va a la fila 11, columna 88, como punto de partida.
3. Se registran en pares los 20 dígitos que se encuentran yendo hacia la derecha, y que son 06, 17, 22, 84, 44 y 55; por comodidad, se baja una línea y se procede al revés para obtener los otros números, o sea, 09, 15, 30 y 59.
4. Se toman los números de los elementos y se llevan a la tabla 4.1 para obtener las correspondientes observaciones: 20, 30, 32, 51, 39, 41, 25, 29, 35 y 42 libras.

Este es un procedimiento aleatorio que equivale a extraer de una bolsa con 100 frijoles marcados con 100 contenidos de grasa de leche, volviendo cada frijol a la bolsa y mezclando bien los frijoles antes de cada extracción. Por esta razón, se dice que el muestreo es con reemplazo. Nótese que cada elemento puede sacarse cualquier número de veces desde 0 hasta 10. El muestreo siempre se hace de la misma población y la probabilidad de sacar cada número de orden es prácticamente la misma. Cualquiera de los dos procedimientos da los mismos resultados que si las extracciones se hicieran de una población infinitamente grande.

Una muestra extraída de esta manera es una muestra *completamente aleatoria*. En trabajos experimentales y encuestas por muestras a menudo se tiene razones válidas para restringir en algún grado la aleatoriedad. El uso de las restricciones se discutirá en capítulos posteriores.

**Ejercicio 2.5.1** Supóngase que una población tiene solamente 50 elementos. ¿Cómo se podría usar la tabla de números aleatorios para obtener una muestra de 10 observaciones con reemplazo? ¿Puede sugerirse otro plan?

**Ejercicio 2.5.2** Supóngase que una población tiene 40 elementos. ¿Cómo se podría extraer una muestra de 5 observaciones? Recuérdese que cada elemento debe tener igual oportunidad de ser escogido para la muestra.

**Ejercicio 2.5.3** Supóngase que una población tiene 75 elementos. ¿Cómo se sacaría una muestra de 10 observaciones?

**Ejercicio 2.5.4** Supóngase que una población tiene 40 elementos y que se desea extraer una muestra de 5 observaciones *sin reemplazo*. ¿Cómo se haría?

**Ejercicio 2.5.5** Si se tiene un terreno plantado con maíz en surcos y se desea tomar una muestra de plantas individuales. ¿cómo se haría la selección? ¿Se puede pensar en otro plan?

**Ejercicio 2.5.6** Si se desea muestrear la flora de una ciénaga mediante observación de áreas de una yarda cuadrada. ¿cómo se seleccionaría su muestra?

**Ejercicio 2.5.7** ¿Cómo se extraería, por ejemplo, una muestra completamente aleatoria de 100 números telefónicos de un directorio telefónico? ¿Podría usarse un plan de dos etapas que implicara menor esfuerzo?

**Ejercicio 2.5.8** Al extraer una muestra de 100 de una población grande con igual número de hombres que de mujeres, ¿sería recomendable una muestra completamente aleatoria o tomar una de 50 hombres y 50 mujeres? (¿Cuáles son los objetivos del muestreo?).

**Ejercicio 2.5.9** Al describir el proceso de extracción de una muestra aleatoria utilizando la tabla A.1. se dijo que la "probabilidad de extraer un elemento cualquiera es prácticamente la misma". ¿Por qué se usa la palabra "prácticamente"?

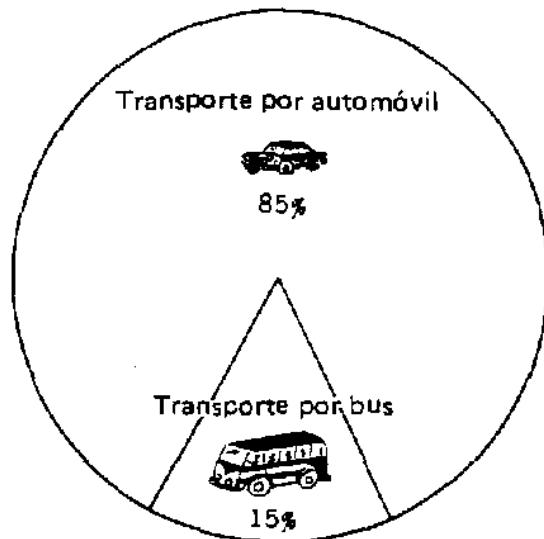


Figura 2.1 Automóvil y bus.  
Pueblonuevo 17,000 habitantes

## 2.6 Presentación, resumen y caracterización de la información

Existen muchas formas de presentar datos, entre ellas el uso de tablas, diagramas y gráficas.

Para datos cualitativos, la enumeración es una forma común de resumir y presentar los resultados. Si un pueblo pequeño realiza una encuesta sobre las formas de transporte, los resultados pueden resumirse y presentarse como porcentajes. Para visualizar la situación, es útil el siguiente diagrama de sectores (Fig. 2.1).

Representaciones tales como los diagramas de sectores y de barras presentan la información respecto a la forma como se gasta el tiempo y el dinero, y a dónde van los impuestos. Son concisos, informativos, fáciles de leer y comprender, y a menudo transmiten la información en forma precisa. Ciertamente son más atractivas que las tablas de frecuencia para presentar el número de personas, de insectos, que poseen ciertas características. Desafortunadamente, pueden no ser confiables. Por ejemplo, la fig. 2.1 puede estar basada en una muestra o en un censo con respecto a millas recorridas o a número de viajes. Son evidentes las posibilidades de engañarse.

Los diagramas pueden ser números o porcentajes efectivos. La fig. 2.2 ilustra el uso de barras o rectas verticales. Si no se indica la escala, el lector ve frecuencias relativas y no

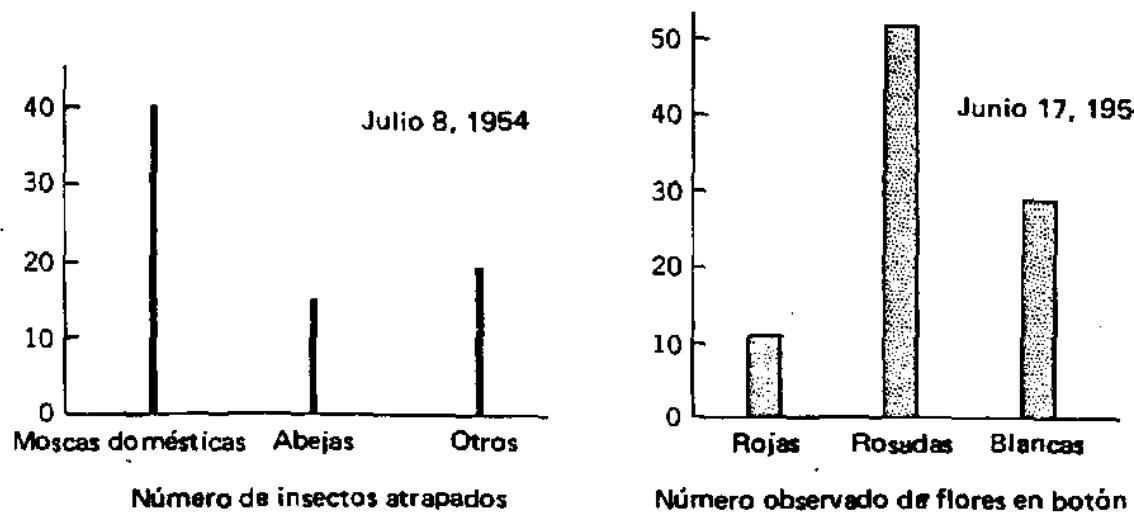


Figura 2.2 Presentación de datos discretos

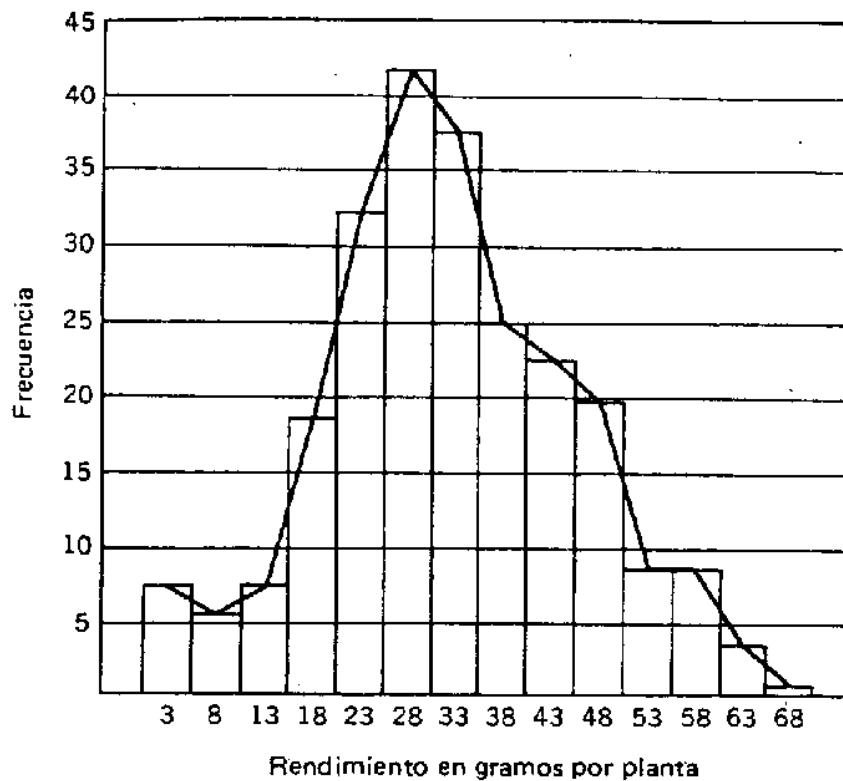


Figura 2.3 Histograma y polígono de datos para los datos de la tabla 2.1.

se dará cuenta del tamaño de la muestra a menos que se le especifique. Cuando se dan los totales, se puede disponer tanto de las frecuencias como de las frecuencias relativas y pueden presentarse mediante el uso de dos escalas. Las frecuencias relativas o proporciones son probabilidades esencialmente, como puede notarse considerando las barras o las columnas como alturas o áreas.

El histograma y el polígono de frecuencia de la fig. 2.3 son métodos usuales para presentar una cantidad considerable de información, recolectada como datos cuantitativos. El *histograma* representa gráficamente los datos con *valores de clase*, los puntos medios de los intervalos de clase, a lo largo del eje de las X y con rectángulos sobre los intervalos de clase para representar la frecuencia. El histograma presenta los datos en una forma fácil de entender de tal manera que de una ojeada se ve la naturaleza general de la distribución. Si se desea comparar una distribución observada con una distribución teórica, se puede superponer la distribución teórica sobre el histograma y apreciar las discrepancias.

El *polígono de frecuencia* se construye localizando el punto medio de cada intervalo de clase y marcando un punto a la altura de la frecuencia correspondiente al intervalo. Estos puntos se unen luego con líneas rectas. El polígono de frecuencias tiende a sugerir a la curva suave de la población de donde se extrajo la muestra. El histograma y el polígono de frecuencia de los datos de la tabla 2.1 se presentan en la fig. 2.3.

Tabla 2.1 Tabla de frecuencia de los rendimientos de 229 plantas de soya espaciadas de Richland

Rendimiento en gr	3 8 13 18 23 28 33 38 43 48 53 58 63 68
Frecuencia	7 5 7 18 32 41 37 25 22 19 6 6 3 1

**Tabla 2.2 Número de caras**

Valor de clase	Frecuencia
5	4
4	15
3	29
2	30
1	17
0	5
Total	100

Es importante, tanto en la elaboración del histograma como del polígono de frecuencia, que el número de clases sea suficientemente grande para que la forma general de la distribución se pueda apreciar fácilmente, pero no tan pequeño que se den demasiados detalles.

Finalmente, los datos se presentan en *tablas de frecuencia*, en las que generalmente hay mayor información para el lector más serio, pero a costa de tener menos número de lectores. Las tablas 2.1 y 2.2 son modelos para datos continuos y discretos. Las tablas de frecuencia se verán más en detalle al comenzar la sec. 2.15.

Los diagramas y gráficas a menudo resumen y caracterizan los datos, pero esto también puede hacerse mediante la presentación de varios números que resuman y caractericen los datos. En particular, nos referimos a un número que sitúe el centro y otro que mida la dispersión de las observaciones.

Con frecuencia, se necesita alguna *medida de posición* o de *tendencia central*, esto es, un número que localice un valor central u ordinario. Además, es preciso tener una *medida del espaciamiento* o de *dispersión* para saber cuán alejados se encuentran los valores más o menos extremos respecto del valor central.

**Ejercicio 2.6.1** En la "La Encuesta Nacional de Caza, Pesca y Vida Silvestre en 1975" realizada por el Servicio de Pesca y Vida Silvestre de los Estados Unidos, se encuestaron más de 2.000 familias por teléfono y se enviaron cuestionarios por lo menos a 1.000 cazadores y pescadores en cada estado. Toda persona de 9 o más años que cazó o pescó al menos un día en 1975 era aceptable para participar en la encuesta postal. De esas personas de 9 años o más, 95.9 millones participaron en alguna actividad relacionada con la vida silvestre. Entre los datos reportados estuvieron los siguientes (a menudo aproximados a partir de una gráfica):

1. 53 millones participaron en pesca; 50 millones en observación de la vida silvestre; 26 millones en recolección de calamares, cangrejos y otros; 21 millones en caza, 17 millones en tiro recreacional; 15 millones en fotografía de vida salvaje y 5 millones en disparo con arco.
2. En miles de millones de días, los estadounidenses del grupo 1 gastaron el siguiente tiempo en las mismas categorías respectivamente así: 1.32, 1.54, 0.24, 0.49, 0.30, 0.19 y 0.12.
3. La participación de hombres y mujeres, en las mismas categorías, fue respectivamente: 57% y 43%, 48% y 52%, 82% y 18%, 58% y 42% y 81% y 19%.
4. El promedio del ingreso en dólares por familia del pescador con caña era: menos de 5,000, 17%; entre 5,000-9,999, 17%; entre 10,000-14,999, 23%; entre 15,000-24,999, 29% y 25,000 o más, 14%.

5. El número, en millones, de pescadores en aguas templadas por tipo de agua, fue: 18 en lagos y embalses, 17 en lagos y lagunas distintos de los Grandes Lagos, 14 en arroyos y ríos, 9 en estanques de granjas, 3 en desembocaduras y 2 en los Grandes Lagos.
6. El número de días, en miles, de participación en pesca en aguas templadas por tipo de agua tal como se especificó en el numeral 5 fue: 279,287, 254,477, 161,427, 93,372, 24,973 y 19,011.
7. El número, en millones, de pescadores en ríos por tipo de agua fue: 2.6 en aguas saladas, 2.5 en corrientes y ríos, 2.0 en estuarios y 1.1 en los Grandes Lagos.
8. El número en miles de días de participación en pesca en ríos por tipo de agua especificada en el numeral 7 fue, respectivamente: 19,478, 18,606, 16,039 y 6,739.
9. De los cazadores de animales grandes, que se cazan en el estado, 58.1% cazaron en propiedad privada, 15.9% en tierras federales, 12.5% en otras zonas silvestres administradas por el estado, 2% en otras zonas de los estados, 8% en tierras públicas no especificadas y 3.5% en tierras de propietario desconocido.
10. De los cazadores de caza mayor que cazan fuera del estado en los Estados Unidos, las categorías del numeral 9 dieron los siguientes porcentajes, respectivamente: 37.1%, 38.6%, 16.6%, 1.7%, 3.8% y 2.2% -

Presentar cada uno de estos 10 conjuntos de datos en forma tabular o gráfica. Trate de variar la presentación.

**Ejercicio 2.6.2** Buscar en una revista técnica o científica nuevas ideas sobre presentación de datos. ¿Está claro, sin hacer referencia al texto, si se trata de una población o de una muestra? ¿En la presentación entran frecuencias relativas? ¿Es completamente comprensible sin hacer referencia al texto?

**Ejercicio 2.6.3** Comparar un ejemplar del informe anual de una compañía con el de una revista técnica. ¿En cuál encuentra más gráficos relativamente? ¿Más tablas? ¿Siempre tienen los gráficos una escala en el eje vertical?

**Ejercicio 2.6.4** Elaborar una tabla de frecuencias, otra de frecuencias relativas y un diagrama de barras para los números enteros en la tabla A.1, ver sec. 2.5

**Ejercicio 2.6.5** Elaborar una tabla de frecuencias, un polígono de frecuencias y un histograma con los datos de la tabla 4.1.

## 2.7 Medidas de tendencia central

Expresiones tales como "estatura media" son vagas pero informativas. Relacionan un individuo con un valor central. Cuando los experimentadores recolectan datos, gastando tiempo, energía y dinero, no pueden darse el lujo de presentar información vaga, necesitan una medida definida de tendencia central.

La medida de tendencia central más común, que a la vez es la mejor en muchos casos, es la *media aritmética* o *promedio aritmético*. Como hay otros tipos de medias, debe quedar claro de qué tipo de media se trata. La media aritmética se representará mediante los símbolos  $\mu$  (la letra griega mu) para la de una población y  $\bar{Y}$  (léase Y barra) para la de una muestra. Es importante hacer la anterior distinción, ya que la media de la población es una cantidad fija, mientras que la media de la muestra es variable, puesto que diferentes muestras extraídas de la misma población tienden a tener diferentes medias. En los escritos estadísticos es posible encontrar otros símbolos. La media se da en las mismas unidades que los datos originales, por ejemplo, centímetros o libras.

Cantidades tales como la media se llaman *parámetros* cuando caracterizan poblaciones, y *estadígrafos*, en el caso de muestras.

Considérese un dado. Sus seis caras tienen uno, dos, tres, cuatro, cinco y seis puntos. Todo el número posible de puntos que aparecen al lanzar el dado constituyen una población finita. Por definición, el parámetro

$$\mu = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3\frac{1}{2} \text{ puntos}$$

Si una muestra de esta población tiene cuatro observaciones, digamos 3, 5, 2 y 2, por definición, el estadígrafo

$$\bar{Y} = \frac{3 + 5 + 2 + 2}{4} = \frac{12}{4} = 3 \text{ puntos}$$

Estos cálculos pueden simbolizarse mediante

$$\bar{Y} = \frac{Y_1 + Y_2 + Y_3 + Y_4}{4}$$

donde  $Y_1$  es el valor de la primera observación, esto es, 3;  $Y_2$  es el valor de la segunda observación, y así sucesivamente. En la situación general con  $n$  observaciones,  $Y_i$  se usa para representar la observación  $i$ -ésima y  $\bar{Y}$  está dada por

$$\bar{Y} = \frac{Y_1 + Y_2 + Y_3 + \cdots + Y_i + \cdots + Y_n}{n}$$

La notación  $\bar{Y}$ , puede abreviarse más

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} \quad o \quad \frac{\sum_i Y_i}{n} \quad (2.1)$$

Esta es una frase con un sujeto  $\bar{Y}$ , un verbo  $=$ , y un objeto  $\sum$  (la letra griega sigma, mayúscula). La frase se lee "Y barra es igual a la suma de las  $Y$  dividida por  $n$ ". Esta es nuestra definición de la media de la muestra. La letra  $i$ , usada para indicar el individuo  $i$ -ésimo, es llamada índice de la sumatoria, y va desde  $i = 1$ , escrito bajo el signo de sumatoria  $\sum$ , hasta  $n$ , escrito en la parte superior. El intervalo de la sumatoria va de 1 hasta  $n$ . Cuando se van a sumar todos los valores, 1 y  $n$  se omiten por lo general.

Se definen como desviaciones de la muestra las diferencias entre las observaciones y la media con sus respectivos signos,  $Y_i - \bar{Y}$ . Para nuestro ejemplo,  $Y_1 - \bar{Y} = 3 - 3 = 0$ ,  $Y_2 - \bar{Y} = 2$ ,  $Y_3 - \bar{Y} = -1 = Y_4 - \bar{Y}$ .

Una propiedad interesante de la media aritmética es que la suma de las desviaciones es cero, o sea,

$$\sum (Y_i - \bar{Y}) = 0 \quad (2.2)$$

Por ejemplo,

$Y_i$	$\bar{Y}$	$Y_i - \bar{Y}$
3	3	0
5	3	+2
2	3	-1
2	3	-1

Esta es una ilustración, no una demostración.

Algunas veces es apropiado ponderar las observaciones que entran en una media. Por ejemplo, si se deben promediar diferentes medias que provienen de muestras con diferente número de observaciones, entonces es conveniente usar ponderaciones que dependen del número de observaciones de cada promedio. Una *media ponderada* se define por

$$\bar{Y}_w = \frac{\sum w_i Y_i}{\sum w_i} \quad (2.3)$$

Otra medida de tendencia central es la *mediana*. También puede usarse para complementar la media. La mediana es el valor en torno al cual quedan de cada lado el 50% de las observaciones, cuando se disponen en orden de magnitud. Si el número de observaciones es par, la mediana es el promedio de dos valores centrales. Por ejemplo, para la muestra 3, 6, 8 y 11, la mediana es  $(6 + 8)/2 = 7$ . Los números 3, 6, 8 y 30 tienen la misma mediana 7. Para datos distribuidos en forma más o menos simétrica, la media y la mediana difieren muy levemente. Sin embargo, cuando se necesita encontrar el promedio del ingreso de un grupo en el que la mayoría tiene ingresos bajos, el ingreso medio puede ser considerablemente mayor que la mediana, e induciría a confusión.

Ciertos tipos de datos muestran una tendencia a concentrarse o formar una cola en los extremos derecho o izquierdo. Tales distribuciones se dice que son *asimétricas* en la dirección de la cola larga, y la media aritmética puede no ser el valor central más informativo.

Otra media de tendencia central es la *moda*, el valor de más frecuente ocurrencia.

Otras medidas de tendencia central son promedios de ciertos *cuartiles*, *deciles* y *percentiles*, puntos que dividen distribuciones de valores ordenados en rangos, en cuartos, en décimos y centésimos respectivamente. Por ejemplo, 10% de las observaciones son menores que el primer decil. La mediana es el segundo cuartil, quinto decil y percentil 50.

Para números positivos, la media geométrica y la media armónica pueden ser útiles. Sus usos principales son, respectivamente, el cálculo de valores lativos tales como números índices, y en el cálculo de promedios de razones y tasas. Se obtienen mediante las siguientes ecuaciones:

$$\begin{aligned} \text{Media geométrica, } G &= \sqrt[n]{Y_1 Y_2 \cdots Y_n} \\ &= (Y_1 Y_2 \cdots Y_n)^{1/n} \end{aligned} \quad (2.4)$$

$$\text{y Media armónica, } \frac{1}{H} = \frac{1}{n} \sum_i \left( \frac{1}{Y_i} \right) \quad (2.5)$$

**Ejercicio 2.7.1** Un maestro le dijo a las niñas de su clase que estimaren su peso (el del instructor). Sus respuestas fueron: 190, 230, 105, 180, 130, 160 y 170 libras. Calcular la media de la muestra. ¿Cuál es la mediana de la muestra?

**Ejercicio 2.7.2** Los muchachos de la clase del ejercicio 2.7.1 estimaron que el peso del instructor era: 150, 195, 175, 147, 175, 170, 195, 170 y 190 libras. Calcular la media de la muestra. ¿Cuál es la mediana de la muestra?

**Ejercicio 2.7.3** ¿La media aritmética puede considerarse como una media ponderada con pesos iguales? ¿Cuáles son estos pesos?

**Ejercicio 2.7.4** Supóngase que tenemos las siguientes medias,  $\bar{Y}_1 = 37$ ,  $\bar{Y}_2 = 41$  y  $\bar{Y}_3 = 28$ , basadas en 50, 20 y 10 observaciones respectivamente. Si hay que escoger una sola media, ¿cuál sería su elección? ¿Por qué? ¿Cuáles son los totales de las muestras originales? ¿Cómo se usarían estos totales para hallar la media aritmética de las 90 observaciones? ¿Es el mismo proceso usado para calcular la media ponderada con pesos iguales a los tamaños de las muestras?

**Ejercicio 2.7.5** Un método de muestrear peces en un lago consiste en matarlos todos con rotenona, recogerlos en baldes y entonces tomar una muestra al azar de los baldes. En un experimento de éstos, se tomó una muestra de 2 baldes de un total de 20 y se midió la longitud de los peces en pulgadas. Los datos fueron: †

Muestra A: 5 peces de 5 pulgadas, 19 de 6, 19 de 7, 8 de 8 y 3 de 9;  $n = 54$

Muestra B: 10 peces de 5 pulgadas, 27 de 6, 15 de 7, 6 de 8 y 3 de 9;  $n = 61$

Para cada muestra, calcular la media. ¿Cuál es la clase modal para cada muestra? ¿Se calculó la media usando 54 (61) observaciones individuales, o como media ponderada? ¿Cuántos valores diferentes de la variable se usan cuando se calcula la media ponderada? ¿Se utilizan las dos medias muestrales para calcular la media ponderada? ¿Por qué se tomaron los pesos utilizados? ¿La media ponderada es la misma que la media aritmética de todas las 115 observaciones?

**Ejercicio 2.7.6** Para la muestra del ejercicio 2.2.2, calcular  $\bar{Y}$  y todas las  $Y_i - \bar{Y}$ s. ¿La  $\sum (Y_i - \bar{Y})$  es igual a cero?

**Ejercicio 2.7.7** ¿Cuál es la mediana de las observaciones de la tabla 2.1? ¿De la tabla 2.2? ¿De la tabla 2.4?

**Ejercicio 2.7.8** Comentar la afirmación: "50% de los estadounidenses tienen una inteligencia por debajo del promedio".

**Ejercicio 2.7.9** Las medias geométricas son útiles al tratar de tasas y razones. Si invierto \$100 y obtengo \$120 al final del año, y reinvierto y obtengo \$144 al final del segundo año, entonces he obtenido el 20% de mi inversión. Claramente la tasa de crecimiento es 1.2. La media geomé-

---

† Datos cortesía de Don W. Hayne, Universidad del Estado de Carolina del Norte.

trica es apropiada y es  $\sqrt{1.2(1.2)} = 1.2$ . ¿Las tasas de crecimiento de la población, para tasas estables de nacimiento y mortalidad, sin migración, dan una situación biológica en la cual la media geométrica es la apropiada? ¿Son los valores constantes o variables? Calcular la media geométrica.

Ejercicio 2.7.10 La producción de crudo en una compañía presentó un incremento anual de 1945 a 1955. En 1945 la producción fue de 3,220 barriles/día y en 1955, 4,780 barriles/día. ¿Cuál es la media geométrica? ¿Para qué año se consideraría esto como una estimación de la producción?

## 2.8 Medidas de dispersión

Una medida de tendencia central sólo proporciona un resumen parcial de la información de un conjunto de datos; es evidente la necesidad de una medida de variación. Se dan a continuación tres conjuntos de datos con una media aritmética común; nótese la diferencia en su variación:

$$\begin{array}{c} 8, 8, 9, 10, 11, 12, 12 \\ 5, 6, 8, 10, 12, 14, 15 \\ 1, 2, 5, 10, 15, 18, 19 \end{array}$$

La media, como otras medidas de tendencia central, no nos dice nada respecto a la variación.

El concepto de medida de dispersión no es fácil. ¿Cuánta información se tiene al decir que los tres conjuntos de datos tienen dispersiones de 4, 10 y 18, o sea,  $Y_7 - Y_1$ ? El segundo grupo de tres conjuntos

$$\begin{array}{c} 8, 9, 10, 10, 10, 11, 12 \\ 5, 7, 9, 10, 11, 13, 15 \\ 1, 5, 8, 10, 12, 15, 19 \end{array}$$

tiene también las dispersiones 4, 10 y 18 y la media también es 10, pero los primeros conjuntos presentan más dispersión en los extremos mientras que en el segundo grupo hay mayor concentración hacia la media. Parece deseable tener una definición que utilice todas las observaciones y de un valor pequeño cuando éstas se encuentran alrededor de la media y un valor grande cuando estén muy dispersas.

Sean los números:

$$5, 6, 8, 10, 12, 14, 15$$

De acuerdo con nuestra definición, estos números no son más variables que los números:

$$105, 106, 108, 110, 112, 114, 115$$

Así, nuestra definición no depende del tamaño de los números en el sentido de relacionar la medida de la dispersión con la media; dará el mismo valor para los dos conjuntos de datos.

La medida numérica de dispersión resultante debe admitir interpretación en términos de las observaciones. Así, la unidad de medida debe ser la misma. Su función deberá servir como una unidad para decidir si una observación es común y corriente o si es un valor no usual para una población dada. La media o un valor hipotético deberán ser el punto de partida para la medición. Por ejemplo, si un hombre está a muchas unidades de dispersión respecto de la media de una población su efigie se puede perpetuar en bronce o bien se le puede recluir, de lo contrario, se trata del hombre de la calle; si un estudiante varón tiene 5 pies y 4 pulgadas de alto, es improbable que pueda pertenecer a una población masculina de jugadores de baloncesto porque en estatura está demasiado distanciado de la media de esa población.

Finalmente, la medida debe poseer propiedades matemáticas, de las cuales no nos ocuparemos por el momento.

La mejor medida de dispersión y la más generalizada es la *varianza* o su raíz cuadrada, la *desviación estándar*. La varianza se representará con dos símbolos:  $\sigma^2$  (la letra griega *sigma*) para la población y  $s^2$  para la muestra. Estas se leen así: Sigma al cuadrado y  $s$  al cuadrado. Las desviaciones estándar de poblaciones y muestras se denotan  $\sigma$  y  $s$  respectivamente.  $\sigma^2$  y  $\sigma$  son parámetros, constantes para una población particular;  $s^2$  y  $s$  son estadígrafos, valores que cambian de muestra a muestra en una misma población. La varianza o *cuadrado medio* se mide en términos de desviaciones al cuadrado:

Supongamos que tenemos una población finita de  $N$  valores, cada uno con la misma probabilidad,  $1/N$ , de ser extraído mediante un proceso aleatorio. Se intenta muestrear esta población con reemplazo, tal como se expuso en la sec. 2.5. En este caso, la varianza poblacional se define como la suma de las desviaciones al cuadrado dividida por el número total. La varianza se mide en las mismas unidades originales pero al cuadrado, por ejemplo, centímetros al cuadrado.

Simbólicamente, la varianza se define mediante

$$\begin{aligned}\sigma^2 &= \frac{(Y_1 - \mu)^2 + (Y_2 - \mu)^2 + \cdots + (Y_N - \mu)^2}{N} \\ &= \frac{\sum_i (Y_i - \mu)^2}{N}\end{aligned}\quad (2.6)$$

(Cuando los valores de la muestra tienen diferente probabilidad de ser tomados, cada desviación al cuadrado se pondera con su probabilidad y entonces no se requiere el divisor  $N$ . Esta idea de ponderar puede extenderse fácilmente a poblaciones infinitas con una variable discreta).

(El símbolo  $s^2$  y la definición con el divisor  $N - 1$  se usa para la varianza muestral cuando el muestreo es sin reemplazo).

Para el ejemplo del dado con una población

$$\begin{aligned}\sigma^2 &= \frac{1}{6} [(1 - 3\frac{1}{2})^2 + (2 - 3\frac{1}{2})^2 + (3 - 3\frac{1}{2})^2 \\ &\quad + (4 - 3\frac{1}{2})^2 + (5 - 3\frac{1}{2})^2 + (6 - 3\frac{1}{2})^2]\end{aligned}$$

Nótese que es posible lanzar un dado más de seis veces. En este caso  $n$  es mayor que  $N$ .

Para variables continuas, la varianza de la población exige matemáticas más refinadas que las que se usan en este texto.

La varianza de la muestra se define en forma similar pero el divisor es  $n - 1$ .

$$\begin{aligned}s^2 &= \frac{(Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2}{n - 1} \\ &= \frac{\sum_i (Y_i - \bar{Y})^2}{n - 1}\end{aligned}\quad (2.7)$$

Nótese

$$(n - 1)s^2 = \sum_i (Y_i - \bar{Y})^2$$

El numerador de  $s^2$  se conoce como la *suma de cuadrados* y a menudo se denota SC. Para los números 3, 6, 8 y 11 la suma de cuadrados es

$$\begin{aligned}(3 - 7)^2 + (6 - 7)^2 + (8 - 7)^2 + (11 - 7)^2 \\ = (-4)^2 + (-1)^2 + (+1)^2 + (+4)^2 = 34,\end{aligned}$$

y la varianza es  $34/3 = 11.33$ .

La raíz cuadrada de la varianza de la muestra se llama *desviación estándar*, y se denota  $s$ . *Desviación media cuadrática* se usa menos, pero es término descriptivo de  $s$ . Para nuestro ejemplo,  $s = \sqrt{34/3} = 3.4$  unidades de las observaciones originales.

La cantidad

$$SC = \sum_i (Y_i - \bar{Y})^2$$

puede llamarse *fórmula de definición* de la suma de cuadrados, pues nos dice que la suma de cuadrados es la suma de los cuadrados de las desviaciones respecto del promedio aritmético.

La fórmula de definición de la suma de cuadrados se reduce a una fórmula de trabajo para los cálculos, es decir:

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i Y_i^2 - \frac{\left(\sum_i Y_i\right)^2}{n} \quad (2.8)$$

En el primer término los valores se elevan al cuadrado y luego se suman, en el segundo término se suman todos los valores y luego se eleva al cuadrado. En muchos computadores,

$$\sum_i Y_i^2 \quad \text{y} \quad \sum_i Y_i$$

puede obtenerse simultáneamente. La cantidad

$$\frac{\left(\sum_i Y_i\right)^2}{n}$$

se llama *término de corrección* o *factor de corrección* o *ajuste de la media* y se representa por  $C$ . El término "corrección de la media" indica que como la SC es una medida de variación con respecto a  $\bar{Y}$ , el término de corrección debe restarse de

$$\sum_i Y_i^2$$

el cual se denomina a menudo sumas de cuadrados sin ajustar. Como corrección, tal cantidad no tiene nada que ver con equivocaciones. La forma  $n\bar{Y}^2$ , que aparece en la ec. (2.9), es menos conveniente para el cálculo ya que introduce la necesidad de redondear en una etapa anterior.

$$\frac{\left(\sum_i Y_i\right)^2}{n} = n\bar{Y}^2 \quad (2.9)$$

La validez de la ec. (2.8) puede demostrarse mediante un ejemplo numérico. Para nuestra muestra ilustrativa

$Y_i$	$Y_i^2$	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
3	9	-4	16
6	36	-1	1
8	64	+1	1
11	121	+4	16
$\sum: 28$		0	34
$\bar{Y} = 7$			

Así:

$$SC = \sum_i (Y_i - \bar{Y})^2 = 34$$

por la fórmula de definición, y

$$SC = \sum_i Y_i^2 - \frac{\left(\sum_i Y_i\right)^2}{n} = 230 - \frac{28^2}{4} = 34$$

por la fórmula práctica. Cuando es necesario redondear valores, tales como la media o los desvíos,  $(Y_i - \bar{Y})$  se presentan pequeñas discrepancias. Es preferible usar la fórmula de trabajo dado que probablemente se vea menos afectada por los errores de redondeo. La ec. (2.9) también puede comprobarse con un ejemplo.

Una propiedad importante de la suma de cuadrados es la de que es un mínimo, es decir, que si se remplaza  $\bar{Y}$  por cualquier otro valor, la suma de los cuadrados de las nuevas desviaciones será un valor mayor. No es factible demostrar esto para todos los valores posibles.

La cantidad  $(n - 1)$  se conoce como *grados de libertad*, denotado por  $gl$  o  $l$ .

Otra medida de dispersión es la *amplitud* o sea la diferencia entre el valor más alto y el más bajo. Ordinariamente, esta medida no es un estadígrafo satisfactorio para una evaluación crítica de los datos, ya que se ve afectada por valores extremos o no usuales. Sin embargo, para muestras de tamaño 2, es un múltiplo de la desviación estándar; y para muestras menores de 10 es a menudo satisfactoria. En ciertas condiciones, algunas técnicas en las que interviene la amplitud son especialmente deseables.

Una medida de variación de carácter intuitivo, porque tiene en cuenta todas las observaciones, es la media de los *valores absolutos*, o como frecuentemente se denomina *desviación media*. Se calcula así

$$\frac{\sum_i |Y_i - \bar{Y}|}{n} \quad (2.10)$$

Las barras verticales nos dicen que tomemos todas las desviaciones como positivas. Para los valores 3, 6, 8 y 11 la desviación media es 2.5.

Otras medidas de dispersión utilizan percentiles. Así la diferencia entre los puntos que se separan el 85 y el 15 por ciento de las observaciones ordenadas tiene su interés; no depende de los valores extremos como pasa con la amplitud.

**Ejercicio 2.8.1** Al considerar una moneda, podemos asignar el valor de 1 a la cara y 0 al sello. Para una moneda equilibrada, estos valores ocurren con igual probabilidad cuando se la lanza. Así, tenemos una población finita con  $N = 2$  valores. Se lanza la moneda en forma repetida de tal manera que podemos tener una muestra tan grande como queremos. Este muestreo es con remplazo. ¿Cuál es la varianza de esta población?

**Ejercicio 2.8.2** Cuando se lanza una moneda dos veces, los eventos son (C,C), (C,S), (S,C) y (S,S). Ocurren con igual frecuencia, con una moneda de verdad.

Supóngase que sumamos el número de caras para obtener 2, 1, 1 y 0 y considérense estos valores como una población. ¿Cuál es la varianza de esta población?

Si decimos que tenemos una variable aleatoria con los valores 2, 1 y 0 pero con probabilidades  $1/4$ ,  $1/2$  y  $1/4$ , respectivamente, ¿con sólo esta información se puede calcular la varianza poblacional?

**Ejercicio 2.8.3** Con los datos del ejercicio 2.7.1, calcular  $s^2$  y  $s$ . Hallar la amplitud. Cuando se multiplica la amplitud por 0.370 ( $n = 7$ ) el resultado es un estimativo no sesgado de  $\sigma$ . Hágase esto y compárese con  $s$ .

**Ejercicio 2.8.4** Para los datos del ejercicio 2.7.2, calcule  $s^2$  y  $s$ . Encuentre la amplitud. Multiplique la amplitud por 0.337 (para  $n = 9$ ) para obtener un estimativo de  $\sigma$ . Compárela con  $s$ .

**Ejercicio 2.8.5** Con los datos de la muestra A del ejercicio 2.7.5 calcular  $s^2$  y  $s$ . Repita esto para la muestra B.

**Ejercicio 2.8.6** Tómense dos números cualesquiera. Considérelos como una muestra y calcule  $s^2$ . Tómese los mismos dos números, elevese al cuadrado su diferencia y divídase por dos. Los dos valores que acaban de calcularse deberán ser idénticos; este resultado está basado en una identidad algebraica.

Leer ahora, cuanto se ha dicho en esta sección respecto a la amplitud como una medida de dispersión cuando  $n = 2$ .

**Ejercicio 2.8.7** Obtenga un pequeño conjunto de datos en el campo que le sea más familiar, o use un conjunto de datos de uno de los ejercicios precedentes y encuentre  $\bar{Y}$  y  $s^2$ . Compruebe las ecs. (2.8) y (2.9) con esos datos. Calcule la desviación promedio mediante la ec. (2.10).

## 2.9 Desviación estándar de las medias

Se ha estado estudiando el muestreo de poblaciones y la caracterización de las muestras. El lector puede haber estado pensando en características tales como estatura, peso, con plantas y animales que dan lugar a poblaciones de interés. También debe recordarse que las medias y las desviaciones estándar de las muestras están sujetas en sí mismas a variación y forman poblaciones de medias de muestras y de desviaciones estándar de muestras.

En el ejercicio 2.7.4, se pidió escoger la mejor media de tres con base en 50, 20 y 10 observaciones. Presumiblemente se escogió la media basada en  $n = 50$ , pero necesitamos algún criterio para hacer la elección. La variabilidad es una elección obvia.

A propósito de observaciones o intuición, se espera que las medias de las muestras sean menos variables que las simples observaciones. En otras palabras, las medias tienden a acumularse más cerca de un valor central que las observaciones simples. Si tomamos dos series de medias, cada una basada en diferente número de observaciones, 10 y 20, por ejemplo, esperamos que la variación entre medias de muestras pequeñas sea mayor que la variación entre muestras grandes. Afortunadamente, existe una relación conocida entre la varianza entre individuos y la varianza entre medias de individuos. Esta relación y la correspondiente para las desviaciones estándar son

$$\sigma_{\bar{Y}}^2 = \frac{\sigma^2}{n} \quad \text{y} \quad \sigma_{\bar{Y}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \quad (2.11)$$

donde  $\sigma_{\bar{Y}}^2$  es la varianza de la población de las  $\bar{Y}$  obtenidas mediante muestreo de una población original de individuos con varianza  $\sigma^2$ . Para valores de muestras se usa la misma relación, o sea

$$s_{\bar{Y}}^2 = \frac{s^2}{n} \quad \text{y} \quad s_{\bar{Y}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}} \quad (2.12)$$

Estas relaciones se ilustran mediante muestreo en el capítulo 4; son válidas para todas las poblaciones.

La necesidad del subíndice  $Y$  es clara. El subíndice  $Y$  también se usa a veces.  $\mu_Y$  es la media de la población de individuos y  $s_Y^2$  es una varianza calculada a partir de una muestra de individuos.

La utilidad de estas relaciones es obvia. Dada  $\sigma^2$ , podemos calcular directamente la varianza de una población de medias de muestras para cualquier tamaño de muestra. Así mismo, de una sola muestra que da una sola  $\bar{Y}$ , podemos encontrar una varianza muestral que estime la varianza de la población de las  $\bar{Y}$ .

La desviación estándar de una media se suele llamar *error estándar* y, menos frecuentemente, *error estándar de una media*. O sea que el término desviación estándar se aplica a observaciones y error estándar se aplica a medias, a menos que se especifique otra cosa. El error estándar es, pues, inversamente proporcional a la raíz cuadrada del número de observaciones en la media. Puede calcularse si se conoce  $s$  o  $s^2$ ; no se requiere más de una  $\bar{Y}$ . Para las observaciones 3, 6, 8 y 11,  $s_Y = s/\sqrt{n} = 3.4/\sqrt{4} = 1.7$ . Por comodidad en los cálculos,  $s_Y$  se calcula usualmente así:

$$\sqrt{\frac{s^2}{n}} = \sqrt{\frac{11.33}{4}} = \sqrt{2.8333} = 1.7$$

Este es un estimativo de  $\sigma^2/n$ , la varianza de la población formada por las medias de muestras tomadas de la población de individuos. Si hubiéramos obtenido varias medias muestrales de cuatro observaciones y si las hubiésemos usado para calcular la varianza de las medias, habríamos obtenido un estimativo de la misma cantidad. Nosotros tenemos una media simple.

**Ejercicio 2.9.1** Con los datos de los ejercicios 2.7.1 y 2.7.2, también utilizados en los ejercicios 2.8.3 y 2.8.4, calcular  $s_Y^2$  y  $s_Y$ .

**Ejercicio 2.9.2** Con las muestras A y B del ejercicio 2.7.5 y usadas nuevamente en el ejercicio 2.8.5, calcular  $s_Y^2$  y  $s_Y$ .

## 2.10 Coeficiente de variabilidad o de variación

Es una cantidad usada por los experimentadores para evaluar los resultados de diferentes experimentos en que interviene la misma característica y posiblemente llevados a cabo por diferentes personas. Se define por la desviación estándar de la muestra, expresada como porcentaje de la media muestral según la siguiente ecuación:

$$CV = \frac{100s}{\bar{Y}} \text{ por ciento} \quad (2.13)$$

Para saber si un determinado coeficiente de variación es insólitamente grande o pequeño, es preciso tener experiencia con datos similares. El CV es una medida relativa de variación, en contraste con la desviación estándar, la cual se expresa en las mismas unidades que las

observaciones originales. Como es la razón de dos promedios, el CV es independiente de las unidades de medida usadas, por ejemplo, da igual que se usen libras o gramos para medir el peso.

**Ejercicio 2.10.1** Para los datos dados por primera vez en los ejercicios 2.7.1 y 2.7.2 calcular el CV.

**Ejercicio 2.10.2** Para las observaciones dadas por primera vez en el ejercicio 2.7.5, calcular el CV.

## 2.11 Ejemplo

Para desarrollar una nueva técnica en Ingeniería Sanitaria, Eliassen (2.2) recolectó y presentó cantidades de sulfuro de hidrógeno provenientes de aguas negras almacenadas durante 42 horas a 37°C en 9 series, como se dan en la tabla 2.3. La técnica se desarrolló para eliminar el

**Tabla 2.3 Sulfuro de hidrógeno producido en la fermentación anaeróbica de aguas negras al cabo de 42 horas a 37°C.**

$i = \text{serie}$	$Y_i = H_2S$ , ppm	$\sqrt{s^2}$	$Y_i - \bar{Y}$
1	210		-8
2	221		+3
3	218		0
4	228		+10
5	220		+2
6	227		+9
7	223		+5
8	224		+6
9	192		-26
Totales	1,963	+35	-34
			+1 (debido al redondeo)

$$\bar{Y} = \frac{\sum Y_i}{9} = \frac{1,963}{9} = 218 \text{ ppm (luego del redondeo)}$$

$$s^2 = \frac{\sum Y_i^2 - (\sum Y_i)^2/9}{9-1} = \frac{429,147 - (1,963)^2/9}{8} = 124.36$$

$$s = \sqrt{124.36} = 11.1 \text{ ppm}$$

$$s_y^2 = \frac{s^2}{9} = \frac{124.36}{9} = 13.82$$

$$s_p = \sqrt{13.82} = 3.7 \text{ ppm}$$

$$CV = \frac{11.1(100)}{218} = 5\% \text{ (aproximadamente)}$$

sulfuro de hidrógeno de un medio de cultivo anaeróbico, mediante un gas inerte y captando el sulfuro de hidrógeno para su análisis cuantitativo con una completa exclusión de aire.

Estos datos constituyen una muestra de 9 de todas las posibles observaciones que se pueden obtener con esta técnica. El experimento fue realizado en el laboratorio, habiendo introducido varios controles para reducir el CV a este valor. Así que la población es bastante reducida. La variabilidad entre las observaciones es causada presumiblemente por cosas tales como las muestras de aguas negras, las muestras del cultivo anaeróbico usado, la técnica del operador y por otros muchos factores conocidos y desconocidos. Es claro que no podemos esperar obtener  $\mu$  y  $\sigma^2$  para esta población abstracta, pero podemos estimarlos. Esto se hace mediante el cálculo de  $\bar{Y}$  y  $s$ ; estos y otros valores se dan en la tabla 2.3.

Una  $\bar{Y}$  es una muestra de una observación tomada de la población de todas las posibles medias de muestras del mismo tamaño, o sea, nueve. La *población derivada* tiene una media  $\mu_{\bar{Y}} (= \mu_Y)$  y  $\sigma_{\bar{Y}}^2 (= \sigma_Y^2/n)$ . A partir de una muestra de observaciones de una población, ha sido posible calcular pues un valor medio  $\bar{Y}$ , una estimación de  $\mu$  y  $\mu_Y$ , y dos varianzas,  $s^2$  y  $s_{\bar{Y}}^2$ , estimativos de  $\sigma^2$  y  $\sigma_{\bar{Y}}^2$ .

## 2.12 Modelo lineal aditivo

Una práctica común en la ciencia es tratar de explicar fenómenos naturales mediante modelos. Por ejemplo, decimos que la tierra es redonda y gira sobre un eje Norte-Sur. Esto nos permite explicar desde un punto de vista rudimentario, observaciones tales como la desaparición de un buque en el horizonte, día y noche, etc. Modelos más perfeccionados dependen de afirmaciones matemáticas y nos permiten explicar, por lo tanto, no sólo hechos observables, sino también posibles sucesos no observados. El aspecto más importante de un modelo matemático es que posee las características intrínsecas del problema del mundo real.

En estadística, un modelo corriente que describa la naturaleza de una observación consta de una media más un error. Este es un modelo lineal aditivo. En la media puede entrar sólo el parámetro  $\mu$ , o puede estar compuesta de una suma de parámetros. Los supuestos respecto a los parámetros y los errores, dependen del problema. Un supuesto mínimo es que los errores son aleatorios, esto es, que la población de las  $Y$  se muestre aleatoriamente. Esto hace que el modelo sea probabilístico en lugar de determinista.

Un modelo semejante se aplica al problema de estimar o hacer inferencias respecto a medias y varianzas poblacionales. El modelo lineal aditivo más simple es

$$Y_i = \mu + \varepsilon_i \quad (2.14)$$

junto con una definición de los símbolos que aparecen en la ecuación. Expresa que la observación  $i$ -ésima es una observación de la media  $\mu$ , pero está sujeta a un error de muestreo,  $\varepsilon_i$  (*epsilon* sub  $i$ ).

Se supone que los  $\varepsilon_i$  pertenecen a una población de  $\varepsilon$  con media cero. Un requisito teórico es que no haya correlación entre los errores de muestreo, lo cual le da validez a las inferencias respecto a una población, y queda asegurado mediante muestreo aleatorio.

La media de la muestra es

$$\bar{Y} = \frac{\sum_i Y_i}{n} = \frac{\sum_i (\mu + \varepsilon_i)}{n} = \mu + \frac{\sum_i \varepsilon_i}{n}$$

En el muestreo aleatorio, es de esperar que

$$\frac{\left( \sum_i \varepsilon_i \right)}{n}$$

sea más pequeño a medida que la muestra aumente, puesto que los valores positivos y negativos tienden a cancelarse y porque el divisor aumenta. Esto es casi lo mismo que decir que se puede esperar que la varianza de las medias sea menor que la varianza de los individuos, o que las medias de muestras grandes varían menos que las medias de muestras pequeñas. En resumen, la media muestral  $\bar{Y}$  es un buen estimativo de la media de la población  $\mu$ , ya que  $\bar{Y}$  debe estar cerca de  $\mu$  si la muestra es suficientemente grande y no poco común.

Aún se requiere de un estimativo de la varianza de los  $\varepsilon$ , los individuos reales. El modelo indica que tenemos una muestra finita de  $\varepsilon$ , y éstos llevan a un estimativo de  $\sigma^2$ . Sin embargo, no podemos calcular los  $\varepsilon_i$  ya que no conocemos  $\mu$ . Por otra parte, podemos estimar los  $\varepsilon_i$  mediante las  $e_i$  correspondientes, calculados como los  $(Y_i - \bar{Y})$ , combinándolos para tener una varianza muestral calculada con la ecuación 2.7. La  $(Y_i - \bar{Y})$  calculable, de preferencia a una desviación con respecto a otro número, tiene la propiedad antes dicha que

$$\sum_i (Y_i - \bar{Y})^2$$

es un mínimo. A veces se dice que  $\bar{Y}$  es una *suma mínima de cuadrados* o una estimación de  $\mu$  por *mínimos cuadrados*

**Ejercicio 2.12.1** Escriba el modelo correspondiente a los datos del ejercicio 2.7.1. Haga lo mismo para el ejercicio 2.7.2. Suponga que las varianzas de los dos modelos son iguales. ¿Puede usted escribir una sola ecuación de un modelo que cubra simultáneamente los dos conjuntos de datos?

### 2.13 Ejemplo

Mediante un ejemplo, se ilustrará la forma de hacer los cálculos y la interpretación de los estadígrafos estudiados. Los datos de la tabla 2.4 corresponden a valores de extracto de malta de cebada Kindred cultivada en 14 localidades, en los viveros de cebada del Valle del Mississippi durante 1948. La población para la cual se ha de hacer alguna inferencia, puede considerarse como valores de extracto de malta de cebada Kindred cultivada duran-

Tabla 2.4 Valores de extracto de malta provenientes de maltas de cebada Kindred cultivada en 14 localidades en los viveros de cebada del Valle del Mississippi durante 1948, expresados en porcentaje de base seca,

Valores de extracto de malta

Original	$Y$	Desviaciones de la media $Y - \bar{Y}$	Desviaciones al cuadrado
77.7	77.7	1.7	2.89
76.0	76.0	0	0
76.9	76.9	0.9	0.81
74.6	74.6	-1.4	1.96
74.7	74.7	-1.3	1.69
76.5	76.5	0.5	0.25
74.2	75.0	-1.0	1.00
75.4	75.4	-0.6	0.36
76.0	76.0	0	0
76.0	76.0	0	0
73.9	73.9	-2.1	4.41
77.4	77.4	1.4	1.96
76.6	76.6	0.6	0.36
77.3	77.3	1.3	1.69
Totales 1,063.2	$\sum_i Y = 1,064.0$	0	$\sum_i (Y - \bar{Y})^2 = 17.38$

Fuente: Datos no publicados obtenidos con permiso de A.D. Dickson, U.S. Department Agriculture Barley and Malt Laboratory, Madison, Wisconsin.

te 1948 en la región que abarcan los viveros del Valle del Mississippi. Un valor original ha sido modificado para facilitar los cálculos.

La primera etapa en los cálculos es encontrar

$$\sum_i Y_i = 1,064 \quad \text{y} \quad \sum_i Y_i^2 = 80,881.38$$

a partir de éstos, se obtiene  $\bar{Y}$  y  $s^2$  por las ecs. (2.1), (2.7) y (2.8). La media de la muestra es

$$\bar{Y} = \frac{\sum_i Y_i}{n} = \frac{1,064}{14} = 76 \text{ por ciento}$$

y el cuadrado medio o varianza es

$$s^2 = \frac{SC}{gl} = \frac{\sum_i Y_i^2 - \left(\sum_i Y_i\right)^2/n}{n-1} = \frac{80,881.38 - (1,064)^2/14}{13} = \frac{17.38}{13} = 1.337$$

La desviación estándar o raíz cuadrada de la varianza es

$$s = \sqrt{s^2} = \sqrt{1.337} = 1.16 \text{ por ciento}$$

El cálculo de la suma de cuadrados mediante la fórmula de definición se ilustra en la última columna de la tabla 2.4.

El promedio de la muestra,  $\bar{Y} = 76\%$ , y la desviación estándar,  $s = 1.16\%$ , nos dan los mejores estimadores de los parámetros de población desconocidos,  $\mu$  y  $\sigma$ .

El coeficiente de variación se calcula con la ec. (2.13).

$$CV = \frac{s(100)}{\bar{Y}} = \frac{1.16(100)}{76.0} = 1.5 \text{ por ciento}$$

la sola media muestral constituye una muestra de tamaño 1 proveniente de una población derivada de medias, cada una basada en una muestra aleatoria de 14 observaciones obtenidas de la población principal. La media de la población derivada se estima también por  $\bar{Y} = 76\%$ , pero la varianza y la desviación estándar deben estimarse por la ec. (2.12). Obtenemos

$$s_{\bar{Y}}^2 = \frac{s^2}{n} = \frac{1.337}{14} = .0955$$

$$s_{\bar{Y}} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{1.337}{14}} \quad \text{o} \quad \frac{s}{\sqrt{n}} = \frac{1.16}{\sqrt{14}} \\ = 0.31 \text{ por ciento}$$

en cualquier caso. Este valor también se llama *error estándar*.

#### 2.14 El uso de codificación en el cálculo de estadígrafos\*

Con frecuencia los cálculos de estadígrafos se pueden facilitar y reducir con *codificación*. La codificación remplaza cada observación por un número en una nueva escala empleando una o más operaciones de adición, substracción, multiplicación y división.

La media aritmética se ve afectada por toda operación de codificación. Por ejemplo, si las observaciones se codifican multiplicando por 10 y luego restando 100, la media de los números codificados debe incrementarse en 100 y dividirse por 100. La regla consiste

---

\* Si se desea, esta sección puede omitirse sin perder continuidad.

en aplicar la operación inversa siendo las inversas de la multiplicación y la substracción, la división y la adición en orden contrario.

La desviación estándar se ve afectada únicamente por la multiplicación y la división. La adición o substracción de un valor a cada observación no afecta a la desviación estándar, ya que un cambio en el origen de las observaciones no afecta a su dispersión. Como la multiplicación y la división cambian la unidad de medida, la desviación estándar calculada con datos codificados se decodifica realizando la operación inversa con los resultados.

Ilustramos el uso de la codificación en el cálculo de la media y la desviación estándar de los datos de la tabla 2.4. Los valores codificados se obtendrán por la resta de 70 a cada una de las observaciones y luego multiplicando el resultado por 10. Puesto que todas las observaciones corresponden a valores un poco por encima de 70, la resta de 70 resulta obvia para obtener valores pequeños y positivos; la multiplicación por 10 elimina el decimal. Así el primer valor,  $Y_1 = 77.7$ , se remplazaría por su valor codificado,  $Y'_1 = (77.7 - 70) \times 10 = 7.0$ . Entonces la media y la desviación estándar de los números codificados se obtienen de  $\sum_i Y'_i = 840$  y  $\sum_i Y'^2_i = 52,138$ . Obtenemos.

$$\bar{Y}' = \frac{\sum_i Y'_i}{14} = \frac{840}{14} = 60$$

y

$$s' = \sqrt{\frac{\sum_i Y'^2_i - \left(\sum_i Y'_i\right)^2/n}{n-1}} = \sqrt{\frac{52,138 - (840)^2/14}{13}} = 11.6$$

Para decodificar  $\bar{Y}'$ , aplíquese la operación inversa en orden inverso, así

$$\bar{Y} = \frac{\bar{Y}'}{10} + 70 = \frac{60}{10} + 70 = 76 \text{ por ciento}$$

como en la sec. 2.13. Para decodificar  $s'$ ,

$$s = \frac{s'}{10} = \frac{11.6}{10} = 1.16 \text{ por ciento}$$

como en la sec. 2.13. La exactitud en los cálculos no se ve afectada por el proceso de codificación; en cambio todo se simplifica por el uso de un conjunto de números más cómodos.

**Ejercicio 2.14.1** Los números 10,807, 10,812, ... , se codificaron como 7, 12, ... La media de 17 observaciones codificadas fue de 14.6. ¿Cuál era la media de las observaciones originales? ¿Qué parte de la codificación afectó la desviación estándar? ¿Cómo?

## 2.15 La tabla de frecuencia†

La tabla de frecuencia se mencionó en la sec. 2.6. Cuando la muestra consiste en un gran número de observaciones, es preferible resumir los datos en una tabla que dé la frecuencia con que ocurre cada valor numérico o la representación de cada clase. Por consiguiente, tanto los datos continuos como los discretos pueden resumirse en tablas de frecuencias. Esto reduce la masa de los datos brutos a una forma más manejable y proporciona una base para su representación gráfica. Los estadígrafos como la media y la desviación estándar se pueden calcular con las tablas de frecuencias con mucho menos trabajo que con los valores originales.

Para variables discretas tales como el número de caras observadas al lanzar cinco monedas, los valores de clase que se han de usar son obvios en general. Así, una tabla de frecuencias del número de caras que se presentan al lanzar 5 monedas 100 veces es la que se da en la tabla 2.2. Cuando el número de clases posibles es grande, este puede ser reducido si se quiere.

Para variables continuas, clases distintas de los valores observados deben escogerse de cierta manera arbitraria. La elección depende de factores tales como el número de observaciones, la amplitud de variación, la precisión que se exija en los estadígrafos a partir de la tabla y del grado de resumen necesario para evitar que las pequeñas irregularidades puedan oscurecer las tendencias generales. Los últimos dos puntos generalmente operan en sentido opuesto; cuanto mayor el número de clases, mayor es la precisión de cualquier cálculo hecho con base en la tabla; pero si el número de clases es muy grande, los datos no quedan resumidos suficientemente.

Una regla para determinar el tamaño del intervalo de clase, cuando se requiere alta precisión en cálculos efectuados a partir de la tabla de frecuencias resultantes, es hacer que el intervalo de clase no sea mayor de un cuarto de la desviación estándar. Si se sigue esta regla estrictamente, a veces los datos no se resumen lo suficiente para la representación gráfica. Si el tamaño del intervalo de clase se incrementa a un tercio o a la mitad de la desviación estándar, la tabla de frecuencias resultante será en general un resumen suficiente para representación gráfica y es adecuada para la mayoría de los datos; la falta de precisión en los estadígrafos calculados por la tabla será bastante pequeña y por lo tanto insignificante.

Como la desviación estándar no se conoce al tiempo de elaborar la tabla de frecuencia, es necesario estimarla. Tippett (2.4) preparó tablas detalladas que indican la relación entre la amplitud de la muestra y la desviación estándar de poblaciones normales. Esto ha sido condensado por Goulden (2.3) en una tabla pequeña que se ha reproducido en la tabla A.2. Esta tabla es útil en la estimación de  $\sigma$ .

**Ejercicio 2.15.1** La amplitud de porcentaje de materia seca en 48 muestras de alfalfa Grimm fue del 7.6%. Estime la desviación estándar

---

† El resto del capítulo tiene que ver con grandes cantidades de datos, su presentación y cálculo de  $\bar{Y}$  y  $s$ . Los métodos de las secciones precedentes se pueden aplicar pero los cálculos resultan muy largos.

## 2.16 Ejemplo

Para ilustrar el uso de la tabla A.2 y la preparación y uso de una tabla de frecuencias, considere los rendimientos en aproximación al gramo, de 229 plantas de soya Richland espaciadas de que informa Drapala (2.1). La amplitud de los rendimientos estuvo entre 1 y 69 gramos. Según la tabla A.2 la razón amplitud/ $\sigma$  para  $n = 229$  es aproximadamente igual a 5.6. Así,  $\sigma$  se estima en  $68/5.6 = 12.2$  g. Un tercio de la estimación es 4 g; la mitad es 6 g. Dado que 12.2 g es una aproximación, el intervalo de clase calculado puede modificarse un poco. Es conveniente hacer que un intervalo de clase sea un número impar en lugar de par ya que el punto medio de tal intervalo cae de modo conveniente entre dos posibles valores de la variable. Tómese 5 g como intervalo de clase. Si se requiere mayor exactitud en el cálculo de los estadígrafos con la tabla, úsese un intervalo de 3 g, aproximadamente un cuarto de la  $\sigma$  estimada. Al fijar los valores de clase, es preferible que el límite inferior de la primera clase sea un poco más pequeño que el menor valor de la muestra. Selecciónese 3 como punto medio del primer intervalo. La tabla 2.1 es la tabla de frecuencias resultante.

Después de elegido el intervalo de clase, determinense las clases necesarias y distribúyanse las observaciones como corresponda. Corrientemente se usan dos métodos. El primero es el método de recuento, ilustrado en segunda, para las primeras clases de la tabla 2.1.

Amplitud de clase	Valor de clase o punto medio	Recuento	Frecuencia
1-5	3	7//	7
6-10	8	7//	5
11-15	13	7//	7

Este método consiste en hacer una marca en la clase correspondiente a cada observación y luego sumarlas para obtener la frecuencia de cada clase. Es costumbre por conveniencia trazar la quinta raya sobre las 4 precedentes, tal como se muestra en la tabla anterior. El método tiene la desventaja de que cuando no hay coincidencia al comprobar las frecuencias es difícil encontrar la fuente de error.

El segundo y quizás el método más seguro consiste en el registro del valor de cada observación en una tarjeta de un tamaño fácil de manejar. Esto debe verificarse cuidadosamente. Las amplitudes de las clases se escriben en otras tarjetas y se disponen en orden en un escritorio y las tarjetas de las observaciones se distribuyen en las clases correspondientes. La comprobación se hace fácilmente examinando las tarjetas de cada clase. La frecuencia de cada clase se determina por recuento de las tarjetas. Si se dispone de equipo de perforación y clasificación de tarjetas, entonces el valor de cada observación se puede perforar en una tarjeta y luego las tarjetas se distribuyen mecánicamente en clases.

Si los rendimientos de la tabla 2.1 se hubieren registrado con aproximación al decagramo, los límites de clase serían 0.5-5.5; 5.5-10.5; 10.5-15.5; etc. Cuando un número par de observaciones es igual a un límite de clase, por ejemplo 5.5, se asigna a cada clase la mitad de las observaciones; si es impar, la observación impar se asigna aleatoriamente a una de las dos clases.

Tabla 2.5 Cálculo de  $\bar{Y}$  y  $s$  en una tabla de frecuencia

Valores de clase o puntos medios de amplitud de clase			Frecuencia multiplicada por los valores de clase codificados	Frecuencia multiplicada por los valores de clase codificados al cuadrado
Efectivo	Codificado	Frecuencia	$f_i Y'_i$	$f_i Y'^2_i$
3	-6	7	-42	252
8	-5	5	-25	125
13	-4	7	-28	112
18	-3	18	-54	162
23	-2	32	-64	128
28	-1	41	-41	41
33	0	37	0	0
38	+1	25	25	25
43	+2	22	44	88
48	+3	19	57	171
53	+4	6	24	96
58	+5	6	30	150
63	+6	3	18	108
68	+7	1	7	49
$\sum f_i = 229 = n$			$\sum f_i Y'_i = -49$	$\sum f_i Y'^2_i = 1,507$

$$\bar{Y} = a + I \frac{\left( \sum_i f_i Y'_i \right)}{\sum_i f_i} \quad s'^2 = \frac{\sum_i f_i Y'^2_i - \left( \sum_i f_i Y'_i \right)^2 / \sum_i f_i}{\sum_i f_i - 1}$$

$$= 33 + 5 \frac{(-49)}{229} \quad = \frac{1,507 - (-49)^2 / 229}{228} = 6.56$$

$$= 31.93 \text{ g} \quad s = I \sqrt{s'^2} = 5 \sqrt{6.56} = 12.80 \text{ g}$$

En la segunda clase de la tabla 2.1 pueden observarse los efectos de agrupación al considerar los cinco valores 6, 7, 7, 9 y 9 g. La media aritmética de estos valores es 7,6 g, pero al hacer los cálculos con base en la tabla 2.1 el valor resultante es 8 g. Si se examina cada clase para este tipo de efecto, aproximadamente la mitad presentaría un error negativo y el resto uno positivo, y así tienden a cancelarse.

## 2.17 Cálculo de la media y la desviación estándar con una tabla de frecuencia

Estos cálculos se ilustrarán con los datos de la tabla 2.1. Se incluye la codificación. Primero prepare una tabla como la tabla 2.5. La primera columna,  $Y_i$ , consiste en los valores efectivos de clase. La segunda columna,  $Y'_i$ , se forma sustituyendo los valores originales

por los codificados. Para un número impar de clases, se asigna cero a la de la mitad para facilitar los cálculos, y en todo caso, el nuevo intervalo es una unidad. La columna 3 es la frecuencia; las últimas dos columnas son necesarias en los cálculos.

Los totales en las tres columnas,  $\sum f_i$ ,  $\sum f_i Y_i$ , y  $\sum f_i Y_i^2$  se necesitan para los cálculos de la media y la desviación estándar. Estos corresponden a

$$n, \quad \sum_i Y_i \quad y \quad \sum_i Y_i^2$$

Nótese que  $i$  puede referirse a la clase de la tabla de frecuencias,  $i = 1, 2, \dots, 14$  o a la observación en los datos  $i = 1, \dots, 229$ .

La media aritmética  $\bar{Y}$  y la desviación estándar  $s$  de valores originales, en lugar de unidades codificadas, se calculan a partir de los valores codificados mediante

$$\bar{Y} = a + I \frac{\sum_i (f_i Y_i)}{\sum_i f_i} \quad (2.15)$$

$$s = I \sqrt{\frac{\sum_i f_i Y_i^2 - (\sum_i f_i Y_i)^2 / \sum_i f_i}{\sum_i f_i - 1}} \quad (2.16)$$

donde  $I$  es el intervalo de clase, y  $a$  es el origen dado, es decir, el valor de  $Y_i$  que corresponde al valor de clase codificado igual a cero.

## 2.18 Representación gráfica de la tabla de frecuencia

Esto ya se expresó en la sec. 2.6. El histograma y el polígono de frecuencia para los datos de la tabla 2.1 se presentan en la fig. 2.3.

## 2.19 Dígitos significativos

Con los cálculos estadísticos viene el problema de exactitud. ¿Cuántas cifras se justifican en el resultado final de una sucesión de cálculos?

Toda observación tiene dos aspectos de información: la unidad de medida y el número de unidades de medida. Cuando se trata una variable continua, el número de unidades es el valor más cercano posible, obtenido de la escala disponible. Así, si un tablero mide 3.7 m de largo, la unidad es 1/10 m y el tablero mide entre 36.5 y 37.5 de esas unidades. Por lo tanto, se dice que el número tiene dos dígitos significantes.

Los cálculos que se hacen en computador son tan simples que es posible que los resultados finales luego de una secuencia de operaciones parezcan más exactos de lo que son en realidad. Hay reglas sobre el número de cifras significativas, resultantes de la aplicación

## 36 BIOESTADISTICA: PRINCIPIOS Y PROCEDIMIENTOS

de operaciones aritméticas, pero son poco prácticas. En la mayoría de los problemas estadísticos es mejor reservar por lo menos dos cifras extras, que pueden ser útiles en el cálculo final para redondear los resultados y obtener así un número significativo de dígitos, en relación con la exactitud de las medias originales.

### Referencias

- 2.1. Drapala, W. J.: "Early generation parent-progeny relationships in spaced plantings of soybeans, medium red clover, barley, sudan grass, and sudan grass times sorghum segregates," Tesis doctoral, University of Wisconsin, Madison, 1949.
- 2.2. Eliassen, Rolf, "Statistical analysis in sanitary engineering laboratory studies," *Biom.*, 6:117-126 (1950).
- 2.3. Goulden, C. H.: *Methods of statistical analysis*, 2a. ed., Wiley, Nueva York, 1952.
- 2.4. Tippett, L. H.: *Biometrika*, 17:386 (1926).

### Ejercicio de laboratorio propuesto en relación con el capítulo 2

Propósito (1) Dar a los estudiantes práctica en la toma de muestras aleatorias y en el cálculo de sus estadígrafos. (2) Obtener, con base en la clase, comprobación empírica respecto de ciertos estadígrafos de las muestras y compararlos con resultados teóricos. (Ha de usarse en relación con los capítulos 3 y 4.).

1. Extraer 10 muestras aleatorias de tamaño 10 de la tabla 4.1 aplicando la tabla de números aleatorios, tabla A.1. Registrar éstas en 10 columnas de dos dígitos, dejando espacio para unas doce entradas más o menos.
2. Para cada muestra calcular
  - a) La suma de las desviaciones
  - b) La media aritmética
  - c) El término de corrección
  - d) La suma de cuadrados (ajustada)
  - e) El cuadrado medio o varianza
  - f) La desviación estándar
  - g) El coeficiente de variación
  - h) La desviación estándar de la media
3. Calcular la media de las medias de las 10 muestras.
4. Calcular  $s_y^2$  y  $s_y$  a partir de las 10  $\bar{Y}$ . Es decir, tratar las 10  $\bar{Y}$  como una muestra de tamaño  $n = 10$  de una población derivada de los  $\bar{Y}$ , cada  $\bar{Y}$  calculada de un conjunto original de 10 observaciones de la población principal.

---

# CAPITULO TRES

---

## PROBABILIDAD

### 3.1 Introducción

Eventos que son comunes o improbables, son aquellos cuyas probabilidades de ocurrencia son grandes o pequeñas, respectivamente. El alcance total de las probabilidades es usado por la mayoría de las personas en una forma u otra. Alguien dice, "El fuego pudo haber sido causado por descuido" cuando no se está seguro de la causa; o dice, "El fuego casi con seguridad fue causado por descuido", cuando tiene idea firme al respecto.

Los estadísticos reemplazan las palabras informativas, pero imprecisas "pudo" y "casi con seguridad" por un número que va de cero a uno, lo cual indica en forma precisa qué tan probable o improbable es el evento. La estadística se usa para razonar de la parte al todo, es decir, de la muestra a la población. Es claro que con información incompleta, no podemos esperar hacer siempre inferencias correctas. El azar juega una parte y las leyes de la causalidad exacta no se cumplen. La estadística ofrece procedimientos que nos permiten saber cuántas veces acertamos en un promedio. Tales enunciados se conocen como enunciados probabilísticos. En este capítulo consideraremos algunas nociones de probabilidades e ilustramos el uso de tablas para obtener probabilidades asociadas con la ocurrencia de eventos estadísticos.

### 3.2 Algunos elementos de probabilidad

El uso de una razón o número para representar una probabilidad no es peculiar de los estadísticos. Los escritores deportivos pueden predecir que un equipo tiene, por ejemplo, tres posibilidades a dos de derrotar al equipo contrario. El lector puede interpretar como es de esperarse que el equipo local aventaje por estrecho margen, o que si ese partido se jugase muchas veces, entonces el equipo local ganaría en aproximadamente tres quintos o 60 % de los juegos. Expresiones tales como tres a dos, a menudo escrito 3:2, son las probabilidades y se convierten en probabilidades formando fracciones cuyos numeradores son esos números y los denominadores son las sumas de esos números, o sea, 3/5 y 2/5.

Los estadísticos asignan números entre cero y uno a las probabilidades de los eventos. Estos números son, básicamente, frecuencias relativas. Así, posibilidades de tres a dos a favor del equipo local, significan que la probabilidad de que el equipo local gane es de  $3/(3 + 2) = 3/5 = 0.6$ , y la probabilidad de que pierda, es decir, que el equipo visitante gane es  $2/(3 + 2) = 2/5 = 0.4$ . La suma de las probabilidades del conjunto completo de eventos es uno, esto es, 1, cuando la ocurrencia de un evento excluye la ocurrencia de todos los otros. Solo un equipo puede ganar si no permitimos el empate. Las probabilidades de cero y uno corresponden a eventos que, respectivamente, no ocurren con certeza y ocurren con certeza.

La ilustración es sencilla, pero con aspectos básicos que permiten la generalización. Tenemos una colección o *conjunto* de eventos. Estos son *mutuamente excluyentes*, esto es, que cuando uno ocurre los otros no pueden ocurrir simultáneamente. Además, el conjunto incluye todos los eventos permisibles o posibles, y por lo tanto es *exhaustivo*.

Como generalización, considérense los posibles resultados del lanzamiento de una moneda dos veces. El conjunto puede representarse simbólicamente así :CC, CS, SC y SS, cada uno es una sucesión ordenada de dos letras que representan caras, C y cruces (sellos), S. Si lanzamos la moneda 10 veces, entonces necesitamos una sucesión de 10 entradas, siendo cada una, C o S, presentada en orden de ocurrencia. Cada entrada debe ser una de las dos letras, así que hay  $2 \times 2 \times \dots \times 2 = 2^{10}$  resultados posibles; todos son diferentes y todos son necesarios para completar el conjunto.

En la ilustración del juego, el evento fue el resultado de un solo ensayo, el juego se desarrolló hasta su conclusión. En el segundo caso, se lanzó una moneda dos veces, así que se representaron dos ensayos que dan como resultado un par ordenado. En el tercer caso, hubo 10 ensayos con un resultado de 10 entradas. Es útil pensar que todos los resultados posibles en cualquier clase de ensayos o sucesión de ensayos son puntos en un espacio. El juego requeriría un espacio de dos puntos los cuales podrán marcarse con L y V; podríamos representarlos en una recta como en la fig. 3.1. Los dos lanzamientos de una moneda exigen cuatro puntos y éstos se distribuyen convenientemente en un espacio bidimensional; ver también la fig. 3.1. Para los diez lanzamientos de una moneda, sería preferible tener un espacio de 10 dimensiones, una dimensión para cada lanzamiento, pero esto sólo puede existir en nuestra imaginación.

Las anteriores ideas respecto a resultados de ensayos representados por partes en diversos espacios han llevado a acuñar los términos de *espacio muestral* como el conjunto de *puntos muestrales*, usados para enumerar todos los resultados posibles de un experimento. Para cada resultado, hay exactamente un *punto muestral*. Usualmente, lo más conveniente, cuando visualizamos por primera vez un espacio muestral, es pensar en un evento simple, como un resultado que *no se puede descomponer*, es decir, que no podemos visualizar otro espacio muestral el cual puede reducirse al que se tiene a mano. Una moneda lanzada dos veces lleva más bien a una sucesión ordenada, no directamente al número de caras observadas, lo cual es un resultado que *se puede descomponer*, donde CS y SC son el mismo resultado.

En cada una de nuestras ilustraciones anteriores podemos describir el curso de un experimento; se jugó una partida según reglas, —se lanzó una moneda dos veces bajo un conjunto de condiciones o una moneda se lanzó 10 veces bajo esas condiciones. Podemos imaginar que el experimento se repite una y otra vez, esencialmente bajo las mismas cir-

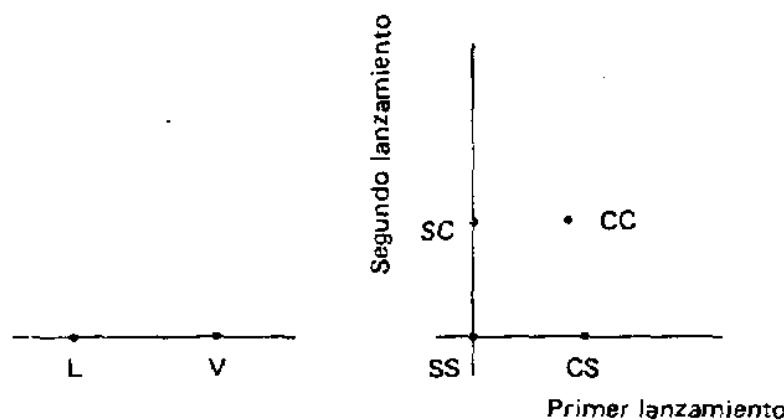


Figura 3.1 Puntos muestrales y espacios muestrales

cunstancias. Sabemos que el resultado debe ser uno de un número posible de éstos, los cuales podemos enumerar de antemano, pero no podemos, con antelación, predecir precisamente un resultado dado. Aquí el azar entra a determinar cual será el posible resultado que se va a presentar y tenemos un *experimento aleatorio*. Los eventos observados, es decir, el resultado del experimento, especifica un punto muestral o ensayo en el espacio muestral.

Los resultados de los experimentos aleatorios que hemos usado como ilustraciones no han sido números. Ahora necesitarnos cuantificar estos resultados; necesitamos una manera de asociar un número con el ensayo o punto muestral. Por ejemplo, podemos asignar 1 a V y 0 a L en el partido; 2 a CC, 1 a CS, 1 a SC y 0 a SS para hablar del número de caras al lanzar una moneda dos veces; y 10, 9, ..., 0, para contar el número de caras para cada uno de los  $2^{10}$  sucesos de los 10 lanzamientos de una moneda.

Toda asociación que asigne un valor real único a cada punto muestral se llama *variable aleatoria*. Los valores asignados son los valores de la variable aleatoria.

Nuestra nueva definición de variable aleatoria es más formal que la que dimos en la sec. 2.2 y es, al mismo tiempo, más útil para el desarrollo ulterior de nuestro tema.

Hay que asignar una probabilidad a cada punto del espacio muestral y también a cada valor de la variable aleatoria. Esto se hace fácilmente cuando la variable aleatoria es discreta en vez de continua, mediante una *función de probabilidad*, que es un conjunto de pares ordenados de valores de una variable aleatoria y la probabilidad correspondiente. Para el juego mencionado, tenemos la función de probabilidad

$$(1, 3/5), (0, 2/5)$$

Para la moneda lanzada dos veces, si la moneda es normal, tenemos

$$(2, 1/4), (1, 1/2), (0, 1/4)$$

Para la moneda lanzada 10 veces, la función de probabilidad no es tan obvia. Consideremos este hecho en la siguiente sección.

Mientras tanto, insistimos en dos hechos respecto a las probabilidades que establecimos anteriormente:

1. La probabilidad de un evento  $E_i$  cae entre 0 y 1, o puede ser 0 ó 1. Simbólicamente,

$$0 \leq P(E_i) \leq 1 \quad (3.1)$$

(El símbolo  $<$  significa "menor que";  $\leq$  significa "menor o igual";  $>$  significa "mayor que";  $\geq$  quiere decir "mayor o igual a").

2. La suma de probabilidades de los ensayos en un conjunto mutuamente excluyente es 1. Simbólicamente,

$$\sum_i P(E_i) = 1 \quad (3.2)$$

Un naípe de poker tiene 26 cartas rojas y 26 negras, con 13 picas, 13 tréboles, todas negras y 13 corazones y 13 diamantes, todas rojas;  $P(\text{pica}) = P(\text{trébol}) = P(\text{corazón}) = P(\text{diamante}) = 1/4$ . El extraer un diamante en una sola prueba excluye la extracción de una pica, un trébol y un corazón. Estos eventos son mutuamente excluyentes y  $P(\text{pica}) + P(\text{trébol}) + P(\text{corazón}) + P(\text{diamante}) = 1$ .

Hasta el momento no se ha presentado nada nuevo al lector, excepto posiblemente el simbolismo, o notación, o definiciones nuevas. Aun las probabilidades de las cartas y las de los dos lanzamientos de la moneda debieron haber sido obvias. Pero en su cálculo, se ha utilizado una definición clásica que es:

Si un ensayo aleatorio se puede presentar de  $n$  formas mutuamente excluyentes e igualmente posibles y si  $m$  ensayos tienen cierta propiedad  $A$ , entonces la probabilidad de  $A$  es la fracción  $m/n$ , o

$$P = \frac{\text{número de éxitos}}{\text{número total de ensayos} (= \text{éxitos} + \text{fracasos})} \quad (3.3)$$

Las probabilidades de ensayos asociados con variables discretas entran en muchos problemas de muestras, por ejemplo, encuestas de opinión, estudios de caracteres genéticos y problemas donde se observan recuentos. Ellas no son aplicables sin modificación, a problemas con variables continuas, tales como el peso.

### 3.3 La distribución binomial

Muchos ensayos presentan sólo dos resultados posibles, por ejemplo, una planta posee o no, cierta característica, una persona vota o no, al lanzar una moneda, puede caer por cara o sello. A tales pruebas se les llama *pruebas binomiales* o de *Bernoulli* y los espacios muestrales apropiados consistirán en dos puntos. El experimento aleatorio que genera observaciones en las mismas circunstancias esencialmente, es fácil de describir.

Los puntos muestrales se presentan convenientemente, a menudo por  $E$ , lo que indica que cierto evento ha corrido y mediante no  $E$ ,  $\bar{E}$  o  $\bar{\bar{E}}$ , para el complemento. La variable aleatoria usual asignará 1 a  $E$  y 0 a no  $E$ .

En pruebas binomiales repetidas, un resultado puede no tener efecto sobre otro, como en el lanzamiento de una moneda; se dice que tales pruebas son *independientes*. Además, la probabilidad de ocurrencia de  $E$  puede permanecer constante de una prueba a otra. Cuando estas dos propiedades se cumplen y el número de pruebas es fijo, tenemos fundamentalmente una distribución binomial. El resultado total de un experimento semejante es una sucesión ordenada de  $E$  y  $\bar{E}$  o de 1 y 0. La variable aleatoria usual asigna un valor igual al número de  $E$  o, lo que es lo mismo, a la suma de 1 y 0. Cuando se asocia una probabilidad con cada uno de los valores de la variable, entonces se tiene una función de probabilidad binomial o *distribución binomial*.

A menudo es posible presentar una fórmula matemática, la cual, en un solo enunciado, da la probabilidad relacionada con todos y cada uno de los eventos aleatorios. Así, para una moneda normal, si es  $Y = 0$  para sello y  $Y = 1$  para cara, la ecuación será:

$$P(Y = Y_i) = 1/2 \quad Y_i = 0, 1 \quad (3.4)$$

(léase: la probabilidad de que la variable aleatoria  $Y$  tome el valor particular  $Y_i$  es un medio para  $Y_i = 0$  y para  $Y_i = 1$ ), constituye una distribución de probabilidad.

Al tirar un dado equilibrado, la distribución de probabilidad sería

$$P(Y = Y_i) = 1/6 \quad Y_i = 1, 2, \dots, 6 \quad (3.5)$$

La tabla A.1 es una muestra muy grande de una población con distribución de probabilidad

$$P(Y = Y_i) = 1/10 \quad Y_i = 0, 1, 2, \dots, 9 \quad (3.6)$$

Si pensamos sólo en números impares y pares, podemos relacionar la tabla A.1 con la ec. (3.4). Las ecs. (3.5) y (3.6), no son, naturalmente, binomiales sino multinomiales.

Consideremos el problema de obtener una ecuación que dé en un solo enunciado todas las probabilidades necesarias de una distribución binomial.

Supóngase que un experimento aleatorio consiste en  $n$  pruebas independientes. Sea  $P(E) = P(1) = p$ , entonces  $P(\bar{E}) = P(0) = 1 - p$ , la ec. (3.2). Un resultado del experimento se representará como una sucesión ordenada de 1 y 0. Así, 5 lanzamientos de una moneda pueden resultar en (0, 0, 1, 1, 0), esto es, dos cruces seguidas por dos caras y al final cruz. La probabilidad de este resultado puede encontrarse, debido a la independencia de las pruebas, multiplicando las probabilidades que entran en cada etapa. Por lo tanto, la probabilidad de que el ensayo descrito ocurra es  $(1 - p)(1 - p)p p(1 - p) = p^2(1 - p)^3$ . Naturalmente, el ensayo ha ocurrido, así que esta probabilidad se aplica antes de realizar el experimento. La probabilidad asociada con cada punto muestral se obtiene de manera parecida. Naturalmente, cuando  $p = 0.5$ , tal como ocurre al lanzar una moneda normal, todos los puntos tienen la misma probabilidad, o sea  $(0.5)^5 = 0.03125$ , aproximadamente

3 posibilidades en 100. Nótese que se ha exigido que las caras ocurran en el tercer y cuarto lanzamientos.

La variable aleatoria que asocia un valor real único con cada punto muestral añadirá las entradas en la secuencia y así asocia el 2 y el número de 1s con el punto muestral de la ilustración. Este no es el único punto muestral que tiene el valor muestral 2; los dos 1s pueden ocurrir en alguna de las dos posiciones (1,2), (1,3), (1,4), (1,5), (2,3), (2,4), (2, 5), (3, 4), (3, 5) y (4, 5), 10 en total. Aquí tenemos especificadas y contadas todas las posibilidades; pero la ec. (3.7) nos permite calcular este valor en forma directa.

$$\binom{n}{Y} = \frac{n!}{Y!(n-Y)!} \quad (3.7)$$

$n$  se lee  $n$  factorial; que se define así:  $n! = n(n - 1)(n - 2) \times \dots \times 2 \times 1$ . Por tanto, para  $Y = 2$ , o sea, dos unos en  $n = 5$  pruebas, es

$$\binom{5}{2} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 3 \cdot 2 \cdot 1} = 10$$

Naturalmente, cero es un posible valor de  $Y$ ; si definimos  $0! = 1$ , no tenemos problemas en la ec. (3.7).

Con una fórmula para contar los puntos muestrales con el mismo valor de  $Y$  y otra que asigne una probabilidad a cada punto muestral, podemos escribir la distribución de probabilidad binomial así

$$P(Y = Y_i | n) = \binom{n}{Y_i} p^{Y_i} (1-p)^{n-Y_i} \quad (3.8)$$

La ecuación (3.8) se lee: "la probabilidad de que una variable aleatoria  $Y$  tome el valor particular  $Y_i$  en un experimento aleatorio con  $n$  pruebas es igual a ...". Recuérdese que  $p^0 = 1$ .

Para la ilustración de la moneda, tenemos

$$P(Y = 2 | 5) = \binom{5}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3$$

$$= 0.3125$$

Ahora resulta que el ensayo dos caras al lanzar una moneda 5 veces no es tan insólito pues, ocurre tres veces de 10, en promedio.

La tabla 3.1 nos da la distribución binomial para  $n = 5$  y varios valores de  $p$ . Nótese que la columna de la izquierda se lee con valores de  $p$  a lo largo de la línea superior, mientras que la columna de la derecha se lee con valores de  $p$  a lo largo de la línea inferior. En cada columna, las probabilidades suman 1. Para  $p = 0.5$ , la distribución es simétrica; a medida que  $p$  se aleja de ese valor, la distribución se hace más asimétrica.

Tabla 3.1 La distribución binomial,  $n = 5$ .

$Y$	Probabilidad	$p = .5$	$p = .4$	$p = .25$	$p = .1$	
0	$\binom{5}{0} p^0 (1-p)^5$	.03125	.07776	.23730	.59049	5
1	$\binom{5}{1} p^1 (1-p)^4$	.15625	.25920	.39551	.32805	4
2	$\binom{5}{2} p^2 (1-p)^3$	.31250	.34560	.26367	.07290	3
3	$\binom{5}{3} p^3 (1-p)^2$	.31250	.23040	.08789	.00810	2
4	$\binom{5}{4} p^4 (1-p)^1$	.15625	.07680	.01465	.00045	1
5	$\binom{5}{5} p^5 (1-p)^0$	.03125	.01024	.00098	.00001	0
		$p = .5$	$p = .6$	$p = .75$	$p = .9$	$Y$

La media y la varianza de una variable aleatoria con distribución binomial son

$$\text{Media : } \mu = np \quad (3.9)$$

$$\text{Varianza: } \sigma^2 = np(1-p) \quad (3.10)$$

Nótese que la varianza se determina con base en  $p$ , así que sólo se necesita un parámetro para caracterizar la distribución binomial;  $n$  es un parámetro observable, de modo que es de una categoría diferente.

La ecuación para la media da valores razonables; si  $p = 1/2$ , esperamos que aproximadamente la mitad de los lanzamientos den caras; si  $p = 0.1$  para algún otro resultado, entonces esperamos que aproximadamente un décimo de las pruebas dé este resultado.

Para la distribución binomial de la tabla 3.1 con  $n = 5$  y  $p = 0.5$ , tenemos que  $\mu = 2.5$ ,  $\sigma^2 = 1.25$ , y  $\sigma = 1.12$ ; para  $p = .4$ ,  $\mu = 2$ ,  $\sigma^2 = 1.2$ , y  $\sigma = 1.10$ ; para  $p = .25$ ,  $\mu = 1.25$ ,  $\sigma^2 = .94$ , y  $\sigma = .97$ ; para  $p = .1$ ,  $\mu = .5$ ,  $\sigma^2 = .45$ , y  $\sigma = .67$ . Nótese que la varianza de la variable aleatoria cambia lentamente a medida que  $p$  comienza a alejarse de 0.5, pero luego cambia rápidamente a medida que se acerca a 0 o a 1.

Las tablas de la A.14A a la A.17G se basan en la distribución binomial y su aplicación; en los caps. 21 a 23 se tratan problemas donde una distribución binomial es, en muchos casos, un supuesto fundamental.

La distribución binomial también se usa como una aproximación de otras distribuciones de variables discretas. Por ejemplo, el muestreo se hace a menudo de poblaciones finitas. Si se representa el tamaño de la población por  $N$ , entonces tenemos una probabilidad de  $1/N$  de extraer un individuo dado en la primera prueba. La probabilidad de extraer ese individuo dada la segunda prueba es dependiente de lo que ocurrió en la primera, ya que estamos muestreando sin remplazo, será 0 si el individuo ya ha sido extraído, pero será  $1/(N - 1)$  si no lo ha sido. Las probabilidades no son constantes de una prueba a otra. Sin embargo, si el tamaño de la muestra no es muy grande en relación con el tamaño de la población, la distribución binomial bien puede ser una aproximación muy satisfactoria para calcular las probabilidades necesarias. Hay aun otras distribuciones de variables discretas para las cuales la distribución binomial suele ser una aproximación razonable.

**Ejercicio 3.3.1** Un dado equilibrado se lanza dos veces. ¿Cuántos puntos tiene el espacio muestral? ¿Cuáles son las probabilidades asociadas con cada punto muestral? Representar el espacio muestral y obtener a partir del mismo la distribución de probabilidad de la suma de los números en los dos lanzamientos. Use la definición de  $\mu$  y  $\sigma^2$  del capítulo 2 para calcular la media y la varianza de esa suma.

**Ejercicio 3.3.2** Repetir el ejercicio 3.3.1 utilizando la media en lugar de la suma. ¿Se hubiera podido aplicar la sec. 2.9 para reducir esfuerzos? Explique su respuesta.

**Ejercicio 3.3.3** Una prueba tiene 10 preguntas de elección múltiple, cada una de las cuales tiene cuatro opciones. Un estudiante se propone adivinar en el examen y se pregunta cuál será la probabilidad de éxito en la prueba. ¿Cuál sería la prueba básica de Bernoulli? ¿Qué probabilidades se han de asociar con el espacio muestral para una prueba simple? ¿Cuántas pruebas deben hacerse? ¿Cuál es la distribución del número de preguntas contestadas correctamente? ¿Cuál es la probabilidad de responder correctamente cinco o más preguntas?

**Ejercicio 3.3.4** Con frecuencia los investigadores calculan intervalos para incluir medias de poblaciones. Como usan datos muestrales, no pueden tener certeza de que los intervalos contengan las medias. Sin embargo, sus técnicas son tales, que pueden decir con qué probabilidad una aplicación particular tendrá éxito en producir un intervalo que contenga la media. La probabilidad se fija a menudo en 0.95.

Supóngase que un investigador lleva a cabo 20 investigaciones independientes, cada una de las cuales resulta en un solo intervalo. ¿Cuál es la probabilidad de que no más de dos intervalos incluyan la media de la población que buscaba?

XVII

### 3.4 Funciones de probabilidad para variables continuas

No todas las probabilidades tienen que ver con variables aleatorias discretas. Para variables continuas, las tablas de frecuencia y los histogramas pueden dar probabilidades aproximadas. Una probabilidad encontrada en esta forma (calculando una frecuencia relativa para cada intervalo, por ejemplo) sería una aproximación de la probabilidad verdadera de una variable aleatoria que toma un valor en un intervalo. Hay que buscar un enfoque diferente para describir una función de probabilidad.

Considérese la ruleta de la fig. 3.2. El punto de parada se define como el punto que queda frente a la flecha fija. ¿Cuántos puntos de parada hay allí? La rueda se podría dividir en 10 sectores y se define el punto de parada como el número más cercano a la flecha. Pero cada sector se podría dividir en más de 10 subsectores para obtener 100 puntos de parada, y así sucesivamente. Es claro que no hay un número finito de punto de parada, y, como resultado, nuestra definición clásica de probabilidad, ec. (3.3), no opera porque no tenemos un número para el denominador.

Para una variable continua con un número indefinidamente grande de ensayos, no se puede asignar una probabilidad para cada valor. Para la ruleta dividida en 10 sectores con los números 0, 1, 2, ..., 9 marcados en líneas de división sucesivas, no podemos hablar de probabilidades asociadas con puntos sino de probabilidades asociadas con sectores o intervalos. El puntero debe, naturalmente, parar en uno de los infinitos puntos. Para probabilidades asociadas con tales intervalos, utilizamos una expresión apropiada en  $Y$  o una función de  $Y$ , que se escribe  $f(Y)$  y se llama *función de densidad de probabilidad*. El símbolo  $f(Y)$  es un término genérico, tal como "manzana"; además se requiere de algo adicio-

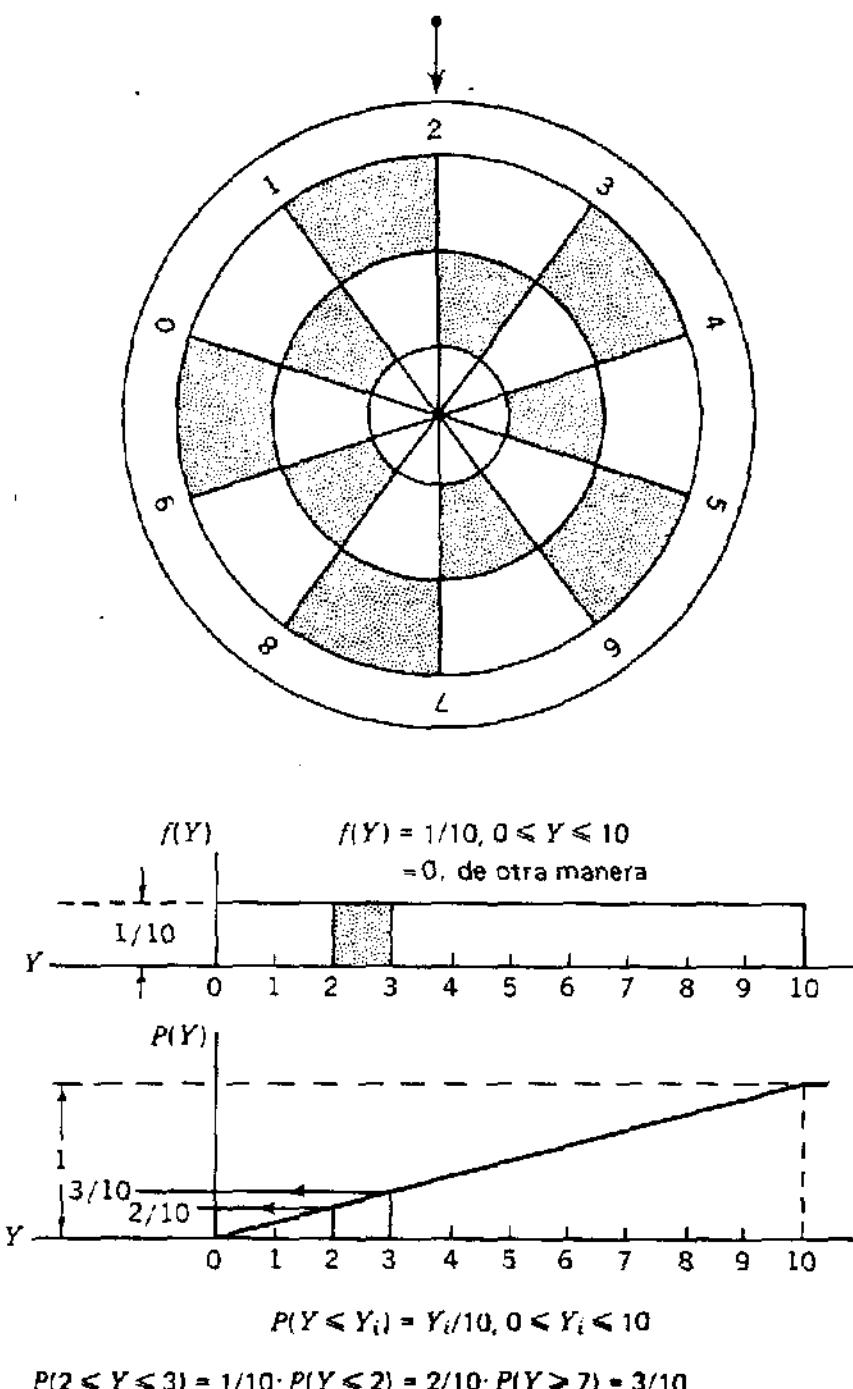


Figura 3.2 Una rueda de la fortuna y su correspondiente distribución de probabilidad.

nal para informarnos plenamente. Para  $f(Y)$ , necesitamos una ecuación, por ejemplo, ec. (3.11); para manzana necesitaríamos un nombre, MacIntosh, por ejemplo.

Una función de densidad de probabilidad se interpreta fácilmente en forma gráfica. Refiriéndonos a las fig. 3.2, vemos que

$$f(Y) = 1/10 \quad 0 \leq Y \leq 10 \quad (3.11)$$

es una función que describe una densidad de probabilidad. Esta es la distribución *uniforme*. Aquí, todo valor de  $Y$  entre 0 y 10 es posible. Las áreas bajo esta curva (el término

puede incluir líneas rectas) están asociadas con probabilidades. Por ejemplo, el área total es  $(1/10)10$ , o 1; el área sombreada entre 2 y 3 es  $1/10$ ; no existe área bajo la curva para valores menores que 0 o mayores a 10. (Los números 0 y 10 son lo mismo para este ejemplo). La *función de distribución acumulada*  $P(Y)$  o su gráfica se usan para encontrar *probabilidades*. Por ejemplo, para hallar la probabilidad de que el puntero se detenga entre los valores 2 y 3, esto es,  $P(2 \leq Y \leq 3)$ , léase de  $Y = 3$  hacia arriba hasta encontrar la recta inclinada, luego hasta contar con el eje de  $P(Y)$  para obtener el valor tres décimos; Repetir para  $Y = 2$  para obtener  $2/10$ . Ahora restar la probabilidad de obtener un valor menor que 2 de la probabilidad de obtener un valor menor que 3 para obtener así la probabilidad de encontrar un valor entre 2 y 3, o sea,

$$P(2 \leq Y \leq 3) = P(Y \leq 3) - P(Y \leq 2)$$

$\frac{3}{10} - \frac{2}{10}$

Las funciones de densidad de probabilidad se caracterizan por sus variables aleatorias. La media y la varianza,  $\mu$  y  $\sigma^2$ , son los parámetros más usados. El cálculo de  $\mu$  y  $\sigma^2$  para variables continuas se sale del dominio de este texto. Cuando se desconocen, se estiman mediante estadígrafos muestrales.

A medida que avancemos tendremos mucho que ver con funciones de distribuciones acumuladas, o  $P(Y)$ . Para nosotros, esto significará tomar un número de una tabla. Usamos poco las funciones de densidad  $f(Y)$ , excepto el nombre.

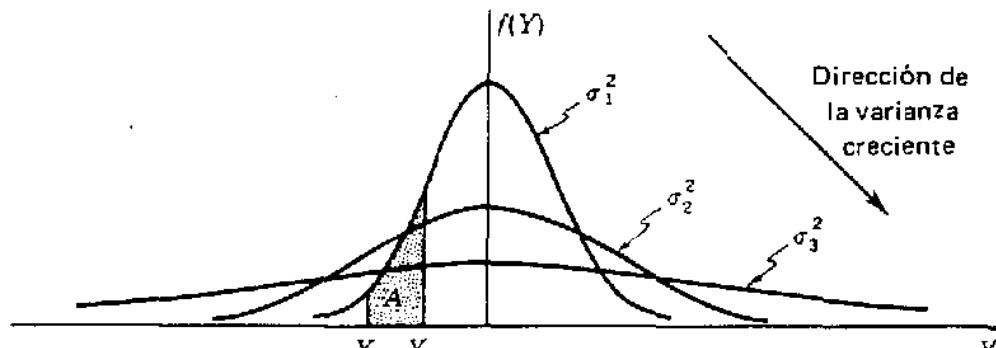
**Ejercicio 3.4.1** Dada  $f(Y) = 1/10, 0 \leq Y \leq 10$ , como en la ec. (3.11). Con la fig. 3.2 determinar:  $P(2 \leq Y \leq 7)$ ;  $P(1 \leq Y \leq 9)$ ;  $P(2 \leq Y \leq 4 \text{ o } 6 \leq Y \leq 8)$ ;  $P(2 \leq Y \leq 4 \text{ y } 3 \leq Y \leq 7$ , simultáneamente).

### 3.5 La distribución normal

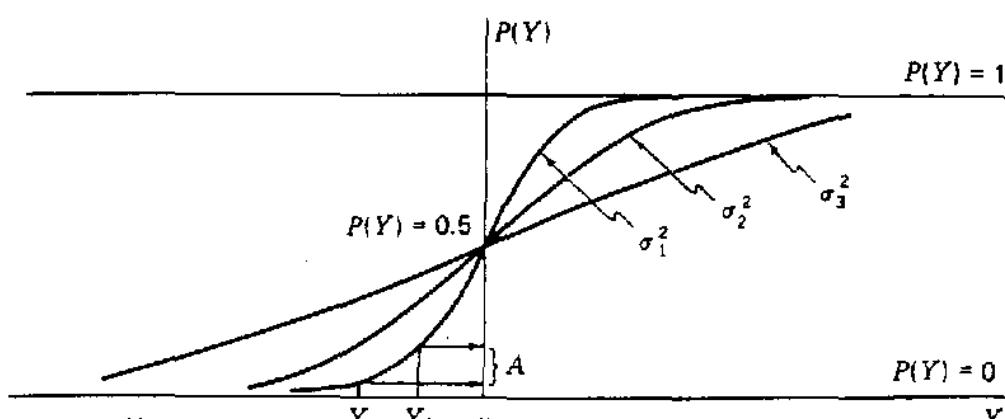
La *distribución normal* es importante en la teoría y en la práctica de la estadística. Muchos fenómenos biológicos presentan datos distribuidos de manera tan suficientemente normal que su distribución es la base de gran parte de la teoría estadística usada por los biólogos. En efecto, lo mismo se verifica en muchos otros campos de aplicación. La gráfica de la distribución normal, la curva normal, también llamada laplaciana o gaussiana, tiene forma de campana. La localización del centro de la curva la da  $\mu$ ; la cantidad de la dispersión está dada por el tamaño de  $\sigma^2$ : una  $\sigma^2$  pequeña da una protuberancia más alta que una  $\sigma^2$  grande. Ver fig. 3.3.

Por un momento considérese la distribución binomial para  $n = 10$  y  $p = .5$ . Sea la variable aleatoria el número de 1 observados, esto es, el número de caras al lanzar una moneda normal 10 veces. Las probabilidades para los valores se dan en la tabla 3.2 y se representan en la fig. 3.4.

Como la distribución normal, la binomial con  $p = 0.5$  es una distribución simétrica, pero sólo se necesita de barras para representar las probabilidades. Podríamos construir un diagrama de barras que se pareciera a un histograma con columnas de una unidad de ancho, como se indica en líneas de puntos; pero por implicación, un valor de la variable aleatoria se extiende ahora media unidad a cada lado del valor efectivo.



$$A = \text{área} = P(Y_1 < Y < Y_2)$$



$$A = \text{longitud} = P(Y_1 < Y < Y_2)$$

Figura 3.3 Distribución normal  $f(Y)$ , y normal acumulada,  $P(Y)$

Con la imaginación podemos ver este histograma comparable a una de las funciones de densidad normal de la fig. 3.3; al histograma sólo le falta suavidad. Ahora imaginemos que  $n$  aumenta mientras continuamos construyendo el histograma como el de la fig. 3.4, usando la misma longitud de la línea base. Será necesario reducir la unidad del intervalo y las columnas del histograma. A medida que  $n$  crece, se necesitan más y más probabilidades. Por lo tanto, éstas se hacen más pequeñas en los puntos de localización correspondientes como el centro o las colas. Los escalones del histograma se hacen más pequeños y la figura se hace cada vez más parecida a una curva continua, en efecto, como la curva normal.

Tabla 3.2 Probabilidades binomiales para  $n = 10, p = 0.5$

$Y$	0	1	2	3	4	5†	5‡
$\binom{10}{r}$	1	10	45	120	210	252	
$P(Y)$	.00098	.00977	.04395	.11719	.20508	.24609	
	$\mu = 5, \sigma^2 = 2.5, \sigma = 1.58.$						

† Para completar la tabla, nótese que  $\binom{10}{r} = \binom{10}{10-r}$  y  $P(Y) = P(10 - Y)$ .

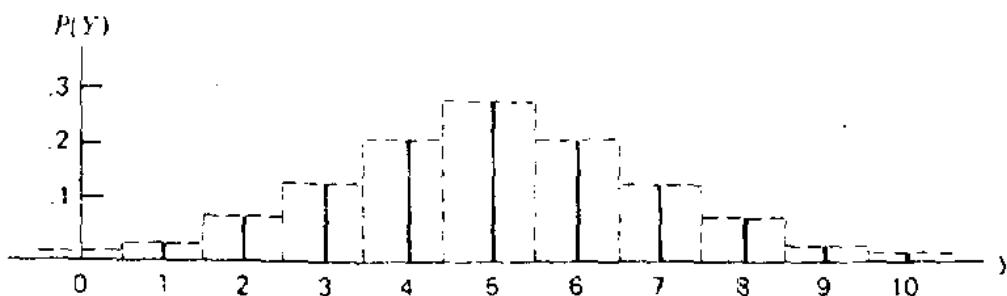


Figura 3.4 Barras verticales, probabilidades binomiales para  $n = 10$ . Las líneas punteadas dan la representación mediante el histograma

Llega el momento en el que la probabilidad de un valor particular de la variable aleatoria se hace muy pequeña debido a un divisor grande,  $2n$ , y al número posible de valores,  $n + 1$ ; por lo tanto, se hace poco práctico calcular y registrar distribuciones de probabilidades. Necesitamos conocer la probabilidad de un conjunto de resultados, por ejemplo, de que al lanzar una moneda 100 veces, la probabilidad de que el número de caras se encuentre entre 45 y 55 o de que el número de caras sea inferior a 25. Fue en el estudio de una distribución aproximada que diera tales probabilidades que de Moivre descubrió la curva normal. Esta nos permite calcular las probabilidades para tal conjunto de valores de  $Y$  y es satisfactoria para valores de  $p$  diferentes de 0.5 con tal que  $n$  sea suficientemente grande.

La fórmula matemática de la función de densidad normal no da probabilidades directamente, sino que describe las curvas de la parte superior de la fig. 3.3; está completamente determinada por los dos parámetros  $\mu$  y  $\sigma^2$  o  $\sigma$ . La magnitud del área  $Y_1$  e  $Y_2$ , es decir,  $P(Y_1 \leq Y \leq Y_2) = A$ . El área total bajo la curva y por encima del eje de la  $Y$  es 1. La curva es simétrica y la mitad del área se encuentra a cada lado de  $\mu$ .

Las funciones de distribución normal acumulada, curvas que dan la probabilidad de que un  $Y$  al azar sea menor que un valor dado, es decir,  $P(Y \leq Y_i)$ , se presentan en la porción inferior de la figura para las curvas de la parte superior. Leáse hacia arriba de  $Y_i$  y hasta el eje de  $P(Y)$  para encontrar las probabilidades. Como las probabilidades acumuladas están dadas, es necesario obtener dos valores de  $P(Y)$  para evaluar expresiones como  $P(Y_1 \leq Y \leq Y_2)$ . No existe expresión matemática simple de  $P(Y)$  para la distribución normal.

### 3.6 Probabilidades de una distribución normal. Uso de una tabla de probabilidades

En vez de una curva, se usa una tabla para obtener las probabilidades de una distribución normal. La tabla A.4 de las probabilidades para una distribución normal con  $\mu = 0$  y  $\sigma^2 = 1$ . La variable se denota  $Z$ . Los valores de  $Z$  se dan con un decimal en la columna de  $Z$  y hasta dos decimales en la fila de  $Z$ . Esta tabla se puede usar para obtener probabilidades asociadas con una distribución normal si se conocen la media y la varianza. Su uso se ilustrará ahora por medio de varios problemas en que interviene una distribución normal con  $\mu = 0$  y  $\sigma^2 = 1$ .

**Caso 1** Encontrar la probabilidad de que un valor aleatorio de  $Z$  sea mayor que un valor positivo de  $Z_1$ , esto es,  $P(Z \geq Z_1)$ . Encontremos  $P(Z \geq 1.17)$ .

*Procedimiento* Encontrar 1.1 en la columna de  $Z$  y 0.07 en la fila de  $Z$ ,  $1.17 = 1.1 + 0.07$ . La probabilidad se encuentra en la intersección de fila y columna, así  $P(Z \geq 1.17) = .1210$ . O sea que aproximadamente 12 por ciento de las veces, en promedio, es de esperar extraer un valor de  $Z$  mayor que 1.17 (Ver la fig. 3.5a).

**Caso 1a** Encontrar la probabilidad de que un valor aleatorio de  $Z$  sea menor que un valor positivo  $Z_1$ , esto es,  $P(Z \leq Z_1)$ . Encontremos  $P(Z \leq 1.17)$ .

*Procedimiento* Dado que el área bajo la curva es 1,

$$P(Z \leq 1.17) = 1 - .1210 = .8790$$

(Ver fig. 3.5 a).

**Caso 1b** Encontrar la probabilidad de que un valor aleatorio de  $Z$  sea menor que un valor negativo  $Z_1$ , esto es,  $P(Z \leq Z_1)$  donde  $Z_1 \leq 0$ . Encontramos  $P(Z \leq -1.17)$ .

*Procedimiento* La curva normal es simétrica. Como estamos viendo una distribución normal con media cero,

$$P(Z \leq -1.17) = P(Z \geq 1.17) = .1210$$

Ver fig. 3.5d para  $Z = -1.05$  y  $+1.05$ .

**Caso 2** Encontrar la probabilidad de que un valor aleatorio de  $Z$  caiga en un intervalo  $(Z_1, Z_2)$ , a la derecha del origen, esto es,  $P(Z_1 \leq Z \leq Z_2)$  donde  $Z_1 \geq 0$  y  $Z_2 > 0$ . Busquemos  $P(.42 \leq Z \leq 1.61)$ .

*Procedimiento* Encontrar  $P(Z \geq .42)$  y  $P(Z \geq 1.61)$ . Ahora

$$\begin{aligned} P(.42 \leq Z \leq 1.61) &= P(Z \geq .42) - P(Z \geq 1.61) \\ &= .3372 - .0537 = .2835 \end{aligned}$$

(Ver fig. 3.5b).

**Caso 2a** Encontrar la probabilidad de que un valor aleatorio de  $Z$  caiga en un intervalo a la izquierda del origen, esto es,  $P(Z_1 \leq Z \leq Z_2)$  donde  $Z_1 < 0$  y  $Z_2 \leq 0$ . Encontremos  $P(-1.61 \leq Z \leq -.42)$ . (Nótese que no es necesario colocar ningún signo antes de la letra  $Z$  a menos que se use algún valor especial como ilustración).

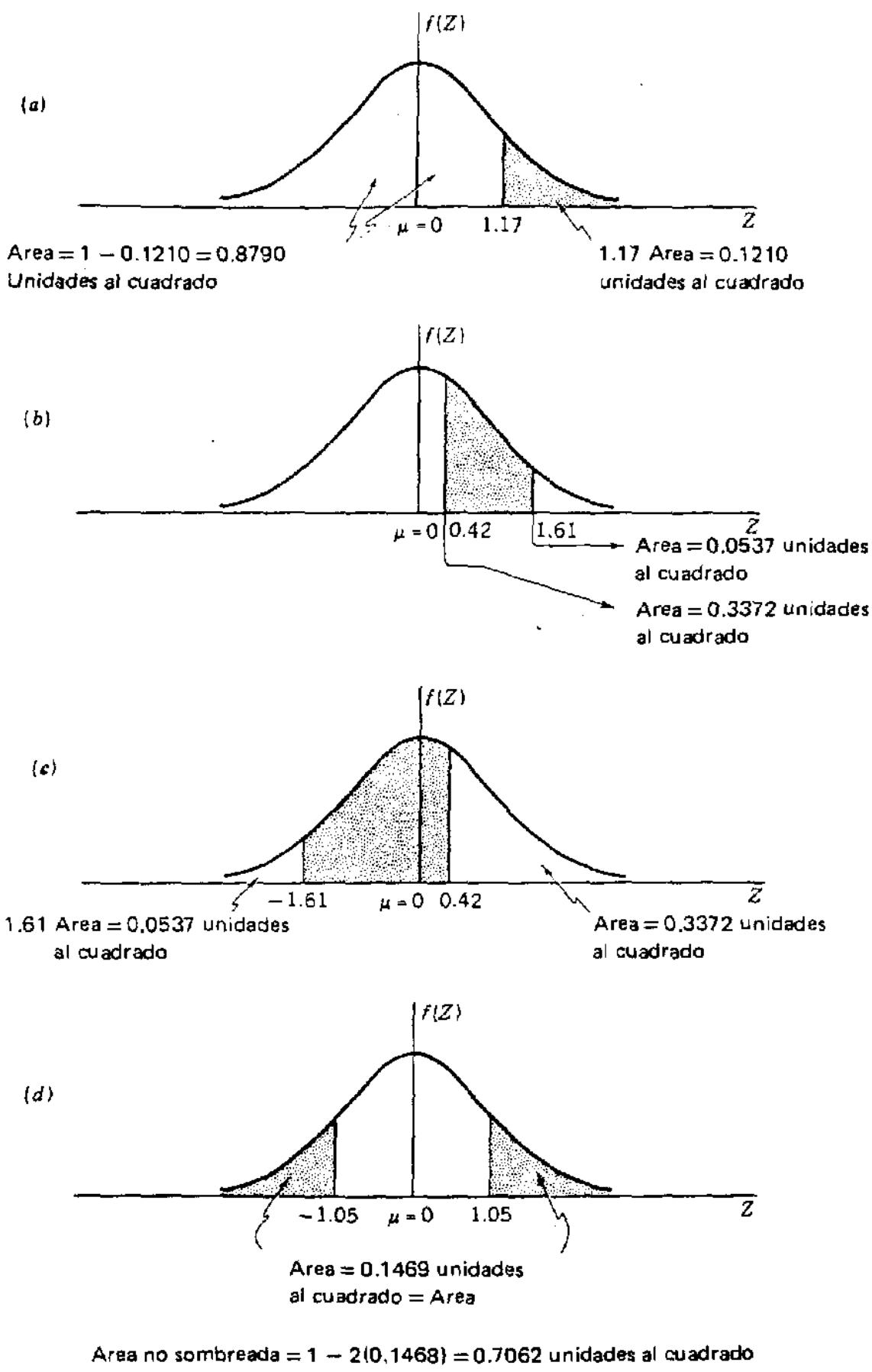


Figura 3.5 Algunas ilustraciones de probabilidades

*Procedimiento* Nuevamente porque la curva es simétrica y porque  $\mu = 0$ ,

$$P(-1.61 \leq Z \leq -0.42) = P(0.42 \leq Z \leq 1.61) = .2835$$

**Caso 2b** Para encontrar la probabilidad de que un valor aleatorio de  $Z$  caiga en un intervalo que incluya el origen, esto es  $P(Z_1 \leq Z \leq Z_2)$ , donde  $Z_1 < 0$  y  $Z_2 > 0$ . Encuentre  $P(-1.61 \leq Z \leq 0.42)$ .

*Procedimiento* Como el área bajo la curva es 1, encontremos las áreas fuera del intervalo y restemos su suma de 1.  $P(Z \leq -1.61) = P(Z \geq 1.61) = .0537$  y  $P(Z \geq 0.42) = .3372$ . Así:

$$P(-1.61 \leq Z \leq 0.42) = 1 - (.0537 + .3372) = .6091$$

(Ver fig. 3.5c)

Una probabilidad que a menudo se necesita en aplicaciones estadísticas es la de encontrar un valor aleatorio numéricamente mayor que un número dado.

**Caso 3** Para encontrar la probabilidad de que un valor aleatorio  $Z$  exceda numéricamente a  $Z_1$ , es decir, que caiga fuera del intervalo  $(-Z_1, Z_1)$ , es necesario que  $P(|Z| \geq Z_1)$ . Encontremos  $P(|Z| \geq 1.05)$ .

*Procedimiento* Debido a la simetría,

$$P(|Z| \geq 1.05) = 2P(Z \geq 1.05) = 2(.1469) = .2938$$

(Ver fig. 3.5d). Este método es más corto que el usado en el caso 2b.

**Caso 3a** Para encontrar la probabilidad de que un valor aleatorio de  $Z$  sea numéricamente inferior a  $Z_1$ , es decir, que caiga dentro del intervalo  $(-Z_1, Z_1)$ , es necesario que  $P(|Z| \leq Z_1)$ . Encontremos  $P(|Z| \leq 1.05)$ .

*Procedimiento* Puesto que el área bajo la curva es 1,

$$\begin{aligned} P(|Z| \leq 1.05) &= 1 - P(|Z| \geq 1.05) \\ &= 1 - 2(.1469) = .7062 \end{aligned}$$

(Ver fig. 3.5d). También se puede usar el procedimiento del caso 2b.

En estadística, a menudo es necesario encontrar valores de un estadígrafo tales como valores aleatorios que lo sobrepasen en una proporción dada de casos, esto es, con una probabilidad dada. Esto equivale a la construcción de una tabla propia con los valores deseados de  $P$ , digamos en los márgenes, y valores de la variable en el cuerpo de la tabla.

Por ejemplo, al operar con la distribución normal (o cualquier distribución simétrica), el 50 por ciento de los valores aleatorios de  $Z$  serán mayores que la media.

**Caso 4** Para ilustrar, encontremos el valor de  $Z$  que sea excedido con una probabilidad dada (es decir, un valor aleatorio debe caer a la derecha del valor requerido con una probabilidad dada); por ejemplo, encontremos  $Z_1$  tal que  $P(Z \geq Z_1) = .25$ .

*Procedimiento* Buscar en el cuerpo de tabla la probabilidad 0.2500. Está en la línea  $Z = 0.6$ , aproximadamente en la mitad entre las columnas 0.07 y 0.08. El valor de  $Z$  está entre 0.67 y 0.68. Así  $P(Z \geq .67) = 0.25$  (aproximadamente).

También se requiere a menudo un valor de  $Z$  que sea excedido numéricamente o no, por una probabilidad dada.

**Caso 5** Encontrar el valor de  $Z$ , digamos  $Z_1$ , tal que  $P(|Z| \geq Z_1)$  sea igual a un valor dado [esto es, que el valor aleatorio caiga fuera del intervalo  $(-Z_1, Z_1)$ ]. Encontremos  $Z_1$  tal que  $P(|Z| \geq Z_1) = .05$ .

*Procedimiento* Dado que la curva es simétrica, hállese  $Z_1$  tal que  $P(Z \geq Z_1) = 0.05/2 = 0.025$ . El mismo procedimiento que para el caso 4 dado  $1.9 + 0.06 = 1.96$ . Por lo tanto,  $P(|Z| \geq 1.96) = 0.05$ .

**Caso 5a** Encontrar el valor de  $Z$ , digamos  $Z_1$ , tal que  $P(-Z_1 \leq Z \leq Z_1)$  sea igual a un valor dado [esto es, que el valor aleatorio caiga dentro del intervalo  $(-Z_1, Z_1)$ ]. Encontremos  $Z_1$  tal que  $P(-Z_1 \leq Z \leq Z_1) = .99$ .

*Procedimiento* Dado que el área bajo la curva es 1, nos referimos al caso 5 y observamos que  $P(-Z_1 \leq Z \leq Z_1) = 1 - P(|Z| \geq Z_1)$ . Así,

$$\begin{aligned} 1 - P(-Z_1 \leq Z \leq Z_1) &= P(|Z| \geq Z_1) \\ &= 1 - .99 = .01 \end{aligned}$$

Como en el caso 5, encuentre  $Z_1$ , tal que  $P(Z \geq Z_1) = .005$ ;  $Z_1$  cae entre 2.57 y 2.58. (Para tres dígitos decimales  $Z_1 = 2.576$ .) Por lo tanto,

$$P(-2.576 \leq Z \leq 2.576) = .99$$

**Ejercicio 3.6.1** Dada una distribución normal con media cero y varianza uno, encontrar  $P(Z \geq 1.70)$ ;  $P(Z \geq .96)$ ;  $P(Z \leq 1.44)$ ;  $P(Z \leq -1.44)$ ;  $P(-1.01 \leq Z \leq .33)$ ;  $P(-1 \leq Z \leq 1)$ ;  $P(|Z| \leq 1)$ ;  $P(|Z| \geq 1.65)$ ;  $P(.45 \leq Z \leq 2.08)$ .

**Ejercicio 3.6.2** Encontrar  $Z_0$  tal que  $P(Z \geq Z_0) = .3333$ ;  $P(Z \leq Z_0) = .6050$ ;  $P(1.00 \leq Z \leq Z_0) = .1000$ ;  $P(|Z| \geq Z_0) = .0100$ ;  $P(|Z| \leq Z_0) = .9900$ .

### 3.7 La distribución normal con media $\mu$ y varianza $\sigma^2$

La distribución normal con  $\mu = 0$  y  $\sigma^2 = 1$  es una y sólo una, y usted puede haberse preguntado acerca de las tablas para las muchas posibles combinaciones de valores de  $\mu$  y  $\sigma^2$ . Como veremos, no son necesarias más tablas.

*Ejemplo* Supóngase que estamos muestreando una población normal con  $\mu = 12$  y  $\sigma^2 = 1$ , y se requiere  $P(Y \geq 13.15)$ .

La distribución es la misma de la sección anterior, salvo que se ha desplazado de modo que  $\mu = 12$  en vez de 0. Por tanto

$$\begin{aligned} P(Y \geq 13.15) &= P(Y - \mu \geq 13.15 - 12) \\ &= P(Z \geq 1.15) = .1251 \end{aligned}$$

En general, deseamos encontrar  $P(Y_1 \leq Y \leq Y_2)$ . Por ejemplo,

$$\begin{aligned} P(11.20 \leq Y \leq 13.44) &= 1 - P(Y \text{ está fuera del intervalo}) \\ &= 1 - [P(Y \leq 11.20) + P(Y \geq 13.44)] \\ &= 1 - [P(Z \leq -.80) + P(Z \geq 1.44)] \\ &= 1 - [P(Z \geq .80) + P(Z \geq 1.44)] \\ &= 1 - (.2119 + .0749) = .7132 \end{aligned}$$

(Ver fig. 3.6). El método, como se ve, consiste en pasar de una variable  $Y$  con media diferente de 0 a una variable  $Z$  con media cero restando  $\mu$ .

Para el caso más general, o sea, cuando  $\mu \neq 0$  y  $\sigma^2 \neq 1$  (el símbolo  $\neq$  quiere decir "diferente de"), se usa la tabla A.4 calculando

$$Z = \frac{Y - \mu}{\sigma} \quad (3.12)$$

una desviación con respecto a la media expresada en unidades de desviación estándar (ver fig. 3.7). Así, transformamos una variable normal en otra con media cero y varianza uno, ya que  $\sigma$  es la nueva unidad de medida.

*Ejemplo* Si se hace el muestreo de una distribución normal con  $\mu = 5$  y  $\sigma^2 = 4$  o  $\sigma = 2$ , encontrar la probabilidad de un valor muestral mayor que 7.78.

Buscar en la tabla A.4

$$Z = \frac{Y - \mu}{\sigma} = \frac{7.78 - 5}{2} = 1.39$$

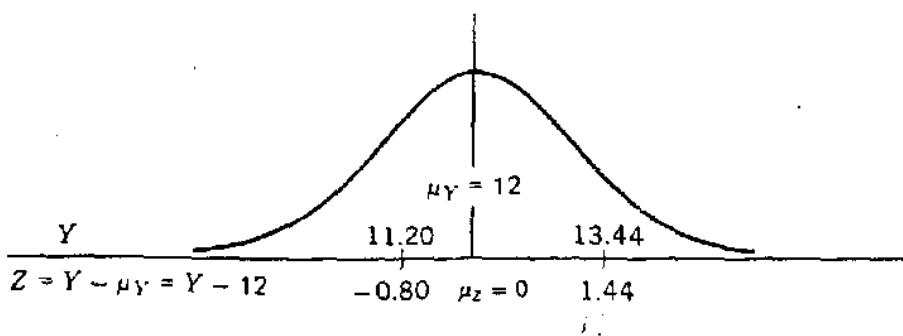


Figura 3.6 Probabilidades para una distribución normal con  $\mu = 12$ ,  $\sigma^2 = 1$ .

Entonces

$$\begin{aligned} P(Y \geq 7.78) &= P\left(\frac{Y - \mu}{\sigma} \geq \frac{7.78 - 5}{2}\right) \\ &= P(Z \geq 1.39) = .0823 \end{aligned}$$

En este caso,  $\sigma$  es la unidad de medida. La variable  $Z$  es una desviación respecto de la media, o sea,  $Y - \mu$ , medida en unidades de desviaciones estándar, o sea,  $(Y - \mu)/\sigma$ . La anterior expresión probabilística da la probabilidad de que un valor aleatorio de  $Z$  sea mayor que 1.395 o de que un valor aleatorio de  $Y$  esté más a la derecha de  $\mu$  que  $1.39\sigma$ . La variable  $Z$  se llama *variable normal estandarizada*. Si no se sabe que una distribución es normal, entonces esa variable es una *variable estandarizada*.

El enunciado

$$P\left(-1 \leq \frac{Y - \mu}{\sigma} \leq +1\right) = 1 - 2(.1587) = .6826$$

quiere decir que la probabilidad de encontrar un valor aleatorio de  $(Y - \mu)/\sigma$  entre -1 y +1 es aproximadamente dos tercios; o que aproximadamente dos tercios de todos los valores de  $(Y - \mu)/\sigma$  se encuentran entre -1 y +1.

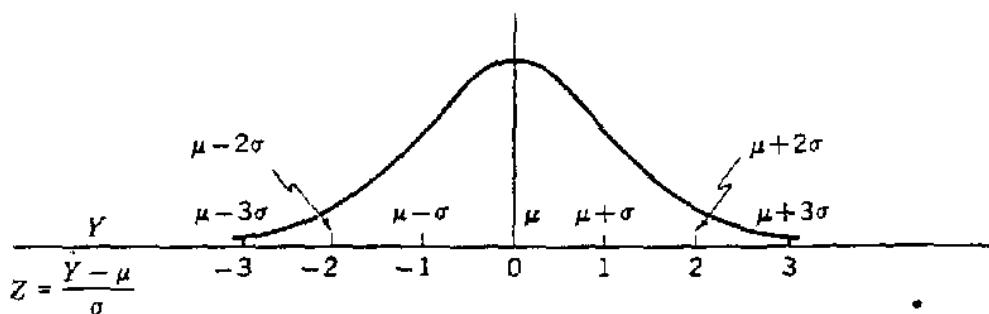


Figura 3.7 Relación entre  $Y$  y  $Z$  para cálculo de probabilidades de la distribución normal.

**Ejercicio 3.7.1** Dada una distribución normal de  $Y$  con media 5 y varianza 16, encontrar  $P(Y \leq 10)$ ;  $P(Y \leq 0)$ ;  $P(0 \leq Y \leq 15)$ ;  $P(Y \geq 5)$ ;  $P(Y \geq 15)$ .

**Ejercicio 3.7.2** Dada una distribución normal de  $Y$  con media 20 y varianza 25, encontrar  $Y_0$  tal que  $P(Y \leq Y_0) = .025$ ;  $P(Y \leq Y_0) = .01$ ;  $P(Y \leq Y_0) = .95$ ;  $P(Y \geq Y_0) = .90$ .

### 3.8 Distribución de medias

Cuando se muestrea una población, es costumbre resumir los resultados mediante el cálculo de  $\bar{Y}$  y otros estadígrafos. Un muestreo continuado genera una población de  $\bar{Y}$  con su propia media y varianza;  $\bar{Y}$  es una muestra de una observación de esa nueva población. La población de la cual se hizo el muestreo se suele llamar *población principal* o *distribución principal*; una población de medias muestrales, como otras poblaciones estadísticas, se llama *distribución derivada*, ya que se obtiene por muestreo de una población principal. Ya se ha dicho que la media y la varianza de una población de medias de  $n$  observaciones son la media y  $1/n$ -ésimo de la varianza de la población principal, es decir,  $\mu_{\bar{Y}} = \mu$  y  $\sigma_{\bar{Y}}^2 = \sigma^2/n$ . La ausencia de subíndice indica un parámetro de la población principal.

Considérese un tipo de problema que se presenta en experimentos de muestreo de poblaciones normales; a saber: ¿cuál es la probabilidad de que la media de una muestra sea mayor que un valor dado?

*Ejemplo* Dada una muestra aleatoria con  $n = 16$  observaciones, extraída de una población normal con  $\mu = 10$  y  $\sigma^2 = 4$ , hallar  $P(\bar{Y} \geq 11)$ .

La media muestral es una muestra de tamaño tomada de una población normal con

$$\mu_{\bar{Y}} = \mu = 10$$

$$\sigma_{\bar{Y}}^2 = \frac{\sigma^2}{n} = \frac{4}{16} = 1/4 \quad \text{y} \quad \sigma_{\bar{Y}} = \sqrt{1/4} = 1/2$$

Hallar en la tabla A.4

$$Z = \frac{\bar{Y} - \mu_{\bar{Y}}}{\sigma_{\bar{Y}}} = \frac{11 - 10}{1/2} = 2$$

Ahora  $P(Z \geq 2) = 0.0228$ , o sea,  $P(\bar{Y} \geq 11) = 0.0228$ , porque 11 corresponde a dos desviaciones estándar a la derecha de la media de la población de las  $\bar{Y}$ . También  $P(|Z| \geq 2) = 2(0.0228) = 0.0456$ . Por tanto

$$P(-2 \leq Z \leq 2) = P(9 \leq \bar{Y} \leq 11) = 1 - .0456 = .9544.$$

Una  $\bar{Y}$  muestral menor que 9 o mayor que 11, probablemente ocurra sólo 5 veces, 100(0.0456), en 100 o 1 vez en 20, en promedio, lo cual se considera generalmente como no usual. Aproximadamente 95 veces en 100 ó 19 veces en 20, en promedio, las  $\bar{Y}$  aleatorias de esta población se encuentran en el intervalo (9, 11).

En el párrafo anterior hemos definido “no usual” para propósitos estadísticos. Si un evento aleatorio sólo ocurre aproximadamente una vez en 20, estamos de acuerdo en considerar el evento como no usual. Ocasionalmente revisamos nuestra definición para circunstancias particulares, pero esta definición es simplemente la más común.

**Ejercicio 3.8.1** Dada  $Y$  con distribución normal, con media 10 y varianza 36, se extrae una muestra de 25 observaciones. Encontrar  $P(\bar{Y} \geq 12)$ ;  $P(\bar{Y} \leq 9)$ ;  $P(8 \leq \bar{Y} \leq 12)$ ;  $P(\bar{Y} \geq 9)$ ;  $P(\bar{Y} \leq 11.5)$ .

**Ejercicio 3.8.2** Dada  $Y$  con distribución normal con media 2 y varianza 9, para una muestra de 16 observaciones encontrar  $\bar{Y}_0$  tal que  $P(\bar{Y} \leq \bar{Y}_0) = .75$ ;  $P(\bar{Y} \leq \bar{Y}_0) = .20$ ;  $P(\bar{Y} \geq \bar{Y}_0) = .66$ ;  $P(\bar{Y} \geq \bar{Y}_0) = .05$ .

### 3.9 Distribución $\chi^2$

Ahora se expondrá la distribución  $\chi^2$  (letra griega, Ji, léase ji cuadrada) debido a su relación con  $s^2$  y la muy importante distribución  $t$  de Student, que será el tema de la sección siguiente. La ji cuadrada se define como la suma de los cuadrados de variables independientes, normalmente distribuidas con medias 0 y varianzas 1. La sec. 3.6 se ocupó exclusivamente de una variable normal con media 0 y varianza 1, mientras la sec. 3.7 indica cómo transformar una variable normal en otra de media 0 y varianza 1. Por lo tanto tenemos

$$\chi^2 = \sum_i Z_i^2 = \sum_i \left( \frac{Y_i - \mu_i}{\sigma_i} \right)^2 \quad (3.13)$$

La ecuación (3.13) es más general de lo que necesitamos actualmente, pues estamos viendo el muestreo de una sola población con  $\sigma$  constante. Al muestrear una distribución normal, la cantidad  $SC = (n - 1)s^2$  consiste en la suma de los cuadrados de  $(n - 1)$  desviaciones independientes, tal como se dijo en la sec. 2.8. Se puede demostrar que tales desviaciones tienen medias cero; la división por la  $\sigma$  común asegura que tengan varianzas unitarias. Así

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2} \quad (3.14)$$

es un caso particular de la ec. (3.13) y es la ecuación que nos interesa ahora.

La distribución  $\chi^2$  depende del número de desviaciones independientes, es decir, de los grados de libertad. Para cada número de grados de libertad hay una distribución  $\chi^2$ . Algunas curvas de ji cuadrada se presentan en la fig. 3.8. Obviamente  $\chi^2$  no puede ser negativa, ya que es una suma de números al cuadrado. Se ve que mientras los máximos se desfasan hacia la izquierda de los grados de libertad, las curvas tienden a ser más simétricas, al aumentar los grados de libertad. La media y la varianza de una distribución  $\chi^2$  son los grados de libertad y dos veces los grados de libertad respectivamente.

Se acostumbra tabular solamente unos cuantos valores de cada una de muchas curvas. Así, tenemos la tabla A.5. Las probabilidades se dan en la parte superior de la tabla, los grados de libertad en la columna de la izquierda y los valores de  $\chi^2$  en el cuerpo de la tabla para las combinaciones dadas de  $P$  y gl.

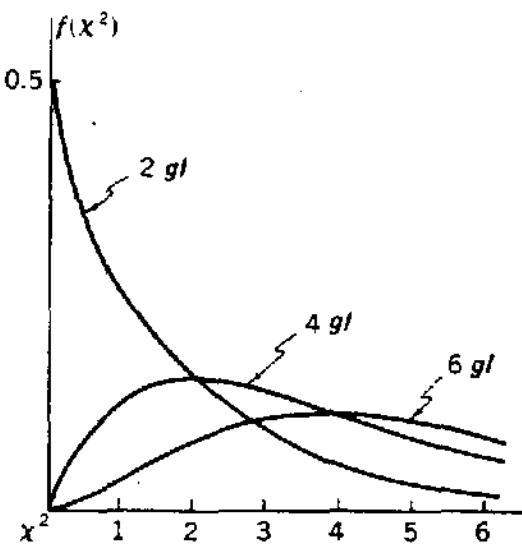


Figura 3.8 Distribución de  $\chi^2$  para 2, 4 y 6 grados de libertad.

*Ejemplo* Encontrar el valor aleatorio de  $\chi^2$  con 15 grados de libertad que sea excedido con una probabilidad de 0.25, esto es, encontrar  $\chi_1^2$  tal que

$$P(\chi^2 \geq \chi_1^2) = .25$$

Utilice la tabla A.5 para los 15 grados de libertad y lea bajo la columna encabezada 0.250. Allí  $\chi^2 = 18.2$  y  $P(\chi^2 \geq 18.2) = 0.25$ .

*Ejemplo* Encontrar la probabilidad de que se exceda un valor observado  $\chi^2 = 13.1$  con 10 grados de libertad.

Hallar 10 grados de libertad en la tabla A.5 y buscar el número 13.1. El valor cae entre 12.5 y 16.0, valores de  $\chi^2$  que son excedidos con probabilidades entre 0.25 y 0.10. Así pues  $P(\chi^2 \geq 13.1) > .10$ .

Estos ejemplos ilustran los problemas que más a menudo se presentan. En ellos sólo entra el uso de la cola derecha de la distribución, a diferencia de la distribución de  $Z$  donde pueden interesar tanto ambas colas como una sola.

La distribución de  $\chi^2$  con un grado de libertad se relaciona directamente con la distribución normal. Considérese  $\chi^2$  con un grado de libertad para  $P = 0.10$ . En nuestra forma abreviada,  $P(\chi^2 \geq 2.71) = 0.10$ , según la tabla A.5. Como por definición este  $\chi^2$  es el cuadrado de una simple desviación normal con media cero y varianza uno,  $\sqrt{\chi^2}$  debe ser una desviación normal con media cero y varianza uno. Así si vamos a la tabla A.4 con  $Z = \sqrt{2.71} = 1.645$ , hallaríamos la probabilidad de obtener que un mayor valor absoluto sea  $(1/2)(0.10) = 0.05$ . A partir de la tabla A.4,  $P(Z \geq 1.64) = 0.0505$  y  $P(Z \geq 1.65) = 0.0495$ . Así, toda la tabla normal A.4 está condensada en una línea de la tabla A.5, la correspondiente a un grado de libertad. Nótese que los valores de  $Z$  de ambas colas de la distribución normal van a la cola superior de  $\chi^2$  con un grado de libertad debido que al elevar el cuadrado desaparece el signo menos, mientras que valores de  $Z$  cercanos a cero, sean positivos o negativos, van a un  $\chi^2$  con un grado de libertad en la cola cercana a cero.

Generalmente los valores cercanos a cero no son de especial interés, así que la tabla de  $\chi^2$  se usa ordinariamente con más énfasis en valores grandes.

**Ejercicio 3.9.1** Encontrar un  $\chi^2_0$  tal que:  $P(\chi^2 \geq \chi^2_0) = 0.05$  para 10 grados de libertad;  $P(\chi^2 \geq \chi^2_0) = 0.01$  para 12 grados de libertad;  $P(\chi^2 \geq \chi^2_0) = 0.50$  para 25 grados de libertad;  $P(\chi^2 \leq \chi^2_0) = 0.025$  para 18 grados de libertad

**Ejercicio 3.9.2** Encontrar  $P$  tal que:  $P(\chi^2 \geq 17.01)$  para 6 grados de libertad;  $P(\chi^2 \geq 6.5)$  para 10 grados de libertad;  $P(\chi^2 \geq 20)$  para 4 grados de libertad.  $P(\chi^2 \leq 3.8)$

### 3.10 Distribución $t$ de Student

William Sealy Gosset, 1876-1937, cervecero o estadístico según se mire, escribió muchos estudios estadísticos bajo el seudónimo de Student. Reconoció que el uso de  $s$  en vez de  $\sigma$  en el cálculo de los valores  $Z$  para su empleo en las tablas normales no era de confiar en el caso de muestras pequeñas, y que se necesitaría otra tabla. Se interesó por una variable muy relacionada con la variable  $t = (\bar{Y} - \mu)/s_{\bar{Y}}$  expresión en que entran dos estadígrafos,  $\bar{Y}$  y  $s_{\bar{Y}}$ , en vez de  $Z = (\bar{Y} - \mu)/\sigma_{\bar{Y}}$ , con uno. Ahora, el estadígrafo

$$t = \frac{\bar{Y} - \mu}{s_{\bar{Y}}} = \frac{\bar{Y} - \mu}{\sqrt{s^2/n}} \quad (3.15)$$

para muestras de distribuciones normales, se conoce universalmente como  $t$  de Student.

Como la  $\chi^2$ ,  $t$  tiene una distribución diferente para cada valor de los grados de libertad. De nuevo, nos contentamos con una tabla abreviada, la tabla A.3, con valores de  $t$  en vez de probabilidades, en el cuerpo de la tabla. En la parte superior la tabla A.3 da las probabilidades para mayores valores de  $t$  sin tener en cuenta el signo. Estas son las que a menudo se llama probabilidades de dos colas. Por ejemplo, para una muestra aleatoria de tamaño 16, en la línea de los  $gl = 16 - 1 = 15$  y la columna encabezada por 0.05, encontramos que  $P(|t| \geq 2.131) = 0.05$ . La tabla A.3 da probabilidades de encontrar valores mayores de  $t$ ; estas se pueden llamar probabilidades de una cola. Así, para una muestra aleatoria de tamaño 16, en la línea para 15 gl y la columna con 0.025 en la parte inferior, encontramos que  $P(t \geq 2.131) = 0.025 = P(t \leq -2.131)$ .

La curva de  $t$  es simétrica, como se puede deducir por los anteriores ejemplos. Es un poco más aplanada que la distribución de  $Z = (\bar{Y} - \mu)/\sigma_{\bar{Y}}$  situándose un poco por debajo de  $Z$  en el centro y por encima de ella en las colas. A medida que crecen los grados de libertad, la distribución de  $t$  se approxima a la normal. Esto puede verse luego de un examen de las entradas de la tabla A.3, ya que la última fila,  $gl = \infty$  es la de una distribución normal, y los valores en toda columna se acercan evidentemente al valor correspondiente de esta distribución.

Una propiedad importante de  $t$  para muestras de poblaciones normales es que sus componentes, esencialmente  $\bar{Y}$  y  $s$ , no muestran indicios de una variación conjunta. O sea que si se recolectan muchas muestras del mismo tamaño, se calculan  $\bar{Y}$  y  $s$  y se representan gráficamente los pares de valores resultantes con  $\bar{Y}$  y  $s$  en los ejes, los puntos se dispersan

de tal forma que no dan muestras de relación alguna, tal que grandes medias estén asociadas con desviaciones estándar grandes. Para una distribución distinta de la normal, se presenta cierto tipo de relación entre los valores muestrales de  $\bar{Y}$  y  $s$  en un muestreo repetido.

**Ejercicio 3.10.1** Encontrar  $t_0$  tal que  $P(t \geq t_0) = 0.025$  para 8 grados de libertad;  $P(t \leq t_0) = 0.01$  para 15 grados de libertad;  $P(|t| \geq t_0) = 0.01$  para 15 grados de libertad; 0.10 para 12 grados de libertad;  $P(-t_0 \leq t \leq t_0) = 0.80$  para 22 grados de libertad.

**Ejercicio 3.10.2** Encontrar  $P(t \geq 2.6)$  para 8 grados de libertad;  $P(t \leq 1.7)$  para 15 grados de libertad;  $P(t \leq 1.1)$  para 18 grados de libertad;  $P(-1.1 \leq t \leq 2.1)$  para 5 grados de libertad;  $P(|t| \geq 1.8)$  para 6 grados de libertad.

### 3.11 Estimación e inferencia

Lo visto hasta el momento ha tenido que ver con muestreo de poblaciones conocidas. En general, se desconocen parámetros de población aunque se pueden plantear hipótesis respecto a sus valores. La estadística se ocupa en gran medida con la toma de inferencias de parámetros poblacionales, inferencias que son inciertas debido a que se basan en comprobaciones obtenidas de las muestras.

Considérese el problema de la estimación de parámetros. Por ejemplo, se puede desear conocer la producción media de una variedad de trigo en su madurez, o el tiempo promedio para recuperarse de un resfriado. Es bien claro que  $\bar{Y}$  es un estimativo de  $\mu$  y que  $s^2$  es un estimativo de  $\sigma^2$ , especialmente si aceptamos la idea, propuesta en la sec. 2.12, del modelo aditivo lineal. Naturalmente, éstos no son los únicos estimativos de esos parámetros. ¿Hasta dónde son estos estadígrafos buenos estimativos en esos parámetros? A menos que se conozcan los parámetros, no se puede saber la bondad de estimación que alcanza un parámetro con un estadígrafo muestral; hay que conformarse con saber lo bueno que es ese estimativo en promedio, esto es, saber lo bien que se comporta en un muestreo repetido, o saber cuantos valores muestrales se puede esperar que caigan dentro de un intervalo dado en torno a un parámetro. Por ejemplo, considérense 3 estadígrafos o fórmulas estimativas de  $\mu$ . (La mediana, la media, y el punto medio de la amplitud, son tres estimadores aunque no necesariamente los del ejemplo). Denótense esos estimadores con  $\hat{\mu}_1$ ,  $\hat{\mu}_2$  y  $\hat{\mu}_3$  donde  $\hat{\phantom{x}}$  (llámelo "sombrero") indica un estimativo y no el propio parámetro. Estas fórmulas se conocen como *estimadores* o también estadígrafos. Todos los posibles valores de  $\hat{\mu}_1$ ,  $\hat{\mu}_2$  y  $\hat{\mu}_3$  pueden generar distribuciones tales como las de la fig. 3.9, donde  $\hat{\mu}_1$  da valores bastante coherentes, o sea, que tienen una varianza relativamente pequeña, pero no están centrados en  $\mu$ ;  $\hat{\mu}_2$  da valores consistentes centrados en  $\mu$ ;  $\hat{\mu}_3$  da valores centrados en  $\mu$ , pero más bien muy dispersos. El problema ahora consiste en seleccionar la "mejor" fórmula, pero primero debemos definir el concepto "mejor".

En vez de definir "mejor", considérense varias propiedades deseables y tratése de tener el mayor número de ellas asociadas con la elección de un estimador. Por ejemplo, *insesgamiento* exige que la media de todas las posibles estimaciones que da un estimador, es decir, la media de la población de estimativos, sea el parámetro que se estima. La media de una población de  $\bar{Y}$  es  $\mu$ , el parámetro que se estima para la población principal, de modo que  $\bar{Y}$  es un estimador no sesgado de  $\mu$ . La media de una población es  $s^2$ , o sea,  $\mu_{s^2}$ , es  $\sigma^2$ , así que  $s^2$  es un estimador no sesgado de  $\sigma^2$ . Sin embargo, si el divisor de la suma

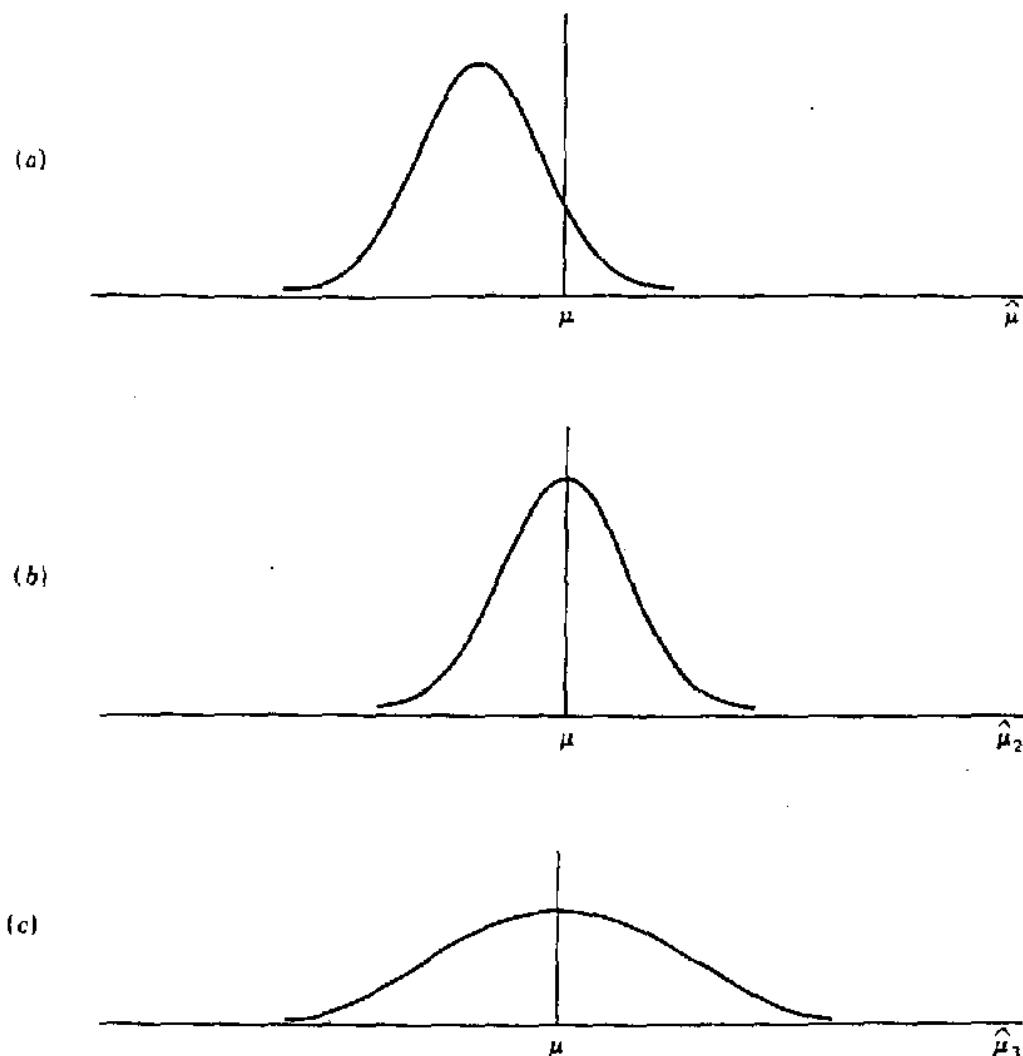


Figura 3.9 Ilustración de la selección de un estimador.

de cuadrados ha sido  $n$  en vez de  $n - 1$ , entonces la estimación es sesgada. El sesgo no es un problema grave si se conoce su magnitud. Este sería el caso si  $n$  fuese el divisor en la estimación de  $\sigma^2$ . El sesgo es serio cuando se desconoce, ya que no se puede hacer ningún tipo de corrección para el mismo.

Otra propiedad deseable, es la de tener una varianza pequeña. Lo ideal *varianza mínima*. En la fig. 3.9,  $\hat{\mu}_3$  presenta una varianza muy grande, mientras que  $\hat{\mu}_1$  y  $\hat{\mu}_2$  tienen varianzas comparativamente pequeñas y se prefieren con esa base. Se dice, entonces, que los estimadores  $\hat{\mu}_1$  y  $\hat{\mu}_2$  son más eficientes que  $\hat{\mu}_3$ .

La sencillez del cálculo constituye otra propiedad deseable. Toda estimación que se encuentra mediante adición y substracción de múltiplos de observaciones se llama *función lineal* de ellas. La media es una función lineal que requiere de un  $n$ -ésimo de cada observación; la varianza y la desviación estándar no son funciones lineales. Es claro que las funciones lineales son fáciles de calcular.

Si para un parámetro podemos encontrar estimadores lineales insesgados y si entre todos esos estimadores hay uno con varianza mínima, entonces se dice que es el mejor

*estimador lineal insesgado*, o el *m.e.l.i.* Ya hemos visto que no siempre insistimos en tener tales estimadores.

A pesar de que  $\bar{Y}$  y  $s^2$  son estimaciones de  $\mu$  y  $\sigma^2$ , sería sorprendente que fueran realmente  $\mu$  y  $\sigma^2$  en vez de encontrarse en la vecindad de éstos. Esto sugiere que puede ser más apropiado dar un intervalo en torno a  $\bar{Y}$  o a  $s^2$  y decir que estamos razonablemente confiados en que  $\mu$  y  $\sigma^2$ , se encuentran en ese intervalo. Esto puede hacerse con la ayuda de la distribución  $t$  de la sec. 3.10.

Para  $\mu$  y  $\sigma^2$  dados, es posible definir un intervalo sobre el eje  $\bar{Y}$  y dar la probabilidad de obtener un valor aleatorio de  $\bar{Y}$  en el intervalo; deseamos invertir el proceso y para  $\bar{Y}$  y  $s^2$  dados definir un intervalo y establecer la probabilidad de que  $\mu$  se encuentre en ese intervalo. Como  $\mu$  podrá estar o no en el intervalo, esto es,  $p = 0$  ó  $1$ , la probabilidad efectivamente será una medida de la confianza puesta en el procedimiento que llevó a la afirmación anterior. Es como lanzar un anillo a un poste fijo; no cae en la misma posición ni tampoco cae en el poste todas las veces. Sin embargo, podemos decir que es posible ensartarlo en el poste 9 veces de un total de 10, o el valor que sea la medida de la confianza en nuestra habilidad.

Para invertir el proceso, comenzamos con un enunciado probabilístico como el siguiente

$$P\left(-t_{.025} \leq \frac{\bar{Y} - \mu}{s_{\bar{Y}}} \leq t_{.025}\right) = .95 \quad (3.16)$$

con respecto a la variable aleatoria  $t = (\bar{Y} - \mu)/s_{\bar{Y}}$ . Dice que la probabilidad  $P$  de que la variable aleatoria  $t = (\bar{Y} - \mu)/s_{\bar{Y}}$  se encuentre en  $-t_{.025}$  y  $+t_{.025}$  es 0.95. Primero nótese que  $t$  es tal cual la ec. (3.12) pero reemplazando la desviación estándar poblacional por su estimativo a partir de la muestra. Esto deja únicamente en la variable aleatoria, el parámetro desconocido, o sea  $\mu$ . Segundo, el subíndice de  $t$  se refiere a la probabilidad de que un valor aleatorio de  $t$  caiga a la derecha del valor tabulado  $t_{.025}$ ; así la probabilidad de que caiga a la derecha de  $t_{.025}$  o la probabilidad de que caiga a la izquierda de  $-t_{.025}$  es 0.025.

Algebraicamente nos permite escribir la ec. (3.16) así

$$P(\bar{Y} - t_{.025} s_{\bar{Y}} < \mu < \bar{Y} + t_{.025} s_{\bar{Y}}) = .95 \quad (3.17)$$

Ahora, el enunciado dice que la probabilidad de que  $\mu$  se encuentra en el intervalo aleatorio  $(\bar{Y} - t_{.025} s_{\bar{Y}}, \bar{Y} + t_{.025} s_{\bar{Y}})$  es 0.95. El intervalo está todavía por verse;  $\mu$  continúa fijo.

Un ejemplo puede aclarar más esto. Se extrajo una muestra, número 1, tabla 4.3, de una población conocida y se calcularon los estadígrafos  $\bar{Y}$ ,  $s^2$  y  $t$ . La ec. (3.17) conduce al enunciado

$$P(24.77 \leq \mu \leq 48.03) = .95$$

En efecto,  $\mu = 40$ , así que sabemos con certeza ( $p = 1$ ) que  $\mu$  se encuentra en ese intervalo particular. Por lo tanto, si aceptamos que el enunciado es probabilístico, entonces debemos interpretarlo diciendo que la probabilidad 0.95 mide la confianza que se tiene en el proceso que nos conduce a tales intervalos. A 0.95 lo llamaremos *coeficiente de confianza*. La costumbre de escribir  $P_F$  en vez de  $P$  cuando se usan resultados de muestras, llama la atención sobre la naturaleza de la probabilidad que interviene aquí; la letra  $F$  se refiere a *fiducial* o sea de confianza.

En una situación real, un experimentador no conoce la media de la población que se muestrea; el problema consiste en estimar la media de la población y establecer el grado de confianza asociado con el estimativo. El procedimiento del intervalo de confianza es una solución a este problema.

*Ejemplo* En un ensayo sobre prolactina presentado por Finney (3.1), se midió la respuesta como el peso del producido glandular en palomas, en 0.1g. Para una dosis de la preparación de prueba de 0.125 se obtuvieron 4 pesos: 28, 65, 35 y 36. El problema consiste en estimar la media poblacional.

Supóngase que los datos constituyen una muestra aleatoria de una población normal con media y varianza desconocidas, o sea, la población de pesos glandulares producidos por todas las posibles palomas que se pudieran mantener confinadas bajo condiciones de laboratorio para el experimento y con una dosis de 0.125 mg de la preparación de la prueba. Los estadígrafos de la muestra son  $\bar{Y} = 41$  dg,  $s^2 = 269$ ,  $s_y = 8.2$  dg. El  $t_{0.025}$  tabulado con 3 grados de libertad es 3.18. Una estimación del intervalo de confianza para la media poblacional es  $\bar{Y} \pm t s_y = 41 \pm (3.18) \times (8.2) = (15, 67)$  dg; usamos 0.95 como medida de nuestra confianza en la afirmación de que  $\mu$  se encuentra en el intervalo (15, 67) dg al menos que la muestra aleatoria dada sea poco común. Una muestra poco común lleva a una afirmación falsa respecto a la localización de  $\mu$ ; esto ocurre una vez en 20 en promedio, es decir, 5 por ciento de las veces. Por lo tanto, si  $\mu$  no se encuentra dentro de (15, 67) dg es porque nuestra muestra es poco común.

En nuestro ejemplo, tenemos un *coeficiente de confianza* de 0.95 y una *tasa de error* de 5 por ciento. La elección de la tasa de error depende obviamente de la gravedad de una decisión incorrecta.

Si existiesen razones a priori para esperar que la media poblacional fuese 20 dg, esta muestra no constituiría evidencia al 5 por ciento de *nivel de significancia*, para tener alguna duda sobre tal expectativa. Si existiesen razones a priori para esperar una media poblacional de 80 dg, esta muestra sería prueba que tendería a negar tal expectativa al *nivel de significancia* de 5 por ciento.

Cuando se conoce  $\sigma^2$ , es posible escoger un tamaño de muestra tal que el intervalo de confianza, una vez obtenido, sea de una longitud predeterminada para la tasa de error escogida. Cuando se desconoce  $\sigma^2$ , es posible determinar un tamaño de muestra tal que el experimentador puede estar razonablemente confiado (el grado de confianza puede fijarse) de que el intervalo de confianza que se va a calcular, para una tasa de error escogida, no será mayor que un valor predeterminado. Además, se puede usar un método de muestreo, llamado secuencial, para obtener un intervalo de confianza de longitud establecida.

Ejercicio 3.11.1 Finney (3.1) también informó sobre los pesos de producción glandular en palomas, en 0.1 g para dosis de preparación de prueba de 0.250 y 0.500 mg. Los pesos resultantes fueron 48, 47, 54, 74 y 60, 130, 83 y 60 respectivamente. Definir las poblaciones para estos dos conjuntos de datos y estimar las medias poblacionales usando intervalos de confianza del 95 y 99 por ciento. Nótese la diferencia de longitud para cada una de las poblaciones.

### 3.12 Predicción de resultados de muestras

Suele oírse la afirmación errónea de que el enunciado probabilístico en relación con la media de la población dice algo sobre la distribución de futuras medias muestrales, por ejemplo, que el 95 por ciento de las medias de las muestras futuras se encuentren dentro de un intervalo dado. La mayoría de estos tipos de afirmaciones son desorientadoras e incorrectas. Sin embargo, es posible hacer una afirmación respecto a una observación muestral futura, que puede ser muy útil.

Considérese el problema de una predicción hidrológica. El uso y desarrollo de recursos de agua ha creado una seria demanda de estimaciones anticipadas de los caudales de las corrientes de agua en una cuenca. Es decir, se desea predecir un evento u observación futuros. La población de todos los posibles caudales suministrados por una corriente determinada en una cuenca, naturalmente, se desconoce, pero sus valores correspondientes para los años procedentes constituyen una muestra de la población. Obviamente ésta no es una muestra aleatoria de observaciones independientes, pero podemos empezar por suponer que lo es. (La experiencia de comprobar predicciones con resultados es un método para juzgar la medida de fiabilidad puesta en nuestras afirmaciones. Otro método es usar todas las observaciones excepto la más reciente al hacer la predicción, luego comprobar la predicción con la observación no utilizada. Si esto se puede hacer con un número suficiente de cuencas similares, entonces se puede aceptar lo dicho sobre la fiabilidad de la tasa de error enunciada antes de hacer pública la afirmación probabilística).

Si se conoce la media de una población, podemos usarla como valor de predicción, un punto estimativo únicamente; no tendría sentido adivinar cuál sería el valor aleatorio en sí mismo. Sin embargo, una estimación de intervalo o la predicción del siguiente valor aleatorio puede hacerse mediante el uso de la tabla normal o la tabla de  $t$ . En el uso de un intervalo se trata de dejar margen al error aleatorio  $\epsilon$  exigido en el modelo lineal aditivo, ec. (2.13) y que con seguridad siempre está presente.

Si  $\mu$  y  $\sigma^2$  se conocieran, entonces el intervalo y el enunciado de confianza podrían ser

3/9  
A/0 50

$$P(\mu - Z_{.005} \sigma \leq Y \leq \mu + Z_{.005} \sigma) = .99$$

dado que el 99 por ciento de las observaciones aleatorias caen dentro de  $Z_{.005}$  desviaciones estándares a partir de la media (ver sec. 3.7).

En el problema corriente,  $\mu$  y  $\sigma^2$  se desconocen, pero se dispone de estimaciones  $\bar{Y}$  y  $s^2$ . La predicción del próximo  $Y$ , por ejemplo, caudal del año venidero, es necesariamente  $\bar{Y}$ . La varianza apropiada para este valor de predicción de  $Y$  es la suma de la media muestral observada más una componente aleatoria, esto es,  $(s^2/n) + s^2 = [(n+1)/n]s^2$ .

Así, el enunciado de confianza apropiado es

$$P\left[\bar{Y} - t_{.005} \sqrt{s^2 \left(\frac{n+1}{n}\right)} \leq Y \leq \bar{Y} + t_{.005} \sqrt{s^2 \left(\frac{n+1}{n}\right)}\right] = .99 \quad (3.18)$$

donde  $t_{.005}$  es el valor tabulado de la  $t$  de Student, tal que la probabilidad de encontrar un valor positivo mayor es 0.005.

**Ejercicio 3.12.1** En la sección 3.4 se usa el término *función de  $Y$* . Dar 6 funciones de  $Y$  usadas en este capítulo que parezcan de importancia en la estadística

**Ejercicio 3.12.2** Distinguir claramente entre estimación de la media de una población y predicción del valor que va a ser observado. Tratar de especificar situaciones para las cuales cada procedimiento sea más apropiado.

## Referencias

3.1. Finney, D. J.: *Statistical method in biological assay*, Hafner, Nueva York, 1952, Tabla 12.1.

## Ejercicio de laboratorio propuesto en relación con el capítulo 3

**Propósito** (1) Dar práctica al estudiante en el cálculo de algunos estadígrafos corrientes y en el uso de las tablas respectivas. (2) Construir más comprobación empírica respecto a los estadígrafos muestrales para comparación con resultados teóricos. (Ver los ejercicios de laboratorio de los capítulos 2 y 4).

Para cada una de sus 10 muestras aleatorias (ver ejercicio de laboratorio del capítulo 2) calcular

a)  $Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$ , usando  $\mu = 40$ ,  $\sigma = 12$

b)  $t = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$

c) El intervalo de confianza del 95 por ciento para  $\mu$

d) El intervalo de confianza del 99 por ciento para  $\mu$

e)  $\chi^2 = [(n-1)s^2]/\sigma^2$

---

CAPITULO  
**CUATRO**

---

## MUESTREO DE UNA POBLACION NORMAL

### 4.1 Introducción

Se ha estudiado el cálculo de los estadígrafos corrientes  $\bar{Y}$ ,  $s^2$  y  $s$  como medidas de tendencia central y dispersión. Cuando se calculan a partir de muestras aleatorias,  $\bar{Y}$  y  $s^2$  son estimadores no sesgados (sec. 3.11) de los parámetros de  $\mu$  y  $\sigma^2$  de la población principal y  $s$  es un estimador sesgado de  $\sigma$ .

La desviación estándar de las medias puede estimarse a partir de una muestra de observaciones mediante fórmula  $s_{\bar{Y}} = s/\sqrt{n}$  donde  $s$  es la desviación estándar de la muestra. En la sección 3.11 se vio el uso de  $\bar{Y}$ ,  $s_{\bar{Y}}$  y de un valor tabulado de la  $t$  de Student para establecer el intervalo de confianza para la media de la población. El número promedio, expresado en fracción decimal, de intervalos que contiene a  $\mu$ , se llama *probabilidad de confianza* o *coeficiente de confianza*.

Los resultados hasta ahora se basan en teoremas y principios matemáticos. Tales resultados pueden demostrarse con un grado razonable de exactitud mediante procedimientos de muestreo a gran escala. Este es un *método empírico*. En este capítulo, se usa el muestreo como un método de examen de la distribución de varios estadígrafos y el procedimiento del intervalo de confianza.

### 4.2 Una población con distribución normal

Una población normal tiene una variable continua en un intervalo infinito; por consiguiente, una observación puede tomar cualquier valor real, positivo o negativo. La tabla 4.1 consiste en los rendimientos en libras de grasa de leche de 100 vacas Holstein, con los datos originales un tanto modificados para formar una distribución aproximadamente normal. En dos aspectos principales se apartan los datos de la normalidad: la variable tiene un intervalo finito y es discreta o discontinua. Los efectos debidos al intervalo finito y lo discreto de los datos son pequeños en comparación con la variación muestral y en consecuencia tendrán poco efecto sobre las inferencias basadas en las muestras.

**Tabla 4.1 Ordenamiento en libras, de la grasa de leche producida por 100 vacas Holstein durante un mes**

Los datos originales se modificaron para que se aproximan a una distribución normal con  $\mu = 40$  lb y  $\sigma = 12$  lb.

Unidad	Libras	Unidad	Libras	Unidad	Libras	Unidad	Libras
00	10	25	33	50	40	75	47
01	12	26	33	51	40	76	48
02	14	27	34	52	41	77	48
03	15	28	34	53	41	78	48
04	17	29	34	54	41	79	49
05	18	30	35	55	41	80	49
06	20	31	35	56	42	81	49
07	22	32	35	57	42	82	50
08	23	33	36	58	42	83	50
09	25	34	36	59	42	84	51
10	26	35	36	60	43	85	51
11	27	36	37	61	43	86	52
12	28	37	37	62	43	87	52
13	28	38	37	63	43	88	53
14	29	39	37	64	44	89	54
15	29	40	38	65	44	90	55
16	30	41	38	66	44	91	57
17	30	42	38	67	45	92	58
18	31	43	38	68	45	93	60
19	31	44	39	69	45	94	62
20	31	45	39	70	46	95	63
21	32	46	39	71	46	96	65
22	32	47	39	72	46	97	66
23	32	48	40	73	47	98	68
24	33	49	40	74	47	99	70
						Total	4,000

Las características sobresalientes de la distribución se pueden observar en las figs. 4.1 y 4.2. La fig. 4.1 es un histograma de los datos con libras en el eje horizontal y frecuencias en el vertical. Los valores se concentran en el centro y se hacen menos densos simétricamente a ambos lados, al principio rápido y después más lentamente. La fig. 4.2 muestra los 100 valores acumulados. Por ejemplo, para encontrar el número de observaciones (posición en el cuadro) menores que cierto peso en libras, basta trazar una vertical desde ese peso hasta la curva suave de puntos que representan las observaciones, y luego una horizontal hasta el eje de las ordenadas donde se puede leer el número. La relación del histograma con el cuadro es que la altura del rectángulo en toda clase del histograma es proporcional al número de puntos que caen entre el par correspondiente de las verticales del cuadro.

La tabla 4.2 corresponde a una distribución en libras de grasa de leche para los datos de la tabla 4.1. Cada clase tiene amplitud 5 lb.

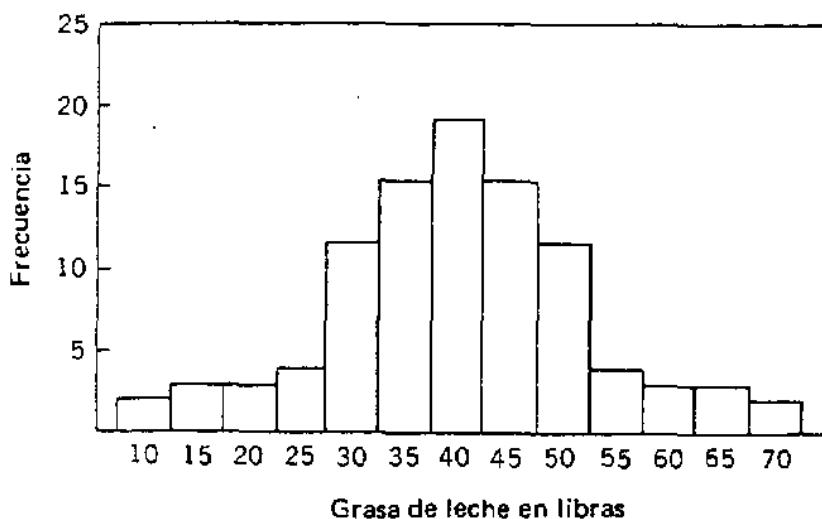


Figura 4.1 Histograma de la distribución de las libras de grasa de leche de 100 vacas Holstein.

**Ejercicio 4.2.1** En una distribución normal, ¿la variable es discreta? ¿Continua? ¿Cualitativa? ¿Cuantitativa? ¿De intervalo finito? ¿Existe un valor mínimo?

**Ejercicio 4.2.2** Para muestras aleatorias de distribuciones normales, ¿la  $\bar{Y}$  de la ec. (2.1),  $s^2$  de la ec. (2.7) y  $s = \sqrt{s^2}$  dan estimativos no sesgados de  $\mu$ ,  $\sigma^2$ , y  $\sigma$ ?

**Ejercicio 4.2.3** ¿Cuál es la distribución de las  $\bar{Y}$  aleatorias donde  $Y_i$  proviene de una distribución normal? ¿Cómo se relaciona la media y la varianza de una población de  $\bar{Y}$  con la media y la varianza de la población principal?

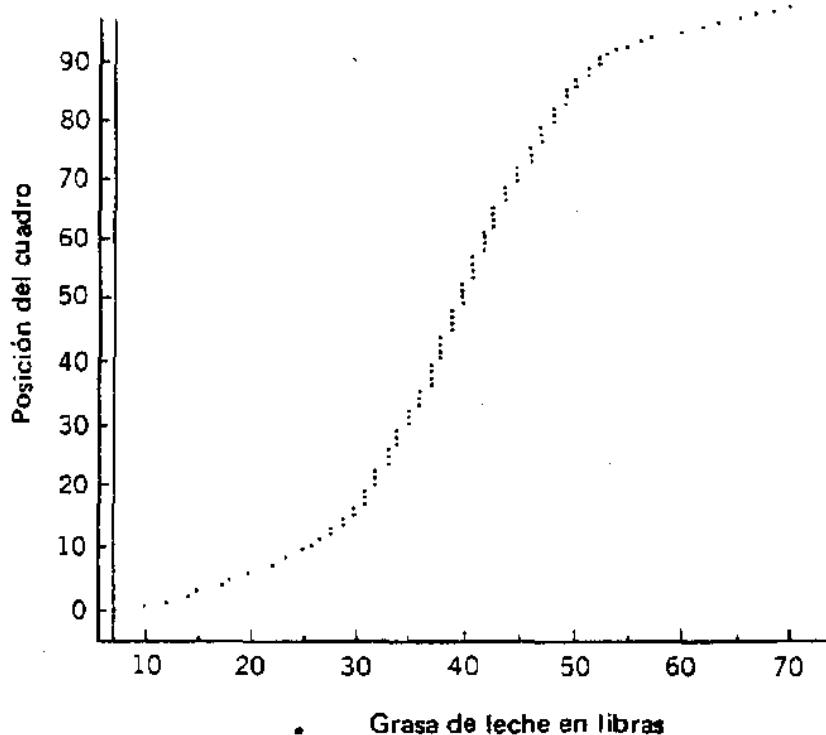


Figura 4.2 Representación gráfica del cuadro de las libras de grasa de leche de 100 vacas Holstein .

**Tabla 4.2 Una distribución de frecuencia de libras de grasa de leche de 100 vacas Holstein**

Punto medio o marca de clase	10	15	20	25	30	35	40	45	50	55	60	65	70
Frecuencias	2	3	3	4	12	16	20	16	12	4	3	3	2

### 4.3 Muestras aleatorias de una distribución normal

La toma de muestras aleatorias no puede dejarse a discreción subjetiva, sino que debe ser el resultado de métodos objetivos y preferiblemente mecánicos. Una tabla de números aleatorios, tal como la tabla A.1, sirve para introducir objetividad. Para facilitar la extracción de muestras aleatorias con una tabla semejante, se asignan números consecutivos a los individuos de una población. Por ejemplo, a las 100 producciones de la tabla 4.1 se les han asignado los números 00 y 99, citados como unidades.

El uso de una tabla de números aleatorios se ilustró en la sección 2.5. Para muestras sucesivas, los dos últimos pares de enteros se pueden usar para localizar la iniciación de las siguientes filas y columnas. Aplicado a la tabla 4.1 este procedimiento de muestreo, garantiza que cada unidad o rendimiento puede extraerse cualquier número de veces. El muestreo es siempre de la misma población y la probabilidad de extraer una unidad particular es la misma para todas las unidades. El procedimiento es esencialmente el mismo, tal como si se muestreara de una población infinita.

La tabla 4.3 da 5 muestras aleatorias, junto con ciertos cálculos pertinentes, que se obtuvieron mediante dicho proceso. Esas 5 muestras provienen de 500 muestras aleatorias de 10 observaciones de la tabla 5.1 usadas para la exposición en el resto del capítulo.

**Ejercicio 4.3.1** Los dos métodos de muestreo aleatorio dados en esta sección no son precisamente equivalentes. ¿Por qué? (Sugerencia: Ver la sección 2.5 para la distribución de dígitos en la tabla A.1).

### 4.4 Distribución de medias muestrales

De la tabla 4.1 se extrajeron 500 muestras aleatorias de 10 observaciones. La distribución de frecuencias de las 500 medias se presenta en la tabla 4.4 para un intervalo de clase de 1.5 lb e ilustra varios aspectos básicos del muestreo. Primero, la distribución de las medias es aproximadamente normal. La teoría dice que la distribución derivada de medias muestrales de observaciones aleatorias procedentes de una población normal es también normal. La teoría también dice que aun si la distribución principal es considerablemente anormal, la distribución de medias muestrales aleatorias se aproxima a la distribución normal, a medida que aumenta el tamaño de la muestra. Esto es muy importante en la práctica, porque la forma de la distribución principal rara vez se conoce. Segundo, el promedio de las 500 medias, 39.79 lb, es muy cercano a  $\mu = 40$  lb, la media de la población principal. Esto ilustra el insesgamiento. La media de la muestra se dice insesgada porque la media de todas las posibles medias de muestras es la media de la población principal. Tercero, la variación de las medias es mucho menor que la de los individuos, siendo 27 lb la

Tabla 4.3 Cinco muestras aleatorias de 10 observaciones tomadas de la tabla 4.1 junto con los estadígrafos muestrales

Número de observaciones y fórmulas		Número de muestra				
		1	2	3	4	5
	Unidad Rendimiento					
1	96	65	39	37	63	39
2	37	51	40	59	42	43
3	04	17	34	54	41	84
4	29	34	81	49	47	39
5	84	51	34	36	41	38
6	05	18	49	40	81	49
7	71	46	47	39	09	25
8	35	36	75	47	03	15
9	03	15	23	32	15	97
10	69	45	96	65	87	52
Suma = $\sum Y$	364		421		393	
Media = $\bar{Y}$	36.4		42.1		39.3	
$\sum Y^2$	15,626.00		18,541.00		17,175.00	
$CY = (\sum Y)^2/10$	13,249.60		17,724.10		15,444.90	
$SC = \sum Y^2 - (\sum Y)^2/10$	2,376.40		816.90		1,730.10	
$s^2 = SC/9$	264.04		90.77		192.23	
$s = \sqrt{s^2}$	16.2		9.5		13.9	
$s_y = \sqrt{s^2/10}$	5.14		3.01		4.38	
$t = (\bar{Y} - 40)/s_y$	-0.70		+0.70		-0.16	
$t_{0.025}s_y = 2.262s_y$	11.6		6.8		9.9	
L.C. (inferior) = $\bar{Y} - t_{0.025}s_y$	24.8		35.3		29.4	
L.C. (superior) = $\bar{Y} + t_{0.025}s_y$	48.0		48.9		49.2	

Para la población principal,  $\mu = 40$ ,  $\sigma^2 = 144$ .

**Tabla 4.4 Distribución de frecuencia de 500 medias de muestras aleatorias de 10 unidades tomadas de la tabla 4.1**

Marcia de clase *lb.	Frecuencia observada	Frecuencia teórica	Frecuencia acumulada observada	Frecuencia acumulada teórica
26.5	1	0	1	0.2
28.0	0	0.5	1	0.8
29.5	2	2	3	2.6
31.0	2	5	5	7.5
32.5	14	11	19	18.8
34.0	20	23	39	41.9
35.5	47	39	86	80.6
37.0	65	58	151	138.8
38.5	74	72	225	210.4
40.0	71	79	296	289.6
41.5	78	72	374	361.2
43.0	49	58	423	419.4
44.5	40	39	463	458.1
46.0	24	23	487	481.2
47.5	8	11	495	492.5
49.0	4	5	499	497.4
50.5	0	2	499	499.2
52.0	0	0.5	499	499.8
53.5	1	0	500	500.0
<hr/>				
Totales	500	500		
Media de medias: $\bar{Y} = 39.79$				

\* El centro de un intervalo de clase.

fluctuación de las medias y 60 lb la de los individuos. La teoría establece que  $\sigma_{\bar{Y}}^2 = \sigma^2/n$ ; aquí  $\sigma_{\bar{Y}}^2 = 14.4$  y  $\sigma_{\bar{Y}} = 3.79$  lb. La correspondiente relación para las muestras es  $s_{\bar{Y}}^2 = s^2/n$ . Al aplicar esto a las medias de las 500 varianzas muestrales, es decir,  $s^2$  (ver tabla 4.5), obtenemos  $s_{\bar{Y}}^2 = 140.4/10 = 14.04$  y  $s_{\bar{Y}} = 3.75$  lb. Los cálculos con las 500 medias dan  $s_{\bar{Y}} = \sqrt{[\sum \bar{Y}^2 - (\sum \bar{Y})^2/500]/499} = 3.71$  lb. En la tabla 4.4 se comparan las frecuencias observadas y teóricas. Las frecuencias teóricas son para una distribución normal con  $\mu = 40$  lb y  $\sigma = \sqrt{144/10} = 12/\sqrt{10} = 3.79$  lb, lo cual corresponde a  $\sigma_{\bar{Y}}$  en nuestro problema. En la columna de la frecuencia teórica la unidad usada fue 1/2; en la de frecuencia teórica acumulada la unidad fue 1/10. Esto explica las discrepancias entre columnas.

**Ejercicio 4.4.1** Dada una distribución normal principal y una distribución de medias derivada de 100 observaciones, ¿cuál es la relación entre las medias de las dos poblaciones? ¿Entre las varianzas? ¿Las amplitudes de las poblaciones son las mismas? Si se tuviera una muestra para cada población, una de 50 observaciones y la otra de 50 medias, ¿cómo esperaríamos que se compararán las dos amplitudes?

**Ejercicio 4.4.2** Para obtener un estimativo de  $\sigma_{\bar{Y}}^2$ , un método usado antes fue promediar las varianzas y luego dividir por  $n$ ; ejemplo, promediar las 500 varianzas muestrales y luego dividir

por 10. Compárese este procedimiento con el que en cada varianza  $s^2$  se divide por 10 y luego se promedian los 500 resultados.

#### 4.5 Distribución de varianzas muestrales y desviaciones estándar

Para cada una de las 500 muestras, se calcularon la varianza y la desviación estándar. El procedimiento y los resultados se dan para cinco muestras en la tabla 4.3.

La distribución de las 500 varianzas muestrales se da en la tabla 4.5. Hay una concentración de varianzas a la izquierda de su media, denotada por  $\bar{s}^2$ , y una menor concentración a la derecha. La distribución es simétrica. Las cantidades  $(n - 1)s^2/\sigma^2 = 9s^2/144$  se distribuyen como una  $\chi^2$  con  $(n - 1) gl = 9 gl$ . La varianza media es  $\bar{s}^2 = 140.4 \text{ lb}^2$ , muy aproximada a la varianza poblacional  $\sigma^2 = 144$ . Esto ilustra el insesgamiento de  $s^2$  como estimación de  $\sigma^2$ . Las varianzas individuales  $s^2$  van de 20 a 380  $\text{lb}^2$ .

En la tabla 4.6 se da la distribución de las desviaciones estándar. Nótese que al tomar la raíz cuadrada se elimina gran parte de la asimetría presentada por las varianzas. El examen de las  $s$  espaciadas igualmente y sus correspondientes  $s^2$  muestra que las varianzas por encima de la media aumentan más rápidamente que las que están debajo y que las diferencias entre  $s^2$  sucesivas son números impares consecutivos.

$s$	10	11	12	13	14
$s^2$	100	121	144	169	196

(Esta es una observación conveniente porque indica la importancia de escoger una escala de medida en toda investigación, ya que la distribución de las observaciones depende mucho de la escala. Si la distribución es normal o puede aproximarse a ella mediante una transformación, o sea, mediante la selección de una escala de medida, entonces se pueden aplicar las técnicas estadísticas basadas en la distribución normal; en otro caso son sólo aproximaciones).

El promedio de las 500 desviaciones estándar, denotado por  $\bar{s}$ , es 11.47 lb, en comparación con  $\sigma = 12$  lb. La raíz cuadrada del promedio de las 500 varianzas, esto es,  $\sqrt{\bar{s}^2}$ , es aquí,  $\sqrt{140.4} = 11.85$  lb. No sorprende que  $\bar{s}$  sea inferior a  $\sqrt{\bar{s}^2}$  ya que  $s$  subestima a  $\sigma$ ,

Tabla 4.5 Distribución de frecuencias de 500 varianzas  $s^2$  para muestras aleatorias de tamaño 10 obtenidas de la tabla 4.1.

Marca de clase	20 40 60 80 100 120 140	140.4 ↑	160 180 200 220 240 260 280 300 320 340 360 380
Frecuencia	11 27 40 46 59 62	55 ↓ 51 43 27 21 16 18 7 5 7 3 1 1	$s^2 = 140.4 \text{ lb}^2$ , usando 9 grados de libertad como divisor.
	• •	• •	$= 126.4 \text{ lb}^2$ , usando el tamaño de la muestra, 10 como divisor.)
			$\sigma^2 = 144 \text{ lb}^2$

**Tabla 4.6 Distribución de frecuencia de 500 desviaciones estándar  $s$ , correspondientes a las varianzas de la tabla 4.5.**

Marca de clase	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Frecuencia	1	10	14	23	37	42	55	69	66	63	40	26	30	11	11	2
									↓							
									$\bar{s} = 11.47 \text{ lb}$							
										$\sqrt{\bar{s}^2} = 11.85 \text{ lb}$						
											$\sigma = 12 \text{ lb}$					

Para un estimativo insesgado de  $\sigma$ , calcúlese  $\{1 + 1/[4(n - 1)]\}s$ , esto es,  $(1 + 1/36)s = 1.0285$ , para  $\sigma = 10$ . Esta es una aproximación, bastante buena, aun para  $n$  pequeño.

**Ejercicio 4.5.1** Construir histogramas con los datos de las tablas 4.5 y 4.6 y observar la simetría en cada caso.

#### 4.6 Insesgamiento de $s^2$

Se ha dicho que  $s^2 = \sum (Y_i - \bar{Y})^2/(n - 1)$  es un estimativo no sesgado de  $\sigma^2$ . El promedio de las 500  $s^2$ ,  $\bar{s}^2$ , es  $140.4 \text{ lb}^2$ .

Si la varianza muestral se define como  $\sum (Y_i - \bar{Y})^2/n$  entonces tenemos una estimación sesgada de  $\sigma^2$ , y la media de la población de tales valores es  $(n - 1)\sigma^2/n$ . Podríamos reconstruir cada suma de cuadrados mediante el uso de  $(n - 1)s^2 = \sum (Y_i - \bar{Y})^2$ . Sin embargo, como los grados de libertad son los mismos para todas las muestras, el promedio de las 500 varianzas calculadas con  $n = 10$  como divisor, es  $9\bar{s}^2/10 = 126.4 \text{ lb}^2$ , un valor mucho menor que 140.4. La diferencia entre los valores obtenidos usando  $n$  y  $n^{-1}$  disminuyen evidentemente a medida que  $n$  aumenta.

#### 4.7 Desviación estándar de la media o error estándar

La desviación estándar de la media es uno de los estadígrafos más útiles. Se calcula así:  $s_{\bar{Y}} = s/\sqrt{n}$  o  $s_{\bar{Y}} = \sqrt{s^2/n}$  y es un estimador sesgado de  $\sigma_{\bar{Y}}$  es decir la desviación estándar de la media de muestras aleatorias de tamaño  $n$  a partir de la población principal con desviación estándar  $\sigma$ . Así para una muestra de tamaño 10 tomada de la tabla 4.1,  $s$  es un estimativo de  $\sigma_{\bar{Y}} = \sigma/\sqrt{10} = 12/\sqrt{10} = 3.79 \text{ lb}$ . Para un estimativo de  $\sigma_{\bar{Y}}$  de 500 muestras, extraiga la raíz cuadrada del promedio de las varianzas dividida por  $n = 10$ . Llamándole  $(s_{\bar{Y}})$  tenemos

$$(s_{\bar{Y}}) = \sqrt{\bar{s}^2/n} = \sqrt{\frac{140.4}{10}} = 3.75 \text{ lb}$$

Este es un procedimiento para estimar  $\sigma_Y$  a partir de un conjunto de  $s^2$  mejor que el de dividir el promedio de las 500 desviaciones estándar, una media de estimativos sesgados, por la raíz cuadrada de 10. Si el último se llama  $(s_Y)''$  entonces tenemos

$$(s_Y)'' = \frac{\bar{s}}{\sqrt{n}} = \frac{11.47}{\sqrt{10}} = 3.63 \text{ lb}$$

Para más justificación en la obtención de  $s_Y$  a partir de  $s$ , usamos las 500 medias para estimar  $\sigma_Y$  por comparación. Encontramos

$$s_Y = \sqrt{\frac{\sum \bar{Y}^2 - (\sum \bar{Y})^2/500}{499}} = 3.71 \text{ lb}$$

El estrecho acuerdo entre ésta y  $(s_Y)'' = 3.75$  lb nos permite decir con más confianza que la relación  $\sigma_Y = \sigma/\sqrt{n}$  es válida ciertamente, y que, por lo tanto, cada muestra aleatoria proporciona un estimativo,  $s_Y$ , del error estándar del promedio  $\sigma_Y$ .

Es importante darse cuenta que la varianza, población o muestra, de una media decrece inversamente a  $n$  mientras la desviación estándar de una media decrece inversamente con  $\sqrt{n}$ . Esto se muestra claramente mediante un ejemplo así como con una fórmula.

$n$	$\sigma_Y^2$	$\sigma_Y$
4	$\frac{\sigma^2}{4} = \frac{144}{4} = 36$	$\frac{\sigma}{\sqrt{4}} = \frac{12}{\sqrt{4}} = 6$
8	$\frac{\sigma^2}{8} = \frac{144}{8} = 18$	$\frac{\sigma}{\sqrt{8}} = \frac{12}{\sqrt{8}} = 4.24$
16	$\frac{\sigma^2}{16} = \frac{144}{16} = 9$	$\frac{\sigma}{\sqrt{16}} = \frac{12}{\sqrt{16}} = 3$

#### 4.8 La distribución $t$ de Student

La distribución de la  $t$  de Student y la  $t$  de Student se vieron en la sección 3.10. Ahora estamos listos para demostrar que la distribución de los valores de  $t$  para las 500 muestras se approxima a la distribución teórica de  $t$  con 9 grados de libertad.

Para cada una de las 500 muestras, se calculó  $t = (\bar{Y} - \mu)/s_Y = (\bar{Y} - 40)/s_Y$ . Se observa que  $t$  es la desviación de la media muestral obtenida de la media de la población expresada en unidades de desviación estándar de medias, unidad de medida comúnmente usada para la toma de decisiones respecto a la utilidad o no de una desviación. Como una población de medias muestrales se distribuye simétricamente en torno a  $\mu$ , aproximadamente la

Tabla 4.7 Valores de  $t$  muestrales y teóricos para 5 grados de libertad

$$t = \frac{\bar{Y} - \mu}{s_y}$$

Intervalo de $t$			Muestral			Teórico		
De	A	Frecuencia	Frecuencia	Acumulada		Frecuencia	Acumulada	
			Porcentual	Una cola*	Dos colas**	Porcentual	Una cola*	Dos colas**
—	-3.250	2	0.4	100.0		0.5	100.0	
-3.250	-2.821	2	0.4	99.6		0.5	99.5	
-2.821	-2.262	7	1.4	99.2		1.5	99.0	
-2.262	-1.833	12	2.4	97.8		2.5	97.5	
-1.833	-1.383	29	5.8	95.4		5.0	95.0	
-1.383	-1.100	21	4.2	89.6		5.0	90.0	
-1.100	-0.703	63	12.6	85.4		10.0	85.0	
-0.703	0.0	116	23.2	72.8		25.0	75.0	
0.0	0.703	133	26.6	49.6	100.0	25.0	50.0	100.0
0.703	1.100	38	7.6	23.0	50.2	10.0	25.0	50.0
1.100	1.383	30	6.0	15.4	30.0	5.0	15.0	30.0
1.383	1.833	23	4.6	9.4	19.8	5.0	10.0	20.0
1.833	2.262	15	3.0	4.8	9.4	2.5	5.0	10.0
2.262	2.821	6	1.2	1.8	4.0	1.5	2.5	5.0
2.821	3.250	1	0.2	0.6	1.4	0.5	1.0	2.0
3.250	—	2	0.4	0.4	0.8	0.5	0.5	1.0
		500		100.0				

\* Porcentaje de valores mayores que la entrada en la primera columna de la izquierda.

\*\* Porcentaje de valores mayores en valor absoluto que la entrada en la primera columna de la izquierda.

mitad de los 500 valores de  $t$  deben ser positivos y el resto negativos; así la media debe ser aproximadamente cero. Encontramos 248 positivos y 252 negativos y la media es  $-0.038$ .

La tabla 4.7 en una distribución de frecuencia de valores de  $t$  observados. Se seleccionaron intervalos de clase desiguales de tal modo que las frecuencias observadas se pudieran comparar con las frecuencias teóricas registradas en la tabla A.3. Así, los límites de clase son idénticos para los  $t$  tabulados a niveles de probabilidad 0.5, 0.2, 0.1, 0.05, 0.02, y 0.01. Para facilitar la comparación, se dan los porcentajes de frecuencia tanto para los valores de  $t$  de la muestra como para los valores de  $t$  teóricos.

En una población de  $t$  valores, 2.5 por ciento son mayores que  $+2.262$  y 2.5 por ciento menores que  $-2.262$ . Esto puede observarse en la frecuencia porcentual teórica. La última columna de la tabla 4.7 combina ambas colas de la distribución al no tener en cuenta el signo de  $t$ . Esta es la columna más frecuentemente usada para niveles de probabilidad. Así, se dice que  $2.262$  es el valor de  $t$  al 5 por ciento de significancia para 9 gl, pero se designa  $t_{0.025}$ . Cuando sólo se considera la cola positiva de la distribución de  $t$ , el 5 por ciento de los valores de  $t$  caen más allá de  $1.833$ . Este valor se designa como  $t_{0.025}$ . Así mismo, cuando se consideran ambas colas, 1 por ciento de los valores de  $t$  caen más allá de  $\pm 3.250$ , el valor de  $t$  al 1 por ciento de nivel de significancia para 9 gl. Para los valores muestrales, 20  $t$ 's superan numéricamente el nivel del 5 por ciento y 4  $t$ 's superan numéricamente el nivel de 1 por ciento, en comparación con lo esperado, que era 25 y 5 respectivamente. Esto muestra coincidencia razonable entre el valor muestral y el teórico. Al

comparar los valores de la muestra y los teóricos a otros niveles de probabilidad, también se halla coincidencia razonable.

**Ejercicio 4.8.1** ¿Cuándo se dice que un estadígrafo es un estimador no sesgado de un parámetro? Clasificar los estadígrafos  $\bar{Y}$ ,  $s^2$ ,  $s$ ,  $s_p^2$ ,  $s_y$  en sesgados y no sesgados.

**Ejercicio 4.8.2** Dada una sola muestra de 20 observaciones aleatorias de una distribución normal, ¿cómo se puede estimar la varianza de la población de las medias muestrales de 20 observaciones? ¿De 40 observaciones?

**Ejercicio 4.8.3** Es la relación  $\sigma_y^2 = \sigma^2/n$  válida cuando la población no es normal?

**Ejercicio 4.8.4** ¿Cómo difiere la  $t$  de Student del criterio  $Z$  normal correspondiente? ¿Existe más de una distribución de  $t$ ? Si se tuvieran dos muestras aleatorias del mismo tamaño de la misma distribución normal y se hubiera calculado  $(\bar{Y} - \mu)/\sqrt{s_y^2/n}$ , se distribuiría como  $t$ ? Si las muestras fuesen de diferente tamaño, ¿ese criterio se distribuiría como  $t$ ? Si las muestras provinieran de diferentes poblaciones, ¿el criterio se distribuiría como  $t$ ? (Sugerencia: Ver la sec. 3.10).

#### 4.9 El enunciado de confianza

Vamos a verificar los enunciados de confianza basados en muestras para ver si la confianza enunciada se justifica. Para cada muestra aleatoria y todo nivel de probabilidad se establece un intervalo de confianza en torno a la media muestral. El procedimiento consiste en resolver las dos ecuaciones  $\pm t = (\bar{Y} - \mu)/s_y$  despejar la media  $\mu$  para obtener  $\mu = \bar{Y} \pm ts_y$  entonces un valor tabulado de  $t$ , uno de la muestra  $\bar{Y}$  y uno de  $s_y$  se reemplazan para obtener dos valores de  $\mu$ , denotados  $l_1$  y  $l_2$  que son los llamados *límites* del intervalo de confianza. Así  $l_1 = \bar{Y} - ts_y$  y  $l_2 = \bar{Y} + ts_y$ .

Para cada una de las 500 muestras aleatorias se calcularon  $l_1 = \bar{Y} - 2.262s_y$  y  $l_2 = \bar{Y} + 2.262s_y$ , es decir, los extremos de un intervalo de confianza al 95 por ciento. También se utilizó el valor de  $t$  al 1 por ciento,  $t_{0.005}$  para 9 grados de libertad. Como  $\mu = 40$  lb, se puede determinar el número de enunciados correctos referentes a  $\mu$ . Para las 500 muestras, los números de intervalos que contenían a  $\mu$  fueron 480 al nivel 5 por ciento y 496 al nivel del 1 por ciento. Estos valores se comparan favorablemente con los valores teóricos 475 y 495, respectivamente. El porcentaje de intervalos no incluyendo  $\mu$  es el mismo que el de los valores  $t$  de la muestra, que sobrepasan al de los  $t$ 's tabulados, en niveles de significancia de 5% y 1%, o sea:  $t_{0.025}$  y  $t_{0.005}$ .

En la práctica el parámetro  $\mu$  no se conoce. Por tanto, un experimentador sólo conoce el porcentaje de inferencias correctas respecto a  $\mu$ , nunca sabe si  $\mu$  cae en un intervalo de confianza dado.

A veces se sostiene erróneamente que un intervalo de confianza al 95 por ciento, con relación a la media muestral, da la amplitud dentro de la cual se encontrará el 95 por ciento de las medias muestrales futuras. Esto es incorrecto ya que la distribución de medias muestrales se centra en la media poblacional y no en una media muestral particular. Con base en una muestra actual, se estudió en la sec. 3.12 un enunciado correcto respecto a una observación futura o media. Para construir un intervalo que incluya una proporción especificada de la población con cierto coeficiente de confianza, es necesario un procedimiento de intervalo de tolerancia; como ejemplo véase Dixon y Massey (4.2).

#### 4.10 Muestreo de diferencias

Un problema que a menudo se presenta al experimentador es el de determinar si hay diferencia real entre las respuestas a dos tratamientos o, si por el contrario, la diferencia observada es suficientemente pequeña para que se la pueda atribuir al azar. Un método empírico de abordar el problema es considerar, como se hace en este capítulo, los resultados de un proceso de muestreo con dos tratamientos ficticios, es decir, muestrear una sola población, pero considerar los datos resultantes como si fueran de dos poblaciones. En esta forma, aprendemos lo que es una diferencia muestral corriente y una diferencia no usual, cuando no existe diferencia de población.

Las diferencias de observaciones, con su signo, extraídas aleatoriamente de una población normal se distribuirán normalmente en torno a una media cero. Las 500 muestras aleatorias de la tabla 4.1 se parearon aleatoriamente y se obtuvieron las diferencias con sus signos. Como las muestras originales se seleccionaron aleatoriamente, es suficiente con parear muestras consecutivas, individuo por individuo, para que no se requiera un proceso aleatorio adicional. Para cada una de las 250 muestras resultantes de 10 diferencias  $D$ , se calcularon los siguientes estadígrafos: la diferencia media  $D$ , la varianza de las diferencias  $s_D^2$ , la desviación estándar de las diferencias  $s_D$ , la desviación estándar de la diferencia de la media  $s_D$ , el valor  $t$ , y los límites de confianza para el promedio de la población de diferencias, valor que se sabe es  $\mu_D = 0$  en este caso. Esto se ilustra en la tabla 4.8, similar a la tabla 4.3 para valores individuales. La tabla 4.9 es una distribución de frecuencias de las 250 diferencias de promedios  $\bar{D}$ . La distribución observada es aproximadamente simétrica con 118 de las diferencias de promedios mayores que cero y 132 menores que cero. Estos valores se obtuvieron de la tabla 4.9, junto con la información adicional de que la clase con marca de clase igual a cero tiene 14  $\bar{D}$  positivas y 19 negativas. La media de las 250 es  $-0.533$ , muy cercana a cero.

La notación  $Y_i$  para la observación  $i$ -ésima de una muestra es inadecuada para distinguir entre observaciones de varias muestras. Así que introducimos un segundo subíndice y denotamos la observación mediante  $Y_{ij}$ . Entonces  $Y_{ij}$  se refiere a la  $j$ -ésima observación de la  $i$ -ésima muestra. Por ejemplo,  $Y_{25}$  (léase  $Y$  sub 2,5; se usan comas entre los subíndices sólo cuando se necesita claridad) corresponde a la quinta observación de la segunda muestra; y  $Y_{13} - Y_{23}$  corresponde a la diferencia con signo resultante de restar la tercera observación de la muestra dos, de la tercera observación de la muestra uno.

Las tablas 4.10 y 4.11 son distribuciones de frecuencias de las 250 varianzas y desviaciones estándar muestrales de 10 diferencias. Las formas de estas distribuciones son similares a las de las tablas 4.5 y 4.6 para  $s^2$  y  $s$ . Compárense las tablas apropiadas y nótese que las amplitudes son considerablemente mayores para las diferencias que para los individuos. La razón de esto es clara cuando se considera la posible amplitud de las diferencias. La amplitud posible va desde  $(10 - 70) = -60$  lb hasta  $(70 - 10) = +60$  lb, el doble que para los individuos. El promedio de las 250 varianzas  $s_D^2$  es 272.7; de la tabla 4.5,  $2s^2 = 2(140.4) = 280.8$ ; ambos están razonablemente cercanos a  $2\sigma^2 = 2(144) = 288$ . Los datos ilustran un importante teorema

La varianza  $\sigma_D^2$  de las diferencias de observaciones pareadas es el doble de la de las observaciones de la población principal.

Tabla 4.8 Tres muestras de diferencias entre observaciones aleatorias según la tabla 4.1

Número de la unidad	Observaciones pareadas $Y_{ij} Y_{tj}$		Diferencias $D_j = Y_{tj} - Y_{ij}$	Número de pareadas $Y_{3j} Y_{4j}$	Observaciones pareadas $Y_{4j} Y_{6j}$	Número de pareadas $Y_{4j} Y_{6j}$	Diferencias $D_j = Y_{3j} - Y_{4j}$	Número de la unidad	Observaciones pareadas $Y_{ij} Y_{6j}$	Diferencias $D_j = Y_{3j} - Y_{6j}$				
	Observaciones	Diferencias		Número de la unidad										
97	78	66	48	18	66	72	44	46	-2	21	14	32	29	3
74	69	47	45	2	62	28	43	34	9	63	28	43	34	9
*58	81	42	49	-7	15	64	29	44	-15	98	42	68	38	30
48	83	40	50	-10	28	37	34	37	-3	86	05	52	18	34
44	43	39	38	1	00	05	10	18	-8	77	94	48	62	-14
73	15	47	29	18	73	07	47	22	25	79	93	49	60	-11
73	81	47	49	-2	56	57	42	42	0	51	29	40	34	6
93	91	60	57	3	04	25	17	33	-16	99	66	70	44	26
79	46	49	39	10	92	53	58	41	17	39	06	37	20	17
63	21	43	32	11	34	94	36	62	-26	17	62	30	43	-13
<b>Suma = <math>\sum D</math></b>		<b>44</b>		<b>-19</b>		<b>-19</b>		<b>87</b>		<b>87</b>				
<b>Media = <math>\bar{D}</math></b>		<b>4.4</b>		<b>-1.9</b>		<b>-1.9</b>		<b>8.7</b>		<b>8.7</b>				
<b>F.C. = <math>(\sum D)^2 / 10</math></b>		<b>1,036.00</b>		<b>2,229.00</b>		<b>3,633.00</b>		<b>3,633.00</b>		<b>3,633.00</b>				
<b>S.C. = <math>\sum D^2 - (\sum D)^2 / 10</math></b>		<b>193.60</b>		<b>36.10</b>		<b>756.90</b>		<b>756.90</b>		<b>756.90</b>				
<b><math>s_D^2 = S.C./9</math></b>		<b>842.40</b>		<b>2,192.90</b>		<b>2,876.10</b>		<b>2,876.10</b>		<b>2,876.10</b>				
<b><math>s_D = \sqrt{S.C./9}</math></b>		<b>93.6</b>		<b>243.66</b>		<b>319.57</b>		<b>319.57</b>		<b>319.57</b>				
<b><math>s_D = \sqrt{s_D^2/10}</math></b>		<b>9.8</b>		<b>15.6</b>		<b>17.9</b>		<b>17.9</b>		<b>17.9</b>				
<b><math>t = (D - \bar{D})/s_D</math></b>		<b>3.06</b>		<b>4.94</b>		<b>5.65</b>		<b>5.65</b>		<b>5.65</b>				
<b><math>t_{.015} s_D = 2.262 s_D</math></b>		<b>1.44</b>		<b>-0.38</b>		<b>1.54</b>		<b>1.54</b>		<b>1.54</b>				
<b>L.C. = <math>\begin{cases} 1, &amp; t &gt; t_{.015} s_D \\ 0, &amp; t \leq t_{.015} s_D \end{cases}</math></b>		<b>6.92</b>		<b>11.15</b>		<b>12.78</b>		<b>12.78</b>		<b>12.78</b>				
<b>L.C. = <math>\begin{cases} 1, &amp; t &gt; t_{.015} s_D \\ 0, &amp; t \leq t_{.015} s_D \end{cases}</math></b>		<b>-2.5</b>		<b>-13.1</b>		<b>-4.1</b>		<b>-4.1</b>		<b>-4.1</b>				
<b>L.C. = <math>\begin{cases} 1, &amp; t &gt; t_{.015} s_D \\ 0, &amp; t \leq t_{.015} s_D \end{cases}</math></b>		<b>11.3</b>		<b>9.3</b>		<b>21.5</b>		<b>21.5</b>		<b>21.5</b>				

**Tabla 4.9 Distribución de frecuencia de 250 medias de diferencias  $\bar{D}$ , para muestras de 10 diferencias.**

Marca de clase	-12	-10.5	-9	-7.5	-6	-4.5	-3	-1.5	0	1.5	3	4.5	6	7.5	9	10.5	12	13.5	15
frecuencias	4	7	7	8	12	16	30	29	33	21	28	17	13	10	8	4	2	0	1

**Tabla 4.10 Distribución de frecuencias de las varianzas  $s_D^2$  de 250 muestras aleatorias de 10 diferencias según la tabla 4.1**

Marca de clase	60	100	140	180	220	260	300	340	380	420	460	500	540	580	620	660	700	740
Frecuencias	8	14	24	37	40	34	19	16	12	15	13	7	1	4	2	3	0	1

$$\bar{s_D^2} = 272.7 \quad \bar{2s^2} = \bar{s^2} = 280.8 \quad 2\sigma^2 = 288$$
**Tabla 4.11 Distribución de frecuencias de las desviaciones estándar  $s_D$  de 250 muestras aleatorias de 10 diferencias según la tabla 4.1.**

Marca de clase	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Frecuencias	1	5	4	7	8	17	24	28	29	26	19	13	19	10	16	10	4	4	3	2	1

$$\bar{s_D} = 16.04 \text{ lb} \quad \sqrt{\bar{2s^2}} = \sqrt{280.8} = 16.76 \text{ lb} \quad \sqrt{\bar{s_D^2}} = \sqrt{272.7} = 16.51 \text{ lb} \quad \sqrt{2\sigma^2} = \sqrt{288} = 16.97 \text{ lb}$$

En consecuencia,

La varianza muestral  $s_D^2$  de diferencias de observaciones pareadas es una estimación no sesgada de  $2\sigma^2$ .

Nótese que 20 observaciones han proporcionado 10 diferencias y que como consecuencia de hacer las diferencias quedan solo 9 grados de libertad relacionados con la estimación de  $s_D^2$ . En la práctica, cuando la varianza de la diferencia entre dos medias se necesita a menudo, no se hace el pareamiento aleatorio y las diferencias no se usan para calcular  $\bar{D}$  y  $s_D^2$ . Mediante un reordenamiento de la parte aritmética, es claro que  $\bar{D} = \bar{Y}_1 - \bar{Y}_2$ . A partir de una  $s^2$ , se estima la varianza  $2\sigma^2$  por  $2s^2 = s_D^2$ .

Los promedios de las desviaciones estándar de las diferencias son  $\bar{s_D} = 16.04$  y  $\sqrt{\bar{s_D^2}} = \sqrt{272.7} = 16.51$  lb. De nuevo el promedio directo de las desviaciones estándar es menor que la raíz cuadrada del promedio de las varianzas, pero ambos son razonablemente cercanos a  $\sigma_D = \sqrt{2\sigma^2} = \sqrt{288} = 16.97$  lb. las desviaciones estándar presentan un leve sesgo; las varianzas son sesgadas.

Se ha dicho que  $\sigma_{\bar{Y}}^2 = \sigma^2/n$  y que  $\sigma_D^2 = 2\sigma^2$ . Ambos teoremas dicen que la varianza de una diferencia entre dos medias denotada por  $\sigma_D^2$  es igual a  $2\sigma^2/n$  cuando cada media

contiene  $n$  observaciones. Así,  $\sigma_D = \sqrt{288/10} = 5.37$  lb. Para las 250 muestras,  $s_D = \sqrt{s_D^2/n} = \sqrt{272.7/10} = 5.22$  lb. o  $s_D = \sqrt{2s^2/n} = \sqrt{280.8/10} = 5.30$  lb.

A partir de un solo valor de  $s^2$ , se pueden obtener estimaciones de los siguientes parámetros importantes:  $\sigma^2$ ,  $\sigma_Y^2$ ,  $\sigma_D^2$ ,  $\sigma_{\bar{D}}^2$ ,  $\sigma$ ,  $\sigma_Y$ ,  $\sigma_D$  y  $\sigma_{\bar{D}}$ . Las interrelaciones en los términos estadísticos se muestran diagramáticamente en la fig. 4.3.

Para cada una de las 250 muestras de 10 diferencias, se calculó  $t$  como  $(\bar{D} - 0)/s_D$  una desviación respecto de la media poblacional expresado en unidades de desviación estándar de  $\bar{D}$ . Nótese que  $t = \bar{D}/s_D = (\bar{Y}_1 - \bar{Y}_2)/s_{\bar{Y}} = -\bar{Y}_2$ , ya que  $\bar{D} = \bar{Y}_1 - \bar{Y}_2$ . La distribución de estos valores de  $t$  se da en la tabla 4.12 y es similar a la de la tabla 4.7 para  $t = (\bar{Y} - \mu)/s_{\bar{Y}}$ . De los valores de  $t$ , 118 son positivos y 132 son negativos; su media es  $-0.00013$ . Catorce valores de  $t$  exceden el 5 por ciento de nivel de significancia en comparación con un número esperado de 12.5; 4 exceden el nivel del 1 por ciento en comparación con 2.5.

Volviendo al problema de determinar si hay diferencia real entre las respuestas a dos tratamientos, vemos que el método de muestreo expuesto ha demostrado que se ha de esperar cuando no hay diferencia real y los resultados son atribuibles sólo al azar. Los valores  $t = \bar{Y}/s_{\bar{Y}}$  de la tabla 4.12 se ve que ajustan bien a la distribución de la  $t$  de Student. En la situación real, es, pues, necesario calcular solamente el estadígrafo  $t$  y encontrar la probabilidad de un valor mayor o igual cuando el muestreo es aleatorio y cuando se realiza en una población cuya media es cero. Si la probabilidad de encontrar un mayor valor es pequeña y el experimentador no está seguro de que el muestreo se hizo en una población con media cero, probablemente decidirá que existe una diferencia real entre las respuestas a los dos tratamientos. El cálculo de intervalos de confianza conduce al mismo tipo de inferencia, porque si el  $t$  muestral es mayor que el 5 por ciento del nivel de probabilidad o  $t_{0.025}$ , por ejemplo, entonces el intervalo de confianza al 95 por ciento no contendrá el cero. Si el intervalo de confianza, no contiene el cero, el investigador puede tener escasa confianza en una afirmación de que no existe diferencia entre las respuestas a los dos tratamientos.

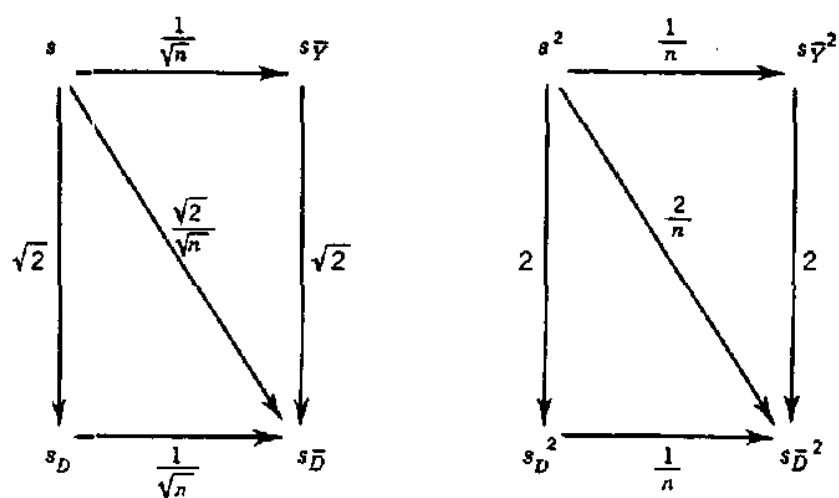


Figura 4.3 Ilustración de las relaciones entre las desviaciones estándar y las varianzas. (El diagrama no tiene que ver con el teorema de Pitágoras: en un triángulo rectángulo la hipotenusa al cuadrado es igual a la suma de los cuadrados de los otros dos lados)..

Tabla 4.12 Valores muestrales y teóricos de  $t = \bar{D}/s_D$ , 9 gl,  $\mu_D = 0$ 

Intervalo del $t$		Muestral		Teórico		
De	A	Frecuencia	Frecuencia porcentual	Frecuencia porcentual	Acumulada	
					Una cola	Dos colas
—	-3.250	1	0.4	0.5	100.0	
-3.250	-2.821	0	0.0	0.5	99.5	
-2.821	-2.262	7	2.8	1.5	99.0	
-2.262	-1.833	5	2.0	2.5	97.5	
-1.833	-1.383	14	5.6	5.0	95.0	
-1.383	-1.100	13	5.2	5.0	90.0	
-1.100	-0.703	21	8.4	10.0	85.0	
-0.703	0.0	71	28.4	25.0	75.0	
0.0	0.703	62	24.8	25.0	50.0	100.0
0.703	1.100	24	9.6	10.0	25.0	50.0
1.100	1.383	10	4.0	5.0	15.0	30.0
1.383	1.833	7	2.8	5.0	10.0	20.0
1.833	2.262	9	3.6	2.5	5.0	10.0
2.262	2.821	2	0.8	1.5	2.5	5.0
2.821	3.250	1	0.4	0.5	1.0	2.0
3.250	—	3	1.2	0.5	0.5	1.0
		250	100.0			

Ejercicio 4.10.1 Dadas dos muestras de observaciones pareadas:

I    10, 15, 13, 12, 11

II    12, 14, 14, 15, 13

¿Cuál es el valor de  $Y_{11}$ ?  $Y_{22}$ ?  $Y_{14}$ ?  $Y_{12} - Y_{22}$ ?  $\bar{Y}_1 - \bar{Y}_2$ ?

Ejercicio 4.10-2

Dada  $s^2 = 36$ . ¿cuál es el valor de  $s_D^2$ ?  $s_D^1$ ?

Dada  $s_Y^2 = 12$ . ¿cuál es el valor de  $s_D^2$ ?  $s^2$ ?

dada  $s_D^2 = 50$ . ¿cuál es el valor de  $s^2$ ?  $s_Y^2$ ?  $s_D^1$ ?

## 4.11 Resumen sobre muestreo

En la tabla 4.13 se presenta un resumen de los resultados obtenidos mediante el experimento de muestreo. Este resumen indica claramente que con el muestreo ha sido posible demostrar varias características y teoremas importantes relacionados con poblaciones de distribución normal. En particular:

1. Las medias de muestras aleatorias de  $n$  observaciones se distribuyen normalmente con media  $\mu$  y desviación estándar  $\sigma/\sqrt{n}$ . (Este teorema es aproximadamente cierto

Tabla 4.13 Un resumen de la información de:

1.500 muestras de 10 observaciones

Muestra	Símbolo	Valor	$s^2$		$s$		$s_p$	
			Divisor		$\sqrt{s^2}$	$s$	$s_p$	$\sqrt{s^2/10}$
			$n - 1 = 9$	$n = 10$				
		39.79	140.42	126.38	11.85	11.47	3.71	3.75
Población	Valor símbolo	40.00	$144$		$12$		$3.79$	
		$\mu$	$\sigma^2$		$\sigma$		$\sigma_p$	

2. 250 muestras de 10 diferencias

Muestra	Símbolo	Valor	$D$	$s_D$		$s_B$		$s_{\bar{D}}$	
				$s_D$	$\bar{s}^2$	$s_B$	$\sqrt{s_D^2}$	$\sqrt{\bar{s}^2}$	$s_{\bar{D}}$
			-0.53	272.71	280.84	16.04	16.51	16.76	5.16
Población	Valor símbolo	0	$288$		$16.97$		$5.37$		
		$\mu$	$\sigma_D$		$\sigma_B$		$\sigma_{\bar{D}}$		

3. Valores de  $t$ 

Número de muestras	Media	Número	Sin tener en cuenta el signo					
			Número mayor que $t_{0.25} = 2.262$		Número mayor que $t_{0.01} = 3.250$			
			Más	Menos	Observado	Esperado	Observado	Esperado
500	-0.038	248	252	20	25	4	5	
250	-0.00013	118	132	14	12.5	4	2.5	

con respecto a la normalidad de las medias cuando el muestreo se hace en poblaciones normales y siempre es cierto con respecto a la desviación estándar).

- Las medias de las diferencias de muestras aleatorias de  $n$  observaciones se distribuyen normalmente con media cero y desviación estándar  $\sqrt{2\sigma^2/n}$ .
- Una muestra aleatoria proporciona estimaciones no sesgadas de  $\mu$ ,  $\sigma^2$ ,  $\sigma_p^2$ ,  $\sigma_B^2$ ,  $\sigma_{\bar{D}}^2$ .
- El estadígrafo  $t = (\bar{Y} - \mu)/s_p$  o  $t = (\bar{D} - 0)/s_B = (\bar{Y}_1 - \bar{Y}_2)/s_{\bar{Y}_1 - \bar{Y}_2}$  se distribuye simétricamente en torno a la media cero y sigue la distribución tabulada de la  $t$  de Student.

### Ejercicio propuesto de laboratorio en relación con el capítulo cuarto

**Propósito** Obtener comparaciones relativas a la distribución de observaciones pareadas aleatoriamente, es decir, de su media y su varianza.

1. Hacer pares de 10 muestras aleatorias para obtener 5 muestras de 10 pares. (Como las 10 muestras son aleatorias, no se necesita de más proceso mecánico para asegurar la aleatoriedad en la formación de pares. Es suficiente con colocar muestras consecutivas adyacentes unas de otras).
2. Para cada uno de los cinco pares de muestras, calcular:
  - a) Las diferencias con sus signos.
  - b) La media de esas diferencias. (Que también es la diferencia de las medias).
  - c) La varianza y la desviación estándar.
  - d)  $t = \bar{D}/s_D$ .
  - e) Los intervalos de confianza el 95 y 99 por ciento para las diferencias medias.
3. Registrar datos de clases en hojas de resumen y resumir como se hizo en el texto.
4. Representar gráficamente los pares obtenidos ( $Y, s$ ) en el ejercicio de laboratorio del capítulo 3 para ver si existe alguna relación (ver sec. 3.10).

### Referencias

- 4.1. Baker, G. A., y R. E. Baker: "Strawberry uniformity yield trials," *Biom.*, 9:412-421 (1953).
- 4.2. Dixon, W. J., y F. J. Massey, Jr.: *Introduction to statistical analysis*, 3a. ed., McGraw-Hill, Nueva York, 1969, Sec. 9.8, Tabla 8b.

## COMPARACIONES ENTRE DOS MEDIAS MUESTRALES

### 5.1 Introducción

La mayoría de las personas usan la estadística de alguna manera; se hacen enunciados de confianza respecto a muchos aspectos de la vida diaria, mediante un adjetivo, un adverbio o una frase. Los experimentadores hacen enunciados de confianza basados en investigaciones, asignándole a cada uno de ellos una fracción decimal entre cero y uno, que es la medida de la confianza que se ha de tener en ellas. Los enunciados de confianza respecto de medias de población se expusieron en la sec. 3.11.

A veces se cambia el problema y se pregunta si una media poblacional puede o no tener un valor dado. Por ejemplo, podemos en forma regular calcular el recorrido de un auto por tanque de gasolina. Supóngase que el valor promedio de esos valores es 12.5 millas/galón y que durante un período de tiempo, tomamos ese valor como un parámetro. Ahora bien, debido al reclamo publicitario de los beneficios de un aditivo en una marca de la competencia, podemos decidir luego de probar algunos tanques de esa marca y determinar si las 12.5 millas/galón es todavía el parámetro. Eso es un problema de sometimiento a prueba de hipótesis.

En este capítulo trata pruebas de hipótesis respecto de una y dos medias de población. La *t* de Student es el criterio principal para esto, aunque se ha introducido *F* en anticipación a una generalización a más de dos medias y del análisis de la varianza. Se considera el problema del tamaño muestral.

### 5.2 Pruebas de significancia

En la sección 3.11 se construyó un intervalo de confianza para una media poblacional  $\mu$ ; el procedimiento es tal que un porcentaje dado de los intervalos contendrán el parámetro. Todo valor dentro de un intervalo de confianza es un candidato aceptable a  $\mu$ ; ninguno de esos valores se puede eliminar.

Es claro entonces que el procedimiento del intervalo de confianza puede usarse para probar una hipótesis referente a la localización precisa de un parámetro. Se construye el intervalo de confianza; si el valor hipotético cae dentro del intervalo, entonces la hipótesis debe ser aceptable, ya que para los valores de  $\mu$  dentro del intervalo de confianza, el azar y la hipótesis ofrecen una adecuada explicación de los datos.

En el capítulo 4, se tomaron 500 muestras de 10 observaciones y se construyeron intervalos de confianza al 95 y 99 por ciento. Se encontró que 480, aproximadamente el 95 por ciento, y 495, aproximadamente el 99 por ciento, de los intervalos contenían a  $\mu$ . En la construcción se utilizaron valores tabulados de  $t$ , para  $t_{0.025}$  y  $t_{0.005}$ , con 9 gl y se señaló que el número de valores muestrales  $t$ , sin considerar el signo, que excedieron a éstos, fueron  $500 - 480 = 20$  y  $500 - 495 = 4$ , respectivamente. En efecto, esas muestras que produjeron intervalos de confianza que no contenían  $\mu$  fueron las mismas que dieron los valores de  $t$  muestreados que excedieron de  $t_{0.025}$  y  $t_{0.005}$ , en valor absoluto. Eso sugiere la posibilidad de usar el  $t$  muestral directamente en pruebas de hipótesis.

El procedimiento de la prueba es pues obvio. Si  $t$  de la muestra, sin tener en cuenta el signo, es mayor que el  $t$  tabulado,  $t_{0.025}$ , por ejemplo, entonces los valores mayores que los observados deben ocurrir con probabilidad inferior a  $2(0.025) = 0.05$ . Esencialmente, el valor observado es mayor que lo que estamos dispuestos a aceptar que se deba al azar y a la hipótesis. Concluimos que la hipótesis es falsa aunque admitiendo que esa decisión puede ser errónea si tenemos una muestra aleatoria no usual, lo cual es probable que ocurra 5 veces en 100 ó 1 en 20, si hemos escogido 0.05 como nuestra tasa de error aceptable.

Examinemos de cerca el problema de la prueba. Nos proponemos calcular un  $t$  muestral como criterio de prueba, mediante

$$t = \frac{\bar{Y} - \mu}{s_y} \quad (5.1)$$

En éste se mide la distancia entre nuestra estimación de  $\mu$ , sea  $\bar{Y}$ , y el valor hipotético, sea  $\mu_0$ ; y como unidad de medida se toma la desviación estándar de  $\bar{Y}$ . Así, el  $t$  muestral es la desviación de una variable normal  $\bar{Y}$  respecto de su media hipotética medida en unidades de error estándar, o es el número de desviaciones estándar aplicable a  $\bar{Y}$  que separa a  $\bar{Y}$  y  $\mu_0$ . Obviamente, tenemos que especificar  $\mu_0$  para poder calcular un valor del criterio de la prueba, aunque podemos pensar que el enunciado, un tanto vago, "en la vecindad de  $\mu$ " sería una hipótesis adecuada. Cuando planteamos una hipótesis respecto a un valor de un parámetro de tal forma que podemos calcular un criterio de prueba, entonces tenemos una *hipótesis nula*, denotada por  $H_0$ . Ahora podemos simplemente escribir  $H_0: \mu = \mu_0$ .

Si la hipótesis nula es correcta, entonces el  $t$  resultante no deberá parecer fuera de lugar comparado con valores ordinarios de  $t$ , por ejemplo, entre,  $-t_{0.025}$  y  $t_{0.025}$ . Sin embargo, a veces rechazamos erróneamente una hipótesis nula porque simplemente ocurre que el  $t$  es muy grande numéricamente. Estamos razonando de la muestra a la población, haciendo una inferencia incierta y necesariamente tendremos la muestra no usual ocasional. Afortunadamente, el porcentaje alto de muestras raras, en una serie larga, con inferencias falsas, se puede controlar. Este porcentaje es nuestra *tasa de error* o *nivel de significación*.

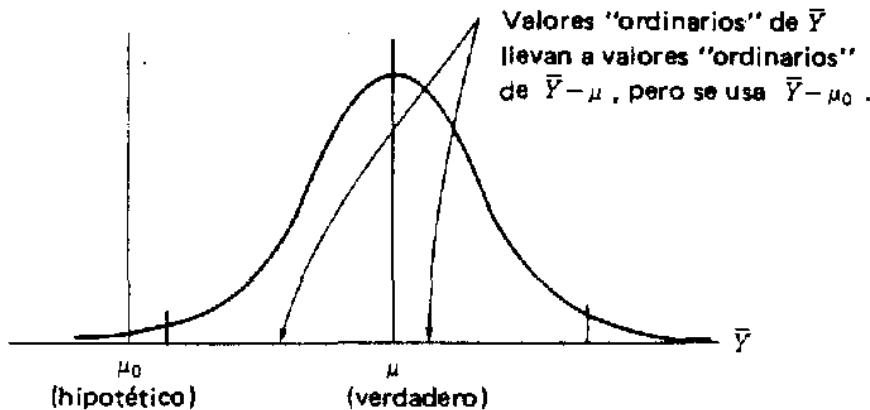


Figura 5.1 Consideración de  $\bar{Y} - \mu$  y  $\bar{Y} - \mu_0$  para una distribución dada de  $\bar{Y}$ .

cancia, generalmente designado por  $\alpha$ ; representa lo que cuesta tener que hacer inferencias inciertas. Por otra parte, el rechazar una hipótesis nula que es cierta, se dice que cometemos un *error de tipo I*, o un *error de primera clase*. Esto conlleva a que aceptemos  $H_0$ , si es cierta, con una probabilidad  $1 - \alpha$ .

Cuando se rechaza la hipótesis nula, necesitamos entonces una *hipótesis alternativa*,  $H_1$  o  $H_A$ . Esta, usualmente, no especificará un solo valor sino sencillamente establecerá como alternativa que  $\mu$  no es  $\mu_0$ , que  $\mu$  es mayor que  $\mu_0$ , o que  $\mu$  es menor que  $\mu_0$ . En forma abreviada, usamos bien sea  $H_1: \mu \neq \mu_0$ ,  $H_1: \mu > \mu_0$ , o  $H_1: \mu < \mu_0$ . Nótese que la elección de la hipótesis alternativa afecta a  $H_0$  hasta cierto punto. Para estos tres casos, podemos escribir ahora,  $H_0: \mu = \mu_0$ ,  $H_0: \mu \leq \mu_0$ ,  $H_0: \mu \geq \mu_0$ , respectivamente.

Si la hipótesis nula es falsa, entonces  $\mu$  no es igual a  $\mu_0$  sino a otro valor. Todavía calculamos  $t$  con la ec. (5.1) con  $\mu = \mu_0$ , pero nótese que  $\bar{Y} - \mu_0$  no es la desviación que exige la  $t$  de Student, sino que posiblemente es mayor que la desviación  $\bar{Y} - \mu$ . Esto se ve claramente en la fig. 5.1. Aquí vemos que  $\bar{Y}$  se distribuye respecto a  $\mu$  con varianza  $\sigma^2/n$ . La  $t$  de Student exige calcular  $\bar{Y} - \mu$  como numerador. Sin embargo, tomemos  $\mu_0$  como  $\mu$  y así calculamos  $\bar{Y} - \mu_0$ . Solamente si  $\bar{Y}$  se encuentra bien en la cola izquierda de su distribución, entonces  $\bar{Y} - \mu_0$  es pequeña y producirá pues un valor de  $t$  "ordinario"; para nuestra ilustración, esto ocurre con una probabilidad pequeña. O sea que si el verdadero  $\mu$  no es  $\mu_0$ , entonces lo más probable es que el valor muestral sea más grande y así con frecuencia descartaremos la hipótesis nula, tal como se desearía hacerlo. Sin embargo, cuando  $\mu$  no es  $\mu_0$ , esto es, cuando  $H_1$  es verdadera, aceptarímos algunas veces erróneamente  $H_0$  en vez de descartarla. Al hacer esto, cometemos un error de *tipo II* o *error de segunda clase*. La probabilidad de cometer este error se representa por  $\beta$ . Claro que esta probabilidad será pequeña cuando  $\mu_0$  y la verdadera  $\mu$  están bien separadas, y aumentará en la medida que se acerquen el uno al otro.

Finalmente, nos interesa poder detectar  $H_1$  cuando  $H_1$  es verdadera. Es claro que esto supone no cometer el error de tipo II, así que la probabilidad de que eso ocurra es  $1 - \beta$ . Esta habilidad para detectar  $H_1$  cuando  $H_1$  es verdadera se llama potencia de la prueba

En resumen, para probar una hipótesis,

1. Tener la población de interés claramente definida y formular una hipótesis con sentido, para la cual se pueda calcular un estadígrafo de prueba. Esta es la hipótesis nula  $H_0$ . Plantear una hipótesis alternativa  $H_1$  en caso de que lo que se compruebe no corrobore la  $H_0$ .
2. Escoger una probabilidad  $\alpha$  con base en la gravedad de rechazar una  $H_0$  cuando es verdadera; la probabilidad de aceptar  $H_0$  cuando es verdadera será  $1 - \alpha$ . Simultáneamente, tener en cuenta que es posible aceptar  $H_0$  cuando  $H_1$  es verdadera. Cuando esta tiene probabilidad  $\beta$ , entonces la probabilidad de aceptar  $H_1$  cuando es verdadera será  $1 - \beta$ . La manera de balancear estos dos errores se trata en la sec. 5.12.

Ahora, se lleva a cabo el experimento y se obtienen los datos. Las probabilidades que se aplican en las etapas 1 y 2, la fase de planeación, no se aplicarán cuando se estén analizando los datos. El investigador sacará entonces una conclusión que será cierta o falsa según cuál sea la hipótesis que representa la situación verdadera, las probabilidades reales de aceptar el parámetro verdadero serán cero o uno.

3. Calcular el valor muestral del estadígrafo de prueba y hallar la probabilidad de obtener, por azar, un valor más extremo que el observado. (Hasta ahora, hemos pensado en valores grandes como extremos, pero los pequeños serán no usuales con  $H_0$  para algún criterio de prueba). Como alternativa para calcular esta probabilidad, hallar el valor tabulado del criterio de la prueba de modo que la probabilidad de ocurrencia, por azar, dé un valor más extremo, sea el  $\alpha$  escogido en la etapa 2.
4. Si la probabilidad en la etapa 3 es tal que el azar parezca inadecuado para explicar el resultado, entonces se concluye que la hipótesis nula es incorrecta y se la descarta, si la probabilidad es menor que el  $\alpha$  escogido en la etapa 2; si no, se acepta  $H_0$ . Si se encuentra un valor tabulado del criterio de prueba, se reduce  $H_0$  si el valor muestral de este criterio es más extremo que el valor tabulado; si no, se la acepta.

El rechazo de una hipótesis nula es una afirmación bastante fuerte. Estamos eliminando definitivamente el valor específico del parámetro escogido para  $H_0$  en favor de algún valor no especificado incluido en el conjunto de alternativas. Por otra parte, la aceptación de  $H_0$  no quiere decir que el valor específico  $\mu_0$  sea el único aceptable para el parámetro; no hemos logrado rechazar  $H_0$  pero existen otros valores, presumiblemente en la vecindad de  $\mu_0$ , que pudieran haber servido, igualmente bien, como hipótesis nulas.

Por el momento, estamos hablando más particularmente sobre  $t$  como un criterio de prueba. Así, si deseamos comparar rendimiento en *bushels* por acre de un cultivo nuevo y de uno corriente de maíz, entonces, en la etapa 1 podíamos proponer la hipótesis nula de que no hay diferencia entre las medias de las poblaciones para los dos cultivos:  $H_0: \mu = 0$ . Podemos proponer como hipótesis alternativa que la media del nuevo cultivo es mayor, o que simplemente es diferente:  $H_1: \mu > 0$   $H_1: \mu \neq 0$ .

En la etapa 2, hagamos  $\alpha = 0.05$ . Estamos preparados para rechazar la hipótesis nula, aun si es verdadera, alrededor de 1 vez en 20, en promedio. Por lo tanto, aceptaremos  $H_0$  unas 19 veces de un total de 20 cuando es verdadera. Nos damos cuenta que al aceptar  $H_0$ , nos equivocaremos si existe una diferencia real de rendimiento entre los cultivos; pudimos no haber evaluado este riesgo.

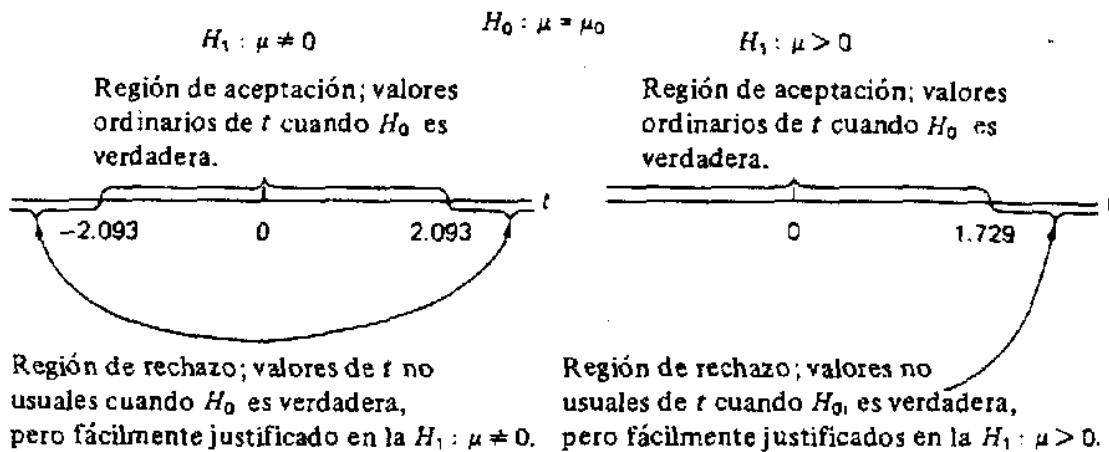


Figura 5.2 Regiones de aceptación y rechazo para la ilustración del texto.

Ahora se realiza el experimento. Se miden varias diferencias, por ejemplo, (rendimiento del nuevo cultivo) – (rendimiento del corriente). Estas diferencias son aleatorias, como resultado de cierto procedimiento y provienen de una población que se supone normal con media  $\mu$  desconocida y desviación estándar  $\sigma$  desconocida. A partir de los datos, calculamos un valor de  $t$  con la ec. (5.1), remplazando  $\mu$  por el valor hipotético  $\mu = 0$ . Esto es parte de la etapa 3.

Como todavía no estamos usando un ejemplo numérico, no tenemos un  $t$  muestral para el cual podamos calcular la probabilidad de encontrar un valor más extremo. Supondremos que se dispone de 20 diferencias aleatorias y encontramos el valor crítico cuando  $\alpha = 0.05$ .

Lo que constituye un valor extremo de  $t$  dependerá de  $H_1$ . Así, si  $H_1: \mu \neq 0$ , entonces cualquier valor de  $t$  que sea numéricamente grande es un valor extremo y es suficiente razón para rechazar  $H_0$ . En la tabla A.3 y para  $20 - 1 = 19$  gl, encontramos que  $t = 2.093$  es tal que  $P(|t| > 2.093) = 0.05$ ; éste es el  $t$  tabulado que en la etapa 3 resulta ser el valor crítico. Aquí tenemos una *prueba de dos colas*, con referencia a las alternativas.

Por otro lado, si  $H_1: \mu > 0$ , entonces, estamos buscando valores grandes positivos de  $t$  para respaldar  $H_1$  y eliminar  $H_0$ . En la tabla A.3 para  $20 - 1 = 19$  gl, encontramos que  $t = 1.729$  es tal que  $P(t > 1.729) = 0.05$ ; éste es el valor tabulado de  $t$  apto para la etapa 3 si  $H_1: \mu > 0$ . Aquí tenemos una *prueba de una cola*.

Finalmente, aceptamos o rechazamos  $H_0$  de acuerdo con el procedimiento de la etapa 4.

La figura 5.2 muestra diagramáticamente lo que sucede en la anterior ilustración. El valor tabulado del criterio de prueba, que corresponda al nivel de significancia elegido, separa los posibles valores de  $t$  en dos clases: la *región de aceptación* y la *región de rechazo o región crítica*. A los valores de  $t$  de  $\pm 2.093$  y  $1.729$  para esta ilustración se les llama *valores críticos*.

En muchos campos de experimentación, se acostumbra usar los niveles de significancia del 5 y del 1 por ciento. Si por azar un valor más discrepante del criterio de prueba que el obtenido ocurre probablemente menos del 5 por ciento de las veces pero no menos del 1 por ciento cuando la hipótesis nula es cierta, entonces se dice que la diferencia es significante, y el valor del criterio de la prueba se marca con un asterisco. Si un valor más discrepante del criterio de prueba que el obtenido ocurre con probabilidad menor que el

Tabla 5.1 Decisiones y sus resultados

El criterio de prueba muestral está en	La decisión es	Los datos son de una población para la cual	
		$H_0$ es cierta $H_1$ es falsa	$H_0$ es falsa $H_1$ es cierta
La región de aceptación o sea que no es significante	Aceptar $H_0$	Decisión correcta La probabilidad debe ser alta: Símbolo: $1 - \alpha =$ coeficiente de confianza.	Decisión incorrecta Se comete error de tipo II. La probabilidad deberá ser baja Símbolo: $\beta =$
	Rechazar $H_1$	Decisión incorrecta Se comete error de tipo I La probabilidad debe ser baja Símbolo: $\alpha =$ Nivel de significancia. Símbolo: $1 - \beta =$	Decisión correcta La probabilidad debe ser alta:
La región de rechazo, o sea que es significante	Rechazar $H_0$	Decisión incorrecta Se comete error de tipo I La probabilidad debe ser baja Símbolo: $\alpha =$ Nivel de significancia. Símbolo: $1 - \beta =$	Decisión correcta La probabilidad debe ser alta:
	Aceptar $H_1$	Decisión incorrecta Se comete error de tipo II. La probabilidad deberá ser baja Símbolo: $\beta =$	Decisión correcta La probabilidad deberá ser alta Símbolo: $1 - \alpha =$

I por ciento de las veces cuando la hipótesis nula es verdadera, se dice que la diferencia es *altamente significante* y el valor muestral del criterio de prueba se señala con dos asteriscos. La aceptación de la hipótesis nula puede indicarse con las letras *ns*.

Los niveles del 5 y del 1 por ciento son arbitrarios, pero parecen haberse escogido adecuadamente en el campo de la agricultura donde fueron usados por primera vez. En el caso de experimentos de tamaño pequeño, es posible que la hipótesis nula no sea rechazada probablemente a esos niveles, a menos que exista una diferencia real más grande. Esto sugiere la elección de otro nivel de significancia, quizás el 10 por ciento, en experimentos pequeños. Si un experimentador usa niveles diferentes del 5 y del 1 por ciento, esto debe quedar claramente establecido.

Los resultados de posibles decisiones en relación con posibles hipótesis se resumen en la tabla 5.1.

### 5.3 Prueba de hipótesis de que una media poblacional es un valor dado

Sean  $\mu$  y  $\sigma^2$  el promedio y la varianza de una población. Se extrae una muestra aleatoria de tamaño  $n$  y se calculan la media y la varianza de la muestra,  $\bar{Y}$  y  $s^2$ . Para probar la hipótesis  $H_0: \mu = \mu_0$  suponiendo que la población tiene distribución normal, el criterio de prueba se da mediante la ec. (5.1) como

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} \quad (5.2)$$

(Un criterio en que interviene un estimativo de  $\sigma^2$  fue primeramente estudiado por Student en 1908. Elaboró la distribución de un estadígrafo relacionado, al cual llamó  $Z$ . Más tarde R.A. Fisher hizo la distribución de  $t$ , la base de la tabla A.3).

En la primera ilustración, supóngase que tenemos un valor de  $\mu = \mu_0$  como hipótesis nula y que la alternativa es simplemente  $\mu \neq \mu_0$ .

Considérese la muestra 1 de la tabla 4.3. Sabemos que  $\mu = 40$ , así que probemos  $H_0: \mu = 40$  contra  $H_1: \mu \neq 40$ . Tenemos  $\bar{Y} = 36.4$  y  $s^2 = 264.04$ . Luego

$$t = \frac{36.4 - 40}{\sqrt{264.04/10}} = -0.701, 9 \text{ gl}$$

En la tabla A.3, con 9 gl encontramos que  $P(|t| > .701) > 0.5$ , pero muy cerca a 0.5. De otro modo, obsérvese que  $-0.701$  cae entre  $-2.262$  y  $2.262$ . No hay razón para pensar que  $-0.701$  sea un valor no usual de  $t$  si  $\mu = 40$ . Aceptamos la hipótesis nula y rechazamos la alternativa.

En este caso, sabemos que hemos aceptado la hipótesis nula, correctamente y con certeza.

Ahora ilustremos la prueba de una cola usando los datos de los corderos en la primera columna de la tabla 5.2. Esos datos son coeficientes de digestibilidad de la materia seca, alimentados con ensilaje de maíz, medidos en porcentajes para 7 corderos.

**Tabla 5.2 Coeficientes de digestibilidad de materia seca, alimentados con ensilaje de maíz, en porcentaje.**

$Y_1$ (Corderos)	$Y_2$ (Novillos)
57.8	64.2
56.2	58.7
61.9	63.1
54.4	62.5
53.6	59.8
56.4	59.2
53.2	
$\sum Y$	393.5
$\sum Y^2$	22,174.41
$\bar{Y}$	56.21 por ciento
	367.5
	22,535.87
	61.25 por ciento

$$\sum (Y_{1j} - \bar{Y}_1)^2 = \sum Y_1^2 - (\sum Y_1)^2/n_1 = 22,174.41 - 22,120.32 = 54.09 = (n_1 - 1)s_1^2$$

$$\sum (Y_{2j} - \bar{Y}_2)^2 = \sum Y_2^2 - (\sum Y_2)^2/n_2 = 22,535.87 - 22,509.37 = 26.50 = (n_2 - 1)s_2^2$$

$$s^2 = \frac{\sum Y_1^2 - (\sum Y_1)^2/n_1 + \sum Y_2^2 - (\sum Y_2)^2/n_2}{(n_1 - 1) + (n_2 - 1)} = \frac{54.09 + 26.50}{6 + 5} = 7.33,$$

un estimativo de la  $\sigma^2$  común

$$gl = (n_1 - 1) + (n_2 - 1) = 11$$

$$s_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{s^2 \frac{(n_1 + n_2)}{n_1 n_2}} = \sqrt{7.33 \frac{(7 + 6)}{42}} = \sqrt{2.27} = 1.51 \text{ por ciento, la desviación}$$

estándar apropiada para la diferencia entre las medias muestrales.

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s_{\bar{Y}_1 - \bar{Y}_2}} = \frac{56.21 - 61.25}{1.51} = \frac{-5.04}{1.51} = -3.33^{**}, gl = 11$$

Para el intervalo de confianza el 95 por ciento para  $\mu_2 - \mu_1$ ,  $\bar{Y}_2 - \bar{Y}_1 \pm t_{0.025} s_{\bar{Y}_1 - \bar{Y}_2} = 5.04 \pm 2.201(1.51) = 5.04 \pm 3.32$ ;  $l_1 = 1.72$  por ciento y  $l_2 = 8.36$  por ciento

Supongamos que un segundo investigador esté trabajando en el mismo problema general, pero en otra raza de corderos. Este investigador se dió cuenta que la investigación estaba en progreso y decidió que cuando los datos estuvieran disponibles, él probaría la hipótesis nula  $H_0: \mu = 54$  por ciento con la alternativa  $H_1: \mu > 54$  por ciento, pues su gran experiencia le hizo pensar como valor del parámetro el 54 por ciento. También, se habían acumulado pruebas de que existía una buena posibilidad de que la raza en investigación tuviera una media mayor a este coeficiente.

Cálculos

$$t = \frac{56.21 - 54.00}{\sqrt{(54.09/6)/7}} = 1.947, 6 \text{ gl}$$

Obsérvese que la diferencia está en la dirección esperada con  $H_1$ ; de otra manera aceptaríamos la hipótesis nula. En la tabla A.3 con 6 gl, encontramos  $P(t > 1.947) < 0.5$ , aunque se encuentra casi exactamente en el valor crítico de  $t = 1.943$ . Rechazamos  $H_0$  y aceptamos  $H_1$ .

Si un intervalo de confianza es de interés, tal vez se quiera conocer la localización del mínimo valor aceptable de  $\mu$ , ya que la alternativa es  $H_1: \mu > 54$  por ciento.

Comenzamos con el enunciado probabilístico

$$P\left(\frac{\bar{Y} - \mu}{s_{\bar{Y}}} < t_{.05}\right) = .95 \quad (5.3)$$

La solución algebraica de este enunciado con respecto a la variable aleatoria  $(\bar{Y} - \mu)/s_{\bar{Y}}$  lleva a

$$P(\mu > \bar{Y} - t_{.05} s_{\bar{Y}}) = .95 \quad (5.4)$$

Este es un enunciado probabilístico con relación a la localización del límite inferior,  $\bar{Y} - t_{.05} s_{\bar{Y}}$ , para el estimativo del parámetro  $\mu$ . Encontramos que este límite inferior es

$$56.21 - 1.943 \sqrt{\frac{(54.09/6)}{7}} = 54.005$$

Decimos que a menos que tengamos una muestra no usual, el parámetro  $\mu$  no deberá ser menor de 54.005.

El modelo lineal aditivo para una sola muestra se vió en la sec. 2.12.

- ~ Ejercicio 5.3.1 Los pesticidas que se aplican a los cultivos pueden afectar a los seres humanos. Un síntoma del efecto de un pesticida es la reducción en el cerebro de la actividad de la acetilcolinesterasa (ACE) y una reducción grave puede ser peligrosa para las funciones del cuerpo. Cuando se asperja el algodón con un pesticida, un criterio de la existencia de tal reducción es ver si se presenta o no reducción en la actividad del ACE en codornices en los límites de los cultivos.

En una recolección se hicieron las siguientes 6 observaciones, promedios de dos determinaciones, en la actividad cerebral del ACE en codornices: 86.03, 83.67, 95.21, 92.94, 83.12 y 80.22\*. En el supuesto de que las aves constituyan una muestra aleatoria, construir el límite superior, usando  $\alpha = 0.01$  para la localización del parámetro  $\mu$ .

En otra recolección se hicieron las siguientes observaciones: 23.76, 34.59, 56.22 y 68.22. De nuevo, encuentre el límite superior con  $\alpha = 0.05$  para la localización del parámetro  $\mu$ .

**Ejercicio 5.3.2** A siete observadores se les mostrió, durante un lapso corto, una red con 161 moscas y se les pidió que estimaran el número. Los resultados los da Cochran (5.2). Basadas en 5 estimaciones los valores fueron: 183.2, 149.0, 154.0, 167.2, 187.2, 158.0 y 143.0. Defina una población razonable de la cual pudieron haberse obtenido estas medias.

Probar la hipótesis nula de que la media de la población es 161 moscas, usando  $\alpha = 0.05$ . Construir un intervalo de confianza para  $\mu$  del 95 por ciento.

**Ejercicio 5.3.3** Los rendimientos de 10 plantas de fresas en un ensayo de uniformidad los presenta Baker y Baker (5.1) así: 239, 176, 235, 217, 234, 216, 318, 190, 181 y 225 g. Al 95 y 99 por ciento, calcule los intervalos de confianza para la media poblacional. Probar la hipótesis  $\mu = 205$  (escogida arbitrariamente) con la alternativa  $\mu \neq 205$  al 5 por ciento de nivel de significancia.

**Ejercicio 5.3.4** Supóngase que un fabricante de llantas mide en miles de millas, el período de vida de 10 llantas. Encuentra que  $\bar{Y} = 26.68$  y  $s^2 = 12$ .

Probar la hipótesis nula de que  $\mu = 25.0$  con  $H_0: \mu > 25.0$ . Construir un intervalo de confianza del 95 por ciento para  $\mu$  de un solo lado de tal forma que dé un límite inferior del parámetro.

**Ejercicio 5.3.5** Las larvas de algunas mariposas monarcas concentran glucósidos cardíacos a partir de plantas de algodoncillo, que las hacen repugnantes para los pájaros, los cuales las evitan después de un primer encuentro.

Supóngase que las mariposas han sido recolectadas en una localidad y que se han medido las concentraciones de glucósidos en relación a sus pesos deseados. Supóngase que los datos resultantes son  $\bar{Y} = 0.200$  y  $s^2 = 0.012$  para  $n = 75$ .

Construir un intervalo de confianza del 95 por ciento para la verdadera media de la población. Probar la hipótesis nula de que  $\mu = 0.150$  con  $H_1: \mu \neq 0.150$ ;  $\mu = 0.150$  puede considerarse como el parámetro para otra localidad.

#### 5.4 Pruebas de dos o más medias

Supóngase que tenemos dos poblaciones con medias  $\mu_1$  y  $\mu_2$ . Se extrae una muestra aleatoria de cada población para probar la hipótesis de que  $\mu_1$  y  $\mu_2$  están separadas por una cantidad específica, que usualmente se toma como cero.

Para la hipótesis nula de no diferencia,  $t$  se define como

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_{\bar{Y}_1 - \bar{Y}_2}}} \quad (5.5)$$

\* Datos usados con el permiso de P.C. Smithson y O.T. Sanders. También ver la referencia 5.11.

Aquí,  $s_{\bar{Y}_1 - \bar{Y}_2}$  es la desviación estándar apropiada para una diferencia entre dos medias aleatorias de una población normal.

Una vez más, obsérvese que  $t$  mide la distancia de una variable aleatoria o una media hipotética en unidades de desviación estándar de la variable aleatoria. Cuando las distribuciones subyacentes son normales con una varianza común, este estadígrafo se distribuye como la  $t$  de Student, tabla A.3, si el valor hipotético es el verdadero valor, pero no si la hipótesis es falsa. Si el  $t$  muestral no puede atribuirse razonablemente al azar y a la hipótesis nula, concluimos que  $\bar{Y}_1 - \bar{Y}_2$  es muy grande porque  $\mu_1 \neq \mu_2$ .

El cálculo de  $s_{\bar{Y}_1 - \bar{Y}_2}$  depende de si

1. Las dos poblaciones tienen una varianza común  $\sigma^2$
2. Los valores de los  $\sigma^2$ , o el  $\sigma^2$  común, se conocen o se estiman
3. Las dos muestras son del mismo tamaño y
4. Las observaciones son pareadas

La elección de una región de rechazo depende de

1. El nivel de significancia escogida
2. El tamaño de la muestra
3. La prueba necesaria, esto es, si es de una cola o de dos colas.

La prueba está diseñada para dos medias y obviamente no es adecuada para generalizar a más de dos medias, ya que no tenemos forma clara de definir un numerador para la generalización. Se necesita un nuevo enfoque que ahora vamos a considerar.

La varianza en una población de medias muestrales es  $\sigma^2/n$ , donde  $\sigma^2$  es la varianza de los individuos en una población principal y todas las muestras son de tamaño  $n$ . Esto implica que pueden usarse las medias para estimar  $\sigma^2$ . Así, usando las medias de dos muestras, calculamos un estimativo de  $\sigma^2/n$  mediante  $\sum (\bar{Y}_i - \bar{Y})^2/(2 - 1)$  lo cual al multiplicar por  $n$  estima a  $\sigma^2$ . Un segundo estimativo de  $\sigma^2$  puede encontrarse directamente a partir de los individuos de cada muestra.

Ahora, cuando  $\mu_1 \neq \mu_2$ , pero las poblaciones tienen la misma varianza, el estimativo de  $\sigma^2$  basado en medias muestrales tenderá a sobreestimar  $\sigma^2$  debido a que la diferencia entre las  $\bar{Y}$  incluirá una contribución atribuible a la diferencia entre las medias poblacionales lo mismo que cualquier diferencia aleatoria. Así, en general, si las  $\mu_i$  difieren, se espera que las  $\bar{Y}$  sean más variables que cuando sólo actúa el azar. Por otra parte, un estimativo basado en los individuos se calcularía dentro de las muestras individuales usando desviaciones de cada estimativo respecto de una  $\mu_i$ . En esta forma, las variaciones entre las  $\mu_i$  podrían no contribuir a la estimación de  $\sigma^2$ .

Por tanto, en una prueba de significancia de una diferencia, esto es, una prueba  $H_0: \mu_1 - \mu_2 = 0$  con  $H_1: \mu_1 - \mu_2 \neq 0$  puede entrar en juego la razón de dos de tales estimativos de  $\sigma^2$ . En particular, la ecuación (5.6) define un criterio de prueba común

$$F = \frac{\text{estimativo de } \sigma^2 \text{ a partir de las medias}}{\text{estimativo de } \sigma^2 \text{ a partir de los individuos}} \quad . \quad (5.6)$$

Los valores de  $F$  se dan en la tabla A.6 para 5 valores de probabilidades y varios pares de grados de libertad de numerador y denominador. Para dos medias, el estimativo de  $\sigma^2$  para el numerador se basa en un sólo grado de libertad; en este caso, la raíz cuadrada de  $F$  tiene distribución de  $t$  de Student.

### 5.5 Comparación de dos medias muestrales, muestras independientes y varianzas iguales

Sean  $\mu_1$  y  $\mu_2$  dos medias de dos poblaciones. Se toma una muestra aleatoria de cada población. Sean las medias de las muestras, sus varianzas y tamaños:  $\bar{Y}_1$ ,  $\bar{Y}_2$ ,  $s_1^2$ ,  $s_2^2$ ,  $n_1$ , y  $n_2$ , respectivamente. Con estas muestras aleatorias vamos a probar la hipótesis nula  $H_0: \mu_1 = \mu_2$  suponiendo que las poblaciones se distribuyen normalmente y tienen una varianza común pero desconocida.

El criterio de prueba es

$$t = \frac{(\bar{Y}_1 - \mu_1) - (\bar{Y}_2 - \mu_2)}{s_{\bar{Y}_1 - \bar{Y}_2}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{s_{\bar{Y}_1 - \bar{Y}_2}} \quad (5.7)$$

el cual para  $H_0: \mu_1 = \mu_2$  se convierte en la ec. (5.5). En general la diferencia  $\mu_1 - \mu_2$  puede hacerse igual al valor hipotético que el experimentador suponga.

En el criterio de la prueba,  $s_{\bar{Y}_1 - \bar{Y}_2}$  es un estimativo de  $\sigma_{\bar{Y}_1 - \bar{Y}_2}$ , sujeto a la variación de muestreo. Primero estimamos  $\sigma^2$  mediante la combinación de las sumas de cuadrados de las dos muestras y dividimos por los grados de libertad combinados:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \quad (5.8)$$

Este es un *promedio ponderado* de las varianzas muestrales y es superior al promedio aritmético, el cual da igual ponderación a las varianzas muestrales. El promedio ponderado y el promedio aritmético coinciden cuando las muestras tienen igual tamaño. El criterio  $t$ , cuando  $H_0$  es verdadera, se distribuye con la  $t$  de Student para muestras aleatorias de poblaciones normales, pero las discrepancias considerables respecto de la normalidad pueden no afectar seriamente a la distribución, especialmente en las cercanías de los valores críticos, 5 y 1 por ciento, usados generalmente.

**Caso 1, La prueba cuando  $n_1 \neq n_2$ .** Calcúlese  $s_{\bar{Y}_1 - \bar{Y}_2}$  mediante

$$s_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{s^2 \left( \frac{n_1 + n_2}{n_1 n_2} \right)} \quad (5.9)$$

Aquí  $s^2$  es el promedio ponderado de las varianzas muestrales, ec. (5.8).

Se da un ejemplo en la tabla 5.2 para datos tomados de Watson et al (5.13). El intervalo de confianza para  $\mu_2 - \mu_1$  en vez de  $\mu_1 - \mu_2$  se calculó solamente porque  $\bar{Y}_2 - \bar{Y}_1$  es positivo.

Tabla 5.3 Análisis de la varianza de los datos de la tabla 5.2

Fuente de variación	gl	Suma de cuadrados	Cuadrado medio	F
Corderos o novillos	1	81.93	81.93	11.18**
Entre corderos + entre novillos	6 + 5	54.09 + 26.50	7.33	
Total	12	162.52		

Cuando se usa el criterio  $F$ , los resultados se presentan generalmente en una tabla de *análisis de la varianza*, donde cada varianza o cuadrado medio es un estimativo de la misma  $\sigma^2$  si la hipótesis nula es verdadera (ver tabla 5.3). Para calcular el numerador de  $F$ , recordar que  $(\bar{Y}_1 - \bar{Y}_2)^2$  estima  $2\sigma_y^2 = 2\sigma^2/n$ . Para estimar  $\sigma^2$ , hay que multiplicar por  $n/2$ .

Cuando las medias provienen de muestras de diferente tamaño, el cuadrado de las diferencias es un estimativo de  $\sigma^2(n_1 + n_2)/n_1 n_2$ ; por tanto, para tener un estimativo de  $\sigma^2$ , multiplicamos por  $n_1 n_2/(n_1 + n_2)$ . Así  $(56.21 - 61.25)^2(6)(7)/(6 + 7) = 81.93$  es un estimativo de  $\sigma^2$ , basado en medias. (El cálculo efectivo se hizo con totales; el que se hace con las medias difiere un poco debido a los errores de redondeo). Nótese que  $n_1 n_2/(n_1 + n_2)$  es el inverso del multiplicador de  $s^2$  cuando  $t$  es el criterio y se calcula  $s_{\bar{Y}_1 - \bar{Y}_2}$ . Ahora  $t^2 = F$ , siempre que  $F$  tenga un grado de libertad asociado con el numerador, así  $(3.33)^2$  y 11.19 son iguales dentro de los errores de redondeo. Nótese también que los grados de libertad y las sumas de cuadrados en el cuerpo de la tabla dan la suma total. La tabla A.6 es la de los valores teóricos de  $F$  para varios niveles de probabilidad. La primera columna es para los casos en los cuales sólo entran dos medias muestrales.  $F(1,11) = 4.84$  al nivel del 5 por ciento y 9.65 al del 1 por ciento. Como el  $F = 11.18$  muestral es mayor que 9.65, se concluye que la diferencia es altamente significante. A la misma conclusión se llega con  $t$  como criterio de prueba.

**Caso 2. La prueba cuando  $n_1 = n_2 = n$ .** También se aplica el procedimiento dado en el caso 1. El criterio es  $t$  en la ec. (5.7); la ec. (5.9) se reduce a

$$s_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{2s^2}{n}} \quad (5.10)$$

Los grados de libertad son  $2(n - 1)$ . En la tabla 5.4 se opera con un ejemplo usando datos de Ross y Knott (5.6). Como la diferencia observada entre medias sólo es significante al nivel del 5 por ciento, el intervalo de confianza no contiene el cero, aunque lo contiene el intervalo de confianza al 99 por ciento.

*Cuando  $\sigma^2$  es conocida*, el criterio de la prueba se convierte en

$$z = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\sigma^2(n_1 + n_2)/n_1 n_2}} \quad (5.11)$$

Tabla 5.4 Ganancia de peso en novillas Holstein

$Y_1$ , control	$Y_2$ , vitamina A
175	142
132	311
218	337
151	262
200	302
219	195
234	253
149	199
187	236
123	216
248	211
206	176
179	249
206	214
$\sum Y_1$	2,627
$\sum Y_2^2$	511,807
$\bar{Y}_1$	187.6 lb
$\bar{Y}_2$	235.9 lb

$$n_1 = n_2 = n = 14$$

$$\sum (Y_{1j} - \bar{Y}_1)^2 = \sum Y_1^2 - (\sum Y_1)^2/n = 511,807 - 492,938 = 18,869 = (n_1 - 1)s_1^2$$

$$\sum (Y_{2j} - \bar{Y}_2)^2 = \sum Y_2^2 - (\sum Y_2)^2/n = 817,583 - 779,272 = 38,311 = (n_2 - 1)s_2^2$$

$$s^2 = \frac{\sum Y_1^2 - (\sum Y_1)^2/n + \sum Y_2^2 - (\sum Y_2)^2/n}{2(n - 1)} = \frac{57,180}{26} = 2,199,$$

Un estimativo de la  $\sigma^2$  común

$$g^1 = 2(n - 1) = 26$$

$$s_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{2s^2}{n}} = \sqrt{\frac{2(2,199)}{14}} = 17.7 \text{ lb.}, \text{ la desviación estándar apropiada}$$

para la diferencia entre medias muestrales

$$t = \frac{\bar{Y}_2 - \bar{Y}_1}{s_{\bar{Y}_1 - \bar{Y}_2}} = \frac{187.6 - 235.9}{17.7} = \frac{-48.3}{17.7} = -2.73^*; (t_{.005} = 2.78)$$

Para el intervalo de confianza al 95 por ciento,  $\bar{Y}_2 - \bar{Y}_1 \pm t_{.025}s_{\bar{Y}_1 - \bar{Y}_2} = 48.3 \pm 2.056(17.7) = 48.3 \pm 36.4$ ;  $l_1 = 11.9$  y  $l_2 = 84.7$  lb.

Para el intervalo de confianza al 99 por ciento,  $\bar{Y}_2 - \bar{Y}_1 \pm t_{.001}s_{\bar{Y}_1 - \bar{Y}_2} = 48.3 \pm 2.779(17.7) = 48.3 \pm 49.2$ ;  $l_1 = -0.9$  y  $l_2 = 97.5$  lb.

el cual se compara con valores tabulados en la última línea de la tabla de la  $t$  de Student. Estos valores se toman de las tablas de la distribución normal.

Ejercicio 5.5.1 En un programa de salud, muchos participantes miden su progreso mediante el tiempo que les toma correr determinada distancia. Un predictor de ese tipo lo constituye la tasa de recuperación cardíaca (TRC) ideada en la Universidad de Harvard.

Los datos siguientes son tiempos, en minutos y segundos, para una carrera de 1.5 millas para hombres que han estado corriendo por algunos años y comenzaban una nueva estación.

Los hombres se han repartido en categorías. Los de TRC de 40-49 ó 50-59, TRC1s o TRC2s respectivamente\*

TRC1 Tiempos: 12:24, 12:45, 11:04, 11:22, 11:58, 8:34, 11:16, 11:52, 8:28, 12:01, 11:03, 12:01, 11:31

TRC2 Tiempos: 14:33, 10:35, 12:51, 11:28, 11:48, 14:05, 10:51, 18:50, 18:11

Usar  $\alpha = 0.05$  y probar la hipótesis de que no hay diferencia entre los tiempos medios de las dos poblaciones. Calcular un intervalo de confianza al 95 por ciento para las diferencias entre las medias de las dos poblaciones. ¿Cómo se hubiera podido utilizar el intervalo de confianza para llegar a las mismas conclusiones que con la prueba de significancia?

**Ejercicio 5.5.2** Datos similares a los del ejercicio 5.5.1 pero para nuevos participantes son\*

TRC1 Tiempos: 10:22, 9:33, 9:16, 11:28, 10:59, 13:55, 10:10, 11:46

TRC2 Tiempos: 10:36, 10:40, 11:31, 12:55, 12:58, 10:54, 11:34, 11:15, 13:43

Usar  $\alpha = 0.05$  y probar la hipótesis de diferencia nula entre los tiempos medios de carrera para las dos poblaciones con la alternativa TRC2 mayor que TRC1. Estimar la diferencia media  $\mu_2 - \mu_1$ , usando un intervalo de confianza de un sólo lado que proporcione un límite inferior para la diferencia. ¿Pueden los dos procedimientos dar la misma información con respecto a si la diferencia es o no cero?

**Ejercicio 5.5.3** Sería de esperar que una población con mayor experiencia produjera menores tiempos de carrera. Probar la hipótesis usando los dos conjuntos de tiempos de carrera para TRC1 individuales de los ejercicios 5.5.1 y 5.5.2.

**Ejercicio 5.5.4** Repítase el ejercicio 5.5.3 para los dos conjuntos de datos TRC2.

**Ejercicio 5.5.5** de un cultivo de una variedad de guayule, se seleccionaron 54 plantas al azar. De éstas, 15 fueron atípicas y 12 aberrantes. Los porcentajes de caucho para estas plantas fueron:\*\*

Atípicas: 6.21, 5.70, 6.04, 4.47, 5.22, 4.45, 4.84, 5.88, 5.82, 6.09, 5.59, 6.06, 5.59, 6.74, 5.55  
Aberrantes: 4.28, 7.71, 6.48, 7.71, 7.37, 7.20, 7.06, 6.40, 8.93, 5.91, 5.51, 6.36

Probar la hipótesis de diferencia nula entre las medias de las poblaciones de los porcentajes de caucho. Calcular un intervalo de confianza del 95 por ciento para la diferencia entre medias de las dos poblaciones. Presentar los resultados en una tabla de análisis de varianza. Compárese  $F$  con  $t^2$ .

**Ejercicio 5.5.6** Los pesos en gramos de 10 machos y 10 hembras jóvenes de faisanes de cuello anillado atrapados en enero en el jardín botánico de la Universidad de Wisconsin fueron:\*

Machos: 1293, 1380, 1614, 1497, 1340, 1643, 1466, 1627, 1383, 1711

Hembras: 1061, 1065, 1092, 1017, 1021, 1138, 1143, 1094, 1270, 1028

\* Datos obtenidos por cortesía de A.C. Linnerud, North Carolina State University.

\*\* Datos cortesía de W.T. Federer, Cornell University, Ithaca, Nueva York.

\* Datos cortesía de Departament of Wildlife Management, University of Wisconsin, Madison, Wisconsin.

Probar la hipótesis de que hay una diferencia de 350 g (valor escogido solamente por razones ilustrativas) entre las medias poblacionales en favor de los machos con la alternativa de que la diferencia es mayor de 350. Esta es una prueba de una cola.

**Ejercicio 5.5.7** Si se tuvieran más muestras ( $k > 2$ ), ¿cómo se podría combinar la información para estimar una  $\sigma^2$  común? *Sugerencia:* Generalícese la ec. (5.8).

## 5.6 Modelo lineal aditivo

El modelo lineal aditivo (ver sec. 2.12) trata de explicar una observación como una media más un elemento aleatorio de variación, donde la media puede ser la suma de varios componentes asociadas con diversos efectos de fuentes de variación. Para muestras de dos poblaciones con medias posiblemente diferentes pero con varianza común, la composición de una observación está dada por

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad (5.12)$$

donde  $i = 1, 2$  y  $j = 1, \dots, n_1$  para  $i = 1$  y  $j = 1, \dots, n_2$  para  $i = 2$ . ( $\tau$  corresponde a la letra griega tau). El modelo explica la observación  $i$ -ésima de la población  $i$ -ésima como compuesta de una media general  $\mu$ , más una componente  $\tau_i$  para la población de que se trata, más de un elemento aleatorio de variación. En términos de la sec. 5.5  $\mu + \tau_1 = \mu_1$  y  $\mu + \tau_2 = \mu_2$ . Por comodidad, cuando  $n_1 = n_2$  hacemos  $\mu = (\mu_1 + \mu_2)/2$  de modo que  $\tau_1 + \tau_2 = 0$  o  $\tau_2 = -\tau_1$ , o sea que las  $\tau$  se miden como desviaciones. Así, si  $\tau_1$  representa un incremento, entonces  $\tau_2 = -\tau_1$  representa un decremento igual. Si bien esto puede no ser cierto, no afecta la diferencia entre las medias, es decir,  $2\tau$ ; y la diferencia es el aspecto importante del problema, mientras que  $\mu$  es simplemente un punto de referencia conveniente. Cuando  $n_1 \neq n_2$  podemos hacer  $n_1 \tau_1 + n_2 \tau_2 = 0$ . Se supone que los  $\varepsilon$  provienen de una sola población de  $\varepsilon$  con media  $\mu = 0$  y varianza  $\sigma^2$ . Puede estimarse, entonces,  $\sigma^2$  a partir de una o de ambas muestras. La diferencia entre este modelo y el de la sec. 2.12 es que éste es más general; nos permite describir simultáneamente dos poblaciones de individuos.

Examinemos las observaciones, los totales y las medias muestrales. Obsérvese que la notación para las sumas,  $Y_{i\cdot}$ , y  $\varepsilon_{i\cdot}$ , y para las medias,  $\bar{Y}_{i\cdot}$  y  $\bar{\varepsilon}_{i\cdot}$ . La notación con puntos es otra manera de expresar  $\sum$ . La sumatoria se extiende a todos los valores del subíndice para el lugar ocupado por el punto.

Muestra 1	Muestra 2
$Y_{11} = \mu + \tau_1 + \varepsilon_{11}$	$Y_{21} = \mu + \tau_2 + \varepsilon_{21}$
$Y_{12} = \mu + \tau_1 + \varepsilon_{12}$	$Y_{22} = \mu + \tau_2 + \varepsilon_{22}$
$\vdots$	$\vdots$
$Y_{1n_1} = \mu + \tau_1 + \varepsilon_{1n_1}$	$Y_{2n_2} = \mu + \tau_2 + \varepsilon_{2n_2}$
$\sum_j Y_{1j} = Y_{1\cdot} = n_1 \mu + n_1 \tau_1 + \varepsilon_{1\cdot}$	$\sum_j Y_{2j} = Y_{2\cdot} = n_2 \mu + n_2 \tau_2 + \varepsilon_{2\cdot}$
$\bar{Y}_{1\cdot} = \mu + \tau_1 + \bar{\varepsilon}_{1\cdot}$	$\bar{Y}_{2\cdot} = \mu + \tau_2 + \bar{\varepsilon}_{2\cdot}$

Para obtener  $s^2$ , se calcula

$$\sum_j (Y_{1j} - \bar{Y}_{1\cdot})^2 = (n_1 - 1)s_1^2$$

Esta suma de cuadrados se asocia solamente con los  $\varepsilon$ , pues  $\mu + \tau_1$  es común a todas esas observaciones y no afecta por lo tanto a su variación. También calculamos

$$\sum_j (Y_{2j} - \bar{Y}_{2\cdot})^2$$

asociada solamente con los  $\varepsilon$ . Como las poblaciones tienen una varianza común  $\sigma^2$ , es decir, puesto que existe una sola población de  $\varepsilon$ , estas sumas de cuadrados se combinan para estimar a  $\sigma^2$ , como se ve en la ec. (5.8). A la estimación de muestreo  $\sigma^2$  no contribuyen los  $\tau$ .

Ahora considérese  $(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) = (\tau_1 - \tau_2) + (\bar{\varepsilon}_{1\cdot} - \bar{\varepsilon}_{2\cdot})$ . Este es el numerador de  $t$ ; se utiliza en el cálculo del numerador de  $F$ . En la hipótesis nula, las poblaciones tienen la misma media y  $\tau_1 = \tau_2 = 0$  también,  $\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}$  es una diferencia de dos medias de observaciones de la misma población y tiene una varianza que es múltiplo de  $\sigma^2$ . Así si los  $\varepsilon$  se distribuyen normalmente, ambos,  $t$  y  $F$ , son criterios apropiados para probar hipótesis nulas.

La cantidad  $(\bar{\varepsilon}_{1\cdot} - \bar{\varepsilon}_{2\cdot})$  deberá ser pequeña pues es la suma algebraica de dos cantidades, cada una con media cero y con varianza pequeña,  $\sigma^2/n_i$ . Cuando la hipótesis nula es falsa, la cantidad  $(\tau_1 - \tau_2) = 2\tau_1 \neq 0$  dado que  $\tau_2 = -\tau_1 \neq 0$ ; una alternativa es verdadera y el numerador de  $t$  o  $F$  se amplía, dando lugar a un valor mayor del criterio de lo que se esperaría aleatoriamente si la hipótesis nula fuera verdadera. Valores grandes del criterio son pues no usuales si la hipótesis nula es verdadera, así que la probabilidad de detectar una diferencia, se ha visto que aumenta a medida que la diferencia real aumenta.

## 5.7 Comparación de medias muestrales; observaciones pareadas de importancia

Las observaciones se parean a menudo. Por ejemplo, dos raciones alimenticias pueden compararse utilizando dos animales de cada una de 10 camadas de cerdos, asignando al azar animales de cada camada, uno a cada ración; o puede compararse el porcentaje de aceite de dos variedades de soya producida en parcelas pareadas en 12 localidades. El pareamiento se ha hecho antes de comenzar el experimento con base en respuestas similares cuando no hay efectos de tratamiento. Si los miembros de los pares tienden a correlacionarse positivamente, es decir, si los miembros de cada par tienden a ser grandes o pequeños conjuntamente, entonces puede aumentar la capacidad del experimento para detectar una pequeña diferencia. La información sobre el pareamiento se usa para eliminar una fuente de varianza extraña que es la que existe de un par a otro. Esto se hace calculando la varianza de las diferencias en vez de la de los individuos dentro de cada muestra. El número de grados de libertad en que se basa la estimación de  $\sigma_D^2$  es el número de pares menos 1. Si se va a incrementar la capacidad de un experimento para detectar una diferencia real, entonces la varianza de diferencias debe ser lo suficientemente menor que

la varianza de individuos, esto es,  $\sigma_D^2$  debe ser menor que  $2\sigma^2$  para compensar la pérdida de grados de libertad debido al pareamiento; para el pareamiento aleatorio,  $\sigma_D^2 = 2\sigma^2$ .

El criterio de prueba es  $t$ , como en la ec. (5.2) remplazando a  $\bar{Y}$  por  $\bar{D}$  y calculando  $s$  de la siguiente forma:

$$s = \sqrt{\frac{\sum_j (Y_{1j} - Y_{2j})^2 - \left[ \sum_j (Y_{1j} - Y_{2j}) \right]^2 / n}{(n - 1)}} = \sqrt{\frac{\sum_j D_j^2 - \left( \sum_j D_j \right)^2 / n}{(n - 1)}} \quad (5.13)$$

En el divisor,  $n - 1$  son los grados de libertad y  $n$  es el número de diferencias muestrales de pares.

Se considera que las pruebas de hipótesis con observaciones pareadas importantes se reducen a probar que la media de las diferencias es un número dado, a menudo igual a cero. En la tabla 5.5 se opera con un ejemplo basado en datos de Shuel (5.10).

La hipótesis nula que se prueba es la de que la media de la población de diferencias es cero; las alternativas son que la media no es cero. El criterio de prueba se distribuye como la  $t$  cuando el supuesto de que las diferencias se distribuyen normalmente es correcto y la hipótesis nula es verdadera. El  $t$  tabulado  $t_{0.005}$  para 9 grados de libertad y una prueba de dos colas con  $\alpha = 0.01$  es 3.3. Aquí es difícil explicar la diferencia observada con base en el muestreo aleatorio de la población asociada con la hipótesis nula. Se rechaza la hipótesis nula con base en la evidencia presentada.

Cuando se conoce  $\sigma^2$ , se compara el  $t$  observado con el  $t$  tabulado que aparece en la última línea de la tabla  $t$ . El criterio  $F$  para observaciones pareadas se estudia en el capítulo 8.

**Ejercicio 5.7.1** Las constantes de enfriamiento de ratones recien sacrificados y las de los mismos ratones recalentados hasta la temperatura del cuerpo fueron determinadas por Hart (5.4) así:

Recien sacrificados: 573, 482, 377, 390, 535, 414, 438, 410, 418, 368, 445, 383, 391, 410, 433, 405, 340, 328, 400

Recalentados: 481, 343, 383, 380, 454, 425, 393, 435, 422, 346, 443, 342, 378, 402, 400, 360, 373, 373, 412

Probar la hipótesis de que no hay diferencia entre las medias poblacionales.

**Ejercicio 5.7.2** En septiembre se obtuvieron los tiempos de los corredores del ejercicio 5.5.1. Los tiempos fueron registrados de nuevo en mayo siguiente. Para el grupo TRC1 en el mismo orden en mayo los tiempos fueron: 11:16, 12:30, 11:30, 11:06, 11:28, 8:18, 11:44, 12:02, 8:28, 11:55, 11:27, 11:31 y 11:46.\*

Con  $\alpha = 0.05$ , probar la hipótesis nula de que las medias de las poblaciones no han cambiado. Construir un intervalo de confianza del 95 por ciento para las diferencias entre las medias de la población.

**Tabla 5.5 Concentración de azúcar de néctar en medias cabezas de trébol rojo a diferentes presiones de vapor durante 8 horas**

Presión de vapor		
4.4 mmHg $Y_1$	9.9 mmHg $Y_2$	Diferencia $D = Y_1 - Y_2$
62.5	51.7	10.8
65.2	54.2	11.0
67.6	53.3	14.3
69.9	57.0	12.9
69.4	56.4	13.0
70.1	61.5	8.6
67.8	57.2	10.6
67.0	56.2	10.8
68.5	58.4	10.1
62.4	55.8	6.6
$\sum Y$	670.4	561.7
		108.7
		$1,226.07 = \sum (Y_{1j} - Y_{2j})^2 = \sum D^2$
$\bar{Y}$	67.0	56.2
		10.8
	$s_D^2 = \frac{\sum D^2 - [\sum D]^2/n}{n-1} = \frac{1,226.07 - 1,181.57}{9} = 4.944$	
	$s_D = \sqrt{s_D^2} = \sqrt{\frac{4.944}{10}} = 0.4944$	$s_D = .703$
	$t = \frac{\bar{D}}{s_D} = \frac{10.8}{.703} = 15.4^{**}$	para 9 gl

El intervalo de confianza del 99 por ciento de la diferencia media poblacional se calcula así:

$$\bar{D} \pm t_{0.005} s_D = 10.8 \pm 3.3(.703). \text{ Así que } l_1 = 8.5 \text{ y } l_2 = 13.1 \text{ por ciento.}$$

Obsérvese que  $\sum (Y_{1j} - Y_{2j}) = \sum Y_{1j} - \sum Y_{2j}$  y que  $Y_1 - Y_2 = \bar{Y}_1 - \bar{Y}_2$ .

**Ejercicio 5.7.3** Los tiempos para los corredores del grupo TRC2 del ejercicio 5.5.2 también se registraron en septiembre. En mayo siguiente, los tiempos correspondientes fueron 11:34, 10:39, 11:47, 12:12, 12:48, 10:30, 11:20, 11:11 y 11:00.†

Probar la hipótesis nula de que la media de la población en mayo es menor que la de la población correspondiente en septiembre. Calcular el intervalo de confianza que se considere más apropiado, de un solo lado al 95 por ciento para la diferencia entre las dos medias de la población.

**Ejercicio 5.7.4** Peterson (5.5) estudió el efecto de la exposición de flores de alfalfa a diferentes condiciones ambientales. Escogió 10 plantas vigorosas con las flores expuestas libremente en la

† Datos cortesía de A.C. Linnerud, North Carolina State University.

parte alta, y flores escondidas, en lo posible en la base. Finalmente determinó el número de semillas producidas por cada dos vainas en cada localidad. Los datos fueron:

Planta	1	2	3	4	5	6	7	8	9	10
Flores en la parte superior	4.0	5.2	5.7	4.2	4.8	3.9	4.1	3.0	4.6	6.8
Flores en la parte inferior	4.4	3.7	4.7	2.8	4.2	4.3	3.5	3.7	3.1	1.9

Probar la hipótesis de que no hay diferencia de las medias poblacionales con la alternativa de que las flores de la parte superior producen más semillas. Calcular el intervalo de confianza apropiado de un solo lado.

### 5.8 El modelo lineal aditivo para las comparaciones pareadas

La expresión lineal aditivo para la composición de cualquier observación está dada por

$$Y_{ij} = \mu + \tau_i + \rho_j + \varepsilon_{ij} \quad (5.14)$$

$Y_{ij}$  es la observación en la muestra  $i$ -ésima para el par  $j$ -ésimo,  $i = 1, 2, \dots, n$  y  $j = 1, 2, \dots, n$ . De nuevo tenemos una media general  $\mu$  una componente  $\tau_i$  peculiar a la muestra y un elemento aleatorio  $\varepsilon_{ij}$  además, existe una componente  $\rho_j$  peculiar al par de observaciones. Los  $\tau_i$  y  $\rho_j$  contribuyen a la variabilidad de las  $Y$ , a condición de que no todos sean iguales a cero. Por conveniencia, hacemos  $\sum \tau_i = 0$  y  $\sum \rho_j = 0$ . Este modelo admite una media poblacional diferente para cada observación, pero estas medias están estrechamente relacionadas por construcción;  $Y_{ij}$  es la única observación de la población con media  $\mu + \tau_i + \rho_j$ , pero  $\tau_i$  está presente en otras  $n - 1$  observaciones y  $\rho_j$  en otra observación. Los  $\varepsilon$  provienen, como máximo, de dos poblaciones correspondientes al subíndice  $i$ , pero no es necesario suponer que provienen de una sola población como en el caso del modelo de la ec. (5.12). Debido a las relaciones entre las medias poblacionales, o sea, los  $(\mu + \tau_i + \rho_j)$  es posible estimar una varianza apropiada para probar la diferencia de las medias muestrales. Para ver esto claramente, se establece una tabla como la siguiente:

Muestra 1 $Y_{1j}$	Muestra 2 $Y_{2j}$	Diferencia $D = Y_{1j} - Y_{2j}$
$Y_{11} = \mu + \tau_1 + \rho_1 + \varepsilon_{11}$	$Y_{21} = \mu + \tau_2 + \rho_1 + \varepsilon_{21}$	$(\tau_1 - \tau_2) + (\varepsilon_{11} - \varepsilon_{21})$
.....	.....	.....
$Y_{1n} = \mu + \tau_1 + \rho_n + \varepsilon_{1n}$	$Y_{2n} = \mu + \tau_2 + \rho_n + \varepsilon_{2n}$	$(\tau_1 - \tau_2) + (\varepsilon_{1n} - \varepsilon_{2n})$
Totales: $Y_{1\cdot} = n\mu + n\tau_1 + \sum \rho_j + \varepsilon_{1\cdot}$	$Y_{2\cdot} = n\mu + n\tau_2 + \sum \rho_j + \varepsilon_{2\cdot}$	$n(\tau_1 - \tau_2) + (\varepsilon_{1\cdot} - \varepsilon_{2\cdot})$
Medias: $\bar{Y}_{1\cdot} = \mu + \tau_1 + \bar{\varepsilon}_{1\cdot}$	$\bar{Y}_{2\cdot} = \mu + \tau_2 + \bar{\varepsilon}_{2\cdot}$	$(\tau_1 - \tau_2) + (\bar{\varepsilon}_{1\cdot} - \bar{\varepsilon}_{2\cdot})$

Es obvio que las diferencias en la última columna tienen una variación relacionada únicamente con las diferencias (sumas algebraicas) de los  $\varepsilon$ , dado que  $(\tau_1 - \tau_2)$  es una

constante en todas. La varianza de las diferencias muestrales es un estimativo de  $\sigma_1^2 + \sigma_2^2$  si los  $\varepsilon_1$  y  $\varepsilon_2$  tienen varianza diferente, o de  $2\sigma^2$  si existe una varianza común. En el numerador del criterio de prueba, bien sea  $t$  o  $F$ , entra la media de las diferencias y, por consiguiente, tiene una contribución proveniente de la diferencia entre las medias de los tratamientos,  $\tau_1 - \tau_2 = 2\tau_1$ , si existe, además de una contribución de los  $\varepsilon$ . Si existe, entonces el numerador es mucho mayor que el denominador, ello se atribuye corrientemente a una diferencia real entre los tratamientos y no a un suceso de azar no usual. Nótese que si los  $\rho_j$  son reales, contribuirán a toda varianza calculada directamente a partir de cualquiera de las dos muestras.

Esta prueba tiene una propiedad importante que no poseen las pruebas anteriores en las que interviene la hipótesis de una diferencia entre las medias poblacionales. La teoría nos dice que la suma algebraica de variables con distribución normal tiene una distribución normal. La aplicación de esto al caso presente es al efecto de que las diferencias se distribuyen normalmente si los errores también lo están, *independientemente de que los  $\varepsilon_1$  y los  $\varepsilon_2$  tengan o no una varianza común*. En la sec. 5.10 se da una explicación adicional.

Un valor del pareamiento no mencionado antes tiene que ver con el alcance de la inferencia. Se ha visto que la variación de un par a otro puede ser grande. Si deliberadamente hacemos grande esta variación, ampliamos el alcance de nuestra inferencia. Por lo tanto, nuestros pares de cerdos pueden proceder de muchas camadas engendrados por diferentes hembras y machos, y posiblemente, diferentes razas; nuestra soya pudo haberse cultivado en diferentes localidades, nuestra inferencia es amplia como resultado.

### 5.9 Muestras independientes y varianzas desiguales

Dada una muestra de cada una de dos poblaciones con  $\sigma_1^2 \neq \sigma_2^2$  es decir, con varianzas desiguales vamos a probar la hipótesis de que  $\mu_1 = \mu_2$  usando las estimaciones muestrales de las varianzas.

La  $s_{\bar{Y}_1 - \bar{Y}_2}$  apropiada se calcula así:

$$s_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (5.15)$$

Ni las sumas de cuadrados ni los grados de libertad se combinan como cuando  $s_1^2$   $s_2^2$  eran estimativos de una  $\sigma^2$  común y las observaciones no eran pareadas como en la ec. (5.8). Ahora calculamos  $t'$  con la ec. (5.16). La prima indica que el criterio no se distribuye estrictamente como la  $t$  de Student, sino como una aproximación de la misma.

$$t' = \frac{\bar{Y}_1 - \bar{Y}_2}{s_{\bar{Y}_1 - \bar{Y}_2}} \quad (5.16)$$

Para determinar el valor crítico de  $t'$ , se usa una suma ponderada de valores de  $t$  tabulados con  $gl = n_i - 1$  y ponderaciones  $w_i = s_i^2/n_i$ ,  $i = 1, 2$ , y el resultado de esa suma

se divide por la suma de las ponderaciones [Cochran y Cox (5.3)], o se usa un  $t$  tabulado con "gl efectivos" calculados con la ec. (5.17). [Ver Satterthwaite (5.8) y Searle (5.9)].

$$\text{gl efectivo} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1 - 1)] + [(s_2^2/n_2)^2/(n_2 - 1)]} \quad (5.17)$$

Es improbable que los grados de libertad efectivos sean un entero, así que es de esperar tener que redondear o interpolar. La prueba resultante tiene una tasa de error,  $\alpha$ , aproximadamente igual al del  $t$  o los  $t$  tabulados usados en la determinación del valor crítico. También se puede calcular un intervalo de confianza con un coeficiente de confianza aproximadamente igual a  $1 - \alpha$ .

Los datos de la tabla 5.6 tomados de Rowles (5.7) se usan para dar un ejemplo del proceso.

La ecuación (3.15) parece definir la  $t$  de Student en una forma tal, que se puede generalizar; la  $t$  mide una distancia entre una variable aleatoria y su media poblacional en desviaciones estándar aplicables. ¿Por qué ha sido necesario cambiar  $t$  por  $t'$  cuando  $\sigma_1^2 \neq \sigma_2^2$ ?

Una definición más completa de  $t$  la describe como la razón de un valor de  $Z$  a un valor  $\sqrt{\chi^2/\text{gl}}$ . Dado que  $Z$  y  $\chi^2$  exigen variables con medias cero y varianzas unitarias, parece necesario usar  $\sigma^2$  en toda aplicación práctica. Cuando  $\sigma^2$  es la misma para ambos,  $Z$  y  $\chi^2$ , como cuando  $\bar{Y}$  y  $s^2$  se calculan a partir de una sola muestra, no es necesario conocer este valor ya que se cancela en numerador y denominador. En el presente problema, el  $Z$  del numerador debe ser  $(\bar{Y}_1 - \bar{Y}_2)/\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$  mientras que  $\chi^2$  en el denominador debe ser  $(n_1 - 1)s_1^2/\sigma_1^2 + (n_2 - 1)s_2^2/\sigma_2^2$ . En  $t$ , la cancelación no es posible, excepto en el caso improbable en que la verdadera razón  $\sigma_1^2/\sigma_2^2$  se conozca. Esta razón puede llamarse un parámetro de incomodidad pues crea un problema considerable en proporcionar criterios como  $t'$ .

**Ejercicio 5.9.1** En un experimento con avena, la media fue 52.8 bushels por acre y la varianza del error experimental fue de 20.31 con 36 gl. Un experimento análogo en circunstancias algo diferentes podría dar  $\bar{Y} = 48.4$  y  $s^2 = 45.06$  con 24 gl.

Probar la hipótesis nula de que las dos medias poblacionales tienen el mismo valor, suponiendo que las varianzas verdaderas son diferentes.

**Ejercicio 5.9.2** Un experimento con conejillos de indias dió una ganancia media de peso de 105.3 g. La varianza del error experimental fue 427.80. En un experimento similar pero con diferente forraje, la media de ganancia de peso fue de 175.0 g. Supongamos que  $s^2 = 726.20$ .

Probar la hipótesis nula de que las dos medias poblacionales tienen el mismo valor. Supóngase que las varianzas poblacionales son diferentes y que los gl son 15 y 7 respectivamente.

## 5.10 La media y la varianza de una función lineal

A medida que hemos avanzado en los problemas de muestreo, se ha respondido con fórmulas apropiadas a la necesidad de conocer las varianzas de las medias, de las diferencias aleatorias y de diferencias entre dos medias muestrales con varios supuestos. Estas fórmulas tienen una base común que ahora examinaremos. Los resultados tienen otras aplicaciones en capítulos posteriores.

Tabla 5.6 Porcentaje de grava fina en suelos superficiales

	Buen suelo	Suelo pobre
	5.9	7.6
	3.8	0.4
	6.5	1.1
	18.3	3.2
	18.2	6.5
	16.1	4.1
	7.6	4.7
$\sum Y$	76.4	27.6
$\sum Y^2$	1,074.60	150.52
$\bar{Y}$	10.91	3.94

$$\sum (Y_i - \bar{Y}_i)^2 = \sum Y_i^2 - (\sum Y_i)^2/n_1 = 1,074.60 - 833.85 = 240.75$$

$$s_1^2 = \frac{\sum (Y_1 - \bar{Y}_1)^2}{n_1 - 1} = \frac{240.75}{6} = 40.12$$

$$\sum (Y_2 - \bar{Y}_2)^2 = \sum Y_2^2 - (\sum Y_2)^2/n_2 = 150.52 - 108.82 = 41.70$$

$$s_2^2 = \frac{\sum (Y_2 - \bar{Y}_2)^2}{n_2 - 1} = \frac{41.70}{6} = 6.95$$

$$s_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{40.12}{7} + \frac{6.95}{7}} = \sqrt{6.72} = 2.59\%$$

$$t' = \frac{\bar{Y}_1 - \bar{Y}_2}{s_{\bar{Y}_1 - \bar{Y}_2}} = \frac{10.91 - 3.94}{2.59} = \frac{6.97}{2.59} = 2.69^*$$

$$gl \text{ efectivos} = \frac{(40.12/7 + 6.95/7)^2}{\frac{(40.12/7)^2}{6} + \frac{(6.95/7)^2}{6}} \approx 8$$

Comparar  $t'$  con el  $t$  tabulado con 8 gl ( $= 2.37$  para  $\alpha = 0.05$ ). La evidencia es que hay una diferencia real en el porcentaje de grava fina.

El intervalo de confianza al 95 por ciento es  $\bar{Y}_1 - \bar{Y}_2 \pm t_{0.025} s_{\bar{Y}_1 - \bar{Y}_2} = 6.97 \pm 2.306(2.59)$   
 $I_1 = 1.00$  por ciento,  $I_2 = 12.94$  por ciento.

Supongamos que comenzamos a buscar esta base común mediante la construcción de una función lineal de observaciones aleatorias:

$$L = \sum c_i Y_i \quad (5.18)$$

Tales funciones son comunes. Cuando  $L = \bar{Y} = \sum Y_i/n$ ,  $c_i = 1/n$  para todo  $i$ . Cuando  $L = \bar{Y}_1 - \bar{Y}_2 = \sum Y_1/n_1 - \sum Y_2/n_2$ ,  $c_1 = 1/n_1$  para las  $Y$  de la muestra 1 y  $c_2 = -1/n_2$  para las  $Y$  de la muestra 2, o simplemente  $c_1 = 1$  y  $c_2 = -1$  si no definimos cada media en términos de las  $Y_{ij}$ .

Considérese  $L$  como una variable aleatoria. Puede generarse por muestreo repetido a partir de una o más poblaciones principales de  $Y$  y del cálculo de  $L$  para cada muestra.

Deseamos conocer la media o *esperanza* o *valor esperado* de esta población derivada; escribimos  $\mu_L = E(L)$ . Escribir  $E$  supone que se efectúa o se ha efectuado una operación de cálculos de medias, muy similar al cálculo de la suma de todos los posibles valores y dividiendo luego por  $N$ , excepto que ahora estamos hablando de promediar una población que no tiene necesariamente un número finito de observaciones. Si  $E(Y_i) = \mu_i$  por ejemplo, para cada  $Y_i$  en la ec. (5.18), lo cual implica que cada  $Y_i$  proviene de una población diferente, entonces la ec. (5.19) es un teorema que da el valor esperado.

$$\begin{aligned}\mu_L &= E(L) \\ &= E\left(\sum c_i Y_i\right) = \sum c_i E(Y_i) \\ &= \sum c_i \mu_i\end{aligned}\tag{5.19}$$

Obsérvese que podemos intercambiar el orden del valor esperado y la sumatoria. Esto es posible para los problemas que tenemos que tratar.

Apliquemos esto a la  $\bar{Y}$  calculada para una muestra de una población con  $E(Y) = \mu$ .

$$\begin{aligned}\mu_{\bar{Y}} &= E(\bar{Y}) \\ &= E\left(\frac{\sum Y_i}{n}\right) = \frac{1}{n} \sum E(Y_i) \\ &= \frac{1}{n} \sum \mu = \frac{n\mu}{n} = \mu\end{aligned}$$

Aquí simplemente hemos sacado  $1/n$  como factor constante; también para  $\sum \mu$  hay que escribir  $\mu$  para cada uno de los  $n$  casos.

Por otra parte, para dos muestras de tamaño  $n_1$  y  $n_2$ , cada una de una población posiblemente diferente con  $E(Y_{1j}) = \mu_1$  y  $E(Y_{2j}) = \mu_2$ , tenemos que

$$\begin{aligned}\mu_{\bar{Y}_1 - \bar{Y}_2} &= E(\bar{Y}_{1.} - \bar{Y}_{2.}) \\ &= E(\bar{Y}_{1.}) - E(\bar{Y}_{2.}) \\ &= \mu_1 - \mu_2\end{aligned}$$

Aquí, primero hemos cambiado el orden de sacar el valor esperado de una diferencia por la diferencia de dos valores esperados y luego hemos usado el resultado previo de que  $E(\bar{Y}) = \mu$ . Nótese que  $\mu_1 = \mu_2$ , entonces  $E(\bar{Y}_{1.} - \bar{Y}_{2.}) = 0$  que son las hipótesis de las secciones 5.5, 5.7 y 5.9.

Una varianza se define como un valor esperado por

$$\begin{aligned}\sigma^2 &= E[Y - E(Y)]^2 \\ &= E(Y - \mu)^2\end{aligned}\tag{5.20}$$

Para determinar la varianza de  $L$ , se aplica la ec. (5.19) y se obtiene

$$\begin{aligned}\sigma_L^2 &= E[L - E(L)]^2 \\ &= E(\sum c_i Y_i - \sum c_i \mu_i)^2 = E[\sum c_i(Y_i - \mu_i)]^2\end{aligned}\quad (5.21)$$

El orden de la esperanza y la sumatoria no se pueden intercambiar aquí. Primero hay que efectuar la operación de elevar al cuadrado, en la que aparecen productos cruzados. Ahora necesitamos una definición de su valor esperado, llamado covarianza;

$$\sigma_{12} = E(Y_{1j} - \mu_1)(Y_{2j} - \mu_2) \quad (5.22)$$

Si hay asociación entre valores altos y positivos de  $Y_1$  y  $Y_2$ , y así sucesivamente, entonces la covarianza será positiva. Cuanto más se parezcan los miembros de un par comparado con los miembros de pares diferentes, mayor será la covarianza.

Finalmente, presentamos una fórmula muy general para la varianza de  $L$ , ec. (5.23), que es un teorema respecto de la varianza de  $L$ .

$$\sigma_L^2 = \sum c_i^2 \sigma_i^2 + 2 \sum_{i < j} c_i c_j \sigma_{ij} \quad (5.23)$$

Usualmente entra en juego una varianza común, es decir,  $\sigma_i^2 = \sigma^2$  para todo  $i$ , y el primer término se convierte en  $(\sum c_i^2)\sigma^2$ . A menudo,  $\sigma_{ij} = 0$  ya que el muestreo es aleatorio; decimos que  $Y_{1j}$  y  $Y_{2j}$  son independientes.

Ahora supongamos que tenemos una muestra aleatoria de  $n$  observaciones de una población con media  $\mu$  y varianza  $\sigma^2$ . Calcúlese  $\bar{Y} = \sum Y_i/n$ . Para encontrar  $\sigma_{\bar{Y}}^2$  úsese la ec. (5.23) con  $c_i = 1/n$  para todo  $i$ , así que  $c_i^2 = 1/n^2$ . Como tenemos una sola población,  $\sigma_i^2 = \sigma^2$ , y por el muestreo aleatorio queda asegurada la independencia,  $\sigma_{ij} = 0$ . Por lo tanto, tenemos

$$\sigma_{\bar{Y}}^2 = \sum \left( \frac{1}{n^2} \right) \sigma^2 = \frac{\sigma^2}{n} \quad (5.24)$$

De nuevo,  $\sum$  ha requerido que escribamos la expresión  $n$  veces; por consiguiente, el coeficiente final de  $\sigma^2$  es  $1/n$ .

Para la diferencia entre dos medias de muestras independientes, la ec. (5.23) da

$$\begin{aligned}\sigma_{\bar{Y}_1 - \bar{Y}_2}^2 &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad \text{en general} \\ &= \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \quad \text{común } \sigma^2 \\ &= \frac{2\sigma^2}{n} \quad \text{común } \sigma^2 \text{ y } n_1 = n_2\end{aligned}$$

En estos casos, los  $c_i$  fueron  $+1$  y  $-1$ . Nótese que como  $c_i$  está siempre al cuadrado como coeficiente de  $\sigma^2$ , nunca tenemos  $\sigma^2$  con coeficientes negativos. La varianza de una diferencia debe ser siempre una suma de varianzas.

La sec. (5.7) se ocupó de observaciones pareadas significativas y por lo tanto dependientes. En particular, era de esperar una covarianza positiva.

Supóngase que extraemos un par de observaciones aleatorias, por ejemplo, podemos extraer un individuo al azar y medir su estatura,  $Y_1$ , y su peso  $Y_2$ . Entonces tenemos lo siguiente, incluyendo la ec. (5.25).

$$\begin{aligned} E(Y_{1j} - Y_{2j}) &= \mu_1 - \mu_2 \\ \sigma_{Y_{1j}-Y_{2j}}^2 &= \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} \\ \sigma_{\bar{Y}_1 - \bar{Y}_2}^2 &= n \left( \frac{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}{n^2} \right) \\ &= \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - \frac{2\sigma_{12}}{n} \end{aligned} \quad (5.25)$$

Si  $\sigma_{12}$  es positiva, entonces  $\sigma_{\bar{Y}_1 - \bar{Y}_2}^2$  es menor que si no hubiera pareamiento. Esta es naturalmente la razón del pareamiento. Estamos tratando de detectar una diferencia entre dos medias menor de lo que sería posible sin pareamiento. En general, tenemos dos opciones: podemos incrementar el tamaño de la muestra y disminuir así toda  $\sigma_{\bar{Y}}^2 = \sigma^2/n$ ; podemos controlar el experimento de modo que se reduzca la  $\sigma^2$ . El último enfoque es posible con un pareamiento con sentido cuando  $\sigma_{12}$  es positivo.

En la sección 5.8 la ecuación del modelo para observaciones pareadas con sentido se escribió como:

$$\begin{aligned} Y_{ij} &= \mu + \tau_i + \rho_j + \varepsilon_{ij} \quad i = 1, 2 \\ j &= 1, \dots, n \end{aligned}$$

con  $\varepsilon_{ij}$  distribuidos normal e independientemente con varianza  $\sigma_i^2$ . Vemos que la covarianza es atribuible a la contribución del par,  $\rho$ . Esta contribución es una fuente de variación en las observaciones, pero sin que contribuya a las diferencias entre las medias de los tratamientos; cada media de tratamiento tiene  $\sum \rho_j / n$  una constante. Por tanto, de la unidad de media que nos ayuda a decidir si existe o no diferencia grande  $\bar{Y}_1 - \bar{Y}_2$ , debemos eliminar cualquier contribución que tenga que ver con la variabilidad entre pares. Esto puede hacerse mediante el uso directo de la covarianza o, indirectamente, tomando las diferencias. En el último caso,  $Y_{1j} - Y_{2j} = (\tau_1 - \tau_2) + (\varepsilon_{1j} - \varepsilon_{2j})$ , donde tanto  $\mu$  como  $\rho_j$  desaparecen. En la sec. (5.7) escogimos este último camino.

Una covarianza muestral para  $n$  pares de observaciones se define por

$$s_{12} = \frac{\sum (Y_{1j} - \bar{Y}_1)(Y_{2j} - \bar{Y}_2)}{n - 1} \quad (5.26)$$

Se calcula mediante

$$s_{12} = \frac{\sum Y_{1j} Y_{2j} - (Y_{1\cdot} Y_{2\cdot})/n}{n - 1} \quad (5.27)$$

Nótese que la definición y fórmulas de cálculo para la varianza son aplicaciones particulares de las ecs. (5.26) y (5.27).

**Ejercicio 5.10.1** Supóngase una población de diferencias  $D_j = Y_{1j} - Y_{2j}$  derivada de dos poblaciones  $Y$  con varianzas,  $\sigma_1^2$  y  $\sigma_2^2$  respectivamente, en tal forma que la covarianza es  $\sigma_{12} \neq 0$ .

Demuéstrese que la ec. (5.23) se convierte en  $\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$ .

**Ejercicio 5.10.2** Para los datos de la tabla 5.5, calcular  $s_1^2$ ,  $s_2^2$  y  $s_{12}$  estimativos muestrales de  $\sigma_1^2$ ,  $\sigma_2^2$  y  $\sigma_{12}$ .

Demuéstrese que  $s_D^2 = s_1^2 + s_2^2 - 2s_{12}$  para estos datos.

**Ejercicio 5.10.3** Demuéstrese que los totales de  $n$  observaciones aleatorias a partir de una sola población con varianza  $\sigma^2$ , tienen varianza  $n\sigma^2$ .

## 5.11 Prueba de hipótesis de igualdad de varianzas

En la sec. 5.2, la elección de una región crítica o de rechazo, como se ha visto, depende, en parte, del conjunto de hipótesis alternativas. Las pruebas que se acaban de exponer han sido tratadas desde el punto de vista de una prueba de dos colas, y el conjunto de hipótesis alternativas, que existía una diferencia. Para las pruebas de una cola de tales hipótesis, el criterio de prueba es el mismo, pero la región de rechazo difiere.

El criterio de prueba  $t$  puede considerarse como una comparación de desviaciones estándar y su cuadrado como una comparación de dos varianzas. Tal prueba, la razón de dos varianzas, puede generalizarse de modo que cualesquiera dos varianzas independientes, sin tener en cuenta los grados de libertad de cada una, pueden compararse en la hipótesis nula de que ellas son varianzas muestrales provenientes de poblaciones con varianza común.

Una prueba semejante es útil para decidir si es legítimo o no combinar varianzas, como se ha hecho en la sec. 5.5 al calcular una varianza para probar la hipótesis de igualdad de medias poblacionales, usando muestras con observaciones no pareadas. Un criterio apropiado para probar la hipótesis de homogeneidad se denota mediante  $F$  o  $F(m, n)$ , donde  $m$  y  $n$  son los grados de libertad para los estimativos de la varianza de numerador y denominador, respectivamente.

Si examinamos  $F$  en la ec. (5.6), nos damos cuenta que los valores grandes ocurren cuando se cumple el conjunto de alternativas  $\mu_1 \neq \mu_2$  y que el criterio no distingue entre  $\mu_1 < \mu_2$  y  $\mu_1 > \mu_2$ . Así, nuestro  $F$  es de una cola relativa a  $F$  y a varianzas, pero de dos colas en relación con las medias, es decir,  $H_1: \sigma^2(\text{basado en medias}) > \sigma^2(\text{basado en individuos})$  es equivalente a  $H_1: \mu_1 > \mu_2$  o  $\mu_2 > \mu_1$ , o sea, equivalente a  $H_1: \mu_1 \neq \mu_2$ . Valores de  $t$  cercanos a cero cuando se elevan al cuadrado se hacen positivos, valores de  $F$  cercanos a cero; claramente éstos exigen la aceptación en vez del rechazo de la hipótesis nula. Así  $F$  puede ser pequeño y significante, si el numerador fuese pequeño con relación al denominador. Para nuestro ejemplo debemos explicar tales valores atribuyéndolos al

azar. En ciertos problemas de pruebas de la homogeneidad de dos varianzas, puede no haber razón para suponer cuál varianza será mayor, por ejemplo  $\sigma_1^2$  o  $\sigma_2^2$  para las poblaciones de la tabla 5.6; así, no hay razón para calcular  $s_1^2/s_2^2$  en vez de  $s_2^2/s_1^2$ . Este es un caso en el que las alternativas son  $\sigma_1^2 \neq \sigma_2^2$ . Esto claramente pide una prueba de dos colas.

Considérese la prueba de la hipótesis nula  $\sigma_1^2 = \sigma_2^2$  con el conjunto de alternativas  $\sigma_1^2 \neq \sigma_2^2$ . Determinar  $s_1^2$  y  $s_2^2$  y calcular  $F$  por

$$F = \frac{\text{La } s^2 \text{ mayor}}{\text{La } s^2 \text{ menor}} \quad (5.28)$$

Este  $F$  se compara con valores tabulados en la tabla A.6, donde los grados de libertad para el cuadrado medio mayor se dan en la parte superior de la tabla, y los gl para el menor, en la parte lateral. Para el conjunto de alternativas  $\sigma_1^2 \neq \sigma_2^2$  los niveles de significancia dados se doblan. Si el  $F$  calculado es mayor que  $F_{0.025}$  hay significancia al nivel del 5 por ciento; si es mayor que  $F_{0.005}$ , hay significancia al nivel del 1 por ciento, y así sucesivamente. Esta prueba es una prueba de dos colas con respecto a  $F$ , ya que no especificamos cuál de las  $\sigma^2$  se espera que sea mayor. Desafortunadamente esta prueba también es sensible a la normalidad.

Para el ejemplo de la sec. 5.9, comparar las dos varianzas muestrales con  $F = 40.12/6.95 = 5.77$  para 6 grados de libertad en el numerador y en el denominador. El  $F$  tabulado es 5.82 al nivel del 5 por ciento para las alternativas deseadas ( $F_{0.025} = 5.82$ ) y la prueba casi no llega a ser significante.

Las tablas de  $F$  se tabulan por comodidad al efectuar pruebas de una cola dado que las alternativas asociadas son más comunes; en las pruebas de  $t$  de este capítulo, se ha visto que se esperaba que el numerador fuese mayor cuando la hipótesis nula era falsa, esto es, que la varianza del numerador tenía que ser mayor para eliminar la hipótesis nula. Si se eleva al cuadrado cualquier valor tabulado de  $t_{0.025}$  o  $t_{0.005}$  se encontrará este valor en la tabla de  $F$  en la columna encabezada con 1 grado de libertad frente a los grados de libertad apropiado a 0.05 ó 0.01.

**Ejercicio 5.11.1** Probar la homogeneidad de las varianzas de las dos poblaciones muestreadas en las tablas 5.2, y 5.4, y en los ejercicios 5.5.1 y 5.5.2. ¿Parece justificarse en todos los casos el supuesto de varianza común?

**Ejercicio 5.11.2** Probar la homogeneidad de las varianzas de las dos poblaciones TRC1 de tiempos dados en los ejercicios 5.5.1 y 5.5.2. Repetir lo mismo para las dos poblaciones de TRC2 de los ejercicios 5.5.1 y 5.5.2.

## 5.12 Poder, tamaño de la muestra y determinación de diferencias

Las ideas de la sec. 5.2, o sea, los errores de tipo I y tipo II y el poder de la prueba, son esenciales para una clara comprensión del problema del tamaño de la muestra. La fig. 5.3 ilustra las ideas cuando son apropiadas las alternativas de una cola. Ahora consideraremos el problema con más detenimiento.

Tomemos una muestra aleatoria de una población con  $\mu$  desconocida pero  $\sigma^2$  conocida con el propósito de probar la hipótesis,  $H_0: \mu = \mu_0$  con  $H_1: \mu > \mu_0$ . Se usará una

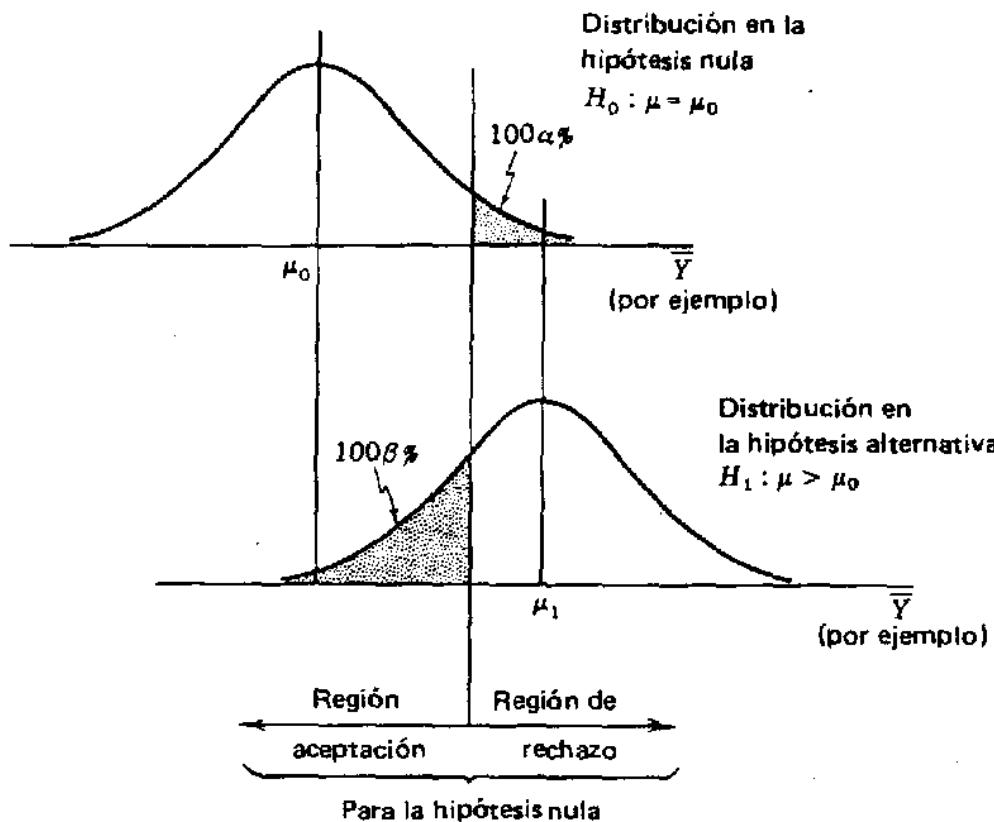


Figura 5.3 Una ilustración relativa a los errores de tipo I y tipo II.

tasa de error  $\alpha$  y el criterio de prueba será  $Z = (\bar{Y} - \mu_0)/\sigma_{\bar{Y}}$  nótese que esto nos permite considerar las regiones de aceptación y rechazo sobre el eje de los  $\bar{Y}$ . Se calcula el valor único crítico con base en que  $H_0$  sea verdadera y es  $\mu_0 + Z_\alpha \sigma_{\bar{Y}}$ . La parte superior de la fig. 5.3 es la que corresponde.

Supóngase que la verdadera media es  $\mu_1$ , tal como se muestra en la parte inferior de la fig. 5.3. Si  $\bar{Y}$  cae en la región de aceptación, entonces se comete un error de tipo II con probabilidad  $\beta$  y el poder de la prueba es  $1 - \beta$ . Nótese que el poder frente a la alternativa, se determinó en el momento en que se fijó el valor crítico y éste a su turno, fue determinado por  $\alpha$  y por la naturaleza unilateral de  $H_1$ .

Independientemente del valor de  $\mu$ , la ec. (5.29) se cumple. Si el verdadero valor  $\mu = \mu_0$  entonces la ec. 5.29 da el valor de  $\alpha$ .

$$P(\text{de rechazo de } H_0) = P\left(\frac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}} > Z_\alpha\right) \quad (5.29)$$

Si el verdadero valor  $\mu = \mu_1$  entonces  $(\bar{Y} - \mu_0)/\sigma_{\bar{Y}}$  realmente no se distribuye como  $Z$ . Sin embargo, podemos reordenar la ecuación para obtener

$$\begin{aligned} P\left(\frac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}} > Z_\alpha\right) &= P\left(\frac{\bar{Y} - \mu_1}{\sigma_{\bar{Y}}} + \frac{\mu_1 - \mu_0}{\sigma_{\bar{Y}}} > Z_\alpha\right) \\ &= P\left(Z > Z_\alpha - \frac{\mu_1 - \mu_0}{\sigma_{\bar{Y}}}\right) \end{aligned} \quad (5.30)$$

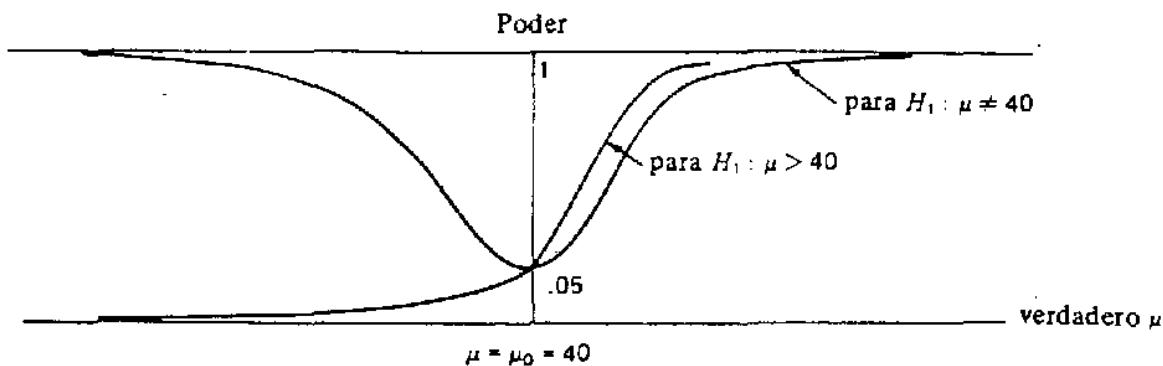


Figura 5.4 Curvas de poder para la ilustración del texto.

Nótese que  $(\bar{Y} - \mu_1)/\sigma_y$  es un verdadero valor de  $Z$  y que  $(\mu_1 - \mu_0)/\sigma_y$  es una constante, lo que hace posible el cálculo del poder a partir de la tabla  $Z$ . Algunos valores, encontrados sin interpolación seria, se presentan en la tabla 5.7 y la curva de poder, basada en estas probabilidades, se muestra en la fig. 5.4.

La curva de poder es útil por cuanto podemos referirnos a ella para tener una mejor impresión del poder de la prueba.

El poder frente a  $H_1: \mu \neq \mu_0$  y con una tasa de error de tipo I,  $\alpha$ , se calcula mediante

$$\begin{aligned} P(\text{de rechazo de } H_0) &= P\left(Z > Z_{\alpha/2} - \frac{\mu_1 - \mu_0}{\sigma_y}\right) \\ &\quad + P\left(Z < -Z_{\alpha/2} - \frac{\mu_1 - \mu_0}{\sigma_y}\right) \end{aligned} \quad (5.31)$$

Uno de los términos, según sea el signo de  $\mu_1 - \mu_0$ , añadirá muy poco a la probabilidad requerida, así que es razonable ignorarlo. En este caso usamos

$$P(\text{de rechazo de } H_0) = P\left(Z > Z_{\alpha/2} - \frac{|\mu_1 - \mu_0|}{\sigma_y}\right) \quad (5.32)$$

La tabla 5.7 da algunos valores de poder para alternativas de dos colas. Estos se representan para dar la curva de poder simétrica de la fig. 5.4.

En muchos experimentos, el experimentador está interesado en determinar si un tratamiento es superior a un control. Para esto es apropiada una prueba de  $t$  de una cola. La tasa de error de tipo I debe fijarse en  $\alpha$  y la tasa de error de tipo II en  $\beta$ . Considérese la fig. 5.5. Supóngase que tenemos una distribución principal con  $\mu$  desconocida y varianza  $\sigma^2$  conocida. Entonces  $a, b, c$  y  $d$  son ilustraciones de posibles distribuciones derivadas de  $\bar{Y}$  con diferentes medias,  $\mu_0, \dots, \mu_3$  y dos varianzas posibles  $\sigma^2/n_i$ ,  $i = 1$  ó  $2$ , según sea el tamaño de la muestra. Las áreas sombreadas se refieren a la población con mayor varianza, es decir, la muestra de menor tamaño. En particular:

**Tabla 5.7 Poder cuando  $\alpha = 0.05$  para ilustración con  $\mu = \mu_0 = 40$**

Verdadero $\mu$	Poder cuando	
	$H_1: \mu > 40^+$	$H_1: \mu \neq 40^+$
24		.9881
26		.9582
28		.8845
.....		
30		.7517
32	.0000	.5596
32.56		.5000
34	.0006	.3522
.....		
36	.0035	.1827
38	.0150	.0828
39	.0281	.0578
40	.0500	.0500
.....		
41	.0838	.0578
42	.2061	.0828
44	.2776	.1827
.....		
46	.4761	.3522
46.23	.5000	
47.44		.5000
48	.6772	.5596
.....		
50	.8389	.7517
52	.9357	.8849
54	.9793	.9582
56	.9945	.9881

† Para  $H_1: \mu > 40$ , el valor crítico es 1.645;  
para  $H_1: \mu \neq 40$ , el valor crítico es 1.96.

- a presenta la distribución de  $\bar{Y}$  si  $H_0$  es verdadera. El área sombreada 100 $\alpha$  por ciento por ejemplo, da la probabilidad de un error de tipo I. La línea que divide las dos áreas separa el eje  $\bar{Y}$  en una región de aceptación y otra de rechazo. El procedimiento de la prueba se basa en  $H_0$  sea cual fuere el verdadero valor de  $\mu$ . Si siempre tenemos datos para los cuales  $H_0$  es verdadera, entonces concluimos erróneamente que  $H_0$  es falsa en 100 $\alpha$  por ciento de las veces.
- b presenta la distribución de  $\bar{Y}$  cuando  $H_0$  es falsa y  $\mu_1$ , es el verdadero parámetro. Si concluimos falsamente que  $H_0$  es verdadera, cometemos un error de tipo II. Cometemos este error cuando  $\bar{Y}$  cae en la región de aceptación determinada por el criterio de prueba usado para probar  $H_0$ . El área sombreada de la curva mide a  $\beta$ , la probabilidad de cometer error de tipo II, por encima del 50 por ciento en este caso.

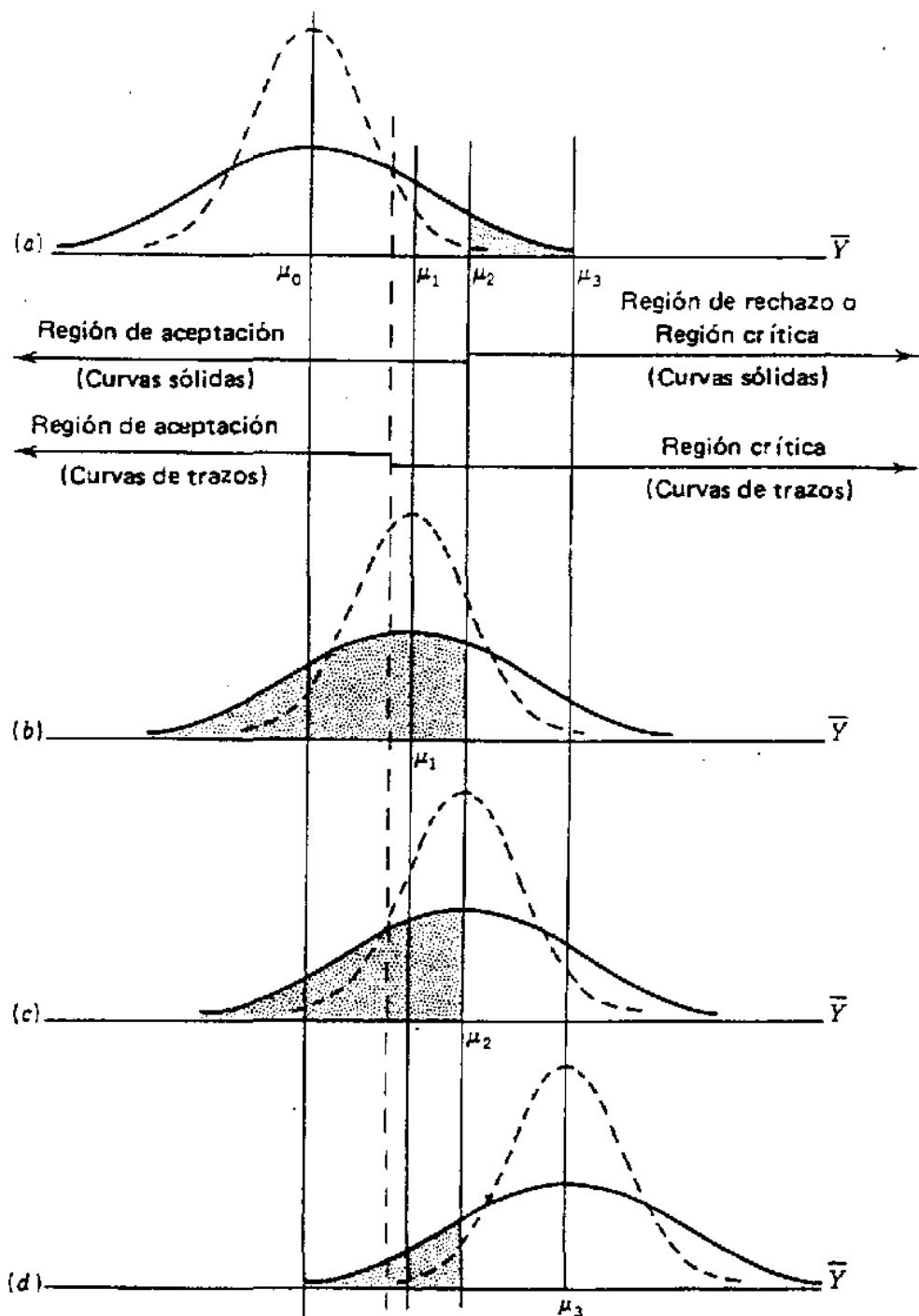


Figura 5.5 Una ilustración relacionada con el tamaño de la muestra.

El poder de la prueba,  $1 - \beta$ , con respecto a la alternativa  $b$  se mide por el área sombreada bajo  $b$ . Esta está asociada con la región crítica determinada por la prueba de  $H_0$ . Cuando, en efecto,  $\mu = \mu_1$ , esta área es la medida de capacidad de la prueba para detectar  $\mu_1$ . En este caso, si el experimentador cuenta con datos para los cuales  $\mu_1$  es verdadera, entonces detectará  $\mu_1$ , con una frecuencia menor al 50 por ciento de las veces. Si fuera importante detectar una diferencia real tan grande como  $\mu_1 - \mu_0$ , entonces sería mejor, en promedio, lanzar una moneda para escoger entre  $\mu_0$  y  $\mu_1$ , pues usando la moneda declararemos el 50 por ciento de las veces a favor de  $\mu_1$ .

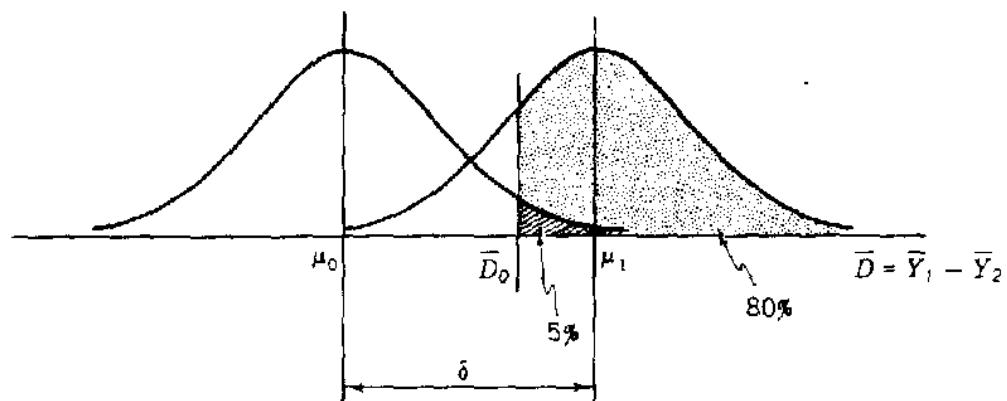


Figura 5.6 Elección de un tamaño de muestra para una protección deseada.

- c Ahora la verdadera  $\mu$  es  $\mu_1$ . Está suficientemente alejado de  $\mu_0$  para ser detectada en el 50 por ciento de las veces. Si  $\mu = \mu_1$  el lanzamiento de la moneda, en relación con el experimento sin recolectar datos es todavía un procedimiento de prueba tan satisfactorio como una basada en  $H_0$ .
- d Si  $\mu = \mu_3$ , tenemos buenas posibilidades de detectar este hecho tal como se indica por el área no sombreada bajo la curva. El poder de la prueba es alto y la probabilidad de un error de tipo II es baja.

Si deseamos una seguridad razonable de detectar  $\mu_1$  cuando  $\mu = \mu_1$ , véase a, será necesario tener una muestra más grande. Esto lleva a una nueva distribución de  $\bar{Y}$  con menor varianza ya que  $\sigma_{\bar{Y}}^2 = \sigma^2/n$ , por ejemplo, una de las distribuciones indicadas por curvas de trazos. La región de aceptación ya no se extiende tanto hacia la derecha. Véase la curva de trazos y la línea vertical de  $a$ .

Ahora la región de rechazo incluye una nueva porción del eje  $\bar{Y}$  a la izquierda de la región original de rechazo. Mejoran así nuestras oportunidades de detectar una alternativa a  $\mu_0$  cuando una alternativa es verdadera. Esto puede verse en las curvas de trazos de b, c y d. El error de tipo II decrece.

Al escoger un tamaño de muestra para detectar una diferencia dada, se debe admitir la posibilidad de un error de tipo I o un error de tipo II, y escoger el tamaño de la muestra en consecuencia. Considérese la fig. 5.6. Aquí hay dos distribuciones de  $\bar{D}$ , una para la cual  $H_0: \mu = \mu_0$  es verdadera y otra para la cual  $H_1: \mu = \mu_1$  es verdadera. El tamaño de la muestra se ha ajustado para dar una varianza  $\sigma_D^2$  tal que el punto  $\bar{D}_0$  divida el eje  $\bar{D}$  para dar un área del 5 por ciento a la derecha de  $\bar{D}_0$  y bajo la distribución de  $H_0$  y un área de 80 por ciento a la derecha de  $\bar{D}_0$  y bajo la distribución de  $H_1$ . Es decir que tenemos una prueba de una cola con una tasa de error del 5 por ciento asociado a  $H_0$  y aceptaremos  $H_1$  en el 80 por ciento de las veces cuando es verdadera; la probabilidad de un error de tipo I es 0.05, la de un error tipo II es 0.20, y el poder de la prueba es 0.80. El problema es encontrar un tamaño de muestra para lograr nuestro objetivo. En la práctica, raramente conocemos  $\sigma^2$ , por lo que debemos confiar en un estimativo.

Cuando entran dos muestras, la variable aleatoria usualmente será  $\bar{Y}_1 - \bar{Y}_2$  y  $H_0$  será  $\mu_1 - \mu_2 = 0$ , mientras que  $H_1$  se convierte en el tamaño de  $\mu_1 - \mu_2 = \delta$  que se desea detectar.

La fórmula aproximada para calcular  $n$ , el número de observaciones en cada tratamiento cuando las alternativas son unilaterales, se da en la ec. (5.33). Como es probable que se obtenga una fracción, úsese el entero siguiente más alto.

$$n = \frac{(Z_\alpha + Z_\beta)^2 \sigma_D^2}{\delta^2} \quad (5.33)$$

Recuérdese que  $\alpha$  y  $\beta$  se refieren a las probabilidades asociadas con una sola cola de la distribución normal. Para unidades experimentales pareadas con sentido,  $\sigma_D^2$  es la varianza de las diferencias, mientras que para muestras independientes es  $2\sigma^2$ .

Esta solución tiene varias dificultades obvias. La primera es que rara vez conocemos  $\sigma^2$ , así que debemos estimarla. Si  $\sigma^2$  se subestima,  $n$  es demasiado pequeño y el poder de la prueba se sobreestima; si  $\sigma^2$  se sobreestima, entonces  $n$  es demasiado grande y el poder de la prueba se subestima.

El problema puede obviarse al definir  $\delta$  en función de  $\sigma$ . Por ejemplo, podemos desear detectar una diferencia de tamaño  $\delta$  con una probabilidad especificada de  $1 - \beta$  si  $\delta$  es la magnitud de una desviación estándar. Tenemos entonces  $\delta = \sigma$ , de modo que  $\sigma/\delta = 1 = \sigma^2/\delta^2$  para usarse en la ec. (5.33).

El raro conocimiento de  $\sigma^2$  también significa que no podemos usar correctamente la tabla de  $Z$ , por lo que nos gustaría reemplazar  $Z_\alpha$  y  $Z_\beta$  por valores de  $t$ . Un procedimiento razonable consiste en usar la ec. (5.33) para lograr un valor aproximado de  $n$ . Multiplíquese este valor por  $(gl \text{ del error} + 3)/(gl \text{ del error} + 1)$ . Para observaciones pareadas, los gl del error serán  $n - 1$ ; para muestras independientes, los gl del error serán  $2(n - 1)$ .

Cuando las alternativas son de dos colas, la ec. (5.34) es una aproximación satisfactoria.

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2 \sigma_D^2}{\delta^2} \quad (5.34)$$

De nuevo, úsese el multiplicador para ajustar este valor al hecho de usar valores de  $Z$ .

Ahora ilustremos el procedimiento.

En la sec. 5.5, para los datos de digestibilidad de materia seca la situación correspondió a muestras independientes de corderos y novillos. Es razonable suponer que los investigadores con animales pueden esperar que los novillos tengan mayores coeficientes promedios de digestibilidad y por lo tanto se propongan probar su hipótesis con alternativas unilaterales. Supóngase que en un nuevo experimento, deseamos detectar, con  $P = 0.8$ , una diferencia real entre medias de una desviación estándar con  $\alpha = 0.05$ . Parecería que ésta puede ser una diferencia real del orden del 2.5 al 3 por ciento ya que  $1(s) = \sqrt{7.33} = 2.7$  por ciento.

En la ec. (5.33),  $\sigma_D^2 = 2\sigma^2$ . Por lo tanto  $\sigma_D^2/\delta^2 = 2\sigma^2/\delta^2 = 2$ . Para una prueba de una cola  $\alpha = 0.05$ ,  $Z_\alpha = 1.65$ . Como  $1 - \beta = 0.80$ ,  $Z_\beta = Z_{0.20} = 0.85$ . Nótese que cada valor de  $Z$  se ha escogido un poco grande en lugar de interpolar en la tabla A.4. Para esos valores prudenciales de  $Z$ ,  $n = (1.65 + 0.85)^2 = 12.5$  y redondeamos a 13.

Con 13 observaciones en cada tipo de animal, el error tendrá  $2(12) = 24$  grados de libertad. El factor de ajuste necesario es  $27/25 = 1.08$  con lo cual se obtiene un valor de  $n$  corregido de  $12.5(1.08) = 13.5$ . Si usamos 14 unidades de cada tipo, la probabilidad de

detectar una diferencia real de una desviación estándar en la magnitud entre medias poblacionales deberá ser de por lo menos 0.80.

En la sección 5.6, los datos correspondieron a concentraciones de azúcar para medias cabezas pareadas de trébol a diferentes presiones de vapor. Supongamos que en el siguiente experimento, quizás con otras presiones, el investigador quedará satisfecho si puede detectar una verdadera diferencia de una  $\sigma$ , basada en diferencias, entre  $\mu_1$  y  $\mu_2$  con probabilidad 0.80. Supóngase que la prueba se hace con dos alternativas bilaterales con  $\alpha = 0.05$ . Aquí usamos la ec. (5.34).

Esta vez,  $\delta/\sigma = 1$ , así  $\sigma_D^2/\delta^2 = 1$ . También necesitamos  $Z_{\alpha/2} = Z_{0.025} = 1.96$  y  $Z_\beta = Z_{0.20} = 0.85$ . Ahora  $n = (1.96 + 0.85)^2 = 7.9$  y redondeando a 8.

Como estamos operando con diferencias, sólo se tienen 7 grados de libertad para el error. El factor de ajuste necesario es  $10/8 = 1.25$  y el valor corregido de  $n$  es  $7.9(1.25) = 9.9$ , y redondeamos a 10. Entonces se necesitarán para el experimento 10 cabezas de trébol, cada una dividida en dos.

### 5.13 Muestras bietápicas de Stein

Cuando se concentra la atención en la estimación más que en la prueba, se usa un procedimiento de Stein (5.12) para determinar el número necesario de observaciones cuando se trata de datos continuos.

El problema consiste en determinar el tamaño de la muestra necesaria para estimar una media por un intervalo de confianza con la garantía que no sea de longitud mayor que la prescrita; el procedimiento consiste en tomar una muestra, estimar la varianza y entonces calcular el número total de observaciones necesarias. Se obtienen entonces las observaciones adicionales y se calcula una media basada en todas las observaciones.

La longitud de un intervalo de confianza es  $2t_{\alpha/2}s_T$  cuando  $\sigma^2$  se conoce, no se necesita una muestra inicial para estimar el tamaño muestral. Este puede calcularse tan pronto se decide la longitud del intervalo de confianza requerido, fijando la longitud igual a  $2z_{\alpha/2}\sigma/\sqrt{n}$  y despejando  $n$ ). Cuando  $\delta^2$  se desconoce, tomamos una muestra y estimamos  $n$  con la ecuación

$$n = \frac{t_1^2 s^2}{d^2} \quad (5.35)$$

$t_1$  es el valor de  $t$  tabulado para el nivel de confianza deseado y los grados de libertad de la muestra inicial y  $d$  es la mitad de la anchura del intervalo de confianza deseado, lo cual da el número total de observaciones necesario. Se obtienen las observaciones adicionales y se calcula una nueva  $\bar{Y}$ . Esta  $\bar{Y}$  estará dentro de la distancia de  $\mu$  a menos que el procedimiento total haya producido una muestra poco usual, tan poco usual que valores más extremos sólo se han de encontrar en un porcentaje no mayor al  $100\alpha$  de las veces, debido al azar. El procedimiento se ha hecho aplicable para obtener el tamaño de la muestra para un intervalo de confianza de una diferencia entre medias poblacionales mediante la multiplicación del numerador de la ec. (5.35) por 2.

*Ejemplo* En un estudio de campo, R.T. Clausen\* deseaba obtener un intervalo de confianza al 95 por ciento de no más de 10 mm para la longitud de hojas maduras de *Sedum lucidum* en varias plantas individuales. En una planta cerca a Orizaba, México, una muestra de cinco hojas dio las longitudes 22, 19, 13, 22, y 23 mm, para las cuales  $\bar{Y} = 19.8$  mm y  $s = 4.1$  mm. Así que el tamaño total de la muestra debe ser

$$\downarrow \\ n = \frac{(7.71)(16.7)}{25} = 5.2 \text{ observaciones}$$

donde  $7.71 = F(1,4) = t^2_{.025}$ ,  $16.7 = s^2$  y  $25 = (10/2)^2$ . El intervalo de confianza escasamente se salió de la norma de longitud aceptable; con una observación más y con una nueva media de todas las 6 observaciones, simplemente diríamos que la nueva media  $\bar{Y}$  estaba dentro de 5 mm de  $\mu$  con un coeficiente de confianza de 0.95.

## Referencias

- 5.1. Baker, G. A., y R. E. Baker: "Strawberry uniformity yield trials" *Biom* 9:412-421 (1953).
- 5.2. Cochran, W. G.: "The combination of estimates from different experiments" *Biom* 10: 101-129 (1954).
- 5.3. Cochran, W. G. y G. M. Cox: *Experimental Designs*, 2a. ed, Wiley, Nueva York, 1957.
- 5.4. Hart, J. S.: "Calorimetric determination of average body temperature of small mammals and its variation with environmental conditions." *Can. J. Zool.*, 29: 224-233 (1951).
- 5.5. Peterson, H. L.: "Pollination and seed set in lucerne," *Roy. Vet. Agr. Coll. Yearb.*, 138-169, (1954).
- 5.6. Ross, R. H., y C. B. Knott: "The effect of supplemental vitamin A upon growth, blood plasma, carotene, vitamin A, inorganic calcium, and phosphorus content of Holstein heifers," *J. Dairy Sci*, 31: 1062-1067 (1948).
- 5.7. Rowles, W.: "Physical properties of mineral soils of Quebec," *Can. J. Res.*, 16: 277-287 (1938).
- 5.8. Satterthwaite, F. W.: "An approximate distribution of estimates of variance components," *Biom. Bull.*, 2: 110-114 (1946).
- 5.9. Searle, S.R.: *Linear models*, Wiley, Nueva York, 1971.
- 5.10. Shuel, R. W.: "Some factors affecting nectar secretion in red clover," *Plant Physiol.*, 27: 95-110 (1952).
- 5.11. Smithson, P. C., y O. T. Sanders, Jr.: "Exposure of bobwhite quail and cottontail rabbits to methyl parathion," *32d Ann. Conf. Southeast Ass. Game Fish Comm.* (1978).
- 5.12. Stein, C.: "A two-sample test for a linear hypothesis whose power is independent of the variance," *Ann. Math. Statist.*, 16: 243-258 (1945).
- 5.13. Watson, C. J., et al.: "Digestibility studies with ruminants XII. The comparative digestive powers of sheep and steers," *Sci. Agr.*, 28: 357-374 (1948).

## PRINCIPIOS DE DISEÑO EXPERIMENTAL

### 6.1 Introducción

Este capítulo es una introducción al planeamiento y conducción de los experimentos en relación con los objetivos, el análisis y la eficiencia.

Si aceptamos la premisa de que el conocimiento nuevo se obtiene muy frecuentemente a través de análisis e interpretación cuidadosos de los datos, entonces es muy importante que se deba dedicar tiempo y esfuerzo considerables al planeamiento y recolección de los mismos con el objeto de obtener la máxima información con el menor costo de recursos. Quizá la función más importante del estadístico consultor sea dar asesoría en el diseño de experimentos eficientes que capaciten al experimentador para obtener estimaciones insesgadas de medias y diferencias de tratamientos y del error experimental.

No es posible destacar demasiado el hecho de que el estadístico pueda hacer una contribución real en la etapa de planeamiento de los experimentos. Con frecuencia se presentan al estadístico datos que sólo dan estimaciones sesgadas de las medias y diferencias de los tratamientos y del error experimental; datos que no proporcionan respuestas a las preguntas planteadas al principio; datos para los cuales ciertos tratamientos no dan la información pertinente; para los cuales las conclusiones sacadas no se aplican a la población que el experimentador tenía en mente; y tales que la precisión del experimento no es suficientemente grande para detectar diferencias importantes. A menudo, con un leve cambio en el diseño y con menor esfuerzo, el experimento hubiera dado la información deseada. El experimentador que consulte al estadístico consultor en la etapa de planeamiento del experimento, en lugar de al final del mismo, mejora las posibilidades de conseguir sus objetivos.

### 6.2 ¿Qué es un experimento?

Existen diferentes definiciones de la palabra *experimento*. Para nuestro propósito, consideramos un experimento como una búsqueda planeada para obtener nuevos conocimientos

o para confirmar o no resultados de experimentos previos, con lo que tal indagación ayudará en la toma de decisiones administrativas, tales como la recomendación de una variedad, un procedimiento o un pesticida. Tales experimentos caen aproximadamente dentro de tres categorías, esto es, preliminares, críticos y demostrativos, cada una de las cuales puede llevar a otra. En un experimento preliminar, el investigador prueba un número grande de tratamientos con el objeto de obtener indicios para futuros trabajos; la mayoría de los tratamientos aparecen solamente una vez. En un experimento crítico, el investigador compara las respuestas a diferentes tratamientos usando un número suficiente de observaciones de las respuestas para tener seguridad razonable de detectar diferencias significantes. Los experimentos demonstrativos se llevan a cabo cuando los trabajadores de extensión comparan uno o más tratamientos nuevos con un patrón. En este texto, estamos interesados casi completamente en el tipo de experimentos críticos. En tal experimento, es necesario que definamos la población a la cual se ha de aplicar las inferencias, que diseñemos el experimento en concordancia con eso, y que hagamos medidas de las variables bajo estudio.

Se dispone cada experimento para proporcionar respuestas a una o más preguntas. Con esto en mente, los investigadores deciden qué comparaciones de tratamientos proporcionarán información relevante. Entonces realizan un experimento para medir o probar hipótesis que tienen que ver con diferencias entre tratamientos en condiciones comparables. Toman mediciones y observaciones sobre el material experimental. A partir de la información obtenida en un experimento que se ha completado con éxito, responden a las preguntas planteadas al comienzo. En este capítulo nos importan principalmente los procedimientos.

Para el estadístico, el experimento es un conjunto de reglas usadas para sacar la muestra de una población. Esto hace que la definición de la población sea lo más importante. El conjunto de reglas es el procedimiento experimental o diseño de experimento. Por ejemplo, el uso de observaciones no pareadas y pareadas son diseños experimentales para experimentos de dos tratamientos.

### 6.3 Objetivos de un experimento

Al diseñar un experimento, se establecen claramente los objetivos como preguntas que han de responderse, hipótesis que han de probarse, y efectos que han de estimarse. Es aconsejable clasificar los objetivos como mayores y menores ya que ciertos diseños experimentales dan más precisión para ciertas comparaciones de tratamientos que otros.

La precisión, sensibilidad o cantidad de información, se mide por el inverso de la varianza de una media. Si  $I$  representa la cantidad de información, entonces  $I = 1/\sigma^2 = n/\sigma^2$ . A medida que  $\sigma^2$  aumenta, la cantidad de información decrece, también a medida que  $n$  aumenta la cantidad de información aumenta. Una comparación de dos medias muestrales se hace más sensible, esto es, puede detectar una diferencia más pequeña entre medias poblacionales, a medida que el tamaño de la muestra crece.

Es de gran importancia definir la población para la cual deben extraerse inferencias y tomar la muestra de esa población en forma aleatoria. Supóngase que el objetivo principal de un experimento es comparar los valores de varias raciones para cerdos en cierta región. Supóngase, también, que los granjeros en esa región crían diferentes razas de /

cerdos, que algunos usan comederos mecánicos y otros los alimentan manualmente. Si el experimentador usa solamente una raza de cerdos en el experimento o los alimenta solamente con comederos mecánicos, la muestra difícilmente puede considerarse representativa de la población, a menos que haya información previa de que el tipo de comederos y de alimentación tienen poco o ningún efecto sobre diferencias debidas a las raciones. Si no hay información disponible sobre el efecto de raza sobre el método de alimentación, sería muy aventurado hacer inferencias a partir de un experimento basado en una raza y un método de alimentación para todas las razas y todos los métodos. Para hacer tales recomendaciones, el experimentador debe incluir todas las razas locales y todos los métodos de alimentación como factores del experimento. Al hacer esto, se amplía el alcance del experimento.

Otro ejemplo es el de un experimento planeado para comparar la eficacia de diferentes fungicidas en el control de una enfermedad en la avena. Supóngase que en la región donde se hayan de aplicar las recomendaciones, diversas variedades cultivadas artificialmente o "cultivares" son de uso común y que varía la cantidad de las semillas. Para poder dar recomendaciones adecuadas, el experimentador debe comparar los fungicidas en varios cultivares y con varias calidades de semillas por cada cultivar. Esto es esencial si el experimentador va a determinar si se puede recomendar o no un solo fungicida para todos los cultivares y calidades de semilla en la región. En general, inferencias de valor respecto a una población extensa no pueden obtenerse a partir de un solo experimento.

#### 6.4 Unidad experimental y tratamiento

Una unidad experimental, o parcela experimental, es la unidad de material a la cual se aplica un tratamiento; el tratamiento es el procedimiento cuyo efecto se mide y se compara con otros tratamientos. Se ve, pues, que estos términos son muy generales. La unidad experimental puede ser un animal, 10 pollos en corral, media hora, etcétera; el tratamiento puede ser una ración normal, un programa de aspersión, una combinación temperatura-humedad, u otros. Cuando se mide el efecto de un tratamiento, se mide en una unidad de muestreo, cierta fracción de la unidad experimental. Por lo tanto la unidad de muestreo puede ser la unidad completa, tal como un animal sometido a una ración de tratamiento, o una muestra aleatoria de hojas de un árbol tratado o la cosecha de 6 pies del surco central de una unidad experimental de 3 surcos. En algunos casos, la unidad experimental será tan grande que su uso no sea práctico como unidad de muestreo, en tanto que una sola unidad de muestreo pequeña es inadecuada. En tales casos, se miden dos o más subdivisiones aleatorias de la unidad experimental. Por ejemplo, al estudiar prácticas de cultivo para establecer la densidad de especies forrajeras se cuentan separadamente las plántulas en dos o más áreas pequeñas, aleatorias dentro de una unidad experimental. Así mismo, cuando se debe destruir una unidad de muestreo, como en el caso de la calidad de frutos, legumbres o un producto alimenticio para el mercado, se toman unidades de muestreo dentro de la unidad experimental.

Al seleccionar un conjunto de tratamientos, es importante definir cada tratamiento cuidadosamente y considerarlo con respecto a cada uno de los demás tratamientos para asegurarse, en lo posible, que el conjunto dé respuestas eficientes relacionadas con los objetivos del experimento.

## 6.5 Error experimental

Una característica de todo material experimental es la variación. El error experimental es una medida de la variación existente entre observaciones sobre medidas experimentales tratadas en forma similar. Esta afirmación es más sutil de lo que puede parecer a primera vista y se relaciona estrechamente con la definición de la sección precedente. Por ejemplo, si a 50 gallinas se las enjaula juntas y se les alimenta con la misma ración, la unidad experimental consiste en las 50 gallinas. Se necesitan otras jaulas de 50 gallinas antes de poder medir la variación entre unidades tratadas en forma semejante. Esto es cierto aun si una medida como el peso del cuerpo se mide en cada gallina en forma individual. El punto está en que si dos tratamientos se han de comparar, cualquier diferencia observada será en parte atribuible simplemente a la diferencia entre jaulas de 50 gallinas y esto es muy probable que sea de mayor magnitud que las diferencias entre gallinas de la misma jaula. Por otra parte, si se enjaulan 10 ratones juntos como una unidad para alimentar, se aplica el mismo argumento. Sin embargo, si la mitad de los animales son machos y la mitad hembras, y si estamos interesados en una posible respuesta diferencial debido al sexo, entonces tenemos unidades experimentales pareadas con sentido para un experimento dentro de otro. Aquí se presenta un segundo error experimental para responder a una segunda pregunta. Para responder a toda pregunta respecto a la posible presencia de una verdadera diferencia de tratamientos, siempre debemos tratar con la unidad a la cual se aplicó el tratamiento al azar. Estas y otras preguntas relativas a la elección de un error experimental apropiado se ilustran y se exponen en capítulos posteriores.

La variación proviene de dos fuentes principales. Primero, existe la variabilidad inherente al material experimental al cual se aplican los tratamientos. Segundo, existe una variación resultante de cualquier falta de uniformidad en la realización física del experimento. En un experimento de nutrición con ratas como material experimental, los individuos tendrán constitución genética diferente a menos que haya una alta endogamia; ésta es variabilidad inherente al material experimental. Las ratas se colocarán en jaulas sujetas a diferencias de calor, luz y otros factores, esto constituye una falta de uniformidad en la realización física del experimento. Las magnitudes relativas de la variación de estas dos fuentes serán bastante diferentes según los varios campos de investigación.

La magnitud de un intervalo de confianza y el poder de una prueba dependen en definitiva de  $V(\bar{Y}) = \sigma^2/n$ . Así que para obtener intervalos cortos o alto poder de la prueba, sólo hay dos puntos que tener en cuenta. En consecuencia, es importante hacer todo el esfuerzo posible para reducir el error experimental con el fin de mejorar el poder de la prueba, para disminuir el tamaño de los intervalos de confianza o para lograr otro objetivo deseable. Esto puede lograrse atendiendo a las dos principales fuentes de error experimental. Así podemos

1. Manejar el material experimental de tal manera que se logre reducir los efectos debidos a la variabilidad inherente.
2. Refinar la técnica experimental.

Estos métodos se verán en las secciones subsiguientes a la exposición de la repetición y de los factores que la afectan al número de repeticiones.

## 6.6 Repeticiones y sus funciones

Cuando un tratamiento aparece más de una vez en un experimento se dice que está *repetido*. Las funciones de la repetición son:

1. Permitir una estimación del error experimental.
2. Mejorar la precisión de un experimento mediante la reducción de la desviación estándar de una media de tratamiento.
3. Aumentar el alcance de la inferencia del experimento a través de la selección y del uso apropiado de unidades experimentales más variables.
4. Ejercer control sobre la varianza del error.

Para las pruebas de significancia y para la estimación del intervalo de confianza, es necesario un estimativo del error experimental. De un experimento en el cual cada tratamiento aparece sólo una vez, se dice que consiste en una repetición simple. De un experimento como éste, no es posible estimar el error experimental. Aquí es posible explicar una diferencia observada como una diferencia entre tratamientos o entre unidades experimentales; es imposible saber con seguridad objetiva cuál explicación es la correcta. O sea, cuando no existe un método para estimar el error experimental, no hay manera de determinar si las diferencias observadas indican diferencias reales o si se deben a la variación inherente. El experimento no es autosuficiente ya que toda inferencia debe basarse en experiencias previas. (*Excepción*. Se puede utilizar una sola repetición, o aun una fracción particular de una repetición de un experimento con un número grande de factores o tipos de tratamientos, esto permitirá un estimativo del error experimental cuando se cumplan ciertos supuestos).

A medida que el número de repeticiones aumenta, las estimaciones de las medias poblacionales, esto es, las medias observadas de los tratamientos, se hacen más precisas. Si se detecta una diferencia de cinco unidades usando 4 repeticiones, un experimento de aproximadamente 16 repeticiones detectará la mitad de esa diferencia, o sea 2,5 unidades, pues las desviaciones estándar están en proporción 2:1, siendo  $\sigma/\sqrt{4}$  y  $\sigma/\sqrt{16}$ , respectivamente. Se usa la palabra "aproximado" porque la precisión especialmente en experimentos pequeños depende en parte de los grados de libertad disponibles para estimar el error experimental. También, aumentar el número de repeticiones puede exigir el uso de material experimental menos homogéneo o una técnica menos cuidadosa, dando así una nueva población principal con un mayor error experimental. Sin embargo, el aumento de las repeticiones por lo general mejora la precisión, disminuyendo las longitudes de los intervalos de confianza y aumentando el poder de las pruebas estadísticas.

En ciertos tipos de experimentos, la repetición es un medio de aumentar el alcance de la inferencia de un experimento; la población muestreada es menos restringida en su definición y así se amplía la inferencia. Por ejemplo, supóngase que deseamos determinar si existe una diferencia real en el rendimiento de dos variedades de un cultivo en una región, y que hay dos tipos principales de suelos en esta área. Si el objetivo del experimentador es sacar inferencias para ambos tipos de suelo, es obvio que ambos deben entrar en el experimento. También es importante que el área incluida dentro de cada repetición, esto es, en cada par de parcelas en las cuales se siembran las variedades, sea de un tipo de

suelo tan uniforme como sea posible; no es necesario y puede ser indeseable tener condiciones muy uniformes entre repeticiones, especialmente si la población en cuestión es bastante amplia.

En muchos experimentos sobre el terreno, el experimento se repite en un período de años. La razón es obvia, ya que las condiciones varían año a año y es importante conocer el efecto de los años sobre las diferencias entre los tratamientos, porque las recomendaciones se suelen hacer para años futuros. De la misma manera, se usan diferentes localidades para evaluar los tratamientos en las diferentes condiciones ambientales presentes en la población, esto es, en la región para la cual van a hacerse las recomendaciones. Tanto las repeticiones en tiempo (años) como en espacio (localidades) pueden considerarse tipos amplios de repetición. El propósito es aumentar el alcance de la inferencia. Frecuentemente se usa el mismo principio en experimentos de laboratorio, o sea que el experimento completo se repita varias veces, posiblemente por personas diferentes, para determinar la repetibilidad de los tratamientos en condiciones posiblemente diferentes que puedan existir en el laboratorio de tiempo en tiempo.

Finalmente, la repetición nos permite agrupar unidades experimentales de acuerdo con la respuesta esperada en ausencia de tratamientos. El objeto es repartir la variación total entre las unidades experimentales de tal manera que se maximice entre los grupos y se minimice simultáneamente dentro de ellos. Ahora bien, si se aplica un conjunto de tratamientos dentro de cada grupo de éstos, las diferencias observadas medirán las verdaderas diferencias de tratamientos mejor que si la agrupación se hubiera pasado por alto. Es claro que no debe inflarse el error experimental, que es nuestra medida para detectar diferencias reales entre tratamientos, debido a diferencias entre grupos. Se cuenta con las matemáticas apropiadas para lograr este propósito.

## 6.7 Factores que afectan el número de repeticiones

El número de repeticiones de un experimento depende de varios factores, de los cuales el más importante es el grado de precisión deseada. Cuanto más pequeña sea la discrepancia con respecto a la hipótesis nula que se ha de medir o detectar, mayor será el número de repeticiones requeridas.

En un experimento, es muy importante tener una cuantía correcta de precisión. No tiene objeto usar 10 repeticiones para detectar una diferencia que se puede detectar con 4 en la mayoría de los casos; tampoco es de utilidad llevar a cabo un experimento en el que el número de repeticiones sea insuficiente para detectar diferencias importantes, excepto ocasionalmente.

Para medir cualquier discrepancia con respecto a la hipótesis nula, el error experimental, esto es, la variación entre observaciones experimentales tratadas de manera análoga, debe proporcionar la unidad de medida. A veces no es práctico hacer una observación en la unidad experimental completa; por ejemplo, en el caso de determinaciones químicas, como la del contenido proteínico de un forraje, se toman muestras de la unidad experimental. Usualmente la variación entre unidades experimentales es grande en comparación con la variación entre muestras de una sola unidad. No hay razón para detectar un número grande de determinaciones químicas en cada unidad experimental, ya que el error ex-

perimental debe estar basado en la variación entre las unidades experimentales, no en la variación entre las muestras de esas unidades.

Ciertos materiales son naturalmente más variables que otros. Considérese el problema de heterogeneidad del suelo. Ciertos suelos son más uniformes que otros y, para la misma precisión, se necesitan menos repeticiones en suelos uniformes que en los heterogéneos. Igualmente, diferentes cultivos desarrollados en la misma localidad presentan variabilidad desigual.

El número de tratamientos afecta la precisión de un experimento y el número de repeticiones necesarias para un grado de precisión dado. Por ejemplo, si aumentamos el número de tratamientos y mantenemos constante el número de repeticiones para cada uno, entonces aumentamos el tamaño del experimento y el número de grados de libertad para la estimación de  $\sigma^2$ . Aún tenemos  $s_y^2 = s^2/n$  pero es mejor la estimación de  $\sigma^2$  y, por tanto, es mejor la precisión. El número de repeticiones puede reducirse si no se requiere un aumento de precisión. Por otra parte, si mantenemos constante el tamaño del experimento, entonces un mayor número de tratamientos implica menos repeticiones de cada uno de ellos y menos grados de libertad para estimar  $\sigma^2$ . Ahora  $s_y^2 = s^2/n$  tiene un menor  $n$  y una estimación más deficiente de  $\sigma^2$ . Como resultado se tiene menor precisión. El número de repeticiones debe aumentarse para lograr una precisión fijada. Todo este razonamiento es más apropiado para experimentos pequeños, por ejemplo, con menos de 20 grados de libertad en el error.

El diseño experimental también afecta la precisión de un experimento y el número de repeticiones necesarias. Cuando el número de tratamientos es grande y es necesario usar unidades experimentales más heterogéneas, entonces aumenta el error experimental por unidad. Los diseños experimentales apropiados pueden controlar parte de esta variación.

Desafortunadamente, el número de repeticiones a menudo está determinado en gran parte por los fondos y el tiempo disponible para el experimento. No se justifica un experimento si no puede obtenerse la necesaria precisión con los fondos de que se dispone. La solución es posponer el experimento hasta que se tengan los fondos suficientes, o reducir el número de tratamientos de modo que se tengan el número de repeticiones necesarias y la precisión para los tratamientos restantes. El número práctico de repeticiones para un experimento se alcanza cuando el costo del experimento en material, tiempo, etc., ya no se compensa con un aumento de la información.

Vale la pena mencionar el hecho de que las repeticiones no reducen el error debido a técnicas defectuosas. Además, la sola significancia estadística puede no dar información sobre la importancia práctica de una discrepancia con respecto a la hipótesis nula. La magnitud de una discrepancia real de significancia práctica puede juzgarse mediante el conocimiento técnico del tema en estudio, esta magnitud junto con una medida de lo deseable que sea detectarla sirve para determinar la precisión requerida y, en definitiva, el número de repeticiones necesarias.

Como se ha dicho tanto sobre el número apropiado de repeticiones, se plantea la pregunta respecto al método para garantizar ese número. El problema de la determinación del tamaño de la muestra se estudió en las secs. 5.13 y 5.14. Exposición más detallada se da en la sec. 9.15. Con la información apropiada, el experimentador puede, por lo general, encontrar un método para determinar las repeticiones necesarias.

## 6.8 Precisión relativa de diseños con pocos tratamientos

La precisión o la cantidad de información de un experimento se mide por  $I = n/\sigma^2$ . Entonces una estimación de la información depende de lo bien que  $s^2$  estima a  $\sigma^2$  y esto está afectado por el número de grados de libertad disponibles para la estimación. Los grados de libertad dependen del número de repeticiones, del número de tratamientos y del diseño experimental. Cuando el número de grados de libertad es inferior a 20, bien vale la pena incrementar la precisión. Así, obsérvese que  $t_{0.025}$  para 5 grados de libertad, es 2.57; para 10 grados de libertad, 2.23; para 20 grados de libertad, 2.09 y para 60 grados de libertad, 2.00. El valor de  $t$  a todo nivel de probabilidad disminuye notablemente por cada grado más de libertad hasta llegar a 20; más allá de este valor, la disminución es lenta.

Para comparar dos diseños experimentales, se compara la cantidad de información. Esta da la *eficiencia relativa*. El procedimiento de Fisher (6.5), tal como lo presenta Cochran y Cox (6.2), estima la eficiencia del diseño 1 en relación con el diseño 2 mediante

$$RE = \frac{(n_1 + 1)/(n_1 + 3)s_1^2}{(n_2 + 1)/(n_2 + 3)s_2^2} = \frac{(n_1 + 1)(n_2 + 3)s_2^2}{(n_2 + 1)(n_1 + 3)s_1^2} \quad (6.1)$$

donde  $s_1^2$  y  $s_2^2$  son los cuadrados medios del error de los diseños uno y dos respectivamente, y  $n_1$  y  $n_2$  son sus grados de libertad. Si el número de observaciones en una media de un tratamiento difiere en los dos experimentos en comparación, sustitúyase  $s_1^2$  y  $s_2^2$  por  $s_y^2$  y  $s_{\bar{Y}_2}^2$ .

Supóngase que deseamos comparar un diseño que tiene 5 grados de libertad para estimar  $\sigma^2$  con otro que tiene 10 grados de libertad; esto podría ser una comparación de un experimento pareado con uno no pareado de dos tratamientos y seis repeticiones. El diseño pareado es más preciso que el no pareado sólo si la eficiencia del primero con relación al último es mayor que 1, esto es, si la eficiencia relativa

$$\frac{(n_1 + 1)(n_2 + 3)s_2^2}{(n_2 + 1)(n_1 + 3)s_1^2} = \frac{(6)(13)s_2^2}{(11)(8)s_1^2} = \frac{.886 s_2^2}{s_1^2}$$

es mayor que 1. O sea que el pareamiento se justifica si  $0.886 s_2^2$  es mayor que  $s_1^2$ .

## 6.9 Control del error

El control del error puede lograrse mediante

1. El diseño experimental
2. El uso de observaciones concomitantes
3. La elección del tamaño y la forma de las unidades experimentales.

1 *El diseño experimental* El uso del diseño experimental como medio de controlar el error experimental ha sido ampliamente investigado, desde más o menos el final del primer cuarto del siglo presente. Este es un tema amplio, así que sólo se expondrán aquí los principios básicos. Para un tratamiento más extenso del tema, se recomienda Cox (6.3), Cochran y Cox (6.2), Federer (6.4), John (6.7), Kempthorne (6.8), Youden (6.9) y otros.

El control del error experimental consiste en el diseño de un experimento de tal manera que parte de la variación natural entre el conjunto de unidades experimentales se trate materialmente de modo que no contribuya en nada a las diferencias entre medias de tratamientos. Por ejemplo, considérese un experimento de dos tratamientos en el que se utilizan dos cerdos de cada una de 10 camadas. Si aplicamos los tratamientos, uno a cada miembro de un par, entonces la descripción matemática es  $Y_{ij} = \mu + \tau_i + \rho_j + \varepsilon_{ij}$  y la diferencia entre las medias es  $\bar{Y}_1 - \bar{Y}_2 = (\tau_1 - \tau_2) + (\bar{\varepsilon}_1 - \bar{\varepsilon}_2)$ ; no hay contribución debida a la variación entre los pares, esto es, debida al conjunto de  $\rho$ . De otra manera, si no existe pareamiento y los 10 cerdos se escogen simplemente al azar entre los 20 y se asignan a un tratamiento específico, entonces la descripción es  $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ , donde cada  $\varepsilon_{ij}$  incluye ahora una contribución proveniente de las camadas y es un elemento de una población con una varianza presumiblemente mayor que la relacionada con los  $\varepsilon$  del primer modelo. Así, la elección de un diseño con pares no aleatorios ha controlado el error experimental  $\sigma^2$ , siempre que los pares sean una fuente adicional de variación, esto es, siempre que la varianza natural entre las observaciones sobre individuos del mismo par sea menor que la de las observaciones entre individuos de pares diferentes. Se ha visto que el sentido común y la agudeza para reconocer fuentes de variación son básicos en la elección de un diseño.

Cuando se agrupan las unidades experimentales en bloques completos, esto es, donde cada bloque contiene todos los tratamientos, de modo que la variación entre unidades dentro de un bloque sea menor que la variación entre unidades en bloques diferentes, la precisión del experimento aumenta como resultado del control del error. Tales bloques de resultados semejantes se llaman también *repeticiones*. El diseño se llama *diseño de bloques completos aleatorizados*. El error experimental se basa en la variación entre unidades dentro de una repetición después del ajuste de cualquier efecto de tratamiento global observado. Es decir, la variación entre repeticiones y la variación entre tratamientos no entran en el error experimental.

A medida que el número de tratamientos en un experimento aumenta, el número de unidades experimentales necesarias para una repetición también aumenta. En la mayoría de los casos, ésto lleva a un incremento en el error experimental, o sea, un aumento en la varianza de la población principal. Se cuenta con diseños en los cuales se subdivide el bloque completo en varios bloques incompletos de modo que cada bloque incompleto contenga sólo una porción de los tratamientos. La subdivisión en bloques incompletos se hace de acuerdo con ciertas reglas, de tal manera que el error experimental pueda estimarse entre unidades dentro de los bloques incompletos. La precisión se aumenta hasta el punto en que las unidades experimentales dentro de un bloque incompleto sean más uniformes que los bloques incompletos dentro de una repetición. Tales diseños se llaman *diseños de bloques incompletos*. Se remite al lector a Cochran y Cox (6.2) y Federer (6.4).

El diseño de *parcelas divididas* es un diseño de bloques incompletos en el que la precisión de ciertas comparaciones se aumenta a expensas de otras. La precisión total es la misma que la del diseño básico usado. En el capítulo 16 se estudian algunos diseños de parcelas divididas.

En experimentos en los que las comparaciones entre todos los tratamientos son esencialmente de la misma importancia, se usa un tipo diferente de diseño de bloques incompletos. Los tratamientos se asignan a las unidades experimentales de manera que cada tratamiento se presenta el mismo número de veces con cada uno de los otros trata-

mientos dentro de su conjunto completo de bloques incompletos. Tales diseños se conocen como *bloques incompletos balanceados*. Hay otro grupo, conocido como de *látices parcialmente balanceados*, donde cada tratamiento se presenta solamente con ciertos tratamientos, no con todos, en su conjunto de bloques incompletos.

El mejor diseño en una situación dada es el diseño más simple disponible que proporcione la precisión requerida. No vale la pena usar un diseño complicado si no se obtiene un aumento de precisión.

**2 El uso de observaciones concomitantes** En muchos experimentos, se puede aumentar la precisión mediante el uso de observaciones complementarias y una técnica aritmética llamada *análisis de covarianza*. Se usa este análisis cuando la variación entre unidades experimentales se debe, en parte, a variación en alguna otra u otras características medibles y no suficientemente controlables para que sean útiles al asignar unidades experimentales a bloques completos o incompletos con base en resultados semejantes. El análisis de la covarianza se expone en el capítulo 17.

**3 Tamaño y forma de las unidades experimentales** Como regla general, las unidades experimentales grandes presentan menos variación que las pequeñas. Esta regla se cumple, en particular, cuando los errores aleatorios se distribuyen normalmente con varianza común, y es una consecuencia de la relación  $\sigma_y^2 = \sigma^2/n$ . Sin embargo, un aumento en el tamaño de la unidad experimental trae a menudo como consecuencia una reducción en el número de repeticiones que pueden tenerse debido a lo limitado del material experimental que suele disponerse en un experimento dado. Generalmente es más útil lograr una repetición adecuada de parcelas pequeñas que de parcelas grandes.

En experimento con parcelas sobre el terreno, el tamaño y forma de la unidad experimental, o parcela, lo mismo que el tamaño y forma de un bloque completo o incompleto, son importantes respecto a la precisión. Los estudios de *ensayos de uniformidad*, es decir, estudios de datos obtenidos de experimentos donde no se aplican tratamientos, efectuados con muchos cultivos y en muchos países, han demostrado que las parcelas individuales deben ser relativamente largas y angostas para lograr la máxima precisión; el bloque, completo o incompleto, debe ser aproximadamente cuadrado. Para una varibilidad dada entre unidades experimentales, este proceder tiende a maximizar la variación entre bloques, a la vez que minimiza la variación entre parcelas dentro de los bloques. Una variación grande entre bloques indica que su uso ha sido útil porque esta variación se elimina del error experimental y, además, no contribuye a las diferencias entre las medias de tratamiento. Cuando los bloques son cuadrados, las diferencias entre bloques tienden a ser grandes. Es conveniente tener una variación pequeña entre parcelas dentro de bloques y, al mismo tiempo, tener la parcela representativa del bloque. En terrenos donde se presentan curvas de nivel de fertilidad definidas, la mayor precisión se obtiene cuando los lados largos de las parcelas son perpendiculares a las curvas o paralelos a la dirección de el gradiente.

Para algunos tipos de experimentos, las unidades experimentales se seleccionan cuidadosamente de modo que sean tan uniformes como sea posible; por ejemplo, ratas provenientes de una línea de endogamia pueden ser las unidades experimentales en un experimento de nutrición. Como consecuencias de esto, se reduce el error experimental,

pero al mismo tiempo el alcance de la inferencia se reduce. La respuesta que se obtiene con unidades experimentales seleccionadas puede diferir ampliamente de la obtenida con unidades no seleccionadas, y las inferencias deben hacerse teniendo esto en cuenta.

### 6.10 Elección de los tratamientos

En ciertos tipos de experimentos, los tratamientos tienen efecto sustancial sobre la precisión. Esto es especialmente cierto en el caso de los experimentos factoriales, de que se trata en el capítulo 15.

En ciertos tipos de experimentos, la cantidad o proporción de cierto factor es importante. Supóngase que el experimentador está midiendo el efecto del aumento de los niveles de un nutriente sobre la respuesta de una planta. Es importante incluir varios niveles para determinar si la respuesta es de naturaleza lineal o curvilinea. Aquí, la elección del número de niveles así como su espaciamiento es importante para lograr respuestas apropiadas a las preguntas planteadas.

En general, a mayor conocimiento del experimentador en cuanto a tratamientos, mejor es el procedimiento de contraste que el estadístico puede idear. Este conocimiento a menudo impone la clase y cantidad de un tratamiento particular en un conjunto de tratamientos. Esto, a su vez, puede influir en la precisión del experimento. Algunos aspectos de este problema se verán en el capítulo 15.

### 6.11 Refinamiento de la técnica

La importancia de una técnica cuidadosa en la conducción de un experimento es evidente por sí misma. Es responsabilidad del experimentador, ver que se haga todo lo posible para asegurar una técnica cuidadosa porque ningún tipo de análisis estadístico, o de otra clase, puede mejorar los datos obtenidos mediante experimentos mal efectuados. En general, la variación proveniente de una técnica deficiente no es aleatoria, y no está sujeta a las leyes del azar en las cuales se basa la inferencia estadística. Esta variación puede denominarse inexactitud, en contraste con la falta de precisión. Desafortunadamente, la exactitud en la técnica no siempre produce alta precisión ya que ésta también tiene que ver con la variabilidad aleatoria entre las unidades experimentales, la cual puede ser bastante grande.

A continuación se consideran algunos puntos de técnica que deben tenerse en cuenta al efectuar un experimento. Es importante, tener uniformidad en la aplicación de los tratamientos. Esto se aplica tanto en la aspersión de fertilizantes o en la aspersión de árboles frutales, como en el corte de forraje a una altura fija en todas las parcelas y el llenado de tubos de ensayo a un nivel dado. Deberá ejercerse control sobre las influencias externas, de tal manera que todos los tratamientos produzcan sus efectos en condiciones comparables y deseables. Por ejemplo, para comparar los efectos de varios tratamientos puede ser necesario crear una epidemia mediante el uso de un inóculo artificial. Deben hacerse esfuerzos para lograr la mayor uniformidad posible en la epidemia. Por otra parte, si es imposible lograr condiciones uniformes, puede ser posible todavía lograr que lo sean dentro de cada uno de los bloques. Por ejemplo, en experimentos de cultivos, si no se puede efectuar un experimento completo o si no se puede cosechar en un día, es conveniente hacer bloques completos en un día; en experimentos de laboratorio donde es nece-

sario emplear varios técnicos, es aconsejable que cada uno de ellos se encargue de uno o más conjuntos completos de tratamientos.

Se debe contar con medidas adecuadas no sesgadas de los efectos de los diversos tratamientos o de las diferencias entre ellos. A menudo, las mediciones requeridas son obvias y fáciles de realizar; en otros casos, se necesita considerable investigación para estar seguros de lograr una medida confiable para expresar los efectos de los tratamientos. Así, los datos de ingeniería sanitaria de la tabla 2.3 se obtuvieron luego de ensayos repetidos y de revisiones de la técnica porque se había decidido que no se utilizaría técnica alguna hasta no tener un coeficiente de variación del orden del 5 por ciento. Siempre debe tenerse cuidado para evitar errores grandes que se presenten durante la experimentación; la adecuada supervisión de los asistentes y un riguroso escrutinio de los datos servirán de mucho para evitarlos.

La técnica defectuosa puede aumentar el error experimental de dos maneras. Puede introducir fluctuaciones adicionales de una naturaleza más o menos aleatoria y, posiblemente, sujetas a las leyes del azar. Tales fluctuaciones, si son grandes, deberán revelarse por sí mismas en la estimación del error experimental, posiblemente mediante la observación del coeficiente de variación. Si los experimentadores encuentran que sus estimaciones del error experimental son consistentemente más altas que las de otros investigadores en el mismo campo, deben examinar cuidadosamente sus técnicas para determinar el origen de los errores. La otra manera como la técnica defectuosa puede aumentar el error experimental es mediante errores no aleatorios. Estos no están sujetos a las leyes del azar y no siempre puede detectarse mediante la observación de las medidas individuales. Se dispone de pruebas estadísticas para detectarlos, pero no se estudiarán en este texto. La técnica defectuosa también puede producir medidas consistentemente sesgadas, lo que no afecta el error experimental en las diferencias entre las medias de tratamientos, pero sí los valores de las medias de tratamientos. El error experimental no puede detectar sesgos. El error experimental estima la precisión o la repetibilidad de las medidas, pero no estima la exactitud.

Un punto que algunas veces se pasa por alto es que una medida puede estar sujeta a dos principales fuentes de variación, una de las cuales es considerablemente mayor que la otra. Como ilustración, considérese la cantidad de proteína por acre producida por un cultivo forrajero. La cantidad es una función del rendimiento por acre y del porcentaje de proteína. La variación entre medidas del rendimiento del forraje es mucho mayor que la variación entre las determinaciones del porcentaje de proteína. Como consecuencia, hay que esforzarse más en reducir la parte del error experimental asociada con el rendimiento del forraje, que a la asociada con el porcentaje de proteína. En general, cuando están presentes varias fuentes de variación es más útil tratar de controlar la fuente más grande.

## 6.12 Aleatorización

La función de la aleatorización consiste en asegurarse que obtengamos un estimativo válido o insesgado del error experimental, de las medias de los tratamientos y de las diferencias entre las mismas. La aleatorización es una de las pocas características del diseño experimental moderno que es realmente nueva; la idea se debe a R.A. Fisher. La aleatorización generalmente supone el empleo de un dispositivo de azar, tal como el lanzamiento de una

moneda o el uso de tablas de números aleatorios. Aleatoriedad y azar no son equivalentes; la aleatorización no puede superar a la técnica deficiente.

Para evitar el sesgo en las comparaciones entre medias de tratamientos, es necesario disponer de alguna manera de asegurar que un tratamiento particular no resulte favorecido en forma consistente en repeticiones sucesivas por alguna fuente externa de variación conocida o desconocida. O sea que cada tratamiento debe tener igual oportunidad de ser asignado a una unidad experimental, sea favorable o desfavorable. La aleatoriedad ofrece el procedimiento de igual oportunidad. Cochran y Cox (6.2) dicen que "la aleatorización es de algún modo análoga a un seguro, en cuanto es una precaución contra percances que pueden ocurrir o no, y que pueden ser graves o no si ocurren".

Los diseños sistemáticos, en los cuales los tratamientos se aplican a las unidades experimentales de una manera no aleatoria y seleccionada, a menudo producen ya sea una subestimación o bien una sobreestimación del error experimental. También pueden dar lugar a desigualdades de precisión en las diversas comparaciones entre medias de tratamiento. Esto es especialmente evidente en muchos experimentos sobre el terreno. Numerosos estudios han demostrado que las parcelas adyacentes tienden a presentar una productividad más semejante que aquellas que están más separadas entre sí. Se dice que tales parcelas dan componentes o residuos de error correlacionados. Como resultado de este hecho, si los tratamientos se disponen en el mismo orden sistemático en cada repetición, entonces puede haber diferencias considerables en la precisión de las comparaciones en que entran diferentes tratamientos. La precisión de comparaciones entre tratamientos que se encuentran físicamente más cerca, es mayor que la de los que están más alejados. La aleatorización tiende a eliminar la correlación entre los errores y a hacer que sean válidas las pruebas de significancia.

### 6.13 Inferencia estadística

Como hemos visto, el objetivo de los experimentos es determinar si hay diferencias reales entre nuestras medias de tratamiento y estimar la magnitud de tales diferencias si existen. Una inferencia estadística respecto a tales diferencias supone la asignación de una medida de la probabilidad a la inferencia. Pero ello, es necesario que la aleatorización y la repetición se introduzcan en el experimento de una manera apropiada

Las repeticiones nos aseguran la manera de calcular el error experimental.

La aleatorización nos asegura una medida válida del error experimental.

La elección entre un experimento con una aleatorización adecuada y uno sistemático con aparente mayor precisión pero que no se puede medir, es como la elección entre una vía de longitud y condiciones conocidas y otra de longitud y condiciones desconocidas, de la cual sólo se sabe que es más corta. Puede ser más satisfactorio saber qué tan lejos hay que ir y cómo será la ida, que partir por una vía de condiciones y longitud desconocidas, con la sola seguridad de que es más corta. Hasta cuando no se disponga de estudios más detallados de los diseños sistemáticos, parece aconsejable evitar utilizarlos.

## Referencias

- 6.1. Brownlee, K. A.: "The principles of experimental design," *Ind. Qual. Cont.*, 13:12-20 (1957).
- 6.2. Cochran, William G., y Gertrude M. Cox: *Experimental designs*, 2a. ed., Wiley, Nueva York, 1957.
- 6.3. Cox, D. R.: *Planning of experiments*, Wiley, Nueva York, 1958.
- 6.4. Federer, W. T.: *Experimental design*, Macmillan, Nueva York, 1955.
- 6.5. Fisher, R. A.: *The design of experiments*, 4a. ed., Oliver y Boyd, Edinburgo, 1947.
- 6.6. Greenberg, B. G.: "Why randomize?" *Biom.*, 7: 309-322 (1951).
- 6.7. John, Peter W. M.: *Statistical design and analysis of experiments*, Macmillan, Nueva York, 1971.
- 6.8. Kempthorne, O.: *The design and analysis of experiments*, Wiley, Nueva York, 1952.
- 6.9. Youden, W. J.: "Comparative tests in a single laboratory," *ASTM Bull.*, 116:48-51 (1950).

---

## CAPITULO SIETE

---

### ANALISIS DE LA VARIANZA I: CLASIFICACION DE UNA VIA

#### 7.1 Introducción

El análisis de la varianza fue ideado por Sir Ronald A. Fisher y es esencialmente un procedimiento aritmético que descompone una suma total de cuadrados en componentes asociados con fuentes de variación reconocida. Se ha usado con provecho todos los campos de investigación en los que los datos se miden cuantitativamente.

En el capítulo 7, consideramos el análisis de la varianza en donde el único criterio de clasificación de los datos es el tratamiento. Estudiamos el modelo lineal aditivo y las componentes de la varianza, los supuestos en que se basa el análisis de la varianza y las pruebas de significancia, el cuadrado medio residual o error experimental y el error de muestreo.

En los capítulos, 9, 15, 16 y 18 se estudian otros diseños y aspectos del análisis de la varianza.

#### 7.2 El diseño completamente aleatorio

Este diseño es útil cuando las unidades experimentales son esencialmente homogéneas, es decir, cuando la variación entre ellas es pequeña y agruparlas en bloques sería poco más que un proceso aleatorio. Este es el caso en muchos tipos de experimentos de laboratorio, en los que una cantidad de material está completamente mezclada y luego se divide en porciones pequeñas para formar las unidades experimentales a las cuales se asignan los tratamientos en forma aleatoria, o en experimentos con animales y plantas con condiciones ambientales muy parecidas.

La aleatorización, el proceso que hace aplicables las leyes del azar, se logra asignando tratamientos a las unidades experimentales de manera completamente aleatoria. No se imponen restricciones a la aleatorización como cuando se necesita que un bloque contenga todos los tratamientos. La elección del número de observaciones que han de hacerse

sobre los diversos tratamientos, no se considera una restricción a la aleatorización. Cada unidad experimental tiene la misma probabilidad de recibir un tratamiento, esto es, que si hay  $n$  unidades experimentales entonces cualquiera de los  $n$  tratamientos, no todos diferentes, claro está, tiene la misma probabilidad de caer en cualquier unidad experimental.

La aleatorización se lleva a cabo mediante el uso de una tabla de números aleatorios. Supóngase que 15 unidades experimentales van a recibir cinco repeticiones de cada uno de tres tratamientos. Asignese los números 1 a 15 a las unidades experimentales, en una forma conveniente, por ejemplo, en forma consecutiva. Localícese un punto de partida en una tabla de números aleatorios, por ejemplo, la fila 10 y columna 20 de la tabla A.1 y selecciónense 15 números de 3 dígitos. Al leer verticalmente obtenemos:

118	701	789	965	688	638	901	841	396	802	687	938	377	392	848
1	8	9	15	7	5	13	11	4	10	6	14	2	3	12

Entonces se asignan rangos a los números; así, 118 es el menor, le corresponde el rango 1 y 965 el mayor, le corresponde el rango 15. Se considera que estos rangos son una permutación aleatoria de los números de 1 a 15 y que los 5 primeros son los números de unidades experimentales correspondientes al tratamiento uno. Por tanto, las unidades 1, 8, 9, 15 y 7 reciben el tratamiento 1, y así sucesivamente.

El procedimiento también es aplicable cuando los tratamientos replican desigual número de veces, por ejemplo, 6, 6 y 3. Se usan los números de 3 dígitos ya que es menos probable que incluya empates como puede ocurrir con números de 2 dígitos. En todo caso, los empates se pueden deshacer mediante el uso de más números.

Las tablas de números aleatorios pueden usarse de otra manera. Por ejemplo, los números de 2 dígitos menores de 90 pueden dividirse por 15 y se registra el residuo, lo que da los números 00, 01, ..., 14 con igual frecuencia. Se descartan los duplicados y se obtienen otros para reemplazarlos. Los números 90, 91, ..., 99 no se usan, ya que harían que 00, 01, ..., 09, se presentaran con mayor frecuencia que 10, 11, ..., 14.

Si, durante el curso del experimento, todas las unidades experimentales se han de tratar de manera análoga, por ejemplo, la siembra en parcelas de terreno, debe hacerse en orden aleatorio si es posible que el orden afecte los resultados, como cuando una técnica mejora debido a la práctica.

El análisis de un diseño completamente aleatorio también es aplicable a datos en los que el "tratamiento" implica simplemente una variable de clasificación y cuando hasta puede ser necesario suponer la aleatoriedad. Por ejemplo, se puede medir el peso de los adultos de ciertas especies de peces obtenidos en varios lagos (tratamientos), y, se desea saber si los pesos de peces adultos cambian de un lago a otro.

**Ventajas** El diseño completamente aleatorio es flexible en cuanto a que el número de tratamientos y de repeticiones sólo está limitado por el número de unidades experimentales disponibles. El número de repeticiones puede variar de un tratamiento a otro, aunque generalmente lo ideal sería tener un número igual por tratamiento. El análisis estadístico es simple aún en el caso en que el número de repeticiones difiera con el tratamiento y si los diversos tratamientos están sujetos a varianzas desiguales, lo cual se conoce

como la falta de homogeneidad del error experimental. Sin embargo, las pruebas de significancia y la construcción del intervalo de confianza requieren atención especial cuando hay heterogeneidad de la varianza. La sencillez del análisis no se pierde si algunas unidades experimentales o tratamientos enteros faltan o se descartan.

La pérdida de información debida a estos faltantes es pequeña en relación con las pérdidas sufridas con otros diseños. El número de grados de libertad para estimar el error experimental es máximo; esto mejora la precisión del experimento y es importante con experimentos pequeños, o sea, en aquellos en los que los grados de libertad para el error experimental son menos de 20 (ver sec. 6.8).

**Desventajas** La principal objeción al diseño completamente aleatorio es su frecuente ineeficiencia. Como la aleatorización no tiene restricciones, el error experimental incluye toda la variación entre las unidades experimentales excepto la debida a los tratamientos. En muchas situaciones es posible agrupar las unidades experimentales de modo que la variación entre unidades dentro de los grupos sea menor que la variación entre las unidades de diferentes grupos. Ciertos diseños sacan ventaja de tal agrupamiento, excluyen la variación del error experimental entre grupos y aumentan la precisión del experimento. Algunos de esos diseños se estudian en los caps. 9, 15 y 16.

### 7.3 Datos con un sólo criterio de clasificación: El análisis de la varianza para cualquier número de grupos con igual número de repeticiones

En la tabla 7.1 se da el contenido de nitrógeno, en miligramos, de plantas de trébol rojo inoculadas con cultivos de *Rhizobium trifolii* más un compuesto de cinco cepas de *Rhizobium meliloti*, tal como lo reporta Erdman (7.9). Cada uno de los cinco cultivos de trébol *R. trifolii* se sometió a prueba individualmente con un compuesto de cinco cepas de alfalfa, *R. meliloti*, y un compuesto de trébol rojo también se sometió a prueba en el compuesto de las cepas de alfalfa, lo que da seis tratamientos en total. El experimento se realizó en un invernadero empleando un diseño completamente aleatorio con cinco materas por tratamiento.

**Cálculos** Disponer los datos como en la tabla 7.1. Sea  $Y_{ij}$  la observación  $j$ -ésima bajo el tratamiento  $i$ -ésimo,  $i = 1, 2, \dots, t$  y  $j = 1, 2, \dots, r$ . Los totales de los tratamientos requieren un subíndice y el total para el tratamiento  $i$ -ésimo puede denotarse  $Y_{i\cdot}$ , donde el punto indica que todas las observaciones para el tratamiento  $i$ -ésimo se han sumado para obtener este total. Las letras  $t$  y  $r$  se usan para designar el número de tratamientos y el número de repeticiones de cada tratamiento; aquí  $t = 6$  y  $r = 5$ .

Para cada tratamiento obtener simultáneamente  $Y_{i\cdot}$  y  $\sum_j Y_{ij}^2$  con una máquina calculadora, tal como se presenta en las líneas 1 y 2 de los cálculos de la tabla 7.1. Estos valores se totalizan luego; por lo tanto:

$$\sum_i Y_{i\cdot} = Y_{..} \quad \text{y} \quad \sum_i \left( \sum_j Y_{ij}^2 \right) = \sum_{i,j} Y_{ij}^2 .$$

**Tabla 7.1 Contenido de nitrógeno de plantas de trébol rojo inoculadas con combinaciones de cultivos de cepas de *Rhizobium trifolii* y cepas de *Rhizobium meliloti*, mg.**

Cepa de *R. trifolii*

Cálculo	3DOK1	3DOK5	3DOK4	3DOK7	3DOK13	Com- puesto	Total
	19.4	17.7	17.0	20.7	14.3	17.3	
	32.6	24.8	19.4	21.0	14.4	19.4	
	27.0	27.9	9.1	20.5	11.8	19.1	
	32.1	25.2	11.9	18.8	11.6	16.9	
	33.0	24.3	15.8	18.6	14.2	20.8	
$\sum_i Y_{ij} = Y_{i\cdot}$	144.1	119.9	73.2	99.6	66.3	93.5	$596.6 = Y_{..}$
$\sum_j Y_{ij}^2$	4,287.53	2,932.27	1,139.42	1,989.14	887.29	1,758.71	12,994.36
$(Y_{i\cdot})^2/r$	4,152.96	2,875.20	1,071.65	1,984.03	879.14	1,748.45	12,711.43
$\sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$	134.57	57.07	67.77	5.11	8.15	10.26	282.93
$\bar{Y}_{i\cdot}$	28.8	24.0	14.6	19.9	13.3	18.7	

En la línea 3, cada total de tratamiento se eleva al cuadrado y se divide por  $r = 5$ , el número de repeticiones por tratamiento.

Obtener el factor de corrección FC de la ec. (7.1) y la suma de cuadrados total (ajustada para la media) de la ec. (7.2). El factor de corrección es el cuadrado de la suma de todas las observaciones dividido por su número. Para estos datos

$$C = \frac{Y_{..}^2}{rt} = \frac{\left(\sum_{i,j} Y_{ij}\right)^2}{rt} \quad (7.1)$$

$$= \frac{(596.6)^2}{(5)(6)} = 11,864.38$$

$$\begin{aligned} \text{Total SC} &= \sum_{i,j} Y_{ij}^2 - C \\ &= 12,994.36 - 11,864.38 = 1,129.98 \end{aligned} \quad (7.2)$$

La suma de cuadrados atribuible a la variable de clasificación, esto es, a los tratamientos, que suele llamarse *suma de cuadrados entre grupos* o *suma de cuadrados de tratamientos* se calcula así:

$$\begin{aligned} \text{SC tratamientos} &= \frac{Y_{1\cdot}^2 + \cdots + Y_{r\cdot}^2}{r} - C \\ &= \frac{(144.1)^2 + \cdots + (93.5)^2}{5} - 11,864.38 = 847.05 \end{aligned} \quad (7.3)$$

La suma de cuadrados entre individuos tratados en forma similar también se llama *suma de cuadrados dentro de grupos*, *suma de cuadrados residual*, *suma de cuadrados de error* o *discrepancia*, y generalmente se obtiene restando del total la suma de cuadrados de tratamiento, como en la ec. (7.4). Esto es posible por la propiedad de aditividad de las sumas de cuadrados.

$$\begin{aligned} \text{SC Error} &= \text{SC Total} - \text{SC Tratamientos} \\ &= 1,129.98 - 847.05 = 282.93 \end{aligned} \quad (7.4)$$

La suma de cuadrados de error también puede encontrarse combinando las sumas de cuadrados dentro de tratamientos como se indica en la ec. (7.5). Estas sumas de cuadrados se dan en la penúltima línea de la tabla 7.1. Cada componente tiene  $r - 1 = 4$  grados de libertad. La suma de las componentes es 282.93. La naturaleza aditiva de las sumas de cuadrados queda demostrada. Esta es una excelente verificación del cálculo. Además, proporciona información relacionada con la homogeneidad de la varianza del error

$$\begin{aligned} \text{SC Error} &= \sum_i \left( \sum_j Y_{ij}^2 - \frac{Y_i^2}{r} \right) \\ &= \left( 4,287.53 - \frac{144.1^2}{5} \right) + \cdots + \left( 1,758.71 - \frac{93.5^2}{5} \right) = 282.93 \end{aligned} \quad (7.5)$$

Los resultados numéricos de un análisis de la varianza generalmente se presentan en una tabla de análisis de la varianza tal como la tabla 7.2, con símbolos, o la tabla 7.3, ejemplo, para los datos que se acaban de exponer.

El cuadrado medio del error se denota  $s^2$  y frecuentemente se denomina término generalizado del error ya que es un promedio de las componentes aportadas por las diferentes poblaciones o tratamientos. Es un estimativo de una  $\sigma^2$  común, la variación entre las observaciones tratadas en forma análoga. Que exista una  $\sigma^2$  es un supuesto, y  $s^2$  es una estimación válida de  $\sigma^2$  sólo si este supuesto es verdadero. Las componentes individuales se basan en sólo unos pocos grados de libertad, así que pueden variar ampliamente en torno a  $\sigma^2$  y, por lo tanto, no son tan buenas estimaciones como la estimación combinada. El principio es el mismo que para las medias: una media de 24 observaciones es mejor estimación de  $\mu$  que una de 4 observaciones porque la primera tiene menor varianza. Análogamente, una varianza muestral basada en 24 observaciones es mejor estimación de  $\sigma^2$  que una basada en 4 observaciones ya que la primera tiene menos varianza. La validez del supuesto de que cada una de las componentes del error es una estimación de la misma  $\sigma^2$ , se puede comprobar con la prueba de homogeneidad,  $\chi^2$ , que se estudia en la sec. 17.3.

El cuadrado medio de los tratamientos es una estimación independiente de  $\sigma^2$  cuando la hipótesis nula es verdadera. Las varianzas entre medias estiman a  $\sigma^2/r$ . Así, el  $r$  en la fórmula de definición de la suma de cuadrados nos asegura que el cuadrado medio de tratamientos estima  $\sigma^2$  en vez de  $\sigma^2/r$ . Razonamiento parecido se aplica a la fórmula de operación donde se usan totales de los tratamientos. La aplicación de la sec. 5.10-muestra que las varianzas entre totales (ver ejercicio 5.10.3) estiman  $r\sigma^2$ ; ahora bien, es necesario un divisor  $r$  si este cálculo ha de dar una estimación de  $\sigma^2$  en vez de  $r\sigma^2$ . Como  $F$  se define

**Tabla 7.2 Análisis de la varianza: clasificación de una vía, con igual número de repeticiones  
(Con símbolos)**

Fuente de variación	gl	Sumas de cuadrados		Cuadrados medios†	F
		Definición	Operación		
Tratamiento	$t - 1$	$r \sum_i (\bar{Y}_i - \bar{Y}_{..})^2$	$\sum_i \frac{\bar{Y}_i^2}{r} - \frac{\bar{Y}_{..}^2}{rt}$	✓	✓
Error	$t(r - 1)$	$\sum_{i,j} (Y_{ij} - \bar{Y}_{..})^2$	Por sustracción	✓	
Total	$rt - 1$	$\sum_{i,j} (Y_{ij} - \bar{Y}_{..})^2$	$\sum_{i,j} Y_{ij}^2 - \frac{\bar{Y}_{..}^2}{rt}$		

†Un cuadrado medio es una suma de cuadrados dividida por los grados de libertad respectivos.

como la razón de dos estimaciones independientes de la misma  $\sigma^2$ , entonces es necesaria la estimación del cuadrado medio de los tratamientos,  $\sigma^2$ . Decimos que ambos cuadrados medios se calcularon por cada observación.

El valor  $F$  se obtiene dividiendo el cuadrado medio de los tratamientos por el cuadrado medio del error, esto es,  $F = 169.4/11.79 = 14.37^{**}$ . Por lo tanto, el valor de  $F$  da el cuadrado medio de tratamientos como un múltiplo del cuadrado medio del error. Estos cuadrados medios son comparables ya que cada uno estima la variación entre observaciones individuales. El  $F$  calculado se compara con el  $F$  tabulado para 5 y 24 grados de libertad para decidir si aceptamos o no la hipótesis nula de que no hay diferencia entre las medias poblacionales o aceptar la hipótesis alterna de que hay diferencia. Los valores tabulados de  $F$  para 5 y 24 grados de libertad son 2.62 y 3.90 a los niveles de probabilidad del 0.05 y 0.01, respectivamente. Puesto que el  $F$  calculado excede el 1 por ciento del  $F$  tabulado, concluimos que el experimento comprueba que hay diferencia real entre las medias de los tratamientos.

Para usar las tablas de  $F$  para el análisis de la varianza, se buscan a lo largo de la parte superior de la tabla los grados de libertad del numerador, esto es, los grados de libertad del tratamiento, y los grados de libertad del error se buscan en el margen. Esto se debe a que el conjunto de hipótesis alternativas sólo admite que existen diferencias entre tratamientos y, por tanto, aumenta el estimativo de la varianza basado en medias o tota-

**Tabla 7.3 Análisis de la varianza de los datos de la tabla 7.1**

Fuente de variación	gl	Suma de cuadrados	Cuadrado medio	F
Entre los cultivos <sup>2</sup>	5	847.05	169.41	14.37**
Dentro de los cultivos	24	282.93	11.79	
Total	29	1,129.98		

les de tratamientos, así que sólo los valores grandes del criterio de prueba se juzgan significantes. Si el cuadrado medio de los tratamientos es menor que el error, el resultado se declara no significante por pequeña que sea la razón. Una  $F$  significante implica que las pruebas son suficientemente sólidas como para indicar que ningún tratamiento pertenece a poblaciones con una  $\mu$  común. Sin embargo, no indica cuáles diferencias se pueden considerar estadísticamente significantes.

Obsérvese que tanto los grados de libertad como las sumas de cuadrados dentro de la tabla dan la suma de los valores correspondientes en la línea del total. Los cuadrados medios no son aditivos. La propiedad de aditividad de las sumas de cuadrados es característica de experimentos bien planeados y ejecutados. Permite ciertas operaciones abreviadas en el análisis de la varianza, como el cálculo de la suma de cuadrados del error restando de la suma de cuadrados total la suma de cuadrados de tratamientos como se ve en la tabla 7.3. Experimentos que no se planean y ejecutan de modo que posean esa propiedad, generalmente suponen muchos más cálculos y tienen menor precisión por observación.

El error estándar de una media de tratamiento y el de una diferencia entre medias de tratamientos se calculan con las ecs. (7.6) y (7.7), respectivamente. Los valores numéricos corresponden a los datos de la tabla 7.1.

$$s_y = \sqrt{\frac{s^2}{r}} = \sqrt{\frac{11.79}{5}} = 1.54 \text{ mg} \quad (7.6)$$

$$s_{\bar{Y}_i - \bar{Y}_{i'}} = \sqrt{\frac{2s^2}{r}} = \sqrt{\frac{2(11.79)}{5}} = 2.17 \text{ mg} \quad i \neq i' \quad (7.7)$$

Estos estadígrafos son útiles para comparar diferencias entre medias de tratamientos, como se verá en el cap. 8, y para el cálculo de intervalos de confianza para medias de tratamientos y diferencias de medias de tratamientos. Otras aplicaciones se ven en la sec. 3.11 y el cap. 5. El coeficiente de variación se calcula por

$$CV = \frac{\sqrt{s^2}}{\bar{Y}} 100 = \frac{\sqrt{11.79}}{19.89} 100 = 17.3 \text{ por ciento} \quad (7.8)$$

Se mostró en la sec. 5.5 que el análisis de la varianza podía usarse en vez de la prueba  $t$  para comparar dos tratamientos en los que el diseño fuera completamente aleatorio.

La prueba de  $F$  de una cola con 1 y  $n$  grados de libertad corresponde a la  $t$  de dos colas con  $n$  grados de libertad. Esta prueba de  $t$  no especifica la dirección de la diferencia entre dos medias de tratamientos para la hipótesis alterna; así se parece a la prueba de  $F$  de una cola que especifica cuál cuadrado medio ha de ser mayor como resultado de diferencias de dirección no especificada entre tratamientos. Se puede demostrar que algebraicamente estas pruebas son equivalentes; en particular  $t^2 = F$ . La relación se muestra gráficamente en la fig. 7.1. Valores pequeños de  $t$ , cuando se elevan al cuadrado, se convierten en valores pequeños de  $F$ , que son positivos. Valores grandes de  $t$ , elevados al cuadrado, se convierten en valores grandes de  $F$ .

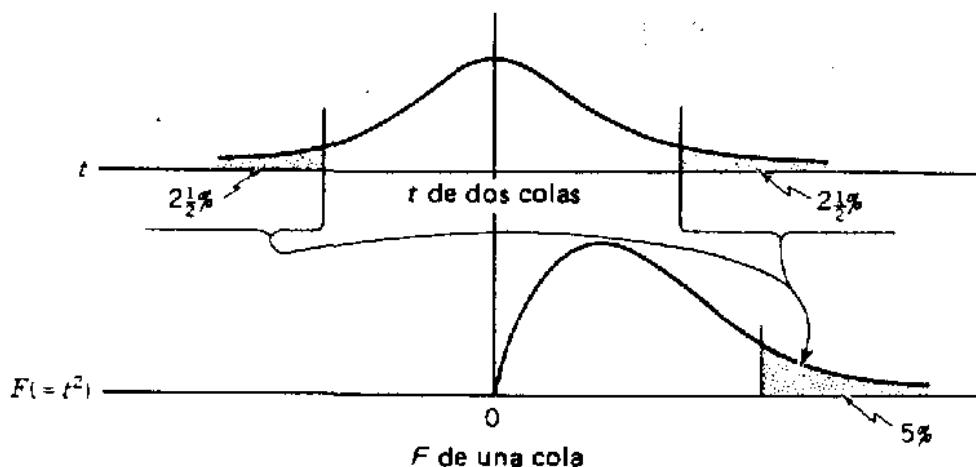


Figura 7.1 Relación entre el  $t$  de dos colas y el  $F$  de una cola (las curvas son aproximadas solamente).

**Ejercicio 7.3.1** F.R. Urey, del Departamento de Zoología, de la Universidad de Wisconsin, llevó a cabo un experimento de estrógeno de varias soluciones que habían estado sujetas a una técnica de inactivación in vitro. El peso del útero de ratones hembras se usó como medida de la actividad estrogénica. En la tabla siguiente, se presentan los pesos en miligramos de cuatro úteros de cuatro ratones hembras para cada una de las soluciones, una de control y seis diferentes.

Control	1	2	3	4	5	6
89.8	84.4	64.4	75.2	88.4	56.4	65.6
93.8	116.0	79.8	62.4	90.2	83.2	79.4
88.4	84.0	88.0	62.4	73.2	90.4	65.6
112.6	68.6	69.4	73.8	87.8	85.6	70.2

Calcular las medias y efectuar el análisis de la varianza de estos datos. Calcular las sumas de cuadrados del error en forma directa y demostrar la aditividad de las sumas de cuadrados. Calcular, además, el coeficiente de variación.

**Ejercicio 7.3.2** Calcular la suma de cuadrados entre los tratamientos de 1 a 6, esto es, sin tener en cuenta el control. Esta tiene 5 grados de libertad. Calcular la suma de cuadrados para comparar el control con la media de todas las demás observaciones. (*Indicación:* leer la sec. 5.5). Esta tiene un grado de libertad. Obsérvese que la suma de estas dos sumas de cuadrados es la suma de cuadrados de tratamientos en el análisis de la varianza. Se dice que tales comparaciones son ortogonales; la ortogonalidad se estudia en el cap. 8.

**Ejercicio 7.3.3** Individuos de la población general de la Prisión Central, Raleigh, Carolina del Norte, se sometieron voluntariamente a un experimento para estudiar la experiencia del "aislamiento". (En este establecimiento están los reos de crímenes mayores.) El experimento fue dirigido por Jordan (7.12). En el tratamiento experimental, los presidiarios fueron expuestos a una combinación de restricciones sensorial y de sugestión. El propósito era reducir las puntuaciones en la escala T de desviación psicótica (DP) en la prueba MMPI (Minnesota Multiphasic Personality Inventory).

Brevemente, los tratamientos consistieron en:

1. Cuatro horas de restricción sensorial más 15 minutos de una grabación "terapéutica", que informa que hay ayuda profesional.
2. Cuatro horas de restricción sensorial más 15 minutos de una grabación "emocionalmente neutra" sobre adiestramiento de perros de caza.
3. Cuatro horas de restricción sensorial pero sin mensaje.

El MMPI se administró antes y después del experimento. Los valores de las puntuaciones de T DP se dan en seguida para 14 individuos en cada tratamiento.

Calcular las medias de los tratamientos y el análisis de la varianza de las puntuaciones antes de la prueba. Probar la hipótesis nula de que no hay diferencias entre grupos antes de efectuar el experimento. Calcular  $s_y$ ,  $s_{y_1-y_2}$  y el coeficiente de variación.

Tratamiento					
1		2		3	
Prueba					
Pre	Post	Pre	Post	Pre	Post
67	74	88	79	86	90
86	50	79	81	53	53
64	64	67	83	81	102
69	76	83	74	69	67
67	64	79	76	81	76
79	81	76	69	76	81
67	74	71	71	74	69
67	50	67	75	60	60
69	60	69	64	67	69
57	57	67	64	86	83
76	62	67	64	86	107
90	76	74	71	74	71
71	71	81	74	71	71
93	76	81	64	71	81

Ejercicio 7.3.4 Calcular las medias de los tratamientos y el análisis de la varianza de las puntuaciones de la prueba posterior (post test) dadas en el ejercicio 7.3.3. Probar la hipótesis nula de que no hay diferencias entre los grupos de tratamientos terminado el experimento. Calcular  $s_y$ ,  $s_{y_1-y_2}$  y el coeficiente de variación. ¿Se puede pensar que la prueba F calculada en el análisis de la varianza cumple la exigencia de la prueba que supone el investigador. Explicar.

#### 7.4 Datos con un solo criterio de clasificación: El análisis de la varianza para cualquier número de grupos con número desigual de repeticiones

Cuando el número de observaciones por tratamiento varía, tenemos un caso general para el cual la sec. 7.3 proporciona un ejemplo especial.

El análisis de la varianza se ilustrará ahora con los datos de la tabla 7.4. Las observaciones originales se tomaron en un experimento de invernadero en Ithaca, Nueva York, y se trataba de determinar si hay o no diferencias genéticas entre las plantas, bien sea entre o dentro de las localidades, donde localidad se refiere al origen de la planta; los

**Tabla 7.4 Longitud de la hoja de *Sedum oxypetalum* al tiempo de la floración (suma de tres hojas) provenientes de seis localidades en el cinturón volcánico trans-mexicano, en mm**

Localidad	$r_i$	$\sum_j Y_{ij}$	$\bar{Y}_i$	$\sum_j Y_{ij}^2$	$\bar{Y}_i^2/r_i$	$\sum_j (Y_{ij} - \bar{Y}_i)^2$
H	1	147	147.00	21,609	21,609.00	
LA	1	70	70.00	4,900	4,900.00	
R	6	634	105.67	71,740	66,992.67	4,747.33
SN	1	75	75.00	5,625	5,625.00	
Tep	3	347	115.67	40,357	40,136.33	220.67
Tis	2	170	85.00	14,500	14,450.00	50.00
Totales	14	1,443	( $\bar{Y}_{..} = 103.07$ )	158,731	153,713.00	5,018.00

Fuente: Datos sin publicar usados con permiso de R.T. Clausen, Universidad de Cornell, Ithaca, Nueva York.

datos en la Tabla 7.4 son apropiados para hacer comparaciones entre localidades (más generalmente, tratamientos) y son para sólo un carácter de los muchos que fueron comprobados.

Calcular

$$\sum_j Y_{ij} = Y_i \quad \text{y} \quad \sum_j Y_{ij}^2,$$

o sea, la suma y la suma de cuadrados de las observaciones para cada tratamiento; éstas se presentan también en la tabla 7.4. La ec. (7.9) es el único procedimiento nuevo de cálculo; la ec. (7.3) resulta de esta ecuación cuando todos los  $r_i$  son iguales.

$$FC = \frac{\bar{Y}_{..}^2}{\sum r_i} = \frac{(1,443)^2}{14} = 148,732.07$$

$$SC \text{ Total} = \sum_{i,j} Y_{ij}^2 - C = 158,731.00 - 148,732.07 = 9,998.93$$

$$\begin{aligned} SC \text{ Tratamientos} &= \sum_i \frac{Y_i^2}{r_i} - C = \frac{Y_1^2}{r_1} + \cdots + \frac{Y_r^2}{r_r} - C \\ &= 153,713.00 - 148,732.07 = 4,980.93 \end{aligned} \tag{7.9}$$

$$SC \text{ Error} = SC \text{ Total} - SC \text{ Tratamientos} = 5,018.00$$

Obsérvese que  $\bar{Y}_{..} = Y_{..}/\sum r_i = \sum r_i \bar{Y}_i / \sum r_i$  es una media ponderada de las medias de los tratamientos.

También, la ec. (7.9) tiene la fórmula de definición correspondiente dada por

**Tabla 7.5 Análisis de la varianza de los datos resumidos en la tabla 7.4**

Fuente de variación	gl	SC	CM	F
Entre localidades	5	4,980.93	996.19	1.59
Dentro de localidades	8	5,018.00	627.25	
Total	13	9,998.93		

$s_{Y_i - Y_j} = \sqrt{627.25 \left( \frac{1}{r_i} + \frac{1}{r_j} \right)}$  mm donde  $r_i$  y  $r_j$  son el número de observaciones en las medias de las comparaciones deseadas.

$$\sum \frac{Y_i^2}{r_i} - C = \sum r_i (\bar{Y}_i - \bar{Y}_{..})^2 \quad (7.10)$$

Esto muestra que estamos calculando realmente una suma ponderada de cuadrados de las desviaciones de las medias de los tratamientos de la media total. Las ponderaciones son proporcionales a la información en toda media, como lo son los inversos de las varianzas,  $\sigma^2/r_i$ , excepto para la  $\sigma^2$  común.

Finalmente, la suma de cuadrados del error puede calcularse combinando las sumas de cuadrados dentro de los tratamientos dados en la última columna de la tabla 7.4. Es de interés observar que tres localidades no contribuyen en nada a la varianza del error, ya que cada una proporciona sólo una observación.

El análisis de la varianza está dado en la tabla 7.5.  $F$  no es significante; el  $F$  tabulado,  $F(0.05) = 3.69$  para 5 y 8 grados de libertad. Lo comprobado no está en favor de la diferencia entre localidades.

**Ejercicio 7.4.1** Wexelsen (7.18) estudió los efectos de la endogamia sobre el peso de la planta en el trébol rojo. Se dan a continuación los pesos medios por planta en gramos de líneas ( $F_1$ ), no endogámicas y tres grupos de familias endogámicas en orden creciente de endogamia.

$F_1$ : 254, 263, 266, 337, 274, 289, 244, 265

Ligeramente endogámicas: 236, 191, 209, 212, 224

$F_2$ : 253, 192, 141, 160, 229, 221, 150, 215, 232, 234, 193, 188

$F_3$ : 173, 164, 183, 138, 146, 125, 178, 199, 170, 177, 172, 198

Calcular el análisis de la varianza para estos datos. Obtener la suma de cuadrados del error directamente y demostrar la actividad de las sumas de cuadrados. Calcular el coeficiente de variación.

**Ejercicio 7.4.2** Los datos del ejercicio 7.3.3 son incompletos. También se obtuvieron las siguientes observaciones.

- a) Efectuar el análisis de la varianza de las puntuaciones de la prueba previa usando todos los datos. Probar la hipótesis nula de que no hay diferencias entre las medias poblacionales. Presentar las medias,  $s$  y C.V.

## Tratamiento

		1		2		3	
		Prueba					
Pre	Post	Pre	Post	Pre	Post		
81	83	53	81				
86	88	81	71			No más	
69	60	83	79				
83	74						

- b) Efectuar el análisis de la varianza de las puntuaciones de la prueba posterior usando todos los datos. Probar la hipótesis nula de que no hay diferencias entre las medias. Calcular las medias,  $s$  y C.V.

Ejercicio 7.4.3 En un estudio experimental de los efectos del Aroclor 1254, en PCB, Sanders (7.14) incorporó la sustancia en las dietas de ratones caseros, machos, albinos de un genotipo aleatorio alimentados *ad libitum*. Las proporciones fueron: 0, 62.5, 250, 1,000 y 4,000 ppm. Luego de dos semanas, se inyectaron los ratones con Nembutal y se registraron sus tiempos de sueño. Estos tiempos constituyen una medida de la actividad enzimática microsómica hepática.

Los tiempos de sueño de los ratones sobrevivientes fueron

## Proporciones Tiempos de sueño

0	67, 69, 72, 79
62.5	96, 98, 130, 65
250	74, 24, 15, 33, 17
1000	49, 46

Calcular el análisis de la varianza y probar la hipótesis nula de que no hay diferencias entre las medias poblacionales de las cuatro raciones. Calcular las medias,  $s$  y el C.V. de las proporciones.

Ejercicio 7.4.4 Sanders (7.14) también midió el peso final en el experimento del ejercicio 7.4.3, tal como se da en la tabla.

## Proporciones Pesos finales

0	55, 47, 46, 53
62.5	47, 51, 40, 44
250	49, 44, 46, 51, 48
1000	36, 41

Repetir el ejercicio 7.4.3 con estos datos

## 7.5 El modelo lineal aditivo

Se ha definido una observación como la suma de los componentes, una media, y un elemento aleatorio, donde la media, a su turno, puede ser una suma de componentes. Otro supuesto básico en el análisis de la varianza, cuando se llevan a cabo pruebas de significancia, es que las componentes aleatorias se distribuyen independiente y normalmente en torno a la media cero y con una varianza común. Para la clasificación de una vía, el modelo lineal comienza con

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad i = 1, \dots, t \quad j = 1, \dots, r_i \quad (7.11)$$

los  $\varepsilon_{ij}$  son componentes aleatorias.

También deben hacerse supuestos respecto a los  $\tau$  para calcular el modelo experimental. Así tenemos:

*El modelo fijo o modelo I* Los  $\tau$  son fijos y

$$\sum_i \tau_i = 0$$

constituyen una población finita y son los parámetros de interés además de  $\sigma^2$ .

*El modelo aleatorio o modelo II* Los  $\tau$  son una muestra aleatoria de una población de  $\tau$  para la cual la media es cero y la varianza es  $\sigma_\tau^2$ , el parámetro de interés además de  $\sigma^2$ .

La diferencia consiste en que para el modelo fijo, una repetición del experimento producirá el mismo conjunto de  $\tau$  en el nuevo experimento; concentraremos la atención en esos  $\tau$ . Un experimento en fertilización ilustra normalmente un modelo fijo. Para un modelo aleatorio, una repetición producirá un nuevo conjunto de  $\tau$  de la misma población de  $\tau$  y, con el tiempo, la población completa de las  $\tau$  aparecerá en la posición  $i$ -ésima, nos interesaremos por la variabilidad de las  $\tau$  pues no estamos en situación de continuar con nuestro interés en un conjunto específico. Un experimento que examina el efecto genético de vacas lecheras en la producción de leche, ilustra bien el modelo aleatorio. Para el modelo fijo, extraemos inferencias acerca de los tratamientos particulares; para el modelo aleatorio, extraemos una inferencia respecto de la población de tratamientos.

La ecuación para el modelo lineal puede haberse escrito  $Y_{ij} = \mu_i + \varepsilon_{ij}$ . En este caso, es claro que se pueden estimar todas las  $\mu_i$ . Ahora bien, si escribimos  $\mu_i = \mu + \tau_i$  podemos estimar  $\mu + \tau_i$  pero no  $\mu$  y  $\tau_i$ . Esto también es claro si consideramos el muestreo de una sola población tratada. ¿Qué tanto de la respuesta promedio se ha de asignar a la población original y qué tanto de tratamiento? Necesitamos un punto de referencia, del cual no se dispone ordinariamente. Más generalmente, no podemos estimar un  $\tau_i$  o  $\sum \tau_i$  cuando tenemos  $t$  poblaciones tratadas. El presente modelo está *sobreparametrizado*.

Esta dificultad se obvia mediante la imposición de algunas condiciones a los parámetros, tal como hacer que  $\sum \tau_i = 0$ ; los economistas a menudo hacen  $\tau_i = 0$ .

**Tabla 7.6 Valores promedios de los cuadrados medios para los modelos I y II, con igual número de repeticiones**

Fuentes de variación	gl	El cuadrado medio es un estimativo de	
		Modelo I	Modelo II
Tratamientos	$t - 1$	$\sigma^2 + r \sum \tau_i^2 / (t - 1)$	$\sigma^2 + r\sigma_t^2$
Individuos dentro de tratamientos	$t(r - 1) = tr - t$	$\sigma^2$	$\sigma^2$
Total	$tr - 1$		

Las condiciones impuestas sólo con el propósito de lograr una solución se llaman *constricciones sobre la solución*, pero cuando hacen parte integral del modelo, esto es, cuando son supuestos, como aquí se les llama *restricciones sobre el modelo*.

Hacer  $\sum \tau_i = 0$  y, por tanto,  $\sum \hat{\tau}_i = 0$ , es simplemente medir los efectos de los tratamientos como desviaciones respecto de una media global. Entonces, se establece fácilmente la hipótesis nula, así  $H_0: \tau_i = 0, i = 1, \dots, t$  y la alterna  $H_1: \text{algun } \tau_i \neq 0$ . Si  $\sum \hat{\tau}_i = 0$  ha sido una condición, entonces  $H_0$  sería simplemente que todos los efectos de tratamientos son iguales. Evidentemente, tenemos el modelo fijo en mente en nuestra discusión cuando desarrollamos la aritmética del análisis de la varianza. Sin embargo, posteriormente lo aplicamos al modelo aleatorio en que tenemos que  $H_1: \sigma_\tau^2 = 0$ .

Una vez decidido el modelo particular, es posible establecer qué parámetros se están estimando cuando se calculan los diferentes cuadrados medios y el valor de  $F$ . Para los modelos I y II, si el experimento se repitiera indefinidamente y se promediaran los cuadrados medios observados, tendríamos los valores dados en la tabla 7.6 – valores esperados. Para un experimento individual, los cuadrados medios son estimativos de los correspondientes valores esperados en la tabla 7.6.

Considérese que los datos de *Rhizobium* de la tabla 7.1 son los datos a los cuales se aplica el modelo I. Entonces, por las tablas 7.3 y 7.6, es claro que para estimar  $\sum \tau_i^2 / (t - 1)$ , necesitamos restar el cuadrado medio dentro de cultivos, el cuadrado medio entre cultivos y luego dividir por  $r = 5$ . Tenemos que

$$\sigma^2 \text{ se estima por } 11.79$$

$$\frac{\sum \tau_i^2}{t - 1} \text{ se estima por } \frac{169.41 - 11.79}{5} = 31.52$$

El valor 31.52 es una estimación de la variabilidad del conjunto fijo de los  $\tau$  y cualquier variabilidad en esta estimación, de experimento a experimento, proviene de la variabilidad en la estimación de  $\sigma^2$ ; no hay diferencia en las  $\tau$  de una muestra a otra.

Para un experimento al cual se aplica el modelo II, la variabilidad del estimativo de  $\sigma_\tau^2$  proviene de la variabilidad en el estimativo de  $\sigma^2$  y de las diferencias de los  $\tau$  de muestra a muestra.

**Tabla 7.7 Valor promedio del cuadrado medio de los tratamientos para el modelo I con desigual número de repeticiones**

Condición impuesta El cuadrado medio de los tratamientos es un estimativo de

$\sum \tau_i = 0$	$\sigma^2 + \frac{\sum r_i \tau_i^2 - (\sum r_i \tau_i)^2 / \sum r_i}{(t - 1)}$
$\sum r_i \tau_i = 0$	$\sigma^2 + (\sum r_i \tau_i^2) / (t - 1)$
Note que	$\sum r_i \tau_i^2 - \frac{(\sum r_i \tau_i)^2}{\sum r_i} = \sum r_i (\tau_i - \bar{\tau})^2$ donde $\bar{\tau} = (\sum r_i \tau_i) / (\sum r_i)$ .

Cuando hay repeticiones diferentes para los tratamientos, el coeficiente correspondiente a  $r$  para el valor promedio del cuadrado medio de los tratamientos es, para el modelo II,

$$r_0 = \left( \sum r_i - \frac{\sum r_i^2}{\sum r_i} \right) \frac{1}{t - 1} \quad (7.12)$$

Para el modelo I, el valor promedio para el cuadrado medio de los tratamientos depende de si los  $\tau$  van a ser simples desviaciones respecto de una media o de si su media ponderada, y por tanto su suma ponderada, ha de ser cero. Los valores se dan en la tabla 7.7. El valor medio del cuadrado medio del error es  $\sigma^2$ , sea cual sea la restricción que se imponga a las  $\tau$ .

Que los cuadrados medios calculados para igual número de replicaciones, tienen los valores esperados dados en la tabla 7.6, se demuestra fácilmente. El cuadrado medio para individuos dentro de tratamientos se obtiene combinando estimativos como el que se obtiene por  $Y_{i1} = \mu + \tau_i + \varepsilon_{i1}, \dots, Y_{ir} = \mu + \tau_i + \varepsilon_{ir}$  donde solo varían los  $\varepsilon$ . Evidentemente, esto nos lleva a un estimativo de  $\sigma^2$ , lo mismo puede decirse si  $r$  es también una variable. El cuadrado medio de tratamientos se basa en totales

$$Y_{1.} = r\mu + r\tau_1 + \varepsilon_{1.}, \dots, Y_{r.} = r\mu + r\tau_r + \varepsilon_{r.}$$

o en medias

$$\bar{Y}_{1.} = \mu + \tau_1 + \bar{\varepsilon}_{1.}, \dots, \bar{Y}_{r.} = \mu + \tau_r + \bar{\varepsilon}_{r.}$$

Si la hipótesis nula es verdadera, entonces los totales son de la forma  $r\mu + \varepsilon_r$  y su varianza es una estimación de  $r\sigma^2$ . Sin embargo, el divisor de  $r$  se usa en el cálculo de tal manera que el cuadrado medio en la tabla de análisis de la varianza sea un estimativo de  $\sigma^2$ . Si la hipótesis nula es falsa, entonces los totales son de la forma  $r\mu + r\tau_i + \varepsilon_i$ , los  $\tau$  y los  $\varepsilon$  contribuyen a la variabilidad de los totales. La contribución atribuible a los  $\varepsilon$  es aún  $\sigma^2$ . Al elevar al cuadrado los totales, se obtienen términos como  $r^2\tau_i^2$ , que dan  $r\tau_i^2$  al dividir por  $r$ . Así, la contribución atribuible a los  $\tau$  es  $\sum \tau_i^2 / (t - 1)$  o una estimación de  $r\sigma^2$ .

Como suponemos que los  $\tau$  y los  $\varepsilon$  son independientes, entonces no hay contribución en que entren productos de  $\tau$  y  $\varepsilon$ .

Cuando se tiene un número desigual de repeticiones, los totales son de la forma  $r_i \mu + r_i \tau_i + \varepsilon_i$ , y las medias de la forma  $\mu + \tau_i + \bar{\varepsilon}_i$ . El cálculo directo de un cuadrado medio de tratamientos a partir de totales no es válido ya que  $r_i \mu$  es variable e introduciría una contribución no debida ni a los  $\tau$  ni a los  $\varepsilon$ ; los cálculos con medias no introducen tal dificultad. Puede demostrarse que

$$\sum_i \frac{Y_i^2}{r_i} - \frac{Y_{..}^2}{\sum r_i} = \sum_i r_i (\bar{Y}_i - \bar{Y}_{..})^2$$

donde el primer miembro es la fórmula para calcular la suma de cuadrados de tratamientos. El segundo miembro es una suma ponderada de cuadrados de las desviaciones de las medias de los tratamientos respecto de la media general definida como  $\bar{Y}_{..}/(\sum r_i)$  o su equivalente  $(\sum r_i \bar{Y}_i)/(\sum r_i)$ . Las ponderaciones asignadas a las desviaciones al cuadrado son, como se ve, inversamente proporcionales a la varianza de cada media de tratamiento, o sea que una varianza grande lleva a una ponderación pequeña y viceversa, tal como debe ser el caso. Ahora es claro que si no hay diferencias debidas a los tratamientos, el cuadrado medio de los tratamientos conduce a un estimativo de  $\sigma^2$ ; si hay efectos de los tratamientos, también hay una contribución debida a las  $\tau$  (ver tabla 7.7).

Cuando  $F$  es el criterio de prueba para probar entre medias de tratamientos en un análisis de la varianza, la hipótesis nula es la de que no hay diferencias en las medias poblacionales de los tratamientos y la alterna es que existen diferencias lo más a menudo sin especificar con respecto al tamaño y a la dirección. Diferencias verdaderas pueden ser resultado de un conjunto particular de efectos fijos o de un conjunto aleatorio de efectos provenientes de una población de efectos. La base del criterio de prueba es la comparación de dos varianzas independientes que son estimaciones de una varianza común si la hipótesis nula es verdadera. Si hay diferencias reales entre tratamientos, tenemos

$$F = \frac{\widehat{\sigma^2} + r(\sum \tau_i^2)/(t-1)}{s^2} \quad F = \frac{\widehat{\sigma^2} + r\sigma_\tau^2}{s^2}$$

según sea el modelo. Los  $\widehat{\phantom{x}}$ , llamados simplemente sombreros, indican que no hay una manera de estimar las componentes de varianza con sólo el numerador. En promedio, entonces, el numerador será mayor que el denominador si hay diferencia real entre tratamientos. Estamos tratando de detectar cantidades poblacionales  $(\sum \tau_i^2)/(t-1)$  o  $\sigma_\tau^2$  según sea el modelo. Formalmente, tenemos  $H_0: \tau_i = 0, i = 1, \dots, t$  versus  $H_1: \text{algun } \tau_i \neq 0$  para el modelo fijo con la restricción  $\sum \tau_i = 0$ ; por otra parte, para el modelo aleatorio,  $H_0: \sigma_\tau^2 = 0$  versus a  $H_1: \sigma_\tau^2 > 0$ . Esta prueba se llama de una cola, donde la referencia a las colas es respecto de la distribución  $F$ .

**Ejercicio 7.5.1** Con los datos del ejercicio 7.3.1., estimar la contribución de los efectos de los tratamientos a la variabilidad medida con el cuadrado medio de los tratamientos. Decidir qué modelo se aplica y definir esta contribución en términos de las componentes del modelo.

Ejercicio 7.5.2 Repetir el ejercicio 7.5.1 con los datos después del tratamiento del ejercicio 7.3.3.

Ejercicio 7.5.3 Repetir el ejercicio 7.5.1 con los datos del ejercicio 7.4.1 y los datos después del tratamiento del ejercicio 7.4.2

## 7.6 Análisis de la varianza con submuestras. Número igual de submuestras

En algunas situaciones experimentales, se pueden tomar varias observaciones dentro de la unidad experimental, la unidad a la cual se aplica el tratamiento. Tales observaciones se hacen en *submuestras* o unidades de *muestreo*. Las diferencias entre submuestras dentro de una unidad experimental son diferencias de observación más que diferencias de unidad experimental. Por ejemplo, considérense los datos de la tabla 7.8. Un grupo grande de plantas se asignaron aleatoriamente a unas materas, cuatro por matera, la unidad experimental; los tratamientos se asignaron al azar a las materas, tres materas por tratamiento. Todas las materas se aleatorizaron completamente con respecto a su localización durante el tiempo transcurrido bajo luz del día, y cada grupo de materas se aleatorizó completamente dentro de los niveles (bajo y alto) de temperatura en invernadero durante el período de obscuridad. Las observaciones se hicieron en plantas individuales.

Dos fuentes de variación que contribuyen a la varianza aplicable a las comparaciones entre medias de tratamientos son:

1. La variación entre plantas tratadas de igual manera, esto es, entre plantas dentro de materas. Como se aplican a diferentes tratamientos, a materas diferentes y, por consiguiente, a plantas diferentes, la variación entre plantas aparecerá en las comparaciones entre medias de tratamientos.
2. La variación entre plantas en diferentes materas tratadas en forma semejante, es decir, variación entre materas dentro de tratamientos. Esta variación puede no ser mayor que la variación entre plantas tratadas análogamente, pero lo corriente es que lo sea. El investigador generalmente sabe si se puede esperar o no que ésta sea una fuente real de variación; por ejemplo, puede ser que los nutrientes, la luz, la humedad, etc., varíen más de matra a matra que dentro de una matra. Una medida de la variación calculada a partir de totales o medias de materas para el mismo tratamiento contendrá, naturalmente, ambas fuentes de variación.

Los cuadrados medios de los dos tipos de variación mencionados anteriormente se llaman en general *error de muestreo* y *error experimental*, respectivamente. Si la segunda fuente de variación no es real, entonces los dos cuadrados medios serán aproximadamente iguales. De otra manera, es de esperar que el *error experimental* sea mayor en muestreo aleatorio puesto que contiene una fuente de variación más.

Situaciones comparables se encuentran en muchos campos de experimentación. Un agrónomo puede estar interesado en determinar el número de plantas por unidad de área en un ensayo varietal en trébol rojo, o un entomólogo puede querer determinar el número de piojos en el ganado en un experimento para evaluar insecticidas para el control de esos parásitos. En ambos casos, las submuestras se pueden tomar en cada unidad experimental. En análisis químicos, determinaciones duplicadas y triplicadas se pueden hacer en cada

Tabla 7.8. Crecimiento en una semana de tallos de plantas de menta cultivadas en una solución nutritiva

Número de plantas	Horas de luz diurna												Altas temperaturas nocturnas												
	Bajas temperaturas nocturnas												Alta temperatura nocturna												
	8			12			16			8			12			16			8			12			
Matera No.	Matera No.			Matera No.			Matera No.			Matera No.			Matera No.			Matera No.			Matera No.			Matera No.			
1	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	
1	3.5	2.5	3.0	5.0	3.5	4.5	5.0	5.5	5.5	8.5	6.5	7.0	6.0	6.0	6.5	7.0	6.0	6.5	7.0	6.0	6.5	7.0	6.0	11.0	
2	4.0	4.5	3.0	5.5	3.5	4.0	4.5	6.0	4.5	6.0	7.0	7.0	5.5	8.5	6.5	9.0	7.0	7.0	7.0	7.0	7.0	7.0	7.0	7.0	7.0
3	3.0	5.5	2.5	4.0	3.0	4.0	5.0	5.0	6.5	9.0	8.0	7.0	3.5	4.5	8.5	8.5	7.5	7.5	8.5	7.0	9.0	8.5	7.0	9.0	9.0
4	4.5	5.0	3.0	3.5	4.0	5.0	4.5	5.0	5.5	8.5	6.5	7.0	7.0	7.5	7.5	7.5	8.5	7.0	8.5	7.0	8.0	8.5	7.0	8.0	8.0
Total de materia	15.0	17.5	11.5	18.0	14.0	17.5	19.0	21.5	22.0	32.0	28.0	28.0	22.0	26.5	29.0	22.0	26.5	29.0	33.0	27.0	35.0				
$= Y_{ij}$																									
Total de tratamiento	44.0			49.5			62.5			88.0			88.0			77.5			95.0						
$= Y_{i..}$																									
Medias de tratamiento	3.7			4.1			5.2			7.3			7.3			6.5			7.9						
$= \bar{Y}_{i..}$																									

Fuente: Estos datos son parte de un conjunto mayor de datos y se han utilizado con permiso de R. Rabson, Universidad de Cornell, Ithaca, Nueva York.

muestra. En experimentos sobre el terreno y con animales, la variación entre submuestras o unidades de muestreo dentro de una unidad experimental es una medida de homogeneidad de la unidad, mientras que en análisis químicos esa variación está a menudo asociada con la repetibilidad de la técnica. Esta es la variación que lleva al *error de muestreo*.

Al probar una hipótesis sobre medias de tratamientos, el divisor apropiado para  $F$  es el cuadrado medio del error experimental ya que éste incluye la variación proveniente de todas las fuentes que contribuyen a la variabilidad de las medias de los tratamientos exceptuados los tratamientos.

De los datos que pueden clasificarse según un sistema consistente en un orden único de criterios de clasificación, en que cada criterio es aplicable dentro de todas las categorías del criterio precedente, se dice que están en una *clasificación jerárquica* o *clasificación anidada*. Así, los datos de la tabla 7.8 pueden clasificarse de acuerdo con los tratamientos, la combinación de temperatura y tiempo, luego dentro de tratamientos de acuerdo con la materia y finalmente dentro de las materas según las plantas. No existe otro orden razonable. Los mismos tratamientos pueden clasificarse en uno de dos órdenes, bien sea horas dentro de temperaturas o temperaturas dentro de horas. Los tratamientos forman una clasificación de dos factores, que se estudiará en el cap. 9.

*Cálculos* Sea  $Y_{ijk}$  el crecimiento semanal de la planta  $k$  en la materia  $j$  con el tratamiento  $i$ ,  $i = 1, \dots, 4$ ,  $j = 1, 2, 3$ ,  $k = 1, \dots, 6$ . Así, la planta 2 de la materia 3 con el tratamiento 6 (alta temperatura nocturna luego de 16 horas de luz diurna), presenta en una semana un crecimiento del tallo de 7.0 cm., esto es,  $Y_{632} = 7.0$  cm. Los números de planta y materas son para comodidad en la identificación de las observaciones. Así, las plantas con el número 2 no tienen nada en común excepto el número; y las materas con el número 3 no tienen nada en común, excepto el número. Por otra parte, las plantas o materas con el mismo número de tratamiento reciben el mismo tratamiento, así que se espera que respondan en forma similar, mientras que plantas y materas con número diferente de tratamiento pueden responder en forma diferente.

Denótense los totales de materas por  $Y_{ij.}$ , totales para una combinación particular de tratamiento y materia. De este modo, el crecimiento de plantas en la materia 2 con el tratamiento 3 en una semana (baja temperatura nocturna, seguida de 16 horas de luz diurna) totaliza 21.5 cm., esto es,  $Y_{32.} = 21.5$  cm. Denótense los totales de tratamientos con  $Y_{i..}$ . El crecimiento total en una sola semana del tallo de plantas con el tratamiento 4 es 88.0 cm.; esto es  $Y_{4..} = 88.0$  cm. Denótese el gran total por  $Y_{...}$ ;  $Y_{...} = 416.5$  cm.

La notación con puntos ( $Y_{ij.}$ ,  $Y_{i..}$ ,  $Y_{...}$ ) es una abreviatura útil y común que puede generalizarse fácilmente a muchas situaciones experimentales. El punto remplaza un subíndice e indica que todos los valores que cubre el subíndice se han sumado, la información particular suministrada antes por los individuos, tal como lo indica el subíndice, no se ha cambiado por un resumen en forma de un total; el subíndice ya no se necesita y se remplaza por punto.

Por lo tanto,

$$Y_{11.} = 3.5 + 4.0 + 3.0 + 4.5 = 15.0 \text{ cm},$$

$$Y_{1..} = 3.5 + 4.0 + \cdots + 3.0 = 15.0 + 17.5 + 11.5 = 44.0 \text{ cm},$$

y así sucesivamente.

Obtener los 18 totales y sumas de cuadrados de materas, es decir,

$$\sum_k Y_{ijk} = Y_{ij}, \quad \text{y} \quad \sum_k Y_{ijk}^2$$

para las 18 combinaciones de  $i$  y  $j$ . De estos subtotales obténgase el factor de corrección  $FC$ , la suma total de cuadrados para las materas. Estos son, respectivamente,

$$C = \frac{Y_{..}^2}{srt} = \frac{(416.5)^2}{4(3)6} = 2,409.34$$

donde  $s$  es el número de submuestras por parcela, o sea, el número de plantas por materia,  $r$  es el número de repeticiones de un tratamiento, o sea, el número de materas por tratamiento, y  $t$  es el número de tratamientos.

$$\begin{aligned} \text{SCTotal} &= \sum_{i,j,k} Y_{ijk}^2 - C \\ &= (3.5)^2 + (4.0)^2 + \dots + (8.0)^2 - C = 255.91 \quad \text{con 71 gl} \end{aligned}$$

$$\begin{aligned} \text{SCMateras} &= \frac{\sum_{i,j} Y_{ij}^2}{s} - C \\ &= \frac{(15.0)^2 + \dots + (35.0)^2}{4} - C = 205.47 \quad \text{con 17 gl} \end{aligned}$$

Las sumas de cuadrados atribuibles a las submuestas (entre plantas dentro de materas) se puede encontrar por diferencia:

$$\begin{aligned} \text{SCdentro de materas} &= \text{SCTotal} - \text{SCmateras} \\ &= 255.91 - 205.47 \\ &= 50.44 \text{ con } 71 - 17 = 54 \text{ gl} \end{aligned}$$

Los cálculos hasta este punto se pueden colocar ahora en una tabla de análisis de la varianza (ver tabla 7.9).

La suma de cuadrados de materas mide la variación debido a los tratamientos, así como la variación entre materas tratadas de manera análoga. Al particionar esta suma de cuadrados, tenemos

$$\begin{aligned} \text{SCtratamientos} &= \frac{\sum_i Y_{i..}^2}{sr} - C \\ &= \frac{(44.0)^2 + \dots + (95.0)^2}{4(3)} - C = 179.64 \quad \text{con 5 gl} \end{aligned}$$

Tabla 7.9 Análisis de la varianza de los datos de plantas de menta, tabla 7.8

Fuente de variación	gl	SC	CM	CM esperado
Entre materas	17	205.47		
Tratamientos	5	179.64	35.93	$\sigma^2 + 4\sigma_t^2 + \left(12\sigma_e^2 \text{ o } 12 \frac{\sum t_i^2}{5}\right)$
Entre materas dentro de tratamiento = error experimental	12	25.83	2.15	$\sigma^2 + 4\sigma_t^2$
Entre plantas dentro de materas = error de muestreo	54	50.44	.93	$\sigma^2$
Total	71	255.91		
$s^2 = .93, s_t^2 = \frac{2.15 - .93}{4} = .30, \frac{\sum t_i^2}{5} = \frac{35.93 - 2.15}{12} = 2.82 \text{ (o } s_e^2 \text{ para el Modelo II)}$				
donde $t_i$ es una estimación de $\tau_i$ .				

Nótese que hay  $s_r = 12$  observaciones en cada total de tratamiento,  $Y_{i..}$ , que es, pues, el divisor para obtener la suma de cuadrados de tratamientos. Finalmente,

$$\begin{aligned} \text{SC materas dentro de tratamientos} &= \text{SCmateras} - \text{SCtratamientos} \\ &= 205.47 - 179.64 = 25.83 \text{ con } 17 - 5 = 12 \text{ gl.} \end{aligned}$$

La tabla de análisis de la varianza se puede completar ahora, incluyendo los cuadrados medios (ver tabla 7.9). Este tipo de análisis de la varianza se presenta a menudo sin la línea de las materas o bien solo indicada como un subtotal debajo de la línea del error experimental. Nótese que los diferentes divisores expresan cada cuadrado medio por observación, o sea, con base en las submuestras. Las medias, también deben expresarse por observación al calcularse.

La suma de cuadrados obtenida por diferencia también puede calcularse directamente. El procedimiento directo indica muy claramente qué fuente de variación entra en cada etapa y cómo resultan los grados de libertad.

El error de muestreo se refiere a las muestras provenientes de la unidad experimental, esto es, las plantas dentro de materas. Así, la primera matra contribuye con la siguiente suma de cuadrados:

$$\begin{aligned} \text{SCplantas de la matra 1, tratamiento 1} &= (3.5)^2 + \cdots + (4.5)^2 - \frac{(15.0)^2}{4} \\ &= 1.25 \text{ con } 3 \text{ gl.} \end{aligned}$$

Cálculos semejantes se efectúan para cada una de las 18 materas, lo que da lugar a 18 sumas de cuadrados, cada una con 3 grados de libertad. Su total es

$$\begin{aligned} \text{SC entre plantas tratadas análogamente} &= 1.25 + \cdots + 8.75 \\ &= 50.45 \text{ con } 18(3) = 54 \text{ gl.} \end{aligned}$$

Las sumas de cuadrados y los grados de libertad se combinan para estimar una varianza muestral. Suponemos que existe una varianza común dentro de materas, sin tener en cuenta el tratamiento.

El error experimental se refiere a la unidad experimental tratada en forma análoga, materas dentro de tratamientos. Así, el tratamiento 1 contribuye con la siguiente suma de cuadrados al error experimental:

SC materas con el tratamiento 1

$$= \frac{(15.0)^2 + (17.5)^2 + (11.5)^2 - (15.0 + 17.5 + 11.5)^2/3}{4}$$

$$= 4.54 \quad \text{con } 2 \text{ gl.}$$

Cálculos parecidos se hacen para los otros cinco tratamientos y se suman los resultados. Obtenemos

$$\text{SC materas tratadas análogamente} = 4.54 + \dots + 8.67$$

$$= 25.83 \quad \text{con } 6(2) = 12 \text{ gl}$$

Aquí suponemos que la varianza entre materas es la misma para todos los tratamientos. Las sumas de cuadrados y los grados de libertad se combinan para dar una estimación de esta varianza común.

La variación de planta a planta, que se usa para medir el error de muestreo, también está presente en la variación entre materas, ya que las diferentes materas contienen diferentes plantas, y por ello también está presente en la variación entre tratamientos. Por lo tanto, ambos, la varianza de tratamientos y la varianza de materas dentro de tratamientos, contienen una varianza de planta a planta. La variación entre tratamientos también contiene una contribución atribuible a la variación entre materas tratadas de igual manera, si la hay, ya que los efectos de los diferentes tratamientos se miden en diferentes materas. Se ve, entonces, que el error experimental es apropiado para hacer comparaciones en que entran diferentes tratamientos, en tanto que el error de muestreo no lo es, ya que el cuadrado medio de tratamiento tiene sólo una posible fuente adicional de variación no contenida en el error experimental, que es la debida a los tratamientos mismos. Una prueba válida de la hipótesis nula de que no hay diferencias entre tratamientos es la dada por

$$F = \frac{\text{Cuadrado medio de tratamientos}}{\text{Cuadrado medio del error experimental}}$$

$$= \frac{35.93}{2.15} = 16.7^{**} \quad \text{con } 5 \text{ y } 12 \text{ gl}$$

El error experimental puede o no contener variación además de la que hay entre submuestras. Esto depende de las diferencias ambientales que existen entre una unidad

experimental y otra y de si son o no mayores que las existentes dentro de la unidad, la materia, en nuestro caso. Se dispone de una prueba en

$$F = \frac{\text{Cuadrado medio del error experimental}}{\text{Cuadrado medio del error de muestreo}}$$

$$= \frac{2.15}{.93} = 2.3^* \quad \text{con } 12 \text{ y } 54 \text{ gl.}$$

El error estándar de una media de tratamiento es  $s_{\bar{Y}} = \sqrt{2.15/12} = 0.42 \text{ cm.}$ , y el error estándar de una diferencia entre medias de tratamientos es  $s_{\bar{Y}_i - \bar{Y}_j} = \sqrt{2(2.15)/12} = 0.60 \text{ cm.}$  El coeficiente de variación es  $(s/\bar{Y}) \times 100 = (1.47/5.78)100 = 25 \text{ por ciento.}$

Los cálculos por clasificaciones jerárquicas con un número creciente de criterios de clasificación son la generalización obvia de los de esta sección.

**Ejercicio 7.6.1** Comprobar los cálculos con los datos de la tabla 7.8. Obtener las sumas de cuadrados del error de muestreo y la del error experimental en forma directa y comprobar así la actividad de las sumas de cuadrados en el análisis de la varianza.

**Ejercicio 7.6.2** Hacer la diferencia entre la variabilidad de una media de materia y la variabilidad entre medias de materias (ver también sec. 7.9).

## 7.7 Modelo lineal para submuestreo

Lo visto con los datos de las plantas de menta, tabla 7.8, hace evidente la expresión matemática propuesto para explicar los datos, así:

$$Y_{ijk} = \mu + \tau_i + \varepsilon_{ij} + \delta_{ijk} \quad (7.13)$$

donde con cada observación se busca ofrecer información acerca de la media de la población de tratamientos muestreados, es decir,  $\mu + \tau_i$ . Si tratamos las  $\tau$  como efectos fijos como obviamente lo son, entonces los medimos como desviaciones tales que  $\sum \tau_i = 0$ ; si los consideramos como efectos aleatorios, entonces se puede suponer que provienen de una población con media cero y varianza  $\sigma_\tau^2$ . Se obtienen dos elementos aleatorios con cada observación. Los  $\varepsilon_{ij}$  son la unidad experimental, o sea, la materia, y se supone se distribuye normal e independientemente con media cero y varianza  $\sigma_\varepsilon^2$ ; los  $\delta_{ijk}$  son la contribución de la submuestra, o sea de las plantas, y se supone que siguen una distribución normal independiente con media cero y varianza  $\sigma^2$ . Los  $\varepsilon$  y los  $\delta$  se suponen no relacionados entre sí, o sea que al tomar un valor particular de  $\delta$  no se afecta la probabilidad de tomar un valor particular cualquiera de  $\varepsilon$ .

Ahora considérense los totales de materas usados en los cálculos del error experimental. Para nuestro conjunto particular de datos, están dados por

$$4\mu + 4\tau_i + 4\varepsilon_{ij} + \sum_k \delta_{ijk}$$

para las 18 combinaciones de  $i$  y de  $j$ . Una varianza calculada con estas sumas, con  $i$  fijo, contendría un  $4\sigma^2$  y un  $(4)^2\sigma_e^2$ ; la primera cantidad entrando en la varianza de las sumas de los  $\delta$  y la segunda en la varianza de un múltiplo de cada  $\varepsilon$ , lo que en esencia es una operación de codificación. Por lo tanto los coeficientes son 4 y  $4^2$ , respectivamente. El proceso de cálculo empleado exige dividir toda suma de cuadrados por el número de observaciones en cada cantidad elevada al cuadrado. Por consiguiente, el cuadrado medio del error experimental calculado estima  $\sigma^2 + 4\sigma_e^2$ . Un razonamiento parecido nos lleva a los coeficientes de los componentes del cuadrado medio de tratamientos. Los resultados se presentan en la tabla 7.9. Ahora se ve claro qué es lo que busca un valor de  $F$ . Por el valor de  $F$ , es evidente que tanto  $\sigma_e^2$  como una contribución de tratamiento están presentes en nuestros datos. Como la componente de tratamiento está presente a menos que tengamos una muestra poco común, concluimos que la hipótesis nula  $H_0: \tau_1 = \tau_2 = \dots = \tau_6 = 0$ , apropiada para el modelo fijo, es falsa y que por lo menos un  $\tau$  difiere de los otros. (Sin la restricción  $\sum \tau_i = 0$ , probamos la igualdad de todas las contribuciones de tratamientos).

Obsérvese que la unidad usada fue  $\frac{1}{2}$  cm y que el error de muestreo fue de 0.93, con una desviación estándar de 0.96 cm. Recuérdese que en la sec. 2.15 se recomendó un intervalo de clase no mayor que un cuarto de la desviación estándar para mantener baja la pérdida de información. Así, para una estimación del error de muestreo con baja pérdida de información debida a la elección del tamaño de la unidad de medida, la unidad escogida no fue satisfactoria. En este caso, es cosa afortunada el relativo poco uso del error de muestreo. Otro punto que se puede destacar en relación con la elección de unidad de muestreo es éste: un tratamiento como la temperatura nocturna alta con 8 horas de luz de día, figura 3, donde todas las observaciones son idénticas, no contribuye en nada a la suma de cuadrados del error de muestreo, pero contribuye con 3 grados de libertad. En casos como éste, se recomienda a veces dejar estos 3 grados de libertad por fuera del total.

**Ejercicio 7.7.1** Mediante muestreo, construir datos para un experimento que incluya tanto el error experimental como el de muestreo. Por sencillez, tómense tres tratamientos, dos unidades experimentales por tratamiento y tres observaciones por unidad. Esbozar un posible plan de tal experimento, identificando cada unidad con una ecuación que describa la composición de la observación correspondiente. Ahora, decidase en cuanto a una media general,  $\mu = 80$ , por ejemplo. En segundo lugar, elijase un conjunto de efectos fijos para tratamientos, por ejemplo,  $\tau_1 = -4$ ,  $\tau_2 = -2$  y  $\tau_3 = 6$ . Tercero, obténgase un conjunto de 6 elementos aleatorios para las 6 unidades experimentales. Para éstos, hágase un muestreo de la tabla 4.1 y calcúlese  $\varepsilon_{ij} = (Y - 40)/4$ , para  $Y$  por la tabla, de modo que la población muestreada tenga  $\mu_\varepsilon = 0$  y  $\sigma_\varepsilon = 3$ . Nótese que la muestra de los  $\varepsilon$  probablemente no tendrá  $\bar{\varepsilon} = 0$  o  $s_\varepsilon = 3$ . Finalmente, obténganse 18 elementos aleatorios para las 18 unidades experimentales. Para éstos, hacer un muestreo en la tabla 4.1 y calcular  $\delta_{ijk} = (Y - 40)/12$  de modo que la población muestreada tenga  $\mu_\delta = 0$  o  $s_\delta = 1$ .

¿Por qué es cierto que  $\varepsilon_{ij}$  tiene  $\mu = 0$  y  $\sigma = 3$ ?

Presente el análisis de la varianza con las fuentes de variación, grados de libertad y parámetros que van a estimarse. Como todos los parámetros tienen valores conocidos, es posible indicar respuestas numéricas. Hacer todos los cálculos del análisis de la varianza y comparar todas las estimaciones con los parámetros correspondientes. (Si resulta cómodo, presentar los resultados con base en clases, en que cada individuo haga su propio muestreo. Calcúlese el promedio de los valores de las distintas estimaciones obtenidas y compárense con los parámetros).

**Ejercicio 7.7.2** Repetir el ejercicio 7.7.1 para el modelo aleatorio. Para remplazar los efectos fijos, obtener  $\tau_i$  por muestreo de la tabla 4.1 con  $\mu = 40$  y  $\sigma = 12$ .

**Tabla 7.10 Observaciones sobre la calidad de un producto obtenido en ocho manufactureras en tres zonas**

Zona	A			B			C	
	I	II	III	I	II	III	I	II
Plantas	6	6, 8	6, 7, 8	5, 7	6, 7	6	7	7, 9
Observaciones								

Con un razonamiento parecido al de esta sección, considerar las sumas de los tratamientos y la varianza calculada a partir de ellas y determinar qué estima el cuadrado medio de tratamientos.

### 7.8 Análisis de la varianza con submuestras: Desigual número de submuestras

Cuando las muestras tienen un número desigual de submuestras, el análisis básico es el de la sec. 7.4; en los cálculos, el cuadrado de cualquier total se divide por el número de observaciones en el total. Por ejemplo, considérense los números de la tabla 7.10. Estos podrían ser observaciones sobre el producto de 3 plantas manufactureras en dos zonas, A y B, y de dos plantas en la zona C. Para las 14 observaciones,  $n = 14$ ,  $\sum Y = 95$ ,  $\sum Y^2 = 659$ ,  $\sum (Y - \bar{Y})^2 = 14.36$ , cm con 13 grados de libertad.

Procedimiento como en la sec. 7.4, tenemos

$$\begin{aligned} SC_{\text{plantas}}(\text{ignorando las zonas}) &= 6^2 + \frac{(6+8)^2}{2} + \cdots + \frac{(7+9)^2}{2} \\ &\quad - \frac{(6+6+8+\cdots+7+9)^2}{14} \\ &= 5.86 \quad \text{con 7 gl.} \end{aligned}$$

$$\begin{aligned} SC_{\text{residual}} &= SC_{\text{total}} - SC_{\text{plantas}} \\ &= 14.36 - 5.86 = 8.50 \quad \text{con } 13 - 7 = 6 \text{ gl.} \end{aligned}$$

Las sumas de cuadrados de plantas se pueden partitionar aún más en una componente asociada con las zonas y otra asociada con plantas dentro de las zonas.

$$\begin{aligned} SC_{\text{zonas}} &= \frac{(6+6+\cdots+8)^2}{6} + \frac{(5+\cdots+6)^2}{5} + \frac{(7+7+9)^2}{3} - C \\ &= 4.07 \quad \text{con 2 gl.} \end{aligned}$$

$$\begin{aligned} SC_{\text{plantas dentro de zonas}} &= SC_{\text{plantas}} - SC_{\text{zonas}} \\ &= 5.86 - 4.07 = 1.79 \quad \text{con } 7 - 2 = 5 \text{ gl.} \end{aligned}$$

Tabla 7.11 Análisis de la varianza de los "datos" de la tabla 7.10

Fuente	gl	SC	CM	CM es una estimación de
Zonas	2	4.07	2.03	$\sigma^2 + 1.90\sigma_e^2 + 4.50\sigma_r^2$
Plantas dentro de zonas = error experimental	5	1.79	0.36	$\sigma^2 + 1.64\sigma_e^2$
Observaciones dentro de plantas = error de muestreo	6	8.50	1.42	$\sigma^2$
Total	13	14.36		
		$s^2 = 1.42$	$s_e^2 = \frac{0.36 - 1.42}{1.64} < 0$	

El análisis de la varianza resultante se da en la tabla 7.11.

Para estos "datos", no hay evidencia de que la variación entre plantas sea de un orden de magnitud diferente al de la variación entre observaciones, ya que  $F = 0.36/1.42 < 1$ . En tal caso se dice que  $s_e^2 = 0$  en vez de un valor negativo, y especialmente si los grados de libertad para el error experimental son pocos, pueden combinarse los dos errores para obtener una nueva estimación de una varianza apropiada para probar las zonas. Aquí, la nueva estimación sería  $(1.79 + 8.50)/(5 + 6) = 0.94$ , con 11 grados de libertad.

Cuando se tiene un número desigual de submuestras, el cálculo de los coeficientes de las componentes de la varianza es mucho menos claro que en el caso de igual número. Primero, definamos  $r_{ij}$  como el número de observaciones en la planta  $j$ -ésima en la zona  $i$ -ésima; por ejemplo,  $r_{13} = 3$ . Ahora,  $r_i$  es el total de observaciones realizadas en la zona  $i$ -ésima por ejemplo,  $r_1 = 1 + 2 + 3 = 6$ . Finalmente  $r_{..}$  es el número total de observaciones,  $r_{..} = 14$ . Sea  $k$  igual al número de zonas; aquí  $k = 3$ .

El coeficiente de  $\sigma_e^2$  depende de la línea en el análisis de la varianza. Así, para plantas dentro de zonas, el coeficiente de  $\sigma_e^2$  es

$$\frac{\left(r_{..} - \sum_i \left(\sum_j r_{ij}^2/r_i\right)\right)}{gl \text{ (error experimental)}} \quad (7.14)$$

$$= \frac{14 - [(1^2 + 2^2 + 3^2)/6 + (2^2 + 2^2 + 1^2)/5 + (1^2 + 2^2)/3]}{5} = 1.64$$

Para las zonas, el coeficiente de  $\sigma_e^2$  es

$$\frac{\sum_i \left(\sum_j r_{ij}^2/r_i\right) - \left(\sum_i r_{ij}^2\right)/r_{..}}{gl \text{ (zonas)}} \quad (7.15)$$

$$= \frac{(1^2 + 2^2 + 3^2)/6 + (2^2 + 2^2 + 1^2)/5 + (1^2 + 2^2)/3 - (1^2 + \dots + 2^2)/14}{2},$$

$$= \frac{5.8 - 2}{2} = 1.90$$

El coeficiente de  $\sigma_e^2$  es

$$\frac{r_{..} - \sum_i r_i^2/r_{..}}{\text{gl(zonas)}} = \frac{14 - (6^2 + 5^2 + 3^2)/14}{2} = 4.50$$

Cuando el error de muestreo es mayor que el del error experimental, y la única explicación a esto parece ser el azar, se acostumbra estimar como cero a  $\sigma_e^2$ , o sea,  $s_e^2 = 0$ . En la situación más general, esto es, cuando  $s_e^2 > 0$ , se presenta un problema real para probar el efecto de las zonas para una componente  $\sum \tau_i^2$  o  $\sigma_e^2$  ya que ninguna línea en el análisis de la varianza difiere de la línea del cuadrado medio de zona en un múltiplo de  $\sum \tau_i^2$  o  $\sigma_e^2$  solamente. Este problema se presenta por la desigualdad en el tamaño de las submuestras.

Una solución aproximada se obtiene como sigue. En la tabla 7.11 sean los tres cuadrados medios de arriba a abajo CM(A), CM(P) y CM(0), donde A corresponde a zonas y así sucesivamente. Entonces

$$\hat{\sigma}_e^2 = \frac{CM(P) - CM(0)}{1.64} = \frac{0.36 - 1.42}{1.64}$$

Desafortunadamente, este valor es negativo en nuestro ejemplo y suficiente razón para no continuar. Haremos como si  $\hat{\sigma}_e^2$  fuera positivo.

Ahora construimos un "término de error" para probar zonas como sigue

$$\begin{aligned}\hat{\sigma}^2 + 1.90\hat{\sigma}_e^2 &= CM(0) + 1.90 \frac{CM(P) - CM(0)}{1.64} \\ &= \left(1 - \frac{1.90}{1.64}\right) CM(0) + \frac{1.90}{1.64} CM(P)\end{aligned}$$

Probamos la hipótesis nula de que no hay diferencias entre zonas usando esta función lineal de cuadrados medios independientes como denominador de una razón  $F$  sintetizada con 2 gl para el numerador y un número incierto para el denominador. Esto puede estimarse como sigue.

De manera más general, sea CM una función lineal de cuadrados medios independientes,  $CM_i$ , con  $gl = f_i$ ,  $i = 1, \dots, k$ , dada por

$$CM = \sum c_i CM_i \quad (7.16)$$

Un número de grados de libertad aproximado para la  $\chi^2$  de aproximación lo da Satterthwaite (7.15) así

$$GL \text{ efectivos} = \frac{(\sum c_i CM_i)^2}{\sum [(c_i CM_i)^2/f_i]} \quad (7.17)$$

La ecuación (5.17) se considerará como una aplicación de esta ecuación. La ecuación apropiada para los  $g_i$  se da remplazando cada  $CM_i$  por  $E(CM_i)$ , pero en la práctica no se dispondrá de valores numéricos para éstos.

**Ejercicio 7.8.1** Un método de muestreo de pesca se dio en el ejercicio 2.7.5. El muestreo, en ese caso, continuó durante la tarde. He aquí una parte de los resultados. Los datos son frecuencias de peces en varias categorías de longitud. Las designaciones de las muestras como A y B no tienen significado especial

Hora de recolección	Muestra	Categorías de longitud, en							$n$
		3	4	5	6	7	8	9	
1:50	A			5	19	19	8	3	54
	B			10	27	15	6	3	61
3:20	A		4	11	26	10	11	3	65
	B		3	11	29	13	8	1	65
4:40	A	2	8	16	44	15	8		93
	B	1	6	15	35	12	7	2	78

\* Datos por cortesía de Don Hayne, Universidad del Estado de Carolina del Norte.

Calcular un análisis de la varianza de la variable longitud. (*Fuentes:* Hora de recolección, muestras dentro de tiempos, individuos dentro de muestras). Probar la hipótesis nula obvia y discutir los resultados del experimento. Comentar sobre la unidad de medida.

## 7.9 Componentes de la varianza en experimentos planeados con submuestras

En experimentos planeados con submuestreo, se plantea el problema de la distribución del tiempo y dinero disponibles, en particular, de decidir si concentrarse en muchas muestras con pocas submuestras o pocas muestras con más submuestras. La respuesta depende de la magnitud relativa de los errores experimentales y de muestreo, y de los costos. Así, el uso de submuestras, donde se hace la medida, puede suponer costosos análisis químicos, procedimientos dispendiosos o ensayos destructivos de artículos costosos, mientras que obtener muestras puede ser de una magnitud trivial. Por otra parte, puede ocurrir que la obtención de muestras exija equipo, parcelas, animales o viajes costosos, suplementarios, en tanto que el submuestreo no suponga nada de esto. Probablemente la verdadera situación será intermedia.

Cuando se dispone de datos, se pueden usar en el planeamiento de experimentos futuros. Los del experimento con plantas de menta que se presentan en las tablas 7.8 y 7.9 se usarán para ilustrar el procedimiento. El error experimental, el criterio para juzgar la significancia de las comparaciones entre medias de tratamientos, consiste en dos fuentes de variación, variación entre plantas tratadas en forma similar,  $\sigma^2$ , y variación resultante de diferencias en el ambiente de materas tratadas análogamente,  $\sigma_e^2$ . Ambas fuentes de variación contribuyen a la varianza entre medias de tratamientos. Las estimaciones de  $\sigma^2$

y  $\sigma_e^2$  son  $s^2 = 0.93$  y  $s^2 = (2.15 - 0.93)/4 = 0.30$ . Nótese que  $s_e^2$ , la variación entre medias de materas tratadas de modo semejante, por encima y por debajo de la debida a plantas dentro de una materia, es mayor que la varianza de la media de materas basada en el error de muestreo, es decir, en  $s_e^2/4 = 0.93/4 = 0.23$ .

Supóngase que en lugar de 4 plantas por materia y 3 materas por tratamiento, se tuvieran 3 plantas por materia y 4 materas por tratamiento, el mismo número de plantas por tratamiento. El error experimental sería una estimación de  $\sigma^2 + 3\sigma_e^2$ . Un experimento semejante conduciría a una varianza de error del orden de

$$s^2 + 3s_e^2 = .93 + 3(.30) = 1.83$$

con  $3 \times 6 = 18$  grados de libertad, comparada con 2.15 y 12 grados de libertad para el experimento realizado. La varianza de una media de tratamiento es el valor importante que ha de considerarse y es  $s_y^2 = 1.83/12 = 0.1525$ , comparada con 0.1792 para el experimento realizado. Naturalmente aquí es suficiente considerar las varianzas siendo fijo el número de plantas por tratamiento. A veces, sería bueno examinar los números de las plantas.

La tabla 7.12 presenta  $s_y^2$  para otras asignaciones de 12 plantas por tratamiento. Es claro que la varianza de una media de tratamientos decrece en la medida en que aumentamos el número de materas a costa de las plantas dentro de materas. Recuérdese que con sólo una planta por materia no hay modo de estimar el error de muestreo. Nótese también el incremento de los grados de libertad para estimar el error experimental.

También se debe considerar el costo. Supóngase que conseguir una planta de menta para el experimento es bajo, y que cuesta 0.1 hora-hombre pero que el costo de la preparación de las materas es de 0.5 hora-hombre. Entonces el costo de disponer el experimento original es de  $(12 \times 0.1) + (3 \times 0.5) = 2.7$  horas-hombre por tratamiento. Es claro que el costo debe aumentar si el número de plantas por materia disminuye y el número de materas aumenta, pues las unidades experimentales suponen mayores gastos. Se presentan los costos de otros experimentos posibles con 12 plantas por tratamiento. Recuérdese que el costo adicional va acompañado de una disminución de  $s_y^2$ .

Se consideran, además, otros costos posibles de factores. Cuando el costo de cultivar una planta sube hasta 1 hora-hombre, entonces es relativamente más lento el aumento del costo de emprender experimentos con menos plantas por materia, pero con más materas.

Tabla 7.12 Algunas varianzas y costos por tratamiento con 12 observaciones

Plantas por materia	Materas	gl del error experimental	$V(\bar{Y}$ de trat.)	Costos†		
				(1, .5)	(.5, .5)	(1.0, .5)
4	3	2t	$2.15/12 = .1792$	2.7	7.5	13.5
3	4	3t	$1.83/12 = .1525$	3.2	8.0	14.0
2	6	5t	$1.53/12 = .1275$	4.2	9.0	15.0
1	12	11t	$1.23/12 = .1025$	7.2	12.0	18.0

† La primera cifra corresponde a costos por plantas en horas-hombre; la segunda a costo por materia.

Finalmente podemos considerar otras formas de distribución del esfuerzo. Por ejemplo, podemos considerar dos plantas por materia con tres materas por tratamiento y aumentar el número de tratamientos a doce. Este experimento tendría el mismo número de observaciones con un número adecuado de grados de libertad para estimar el error experimental.

Como refinamiento adicional, podríamos introducir también la idea de precisión del experimento, como se vio en la sec. 6.8.

En muchos casos, un presupuesto fijo determinará la cantidad permisible de esfuerzo. Aquí se tratará de distribuir este esfuerzo entre unidades experimentales y unidades de muestreo con el fin de minimizar la varianza de una media de tratamiento. O bien, se puede fijar la varianza deseada y hacer asignación de tal forma que se minimice el costo. (La asignación óptima se estudia en el cap. 25). Los dos enfoques llevan a la misma solución.

Cochran (7.4) da la solución de asignación óptima así

$$n_2 = \sqrt{\frac{c_1 s^2}{c_2 s_e^2}} \quad (7.18)$$

donde  $n_2$  = número de unidades de muestreo o plantas por materia

$c_1$  = costo por materia

$c_2$  = costo por planta

$s^2$  y  $s_e^2$  = valores definidos en la tabla 7.11

Obsérvese que las razones de costos y las razones de varianzas son adecuadas para despejar  $n_2$ .

Para  $c_1 = 0.05$  y  $c_2 = 0.1$

$$n_2 = \sqrt{\frac{.5}{.1} \cdot \frac{.93}{.30}} = \sqrt{15.5} \approx 4$$

El número de unidades de muestreo en este experimento fue 4.

El investigador necesitará comparar costos, mano de obra y objetivos, así como la eficiencia al elegir entre los posibles diseños.

**Ejercicio 7.9.2** ¿Cuántos grados de libertad se tienen para estimar el error de muestreo para cada experimento alternativo sugerido en la tabla 7.12?

**Ejercicio 7.9.2** Supóngase que el experimento de plantas de menta se fuera a efectuar con seis plantas por materia y dos materas por tratamiento, o sea, el mismo número de plantas por tratamiento. ¿Cuántos grados de libertad se tienen para estimar el error experimental en un experimento de tratamientos  $t$ ? ¿Para estimar el error de muestreo?

Estime  $V(Y \text{ de tratamiento})$  para esta asignación. Para cada una de las tres sugerencias sobre costos, elaborar en función de horas-hombres los costos de efectuar el experimento.

**Ejercicio 7.9.3** Repetir el ejercicio 7.9.2 con dos plantas por materia y tres materas por tratamiento. Con tres plantas por materia y tres plantas por tratamiento. Con dos plantas por materia y cuatro materas por tratamiento.

**Ejercicio 7.9.4** Encontrar la eficiencia para todas las alternativas sugeridas relativas al experimento efectuado.

**Ejercicio 7.9.5** Encontrar  $n_2$  para cada una de las razones de costos propuestas en la tabla 7.12.

### 7.10 Supuestos en que se fundamenta el análisis de la varianza

Los supuestos básicos en el análisis de varianza cuando se hacen pruebas de hipótesis son:

1. Los tratamientos y los efectos ambientales son aditivos
2. Los errores experimentales son aleatorios y se distribuyen normal e independiente-mente en torno a una media cero y con una varianza común.

El supuesto de normalidad no es necesario para estimar las componentes de la varianza. Cuando el supuesto no se hace, es necesario que los errores no estén correlacionados en vez de ser independientes. En la práctica, nunca estamos seguros de que todos estos supuestos se cumplan; a menudo hay razón para creer que algunos son falsos. Excelentes exposiciones de estos supuestos, las consecuencias cuando son falsos y las medidas remediales, se encuentran en Eisenhart (7.8), Cochran (7.3) y Bartlett (7.2). A continuación se da una breve discusión al respecto.

El incumplimiento de uno o más supuestos puede afectar tanto el nivel de significancia como la sensibilidad de  $F$  o  $t$  a las discrepancias reales respecto de la hipótesis nula. En el caso de no normalidad, el verdadero nivel de significancia es corrientemente, pero no siempre, mayor que el nivel aparente. Esto lleva al descarte de la hipótesis nula cuando es verdadera con mayor frecuencia de lo que prescribe el nivel de probabilidad, esto es, se presentan demasiadas diferencias significantes que no existen. Los experimentadores pueden pensar que están usando el nivel del 5 por ciento cuando el nivel puede ser de 7 u 8 por ciento en realidad. En algunos casos, una prueba puede detectar la no normalidad en vez de una hipótesis alterna verdadera. Se presenta pérdida de sensibilidad para las pruebas de significancia y la estimación de efectos, ya que hubiera podido combinarse una prueba más poderosa si se hubiera conocido el modelo matemático exacto. Es decir, si se conociera la verdadera distribución de los errores y la naturaleza de los efectos, su aditividad o no aditividad, entonces podrían construirse una prueba más capaz de detectar o estimar efectos reales.

Para la mayoría de datos biológicos, la experiencia indica que las perturbaciones debidas a que los datos no cumplen los anteriores requisitos no son de importancia. Hay casos excepcionales y se verán procedimientos para analizar tales datos. En todo caso, la mayoría de los datos no cumplen exactamente con los requisitos del modelo matemático y los procedimientos de pruebas de hipótesis y de estimación de intervalos de confianza no deben considerarse exactos sino aproximados.

Considérese el supuesto de que los *tratamientos* y los *efectos ambientales* son *aditivos*. Como ilustración véase la ec. (5.14) para unidades experimentales pareadas con sentido. Una forma común de no aditividad se presenta cuando tales efectos son multiplicativos. Considérese un caso simple en que dos ambientes, llamados pares en la ilustración que sigue, y dos tratamientos tienen efectos que son multiplicativos. Una comparación de modelos aditivos y multiplicativos se da en la tabla 7.13, en la cual se hace caso omiso de errores experimentales.

Tabla 7.13 Modelos aditivos y multiplicativos

Modelo	Aditivo		Multiplicativo		Log(los multiplicativos se convierte en aditivo)	
	1	2	1	2	1	2
Tratamiento 1	10	20	10	20	1.00	1.30
Tratamiento 2	30	40	30	60	1.48	1.78

Para el modelo aditivo, el aumento del bloque 1 al bloque 2 es una cantidad fija independientemente del tratamiento; lo mismo se cumple para los tratamientos. Para el modelo multiplicativo, el aumento del bloque 1 a 2 es un porcentaje fijo independientemente del tratamiento; lo mismo ocurre para tratamientos. Cuando los efectos son multiplicativos, los logaritmos de los datos manifiestan los efectos de manera aditiva y entonces es apropiado un análisis de la varianza de los logaritmos. Los datos nuevos, los logaritmos, se llaman datos transformados; el proceso de cambio de los datos es pues una *transformación*. Para otros tipos de no aditividad, se dispone de otras transformaciones. La transformación de los datos implica que los errores experimentales se distribuyan normal e independientemente en la escala transformada si se contempla hacer pruebas de significancia. En la sec. 15.8 se presenta una prueba de aditividad.

La presencia de no aditividad en los datos conduce a una aparente heterogeneidad del error debido a supuestos falsos cuando no se efectúa una transformación antes del análisis. Las componentes de la varianza del error aportadas por las diversas observaciones no dan estimaciones de una varianza común. La varianza combinada del error que resulta puede ser un tanto ineficaz para hacer estimaciones del intervalo de confianza de los efectos de tratamientos y puede dar niveles de significancia falsos para ciertas comparaciones específicas de medias de tratamientos, mientras que el nivel de significancia para la prueba *F* en que entran todas las medias de tratamientos, puede ser muy poco afectado.

Nuestro segundo supuesto no es independiente del primero, como se ha visto, y en realidad es un conjunto de supuestos. Considérese la *independencia de los errores* o más generalmente el supuesto de correlación cero entre ellos. Para experimentos en el terreno, las respuestas de cultivos en parcelas adyacentes tienden a ser más parecidas que las respuestas de parcelas no adyacentes; la anterior también es cierto para observaciones de experimentos de laboratorio realizadas por la misma persona o más o menos al mismo tiempo. La consecuencia es que las pruebas de significancia pueden ser engañosas si no se intenta superar la dificultad. En la práctica, los tratamientos se asignan a las unidades experimentales aleatoriamente o bien el orden de las observaciones se determina aleatoriamente; el efecto del proceso de aleatorización es hacer que los errores sean independientes unos de otros.

En experimentos sobre el terreno, los diseños sistemáticos convenientes colocan los mismos tratamientos adyacentes unos a otros en todos los bloques. Como las parcelas adyacentes tienden a parecerse más, la precisión de una comparación es mayor para tratamientos que caen en parcelas cercanas que para aquéllas que caen en parcelas alejadas. Un análisis de la varianza de tales datos da un término de error generalizado demasiado grande para ciertas comparaciones y demasiado pequeño para otras. No se dispone de térmi-

nos de error apropiados para comparaciones individuales. Cochran y Cox (7.4) resumen bien la necesidad de la aleatorización en la siguiente afirmación:

La aleatorización es un tanto análoga al seguro, en cuanto es una precaución contra perturbaciones que pueden o no ocurrir y que pueden ser o no graves si ocurren. En general, es aconsejable tomarse el trabajo de aleatorizar aun cuando no se espere un sesgo serio por falta de la aleatorización. De esta manera, el experimentador queda protegido ante sucesos insólitos que trastornen sus expectativas.

*El error experimental se debe distribuir normalmente* Este supuesto se aplica particularmente a pruebas de significancia y no a la estimación de componentes de la varianza. Cuando la distribución de los errores experimentales es decididamente asimétrica, la componente del error de un tratamiento tiende a ser una función de la media de tratamiento. Esto produce una heterogeneidad en el término del error. Si se conoce la relación funcional, se puede encontrar una transformación que dé errores que se distribuyan de manera más cercana a la normal. De esta forma se puede efectuar un análisis de la varianza con los datos transformados de modo que el término de error sea esencialmente homogéneo. Transformaciones comunes y útiles son las logarítmicas, raíz cuadrada y arco seno; en la sec. 9.16 se estudia su uso.

*Los errores experimentales deben tener una varianza común* Por ejemplo, en un diseño completamente aleatorizado, las componentes de error provenientes de los diferentes tratamientos, deben ser todas estimaciones de una varianza de población común. Aquí, la heterogeneidad del error puede ser resultado del comportamiento errático de la respuesta a ciertos tratamientos. En experimentos como los destinados a determinar la eficacia de diferentes insecticidas, fungicidas o herbicidas, puede incluirse un control no tratado para medir el nivel de infestación y proporcionar una base para determinar la efectividad de los tratamientos. La variación de las observaciones individuales en el control de prueba puede ser considerablemente mayor que en los otros tratamientos, ante todo porque el control puede tener una media más alta y, así, una mayor base de variación. En estas situaciones el término de error puede no ser homogéneo. Un remedio para esto es repartir el término del error en componentes homogéneas para probar comparaciones de tratamientos específicos. Algunas veces, si las medias de uno o dos tratamientos son mucho mayores que las otras y tienen una variación significativa mayor, entonces tales tratamientos se pueden excluir del análisis.

El que no se cumplan algunos de los demás supuestos puede dar lugar a heterogeneidad en el error experimental. Se han hecho sugerencias para remediar esta situación, y dependen de la naturaleza del no cumplimiento.

## Referencias

- 7.1. Anderson, R. L., y T. A. Bancroft: *Statistical theory in research*, McGraw-Hill, Nueva York, 1952
- 7.2. Bartlett, M. S.: "The use of transformations," *Biom.* 3:39-52 (1947).
- 7.3. Cochran, William G.: "Some consequences when the assumptions for the analysis of variance are not satisfied," *Biom.* 3:22-38 (1947).
- 7.4. Cochran, William G.: *Sampling techniques*, 2a. ed., Wiley, Nueva York, 1965.

- 7.4. Cochran, William G., y Gertrude M. Cox: *Experimental designs*, 2a. ed., Wiley, Nueva York, 1957.
- 7.5. Duncan, D. B.: "A significance test for differences between ranked treatments in an analysis of variance," *Va. J. Sci.*, 2: 171-189 (1951).
- 7.6. Duncan, D. B.: "Multiple range and multiple *F* tests," *Biom.*, 11: 1-42 (1955).
- 7.7. Dunnett, C. W.: "A multiple comparisons procedure for comparing several treatments with a control," *J. Amer. Statist. Ass.* 50: 1096-1121 (1955).
- 7.8. Eisenhart, C.: "The assumptions underlying the analysis of variance," *Biom.*, 3: 1-21 (1947).
- 7.9. Erdman, Lewis W.: "Studies to determine if antibiosis occurs among *Rhizobia*. I. Between *Rhizobium meliloti* and *Rhizobium trifolii*," *J. Amer. Soc. Agron.*, 38: 251-258 (1946).
- 7.10. Harter, H. L.: "Error rates and sample sizes for range tests in multiple comparisons," *Biom.*, 13: 511-536 (1957).
- 7.11. Hartley, H. O.: "Some recent developments in analysis of variance," *Comm. Pure Appl. Math.*, 8: 47-72 (1955).
- 7.12. Jordan, H. G.: "Sensory restriction and suggestion: A proposed treatment modality for selected prison inmates," Tesis doctoral, North Carolina State University, Raleigh, NC, 1971.
- 7.12. Kramer, C. Y.: "Extension of multiple range tests to group means with unequal numbers of replication," *Biom.*, 12: 307-310 (1956).
- 7.13. Newman, D.: "The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation," *Biometrika*, 31: 20-30 (1939).
- 7.14. Sanders, O. T., Jr.: "Effects of a polychlorinated biphenyl in mice caged at different densities and fed at different nutritional levels," Tesis doctoral, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1974.
- 7.15. Satterthwaite, F. E.: "An approximate distribution of estimates of variance components," *Biom.*, 2: 110-114 (1946).
- 7.16. Snedecor, G. W.: *Statistical methods*, 5a. ed., Iowa State College Press, Ames, Iowa, 1956.
- 7.17. Tukey, J. W.: "The problem of multiple comparisons," Princeton University, Princeton, NJ, 1953 (Ditto.)
- 7.18. Wexelsen, H.: "Studies in fertility, inbreeding, and heterosis in red clover (*Trifolium pratense L.*)," *Norske videnskaps-akad. I Oslo, Mat.-Natur. klasse*, 1945.

---

## CAPITULO OCHO

---

### COMPARACIONES MULTIPLES

#### 8.1 Introducción

En el capítulo 7, se usó la prueba de  $F$  para probar diferencias reales entre tratamientos. Cuando no se rechaza la hipótesis nula, parecería innecesario plantearse más preguntas. Sin embargo, considerar el conjunto de tratamientos en el experimento de las plantas de menta hace pensar que ésta es una simplificación exagerada. En este experimento se compararon seis tratamientos. Supóngase que se hayan declarado no diferentes de modo significante. ¿No hubiéramos podido preguntarnos si alguna diferencia real entre temperaturas pudo haberse perdido al promediar con las otras posibles comparaciones?

Si se rechaza la hipótesis nula cuando se usa la prueba  $F$ , entonces qué tan satisfecho está el investigador? ¿Dónde están las diferencias reales? En el experimento con plantas de menta, ¿debió haber planeado buscar posibles diferencias debidas a diferentes temperaturas nocturnas y a diferentes horas de luz diurna? Si tales preguntas obvias no se formulan, entonces es razonablemente claro que la máxima diferencia observada puede declararse significante ya que  $H_0$  se ha rechazado. Además, ¿qué puede decirse respecto a otras diferencias?

Con la prueba de  $t$ , se comete un error de tipo I al rechazar erróneamente una simple hipótesis respecto a un solo parámetro o diferencia. Con la prueba de  $F$ , la hipótesis incluye muchos parámetros dentro de un experimento; esto se puede considerar como la prueba simultánea de hipótesis respecto a muchas diferencias en las que el rechazo general depende de una o más diferencias no especificadas. ¿Ha cambiado nuestro concepto de error de tipo I?

Este capítulo se ocupa de los procedimientos y tasas de error de comparaciones múltiples. El mayor énfasis se hace sobre casos con igual número de repeticiones, pero el problema de repetición desigual se considera en la sec. 8.11.

## 8.2 La diferencia mínima significante

Supongamos que al planear el experimento sobre *Rhizobium* de la sec. 7.3, el investigador había decidido comparar medias de tratamientos de 3DOk1 y 3DOk5, 3DOk4 y 3DOk7, y 3DOk13 y la combinación con un nivel de significancia del 5 por ciento, usando tres pruebas *t*; las medias se dan al pie de la tabla 7.1. O bien el investigador pudo calcular la diferencia más pequeña que se pudiera declarar significante y comparar con ella el valor absoluto de cada una de las diferencias observadas. Para ésto, el valor muestral de *t* tendría que ser igual o mayor que el valor tabulado de *t* usado como valor crítico; esto es, para una prueba al nivel  $\alpha$  frente a alternativas bilaterales, declarará significante la prueba cuando

$$\frac{|\bar{Y}_i - \bar{Y}_{i'}|}{s_{\bar{Y}_i - \bar{Y}_{i'}}} \geq t_{\alpha/2}$$

o cuando  $|\bar{Y}_i - \bar{Y}_{i'}| \geq t_{\alpha/2} s_{\bar{Y}_i - \bar{Y}_{i'}}$ . El criterio de prueba para examinar directamente las diferencias entre medias se llama *diferencia mínima significante*, o dms, dada por

$$\begin{aligned} \text{dms} &= t_{\alpha/2} s_{\bar{Y}_i - \bar{Y}_{i'}} \\ &= t_{\alpha/2} s \sqrt{\frac{2}{r}} \quad \text{para } r \text{ igual para todos los tratamientos} \end{aligned} \quad (8.1)$$

donde *s* es la raíz cuadrada de la varianza del error combinada. Como la dms sólo se calcula una vez y se requiere de la varianza del error combinada, su uso es una comodidad en relación con la ejecución de pruebas *t* individuales.

Para los datos de *Rhizobium*,

$$\text{dms}(0.05) = t_{0.025} s_{\bar{Y}_i - \bar{Y}_{i'}} = 2.064 \sqrt{\frac{2(11.79)}{5}} = 4.5 \text{ mg}$$

$$\text{dms}(0.01) = t_{0.005} s_{\bar{Y}_i - \bar{Y}_{i'}} = 2.797 \sqrt{\frac{2(11.79)}{5}} = 6.1 \text{ mg}$$

Las diferencias observadas son  $\bar{Y}_1 - \bar{Y}_2 = 28.8 - 24.0 = 4.8$ ;  $\bar{Y}_3 - \bar{Y}_4 = 14.6 - 19.9 = -5.3$ ; y  $\bar{Y}_5 - \bar{Y}_6 = 13.3 - 18.7 = -5.4$  mg; todas caen dentro 4.5 y 6.1 en valor absoluto, así que esas diferencias son significantes al nivel del 5 por ciento, pero no al nivel del 1 por ciento. Aquí tenemos un ejemplo de uso válido de la dms.

Pruebas con alternativas unilaterales tal como se vieron en el cap. 5, también pueden llevarse a cabo.

Como la dms puede usarse incorrectamente, y a menudo así ocurre, algunos estadísticos vacilan en recomendarla. El uso incorrecto más común es hacer comparaciones sugeridas por los datos, comparaciones que no se han planeado inicialmente. En una comparación planeada se les da nombres a los tratamientos de antemano; no es necesario esperar los resultados, por ejemplo, para saber cuál tratamiento está asociado con la mayor respuesta. La realización de comparaciones no planeadas se suele llamar de "efectos

sugeridos por los datos" o "razonamientos posteriores a los hechos". Para que los niveles de significancia tabulados sean válidos, la dms sólo deberá usarse para comparaciones planeadas antes de examinar los datos.

Cochran y Cox (8.3) indican que en la situación extrema en que el experimentador sólo compara la diferencia entre la más alta y la más baja de las medias de los tratamientos, mediante la prueba  $t$  o la dms, esta diferencia será probablemente considerable aun cuando no exista efecto. Puede demostrarse que con 3 tratamientos el valor observado de  $t$  para la máxima diferencia será mayor que el valor tabulado al nivel del 5 por ciento en el 13 por ciento de las veces; con seis tratamientos, la cifra es 40 por ciento, con diez tratamientos, 60 por ciento y con veinte tratamientos, 90 por ciento. Así, cuando los experimentadores piensan que hacen una prueba  $t$  al 5 por ciento, en realidad lo están haciendo a un nivel del 13 por ciento para tres tratamientos, 40 por ciento para seis tratamientos, y así sucesivamente.

Sin embargo, el nivel de probabilidad tabulado para  $t$  es correcto cuando se usa la dms para hacer comparaciones planeadas múltiples de medias pareadas. Para ver esto, supóngase que son correctos todos los supuestos que exige el modelo; que todas las respuestas atribuibles a los tratamientos son iguales, es decir, que todas las hipótesis nulas son verdaderas, y que se planea hacer varias comparaciones. Entonces si el experimento se repite indefinidamente, a menudo y en cada oportunidad se le pide a  $A$  que sólo pruebe la primera hipótesis, debe encontrar que el 100 $\alpha$  por ciento de sus diferencias son mayores que la dms; y si en cada oportunidad se le pide a  $B$  que pruebe solamente la segunda hipótesis, debe encontrar que 100 $\alpha$  por ciento de sus diferencias son significantes; y así sucesivamente. Es claro que el 100 $\alpha$  por ciento de todas las comparaciones hechas se declaran significantes en forma errónea. Esto sería cierto todavía si se hicieran todas las comparaciones de todos los pares de medias posibles. Nótese que no se ha hecho una prueba de  $F$  inicial; cada prueba de  $t$  dentro del experimento se hace directamente como si fuera el único y todas se hacen con la misma varianza del error.

La tasa de error ha sido definida por implicación; el número de inferencias erróneas se ha comparado con el número de inferencias hechas. Esta es una *tasa de error por comparación* y es realmente el valor dado por aproximación en la ec. (8.2) en experimentación repetida.

#### Tasa de error por comparación ( $H_0$ verdadera)

$$\frac{\text{Número de diferencias erróneas}}{\text{Número de inferencias hechas}} \quad (8.2)$$

Nótese en particular que antes del experimento, establecemos las comparaciones que se han de efectuar; ninguna ha sido sugerida por los datos. Si se hubiere pedido a  $A$  que probara la máxima diferencia observada, esta hubiera sido una comparación planeada entre los tratamientos con denominación previa y se habría declarado significancia con mayor frecuencia que el 100 $\alpha$  por ciento de las veces. Por otra parte, una persona asignada a la labor de probar el par de medias más cercanas declararía significante esta diferencia con una probabilidad tabulada para  $t$ .

Los intervalos de confianza también pueden construirse y a menudo son más útiles que las pruebas. La probabilidad de que un intervalo de confianza contenga la diferencia

que se estima es  $1 - \alpha$ , así a la larga el  $(1 - \alpha) 100$  por ciento de los intervalos de confianza contendrán la verdadera diferencia. También pueden construirse intervalos de confianza unilaterales.

En resumen, el uso de la dms es una prueba válida para comparaciones planeadas. La tasa de error implicada es por comparación y son apropiados los niveles de probabilidad tabulados para  $t$ . Como la dms sólo necesita calcularse una vez y se aprovecha la varianza combinada del error, puede ser conveniente en cuanto al uso de prueba de  $t$  múltiples. Muchos investigadores la consideran como la prueba más apropiada cuando las comparaciones han sido planeadas con sentido en términos de la naturaleza de los tratamientos, pero no la usarían para comparar todos los posibles pares de medias.

*Todas las posibles comparaciones pareadas*, lo cual significa comparaciones de todos los posibles pares de medias, merecen la atención en cuanto a presentación. Las siguientes sugerencias también pueden usarse con los procedimientos para probar que siguen para todos los posibles pares de medias.

Primeramente, ordenar las medias de menor a mayor, o a la inversa. Puede ser útil espaciarlas de una manera que se aproximen a las diferencias entre medias.

3DOk13	3DOk4	Combinación	3DOk7	3DOk5	3DOk1	
13.3(1)	14.6(2)		18.7(3)	19.9(4)	24.0(5)	28.8(6)

En seguida hacer un listado de todas las diferencias y probar. Encontramos

- (6) – (1) = 15.5 > 4.5; significante
- (6) – (2) = 14.2 > 4.5; significante
- .....
- (6) – (5) = 4.8 > 4.5; significante
- (5) – (1) = 10.7 > 4.5; significante
- (5) – (2) = 9.4 > 4.5; significante
- .....
- (2) – (1) = 1.3 < 4.5; no significante

El patrón es claro. Como alternativa a las palabras "significante" y "no significante" se puede usar un "\*" y un "ns" para presentar la misma información.

Una presentación obvia que permite ahorrar espacio enumera las diferencias como en la tabla 8.1.

Finalmente, Duncan (8.6) sugiere lo siguiente, que bien puede ser el método más popular de presentación

13.3(1)	14.6(2)	18.7(3)	19.9(4)	24.0(5)	28.8(6)
---------	---------	---------	---------	---------	---------

Las medias se colocan aproximadamente a escala. Todo par de medias no subrayado por la misma línea son significativamente diferentes. Por ejemplo, 18.7 y 19.9 están subraya-

**Tabla 8.1 Diferencias entre medias de nitrógeno en un experimento con *Rhizobium***

	(6)	(5)	(4)	(3)	(2)
(1)	15.5*	10.7*	6.6*	5.4*	1.3
(2)	14.2*	9.4*	5.3*	4.1	
(3)	10.1*	5.3*	1.2		
(4)	8.9*	4.1			
(5)	4.8*				

dos por la misma línea, así que la evidencia es que las dos poblaciones en cuestión no se pueden distinguir; 14.6 y 19.9 no están subrayadas por la misma línea, así que las dos poblaciones parecen tener diferente localización; y 28.8 aparece solo, así que su población principal es distinta de todas las otras.

Para esta presentación, súmese 4.5, la dms, a 13.3. El resultado es 17.8, así que subrayense los valores 13.3 y 14.6 pero no 18.7. A 14.6 súmese 4.5 y continúese el proceso. Nunca debe haber una linea que quede completamente dentro de otra más larga; la más larga indica ya que sus medias son homogéneas, esto es, que pertenecen a una sola población.

Para quienes deseen utilizar la dms para hacer todas las posibles comparaciones por pares de medias, se recomienda como medida prudente primero efectuar una prueba de tratamientos y sólo si  $F$  resulta significante, entonces proceder a hacerlas. Este procedimiento se conoce como la dms (*protegida*) de Fisher. Como  $F$  nos conduce a aceptar o descartar una hipótesis en que entran simultáneamente todas las medias, la unidad para juzgar la tasa de error ya no es la comparación, sino el experimento. Una tasa de error semejante se estudia en la sec. 8.4.

Carmer y Swanson (8.1 y 8.2) consideran las propiedades de estas pruebas y otras introducidas en este capítulo. Sus estudios se refieren a la idoneidad de las pruebas para detectar diferencias reales y otras cosas.

**Ejercicio 8.2.1** MacDonald (8.10) estudió la aptitud de un escollo artificial en forma de barco de transporte para atraer y mantener epifauna macrobentídica. Las variables de mayor interés fueron peso seco y densidad medidas por el número de organismos por  $400\text{cm}^2$ . Examinaremos algunos datos de densidad.

Los datos de densidad heterogénea así que MacDonald consideró que  $Y = \sqrt{\text{recuento}} + 0.5$ , una transformación común para recuentos (ver sec. 9.16). Al terminar es deseable efectuar una transformación inversa sobre las medias  $Y$  elevando al cuadrado y restando 0.5. Estas serán algo más pequeñas que las medias de los datos originales y, por lo tanto, estimaciones prudentes de la densidad media.

Las siguientes medias ordenadas se obtuvieron con los datos transformados en la familia Ostreidae, ostras. Cada media se basa en doce observaciones;  $s^2 = 10.697$  con  $66\text{ gl}$ .

Lugares	Pisos de las bodegas	Cubierta de estribor	Costado de estribor	Costado de babor	Costados de las bodegas	Lados de las bodegas
Medias	4.05	7.76	7.85	10.48	10.54	11.28

Calcular la dms, probar todos los pares de medias y presentar los resultados con la técnica de subrayado de Duncan. Elaborar una tabla como la tabla 8.1. Calcular las correspondientes medias de las transformaciones inversas.

**Ejercicio 8.2.2** Repetir el ejercicio 8.2.1 con los siguientes datos de densidad transformados obtenidos por MacDonald (8.10). Estos son de la familia Arbaciidae, erizo marino. Cada media proviene de doce observaciones;  $s^2 = 0.218$  con 66 gl.

Lugares	Pisos de las bodegas	Costados de las bodegas	Cubierta de estribor	Cubierta de babor	Costado estribor	Costado de babor
Medias	0.79	0.98	1.16	1.30	1.36	1.44

**Ejercicio 8.2.3** Con los datos del ejercicio 7.3.1, calcular la dms, probar todos los pares de medias y presentar los resultados con la técnica de subrayado de Duncan. Así mismo, elaborar una tabla como la tabla 8.1.

**Ejercicio 8.2.4** Considerar los datos de la prueba posterior del ejercicio 7.3.3. Calcular la dms y probar todos los pares de medias.

### 8.3 Comparaciones

En la sección 8.2, se recomendó la dms para comparaciones planeadas, en las que intervienen pares de medias. Una recomendación más general es el uso de las pruebas  $t$  para comparaciones planeadas en que entran funciones lineales de observaciones.

En la sección 5.10, se da la media y la varianza de una función lineal de observaciones; en particular, ver las ecs. (5.8), (5.19) y (5.23) aplicables en el siguiente desarrollo.

*Las comparaciones o contrastes*, definidas por la ec. (8.3), son un subconjunto de funciones lineales.

$$Q = \sum c_i Y_i \quad \text{con} \quad \sum c_i = 0 \quad (8.3)$$

Las  $Y_i$  pueden ser totales o medias de tratamientos; los primeros son más convenientes para las pruebas, los últimos para la estimación de intervalos de confianza. Por comodidad, los  $c_i$  son generalmente enteros; es esencial la restricción de que  $\sum c_i = 0$ . Una comparación siempre tiene un solo grado de libertad, así que  $t$  es la prueba apropiada, pero puede usarse  $F$  también.

Supóngase que los  $Y_i$  de la ec. (8.3) sean medias muestrales de  $r$  observaciones provenientes de poblaciones con  $\mu_i$  posiblemente diferentes pero con una varianza común. Entonces las ecs. (5.19) y (5.23) dan

$$Q = \sum c_i Y_i \quad E(Q) = \sum c_i \mu_i \quad \sigma_Q^2 = \frac{(\sum c_i^2) \sigma^2}{r}$$

Para probar  $H_0: \sum c_i \mu_i = \sum c_i \tau_i = 0$ , calcúlese  $t$  por

$$t = \frac{Q}{s_Q} = \frac{Q}{s \sqrt{(\sum c_i^2)/r}} \quad (8.4)$$

donde  $s$  es la raíz cuadrada del cuadrado medio del error experimental y  $r$  es el número de observaciones por tratamiento. Cuando la unidad experimental se muestrea,  $r$  es un múltiplo del número de repeticiones. Obsérvese que  $Q$  es un estimativo de  $\sum c_i \mu_i = \sum c_i \tau_i$ , así que un intervalo de confianza se construye fácilmente como  $Q \pm ts_Q$ .

Para probar hipótesis nulas, generalmente es más conveniente usar  $F$  y calcular el numerador a partir de los totales de tratamientos. Sea la ec. (8.5) la que define una comparación usando totales.

$$Q = \sum c_i Y_i \quad \text{con} \quad \sum c_i = 0 \quad (8.5)$$

Entonces  $E(Q) = r \sum c_i \mu_i$  y  $\sigma_Q^2 = r (\sum c_i^2) \sigma^2$ . En consecuencia, la suma de cuadrados y el cuadrado medio, por cada observación, atribuible al contraste definido en la ec. (8.5), se define por

$$SC(Q) = CM(Q) = \frac{Q^2}{r \sum c_i^2} \quad (8.6)$$

que puede usarse directamente como denominador de una prueba de  $F$  con  $s^2$  como denominador para probar  $H_0: \sum c_i \mu_i = 0$ .

Considérense de nuevo los datos de plantas de menta de la tabla 7.8. Supóngase que antes de realizar el experimento, se decidió examinar las siguientes diferencias entre respuestas:

1. Entre las temperaturas nocturnas alta y baja.
2. Entre 8 y 16 horas de luz diurna. Como éstos son valores extremos, se espera una diferencia máxima de medias de exposición pareadas.
3. Entre el promedio de 8 h más 16 h y 12 h. La media de 8 más 16 es 12. Si no se observa respuesta en 2, entonces no es probable una respuesta en esta etapa; las plantas no presentan una respuesta diferencial debido a la variación en longitud del día en este intervalo. Si se observa una respuesta en 2, entonces una respuesta ahora indica que la tasa de crecimiento varía dentro de todo el intervalo.
4. La respuesta en 2, si es real, puede tener componentes que difieren en magnitud con la temperatura nocturna. En consecuencia, examinamos la diferencia entre dos diferencias, entre las 8 y 16 horas para temperaturas nocturnas alta y baja.
5. Lo que se dijo en 4 respecto a 2, también se aplica a 3.

Enumérense los tratamientos de la tabla 7.8, de izquierda a derecha, como 1[1]6. Formalmente, entonces, se proponen las siguientes hipótesis para probar las anteriores diferencias en cuanto a significancia:

1.  $H_0: (\mu_4 + \mu_5 + \mu_6)/3 = (\mu_1 + \mu_2 + \mu_3)/3, \text{ o}$   
 $H_0: \mu_1 + \mu_2 + \mu_3 - \mu_4 - \mu_5 - \mu_6 = 0$
2.  $H_0: (\mu_1 + \mu_4)/2 = (\mu_3 + \mu_6)/2, \text{ o}$   
 $\mu_1 + \mu_4 - \mu_3 - \mu_6 = 0$

3.  $H_0: (\mu_1 + \mu_4 + \mu_3 + \mu_6)/4 = (\mu_2 + \mu_5)/2, \text{ o}$   
 $H_0: \mu_1 - 2\mu_2 + \mu_3 + \mu_4 - 2\mu_5 + \mu_6 = 0$
4.  $H_0: \mu_1 - \mu_3 = \mu_4 - \mu_6, \text{ o}$   
 $H_0: \mu_1 - \mu_3 - \mu_4 + \mu_6 = 0$
5.  $H_0: (\mu_1 + \mu_3)/2 - \mu_2 = (\mu_4 + \mu_6)/2 - \mu_5, \text{ o}$   
 $H_0: \mu_1 - 2\mu_2 + \mu_3 - \mu_4 + 2\mu_5 - \mu_6 = 0$

En 3, nótese que  $\mu_1 - 2\mu_2 + \mu_3 = (\mu_1 - \mu_2) - (\mu_2 - \mu_3)$  es una comparación entre el crecimiento en las primeras 4 horas más allá de 8, y en las segundas 4. En consecuencia, es una medida de la diferencia entre las tasas de crecimiento con temperatura nocturna baja. A su turno, la función lineal de la hipótesis 3 combina esta respuesta con ambas temperaturas. En la hipótesis 5, se examina la diferencia. Las funciones lineales en la hipótesis 2 y 4 son semejantes, pero la primera es una combinación más simple y la otra corresponde a respuestas diferenciales.

Nótese la función lineal de los  $\mu_i$  dada en la segunda expresión formal de cada hipótesis nula. Cada función puede estimarse mediante la misma función de la  $\bar{Y}_{..}$ , o bien estimamos un múltiplo de cada función mediante las  $Y_{ij}$ . Cada estimativo se considera una estimación. Si los estimativos se denotan por  $Q_1, \dots, Q_5$  entonces las sumas de cuadrados atribuibles a cada contraste, por cada observación, están dadas por la ec. (8.6) así

$$\begin{aligned} SC(Q_1) &= (44.0 + 49.5 + 62.5 - 88.0 - 77.5 - 95.0)^2/4(3)6 \\ &= (-104.5)^2/72 = 151.67 \quad F = 70.54^{**} \end{aligned}$$

$$\begin{aligned} SC(Q_2) &= (44.0 - 62.5 + 88.0 - 95.0)^2/4(3)4 \\ &= (-25.5)^2/48 = 13.55 \quad F = 6.30^* \end{aligned}$$

$$\begin{aligned} SC(Q_3) &= [44.0 - 2(49.5) + 62.5 + 88.0 - 2(77.5) + 95.0]^2/4(3)12 \\ &= (35.5)^2/144 = 8.75 \quad F = 4.07 \end{aligned}$$

$$\begin{aligned} SC(Q_4) &= (44.0 - 62.5 - 88.0 + 95.0)^2/4(3)4 \\ &= (-11.5)^2/48 = 2.76 \quad F = 1.28 ? \end{aligned}$$

$$\begin{aligned} SC(Q_5) &= [44.0 - 2(49.5) + 62.5 - 88.0 + 2(77.5) - 95.0]^2/4(3)12 \\ &= (-20.5)^2/144 = 2.92 \quad F = 1.36 \end{aligned}$$

Cada estimativo se ha contrastado con  $F$ , usando el error experimental con 12 grados de libertad.

La información sobre las comparaciones puede presentarse en forma conveniente en una tabla tal como la tabla 8.2. Los totales de los tratamientos se encuentran en la parte superior de la tabla, los coeficientes de los contrastes se encuentran en el cuerpo de la tabla, y a la derecha se completan los cálculos. Obsérvese que el único efecto de un cambio completo de signos para cualquier contraste es un cambio en el signo de  $Q$ .

Ahora es claro que la mayor parte de variación entre las medias de tratamientos está asociada con el efecto de las temperaturas nocturnas alta y baja; la observación de las medias o el signo del contraste indica que el crecimiento es máximo con la temperatura alta. También una diferencia significante en crecimiento debida a las horas de luz diurna, con mayor crecimiento en el día más largo. El tercer contraste sugiere que la tasa de creci-

**Tabla 8.2 Información de comparación para los datos de las plantas de menta de la tabla 7.8**

Contraste	Nombres de tratamientos y totales						$Q_i$	$r \sum c_i^2$	$SC(Q) = Q^2 / r \sum c_i^2$	$F$
	L8	L12	L16	H8	H12	H16				
1	1	1	1	-1	-1	-1	-104.5	12(6)	151.67	70.54**
2	1	0	-1	1	0	-1	-25.5	12(4)	13.55	6.30**
3	1	-2	1	1	-2	1	35.5	12(12)	8.75	4.07
4	1	0	-1	-1	0	1	-11.5	12(4)	2.76	1.28
5	1	-2	1	-1	2	-1	-20.5	12(12)	2.92	1.36

Tab  $F(1, 12)$ :  $F_{0.05} = 4.75$ ;  $F_{0.01} = 9.33$

miento no es constante a medida que aumenta la longitud del día, pues  $F_{1, 12}(0.05) = 4.75$ . Puesto que sólo había 12 grados de libertad en el error, valdría la pena estudiarse en una futura investigación este aspecto.

Preguntas que en esta etapa se formulan a menudo son: ¿qué tan estrechamente relacionadas están estas comparaciones con sentido? Parece que existe una buena cantidad de superposiciones con algunas medias usadas en total. ¿Se ha sacado toda la información disponible respecto a los tratamientos?

La *ortogonalidad* entre dos contrastes se define en términos de los coeficientes. Si  $Q_1 = \sum c_{1i} Y_i$  y  $Q_2 = \sum c_{2i} Y_i$ , entonces son ortogonales si se cumple la ec. (8.7).

$$\sum c_{1i} c_{2i} = 0 \quad (8.7)$$

Para las comparaciones 1 y 2 sobre los datos de la menta, tenemos

$$1(1) + 1(0) + 1(-1) + (-1)1 + (-1)0 + (-1)(-1) = 0$$

Estas comparaciones son ortogonales, como lo son todos los pares en el conjunto escogido. Nótese lo conveniente que resulta la tabla 8.2 para comprobar la ortogonalidad.

En general,  $t$  tratamientos piden  $t - 1$  gl y las sumas de cuadrados de tratamientos pueden repartirse en  $t - 1$  componentes con 1 solo grado de libertad cada una, cuya suma es igual a esta suma de cuadrados. Sin embargo, las  $t - 1$  componentes deben derivarse de un conjunto de  $r - 1$  contrastes ortogonales. De tal conjunto de contrastes puede considerarse que incluye toda la información disponible en los datos. Esta afirmación se refuerza más si observamos que como  $\sum c_{1i} = 0 = \sum c_{2i}$ , la ec. (8.7) no es otra cosa que el numerador de una covarianza, tal como se define en la ec. (5.27). Pero depende del investigador decidir si los contrastes tienen sentido en términos de la naturaleza de los tratamientos, y si esto es lo que se quería de los datos.

Muchos estadísticos recomiendan que las pruebas con una tasa de error se usen para los contrastes ortogonales.

Cuando  $\sum c_{1i} c_{2i} \neq 0$ , los contrastes no son ortogonales. Un conjunto no ortogonal de  $t - 1$  contrastes con un solo grado de libertad cada uno no dará sumas de cuadrados cuya suma total sea igual a la suma de cuadrados de tratamientos.

**Ejercicio 8.3.1** Demostrar que los cinco contrastes utilizados en el análisis de los datos de plantas de menta son ortogonales dos a dos.

**Ejercicio 8.3.2** Demuéstrese que las sumas de cuadrados para el conjunto ortogonal de comparaciones dan la suma de cuadrados de tratamientos.

**Ejercicio 8.3.3** En el ejercicio 7.3.1, escribir los coeficientes necesarios para comparar el "control" con cada tratamiento. ¿Cuántos grados de libertad supone el conjunto? ¿Cuántos grados de libertad hay para los tratamientos? ¿Son los contrastes ortogonales? Calcular la suma de cuadrados para cada contraste. Comparar la suma de las seis sumas de cuadrados de los contrastes con la suma de cuadrados de tratamiento.

**Ejercicio 8.3.4** En el ejercicio 7.3.2, escribir los coeficientes para el contraste que compara el "control" con la media de todas las demás observaciones. Emplear estos coeficientes en el cálculo de la suma de cuadrados para el contraste. Compárese este resultado con el obtenido en el ejercicio 7.3.2.

**Ejercicio 8.3.5** Los contrastes 1, 2 y 3 tal como se escribieron primero como hipótesis nulas son comparaciones de dos medias evidentemente. En 3, éstas no componen números iguales. El procedimiento general para encontrar las sumas de cuadrados por observación para desigual número de repeticiones se da en la sec. 7.8. Con este procedimiento, demostrar que se obtienen los mismos resultados que por el método usado en la presente sección.

**Ejercicio 8.3.6** En el ejercicio 8.2.1, McDonald (8.10) se propuso hallar semejanzas y diferencias de respuesta en varias localidades. Supóngase que planeó realizar los siguientes contrastes:

1. Pisos en costados de las bodegas
2. Babor con estribor
3. Cubiertas con costados
4. La diferencia entre los componentes de babor con los de estribor, esto es, los costados y las cubiertas.
5. La diferencia entre las componentes de costados y cubiertas, esto es, entre las de babor y estribor.

Disponer una tabla como la tabla 8.2 y probar la hipótesis nula de que la media poblacional para cada contraste es cero. ¿Constituyen estos contrastes un conjunto ortogonal?

**Ejercicio 8.3.7** Repetir el ejercicio 8.3.6 con los datos del ejercicio 8.2.2 sobre erizos de mar.

**Ejercicio 8.3.8** En el ejercicio 7.3.3, el "tratamiento" se define como una combinación de una restricción sensorial y sugestión. Así que los tratamientos 2 y 3 se deben considerar como controles.

Para los datos de la prueba posterior, dos contrastes de interés pudieron haberse planeado así (1) comparación de los dos controles (2) una comparación del "tratamiento" con la media de los controles.

Probar la hipótesis nula de que la media poblacional de cada contraste es cero. ¿Son ortogonales los contrastes? ¿Es  $SC(Q_1) + SC(Q_2) = SC(\text{tratamientos})$ ?

#### 8.4 Prueba de efectos sugeridos por los datos

Cuando se conoce tan poco, respecto a la naturaleza de los tratamientos que se vacila en proponer comparaciones con sentido, entonces se necesitan técnicas para probar efectos

sugeridos por los datos. Tales técnicas se usan mucho para comparar todos los posibles pares de medias, aunque este conjunto puede considerarse como planeado. En todo caso, cuando la hipótesis nulas van a ser sugeridas por los datos o incluyen tanto que entran más del número de grados de libertad para los tratamientos, entonces se debe tener mucho cuidado en el procedimiento de prueba.

Quizás la comparación más obvia sugerida por los datos es la de la media máxima comparada con la mínima. Para esto, la distribución necesaria es la amplitud de las medias  $t$ , la cual ha sido "estudentizada" para tener en cuenta toda desviación estándar. Ver el encabezamiento de la tabla A.8.

Naturalmente si la diferencia máxima es significante y se ve que otras amplitudes dentro del conjunto superan el valor crítico utilizado, éstas también pueden considerarse como prueba de que las medias de población correspondientes no son homogéneas.

Una comparación sugerida por los datos supone el experimento como la unidad natural o conceptual básica. Esto lleva a una *tasa experimental de error* dada por la ec. (8.8) en experimentación repetida.

Tasa experimental de error ( $H_0$  verdadera)

$$= \frac{\text{número de experimentos con una inferencia errónea por lo menos}}{\text{número de experimentos realizados}} \quad (8.8)$$

Para un experimento en el cual se declara una diferencia significante, independientemente de cuantas diferencias más se encuentren, en el numerador de la ec. (8.8) va exactamente el valor uno.

Esta no es la única definición de tasa de error alterna a la definición por comparación. En particular, esa definición no debe confundirse con la tasa de error por experimento cuya definición es

$$\text{Razón de error por experimento } (H_0 \text{ verdadera}) = \frac{\text{número de inferencias erróneas}}{\text{número de experimentos}}$$

No usaremos esta tasa de error.

Para una tasa de error experimental fija, por ejemplo, con  $\alpha = 0.05$ , el valor crítico debe aumentar a medida que aumenta el número de tratamientos. Esto la hace prudente o cautelosa cuando la hipótesis nula es verdadera. También reduce la idoneidad de la prueba para detectar diferencias reales, y algunos investigadores sugieren que la tasa de error del 5 por ciento, común en pruebas con tasas de error por comparación, es demasiado cautelosa. En secciones posteriores se muestra como una u otra prueba merecen la misma crítica hasta cierto punto.

Una verdadera tasa experimental de error debe permitir claramente probar todas y cada una de las hipótesis. Por lo común, se desea probar sólo un subconjunto o *familia* de hipótesis nulas; el conjunto de todos los posibles pares de comparaciones es realmente una familia, cada tratamiento con control es otra, y un conjunto de comparaciones con sentido como el del experimento de plantas de menta también constituye una familia. Generalmente, una sola familia está asociada con un experimento.

Si nos limitamos a una familia de hipótesis dentro de las posibles, un valor crítico más pequeño es suficiente. Es importante, entonces, tener una definición más. Una *tasa de error por familia* se define como el valor al cual se aproxima

Tasa de error por familia ( $H_0$  verdadera)

$$= \frac{\text{número de familias con una inferencia errónea por lo menos}}{\text{número de familias probadas}} \quad (8.9)$$

De nuevo obsérvese que la tasa de error por familia se definiría de modo diferente.

En general, los procedimientos de comparaciones múltiples también tienen técnicas correspondientes para construir intervalos de confianza. Estas exigen que todos los enunciados sean simultáneamente correctos con un coeficiente de confianza, familiar o experimental, establecido. Una vez más la tasa de error se toma como la probabilidad de que por lo menos un enunciado componente sea falso.

### 8.5 Prueba de Scheffé

El método de Scheffé (8.12) es muy general en el sentido de que todas las posibles comparaciones pueden probarse en cuanto a significancia o bien pueden construirse intervalos de confianza para las correspondientes funciones lineales de parámetros. Esto quiere decir que son permisibles infinito número de pruebas simultáneas, aunque sólo se lleve a cabo un número finito, lo que da como resultado una tasa de error no mayor que la planeada; el conjunto de intervalos de confianza tendrá un coeficiente de confianza tan grande al menos como el dado.

Necesariamente, la prueba debe tener un valor crítico alto para toda comparación. Por tanto, es prudente en este sentido, y el poder puede ser bajo. Su uso parece más apropiado para "rastreo de datos" —busca de comparaciones sugeridas por los datos— como a menudo se hace en análisis de encuestas.

El valor crítico para una comparación,  $Q$ , exige el cálculo de  $S$

$$S = \sqrt{f_t F_\alpha(f_t, f_e)} \quad (8.10)$$

donde  $f_t$  y  $f_e$  son los grados de libertad para los tratamientos y el error experimental y  $F$  es el valor tabulado para una tasa de error de  $\alpha$ .

El valor crítico se calcula mediante

$$\text{Valor de Scheffé} = S s_Q \quad (8.11)$$

Para ilustrar el uso de esta prueba, considérese  $Q_1$  de la sección 8.3.  $Q_1 = -104.5$ , basado en totales. Cada total tiene 12 observaciones,  $f_t = 5$ ,  $f_e = 12$ ,  $s^2 = 2.15$  y  $\sum c_i^2 = 6$ . En consecuencia,  $s_Q = \sqrt{6(12)2.15}$ ; para  $\alpha = 0.05$ ,  $F = 3.11$ . Con las ecs. (8.10) y (8.11), encontramos

$$\begin{aligned} \text{Valor de Scheffé} &= \sqrt{5(3.11)6(12)2.15} \\ &= 49.06 \end{aligned}$$

Como el valor observado es numéricamente mayor, queda probado que la hipótesis nula es falsa.

Podemos proceder también a probar con  $\alpha = 0.05$ , hipótesis nulas con respecto a  $Q_2, \dots, Q_5$ , y cualesquiera otras funciones lineales que escojamos, posiblemente sólo porque los datos mismos sugieren esas funciones. El nivel  $\alpha$  de 0.05 se aplica a la totalidad de las hipótesis probadas. O sea que si las hipótesis nulas son verdaderas, entonces todas se aceptarán con probabilidad  $1 - \alpha = 0.95$ , o al menos una se rechazará erróneamente con probabilidad  $\alpha = 0.05$ .

Muchos estadísticos piensan que esta técnica sobreprotege contra el riesgo de cometer errores de tipo I, cuando se efectúan comparaciones ortogonales y no la recomiendan en este caso.

Para construir un intervalo de confianza para  $\mu_1 + \mu_2 + \mu_3 - \mu_4 - \mu_5 - \mu_6$ , son apropiadas las medias de modo que esta función lineal se estima mediante  $-104.5/12 = -8.71$ ; su varianza es  $(\sum c_i^2)s_y^2 = 6(2.15/12)$ ; y el intervalo de confianza es

$$Q(\text{usando las medias}) \pm S_{s_Q} = -8.71 \pm \sqrt{\frac{5(3.11)6(2.15)}{12}} = (-12.80, -4.62)$$

Por lo demás, éste y todos los posibles intervalos de confianza construidos con esta técnica deben incluir simultáneamente las correspondientes funciones lineales de los parámetros con una probabilidad de  $1 - \alpha = 0.95$  por lo menos, ya que solamente se construirá un número finito.

Como segunda y posiblemente más apropiada ilustración, compárense todos los posibles pares de medias de los datos de nitrógeno del trébol rojo, tabla 7.1, calculando el valor crítico  $S_{s_{\bar{Y}_t - \bar{Y}_r}}$ , aplicable directamente a diferencias entre medias. Para esto,  $f_t = t - 1 = 5$ ,  $f_r = 24$ ,  $r = 5$ , y  $s^2 = 11.79$ .

$$\begin{aligned} S_{s_{\bar{Y}_t - \bar{Y}_r}} &= \sqrt{f_t F_x(5, 24) s_{\bar{Y}_t - \bar{Y}_r}^2} \\ &= \sqrt{5(2.62)2(11.79)/5} \\ &= 7.9 \text{ for } \alpha = .05 \end{aligned}$$

La dms solamente fue igual a 4.5.

La presentación de los resultados de la prueba mediante subrayado es como sigue:

13.3	14.6	<u>18.7</u>	<u>19.9</u>	<u>24.0</u>	28.8
------	------	-------------	-------------	-------------	------

Para  $\alpha = 0.01$ , el valor crítico de Scheffé es 9.6 y la dms es 6.1.

**Ejercicio 8.5.2** Para cada una de las funciones lineales de  $\mu$  construidas usando los coeficientes mientas para los datos de la menta, calcular el valor crítico de Scheffé para probar la hipótesis nula correspondiente, tablas 7.8 y 8.2.

**Ejercicio 8.5.2** Para cada una de las funciones lineales de  $\mu$  contruidas usando los coeficientes de la tabla 8.2, calcular un intervalo de confianza por el método de Scheffé usando  $1 - \alpha = 0.95$  y también  $1 - \alpha = 0.99$ . ¿A qué se aplica el coeficiente de confianza? Cuando todos los supuestos necesarios para el modelo son ciertos y se calculan correctamente los cinco intervalos de confianza, ¿el coeficiente de confianza se precisamente igual a 0.95 o a 0.99?

**Ejercicio 8.5.3** Aplicar ambos procedimientos de prueba y de intervalo de confianza a los datos del ejercicio 7.3.1. Usese  $1 - \alpha = 0.95$ .

**Ejercicio 8.5.4** Refiérase a las comparaciones del ejercicio 8.3.6. Probar la hipótesis nula de que las medias de población de las comparaciones dadas son cero. Construir intervalos de confianza de las medias de población para esas comparaciones. Usar la técnica de Scheffé con  $\alpha = 0.05$ .

**Ejercicio 8.5.5** Repetir el ejercicio 8.5.4 con los datos del ejercicio 8.2.2 de erizos de mar.

## 8.6 Procedimiento $w$ de Tukey

El procedimiento de Tukey (8.13) hace uso de la amplitud "studentizada" y es aplicable a pares de medias; necesita de un solo valor para juzgar la significancia de todas las diferencias, y por lo tanto es rápido y fácil de usar. Ya que sólo se hacen comparaciones por pares, el valor crítico es menor que el exigido por el método de Scheffé. Todos los pares de medias constituyen una familia y la tasa de error es familiar, como lo es el coeficiente de confianza cuando se construyen estimaciones de intervalo de diferencias.

El procedimiento consiste en el cálculo de un valor crítico mediante la ec. (8.12) y su aplicación a diferencias entre todos los pares de medias.

$$w = q_\alpha(p, f_e)s_p \quad (8.12)$$

donde  $q_\alpha$  se obtiene de la tabla A.8,  $p = t$  es el número de tratamientos y  $f_e$  corresponde a los grados de libertad del error.

Para los datos de *Rhizobium*,  $p = 6$ ,  $f_e = 24$ ,  $q_{0.05} = 4.37$  y  $s_p = \sqrt{11.79/5} = 1.54$ . Por lo tanto,  $w = 4.37(1.54) = 6.7$  mg. Resumiendo los resultados de las pruebas mediante el subrayado se tiene

13.3	14.6	18.7	19.9	24.0	28.8
------	------	------	------	------	------

Para el método de Scheffé, el valor crítico es 7.9 mg, pero el resumen resulta el mismo. En general, no es así.

La tasa de error de  $\alpha = 0.05$  se aplica a la familia de todos los pares de comparaciones. Así, en experimentación repetida cuando todas las medias poblacionales son iguales, el 5 por ciento de las familias o conjuntos de diferencias tendrían una o más declaraciones de significancia falsas y el 95 por ciento de las familias, no se harían declaraciones de una diferencia significante.

El  $w$  de Tukey también puede usarse para calcular un conjunto de intervalos de confianza para las diferencias. La verdadera diferencia entre medias poblacionales estimadas por  $\bar{Y}_i$  y  $\bar{Y}_r$  se estima mediante

$$IC = \bar{Y}_i - \bar{Y}_r \pm w \quad \text{para } w \text{ como aparece en la ec. (8.13)} \quad (8.13)$$

El coeficiente de confianza se aplica a la familia de intervalos. Con una probabilidad de  $1 - \alpha$ , todos los intervalos contienen la  $\mu_i - \mu_r$  que se estima; con probabilidad  $\alpha$ , por lo menos uno no la contiene.

El procedimiento también puede generalizarse para comparaciones lineales.

**Ejercicio 8.6.1** Calcular intervalos de confianza para diferencias por pares entre medias con los datos de *Rhizobium*. Usar  $1 - \alpha = 0.95$  y  $0.99$  con el procedimiento de Tukey.

**Ejercicio 8.6.2** Probar todos los pares de medias con los datos de las plantas de menta, tabla 7.8, usando el procedimiento w.d de Tukey y un  $\alpha = 0.05$ .

Cuando se llega a la extracción de la información de los datos, ¿qué puede pensar de este conjunto de comparaciones frente al conjunto propuesto en la sec. 8.3?

**Ejercicio 8.6.3** Con los datos de ostras del ejercicio 8.2.1, aplicar el procedimiento de Tukey para probar todos los pares de medias. Construya intervalos de confianza de las diferencias entre todas las medias de la población.

**Ejercicio 8.6.4** Repetir el ejercicio 8.6.3 con los datos del ejercicio 8.2.3 sobre erizos de mar.

## 8.7 Prueba de Student-Newman-Keuls o S-N-K

Cada una de las tres personas mencionadas contribuyeron al desarrollo de esta prueba, llamada también prueba de Newman-Keuls (8.11), o simplemente método de Keul. No es tan prudente como la prueba de Tukey; usa amplitudes múltiples para probar. Sin embargo, esto lo hace un procedimiento guiado por los resultados, lo cual hace difícil la descripción de la tasa de error. El lector interesado puede remitirse a otras fuentes, como Hartley (8.15). Los intervalos de confianza no son apropiados.

Designase un conjunto de medias de tratamientos en orden de magnitud así  $\bar{Y}_{(1)}, \dots, \bar{Y}_{(t)}$ , siendo  $\bar{Y}_{(1)}$  la media menor. El procedimiento de prueba S-N-K se inicia como la prueba de Tukey, comparando las medias máxima y mínima. Si la amplitud no es significante, no se hacen más pruebas y se declara homogéneo el conjunto de medias. Si esta diferencia máxima se declara significante, se concluye que  $\mu_{(1)} \neq \mu_{(t)}$  y se continúa comprobando como si esto fuera un hecho. O sea que, en la etapa siguiente, se prueba  $\bar{Y}_{(1)} \text{ vs. } \bar{Y}_{(t-1)}$  y  $\bar{Y}_{(2)} \text{ vs. } \bar{Y}_{(t)}$ , usando un criterio de prueba para  $t - 1$  medias. En cualquier etapa, donde una diferencia no sea significante, la prueba se detiene y se declara el conjunto homogéneo. De otra manera se continúa con la prueba.

El procedimiento consiste en calcular un conjunto de valores críticos mediante

$$W_p = q_\alpha(p, f_e) s_p \quad p = t, t-1, \dots, 2 \quad (8.14)$$

donde  $q_\alpha$  se toma de la tabla A.8,  $p = t$  = el número total de tratamientos, y  $f_e$  = grados de libertad del error. Para los datos de *Rhizobium* esos valores son los siguientes.

$p$	2	3	4	5	6
$q_{0.05}(p, 24)$	2.92	3.53	3.90	4.17	4.37
$W_p$	4.5	5.4	6.0	6.4	6.7

Obsérvese que  $W_6$  es el único valor usado en el método de Tukey.

El siguiente es el resumen de los resultados usando subrayados.

<u>13.3</u>	<u>14.6</u>	<u>18.7</u>	19.9	24.0	28.8
-------------	-------------	-------------	------	------	------

Resultan más declaraciones de significancia que con el método de Tukey pero menos que con la dms.

Ejercicio 8.7.1 Aplicar el procedimiento de S-N-K al problema de probar todos los pares de medias para los datos de ostras el ejercicio 8.2.1.

Ejercicio 8.7.2 Repetir el ejercicio 8.7.1 para los datos de los erizos marinos del ejercicio 8.2.2.

Ejercicio 8.7.3 Aplicar el método de S-N-K a los datos de menta, tabla 7.8. Para esos tratamientos, ¿se puede pensar que la prueba de pares de medias es tan informativa como la prueba de conjunto de comparaciones ortogonales con sentido dado en la sección 8.3?

## 8.8 Nueva prueba de amplitud múltiple de Duncan

En 1955, Duncan (8.6) desarrolló una *nueva prueba de amplitud múltiple*, que aunque no es tan potente como una anterior (8.5), tiene la ventaja de su sencillez.

Se parece a la prueba de S-N-K en cuanto usa amplitudes múltiples y está guiada por los resultados. Sin embargo, es menos prudente. Los intervalos de confianza no son apropiados; la noción de confianza se remplaza por la de *niveles de protección* ante el riesgo de hallar diferencias significantes erróneas en etapas de la prueba. La profundización de esta idea se deja al lector. La prueba es bastante popular. Desafortunadamente se ha usado en casos en los que parecía apropiado utilizar comparaciones planeadas.

La prueba se aparta del procedimiento S-N-K en el cual se emplea un nivel de significancia constante en todas las etapas de la prueba. Utiliza un nivel de significancia variable que depende del número de medias que entran en una etapa. La idea es que a medida que el número de medias que se prueban aumenta, menor es la probabilidad de que se asemejen. Si  $t = 2$  medias, entonces úsese algún  $\alpha$  en general aceptable como 0.05. Sin embargo, para tres medias,  $1 - (1 - \alpha)^2 = 0.0975$ , se sugiere usar  $\alpha = 0.05$ ; para cuatro medias, úsese  $1 - (1 - \alpha)^3 = 0.14$ ; ... ; para  $t$  medias, úsese  $1 - (1 - \alpha)^{t-1}$ . La tabla A.7 se construye de acuerdo con eso a partir de la distribución de amplitud "estudiantizada"

El procedimiento de prueba se ilustra con los datos de *Rhizobium*, tabla 7.1, y discurre tal como la prueba S-N-K, pero usando la tabla A.7. Se comienza por el cálculo de las *amplitudes mínimas significantes de R<sub>p</sub>*, mediante

$$R_p = q_{\alpha} s_p \quad \alpha' = 1 - (1 - \alpha)^{p-1} \quad p = 2, 3, \dots, t \quad (8.15)$$

$p$	2	3	4	5	6
$q_{\alpha}(6, 24)$	2.92	3.07	3.15	3.22	3.28
$R_p$	4.5	4.7	4.9	5.0	5.1

Un resumen de los resultados utilizando subrayados es el siguiente

<u>13.3</u>	<u>14.6</u>	<u>18.7</u>	<u>19.7</u>	<u>24.0</u>	<u>28.8</u>
-------------	-------------	-------------	-------------	-------------	-------------

Se ve que los resultados son los mismos cuando se usa la dms, si bien no siempre es así. Obsérvese que para  $p = 2$ , el valor crítico es el mismo que para la prueba dms y el S-N-K.

Ejercicio 8.8.1 Aplique la prueba nueva de amplitud múltiple de Duncan al problema de todos los pares de medias con los datos de ostras del ejercicio 8.2.1.

Ejercicio 8.8.2 Repetir el ejercicio 8.8.1 con los datos de erizos marinos del ejercicio 8.2.2.

Ejercicio 8.8.3 Aplicar la prueba de Duncan a las medias de datos de menta, tabla 7.8. Para esos tratamientos, ¿se puede pensar que el procedimiento de Duncan produce tanta información como el de hacer el conjunto de comparaciones con sentido en la sec. 8.3?

Ejercicio 8.8.4 Aplicar la prueba de Duncan, haciendo comparaciones por pares de las medias calculadas en el ejercicio 7.3.1.

## 8.9 Comparación de todas las medias con un control

El objetivo de un experimento es a veces localizar tratamientos que son diferentes o mejores que un tratamiento estándar, pero sin compararlos; la comparación se deja para un experimento posterior. Una familia de comparaciones semejante, esto es, control con cada tratamiento, no es un conjunto independiente, pero sí merece consideración especial. Al restringir las comparaciones a esta familia, la técnica resultante exige un valor crítico menor que la prueba de Tukey.

El procedimiento de Dunnett (8.8) requiere de un solo valor para juzgar la significancia de las diferencias observadas entre cada tratamiento y el control. Se dispone de la tabla A.9 para comparaciones con alternativas unilaterales y bilaterales. La tasa de error es familiar y pueden construirse intervalos de confianza. Ahora el procedimiento se aplica a los datos de *Rhizobium* con el compuesto como patrón de comparación o control.

Considérense las medias de los datos de *Rhizobium* como  $p$  medias de tratamiento y un control con  $p$  comparaciones que han de efectuarse. Se acude a la tabla A.9 para el conjunto apropiado de alternativas y la tasa de error, con  $p$  tratamientos y  $f_e$  grados de libertad para el error, y se obtiene un valor  $t$  de Dunnett. Para la ilustración, cada comparación va a ser de dos colas con  $\alpha = 0.05$ ,  $p = 5$ , y  $f_e = 24$ . El valor crítico está dado por

$$\begin{aligned} d' &= t(\text{Dunnett})s_{\bar{Y}_1 - \bar{Y}_2} \\ &= 2.76\sqrt{2(11.79)/5} = 5.99 \text{ mg}, \end{aligned} \quad (8.16)$$

para el ejemplo. La única diferencia que debe declararse significante es la existente entre 3DOk1 y el compuesto. Los procedimientos de Tukey, S-N-K y Duncan también declararon significante esa diferencia, si bien cada uno requirió de un valor crítico diferente para juzgar la significancia. El método S-N-K declara significante la diferencia entre 3DOk13 y

el compuesto cuando sólo se tiene una cifra decimal. El método de Duncan también declara significante las diferencias entre 3DOk13 y el compuesto y entre 3DOk5 y el compuesto; éstas resultan muy cercanas a la significancia por el procedimiento de Dunnett. El valor crítico de Dunnett es menor que el del procedimiento de Tukey, en el que la familia contiene más comparaciones.

Obsérvese que en la ec. (8.16) se necesita  $s_{Y_i - Y_j}$ , así que se requiere un  $\sqrt{2}$ . Las pruebas basadas en la distribución de la amplitud exigen  $s_Y$ .

Las ecuaciones (8.17) y (8.18) dan intervalos de confianza simultáneos, es decir, intervalos de confianza incluidos simultáneamente en un solo coeficiente de confianza, para verdaderas diferencias,  $\mu_i - \mu_0$

$$IC = (\bar{Y}_i - \bar{Y}_0) \pm ts \sqrt{\frac{2}{r}} \quad (8.17)$$

donde  $\bar{Y}_0$  es la media del control y  $t$  viene de la tabla de Dunnett para comparaciones bilaterales, para un intervalo de dos lados.

$$IC = (\bar{Y}_i - \bar{Y}_0) - ts \sqrt{\frac{2}{r}}, \infty \quad (8.18)$$

donde  $t$  se encuentra en la tabla de comparaciones unilaterales, tabla A.9A. Para un intervalo unilateral, establecemos que el verdadero  $\mu_i - \mu_0$  es por lo menos tan grande como el valor calculado por la ec. (8.18). El símbolo  $\infty$  indica que no existe límite por la derecha. Cuando valores más pequeños que el control representan mejor desempeño, remplácese  $\bar{Y}_i - \bar{Y}_0$  por  $\bar{Y}_0 - \bar{Y}_i$  para tener una mejor presentación de los resultados.

El coeficiente de confianza cubre la familia de enunciados. Todos los intervalos contienen simultáneamente las correspondientes diferencias verdaderas con probabilidad  $1 - \alpha$ .

**Ejercicio 8.9.1** Aplicar la prueba de Dunnett a los datos del ejercicio 7.3.1. Probar con alternativas bilaterales. Usando la ec. (8.17), constrúyase un conjunto de intervalos de confianza simultáneo.

**Ejercicio 8.9.2** Suponiendo que el investigador pudiera haber esperado una respuesta decreciente debido a los tratamientos del ejercicio 7.3.1, aplicar el procedimiento de la prueba de Dunnett frente a alternativas unilaterales. Construir un conjunto de intervalos de confianza simultáneos para todas las diferencias entre tratamientos y control. Para la construcción, ¿qué se diría en cuanto a dónde quedan las verdaderas diferencias con respecto a los valores calculados?

**Ejercicio 8.9.3** Aplicar el procedimiento de Dunnett a los datos del ejercicio 7.3.3. Usar el tratamiento 1 como "control" de Dunnett. El tratamiento 1 se compara con dos tratamientos, 2 y 3, descritos por Jordan (7.12) como controles. Como el tratamiento 1 buscaba reducir las puntuaciones de T, estudiar una prueba apropiada unilateral y construir intervalos de confianza unilaterales. Usar  $\alpha = 0.05$ ,  $1 - \alpha = 0.95$ .

**Ejercicio 8.9.4** Aplicar el procedimiento de prueba de Dunnett a los datos de la menta, tabla 7.8. Usar "temperatura baja, 8h" como control. Presumiblemente, un investigador familiarizado con el material esperaría que este tratamiento diera la respuesta menor. Efectuar pruebas unilaterales con  $\alpha = 0.05$  y construir intervalos de confianza con  $1 - \alpha = 0.95$ .

### 8.10 Prueba de $t$ de razón $k$ bayesiana de Waller-Duncan

El interés de Duncan (8.7) en procedimientos de comparaciones múltiples ha continuado con varios enfoques. La prueba en esta sección es la que se espera que reemplazará a otra usada por pares de comparaciones. Una presentación completa de este procedimiento y solución exige esfuerzo considerable, así que sólo podemos dar un tratamiento somero. Se pueden calcular estimaciones de intervalos, pero no se ilustran aquí; ver la referencia 8.4.

Ante todo, no interviene un nivel de significancia, en lugar de esto, se escoge una *gravedad de error* o *peso de error de tipo I o tipo II*. Esto es difícil para muchos de nosotros; se nos aconseja que razones  $k$  de 50:1, 100:1 y 500:1, pueden considerarse en sentido amplio, en vez de  $\alpha = 0.10$ , y 0.05 y 0.01. La tabla A.21 contiene valores  $t$  de riesgo promedio mínimo para  $k = 100$ . Estos son los únicos de estos valores presentados aquí.

La tabla se consulta con base en el valor para el contraste  $F$  de tratamientos. Para ilustración referirse a los datos de *Rhizobium* de la tabla 7.1, donde el valor  $F$  es 14.37. Casi con seguridad, se requerirá interpolación; aquí, se dan los  $F$  tabulados de 10.00 y 25.0. A estas partes de la tabla se va con  $q = f_t = gl$  de tratamientos y  $f = f_e gl$  de error;  $q = 5$  y  $f = 24$ . Los valores  $t$  de riesgo promedio mínimo tabulados son 1.96 para  $F = 10$  y 1.88 para  $F = 25$ . La nota al pie de la tabla nos dice que interpolemos con respecto a  $b$ , un valor dado con  $F$ ; para  $F = 10$ ,  $b = 1.054$ , y para  $F = 25$ ,  $b = 1.021$ . La muestra o experimento  $b$  es  $b = [F/(F - 1)]^{1/2} = (14.37/13.37)^{1/2} = 1.037$ . La interpolación para obtener el  $t$  necesario exige que  $1.96 - 1.88$  sea a  $1.054 - 1.021$  como  $1.96 - t$  es a  $1.054 - 1.037$ , o bien

$$t = 1.96 - \frac{1.96 - 1.88}{1.054 - 1.021} (1.054 - 1.037) = 1.92$$

El valor crítico o DMS está dado por la ec. (8.19). Obsérvese el uso de mayúsculas para diferenciarlo de dms de la sec. 8.2.

$$\begin{aligned} \text{DMS} &= ts_{\bar{Y}_1 - \bar{Y}_2} \\ &= 1.92 \sqrt{2(11.79)/5} = 4.2 \text{ mg} \end{aligned} \quad (8.19)$$

para el ejemplo. La dms (0.05) fue de 4.5 mg.

Al resumir los resultados por subrayado se tiene

<u>43.3</u>	<u>14.6</u>	<u>18.7</u>	<u>19.9</u>	<u>24.0</u>	<u>28.8</u>
-------------	-------------	-------------	-------------	-------------	-------------

Se ve que los resultados son los mismos que los obtenidos con la nueva prueba de amplitud múltiple de Duncan. No tenía que ser así. En el procedimiento de Waller-Duncan (8.14), no se aplica nivel de significancia.

Ejercicio 8.10.1 Aplicar el procedimiento de Waller-Duncan para probar las medias pareadas de los datos del ejercicio 8.2.1.

**Ejercicio 8.10.2** Aplicar el procedimiento de Waller-Duncan para probar medias pareadas de los datos de erizos de mar del ejercicio 8.2.2.

**Ejercicio 8.10.3** Aplicar el procedimiento de Waller-Duncan para probar medias pareadas de los datos de pesos del ejercicio 7.3.1.

**Ejercicio 8.10.4** Aplicar el procedimiento de Waller-Duncan para probar medias pareadas de los datos de la menta, tabla 7.8.

**Ejercicio 8.10.5** Revisar los procedimientos de las pruebas de comparaciones múltiples para comparar todos los pares de medias. Comparar las conclusiones para los conjuntos de datos donde se ha usado más de un procedimiento.

## 8.11 Pruebas de medias con número desigual de repeticiones

El problema de probar comparaciones con un solo grado de libertad se hace más incómodo cuando hay repetición variable. Si se desea hacer tales pruebas, se sugieren los siguientes procedimientos, sobre todo para pares de medias.

*Para la dms* Calcular  $s_{\bar{y}_i - \bar{y}_{i'}}$  como

$$s_{\bar{y}_i - \bar{y}_{i'}} = \sqrt{s^2 \left( \frac{1}{r_i} + \frac{1}{r_{i'}} \right)} = \sqrt{s^2 \frac{r_i + r_{i'}}{r_i r_{i'}}} \quad (8.20)$$

Ahora la  $dms = ts_{\bar{y}_i - \bar{y}_{i'}}$ , donde  $t$  es la  $t$  de Student para el nivel de significancia escogido y los grados de libertad del error, y,  $r_i$  y  $r_{i'}$ , son el número de observaciones en las dos medias que se comparan. También puede usarse  $s_{\bar{y}_i - \bar{y}_{i'}}$  para calcular intervalos de confianza. El procedimiento es estadísticamente correcto.

*Para contrastes* Sea  $Y_i$ , el total del tratamiento  $i$ -ésimo con  $r_i$  observaciones. Un contraste, utilizando totales, se define por

$$Q = \sum c_i Y_i = \sum r_i c_i \bar{Y}_i \quad \text{con} \quad \sum r_i c_i = 0 \quad (8.21)$$

Ahora  $E(Q) = \sum r_i c_i \mu_i$  y  $\sigma_Q^2 = (\sum r_i c_i^2) \sigma^2$ . Para probar  $H_0: \sum r_i c_i \mu_i = 0$ , úsese

$$t = \frac{Q}{s_Q} = \frac{Q}{s \sqrt{\sum r_i c_i^2}} \quad (8.22)$$

Cuando se usa  $F$ ,  $Q^2$  exige un divisor de  $\sum r_i c_i^2$  para que sea una base por observación  
Dos contrastes  $\sum c_{1i} Y_i$  y  $\sum c_{2i} Y_i$  se dicen ortogonales si

$$\sum r_i c_{1i} c_{2i} = 0 \quad (8.23)$$

*Para la prueba de Scheffé* No se necesita tratamiento especial porque fue definida para comparaciones y éstos acaban de generalizarse precisamente.

*Para el procedimiento w de Tukey* Obtener un valor  $w' = q_\alpha(p, f_e)s$  como en la ec. (8.12), pero remplazando  $s_y$  por  $s$ . Para cualquier comparación que se desee hacer, multiplicar  $w'$  por el valor

$$\sqrt{\frac{1}{2} \left( \frac{1}{r_i} + \frac{1}{r_j} \right)} \quad (8.24)$$

La validez de este procedimiento no ha sido verificada.

*Para la prueba de S-N-K* Obtener valores  $W'_p = q_\alpha(p, f_e)s$  como en la ec. (8.14). Para cualquier comparación deseada, multiplicar por el valor apropiado dado por la ec. (8.24). La validez de este procedimiento no ha sido verificada aún.

*Para la nueva prueba de amplitud múltiple de Duncan* Obtener amplitudes "estudentizadas" significantes y multiplicar por  $s$  en vez de  $s_y$  para lograr un conjunto de amplitudes intermedias significantes. Para cualquier comparación deseada, multiplíquese el valor intermedio apropiado por la cantidad necesaria calculada mediante la ec. (8.24).

Kramer (8.9) ha propuesto este procedimiento para tratamientos con un número desigual de repeticiones, su validez no ha sido verificada.

*Para el procedimiento de Dunnett* Obtener los  $t$  de la tabla de Dunnett. Para cualquier comparación deseada, multiplíquese por

$$\sqrt{\frac{1}{r_i} + \frac{1}{r_j}}$$

La validez de este procedimiento no ha sido verificada.

**Ejercicio 8.11.1** Hacer comparaciones por pares de medias de los datos de la prueba posterior en el ejercicio 7.4.2 (incluyendo datos del ejercicio 7.3.3) usando

- a) La dms
- b) La prueba de Scheffé
- c) La prueba de Tukey
- d) La prueba de S-N-K
- e) La prueba de Duncan
- f) La prueba de Dunnett (usando el tratamiento 1 como control y probar con alternativas unilaterales apropiadas, que exigen una puntuación reducida con 1).

**Ejercicio 8.11.2** Hacer comparaciones por pares de medias de los datos del ejercicio 7.4.1 usando los procedimientos enumerados en el ejercicio 8.11.1.

Para la prueba de Dunnett, usar  $F_1$  como control y probar con las alternativas unilaterales apropiadas.

**Ejercicio 8.11.3** Supóngase que podemos tomar 20 observaciones para probar la hipótesis nula de que no hay diferencia entre las medias de dos poblaciones con varianza común.

¿Cuál será la varianza de la diferencia entre las medias si tomamos 10 observaciones de cada población? ¿Si tomamos 12 de la primera y 8 de la segunda? ¿15 y 5 respectivamente? ¿18 y 2 respectivamente? ¿Cuántos grados de libertad se tienen para estimar  $\sigma^2$  en cada caso?

¿Qué conclusión se puede sacar de los cálculos? [Este razonamiento relacionado con la mejor asignación de observaciones no se aplica al problema de comparar tratamientos con control. Ver Dunnett (8.8)].

## Referencias

- 8.1. Carmer, S. G., y M. R. Swanson: "An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods," *J. Amer. Statist. Ass.*, 68:66-74 (1973).
- 8.2. Carmer, S. G., y M. R. Swanson: "Detection of differences between means: A Monte Carlo study of five pairwise multiple comparison procedures," *Agron. J.* 63:940-945 (1971).
- 8.3. Cochran, William G., y Gertrude M. Cox: *Experimental designs*, 2a. ed., Wiley, Nueva York, 1957.
- 8.4. Dixon, Dennis O., y David B. Duncan. "Minimum Bayes risk *t*-intervals for multiple comparisons," *J. Amer. Statist. Ass.*, 70:822-831 (1975).
- 8.5. Duncan, D. B.: "A significance test for differences between ranked treatments in an analysis of variance," *Va. J. Sci.*, 2:171-189 (1951).
- 8.6. Duncan, D. B.: "Multiple range and multiple *F* tests," *Biom.*, 11:1-42 (1955).
- 8.7. Duncan, D. B.: "T tests and intervals for comparisons suggested by the data," *Biom.*, 31:339-359(1975).
- 8.8. Dunnett, C. W.: "A multiple comparisons procedure for comparing several treatments with a control," *J. Amer. Statist. Ass.*, 50:1096-1121 (1955).
- 8.9. Kramer, C. Y.: "Extension of multiple range tests to group means with unequal numbers of replication," *Biom.*, 12:307-310 (1956).
- 8.10. McDonald, Michael E.: "The standing crop, distribution, and production of the macrobenthic epifauna on an artificial reef off the coast of North Carolina," Tesis de maestría, North Carolina State University, Raleigh, NC, 1978.
- 8.11. Newman, D.: "The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation," *Biometrika*, 31:20-30 (1939).
- 8.12. Scheffé, Henry: *The analysis of variance*, Wiley, Nueva York, 1959.
- 8.13. Tukey, J. W.: "The problem of multiple comparisons," Princeton University, Princeton, NJ, 1953 (Ditto.).
- 8.14. Waller, Roy A., y David B. Duncan: "A Bayes rule for the symmetric multiple comparisons problem," *J. Amer. Statist. Ass.*, 64:1484-1503 (1969).
- 8.15. Hartley, H. O.: "Some recent developments in analysis of variance," *Comm. Pure Appl. Math.*, 8:47-72 (1955).

---

## CAPITULO NUEVE

---

### ANALISIS DE LA VARIANZA II: CLASIFICACIONES MULTIPLES

#### 9.1 Introducción

El diseño completamente aleatorio es apropiado cuando se sabe que no hay fuentes de variación fuera de los efectos de tratamiento. En muchas situaciones se sabe de antemano que ciertas unidades experimentales, si se tratan de modo parecido se comportan en forma diferente. Por ejemplo, en experimentos sobre el terreno, las parcelas adyacentes suelen responder en forma más similar que las que están a cierta distancia; de la misma manera, los animales más pesados entre un grupo de la misma edad pueden presentar una tasa de crecimiento diferente que animales más livianos; también, las observaciones efectuadas en un día dado o usando cierto equipo pueden parecerse más que las hechas en días diferentes con diferentes equipos. En tales situaciones, cuando puede anticiparse en parte el comportamiento de unidades individuales y por consiguiente clasificarlas, pueden construirse diseños o planes de tal modo que la parte de la variabilidad atribuible a una fuente reconocida pueda medirse y excluirse así del error experimental; al mismo tiempo las diferencias entre medias de tratamientos no contendrán contribución alguna atribuible a la fuente reconocida. En el cap. 5, se usó este principio en la comparación de dos tratamientos con observaciones pareadas.

Este capítulo trata del análisis de la varianza cuando hay dos o más criterios de clasificación. Se dan los análisis del bloque completo aleatorizado y del cuadrado latino. Además, se define la interacción y se estudia el uso de transformaciones.

#### 9.2 El diseño de bloque completo al azar

Este diseño puede usarse cuando las unidades experimentales pueden agruparse; generalmente el número de unidades por grupo es igual al número de tratamientos. Tal tipo de grupo se llama *bloque* o *repetición*. El objetivo del agrupamiento es lograr que las unidades en un bloque sean tan uniformes como sea posible, de modo que las diferencias obser-

vadas se deban en gran parte a los tratamientos. En promedio, la variabilidad entre unidades de diferentes bloques será mayor que la variabilidad entre unidades del mismo bloque si no van a aplicarse tratamientos. Idealmente, la variabilidad entre unidades experimentales se controla de tal forma que se maximice la variación entre bloques, mientras que la variación dentro de ellos se minimice. La variación entre bloques no afecta claramente a las diferencias entre medias de tratamientos, ya que cada tratamiento aparece el mismo número de veces en cada bloque.

En experimentos sobre el terreno, usualmente cada bloque consiste en un grupo compacto de parcelas aproximadamente cuadradas. De igual manera, en muchos experimentos con animales, los animales se colocan en grupos de resultados o bloques con base en características tales como peso inicial, condición del animal, raza, sexo o edad, o como etapa de lactancia y producción de leche en el ganado, y como camadas en cerdos.

Durante el transcurso del experimento, todas las unidades de un bloque deben tratarse tan uniformemente como sea posible en todo aspecto diferente del tratamiento. Todo cambio en la técnica u otra condición que pueda afectar los resultados debe hacerse en todo el bloque. Por ejemplo, si la cosecha de las parcelas de terreno se extiende en un período de tiempo, todas las parcelas de un bloque deberán cosecharse el mismo día. Así mismo, si personas diferentes hacen observaciones en el material experimental, y si existe alguna posibilidad de que las observaciones hechas en la misma parcela difiera de una persona a otra, y si sólo se ha de hacer una observación en cada unidad experimental, entonces una persona debe hacer todas las observaciones en un bloque dado. Por otra parte, si el número de observaciones por unidad es igual al número de observadores, entonces cada observador debe hacer una observación por unidad. Estas prácticas ayudan a controlar la variación dentro de bloques y por tanto el error experimental; al mismo tiempo, no contribuyen en nada a las diferencias entre medias de tratamiento. La variación entre bloques se elimina aritméticamente del error experimental.

Obsérvese el balance que existe en este diseño. Cada observación se clasifica de acuerdo con el bloque que contiene la unidad experimental y al tratamiento aplicado, dando lugar a una clasificación de dos vías. Cada tratamiento aparece un número igual de veces, usualmente una vez, en cada bloque y cada bloque contiene todos los tratamientos. Bloques y tratamientos son ortogonales entre sí. Esta propiedad es la que lleva a los sencillos cálculos aritméticos que entran en el análisis de los datos resultantes. Este diseño se usa con mayor frecuencia que cualquier otro, y si da precisión satisfactoria, no hay objeto en usar otro diferente.

**Aleatorización** Cuando se han asignado las unidades experimentales a los bloques, se numeran en cierto orden conveniente. Los tratamientos también se numeran y luego se asignan aleatoriamente a las unidades dentro de un bloque. Una nueva aleatorización se efectúa en cada bloque.

El procedimiento puede ser alguno de los dados en la sec. 7.2. Por ejemplo, si tenemos ocho tratamientos, obtenemos 8 números de tres dígitos y observamos sus rangos. Los rangos se consideran como permutaciones aleatorias de los números 1, ..., 8. Por ejemplo, podemos obtener, 1, 8, 7, 5, 4, 6, 2 y 3 como la permutación aleatoria. Ahora el tratamiento 1 se aplica a la unidad 1, el tratamiento 8 a la unidad 2, ..., y el tratamiento 3 a la unidad 8. O bien, podríamos aplicar el tratamiento 1 a la unidad 1, el tratamien-

to 2 a la unidad 8, ..., y el tratamiento 8 a la unidad 3. Los otros procedimientos de aleatorización de la sec. 7.2 se generalizan con igual facilidad.

El diseño de bloque completo al azar tiene muchas ventajas sobre otros diseños. En general, es posible agrupar las unidades experimentales de modo que se logre mayor precisión que con el diseño completamente aleatorizado. No hay restricción en cuanto al número de tratamientos o de bloques. Si se desea usar repeticiones adicionales para ciertos tratamientos, éstos se pueden aplicar a dos o más unidades por bloque con aleatorización adecuada para dar un diseño de bloque completo al azar generalizado. También son posibles diseños que incluyan tratamientos no repetidos; ver Federer (9.26). El análisis estadístico de los datos es simple. Si, como resultado de un contratiempo, los datos de un bloque completo para ciertos tratamientos son inutilizables, estos datos pueden omitirse sin complicación en el análisis. Si faltan datos de unidades individuales, pueden estimarse fácilmente de tal manera que no se pierda la comodidad en los cálculos. Si el error experimental es heterogéneo, pueden obtenerse componentes no sesgadas aplicables a la prueba de comparaciones específicas.

La principal desventaja de los bloques completos al azar es que cuando la variación entre unidades experimentales dentro de un bloque es grande, resulta un término de error considerable. Esto ocurre frecuentemente cuando el número de tratamientos es grande; así puede no ser posible asegurar grupos de unidades suficientemente uniformes para los bloques. En tales situaciones, se dispone de otros diseños para controlar una mayor proporción de la variación.

### 9.3 Análisis de la varianza para cualquier número de tratamientos; diseño de bloque completo al azar

En la tabla 9.1 se da un resumen simbólico de las fórmulas de definición y operación para las sumas de cuadrados y grados de libertad en el análisis de la varianza de datos del diseño de bloques completos al azar.

Sea  $Y_{ij}$  la observación de  $j$ -ésimo bloque bajo el tratamiento  $i$ -ésimo,  $i = 1, 2, \dots, t$  tratamientos y  $j = 1, 2, \dots, r$  bloques. La notación de puntos se usa siempre que sea posible. Así,  $\sum_j Y_{ij}^2$  quiere decir que se obtienen las sumas

$$Y_j = \sum_i Y_{ij}$$

para cada valor de  $j$ , se elevan al cuadrado y se suman para todos los valores de  $j$ . Representese la media general por  $\bar{Y}_{..}$ . Como la varianza de medias de  $n$  observaciones es  $\sigma^2/n$ , resultan los multiplicadores de  $t$  y  $r$  que aparecen en la columna SC de definición en todos los cuadrados medios que estiman la misma  $\sigma^2$  cuando no hay efectos de bloques o tratamientos. Con el mismo razonamiento, basado en totales, se explican los divisores  $t$  y  $r$  en la columna de operación de las SC.

La tabla 9.2 da en porcentajes los contenidos de aceite de semillas de lino para parcelas localizadas en Winnipeg, e inoculadas usando varias técnicas, con suspensiones de esporas de *Septoria linicola*, el organismo que causa pasmo en el lino. Los datos fueron reportados por Sackston y Carson (9.16). Los datos originales han sido codificados restando 30 de cada observación. Puede codificarse todavía más, multiplicando cada número por 10

**Tabla 9.1 Fórmulas para el análisis de la varianza de  $t$  tratamientos organizados en un diseño de bloque completo al azar de  $r$  bloques**

Fuente de variación	gl	Sumas de cuadrados	
		Definición	Operación
Bloques	$r - 1$	$t \sum_j (Y_{ij} - \bar{Y}_j)^2 = \frac{\sum_i Y_{ij}^2}{t} - C$	$\sum_i Y_{ij}^2$
Tratamientos	$t - 1$	$r \sum_i (Y_{i.} - \bar{Y}_{..})^2 = \frac{\sum_i Y_{i.}^2}{r} - C$	$\sum_i Y_{i.}^2$
Error	$(r - 1)(t - 1)$	$\sum_{i,j} (Y_{ij} - \bar{Y}_j - \bar{Y}_{i.} + \bar{Y}_{..})^2 = SC(\text{total}) - SC(\text{bloques}) - SC(\text{tratamientos})$	
Total	$rt - 1$	$\sum_{i,j} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i,j} Y_{ij}^2 - C$	

para eliminar los puntos decimales; pero la codificación sin multiplicación evita la decodificación de las varianzas o de las desviaciones estándar. Obsérvese que da una comprobación del cálculo de las sumas de cuadrados total.

En detalle, los cálculos son como sigue

*Paso 1* Disponer los datos como se hace en la tabla 9.2. Obtener los totales de tratamientos  $Y_{i.}$ , totales de bloques  $Y_{j.}$ , y el gran total  $Y_{..}$ . Simultáneamente, obténgase  $\sum Y^2$  cada tratamiento y bloque, es decir,

$$\sum_j Y_{ij}^2 \quad i = 1, \dots, t$$

$$\sum_i Y_{ij}^2 \quad j = 1, \dots, r$$

Obténgase el gran total sumando los totales de tratamientos y los totales de bloque, separadamente. Al mismo tiempo, obténganse las sumas de cuadrados de esos totales. Estas no se indican en la tabla 9.2, donde 788.23 es

$$\sum_{i,j} Y_{ij}^2$$

*Paso 2* Obténganse las sumas de cuadrados (ajustadas) de la forma siguiente

$$\begin{aligned} \text{Factor de corrección } &= C = \frac{Y_{..}^2}{rt} \\ &= \frac{(132.7)^2}{24} = 733.72 \end{aligned} \tag{9.1}$$

**Tabla 9.2 Contenido de aceite de semillas de lino Redwing inoculadas en diferentes estados de crecimiento con *S. Linicola*, Winnipeg, 1947, en porcentajes**  
 Observación original = 30 + observación tabulada

Tratamiento (estado en que se inocula)	Bloques				Totales de tratamientos			medias de tratamientos descodificados
	1	2	3	4	$\bar{Y}_i$	$\sum_j Y_{ij}^2$		
Plántula	4.4	5.9	6.0	4.1	20.4	106.98	35.1	
Florecimiento temprano	3.3	1.9	4.9	7.1	17.2	88.92	34.3	
Florecimiento completo	4.4	4.0	4.5	3.1	16.0	65.22	34.0	
Florecimiento completo (1/100)	6.8	6.6	7.0	6.4	26.8	179.76	36.7	
Maduración	6.3	4.9	5.9	7.1	24.2	148.92	36.0	
Sin inocular	6.4	7.3	7.7	6.7	28.1	198.43	37.0	
Totales	31.6	30.6	36.0	34.5	132.7		35.5	
bloque $\left\{ \begin{array}{l} \bar{Y}_i \\ \sum_i Y_{ij}^2 \end{array} \right.$	176.50	175.28	223.36	213.09		788.23		

#### Análisis de la varianza

Fuentes de variación	gl	SC	CM	F
Bloques	$r - 1 = 3$	3.14	1.05	
Tratamientos	$t - 1 = 5$	31.65	6.33	4.83**
Error	$(r - 1)(t - 1) = 15$	19.72	1.31	
Total	$rt - 1 = 23$	54.51		

$$SC \text{ total} = \sum_{i,j} Y_{ij}^2 - C \quad (9.2)$$

$$= 106.98 + \dots + 198.43 - 733.72 = 54.51$$

$$\circ = 176.50 + \dots + 213.09 - 733.72 = 54.51$$

$$SC \text{ bloques} = \frac{\sum_j Y_{ij}^2}{t} - C \quad (9.3)$$

$$= \frac{31.6^2 + \dots + 34.5^2}{6} - 733.72 = 3.14$$

$$SC \text{ tratamientos} = \frac{\sum_i Y_i^2}{r} - C \quad (9.4)$$

$$= \frac{20.4^2 + \dots + 28.1^2}{4} - 733.72 = 31.65$$

$$\begin{aligned} \text{SC error} &= \text{SCtotal} - \text{SCtratamientos} - \text{SCbloques} \\ &= 54.51 - 3.14 - 31.65 = 19.72 \end{aligned} \quad (9.5)$$

El valor de  $F$  para probar la hipótesis nula de que no hay diferencia entre tratamientos es  $6.33/1.31 = 4.83^{**}$  con 5 y 15 grados de libertad es significante al 1 por ciento. Esto comprueba que hay diferencias reales entre las medias de los tratamientos. Para determinar dónde se encuentran las diferencias, pueden usarse procedimientos generales como los vistos en el cap. 8. El procedimiento apropiado se determinará mediante las preguntas que inicialmente se plantea el experimentador. Otros procedimientos se discuten en el cap. 15.

El error estándar muestral de la diferencia entre dos medias de tratamiento igualmente repetido da  $s_{Y_1 - Y_2} = \sqrt{2s^2/r}$ . En esta fórmula  $s^2$  es el cuadrado medio del error y  $r$  es el número de bloques. Para los datos de la tabla 9.2,  $s_{Y_1 - Y_2} = \sqrt{2(1.31)/4} = 0.81$  de porcentaje de aceite, donde el porcentaje de aceite es la unidad de media. El coeficiente de variación es  $CV = s(100)/\bar{Y}_.. = 1.14(100)/35.5 = 3.2$  por ciento, donde el porcentaje ya no se refiere a la unidad de medida. Si ciertos tratamientos reciben repeticiones extras, la fórmula está dada por  $s_{Y_1 - Y_2} = \sqrt{s^2(1/r_1 + 1/r_2)}$ . La repetición extra en un experimento de bloques completos al azar implica que algún tratamiento aparece un número igual de veces en todas las repeticiones, variando el número de tratamiento a tratamiento;  $r_1$  y  $r_2$  serán múltiplos de  $r$ .

La variación entre bloques también puede probarse. En nuestro ejemplo, no es significante. La prueba  $F$  de bloques es válida, pero su interpretación debe hacerse con cuidado. En la mayoría de los experimentos, la hipótesis nula de que no hay diferencias entre bloques no es de importancia particular ya que los bloques son una fuente de variación reconocida, a menudo, con base en experiencia pasada, que se espera considerable. En algunos experimentos, los bloques pueden medir diferencia en el orden de ejecución de un conjunto de operaciones, en partes de un equipo, en personas, etc. En tales casos, la prueba  $F$  para bloques puede tener significado especial.

Si los efectos de los bloques son significantes, ello indica que la precisión del experimento ha aumentado debido al uso del diseño en relación con el diseño completamente aleatorizado. En efecto, la ganancia en eficiencia puede ser de más interés que los resultados de una prueba de significancia; la eficiencia se estudia en la ec. (9.7). También el alcance de un experimento puede haber aumentado cuando los bloques son significativamente diferentes, ya que los tratamientos han sido probados en condiciones experimentales más amplias. Una palabra de cautela es pertinente aquí: si las diferencias de bloques son muy grandes, puede haber un problema de heterogeneidad de error. Este problema se discute en las secs. 7.10, 9.5 y 9.16. Si los efectos de bloque son pequeños, ello indica o que el experimentador no tuvo éxito en reducir la varianza del error agrupando las unidades individuales o que las unidades experimentales eran esencialmente homogéneas desde un principio.

**Ejercicio 9.3.]** Tucker et al. (9.18) determinaron el efecto de lavar y eliminar el exceso de humedad secando o mediante corriente de aire sobre el contenido de ácido ascórbico de nabos. En la tabla siguiente se presentan los datos en miligramos por 100 gramos de peso seco.

Tratamiento	Bloque				
	1	2	3	4	5.
Control	950	887	897	850	975
Lavado y secado con un absorbente	857	1,189	918	968	909
Lavado y secado en corriente de aire	917	1,072	975	930	954

Efectuar un análisis de la varianza de estos datos. Usar el procedimiento de Dunnett para probar las diferencias entre el control y medias de los tratamientos. ¿Qué tasa de error se aplica?

Como alternativa al método de Dunnett, el investigador puede desear comparar

1. El control con las medias de los otros dos tratamientos
2. Los datos de lavado.

Disponer las comparaciones necesarias. Escribir las hipótesis nulas que intervienen. ¿Son ortogonales las comparaciones? Probar las hipótesis nulas. ¿Qué tipo de tasa de error se debe usar? Comparar la suma de las sumas de cuadrados con la suma de cuadrados de tratamientos. Asegúrese de que las sumas de cuadrados se hagan por observación.

Otra alternativa sería probar las dos hipótesis anteriores en forma simultánea. Hacer estas pruebas con el procedimiento de Scheffé. ¿Qué tipo de tasa de error se aplica? Construir intervalos de confianza conjuntos para las dos funciones lineales implicadas por las pruebas; ¿cuáles son estas dos funciones lineales? ¿A qué se aplica el coeficiente de confianza?

**Ejercicio 9.3.2** Bing (9.4) comparó el efecto de varios herbicidas sobre el peso de las flores de gladiolos. El peso promedio por inflorescencia en onzas se da a continuación para los cuatro tratamientos.

Tratamiento	Bloque			
	1	2	3	4
Control	1.25	1.73	1.82	1.31
2,4-D TCA	2.05	1.56	1.68	1.69
DN/Cr	1.95	2.00	1.83	1.81
Sesin	1.75	1.93	1.70	1.59

Analizar los datos. Usar el procedimiento de Dunnett para probar las diferencias entre el control y cada una de las medias de tratamiento. Construir un conjunto de intervalos de confianza para esas diferencias. ¿A qué se aplica  $1 - \alpha$ ?

Supóngase que el investigador desea comparar el control con la media de los otros tres tratamientos. ¿Cuál es la comparación necesaria? Formular la hipótesis nula que se ha de probar. ¿Cuál es la suma de cuadrados, por observación, atribuible a esta comparación? Pruebe la hipótesis nula.

Probar la hipótesis nula de que las tres medias de tratamientos diferentes al control, constituyen un conjunto homogéneo. Para hacer esto, calcular la suma de cuadrados entre estos tres

tratamientos; tendrá dos gl y debe ser con base en observación. El cuadrado medio del error experimental es un divisor apropiado para la prueba  $F$ .

Sumar las sumas de cuadrados y los grados de libertad para las comparaciones control con no control y entre los no controles. Comparar esta suma con la suma de cuadrados de tratamientos.

**Ejercicio 9.3.3** Los datos de la tabla 5.5 corresponden a una clasificación de dos vías en que cada línea corresponde a un bloque con dos mitades de la misma cabeza de trébol rojo.

Analizar los datos con los procedimientos de este capítulo. ¿Se llega a la misma conclusión respecto a la hipótesis nula? ¿Es el CM(error) el mismo que se usó en la ilustración del cap. 5? ¿Es un error un múltiplo entero del otro? ¿Cómo se explica esto?

Obsérvese que el término del error en la ilustración fue medido como la variabilidad entre las diferencias en respuesta a los dos tratamientos en el mismo bloque o como el no haber logrado que las respuestas diferenciales a los tratamientos fuesen análogos. Tener en cuenta esto al leer la sec. 9.4.

**Ejercicio 9.3.4** Repetir el ejercicio 9.3.3 con los datos del ejercicio 5.7.1. Además con los datos del ejercicio 5.7.2.

**Ejercicio 9.3.5** Linthurst llevó a cabo un experimento de invernadero sobre el crecimiento de *Spartina alterniflora*, planta ecológicamente importante de las salinas, para evaluar los efectos de la salinidad, el nitrógeno y aireación. La variable reportada aquí es la biomasa, el peso seco de toda la parte área de la planta.

Bloque	Tratamiento											
	1	2	3	4	5	6	7	8	9	10	11	12
1	11.8	18.8	21.3	83.3	8.8	26.2	20.4	50.2	2.2	8.8	1.4	25.8
2	8.1	15.8	22.3	25.3	8.1	19.5	8.5	47.7	3.3	7.6	15.3	22.6
3	22.6	37.1	19.8	55.1	2.1	17.8	8.2	16.4	11.1	6.0	10.2	17.9
4	4.1	22.1	49.0	47.6	10.0	20.3	4.8	25.8	2.7	7.4	0.0	14.0

*Fuente:* Datos no publicados, cortesía de R.A. Linthurst y E.D. Seneca, Universidad de Carolina del Norte, Raleigh, NC. Título planeado: Aeration, nitrogen, and salinity as determinants of *Spartina alterniflora* growth response.

Número	Tratamiento codificado											
	1	2	3	4	5	6	7	8	9	10	11	12
Salinidad partes/mil	15	15	15	15	30	30	30	30	45	45	45	45
Nitrógeno kg/hectárea	0	0	168	168	0	0	168	168	0	0	168	168
Aireación (0 = ninguna, 1 = saturación)	0	1	0	1	0	1	0	1	0	1	0	1

Presentar un análisis de la varianza para estos datos. Calcular el coeficiente de variación.

Probar la hipótesis nula de que no hay diferencias en respuesta atribuibles a diferencias en la salinidad. ¿Cuántos de los datos son utilizables para esta prueba?

Probar la hipótesis nula de que hay diferencias en respuesta atribuibles a los tratamientos de nitrógeno. ¿Cuántos de los datos son utilizables para esta prueba? ¿Cuál es la probabilidad de obtener un valor mayor del criterio de prueba que el observado en la hipótesis nula?

¿Son ortogonales las comparaciones de nitrógeno y aireación?

Las comparaciones con sentido propuestas, ¿dan razón de todos los grados de libertad de los tratamientos? Comentar.

#### 9.4 La naturaleza del término de error

En el análisis de la varianza para un diseño de bloque completo aleatorizado, la suma de cuadrados del error se encontró, restando de la suma de cuadrados total, las sumas de cuadrados de bloques y tratamientos. Esto es posible ya que las sumas de cuadrados son aditivas. La suma de cuadrados del error puede obtenerse directamente por

$$\text{SC error} = \sum_{i,j} (Y_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y}_{..})^2 \quad (9.6)$$

Esta fórmula de definición proviene del modelo que define las medias de las varias poblaciones muestreadas. Hay  $rt$  medias en el caso del diseño de bloque completo al azar, uno por celda, con sólo una observación necesariamente hecha en cada población. Una medida o valor esperado se define en términos de una media general  $\mu$ , una contribución de tratamiento  $\tau_i$ , y una contribución de bloque  $\beta_j$ ; esto es, la media de la celda  $i, j$ -ésima es  $\mu + \tau_i + \beta_j$ . Una observación está sujeta a un error aleatorio, donde los errores provienen de una sola población con media cero y varianza fija pero desconocida. Así

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}$$

Al usar  $t$  y  $b$  como estimativos, se requiere que  $\sum t_i = 0 = \sum b_j$  como consecuencia de una restricción semejante sobre el modelo (ver sec. 7.5), esto es, que los efectos de tratamientos y bloques se midan como desviaciones. Obtenemos

$$t_i = \bar{Y}_i - \bar{Y}_{..} \quad \text{y} \quad b_j = \bar{Y}_j - \bar{Y}_{..}$$

En realidad, éstas son las estimaciones por mínimos cuadrados con  $\mu$  estimada por  $\bar{Y}_{..}$ . En otras palabras, nuestra estimación de la media de celda  $\mu_{ij} = \mu + \tau_i + \beta_j$  está dada por

$$\hat{\mu}_{ij} = \bar{Y}_{..} + (\bar{Y}_i - \bar{Y}_{..}) + (\bar{Y}_j - \bar{Y}_{..}) = \bar{Y}_i + \bar{Y}_j - \bar{Y}_{..}$$

y

$$\sum [Y_{ij} - (\bar{Y}_i + \bar{Y}_j - \bar{Y}_{..})]^2 = \sum (Y_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y}_{..})^2$$

es la suma de cuadrados más pequeña posible, dada una completa libertad para escoger estimaciones de  $\mu$ , los  $\tau_i$  y  $\beta_j$  con la restricción de que  $\sum \tau_i = 0 = \sum \beta_j$ .

Tabla 9.3 Naturaleza del término de error en el análisis de la varianza

Valores observados, $Y_{ij}$			Estimaciones de las medias $\hat{\mu}_{ij}$			Residuos = diferencias = $Y_{ij} - \hat{\mu}_{ij}$		
Tratamientos	Bloque		Totales	Medias	Bloque	Medias	Tratamiento	Bloque
	1	2	3	1	2	3	1	2
1	5	4	3	12	4	4	3	5
2	4	5	6	15	5	5	4	6
3	6	3	9	18	6	6	5	7
4	7	6	8	21	7	7	6	8
5	3	2	4	9	3	3	2	4
Total	25	20	30	75			Total	
Medias	5	4	6	5	5	4	6	5

## Análisis de la varianza

Fuente de variación	gl	SC
Bloques	2	10
Tratamiento	4	30
Error	8	16
Total	14	56

La tabla 9.3 muestra los cálculos de estimaciones de las *rt* medias y residuos para datos sintéticos. Una media estimada de celda se denota por  $\hat{\mu}_{ij}$ , donde

$$\hat{\mu}_{ij} = \bar{Y}_i + \bar{Y}_j - \bar{Y}_{..} \quad (9.7)$$

La suma de cuadrados para los residuos es

$$\sum_{i,j} (Y_{ij} - \hat{\mu}_{ij})^2$$

se ve que resulta idéntica a la obtenida en el análisis de la varianza. Finalmente,  $Y_{ij} - \hat{\mu}_{ij} = e_{ij}$  es una estimación de  $\varepsilon_{ij}$ . El término de error resultante mide el grado hasta donde las diferencias de tratamientos no son las mismas en todos los bloques; nótese que para los tratamientos 4 y 5 las diferencias son las mismas en todos los bloques, de modo que aquí no se contribuye en nada al error. El término de error también se ve que mide el fracaso de las observaciones a ser iguales a las estimaciones de sus valores esperados.

Nótese que las sumas de residuos son cero para cada fila y columna. Esto indica por qué tenemos 8 grados de libertad en el error. Si se fijan los totales de residuos de filas y columnas, entonces la tabla de residuos no se puede llenar completamente a voluntad. Es necesario resérvar todos los espacios de la última fila y columna, por ejemplo, para los valores que sean necesarios para que los totales de fila y columna sean cero. Esencialmente, esto quiere decir que tenemos libertad de escoger solamente  $(r - 1)(t - 1)$  residuos; tenemos sólo  $(r - 1)(t - 1)$  grados de libertad.

## 9.5 Partición del error experimental

Se ha supuesto que los errores aleatorios provienen de sólo una población. A veces, este supuesto es falso y la varianza del error será mayor para algunos tratamientos que para otros. Esto puede crear un problema en la prueba de hipótesis, por ejemplo, como se vio en la sec. 5.9 con observaciones no pareadas y varianzas desiguales. A veces, se dispone de una solución sencilla para el problema como cuando se utilizan diferencias para observaciones pareadas con sentido sin importar si la varianza está o no relacionada con los tratamientos (ver secs. 5.7 y 5.8). Esta última solución puede generalizarse para aplicarla a comparaciones para este tipo de heterogeneidad de la varianza.

Hoppe registró observaciones con el objeto de comparar 7 fungicidas para semillas y un control sin tratamiento respecto a la emergencia de plántulas de maíz infectado con *Diplodia spp.* El experimento se llevó a cabo en un invernadero con seis bloques en un diseño en bloques completos al azar. Cada unidad experimental consistía en 25 semillas. Los datos y tratamientos codificados se dan en la tabla 9.4 junto con el análisis de la varianza convencional.

Para la partición de las sumas de cuadrados y 7 grados de libertad para los tratamientos, es aconsejable dar un conjunto ortogonal de comparaciones. Las comparaciones de interés se hacen entre las medias de: el control y los 7 fungicidas, fungicidas mercúricos y no mercúricos, los dos fungicidas mercúricos, fungicidas de las compañías I y II, fungicidas de la compañía I, la formulación original y las nuevas de la compañía II y las nuevas fórmulas de la compañía II.

**Tabla 9.4 Densidad de semillas de maíz en invernaderos, infectadas con *Diplodia spp.*, en tratamiento con varios fungicidas**

Bloque	Tratamiento							Totales de bloques	
	A	B	C	D	E	F	G		
1	8	16	14	10	8	8	7	12	83
2	8	19	16	11	7	8	6	19	94
3	9	24	14	12	1	3	6	9	78
4	7	22	13	8	1	3	6	11	71
5	7	19	14	7	3	3	4	9	66
6	5	19	13	3	2	7	4	5	58
Totales de tratamiento	44	119	84	51	22	32	33	65	$G = 450$

Símbolo	Tratamiento
A	control sin tratamiento
B y C	fungicidas mercúricos
D y H	fungicidas no mercúricos, compañía I
E, F y G	fungicidas no mercúricos, compañía II, donde F y G son formulaciones nuevas de E.

Análisis de la varianza				
Fuente	gl	SC	Cuadrado medio	F
Bloque	5	102.50	20.50	
Tratamientos	7	1,210.58	172.94	29.92**
Error	35	202.17	5.78	
Total	47	1,515.25		

*Fuente:* Datosos usados con el permiso de P.E. Hoppe, Universidad de Wisconsin, Madison, Wisconsin.

La tabla 9.5 presenta las comparaciones de los tratamientos, coeficientes, divisores y sumas de cuadrados. Para las comparaciones se usan enteros, en vez de fracciones. La suma de los coeficientes en cualquier fila es cero y la suma de los productos cruzados de los coeficientes de cualquier par de fila es cero, así tenemos comparaciones de acuerdo con nuestra definición y son ortogonales.

Con relación a los coeficientes de la tabla 9.5, todos los signos en cualquier línea o líneas puede combinarse sin efectuar ninguna suma de cuadrados. Para la comparación I, tenemos

$$-7(44) + 1(119 + 84 + 51 + 22 + 32 + 33 + 65) = (-308 + 406) = 98,$$

la diferencia entre 7 veces el total del tratamiento A y el total de los otros siete tratamientos o, alternativamente, 7 veces la diferencia entre la media del control y la media de los 7 fungicidas. Para esta comparación,  $\sum c_i^2 = (-7)^2 + (1)^2 + \dots + (1)^2 = 56$ .

Tabla 9.5 Información pertinente para las siete comparaciones ortogonales

Tratamientos	A	B	C	D	E	F	G	H		
Total de tratamientos										
$T_i$	44	119	84	51	22	32	33	65	$Q$	$(\sum c_i^2)r$
Comparación y no:										SC
1 A con el resto	-7	+1	+1	+1	+1	+1	+1	+1	98	56(6)
2 BC vs. DEFGH	0	+5	+5	-2	-2	-2	-2	-2	609	70(6)
3 B vs. C	0	+1	-1	0	0	0	0	0	35	2(6)
4 DH vs. EFG	0	0	0	+3	-2	-2	-2	+3	174	30(6)
5 D vs. H	0	0	0	+1	0	0	0	-1	-14	2(6)
6 E vs. FG	0	0	0	0	+2	-1	-1	0	-21	6(6)
7 F vs. G	0	0	0	0	0	+1	-1	0	-1	2(6)
Total										28.58*
										1,210.57

Cada suma de cuadrados tiene sólo un grado de libertad y la suma de las sumas de cuadrados es igual a la suma de cuadrados de tratamiento. Cada suma de cuadrados se prueba con error experimental y se compara con  $F(1, f_e)$  donde  $f_e$  es el número de grados de libertad en el error.

La suma de cuadrados del error con 35 grados de libertad puede particionarse en 7 comparaciones ortogonales. El procedimiento es obtener los  $Q$  para cada comparación de la tabla 9.5, para cada bloque, tal como se obtuvieron las diferencias por pares o bloques en la tabla 5.5. Estas se presentan en la tabla 9.6. Para la comparación 1, bloque 1,  $Q = -7(8) + 1(16 + 14 + 10 + 8 + 8 + 7 + 12) = 19$ ; para la comparación 2, bloque 2,  $Q = 5(6 + 14) - 2(10 + 8 + 8 + 7 + 12) = 60$ ; y así sucesivamente. Los totales para cada comparación dentro de bloques se dan en la tabla 9.6. El total, sobre los bloques, de cualquiera de estas comparaciones es el total dado en la tabla 9.5 para la misma comparación. Simplemente se han reordenado los cálculos.

Ninguna comparación dentro de un bloque se ve afectada por el nivel general del bloque si el modelo de bloques completos al azar es válido. Así, si sumamos 10 a cada observación en un bloque, la comparación para ese bloque no cambiará. Esto se debe a que simplemente hemos añadido a cada comparación  $(\sum c_i)10 = 0$ , puesto que  $\sum c_i = 0$ . En consecuencia, la varianza entre los valores de los seis bloques de cualquier comparación deberá ser una varianza aceptable para probar una hipótesis con respecto a la media de la comparación, esto es, respecto al total. Por ejemplo, para la comparación 1, tenemos la suma de cuadrados

$$\frac{19^2 + 30^2 + \dots + 18^2}{56} - \frac{98^2}{6(56)} = 34.75 - 28.58 = 6.17 \quad \text{con 5 gl}$$

Así, para cualquier comparación que se va a probar mediante su propia componente de error, comenzamos por sintetizar seis observaciones como combinaciones lineales de las observaciones en cualquier bloque. Las diferencias entre bloques no son una fuente de variación entre las observaciones sintetizadas. Entonces calculamos una varianza entre las

**Tabla 9.6 Diferencias para siete comparaciones por bloques y la suma de cuadrados de el error para las mismas**

Bloque	Tratamiento						
	1	2	3	4	5	6	7
1	19	60	2	20	-2	1	1
2	30	73	3	48	-8	0	2
3	6	128	10	43	3	-7	-3
4	15	117	9	37	-3	-7	-3
5	10	113	5	28	-2	-1	-1
6	18	118	6	-2	-2	-7	3
Total	98	609	35	174	-14	-21	-1
Divisor	56	70	2	30	2	6	2

Componentes del error (como sumas de cuadrados)							
Comparación	1	2	3	4	5	6	7
Sumas de cuadrados entre bloques	34.75	938.50	127.50	223.67	47.00	24.83	16.50
Factor de corrección $Q^2/(\sum c_i^2)r$	28.58	883.05	102.08	168.20	16.33	12.25	0.08
Componente de error	6.17	55.45	25.42	55.47	30.67	12.58	16.42

seis desviaciones y la usamos para probar la hipótesis nula de que la media de la población de tales observaciones es cero.

Estas sumas de cuadrados de componentes se calculan todas de la misma manera, tal como se ve en la tabla 9.6. El divisor  $\sum c_i^2$  se usa para dar componentes con base por observación. La suma de las siete componentes es 202.18 en comparación con el término de error 202.17 obtenido por diferencia en el análisis inicial. La pequeña diferencia se debe a errores de redondeo. La prueba de homogeneidad  $j_i$ -cuadrado de Bartlett, sec. 20.3, da un valor de 7.95 con 6 grados de libertad. Este valor será mayor en el muestreo aleatorio de las varianzas cuando la hipótesis nula de varianza homogénea es verdadera, con probabilidad entre 0.30 y 0.20. Se concluye que las componentes de la varianza del error no son heterogéneas. La partición del error para probar las comparaciones propuestas parece justificada.

**Ejercicio 9.5.1** ¿Cuántas sumas de productos de coeficientes en la tabla 9.5 se deben verificar para estar seguros de que las siete comparaciones son ortogonales?

**Ejercicio 9.5.2** Revise varios pares de las comparaciones de la tabla 9.5 para ver si son ortogonales.

**Ejercicio 9.5.3** Calcular la suma de cuadrados entre los seis valores dados en la tabla 9.6 bajo la comparación 2. Probar la hipótesis nula de que la media de la población de la cual provienen estas observaciones es cero, esto es, tratar la columna 2 como si fueran todos los datos disponibles.

Calcular  $F$  para probar la hipótesis nula para la comparación 2, dividiendo el cuadrado medio (tabla 9.5 o el factor de corrección en la tabla 9.6) por el cuadrado medio de la componente de error correspondiente. Usar una prueba de  $t$  para probar la misma hipótesis nula. ( $F$  y  $t^2$  sólo deben diferir por errores de redondeo).

**Ejercicio 9.5.4** Considérese la alternativa al procedimiento de Dunnett, ejercicio 9.3.1. Partitionar el error como se describió en esta sección y probar las dos comparaciones. ¿Qué tipo de tasa de error se aplica ahora? ¿Parece haber sido apropiada la partición?

**Ejercicio 9.5.5** Considérense los datos del ejercicio 9.3.5. Encontrar una componente del error experimental aplicable estrictamente a la comparación de nitrógeno. A la comparación con aireación. ¿Cuántos grados de libertad tiene cada componente? ¿Son homogéneas estas componentes? En la hipótesis nula, ¿cuál es la probabilidad de encontrar un valor más extremo del criterio de prueba que el observado?

## 9.6 Datos faltantes

A veces faltan, o son inútiles, datos de ciertas unidades, como cuando un animal se enferma o muere pero no a consecuencia del tratamiento, o cuando los roedores destruyen una parcela en un ensayo sobre el terreno, o cuando se rompe una botella en un invernadero, o cuando ha habido un manifiesto error de registro. Se dispone de un método desarrollado por Yates (9.20) para estimar tales datos faltantes. Una estimación de un valor faltante no proporciona información adicional al experimento; sólo facilita el análisis de los datos restantes.

Cuando falta un *solo valor* en un experimento con bloques completos al azar, calcúlese una estimación del valor faltante por

$$Y = \frac{rB + tT - G}{(r-1)(t-1)} \quad (9.8)$$

donde  $r$  y  $t$  = número de bloques y tratamientos, respectivamente

$B$  y  $T$  = totales de las observaciones observadas en el bloque y tratamiento que contiene la unidad faltante.

$G$  = gran total de observaciones observadas.

El valor estimado se lleva a la tabla con valores observados y se efectúa el análisis de la varianza en la forma usual, restando un grado de libertad a los grados de libertad del total y del error. El valor estimado es tal que la suma de cuadrados del error en el análisis de la varianza es mínimo. La suma de cuadrados de tratamientos está sesgada hacia arriba en una cantidad

$$\text{Sesgo} = \frac{[B - (t-1)Y]^2}{t(t-1)} \quad (9.9)$$

donde  $Y$  se determina según la ec. (9.8).

El error estándar de la diferencia entre la media del tratamiento con un valor faltante y la de cualquier otro tratamiento es

$$s_{\bar{Y}_t - \bar{Y}_c} = \sqrt{s^2 \left[ \frac{2}{r} + \frac{t}{r(r-1)(t-1)} \right]} \quad (9.10)$$

Para ilustrar el procedimiento, supóngase que los datos corresponden a los últimos tres tratamientos de la tabla 9.2 y que la observación faltante sea la de "maduración" en el segundo bloque, eliminamos la observación 4.9 y seguimos.

Los totales observados de los bloques se convierten en 19.5, 13.9, 20.6 y 20.2; los totales observados de los tratamientos se convierten en 26.8, 19.3 y 28.1; el nuevo gran total es 74.2. Usese la ec. (9.8) para obtener

$$Y = \frac{4(13.9) + 3(19.3) - 74.2}{3(2)} = 6.55$$

El valor calculado 6.55 se trata como un valor observado en el análisis de la varianza. Entonces el total de bloque 13.9 pasa a ser 20.45, el total del tratamiento 19.3 pasa a ser 25.85 y el gran total pasa a ser 80.75. El análisis de varianza resultante se da en la tabla 9.7.

El sesgo hacia arriba en la suma de cuadrados de tratamientos se calcula mediante la ec. (9.9), así  $[13.9 - 2(6.55)]^2/3(2) = 0.1067$ , así que la estimación sesgada es 0.53125.

Estos datos con una observación faltante, han perdido el balance esperado en un experimento con bloques completos al azar. Las tres medias de tratamientos según se ve son estimaciones de

$$\mu + \tau_1 + \sum_1^4 \beta_j \quad \mu + \tau_2 + \sum_{j \neq 2} \beta_j \quad y \quad \mu + \tau_3 + \sum_1^4 \beta_j$$

La variación entre las medias observadas no se debe solamente a los  $\tau$  y las componentes aleatorias, porque la media del tratamiento 2 no tiene el conjunto completo de  $\beta$ . Los bloques y los tratamientos no son ortogonales. Mediante el procedimiento de la parcela

**Tabla 9.7 Análisis de varianza para una porción de un conjunto de datos con una observación faltante**

Fuente	gl	SC	CM
Bloques	3	.2373	.0791
Tratamientos	2	.6379†	.3190
Error	6 - 1	1.7471	.3494
Total	11 - 1	2.6223	

4.

3

† Sesgo = 0.1067; SC tratamientos insesgados  
 $= 0.6379 - 0.1067 = 0.53125$

**Tabla 9.8 Análisis de la varianza distinto del de la tabla 9.7**

Fuente	gl	SC	CM
Bloques y tratamientos	5	.8402	
Bloques, sin los tratamientos	(3)	.3089	
Tratamientos, sin los bloques.	(2)	.5313	.2656
Error	5	1.7471	.3494
Total	10	2.5873	

faltante se ha estimado una media de celda  $\mu + \tau_2 + \beta_2$ , por el método de mínimos cuadrados y el análisis se ha proseguido con ese valor. El valor estimado no tiene componente aleatoria, así que los grados de libertad en el total y el error se reducen en uno.

En análisis posiblemente más informativo se presenta en la tabla 9.8. La SC(total) se calcula a partir de las 11 observaciones, por lo tanto difiere del de la tabla 9.7. La SC(error) se toma de la tabla 9.7 y es una estimación insesgada de  $f_e \sigma^2$ , donde  $f_e = 5$ ; otros procedimientos de cálculos se dan en los caps. 14 y 18. La diferencia, SC(total) – SC(error), debe ser la suma de cuadrados asociada con la presencia de efectos de bloques y tratamientos en el modelo. Estos efectos no son ortogonales, así que en seguida se calcula una suma de cuadrados de bloques dejando de lado los tratamientos, como si se tratara de una clasificación simple con bloques como categorías; cada suma de cuadrados debe tener su propio divisor. Esta SC(bloques sin tratamientos) restada de SC(bloques y tratamientos) dice qué tanto de la última SC puede atribuirse a efectos de tratamientos después de tener en cuenta los bloques. Se llama suma de cuadrados para *tratamientos ajustados por bloques*, o SC(tratamientos | bloques), y es la misma que la suma de cuadrados insesgada, calculada previamente.

Cuando hay *varios valores faltantes*, primero se aproximan los valores para todas las unidades excepto una. Aproximaciones razonables de estos valores se pueden obtener calculando  $(\bar{Y}_i + \bar{Y}_j)/2$ , donde  $\bar{Y}_i$  y  $\bar{Y}_j$  son las medias de los valores conocidos para tratamiento y bloque, respectivamente, que contienen alguno de los valores perdidos. También pueden obtenerse esos valores por inspección. Entonces se usa la ec. (9.8) para obtener una aproximación para el valor restante. Con esta aproximación y los valores previamente asignados a todas excepto una de las parcelas faltantes, se usa de nuevo la ec. (9.8) para aproximar ésta.

Luego de completar el ciclo, se encuentra una segunda aproximación para todos los valores en el orden usado previamente. Se continúa hasta que las nuevas aproximaciones no sean materialmente diferentes de las halladas en el ciclo anterior. Por lo general, dos ciclos son suficientes. Los valores estimados se llevan a la tabla con los valores observados, y se completa el análisis de la varianza. Por cada valor faltante, se resta un grado de libertad de los grados de libertad del total y del error. Esto se debe a que los valores estimados no aportan contribución a la suma de cuadrados del error.

Para ilustrar el procedimiento supóngase que faltan los dos valores *a* y *b* en la tabla 9.2 (ver tabla 9.9).

Tabla 9.9 Técnicas de la parcela faltante

Tratamientos	Bloques				Totales de tratamientos	
	1	2	3	4	Valores observados	Todos los valores
1	4.4	5.9	6.0	4.1	20.4	
2	( $a = 4.5$ )	1.9	4.9	7.1	13.9	18.4
3	4.4	4.0	4.5	3.1	16.0	
4	6.8	6.6	( $b = 7.2$ )	6.4	19.8	27.0
5	6.3	4.9	5.9	7.1	24.2	
6	6.4	7.3	7.7	6.7	28.1	
Totales   Valores observados	28.3	30.6	29.0	34.5	122.4	
bloque   Todos los valores	32.8		36.2			134.1

Los cálculos son como sigue

1. Estimar  $b$  así

$$b = \frac{\bar{Y}_i + \bar{Y}_j}{2} = \frac{19.8/3 + 29.0/5}{2} = 6.2$$

2. Estimar  $a$ , primer ciclo, mediante la ec. (9.8)

$$a_1 = \frac{rB + tT - G}{(r-1)(t-1)} = \frac{4(28.3) + 6(13.9) - (122.4 + 6.2)}{(4-1)(6-1)} = 4.5$$

3. Estimar  $b$ , primer ciclo, así

$$b_1 = \frac{4(29.0) + 6(19.8) - (122.4 + 4.5)}{(4-1)(6-1)} = 7.2$$

4. Estimar  $a$ , segundo ciclo, así

$$a_2 = \frac{4(28.3) + 6(13.9) - (122.4 + 7.2)}{(4-1)(6-1)} = 4.5$$

5. Estimar  $b$ , segundo ciclo, así

$$b_2 = \frac{4(29.0) + 6(19.8) - (122.4 + 4.5)}{(4-1)(6-1)} = 7.2$$

Aquí  $G = 122.4 + 4.5 = 126.9$  como en el primer ciclo, ya que  $a_1$  y  $a_2$  son los mismos cuando se redondea a una cifra decimal.

Los valores estimados de la población de medias para las celdas faltantes están dados por la ec. (9.7). Para los valores faltantes de las celdas  $a$  y  $b$ , estas estimaciones son  $32.8/6 + 18.4/4 - 134.1/24 = 4.5$  y  $36.2/6 + 27.0/4 - 134.1/24 = 7.2$ , iguales que los valores faltantes redondeados a un lugar decimal. Las fórmulas para los valores faltantes proporcionan las estimaciones de las medias poblacionales, y así las estimaciones no contribuyen a los grados de libertad del error.

El análisis de la varianza se calcula como de costumbre una vez que los valores estimados se han involucrado en los datos. Los grados de libertad para el total y el error son 21 y 13 respectivamente. Las sumas de cuadrados para bloques, tratamientos y error son 0.96, 5.86 y 1.45 respectivamente. El cuadrado medio del error es un estimador no sesgado de  $\sigma^2$ ; el cuadrado medio de tratamientos está sesgado hacia arriba. Es improbable que la prueba de  $F$  para la hipótesis nula de no diferencia entre medias de tratamientos contenga considerable error a menos que falte un número sustancial de valores. Este procedimiento de aproximación y un poco de sentido común son elementos suficientes para resolver la mayoría de los problemas. Un método exacto de estimación de los valores perdidos lo da Yates (9.20). Un método fácil y exacto lo da también la covarianza, sec. 17.11. El cap. 18 se ocupa de este problema en el caso general.

Para obtener un error estándar para la comparación de dos medias de tratamiento, cada uno con una unidad faltante, usamos una aproximación debida a Taylor (9.24). En

$$s_{\bar{y}_i - \bar{y}_{i'}} = s \sqrt{\frac{1}{r_i} + \frac{1}{r_{i'}}}$$

se reemplazan  $r_i$  y  $r_{i'}$  por un "número efectivo de repeticiones" calculado como sigue: para  $r_i$ , cuéntese 1 si ambos tratamientos están presentes, cuéntese  $(t - 2)/(t - 1)$  si está presente  $i$  pero no  $i'$ , y cuéntese 0 si el tratamiento  $i$  falta.

Sean  $a$  y  $b$  datos faltantes en la tabla 9.9. Supóngase que  $s_{\bar{y}_2 - \bar{y}_4}$  es el error estándar requerido, así que deben calcularse  $r_2$  y  $r_4$ . Para  $r_2$ : en el bloque 1 asignar 0; en los bloques 2 y 4 asignar 1; en el bloque 3, asignar  $(6 - 2)/(6 - 1) = 0.8$ . Entonces  $r_2 = 2.8$ . Para  $r_4$ : en el bloque 1, asignar  $(6 - 2)/(6 - 1) = 0.8$ ; en los bloques 2 y 4 asignar 1; en el bloque 3, asignar 0. Entonces  $r_4 = 2.8$  y

$$s_{\bar{y}_2 - \bar{y}_4} = \sqrt{s^2 \left( \frac{1}{2.8} + \frac{1}{2.8} \right)}$$

De nuevo se necesita tanto  $s_{\bar{y}_2 - \bar{y}_3}$  como  $r_2$  y  $r_3$ . Para  $r_2$ : en el bloque 1 asignar 0; a los bloques 2, 3 y 4 asignar 1. Entonces  $r_2 = 3$ . Para  $r_3$ : en el bloque 1, asignar  $(6 - 2)/(6 - 1) = 0.8$ ; en los bloques 2, 3 y 4, asignar 1. Entonces,  $r_3 = 3.8$  y

$$s_{\bar{y}_2 - \bar{y}_3} = \sqrt{s^2 \left( \frac{1}{3} + \frac{1}{3.8} \right)}$$

**Ejercicio 9.6.1** A partir de los datos de la tabla 9.2, omitir el valor bajo maduración en el bloque 2 y considerar los datos resultantes como si faltara una parcela. Calcular un valor para la parcela faltante y completar el análisis de la varianza. Calcular el sesgo y ajustar la suma de cuadrados de tratamientos. Comparar los resultados con los de esta sección donde sólo se usaron tres tratamientos.

**Ejercicio 9.6.2** A partir de los datos del ejercicio 9.3.1 ó 9.3.2, omitir dos o tres valores para considerarlos como parcelas faltantes. Calcular los valores de las parcelas faltantes y completar el análisis de la varianza. Calcular también el número efectivo de repeticiones para la comparación de dos medias de tratamientos cada una con una parcela faltante; para dos medias de tratamientos, sólo una de las cuales tiene una parcela faltante.

Tomar la SC(error) de este análisis y generalizar el método exacto de análisis descrito para una sola observación faltante para encontrar la SC(tratamientos ajustada por bloques). Esto implica que la SC(total) de los datos incompletos, SC(bloques + tratamientos) obtenida por diferencia, y la SC(bloques, ignorando los tratamientos) también a partir de los datos incompletos.

**Ejercicio 9.6.3** En los datos del ejercicio 9.3.5, la observación para el tratamiento II en el bloque 4 es cero. Esto sugiere que algo salió mal en el manejo del material experimental. Tratar esta "observación" como si faltara. Calcular un valor faltante y calcular de nuevo el análisis de la varianza. Calcular el error estándar de la diferencia entre las dos medias del nitrógeno. Entre las dos medias de aireación.

Probar la hipótesis nula de que no hay diferencia real en la respuesta a los niveles de nitrógeno. ¿Cuál es la probabilidad de encontrar un valor más extremo para el criterio de prueba, bajo la hipótesis nula, que el observado? Comparar este resultado con el observado para la misma pregunta del ejercicio 9.3.5.

Repetir lo anterior para los niveles de aireación

## 9.7 Estimación de la ganancia en eficiencia

Siempre que se use un diseño de bloques completos al azar, es posible estimar la eficiencia relativa a la esperada de un diseño completamente al azar. Se empieza por estimar qué error experimental, CME(CA), pudo haberse obtenido si hubiera usado un diseño completamente aleatorizado. El CME(CA) se estima por

$$\widehat{\text{CME(CA)}} = \frac{f_b \text{CMB} + (f_t + f_e) \text{CME}}{f_b + f_t + f_e} \quad (9.11)$$

donde CMB y CME son los cuadrados medios de los bloques y del error, y  $f_b$ ,  $f_t$  y  $f_e$  son los grados de libertad de bloques, tratamientos y error. Obsérvese que el peso asignado al CME son los grados de libertad que se hubieran obtenido si no hubiera habido tratamientos. El razonamiento viene de una consideración de las propiedades de la aleatorización en pruebas de uniformidad, donde no hay tratamientos. Si los grados de libertad del error de bloque completo al azar son menos de 20, es importante considerar la pérdida de precisión debido a los menos grados de libertad con los cuales se estima el cuadrado medio del error de los bloques al azar comparado con el diseño completamente al azar. Esto se logra multiplicando el factor de corrección una vez obtenido por  $(f_1 + 1)(f_2 + 3)(f_2 + 1)(f_1 + 3)$  (que se estudió en la sec. 6.8), donde  $f_1$  y  $f_2$  son los grados de libertad asociados con el error para los diseños de bloques completos al azar y el completamente aleatorizado, respectivamente.

Al aplicar la sec. (9.11) a los datos de la tabla 9.2 obtenemos

$$\widehat{\text{CME(CA)}} = \frac{3(1.05) + (5 + 15)1.31}{3 + 5 + 15} = 1.28$$

Con la ec. (6.1), encontramos la eficiencia del diseño de bloques en relación con la del diseño completamente aleatorizado,  $\text{ER(BCA a CA)}$ .

$$\begin{aligned}\text{ER(BCA a CA)} &= \frac{(f_1 + 1)(f_2 + 3)}{(f_2 + 1)(f_1 + 3)} \frac{\widehat{\text{CME(CA)}}}{\text{CME(BCA)}} 100 \\ &= \frac{(15 + 1)(18 + 3)}{(18 + 1)(15 + 3)} \frac{1.28}{1.31} 100 = 96 \text{ por ciento}\end{aligned}$$

donde  $f_2 = f_b + f_e = 18$ .

En este caso se sacrifica información, en teoría, al usar el diseño de bloques al azar, pues 96 repeticiones en un diseño completamente al azar dan tanta información como los 100 bloques o repeticiones en un diseño de bloques completos al azar. Cochran y Cox (9.8) dan una demostración de la ec. (9.11).

**Ejercicio 9.7.1** Calcular la eficiencia del diseño de bloques completos al azar en relación con la del diseño de bloques completos al azar en relación con la del diseño completamente aleatorizado para los datos usados en los ejercicios 9.3.1 y 9.3.2.

**Ejercicio 9.7.2** El uso de bloques en experimentos en invernadero no siempre controla la varianza. Calcular la eficiencia del diseño de bloques completos al azar en relación con la del diseño completamente al azar para el experimento descrito en el ejercicio 9.3.5.

## 9.8 El diseño de bloques completos al azar: más de una observación por tratamiento por bloque

Las clasificaciones de dos vías con observaciones múltiples en las celdas son comunes. Se presentan en dos formas. En un caso, cada observación se hace en una unidad experimental en la que todo el conjunto entra en la asignación aleatoria de tratamientos. Es claro que habrá más de  $t$  unidades experimentales por bloque, así que algunos o todos los tratamientos aparecen más de una vez en un bloque. Cuando existen  $r_i$  unidades experimentales por tratamiento en el bloque  $i$ -ésimo, el experimento es un *diseño en bloques (completos) al azar generalizado*. El cap. 18 incluye una discusión del caso en que  $r_i$  depende del bloque, el caso de números de subclases proporcionales. Addelman (9.27) da una recomendación para un mayor uso del diseño de bloques al azar generalizado, junto con algunas referencias adicionales.

En el otro caso, hay por lo general solo  $rt$  unidades experimentales con todos los  $t$  tratamientos asignados aleatoriamente dentro de cada bloque. Sin embargo, la unidad experimental completa no es la unidad medida, ni existe una sola unidad muestral (ver sec. 7.6) por unidad experimental. En cambio, hay más de una unidad muestral por unidad

**Tabla 9.10 Cálculo de sumas de cuadrados para un diseño de bloques completos al azar con varias observaciones por unidad experimental**

Fuente	gl	Sumas de cuadrados	
		Definición	Operación
Bloques	$r - 1$	$ts \sum_j (\bar{Y}_{j..} - \bar{Y}_{...})^2$	$\frac{\sum_j Y_{j..}^2}{ts} - \frac{\bar{Y}_{...}^2}{rts}$
Tratamientos	$t - 1$	$rs \sum_i (\bar{Y}_{i..} - \bar{Y}_{...})^2$	$\frac{\sum_i Y_{i..}^2}{rs} - \frac{\bar{Y}_{...}^2}{rts}$
Error experimental $(r - 1)(t - 1)$	$s \sum_{i,j} (\bar{Y}_{ij.} - \bar{Y}_{i..})^2$	$\frac{\sum_{i,j} Y_{ij.}^2}{s} - \frac{\sum_i \bar{Y}_{i..}^2}{rs}$	$= SC(\text{bloques}) - SC(\text{trat.}) = SC(\text{bloques}) - SC(\text{trat.})$
Error de muestreo	$rt(s - 1)$	$\sum_{i,j,k} (Y_{ijk} - \bar{Y}_{ij.})^2$	$\sum_{i,j} \left( \sum_k Y_{ijk}^2 - \frac{\bar{Y}_{ij.}^2}{s} \right)$
Total	$rts - 1$	$\sum_{i,j,k} (Y_{ijk} - \bar{Y}_{...})^2$	$\sum_{i,j,k} Y_{ijk}^2 - \frac{\bar{Y}_{...}^2}{rts}$

experimental; ahora es necesario tener tanto error de muestreo como error experimental. *Las unidades de observación o de muestreo* se seleccionan al azar dentro de la unidad experimental. Para lograr la mayor información por observación y para mayor comodidad en los cálculos, se necesita de un número igual de observaciones por unidad experimental, siempre que sea posible. Este es el caso que consideramos.

Sea  $Y_{ijk}$  la observación  $k$ -ésima hecha en el bloque  $j$ -ésimo bajo el tratamiento  $i$ -ésimo,  $i = 1, \dots, t, j = 1, \dots, r$  y  $k = 1, \dots, s$  observaciones. Aquí  $i$  y  $j$  se refieren a criterios de clasificación mientras que  $k$  es un índice necesario, pero no sirve como tal criterio. En otras palabras, la observación  $k$ -ésima en la  $ij$ -ésima unidad no es más que la  $k$ -ésima en otra unidad como la  $(k + 1)$ -ésima o cualquier otra observación. Así que los únicos totales con sentido son el gran total  $\bar{Y}_{...}$ , los totales de las unidades experimentales o celdas  $\bar{Y}_{ij.}$ , los totales de bloques  $\bar{Y}_{j..}$ , y los totales de tratamiento  $\bar{Y}_{i..}$ .

Los cálculos necesarios para las sumas de cuadrados en el análisis de la varianza se presentan simbólicamente en la tabla 9.10. Comparar las fórmulas con las de la tabla 9.1; las de la tabla 9.10 son  $s$  veces las de la tabla 9.1. Si fuésemos a calcular bien sea los totales o las medias de las unidades experimentales y las analizáramos sin referencia a  $s$ , entonces las tres primeras líneas serían múltiplos del análisis por observación de la tabla 9.10,  $s$  veces para los totales,  $1/s$  veces para las medias y no podríamos estimar el error de muestreo. Sin embargo, este análisis sería suficiente para pruebas de hipótesis relacionados con tratamientos y bloques porque los valores serían idénticos.

Las sumas de cuadrados del error de muestreo se puede calcular restando de la suma de cuadrados total las sumas de cuadrados de bloques, tratamientos y error experimental o como aparece en la tabla 9.10. El método indicado es obtener la suma de cuadrados

entre las observaciones dentro de cada celda y sumar para la celda  $i, j$ -ésima, la suma de cuadrados entre observaciones es

$$\sum_k Y_{ijk}^2 - \frac{Y_{ij\cdot}^2}{s}$$

El error de muestreo mide el fracaso de las observaciones hechas en una unidad experimental y deben ser precisamente semejantes.

A menudo se espera que el error experimental sea mayor que el error de muestreo, o sea que se espera a menudo que la variación entre unidades experimentales sea mayor que la variación entre submuestras de la misma unidad. Cuando se supone que ambas fuentes de variación son aleatorias, el error experimental es el error apropiado para probar la hipótesis referente a tratamientos y bloques.

Cuando se supone que los efectos de bloques y tratamientos son fijos, no se supone necesariamente que la llamada línea del error experimental, se base en efecto solo en componentes aleatorias. Cuando se supone cierta falta de aleatoriedad, decimos esencialmente que hay efectos fijos y que son fijos porque bloques y tratamientos son fijos para cada combinación de bloque y tratamiento además por encima de las contribuciones aditivas de tratamiento y bloque. Esto también puede enunciarse diciendo que las diferencias en respuestas a los diversos tratamientos no son del mismo orden de magnitud de bloque a bloque. Estas respuestas adicionales, entonces, se suman a las componentes aleatorias. Tales situaciones no son infrecuentes. Se estudian bajo el nombre de *interacciones* en el cap. 15. (Aunque no se usó el término, las interacciones se consideraron en la sec. 8.3 al estudiar ciertos datos de plantas de menta). Por ejemplo, supóngase que se efectúa un experimento sobre el terreno, por elección, en una ladera con la cima bastante seca y la parte baja razonablemente húmeda. Los bloques se escogen de acuerdo con los efectos de los bloques considerados como fijos. El experimento es una prueba de variedades de varios cultivares de un forraje cultivados localmente, uno de los cuales se considera resistente a la sequía. Los efectos de los tratamientos también se consideran fijos. Ahora bien, la variedad resistente a la sequía debe ser notable en el bloque seco en relación con su desempeño en el bloque húmedo. Es decir que la magnitud de la diferencia en respuestas entre ésta y cualquier otra variedad depende del bloque. Este no puede llamarse efecto aleatorio. Por tanto, lo que hemos estado llamando error experimental no es puramente aleatorio; lo llamaremos interacción más error. Ninguna línea en el análisis de la varianza de un término de error apropiado para probar hipótesis relacionadas con la interacción o hipótesis relacionadas con efectos de los tratamientos y bloques.

**Ejercicio 9.8.1** Los datos en la siguiente clasificación de dos factores son tiempos, en segundos, de varios corredores en una distancia de 1.5 millas. Los corredores se clasifican en tres grupos de medidas y tres categorías de estado físico, siendo estas últimas funciones de diversas variables.<sup>†</sup>

¿Cómo difiere el diseño de este experimento del de un experimento de nutrición animal en el que los animales están encerrados o enjaulados por pares y los tratamientos se asignan al azar

---

<sup>†</sup> Datos por cortesía de A. C. Linnerud, Universidad del Estado de Carolina del Norte

	Estado físico		
	Bajo	Medio	Alto
Edad: 40	669	602	527
	671	603	547
50	775	684	571
	821	687	573
60	1009	824	688
	1060	828	713

a los pares enjaulados? ¿Cuál es la unidad experimental en el experimento animal? ¿Cuál en los datos de los corredores?

Para el análisis de la varianza de los datos de los corredores, ¿cómo deberán llamarse las líneas en la tabla 9.10? Completar el análisis de la varianza. Probar la hipótesis nula de que las medias de la población de tiempos de carrera son las mismas para los tres grupos de edades. Probar la hipótesis de que las medias de la población de los tiempos de carrera no dependen de las categorías de estado físico.

¿De qué información se dispone en la línea llamada originalmente "error experimental"? (Considerar cómo se calcula éste).

## 9.9 Modelos lineales y el análisis de la varianza

Para una adecuada evaluación de los datos experimentales, debe establecerse el modelo de manera específica. Dos modelos corrientes son el de *efectos fijos* o *modelo I* y el de *efectos aleatorios*, o *modelo II* (ver sec. 7.5).

Otro modelo común es el *modelo mixto*, que exige por lo menos un criterio de clasificación para incluir efectos fijos y otro para incluir efectos aleatorios. También son posibles otros modelos.

A partir de la tabla 9.11, es evidente el tipo de conclusiones que pueden obtenerse de las pruebas de significancia, en la forma de razones de cuadrados medios. Con respecto a la tabla 9.11 recuérdese que los valores esperados son promedios de los resultados que se obtendrían en experimentación repetida con el modelo dado. Así, para los efectos fijos, se repetirían los mismos efectos en las mismas posiciones en experimentos sucesivos. Para efectos aleatorios, se toma una nueva muestra cada vez y, si bien los mismos efectos terminarán por presentarse de nuevo, puede aparecer en cualquier posición.

Para el *modelo aleatorio* sin interacción, tratamientos y bloques se toman al azar a partir de poblaciones de efectos de tratamientos y de bloques. Las inferencias se sacan respecto de las poblaciones más bien que respecto de tratamientos y bloque particulares. Muy a menudo es deseable tener bloques cuando las generalizaciones referentes a los efectos de los tratamientos se han de hacer para algún conjunto variable de condiciones, tales como municipios o departamentos. Si puede justificarse razonablemente que los bloques son representativos de la población, puede suponerse la aleatoriedad. Los efectos aleatorios de tratamiento son apropiados en muchos experimentos con plantas y animales y en estudios de producción donde puede presentarse variación de un lote a otro o de un día a otro. El cuadrado medio residual es el cuadrado medio del error apropiado para probar

Tabla 9.11 Valores esperados de cuadrados medios para un análisis de bloques completos

Fuente	gl	Modelo aleatorio		
		Sin muestreo	Con muestreo †	
Bloques	$r - 1$	$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$	$Y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ij} + \delta_{ijk}$	
Tratamientos	$t - 1$	$\sigma_e^2 + t\sigma_\beta^2$	$\sigma_e^2 + s\sigma_\tau^2 + r\sigma_\beta^2$	
Residuo	$(r - 1)(t - 1)$	$\sigma_e^2 + r\sigma_\tau^2$	$\sigma_e^2 + rs\sigma_\tau^2$	
Error de muestreo	$rt(s - 1)$	$\sigma_e^2$	$\sigma_e^2 + s\sigma_\tau^2$	$\sigma^2$
Modelo fijo, sin interacción				
Sin muestreo	Con muestreo †	Sin muestreo	Con muestreo †	
$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$	$Y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ij} + \delta_{ijk}$	$Y_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ij}$	$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ij} + \delta_{ijk}$	
$\sigma_e^2 + t \sum \beta_j^2/(r - 1)$	$\sigma_e^2 + s\sigma_\tau^2 + st \sum \beta_j^2/(r - 1)$	$\sigma_e^2 + t \sum \beta_j^2/(r - 1)$	$\sigma_e^2 + s\sigma_\tau^2 + st \sum \beta_j^2/(r - 1)$	
$\sigma_e^2 + r \sum \tau_i^2/(t - 1)$	$\sigma_e^2 + s\sigma_\tau^2 + sr \sum \tau_i^2/(t - 1)$	$\sigma_e^2 + r \sum \tau_i^2/(t - 1)$	$\sigma_e^2 + s\sigma_\tau^2 + sr \sum \tau_i^2/(t - 1)$	
$\sigma_e^2$	$\sigma_e^2 + s\sigma_\tau^2$	$\sigma_e^2 + \sum (\tau\beta)_{ij}^2/(r - 1)(t - 1)$	$\sigma_e^2 + s\sigma_\tau^2 + s \sum (\tau\beta)_{ij}^2/(r - 1)(t - 1)$	
		$\sigma^2$	$\sigma^2$	
Modelo mixto, sin interacción				
Sin muestreo	Con muestreo †	Sin muestreo	Con muestreo †	
$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$	$Y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ij} + \delta_{ijk}$	$Y_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ij}$	$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ij} + \delta_{ijk}$	
$\sigma_e^2 + t\sigma_\beta^2$	$\sigma_e^2 + s\sigma_\tau^2 + st\sigma_\beta^2$	$\sigma_e^2 + t\sigma_\beta^2$	$\sigma_e^2 + s\sigma_\tau^2 + st\sigma_\beta^2$	
$\sigma_e^2 + r \sum \tau_i^2/(t - 1)$	$\sigma_e^2 + s\sigma_\tau^2 + rs \sum \tau_i^2/(t - 1)$	$\sigma_e^2 + t\sigma_\beta^2/(t - 1) + r \sum \tau_i^2/(t - 1)$	$\sigma_e^2 + s\sigma_\tau^2 + st\sigma_\beta^2/(t - 1) + sr \sum \tau_i^2/(t - 1)$	
$\sigma_e^2$	$\sigma_e^2 + s\sigma_\tau^2$	$\sigma_e^2 + t\sigma_\beta^2/(t - 1)$	$\sigma_e^2 + s\sigma_\tau^2 + st\sigma_\beta^2/(t - 1)$	
		$\sigma^2$	$\sigma^2$	
Modelo mixto, con interacción				
Sin muestreo	Con muestreo †	Sin muestreo	Con muestreo †	
$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$	$Y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ij} + \delta_{ijk}$	$Y_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ij}$	$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ij} + \delta_{ijk}$	
$\sigma_e^2 + t\sigma_\beta^2$	$\sigma_e^2 + s\sigma_\tau^2 + st\sigma_\beta^2$	$\sigma_e^2 + t\sigma_\beta^2$	$\sigma_e^2 + s\sigma_\tau^2 + st\sigma_\beta^2$	
$\sigma_e^2 + r \sum \tau_i^2/(t - 1)$	$\sigma_e^2 + s\sigma_\tau^2 + rs \sum \tau_i^2/(t - 1)$	$\sigma_e^2 + t\sigma_\beta^2/(t - 1) + r \sum \tau_i^2/(t - 1)$	$\sigma_e^2 + s\sigma_\tau^2 + st\sigma_\beta^2/(t - 1) + sr \sum \tau_i^2/(t - 1)$	
$\sigma_e^2$	$\sigma_e^2 + s\sigma_\tau^2$	$\sigma_e^2 + t\sigma_\beta^2/(t - 1)$	$\sigma_e^2 + s\sigma_\tau^2 + st\sigma_\beta^2/(t - 1)$	
		$\sigma^2$	$\sigma^2$	

† Recuérdese que  $s$  es igual a 1, en cuyo caso no puede estimarse  $\sigma^2$ .

efectos de bloque y tratamiento con o sin muestreo. Las pruebas pueden efectuarse aunque sólo se tome una observación de muestra en la unidad experimental; sin embargo,  $\sigma_\beta^2$  no puede estimar cuando  $s = 1$ .

Para el modelo aleatorio con interacción, se supone una componente adicional,  $(\tau\beta)_{ij}$ . El símbolo  $(\tau\beta)$  llama la atención sobre la fuente de la componente; el subíndice por fuera del paréntesis evita designar la interacción como un efecto multiplicativo sin excluir esa posibilidad. Para un modelo con efectos aleatorios de bloques y tratamientos, se puede suponer también que los efectos de interacción sean aleatorios con media cero y varianza,  $\sigma_{\tau\beta}^2$ . Esto añade  $\sigma_{\tau\beta}^2$  al modelo sin muestro y  $s\sigma_{\tau\beta}^2$  al modelo con muestreo con todas las líneas, excepto la del error muestral. Las pruebas de hipótesis pueden efectuarse aunque  $\sigma^2$  y  $\sigma_{\tau\beta}^2$  no puedan estimarse separadamente. Este modelo con interacción no se necesita con frecuencia.

Cuando es apropiado el *modelo con efectos fijos* sin interacción, todos los tratamientos respecto a los cuales van a hacerse inferencias se incluyen en el experimento. Lo mismo vale para los bloques. No se pretende que inferencias respecto a tratamientos o bloques se apliquen a tratamientos o bloques no incluidos en el experimento. La prueba se hace sin ningún problema, con o sin muestreo. Este modelo no es raro debido a la naturaleza misma de las categorías en un sistema de clasificación.

Para un modelo de efectos fijos con efectos de interacción, estos últimos serán fijos, ya que los efectos de tratamientos y bloques son fijos. No pueden efectuarse pruebas de hipótesis satisfactorias, con o sin muestreo, así que es importante tratar de saber algo más sobre la naturaleza de la misma.

El *modelo mixto* es ciertamente un modelo común. Lo más frecuente es que los bloques suministren los efectos aleatorios y puede suponerse que son representativos de una población de bloques que cubre una gama de condiciones a las cuales van a aplicarse las inferencias respecto de los tratamientos. Los tratamientos son los únicos de interés con respecto a las inferencias, así que se consideran fijos en experimentación repetida. En el modelo sin interacción, las pruebas, como se ve, no ofrecen problemas especiales con o sin muestreo.

Para el modelo fijo con interacción, considérese que los efectos de interacción están en filas ilimitadas, donde cada columna es una población finita de  $t$  elementos que suman cero, en tanto que la esperanza de una columna es cero. Los elementos que entran en un experimento constituyen una muestra aleatoria de  $r$  columnas, una para cada bloque. Los valores esperados de los cuadrados medios con estos supuestos son tal como se indican. No hay problema para probar tratamientos utilizando el error experimental, independientemente de que haya o no muestreo. No todos los estadígrafos requieren que la suma de elementos de una columna sea cero. En este caso, el coeficiente  $t/(t - 1)$  para  $\sigma_{\tau\beta}^2$  se reemplazará por 1. Los procedimientos de prueba no se verán afectados pero sí las estimaciones de  $\sigma_{\tau\beta}^2$ .

## 9.10 Agrupamiento doble: cuadrados latinos

En el diseño cuadrado latino, se disponen los tratamientos de dos maneras diferentes, por filas y por columnas. Para cuatro tratamientos la disposición puede ser

A	D	C	B
B	C	A	D
D	A	B	C
C	B	D	A

Cada tratamiento se presenta una y sólo una vez en cada fila y columna; cada fila así como cada columna, es un bloque completo. Mediante un análisis apropiado, es posible eliminar del error la variabilidad debida a diferencias tanto en filas como en columnas.

Este diseño se ha usado con ventaja en muchos campos de investigación donde hay dos fuentes principales de variación en la realización de un experimento. En experimentos sobre el terreno, el esquema suele ser cuadrado, permitiendo así la eliminación de la variación proveniente de diferencias en el suelo en dos direcciones. En el invernadero o en el terreno, si hay un gradiente en una sola dirección, el experimento se puede disponer así:

A	D	C	B		B	C	A	D		D	A	B	C		C	B	D	A
---	---	---	---	--	---	---	---	---	--	---	---	---	---	--	---	---	---	---

Aquí las filas son bloques, y las columnas, posiciones en los bloques. Los experimentos de mercadeo se ajustan a este esquema, con días como filas y almacenes como columnas.

Como filas y columnas son términos generales que se refieren a criterios de clasificación, pueden ser una clase de tratamiento. Cuando no hay interacción entre dos cualesquiera o entre todos los criterios —filas, columnas y tratamientos— el *F* calculado no se distribuye como el *F* tabulado, y no son posibles pruebas válidas de significancia. En aquellos casos en los que el experimentador no esté preparado para suponer la ausencia de interacción, no deberá usarse el cuadrado latino.

Babcock (9.2) usó el diseño de cuadrado latino en un experimento para determinar si había diferencias entre las cantidades de leche producidas por los cuatro cuartos de las ubres de las vacas, siendo los cuartos los tratamientos o sea las letras del cuadrado. En este experimento, los tiempos de ordeño eran las filas del cuadrado latino y las órdenes de ordeño eran las columnas —donde orden quiere decir, posición en el tiempo—. El cuadrado latino ha sido usado con ventaja en el laboratorio, en la industria y en las ciencias sociales.

La principal desventaja del cuadrado latino es que el número de filas, columnas y tratamientos debe ser el mismo. Así, si hay muchos tratamientos, el número de parcelas pronto se hace impráctico. Los cuadrados más comunes van de  $5 \times 5$  a  $8 \times 8$ ; cuadrados mayores de  $12 \times 12$  se usan muy rara vez. En los cuadrados latinos, como en los bloques al azar, a medida que aumenta el tamaño del bloque, el error experimental por unidad probablemente aumente. Los cuadrados latinos pequeños proporcionan pocos grados de libertad para estimar el error experimental, y así debe lograrse una disminución sustancial en el error para compensar el corto número de grados de libertad. Sin embargo, puede usarse más de un cuadrado latino en el mismo experimento; por ejemplo, dos cuadrados latinos  $4 \times 4$  darán 15 grados de libertad para el error si los tratamientos responden de

manera análoga en ambos cuadrados. Los grados de libertad para el análisis de  $s$ , para cuadrados  $r \times r$  son como se indican en la tabla.

Fuente	gl
Quadrados	$s - 1 = 1$
Filas dentro de cuadrados	$s(r - 1) = 6$
Columnas dentro de cuadrados	$s(r - 1) = 6$
Tratamientos	$r - 1 = 3$
Error	$s(r - 1)(r - 2) + (s - 1)(r - 1) = 15$
Total	$sr^2 - 1 = 31$

La aleatorización en el cuadrado latino consiste en elegir un cuadrado al azar entre todos los cuadrados latinos posibles. Fisher y Yates (9.11) dan el conjunto completo de cuadrados latinos desde  $4 \times 4$  hasta  $6 \times 6$ , y muestran cuadrados hasta de tamaño  $12 \times 12$ . Cochran y Cox (9.8) dan cuadrados latinos de muestra desde  $3 \times 3$  hasta  $12 \times 12$ . Un modo de aleatorización indicado por Cochran y Cox es el que sigue.

*Cuadrado  $3 \times 3$*  Asignar letras a los tratamientos; esto no tiene que ser al azar. Tratar un cuadrado  $3 \times 3$  y eleatorizar el arreglo de las tres columnas y luego las de las dos últimas filas.

*Cuadrado  $4 \times 4$*  Aquí se tienen cuatro cuadrados, así que no se puede obtener uno de ellos a partir de otro simplemente por reordenación de filas y columnas. Entonces seleccionamos al azar uno de los cuatro cuadrados posibles y distribuimos al azar todas las columnas y las tres últimas filas.

*Cuadrado  $5 \times 5$  y cuadrados mayores.* Ahora hay muchos cuadrados, así que no se puede obtener uno de ellos a partir de otro con reorganizar las filas y columnas. Asignar letras o los tratamientos al azar. Aleatorizar todas las columnas y todas las filas.

### 9.11 Análisis de la varianza del cuadrado latino

Los grados de libertad y las fórmulas para las sumas de cuadrados para un cuadrado latino  $r \times r$  se dan en la tabla 9.12. Aquí,  $Y_{ij}$  representa la observación en la intersección de la fila  $i$ -ésima y la columna  $j$ -ésima. Las sumas de filas y medias se representan como  $Y_i$  y  $\bar{Y}_i$  para  $i = 1, \dots, r$  y las sumas de columnas y medias con  $Y_j$  y  $\bar{Y}_j$ ,  $j = 1, \dots, r$ . Si bien esta notación es adecuada para localizar una observación, no dice nada respecto al tratamiento recibido. Hemos usado  $Y_t$  y  $\bar{Y}_t$  para denotar totales y medias de tratamientos,  $t = 1, \dots, r$ .

El análisis estadístico de un cuadrado latino  $4 \times 4$  se ilustra mediante los datos de rendimiento en una prueba de evaluación de una variedad de trigo efectuada por Ali A. El Khishen, Escuela de Agricultura, Universidad de Alejandría, Alejandría, Egipto. Los datos, el análisis y el plan de campo se presentan en la tabla 9.13. Las variedades están representadas por letras  $A =$  Baladi 16,  $B =$  Mokhtar,  $C =$  Giza 139, y  $D =$  Thatcher. Los rendimientos están en kilogramos por parcela de tamaño  $42 \text{ m}^2$ . El procedimiento de cálculo es como sigue.

Tabla 9.12 Análisis de la varianza para un cuadrado latino  $r \times r$ 

Fuente de variación	gl	Sumas de cuadrados	
		Fórmulas de definición	Fórmulas de cálculo
Filas	$r - 1$	$r \sum_i (Y_{i.} - \bar{Y}_{..})^2$	$\frac{\sum_i Y_{i.}^2}{r} - C$
Columnas	$r - 1$	$r \sum_j (Y_{.j} - \bar{Y}_{..})^2$	$\frac{\sum_j Y_{.j}^2}{r} - C$
Tratamientos	$r - 1$	$r \sum_t (Y_t - \bar{Y}_{..})^2$	$\frac{\sum_t Y_t^2}{r} - C$
Error	$(r - 1)(r - 2)$	$\sum_{i,j} (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$	Por sustracción
Total	$r^2 - 1$	$\sum_{i,j} (Y_{ij} - \bar{Y}_{..})^2$	$\sum_{i,j} Y_{ij}^2 - C$

Paso 1 Calcular los totales de filas  $Y_{i.}$ , de columna  $Y_{.j}$ , de tratamientos  $Y_t$  y el gran total  $Y_{..}$ . Simultáneamente hallar

$$\sum_j Y_{ij}^2 \quad \text{y} \quad \sum_i Y_{ij}^2$$

para cada valor  $i$  y  $j$ , respectivamente. La suma de las cuatro cantidades resultantes para las  $i$ , será igual a la de las  $j$  y es una comprobación de los cálculos. Esta es la suma cuadrados total no ajustada.

Paso 2 Hallar el término de corrección y las sumas de cuadrados (ajustadas)

$$\text{Factor de corrección} = FC = \frac{Y_{..}^2}{r^2} = \frac{167.2^2}{4^2} = 1,747.24$$

$$\text{SC Total} = \sum_{i,j} Y_{ij}^2 - C = 1,837.64 - 1,747.24 = 90.40$$

$$\text{SC Filas} = \frac{\sum_i Y_{i.}^2}{r} - C = \frac{43.4^2 + \dots + 40.0^2}{4} - 1,747.24$$

$$= 1.95$$

$$\text{SC Columnas} = \frac{\sum_j Y_{.j}^2}{r} - C = \frac{39.0^2 + \dots + 44.6^2}{4} - 1,747.24$$

Tabla 9.13 Plan de campo con los rendimientos de trigo, en kilogramos por parcela dispuestos en cuadrado latino  $4 \times 4$ .

Fila	Columna				Totales de fila	
	1	2	3	4	$Y_i$	$\sum_j Y_{ij}^2$
1	$C = 10.5$	$D = 7.7$	$B = 12.0$	$A = 13.2$	43.4	487.78
2	$B = 11.1$	$A = 12.0$	$C = 10.3$	$D = 7.5$	40.9	429.55
3	$D = 5.8$	$C = 12.2$	$A = 11.2$	$B = 13.7$	42.9	495.61
4	$A = 11.6$	$B = 12.3$	$D = 5.9$	$C = 10.2$	40.0	424.70
Totales de columna	39.0	44.2	39.4	44.6	$\sum_i Y_{ij}$	$\sum_{i,j} Y_{ij}^2$
	$\sum_i Y_{ij}^2$	401.66	503.42	410.34	522.22	= 167.2 = 1,837.64

Totales y medias de las variedades			
	$A$	$B$	$C$
Totales = $Y_i$	48.0	49.1	43.2
Medias = $\bar{Y}_i$	12.0	12.3	10.8

Análisis de la varianza					
Fuentes de variación	gl	SC	CM	F	
Filas	$(r - 1) = 3$	1.95	0.65	1.44	
Columnas	$(r - 1) = 3$	6.80	2.27	5.04	
Variedades	$(r - 1) = 3$	78.93	26.31	58.47**	
Error	$(r - 1)(r - 2) = 6$	2.72	0.45		
Total	$(r^2 - 1) = 15$	90.40			
$s = 0.67 \text{ kg}$ $s_{\bar{Y}} = 0.34 \text{ kg}$ $s_{\bar{Y}_i - \bar{Y}_{ij}} = 0.47 \text{ kg}$ CV = 6.4 por ciento					

$$= 6.80$$

$$\text{SC Tratamientos} = \frac{\sum_i Y_i^2}{r} - C = \frac{48.0^2 + \dots + 26.9^2}{4} - 1,747.24$$

$$= 78.93$$

$$\begin{aligned} \text{SC error} &= \text{SC(total)} - (\text{filas}) - \text{SC(columnas)} - \text{SC(tratamientos)} \\ &= 90.40 - (1.95 + 6.80 + 78.93) = 2.72 \end{aligned}$$

Las sumas de cuadrados se llevan a una tabla de análisis de la varianza y luego se encuentran los cuadrados medios. El valor  $F$  para las variedades (tratamientos) es igual a  $26.31/0.45 = 58.47^{**}$ , con 3 y 6 grados de libertad; es mucho mayor que el valor tabulado, el 1 por ciento, 9.28. Entonces se dice que hay diferencia altamente significante entre los rendimientos de las variedades.

El error estándar muestral para una media de tratamiento es  $s_{\bar{y}} = \sqrt{s^2/r} = 0.34$  kg., donde  $s^2$  es el cuadrado medio del error y  $r$  es el número de unidades experimentales por tratamiento. El error estándar de una diferencia entre dos medias de tratamiento es  $s_{\bar{y}_1 - \bar{y}_2} = \sqrt{2s^2/r} = 0.47$  kg. Si se sospecha heterogeneidad de error, éste no puede dividirse tan fácilmente como en el caso de diseño de bloques completos al azar. Cochran y Cox (9.8) ilustran el procedimiento.

El método para determinar cómo difieren las variedades depende de los objetivos del experimento y del conocimiento que se tenga de las variedades, por ejemplo, base genética. En el cap. 8 se han estudiado métodos válidos.

**Ejercicio 9.11.1** Peterson y otros (9.15) presentan el contenido de humedad de un tipo de nabo y otros datos de un experimento efectuado como un cuadrado latino. Los datos aparecen en la tabla adjunta de por cientos (contenido de humedad - 80). Los tratamientos corresponden a tiempos de pasado, dado que las medias de humedad pueden anticiparse en un laboratorio a 70°F a medida que progresó el experimento.

Planta	Tamaño de la hoja ( $A =$ la más pequeña, $E =$ la más grande)				
	$A$	$B$	$C$	$D$	$E$
1	6.67(V)	7.15(IV)	8.29(I)	8.95(III)	9.62(II)
2	5.40(II)	4.77(V)	5.40(IV)	7.54(I)	6.93(III)
3	7.32(III)	8.53(II)	8.50(V)	9.99(IV)	9.68(I)
4	4.92(I)	5.00(III)	7.29(II)	7.85(V)	7.08(IV)
5	4.88(IV)	6.16(I)	7.83(III)	5.83(II)	8.51(V)

Calcular el análisis de la varianza. ¿Cuál es la desviación estándar aplicable a una diferencia entre medias de tratamiento? ¿Entre medias de tamaño de hoja?

**Ejercicio 9.11.2** Un experimento con seis novillas de un año en una vaquería se efectuó en dos cuadros latinos. Los tratamientos fueron tres raciones seleccionadas con base en la localidad y características físicas diversas y se alimentaron *ad libitum*. Cada animal recibió las tres raciones sucesivamente, con una semana para cada una. La variable dada aquí es  $Y$  = libras de materia seca consumida por 100 libras de peso corporal. Los datos se presentan a continuación. Los números dentro de ( ) indican tratamientos. Los tratamientos fueron (1) forraje de alfalfa, (2) maíz ensilado, (3) pastillas de pasto azul.

Novilla	Cuadrado 1		Cuadrado 2			
	1	2	3	4	5	6
1	2.7(1)	2.6(2)	1.9(3)	3.3(1)	2.3(2)	0.1(3)
2	2.1(2)	0.2(3)	2.3(1)	1.7(3)	2.8(1)	1.8(2)
3	1.9(3)	2.1(1)	2.4(2)	2.1(2)	1.7(3)	2.7(1)

*Fuente:* Datos por cortesía de A.C. Linnerud, Universidad del Estado de Carolina del Norte, Raleigh, NC.

Calcular un análisis de la varianza para cada cuadrado separadamente.

¿Difieren significativamente para cada cuadrado las medias de las ecuaciones? ¿Las medias de las novillas para cada cuadrado? ¿Las medias semanales para cada cuadrado?

**Ejercicio 9.11.3** Un cuadrado latino  $3 \times 3$  sólo tiene 2 grados de libertad para estimar el error. Una manera de mejorar la situación es combinar los resultados de experimentos semejantes.

Es claro que la combinación se puede hacer por líneas de los análisis individuales. Por ejemplo, para los cuadrados precedentes, agregar las SC(semanas) para obtener SC(semanas dentro de cuadrados) con 4 grados de libertad. Procédase en forma análoga con novillas, tratamientos y error. El grado de libertad faltante es el de los "cuadrados" que se calcula fácilmente. Este análisis no es muy bueno y deja por fuera algunos puntos.

En lugar de lo anterior, combine la información de dos cuadros del ejercicio precedente de la siguiente forma:

1. Calcular SC(cuadrados) usando los totales de los dos cuadrados. Esto se sugirió arriba.
2. Calcular la SC(semanas) usando los totales de las 3 semanas para ambos cuadrados. Esto es diferente a la sugerencia precedente y tiene sentido si las semanas son las mismas. De otra manera, la sugerencia precedente es satisfactoria.
3. Calcular SC(novillas dentro de cuadrados) mediante la combinación de SC(novillas) a partir de los dos cuadros. Esto tiene 4 grados de libertad. Es el procedimiento sugerido en el anterior parágrafo. (Hay 5 gl entre novillas. El grado de libertad faltante es para las "diferencias entre novillas en diferentes cuadrados" o simplemente "cuadrados".)
4. Calcular SC(tratamientos) usando los tres totales de tratamiento en los cuadros. Esta es evidentemente la forma inteligente de buscar diferencias entre tratamientos.
5. Preparar una tabla de  $3 \times 2$  de totales de semanas por cuadrado. Analizar esto por observación como una clasificación de dos factores. SC(cuadrados) y SC(semanas) ya están en el análisis. La SC(residuales) con 2 grados de libertad es un candidato para el error.
6. Repetir el paso 5 para tratamientos. La SC(residuales) es también aquí candidato para error.
7. Completar la tabla de análisis de la varianza. Sumar las columnas de grados de libertad y SC. Con esto el análisis queda completo.
8. Hallar SC(total). ¿Cuántos gl tiene? ¿Es igual al total en la columna SC del 7?

## 9.12 Parcelas faltantes en el cuadrado latino

El principio que supone la estimación de *valores faltantes* se ilustra en la sec. 9.6 para el diseño en bloques completos al azar. La fórmula para una sola observación faltante es

$$Y = \frac{r(R + C + T) - 2G}{(r-1)(r-2)} \quad (9.12)$$

donde  $R$ ,  $C$  y  $T$  son los totales de los valores observados para fila, columna y tratamiento que contienen el valor faltante y  $G$  es el gran total de los valores observados. El análisis de la varianza se lleva a cabo en la forma usual, restando un gl del total y del error por cada valor faltante.

Como en el caso del diseño en bloques completos aleatorios, la suma de cuadrados de tratamientos está sesgada hacia arriba, en esta oportunidad en una cantidad dada por

$$\text{Sesgo} = \frac{[G - R - C - (r-1)T]^2}{[(r-1)(r-2)]^2} \quad (9.13)$$

El error estándar muestral de la diferencia entre la media del tratamiento,  $A$ , con la unidad faltante y una media del tratamiento,  $B$ , con todas sus unidades presentes es

$$s_{\bar{Y}_A - \bar{Y}_B} = \sqrt{s^2 \left[ \frac{2}{r} + \frac{1}{(r-1)(r-2)} \right]} \quad (9.14)$$

El procedimiento sugerido como más informativo para un experimento de bloques aleatorios con una parcela faltante puede adaptarse al cuadrado latino; sin embargo, es muy largo.

Si faltan *varias unidades* que no constituyen toda una fila, columna o tratamiento, el procedimiento es hacer aplicaciones repetidas de la ec. (9.12) como se hizo en la ec. (9.8). Cuando se han estimado todas las unidades faltantes, se efectúa el análisis de la varianza de la manera usual, restando grados de libertad del total y del error. El procedimiento exacto para obtener el error estándar entre dos medias es complicado. Una aproximación útil dada por Yates (9.20) puede usarse para determinar el número de "replicaciones efectivas" para las dos medias de tratamientos que se comparan. El número efectivo de repeticiones para un tratamiento que se compara con otro se determina sumando los valores asignados como sigue:

- 1 si el otro tratamiento está presente en la columna y fila correspondiente
- 2/3 si el otro falta en una fila o en una columna, pero no en ambas
- 1/3 si el otro falta tanto en la fila como en la columna
- 0 cuando falta el tratamiento en cuestión

*Colas*

Esto se ilustra con cuadrados latinos  $5 \times 5$  dado enseguida con tres unidades faltantes, una para cada tratamiento  $A$ ,  $B$  y  $C$ . Las unidades faltantes se indican entre paréntesis.

1	2	3	4	5
$B$	( $A$ )	$C$	$E$	$D$
$D$	$B$	$E$	$A$	( $C$ )
$E$	$D$	$A$	$C$	$B$
$C$	$E$	$B$	$D$	$A$
$A$	$C$	$D$	( $B$ )	$E$

El número efectivo de repeticiones para los tratamientos  $A$  y  $B$  en la comparación de sus medias se encuentra como sigue: comenzando con la columna 1, hallar  $A$ .  $B$  está en la misma columna, pero no en la misma fila que  $A$ ; asignar el valor  $\frac{2}{3}$ . Para la columna 2, falta  $A$ ; asignar el valor 0. Para la columna 3,  $B$  está presente con  $A$  tanto en la columna como en la fila; asignar el valor 1. Para la columna 4,  $B$  no aparece en la columna, pero sí en la fila con  $A$ ; asignar el valor  $\frac{2}{3}$ . Para la columna 5,  $B$  está con  $A$  en columna y fila; asignar el valor 1. Para  $B$ , se hace lo mismo. Los valores asignados y los números efectivos de repeticiones son

$$\text{Para } A: \quad \frac{2}{3} + 0 + 1 + \frac{2}{3} + 1 = \frac{10}{3}$$

$$\text{Para } B: \quad \frac{2}{3} + \frac{2}{3} + 1 + 0 + 1 = \frac{10}{3}$$

El error estándar de la diferencia entre las medias de los tratamientos  $A$  y  $B$  es entonces  $\sqrt{s^2(1/r_A + 1/r_B)} = \sqrt{s^2(\frac{1}{f_D} + \frac{1}{f_D})}$  donde  $r_A$  y  $r_B$  se refiere a las repeticiones efectivas. El número efectivo de repeticiones para un tratamiento puede diferir con la comparación que se hace.

Cuando faltan todos los valores de una o más filas, columnas o tratamientos, usualmente el análisis se complica. El método para obviar el problema cuando falta una fila, una columna o un tratamiento está dado por Yates (9.21); si faltan dos o más, ver Yates y Hale (9.22). De Lury (9.9) también da esos métodos. Youden (9.23) presenta la construcción de cuadrados latinos incompletos como diseños experimentales.

En algunas oportunidades, los investigadores pueden tener nuevos materiales que desean incluir entre sus tratamientos, pero en cantidades insuficientes por replicación. También es posible reemplazar un tratamiento en un cuadrado latino por  $r$  tratamientos no replicados y contar aún con un análisis. Los cuadrados latinos también son la base de diseños aumentados en que cada parcela se divide de tal modo que el experimento permita  $r^2$  tratamientos adicionales no repetidos. Finalmente, pueden añadirse filas seleccionadas apropiadamente al cuadrado latino para lograr diseños en los cuales puedan medirse efectos residuales cuando ha sido necesario usar cada unidad experimental para una secuencia de tratamientos. Experimentos sobre ganado lechero a menudo usan tales diseños. Como ejemplo, ver Lucas (9.25).

**Ejercicio 9.12.1** Descartar una de las observaciones dadas en el ejercicio 9.11.1. Calcular un valor faltante para la observación descartada y completar el análisis de la varianza. Calcular el sesgo en la suma de cuadrados de tratamientos. ¿Cuál es la desviación estándar aplicable a la diferencia entre dos medias de tratamientos, una con la observación faltante?

**Ejercicio 9.12.2** Para el ejercicio anterior, encontrar la suma de cuadrados total de las 24 observaciones no descartadas. Use ésta y la suma de cuadrados calculada previamente para calcular la SC(filas, columnas y tratamientos) con 12 gl.

Ahora considere que las 24 observaciones provienen de un diseño de bloques con las filas y las columnas correspondientes a bloques y tratamientos. Calcular un valor faltante con la ec. (9.8). Completar el análisis de la varianza. Use el método recomendado como informativo para calcular la SC(filas, columnas) = SC(filas, columnas sin tener en cuenta tratamientos).

Ahora  $SC(\text{filas, columnas y tratamientos}) - SC(\text{filas y columnas, sin tener en cuenta tratamientos}) = SC(\text{tratamientos ajustados por filas y columnas})$ . Esta SC deberá ser igual a  $SC(\text{tratamientos})$  del análisis del ejercicio 9.12.1, una vez se haya reducido por el sesgo. También pudo haberse descrito como una reducción adicional en la SC total atribuible a los efectos de tratamiento en un modelo que previamente sólo incluyó filas y columnas.

### 9.13 Estimación de la ganancia en eficiencia

Se puede estimar la precisión relativa de un cuadrado latino respecto a experimento en bloques completos al azar. Se pueden obtener dos estimaciones de la eficiencia relativa, una cuando las filas se consideran como bloques y otra cuando las columnas se consideran como bloques. Estimamos el cuadrado medio del error para los bloques completos al azar si las filas son los únicos bloques, así

$$\widehat{CME(BCA)} = \frac{f_c CMC + (f_t + f_e) CME}{f_c + f_t + f_e} \quad (9.15)$$

donde CMC y CME son los cuadrados medios de columnas y error en el cuadrado latino y  $f_c$ ,  $f_t$  y  $f_e$  son los grados de libertad de columnas, tratamientos y error del cuadrado latino. Esto quiere decir que las columnas se deben dejar de lado como fuente de variación. Si las columnas son los únicos bloques, y se hace caso omiso de las filas reemplácese  $f_c$  y CMC en la ec. 9.15 por  $f_t$  y CMF, que son los grados de libertad y el cuadrado medio de las filas en el cuadrado latino.

Como se dispone de más grados de libertad para estimar el cuadrado medio del error de bloques al azar que de cuadrados latinos, esto debe tenerse en cuenta al calcular la eficiencia relativa de los dos diseños si el número de grados de libertad en el error del cuadrado latino no es inferior a 20. Esto se logra mediante la incorporación de  $(f_1 + 1)(f_2 + 3)$  ( $f_2 + 1$ ) ( $f_1 + 3$ ) en la fórmula para el factor de eficiencia donde  $f_1$  y  $f_2$  son los grados de libertad para los términos de error del cuadrado latino y de los bloques completos al azar.

Los datos de la tabla 9.13 se usarán para ilustrar el procedimiento. El cuadrado medio del error estimado con filas como bloques y sin columnas es  $3(2.27) + (3 + 6)(0.45)/(3 + 3 + 6) = 0.91$ ; con las columnas como bloques y las filas eliminadas es  $3(0.65) + (3 + 6)(0.45)/(3 + 3 + 6) = 0.50$ . El ajuste de las diferencias en grados de libertad para los dos diseños es  $(6 + 1)(9 + 3)/(9 + 1)(6 + 3) = 0.933$ . Así, la precisión relativa estimada, usando filas como bloques, de modo que comparados los diseños con y sin columnas en la presencia de las filas, es

$$\begin{aligned} \text{ER(CL con respecto a BCA)} &= \frac{(f_1 + 1)(f_2 + 3)}{(f_2 + 1)(f_1 + 3)} \frac{\text{CME (BCA)}}{\text{CME(CL)}} 100 \\ &= .933 \frac{0.91}{0.45} 100 = 189 \text{ por ciento} \end{aligned}$$

Las columnas son evidentemente una fuente de variación que un experimentador deseará controlar.

Si las columnas fueran bloques, de modo que el diseño con y sin filas se pudieran comparar,

$$\text{ER(CL con respecto a BCA)} = .933 \frac{0.50}{0.45} 100 = 104 \text{ por ciento}$$

La agrupación en filas no ha logrado un control apreciable de la variación y el investigador puede desear dejar de lado cuando diseñe un nuevo experimento en circunstancias parecidas.

La agrupación en columnas aumentó la precisión en un porcentaje estimado del 89 por ciento y las filas, en un porcentaje estimado del 4 por ciento. Así, si se hubiera usado el diseño en bloques completos al azar con las filas del cuadrado latino como bloques y se hubiesen omitido las columnas, se habría necesitado un porcentaje estimado de más de 89 de repeticiones para detectar diferencias de la misma magnitud que las detectadas por el cuadrado latino, mientras que se habría necesitado un 4 por ciento más si las columnas hubieran sido bloques y se hubieran omitido las filas. Las filas parecen ser no eficientes como bloques, mientras que las columnas son razonablemente eficientes. En todo campo

de investigación, se necesitan varias de tales comparaciones antes de poder concluir que los cuadrados latinos son probablemente más precisos, en promedio, que bloques al azar.

**Ejercicio 9.13.1** Calcular la eficiencia del cuadrado latino en relación con la de los bloques completos al azar para los datos del ejercicio 9.11.1, usando el tamaño de las hojas como bloques. Repetir usando las plantas como bloques.

#### 9.14 El modelo lineal para el cuadrado latino

Sea  $Y_{ijt}$  la observación en la intersección de la fila  $i$ -ésima con la columna  $j$ -ésima. Esto ubica cualquier observación, pero no dice nada respecto al tratamiento aplicado. Un tercer subíndice puede desorientar, haciendo pensar que se tiene  $r^3$  en vez de  $r^2$  observaciones. Por ejemplo, el tratamiento aparece una vez en cada una de las  $r$  filas, una vez en cada una de las  $r$  columnas, pero solamente  $r$  veces en total; así que  $t = 1$  supone un conjunto de variables  $i, j$ , con un número  $r$ . Lo mismo puede decirse para los otros valores de  $t$ .

Expresamos una observación mediante

$$Y_{ij(t)} = \mu + \beta_i + \kappa_j + \tau_{(t)} + \varepsilon_{ij} \quad (9.16)$$

Esto implica, al usar  $(t)$ , que no se trata de una clasificación ordinaria de tres vías.

Para sacar conclusiones válidas, los  $\varepsilon$  deben ser aleatorios y no correlacionados, y si se van a efectuar pruebas de significancia o se van a construir límites de confianza, por procedimientos ya expuestos, también deben distribuirse normalmente. Un modelo con interacciones no lleva a un error válido para probar hipótesis, y el cuadrado latino no es un diseño apropiado cuando hay interacciones presentes.

Pueden hacerse varios supuestos acerca de los componentes de las medias, esto es, las  $\beta$ ,  $\kappa$  y  $\tau$ . Los valores promedios de los diferentes cuadrados medios se dan en la tabla 9.14 de acuerdo con el modelo supuesto.

No se presenta el modelo mixto, pero los valores promedios de los cuadrados medios se obtienen reemplazando la contribución apropiada por la distribución correspondiente al cambiar de efectos fijos aleatorios, por ejemplo, reemplazando  $(\sum \beta_i^2)/(r - 1)$  por  $\sigma_\beta^2$ .

**Tabla 9.14 Valores promedios de los cuadrados medios para un análisis de un cuadrado latino**

Fuente	gl	Valores promedios de los cuadrados medios	
		Modelo I (fijo)	Modelo II (aleatorio)
Filas	$r - 1$	$\sigma^2 + r(\sum \beta_i^2)/(r - 1)$	$\sigma^2 + r\sigma_\beta^2$
Columnas	$r - 1$	$\sigma^2 + r(\sum \kappa_j^2)/(r - 1)$	$\sigma^2 + r\sigma_\kappa^2$
Tratamientos	$r - 1$	$\sigma^2 + r(\sum \tau_t^2)/(r - 1)$	$\sigma^2 + r\sigma_\tau^2$
Residuo	$(r - 1)(r - 2)$	$\sigma^2$	$\sigma^2$

### 9.15 El tamaño de un experimento

El tamaño de la muestra se estudió en las secc. 5.13, 5.14 y 6.7. Estas secciones tratan de la obtención de intervalos de confianza no mayores que una longitud dada y de la manera de detectar diferencias de magnitud dada; no se consideran problemas en que intervengan más de dos tratamientos, si bien la mayoría de los experimentos incluyen más de dos tratamientos.

Entre los problemas del enfoque general del tamaño de un experimento están los de Cochran y Cox (9.8), Harris y otros (9.12), Harter (9.13), Tang (9.17) y Tukey (9.19). Se estudiarán dos de ellos.

El cálculo del número de repeticiones necesarias depende de

1. Una estimación de  $\sigma^2$
2. La magnitud de la diferencia que se va a detectar.
3. De la seguridad con que se desea detectar la diferencia (poder de la prueba =  $1 - \beta$ ).
4. Del nivel de significancia que se va a usar en el experimento (error tipo I).
5. Del tipo de prueba requerido, bien sea de una o de dos colas.

La sección 5.13 da una solución razonable e ilustra el procedimiento. Esta solución exige una tasa de error por comparación.

En la práctica, se puede obtener una estimación de  $\sigma^2$  por experimentos previos, o eludir su necesidad expresando  $\delta$  como un múltiplo de la verdadera desviación estándar,  $\sigma$ . Para un diseño de bloques con  $t$  tratamientos,  $\sigma_D^2$  del cap. 5 será  $2\sigma^2$ . Así para hipótesis alternativas de dos colas, la ec. (5.33) que puede escribirse

$$r \geq 2(Z_{\alpha/2} + Z_{\beta})^2 \left( \frac{\sigma}{\delta} \right)^2 \quad (9.17)$$

Los subíndices en cada  $Z$  se basan en errores de tipo I y de tipo II aceptables, pero se refieren específicamente a la cantidad de probabilidad en una sola cola de la distribución de  $Z$ .

Para ilustrar el uso de la ec. (9.17), supóngase que deseamos llevar a cabo un experimento como el que dio lugar a los datos de la tabla 9.2, usando nuevamente seis tratamientos. Deseamos detectar diferencias, sin importar la dirección de no más del 2.5 por ciento de aceite al nivel del 95 por ciento con una seguridad del 90 por ciento de detectar una verdadera diferencia de tal magnitud, así  $\alpha = 0.05$  y  $\beta = 0.10$ .

La tabla 9.2 da  $s^2 = 1.31$ , que es una estimación de  $\sigma^2$ ;  $Z_{\alpha/2} = Z_{0.025} = 1.96$  y  $Z_{\beta} = Z_{0.10} = 1.28$ . Por consiguiente,

$$r \geq \frac{2(1.96 + 1.28)^2 1.31}{2.5^2} = 4.4, \text{ ó } 5, \text{ bloques}$$

El experimento con 6 tratamientos y 5 bloques tendrá  $(6 - 1) \times (5 - 1) = 20$  gl. El factor de ajuste necesario es  $(20 + 3)/(20 + 1)$  y el número de bloques ajustado viene a ser  $4.4 \times (23)/21 = 4.8$ , o sea 5 bloques.

El procedimiento (9.19) de Tukey da el tamaño muestral necesario para lograr un conjunto de intervalos de confianza no mayores de un tamaño especificado por todas las

posibles diferencias entre medias verdaderas de tratamientos. El investigador escoge el nivel de significancia y la tasa de error se aplica a un error de tipo I con base experimental.

Como un experimento es una muestra, es imposible dar completa seguridad de que, por ejemplo, el 95 por ciento de los intervalos sean invariablemente menores que el tamaño especificado. Por lo tanto, resulta necesario establecer con qué frecuencia, o con qué seguridad, se desea tener intervalos de confianza menores que la longitud especificada.

Tukey da la siguiente fórmula para el cálculo del tamaño de un experimento.

$$r = \frac{s_1^2 q_\alpha^2(p, f_2) F_\beta(f_2, f_1)}{d^2} \quad (9.18)$$

donde  $s_1^2$  es una estimación de  $\sigma^2$ , basada en  $f_1$  grados de libertad que se obtiene de la tabla A.8 para el coeficiente de confianza deseado y los grados de libertad,  $f_2$ , para el cuadrado medio del error en el experimento que se planea;  $F_\beta$  se obtiene en la tabla A.6 (una cola) para el par de grados de libertad indicados; y  $\beta$  se define de tal modo que  $1 - \beta$  sea la seguridad que deseamos tener de que los intervalos de confianza sean menores que  $2d$ . Es decir,  $d$  se define como la mitad de la longitud del intervalo de semiconfianza deseado. Nótese que  $q$  y  $F$  dependen del valor que se busca. Esto implica que quizá haya que aplicar la fórmula varias veces.

Para ilustrar el empleo de la ec. (9.18), suponga que usamos de nuevo los datos de la tabla 9.2. Si se decide usar un conjunto de intervalos de confianza al nivel del 95 por ciento e incluir los mismos 6 tratamientos, el número de tratamientos es necesario cuando se va a obtener  $q$ . Deseamos que todos los intervalos de semiconfianza al nivel del 95 por ciento tengan una longitud de no más del 2.5 por ciento de aceite con una seguridad de 0.90.

Según la tabla 9.2,  $s_1^2 = 1.31$  y  $f_1 = 15$  grados de libertad. Para obtener  $q$ , hay que conjutar el valor de  $f_2$ . Si suponemos que  $\beta$  será 5 repeticiones y si se planea un diseño de bloques completos, entonces  $f_2 = (5 - 1)(6 - 1) = 20$  grados de libertad y  $q_\alpha(p, f_2) = 4.45$ . Encontramos  $F_{0.10}(20, 15) = 1.92$  y tenemos un conjunto de  $d = 2.5$  de porcentaje de aceite. Así,

$$r = 1.31(4.45)^2(1.92)/(2.5)^2 = 8.0 \text{ bloques}$$

Dado que 8 bloques es bastante más de lo esperado, subestimamos  $f_2$ . Entonces vale la pena estimar de nuevo a  $r$ , usando 7 u 8 para determinar  $f_2$ . Para 7 bloques,  $f_2 = (7 - 1)(6 - 1) = 30$  grados de libertad. Ahora

$$r = 1.31(4.30)^2(1.87)/(2.5)^2 = 7.2 \text{ bloques}$$

Ahora resulta que 7 bloques son suficientes, pero que 8 bloques son algo más que adecuados.

El procedimiento de Stein de dos muestras (sec. 5.14) ha sido generalizado por Healy (9.14) para obtener intervalos de confianza conjuntos de longitud y coeficientes de confianza fijos para todas las posibles diferencias entre medias poblacionales. El procedimiento es aplicable cuando el investigador puede continuar el mismo experimento y no

necesita preocuparse acerca de la heterogeneidad de la varianza para las dos etapas del experimento.

### 9.16 Transformaciones

La aplicación válida de las pruebas de significancia en el análisis de la varianza exige que los errores experimentales se distribuyan normal e independientemente con una varianza común. Además, la escala de medida debe ser tal que cumpla el modelo lineal aditivo. Es costumbre confiar en la aleatorización para romper cualquier correlación de los errores experimentales, siempre que sea posible incorporar un proceso de aleatorización en la investigación. La aditividad se puede probar con un método dado en la sec. 15.8. La otra condición a la cual han concedido mucha atención los estadísticos, es la de la estabilización de la varianza.

La heterogeneidad del error, o heterocedasticidad, puede clasificarse como irregular o regular. De tipo *irregular*, se caracteriza por ciertos tratamientos que tienen una variabilidad considerablemente mayor que otros, sin relación aparentemente manifiesta entre medias y varianzas. Las diferencias en variabilidad pueden o no esperarse de antemano. Por ejemplo, al comparar insecticidas, a menudo se incluyen unidades sin tratar o de control. Los números de insectos presentes en las unidades de control son probablemente mucho mayores y más variables que las de las unidades en que un insecticida ofrece un control considerable. Así, las unidades de control contribuyen al cuadrado medio del error en mayor proporción que las unidades tratadas. En consecuencia, las desviaciones estándar, basadas en cuadrados medios del error combinado, serán demasiado grandes para comparación entre insecticidas y pueden no lograr detectar diferencias reales. En otros casos, ciertos tratamientos pueden presentar más variación que otros sin razón aparente. Esta parte del experimento no está bajo control estadístico.

Cuando la heterogeneidad del error es del tipo irregular, el mejor procedimiento es omitir ciertas porciones de los datos en el análisis o subdividir el cuadrado medio del error en componentes aplicables a las diversas comparaciones de interés. El último procedimiento se estudia en la sec. 9.5 con cierta extensión.

El tipo de heterogeneidad *regular* se origina por lo general en algún tipo de no normalidad en los datos, al estar relacionada la variabilidad dentro de los diversos tratamientos con las medias de tratamientos de alguna manera razonable. Si se conoce la distribución de la población principal, entonces se conoce la relación entre las medias de los tratamientos y las varianzas de los tratamientos calculados con base en los tratamientos individuales. Los datos pueden transformarse o medirse en una nueva escala de medida de tal manera que los datos transformados se distribuyan de forma aproximadamente normal. Tales transformaciones también se proponen hacer que las medias y las varianzas sean independientes, y que las varianzas resultantes sean homogéneas. Este resultado no siempre se consigue. Cuando es imposible hallar una transformación que haga que las medias y las varianzas sean independientes y la varianza estable, entonces deben usarse otros métodos de análisis, tales como el análisis ponderado.

Las transformaciones más comunes son la raíz cuadrada, la-logarítmica y la angular o transformación arco-seno. Se estudian brevemente a continuación.

*Transformación raíz cuadrada.*  $\sqrt{Y}$  Cuando los datos consisten en números enteros pequeños, por ejemplo el número de colonias de bacterias en un recuento de placa, el número de plantas o insectos de una especie determinada en una zona dada, tales datos siguen a menudo la distribución de Poisson, en la cual la media y la varianza son iguales. (Ver sec. 23.6). El análisis para tales datos enumerativos se logra mejor a menudo transformándolos sacando la raíz cuadrada de cada observación antes de proceder al análisis de la varianza.

Los datos porcentuales basados en recuentos y un denominador común, donde el intervalo de porcentaje va de 0 a 20 por ciento o de 80 a 100 por ciento, pero no de ambos, también pueden analizarse con la transformación raíz cuadrada. Los porcentajes entre 80 y 100 por ciento deberán restarse de 100 antes de hacer la transformación. La misma transformación es útil para porcentajes en los mismos intervalos en que las observaciones están en una escala continua, ya que medias y varianzas pueden ser aproximadamente iguales.

Cuando intervienen valores muy pequeños,  $\sqrt{Y}$  tiende a corregir en exceso, así que el intervalo de los valores transformados que dan una media pequeña, puede ser mayor que el intervalo de valores transformados que dan una media mayor. Por esta razón se recomienda a  $\sqrt{Y + \frac{1}{2}}$  como transformación adecuada cuando algunos de los valores está por debajo de 10 o aún por debajo de 15 y especialmente cuando hay ceros.

Las medias apropiadas para una tabla de medias de tratamientos se encuentra elevando al cuadrado la media de los tratamientos calculada con los valores  $\sqrt{Y}$ . Estas serán menores que las medias de los datos originales, situación que se corrige aproximadamente mediante la adición del cuadrado medio del error sin convertir a cada una. Debe recordarse también que el modelo aditivo debe cumplir para la escala transformada y no para la original. En la sec. 15.8 se da una prueba de aditividad.

*Transformación logarítmica,*  $\log Y$  Cuando las varianzas son proporcionales a los cuadrados de las medias de los tratamientos o las desviaciones estándar son proporcionales a las medias, la transformación logarítmica equilibra las varianzas. Se usa la base 10 por comodidad, si bien cualquier base es satisfactoria. Los efectos que son multiplicativos en la escala original de medida, se vuelven aditivos en la escala logarítmica, por ejemplo, si un tratamiento da consistentemente una respuesta 20 por ciento mayor que otro; ver tabla 7.12. Las medias obtenidas volviendo a la escala original mediante los antilog de las medias de los valores de  $\log Y$  son medias geométricas de los datos originales. Cuando se usan pruebas como la dms con los datos transformados son equivalentes a las *razones mínimas significativas* de medias que han sido transformadas de nuevo a la escala original. Esto proporciona un método adecuado de presentar los resultados experimentales.

La transformación logarítmica se usa con números enteros positivos que cubren un amplio intervalo. No puede usarse directamente para valores cero y cuando algunos de los valores son menores que 10, es deseable contar con una transformación que opere como la raíz cuadrada para valores pequeños y como logarítmica para valores grandes. La suma del valor 1 a cada número antes de tomar los logaritmos tiene el efecto deseado. Esto es,  $\log(Y + 1)$  se comporta como la transformación raíz cuadrada para números hasta 10 y difiere poco de  $\log Y$  para valores mayores de 10.

En algún tipo de trabajo experimental, se desea efectuar un análisis de la varianza en que la variable es la varianza. Es apropiada la transformación logarítmica de las varianzas antes del análisis, esto es, analizamos  $\log s^2$ . Tales varianzas deben tener 10 gl o más para

que un análisis de la varianza de los datos transformados sea aprobado. La prueba de Bartlett de la homogeneidad de un conjunto de varianzas, sec. 20.3, también exige el empleo de  $\log s^2$ .

*Transformación angular o transformación arcosen*  $\sqrt{\bar{Y}} \text{ o } \sin^{-1} \sqrt{\bar{Y}}$  Esta transformación es aplicable a datos binomiales expresados como fracciones decimales o porcentajes, y se recomienda especialmente cuando los porcentajes cubren un intervalo amplio de valores. La mecánica de la transformación requiere fracciones decimales, pero las tablas de la transformación arco-seno están generalmente en porcentajes (ver tabla A.10). Los valores tabulados o valores arco-seno se expresan en grados o radianes. La varianza de las observaciones resultantes es aproximadamente constante, y es  $821/n$  cuando los datos transformados se expresan en grados y  $1/(4n) = 0.25/n$  cuando están en radianes, siendo  $n$  el denominador común de todas las fracciones. Esta varianza aclara que todos los porcentajes se deben basar en un número igual de observaciones. Sin embargo, la transformación se usa a menudo cuando los denominadores son desiguales, y especialmente si son aproximadamente iguales; de otra manera, se necesita un análisis ponderado. Bartlett (9.3) sugiere que 0 se sustituya por  $25/n$  por ciento y que 100 por ciento se sustituya por  $100 - 25/n$  (siendo  $n$  el divisor).

La transformación raíz cuadrada ya ha sido recomendada para porcentajes entre 0 y 20 o entre 80 y 100, debiéndose restar estos últimos de 100 antes de hacer la transformación. Si el intervalo de porcentaje es 30 a 70, es dudosa la necesidad de transformación.

*Algunas consideraciones generales* Siempre es útil examinar los datos discretos para asegurarse si existe o no correlación entre las medias de los tratamientos y sus varianzas dentro de los tratamientos. Desafortunadamente, esto no es posible para la mayoría de los diseños. Si la variación muestra poco cambio, es dudoso el valor de cualquier transformación. Cuando hay cambio en la variación, no siempre es claro cuál es la mejor transformación.

Cuando hay duda sobre la transformación apropiada, a veces es útil transformar los datos de varios tratamientos, incluyendo algunos con medias pequeñas, intermedias y grandes en la escala original, y de nuevo examinar varianzas y medias para ver la posibilidad de alguna relación en la escala transformada. La transformación para la cual la relación es mínima, probablemente sea la más apropiada.

Cuando se hace una transformación, todas las comparaciones o las estimaciones de los intervalos de confianza se hace en la escala transformada. Si no se desea presentar resultados en la escala transformada, entonces las medias deben transformarse volviendo a la escala original. Cuando se presentan los resultados de un análisis de datos transformados en la escala original, ésta se le debe aclarar al lector. No es apropiado transformar desviaciones estándar o varianzas, provenientes de datos transformados, volviéndolas a la escala original.

El experimentador que esté interesado en mayor información en cuanto a los supuestos en que se fundamenta el análisis de la varianza y sobre transformaciones, remítase a los escritos de Eisenhart (9.10), Cochran (9.5, 9.6 y 9.7), y Bartlett (9.3). Tales escritos dan muchas referencias suplementarias.

**Ejercicio 9.16.1** En la evaluación de insecticidas, se observó el número de gorgojos adultos vivos del ciruelo que salen de áreas separadas encajonadas de suelo tratado. Los resultados que se presentan en la tabla siguiente se deben a la cortesía de C.B. McIntyre, Departamento de Entomología, Universidad de Wisconsin. Obsérvese que éste es un diseño de bloques completos al azar y que no podemos medir directamente las varianzas dentro de tratamiento.

Bloque	Tratamientos					
	Lindane	Dieldrin	Aldrin	EPN	Chlordane	Check
1	14	7	6	95	37	212
2	6	1	1	133	31	172
3	8	0	1	86	13	202
4	36	15	4	115	69	217

¿Qué recomendación sería apropiada para estos datos? Analizar los datos utilizando como transformación  $\log(Y + 1)$ . Usar los dos procedimientos de Duncan para probar todos los pares posibles de medias de tratamientos con exclusión de la del control.

**Ejercicio 9.16.2** Aughtry (9.1) presenta los siguientes datos sobre la simbiosis del cruce de *Medicago sativa* (53) – *M. Falcata* (50) cruzados con la cepa B. Los datos son porcentajes de plantas con nódulos de un total de 20 por celda. El experimento fue realizado como un diseño de bloques completos al azar.

Bloque	Padres		$F_1$	Lotes de $F_2$ de cada $F_1$				
	53	50		53 × 50	114-1	114-2	114-3	114-4
1	11	65	47		31	22	16	70
2	16	67	32		40	16	19	63
3	6	76	40		27	20	20	52

¿Qué transformación se recomendaría para el análisis de estos datos? Usar la transformación angular y efectuar el análisis de la varianza. ¿Cómo se compara la varianza observada con la teórica?

## Referencias

- 9.1. Aughtry, J. D.: "The effects of genetic factors in *Medicago* on symbiosis with *Rhizobium*," *Cornell Univ. Agr. Exper. Sta. Memo.* 280, 1948.
- 9.2. Babcock, S. M.: "Variations in yield and quality of milk," *6th Ann. Re. Wis. Agr. Exper. Sta.*, 6:42-67 (1889);
- 9.3. Bartlett, M. S.: "The use of transformations," *Biom.*, 3:39-52 (1947).
- 9.4. Bing, A.: "Gladiolus control experiments, 1953," *Gladiolus* 1954, New England Gladiolus Society, Inc.
- 9.5. Cochran, W. G.: "Some difficulties in the statistical analysis of replicated experiments," *Empire J. Exper. Agr.*, 6:157-175 (1938).
- 9.6. Cochran, W. G.: "Analysis of variance for percentages based on unequal numbers," *J. Amer. Statist. Ass.*, 38:287-301 (1943).

## 230 BIOESTADISTICA: PRINCIPIOS Y PROCEDIMIENTOS

- 9.7. Cochran, W. G.: "Some consequences when the assumptions for the analysis of variance are not satisfied," *Biom.*, 3:22-38 (1947).
- 9.8. Cochran, W. G., y J. G. M. Cox: *Experimental designs*, 2a, ed., Wiley, Nueva York, 1957.
- 9.9. DeLury, D. B.: "The analysis of latin squares when some observations are missing," *J. Amer. Statist. Ass.*, 41:370-389 (1946).
- 9.10. Eisenhart, C.: "The assumptions underlying the analysis of variance," *Biom.*, 3:1-21 (1947).
- 9.11. Fisher, R. A., y F. Yates: *Statistical tables for biological, agricultural and medical research*, 5a. ed. Hafner, Nueva York, 1957.
- 9.12. Harris, M., D. G. Horvitz, y A. M. Mood: "On the determination of sample sizes in designing experiments," *J. Amer. Statist. Ass.*, 43:391-402 (1948).
- 9.13. Harter, H. L.: "Error rates and sample sizes for range tests in multiple comparisons," *Biom.*, 13:511-536 (1957).
- 9.14. Healy, W. C., Jr.: "Two-sample procedures in simultaneous estimation," *Ann. Math. Statist.*, 27:687-702 (1956).
- 9.15. Peterson, W. J., H. P. Tucker, J. T. Wakeley, R. E. Comstock, y F. D. Cochran: "Variation in moisture and ascorbic acid content from leaf to leaf and plant to plant in turnip greens," No. 2 in *Southern Coop. Ser. Bull.* 10, págs. 13-17 (1951).
- 9.16. Sackston, W. E., y R. B. Carson: "Effect of pasmo disease of flax on the yield and quality of linseed oil," *Can. J. Bot.*, 29:339-351 (1951).
- 9.17. Tang, P. C.: "The power function of the analysis of variance tests with tables and illustrations of their use," *Statist. Res. Mem.*, 2:126-157 (1938).
- 9.18. Tucker, H.P., J. T. Wakeley, y F. D. Cochran: "Effect of washing and removing excess moisture by wiping or by air current on the ascorbic acid content of turnip greens," No. 10 in *Southern Coop. Ser. Bull.*, 10 págs 54-56 (1951).
- 9.19. Tukey, J. W.: "The problem of multiple comparisons," Ditto, Princeton University, Princeton. NJ, 1953.
- 9.20. Yates, F.: "The analysis of replicated experiments when the field results are incomplete," *Empire J. Exper. Agr.* 1:19-142 (1933).
- 9.21. Yates, F.: "Incomplete latin squares," *J. Agr. Sci.*, 26:301-315 (1936).
- 9.22. Yates, F., o R. W. Hale: "The analysis of latin squares when two or more rows, columns, or treatments are missing," *J. Roy Statist. Soc. Suppl.*, 6:67-69 (1939).
- 9.23. Youden, W. J.: *Statistical methods for chemists*, Wiley, Nueva York, 1951.
- 9.24. Taylor, J.: "Errors of treatment comparisons, when observations are missing," *Nature*, 162: 262-263 (1948).
- 9.25. Lucas, H. L.: "Extra-period latin square change-over designs," *J. Dairy Sci.*, 40:225-239 (1957).
- 9.26. Federer, W. T.: "Augmented designs with one-way elimination of heterogeneity," *Biom.*, 17: 447-473 (1961).
- 9.27. Addelman, Sidney: "The generalized randomized block design," *Amer. Statist.*, 23(4):35-36 (1969).

## REGRESION LINEAL

### 10.1 Introducción

En capítulos anteriores, desarrollamos la idea de que una observación es la suma de una media poblacional y una componente aleatoria. Se disponía de un conjunto más de componentes para las medias y cada componente estaba presente o ausente de una media particular. Por ejemplo, en un diseño completamente aleatorio toda media poblacional contenía  $\mu$ , en tanto que la  $i$ -ésima contenía  $\tau_i$ , pero no otra  $\tau$ . Ahora consideramos medias de población con una componente, un múltiplo fijo de una cantidad variable y medible, llamada variable *concomitante*.

En este capítulo se exponen los usos de una observación concomitante, con los supuestos de los varios usos y con los cálculos necesarios. Los capítulos 11, 14, 17 y 19 también tratan estos problemas generales.

### 10.2 La regresión lineal de $Y$ con respecto a $X$

En la regresión lineal, los valores de  $Y$  se obtienen de varias poblaciones, cada una determinada por un valor correspondiente de  $X$ . Para que la teoría probabilística sea aplicable, es esencial que  $Y$  sea aleatoria. Así mismo, se supone que las poblaciones  $Y$  son normales y tienen una varianza común.

La variable  $Y$  se llama variable *dependiente* pues todo valor de  $Y$  depende de la población muestreada. La variable  $X$  se llama variable *independiente* o *argumento*.

Como ilustración se usan los datos de la tabla 10.1, registros de corral provenientes de una prueba de hipótesis hecha en Nueva York entre 1953-1954. La variable dependiente corresponde al alimento consumido  $Y$ ; depende la variable peso del cuerpo,  $X$ . Los diez pares de valores de la tabla 10.1 se representan en la fig. 10.1. Hay una relación bastante definida entre las dos variables. En particular, una recta como la trazada entre esos puntos podría servir como promedio móvil de los valores de  $Y$ . Siempre es buena idea representar

**Tabla 10.1 Peso promedio  $X$  y consumo de alimento  $Y$  de 50 gallinas provenientes de 10 razas White Leghorn**

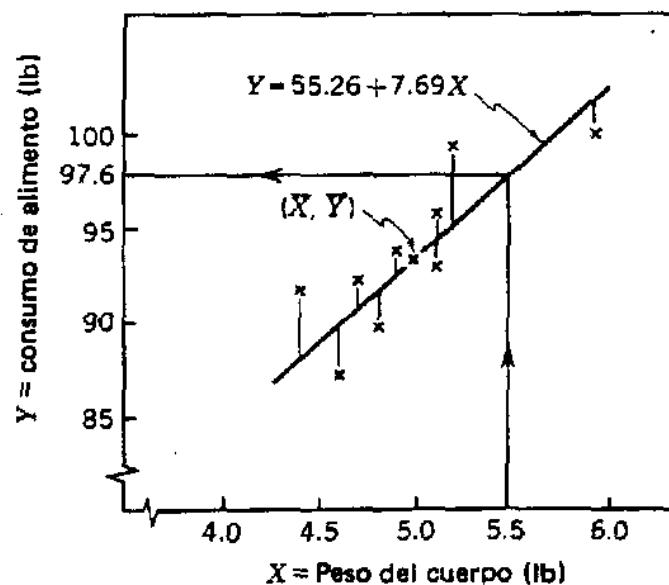
Período de 350 días

Peso del cuerpo		Consumo de alimento	
$X$	$X' = X - 4.0$	$Y$	$Y' = Y - 80$
4.6	-0.6	87.1	7.1
5.1	1.1	93.1	13.1
4.8	0.8	89.8	9.8
4.4	0.4	91.4	11.4
5.9	1.9	99.5	19.5
4.7	0.7	92.1	12.1
5.1	1.1	95.5	15.5
5.2	1.2	99.3	19.3
4.9	0.9	93.4	13.4
5.1	1.1	94.4	14.4
$\sum (X - \bar{X})^2 = 1.536$		$\sum (Y - \bar{Y})^2 = 135.604$	
* cada una con 9 gl			

Fuente: Datos cortesía de S.C. King, ahora en la Universidad de Purdue, Lafayette, Indiana.

tales datos para tener una indicación de la intensidad de una relación, de su discrepancia respecto de la linealidad y de las observaciones extremas o insólitas. Consideraremos ahora una gráfica en línea recta y su ecuación.

La ecuación de una recta puede escribirse,  $Y = a + bX$  (ver fig. 10.2). Todo punto  $(X, Y)$  de esta recta tiene una coordenada  $X$  o *abscisa* y una coordenada  $Y$  u *ordenada*, cuyos valores satisfacen a la ecuación. Las ordenadas de puntos que no están en la recta no satisfacen a la ecuación. Cuando  $X = 0$ ,  $Y = a$ , así que  $a$  es el punto donde la recta corta el eje de las  $Y$ , esto es,  $a$  es el *intercepto de Y*. Cuando  $a$  es 0, la recta pasa por el



**Figura 10.1 Regresión del consumo de alimento  $Y$  respecto del peso del cuerpo  $X$  para 10 razas de White Leghorn (promedios para 50 gallinas).**

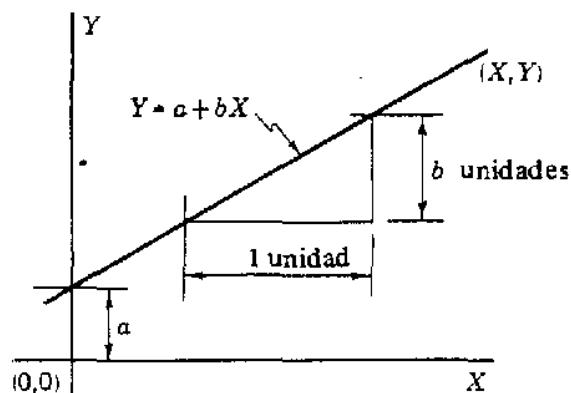


Figura 10.2 Gráfica de una recta.

origen. Una unidad de variación en  $X$  produce una variación de  $b$  unidades en  $Y$ , así que  $b$  es una medida de la *pendiente* de la recta. Cuando  $b$  es positivo, ambas variables aumentan o disminuyen juntas; si  $b$  es negativo, una variable aumenta cuando la otra disminuye. Para una línea recta, cualquier par de puntos o la pendiente y el intercepto determinan la posición de la recta en forma única. Los matemáticos llaman a las relaciones como  $Y = a + bX$  *relaciones funcionales*. Para un valor de  $X$ , una relación funcional le da un valor a  $Y$ ; una fórmula matemática relaciona las dos variables. Este tipo de relación se ha usado en el cap. 3 para asignar probabilidades, en las ecs. (3.4) a (3.6) y determinan ordenadas de una curva, como en la ec. (3.11).

El investigador científico que trabaja con observaciones bivariantes no tendrá datos que caigan en una línea recta. Aun en el caso en que las temperaturas se miden en escala Fahrenheit y Celsius, habrá variación aleatoria atribuible a errores de medida; esto impide una relación perfecta aun cuando exista una relación funcional entre variables. Más a menudo, los datos bivariantes siguen una *ley estadística*, que se cumple en promedio. Este es el caso de los datos de la tabla 10.1. Es claro que, para un peso dado  $X$ , no existe un solo valor de consumo de alimento  $Y$ ; en vez de esto, hay una población de  $Y$  y las observaciones aleatorias se extraen de esa población.

Cuando se va a ajustar una recta a datos consistentes en más de dos pares de valores, se elige la recta que mejor se ajuste a los datos, esto es, aquella que corresponda al mejor promedio móvil. Nuestro criterio de lo que es mejor es el criterio de los mínimos cuadrados, que exige que la suma de los cuadrados de las desviaciones de los puntos observados con respecto al promedio móvil de la línea recta para un mismo valor  $X$  sea mínima. En tal recta ajustada,  $b$  se llama *coeficiente de regresión*; la recta se llama recta de regresión y su ecuación se denomina *ecuación de regresión*.

Para determinar el coeficiente de regresión  $b$ , necesitamos la suma de *productos cruzados* de las desviaciones de las observaciones respecto de sus medias correspondientes y la suma de cuadrados de  $X$ . Las fórmulas de definición y de cálculo para una suma de productos cruzados se dan en las ecs. (10.1) y (10.2), respectivamente

$$PC = \sum_i (X_i - \bar{X})(Y_i - \bar{Y}) \quad \text{definición} \quad (10.1)$$

$$= \sum_i X_i Y_i - \frac{\sum_i X_i \sum_i Y_i}{n} \quad \text{cálculo} \quad (10.2)$$

Para los valores codificados de  $X$  e  $Y$ , esto es,  $X'$  e  $Y'$ , obtenemos

$$\sum (X - \bar{X})(Y - \bar{Y}) = 144.70 - \frac{(9.8)(135.6)}{10} = 11.812$$

La codificación mediante suma o resta de una constante para cada variable no afecta la suma de los productos cruzados. Las siguientes relaciones también son útiles:

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i(Y_i - \bar{Y}) = \sum (X_i - \bar{X})Y_i$$

La semejanza entre una suma de productos cruzados y una suma de cuadrados, tanto la fórmula de definición como la operación, es evidente si se reemplazan  $Y_i$  y  $\bar{Y}$  por  $X$  y  $\bar{X}$  en la fórmula de productos cruzados. Las sumas de cuadrados y productos cruzados suelen llamarse simplemente *sumas de productos*.

Cuando  $\sum (X - \bar{X})(Y - \bar{Y})$  se divide por los grados de libertad, recibe el nombre de *covarianza* (ver sec. 5.10). Tenemos  $\sum (X - \bar{X})(Y - \bar{Y})/(n - 1) = 11.812/9 = 1.312$ . Una covarianza es una medida de la variación conjunta de dos variables y puede ser positiva o negativa. Es simétrica en  $X$  o en  $Y$ , y las variables no necesitan especificarse como dependientes o independientes.

El coeficiente de regresión se determina mediante

$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} \quad (10.3)$$

$$= \frac{11.812}{1.536} = 7.69 \text{ lbs de alimento por libra de gallina}$$

Para un aumento de una libra en el peso del cuerpo, el consumo de alimento es de 7,69 libras.

La ecuación de regresión puede escribirse como la ec. (10.4) pues la recta de regresión pasa por la media muestral.

$$\hat{Y} - \bar{Y} = b(X - \bar{X}) \quad \text{o} \quad \hat{Y} = \bar{Y} + b(X - \bar{X}) \quad (10.4)$$

El  $\hat{Y}$  indica simplemente que ésta es la recta de regresión de la muestra, no un valor observado de  $Y$ .

$$\hat{Y} - 93.56 = 7.69(X - 4.98)$$

o bien

$$\hat{Y} = 93.56 + 7.69(X - 4.98) \quad \text{o} \quad \hat{Y} = 55.26 + 7.69X$$

(ver fig. 10.1). El intercepto de  $Y$ ,  $a$ , es igual a  $\bar{Y} - b\bar{X} = 55.26$ .

Las rectas calculadas en problemas de regresión son rectas en torno a las cuales se acumulan los pares de valores, no son rectas en las cuales quedan los puntos. Un punto sobre una línea de regresión es una estimación de una media de una población de  $Y$ , los  $Y$  que tienen el correspondiente valor de  $X$ .

Ejercicio 10.2.1 Los registros de corral presentados a continuación para gallinas White Leghorns, en una competencia de muestra\* aleatoria en California son similares a los datos de la tabla 10.1.

$Y$ (consumo de alimento en lbs., en 350 días)	87.8, 93.2, 98.0, 89.8, 94.0, 83.0, 88.3, 82.4, 84.8, 80.2
$X$ (peso corporal promedio de 50 gallinas, lbs )	4.15, 4.76, 5.23, 4.75, 5.13, 4.24, 4.66, 4.41, 4.50, 4.23

Calcular la ecuación de regresión del consumo  $Y$  respecto al peso del cuerpo  $X$ . ¿Cuál es el coeficiente de regresión y qué unidad le corresponde? ¿Cuál es la interpretación de este coeficiente de regresión?

**Ejercicio 10.2.2** Demostrar algebraicamente que  $\sum (X - \bar{X})(Y - \bar{Y}) = \sum (X - \bar{X})Y$ . (Sugerencia: Demostrar que cada una es igual a la fórmula de cálculo).

**Ejercicio 10.2.3** En un programa de carreras para la conservación de la salud se tomaron medidas al final de un año de  $X_1$  = tasa de recuperación cardíaca según una prueba de Harvard,  $X_2$  = número de saltos hasta el agotamiento,  $X_3$  = cantidad de colesterol,  $X_4$  = cantidad de ácido úrico y  $Y$  = tiempo en minutos: segundos tiempo de varios hombres para correr 1.5 millas. Algunos de los datos son los siguientes:

$X_1$	52	39	39	51	49	45	54	54	
$X_2$	45	60	40	101	60	51	50	99	
$X_3$	269	279	248	318	318	254	263	320	
$X_4$	43	65	78	73	71	69	67	45	
$Y$	11:16	12:30	11:30	10:17	11:48	11:06	12:02	11:52	
<hr/>									
$X_1$	47	69	35	33	45	39	61	46	42
$X_2$	34	41	251	125	105	40	42	113	25
$X_3$	228	306	303	264	253	281	346	223	266
$X_4$	69	60	65	59	76	40	70	72	60
$Y$	11:28	13:45	8:18	10:23	12:02	13:45	13:41	10:30	23:11

*\*Fuente: Datos cortesía de Ardell Linnerud, Universidad del Estado de Carolina del Norte, Raleigh, N.C.*

Calcular la ecuación de regresión muestral del tiempo de carrera  $Y$  respecto a cada variable  $X$  separadamente. Escribir cada ecuación con intercepto y en la forma de la ec. (10.4). ¿Cuál es la interpretación del coeficiente de regresión en cada caso?

\* Datos cortesía de S.C. King, Universidad de Purdue, Lafayette, Indiana.

**Ejercicio 10.2.4** En un estudio sobre acumulaciones en suelos forestales, se usaron parcelas de  $\frac{1}{16}$  de acre de un tipo de pino en una cuenca de Wyoming. La variable respuesta registrada es  $Y$  = peso de humus en libras por acre. Se usaron varias variables predictoras incluyendo  $X_1$  = área basal en pulgadas cuadradas medida a la altura de la cintura,  $X_2$  = densidad de árboles, algunos de los datos son

$X_1$	3,062	2,347	1,948	3,075	2,899	2,138	2,316	1,968	2,827	1,072
$X_2$	280	203	307	148	79	71	97	51	192	44
$Y$	51,916	47,022	29,945	42,649	43,557	34,964	57,129	21,633	68,620	51,417
$X_1$	2,804	2,358	2,526	2,400	2,507	3,009	3,764	4,070	3,038	3,116
$X_2$	122	136	97	124	404	270	55	88	82	33
$Y$	34,008	52,667	48,575	30,793	29,242	38,174	61,653	71,410	61,625	54,180

Fuente: Datos cortesía de James Reynolds, Universidad del Estado de Carolina del Norte, Raleigh, N.C. También ver J. I. Reynolds y O. H. Knight (10.9).

Calcular la ecuación de regresión muestral del peso de humus respecto de cada variable  $X$ . Escribir cada ecuación con intercepto y en forma de desviaciones con respecto a las medias. ¿Cuál es la interpretación del coeficiente de regresión en cada caso?

**Ejercicio 10.2.5** Reynolds y otros (10.8) estudiaron  $Y$  = (tasa de fotosíntesis) $10^4$  de *Larrea tridentata* y su relación con  $X_1$  = radiación,  $X_2$  = concentración de  $\text{CO}_2$  en el ambiente,  $X_3 = 10^6/X_1$ , y  $X_4$  = resistencia de la hoja al vapor de agua, entre otras variables independientes. Los datos se tomaron en el fitotrón de la Universidad de Duke; algunos de los datos son los siguientes:

$X_1$	294	190	294	550	550	2,000	550	550	550	2,000	2,000
$X_2$	665	671	664	577	577	576	682	614	605	605	545
$X_3$	3,401	5,263	3,401	1,818	1,818	500	1,818	1,818	1,818	1,818	500
$X_4$	990	968	1,868	1,814	2,521	1,516	4,707	1,935	4,675	2,234	1,158
$Y$	348	131	402	731	526	1,346	4,767	635	360	618	1,385
$X_1$	2,000	2,000	800	800	1,200	1,200	1,200	1,600	1,600	400	400
$X_2$	502	521	536	536	556	570	547	582	553	576	568
$X_3$	500	500	1,250	1,250	833	833	833	625	625	2,500	2,500
$X_4$	1,697	646	1,086	998	911	765	1,284	915	1,410	4,111	1,802
$Y$	1,415	1,467	842	927	1,099	1,086	910	1,055	937	349	498

Calcular la ecuación de regresión de la tasa de fotosíntesis ( $\times 10^4$ ) respecto a cada una de las variables  $X$ . Escribir toda ecuación con intercepto y como en la ec. (10.4). ¿Cuál es la interpretación de la regresión en cada caso?

### 10.3 El modelo y la ecuación de regresión lineal

Por definición, la verdadera regresión de  $Y$  con respecto a  $X$  consiste en las medias de las poblaciones de valores  $Y$ , donde una población está determinada por el valor de  $X$ . Una línea de regresión no tiene que ser recta. En el muestreo, es necesario suponer la forma de línea de las medias; de otra manera no sería posible desarrollar un procedimiento de cál-

culo. Hemos supuesto una línea recta o *regresión lineal*. Tales supuestos generalmente se hacen con base en la teoría o la experiencia, o aun después de observar los datos representados gráficamente. Por facilidad en los cálculos, a menudo se escoge una recta como aproximación cuando se ajusta razonablemente bien en el intervalo de  $X$  en cuestión, aun cuando se sepa que la verdadera forma no es lineal.

La definición matemática de una observación está dada por

$$\begin{aligned} Y_i &= \mu_Y \cdot x + \varepsilon_i = \alpha + \beta X_i + \varepsilon_i \\ &= \mu + \beta(X_i - \bar{X}) + \varepsilon_i \end{aligned} \quad (10.5)$$

donde  $\alpha$  y  $\beta$  son parámetros que hay que estimar y  $X$  es un parámetro observable;  $\alpha$  representa el intercepto de la población  $Y$ , el valor de  $\mu_Y$  para  $X = 0$ , esto es,  $\mu_{Y \cdot 0}$ ;  $\beta$  es la pendiente de la recta que pasa por las medias de las poblaciones  $Y$ . En la última forma del modelo dada en la ec. (10.5), con los  $X$  medidos a partir de su media,  $\mu$  está estimado por  $\bar{Y}$ . Se supone que los  $\varepsilon$  pertenecen a una sola población con media 0 y varianza  $\sigma^2$ . Esta varianza es otro parámetro que hay que estimar.

El modelo de regresión puede ser el modelo I con  $X$  fijos, o el modelo II con  $X$  aleatorio.

Para el modelo I, el investigador selecciona los  $X$ , no hay variación muestral aleatoria o relación con ellos. En cambio, los valores de  $Y$  deben ser aleatorios. La selección de los  $X$  puede inducir un conjunto específico de valores o valores que se encuentran simplemente dentro de un intervalo deseado. La respuesta a un insecticida bien puede medirse así mediante una serie específica de diluciones, mientras que el peso del cuerpo humano puede corresponder a un intervalo de estaturas condicionadas por una descripción de la ocupación. Cuando se consideran valores esperados, se usan los mismos  $X$  para definir la repetición del muestreo que es su base. Estos  $X$  deben medirse sin error.

Los valores de la variable independiente, por ejemplo, horas de luz artificial, niveles de temperatura, cantidades de tratamientos y distancias entre plántulas, pueden estar igualmente espaciados o de otra forma conveniente para llevar a cabo el experimento.

La medición de  $Y$  sin error no es requisito teórico, siempre y cuando que el error de medida tenga una distribución media conocida, generalmente 0. La varianza observada de los  $Y$  es, entonces, la suma de la varianza biológica u otra varianza en  $Y$  y la varianza del error de medida. Es importante, naturalmente, mantener al mínimo los errores de las medidas.

Si se hace una observación aleatoria de  $Y$  para  $X = X_i$ , correctamente identificada, entonces el observador puede proceder a usar esta información con sentido. Sin embargo, si se registra incorrectamente el valor de  $X_i$  como  $X'_i$ , entonces los cálculos subsiguientes pueden conducir a conclusiones desorientadoras porque la población de los  $Y$  está identificada incorrectamente. La fig. 10.3 debe aclarar esto.

Supóngase que el modelo I es apropiado y que el problema se especifica más de la forma siguiente.

Primero, supóngase que existe una relación funcional fundamental entre  $X$  y  $Y$ , pero que son posibles errores de observación. El problema está en estimar esa relación. Si sólo se miden los  $X$  sin error, tal como se ve en la fig. 10.4 a, entonces son aplicables los procedimientos de cálculo ya expuestos. Hay una sola recta de regresión, la de  $Y$  con respecto a  $X$ .

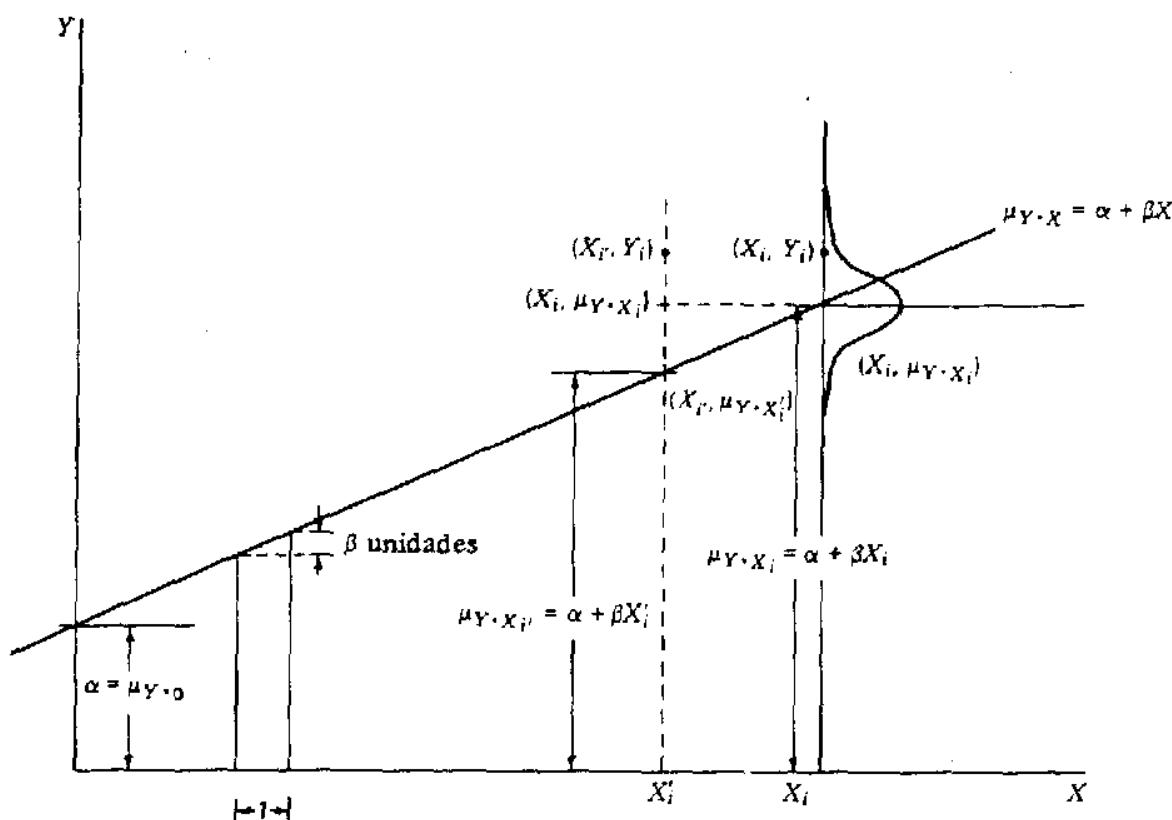


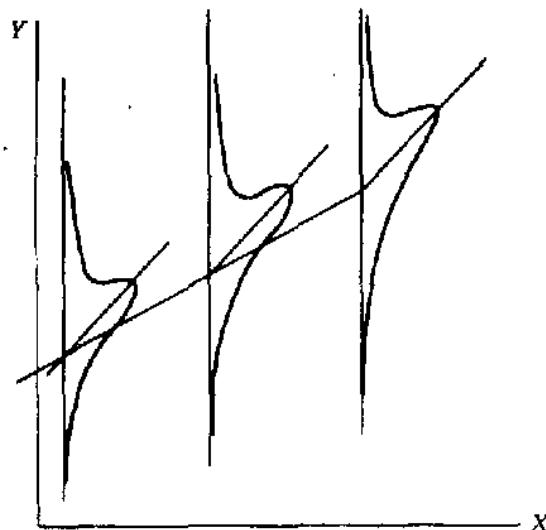
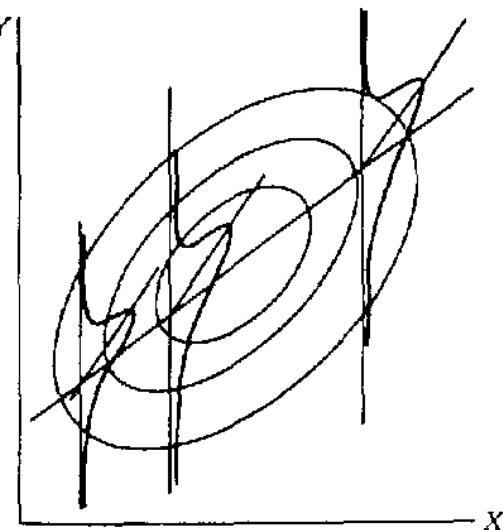
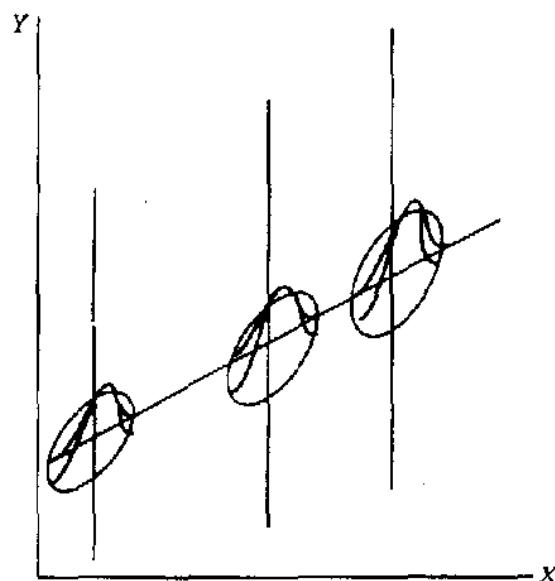
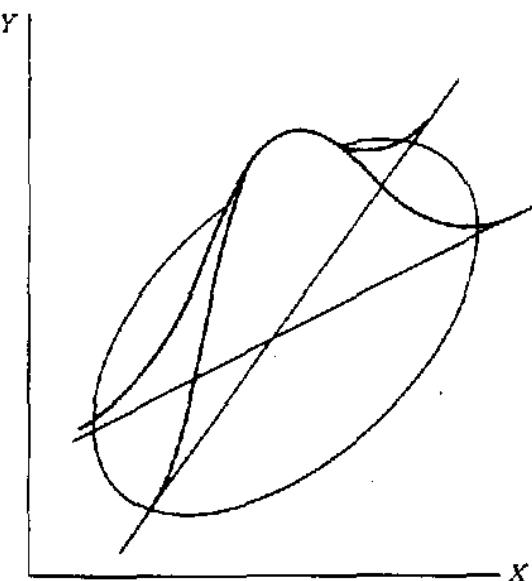
Figura 10.3 Verdadera regresión de  $Y$  con respecto a  $X$ ; error de medición en  $X_i$

Segundo, si los  $X$  también se miden con error tal como aparece en la fig. 10.4.b, entonces se debe visualizar una distribución bivariante en cada punto sobre la verdadera recta. Para estimar esa relación funcional fundamental debe usarse otro procedimiento de cálculo; ver Bartlet (10.6) o Natrella (10.7).

Tercero, existe una relación o asociación estadística entre  $X$  y  $Y$ . Inicialmente, es apropiada una sola distribución bivariante en el plano  $X, Y$ . Sin embargo,  $Y$  es restringida más que aleatoria, tal como se ve en la fig. 10.4.c. En consecuencia, sólo hay una regresión con sentido por estimar, la de  $Y$  respecto de  $X$ . El error de medición en  $X$  o en  $Y$  es probablemente insignificante con respecto al intervalo de los  $X$  escogidos o a la variación aleatoria en  $Y$ . Los métodos de este capítulo son apropiados para lo dicho.

Para el modelo II, tanto  $X$  como  $Y$  son aleatorios como se ve en la fig. 10.4.d. Este es el problema clásico de regresión bivariante, en el que se supone la normalidad. El muestreo aleatorio es de individuos en los cuales se efectúan pares de medias. La elección de cuál ha de ser la variable dependiente la determina el problema. Son posibles dos líneas de regresión la de  $Y$  respecto  $X$ , y la de  $X$  respecto de  $Y$ .

Para la regresión lineal se supone que los  $\epsilon$  se distribuyen normal e independientemente con una varianza común. Cuando es válido el supuesto de una varianza común, se aplica el cuadrado medio del error residual para hacer inferencias probabilísticas válidas respecto a una media poblacional, independientemente del valor de  $X$ . Este cuadrado medio se calcula a partir de las desviaciones respecto de la recta de regresión, también llamados residuos. Si las varianzas no son homogéneas, entonces es necesaria una regresión ponderada o una transformación de los datos de tal manera que las varianzas sean homogéneas;

(a) Errores de observación en  $Y$ (c) Restricciones sobre  $Y$ (b) Errores de observación en  $X$  y  $Y$ 

(d) Superficie bivariante (normal)

Figura 10.4 Modelos gráficos para modelos de regresión.

por ejemplo, los entomólogos calculan análisis prohit basados en porcentajes de mortalidad, donde la varianza es binomial por naturaleza, usan tanto la transformación como la ponderación.

Los datos de las gallinas Leghorn podrían provenir de un modelo fijo o de un modelo aleatorio, siendo el último una posibilidad real. Necesitamos saber si el peso del cuerpo fue restringido de alguna manera para estos datos.

Una vez estimados  $\alpha$  y  $\beta$ , es posible estimar la media de una población de  $Y$  sin haber observado uno solo de los individuos. Por ejemplo, no observamos ningún  $Y$  para

$X = 5.5$  lbs para los datos de las gallinas White Leghorn. Sin embargo, estimamos la media de la población de los  $Y$  para  $X = 5.5$  lbs del peso del cuerpo, usando la ec. (10.4) así

$$\hat{Y}_{5.5} = 55.26 + 7.69(5.5) = 97.6 \text{ lb de alimentos}$$

Como la notación  $\hat{Y}_{5.5}$  puede hacer creer fácilmente que se ha observado una muestra de los  $Y$  con  $X = 5.5$  lbs, se acostumbra otra manera de denotar una estimación de una media poblacional. Usamos  $\hat{Y}_x$  o simplemente  $\hat{Y}$  o  $\hat{\mu}_{Y,x}$ .

Las estimaciones  $\alpha$  y  $\beta$  se escriben como  $\hat{\alpha}$  y  $\hat{\beta}$  o como  $a$  y  $b$ ;  $\hat{\alpha} = a = 55.26$  lbs de alimento y  $\hat{\beta} = b = 7.69$  lbs de alimento por libra de peso del cuerpo.

La solución del problema de la regresión lineal tiene las siguientes propiedades:

1. El punto  $(\bar{X}, \bar{Y})$  se encuentra sobre la recta de regresión muestral.
2. La suma de las desviaciones respecto de la recta de regresión es 0, esto es,  $\sum (Y_i - \hat{Y}_i) = 0$ . Una desviación o residuo asignada entre el valor observado y la estimación correspondiente de la media poblacional. También, la siguiente suma ponderada es cero:  $\sum X_i(Y_i - \hat{Y}_i) = 0$ .
3. La suma de cuadrados de los residuos es un mínimo. Esto es, si reemplazamos la recta de regresión muestral calculada como en la sec. (10.2) por cualquier otra recta, la suma de cuadrados del nuevo conjunto de residuos tendrá un valor mayor.

**Ejercicio 10.3.1** ¿Qué modelo se aplica a los datos del ejercicio 10.2.1?

**Ejercicio 10.3.2** Para los datos del ejercicio 10.2.1, ¿cuál es la variable dependiente? ¿Cuál es la independiente? Considerar el requisito de que "X debe medirse sin error" según se aplica a estos datos.

**Ejercicio 10.3.3** Calcular las diez desviaciones con respecto de la recta de regresión, ejercicio 10.2.1. Demuéstrese que la suma de estos diez residuos es cero dentro de los errores de aproximación. Hallar las sumas de cuadrados de las diez desviaciones respecto de la recta de desviación. Demuéstrese que el punto  $(\bar{X}, \bar{Y})$  está en la recta de regresión.

**Ejercicio 10.3.4** ¿Qué modelo se aplicaría a los datos del ejercicio 10.2.3 cuando  $Y$  se relaciona por regresión con  $X_1$ ?,  $X_2$ ?,  $X_3$ ?,  $X_4$ ? En cada caso, dar alguna justificación de elección de modelo.

**Ejercicio 10.3.5** Repetir el ejercicio 10.3.4 con los datos del ejercicio 10.2.4.

**Ejercicio 10.3.6** Repetir el ejercicio 10.3.4 con los datos del ejercicio 10.2.5.

#### 10.4 Fuentes de variación en la línea de regresión lineal

El modelo de regresión lineal, ec. (10.5), considera una observación como la suma de una, media  $\mu_{Y,x} = \alpha + \beta X$  y una componente aleatoria  $\varepsilon$ . Ya que por azar o intencionalmente se observan valores diferentes de  $X$ , intervienen medias de formación diferentes y contribuyen a la varianza total. Así, las dos fuentes de variación en las observaciones son

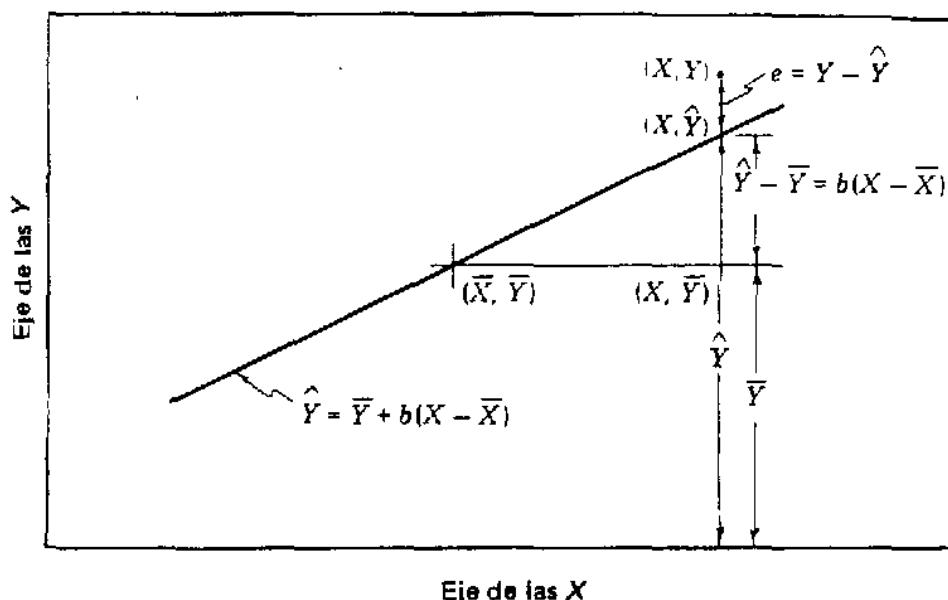


Figura 10.5 Regresión de  $Y$  con respecto a  $X$ ; fuentes de variación en  $Y$ .

medias y componentes aleatorias. La variación atribuible a las medias puede considerarse atribuible a  $X$  ya que  $X$  determina la media.

En términos de la regresión muestral, una observación  $Y$  está compuesta de una media muestral  $\bar{Y}$  determinada por la recta de regresión y una desviación muestral o residual  $e = Y - \hat{Y}$  respecto de esta media (ver fig. 10.5).  $\hat{Y}$  mismo consiste en la media muestral  $\bar{Y}$ , y otra desviación  $\hat{Y} - \bar{Y} = b(X - \bar{X})$ , esta última atribuible a la regresión. El residuo  $e = Y - \hat{Y}$  es una estimación de las desviaciones aleatorias  $\varepsilon$  y también puede escribirse  $e_{Y \cdot X}$ . Así, la ec. (10.4) se convierte en

$$Y - \bar{Y} = (\hat{Y} - \bar{Y}) + (Y - \hat{Y}) = b(X - \bar{X}) + e_{Y \cdot X} \quad (10.6)$$

La suma total de cuadrados no ajustada de los  $Y$  puede particionarse de acuerdo con estas fuentes. La suma de cuadrados atribuible a la media es  $n\bar{Y}^2 = (\sum Y)^2/n$ ; la correspondiente a la regresión es  $b^2 \sum (X - \bar{X})^2 = [\sum (X - \bar{X})(Y - \bar{Y})]^2 / \sum (X - \bar{X})^2$ ; y la correspondiente al azar es  $\sum e_{Y \cdot X}^2$ , la cual se calcula como una suma residual de cuadrados.

La ecuación (10.7) se deriva de la ec. (10.6) y es comparable a su contraparte de población, ec. (10.5).

$$Y = \bar{Y} + b(X - \bar{X}) + e_{Y \cdot X} \quad (10.7)$$

Por la ecuación (10.7) puede demostrar que

$$\sum Y^2 = n\bar{Y}^2 + b^2 \sum (X - \bar{X})^2 + \sum e_{Y \cdot X}^2$$

Esto

$$\sum Y^2 - \frac{(\sum Y)^2}{n} = \frac{[\sum (X - \bar{X})(Y - \bar{Y})]^2}{\sum (X - \bar{X})^2} + \text{SC residual.}$$

La suma de cuadrados residual se calcula por diferencia. Por ejemplo,  $135.604 = 90.836 + 44.768$ . La suma total de cuadrados de  $Y$ ,  $\sum (Y - \bar{Y})^2 = 135.604$ , se ha particionado en una suma de cuadrados atribuible a la regresión y una porción no explicada, la suma residual de cuadrados.

La suma de cuadrados atribuible a la regresión respecto de  $X$  puede escribirse en una de las formas siguientes

$$\begin{aligned} SC(Y|X) &= \frac{[\sum (X - \bar{X})(Y - \bar{Y})]^2}{\sum (X - \bar{X})^2} \\ &= b \sum (X - \bar{X})(Y - \bar{Y}) = b^2 \sum (X - \bar{X})^2 = \sum (\hat{Y} - \bar{Y})^2 \end{aligned}$$

de las cuales la primera (léase como suma de cuadrado de  $Y$  dado  $X$ ) dice que estamos tratando con una suma de cuadrados de  $Y$  atribuible a variación de  $X$ ; la segunda es la forma más usada para los cálculos; la tercera y la cuarta son interesantes, pero muy posiblemente producen errores de redondeo si se usan para hacer cálculos; y la última no es práctica. Esta cantidad tiene un solo grado de libertad.

La suma residual de cuadrados se encuentra por diferencia y tiene  $n - 2$  grados de libertad; el cuadrado medio de los residuos o *varianza respecto de la regresión* es una estimación del error experimental, la varianza común  $\sigma^2$ .

**Ejercicio 10.4.1** Calcular  $SC(Y|X) = CM(Y|X)$  y el cuadrado medio residual para los datos de

- a) Ejercicio 10.2.1
- b) Ejercicio 10.2.3, cuatro regresiones
- c) Ejercicio 10.2.4, dos regresiones
- d) Ejercicio 10.2.5, cuatro regresiones.

## 10.5 Valores de regresión y valores ajustados

Los valores determinados por la ecuación de regresión, *valores de regresión*,  $\hat{Y}$  son estimaciones de parámetros poblacionales, esto es, de  $\mu_{Y|X} = \alpha + \beta X_i$ . Las diferencias entre éstos y los valores observados son estimaciones de la variación en  $Y$ , que no está explicada por la variación en  $X$ . Los residuos observados se presentan en la tabla 10.2 para los datos de las gallinas White Leghorn de la tabla 10.1. La suma de sus cuadrados es de 44.22, que difiere del valor encontrado en la sección anterior, 44.768 debido a errores de redondeo.

El investigador cuidadoso siempre examina las observaciones individuales. En un problema de no regresión, estará particularmente consciente de las observaciones que presentan las más grandes desviaciones con respecto a la media. En un problema de regresión, se consideran desviaciones de la regresión para una información comparable.

Los residuos pueden ser particularmente útiles cuando se representan respecto de  $X$ . Si tienden a ser del mismo signo en ambos extremos de la gráfica y de signos opuestos en el medio, entonces queda comprobado que la respuesta no es lineal. Si sus magnitudes cambian en manera regular, por ejemplo, aumentando con  $X$ , entonces hay evidencia de heterogeneidad de la varianza y, posiblemente, también de su naturaleza. Los valores extremos o alejados pueden detectarse sin gráfica. La representación gráfica respecto de

Tabla 10.2 Consumos de regresión, residuos y consumos ajustados, para los datos de las gallinas White Leghorn †

$X$ Peso del cuerpo lb	$Y$ Consumo de alimento lb	$\hat{Y}$ Consumo de regresión lb	$e_{Y,x} = Y - \hat{Y}$ Desviaciones de la regresión lb	$e_{Y,x}^2$ Residuos al cuadrado	$93.56 + e_{Y,x}$ Consumo ajustado lb
4.6	87.1	90.6	-3.5	12.25	90.0
5.1	93.1	94.5	-1.4	1.96	92.2
4.8	89.8	92.2	-2.4	5.76	91.2
4.4	91.4	89.1	+2.3	5.29	95.9
5.9	99.5	100.6	-1.1	1.21	92.4
4.7	92.1	91.4	+0.7	0.49	94.3
5.1	95.5	94.5	+1.0	1.00	94.6
5.2	99.3	95.3	+4.0	16.00	97.6
4.9	93.4	92.9	+0.5	0.25	94.0
5.1	94.4	94.5	-0.1	.01	93.5
	935.6	935.6	0.0	44.22	935.7

† Los cálculos para esta tabla se hicieron con más decimales que los indicados y se redondearon los resultados.

otras variables, tales como el tiempo, puede ser útil en el estudio de la naturaleza de la respuesta.

A los *valores ajustados*, ec. (10.9), se les ha eliminado la contribución debido a la regresión. Es como si cada  $Y$  se moviera paralelamente a la recta de regresión muestral hasta por encima de  $X$  y se midiese entonces como un valor nuevo ajustado de  $Y$ . La última columna de la tabla 10.2 contiene los consumos ajustados, que son los esperados si todas las gallinas tienen un peso del cuerpo de  $X = 4.98$  lb. Estos se obtienen mediante la adición de los residuos a  $Y = 93.56$  lb.

$$Y \text{ ajustado} = \bar{Y} + e_{Y,x} = Y - b(X - \bar{X}) \quad (10.9)$$

La última forma es conveniente si no importa tener que calcular los residuos; restar la parte de la observación atribuible a regresión, esto es,  $b(X - \bar{X})$ . La media de los  $Y$  y la desviación aleatoria se dejan (ver fig. 10.5). Las diferencias entre los valores ajustados son idénticas a las diferencias entre residuos: solamente hemos cambiado su localización. El uso de los valores ajustados reemplaza el valor estándar móvil  $X$ , el valor de regresión, por un valor estándar fijo, el valor medio.

Las comparaciones entre medias ajustadas son muy útiles. Supóngase que se dispone de dos grupos de observaciones, por ejemplo, los datos para cada uno de los años, dos localidades o dos tipos de vivienda. Puede ser pertinente comparar las medias de grupo si se ajustan a un valor común  $X$ . Será necesario suponer un coeficiente de regresión común, un supuesto que se puede probar como la hipótesis de coeficientes de regresión homogéneos. Los procedimientos para comparar medias ajustadas se dan en el cap. 17 sobre covarianza.

**Ejercicio 10.5.1** D. Kuesel, Universidad de Wisconsin, determinó el porcentaje promedio de sustancias sólidas insolubles en alcohol  $Y$  y las lecturas del tenderómetro  $X$  para 26 muestras de arvejas de Alaska tamizadas. Las observaciones son

$Y, X: 7.64, 72; 8.08, 78; 7.39, 81; 7.45, 81; 9.56, 81; 7.96, 82; 10.81, 83; 10.70, 83; 10.56, 89; 11.75, 93; 11.56, 96; 11.74, 97; 13.72, 99; 15.08, 103; 16.26, 112; 16.79, 115; 15.40, 118; 15.90, 122; 16.30, 122; 17.56, 133; 17.38, 135; 17.90, 139; 18.80, 143; 19.90, 145; 20.10, 161; 22.01, 165$

$$\sum (X - \bar{X})^2 = 18,774.62 \quad \sum (X - \bar{X})(Y - \bar{Y}) = 2,924.50 \quad \sum (Y - \bar{Y})^2 = 489.58$$

Representar gráficamente los 26 pares de puntos. Calcular la ecuación de regresión muestral de  $Y$  respecto de  $X$ . ¿En qué unidades de medida se expresa  $b$ ? Trazar esta recta en la gráfica. Obtener información respecto a cómo se hacen las lecturas del tenderómetro y comentar sobre el supuesto referente a medir  $X$  sin error. ¿Cómo se probaría la hipótesis nula de que no existe variación de  $Y$  atribuible a la variación de  $X$ ?

**Ejercicio 10.5.2** Calcular los valores de regresión sólidos insolubles de alcohol (para  $X = 78$  y  $X = 112$ ) y las desviaciones con respecto a la regresión o residuos. Calcular los valores ajustados de  $Y$  para estos dos valores de  $X$  y compararlos

**Ejercicio 10.5.3** Representar gráficamente los pares de observaciones dados en el ejercicio 10.2.1. Trazar las rectas de regresión en la gráfica. Calcular los valores de regresión para  $X = 4.50$  y  $X = 5.13$ . Encuentre los valores correspondientes de  $b(X - \bar{X})$  y las desviaciones con respecto a la regresión, o residuos.

**Ejercicio 10.5.4** Repetir el ejercicio 10.5.3 con los datos del ejercicio 10.2.3. Para  $X_1 = 35$  y  $X_1 = 54$ . Para  $X_2$ , use  $X_2 = 60$  y  $X_2 = 125$ . Para  $X_3$ , use  $X_3 = 248$  y  $X_3 = 303$ . Para  $X_4$ , use  $X_4 = 60$  y  $X_4 = 70$ .

**Ejercicio 10.5.5** Repetir el ejercicio 10.5.3 con los datos del ejercicio 10.2.4. Para  $X_1 = 2,400$  y  $X_1 = 3,075$ . Para  $X_2$ , use  $X_2 = 79$  y  $X_2 = 270$ .

**Ejercicio 10.5.6** Repetir el ejercicio 10.5.3 con los datos del ejercicio 10.2.5. Para  $X_1 = 1,200$  y  $X_1 = 2,000$ . Para  $X_2$ , use  $X_2 = 545$  y  $X_2 = 605$ . Para  $X_4$ , use  $X_4 = 1,935$  y  $X_4 = 2,521$ .

## 10.6 Desviaciones estándar, intervalos de confianza y pruebas de hipótesis

Una estimación insesgada de la verdadera varianza en torno a la regresión la da el cuadrado medio de los residuos con  $(n - 2)$  grados de libertad. Se denota  $s_{Y \cdot X}^2$  y se define como

$$s_{Y \cdot X}^2 = \frac{\sum (Y - \hat{Y})^2}{n - 2} = \frac{\sum (Y - \bar{Y})^2 - [\sum (X - \bar{X})(Y - \bar{Y})]^2 / \sum (X - \bar{X})^2}{n - 2}$$

$$= \frac{44.77}{8} = 5.60 \text{ lb}^2 \quad (10.10)$$

para los datos de White Loghorn. A su raíz cuadrada se la llama *error estándar* o *desviación estándar de  $Y$  para  $X$  fijo* o la *desviación estándar de  $Y$  manteniendo  $X$  constante*.

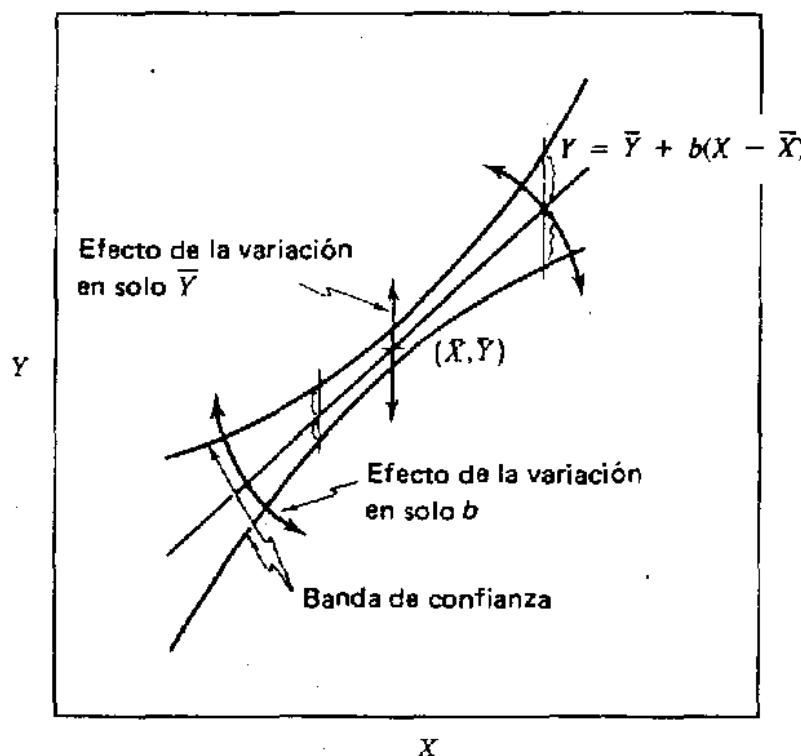


Figura 10.6 Efecto de la variación muestral sobre las estimaciones de regresión de medias poblacionales, para un conjunto fijo de  $X$ .

La figura 10.6 muestra que una sola desviación estándar no se aplica a todos los  $\hat{Y}$ , sino que debe depender del valor de  $X$  que determina la población  $Y$ . Si consideramos muestras de valores de  $Y$  para un conjunto fijo de valores de  $X$ , entonces  $\bar{X}$  es una constante mientras que  $\bar{Y}$  y  $b$  son variables. La variación de  $\bar{Y}$  sube o baja la recta de regresión paralelamente a sí misma el efecto es aumentar o disminuir todas las estimaciones de las medias para un valor fijo. La variación de  $b$  hace girar la recta de regresión en torno al punto  $\bar{X}, \bar{Y}$ , y el efecto sobre una estimación depende de la magnitud de la  $X - \bar{X}$ , que determina la población  $Y$ . La variación de  $b$  no tiene efecto en la estimación de la media si  $X = \bar{X}$ , pero en otro caso lo incrementa en proporción a la magnitud de  $X - \bar{X}$ . Esto se ve fácilmente en la ecuación que estima la media poblacional, esto es,  $\hat{Y} = \bar{Y} + b(X - \bar{X})$ .

Una desviación estándar aplicable a una estimación de una media da margen a la variación tanto en  $\bar{Y}$  como en  $b$  para la distancia  $X - \bar{X}$ . La varianza de  $\bar{Y}$  es simplemente una estimación de  $\sigma^2_{\bar{Y}} \cdot x/n$  o sea,  $s^2_{\bar{Y}} \cdot x/n$ . El coeficiente de regresión  $b$  es una función lineal de los  $Y$ . En particular,  $\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum (X_i - \bar{X})Y_i$  nos permite escribir

$$\begin{aligned}
 b &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum (X_i - \bar{X})Y_i}{\sum (X_i - \bar{X})^2} \\
 &= \sum \frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} Y_i
 \end{aligned} \tag{10.11}$$

Ahora  $b$  se expresa como una función lineal  $L$ , tal como se define en la ec. (5.18) con  $c_i = (X_i - \bar{X})/\sum(X_i - \bar{X})^2$ ;  $b$  es también una comparación, tal como se definió en la ec. (8.3). Distingase entre los dos usos de  $i$ . En el numerador,  $i$  especifica la observación  $i$ -ésima; en el denominador es un índice de sumatoria,  $i = 1, \dots, n$ . Ahora la varianza se obtiene mediante la ec. (5.23); toda  $\sigma_{ij} = 0$ ,  $i \neq j$ , ya que el muestreo es aleatorio. La varianza de  $b$  es la varianza de  $\sigma_{Y \cdot X}^2$  multiplicada por la suma de cuadrados de los coeficientes de los  $Y$ . Así, estimamos la varianza de  $b$  por

$$s_b^2 = \frac{s_{Y \cdot X}^2}{\sum(X - \bar{X})^2} \quad (10.12)$$

La varianza requerida de una estimación,  $\hat{Y} = \bar{Y} + b(X - \bar{X})$ , de una media poblacional está dada por la suma de las varianzas de  $\bar{Y}$  y  $b(X - \bar{X})$ , ya que la covarianza entre  $\bar{Y}$  y  $b$  es cero. [El último hecho no es evidente, pero puede deducirse al examinar los coeficientes de los  $Y$  en  $\bar{Y}$  y  $b$ . Para  $\bar{Y}$ , todos son iguales a la constante  $1/n$  y para  $b$  varían, pero tienen una suma igual a cero. Usese la ec. (8.7) para comprobar la ortogonalidad.] Esta varianza está dada por

$$s_{\hat{Y}}^2 = s_{Y \cdot X}^2 \left[ \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum(X - \bar{X})^2} \right] \quad (10.13)$$

Esta varianza aumenta al aumentar  $X - \bar{X}$ . Si el  $t$  tabulado multiplicado por la desviación estándar se representara junto con la recta de regresión, formaría una banda como se ve en la fig. 10.6. Esta banda es aplicable punto por punto, y la tasa de error y el coeficiente de confianza asociados son del tipo de comparación o puntual.

Una *banda de confianza* para la recta de regresión, en la cual se dice que cae la recta de regresión entera, exige que la desviación típica implicada por la ec. (10.13) se multiplique por  $\sqrt{2F_{2, n-2}}$  para el  $F$  tabulado. En este caso, son aplicables una tasa experimental de error y un coeficiente de confianza.

Para construir un intervalo de confianza para la media poblacional, es aplicable la varianza dada en la ec. (10.3) y el intervalo de confianza del 95 por ciento es

$$IC(\mu_{Y \cdot X}) = \bar{Y} + b(X - \bar{X}) \pm t_{0.025} s_{Y \cdot X} \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum(X - \bar{X})^2}} \quad (10.14)$$

donde  $t$  es la  $t$  de Student para  $n - 2$  grados de libertad. Obsérvese que  $\bar{Y} + b(X - \bar{X})$  puede reemplazarse por  $a + bX$ .

Para los datos de las gallinas White Leghorn, un intervalo de confianza de la media poblacional para  $X = 5.5$  lb es

$$\begin{aligned} IC(\mu_{Y \cdot X=5.5}) &= 93.56 + 7.69(5.5 - 4.98) \pm 2.306(2.37) \sqrt{\frac{1}{10} + \frac{(0.52)^2}{1.536}} \\ &= 97.6 \pm 2.9 \\ &= (94.7, 100.5) \text{ lb} \end{aligned}$$

Para una banda de confianza de 95 por ciento en toda la recta de regresión, se requiere el multiplicador  $\sqrt{2F_{0.05}} (2.8) = \sqrt{2(4.46)} = 2.99$ . Nótese que este valor de  $F$  de una cola genera una banda. Los límites sobre la banda cuando  $X = 5.5$  son

$$93.56 + 7.69(5.5 - 4.98) \pm 2.99(2.37) \sqrt{\frac{1}{10} + \frac{(0.52)^2}{1.536}}$$

$$= 97.6 \pm 3.7 = (93.9, 101.3) \text{ lb}$$

Estos límites están sólo ligeramente más allá de la recta de regresión que los de la estimación de los puntos del puntual  $\mu_{Y \cdot X=5.5}$ . Sin embargo, la anchura de la banda ha pasado de  $2(2.9) = 5.8$  a  $2(3.7) = 7.4$  lbs. En promedio, el 95 por ciento de tales bandas contendrán toda la verdadera recta de regresión.

A veces es deseable construir un intervalo de confianza para  $\beta$ , el parámetro de regresión poblacional estimado por  $b$ . Este intervalo depende de  $s_b^2$  tal como se ve en la ec. (10.12), y está dado por

$$\text{IC}(\beta) = b \pm \frac{t_{0.025} s_{Y \cdot X}}{\sqrt{\sum (X - \bar{X})^2}} \quad (10.15)$$

Para los datos de White Leghorn, tenemos

$$\begin{aligned} \text{IC}(\beta) &= 7.69 \pm \frac{2.306(2.37)}{\sqrt{1.536}} = 7.69 \pm 4.41 \\ &= 3.28, 12.10 \end{aligned}$$

Concluimos que  $\beta$  cae entre 3.28 y 12.10 lbs de alimento por libra de gallina. Obsérvese el efecto de la pequeña  $\sum (X - \bar{X})^2$  sobre los límites de  $\beta$ . El procedimiento que nos lleva a esta conclusión es tal, que, en promedio, el 95 por ciento de tales conclusiones son correctas.

Cualquier hipótesis nula pertinente acerca de una media se puede probar. Para probar la hipótesis nula de que la media de la población de los  $Y$  para los cuales  $X = X_0$  es  $\mu_{Y \cdot X_0}$  calcúlese  $t$  así:

$$t = \frac{\hat{Y}_{X_0} - \mu_{Y \cdot X_0}}{\sqrt{s_{Y \cdot X}^2[(1/n + (X_0 - \bar{X})^2)/\sum (X - \bar{X})^2]}} \quad (10.16)$$

Esta se distribuye como la  $t$  de Student con  $n - 2$  grados de libertad.

Para probar la hipótesis nula que  $\beta = \beta_0$ , calcúlese  $t$  como

$$t = \frac{b - \beta_0}{\sqrt{s_{Y \cdot X}^2 / \sum (X - \bar{X})^2}} \quad (10.17)$$

Esta se distribuye como la  $t$  de Student con  $(n - 2)$  grados de libertad. La prueba  $F$  de la tabla 10.3 es prueba de la hipótesis nula  $\beta = 0$  o de la hipótesis nula de que la variación en  $X$  no contribuye a la variación de  $Y$ .

**Ejercicio 10.6.1** Para los datos del ejercicio 10.2.1, encontrar  $s_{Y \cdot X}$ ,  $s_Y^2$ ,  $s_b^2$ ,  $\hat{Y}$  en  $X = 5.00$ , en  $X = 5.00$ , el intervalo de confianza de 95 por ciento para la media poblacional en  $X = 5.00$ , y los límites de la banda de confianza de 95 por ciento respecto a la verdadera recta de regresión en  $X = 5.00$ . (Recuérdese que los intervalos de confianza tienen una tasa de error por comparación o puntual, mientras que las bandas de confianza tienen tasas experimentales de error. Calcular y comparar las longitudes del intervalo de confianza y de la banda de confianza en  $X = 5.00$ . Pruebe la hipótesis nula de que  $\beta = 0$ ).

**Ejercicio 10.6.2** Repetir el ejercicio 10.6.1 con los datos del ejercicio 10.2.3. Para  $X$  usar  $X_1 = 35$ . Usar  $X_2 = 60$ . Usar  $X_3 = 248$ . Usar  $X_4 = 60$ .

**Ejercicio 10.6.3** Repetir el ejercicio 10.6.1 con los datos del ejercicio 10.2.4. Para  $X$ , usar  $X_1 = 2,400$ . Usar  $X_2 = 79$ .

**Ejercicio 10.6.4** Repetir el ejercicio 10.6.1 con los datos del ejercicio 10.2.5. Para  $X$ , usar  $X_1 = 1,200$ . Usar  $X_2 = 545$ . Usar  $X_3 = 1,250$ . Usar  $X_4 = 1,935$ .

**Ejercicio 10.6.5** Repetir el ejercicio 10.6.1 con los datos del ejercicio 10.5.1. Para  $X$ , usar  $X = 78$ .

## 10.7 Control de la variación por observaciones concomitantes

Las fuentes de variación que afectan a una variable no siempre son controlables mediante un plan experimental. Cuando el plan no puede efectuar el control, es posible medir algunas características de la fuente de variación. Por ejemplo, la cantidad de alimento consumido por las gallinas es una variable de importancia económica. Sería de esperar que se viera afectada por otras variables medibles, tales como el peso del cuerpo y el número y peso de los huevos puestos. Para los datos de la tabla 10.1, el peso del cuerpo fácilmente explica la mayor variabilidad en el alimento consumido. La importancia económica es obvia.

Ahora usamos los datos de la tabla 10.1 para ilustrar el control estadístico de una fuente de variabilidad mediante el uso de una observación concomitante. La desviación estándar de  $Y$  antes del ajuste de la variación en  $X$  es  $\sqrt{\sum (Y - \bar{Y})^2 / (n - 1)} = \sqrt{135.604 / 9} = 3.88$  lbs. Hemos visto que luego de ajustarla es  $s_{Y \cdot X} = 2.37$  lbs.

La parte de la suma de cuadrados de  $Y$  atribuible a variación en  $X$  la da la ec. (10.18), ver también la ec. (10.8).

$$\text{Reducción en SC} = \text{SC} (\text{Regresión} = \text{SC} (Y | X) = \frac{[\sum (X - \bar{X})(Y - \bar{Y})]^2}{\sum (X - \bar{X})^2}$$

$$= \frac{(11.812)^2}{1.536} = 90.836 \quad (10.18)$$

para nuestro ejemplo. Tiene un grado de libertad. También podemos observar que la proporción de la suma de cuadrados de  $Y$  atribuible a variación en  $X$  es

$$\frac{[\sum (X - \bar{X})(Y - \bar{Y})]^2 / \sum (X - \bar{X})^2}{\sum (Y - \bar{Y})^2} = \frac{90.836}{135.604} = 0.67 \text{ (o } 67 \text{ por ciento)}$$

La suma de cuadrados de  $Y$  reducida o residual se encuentra por diferencia y tiene  $n - 2$  grados de libertad.

$$\text{SC(residual para } Y) = 135.604 - 90.836 = 44.768 \quad \text{con 8 gl.}$$

Puede utilizarse una tabla de análisis de la varianza para presentar los resultados. La tabla 10.3 muestra dos posibilidades. La primera de éstas es comparable a las tablas del cap. 8; la segunda es más completa y sirve como base para la generalización a tablas para regresión parcial y múltiple (cap. 14) y covarianza (cap. 17).

**Ejercicio 10.7.1** Presentar un análisis de la varianza para los datos del ejercicio 10.2.1; usar la forma dada en la parte superior de la tabla 10.3. ¿Hay reducción significante en la variación en  $Y$  atribuible a variación en  $X$ ? Comparar el valor de  $F$  con  $t^2$  para el contraste de  $H_0: \beta = 0$  tal como se calcula en el ejercicio 10.6.1

**Ejercicio 10.7.2** Repetir el ejercicio 10.7.1 con los datos en el ejercicio 10.2.3. Para  $t^2$ , véase el ejercicio 10.6.2.

**Ejercicio 10.7.3** Repetir el ejercicio 10.7.1 con los datos en el ejercicio 10.2.4. Para  $t^2$ , véase el ejercicio 10.6.3.

**Ejercicio 10.7.4** Repetir el ejercicio 10.7.1 con los datos del ejercicio 10.2.5. Para  $t^2$ , véase el ejercicio 10.6.4.

**Tabla 10.3 Análisis de la varianza para los datos de la tabla 10.1**

Fuente	gl	SC simbólica	Ejemplo			
			gl	SC	CM	F
X	1	$[\sum (X - \bar{X})(Y - \bar{Y})]^2 / \sum (X - \bar{X})^2$	1	90.836	90.836	16.22**
Residual	$n - 2$	por sustracción	8	44.768	5.60	
Total	$n - 1$	$\sum (Y - \bar{Y})^2$	9	135.604		

Otra presentación (solo el ejemplo)

Fuente	gl	$\sum (X - \bar{X})^2$	$\sum (X - \bar{X}) \times (Y - \bar{Y})$	$\sum (Y - \bar{Y})^2$	SC Regresión = SC(Y X)	Residual		
						SC	CM	F
Total	$n - 1 = 9$	1.536	11.812	135.604	90.836	44.768	5.60	16.22**

Ejercicio 10.7.5 Repetir el ejercicio 10.7.1 con los datos del ejercicio 10.5.1. Para  $t^2$ , véase el ejercicio 10.6.5.

### 10.8 Diferencia entre dos regresiones independientes

Puede ser deseable contrastar la homogeneidad de los  $b$ , es decir, determinar si pueden considerarse o no estimaciones de un  $\beta$  común. Para esto,  $t$  como la ec. (10.19) se distribuye como una  $t$  de Student con  $n_1 + n_2 - 4$  grados de libertad.

$$t = \frac{b_1 - b_2}{\sqrt{s_p^2 [1/\sum (X_{1j} - \bar{X}_{1.})^2 + 1/\sum (X_{2j} - \bar{X}_{2.})^2]}} \quad (10.19)$$

Las cantidades

$$b_1 \quad \text{y} \quad \sum_j (X_{1j} - \bar{X}_{1.})^2$$

son el coeficiente de regresión y la suma de cuadrados para  $X$  de la primera muestra, y análogamente para la segunda muestra

$$s_p^2 = \frac{\left\{ \sum (Y_{1j} - \bar{Y}_{1.})^2 - [\sum (X_{1j} - \bar{X}_{1.})(Y_{1j} - \bar{Y}_{1.})]^2 / \sum (X_{1j} - \bar{X}_{1.})^2 \right\} + \left\{ \sum (Y_{2j} - \bar{Y}_{2.})^2 - [\sum (X_{2j} - \bar{X}_{2.})(Y_{2j} - \bar{Y}_{2.})]^2 / \sum (X_{2j} - \bar{X}_{2.})^2 \right\}}{n_1 - 2 + n_2 - 2}$$

es la mejor estimación de la variación respecto de la regresión. Se considera como las sumas combinadas de cuadrados residuales de las dos regresiones independientes divididas por los grados de libertad combinados.

La homogeneidad de la regresión dice que las dos rectas tienen la misma pendiente, pero no que sean la misma recta. Para decidir al respecto, es necesario contrastar las medias ajustadas. Esto se trata en el cap. 17.

La homogeneidad de la regresión también se puede contrastar con  $F$ . Dicho brevemente, hallar la reducción en la suma de cuadrados atribuible a  $X$  cuando se supone un solo coeficiente de regresión, y la reducción cuando se suponen dos coeficientes. El último puede ser menor que el primero. La diferencia en las dos sumas de cuadrados es una reducción adicional que ordinariamente deberá atribuirse al azar si existe un solo  $\beta$ , pero que será mayor en promedio si existen dos  $\beta$ . La fig. 10.7 ilustra el enfoque en que se basa esta afirmación.

El procedimiento se indica en la tabla 10.4. Los espacios marcados se llenan por los procedimientos de cálculo usuales. En la línea 6, columna 8, la diferencia (línea 5 - línea 4) dará la misma respuesta que en la columna 6. La extensión de este procedimiento para contrastar diferencias entre más de 2  $\beta$  se presenta en el cap. 17.

Ejercicio 10.8.1 H.L. Self (Tesis de doctorado, Universidad de Wisconsin, 1954) registró el espesor de la grasa del lomo  $Y$  y el peso en canal  $X$  en cuatro lotes de cerdos Poland China alimentados con raciones diferentes. Los datos para el tratamiento 3 son:

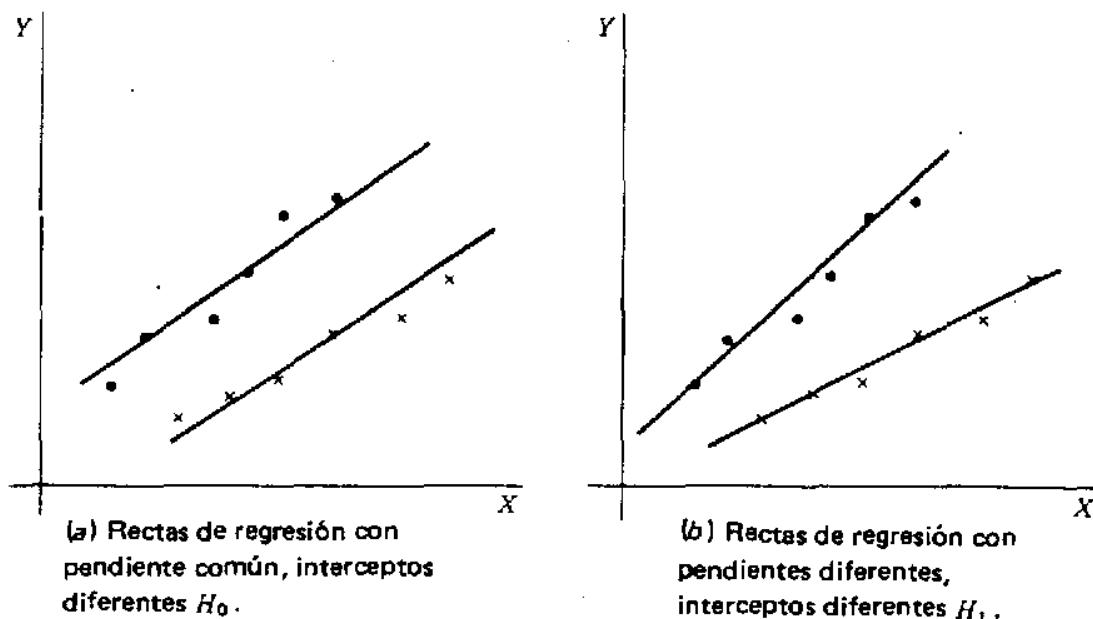


Figura 10.7 Pruebas de regresión independiente para homogeneidad.

$Y$ , mm    42, 38, 53, 34, 35, 31, 45, 43  
 $X$ , lb    206, 261, 279, 221, 216, 198, 277, 250

y, para el tratamiento 4:

$Y$     33, 34, 38, 33, 26, 28, 37, 31  
 $X$     167, 192, 204, 197, 181, 178, 236, 204

Calcular los dos coeficientes de regresión y probar la hipótesis nula de que son estimaciones de un  $\beta$  común.

¿Qué porcentaje de la SC(dentro de muestras) con  $n_1 + n_2 - 2$  grados de libertad puede atribuirse a un solo coeficiente de regresión en el modelo en la  $H_0$ , la hipótesis nula? ¿A dos coeficientes de regresión en la  $H_1$ , la hipótesis alterna?

**Ejercicio 10.8.2** Al suponer una varianza común, se refiere a la varianza con respecto a la regresión. Contrastar la hipótesis nula de una varianza común respecto a la regresión.

**Ejercicio 10.8.3** Los datos siguientes son  $Y$  = tiempos de carrera en minutos y segundos y  $X$  = número de saltos hasta el agotamiento. El primer conjunto corresponde a hombres que han estado en el programa de carrera durante 10 años y están cerca al final de la estación de carreras. †

$Y$	11:16	12:30	11:30	10:17	11:48	9:29	11:06	12:02	11:52	11:28
$X$	45	60	40	101	60	80	51	50	99	34

† Datos cortesía de A.C. Linnerud, Universidad del Estado de Carolina del Norte.

Tabla 10.4 Tabla de análisis de la varianza para las diferencias entre dos  $\beta$ 

Línea \ Columna	Cálculos usados	1	2	3	4	5	6	7	8	9
1	Fuente de variación	$g_1$	$\sum (x - \bar{x})^2$	$\sum (x - \bar{x})(y - \bar{y})$	$\sum (y - \bar{y})^2$		$\frac{\sum (x - \bar{x})(y - \bar{y})^2}{\sum (x - \bar{x})^2}$	$g_1$	$= \text{col } 5 - \text{col } 6$	$g_1$
2	$H_1$	Dentro de la muestra 1	$n_1 - 1$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$n_1 - 2$
3	$H_1$	Dentro de la muestra 2	$n_1 - 1$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$n_2 - 2$
4	$H_1$	(Dos regresiones)					Subtotal	2	Subtotal	$n_1 + n_2 - 4$
5	$H_0$	Dentro de 1 + dentro de 2 (Una regresión)				$\checkmark$				
6	$H_1$ vs. $H_0$	Coeficiente de regresión (Dos regresiones frente a una)				$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	
						Línea 4	Línea 5	1		
						Línea 6, columna 6				
						$F = \frac{\text{Línea 6, columna 6}}{(\text{Línea 4, col 8 entrada})/(n_1 + n_2 - 4)}$				

El segundo conjunto corresponde a hombres que se unieron al programa más tarde, pero que están en la misma etapa de la estación de carreras.

$Y$	11:34	13:21	10:39	10:14	9:27	9:56	11:47	12:12	12:20	11:03	11:24	10:21
$X$	125	40	123	92	93	100	70	57	67	101	70	85

Calcular los dos coeficientes de regresión de  $Y$  respecto a  $X$  y probar  $H_0: \beta_1 = \beta_2$ . Se presume que se ha supuesto una varianza común al efectuar las pruebas precedentes. ¿Respalda los datos tal supuesto?

### 10.9 Una predicción y su varianza

Entre los usos de la regresión, está la predicción tal como se sugirió en la sec. 3.12. A veces se desea decir algo respecto a un valor futuro particular, tal como el tiempo de la primera helada, la máxima precipitación aluvial, etc. O puede ser posible observar un valor de  $X$  e imposible o nada práctico observar el correspondiente valor de  $Y$ . En realidad no es tanto que se deseé decir algo acerca de una media como acerca de cuánto puede discrepar de ella la futura observación. Realmente, se desearía predecir la componente aleatoria y esta, como es natural, no es predecible por definición.

Si se conociera la media de población, esta sería la predicción. Pero asociaríamos con ésta un múltiplo de la desviación estándar de las observaciones para dar un intervalo; el múltiplo dependerá de qué seguridad deseamos tener de que el intervalo contenga la futura observación. Lo más probable es que se desconozca la media, y la predicción será una estimación de ella. Esta es una variable y se debe considerar su variación tanto como la de la componente aleatoria que estamos tratando realmente de tener en cuenta. Así, la varianza de un  $Y$  predicho está dada por

$$\begin{aligned} V(Y \text{ pred.}) &= \sigma_{Y \cdot X}^2 + \frac{\sigma_{Y \cdot X}^2}{n} + (X - \bar{X})^2 \frac{\sigma_{Y \cdot X}^2}{\sum (X - \bar{X})^2} \\ &= \sigma_{Y \cdot X}^2 \left( 1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum (X - \bar{X})^2} \right) \end{aligned} \quad (10.20)$$

Esto se estima mediante

$$s^2(Y \text{ pred.}) = s_{Y \cdot X}^2 \left( 1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum (X - \bar{X})^2} \right) \quad (10.21)$$

Esta varianza se aplica a la siguiente situación: tómese una muestra aleatoria de  $Y$  para un conjunto de  $X$ , calcúlese la ecuación de regresión, y luego tómese un  $Y$  de esta recta en un valor particular  $X$ . Selecciónese también al azar un nuevo  $Y$  con el mismo  $X$ . Nuestro interés está en la variabilidad de diferencias,  $(Y - \hat{Y})$ , cuando todo el proceso se repite indefinidamente, esto es, en  $V(Y - \hat{Y})$ .

Es de interés un intervalo de confianza para un  $\hat{Y}$  predicho. Si bien en general construimos intervalos de confianza para parámetros, muchas situaciones prácticas los hacen apropiados para las predicciones. Para establecer un intervalo de confianza de 95 por ciento de un valor de predicción, calcúlese

$$IC(\hat{Y} \text{ pred.}) = \hat{Y} + b(X - \bar{X}) \pm t_{0.025} s_{\hat{Y} \cdot X} \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum (X - \bar{X})^2}} \quad (10.22)$$

donde  $t_{0.025}$  es el  $t$  tabulado para  $n - 2$  grados de libertad. Téngase presente que este  $\hat{Y}$  está todavía por obtenerse y que no se usó para hallar la ecuación de regresión que dio  $\hat{Y}$ . Para cualquier otro coeficiente de confianza, reemplácese  $t_{0.025}$  por  $t$  para el nivel de probabilidad apropiado.

Considérese la tabla 10.5. Se desea predecir el número de caballos para 1949 en el supuesto de que la disminución en número es lineal a lo largo de los años.

La ecuación de regresión es

$$\hat{Y} = 2,291.2 - 221.5(X - 1946) \quad (10.23)$$

$s_{\hat{Y} \cdot X} = 77.6$  caballos. En la fig. 10.8 se presentan los datos y la ecuación de regresión.

Para predecir el número de miles de caballos para 1949, sustitúyase  $X = 1949$  en la ec. (10.23). Encontramos  $\hat{Y}_{1949} = 1,627$  caballos. La desviación estándar aplicable a este resultado es

$$s_{\hat{Y} \cdot X} \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum (X - \bar{X})^2}} = 77.6 \sqrt{1 + \frac{1}{5} + \frac{3^2}{10}} = 112.5 \text{ caballos}$$

El intervalo de confianza del 95 por ciento es

$$1,627 \pm 3.182(112.5) = (1,269, 1,985)$$

En 1949, el número de caballos era 1,796, lo cual cae dentro del intervalo de confianza, pero  $1,796 - 1,627 = 169$  miles de caballos más que la estimación puntual.

Si nuestra muestra es poco usual y predecimos muchas observaciones, puede ocurrir que ninguna de las futuras observaciones efectivas caigan dentro del intervalo de predicción. Esto pone de presente la falacia de usar una recta de regresión con una tasa de error por comparación o puntual para hacer predicciones una y otra vez. Si el muestreo aleatorio no se repite, lo menos que se puede hacer es revisar la recta de regresión a medida que se acumula experiencia.

Para predecir una media futura de  $n_1$  observaciones, se usa el valor de la regresión para la media predicha. La varianza adecuada para la predicción está dada por el remplazo del 1 del paréntesis de las ecs. (10.20) y (10.21) por  $1/n_1$ .

Al predecir un valor o al estimar una media para un  $X$  fuera del intervalo observado, esto es, al extrapolar, (lo opuesto a interpolar), se supone que la relación se mantiene

**Tabla 10.5 Caballos en fincas canadienses, junio, 1944-1948**

$X = \text{año}$	$(X - \bar{X})$	$Y = \text{número de caballos en fincas canadiense en miles}$
1944	-2	2,735
1945	-1	2,585
1946	0	2,200
1947	+1	2,032
1948	+2	1,904

$$\sum (X - \bar{X})^2 = 10 \quad \sum (X - \bar{X})(Y - \bar{Y}) = -2,215$$

$$\sum (Y - \bar{Y})^2 = 508,702.8$$

$$\bar{Y} = 2,291.2 \text{ caballos}$$

$$b = -221.5 \text{ caballos por año}$$

$$SC(b|a) = \frac{[\sum (X - \bar{X})(Y - \bar{Y})]^2}{\sum (X - \bar{X})^2} = 490,622.5$$

$$CM \text{ residual} = \frac{18,080.3}{3} = 6,026.8$$

Fuente: Datos del *Canadá*, 1950, pág. 127.

lineal. Este supuesto puede no ser acertado, especialmente si se usa la recta como una aproximación y se extrapolan puntos muy alejados del intervalo. En el caso que nos ocupa, la interpolación no sería muy informativa, ya que los datos probablemente se recolectaron durante el año y no en una fecha fija.

**Ejercicio 10.9.1** Supóngase que un lote de 50 gallinas en el concurso de muestra aleatoria de Nueva York (tabla 10.1) da un peso promedio de 5.3 libras. ¿Cuál es el intervalo de confianza de 95 por ciento para predicción de consumo de alimentos entre 150 días para este lote de gallinas? ¿Cuál es el intervalo de confianza de 95 por ciento para el consumo de alimento medio para lotes con un peso promedio de 5.3 libras?

**Ejercicio 10.9.2** ¿Cuál es la desviación estándar de un  $Y$  predicho en  $X = 5.00$  para los datos del ejercicio 10.2.1?

¿De un  $Y$  predicho en cada uno de los valores  $X_1 = 35, X_2 = 60, X_3 = 248$ , y  $X_4 = 60$  para los datos del ejercicio 10.2.3?

¿De un  $Y$  predicho en cada uno de los valores  $X_1 = 2,400$  y  $X_2 = 79$  para los datos del ejercicio 10.2.4?

¿De un  $Y$  predicho en cada uno de los valores  $X_1 = 1,200, X_2 = 545, X_3 = 1,250$  y  $X_4 = 1,935$  para los datos del ejercicio 10.2.5?

¿De un  $Y$  predicho en  $X = 78$  para los datos del ejercicio 10.5.1?

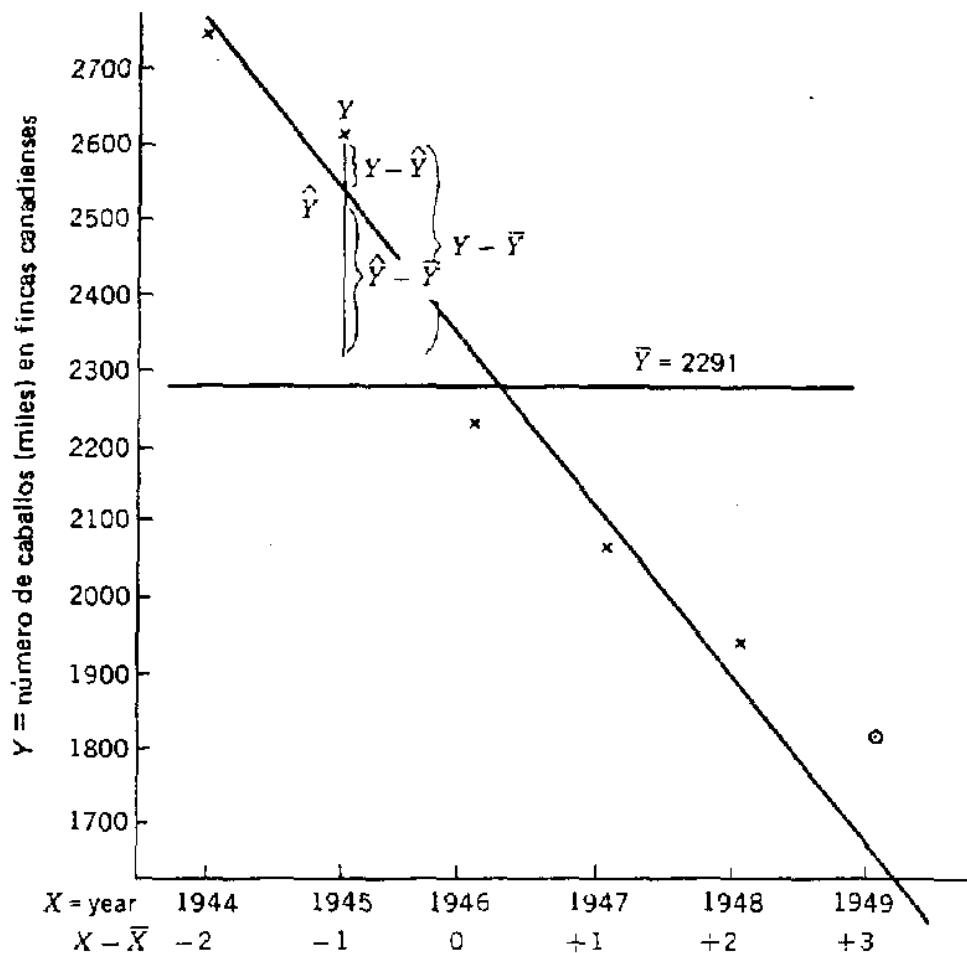


Figura 10.8 Datos y recta de regresión de los datos de la tabla 10.5

### 10.10 Predicción de $X$ , modelo I

A veces es necesario predecir un valor  $X$  o estimar la media de una población de  $X$ , aunque  $X$  sea una variable fija en los datos disponibles. Esta es una situación común en ensayos biológicos y problemas de dosis letales. En tales circunstancias, se predice o estima despejando  $X$  en la ecuación de regresión de  $Y$  con respecto a  $X$ ; así

$$X - \bar{X} = \frac{Y - \bar{Y}}{b} \quad \text{o} \quad X = \bar{X} + \frac{Y - \bar{Y}}{b}$$

Esto da una estimación puntual de  $X$  donde usualmente se prefiere una estimación de intervalo de confianza. Es posible calcular un intervalo de manera directa, pero más larga que el cálculo de un valor de la variable dependiente. El lector interesado en explicaciones adicionales refiérase a Bliss (10.1), Eisenhart (10.2), Finney (10.3), Irwin (10.4) y Natrella (10.7).

### 10.11 Distribuciones bivariantes, modelo II

En muchos casos, se observan al azar pares de valores ( $X, Y$ ); es decir, no se ejerce control

dos, tomando los individuos al azar y haciendo un par de medidas; o considérese un experimento para determinar rendimientos de papa dentro de una variada gama de condiciones de terreno, donde también se observa la tasa de infestación de nemátodos. En tales casos, el muestreo se hace de una distribución bivariante.

Considérense los pares de observaciones de la tabla 10.6 de una sola cepa de guayule, una planta de la cual se extrae caucho. Las variables son peso del arbusto secado al horno y la circunferencia de la corona; los datos son sólo una porción de los originales. Como variable dependiente, elíjase aquella para la cual van a estimarse las medias, predecir valores, o para la cual se desea control estadístico mediante el uso de las otras variables. Así, cuando la circunferencia de la corona es la variable dependiente, tenemos

$$\text{SC regresión} = \text{SC}(b|a) = \frac{(176.84)^2}{8,120.9} = 3.851$$

$$\text{SC residual} = 7.764 - 3.851 = 3.913$$

$$\text{Porcentaje de reducción} = \frac{3.851}{7.764} 100 = 49.6 \text{ por ciento}$$

$$b = \frac{176.84}{8,120.9} = .022 \text{ g/cm}$$

Cuando el peso es la variable dependiente,

$$\text{SC regresión} = \text{SC}(b|a) = \frac{(176.84)^2}{7.764} = 4,027.9$$

$$\text{SC residual} = 8,120.9 - 4,027.9 = 4,093.0$$

$$\text{Porcentaje de reducción} = \frac{4,027.9}{8,120.9} 100 = 49.6 \text{ por ciento}$$

$$b = \frac{176.84}{7.764} = 22.78 \text{ g/cm}$$

Es claro, entonces, que en un muestreo aleatorio de pares de una distribución bivariante, hay dos ecuaciones de regresión. Las dos sumas residuales de cuadrados se miden por perpendiculares a diferentes ejes. Las rectas no coinciden, como puede verse por el hecho de que los  $b$  no son inversos uno a otro. El producto de los  $b$  es la reducción en la suma de cuadrados como fracción decimal. Así,  $(22.78)(0.022) 100 = 50.1$  por ciento, que difiere de 49.6 por ciento debido a errores de redondeo. Ambas rectas pasan por la media de la muestra ( $\bar{X}, \bar{Y}$ ).

La proporción en la suma de cuadrados de la variable dependiente que puede atribuirse a la variable independiente es siempre la misma, sea cual sea la variable independiente. Esto se ve comparando fórmulas; tenemos

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

**Tabla 10.6 Peso seco en horno y circunferencia de la corona de una muestra aleatoria de plantas de guayule**

Peso seco en horno g	Circunferencia de la corona cm
65	6.5
100	6.3
82	5.9
133	6.3
133	7.3
165	8.0
116	6.9
120	8.1
150	8.7
117	6.6
$\bar{X}: 118.1$	7.06
$\sum (X - \bar{X})^2: 8,120.9$	7.764
$\sum (X_1 - \bar{X}_1)(X_2 - \bar{X}_2) = 176.84$	

*Fuente:* Datos cortesía de W.T. Federer, Universidad de Cornell, Ithaca, Nueva York.

Esta cantidad, a menudo se denota por  $r^2$  y se llama *coeficiente de determinación* y se tratará más detalladamente en el capítulo siguiente.

Cuando el muestreo no se hace en una distribución bivariante, aún es imposible estimar las medias y predecir los valores para cualquiera de las variables dependientes o independientes. Esto se hace con una línea de regresión, la regresión de la variable dependiente con relación a la independiente. Las referencias al final de la sec. 10.10 tienen que ver con el problema de un intervalo de confianza para la estimación o predicción de valores de la variable independiente.

**Ejercicio 10.11.1** Calcular el coeficiente de determinación para los datos del ejercicio 10.8.1. Interpretar este coeficiente.

### 10.12 Regresión a través del origen

En algunas situaciones, la teoría exige el paso de una línea recta a través del origen. En tales casos, se nos da un punto sobre la línea, un punto para el cual no existe variación de muestreo. Claramente tal punto debe tratarse en forma diferente a uno observado.

Como ejemplo de una regresión a través del origen, considérense los datos de la tabla 10.7. Dado que la línea de regresión debe pasar a través del origen, la ecuación correspondiente puede escribirse como  $Y = bX$ . El coeficiente de regresión está dado por

$$b = \frac{\sum XY}{\sum X^2} \quad . \quad (10.24)$$

= 3.67 retornos inducidos por dosis

**Tabla 10.7 Reversiones inducidas a la independencia por  $10^7$  células sobrevivientes y por dosis (ergs/bacterias)  $10^{-5} X$  de *Escherichia coli* estreptomiceno-dependiente sometida a radiación ultravioleta monocromática de  $2,967 = \text{Å}$  longitud de onda**

X	Y
13.6	52
13.9	48
21.1	72
25.6	89
26.4	80
39.8	130
40.1	139
43.9	173
51.9	208
53.2	225
65.2	259
66.4	199
67.7	255

$$\begin{aligned}\sum X &= 528.8 \\ \sum X^2 &= 26,062.10 \\ \sum (X - \bar{X})^2 &= 4,552.14 \\ \sum XY &= 95,755.7\end{aligned}\quad \begin{aligned}\sum Y &= 1,929 \\ \sum Y^2 &= 356,259 \\ \sum (Y - \bar{Y})^2 &= 70,025 \\ \sum (X - \bar{X})(Y - \bar{Y}) &= 17,289.9\end{aligned}$$

$$t = a/s_a = .468 \text{ ns}$$

*Fuente:* Datos cortesía de M.R. Zelle, Universidad de Cornell, Ithaca, N.Y.

La línea de regresión es

$$\hat{Y} = 3.67X$$

La suma de las desviaciones con respecto a esta línea no es cero. La reducción de la suma de cuadrados debida a la regresión es  $(\sum XY)^2 / \sum X^2 = 351,819$ . La suma de cuadrados residual es 4,440 con 12 grados de libertad y el cuadrado medio residual para  $Y$  es 370. Dado que no se han hecho ajustes para la media,  $\sum Y^2$  tiene 13 grados de libertad = número de observaciones, y el cuadrado medio residual tiene 12 grados de libertad.

Toda hipótesis a propósito del valor estimado por  $b$  puede probarse con una prueba de  $t$  de Student. Una prueba de la hipótesis  $\beta = 0$  también la da  $F = SC(\text{regresión})/CM$  residual.

Cuando existen reservas en cuanto al supuesto de que la regresión pasa por el origen, puede ser deseable contrastar esto como una hipótesis. Para ello, calcúlese la recta de regresión  $\hat{Y} = \bar{Y} + b(X - \bar{X})$ . Está dada por

$$\hat{Y} = 148.4 + 3.80(X - 40.68) \quad (10.25)$$

La reducción atribuible a la media es  $(\sum Y)^2/n = 286,234$  y a  $b$  es  $[\sum (X - \bar{X})(Y - \bar{Y})]^2 / \sum (X - \bar{X})^2 = 65,670$ . La reducción total es  $(286,234 + 65,670) = 351,904$ ; la suma

de cuadrados residual es 4,355 con 11 grados de libertad y el cuadrado medio residual para  $Y$  es 396. Este es efectivamente mayor que el cuadrado medio para la regresión que pasa por el origen, aunque la suma de cuadrados es necesariamente menor.

Para probar la hipótesis de que la regresión pasa por el origen, calcúlese la regresión adicional debida al ajuste de la media. Esta es  $(286,234 + 65,670) - (351,819) = 85$  o  $4,440 - 4,355 = 85$  con un solo grado de libertad y es claro que no es significante. (El último cálculo es la diferencia entre las dos sumas residuales de cuadrados.) El cuadrado medio del error apropiado es  $4,355/11 = 396$ .

El análisis de la varianza de la tabla 10.8 es resumen de los cálculos. Nótese que SC(total) tiene 13 grados de libertad, ya que no se ha hecho ningún ajuste inicial para la media. La línea superior de la suma de cuadrados asociados con el ajuste de la ecuación de regresión; los cálculos consisten en la adición del par inferior de líneas entre paréntesis que asociamos con el problema usual de la regresión como contribuciones atribuibles a la pendiente en forma sucesiva. El par superior de líneas entre paréntesis tiene que ver con el problema que nos ocupa; primero tenemos en cuenta sólo la pendiente, con la regresión por el origen y luego sigue considerar cuánto puede mejorar el modelo por un intercepto diferente de cero.

Este procedimiento es el general introducido en la sec. 9.6 como método exacto de amplia aplicación y usado de nuevo en la sec. 10.8. Exige el ajuste de dos modelos, uno con  $H_0$ , el otro con  $H_1$ . Se busca entonces la mejoría observando la suma adicional de cuadrados atribuible a la introducción de un parámetro o conjunto adicional en el modelo. Este procedimiento general es informativo aunque no es necesario acá. En vez de esto, podemos usar las ecs. (10.4), (10.13) y (10.14) con  $X = 0$ , como es lo apropiado.

Como en el caso de regresión general, se supone que las desviaciones con respecto a la recta de regresión se distribuyen normalmente con una varianza común. Mayor número de datos de los que presentan aquí indican que la varianza es probablemente una función de la dosis y de la longitud de onda. En este caso, es adecuado un análisis de regresión ponderada.

**Ejercicio 10.12.1** Completar la prueba de  $F$  de  $H_0: \alpha = 0$  utilizando la tabla 10.8.

Calcular el intercepto a partir de la ec. (10.25) y usar la  $t$  de Student para probar la hipótesis nula de que es diferente de cero debido al error de muestreo. Elevar  $t$  al cuadrado y comparar con el valor  $F$  calculado previamente.

**Tabla 10.8 Análisis de la varianza para los datos de la tabla 10.7. ¿Es 0 el intercepto?**

Fuente	Ecuación del modelo	gl	SC	CM	F
Pendiente e intercepto	$Y = \alpha + \beta X + \epsilon$	2	351,904		
Pendiente sin tener en cuenta el intercepto	$Y = \beta X + \epsilon$	1	351,819		
		1	85	85	.215 ns
Intercepto luego de la pendiente					
Intercepto sin tener en cuenta la pendiente	$Y = \mu + \epsilon$	1	286,234		
Pendiente luego del intercepto		1	65,670	65,670	165.8**
Residuo		$n - 2 = 11$	4,355	. 396	
Total		$n = 13$	356,259		

**Ejercicio 10.12.2** En el ejercicio 5.5.1, los datos incluían los tiempos de carrera de corredores experimentados clasificados en dos categorías RHR1s y RHR2s, al comienzo de una nueva temporada. En el ejercicio 5.7.2 se dieron los tiempos para los mismos individuos luego de una temporada de carreras. En el último ejercicio, la hipótesis alterna implicaba que todos los corredores mejoraron igualmente, dentro del muestreo aleatorio.

Ahora, la regresión nos permite proponer un modelo alterno. En  $H_0$ , los tiempos deberán ser iguales cuando la regresión pasa por el origen, es decir,  $Y = X + \epsilon$ . En  $H_1$ , la regresión aún debe pasar por el origen, pero ahora  $Y = \beta X + \epsilon$  donde  $\beta < 1$ .

Probar  $H_0: \beta = 0$  frente a  $H_1: \beta < 1$ .

No suponga nada respecto al valor de  $\beta$ , pero pruebe  $H_0: \alpha = 0$  frente a  $H_1: \alpha \neq 0$ .

**Ejercicio 10.12.3** Repetir el ejercicio 10.12.2 usando los RHR2s de los ejercicios 5.5.1 y 5.7.3.

### 10.13 Análisis de regresión ponderada

A veces los datos de que se dispone provienen de observaciones con varianzas desiguales. Por ejemplo, se puede necesitar la regresión de medias del tratamiento con respecto a una variable concomitante, cuando las medias del tratamiento se basan en muestras de tamaños diferentes, con varianzas  $\sigma^2/n_1, \dots, \sigma^2/n_k$ . El supuesto de varianza homogénea no se justifica y entonces hay que efectuar un análisis de regresión ponderada.

En la mayoría de los análisis de regresión ponderada, las ponderaciones dependen de las cantidades de información en las observaciones o la precisión de las observaciones. Son los inversos de las varianzas, esto es,  $w_i = 1/\sigma_i^2$  donde  $w_i$  representa la ponderación para la observación  $i$ -ésima. Si las observaciones son medias entonces  $w_i = n_i/\sigma^2$ .

Las ponderaciones importantes son las relativas, más bien que las efectivas. Por tanto, cuando las observaciones tienen una varianza común estamos calculando la regresión de las medias con respecto a otra variable, las ponderaciones son los números de observaciones.

Para una regresión ponderada, la suma total de cuadrados de los  $Y$  es ponderada. En particular,

$$\begin{aligned} SC(Y) &= \sum w_i(Y_i - \bar{Y})^2 && \text{definición} \\ &= \sum w_i Y_i^2 - \frac{(\sum w_i Y_i)^2}{\sum w_i} && \text{cálculo,} \end{aligned}$$

donde  $\bar{Y} = (\sum w_i Y_i)/\sum w_i$ , una media ponderada. El coeficiente de regresión es

$$\begin{aligned} b &= \frac{\sum w_i(X_i - \bar{X})(Y_i - \bar{Y})}{\sum w_i(X_i - \bar{X})^2} && \text{definición} \\ &= \frac{\sum w_i X_i Y_i - [(\sum w_i X_i)(\sum w_i Y_i)/\sum w_i]}{\sum w_i X_i^2 - [(\sum w_i X_i)^2/\sum w_i]} && \text{cálculo} \end{aligned}$$

donde  $\bar{X} = (\sum w_i X_i)/\sum w_i$ . La ecuación de regresión es

$$Y = \bar{Y} + b(X - \bar{X})$$

La reducción en la suma de cuadrados atribuible a la regresión es

$$\text{SC regresión} = \text{SC}(b|a)$$

$$= \frac{\left\{ \sum w_i X_i Y_i - [(\sum w_i X_i)(\sum w_i Y_i)/\sum w_i] \right\}^2}{\sum w_i X_i^2 - [(\sum w_i X_i)^2/\sum w_i]} \quad 1 \text{ gl}$$

y la suma residual de cuadrados se obtiene restando del total la reducción. La suma ponderada de las desviaciones respecto de la regresión es cero. Esto es,

$$\sum w_i [Y_i - \bar{Y} - b(X_i - \bar{X})] = 0$$

Así mismo, la suma ponderada de cuadrados,

$$\sum w_i [Y_i - \bar{Y} - b(X_i - \bar{X})]^2$$

es un mínimo; no hay otra recta de regresión que dé una suma ponderada de cuadrados menor.

Como se ve, los cálculos son más largos que los para la regresión corriente. Sin embargo, si se usan las columnas de  $w_i X_i$  y  $w_i Y_i$ , las sumas ponderadas de productos,  $\sum w_i X_i^2$ ,  $\sum w_i Y_i^2$ , y  $\sum w_i X_i Y_i$ , se obtienen fácilmente de pares de las columnas  $w_i X_i$  y  $X_i$ ,  $w_i Y_i$  y  $Y_i$ ,  $w_i X_i$  y  $Y_i$  o  $w_i Y_i$  y  $X_i$ , respectivamente.

En el caso especial en que la regresión es de medias de diferentes números de observaciones respecto a una variable independiente, tenemos  $w_i Y_i = n_i \bar{Y}_i$  el total de observaciones en la media  $i$ -ésima, cantidad de la que ya se dispone. En consecuencia, las partidas de la columna  $w_i Y_i$  ya están calculadas.

## Referencias

- 10.1. Bliss, C. I.: *The statistics of bioassay*, Academic, Nueva York, 1952.
- 10.2. Eisenhart, C.: "The interpretation of certain regression methods and their use in biological and industrial research," *Ann. Math. Statist.*, 10:162-186 (1939).
- 10.3. Finney, D. J.: *Probit analysis*, 2a. ed., Cambridge University Press, Cambridge, 1951.
- 10.4. Irwin, J. O.: "Statistical methods applied to biological assays," *J. Roy. Statist. Soc. Suppl.*, 4:148 (1937).
- 10.5. Winsor, C. P.: "Which regression?" *Biom. Bull.*, 2:101-109 (1946).
- 10.6. Bartlett, M. S.: "Fitting a straight line when both variables are subject to error," *Biom.*, 5:207-212 (1949).
- 10.7. Natrella, G.: *Experimental statistics*, Nat. Bur. Stand. Handb. 91, 1963, chap. 5.
- 10.8. Reynolds, J., G. Cunningham y J. Syvertsen: "A net Co<sub>2</sub> exchange model for *Larrea tridentata*," *Photosyn.*, 13(3): (1979).
- 10.9. Reynolds, J. y D. H. Knight: "The magnitude of snowfall and rainfall interception by litter in lodgepole and spruce-fir forests in Wyoming," *Northwest Sci.*, 47(1):50-60 (1973).

## CORRELACION LINEAL

### 11.1 Introducción

En la sección 10.11 se habló brevemente de las distribuciones bivariantes. Al muestrear en una distribución bivariante, una observación consiste en un par de medias aleatorias. Entonces son posibles y válidas dos regresiones, aunque solo se necesite una regresión. Un resumen de los datos en una muestra proveniente de una distribución bivariante consiste en dos medias, dos varianzas y la covarianza. La covarianza puede reemplazarse sin pérdida de información por el coeficiente de determinación a su raíz cuadrada, el coeficiente de correlación lineal. Este capítulo se ocupa de la correlación lineal.

### 11.2 La correlación y el coeficiente de correlación

La correlación, como la covarianza, es una medida del grado en que dos variables varían conjuntamente o una medida de la intensidad de asociación. Por tanto, debe haber simetría en las dos variables. El coeficiente de correlación lineal muestral, también llamado correlación simple, correlación total y correlación momento-producto, se usa con propósitos descriptivos y se define por

$$\begin{aligned} r &= \frac{\sum (X - \bar{X})(Y - \bar{Y})/(n - 1)}{\sqrt{\sum (X - \bar{X})^2/(n - 1)} \sqrt{\sum (Y - \bar{Y})^2/(n - 1)}} \\ &= \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}} \end{aligned} \quad (11.1)$$

Se supone que en la población existe una relación lineal entre las variables. Este es un supuesto válido cuando el muestreo se hace en una distribución normal bivariante. El coefi-

ciente de correlación  $r$  es una estimación no sesgada del correspondiente coeficiente de correlación poblacional  $\rho$  ( $ro$ ) sólo cuando el parámetro poblacional  $\rho$  es cero. A diferencia de una varianza o de un coeficiente de regresión, el coeficiente de correlación es independiente de las unidades de medida; es una cantidad absoluta o sin dimensión. El uso de  $X$  y  $Y$  ya no implica una variable independiente o dependiente.

Los *diagramas de dispersión* de la fig. 11.1, para los cuales se han ideado los datos expresamente, pueden dar cierta idea para la interpretación de la correlación lineal. En la parte *a* de la figura, los puntos se acumulan en torno a una recta que pasa por  $(\bar{X}, \bar{Y})$  y es paralela al eje de los  $X$  simplemente porque la varianza de  $X$  es mayor que la varianza de  $Y$ . La falta de tendencia a acumularse en torno a una recta distinta de la que pasa por  $(\bar{X}, \bar{Y})$  paralela a un eje, es típica de datos con poca o ninguna correlación lineal. En tales casos, toda recta de regresión se acerca a una recta que pasa por  $(\bar{X}, \bar{Y})$  y es paralela al eje de la variable independiente. La falta de correlación lineal es más notoria cuando en el diagrama de dispersión se utilizan desviaciones estándar como unidades de medida, como ocurre en la parte *b*. Cada dato de la parte *a* se ha dividido por su desviación estándar, esto es,  $(X - \bar{X})/s_X = (X - \bar{X})/6$  y  $(Y - \bar{Y})/s_Y = (Y - \bar{Y})/2$ , así que cada variable tiene varianza unitaria. Las variables con medias cero y varianzas unitarias iguales se llaman *variables estándar*. Los puntos ya no muestran tendencia a acumularse en torno a ningún eje. La covarianza es igual a la correlación.

Cuando la correlación lineal es pequeña,  $r$  se acerca a cero. Esto puede observarse en la fig. 11.2*a*. Los puntos muestrales se distribuyen en forma aproximadamente igual en cada uno de los cuatro cuadrantes, de modo que los productos cruzados  $(X - \bar{X})(Y - \bar{Y})$  de la misma magnitud ocurren con igual frecuencia aproximadamente. Así que  $\sum (X - \bar{X}) \times (Y - \bar{Y})$ , el numerador de  $r$ , tiende a estar en las cercanías de cero.

La figura 11.1*c* y *d* utiliza los mismos números que se usaron en *a* y *b*, pero pareados en forma diferente. Para *c*,  $\sum (X - \bar{X})(Y - \bar{Y}) = 83$ , pero las medias y las varianzas no cambian. Ahora los puntos tienden a acumularse en torno a una recta que no es un eje. Esto es típico de datos con alta correlación. Las rectas de regresión casi coinciden y están inclinadas hacia el eje  $X$  debido a la considerable varianza de  $X$ . La casi coincidencia de las rectas de regresión es típica de datos con alta correlación.

La elevada correlación lineal,  $r$  muy cercana a +1 o -1, se reconoce más fácilmente en diagramas de dispersión donde se han estandarizado las variables, tal como ocurre en la fig. 11.1*d*. Aquí, los puntos se acumulan en torno a una recta que se encuentra aproximadamente equidistante de los ejes. Ambas rectas de regresión están cerca a esta recta y los coeficientes de regresión son iguales al coeficiente de correlación y tienen valores cercanos a +1 o -1. Una unidad de cambio en una variable implica un cambio aproximadamente igual a una unidad en la otra para rectas de regresión con posición cercana a la intermedia. La fig. 11.2*b* y *c* muestra por qué  $r$  será numéricamente grande. Aquí, los productos cruzados con valores aparecen con mayor frecuencia en las esquinas superior derecha e inferior izquierda o superior izquierda e inferior derecha. Esto tiende a dar un valor numéricamente grande para  $\sum (X - \bar{X})(Y - \bar{Y})$ . En la fig. 11.2*b*,  $\sum (X - \bar{X})(Y - \bar{Y})$  será positiva y dará una correlación positiva; en la fig. 11.2*c*,  $\sum (X - \bar{X})(Y - \bar{Y})$  será negativa y dará correlación negativa.

Detectar visualmente una correlación lineal a partir de una gráfica de puntos puede ser difícil. Una elección no adecuada de escala puede tender a ocultar una correlación real o indicar una real donde no existe. Un cambio de escala variará la pendiente aparente de

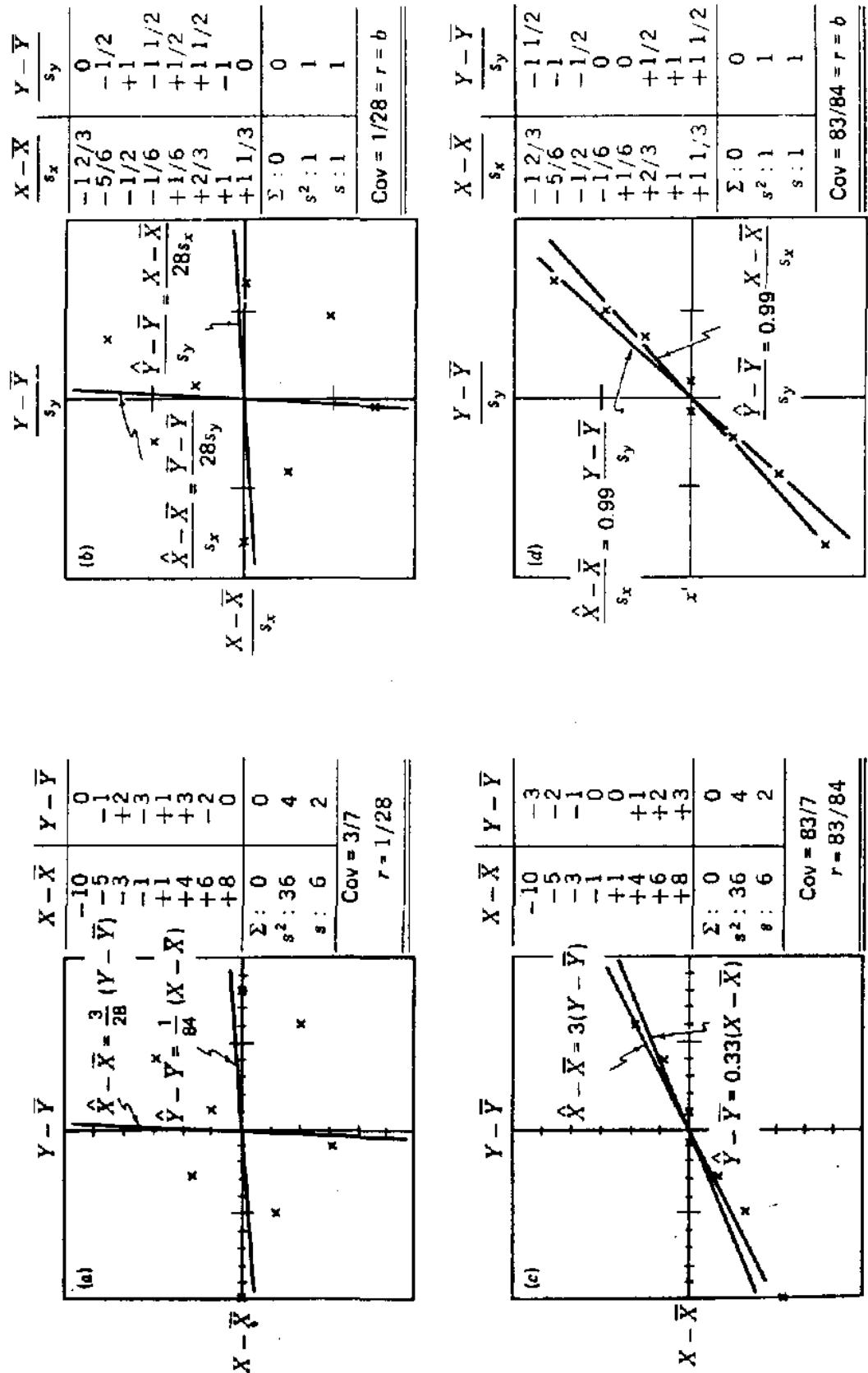


Figura 11.1 Diagramas de dispersión para ilustrar la correlación

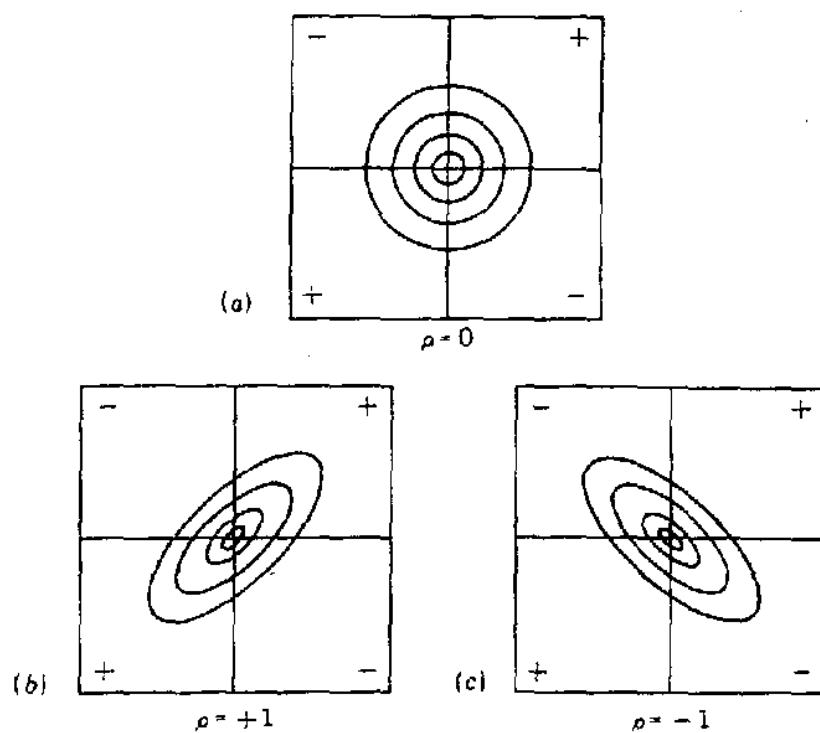


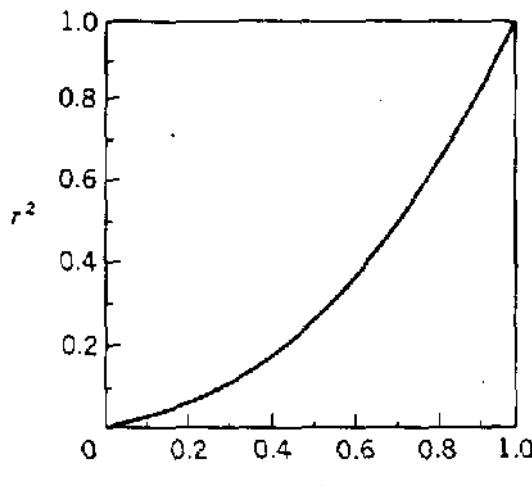
Figura 11.2 Diagrama para ilustrar la correlación. Los signos de las esquinas ( $\pm$ ) se refieren al signo de los productos cruzados para ese cuadrante. Se intenta que la probabilidad de que un punto quede en una región entre círculos o elipses sucesivos sea la misma para cada región. Así la distribución de frecuencias tridimensionales mostraría un amontonamiento sobre la intersección de los ejes y una disminución gradual a medida que se aleja de este punto. Esta presentación se hace en cambio de una figura tridimensional, tal como la fig. 10.4b y d.

la recta de regresión. Si un diagrama de dispersión no usa realmente desviaciones estándar como unidades, como mínimo deberá disponerse de desviaciones estándar junto con el diagrama para evitar conclusiones falsas.

Además de las dificultades debidas a mala elección de escala, la detección visual se ve además estorbada porque la relación entre  $r$  y la proporción de la suma de cuadrados total explicada por la regresión no es lineal. La fig. 11.3 muestra esto gráficamente. Para  $r = 0.1$ , solamente el 1 por ciento de la variación en una variable dependiente se explica por la variación de la variable independiente; para  $r = 0.2$ , el porcentaje es sólo de 4; para  $r = 0.5$ , es sólo de 25. Para  $r > 0.5$ , el porcentaje explicado aumenta más rápidamente. Además del cálculo efectivo de  $r$ , la detección es más probable mediante la selección de valores grandes o pequeños de una variable y observando si los valores correspondientes de la otra variable tienden a estar en un extremo de la escala para esa variable. También dan información los recuentos del número de puntos de cada cuadrante.

Finalmente, la variación muestral en  $r$  es bastante grande para muestras de tamaño pequeño.

Puede demostrarse que  $r$  cae entre  $-1$  y  $+1$ , esto es  $-1 \leq r \leq 1$ . Los valores de  $+1$  y  $-1$  indican una perfecta correlación lineal o una perfecta relación funcional entre las dos variables. Esto lleva las observaciones fuera del ámbito de la estadística. Salvo errores de redondeo, sólo se encuentra perfecta correlación en momentos de descuido, como podría suceder correlacionando la estatura con altura hasta el hombro, más altura del hombro a

Figura 11.3 Relación entre  $r$  y  $r^2$ .

la coronilla. Existen correlaciones extremadamente altas, pero, cuando se observan, debe hacerse un escrutinio muy estricto, ya que puede haberse presentado el descuido mencionado anteriormente y no haberse obtenido correlación perfecta debido solamente a errores de redondeo en los cálculos.

Los coeficientes de determinación y de alienación son dos cantidades estrechamente relacionadas con  $r$ . El *coeficiente de determinación* es  $r^2$ , que es el cuadrado del coeficiente de correlación. Este término también puede usarse en el caso del análisis de regresión, cuando no es aplicable el coeficiente de correlación. En tales casos,  $r^2$  es la proporción de una suma de cuadrados total que es atribuible a otra fuente de variación, la variable independiente.

El *coeficiente de no determinación* está dado por  $1 - r^2 = k^2$  y corresponde a la proporción no explicada de una suma de cuadrados total. Usualmente, constituye la base de un término de error. Su raíz cuadrada  $k$  se llama *coeficiente de alienación*  $k = 1 - k$  ha sido llamado *factor de mejoramiento*. Estos términos no son muy comunes.

**Ejercicio 11.2.1** H. H. Smith recolectó los siguientes datos sobre flores de un cruce de Nicotiana.  $T$  = longitud del tubo,  $L$  = longitud del limbo y  $N$  = longitud de la base del tubo.<sup>†</sup>

$T$ : 49, 44, 32, 42, 32, 53, 36, 39, 37, 45, 41, 48, 45, 39, 40, 34, 37, 35

$L$ : 27, 24, 12, 22, 13, 29, 14, 20, 16, 21, 22, 25, 23, 18, 20, 15, 20, 13

$N$ : 19, 16, 12, 17, 10, 19, 15, 14, 15, 21, 14, 22, 22, 15, 14, 15, 15, 16

Calcular los coeficientes de correlación entre  $T$  y  $L$ ,  $T$  y  $N$ ,  $L$  y  $N$ . ¿Cuáles son los coeficientes de determinación de cada caso?

**Ejercicio 11.2.2** En un estudio sobre el uso de folículos ovulados en la determinación de los huevos puestos por faisán de cuello anillado, C. Kabar et al. (11.3) presentan los siguientes datos de 14 hembras cautivas.

<sup>†</sup> Publicado con permiso de H. H. Smith, Brookhaven National Laboratories.

Huevos puestos: 39, 29, 46, 28, 31, 25, 49, 57, 51, 21, 42, 38, 34, 47  
 Folículos ovulados: 37, 34, 52, 26, 32, 25, 55, 65, 44, 25, 45, 26, 29, 30

Calcular el coeficiente de correlación y el de determinación.

**Ejercicio 12.2.3** En el ejercicio 10.2.3, se dieron medidas para  $X_3$  = contenido de colesterol en la sangre, y  $X_4$  = contenido de ácido úrico en la sangre. Calcular el coeficiente de correlación entre estas dos variables

### 11.3 Correlación y regresión

Ahora tenemos tres métodos para tratar pares aleatorios de observaciones.

1. Dejar de lado cualquier relación entre las variables y analizarla separadamente.
2. Usar un análisis de regresión
3. Examinar la correlación.

Aquí sólo nos interesan los métodos segundo y tercero.

La correlación mide una co-relación, una propiedad conjunta de dos variables. Cuando las variables están afectadas en forma conjunta por causas externas, la correlación puede ofrecer el enfoque más lógico para analizar los datos. La regresión trata sobre todo las medias de una variable y cómo cambia su localización con otra variable. Para la correlación, los pares aleatorios de observaciones se obtienen de una distribución normal bivariante, distribución que no es muy común; para la regresión, solo el elemento dependiente de cada par debe distribuirse en forma aleatoria y normal. La correlación está asociada con técnicas descriptivas; la regresión tiene que ver con una relación entre las medias poblacionales y los valores de una variable concomitante. Así, mientras que un coeficiente de correlación nos dice algo sobre una relación conjunta entre variables, un coeficiente de regresión nos dice que si alteramos el valor de la variable independiente entonces podemos esperar una alteración de cierta magnitud en la variable dependiente en promedio, pero la variación muestral hace improbable que se observe precisamente la magnitud de dicha variación.

Ya se ha observado que

$$\begin{aligned}
 r^2 &= \frac{[\sum (X - \bar{X})(Y - \bar{Y})]^2}{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2} \quad \text{es decir, cuadrado del } r \text{ de definición} \\
 &= \frac{[\sum (X - \bar{X})(Y - \bar{Y})]^2 / \sum (X - \bar{X})^2}{\sum (Y - \bar{Y})^2} = \frac{SC(Y/X)}{SC_{\text{total}}(Y)} \quad (\text{debido a } X) \\
 &= \frac{[\sum (X - \bar{X})(Y - \bar{Y})]^2 / \sum (Y - \bar{Y})^2}{\sum (X - \bar{X})^2} = \frac{SC(X/Y)}{SC_{\text{total}}(X)} \quad (\text{debido a } Y)
 \end{aligned}$$

Además

$$r^2 = \left( \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} \right) \left( \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (Y - \bar{Y})^2} \right) = b_{yx} b_{xy} \quad (11.2)$$

donde  $b_{YX}$  y  $b_{XY}$  son coeficientes de regresión de  $Y$  con respecto a  $X$  y de  $X$  con respecto a  $Y$ . Así, el producto de los coeficientes de regresión es el cuadrado del coeficiente de correlación; o el coeficiente de correlación es la raíz cuadrada del producto de los coeficientes de regresión, o de su media geométrica. Estos resultados son siempre algebraicamente correctos; *tienen significado sólo cuando la muestra consiste en pares aleatorios*.

Si estandarizamos nuestras variables, entonces la ecuación de regresión de  $Y$  con respecto a  $X$  se convierte en

$$\frac{Y - \bar{Y}}{s_Y} = r \frac{X - \bar{X}}{s_X} \quad \text{o} \quad Y' = rX'$$

tal como se estableció en la sec. 11.2. Análogamente

$$\frac{X - \bar{X}}{s_X} = r \frac{Y - \bar{Y}}{s_Y} \quad \text{o} \quad X' = rY'$$

Aquí  $r$  es un coeficiente de regresión. Obsérvese que ninguna de estas ecuaciones puede obtenerse resolviendo la otra; un coeficiente de regresión no es un enunciado simétrico respecto a la relación entre dos variables.

#### 11.4 Distribuciones muestrales, intervalos de confianza y pruebas de hipótesis

Como  $-1 \leq r \leq 1$ , no se puede esperar que la distribución muestral de  $r$  sea simétrica cuando el parámetro poblacional  $\rho$  es diferente de cero. La simetría ocurre solamente para  $\rho = 0$ , y la falta de simetría, o asimetría, aumenta al acercarse  $\rho$  a  $+1$  o  $-1$ . Esto quiere decir que para muestras pequeñas, una aproximación normal no es en general apropiada para calcular un intervalo de confianza para  $\rho$  y que el intervalo de confianza no debe centrarse en  $r$ . El defecto de centrarse en  $r$  aumentará a medida que los valores observados se aproximan a  $+1$  o  $-1$ , ya que estos valores son los límites de valores posibles.

Los  $r$  muestrales son bastante variables para muestras pequeñas, particularmente para  $\rho$  cero o cerca de cero. Un solo par de valores puede causar gran diferencia en el valor de  $r$ . Esto hace difícil detectar valores pequeños, pero reales de  $\rho$  cuando se usan muestras pequeñas. Afortunadamente, valores pequeños de  $\rho$  son de poco uso práctico como se dijo en la sec. 11.2 en cuanto a la relación entre  $r$  y  $r^2$ .

Possiblemente el método más simple de construir un intervalo de confianza para  $\rho$  es usar los diagramas preparados por David (11.1). Tales diagramas se dan en la tabla A.11. Para establecer un intervalo de confianza respecto a  $\rho$ , trácese una vertical por el valor observado de  $r$  en la "escala de  $r$ ". En los puntos de intersección de esta recta con las dos líneas curvas que corresponden al tamaño de la muestra, trácese rectas que corten la "escala de  $\rho$ " perpendicularmente. Los dos valores de  $\rho$  en esta escala son los límites del intervalo de confianza. Estos esquemas de fácil empleo son suficientemente precisos para la mayoría de los propósitos.

Otro método cómodo lo da Fisher (11.2). Para esto, calcúlese la transformación

$$Z' = .5 \ln \frac{1+r}{1-r} \quad (11.3)$$

que tiene aproximadamente distribución normal con una media aproximada y desviación estándar de  $0.5 \ln [(1+\rho)/(1-\rho)]$  y  $1/\sqrt{n-3}$ , independientemente del valor de  $\rho$ . ( $\ln$  es el logaritmo natural, en base  $e$ ). Los valores de  $Z'$  para  $r$  desde 0.00 [0.01] 0.99, se dan en la tabla A.12. Para un intervalo de confianza para  $\rho$ , primero determinese uno para la media poblacional de  $Z'$  y luego se le convierte para  $\rho$ . La fórmula de conversión es  $r = (e^{2Z'} - 1)/(e^{2Z'} + 1)$ , pero la tabla A.12 se puede usar satisfactoriamente.

Para ilustrar lo dicho se observó que la correlación del porcentaje de resina y el porcentaje de contenido de caucho en las plantas de guayule es  $r = 0.527$  para 50 plantas de la cepa 416. Encontramos

$$n - 3 = 47 \quad \sigma_{Z'} = \sqrt{\frac{1}{47}} = .146$$

$$\begin{aligned} Z' &= .5 \ln \frac{1+r}{1-r} = .5 \ln \frac{1.527}{.473} = .5 \ln 3.23 \\ &= .5(1.172) = .586 \end{aligned}$$

$$\begin{aligned} \text{IC } (\mu_{Z'}) &= .586 \pm 1.96(.146) \\ &= .300, .872 \end{aligned}$$

Una interpolación aproximada en la tabla A.12 da el intervalo de confianza para  $\rho$  del 95 por ciento de (0.290, 0.703).

Para probar la hipótesis nula de que  $\rho$  es igual a un valor diferente de cero, transfórmese a normalidad aproximada como se hizo en el párrafo anterior y úsese la prueba normal. También debe transformarse la media en hipótesis.

Si la hipótesis es que  $\rho = 0$ , es más simple y correcto calcular

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

y comparar con la  $t$  de Student con  $n - 2$  grados de libertad. El cuadrado  $t$  es igual al  $F$  del análisis de la varianza para la regresión, y por consiguiente, el procedimiento equivale a probar la hipótesis de que  $\beta = 0$ . Esta prueba no puede usarse para probar la hipótesis  $\rho = \rho_0 \neq 0$ , esto es, que  $\rho$  es igual a una constante diferente de cero.

La tabla A.13, bajo una variable independiente, da valores de  $r$  al 95 y 99 por ciento de niveles de significancia para varios valores de los grados de libertad.

La distinción entre significancia e insignificancia no es problema serio con muestras pequeñas, pero puede serlo con muestras grandes. Esto es especialmente cierto en la corre-

lación, ya que se recolectan muchas muestras bivariantes grandes en ciertos tipos de estudios. En muestras grandes, valores pequeños de  $r$  pueden ser significantes. Sin embargo, si el porcentaje de reducción en la suma de cuadrados total es pequeño para cualesquiera de las variables consideradas como dependientes, entonces la correlación puede ser inútil.

**Ejercicio 11.4.1** Construir un intervalo de confianza para uno de los coeficientes de correlación calculados en el ejercicio 11.2.1. Utilizar el diagrama de David (tabla A.11) y una transformación a  $Z'$  y luego volver a  $r$ . ¿Difieren los resultados de manera apreciable?

**Ejercicio 11.4.2** Para el coeficiente de correlación obtenido en el ejercicio 11.2.2, probar la hipótesis nula de que  $\rho = 0$ .

**Ejercicio 11.4.3** Construir un intervalo de confianza para el parámetro estimado por el coeficiente de correlación del ejercicio 11.2.3. Probar la hipótesis nula de que  $\rho = 0$ .

## 11.5 Homogeneidad de los coeficientes de correlación

La prueba de una diferencia entre dos valores poblacionales de  $\rho$  es simple. Los dos valores de  $r$  se pasan a  $Z'$  y se aplica la prueba normal apropiada para muestras grandes vista en el cap. 5. Para pruebas de hipótesis respecto a uno de dos valores de  $\rho$ , se deben encontrar los valores  $Z'$ , pero no se necesita de reconversión; ésta se necesita sólo en el caso de intervalos de confianza. En la tabla 11.1 se da un ejemplo de la prueba de homogeneidad de dos  $r$ .

En la prueba que se presenta en la tabla 11.2, se compara una diferencia con su desviación estándar, la raíz cuadrada de la suma de las varianzas, y el resultado se compara con valores tabulados en una tabla normal. La probabilidad de encontrar una diferencia mayor debida al azar cuando no hay diferencia, es aproximadamente igual al 38 por ciento para esos dos valores de  $r$ . Obsérvese que  $r^* = 0.310$  no es significante, mientras que  $r = 0.542$  es altamente significante. Sin embargo, la comparación directa es la apropiada.

**Tabla 11.1 Prueba de la hipótesis  $\rho_1 = \rho_2$  para la correlación del contenido de caucho de las ramas respecto del diámetro en guayule**

Cepa	Número de plantas	$r$	$Z'$	$1/(n - 3)$
109	22	0.310	0.32055	0.0526
130	21	0.542	0.60699	0.0556
Diferencia = 0.28644, suma = 0.1082				
$Z^* = \frac{0.28644}{\sqrt{0.1082}} = \frac{0.28644}{0.329} = 0.87 \text{ ns}$				

\* Este  $Z$  es una variable estándar con  $\mu = 0$  y  $\sigma = 1$  en la hipótesis, inferior a nula  $H_0$ .

Fuente: Datos del U.S. Dept. Agr. Tech Bull. 919, 1946, por W.T. Federer.

**Tabla 11.2 Homogeneidad y combinación de los  $r$  para la correlación entre el porcentaje de contenido de resina respecto del porcentaje de contenido de caucho en el guayule**

Cepa	$n_i$	$r_i$	$Z'_i$	$Z'_i - \bar{Z}'_w$	$(n_i - 3)(Z'_i - \bar{Z}'_w)^2$
405	50	0.362	0.379	-0.0913	0.392
407	50	0.419	0.446	-0.0243	0.028
416	50	0.527	0.586	+0.1157	0.629
<u>150</u>					$\chi^2 = 1.049, 2 \text{ gl}$
					$\bar{Z}'_w = \frac{\sum (n_i - 3)Z'_i}{\sum (n_i - 3)} = 0.4703$

Fuente: Datos obtenidos del U.S. Dept. Agr. Tech. Bull. 919, 1946, por W. T. Federer.

A veces se desea probar la homogeneidad de varios coeficientes de correlación y obtener un solo coeficiente si parecen ser homogéneos. Por ejemplo, se puede disponer de mediciones de dos características de un cultivo o de un tipo de animal para varias cepas o razas. Las varianzas de estas cepas o razas pueden ser no homogéneas, así que la combinación de las sumas de productos y el cálculo de un solo coeficiente de correlación no es válido. La prueba de homogeneidad y el método de combinación se ilustran en la tabla 11.2.

Los  $Z'$  se distribuyen en forma aproximadamente normal. Como las varianzas,  $1/(n_i - 3)$ , generalmente serán desiguales, se usa una media ponderada  $\bar{Z}'_w$  en los cálculos. El criterio de prueba es  $\chi^2$ , con un número de grados de libertad igual al número de los  $r$ , sea  $k$ , menos uno. El criterio de prueba de  $\chi^2$  se definió originalmente en la ec. (3.13). En nuestro ejemplo presente,  $Z'_i - \bar{Z}'_w$  reemplaza a  $Y_i$ ; y cero, la medida de la población de todas las posibles desviaciones, reemplaza a  $\mu_i$ . La desviación estándar  $\sigma_i$  se reemplaza por  $1/\sqrt{n_i - 3}$ . Así, la ec. (3.13) se convierte en

$$\chi^2 = \sum_i \left( \frac{Z'_i - \bar{Z}'_w}{1/\sqrt{n_i - 3}} \right)^2 = \sum_i (n_i - 3)(Z'_i - \bar{Z}'_w)^2$$

Como sólo las desviaciones  $k - 1$  son independientes,  $\chi^2$  tiene sólo  $k - 1$  grados de libertad. Aquí,  $\chi^2$  no es significante, así que concluimos que los coeficientes de correlación son homogéneos. La conversión de  $\bar{Z}'_w$  otra vez a  $r$ , da un valor combinado de los varios coeficientes; el valor combinado de  $r = 0.438$ . Para establecer límites de confianza para el  $\rho$ , común, sitúese un intervalo en torno a  $\bar{Z}'_w$  y conviértase a valores de  $r$ . La desviación estándar apropiada es  $1/\sqrt{\sum (n_i - 3)}$ . Como  $n_i - 3$  es la cantidad de información en  $\bar{Z}'_w$ , la cantidad de información en  $Z'_w$  es  $\sum (n_i - 3)$  y su inverso es la varianza apropiada para  $\bar{Z}'_w$ .

Existe un pequeño sesgo en  $Z'$  que puede ser serio si se promedian muchas correlaciones. Ya que aquí sólo hay 3, hacemos la conversión sin vacilar. Este sesgo es igual a

$$\frac{\rho}{2(n - 1)}$$

y es positivo. Dado que  $\rho$  es desconocido, el sesgo no puede eliminarse. Se ha sugerido que el  $r$  promedio obtenido a partir de  $\tilde{Z}'_w$  se use para  $\rho$  en el cálculo del sesgo para cada  $Z'$ . Estos sesgos se restan de sus respectivos  $Z'$  y se calcula un nuevo  $\tilde{Z}'_w$ . Este se convierte en un  $r$  ajustado que debe ser más preciso que el no ajustado. No es apropiado calcular un nuevo  $\chi^2$ .

**Ejercicio 11.5.1** Para mostrar el efecto de la asimetría en una distribución de  $r$ , supongamos que tres  $r$  muestrales de la tabla 11.1 se aumentan en 0.4, 0.762, 0.819 y 0.927. Probar la homogeneidad y comparar los dos valores de  $\chi^2$  y las dos probabilidades de obtener un valor mayor de  $\chi^2$ .

## 11.6 Correlación intraclasses

En ocasiones se desea un coeficiente de correlación donde no hay criterio significativo para asignar un miembro del par a una variable más que a otra. Esto puede ser así cuando se mide la correlación de una característica en gemelos. En muchos de tales casos, podemos obtener un valor del coeficiente a partir del cálculo de ciertas varianzas. Al coeficiente resultante se le llama *correlación intraclass* y se calcula mediante la ec. (11.4) cuando hay  $n$  observaciones por clase.

$$r_I = \frac{\text{CM(entre clases)} - \text{CM(dentro de clases)}}{\text{CM(entre clases)} + (n - 1) \text{CM(dentro de clases)}} \quad (11.4)$$

El mismo resultado se obtiene a partir de

$$r_I = \frac{\widehat{\sigma}_t^2}{\widehat{\sigma}_t^2 + \widehat{\sigma}^2} \quad (11.5)$$

donde  $\widehat{\sigma}^2$  y  $\widehat{\sigma}_t^2$  son estimaciones de las correspondientes componentes de varianza en un análisis entre y dentro de clases. Esta definición de una cantidad muestral se basa en la definición de la *correlación intraclass* poblacional, o sea,

$$\rho_I = \frac{\sigma_t^2}{\sigma_t^2 + \sigma^2}$$

donde  $\sigma_t^2$  es la variación en la población de medias de clase.

El primer párrafo de esta sección constituye una justificación del término correlación intraclass más que una definición. La definición se aplica aun cuando el número de observaciones varíe de una clase a otra. Interesa anotar que es la segunda vez que las varianzas se han usado para determinar la correlación. La razón de una suma de cuadrados de regresión a una suma total de cuadrados es el cuadrado del coeficiente de correlación; así, la razón de las varianzas es un múltiplo de  $r^2$  dependiente de los grados de libertad de la suma de cuadrados total.

En ocasiones, se sabe que las respuestas de los individuos dentro de un grupo no son independientes. Por ejemplo, un número de animales seleccionados al azar pueden ponerse

juntos en una jaula para tener una unidad experimental. Dentro de cada jaula, competirán por una cantidad fija de alimento y los más fuertes obtendrán mayores cantidades. En consecuencia, las respuestas dentro de jaulas se correlacionan negativamente; son más variables que en el caso de que no hubiese competencia. Aquí, una estimación de  $\rho_I$  sería negativa, y si se estima a partir de un análisis de la varianza, esto puede ocurrir sólo si CM(entre clases) < CM(dentro de clases), un resultado no esperado. O bien, los niños en un salón de clase pueden responder en forma similar a un procedimiento de examen hasta el punto que las respuestas se correlacionan positivamente con relación a una muestra puramente aleatoria. Esto puede ocurrir debido a que van a una escuela privada o de vecindario donde la selección ha creado una homogeneidad inusitadamente estrecha.

Supóngase que tenemos un diseño completamente aleatorizado, de ordinario el modelo II, pero ahora con una correlación intraclasa apropiada. Una alternativa razonable al modelo II sería  $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$  donde  $\mu$  es fijo y los  $\tau_i$  se toman en forma aleatoria de una población con media  $\mu_\tau = 0$  y varianza  $\sigma_\tau^2 > 0$ , y los  $\varepsilon_{ij}$  se distribuyen normalmente con media  $\mu_\varepsilon = 0$ , y varianza común  $\sigma^2 > 0$ , pero  $\sigma_{jj'} = \rho_I \neq 0$  para  $j \neq j'$ ; y no se suponen correlaciones cero entre los  $\varepsilon$ . Para este modelo,

$$E[CM(\text{dentro de clases})] = \sigma^2(1 - \rho_I)$$

$$E[CM(\text{entre clases})] = \sigma^2[1 + (n - 1)\rho_I]$$

Las ecuaciones (11.4) y (11.5) ahora se ven razonables.

El valor de  $r_I$  puede ser tan elevado como +1, pero nunca puede caer por debajo de  $-1/(n - 1)$ , el caso en que CM(entre clases) = 0. La correlación momento-producto es una *correlación interclase*. La correlación intraclasa es una medida del parecido fraternal cuando este término sea significativo.

Para los datos de la tabla 11.3

$$\begin{aligned} r_I &= \frac{16,320.5 - 2,199.2}{16,320.5 + 13(2,199.2)} = .314 \\ &= \frac{1,008.7}{1,008.7 + 2,199.2} = .314 \end{aligned}$$

Es evidente ahora una prueba de significancia de una correlación intraclasa. Pruébese la presencia de  $\sigma_\tau^2$  en el análisis de la varianza; el criterio es  $F$  y el numerador y el denominador son las varianzas entre y dentro de clases respectivamente. Por lo general, se efectúa una prueba de significancia antes de calcular  $r_I$ .

Cuando los datos son simplemente observaciones pareadas ( $Y, Y'$ ), como en el primer párrafo, se puede calcular el coeficiente de correlación intraclasa por

$$r'_I = \frac{2 \sum (Y - \bar{Y})(Y' - \bar{Y}')}{\sum (Y - \bar{Y})^2 + \sum (Y' - \bar{Y}')^2} \quad (11.6)$$

**Tabla 11.3 Análisis de la varianza de la ganancia de peso de dos lotes de novillas Holstein**

Fuente	gl	CM	Estimaciones CM	
			(modelo usual)	Componente
Tratamientos	1	16,320.5	$\sigma^2 + 14\sigma_e^2$	$\sigma_e^2 = 1,008.7$
Error	26	2,199.2	$\sigma^2$	$\sigma^2 = 2,199.2$

No hay técnica especial que asigne un miembro de un par a  $Y$  y el otro a  $Y'$ . La ec. (11.6) es equivalente a incluir dos veces cada par, una vez como  $(Y, Y')$  y otra como  $(Y', Y)$ ; esto también se aplica cuando  $Y = Y'$ . Snedecor y Cochran (11.5) ilustran este procedimiento usando datos sobre el número de surcos digitales en gemelos idénticos; también usan el análisis de la varianza.

Los procedimientos de las secs. 11.4 y 11.5 pueden usarse para encontrar  $r'_t$  con un cambio en la varianza aproximada a  $\sigma_Z^2 = 1/(n - \frac{3}{2})$ , donde  $n$  es el número de pares antes de duplicar. Las estimaciones de  $\rho$  son sesgadas. Para una aproximación cercana a un valor no sesgado, añádase  $1/(2n - 1)$  a  $Z'$  antes de transformar de nuevo a  $r$ .

Fisher (11.4, cap. 7) expone con más detalle el coeficiente de correlación intraclase.

## Referencias

- 11.1 David, F. N.: *Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples*, Cambridge, Nueva York, 1938.
- 11.2 Fisher, R. A.: "On the 'probable error' of a coefficient of correlation deduced from a small sample," *Metron*, 1:3-32 (1921).
- 11.3 Kabat, C., I. O. Buss, y R. K. Meyer: "The use of ovulated follicles in determining eggs laid by ring-necked pheasant," *J. Wildlife Manage*, 12:399-416 (1948).
- 11.4 Fisher, R. A.: *Statistical Methods for Research Workers*, 11a. ed., rev., Hafner, Nueva York, 1950.
- 11.5 Snedecor, G. W., y W. G. Cochran: *Statistical Methods*, 6a. ed., Iowa State University Press, Ames, Iowa, 1967.

---

## CAPITULO DOCE

---

### NOTACION MATRICIAL

#### 12.1 Introducción

Un investigador que recolecta datos cuantitativos probablemente usará cálculos estadísticos, los cuales son el resultado de aproximaciones matemáticas al análisis de datos. Por ejemplo, se dice que algunos estimadores son insesgados, de mínima varianza, o estimadores por mínimos cuadrados, con frecuencia las matemáticas han sido necesarias para desarrollar los procedimientos de cálculo, algunos de los cuales son bastante complejos.

Para interpretar y entender procedimientos complejos, el investigador puede buscar la cooperación de un estadístico. Cuando se busca tal ayuda, el investigador debe conocer algo del lenguaje matemático. Entre los instrumentos matemáticos que se suelen emplear, las matrices ocupan un lugar importante. Se usan mucho en la preparación de datos para cálculos mecánicos de alta velocidad y en la presentación de resultados.

Este capítulo se propone dar algunos conocimientos y desarrollar cierta destreza en la manipulación de matrices. Se espera que la presentación de problemas en notación matricial ayude al lector a aclararlos y también a visualizar la naturaleza de la solución.

Para quienes no se inclinan a permitir que algo impida su progreso en la metodología estadística, puede ser suficiente proceder con lo siguiente como bases:

1. Las ilustraciones de matrices dadas en la sec. 12.2 mediante la ec. (12.2).
2. El ejemplo de la multiplicación de matrices que precede a la ec. (12.6).
3. El ejemplo sobre presentación de un conjunto de ecuaciones en notación matricial, ec. (12.7).
4. La definición de la matriz identidad, ec. (12.8).
5. La sección 12.4 sobre inversas hasta la definición del término "inconsistente".

Si se necesita más familiaridad con matrices, los estudiantes pueden proceder con el capítulo o buscar otras fuentes. El texto de Searle (12.1) es ciertamente muy apropiado para biólogos.

## 12.2 Matrices

Una *matriz* es una disposición rectangular de números, llamados también escalares. Por ejemplo,

$$\begin{pmatrix} 2 & 3 & -1 \\ 3 & -8 & 2 \end{pmatrix}$$

es una matriz. Esta puede ser la matriz de coeficientes de un par de ecuaciones lineales en las incógnitas  $x$ ,  $y$ , y  $z$ . Las ecuaciones podrían ser

$$\begin{aligned} 2x + 3y - z &= 8 \\ 3x - 8y + 2z &= 3 \end{aligned} \tag{12.1}$$

La matriz de varianza-covarianza para los datos de la tabla 10.1 es

$$\begin{pmatrix} 1.536/9 & 11.812/9 \\ 11.812/9 & 135.604/9 \end{pmatrix}$$

Las letras mayúsculas en negrilla usualmente simbolizan matrices. La ec. (12.2) define una matriz general

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1c} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2c} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{ic} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{r1} & a_{r2} & \cdots & a_{rj} & \cdots & a_{rc} \end{pmatrix} \tag{12.2}$$

Esta es una matriz  $r \times c$ , o una matriz de *orden* o *dimensiones*  $r$  por  $c$ . Tiene  $r$  filas y  $c$  columnas. El elemento  $a_{ij}$  es la intersección de la  $i$ -ésima fila y la  $j$ -ésima columna; a veces se llama elemento  $ij$ -ésimo. La letra  $i$  es el *índice de fila* y  $j$  es el *índice de columna*.

Los elementos de una fila tendrán algo en común, igual que los elementos de una columna. En la primera ilustración, los elementos de la primera fila son los coeficientes de la primera ecuación, y los elementos de la primera columna son todos coeficientes de  $x$ . Por lo tanto, las matrices son cuadrados de números *dblemente ordenados*.

Finalmente, las matrices deben obedecer a ciertas reglas sobre la combinación de unas con otras. Esto nos permite tener un *álgebra de matrices*, con cada matriz tratada como una entidad aunque los elementos en ella entran en juego.

La *transpuesta* de una matriz es aquella en la cual la primera fila es la primera columna de la original, la segunda fila es la segunda columna, y así sucesivamente. La transpuesta se simboliza  $\mathbf{A}'$ . Se verifica la ec. (12.3).

$$\mathbf{A}' = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{r1} \\ a_{12} & a_{22} & \cdots & a_{r2} \\ \dots & \dots & \dots & \dots \\ a_{1j} & a_{2j} & \cdots & a_{rj} \\ \dots & \dots & \dots & \dots \\ a_{1c} & a_{2c} & \cdots & a_{rc} \end{pmatrix} \quad (12.3)$$

Si  $\mathbf{A}$  es una matriz  $r \times c$  entonces  $\mathbf{A}'$  es  $c \times r$ .

Cuando  $r = c$ ,  $\mathbf{A}$  es una *matriz cuadrada*. Los elementos  $a_{ii}$  están situados en la *diagonal principal*. Si  $a_{ij} = a_{ji}$  entonces  $\mathbf{A}$  es *simétrica* y  $\mathbf{A} = \mathbf{A}'$ .

Una matriz de una columna es un *vector* o un *vector columna*. Por ejemplo,

$$\begin{pmatrix} 8 \\ 3 \end{pmatrix} \quad \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} \quad \text{y} \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

son vectores. El primero es un vector de los términos constantes en las dos ecuaciones, el segundo es una solución de las ecuaciones, y el último es una lista, o vector de las incógnitas.

Una matriz que es una sola fila se llama *vector fila*.

**Ejercicio 12.2.1** Ver la tabla 2.3. La columna  $Y_i$  es un vector de observaciones. La columna  $Y_i - \bar{Y}$  es un vector de desviaciones con signo.

Hallar otras cinco tablas que presenten vectores de observaciones.

**Ejercicio 12.2.2** Veáñse en la tabla 5.5 las columnas encabezadas con 4.4 mm. y 9.9 mm Hg. Estos datos son una matriz de observaciones. ¿Cuáles son los valores de  $r$  y  $c$ ? ¿Qué tienen en común los elementos de una fila? ¿Qué tienen en común los elementos de una columna?

**Ejercicio 12.2.3** Las observaciones en la tabla 10.2 están presentadas como una matriz de datos. ¿Cuáles son sus dimensiones? ¿Qué tienen en común los elementos de una fila? ¿Los elementos de una columna?

Escribir la transpuesta de la matriz de datos. En la transpuesta, ¿qué tienen en común los elementos de una fila? ¿Los elementos de una columna? ¿Cuál es la dimensión de la matriz transpuesta?

**Ejercicio 12.2.4** ¿Qué matriz dada en esta sección es una matriz cuadrada? Nombrar una clase de matrices que siempre deben ser cuadradas.

### 12.3 Operaciones con matrices

El álgebra de matrices combina matrices mediante operaciones de la clase que habitualmente se usan en aritmética, es decir, suma, resta, multiplicación y división. Se necesitan definiciones.

La *adición* de matrices se define de manera natural como una operación elemento por elemento sumando los que están en posiciones correspondientes. La ec. (12.4) define la adición.

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} a_{11} + b_{11} & \cdots & a_{1c} + b_{1c} \\ \cdots & \cdots & \cdots \\ \cdots & a_{ij} + b_{ij} & \cdots \\ \cdots & \cdots & \cdots \\ a_{r1} + b_{r1} & \cdots & a_{rc} + b_{rc} \end{pmatrix} \quad (12.4)$$

Evidentemente, la adición exige que las matrices sean del mismo orden o *conformes para la adición*.

Supóngase que un director de ventas registra las ventas de autos, según que los modelos sean compacto, estándar o de lujo y por períodos de la semana, por ejemplo, de lunes a jueves y el fin de semana. Ha recolectado datos de ventas de un nuevo vendedor durante dos semanas. Desea sumar los datos para tener las ventas del período.

$$\begin{pmatrix} 2 & 3 & 1 \\ 5 & 4 & 3 \end{pmatrix} + \begin{pmatrix} 4 & 3 & 2 \\ 7 & 6 & 3 \end{pmatrix} = \begin{pmatrix} 6 & 6 & 3 \\ 12 & 10 & 6 \end{pmatrix}$$

En esta presentación, una fila representa un período; también se podrían presentar en forma igualmente cómoda los datos transpuestos de modo que las columnas fueran los períodos.

El avicultor que mide el consumo de alimentos, producción de huevos por número y producción de huevos por peso por ave, diariamente, obtiene una matriz diaria de datos  $n \times 3$ . Probablemente sumará estas matrices para tener totales en algún período de tiempo que le interese.

Es fácil visualizar las matrices de datos para animales de laboratorio, plantas, etc. Muchos de éstos serán para intervalos de tiempo breves para acumular los datos de un período de estudio más prolongado.

La aplicación de la definición de adición a  $k$  matrices idénticas lleva a la definición de la *multiplicación de una matriz por un escalar  $k$* ; ver la ec. (12.5).

$$k\mathbf{A} = \underbrace{\mathbf{A} + \cdots + \mathbf{A}}_{k \text{ veces}} = \begin{pmatrix} ka_{11} & \cdots & ka_{1c} \\ \cdots & \cdots & \cdots \\ ka_{r1} & \cdots & ka_{rc} \end{pmatrix} \quad (12.5)$$

Esta definición es también aplicable si  $k$  no es entero.

La *substracción* de matrices se deduce directamente.

Supóngase que un investigador asigna diez ratones machos y diez hembras a cada uno de varios tratamientos, midiendo los pesos iniciales y finales. Su primer conjunto de datos consiste en los pesos iniciales para cada ratón registrado por sexo y por tratamiento que se va a administrar. Estos pueden resumirse en forma conveniente por totales en una matriz  $2 \times t = \text{matriz sexo} \times \text{tratamiento}$ . Las medias se obtienen multiplicando por  $\frac{1}{10}$ . Su conjunto resumen final de datos consiste en pesos finales. La diferencia entre las dos matrices de datos mide aumentos medios de peso.

Naturalmente ahora sigue la construcción de cualquier función lineal de matrices.

La *multiplicación* de matrices se define quizás de manera inesperada. Puede describirse como una operación fila por columna. Por ejemplo;

$$\mathbf{AB} = \begin{pmatrix} 2 & 5 \\ 3 & 8 \end{pmatrix} \begin{pmatrix} 1 & 6 \\ 4 & 7 \end{pmatrix} = \begin{pmatrix} 2(1) + 5(4) & 2(6) + 5(7) \\ 3(1) + 8(4) & 3(6) + 8(7) \end{pmatrix} = \begin{pmatrix} 22 & 47 \\ 35 & 74 \end{pmatrix} = \mathbf{C}$$

Obsérvese que los elementos de la primera fila **A** multiplican a los elementos correspondientes de la primera columna de **B** y luego se suman para dar  $c_{11}$ . Los elementos de la primera fila de **A** multiplican a los elementos correspondientes de la segunda columna de **B** para dar  $c_{12}$ . Para obtener los elementos de la segunda fila de **C**, se multiplican los elementos de la segunda fila de **A** por los de la columna apropiada de **B**.

La ecuación (12.6) define la multiplicación de matrices.

$$\mathbf{A}_{r,s} \mathbf{B}_t = \begin{pmatrix} \sum_i a_{1i} b_{i1} & \sum_i a_{1i} b_{i2} & \cdots & \sum_i a_{1i} b_{it} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_i a_{ri} b_{i1} & \sum_i a_{ri} b_{i2} & \cdots & \sum_i a_{ri} b_{it} \end{pmatrix} \quad (12.6)$$

Los subíndices de **A** y **B** indican el número de filas y columnas en cada una. Nótese que **A** necesita tantas columnas como filas tenga **B** para que las dos sean conformes para el producto. Así mismo, las dimensiones de la matriz producto están dadas por el número de filas de **A** y el número de columnas de **B**; es decir, si **A** es  $r \times s$  y **B** es  $s \times t$ , entonces  $\mathbf{AB} = \mathbf{C}$  es  $r \times t$ .

Es evidente que el orden de las matrices generalmente no puede invertirse en la multiplicación; en efecto, ellas no serán conformes si  $r \neq t$  en la ec. (12.6). Para el ejemplo, cuando **AB** y **BA** existen,

$$\begin{aligned} \mathbf{BA} &= \begin{pmatrix} 1 & 6 \\ 4 & 7 \end{pmatrix} \begin{pmatrix} 2 & 5 \\ 3 & 8 \end{pmatrix} \\ &= \begin{pmatrix} 1(2) + 6(3) & 1(5) + 6(8) \\ 4(2) + 7(3) & 4(5) + 7(8) \end{pmatrix} = \begin{pmatrix} 20 & 53 \\ 29 & 76 \end{pmatrix} \end{aligned}$$

Ningún número de la matriz producto **BA** corresponde a ninguno de los en  $\mathbf{AB} = \mathbf{C}$ .

En el caso de **AB**, decimos que **B** está *premultiplicada* por **A** y **A** está *postmultiplicada* por **B**.

Ahora vemos que las ecs. (12.1) pueden escribirse en notación matricial como,

$$\begin{pmatrix} 2 & 3 & -1 \\ 3 & -8 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 8 \\ 3 \end{pmatrix} \quad (12.7)$$

Aquí tenemos una multiplicación de una matriz de coeficientes  $2 \times 3$  y una matriz de incógnitas  $3 \times 1$ , considerada esta última también como un vector. El resultado es una matriz  $2 \times 1$  o vector de escalares.

Por lo demás, podemos escribir los valores separados de las medias de tratamientos en un diseño completamente aleatorizado como sigue:

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix} = \begin{pmatrix} \mu + \tau_1 \\ \mu + \tau_2 \\ \mu + \tau_3 \end{pmatrix}$$

La importante *matriz identidad I* está definida por

$$I = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \dots & \dots & \dots & \cdots & \dots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \quad (12.8)$$

Los elementos de esta matriz cuadrada son 1's en la diagonal principal y 0's fuera de ella. Si una matriz  $r \times r$  se premultiplica o postmultiplica por una matriz identidad  $r \times r$  la matriz original no se altera. Esto es,

$$AI = A = IA \quad \text{para } I_r \text{ y } A_r$$

Sólo se necesita un subíndice en I para indicar su orden puesto que, por definición, es cuadrada.

I es una matriz identidad para la multiplicación porque no cambia la matriz a la cual multiplica. Es como el 1 de la aritmética o álgebra ordinarias;  $1 \times 4 = 4 = 4 \times 1$ .

Finalmente, es necesaria una analogía de la división como operación matricial. La proporciona el uso de matrices inversas (sec. 12.4), y la matriz identidad se necesita en la definición.

**Ejercicio 12.3.1** En las siguientes matrices, las filas representan tres grupos o bloques de conejillos de indias. Las columnas en orden sucesivo dan los pesos iniciales, forraje consumido, y aumento de peso, todo en gramos, para los conejillos de indias. Las matrices son para suelos no fertilizados y fertilizados, origen del forraje. (Ver tabla 17.12).

$$\begin{pmatrix} 220 & 1,155 & 224 \\ 246 & 1,423 & 289 \\ 262 & 1,576 & 280 \end{pmatrix} \quad \begin{pmatrix} 222 & 1,326 & 237 \\ 268 & 1,559 & 265 \\ 314 & 1,528 & 256 \end{pmatrix}$$

Sumar las dos matrices.

**Ejercicio 12.3.2** Estas matrices son parecidas a las del ejercicio 12.3.1, pero para un suelo de origen diferente.

$$\begin{pmatrix} 198 & 1,092 & 118 \\ 266 & 1,703 & 191 \\ 335 & 1,546 & 115 \end{pmatrix} \quad \begin{pmatrix} 205 & 1,154 & 82 \\ 236 & 1,250 & 117 \\ 268 & 1,667 & 117 \end{pmatrix}$$

Sumar estas dos matrices.

**Ejercicio 12.3.3** Las siguientes matrices de datos corresponden a pesos iniciales y aumentos de peso de varios grupos (filas) de conejillos de indias asignados a raciones derivadas de varias combinaciones de fertilizantes y suelos (columnas). (Ver tabla 17.12).

$$\begin{pmatrix} 220 & 222 & 198 & 205 \\ 246 & 268 & 266 & 236 \\ 262 & 314 & 335 & 268 \end{pmatrix} \quad \begin{pmatrix} 224 & 237 & 118 & 82 \\ 289 & 265 & 191 & 117 \\ 280 & 256 & 115 & 117 \end{pmatrix}$$

Sumar estas matrices. ¿Cuál es la variable medida por el resultado?

**Ejercicio 12.3.4** Las siguientes matrices de datos son aumentos de peso en libras para lechones, machos y hembras, según el corral (filas) en que estaban y la ración (columna) con que fueron alimentados. (Ver tabla 17.7).

$$\begin{pmatrix} 9.52 & 8.51 & 9.11 \\ 8.21 & 9.95 & 8.50 \\ 9.32 & 8.43 & 8.90 \\ 10.56 & 8.86 & 9.51 \\ 10.42 & 9.20 & 8.76 \end{pmatrix} \quad \begin{pmatrix} 9.94 & 10.00 & 9.75 \\ 9.48 & 9.24 & 8.66 \\ 9.32 & 9.34 & 7.63 \\ 10.90 & 9.68 & 10.37 \\ 8.82 & 9.67 & 8.57 \end{pmatrix}$$

Restar la primera matriz de la segunda. ¿En cuántos casos el aumento de peso de las hembras excede o iguala al de los machos?

**Ejercicio 12.3.5** Multiplicar las matrices encontradas en los ejercicios 12.3.1 y 12.3.2 por el escalar  $\frac{1}{2}$ ; observese que los resultados son promedios de las dos raciones en cada caso.

**Ejercicio 12.3.6** Efectuar las siguientes multiplicaciones matriciales.

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \quad \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

**Ejercicio 12.3.7** Efectuar la siguiente multiplicación matricial.

$$AB = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 3 & 2 \\ 2 & 3 \\ 1 & 1 \end{pmatrix}$$

¿Podemos agregar más filas a  $A$  y todavía hacer la multiplicación matricial? ¿Más columnas?

¿Podemos añadir más filas a  $B$  y aún efectuar el producto matricial? ¿Más columnas?

**Ejercicio 12.3.8** Los siguientes datos provienen del ejercicio 7.3.1. Completar la operación matricial indicada.

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 89.8 & 93.8 & 88.4 & 112.6 \end{pmatrix} \begin{pmatrix} 1 & 89.8 \\ 1 & 93.8 \\ 1 & 88.4 \\ 1 & 112.6 \end{pmatrix}$$

En términos generales, ¿qué significan los cuatro números que se tienen ahora?

**Ejercicio 12.3.9** Hallar  $X'X$ , donde

$$X' = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{pmatrix}$$

Hallar  $X\beta$ , donde

$$\beta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

Hallar  $X'Y$ , donde  $Y' = (Y_1, \dots, Y_n)$ .

Las ecuaciones  $(X'X)\hat{\beta} = X'Y$  son las que se resolvieron para dar estimaciones de  $\alpha$  y  $\beta$  en el problema de regresión, cap. 10.

## 12.4 Inversas, dependencia lineal, y rango

Considérese la ec. (12.7), una ecuación en que entra un vector de incógnitas. Si fuera posible dividir ambos miembros por la matriz de coeficientes, la ecuación estaría resuelta y sería de esperar que los valores fueran iguales a las incógnitas.

En aritmética o álgebra, la multiplicación por el inverso de un número es equivalente a la división por tal número; todo número diferente de cero, multiplicado por su inverso da la unidad. Esta idea proporciona la clave para encontrar la operación análoga de la división que se necesita.

La *inversa* de una matriz es como un inverso, en cuanto que una matriz multiplicada por su inversa da la matriz identidad.

En aritmética, la división por cero no está definida, de modo que una definición útil de su inverso tampoco existe. En forma análoga, hay matrices cuadradas para las cuales la inversa no está definida. Por tanto, la inversa está definida sólo para ciertas matrices cuadradas.

Cuando existe una inversa de una matriz cuadrada  $A$  se indica con el símbolo  $A^{-1}$  llamado  $A$  inversa. Tal matriz es única y es inversa tanto a la derecha como a la izquierda, así que se cumple la ec. (12.9).

$$AA^{-1} = I = A^{-1}A \quad (12.9)$$

La ecuación (12.9) define la inversa de  $A$ .

Observando la ecuación (12.9), se ve claramente la naturaleza del problema de hallar una inversa. Supóngase que  $A$  es conocida, pero que los elementos de  $A^{-1}$  no se conocen. Con la ec. (12.9), construimos la siguiente ilustración:

$$\begin{pmatrix} 1 & 2 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

La matriz de letras representa  $A^{-1}$ . Completando las multiplicaciones, tenemos, elemento por elemento,

$$a + 2c = 1 \quad 3a - c = 0 \quad b + 2d = 0 \quad 3b - d = 1$$

Estas ecuaciones simultáneas se resuelven fácilmente por pares para obtener  $a = \frac{1}{7}$ ,  $b = \frac{2}{7}$ ,  $c = \frac{3}{7}$ , y  $d = -\frac{1}{7}$ . Ahora bien

$$\begin{pmatrix} 1 & 2 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{7} & \frac{2}{7} \\ \frac{3}{7} & -\frac{1}{7} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{7} & \frac{2}{7} \\ \frac{3}{7} & -\frac{1}{7} \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & -1 \end{pmatrix}$$

Por otra parte, si elegimos

$$A = \begin{pmatrix} 1 & 2 \\ 4 & 8 \end{pmatrix}$$

las ecuaciones implicadas no se pueden resolver. En particular,  $a + 2c = 1$  de donde  $a = 1 - 2c$ . Sustituyendo en  $4a + 8c = 0$ , obtenemos  $4 - 8c + 8c = 4 \neq 0$ . Las dos ecuaciones no tienen solución y decimos que son inconsistentes. Se encuentra otra contradicción si tomamos las ecuaciones con  $b$  y  $d$ . Esta matriz no tiene inversa.

En la segunda matriz  $A$ , la segunda columna es múltiplo de la primera. Podemos escribir

$$2 \begin{pmatrix} 1 \\ 4 \end{pmatrix} - \begin{pmatrix} 2 \\ 8 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Si consideramos la matriz  $A$  como dos vectores columna de dos elementos cada uno, entonces  $A = (C_1 C_2)$  y

$$2C_1 - C_2 = 0$$

Obsérvese que

$$\mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Este se conoce como *vector nulo* o *cero*.

En general, siempre que una función lineal de los vectores columna de una matriz la hagan igual a cero, estos son *linealmente dependientes*. Algun vector del sistema está compuesto de información que ya dan otros vectores; es pues redundante.

Más especialmente, para  $\mathbf{A} = (\mathbf{C}_1, \dots, \mathbf{C}_c)$  donde  $\mathbf{C}_i$  es una columna de  $r$  elementos, las columnas de  $\mathbf{A}$  son *linealmente dependientes* si hay  $c$  escalares,  $\lambda_1, \dots, \lambda_c$  no todos ceros, tales que

$$\sum \lambda_i \mathbf{C}_i = \mathbf{0} \quad (12.10)$$

Si la ec. (12.10) se cumple sólo cuando todos los  $\lambda_i = 0$  entonces se dice que los vectores son *linealmente independientes*.

De un conjunto de vectores linealmente dependientes, elimíñese uno de ellos con coeficiente diferente de cero. El vector eliminado es una combinación lineal de los otros, así que no tiene información que no esté ya contenida en los otros vectores. Examíñese si hay dependencia lineal de los otros vectores y continúese el proceso hasta obtener un conjunto linealmente independiente. Este conjunto contiene el máximo número de columnas linealmente independientes en  $\mathbf{A}$ . Este máximo se llama *rango* de la matriz; se designa  $r(\mathbf{A})$ . En la práctica, para encontrar el rango de una matriz se emplea un procedimiento con operadores elementales que no expondremos aquí. Este procedimiento está programado para aplicación al computador.

El rango de columna de una matriz es igual al rango de fila, ya que se obtendría el mismo valor si hubiéramos empezado con las filas como vectores. El rango de una matriz es pues único y no puede exceder al  $\min(r, c)$ .

Recuérdese que la matriz debe ser cuadrada para tener inversa, y que no toda matriz cuadrada tiene inversa. La inversa de una matriz existe sólo cuando la matriz es de rango completo, o sea,  $r(\mathbf{A}) = r = c$ . Tal matriz recibe el nombre de *no singular*. Cuando una matriz no es de rango completo, se dice que es *singular* o que tiene una o más singularidades.

Un método para determinar cuándo una matriz cuadrada tiene inversa o es no singular, es evaluar su *determinante*. El determinante de una matriz es un número o escalar simbolizado por  $|\mathbf{A}|$ . Para una matriz  $2 \times 2$ , el determinante está definido por

$$|\mathbf{A}| = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc \quad (12.11)$$

Para las matrices  $2 \times 2$   $\mathbf{A}$  usadas antes en esta sección, los determinantes son:

$$\begin{vmatrix} 1 & 2 \\ 3 & -1 \end{vmatrix} = 1(-1) - 2(3) = -7$$

y

$$\begin{vmatrix} 1 & 2 \\ 4 & 8 \end{vmatrix} = 1(8) - 2(4) = 0$$

Una matriz con determinante diferente de cero es no singular y tiene una inversa; en caso contrario, será singular y su inversa no está definida.

Para una matriz  $3 \times 3$  el determinante es evaluado por

$$|\mathbf{A}| = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - ceg - afh - bdi \quad (12.12)$$

La evaluación es fácil si se escriben primero las columnas 1 y 2 después de  $\mathbf{A}$  en un cuadro de  $3 \times 5$ . Los productos de los elementos de cada una de las tres diagonales de izquierda a derecha dan los términos positivos, en tanto que los mismos para las tres diagonales de derecha a izquierda dan los términos negativos.

No estamos directamente interesados en la evaluación de determinantes para matrices de orden mayor, dando por sentado que éstos se calcularán electrónicamente. Pero la definición del determinante de una matriz  $r \times r$  está dada por

$$|\mathbf{A}| = \begin{cases} \sum_{j=1}^r a_{ij}(-1)^{i+j} |\mathbf{M}_{ij}| & \text{para cualquier } i \\ \sum_{i=1}^r a_{ij}(-1)^{i+j} |\mathbf{M}_{ij}| & \text{para cualquier } j \end{cases} \quad (12.13)$$

Esta definición se llama de *desarrollo por los elementos de la i-ésima fila*, en el primer caso, o por los *elementos de la j-ésima columna* en el segundo caso.

$\mathbf{M}_{ij}$  es un *menor* de  $\mathbf{A}$ . Se halla eliminando la  $i$ -ésima fila y la  $j$ -ésima columna de  $\mathbf{A}$  para obtener una matriz  $(r-1) \times (r-1)$  y tomando luego su determinante. El menor, con su signo correspondiente, se llama *cofactor* de  $a_{ij}$  en  $|\mathbf{A}|$ .

Las ecuaciones (12.13) y (12.12) nos permiten evaluar una matriz  $4 \times 4$ . Con esta evaluación y la ec. (12.13) podemos pasar a matrices  $5 \times 5$  y así sucesivamente.

En general, el desarrollo de una matriz se presenta como un polinomio en los elementos de la matriz. Cada término del desarrollo final será un producto, con el signo apropiado, en el cual hay un único elemento de cada fila y columna de la matriz, es decir, un producto de  $r$  elementos.

Finalmente, la *inversa* de una matriz  $2 \times 2$  está dada por

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \quad (12.14)$$

En general, tenemos

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} \begin{pmatrix} \mathbf{M}_{11} & -\mathbf{M}_{21} & \mathbf{M}_{31} & \cdots & (-1)^{r+1}\mathbf{M}_{r1} \\ -\mathbf{M}_{12} & \mathbf{M}_{22} & -\mathbf{M}_{32} & \cdots & (-1)^{r+2}\mathbf{M}_{r2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ (-1)^{r+1}\mathbf{M}_{1r} & (-1)^{r+2}\mathbf{M}_{2r} & (-1)^{r+3}\mathbf{M}_{3r} & \cdots & \mathbf{M}_{rr} \end{pmatrix} \quad (12.15)$$

Se supone que la determinación de esta inversa, si se requiere, será hecha por un computador electrónico. El método empleado para el cálculo probablemente no será el que supone la ec. (12.15).

**Ejercicio 12.4.1** Hallar las inversas de las matrices

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad \text{y} \quad \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$$

Verificar que ambas son inversas, tanto a la izquierda como a la derecha.

**Ejercicio 12.4.2** Hallar la inversa de  $\mathbf{AB}$ , calculada en el ejercicio 12.3.7.

**Ejercicio 12.4.3** Demostrar que el determinante de

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 4 & 8 \end{pmatrix}$$

es cero.

**Ejercicio 12.4.4** Para  $(\mathbf{X}'\mathbf{X})$  del ejercicio 12.3.9, hallar la inversa. Multiplicar ambos miembros de la ecuación matricial dada allí, para encontrar una solución para  $\hat{\beta}$ . Mostrar que esta solución es la misma obtenida en el cap. 10.

Calcular  $\hat{\beta}'(\mathbf{X}'\mathbf{Y})$ . ¿A qué cantidad conocida es igual esto?

## Referencias

- 12.1. Searle, S. R.: *Matrix Algebra for the Biological Sciences*, Wiley, Nueva York, 1966.

---

## CAPITULO TRECE

---

# REGRESION LINEAL EN NOTACION MATRICIAL

### 13.1 Introducción

En los capítulos siguientes, expondremos la regresión múltiple y problemas de número desproporcionado de subclases. La aritmética de los análisis es muy tediosa, de modo que se deja a los computadores de alta velocidad. Con todo, los conceptos estadísticos fundamentales deben quedar claros en las limitaciones de los programas en computador y sus resultados deben comprenderse totalmente. Como ayuda, confiemos en la notación matricial. Aunque los problemas son complejos, su presentación en esta forma es concisa. Sigue lo mismo con las soluciones.

En este capítulo, el problema de regresión lineal simple se presenta en notación matricial y está relacionado con el cap. 10. Así que el material es conocido. Además, las soluciones desarrolladas en forma de matriz fácilmente se generalizan a los problemas más complejos de los capítulos posteriores. Las ecuaciones de especial importancia están así señaladas.

### 13.2 El modelo y la estimación de mínimos cuadrados

La ecuación para el modelo de regresión lineal se ha dado en la ec. (10.5) como

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n$$

Entonces podemos escribir el conjunto de observaciones como

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ &\cdots \cdots \cdots \\ Y_n &= \beta_0 + \beta_1 X_n + \varepsilon_n \end{aligned} \tag{13.1}$$

Estas  $n$  ecuaciones tienen elementos que caen naturalmente dentro de categorías. A la izquierda de las igualdades están las observaciones propiamente dichas. A la derecha están los coeficientes de los parámetros, los parámetros y las componentes de error. En notación vectorial y matricial, pueden escribirse, respectivamente, como

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

El vector  $\mathbf{Y}$  es el vector aleatorio de observaciones. La *matriz*  $\mathbf{X}$ , llamada también *matriz de diseño*, es una matriz de parámetros observables. Aquí, la columna de los 1 puede considerarse como de valores de una *variable ficticia*  $X_{0i} = 1, i = 1, \dots, n$ . El vector  $\boldsymbol{\beta}$  es un vector de parámetros desconocidos y  $\boldsymbol{\varepsilon}$  es un vector de componentes aleatorias desconocidas.

Ahora las ecuaciones (13.1) pueden escribirse como

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

o, en forma compacta, como la ec. (13.2) matricial.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (13.2)$$

Esta importante expresión de la ecuación modelo debe entenderse perfectamente.

En términos de los datos de la tabla 10.1,

$$\mathbf{Y} = \begin{pmatrix} 87.1 \\ 93.1 \\ \vdots \\ 94.4 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & 4.6 \\ 1 & 5.1 \\ \vdots & \vdots \\ 1 & 5.1 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

Para el modelo, suponemos que los  $\varepsilon$  tienen media poblacional cero y una varianza poblacional común  $\sigma^2$ , y no están correlacionados. Cuando se tengan que hacer contrastes de significancia, los  $\varepsilon$  también deben ser de una población normal.

La mayoría de los aspectos del modelo pueden presentarse en notación matricial. Por el supuesto de que los  $\varepsilon$  tienen media cero, escribimos  $E(\boldsymbol{\varepsilon}) = \mathbf{0}_n$ , el vector nulo. Usamos  $E$  para valor esperado o media, idea inicialmente introducida en la sec. 5.10 y usada repetidamente al hablar acerca de valores esperados de cuadrados medios. Entonces, la ec. (13.2) implica que  $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ ; ésta es la expresión necesaria acerca de la media. Finalmente obsérvese que

$$\mathbf{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \begin{pmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 & \varepsilon_1\varepsilon_3 & \cdots & \varepsilon_1\varepsilon_n \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 & \varepsilon_2\varepsilon_3 & \cdots & \varepsilon_2\varepsilon_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & \varepsilon_n\varepsilon_3 & \cdots & \varepsilon_n^2 \end{pmatrix}$$

Los cuadrados en la diagonal indican varianza; los productos cruzados fuera de la diagonal indican covarianzas. En consecuencia, las suposiciones concernientes a varianza homogénea y errores no correlacionados pueden plantearse como sigue

$$E(\varepsilon\varepsilon') = \begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

$$= I\sigma^2 \quad \text{o} \quad I\sigma_{Y \cdot X}^2$$

A partir de la muestra,  $\beta$  se estima por  $\hat{\beta}$ , por ejemplo, o  $\mathbf{b}$ . Para estimar la media poblacional cuando  $\mathbf{X} = \mathbf{X}_0$ , escribimos  $\mathbf{X}'_0 = (1 \ X_0)$  y luego la ecuación de regresión muestral como

$$\hat{\mu}_Y = \mathbf{X}'_0 \hat{\beta} \quad \text{o} \quad \hat{Y} = \mathbf{X}'_0 \hat{\beta} \quad (13.3)$$

Cuando  $X_0 = X_i$ , uno de los  $X$  observados, entonces  $\hat{Y}$  se llama *valor de regresión* o *valor ajustado* y  $e = Y - \hat{Y}$  es una desviación de la observación respecto de la recta de regresión, o un residuo. La suma de cuadrados de los residuos es una medida del defecto global del modelo para ajustarse a los datos, es decir, es la suma de cuadrados del error. Esto, también, puede escribirse en la nueva notación,

$$\mathbf{e}'\mathbf{e} = (e_1, e_2, \dots, e_n) \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \sum e_i^2$$

Ordinariamente, elegimos  $\hat{\beta}$  o  $\mathbf{b}$  de modo que  $\sum e^2$  sea mínima; es decir, obtenemos *estimaciones de mínimos cuadrados* de los parámetros del modelo. Las ecuaciones de mínimos cuadrados o *ecuaciones normales* que proporcionan estas estimaciones son

$$(\mathbf{Z}'_n \mathbf{Z}_2)_2 \hat{\beta}_1 = \mathbf{Z}'_n \mathbf{Y}_1 \quad (13.4)$$

Esta ecuación matricial es típica para ecuaciones normales.

La matriz  $\mathbf{Z}'\mathbf{Z}$  de nuestro problema es de dimensión  $2 \times 2$  y la de  $\mathbf{Z}'\mathbf{Y}$  es  $2 \times 1$ . Para los datos sin codificar de la tabla 10.1,

$$\mathbf{Z}'\mathbf{Z} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 4.6 & 5.1 & \cdots & 5.1 \end{pmatrix} \begin{pmatrix} 1 & 4.6 \\ 1 & 5.1 \\ \vdots & \vdots \\ 1 & 5.1 \end{pmatrix}$$

$$= \begin{pmatrix} 10 & 49.8 \\ 49.8 & 249.54 \end{pmatrix}$$

Estos números son  $10 = n$ ,  $49.8 = \sum X_i$ , y  $249.54 = \sum X_i^2$ . Así, la ec. (13.5) es válida para regresión lineal.

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix} \quad (13.5)$$

Para los mismos datos,

$$\begin{aligned} \mathbf{X}'\mathbf{Y} &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 4.6 & 5.1 & \cdots & 5.1 \end{pmatrix} \begin{pmatrix} 87.1 \\ 93.1 \\ \vdots \\ 94.4 \end{pmatrix} \\ &= \begin{pmatrix} 935.6 \\ 4671.10 \end{pmatrix} \end{aligned}$$

Estos números son  $935.6 = \sum Y_i$  y  $4,671.1 = \sum X_i Y_i$ . Se verifica la ec. (13.6).

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix} \quad (13.6)$$

Si  $\mathbf{X}'\mathbf{X}$  es no singular, entonces existe su inversa y la solución de la ecuación matricial (13.4) está dada por

$$_2\hat{\mathbf{b}}_1 = _2\hat{\beta}_1 = _2(\mathbf{X}'\mathbf{X})^{-1} _2\mathbf{X}'_{n,n}\mathbf{Y}_1 \quad (13.7)$$

Esta ecuación debe entenderse y memorizarse como la solución de un conjunto de ecuaciones normales cuando  $\mathbf{X}'\mathbf{X}$  es no singular.

La inversa de  $(\mathbf{X}'\mathbf{X})$  puede calcularse usando la ec. (12.14) para obtener

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n \sum X_i^2 - (\sum X_i)^2} \begin{pmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{pmatrix} \quad (13.8)$$

Para la ilustración numérica,

$$\begin{aligned} (\mathbf{X}'\mathbf{X})^{-1} &= \frac{1}{2,495.4 - 2,480.04} \begin{pmatrix} 249.54 & -49.8 \\ -49.8 & 10 \end{pmatrix} \\ &= \begin{pmatrix} 16.24609375 & -3.2421875 \\ -3.2421875 & 0.65104167 \end{pmatrix} \end{aligned}$$

Finalmente, la ec. (13.7) y alguna manipulación algebraica dan

$$\mathbf{b} = \hat{\boldsymbol{\beta}} = \left( \begin{array}{c} \bar{Y} - \hat{\beta}_1 \bar{X} \\ \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2/n} \end{array} \right) \quad (13.9)$$

Estas son las expresiones familiares del cap. 10. Para el problema numérico,

$$\begin{aligned} (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) &= \begin{pmatrix} 16.24609375 & -3.2421875 \\ -3.2421875 & 0.65104167 \end{pmatrix} \begin{pmatrix} 935.6 \\ 4671.10 \end{pmatrix} \\ &= \begin{pmatrix} 55.26 \\ 7.69 \end{pmatrix} \end{aligned}$$

Estas estimaciones son las mismas que se encontraron en el cap. 10.

**Ejercicio 13.2.1** Por multiplicación directa, mostrar que  $(\mathbf{X}'\mathbf{X})^{-1}$ , que se encontró para los datos de Leghorn, es inversa a la derecha y a la izquierda.

**Ejercicio 13.2.2** Redondear los elementos en la matriz inversa anterior a dos cifras decimales y calcular el vector  $\mathbf{b}$  o  $\hat{\boldsymbol{\beta}}$ . Obsérvese la discrepancia entre éste y el resultado dado en el texto. (*Nota:* en cálculos de regresión, consérvense tantas cifras como sea posible o si no los errores de redondeo pueden crear serias dificultades).

**Ejercicio 13.2.3** Para el ejercicio 10.2.1, ¿cuál es el vector  $\mathbf{Y}$ ? La matriz  $\mathbf{X}$ ? El vector  $\hat{\boldsymbol{\beta}}$ ? Calcular  $\mathbf{X}'\mathbf{X}$  y  $\mathbf{X}'\mathbf{Y}$ . Hallar  $(\mathbf{X}'\mathbf{X})^{-1}$ . Calcular  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ .

Comparar  $\hat{\boldsymbol{\beta}}$  con el intercepto y la pendiente halladas para la ecuación en el ejercicio 10.2.1.

**Ejercicio 13.2.4** Repetir el ejercicio 13.2.3 para el ejercicio 10.2.3.

**Ejercicio 13.2.5** Repetir el ejercicio 13.2.3 para el ejercicio 10.2.4.

**Ejercicio 13.2.6** Repetir el ejercicio 13.2.3 para el ejercicio 10.2.5.

**Ejercicio 13.2.7** Repetir el ejercicio 13.2.3 para cada conjunto de datos en el ejercicio 10.8.1.

**Ejercicio 13.2.8** Repetir el ejercicio 13.2.3 para cada conjunto de datos en el ejercicio 10.8.3.

### 13.3 El análisis de la varianza

Para estimar  $\sigma^2$ , se requiere la suma residual muestral de cuadrados; está dada por

$$\begin{aligned} \text{SC(residuos)} &= \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} \quad n - 2 \text{ gl} \end{aligned} \quad (13.10)$$

La importancia de esta ecuación se hará más clara a medida que avancemos. Se la deberá comprender perfectamente a un nivel operacional.

La primera expresión de la derecha para  $\mathbf{e}'\mathbf{e} = \sum e_i^2$  puede deducirse de la ec. (13.2); la expresión final exige un poco más de pericia algebraica que la que se dio en el cap. 12.

La ecuación (13.10) de la suma de cuadrados residual como diferencia entre la suma de cuadrados de las observaciones,  $\mathbf{Y}'\mathbf{Y}$ , y una componente que tiene que ver con el ajuste del modelo, es decir, con los parámetros en  $\hat{\beta}$ , para los datos White Leghorn.

$$SC(\text{total, sin ajuste}) = \mathbf{Y}'\mathbf{Y} = \sum Y_i^2 = 87,670.34 \quad n = 10 \text{ gl}$$

y

$$\begin{aligned} SC(\text{modelo}) &= \hat{\beta}'\mathbf{X}'\mathbf{Y} = (55.26 \quad 7.69) \begin{pmatrix} 935.6 \\ 4671.1 \end{pmatrix} \\ &= 87,625.644 \quad 2 \text{ gl} \end{aligned}$$

En estos cálculos se usaron más cifras decimales que las que se han indicado.

Obsérvese la ec. (13.11).

$$SC(\text{modelo}) = \hat{\beta}'\mathbf{X}'\mathbf{Y} \quad 2 \text{ gl} \quad (13.11)$$

Esta suma de cuadrados tiene dos grados de libertad, el rango de  $\mathbf{X}'\mathbf{X}$ .

Para encontrar la contribución atribuible a la regresión, restese el factor de corrección de la ec. (13.11).

$$\begin{aligned} SC(\text{regresión}) &= \hat{\beta}'\mathbf{X}'\mathbf{Y} - \frac{(\sum Y)^2}{n} \quad 1 \text{ gl} \\ &= 87,625.644 - 87,534.736 = 90.908 \end{aligned}$$

Finalmente, el análisis de la varianza se presenta como en la tabla 13.1. El contraste  $F$  es el de  $H_0: \beta_1 = 0$  frente a  $H_1: \beta_1 \neq 0$ , un contraste frente a alternativas bilaterales; es un contraste para evaluar si la variación de  $X$  contribuye o no a la variación en  $Y$ . La tabla 13.1 es esencialmente la misma tabla 10.3. Se han usado más posiciones decimales que las indicadas, lo cual es necesario cuando los datos están sin codificar.

Tabla 13.1 Análisis de la varianza para regresión

Fuente	gl	SC	CM	F
Regresión, $X$	1	$\hat{\beta}'\mathbf{X}'\mathbf{Y} - (\sum Y)^2/n = 90.908$	90.908	16.3**
Residuo	$n - 2 = 8$	$\mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y} = 44.696$	5.587	
Total	$n - 1 = 9$	$\mathbf{Y}'\mathbf{Y} - (\sum Y)^2/n = 135.604$		

Ejercicio 13.3.1 Usando los procedimientos dados en este capítulo, completar un análisis de la varianza de los datos de la tabla 10.5. Indicar la ecuación de regresión.

Ejercicio 13.3.2 Calcular  $\hat{\beta}'X'Y$ ,  $Y'Y$ ,  $(\sum Y)^2/n$ , y presentar el análisis de la varianza para los datos del ejercicio 10.2.1. Usar los cálculos del ejercicio 13.2.3.

Ejercicio 13.3.3. Repetir el ejercicio 13.3.2 usando los ejercicios 10.2.3 y 13.2.4.

Ejercicio 13.3.4 Repetir el ejercicio 13.3.2 usando los ejercicios 10.2.4 y 13.2.5.

Ejercicio 13.3.5 Repetir el ejercicio 13.3.2 usando los ejercicios 10.2.5 y 13.2.6.

Ejercicio 13.3.6 Repetir el ejercicio 13.3.2 usando los ejercicios 10.8.1 y 13.2.7.

Ejercicio 13.3.7. Repetir el ejercicio 13.3.2 usando los ejercicios 10.8.3 y 13.2.8.

### 13.4 Desviaciones estándar, intervalos de confianza y pruebas de hipótesis

Si bien el análisis de la varianza ofrece una prueba de  $H_0: \beta_1 = 0$ , no da directamente desviaciones estándar de  $\hat{\beta}_0$  y  $\hat{\beta}_1$ . Se tiene una estimación de  $\sigma^2$  y las varianzas y desviaciones estándar pueden hallarse, tediosamente, considerando cada  $\hat{\beta}$  o  $b$  como una combinación lineal de los  $Y$ . Sin embargo, con un enfoque matricial, la matriz de varianza-covarianza para  $\hat{\beta}$  está dada por

$$\begin{aligned} V(\hat{\beta}) &= \begin{pmatrix} \sigma^2(b_0) & \sigma(b_0, b_1) \\ \sigma(b_1, b_0) & \sigma^2(b_1) \end{pmatrix} \\ &= (X'X)^{-1}\sigma_{Y \cdot X}^2 \\ &= \begin{pmatrix} \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X - \bar{X})^2}\right)\sigma_{Y \cdot X}^2 & \frac{-\bar{X}\sigma_{Y \cdot X}^2}{\sum(X - \bar{X})^2} \\ \frac{-\bar{X}\sigma_{Y \cdot X}^2}{\sum(X - \bar{X})^2} & \frac{\sigma_{Y \cdot X}^2}{\sum(X - \bar{X})^2} \end{pmatrix} \end{aligned} \quad (13.12)$$

Las dos primeras líneas de la ec. (13.12) son otras ecuaciones esenciales en la presentación matricial.

Las varianzas están en la diagonal principal y las covarianzas son elementos fuera de la diagonal. En los datos de un problema real,  $\sigma_{Y \cdot X}^2$  se reemplaza por su valor estimado,  $s_{Y \cdot X}^2$ . Nótese que  $\sigma^2(b_1)$  en la última forma de la ec. (13.12) es el parámetro correspondiente a la estimación muestral en la ec. (10.12) y  $\sigma^2(b_0)$  es el parámetro correspondiente a la estimación en la ec. (10.13) cuando  $X = 0$ .

La expresión  $\sigma(b_0, b_1) = \sigma(b_1, b_0)$  es la covarianza entre  $b_0$  y  $b_1$ . No es cero a menos que  $X = 0$ . Cuando la covarianza no es cero, las sumas de cuadrados no son aditivas. Por tanto, calculamos

$$SC(\text{regresión}) = SC(b_1 | b_0)$$

$$\begin{aligned}
 &= SC(\text{modelo}) - SC(b_0 \text{ omitiendo la regresión}) \\
 &= SC(\text{modelo}) - (\sum Y)^2/n
 \end{aligned}$$

O sea que calculamos  $SC(\text{regresión})$  como una suma de cuadrados adicional después de ajustar la media en tanto se *omite la regresión* o no *ajustada* para regresión. Así mismo, calculamos  $SC(\text{intercepto})$  ajustando la regresión por el origen y luego hallando una suma de cuadrados adicional debido a la inclusión de una intersección con y en el modelo. Es decir,

$$\begin{aligned}
 SC(\text{intercepto}) &= SC(b_0 | b_1) \\
 &= SC(\text{modelo}) - SC(b_1 \text{ sin ajuste}) \\
 &= SC(\text{modelo}) - \frac{(\sum XY)^2}{\sum X^2}
 \end{aligned}$$

La motivación se bosquejó en el último párrafo de la sec. 10.12.

Para los datos White Leghorn, se da a continuación la matriz muestra de varianza-covarianza.

$$\begin{aligned}
 V(\hat{\beta}) &= \begin{pmatrix} 16.24609375 & -3.2421875 \\ -3.2421875 & 0.65104167 \end{pmatrix} 5.587 \\
 &= \begin{pmatrix} 90.76643637 & -18.11400389 \\ -18.11400389 & 3.63734611 \end{pmatrix}
 \end{aligned}$$

Para probar  $H_0: \beta_1 = 0$  frente a  $H_1: \beta_1 \neq 0$ , podemos escribir directamente por  $\hat{\beta}$  y  $V(\hat{\beta})$ ,

$$t = \frac{7.69}{\sqrt{3.6373}} = 4.03^{**} \quad 8 \text{ gl}$$

Se descarta la hipótesis nula; lo comprobado no respalda la hipótesis de que la pendiente de la recta de regresión es cero.

Para construir un intervalo de confianza de 95 por ciento, calcúlense los límites como sigue:

$$IC(\beta_1) = 7.69 \pm 2.306\sqrt{3.6373} = (3.29, 12.09)$$

A menos que la muestra obtenida sea de tan baja ocurrencia como 1 vez en 20, en promedio, la verdadera pendiente de la recta de regresión estaría en este intervalo. Difiere sólo en una unidad en la segunda cifra decimal respecto de la que se obtuvo en el cap. 10.

Para probar,  $H_0: \beta_0 = 0$  frente a  $H_1: \beta_0 \neq 0$ , la hipótesis de que la regresión lineal pasa por el origen, calcúlese

$$t = \frac{55.26}{\sqrt{90.7664}} = 5.80^{**} \quad 8 \text{ gl}$$

Concluimos que la regresión no pasa por el origen. Esta hipótesis podría no ser de mucho interés en este problema; sólo se presenta para ilustrar la técnica. Para el intercepto es fácil y más apropiado construir un intervalo de confianza.

$$\text{IC}(\beta_0) = 55.26 \pm 2.306\sqrt{90.7664} = (33.29, 77.23)$$

**Ejercicio 13.4.1** Para los datos de la tabla 10.5, hallar la matriz varianza-covarianza para el vector  $\hat{\beta}$ . Construir estimaciones de intervalos de confianza del 95 por ciento para  $\beta_0$  y  $\beta_1$ . A partir del intervalo de confianza de  $\beta_1$ , ¿qué puede decirse acerca de la hipótesis nula  $H_0: \beta_1 = 0$ ?

**Ejercicio 13.4.2** Repetir el ejercicio 13.4.1 como continuación de los ejercicios 13.2.3 y 13.3.2.

**Ejercicio 13.4.3** Repetir el ejercicio 13.4.1 como continuación de los ejercicios 13.2.4 y 13.3.3.

**Ejercicio 13.4.4** Repetir el ejercicio 13.4.1 como continuación de los ejercicios 13.2.5 y 13.3.4.

**Ejercicio 13.4.5** Repetir el ejercicio 13.4.1 como continuación de los ejercicios 13.2.6 y 13.3.5.

**Ejercicio 13.4.6** Repetir el ejercicio 13.4.1 como continuación de los ejercicios 13.2.7 y 13.3.6.

**Ejercicio 13.4.7** Repetir el ejercicio 13.4.1 como continuación de los ejercicios 13.2.8 y 13.3.7.

### 13.5 Estimación y predicción

Para estimar la media poblacional correspondiente a  $X_0$ , digamos,  $\hat{\mu}_Y$ ,  $\hat{\mu}_{Y \cdot X}$ ,  $\hat{Y}$ , o  $\hat{Y}_X$ , hacemos  $X'_0 = (1 \ X_0)$  y calculamos  $\hat{\mu}_Y = X'_0 \hat{\beta}$ , dada antes como la ec. (13.3). La varianza de esta estimación está dada por

$$\begin{aligned} V(\hat{\mu}_{Y \cdot X}) &= X'_0 V(\hat{\beta}) X_0 \\ &= X'_0 (X'X)^{-1} X_0 \sigma_{Y \cdot X}^2 \end{aligned} \quad (13.13)$$

En la práctica,  $\sigma_{Y \cdot X}^2$  se reemplaza por  $s_{Y \cdot X}^2$  para dar una estimación de la varianza.

Cuando  $X_0 = 5.5 \text{ lb}$ ,

$$\hat{\mu}_Y = (1 \quad 5.5) \begin{pmatrix} 55.26 \\ 7.69 \end{pmatrix} = 97.56 \text{ lb}$$

tal como se obtuvo en la sec. 10.6. La varianza de esta estimación es

$$s_y^2 = (1 - 5.5) \begin{pmatrix} 90.7664 & -18.1140 \\ -18.1140 & 3.6373 \end{pmatrix} \begin{pmatrix} 1 \\ 5.5 \end{pmatrix}$$

$$= 1.5407$$

y  $s_y = 1.2418$  lb.

El intervalo de confianza estimado del 95 por ciento de  $\mu_Y$  en  $X = 5.5$  se calcula como

$$\text{IC}(\mu_Y | X = 5.5) = \hat{\mu}_Y \pm t_{0.025}(8 \text{ gl}) s_y$$

$$= 97.56 \pm 2.306(1.2418)$$

$$= (94.70, 100.42)$$

Cualquier diferencia entre este resultado y el correspondiente obtenido en la sec. 10.6 se atribuirá a errores de redondeo.

Cuando un problema busca predicción en vez de estimación, es decir, cuando tratamos de decir algo acerca de una observación futura, el valor predicho debe ser el mismo que el estimado para la media, pero su varianza será mayor. En principio, esta varianza debe ser la de las observaciones y a ella debemos agregar la varianza de la estimación, que también es el valor predicho. Esta varianza está dada por

$$V[Y(\text{predicho para } X = X_0)] = [1 + X_0'(X'X)^{-1}X_0] \sigma_{Y \cdot X}^2 \quad (13.14)$$

Por otra parte,  $s_{Y \cdot X}^2$  remplaza a  $\sigma_{Y \cdot X}^2$  para dar una estimación de esta varianza.

**Ejercicio 13.5.1** Usando la notación de las ecs. (13.3) y (13.13), estimar la media poblacional de los  $Y$  para los cuales  $X = 5.00$  para el ejercicio 10.2.1. Estimar la varianza. Obtener un intervalo de confianza del 95 por ciento para la media poblacional. Remítase al ejercicio 13.4.2.

**Ejercicio 13.5.2** Repetir el ejercicio 13.5.1 refiriéndose a los ejercicios 10.6.2 y 13.4.3.

**Ejercicio 13.5.3** Repetir el ejercicio 13.5.1 refiriéndose a los ejercicios 10.6.3 y 13.4.4.

**Ejercicio 13.5.4** Repetir el ejercicio 13.5.1 refiriéndose a los ejercicios 10.6.4 y 13.4.5.

**Ejercicio 13.5.5** Usando la notación de las ecs. (13.3) y (13.13) y los datos del tratamiento 3, del ejercicio 10.8.1, estimar la media y la varianza de la población de los  $Y$  en  $X = 235$ . Obtener un intervalo de confianza del 95 por ciento para la media poblacional. Refiérase al ejercicio 13.4.6.

Repetir el ejercicio con los datos del tratamiento 4.

**Ejercicio 13.5.6** Repetir el ejercicio 13.5.5 con los datos de los 10 corredores, con 10 años en el programa, del ejercicio 10.8.3; usar  $X = 80$ . Refiérase al ejercicio 13.4.7.

### 13.6 Variables indicadoras o binarias

(Esta sección puede omitirse sin perder continuidad. Introduce las variables binarias, las cuales amplían el uso de las técnicas de regresión para el análisis de la varianza. Se trata de una matriz singular. La sección puede cubrirse en forma igualmente apropiada al final del cap. 14.)

Las variables de regresión de este capítulo han sido ante todo cuantitativas, aunque se introdujo por comodidad o consistencia de notación una variable ficticia, que toma el valor de 1. Sin embargo, muchas variables de interés son categóricas, por ejemplo, sexo y tratamiento-sexo con dos categorías y tratamiento con  $t \geq 2$ . Introduciendo *variables indicadoras o binarias*, que toman solamente los valores 0 y 1, podemos ampliar las técnicas de regresión matricial para manipular variables categóricas y combinaciones de ambas.

Considérense dos muestras independientes con varianza homogénea y una hipótesis nula de una media común. Por ejemplo, ver los datos de la tabla 5.2. La ecuación usual para el modelo con la  $H_1$  es

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad i = 1, 2, \quad j = 1, \dots, n_i \quad (13.15)$$

Escribamos de nuevo las observaciones cambiando los subíndices así:  $Y_1, \dots, Y_{n_1}$  para la primera muestra y  $Y_{n_1+1}, \dots, Y_{n_1+n_2}$  para la segunda. Luego consideremos

$$Y_i = \beta_0 X_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad i = 1, \dots, n_1 + n_2 \quad (13.16)$$

donde

$$X_i = 1 \text{ para todo } i \quad X_{1i} = \begin{cases} 1 & i = 1, \dots, n_1 \\ 0 & i = n_1 + 1, \dots, n_1 + n_2 \end{cases} \quad X_{2i} = \begin{cases} 0 & i = 1, \dots, n_1 \\ 1 & i = n_1 + 1, \dots, n_1 + n_2 \end{cases}$$

La variable  $X_{1i}$  indica cuando estamos tratando con la muestra 1,  $X_{2i}$  cuando tenemos la muestra 2. Se ve que la ec. (13.16) proporciona la misma información que la ec. (13.15). En particular, obsérvense las siguientes esperanzas.

$$E(Y_i | X_1 = 1, X_2 = 0) = \beta_0 + \beta_1 \quad E(Y_i | X_1 = 0, X_2 = 1) = \beta_0 + \beta_2$$

Hemos cambiado  $\beta_0$  por  $\mu$ ,  $\beta_1$  por  $\tau_1$ , y  $\beta_2$  por  $\tau_2$ . Realmente nada ha cambiado.

Los datos descritos por la ec. (13.16) se presentan en notación matricial a continuación. Comparar con la ec. (13.2).

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_{n_1} \\ Y_{n_1+1} \\ \vdots \\ Y_{n_1+n_2} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{n_1} \\ \varepsilon_{n_1+1} \\ \vdots \\ \varepsilon_{n_1+n_2} \end{pmatrix}.$$

Las ecuaciones normales  $\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{Y}$  son

$$\begin{pmatrix} n_1 + n_2 & n_1 & n_2 \\ n_1 & n_1 & 0 \\ n_2 & 0 & n_2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \sum_1^{n_1+n_2} Y_i \\ \sum_1^{n_1} Y_i \\ \sum_{n_1+1}^{n_1+n_2} Y_i \end{pmatrix}$$

Si sumamos las filas 2 y 3 en esta matriz  $\mathbf{X}'\mathbf{X}$ , obtenemos la fila 1 y nos damos cuenta que  $\mathbf{X}'\mathbf{X}$  es singular y por tanto no tiene inversa. También, si sumamos los elementos 2 y 3 en  $\mathbf{X}'\mathbf{Y}$ , obtenemos el primero. Esto confirma que la primera ecuación es la suma de las otras dos. A su vez, podemos concluir que las tres ecuaciones con tres incógnitas no dan realmente tres partes de información diferente. No podemos resolver en  $\hat{\beta}$  todas las incógnitas.

Eliminemos  $X_2$  y, por lo tanto  $\beta_2$  y la tercera de las tres ecuaciones normales originales. Recuérdese que la primera ecuación incluye esta información. Ahora tenemos  $Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \varepsilon_i$ , o

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_{n_1} \\ Y_{n_1+1} \\ \vdots \\ Y_{n_1+n_2} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{n_1} \\ \varepsilon_{n_1+1} \\ \vdots \\ \varepsilon_{n_1+n_2} \end{pmatrix}$$

Aquí  $E(Y_i | X_1 = 1) = \beta_0 + \beta_1$  y  $E(Y_i | X_1 = 0) = \beta_0$ . Estas esperanzas no son las mismas que se deducen de la ec. (13.16). Originalmente definimos elementos de un vector  $\beta$  de tres parámetros, pero no tuvimos suficiente información para estimarlos. El sistema de ecuaciones tenía exceso de parámetros. Ahora tenemos un vector  $\beta$  de dos parámetros y es de esperar que podamos resolver las ecuaciones. Sin embargo, la solución para los  $\beta$  está definida diferentemente que las del vector  $\beta$  anterior; esto es claro por el cambio de valores esperados.

Las ecuaciones normales en forma matricial son ahora

$$\begin{pmatrix} n_1 + n_2 & n_1 \\ n_1 & n_1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \sum_1^{n_1+n_2} Y_i \\ \sum_1^{n_1} Y_i \end{pmatrix} \quad (13.17)$$

Esta matriz  $\mathbf{X}'\mathbf{X}$  es no singular y podemos despejar las ecuaciones.

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{n_2} & -\frac{1}{n_2} \\ -\frac{1}{n_2} & \frac{n_1+n_2}{n_1 n_2} \end{pmatrix}$$

La solución de la ec. (13.17) está dada por  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_0$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \frac{1}{n_2} & -\frac{1}{n_2} \\ -\frac{1}{n_2} & \frac{n_1 + n_2}{n_1 n_2} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n_1+n_2} Y_i \\ \sum_{i=1}^{n_1} Y_i \end{pmatrix} \quad (13.18)$$

$$= \begin{pmatrix} \sum_{i=1}^{n_1+n_2} Y_i / n_2 \\ \sum_{i=1}^{n_1} Y_i - \sum_{i=1}^{n_1+n_2} Y_i \\ \frac{1}{n_1} - \frac{n_1 + n_2}{n_2} \end{pmatrix} \quad o \quad \begin{pmatrix} \bar{Y}_{2.} \\ \bar{Y}_{1.} - \bar{Y}_{2.} \end{pmatrix} \quad \text{en la notación anterior}$$

Ahora es claro que  $\hat{\beta}_0 = \bar{Y}_{2.}$  y  $\hat{\beta}_1 = \bar{Y}_{1.} - \bar{Y}_{2.}$  Además, es  $\hat{\beta}_1$  lo que se requiere para probar la hipótesis nula de que no hay diferencia entre las medias de población de la que se extrajeron las muestras.

Para los datos de la tabla 5.2,

$$\mathbf{Y} = \begin{pmatrix} 57.8 \\ \vdots \\ 53.2 \\ 64.2 \\ \vdots \\ 59.2 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix} \quad \text{después de descartar } X_2$$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 13 & 7 \\ 7 & 7 \end{pmatrix} \quad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} 761.0 \\ 393.5 \end{pmatrix} \quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{7+6}{7(6)} \end{pmatrix}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \frac{1}{6}(761.0) - \frac{1}{6}(393.5) \\ -\frac{1}{6}(761.0) + \frac{13}{7(6)}(393.5) \end{pmatrix} = \begin{pmatrix} 61.25 \\ -5.04 \end{pmatrix}$$

Para pruebas e intervalos de confianza, se necesitan las varianzas muestrales. Empecemos con  $SC(\text{total no ajustada}) = \mathbf{Y}'\mathbf{Y}$ , y la  $SC(\text{modelo}) = \hat{\beta}'\mathbf{X}'\mathbf{Y}$ .

$$\mathbf{Y}'\mathbf{Y} = (57.8)^2 + \cdots + (59.2)^2 = 44,710.28$$

$$\hat{\beta}'\mathbf{X}'\mathbf{Y} = (61.25 \quad -5.04) \begin{pmatrix} 761.0 \\ 393.5 \end{pmatrix} = 44,629.70$$

Además

$$SC(\text{residuales}) = \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y} = 80.58 \quad \text{con } 11 \text{ gl}$$

$$s^2 = 7.3258 \quad \text{y} \quad s = 2.71 \text{ por ciento}$$

Ahora calculemos  $\hat{V}(\hat{\beta})$ .

$$\begin{aligned} \hat{V}(\hat{\beta}) &= (\mathbf{X}'\mathbf{X})^{-1}s^2 \\ &= \begin{pmatrix} \frac{1}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{13}{42} \end{pmatrix} 7.3258 \\ &= \begin{pmatrix} 1.2210 & -1.2210 \\ -1.2210 & 2.2675 \end{pmatrix} \end{aligned}$$

Se ve que la varianza de  $\hat{\beta}_1$  es 2.2675 y  $s_{\hat{\beta}_1} = 1.5058$  por ciento. Finalmente, para probar  $H_0: \beta_1 = 0$  frente a  $H_1: \beta_1 \neq 0$ , una prueba de si hay o no diferencias reales en los coeficientes de digestibilidad, tenemos

$$t = \frac{-5.04}{1.51} = 3.34^{**} \quad \text{con } 11 \text{ gl}$$

Puesto que  $t_{0.005}(11 \text{ gl}) = 3.106$ , concluimos que existe una diferencia real en las medias poblacionales. En la tabla 5.2 se obtuvo y se presentó el mismo resultado. Finalmente, se construye un intervalo de confianza para la diferencia poblacional.

El investigador también podría desear intervalos de confianza para las dos medias poblacionales. Obsérvese que  $\bar{Y}_{2.} = \hat{\beta}_0$  de la ec. (13.18). También  $\hat{V}(\hat{\beta}_0 = \bar{Y}_{2.}) = 7.3258/6 = 1.2210$  por  $\hat{V}(\hat{\beta})$ . Ahora puede calcularse fácilmente un intervalo de confianza para el promedio de la segunda población.

Por la ecuación (13.18), obsérvese que  $\bar{Y}_{1.} = \beta_0 + \beta_1 = (1 \ 1)\hat{\beta}$ . A su vez,  $V(\bar{Y}_{1.}) = V(\hat{\beta}_0 + \hat{\beta}_1) = V[(1 \ 1)\hat{\beta}]$  está dado por la ec. (13.19), una aplicación de la ec. (13.13).

$$\hat{V}(\hat{\beta}_0 + \hat{\beta}_1) = (1 \ 1)\hat{V}(\hat{\beta}) \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\begin{aligned}
 &= (1 \quad 1) \begin{pmatrix} \frac{1}{n_2} & -\frac{1}{n_2} \\ -\frac{1}{n_2} & \frac{n_1 + n_2}{n_1 n_2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} s^2 \\
 &\qquad\qquad\qquad = \frac{s^2}{n_1}
 \end{aligned} \tag{13.19}$$

Para los datos,  $\bar{Y}_{1 \cdot} = \hat{\beta}_0 + \hat{\beta}_1 = 56.21$  por ciento, y

$$\begin{aligned}
 V(\bar{Y}_{1 \cdot}) &= (1 \quad 1) \begin{pmatrix} 1.2210 & -1.2210 \\ -1.2210 & 2.2675 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
 &= 1.0465
 \end{aligned}$$

Comárese esto con  $s^2/7 = 1.0465$ . Simplemente se han reordenado los cálculos. Ahora puede construirse un intervalo de confianza para esta media poblacional.

Es bien fácil ver cómo el uso de variables indicadoras puede extenderse para cubrir  $t$  tratamientos. Por lo demás, si queremos evitar una matriz singular, es necesario eliminar una variable. Después de esto es necesario considerar los valores esperados para conocer la información disponible en el vector  $\hat{\beta}$ .

**Ejercicio 13.6.1** Usando variables indicadoras, hacer el ejercicio 5.5.1. Continuar con los métodos matriciales de este capítulo para encontrar las varianzas de las medias en ambas muestras.

**Ejercicio 13.6.2** Repetir el ejercicio anterior, pero con los datos del ejercicio 5.5.2.

CAPITULO  
**CATORCE**

**REGRESION Y CORRELACION  
MULTIPLE Y PARCIAL**

#### 14.1 Introducción

En los capítulos 10 y 13 se trató de la regresión de línea recta, donde los valores de la variable dependiente deben obtenerse al azar. El cap. 11 se refirió a la medida de la intensidad de asociación lineal o *correlación simple* o *total* entre dos variables cuando los pares de observaciones debe obtenerse al azar. La regresión y correlación múltiples, que se exponen en este capítulo, consideran relaciones lineales entre más de dos valores.

Las correlaciones simples pueden no ser lo que se desea en situaciones en las que la variable dependiente está afectada por varias variables independientes. Consideremos los datos reportados por Drapala (14.6), donde las correlaciones siguientes son para 152 plantas  $F_2$  de híbridos de pasto de Sudan y sorgo de Sudan.

Producción de forraje verde y desechos de forraje  $r_{PD} = +0.554$

Producción de forraje verde y altura  $r_{PA} = +0.636$

Desechos de forraje y altura  $r_{DA} = +0.786$

Las correlaciones simples se calculan dejando de lado los valores de las demás variables. En consecuencia, no podemos concluir de  $r_{PD}$ , por ejemplo, que la relación entre forraje verde y desechos será la misma si se considera un conjunto diferente de valores de altura. En efecto, se puede mostrar que la correlación entre  $P$  y  $D$  a un valor único de altura, llamada *correlación parcial*, es +0.113 y no es significante. Por lo tanto, cuando se seleccionan todas las plantas para que tengan cierta altura específica, entonces no hay relación clara entre producción y desechos. La correlación simple entre desecho y forraje verde parece ser, en gran parte, un reflejo de la estrecha relación entre desecho y altura, la cual a su vez está estrechamente relacionada con la producción. Es decir, la mayoría de

las causas que producen correlación entre altura y desechos parecen ser las mismas que producen correlación entre altura y producción. La correlación y regresión lineal parciales se exponen dentro del marco de la correlación y regresión lineal múltiples.

Cuando el interés está en primer lugar en la estimación o predicción de valores de una característica a partir del conocimiento de otras características, necesitamos una sola ecuación que relacione la variable dependiente con la independiente. Dicha ecuación sólo tiene que ser útil para predicción o estimación; no tiene que describir una relación física complicada entre las variaciones de la variable independiente y las respuestas en la dependiente. Las técnicas de *regresión múltiple* suministran la ecuación necesaria; la *correlación múltiple* mide el grado de relación entre la variable dependiente y el conjunto de variables independientes. Para los datos en referencia, la correlación múltiple de altura y desecho con producción es 0.642.

#### 14.2 La ecuación lineal y su interpretación en más de dos dimensiones

Ciertos enunciados de la introducción serán más claros si nos referimos a la fig. 14.1, la cual muestra un plano en el espacio tridimensional.

La ecuación (14.1) es la de un plano en tres dimensiones.

$$Y = b_0 + b_1 X_1 + b_2 X_2 \quad (14.1)$$

A continuación se dan las coordenadas de los puntos representados en la fig. 14.1.

Punto No.	$X_1$	$X_2$	$Y$
1	10	2	11
2	10	4	7
3	10	6	3
4	20	2	14
5	20	4	10
6	20	6	6
7	30	2	17
8	30	4	13
9	30	6	9

Para construir un modelo físico de tal superficie, se trazan dos triángulos rectángulos, idénticos paralelos al plano  $X_1, Y$ , y con los ángulos rectos del mismo valor  $X_1$ , en este caso  $X_1 = 35$ ; obsérvese uno de ellos al frente de la figura con un ángulo en  $X_2 = 6$  y base perpendicular al eje  $X_2$ , y el otro en el plano  $X_1, Y$  con un ángulo en  $X_2 = 0$ . (Existe un tercer triángulo idéntico usado más tarde, en el plano  $X_1, Y$  y con ángulo en  $Y = 12$ ). Ahora trácese otros dos triángulos rectángulos idénticos, no necesariamente los mismos del primer par, y colóquenseles paralelos al plano  $X_2, Y$ , descansando sobre los dos triángulos anteriores y con sus ángulos rectos en el mismo valor  $X_2$ ; aquí  $X_2 = 0$ ; obsérvese uno sobre el plano  $X_2, Y$ , y el otro descansando sobre el rectángulo paralelo al plano  $X_2, Y$  en

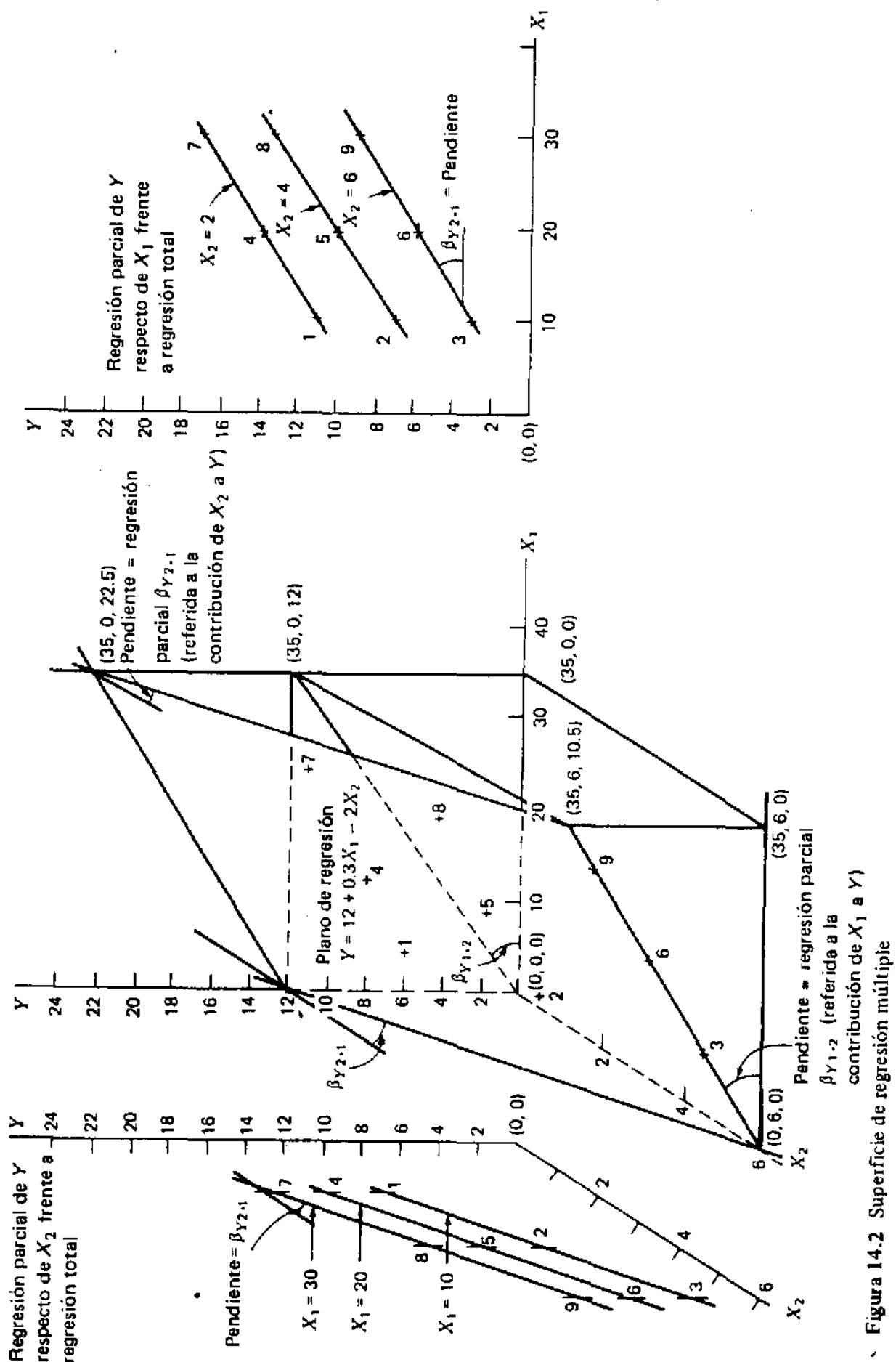


Figura 14.2 Superficie de regresión múltiple

$X_1 = 35$ . El plano que descansa sobre los dos últimos triángulos es el plano deseado. Se inclina hacia el plano  $X_1, X_2$  porque  $b_2$  es negativo. En la fig. 14.1, la ecuación del plano es  $Y = 12 + 0.3X_1 - 2X_2$ . En general, sería de esperar tener que levantar o bajar este plano para tener el intercepto apropiado, pero esto no es necesario en nuestra ilustración.

Por el método de construcción de este plano, es claro que todas las rectas del mismo paralelas al plano  $X_1, Y$  tienen igual pendiente, la del primer par de triángulos. Análogamente, todas las rectas del plano construido paralelo al plano  $X_2, Y$  tienen la misma pendiente, la del segundo par de triángulos, y difieren, probablemente, de la del otro par. Estas propiedades están asociadas con los coeficientes de regresión parcial.

Un punto del plano tiene tres componentes. Ilustramos para el punto (35, 6, 10.5).

1. Un valor base igual al intercepto de la ec. (14.1). En la ilustración,  $\alpha = 12$ . Este es el valor de  $Y$  en el plano de regresión y sobre  $(X_1, X_2) = (0, 0)$ .
2. Una contribución debida a  $X_1$  solamente. Para verlo, manténgase fijo  $X_2$  en  $X_2 = 0$  y muévase  $X_1$  a  $X_1 = 35$ . A medida que  $X_1$  se incrementa,  $Y$  sube una cantidad  $b_1 X_1 = 0.3(35) = 10.5$  hasta  $Y = 12 + 10.5 = 22.5$ . La pendiente  $b_1$  o  $b_{Y1 \cdot 2}$  es un coeficiente de regresión parcial y  $b_{Y1 \cdot 2}$  veces la variación de  $X_1$  mide el incremento en  $Y$  para esa variación en un valor fijo de  $X_2$ . Es decir, que la variación es independiente de  $X_2$ .
3. Una contribución debida únicamente a  $X_2$ . Finalmente, mantengamos  $X_1$  fijo en  $X_1 = 35$ , mientras que  $X_2$  se mueve hasta  $X_2 = 6$ . Ahora  $Y$  se mueve una cantidad  $b_2 X_2 = -2(6) = -12$  para  $Y = 22.5 - 12 = 10.5$ . De nuevo,  $b_2$  o  $b_{Y2 \cdot 1}$  veces la variación de  $X_2$  mide el incremento en  $Y$  para esa variación en un valor fijo de  $X_1$ . Aquí  $b_{Y2 \cdot 1}$  es negativo y  $X_2$  cambia de 0 a 6, de modo que la variación resultante es una disminución de  $Y$ . Esta variación es independiente del valor de  $X_1$ .

### 14.3 Regresión lineal parcial, total y múltiple

Toda ecuación de regresión muestral determina o proporciona una estimación de una media poblacional. Una ecuación de regresión lineal múltiple tiene más de una variable independiente. Para una población, los parámetros tienen que ver entre sí y una ecuación de regresión puede escribirse en la siguiente forma:

$$E(Y|X_1, \dots, X_k) = \mu_{Y \cdot x_1 \dots x_k} = \beta_0 X_0 + \beta_1 X_1 + \dots + \beta_k X_k \quad (14.2)$$

$E(Y|X_1, \dots, X_k)$  es el valor esperado o media de la población de los  $Y$  para un conjunto específico de valores de los  $X_i$ , ( $X_1, \dots, X_k$ ) por ejemplo. Siempre  $X_0 = 1$  y  $\beta_0$  representa el intercepto con  $Y$  o la media poblacional de los  $Y$  cuando  $(X_1, \dots, X_k) = (0, \dots, 0)$ . Para más claridad  $\beta_i$  se escribe como  $\beta_{YX_i \cdot x_1 \dots x_{i-1}, x_{i+1} \dots x_k}$  y puede leerse como la *regresión de Y respecto de  $X_i$  para valores fijos de los otros  $X$*  o como la *regresión parcial de Y respecto de  $X_i$* . Cuando una ecuación semejante resulta de una muestra, escribimos

$$\hat{Y} = \hat{\mu}_{Y \cdot x_1 \dots x_k} = b_0 + b_1 X_1 + \dots + b_k X_k \quad (14.3)$$

donde  $b_i = b_{YX_1 \dots X_{i-1} X_{i+1} \dots X_k}$ . La estimación de  $\beta_0$  es  $\bar{Y} - b_1 \bar{X}_1 - \dots - b_k \bar{X}_k$ . En consecuencia, a veces reemplazamos la ec. (14.3) por

$$\hat{Y} = \bar{Y} + b_1(X_1 - \bar{X}_1) + \dots + b_k(X_k - \bar{X}_k) \quad (14.4)$$

La ecuación (14.2) puede escribirse como sigue:

$$\begin{aligned} E(Y | X_1, \dots, X_k) &= \mu_{Y|X_1 \dots X_k} = \mu_{Y|\bar{X}_1 \dots \bar{X}_k} \\ &\quad + \beta_1(X_1 - \bar{X}_1) + \dots + \beta_k(X_k - \bar{X}_k) \end{aligned}$$

Para una muestra, raramente un punto observado cae sobre el plano de regresión, sino estará arriba o abajo de él. Entonces, a las componentes de un punto de un plano como se dan en la sec. 14.2, añádase una componente aleatoria para tener un punto muestral. O sea que la ilustración usada allí da valores de  $\mu_{Y|X_1 \dots X_k}$ , mientras que la descripción de una observación está dada por

$$Y_i = \beta_0 X_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + e_i \quad (14.5)$$

En la figura 14.1,  $b_{y1 \cdot 2}$  y  $b_{y2 \cdot 1}$  son pendientes o ángulos de los triángulos usados para construir el plano de regresión. Como pendiente que es,  $b_{y1 \cdot 2}$  mide el incremento en  $Y$  por unidad de  $X_1$  para cualquier valor de  $X_2$ . Esta contribución es independiente de  $X_2$  por la naturaleza de un plano. Esta independencia puede verse en la parte derecha de la fig. 14.1, donde los valores  $X_1$ ,  $Y$  están representados para los diferentes valores de  $X_2$ . Lo mismo sucede para  $b_{y2 \cdot 1}$  en la parte izquierda de la fig. 14.1. Los valores de  $b_{y1 \cdot 2}$  y  $b_{y2 \cdot 1}$  se llaman *coeficientes de regresión parcial*.

En la parte derecha de la fig. 14.1, es evidente la distinción entre regresión total o simple y regresión parcial. La figura se construye de modo que todos los puntos queden en el plano de regresión. Una regresión total, por ejemplo, de  $Y$  respecto de  $X_1$ , omite el  $X_2$  observado y se refiere a los nueve puntos en el plano  $X_1$ ,  $Y$ . Es claro que la regresión total de  $Y$  respecto de  $X_1$  no da razón de mucha variación en  $Y$ . También esta regresión puede cambiarse en gran parte sin afectar la regresión parcial. Por ejemplo, si se ha observado  $X_2 = 4$  para  $X_1 = 20, 30, y 40$ , y  $X_2 = 2$  para  $X_1 = 30, 40, y 50$ , y si los  $Y$  estuvieran todavía en el plano de regresión dado, la regresión total de  $Y$  respecto de  $X_1$  sería muy diferente. La parte derecha de la fig. 14.1 se puede modificar para ilustrar este punto.

Comentarios similares son válidos para correlación parcial y total. La *correlación parcial* de  $Y$  respecto de  $X_1$ , denotada por  $r_{y1 \cdot 2}$ , es una medida de la asociación de  $Y$  y  $X_1$  en  $X_2 = 2, X_2 = 4$ , y  $X_2 = 6$ . Evidentemente es una correlación perfecta. La correlación total o simple entre  $Y$  y  $X_1$  omite  $X_2$  y se ve, por la parte derecha, de la fig. 14.1, que es más bien baja. En la introducción el ejemplo fue del tipo opuesto, con una correlación total alta que oculta una correlación parcial baja entre producción y desechos para altura fija.

Un coeficiente de *correlación múltiple* mide lo estrecho de la asociación entre los valores observados de  $Y$  y una función de los valores independientes, que son los valores

de regresión o *valores ajustados*. Estos valores están en el plano de regresión muestral para los valores ( $X_1, \dots, X_k$ ) correspondientes a los observados. El coeficiente de correlación múltiple se denota  $R_{y,12\dots k}$ .

#### 14.4 La ecuación muestral de regresión lineal múltiple

Por definición, una ecuación de regresión muestral da estimaciones de medias poblacionales. En la práctica, también puede usarse para predecir sucesos.

La estimación de los parámetros en la ecuación de regresión es por mínimos cuadrados. En esencia, consideramos todos los posibles valores para cada una de las partidas en  $\beta' = (\beta_0, \beta_1, \dots, \beta_k)$  escogemos el conjunto para el cual la suma de cuadrados de los residuos es mínima. Finalmente se cumple la ec. (14.6).

$$\sum (Y - \hat{Y})^2 = \text{mínimo} \quad (14.6)$$

También es cierto que las sumas de estas desviaciones es cero, es decir,  $\sum (Y - \hat{Y}) = 0$  cuando  $\beta_0$ , un intercepto, está incluido en el modelo.

Para encontrar la estimación apropiada de  $\beta$ , consideremos las observaciones dadas en la ec. (14.5), en notación matricial.

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{01} & X_{11} & X_{21} & \cdots & X_{k1} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{0n} & X_{1n} & X_{2n} & \cdots & X_{kn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

o

$$\mathbf{Y} = \mathbf{X} \beta + \boldsymbol{\varepsilon}$$

Siguiendo el procedimiento de la sec. 13.2, vemos que las ecuaciones normales se convierten en la ecuación matricial (14.7).

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{Y} \quad (14.7)$$

Ahora,  $\mathbf{X}'\mathbf{X}$  es una matriz  $(k+1) \times (k+1)$  y puesto que siempre  $X_{0i} = 1$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum X_{1i} & \sum X_{2i} & \cdots & \sum X_{ki} \\ \sum X_{1i} & \sum X_{1i}^2 & \sum X_{1i}X_{2i} & \cdots & \sum X_{1i}X_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum X_{ki} & \sum X_{ki}X_{1i} & \sum X_{ki}X_{2i} & \cdots & \sum X_{ki}^2 \end{pmatrix}$$

y

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum X_{1i}Y_i \\ \vdots \\ \sum X_{ki}Y_i \end{pmatrix}$$

Realmente nada ha cambiado desde la sec. 13.2, excepto las dimensiones de la ecuación matricial y las dificultades aritméticas. Si  $\mathbf{X}'\mathbf{X}$  es no singular, siempre podemos escribir la solución como

$$k+1 \mathbf{b}_1 = k+1 \hat{\beta}_1 = k+1 (\mathbf{X}'\mathbf{X})_{k+1}^{-1} k+1 \mathbf{X}'_{nn} \mathbf{Y}_1 \quad (14.8)$$

Supongamos que esta solución sea dada ordinariamente por un equipo de computación.

Es fácil mostrar que  $b_0 = \bar{Y} - b_1 \bar{X}_1 - \cdots - b_k \bar{X}_k$ , sustituir esto en las últimas  $k$  ecuaciones normales y reducir éstas a la ec. (14.9) matricial.

$$\begin{pmatrix} \sum (X_{1i} - \bar{X}_{1.})^2 & \sum (X_{1i} - \bar{X}_{1.})(X_{2i} - \bar{X}_{2.}) & \cdots & \sum (X_{1i} - \bar{X}_{1.})(X_{ki} - \bar{X}_{k.}) \\ \sum (X_{2i} - \bar{X}_{2.})(X_{1i} - \bar{X}_{1.}) & \sum (X_{2i} - \bar{X}_{2.})^2 & \cdots & \sum (X_{2i} - \bar{X}_{2.})(X_{ki} - \bar{X}_{k.}) \\ \vdots & \vdots & \ddots & \vdots \\ \sum (X_{ki} - \bar{X}_{k.})(X_{1i} - \bar{X}_{1.}) & \sum (X_{ki} - \bar{X}_{k.})(X_{2i} - \bar{X}_{2.}) & \cdots & \sum (X_{ki} - \bar{X}_{k.})^2 \end{pmatrix} \times \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix} = \begin{pmatrix} \sum (X_{1i} - \bar{X}_{1.})(Y_i - \bar{Y}) \\ \sum (X_{2i} - \bar{X}_{2.})(Y_i - \bar{Y}) \\ \vdots \\ \sum (X_{ki} - \bar{X}_{k.})(Y_i - \bar{Y}) \end{pmatrix} \quad (14.9)$$

A estas ecuaciones también se les llama a veces ecuaciones normales. En la próxima sección, usaremos estas ecuaciones con  $k = 2$  para ilustrar algunos aspectos generales de la regresión múltiple que puede generalizarse fácilmente a  $k > 2$  variables independientes. Puesto que cada suma de cuadrados de productos ha sido *ajustada* para las medias, y  $b_0$  ha sido eliminado de  $\hat{\beta}$ , escribamos estas ecuaciones como la ecuación matricial (14.10)

$$\mathbf{X}'_A \mathbf{X}_A \hat{\beta}_A = \mathbf{X}'_A \mathbf{Y}_A \quad (14.10)$$

Recuérdese que  $\sum (X - \bar{X})(Y - \bar{Y}) = \sum (X - \bar{X})Y$ , de modo que  $\mathbf{X}'_A \mathbf{Y}_A = \mathbf{X}'_A \mathbf{Y}$ .

**Ejercicio 14.4.1** Calcular  $\sum (X_1 - \bar{X}_{1.})(X_2 - \bar{X}_{2.})$ ,  $\sum (X_1 - \bar{X}_{1.})(Y - \bar{Y})$ ,  $\sum (X_2 - \bar{X}_{2.}) \times (Y - \bar{Y})$  para la ilustración de la sec. 14.2. Obsérvese el valor de  $\sum (X_1 - \bar{X}_{1.})(X_2 - \bar{X}_{2.})$ . Esta es la propiedad de ortogonalidad que puede hacer sencilla la solución de las ecs. (14.9) o (14.10) y nos permitirá calcular directamente las contribuciones independientes como en el caso de comparaciones ortogonales.

## 14.5 Regresión lineal múltiple; dos variables independientes

La tabla 14.1 da porcentajes  $X_1$  de nitrógeno,  $X_2$  cloro,  $X_3$  potasio, y log de combustión foliar en segundos,  $Y$ , para 30 muestras de tabaco tomadas en campos de agricultores. De estas cuatro variables, se usarán en esta sección, las dos primeras variables independientes y la dependiente para determinar una ecuación de regresión. Los datos de todas las cuatro variables se usarán para ilustrar el procedimiento con cualquier número de variables independientes a partir de la sec. 14.7.

**Tabla 14.1 Porcentajes de nitrógeno  $X_1$ , cloro  $X_2$ , potasio  $X_3$ , y logaritmo de combustión foliar en segundos,  $Y$ , en muestras de tabaco de campos de agricultores**

Muestra No.	Nitrógeno % $X_1$	Cloro % $X_2$	Potasio % $X_3$	Log. de combustión foliar $Y$ , s
1	3.05	1.45	5.67	0.34
2	4.22	1.35	4.86	0.11
3	3.34	0.26	4.19	0.38
4	3.77	0.23	4.42	0.68
5	3.52	1.10	3.17	0.18
6	3.54	0.76	2.76	0.00
7	3.74	1.59	3.81	0.08
8	3.78	0.39	3.23	0.11
9	2.92	0.39	5.44	1.53
10	3.10	0.64	6.16	0.77
11	2.86	0.82	5.48	1.17
12	2.78	0.64	4.62	1.01
13	2.22	0.85	4.49	0.89
14	2.67	0.90	5.59	1.40
15	3.12	0.92	5.86	1.05
16	3.03	0.97	6.60	1.15
17	2.45	0.18	4.51	1.49
18	4.12	0.62	5.31	0.51
19	4.61	0.51	5.16	0.18
20	3.94	0.45	4.45	0.34
21	4.12	1.79	6.17	0.36
22	2.93	0.25	3.38	0.89
23	2.66	0.31	3.51	0.91
24	3.17	0.20	3.08	0.92
25	2.79	0.24	3.98	1.35
26	2.61	0.20	3.64	1.33
27	3.74	2.27	6.50	0.23
28	3.13	1.48	4.28	0.26
29	3.49	0.25	4.71	0.73
30	2.94	2.22	4.58	0.23
$\sum X_i$	98.36	24.23	139.61	20.58
$\bar{X}$	3.2787	0.8077	4.6537	0.6860
$\sum X_i^2$	332.3352	30.1907	682.7813	20.8074
$\sum X_i X_j$	$\sum X_1 X_2 = 81.5834$ $\sum X_2 X_3 = 120.3950$	$\sum X_1 X_3 = 459.4052$ $\sum X_2 Y = 12.4103$	$\sum X_1 Y = 61.6502$ $\sum X_3 Y = 98.4408$	

Fuente: Datos obtenidos por cortesía de O.J. Attoe, Universidad de Wisconsin, Madison, Wisconsin.

Los datos brutos se tratan en un computador programado para hacer lo siguiente:

1. Calcular sumas, sumas de cuadrados y sumas de productos cruzados para los datos brutos. Estas se dan al pie de la tabla 14.1.
2. Construir las matrices  $\mathbf{X}$ ,  $\mathbf{X}'\mathbf{X}$ ,  $\mathbf{X}'\mathbf{Y}$ , y a su vez, las ecuaciones normales con  $\hat{\beta}' = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (b_0, b_1, b_2)$ .

$$\mathbf{X} = \begin{pmatrix} 1 & 3.05 & 1.45 \\ 1 & 4.22 & 1.35 \\ 1 & 2.94 & 2.22 \end{pmatrix} \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} 30 & 98.36 & 24.23 \\ 98.36 & 332.3352 & 81.5834 \\ 24.23 & 81.5834 & 30.1907 \end{pmatrix}$$

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 20.58 \\ 61.6502 \\ 12.4103 \end{pmatrix}$$

A continuación se dan las ecuaciones normales en forma matricial.

$$\begin{pmatrix} 30 & 98.36 & 24.23 \\ 98.36 & 332.3352 & 81.5834 \\ 24.23 & 81.5834 & 30.1907 \end{pmatrix} \hat{\beta} = \begin{pmatrix} 20.58 \\ 61.6502 \\ 12.4103 \end{pmatrix}$$

3. Resolver las ecuaciones normales. Probablemente la codificación se hace, en alguna fase del cálculo, de modo que la aritmética sea lo más eficiente posible. La solución impresa consistirá en parte o en total de lo siguiente  $(\mathbf{X}'\mathbf{X})^{-1}$ ,  $\hat{\beta}$ , la ecuación de regresión

$$SC(\text{modelo}) = SC(b_0, b_1, b_2) = \hat{\beta}'\mathbf{X}'\mathbf{Y}$$

$$SC(\text{regresión}) = SC(b_1, b_2 | b_0) = SC(X_1, X_2 | X_0) = \hat{\beta}'\mathbf{X}'\mathbf{Y} - \frac{(\sum Y)^2}{n}$$

$$SC(\text{residuos}) = \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y}, \quad \hat{V}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}s^2$$

el análisis de la varianza, pruebas de  $H_0: \beta_1 = \beta_2 = 0$  simultáneamente, y de  $H_0: \beta_1 = 0$  y  $H_0: \beta_2 = 0$ , donde  $\beta_1$  y  $\beta_2$  son coeficientes de regresión parcial. También puede incluirse otra información.

Obsérvese que  $SC(\text{regresión})$  es una diferencia;  $\hat{\beta}'\mathbf{X}'\mathbf{Y}$  mide la parte de la variación total  $\mathbf{Y}'\mathbf{Y}$  que puede asociarse con el ajuste de los parámetros en el modelo lineal completo  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ , mientras que  $(\sum Y)^2/n$  mide lo mismo para el modelo reducido  $Y = \beta_0 + \varepsilon$ . Estas diferencias en sumas de cuadrados mide la reducción adicional atribuible a la adición de los términos lineales,  $\beta_1 X_1$  y  $\beta_2 X_2$  al modelo con una constante y una componente aleatoria solamente. En otras palabras, estamos calculando sumas de cuadrados aditivas u ortogonales, como se expuso para comparaciones en la sec. 8.3.

También, observese que  $\beta_0$  en el vector  $\beta$  no tiene el mismo estimador del modelo reducido. En el modelo completo,  $\hat{\beta}_0 = b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$ , mientras en el modelo reducido,  $\hat{\beta}_0 + b_0 = \bar{Y}$ . Es decir, cuando se añaden términos a un modelo, es de esperar que los estimadores de los parámetros comunes cambien.

- 3a. El punto 3 puede describirse como un enfoque moderno de la solución del problema de regresión. Sin embargo, todavía muchos textos tratan la ecuación matricial (14.10) como las ecuaciones normales; consideraremos su uso con dos variables independientes.

La ecuación del modelo es la misma, o sea,  $E(Y) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$ , pero también se escribe como  $E(Y) = \beta_0 + \beta_1(X_1 - \bar{X}_1) + \dots + \beta_k(X_k - \bar{X}_k)$ . Aquí  $\beta_1, \dots, \beta_k$  son las mismas, pero  $\beta_0$  es el intercepto  $\bar{Y}$  en el primer caso, es decir,  $\beta_0$  (primera forma) =  $E(Y|X_1 = 0, \dots, X_k = 0)$ , mientras que es la media poblacional en  $(\bar{X}_1, \dots, \bar{X}_k)$  en el último caso, o  $\beta_0$  (última forma) =  $E(Y|X_1 = \bar{X}_1, \dots, X_k = \bar{X}_k)$ .

La ecuación de regresión muestral que estima medias poblacionales en la segunda forma puede escribirse como

$$\hat{Y}_Y = \bar{Y} - \bar{Y} + b_1(X_1 - \bar{X}_1) + \dots + b_k(X_k - \bar{X}_k) \quad (14.11)$$

De antemano debemos conocer la forma general de la estimación del intercepto  $\bar{Y}$ , sin especificar las estimaciones de  $\beta_i$ ,  $i = 1, \dots, k$ . La ec. (14.10) queda todavía por resolver.

La tabla 14.2 da sumas ajustadas de cuadrados y productos, para construir  $\mathbf{X}'_A \mathbf{X}_A$  y  $\mathbf{X}'_A \mathbf{Y}_A$ , ( $\mathbf{X}'_A \mathbf{Y} = \mathbf{X}'_A \mathbf{Y}_A$ ). La ec. (14.10) para nuestros datos es

$$\begin{pmatrix} 9.845547 & 2.141307 \\ 2.141307 & 10.620937 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} -5.82476 \\ -4.21148 \end{pmatrix}$$

Estas ecuaciones pueden resolverse por sustitución o calculando  $(\mathbf{X}'_A \mathbf{X}_A)^{-1}$ .

$$(\mathbf{X}'_A \mathbf{X}_A)^{-1} = \begin{pmatrix} 0.106227 & -0.021417 \\ -0.021417 & 0.098471 \end{pmatrix}$$

**Tabla 14.2** Sumas ajustadas de cuadrados y productos y correlaciones simples para los datos de la tabla 14.1

$X_1$	$X_2$	$Y$
$X_1 = \sum (X_1 - \bar{X}_1)^2 = 9.845547$	$\sum (X_1 - \bar{X}_1)(X_2 - \bar{X}_2) = 2.141307$ $r_{12} = 0.209400$	$\sum (X_1 - \bar{X}_1)(Y - \bar{Y}) = -5.82476$ $r_{11} = -0.717729$
$X_2$	$\sum (X_2 - \bar{X}_2)^2 = 10.620937$	$\sum (X_2 - \bar{X}_2)(Y - \bar{Y}) = -4.21148$ $r_{22} = -0.499638$
$Y$		$\sum (Y - \bar{Y})^2 = 6.68952$

Luego, estimamos

$$\hat{\beta}_A = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

Esto requiere  $\mathbf{X}'_A \mathbf{Y}_A$ .

$$\hat{\beta}_A = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{Y}_A = \begin{pmatrix} -0.52855 \\ -0.28996 \end{pmatrix}$$

Si se usa la forma del intercepto de la ecuación  $b_0 = \bar{Y} = b_1 \bar{X}_1 - b_2 \bar{X}_2 = 2.65313$ . La ecuación de regresión se determina a partir de la ec. (14.11) o de la (14.3)

$$\begin{aligned} \hat{Y} &= 0.6860 - .5285(X_1 - 3.2787) - .2900(X_2 - 4.6537) \\ &= 2.6531 - .5285X_1 - .2900X_2 \\ &= (1 \quad X_1 \quad X_2) \begin{pmatrix} 2.6531 \\ -0.5285 \\ -0.2900 \end{pmatrix} \end{aligned}$$

La tabla 14.3 da el análisis de la varianza; la exposición de la partición de la suma de cuadrados de la regresión viene después de las pruebas t que se dan a continuación.

El resultado del computador puede incluir la matriz de varianza-covarianza y casi con certeza incluir pruebas de hipótesis para  $\beta_1$  y  $\beta_2$ .

$$\hat{\mathbf{V}}(\hat{\beta}_A) = (\mathbf{X}'_A \mathbf{X}_A)^{-1} s^2 = \begin{pmatrix} .009401717 & -0.001895498 \\ -0.001895498 & .008715337 \end{pmatrix}$$

Es claro que los coeficientes de regresión parcial no son independientes; sus covarianzas son los elementos fuera de la diagonal en  $\hat{\mathbf{V}}(\hat{\beta}_A)$ , cuando se han dado. A su vez, las sumas de cuadrados para las pruebas F de estos coeficientes no son aditivas.

Para probar  $H_0: \beta_1 = 0$ ,

$$t = \frac{b_1}{s_{b_1}} = \frac{-0.52855}{\sqrt{.009401717}} = -5.45^{**}$$

Análogamente, para probar  $H_0: \beta_2 = 0$ ,

$$t = \frac{b_2}{s_{b_2}} = \frac{-0.28996}{\sqrt{.008715337}} = -3.11^{**}$$

Tabla 14.3 Análisis de la varianza de regresión de  $Y$  con respecto de  $X_1$  y  $X_2$  según la tabla 14.1

Fuente de variación	gl	SC	CM	F
Regresión $  b_0 \quad 0 \quad b_1, b_2   b_0$	$k = 2$	$\hat{\beta}_A X_A Y_A = 4.29985$	2.14992	24.3**
$X_1$ omitiendo $X_2 \quad 0 \quad b_1   b_0$ solamente 1		$\frac{(X_{1,A} Y_A)^2}{X_{1,A} X_{1,A}} = 3.44601$		
$X_2   X_0, X_1 \quad 0 \quad b_2   b_0, b_1$	1	$\hat{\beta}_A X_A Y_A - SC(X_1   X_0) = .85384$	.85384	9.6**
$X_1$ omitiendo $X_2 \quad 0 \quad b_2   b_0$ solamente 1		$\frac{(X_{2,A} Y_A)^2}{X_{2,A} X_{2,A}} = 1.66996$		
$X_1   X_0, X_2 \quad 0 \quad b_1   b_0, b_2$	1	$\hat{\beta}_A X_A Y_A - SC(X_2   X_0) = 2.62989$	2.62989	29.7**
Error = residuo	$n - k - 1 = 27$	$Y_A Y_A - \hat{\beta}_A X_A Y_A = 2.38967$	.08851	
Total	$n - 1 = 29$	$Y_A Y_A = 6.68952$		

Ambas hipótesis nulas se descartan con una prueba  $t$  con una tasa de error por comparación.

También es directa la construcción del intervalo de confianza. Para  $1 - \alpha = 0.95$ ,

$$\begin{aligned} IC(\beta_1) &= b_1 \pm t_{0.025} s_{b_1} = -.52855 \pm 2.052 \sqrt{.009401717} \\ &= (-.7275, -.3296) \end{aligned}$$

$$\begin{aligned} IC(\beta_2) &= b_2 \pm t_{0.025} s_{b_2} = -.28996 \pm 2.052 \sqrt{.008715337} \\ &= (-.4815, -.0984) \end{aligned}$$

Para probar la hipótesis nula  $H_0: \beta = 0$ , una prueba simultánea, empleese la prueba  $F$  de la parte superior de la tabla 14.3;  $F = 2.14992/0.08851 = 24.3^{**}$ . Es posible la construcción de una región conjunta de confianza para  $\beta_1$  y  $\beta_2$  sea rectangular o elíptica. El coeficiente de confianza se aplicaría a la estimación conjunta; es decir en estimación simultánea repetida, la región de confianza contendrá ambos parámetros  $100(1 - \alpha)$  por ciento de las veces y no contendrá a ninguna o sólo a uno el  $100\alpha$  por ciento de las veces.

Las pruebas de hipótesis anteriores son de hipótesis que se refieren a coeficientes de regresión parcial, a pruebas acerca de la regresión de  $Y$  respecto de una variable independiente cuando las otras se mantienen constantes, o a pruebas del valor de introducir una variable independiente en un modelo donde ya se han incluido otras. Al calcular, podemos particionar  $SC(\text{regresión} | b_0) = SC(\text{regresión})$  como en la tabla 14.3. Primero, se calcula  $SC(X_1 \text{ omitiendo a } X_2)$  como en el cap. 10. Para esto, se introduce  $X'_{iA}$ . Por implicación, la ec. (14.9) define a  $X_A$  como una matriz con la  $i$ -ésima columna compuesta de  $n$  desviaciones,  $X_{ij} - \bar{X}_i$ ,  $j = 1, \dots, n$ . Sea  $X'_{iA}$  la  $i$ -ésima columna o vector de desviaciones en  $X_A$ ;  $i$  indica el valor en  $X_A$ ; y  $A$  nos dice que cada  $X$  ha sido ajustado para la media apropiada. A continuación se deduce la ec. (14.12).

$$X'_{iA} = (X_{i1} - \bar{X}_i, X_{i2} - \bar{X}_i, \dots, X_{in} - \bar{X}_i) \quad (14.12)$$

Es claro que  $X'_{iA} X_{iA} = \sum_j (X_{ij} - \bar{X}_i)^2$ ; en forma análoga  $Y'_{iA} Y_{iA} = \sum_j (Y_j - \bar{Y})^2$  y  $X'_{iA} Y_{iA} = \sum_j (X_{ij} - \bar{X}_i)(Y_j - \bar{Y})$ . A su vez, escribimos  $SC(X_1 | X_0) = SC(b_1 | b_0) = (X'_{iA} Y_{iA})^2 / X'_{iA} X_{iA}$ . Esta es la suma de cuadrados, además del factor de corrección y variación atribuible a  $X_1$ .

En el análisis de varianza,  $SC(X_1 | X_0) = SC(b_1 | b_0)$  es la resta de  $SC(\text{regresión} | b_0) = SC(b_1, b_2 | b_0)$  hasta dar  $SC(b_2 | b_0, b_1)$ ; esto es, computar  $SC(b_1, b_2 | b_0) - SC(b_1 | b_0) = SC(b_2 | b_0, b_1)$ . Esta es la suma de cuadrados, además del factor de corrección y variación atribuible a  $X_1$  que puede atribuirse a  $X_2$ . Obsérvese que  $(-3.11)^2 = t^2 = F = 9.65$  dentro de los errores de redondeo, al probar  $H_0: \beta_2 = 0$ .

El cálculo de  $SC(X_2 | X_0)$  es parecido y, a su vez,

$$SC(X_1 | X_2, X_0) = SC(b_1 | b_0, b_2) = SC(b_1, b_2 | b_0) - SC(b_2 | b_0).$$

**Ejercicio 14.5.1** Resolver las ecuaciones citadas en el ejercicio 14.4.1 y mostrar que la ecuación de regresión de la sec. 14.2 es correcta. Comparar los coeficientes de regresión parcial y total, explicar este resultado particular.

**Ejercicio 14.5.2** Birch (14.2) en un estudio de respuesta al fosfato, saturación de bases y relación de sílice en suelos ácidos en el maíz, recolectó los datos que se dan en la tabla siguiente. El porcentaje de respuesta se midió como la diferencia entre rendimiento de parcelas que reciben  $P$  y aquéllas que no lo reciben, dividido por la producción de las parcelas que no reciben  $P$ , y multiplicando por 100. En consecuencia, por el procedimiento de cálculo se ha introducido una correlación entre  $Y$  y  $X_1$ . BEC se refiere a capacidad de intercambio de bases.

$Y = \text{respuesta al fosfato, en porcentaje}$	$X_1 = \text{producción del control, lb grano/acre}$	$X_2 = \text{saturación de BEC, en porcentaje}$	$X_3 = \text{pH del suelo}$
88	844	67	5.75
80	1,678	57	6.05
42	1,573	39	5.45
37	3,025	54	5.70
37	653	46	5.55
20	1,991	62	5.00
20	2,187	69	6.40
18	1,262	74	6.10
18	4,624	69	6.05
4	5,249	76	6.15
2	4,258	80	5.55
2	2,943	79	6.40
-2	5,092	82	6.55
-7	4,496	85	6.50

Considérese como un problema de regresión con  $X_1$  y  $X_3$  como las únicas variables independientes. ¿Cuáles son  $\mathbf{Y}$ ,  $\mathbf{X}$ ,  $\boldsymbol{\beta}$ ? Calcular  $\mathbf{X}'\mathbf{X}$  y  $\mathbf{X}'\mathbf{Y}$ . Escribir las ecuaciones por mínimos cuadrados en notación matricial. Calcular  $\mathbf{X}'\mathbf{X}$  y hallar  $\hat{\boldsymbol{\beta}}$ . Escribir la ecuación de regresión.

Calcular  $\mathbf{Y}'\mathbf{Y}$ , SC(modelo), SC(regresión) y SC(residuos), y presentar los resultados numéricos en un análisis de la varianza semejante al de la tabla 14.3. Usando  $\alpha = 0.05$ , contrastar  $H_0: \beta_1 = 0 = \beta_2$ ;  $H_0: \beta_1 = 0$ ;  $H_0: \beta_2 = 0$ . ¿Qué es  $\alpha = 0.05$  en estas pruebas?

Hallar  $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$ . Calcular  $\hat{Y}$  para  $\mathbf{X}'_0 = (1, 2,000, 6.00)$ . Hallar  $\hat{\mathbf{V}}(\hat{Y})$  y construir un intervalo de confianza del 95 por ciento para la media poblacional que se ha estimado.

**Ejercicio 14.5.3** Repetir el ejercicio 14.5.2 usando los datos del ejercicio 10.2.3. Para las variables independientes, usar dos medidas de ajuste,  $X_1$  y  $X_2$ . Para  $\hat{Y}$ , usar  $\mathbf{X}'_0 = (1, 50, 100)$ . ¿Parece  $\mathbf{X}'_0$  un vector que pueda aplicarse a una persona real?

Repetir el ejercicio usando dos medidas hemáticas  $X_3$  y  $X_4$ . Para  $\hat{Y}$ , usar  $\mathbf{X}'_0 = (1, 280, 50)$ . ¿Parece  $\mathbf{X}'_0$  un vector que puede aplicarse a una persona real?

Repetir el ejercicio usando  $X_1$  y  $X_3$ . Para  $\hat{Y}$ , use,  $\mathbf{X}'_0 = (1, 50, 280)$ . ¿Parece  $\mathbf{X}'_0$  un vector que puede aplicarse a una persona real?

**Ejercicio 14.5.4** Repetir el ejercicio 14.5.2 usando los datos del ejercicio 10.2.4. Para  $\hat{Y}$ , usar  $\mathbf{X}'_0 = (1, 2,400, 150)$ .

**Ejercicio 14.5.5** Repetir el ejercicio 14.5.2 usando los datos del ejercicio 10.2.5. Como variables independientes, tomar  $X_1$  y  $X_2$ . Para  $\hat{Y}$ , usar  $\mathbf{X}'_0 = (1, 1,200, 550)$ .

## 14.6 Correlación parcial y múltiple

Los coeficientes de correlación múltiple y parcial son estrictamente aplicables sólo cuando la observación total, esto es,  $(Y_i, X_{1i}, \dots, X_{ki})$ , es aleatoria. Sin embargo, con indepen-

dencia de aleatoriedad de las observaciones, estos coeficientes de correlación pueden ser útiles, para cálculos y por otras razones.

Las correlaciones parciales se definen como correlaciones entre dos variables cuando las demás están fijas. El símbolo  $r_{Y \cdot X_1 \cdot X_2}$  se usa para la correlación muestral entre  $Y$  y  $X_1$ , cuando  $X_2$  y  $X_3$  son constantes o "ajustadas". Puesto que la observación total deberá ser aleatoria y el coeficiente es fundamentalmente descriptivo, no hay necesidad de referirse a una variable dependiente en particular o de denotarla  $Y$ .

Para calcular los coeficientes de correlación parcial, definase primero  $\mathbf{R}$  como la matriz simétrica de correlaciones simples entre un conjunto de  $k$  variables  $X_1, \dots, X_k$ , ninguna de ellas señalada como independiente. En general,  $\mathbf{R}$  está dada por

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ r_{21} & 1 & r_{23} & \cdots & r_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{k1} & r_{k2} & r_{k3} & \cdots & 1 \end{pmatrix} \quad (14.13)$$

La inversa de  $\mathbf{R}$  está determinada por la ec. (12.15) y también es simétrica; escribámosla como

$$\mathbf{R}^{-1} = \mathbf{C} = \begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1k} \\ C_{21} & C_{22} & \cdots & C_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ C_{k1} & C_{k2} & \cdots & C_{kk} \end{pmatrix} \quad (14.14)$$

Ahora la ec. (14.15) define la correlación parcial entre  $X_i$  y  $X_j$ .

$$r_{ij \cdot 1 \cdots i-1, i+1 \cdots j-1, j+1 \cdots k} = \frac{-C_{ij}}{\sqrt{C_{ii}C_{jj}}} \quad (14.15)$$

En la ecuación (14.15), los elementos de la matriz inversa pueden reemplazarse por los menores con signo o cofactores de  $\mathbf{R}$ . Ver las ecs. (12.13) y (12.15). Los elementos de la inversa son simplemente los cofactores divididos por el determinante de la matriz.

La matriz de correlación para  $X_1$ ,  $X_2$ , y  $Y$  se obtiene a partir de la tabla 14.2;  $X_1$ ,  $X_2$  y  $Y$  se dan para identificación.

$$\mathbf{R} = \begin{matrix} & X_1 & X_2 & Y \\ X_1 & \begin{pmatrix} 1.000000 & 0.209400 & -0.717729 \\ 0.209400 & 1.000000 & -0.499638 \\ -0.717729 & -0.499638 & 1.000000 \end{pmatrix} \\ X_2 \\ Y \end{matrix}$$

La matriz de cofactores se calcula fácilmente.

$$\begin{pmatrix} 1 - r_{Y2}^2 & -r_{12} + r_{1Y}r_{Y2} & -r_{1Y} + r_{12}r_{2Y} \\ -r_{21} + r_{Y1}r_{2Y} & 1 - r_{Y1}^2 & -r_{2Y} + r_{21}r_{1Y} \\ -r_{Y1} + r_{21}r_{Y2} & -r_{Y2} + r_{12}r_{Y1} & 1 - r_{12}^2 \end{pmatrix} = \begin{pmatrix} .750361 & .149205 & .613105 \\ .149205 & .484865 & .349346 \\ .613105 & .349346 & .956151 \end{pmatrix}$$

Ahora calculamos

$$r_{Y1 \cdot 2} = \frac{- .613105}{\sqrt{.750361(.956151)}} = -.723829$$

$$r_{Y2 \cdot 1} = \frac{- .349346}{\sqrt{.484865(.956151)}} = -.513076$$

Una vez más se han utilizado todas las cifras decimales que proporcionó nuestra calculadora. En general, para probar  $H_0: \rho_{ij \cdot 1 \dots i-1, i+1 \dots j-1, j+1 \dots k} = 0$  calcúlese  $t$  por

$$t = \frac{r \sqrt{n-k}}{\sqrt{1-r^2}} \quad (14.16)$$

Con los datos, para probar  $H_0: \rho_{Y2 \cdot 1} = 0$  frente a  $H_1: \rho_{Y2 \cdot 1} \neq 0$ , calcular

$$\begin{aligned} t &= \frac{r_{Y2 \cdot 1} \sqrt{n-3}}{\sqrt{1-r_{Y2 \cdot 1}^2}} \\ &= \frac{-.513076 \sqrt{27}}{\sqrt{1-(-.513076)^2}} = -3.11^{**} \quad \text{con } 27 \text{ gl} \end{aligned}$$

La prueba también puede hacerse frente a alternativas unilaterales. La prueba  $t$  frente a alternativas bilaterales es equivalente a la prueba  $F$  de la misma hipótesis nula;  $t^2 = (-3.11)^2 = 9.65$  se compara con  $F = 9.6$  en el análisis de la varianza.

También pueden construirse intervalos de confianza.

El coeficiente de correlación múltiple, denotado por  $R_{Y \cdot 1 \dots k}$ , mide la exactitud con la cual el plano de regresión se ajusta a los puntos observados. Es decir, es la correlación entre los  $Y$  observados y los  $\hat{Y}$  de regresión, es decir, los  $\hat{Y}$ . Así pues,  $R_{Y \cdot 1 \dots k}$  mide el efecto combinado de todas las variables independientes sobre la variable dependiente  $Y$ . Este coeficiente está definido por

$$R_{Y \cdot 1 \dots k} = \sqrt{\frac{\text{SC(regresión)}}{\text{SC(total, ajustado)}}}$$

$$= \sqrt{\frac{\hat{\beta}' X' Y - (\sum Y)^2/n}{Y' Y - (\sum Y)^2/n}} = \sqrt{\frac{\hat{\beta}' X_A' Y_A}{Y_A' Y_A}} \quad (14.17)$$

El cuadrado de este coeficiente también se llama *coeficiente de determinación (multiple)*.

La ecuación (14.17) puede transformarse para obtener las ecs. (14.18) y (14.19), que son de gran utilidad.

$$SC(\text{regresión}) = \hat{\beta}' X_A' Y_A = R_{Y \cdot 1 \dots k}^2 Y_A' Y_A \quad k \text{ gl} \quad (14.18)$$

$$\begin{aligned} SC(\text{error}) &= Y' Y - \hat{\beta}' X' Y = Y_A' Y_A - \hat{\beta}' X_A' Y_A \\ &= (1 - R_{Y \cdot 1 \dots k}^2) Y_A' Y_A \quad n - k - 1 \text{ gl} \end{aligned} \quad (14.19)$$

La ecuación (14.20) implica un procedimiento secuencial de cálculo.

$$1 - R_{Y \cdot 1 \dots k}^2 = (1 - r_{Y1}^2)(1 - r_{Y2 \cdot 1}^2) \cdots (1 - r_{Yk \cdot 1 \dots k-1}^2) \quad (14.20)$$

Para los datos de tabaco con  $X_1$  y  $X_2$  solamente,

$$R_{Y \cdot 12} = \sqrt{\frac{4.29985}{6.68952}} = .8017 \quad \text{de la ecuación (14.7)}$$

Por la ecuación (14.20),

$$\begin{aligned} 1 - R_{Y \cdot 12}^2 &= [1 - (-.7177)^2][1 - (-.5131)^2] \\ &= .3572 \quad \text{y} \quad R_{Y \cdot 12} = .8017 \end{aligned}$$

Los valores significantes de  $R$  se dan en la tabla A13.

**Ejercicio 14.6.1** Calcular  $r_{Y1 \cdot 2}$ ,  $r_{Y2 \cdot 1}$ ,  $r_{12 \cdot Y}$ , y  $R_{Y \cdot 12}$  para los datos del ejercicio 14.5.2.

**Ejercicio 14.6.2** Construir los intervalos de confianza del 95 por ciento para los parámetros  $\rho_{Y1 \cdot 2}$  y  $\rho_{Y2 \cdot 1}$ , estimados en el ejercicio 14.6.1. ¿Incluyen el cero estos intervalos? A partir de esto, ¿qué conclusión puede sacarse?

**Ejercicio 14.6.3** Probar la hipótesis nula de que el parámetro poblacional estimado por  $R_{Y \cdot 12}$  en el ejercicio 14.6.1 es cero. (Ver tabla A13).

**Ejercicio 14.6.4** Demostrar que la prueba  $F$  de  $H_0: \beta_1 = \beta_2 = 0$  en el análisis de la varianza puede reducirse a

$$F = \frac{R_{Y \cdot 12}^2}{1 - R_{Y \cdot 12}^2} \frac{n - k - 1}{k}$$

**Ejercicio 14.6.5** Los ejercicios 14.6.1 y 14.6.4 pueden repetirse para los ejercicios 14.5.3 hasta 14.5.5.

#### 14.7 Regresión lineal múltiple; resultados impresos para $k$ variables independientes

En la regresión lineal múltiple, el primer paso en la obtención de estimaciones del vector  $\beta$  es establecer las ecuaciones normales  $X'X\hat{\beta} = X'Y$ . Estas pueden resolverse por la técnica directa de eliminación gaussiana, multiplicando una ecuación por una constante y sumando este múltiplo a otra ecuación con el fin de reducir el número de variables. Doolittle hizo un primer intento de estructurar esta técnica para usarla en estadística. El método básico de Doolittle fue presentado en 1878 cuando él era ingeniero del United States Coast and Geodetic Survey. El resultado fue un procedimiento secuencial donde cada paso ofrecía alguna información, pero el problema no quedaba resuelto sino al obtener  $(X'X)^{-1}$ . Los métodos actuales de cálculo por computador generalmente no nos permiten ver lo que está sucediendo, pero, aun así, proporcionan resultados intermedios en varios estados del proceso para obtener la solución.

Primeramente, se disponen los datos para procesarlos en el equipo computador. Sean estos los datos de la tabla 14.1, 30 observaciones multivariantes de 4 valores cada una. Un programa de instrucciones al computador sobre qué ha de hacer y qué ha de imprimir.

A partir de los datos, se construye  $X$ . La ecuación ha de incluir un intercepto, de modo que  $X$  estará conformada por una columna de 30 unos y tres columnas más con los  $X_1$ , los  $X_2$ , y los  $X_3$  de la tabla 14.1. La matriz  $Y$  se obtiene de la columna encabezada  $Y$ . A su vez, se obtienen  $X'X$  y  $X'Y$  y se completan otros cálculos.

El resultado impreso del computador puede incluir una tabla como la 14.1. Esto permite verificar si hay errores de transcripción. Las matrices  $X'X$  y  $(X'X)^{-1}$  pueden imprimirse. Estas se presentan en la tabla 14.4. Si la ecuación del modelo incluye un intercepto, como en este caso, el valor superior izquierdo de  $X'X$  será  $n = 30$ .

**Tabla 14.4** Las matrices  $X'X$  y  $(X'X)^{-1}$  para los datos del tabaco

Combustión foliar en segundos en muestras de tabaco tomadas en campos de agricultores

##### MATRIZ $X'X$

VARIABLE DEPENDIENTE:  $Y$

	INTERCEPTO	X1	X2	X3
INTERCEPTO	30.00000000	98.36000000	24.23000000	139.61000000
X1	98.36000000	332.33520000	81.58340000	459.40520000
X2	24.23000000	81.58340000	30.19070000	120.39500000
X3	139.61000000	459.40520000	120.39500000	682.78130000

##### MATRIZ INVERSA DE $X'X$

	INTERCEPTO	X1	X2	X3
INTERCEPTO	1.71457789	-0.32895337	0.09529059	-0.14605239
X1	-0.32895337	0.10623365	-0.02105557	-0.00050401
X2	0.09529059	-0.02105557	0.11706378	-0.02595908
X3	-0.14605239	-0.00050401	-0.02595908	0.03624479

Tabla 14.5 Análisis de la regresión de los datos de tabaco

COMBUSTION FOLIAR EN SEGUNDOS EN MUESTRAS DE TABACO TOMADAS DE CAMPOS DE AGRICULTORES						
VARIABLE DEPENDIENTE: Y	LOG DE COMBUSTION FOLIAR					
FUENTE	GL	SUMA DE CUADRADOS CUADRADO MEDIO	VALOR F	PR>F	R-CUADRADO	C.V.
MODELO	3	5.50473408	1.83491136	40.27	0.0001	0.822889
ERROR	26	1.18478592	0.04555669		DESV.EST.	31.1178
TOTAL CORREGIDO	29	6.68922000		0.21346824	MEDIA Y	0.68360000
FUENTE	GL	SC SECUENCIAL	VALOR F	PR > F	GI	SC PARCIAL
X <sub>1</sub>	1	3.44600764	75.62	0.0001	1	2.63871240
X <sub>2</sub>	1	0.85384481	18.74	0.0002	1	1.65106256
X <sub>3</sub>	1	1.20488163	26.44	0.0001	1	1.20488163
T PARA HO: PARAMETRO = 0 PR >  T  ERROR ESTANDAR DE ESTIMACION						
PARAMETRO	ESTIMACION	PARAMETRO = 0				
INTERCEPTO	1.8104253	6.48	0.0001		0.27951935	
X <sub>1</sub>	-0.53145530	-7.64	0.0001		0.06957678	
X <sub>2</sub>	-0.43963579	-6.02	0.0001		0.07303727	
X <sub>3</sub>	0.20897531	5.14	0.0001		0.04064022	

La tabla 14.5 es un impresio del SAS; ver A. J. Barr et al. (14.13). En el análisis de la varianza, frente a MODELO, encontramos  $gl = 3$ . Esta da SC(regresión) en nuestra notación puesto que claramente se refiere sólo a las tres variables  $X$  y es  $\hat{\beta}'X'Y - (\sum Y)^2/n = \hat{\beta}'X'_A Y_A$ . La columna encabezada por PR > F da la probabilidad de que un valor aleatorio de  $F$  sea mayor que el observado. Aquí, lo anunciado es que  $P(F > 40.27) \leq 0.0001$ . Esta prueba es de  $H_0: \beta_1 = \beta_2 = \beta_3 = 0$  simultáneamente frente a  $H_1: \text{al menos un } \beta_i \neq 0$  donde los  $\beta$  son coeficientes de regresión parcial. Evidentemente la tasa de error es experimental. R-CUADRADO es el cuadrado del coeficiente múltiple. Es decir,

$$R^2 = [\hat{\beta}'X'Y - (\sum Y)^2/n]/[Y'Y - (\sum Y)^2/n] = 5.504734/6.689520 = .822889.$$

La segunda parte de la tabla incluye pruebas de las  $X$  como fuentes de variación, equivalentes a pruebas de los  $\beta$  asociados. Bajo FUENTE, tenemos la secuencia en la cual las  $X$  se añaden al modelo después de la media. La suma de cuadrados obtenida en esta secuencia están bajo SC SECUENCIAL. Primero está  $X_1$ , que corresponde a una suma adicional de cuadrados de 3.4460; para esto,  $F = 75.62^{**}$ . Sabemos que esto es igual a

$$[\sum (X_1 - \bar{X}_1)(Y - \bar{Y})]^2 / \sum (X_1 - \bar{X}_1)^2.$$

Luego se incluye  $X_2$  en la ecuación y le corresponde una suma de cuadrados adicionales de 0.8538; para este,  $F = 18.74^{**}$ . Este cálculo no es muy obvio para nosotros, puesto que tanto  $Y$  como  $X_2$ , se han tenido que poner en regresión respecto de  $X_1$  primero. A su vez, las desviaciones de  $Y$  resultantes de la regresión respecto de  $X$  se relacionan por regresión respecto de las desviaciones  $X_2$  de la regresión respecto de  $X_1$  en este procedimiento secuencial. Finalmente, se incluye  $X_3$  y corresponde a una suma adicional de cuadrados de 1.2049;  $F = 26.44^{**}$ . En efecto los valores  $F$  prueban secuencialmente  $H_0: \beta_{Y1} = 0, H_0: \beta_{Y2 \cdot 1} = 0$ , y  $H_0: \beta_{Y3 \cdot 12} = 0$ ,

El último  $X$  que se ha de incluir,  $X_3$ , tiene una suma de cuadrados secuencial que permite una prueba de  $X_3$  ajustada por  $X_1$  y por  $X_2$ . Esta suma de cuadrados debe ser también una suma de cuadrados parcial y así prueba  $H_0: \beta_{Y3 \cdot 12} = 0$ . Hasta aquí, ésta es la única prueba de un coeficiente de regresión parcial que figura efectivamente en la ecuación de regresión requerida.

En la columna SC PARCIAL figuran cálculos que han sido repetidos de modo que  $X_1$  puede ser último en la ecuación y, de nuevo, de modo que  $X_2$  puede ser último. Esto conduce a las pruebas usuales de coeficientes de regresión parcial. Las pruebas  $F$  dadas son equivalentes a las pruebas  $t$  y por eso tienen tasas de error por comparación.

Finalmente, se dan estimaciones de los parámetros en  $\beta$ . Las pruebas de significancia son las pruebas  $t$  y equivalen a pruebas  $F$  para sumas parciales de cuadrados. Esto puede verificarse al elevar al cuadrado los valores de  $t$ . Así mismo PR > F y PR > |T| son idénticas para cada  $X$  correspondiente. Los valores finales corresponden a las desviaciones estándar o a los errores estándar. Estos pueden calcularse directamente a partir de  $(X'X)^{-1}$  y  $s^2$ . Por ejemplo, para hallar  $s_{\beta_1}$ , obténgase el valor apropiado de  $(X'X)^{-1}$  y  $s^2$  a partir del análisis de varianza;  $s_{\beta_1} = \sqrt{.106234(0.45569)} = 0.069577$  como en la columna de errores estándar. Los errores estándar se necesitan para fijar intervalos de confianza de estos parámetros.

Los valores de  $F$  para SC PARCIALES indican que todos los  $X$  son importantes en esta ecuación de regresión; no es necesario considerar la eliminación de ellos. Este no siempre es el caso.

Es posible encontrar datos en los que algunas de las variables independientes sean significantes en el procedimiento secuencial y no lo sea ninguno de los coeficientes de regresión parcial. Esto generalmente indicará correlaciones bastante altas entre todas las variables. Esto se denomina *intercorrelación* o *multicolinealidad*, este último término se aplica a una relación perfecta o casi perfecta. En consecuencia, cualquier variable independiente aislada puede eliminarse; las demás estarán relacionadas lo suficientemente para compensar la pérdida. Así, si se hace la regresión peso respecto de estatura y longitud de la pierna, es probable una correlación lo suficientemente alta entre las dos últimas de modo que podemos prescindir de una de ellas pero no de ambas; ambos coeficientes de correlación parcial podrían no ser significantes.

Los resultados impresos del computador pueden incluir valores de regresión correspondientes a todos los valores observados. Los resultados impresos del SAS exigen estos valores predichos, que aquí no se dan. También se calculan residuos, se elevan al cuadrado y se suman para comparar con la suma de cuadrados del error. Estas sumas difieren sólo debido a errores de redondeo.

Como se han dado  $(\mathbf{X}'\mathbf{X})^{-1}$  y  $s^2$ , pueden calcularse pruebas de hipótesis relacionadas con los parámetros e intervalos de confianza en conjuntos especificados de valores  $X$ .

**Ejercicio 14.7.1** Obtener la ecuación de regresión para la regresión de  $Y$  respecto de  $X_1$ ,  $X_2$  y  $X_3$ , con los datos del ejercicio 14.5.2. Asegúrese de obtener  $\mathbf{X}'\mathbf{X}$  y  $(\mathbf{X}'\mathbf{X})^{-1}$  o  $\mathbf{X}_A'\mathbf{X}_A$  y  $(\mathbf{X}_A'\mathbf{X}_A)^{-1}$ .

**Ejercicio 14.7.2** Presentar un análisis de varianza para el ejercicio 14.7.1. Prestar simultáneamente  $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ . ¿Cuál es la probabilidad de observar un valor aleatorio de  $F$  que sea mayor que el observado? ¿Cuáles son  $\bar{Y}$ ,  $s$ , CV, y  $R_{Y \cdot 123}^2$ ?

**Ejercicio 14.7.3** Se considera que  $X_3$  es una alternativa para  $X_2$  para evaluar las respuestas de fosfato. ¿Da  $X_3$  información respecto a  $Y$  no disponible todavía a partir de  $X_1$  y  $X_2$ ? ¿Cuáles de las  $X$  no estaría usted dispuesto a eliminar de ninguna manera de la ecuación de regresión?

**Ejercicio 14.7.4** Estimar la media poblacional correspondiente a (3,000, 80, 6.40). Construir un intervalo de confianza para el parámetro.

Repetir para (1,500, 60, 6.00).

**Ejercicio 14.7.5** Supóngase que los valores del ejercicio 14.7.4 van a hacer predicciones de  $Y$  futuros. ¿Cuáles serían sus errores estándar?

**Ejercicio 14.7.6** Repetir el ejercicio 14.7.1 y 14.7.2 con los datos del ejercicio 10.2.3 usando todos los  $X$ .

¿Qué variable parece menos útil en estimación de medias poblacionales?

Estimar la media poblacional para  $\mathbf{X}' = (1, 50, 75, 280, 70)$ . ¿Cuál es la varianza de esta estimación?

**Ejercicio 14.7.7** Repetir el ejercicio 14.7.6 con los datos del ejercicio 10.2.5. Para  $\mathbf{X}$ , úsese  $\mathbf{X}' = (12,000, 500, 500, 1,600)$ .

#### 14.8 Miscelánea

El problema de seleccionar la “mejor” ecuación de regresión no se considera aquí si bien se insinuó al referirnos a la multicolinealidad. Hay varios procedimientos, Draper y Smith (14.11) discuten el problema con algún detalle.

Si las variables independientes se ordenan de tal modo que las que se sabe que son importantes están primero y luego las restantes de importancia desconocida, entonces se puede probar simultáneamente el último conjunto, lo cual es evidente por la columna SC SECUENCIAL. Simplemente sumamos las sumas de cuadrados de las del conjunto dudosos, que es el último en el procedimiento secuencial, hallamos el cuadrado medio y lo usamos como numerador en una prueba  $F$ .

Es posible estimar conjuntamente los parámetros en el vector  $\beta$  mediante una región de confianza. La frontera será probablemente una extensión de la idea de elipse o elipsoide.

Se puede calcular una *región de confianza* para toda la superficie de regresión. Esta es una extensión del procedimiento para construir una banda de confianza dado en la sec. 10.6.

Debemos evitar hacer *extrapolaciones* o hay que hacerlas con cuidado fuera de la región de confianza definida por el conjunto observado de valores  $(X_1, \dots, X_k)$ . No es suficiente examinar la amplitud de las  $X$  para determinar esta región. Por ejemplo, en dos dimensiones la amplitud implica un rectángulo, pero los  $(X_1, X_2)$  observados pueden estar en una elipse. Una superficie de regresión es, en general, una aproximación a la verdadera situación, y la extrapolación implica que la aproximación sigue siendo válida fuera de la región determinada por el conjunto observado de vectores  $X$ .

Los resultados impresos del computador pueden incluir también *gráficas de residuos* respecto de diferentes variables, por ejemplo, para cada  $X$  por separado. Esto ayudaría a determinar si es adecuada una ecuación lineal en cada  $X$ . Draper y Smith (14.11) presentan una buena exposición de tales representaciones. Puede disponerse de otras representaciones.

Es posible también la *predicción* de una sola observación futura, en vez de la estimación de un parámetro. Aquí hay que sumar 1 al coeficiente del cuadrado medio de error de una estimación de un parámetro. En otros términos,  $\hat{V}(Y \text{ predicho}) = [1 + X'_0(X'X)^{-1} X_0]s^2$ . Para tener una nueva media de  $m$  observaciones sumar  $1/m$  en lugar de 1.

Claro está que las técnicas de este capítulo son apropiadas para modelos *polinómicos*, modelos siempre lineales en los parámetros, aunque no lineales en la variable o variables independientes. La ecuación modelo

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_k X_i^k + \varepsilon_i$$

se llama modelo de *orden k* con una variable independiente. En la sec. 19.4 se ilustra el caso para  $k = 2$ . Para un modelo de segundo orden en dos variables, la ecuación es

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_{11} X_{1i}^2 + \beta_2 X_{2i} + \beta_{22} X_{2i}^2 + \beta_{12} X_{1i} X_{2i} + \varepsilon_i$$

El término  $X_{1i} X_{2i}$  examina la posibilidad de que una variación lineal en  $Y$  por unidad de variación en  $X_1$  pueda depender del valor de  $X_2$ . El hecho de que la respuesta

lineal no sea constante es un ejemplo de *interacción*. El enunciado es simétrico en  $X_1$  y  $X_2$ .

La advertencia concerniente a la extrapolación es obvia para polinomios. Simplemente no sería deseable una estimación de una media poblacional donde  $X_1 = 3$  y  $X_1^2 = 16$ , puesto que  $3^2 \neq 16$ .

Finalmente, cuando la regresión pasa por el origen, no se necesita intercepto en el modelo.

#### 14.9 Coeficientes de regresión parcial estándar

Los coeficientes estándar de regresión parcial son los de ecuaciones donde se han estandarizado todas las variables, esto es, se han medido respecto de sus medias en unidades de desviaciones estándar. Los coeficientes estándar de regresión parcial pueden denotarse por  $b'_i$  o  $b'_{Y_i+1 \dots i-1, i+1 \dots k}$ . La ecuación estándar de regresión es

$$\frac{\hat{Y} - \bar{Y}}{s_Y} = b'_1 \frac{X_1 - \bar{X}_1}{s_1} + \cdots + b'_k \frac{X_k - \bar{X}_k}{s_k} \quad (14.21)$$

o

$$\hat{Y}' = b'_1 X'_1 + \cdots + b'_k X'_k$$

Al comparar la ec. (14.3) o la (14.4) con la (14.21), resulta la ec. (14.22), que relaciona los  $b$  con los  $b'$ .

$$b'_i = b_i \frac{s_i}{s_y} \quad \text{y} \quad b_i = b'_i \frac{s_y}{s_i} \quad (14.22)$$

Puesto que ahora todo  $b'_i$  es adimensional, una comparación de dos cualesquiera de ellos da una medida de la importancia relativa de los dos  $X$  que intervienen. Si  $b'_1$  es el doble de  $b'_2$ , entonces  $X_1$  es aproximadamente 2 veces más importante que  $X_2$ , para estimar o predecir  $Y$  en el sentido de que una unidad de variación en  $X_1$  produce variación doble de  $Y$  para una unidad de variación en  $X_2$ . Es apropiado advertir que las desviaciones estándar de los  $b'$  no son iguales y deben considerarse en el contexto de los otros  $X$  en el modelo.

Para el ejemplo de la sec. 14.5,

$$\begin{aligned} b'_1 &= b_1 \frac{s_1}{s_y} = -0.5285 \sqrt{\frac{9.845547/29}{6.68952/29}} \\ &= -0.6412 \end{aligned}$$

$$\begin{aligned} b'_2 &= b_2 \frac{s_2}{s_y} = -0.2900 \sqrt{\frac{10.620937/29}{6.68952/29}} \\ &= -0.3654 \end{aligned}$$

Cuando la ecuación estándar se obtiene a partir de los datos originales pasando directamente a variables estándar, la matriz  $X'_A X_A$  consta de elementos en la diagonal principal y  $r_{ij}$  como elementos fuera de la diagonal.

Por lo demás, si consideramos la matriz de correlación de todas las variables, incluida la dependiente, de modo que se tenga  $k + 1$  variables y usamos la notación  $R$  y  $R^{-1} = C$  como en las ecs. (14.13) y (14.14), entonces es fácil hallar los coeficientes de regresión parcial estándar por

$$b'_{ij \cdot 1 \dots i-1, i+1 \dots j-1, j+1 \dots k+1} = \frac{-C'_{ij}}{C'_{ii}} \quad (14.23)$$

Nótese que los coeficientes no son simétricos en  $i$  y  $j$ ; en efecto se cumple la ec. (14.24).

$$r_{12 \cdot 3 \dots k+1} = \pm \sqrt{b'_{12 \cdot 3 \dots k+1} b'_{21 \cdot 3 \dots k+1}} \quad (14.24)$$

Ambos  $b$  tendrán el mismo signo y el signo común se da a la raíz cuadrada. Esta ecuación es válida también para los  $b$ .

**Ejercicio 14.9.1** Calcular los tres coeficientes de regresión parcial estándar para los datos del ejercicio 14.5.2.

**Ejercicio 14.9.2** Johnson y Hasler (14.10) estudiaron varios factores que influyen en la producción de trucha arco iris en tres lagos eutróficos. Obtuvieron los siguientes coeficientes de correlación simple.

	$X_2$	$X_3$	$X_4$	$X_5$
$X_1$	0.2206	-0.3284	-0.0910	-0.2160
$X_2$		0.6448	-0.1566	-0.1079
$X_3$			0.0240	-0.2010
$X_4$				-0.7698

$X_1$  = tasa instantánea de crecimiento para el grupo de edad I de truchas para el intervalo de estudio

$X_2$  = índice de densidad de zooplancton (disponibilidad de alimento)

$X_3$  = producción total permanente de trucha (competencia)

$X_4$  = temperatura del agua

$X_5$  = tamaño de truchas de grupo de edad I

Si el computador tiene un programa para invertir matrices, invertir la matriz de correlación anterior.

Calcular  $r_{12 \cdot 345}, r_{13 \cdot 245}, r_{14 \cdot 235}$ , y  $r_{15 \cdot 234}$ .

Calcular todos los coeficientes de regresión parcial estándar y escribir la ecuación de regresión parcial estándar con  $X_1$  como variable dependiente.

## Referencias

- 14.1. Anderson, R. L., y T. A. Bancroft: *Statistical Theory in Research*, McGraw-Hill, Nueva York, 1952.
- 14.2. Birch, H. F.: "Phosphate response, base saturation and silica relationships in acid soils," *J. Agr. Sci.*, 43:229-235 (1953).
- 14.3. Cochran, W. G.: "The omission or addition of an independent variate in multiple linear regression," *J. Roy. Statist. Soc. Suppl.*, 5:171-176 (1938).
- 14.4. Cowden, D. J.: "A procedure for computing regression coefficients," *J. Amer. Statist. Ass.*, 53:144-150 (1958).
- 14.5. Cramer, C. Y.: "Simplified computations for multiple regression," *Ind. Qual. Contr.*, 8:8-11 (1957).
- 14.6. Drapala, W. J.: "Early generation parent-progeny relationships in spaced plantings of soybeans, medium red clover, barley, Sudan grass, and Sudan grass times sorghum segregates," tesis doctoral, University of Wisconsin, Madison, 1949.
- 14.7. Durand, D.: "Joint confidence regions for multiple regression coefficients," *J. Amer. Statist. Ass.*, 49:130-146 (1954).
- 14.8. Dwyer, P. S.: *Linear Computations*, Wiley, Nueva York, 1951.
- 14.9. Friedman, J., and R. J. Foote: "Computational methods for handling systems of simultaneous equations," *U.S. Dep. Agr. Handb.*, 94, 1957.
- 14.10. Johnson, W. E., y A. D. Hasler: "Rainbow trout production in dystrophic lakes," *J. Wildlife Manag.*, 18:113-134 (1954).
- 14.11. Draper, N., y H. Smith: *Applied Regression Analysis*, Wiley, Nueva York, 1966.
- 14.12. Reynolds, J., G. Cunningham y J. Syvertsen: "A net CO<sub>2</sub> exchange model for *Larrea tridentata*," *Photosyn.*, 13(3):In press (1979).
- 14.13. Barr, A. J., J. H. Goodnight, J. P. Sall, y J. T. Helwig: *A User's Guide to SAS 76*, SAS Institute, Raleigh, N.C., 1976.

---

## CAPITULO QUINCE

---

### ANALISIS DE LA VARIANZA III: EXPERIMENTOS FACTORIALES

#### 15.1 Introducción

En este capítulo tratamos experimentos factoriales. Aquí hay varios tratamientos en cada una de varias categorías y definen un marco de tratamientos. Esta elección de diseño de tratamientos conduce a una partición lógica de las sumas de cuadrados de tratamientos en componentes aditivas con pruebas correspondientes de hipótesis.

El tema de las comparaciones independientes, introducido en la sec. 8.3 se trata más profundamente, concediendo atención especial a tratamientos igualmente espaciados y a la regresión. También se da la prueba de no aditividad de Tukey.

#### 15.2 Experimentos factoriales

Un *factor* es una clase de tratamiento, y en experimentos factoriales, todo factor proporcionará varios tratamientos. Por ejemplo, si la dieta es un factor en un experimento, entonces se usarán varias dietas; si temperatura de horneado es un factor, entonces el horneado se hará a varias temperaturas.

El concepto de experimento factorial puede ilustrarse mediante un ejemplo. Considérese un experimento para evaluar rendimientos de variedades de soya. En el caso de un solo factor, todas las variables diferentes a las variedades se mantienen tan uniformes como sea posible, esto es, se escoge solo un nivel de los otros factores. Supóngase que también es de interés un segundo factor, distancia entre surcos. Se puede planear un experimento con dos factores en el que los tratamientos consisten en todas las combinaciones entre las variedades y los espaciamientos elegidos de los surcos, esto es, cada variedad se encuentra presente en todos los espaciamientos de surcos. En un experimento de un solo factor, todas las variedades se plantarán a solo un espaciamiento de un surco, o una sola variedad en todos los espaciamientos entre surcos. En suelos, puede diseñarse un experimento para comparar todas las combinaciones de varios niveles de fertilizantes de fósforo

y potasio. En un experimento de nutrición animal, los factores en consideración pueden ser las cantidades y clases de suplementos de proteínas.

El término *nivel* se refiere a los diferentes tratamientos dentro de un factor. Se deriva de alguno de los primeros experimentos factoriales. Estos trataban de fertilidad de suelos donde las combinaciones de diferentes cantidades, o niveles, de los diferentes fertilizantes eran los tratamientos. Hoy esa palabra tiene un sentido más general, que implica una cantidad o estado dados de un factor. Así, si se comparan 5 variedades de un cultivo, usando tres diferentes prácticas de manejo, el experimento se llama experimento factorial  $5 \times 3$ , con cinco niveles del factor variedad y tres niveles del factor manejo. El número de factores y niveles que pueden compararse en un solo experimento sólo se limita por consideraciones prácticas.

Así, un *experimento factorial* es aquel en el que el conjunto de tratamientos consiste en todas las combinaciones posibles de los niveles de varios factores. En la palabra "factorial" está implicado el concepto de un diseño de tratamientos.

Los experimentos factoriales se usan prácticamente en todos los campos de investigación. Son de gran valor en trabajo exploratorio cuando se sabe poco sobre niveles óptimos de los factores, o ni siquiera cuales son importantes. Considérese un cultivo nuevo para el cual se tienen varias variedades promisorias, pero se sabe poco respecto a fecha y densidad de siembra adecuadas. En este caso es indicado usar un experimento de tres factores. Si se usa el enfoque de un solo factor, se seleccionan una fecha y una densidad y se efectúa un experimento con las variedades. Sin embargo, puede ocurrir que la mejor variedad para la fecha y densidad de siembra escogidas, no sea la mejor para otras fecha y densidad. Cualquier otro factor individual debe implicar sólo una variedad. Otras veces, el experimentador puede estar ante todo interesado en la intersección entre los factores, es decir, saber si las diferencias en respuesta a los niveles de un factor son semejantes o diferentes a niveles diferentes de otro factor o factores. Cuando se dispone de información considerable, el mejor enfoque puede ser comparar sólo un número muy limitado de combinaciones de varios factores o niveles específicos.

Así vemos que el alcance de un experimento, o la población para la cual pueden hacerse inferencias, puede a menudo aumentarse mediante el uso de un experimento factorial. Es particularmente importante hacer esto cuando se requiere información sobre algún factor para el cual se van a hacer recomendaciones con amplio margen de condiciones.

**Notación y definiciones** Los sistemas de notación que se usan en experimentos factoriales generalmente son similares, pero presentan diferencias suficientes como para que el lector tenga que comprobar con cuidado cuando utilice nuevas referencias. Seguimos una notación parecida en muchos aspectos a la sugerida por Yates (15.18). Las tres letras mayúsculas se usan para designar *factores*, por ejemplo, en un experimento en que entran varias dispersiones para controlar insectos, y éstas se aplican mediante varios métodos, podemos denotar un factor dispersión de insecticida por *A* y el factor método por *B*. Las combinaciones de letras minúsculas y subíndices numéricos, o simplemente los subíndices, se usan para denotar *combinaciones de tratamientos* y medias; por ejemplo  $a_1 b_3$  puede indicar la combinación de tratamiento compuesta del primer nivel de *A* y el tercer nivel de *B*, y a la correspondiente media de tratamiento. A menudo, el cero se usa para el primer nivel de un subíndice

Para un experimento con dos factores con dos niveles de cada factor, o sea, para un factorial  $2 \times 2$  ó  $2^2$ , cualquiera de las siguientes notaciones es adecuada. La primera y la tercera pueden generalizarse fácilmente a muchos factores y niveles; la segunda puede extenderse fácilmente a muchos factores, pero sólo a dos niveles de cada factor.

Factor	A						
	Forma completa		Formas abreviadas				
B	Level	$a_1$	$a_2$	$a_1$	$a_2$	$a_1$	$a_2$
	$b_1$	$a_1 b_1$	$a_2 b_1$	(1)	$a$	00	10
	$b_2$	$a_1 b_2$	$a_2 b_2$	$b$	$ab$	01	11

Los tres grados de libertad y las sumas de cuadrados para la varianza entre las cuatro medias de tratamientos en un factorial  $2^2$  pueden particionarse en grados de libertad únicos e independientes y sus correspondientes sumas de cuadrados cuya interpretación general tiene sentido, y es relativamente simple. Para el experimento factorial general, se partitionan los grados de libertad y las sumas de cuadrados en subconjuntos o componentes aditivos, no necesariamente con grados de libertad únicos. Los principios que supone la partición se ilustran mejor mediante el uso de una tabla. La tabla 15.1 es una ilustración de factorial  $2^2$ .

En la tabla 15.1 sean los números promedios o medias de las respuestas medidas a las combinaciones de tratamientos indicadas por los encabezamientos de fila y columna; las medias son para todas las replicaciones.

Las cuatro diferencias,  $a_2 - a_1$ , a cada nivel de  $B$  y  $b_2 - b_1$  a cada nivel de  $A$ , se llaman *efectos simples*. Estos no se incluyen en el resumen usual de un experimento factorial, pero son muy útiles en la interpretación del resumen, como también en sí mismos. Para los datos en I, el efecto simple de  $A$  al primer nivel de  $B$  es 2; en II, el efecto simple de  $B$  al segundo nivel de  $A$  es -6.

Cuando se promedian efectos simples, los resultados se llaman *efectos principales*. Estos se detectan con letras mayúsculas, como los factores. El efecto principal del factor  $A$  para los datos en I es 5; el efecto principal del factor  $B$  para los datos en III es 6. En general, para un factorial  $2^2$ ,  $A$  y  $B$  están dados por las ecs. (15.1) y (15.2).

$$\begin{aligned} A &= \frac{1}{2}[(a_2 b_2 - a_1 b_2) + (a_2 b_1 - a_1 b_1)] \\ &= \frac{1}{2}[(a_2 b_2 + a_2 b_1) - (a_1 b_2 + a_1 b_1)] \end{aligned} \quad (15.1)$$

$$\begin{aligned} B &= \frac{1}{2}[(a_2 b_2 - a_2 b_1) + (a_1 b_2 - a_1 b_1)] \\ &= \frac{1}{2}[(a_2 b_2 + a_1 b_2) - (a_2 b_1 + a_1 b_1)] \end{aligned} \quad (15.2)$$

Los efectos principales se calculan por unidades. Las ecuaciones (15.1) y (15.2) pueden generalizarse fácilmente a factoriales  $2^n$ .

**Tabla 15.1 Ilustración de efectos simples, efectos principales e interacciones**

Factor	I $A = \text{Clase}$				
	Nivel	$a_1$	$a_2$	Media	$a_2 - a_1$
$B = \text{tasa}$	$b_1$	30	32	31	2
	$b_2$	36	44	40	8
	Media	33	38	35.5	5
	$b_2 - b_1$	6	12	9	
Factor	II $A = \text{Clase}$				
	Nivel	$a_1$	$a_2$	Media	$a_2 - a_1$
$B = \text{tasa}$	$b_1$	30	32	31	2
	$b_2$	36	26	31	-10
	Media	33	29	31	-4
	$b_2 - b_1$	6	-6	0	
Factor	III $A = \text{Clase}$				
	Nivel	$a_1$	$a_2$	Media	$a_2 - a_1$
$B = \text{tasa}$	$b_1$	30	32	31	2
	$b_2$	36	38	37	2
	Media	33	35	34	2
	$b_2 - b_1$	6	6	6	

Los efectos principales son promedios en una variedad de condiciones, lo mismo que cualesquiera otras medias de tratamiento. Para un experimento factorial en un diseño de bloques completos al azar o en cuadrado latino, la variedad de condiciones se da dentro de los bloques, así como también entre los bloques; de este modo el factor  $A$  se replica dentro de cada bloque ya que está presente a ambos niveles para cada nivel del factor  $B$ . Esta es una *replicación oculta*. El hecho de promediar implica que las diferencias, esto es, los efectos simples, varían sólo debido al azar de un nivel a otro del factor o factores. Esta es, en efecto, una hipótesis usualmente sometida a una prueba de significancia cuando los tratamientos están dispuestos factorialmente; la hipótesis es la de que no hay interacción entre factores.

Para los datos en I y II, los efectos simples difieren para ambos factores, clase y caso. En III, los efectos simples para  $A$  son iguales, lo mismo que para los de  $B$ ; aquí también son iguales al efecto principal correspondiente. Cuando los efectos simples para un factor difieren más de lo que pueda ser atribuible al azar, esta respuesta diferencial se llama *interacción* de los dos factores. Esta relación es simétrica; esto es, la interacción de  $A$  con  $B$  es lo mismo que la de  $B$  con  $A$ . A partir de la tabla 15.1 se verá que la diferencia en los efectos simples de  $A$  es igual a la de  $B$  en los tres casos; sería imposible construir una tabla de otra forma. En nuestra notación, la interacción de  $A$  y  $B$  se define en

$$\begin{aligned} AB &= \frac{1}{2}[(a_2 b_2 - a_1 b_2) - (a_2 b_1 - a_1 b_1)] \\ &= \frac{1}{2}[(a_2 b_2 + a_1 b_1) - (a_1 b_2 + a_2 b_1)] \end{aligned} \quad (15.3)$$

Se usa el valor  $\frac{1}{2}$  de modo que la interacción, lo mismo que los efectos principales, se da por unidad. Para los datos en I,

$$\begin{aligned} AB &= \frac{1}{2}(8 - 2) && \text{en términos de los efectos simples de } A \\ &= \frac{1}{2}(12 - 6) && \text{en términos de los efectos simples de } B \\ &= 3 \end{aligned}$$

Para los datos en II, encontramos

$$AB = \frac{1}{2}(26 - 36 - 32 + 30) = -6$$

y en III,

$$AB = \frac{1}{2}(38 - 36 - 32 + 30) = 0$$

Observese que la interacción también es la mitad de la diferencia entre las sumas de las dos diagonales de la tabla  $2 \times 2$  que es la mitad de la diferencia entre las sumas de los tratamientos, donde  $A$  y  $B$  están presentes a los niveles más alto y más bajo y de los tratamientos donde sólo uno está presente al nivel más alto. Esto siempre es cierto para el factorial  $2^2$ .

La interacción mide el que no logre el efecto  $A$ , o la respuesta a  $A$ , de ser la misma para cada nivel de  $B$ , o al contrario, el no lograr el efecto  $B$  de ser el mismo para cada nivel de  $A$ . En I, los efectos simples por clase son 2 y 8 en tanto que el efecto principal es 5. La interacción puede definirse como una medida de la discrepancia de los efectos simples con respecto a una ley o modelo aditivo basado sólo en efectos principales.

En I, tabla 15.1, la respuesta a  $A$  o el aumento de  $a_1$  a  $a_2$  es mayor para  $b_2$  que para  $b_1$ , esto es, ha habido variación en la magnitud del incremento. En II, la respuesta a  $A$  es un aumento en presencia de  $b_1$  y una disminución en presencia de  $b_2$ ; ha habido un cambio en la dirección del incremento. En términos de medias de tratamientos presentadas en una tabla de dos factores, variaciones suficientemente grandes en las magnitudes de

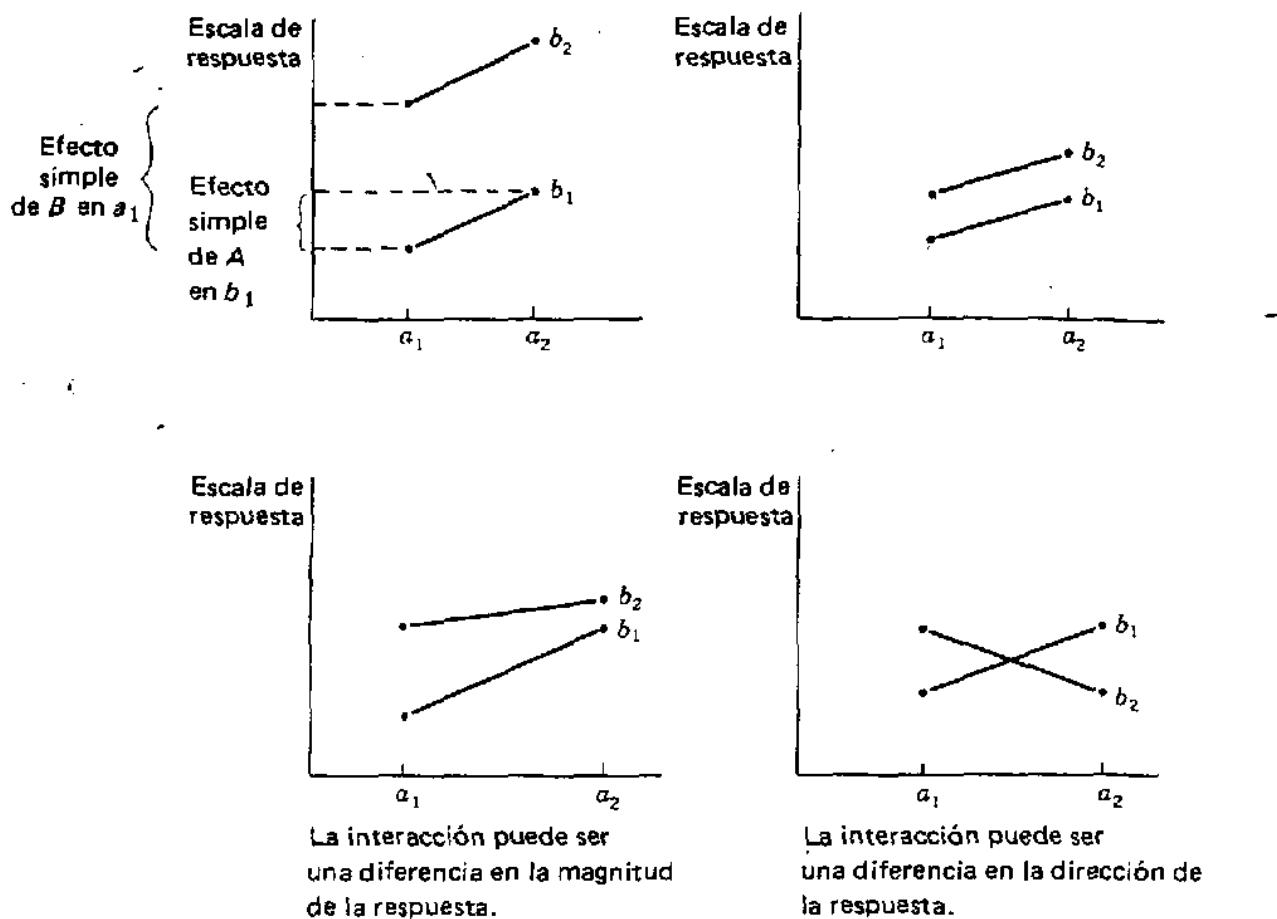


Figura 15.1 Ilustración de la interacción.

las diferencias entre medias de tratamientos en una columna (o fila), al pasar de columna a columna (o de fila a fila), pueden constituir una interacción. Además, los cambios de rango de cualquier media de tratamiento de una columna (o fila), al pasar de columna a columna (o filas), puede constituir una interacción.

La figura 15.1 ilustra gráficamente lo que se entiende por interacción. La presencia o ausencia de efectos principales no nos dice nada respecto a la presencia o ausencia de interacción. La presencia o ausencia de interacción no nos dice nada respecto a la presencia o ausencia de efectos principales, pero nos dice algo acerca de la homogeneidad de los efectos simples.

Una interacción significante es aquélla que es lo suficientemente grande como para que se pueda explicar con base en el azar y la hipótesis nula de que no hay interacción. Cuando la interacción es significante, los factores no son independientes entre sí; los efectos simples de un factor difieren y la magnitud de un efecto simple depende del nivel del otro factor del término de la interacción. Cuando los factores interactúan, un experimento de un solo factor conduce a información desconectada y posiblemente confusa.

Si la interacción es no significante, se puede concluir que los factores en consideración son independientes entre sí; los efectos simples de un factor son los mismos para todos los niveles de los otros factores, dentro de una variación aleatoria medida por el error experimental. El promedio de los efectos simples, esto es, el efecto principal, es lo apropiado y la mejor estimación de la diferencia común. Lo comprobado es que no hay discrepancia con respecto a un modelo aditivo simple con componentes debidas a los nive-

les de solo los factores. Cuando los factores son independientes, el experimento factorial es un ahorro considerable en tiempo y esfuerzo. Esto es así porque los efectos simples son iguales a los efectos principales correspondientes a un efecto principal; en un experimento factorial, se estima tan exactamente cómo sería si todo el experimento se dedicara a ese factor.

En el análisis de un experimento factorial, es correcto particionar los grados de libertad de los tratamientos y la suma de cuadrados en las componentes atribuibles a efectos principales y a interacciones aun cuando la prueba  $F$  total de que no hay diferencias entre tratamientos no sea significante; los efectos principales y comparaciones de interacción son comparaciones planeadas. Es fácil visualizar una situación en que un factor, por ejemplo  $B$ , no hace ni bien ni mal por sí mismo ni tampoco tiene efecto alguno sobre  $A$ , y por lo tanto no contribuye a la suma de cuadrados de tratamientos con más de lo que se puede atribuir al azar; una respuesta significante a  $A$  bien podría perderse en una prueba general de significancia, como el  $F$ . Así, en un factorial  $2^2$  con 3 grados de libertad para los tratamientos, la suma de cuadrados para el efecto real  $A$  con un grado de libertad puede perderse fácilmente cuando se promedia con las sumas de cuadrados para los  $B$  y  $AB$  no reales en sus 2 grados de libertad. El cálculo de la suma de cuadrados para tratamientos comúnmente se usa más como parte de un procedimiento de cómputo que para proporcionar el numerador de una prueba de  $F$ .

Todas las unidades del experimento factorial entran en la medición de cualquier efecto principal o interacción. Esto es evidente por las ecs. (15.1) a (15.3); se ve así, pues, que en un experimento factorial al contrario de lo que ocurre con experimentos de un solo factor, todas las unidades se dedican a medir  $A$  y a su turno a  $B$  o  $AB$ , y así sucesivamente, cuando hay más factores, nada se pierde bien sea en la replicación o en la medición de efectos principales; algo se gana en cuanto medimos un factor cualquiera a diferentes niveles de los otros factores y, en términos de comparaciones de tratamientos, podemos medir también interacciones y probar hipótesis relacionadas con ellos.

Los resultados de un experimento factorial conducen a una explicación relativamente sencilla debido a la variedad y naturaleza de las comparaciones de tratamientos. Si los factores son en gran parte independientes, la tabla de medias de tratamientos y el análisis de la varianza resumen bien los datos. Cuando los factores no son independientes, los datos necesitan de un estudio detallado con la posibilidad de más experimentación. Aquí la dificultad está en la naturaleza compleja de la situación y no en su enfoque factorial. El experimento factorial ha puesto de presente la complejidad del problema, lo cual podría pasarse por alto si se hubiese usado un enfoque de un solo factor.

### 15.3 El experimento factorial $2 \times 2$ : un ejemplo

Wilkinson (15.17) presenta los resultados de un experimento para estudiar la influencia del tiempo de sangría, factor  $A$ , y el dietilestilbestrol (un componente estrogénico), factor  $B$ , sobre el fosfolípido del plasma en corderos. Se asignaron al azar 5 corderos a cada uno de cuatro grupos de tratamiento: las combinaciones de tratamientos consisten en los tiempos de sangría en la mañana y en la tarde y con y sin tratamiento de dietilestilbestrol. Los datos se dan en la tabla 15.2.

La tabla de totales de tratamientos se puede usar en el cálculo de las sumas de cuadrados para probar las hipótesis respecto a los efectos principales y las interacciones. Los

**Tabla 15.2 Influencia del tiempo de sangría y del dietilestilbestrol sobre los fosfolípidos en corderos**

Grupos de tratamiento

$a_1 = \text{A.M.}$		$a_2 = \text{P.M.}$		Totales
$a_1 b_1 =$ control 1	$a_1 b_2 =$ tratado 1	$a_2 b_1 =$ control 2	$a_2 b_2 =$ tratado 2	
8.53	17.53	39.14	32.00	
20.53	21.07	26.20	23.80	
12.53	20.80	31.33	28.87	
14.00	17.33	45.80	25.06	
10.80	20.07	40.20	29.33	
$\sum Y$	66.39	96.80	182.67	139.06
$\sum Y^2$	963.88	1,887.02	6,913.63	3,912.17
$\bar{Y}$	13.28	19.36	36.53	27.81
				484.92
				13,676.70
				24.25

Totales de los tratamientos

Factor	$A = \text{Tiempo}$			Totales
	Nivel	$(a_1) = \text{A.M.}$	$(a_2) = \text{P.M.}$	
$B = \text{estrógeno}$	$(b_1) = \text{control}$	66.39	182.67	249.06
	$(b_2) = \text{control}$	96.80	139.06	235.86
	Totales	163.19	321.73	484.92

totales de los tratamientos se distinguen de las medias de tratamiento por los paréntesis de los símbolos de las combinaciones de tratamiento. Así,  $(a_2 b_2)$  es la suma extendida a todas las replicaciones de las observaciones hechas en la combinación de tratamientos  $a_2 b_1$ , mientras que  $(a_1)$  es la suma extendida a todas las replicaciones de observaciones hechas sobre las combinaciones de tratamientos  $a_1 b_1$  y  $a_1 b_2$ . Por ejemplo, para el efecto principal de  $A$ , tenemos

$$(A) = [(a_2 b_2) - (a_1 b_2) + (a_2 b_1) - (a_1 b_1)] \quad (15.4)$$

Los totales de los tratamientos también pueden presentarse como en la tabla 15.3. Aquí hacemos énfasis en las comparaciones expuestas en la sec. 8.3. Obsérvese lo fácil que es comprobar la ortogonalidad y, a la vez, la aditividad de las sumas de cuadrados. Esta tabla es especialmente cómoda para el cálculo de los efectos principales y las interacciones.

Las etapas en los cálculos se presentan a continuación. Sean  $r$ ,  $a$  y  $b$  el número de replicaciones (número de observaciones por combinación de tratamientos), los niveles de  $A$ , y los niveles de  $B$ , respectivamente. Entonces el número de tratamientos es  $ab$ . El diseño es completamente aleatorio.

*Paso 1* Calcular el análisis de la varianza sin tener en cuenta la disposición factorial de los tratamientos para el diseño usado. Obtenemos

$$\begin{aligned}\text{Factor de Corrección} &= FC = 11,757.37 \\ \text{SC Total} &= 1,919.33 \\ \text{SC Tratamientos} &= 1,539.41 \\ \text{SC Error} &= 379.92\end{aligned}$$

*Paso 2* A partir de los totales de los tratamientos de la tabla 15.2, calcular las sumas de cuadrados para los efectos principales y la interacción así :

$$\begin{aligned}\text{SC}(A) &= \frac{\sum_i (a_i)^2}{rb} - FC \\ &= \frac{163.19^2 + 321.73^2}{5(2)} - \frac{(484.92)^2}{5(4)} = 1,256.75\end{aligned}$$

o usar la ec. (15.5), repetición de la ec. (8.6), y la tabla 15.3

$$\text{SC}(A) = \frac{(A)^2}{r \sum c_i^2} \quad (15.5)$$

$$= \frac{(158.54)^2}{20} = 1,256.75$$

$$\begin{aligned}\text{SC}(B) &= \frac{\sum_j (b_j)^2}{ra} - FC \\ &= \frac{249.06^2 + 235.86^2}{5(2)} - FC = 8.71\end{aligned}$$

Tabla 15.3 Totales de tratamiento y efectos factoriales

Símbolo del efecto	Totales de tratamientos y coeficientes, $c_i$				totales de efectos	$r \sum c_i^2$
	$(a_1 b_1) = 66.39$	$(a_1 b_2) = 96.80$	$(a_2 b_1) = 182.67$	$(a_2 b_2) = 139.06$		
<i>A</i>	-1	-1	+1	+1	158.54	$5(4) = 20$
<i>B</i>	-1	+1	-1	+1	-13.20	$5(4) = 20$
<i>AB</i>	+1	-1	-1	+1	-74.02	$5(4) = 20$

o bien la ec. (15.6) y la tabla 15.3

$$\begin{aligned} \text{SC}(B) &= \frac{(B)^2}{r \sum c_i^2} \\ &= \frac{(-13.20)^2}{20} = 8.71. \end{aligned} \quad (15.6)$$

$$\begin{aligned} \text{SC}(AB) &= \text{SC(trts)} - \text{SC}(A) - \text{SC}(B) \\ &= 1,539.41 - 1,256.75 - 8.71 = 273.95 \end{aligned}$$

o la ec. (15.7) y la tabla 15.3

$$\begin{aligned} \text{SC}(AB) &= \frac{(AB)^2}{r \sum c_i^2} \\ &= \frac{(-74.02)^2}{20} = 273.95 \end{aligned} \quad (15.7)$$

Los resultados se llevan a una tabla de análisis de la varianza, tal como la tabla 15.4, donde también se presentan los grados de libertad para el caso general.

La interacción significante indica que los factores no son independientes, la diferencia entre los efectos simples de  $A$  para los dos niveles de  $B$  es significante y, reciprocamente, la diferencia en los efectos simples de  $B$  a los dos niveles de  $A$  es significante. O sea que la diferencia en las mediciones entre los tiempos de sangría difiere para los grupos de control y tratado o, lo que es lo mismo, la diferencia en las mediciones entre los animales tratados y los de control es distinta para los dos tipos de sangría. Así, todo efecto simple es dependiente del nivel del otro factor en el experimento. Tenemos

$$\begin{aligned} (AB) &= [(a_2 b_2) - (a_1 b_2)] - [(a_2 b_1) - (a_1 b_1)] \\ &= [139.06 - 96.80] - [182.67 - 66.39] = -74.02 \\ &= [(a_2 b_2) - (a_2 b_1)] - [(a_1 b_2) - (a_1 b_1)] \\ &= [139.06 - 182.67] - [96.80 - 66.39] = -74.02 \end{aligned}$$

Tabla 15.4 Análisis de la varianza para los datos de la tabla 15.2

Fuente	gl	SC	Cuadrado medio	F
Tratamientos		(ab - 1 = 3)	(1,539.41)	
$A$	$a - 1 = 1$	1,256.75	1,256.75	53**
$B$	$b - 1 = 1$	8.71	8.71	<1
$AB$	$(a - 1)(b - 1) = 1$	273.95	273.95	11.5**
Error	$ab(r - 1) = 16$	379.92	23.75	
Total	$rab - 1 = 19$	1,919.33		

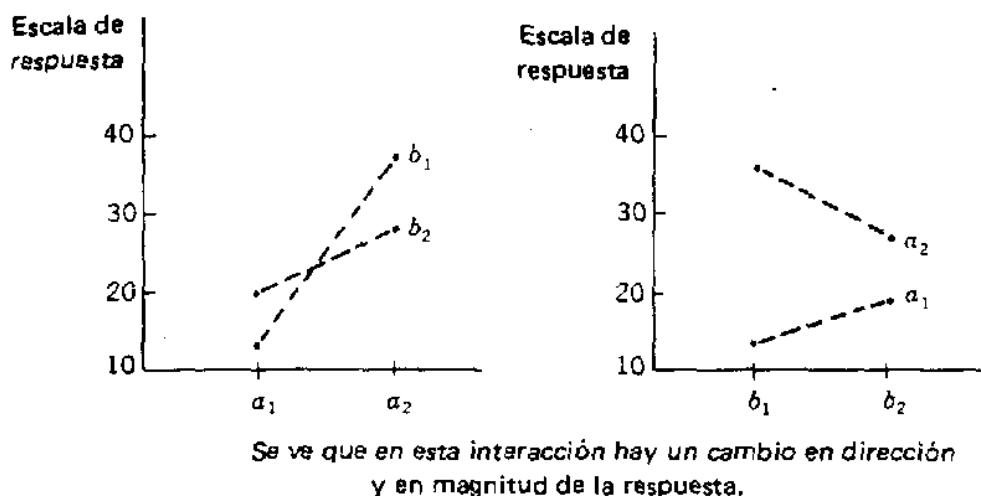


Figura 15.2 Interacción para los datos de la tabla 15.2

donde los paréntesis indican un total sobre replicaciones y los corchetes simplemente llaman la atención sobre los efectos simples a los cuales se refiere el texto. La fig. 15.2 indica la naturaleza de la interacción.

Como resultado de interacción significante  $AB$ , el investigador puede decidir hacer un examen de los efectos simples, ya que se les ha declarado heterogéneos. Las sumas de cuadrados para los efectos simples se calculan como sigue

$$SC(A \text{ dentro de } b_1) = \frac{(182.67 - 66.39)^2}{2 \times 5} = 1,352.10$$

$$SC(A \text{ dentro de } b_2) = \frac{(139.06 - 96.80)^2}{2 \times 5} = 178.59$$

Nótese que la suma de estos es igual a la de  $A$  y  $AB$ , esto es,

$$1,352.10 + 178.59 = 1,530.69 \text{ frente a } 1,530.70 = 1,256.75 + 273.95$$

En el caso de  $A$  y  $AB$ , la información en los dos efectos simples ha sido dispuesta en forma diferente en términos de su media y su varianza. Recuérdese que  $(Y_1 - Y_2)^2/2 = \sum (Y - \bar{Y})^2$  y que la varianza mide dispersión, heterogeneidad u homogeneidad de las observaciones. También

$$SC(B \text{ dentro de } a_1) = \frac{(96.80 - 66.39)^2}{2 \times 5} = 92.48$$

$$SC(B \text{ dentro de } a_2) = \frac{(139.06 - 182.67)^2}{2 \times 5} = 190.18$$

Aquí la suma es igual a la de  $B$  y  $AB$ , esto es,

$$92.48 + 190.18 = 282.66 = 8.71 + 273.95$$

Los resultados pueden presentarse en una tabla auxiliar de sumas de cuadrados de tratamientos. Por ejemplo

Comparación de tratamientos	gl	Cuadrado medio	F
Entre tiempos dentro del control	$a - 1 = 1$	1,352.10	57**
Entre tiempos dentro del tratado	$a - 1 = 1$	178.59	7.5*
Entre niveles de estrógeno, A.M.	$b - 1 = 1$	92.48	3.9
Entre niveles de estrógeno, P.M.	$b - 1 = 1$	190.18	8.0*

Cuando sólo existe un grado de libertad para cada comparación, no se justifica presentar sumas de cuadrados a menos que hagan parte de una tabla, tal como la 15.4. Nótese que la tabla auxiliar contiene 4 grados de libertad para comparaciones de tratamientos, mientras que sólo hay 3 grados de libertad entre medias de tratamientos. Nuestras cuatro comparaciones no pueden ser ortogonales como lo son las comparaciones de efectos principales e interacción de la tabla 15.4. Estas pruebas *F* son esencialmente comparaciones *dm<sub>ij</sub>*, pero efectuadas después de que la interacción significante dió pruebas de diferencias reales entre efectos simples. Por lo tanto, son comparaciones orientadas por los resultados como consecuencia de una prueba de significancia. Críticas muy severas a estas pruebas parecerían injustificadas,

Ejercicio 15.3.1 A. Wojta de los Departamentos de Ingeniería Agrícola y Suelos de la Universidad de Wisconsin, determinó el tiro, en unidades de 10 lbs, necesario para arar un suelo fangoso donde el tractor iba a 2 millas/h. Los datos que siguen corresponden a cada una de tres posiciones de la rueda derecha. Dos tamaños de llanta de la rueda izquierda y dos tipos de enganche dieron lugar a un factorial  $2 \times 2$  para cada una de las tres posiciones de la rueda derecha. (El experimento original fue un factorial  $3 \times 2 \times 2$  dentro de un diseño completamente aleatorio con 3 observaciones por tratamiento).

		Rueda derecha en línea	Rueda derecha en ángulo 1.29°	Rueda derecha		
		Llanta rueda izquierda	Llanta rueda izquierda	Llanta rueda izquierda		
		Enganche, pul 6.50 × 16	7.50 × 16	6.50 × 16	7.50 × 16	6.50 × 16
2		76, 76, 81	76, 76, 75	65, 78, 68	42, 56, 35	74, 85, 79
4		69, 74, 79	50, 63, 62	77, 60, 82	48, 50, 48	60, 88, 88
						74, 69, 74
						66, 62, 57

Considerar estos datos como provenientes de tres experimentos diferentes. Analizar cada conjunto de datos por separado. Preparar una tabla, como la 15.3, para calcular los efectos principales, interacciones y sumas de cuadrados.

¿Qué hipótesis se prueban en las varias entradas de la columna *F* en el análisis de la varianza? Interpretar los resultados.

Ejercicio 15.3.2 En el ejercicio 9.3.5 se tienen datos de un experimento factorial  $3 \times 2 \times 2$ . Calcular los efectos de nitrógeno, aireación e interacción nitrógeno × aireación y sumas de cuadrados para cada nivel de salinidad. Usar el término del error experimental de todo el experimento para probar las diversas hipótesis nulas. ¿Cuáles son estas hipótesis nulas?

**Ejercicio 15.3.3** Para los limitados visualmente, la escucha se considera una actividad de lectura llamada "lectura-audición". Rawls (15.2.1) estudió unas técnicas para mejorar la eficiencia en esta habilidad. Sus sujetos fueron estudiantes de la Governor Morehead School, Raleigh, Carolina del Norte, en los cursos 10 a 12. Todos tenían fuertes deficiencias visuales; todos leían en Braille. Sin embargo, algunos de los estudiantes leían de preferencia impresos.

Dos de los tratamientos incluidos en el estudio fueron: (1) Instrucción en técnicas de audición, más práctica de audición de lectura de textos seleccionados; (2) lo mismo que (1) pero con copias de las lecturas seleccionadas en Braille o impresas, a selección del estudiante, que podían seguir al tiempo que escuchaban.

Se usaron varias medidas del rendimiento que suministraron los datos de la prueba posterior. Estos se dan para la exactitud tal como se mide la Prueba de Lectura Oral de Gilmore. Se da un conjunto limitado de datos de tal manera que se puede mantener el balance y para que puedan aplicarse los métodos de este capítulo. Se suministran los datos de la prueba previa para establecer si hay diferencias en los grupos antes del experimento.

	Datos prueba previa		Datos prueba posterior	
	Tratamiento 1	Tratamiento 2	Tratamiento 1	Tratamiento 2
Braille	89, 82, 88, 94	89, 90, 91, 92	87, 86, 94, 96	84, 94, 97, 93
Impresos	71, 88, 96, 96	89, 99, 84, 87	58, 82, 97, 93	96, 97, 75, 77

Analizar los datos de las pruebas previa y posterior por separado. Calcular las sumas de cuadrados del efecto principal y la interacción para cada uno de los conjuntos de datos. ¿Qué hipótesis pueden probarse ahora? ¿Qué conclusiones pueden sacarse de los análisis?

#### 15.4 Factorial de $3 \times 3 \times 2$ ó $3^2 \times 2$ : un ejemplo

Los datos de la tabla 15.5 son los resultados de un experimento en invernadero efectuado por Wagner (15.15) para determinar la tasa de emergencia de semillas de tres especies de leguminosas, tratadas y no tratadas con un fungicida y sembradas en tres tipos de suelos. (Estos datos son de un experimento con un cuarto factor, profundidad de siembra, a tres niveles). El diseño era un diseño de bloques completos al azar. Los datos de la tabla 15.5 corresponden a un factorial  $3 \times 3 \times 2$  ó  $3^2 \times 2$  donde cada dato es una suma extendida a los bloques.

Los datos brutos se han organizado en una tabla de dos factores, similar a la tabla 9.2, con los tratamientos enumerados al lado y los bloques en la parte superior. El análisis de la varianza inicial se efectúa luego como se describe en la sec. 9.3. Además, los totales de los 18 tratamientos se incluyen en la tabla 15.5 y se obtienen varios subtotales. Esta tabla es necesaria para el cálculo de las sumas de cuadrados de los efectos principales y las interacciones.

En los cálculos que siguen  $a$ ,  $b$   $c$  y  $r$  representan el número de niveles de los factores  $A$ ,  $B$  y  $C$  y el número de bloques o replicaciones. Previamente se generalizaron los símbolos descritos. Así, en un experimento de tres factores ( $a_1 b_2 c_1$ ) es el total de las  $r$  observaciones de la combinación de tratamientos con los niveles más bajos de  $A$  y  $C$  y el segundo nivel de  $B$ ; ( $a_1 b_2$ ) es el total de las  $rc$  observaciones hechas en unidades para las cuales estaba  $A$  al nivel más bajo,  $B$  al segundo nivel; y así sucesivamente.

**Tabla 15.5** Número\* de plantas de tres especies,  $A$ , sembradas en tres tipos de suelos,  $B$ , a una profundidad de  $\frac{1}{2}$  pulgada con semillas tratadas y no tratadas con un fungicida,  $C$ .

Species = $A$	Fungicida = $C$			Franco limoso = $b_1$			Arenoso = $b_2$			Arcilloso = $b_3$			Total = $b_1 + b_2 + b_3$		
	Tipo de suelo = $B$														
Alfalfa = $a_1$	No tratadas = $c_1$	266 = $(a_1, b_1, c_1)$	286 = $(a_1, b_2, c_1)$	66 = $(a_1, b_3, c_1)$	618 = $(a_1, c_1)$										
	Tratadas = $c_2$	276 = $(a_1, b_1, c_2)$	271 = $(a_1, b_2, c_2)$	215 = $(a_1, b_3, c_2)$	762 = $(a_1, c_2)$										
	Total = $c_1 + c_2$	542 = $(a_1, b_1)$	557 = $(a_1, b_2)$	281 = $(a_1, b_3)$	1,380 = $(a_1)$										
Trébol rojo = $a_2$	No tratadas = $c_1$	252 = $(a_2, b_1, c_1)$	289 = $(a_2, b_2, c_1)$	167 = $(a_2, b_3, c_1)$	708 = $(a_2, c_1)$										
	Tratadas = $c_2$	275 = $(a_2, b_1, c_2)$	292 = $(a_2, b_2, c_2)$	203 = $(a_2, b_3, c_2)$	770 = $(a_2, c_2)$										
	Total = $c_1 + c_2$	527 = $(a_2, b_1)$	581 = $(a_2, b_2)$	370 = $(a_2, b_3)$	1,478 = $(a_2)$										
Trébol dulce = $a_3$	No tratadas = $c_1$	152 = $(a_3, b_1, c_1)$	197 = $(a_3, b_2, c_1)$	52 = $(a_3, b_3, c_1)$	401 = $(a_3, c_1)$										
	Tratadas = $c_2$	178 = $(a_3, b_1, c_2)$	219 = $(a_3, b_2, c_2)$	121 = $(a_3, b_3, c_2)$	518 = $(a_3, c_2)$										
	Total = $c_1 + c_2$	330 = $(a_3, b_1)$	416 = $(a_3, b_2)$	173 = $(a_3, b_3)$	919 = $(a_3)$										
Total = $a_1 + a_2 + a_3$	No tratadas = $c_1$	670 = $(b_1, c_1)$	772 = $(b_2, c_1)$	285 = $(b_3, c_1)$	1,727 = $(c_1)$										
	Tratadas = $c_2$	729 = $(b_1, c_2)$	782 = $(b_2, c_2)$	539 = $(b_3, c_2)$	2,050 = $(c_2)$										
	Total = $c_1 + c_2$	1,399 = $(b_1)$	1,554 = $(b_2)$	824 = $(b_3)$	3,777 = $G$										

\* Cada valor es un total de tres repeticiones de 100 semillas cada una.

*Paso 1* Calcular

$$\text{Factor de Corrección} = FC = 264,180.17$$

$$\text{SC Total} = 35,597.67$$

$$\text{SC Bloques} = 356.77$$

$$\text{SC Tratamientos} = 32,041.50$$

$$\text{SC Error} = 3,199.40$$

*Paso 2* La suma de cuadrados de tratamientos se partitiona en componentes atribuibles a efectos principales y a interacciones. En la mayoría de los casos, éstas suponen más de un solo grado de libertad. Definiciones tales como las de las ecs. (15.1) a (15.3) no serán aplicables, ni tampoco las fórmulas de cálculo dadas como ecs. (15.5) a (15.7). Hay que usar fórmulas corrientes de sumas de cuadrados. Así,

$$\text{SC}(A) = \frac{\sum_i (a_i)^2}{rbc} - FC \quad (15.8)$$

$$= \frac{1,380^2 + 1,478^2 + 919^2}{3(3)2} - FC = 9,900.11$$

$$\text{SC}(B) = \frac{\sum_j (b_j)^2}{rac} - FC$$

$$= \frac{1,399^2 + 1,554^2 + 824^2}{3(3)2} - FC = 16,436.11$$

$$\text{SC}(C) = \frac{\sum_k (c_k)^2}{rab} - FC = \frac{[(c_2) - (\bar{c}_1)]^2}{2rab}$$

$$= \frac{1,727^2 + 2,050^2}{3(3)3} - FC = \frac{(2,050 - 1,727)^2}{2(3)3(3)} = 1,932.02$$

$$\text{SC}(AB) = \frac{\sum_{i,j} (a_i b_j)^2}{rc} - FC - \text{SC}(A) - \text{SC}(B) \quad (15.9)$$

$$= \frac{542^2 + \dots + 173^2}{3(2)} - FC - (9,900.11 + 16,436.11) = 658.44$$

$$\text{SC}(AC) = \frac{\sum_{i,k} (a_i c_k)^2}{rb} - FC - \text{SC}(A) - \text{SC}(C)$$

$$\begin{aligned}
 &= \frac{618^2 + \dots + 518^2}{3(3)} - FC - (9,900.11 + 1,932.02) = 194.03 \\
 SC(BC) &= \frac{\sum_{j,k} (b_j c_k)^2}{ra} - FC - SC(B) - SC(C) \\
 &= \frac{670^2 + \dots + 539^2}{3(3)} - FC - (16,436.11 + 1,932.02) = 1,851.14 \\
 SC(ABC) &= \frac{\sum_{i,j,k} (a_i b_j c_k)^2}{r} - FC - SC(A) - SC(B) - SC(C) - SC(AB) \\
 &\quad - SC(AC) - SC(BC) \\
 &= \frac{266^2 + \dots + 121^2}{3} - FC - (9,900.11 + 16,436.11 + 1,932.02 \\
 &\quad + 658.44 + 194.03 + 1,851.14) = 1,069.65
 \end{aligned} \tag{15.10}$$

Obsérvese que la suma de cuadrados para las interacciones  $AB$ ,  $AC$  y  $BC$  no son más que residuos de tablas de dos factores. Análogamente,  $SC(AB)$  es un residuo de una tabla de 3 factores. Se puede decidir construir tablas de dos factores o marcar distintamente, con colores por ejemplo, totales apropiados. Nótese, también, que este ejemplo permite una generalización fácil a un experimento factorial de dimensiones cualesquiera.

Los resultados se presentan en una tabla de análisis de la varianza como la tabla 15.6. Aquí no hemos presentado la suma de cuadrados total para tratamiento, aunque esto hubiere sido muy correcto. Eso se hace a menudo, y entonces la partición en efectos principales e interacciones se inserta debajo de "Tratamientos" en el análisis de la varianza. Ver la tabla 15.4.

A las interacciones en que entran dos factores, se las llama *interacciones de dos factores* o de *primer orden*, por ejemplo,  $AB$ ,  $AC$  y  $BC$ . Las interacciones con 3 factores son *interacciones de tres-factores* o de *segundo orden*.

Tabla 15.6 Análisis de la varianza de los datos de la tabla 15.5

Fuente	gl	SC	Cuadrado medio	F
Bloques	$(r - 1) = 2$	356.77	178.39	1.90
$A =$ especies	$a - 1 = 2$	9,900.11	4,950.06	52.60**
$B =$ tipo de suelo	$b - 1 = 2$	16,436.11	8,218.06	87.33**
$C =$ fungicida	$c - 1 = 1$	1,932.02	1,932.02	20.53**
$AB$	$(a - 1)(b - 1) = 4$	658.44	164.61	1.75
$AC$	$(a - 1)(c - 1) = 2$	194.03	97.02	1.03
$BC$	$(b - 1)(c - 1) = 2$	1,851.14	925.57	9.84**
$ABC$	$(a - 1)(b - 1)(c - 1) = 4$	1,069.65	267.41	2.84*
Error	$(r - 1)(abc - 1) = 34$	3,199.40	94.10	
Total	$abcr - 1 = 53$	35,597.67		

Para estos datos, el análisis de la varianza indica que los tres efectos principales y las interacciones  $BC$  y  $ABC$  son significantes. La interacción significante  $BC$  implica que las diferencias entre las respuestas a  $C$  varían con el nivel de  $B$ , donde las respuestas están medidas en todos los niveles de  $A$ . De otra manera, las diferencias entre las respuestas a los niveles de  $B$  varían para los niveles de  $C$ , donde las respuestas se miden de nuevo como totales o medias sobre todos los niveles de  $A$ . Específicamente, las diferencias en tasas de emergencia, promediadas para todas las especies, entre semillas tratadas y no tratadas no son las mismas para los tres tipos de suelos; o las diferencias entre tasas de emergencia de semillas cultivadas en tres tipos de suelos son las mismas semillas para semillas tratadas y no tratadas.

Algunas veces es difícil interpretar las interacciones de segundo orden o de orden más elevado. La interacción significante  $ABC$  puede considerarse de tres maneras: como una interacción de la interacción  $AB$  con el factor  $C$ , de la interacción  $AC$  con el factor  $B$ , o de la interacción  $BC$  con el factor  $A$ . Aquí la interacción  $BC$  no es coherente para los niveles de  $A$ , y así sucesivamente. La manera de tratar esto dependerá de qué enfoque tiene más sentido y posiblemente de la significancia de las interacciones de dos factores.

Para estos datos, como las interacciones  $BC$  y  $ABC$  son significantes, parece lógico comenzar por el examen de la interacción  $BC$ . Puesto que  $C$  está sólo a dos niveles, comenzemos por considerar los efectos simples de  $C$  a los diferentes niveles de  $B$ . Aparentemente, los efectos simples de  $C$  no son homogéneos para los tres tipos de suelos. Por lo tanto, examinaremos la tabla de dos factores, contenida en la tabla 15.5 y usada en el cálculo de la interacción  $BC$ . Esto se presenta en la tabla 15.7, donde las respuestas al fungicida se comparan para los diferentes tipos de suelos. La diferencia de 254 sobresale inmediatamente. Ahora propondremos una prueba de una comparación sugerida para los datos. Sin embargo, una prueba de significancia ya nos ha alertado sobre la presencia de diferencias reales.

Cada suma de cuadrados tiene un solo grado de libertad y su suma es la igual a la suma para  $C$  y  $BC$ , esto es,  $193.39 + 5.56 + 3,584.22 = 3,783.17 = 1,923.02 + 1,851.14$ .

Las sumas de cuadrados para  $C$  y  $BC$  representan una partición común de una suma de cuadrados de  $n = 3$  observaciones en un término de corrección (el efecto total de  $C$ ) y una varianza o medida de la homogeneidad o heterogeneidad (la interacción  $BC$ ). Puesto que hay una interacción  $BC$  y los efectos simples de  $C$  no son homogéneos, decidimos estudiar la información presentada por las sumas individuales de cuadrados de los efectos simples.

Encontramos que la diferencia de las tasas de emergencia de semillas tratadas y no tratadas, promediadas para todas las especies, no es significante para los suelos franco limoso o arenoso, pero sí lo es para el suelo arcilloso.

Debido a las dificultades en la interpretación de interacciones significantes en el análisis de la varianza, hemos escogido examinar ciertos efectos simples a costa de una pérdida de replicación; cada uno de los efectos simples de  $C$  se mide solo a un nivel de  $B$ .

La interacción  $ABC$  significante implica que la interacción  $BC$  difiere con el nivel de  $A$ . Desde este punto de vista, consideramos a  $ABC$ , ya que  $BC$  es significante. Una vez examinada la interacción  $BC$ , encontramos razonable concluir que la dificultad está ligada a  $C$  en suelo arcilloso y procedemos directamente a un examen de los efectos simples de  $C$  para el suelo tipo arcilloso  $b_3$ , a los diferentes niveles de  $A$ . Una ojeada a las tres tablas

**Tabla 15.7 Examen de la interacción  $BC$  para los datos de la tabla 15.5**

	Fungicida		
	$c_1$	$c_2$	$c_2 - c_1$
Tipo de suelo			
$b_1$	670	729	59
$b_2$	772	782	10
$b_3$	285	539	254
$b_1 + b_2 + b_3$	1,727	2,050	
SC( $C$ dentro de $b_1$ )	$\frac{[(b_1 c_2) - (b_1 c_1)]^2}{2ra}$		
	$= \frac{(729 - 670)^2}{2(3)3} = 193.39\ ns$		
SC( $C$ dentro de $b_2$ )	$\frac{[(b_2 c_2) - (b_2 c_1)]^2}{2ra}$		
	$= \frac{(782 - 772)^2}{2(3)3} = 5.56\ ns$		
SC( $C$ dentro de $b_3$ )	$\frac{[(b_3 c_2) - (b_3 c_1)]^2}{2ra}$		
	$= \frac{(539 - 285)^2}{2(3)3} = 3,584.22^{**}$		

de tipos de suelo para totales de fungicida de la tabla 15.5 parece justificar este enfoque. Así, consideramos la interacción  $AC$  en suelo arcilloso. La tabla apropiada de dos factores y los cálculos correspondientes se presentan en la tabla 15.8.

Algunas de las más importantes conclusiones que se sacan de este experimento son las siguientes. No se encontró diferencia entre las tasas de emergencia de semillas tratadas  $c_2$  y no tratadas  $c_1$  cuando se promediaron para todas las especies en el suelo franco limoso  $b_1$  y arenoso  $b_2$ ; sin embargo, la diferencia fue significativa a favor de las semillas tratadas para el suelo arcilloso. Como la interacción de tres factores fue significante, se hizo otro análisis, el cual indicó que en el suelo arcilloso, las semillas tratadas de alfalfa  $a_1$  y trébol dulce  $a_3$  emergieron mejor que las no tratadas, mientras que no se encontró diferencia para el trébol rojo  $a_2$ . Las diferencias entre tasas de emergencia para semillas tratadas y no tratadas, para cada una de las tres especies, no fueron significantes para los suelos franco limoso y arenoso.

**Ejercicio 15.4.1** Esquematice la interacción de  $BC$  de la tabla 15.7 y las interacciones  $BC$  para cada uno de los niveles de  $A$ . Así se mejora la apreciación de la naturaleza de una interacción de tres factores.

**Tabla 15.8 Examen de una interacción  $AC$ , la de  $b_3$ , para los datos de la tabla 15.5**

Arcilloso = $b_3$			
	$c_1$	$c_2$	$c_2 - c_1$
Alfalfa = $a_1$	66	215	149
Trébol rojo = $a_2$	167	203	36
Trébol dulce = $a_3$	52	121	69

$$\text{SC(C dentro de } a_1 \text{ para } b_3) = \frac{[(a_1 b_3 c_2) - (a_1 b_3 c_1)]^2}{2r}$$

$$= \frac{(215 - 66)^2}{2(3)} = 3,700.17^{**}$$

$$\text{SC(C dentro de } a_2 \text{ para } b_3) = \frac{[(a_2 b_3 c_2) - (a_2 b_3 c_1)]^2}{2r}$$

$$= \frac{(203 - 167)^2}{2(3)} = 216.00 \text{ ns}$$

$$\text{SC(C dentro de } a_3 \text{ para } b_3) = \frac{[(a_3 b_3 c_2) - (a_3 b_3 c_1)]^2}{2r}$$

$$= \frac{(121 - 52)^2}{2(3)} = 793.50^{**}$$

**Ejercicio 15.4.2** Los datos del ejercicio 9.3.5 son de un experimento factorial  $3 \times 2 \times 2$ . Analizar estos datos utilizando esta información. ¿Qué hipótesis se prueban con diferentes pruebas  $F$ ? ¿Cuáles son las conclusiones importantes?

En el ejercicio 15.3.2 se examinaron tres interacciones  $2 \times 2$ . La información sobre éstas debe aparecer en el actual análisis. ¿Dónde aparece? ¿Qué hecho nuevo agrega el presente análisis al conocimiento de la naturaleza de la interacción nitrógeno  $\times$  aireación?

**Ejercicio 15.4.3** A la luz de los nuevos conocimientos de los efectos principales e interacciones reconsiderese el análisis de la varianza llevado a cabo en el ejercicio 9.8.1. ¿A qué conclusiones llevan las diferentes pruebas  $F$ ?

**Ejercicio 15.4.4** Analizar los datos del ejercicio 15.3.1 como si provinieran de un solo diseño completamente aleatorizado como era el caso. ¿Se encuentra el error experimental mediante la combinación de los términos del error original? ¿Qué ha resultado de las tres interacciones calculadas anteriormente?

**Ejercicio 15.4.5** Particionar las sumas de cuadrados de tratamientos de la tabla 7.9 en sumas de cuadrados para efectos principales e interacciones. ¿Qué nuevas hipótesis se pueden probar?

## 15.5 Modelos lineales para experimentos factoriales

En el texto se han expuesto modelos lineales. Si consideramos el diseño en bloques completos al azar como experimento de dos factores, entonces ya se ha expuesto el modelo

lineal para algunos experimentos factoriales. En realidad, estas dos situaciones son diferentes, cosa que es clara por la aleatorización. Anderson (15.23) y Anderson y McLean (15.24) consideran estas diferencias cuidadosamente.

Dos clases de problemas fundamentalmente diferentes se han originado. Los problemas de Clase I, que suponen el *modelo de efectos fijos, Modelo I*. Los problemas de Clase II suponen el *modelo de efectos aleatorios, Modelo II*.

Muchos conjuntos de datos presentan una mezcla de las dos clases de problemas, así que tenemos el *modelo mixto*. Pero aún son posibles otros modelos. En todo caso, los cálculos serán los mismos sea cual fuere el modelo, aunque variará la elección de los términos de error y el tipo de inferencia. Ver también Scheffé (15.10) y Wilk y Kempthorne (15.16).

Los valores promedios o esperados de los cuadrados medios en un experimento de tres factores, en un diseño de bloques completos aleatorizados, se presentan en la tabla 15.9. Las letras mayúsculas se refieren a efectos, esto es, efectos principales o interacciones; las letras minúsculas se refieren a los números de los niveles de los efectos designados por las mayúsculas correspondientes. La varianza del error es  $\sigma^2$ ; otras varianzas tienen subíndices que las relacionan con los efectos correspondientes. Las letras griegas se refieren a las componentes individuales usadas para describir una observación particular; éstas se usan en valores esperados en los que los efectos son fijos, ya que los del experimento constituyen la población completa. Los subíndices en las letras griegas o combinaciones de letras se omiten por comodidad en la presentación de la tabla. Pero la descripción matemática completa de una observación es como sigue

$$Y_{ijkl} = \mu + \rho_i + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} + (\alpha\beta\gamma)_{jkl} + \varepsilon_{ijkl}$$

Se ve fácilmente por la tabla 15.9 que, para el modelo fijo, la varianza del error es un término apropiado para probar la hipótesis acerca de una fuente de variación en el análisis de la varianza. Sin embargo, como hemos visto en los ejemplos, una interacción significante puede llevarnos a perder interés en pruebas de hipótesis relacionadas con efectos principales, e interesarnos en otras pruebas tales como las de efectos simples. Así, en un modelo de efectos fijos, hemos seleccionado todas las combinaciones de tratamientos y estamos interesados en éstas solamente. En consecuencia, nos interesaremos casi con certeza en efectos simples cuando hay interacción. Tal cambio de acento conduce más probablemente a una satisfactoria interpretación de los datos. Sin embargo, para un modelo mixto, podemos no estar interesados en efectos simples para un efecto fijo, ya que se medirán a niveles seleccionados al azar de otro factor y por tanto serán valores de una variable aleatoria.

Para el modelo aleatorio, la elección de un término de error adecuado, cuando todas las fuentes de variación son reales, es más difícil cuando se han de probar hipótesis relacionadas con efectos principales. La tabla 15.9 indica que ese error es adecuado para probar la interacción de tres factores; si  $\sigma_{\alpha\beta\gamma}^2$  es real, el cuadrado medio de ABC es el apropiado para probar las interacciones de dos factores. Para las pruebas de efectos principales, es necesario usar alguna combinación de cuadrados medios. Por ejemplo, para probar  $H_0: \sigma^2 = 0$ , podemos usar un criterio de prueba análogo al F con el CM (C) como numerador y  $CM(AC) + CM(BC) - CM(ABC)$  como denominador. Puesto que la presen-

Tabla 15.9 Valores esperados de los cuadrados medios para experimentos factoriales: experimentos de tres-factores

Fuente	gl	Valor esperado del cuadrado medio	
		Modelo I (fijo)	Modelo II(aleatorio)
Bloques	$r - 1$	$\sigma^2 + abc \sum p^2/(r - 1)$	$\sigma^2 + ab\sigma_p^2$
$A$	$a - 1$	$\sigma^2 + rbc \sum \alpha^2/(a - 1)$	$\sigma^2 + r\sigma_{\alpha p}^2 + r\sigma_{\alpha s}^2 + r\sigma_{\beta p}^2 + r\sigma_{\beta s}^2$
$B$	$b - 1$	$\sigma^2 + rac \sum \beta^2/(b - 1)$	$\sigma^2 + r\sigma_{\beta p}^2 + r\sigma_{\alpha p}^2 + r\sigma_{\beta s}^2 + r\sigma_{\alpha s}^2$
$C$	$c - 1$	$\sigma^2 + rab \sum \gamma^2/(c - 1)$	$\sigma^2 + r\sigma_{\gamma p}^2 + r\sigma_{\alpha p}^2 + r\sigma_{\beta p}^2 + r\sigma_{\gamma s}^2$
$AB$	$(a - 1)(b - 1)$	$\sigma^2 + rc \sum (\alpha\beta)^2/(a - 1)(b - 1)$	$\sigma^2 + r\sigma_{\alpha\beta p}^2 + r\sigma_{\alpha\beta s}^2$
$AC$	$(a - 1)(c - 1)$	$\sigma^2 + rb \sum (\alpha\gamma)^2/(a - 1)(c - 1)$	$\sigma^2 + r\sigma_{\alpha\gamma p}^2 + r\sigma_{\alpha\gamma s}^2$
$BC$	$(b - 1)(c - 1)$	$\sigma^2 + ra \sum (\beta\gamma)^2/(b - 1)(c - 1)$	$\sigma^2 + r\sigma_{\beta\gamma p}^2 + r\sigma_{\beta\gamma s}^2$
$ABC$	$(a - 1)(b - 1)(c - 1)$	$\sigma^2 + r \sum (\alpha\beta\gamma)^2/(a - 1)(b - 1)(c - 1)$	$\sigma^2 + r\sigma_{\alpha\beta\gamma p}^2$
Error	$(r - 1)(abc - 1)$	$\sigma^2$	$\sigma^2$

Modelo Mixto:  $A$  y  $B$  fijos,  $C$  aleatorio

Bloques	$\sigma^2 + ab\sigma_p^2$
$A$	$\sigma^2 + rb \frac{a}{a - 1} \sigma_{\alpha p}^2 + rbc \sum \alpha^2/(a - 1)$
$B$	$\sigma^2 + ra \frac{b}{b - 1} \sigma_{\beta p}^2 + rac \sum \beta^2/(b - 1)$
$C$	$\sigma^2 + rab\sigma_{\gamma p}^2$
$AB$	$\sigma^2 + r \frac{a}{a - 1} \frac{b}{b - 1} \sigma_{\alpha\beta p}^2 + rc \sum (\alpha\beta)^2/(a - 1)(b - 1)$
$AC$	$\sigma^2 + rb \frac{a}{a - 1} \sigma_{\alpha\gamma p}^2$
$BC$	$\sigma^2 + ra \frac{b}{b - 1} \sigma_{\beta\gamma p}^2$
$ABC$	$\sigma^2 + r \frac{a}{a - 1} \frac{b}{b - 1} \sigma_{\alpha\beta\gamma p}^2$
Error	$\sigma^2$

cia de signos negativos en tales funciones lineales puede conducir a dificultades, se ha propuesto que  $CM(C) + CM(AB)$  se use como numerador y  $CM(AC) + CM(BC)$  como denominador. Satterthwaite (15.9) sugirió este último criterio que también ha sido considerado por Cochran (15.3). Los criterios de prueba cuando se han restado cuadrados medios ya se han estudiado. Se recomienda al lector ver Gaylor y Hopper (15.20).

En forma más general, Satterthwaite (15.9) nos ha dado las ecs. (15.11) y (15.12), donde cada  $M_i$  representa un cuadrado medio y un  $M_i$  no debe aparecer tanto en el numerador como denominador de  $F$  o  $F'$ , como a menudo se la designa. Tales razones se llaman *cuasi razones F*. El criterio de prueba se distribuye aproximadamente como  $F$ .

$$F_{p, q} = \frac{M_p + \dots + M_s}{M_u + \dots + M_v} \quad (15.11)$$

donde

$$p = \frac{(M_r + \cdots + M_s)^2}{M_r^2/f_r + \cdots + M_s^2/f_s} \quad (15.12)$$

$f_i$  corresponde a los grados de libertad de  $M_i$ ;  $q$  se define en forma similar a  $p$ . Aquí  $p$  y  $q$  son grados de libertad "efectivos".

Este criterio de prueba es el que se ha usado en la sec. 5.9 para probar dos medias cuando las varianzas son desiguales.

Consideremos cómo se obtienen los valores esperados para el modelo aleatorio. Crump (15.4) da una regla cómoda para hacerlo y Schultz (15.11) da reglas para situaciones más generales, comprendidos modelos aleatorios.

**Regla 1** Para el modelo aleatorio, cualquier efecto tendrá, en el valor esperado de su cuadrado medio, una combinación lineal de  $\sigma^2$  y aquellas varianzas, y no otras, cuyos subíndices contienen todas las letras del efecto. Por ejemplo, el valor esperado del cuadrado medio para  $AC$  incluirá  $\sigma^2$ ,  $\sigma_{\alpha\beta\gamma}^2$ , y  $\sigma_{\alpha\gamma}^2$ . Los coeficientes de las varianzas son: 1 para  $\sigma^2$  y para cualquier otra varianza, el producto del número de replicaciones (bloques en la tabla 15.9) y todas las letras minúsculas que correspondan a las mayúsculas que no están en el subíndice. Por ejemplo, para  $AC$  tenemos  $\sigma^2$ ,  $r\sigma_{\alpha\beta}^2$ , y  $rb\sigma_{\alpha\gamma}^2$ . Nótese que el conjunto completo de letras usadas para los factores aparece con cada varianza diferente de  $\sigma^2$ , bien sea con coeficientes (minúscula) o como subíndice (minúscula griega correspondiente).

**Regla 2** Para el modelo mixto, se empieza con el cálculo de valores esperados de cuadrados medios para el modelo aleatorio y luego se eliminan ciertas varianzas y se remplazan otras por cuadrados medios de efectos de población. La componente con el mismo subíndice del efecto siempre está presente en el cuadrado medio del efecto. En el cuadrado medio para cualquier efecto, considérese cualquier otra componente. En un subíndice, ignórese toda letra usada para denominar el efecto; si otra letra de un subíndice corresponde a un efecto fijo, elimínese la componente de varianza. Por ejemplo, en la tabla 15.9 para el modelo mixto y frente a  $A$ , tenemos que considerar  $\sigma_{\alpha\beta\gamma}^2$ ,  $\sigma_{\alpha\beta}^2$ , y  $\sigma_{\alpha\gamma}^2$ ; en cada subíndice ignórese  $A$ ; para  $\sigma_{\alpha\beta}^2$ , y  $\sigma_{\alpha\beta}^2$ ,  $B$  es fijo, así que ambas varianzas se eliminan; para  $\sigma_{\alpha\gamma}^2$ ,  $C$  es aleatoria, así que  $\sigma_{\alpha\gamma}^2$ , no se elimina; finalmente, como  $A$  es fijo, se reemplaza  $\sigma_{\alpha}^2$  por  $\sum \alpha^2/(a - 1)$ . Por otra parte, para  $AC$ , examinamos solamente a  $B$ ;  $B$  es fijo por tanto eliminamos  $\sigma_{\alpha\beta\gamma}^2$ ; como  $C$  es aleatorio,  $AC$  también es aleatorio y  $\sigma_{\alpha\gamma}^2$  se deja como varianza.

**Regla 3** (Sólo para el modelo mixto). Luego de aplicar la regla 2, si alguna varianza dejada en un valor esperado tiene en su subíndice una o más letras correspondientes a efectos fijos, entonces el coeficiente de la varianza requiere un factor para cada efecto fijo. El factor es la razón del número de niveles del efecto fijo al número de niveles menos uno. Por ejemplo, en la tabla 15.9 para el modelo mixto y frente a  $A$ ,  $\sigma_{\alpha\gamma}^2$  tiene  $A$  como efecto fijo. En consecuencia, el coeficiente requiere del factor  $a/(a - 1)$ . [La mayoría de los investigadores no usan ese factor; por ejemplo Schultz (15.11)].

Una palabra explicativa puede ayudar a recordar estas reglas, especialmente cuando se aplican a interacciones. Considérese un conjunto de  $(\alpha\gamma)$ ' en un experimento

$$\begin{matrix} (\alpha\gamma)_{11} & (\alpha\gamma)_{12} & \cdots & (\alpha\gamma)_{1c} \\ (\alpha\gamma)_{21} & (\alpha\gamma)_{22} & \cdots & (\alpha\gamma)_{2c} \\ \cdots & \cdots & \cdots & \cdots \\ (\alpha\gamma)_{a1} & (\alpha\gamma)_{a2} & \cdots & (\alpha\gamma)_{ac} \end{matrix}$$

Para el modelo fijo, ésta es la población de interés y

$$\sum_i (\alpha\gamma)_{ij} = 0 \quad \text{para todo } j$$

y

$$\sum_j (\alpha\gamma)_{ij} = 0 \quad \text{para todo } i$$

O sea que la suma de filas y columnas es igual a cero.

Considérese  $CM(A)$ . Los cálculos requieren los totales de  $A$ . En esos totales, en la tabla de arriba se suma para todos los niveles de  $C$  para cada nivel de  $A$ , estas sumas son iguales a cero. Por tanto,  $E[CM(A)]$  no contiene ningún componente  $(\alpha\gamma)$ .

Para el modelo aleatorio, considérese una tabla básica de  $\sigma_{xy}^2$ , que es una tabla infinitamente grande de dos factores. Aquí, la media o valor esperado para cualquier fila o columna es cero. Esta tabla se muestra por selección aleatoria de niveles de  $A$  y  $C$ . Los elementos en las intersecciones de estas filas y columnas son los  $\sigma_{xy}^2$  del experimento. La suma de los elementos en cualquier fila o columna de esta tabla es un valor aleatorio distribuido en torno a cero.

Considérese  $CM(A)$ . Ahora cuando sumamos los  $(\alpha\gamma)$  muestreados sobre los niveles de  $C$ , tienen una suma que varía de un experimento a otro en muestreo repetido. En consecuencia,  $E[CM(A)]$  tendrá un término de  $\sigma_{\alpha\gamma}^2$ .

Para el modelo mixto, con  $A$  fijo y  $C$  aleatorio, una tabla básica  $(\alpha\gamma)$  sólo tendrá  $a$  filas pero un número infinito de columnas. Aquí, la suma de los  $(\alpha\gamma)$  sobre los  $a$  elementos es igual a cero en cualquier columna, mientras que la media poblacional o valor esperado para cualquier fila es igual a cero. Los  $(\alpha\gamma)$  en un experimento se obtienen tomando una muestra aleatoria de  $c$  columnas. La suma para cualquier columna es cero; para una fila, la suma es la suma de una muestra aleatoria de  $c$  elementos y así, pues, sólo se distribuye en torno a cero.

Considérese  $CM(A)$ . Como es sobre niveles de  $C$ , una suma de  $A$  es un valor aleatorio con respecto a las componentes  $(\alpha\gamma)$ . Así,  $\sigma_{\alpha\gamma}^2$  es una componente de  $E[CM(A)]$ . Considérese  $CM(C)$ . Ahora la suma se hace sobre los  $a$  niveles de  $A$  y los  $(\alpha\gamma)$  suman cero. Así ningún  $\sigma_{\alpha\gamma}^2$  aparecerá en  $E[CM(A)]$ .

Finalmente, considérense los totales  $A$  necesarios en la fórmula de cálculo para  $SC(A)$ . Expresada en términos del modelo, ésta incluye la suma de  $a$  totales  $(\alpha\gamma)$  al cuadrado. Entonces el valor esperado de  $SC(A)$  debe tener un término que es un múltiplo de  $a\sigma_{\alpha\gamma}^2$ .

Para obtener  $CM(A)$ , dividimos por los grados de libertad =  $a - 1$ . En consecuencia,  $a/(a - 1)$  debe hacer parte del coeficiente.

En general, cuando una componente del modelo no suma cero en los totales usados para calcular el cuadrado medio de una fuente de variación, el coeficiente de la varianza correspondiente a la componente del modelo tendrá un multiplicador como  $a/(a - 1)$  para cada efecto fijo nombrado en su subíndice, siempre que exista también un efecto aleatorio nombrado allí.

Las reglas dadas en esta sección también son aplicables cuando un experimento factorial incluye muestreo, esto es, cuando tenemos tanto un experimento como un factorial uno jerarquizado o muestral. La tabla 15.10 presenta un ejemplo. Una nueva notación se ha introducido para tratar la situación nueva: un subíndice de una varianza puede contener letras entre y fuera del paréntesis; letras dentro del paréntesis indican la posición en la jerarquía en la cual la componente aparece. Por ejemplo,  $\sigma_{\delta(\gamma)(\alpha\beta)}^2$  es la varianza de  $D$  dentro de  $C$ , dentro de  $AB$ . Las letras entre paréntesis no entran en la aplicación de la regla 2, que se aplica a eliminaciones.

Para ilustrar el uso de las reglas con una clasificación jerarquizada, considérese el valor esperado del cuadrado medio para  $A$  en la tabla 15.10 para el modelo aleatorio. Un total de  $A$  incluye variación debida a submuestras y así tiene la componente  $\sigma_{\delta(\gamma)(\alpha\beta)}^2$ , variación debida a las muestras y así tiene  $\sigma_{\alpha\beta}^2$ , variación debida a  $AB$  y así tiene  $\sigma_{\alpha\beta}^2$ , variación debida a  $A$  y, por tanto, contiene a  $\sigma_x^2$ , pero no hay variación debida a  $B$ , pues cada nivel de  $B$  aparece en cada total de  $A$ . El coeficiente apropiado de una  $\sigma^2$  está compuesto de las letras (minúsculas) que no aparecen en el subíndice de la componente. Así, el valor esperado es  $\sigma_{\delta(\gamma)(\alpha\beta)}^2 + d\sigma_{\gamma(\alpha\beta)}^2 + cd\sigma_{\alpha\beta}^2 + bcd\sigma_x^2$ . Hasta aquí, sólo se ha usado la regla uno.

Si el modelo exige que sólo  $A$  sea fijo, entonces todas las otras letras se refieren a efectos aleatorios, así que ninguna componente en el valor esperado del cuadrado medio de  $A$  será eliminada y el único cambio consistirá en el remplazo de  $\sigma_x^2$  por  $\sum \alpha^2/(a - 1)$ .

Si el modelo exige que  $A$  y  $B$  sean fijos, entonces en el valor esperado del cuadrado medio de  $A$ , eliminamos  $\sigma_{\alpha\beta}^2$ , dado que  $B$  se refiere a un efecto y no está entre paréntesis. Así obtenemos el valor dado en la tabla 15.10.

Si el modelo exige que  $A$  sea fijo y  $B$  aleatorio, entonces habría una  $\sigma_{\alpha\beta}^2$  en el valor esperado para  $A$ , ya que  $B$  es aleatorio y su contraparte griega aparece en el subíndice  $\alpha\beta$ . La regla 3 también se aplica aquí.

En ciertos campos de investigación, números iguales de subclases son la excepción más que la regla. Los métodos para obtener valores promedio de cuadrados medios en tales casos los han dado Henderson (15.8), Searle (15.22) y otros.

**Ejercicio 15.5.1** Decidir sobre un modelo apropiado para los datos de las tablas 7.8, 15.2 y 15.5 y el ejercicio 15.4.2. Elaborar una tabla de los valores esperados de los cuadrados medios de acuerdo con sus modelos.

**Ejercicio 15.5.2** En un experimento factorial  $2 \times 2$ , sea  $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$ . La descripción de una observación. En términos de esta ecuación, escribir la suma de cada combinación de tratamiento. Continuar y escribir  $A$ ,  $B$  y  $AB$  tal como se define en las ecs. (15.1) a (15.3). ¿Cuál es el efecto principal  $A$  en términos de las contribuciones de tratamientos? ¿Del efecto principal  $B$ ? ¿De la interacción  $AB$ ?

**Tabla 15.10 Valores esperados de los cuadrados medios para un experimento factorial con submuestreo**

Fuente	gl	Valor esperado del cuadrado medio, modelo aleatorio
A	a - 1	$\sigma_{\delta(\gamma_{AB})}^2 + d\sigma_{\gamma_{(AB)}}^2 + cd\sigma_{z_B}^2 + bcd\sigma_z^2$
B	b - 1	$\sigma_{\delta(\gamma_{AB})}^2 + d\sigma_{\gamma_{(AB)}}^2 + cd\sigma_{z_B}^2 + acd\sigma_z^2$
AB	(a - 1)(b - 1)	$\sigma_{\delta(\gamma_{AB})}^2 + d\sigma_{\gamma_{(AB)}}^2 + cd\sigma_{z_B}^2$
C en AB	(c - 1)ab	$\sigma_{\delta(\gamma_{AB})}^2 + d\sigma_{\gamma_{(AB)}}^2$
D en C en AB	(d - 1)abc	$\sigma_{\delta(\gamma_{AB})}^2$

Modelo mixto		
Fuente	A fijo	A y B fijos
A	$\sigma_{\delta(\gamma_{AB})}^2 + d\sigma_{\gamma_{(AB)}}^2 + cd \frac{a}{a-1} \sigma_{z_B}^2 + bcd \sum z^2/(a-1) + bcd \sum z^2/(a-1)$	$\sigma_{\delta(\gamma_{AB})}^2 + d\sigma_{\gamma_{(AB)}}^2 + bcd \sum z^2/(a-1)$
B	$\sigma_{\delta(\gamma_{AB})}^2 + d\sigma_{\gamma_{(AB)}}^2 + acd\sigma_{\beta}^2$	$\sigma_{\delta(\gamma_{AB})}^2 + d\sigma_{\gamma_{(AB)}}^2 + acd \sum \beta^2/(b-1)$
AB	$\sigma_{\delta(\gamma_{AB})}^2 + d\sigma_{\gamma_{(AB)}}^2 + cd \frac{a}{a-1} \sigma_{z_B}^2 + cd \sum (\alpha\beta)^2/(a-1)(b-1)$	$\sigma_{\delta(\gamma_{AB})}^2 + d\sigma_{\gamma_{(AB)}}^2$
C en AB	$\sigma_{\delta(\gamma_{AB})}^2 + d\sigma_{\gamma_{(AB)}}^2$	$\sigma_{\delta(\gamma_{AB})}^2 + d\sigma_{\gamma_{(AB)}}^2$
D en C en AB	$\sigma_{\delta(\gamma_{AB})}^2$	$\sigma_{\delta(\gamma_{AB})}^2$

Por ejemplo, A puede referirse al tratamiento y B al observador. Se requiere que cada observador haga una observación en cada una de un número de veces C. (No debe haber proceso de aprendizaje o tendencia en el tiempo que implique que las primeras observaciones son más parecidas que cualquier otro conjunto, por ejemplo, de lo que sería un conjunto en que entraran varias veces). Finalmente, D puede implicar un conjunto de submuestras observadas dentro de cada tiempo.

### 15.6 Clasificaciones de $n$ vías y experimentos factoriales; superficies de respuesta

La tabla 15.5 contiene las sumas clasificadas en un sistema de tres vías. Ellas proveen todo el material necesario para los cálculos de sumas de cuadrados, para efectos principales e interacciones (Tabla 15.6). Para calcular las sumas de cuadrados de bloques y error, se necesitan las observaciones individuales.

En general, son comunes las clasificaciones de datos de  $n$  vías, no necesariamente sumas como en la tabla 15.5. Vemos ahora que pueden analizarse, desde un punto de vista puramente computacional, tal como analizamos los totales de los tratamientos de la tabla 15.5. El problema de pruebas de significancia es otro aspecto. Por ejemplo, un investigador de tasas de emergencia de semillas que debe trabajar en el campo, no puede aleatorizar todas las combinaciones de un experimento  $3 \times 3 \times 2$ , sino sólo  $3 \times 2$  de ellos; sus suelos se encontrarán en diferentes localidades. Si considera los suelos como bloques de modo que cada combinación de tratamientos aparezca solo una vez en cada tipo de suelo, entonces no hay replicación de tipos de suelo y no puede probar ninguna hipótesis respecto a ellos.

Supóngase que se propone el siguiente análisis

Fuente de variación	gl
Bloques (tipos de suelos, $B$ )	2
Tratamientos	5
Especies, $A$	2
Fungicidas, $C$	1
Especies $\times$ fungicidas, $AC$	2
Residuo (= error)	10
Total	17

Ya hemos visto que los 10 grados de libertad para el residuo son los asociados con las interacciones  $AB$ ,  $BC$  y  $ABC$  con 4, 2 y 4 grados de libertad; también que  $AB$  no es significante,  $BC$  es altamente significante, y  $ABC$  es significante. En otras palabras, la varianza residual o de error sería un promedio de componentes no homogéneas y no verdaderamente apropiada para probar hipótesis respecto a  $A$ ,  $C$  y  $AC$ . Entonces serían dudosas las conclusiones obtenidas en el análisis anterior; se necesitaría un diseño diferente.

Esto ilustra que la elección de considerar datos en una clasificación de  $n$  vías como un experimento de  $n$  factores con un bloque o como un experimento de bloques completos al azar con un número menor que  $n$  factores no siempre es tan sencillo como mirar el esquema de aleatorización. Esencialmente, es un problema de reconocer fuentes potenciales de variación e incluirlas en el modelo. En muchas situaciones, tales datos se analizan empleando el enfoque factorial, ya que, en el peor de los casos, la partición de la suma de cuadrados total resulta excesiva y, en consecuencia, algo falta de sentido. También es a expensas de grados de libertad en el error, como se llega a una estimación menos precisa de la varianza del error. Aquí, es posible usar la interacción de  $n$  factores como término de error. Aquí, es posible usar la interacción de  $n$  factores como término de error, especialmente si parece no tener interpretación con sentido. En cualquier caso, no puede probarse su significancia. En el experimento de especies-suelos-fungicidas, la interacción de tres factores tiene una clara explicación.

A menudo, clasificaciones y experimentos factoriales de  $n$  vías, en los cuales los niveles de un factor se refieren a cantidades medidas de un tratamiento tal como fertilizantes, insecticidas, temperatura de cocción o componente dietético, pueden considerarse como experimentos planeados para determinar la naturaleza a una *superficie de respuesta*. En particular, se puede estar buscando un valor igual o cercano a un máximo, ya que la mayoría de las respuestas no son puramente lineales. Los efectos principales las interacciones pueden interpretarse fácilmente en términos de tales superficies. Gran parte del interés actual en superficies de respuesta se centra en la investigación de Box y Hunter (15.2) y otros.

Considérese la fig. 15.3 que representa los resultados de un experimento factorial  $4 \times 3$ . Los puntos representados corresponden a las medias de 12 combinaciones de tratamientos y la superficie indicada tiene el propósito de ajustar estos puntos en una forma razonable. Cuatro rectas indican las respuestas a tres cantidades de  $B$  a cuatro niveles de  $A$ ; éstos son comparables a efectos simples. Aquí se han indicado con rectas pero con

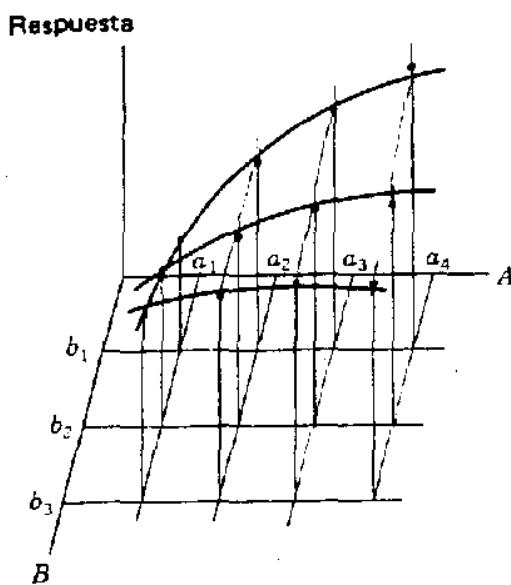


Figura 15.3 Posible superficie de respuesta para un experimento de dos factores.

pendientes que varían de acuerdo con el nivel de  $A$ . Parecería que la respuesta de  $B$  fuese real y más que todo lineal, y que esta respuesta lineal no es homogénea para los diferentes niveles de  $A$ ; las pendientes de la recta de regresión varían. Desde el punto de vista del cálculo se puede examinar la regresión lineal de la respuesta  $B$  o la *componente lineal* de  $B$  a cada nivel de  $A$  y luego probar la hipótesis de homogeneidad. Lo último será una prueba de la interacción entre la componente lineal de  $B$  y el factor  $A$ . Tal interacción parecería existir para la superficie en la fig. 15.3. Si la primera de las cuatro rectas tendiera hacia arriba, la cuarta hacia abajo, y las otras dos fueran intermedias, entonces podríamos encontrar que el efecto principal de  $B$  no mostraría significancia ya que se basa en promedios de  $B$  a todos los niveles de  $A$ ; a la vez la interacción de la componente lineal de  $B$  y el factor  $A$  podría ser significativo, ya que una interacción mide homogeneidad de respuesta.

El factor  $A$  muestra una tendencia hacia la curvilinealidad, hecho que puede detectar una prueba de significancia. Como las curvaturas son bastante leves, la componente lineal de  $A$  probablemente sería significante también para una respuesta como la de la fig. 15.3. Ninguna de esas componentes parece no estar afectada por el nivel de  $B$ , así que podemos esperar que las interacciones de la componente lineal de  $A$  con  $B$  y de la componente curvilinea de  $A$  con  $B$  sean significante, ya que ninguna componente parece ser homogénea para los niveles de  $B$ . Es posible hacer más particiones de los seis grados de libertad y sumas de cuadrados para esta interacción; se pueden obtener, probar e interpretar seis grados de libertad independientes y sumas de cuadrados.

Como un conjunto de comparaciones con sentido ha de suponer regresión, la siguiente sección trata de las componentes de regresión de una suma de cuadrados de tratamientos y su homogeneidad. En secciones posteriores se presentan algunas ilustraciones de superficies de respuesta.

### 15.7 Grados de libertad individuales; tratamientos igualmente espaciados

Muchos experimentos se planean para determinar la naturaleza de curva o superficie de respuesta donde los niveles de un factor se refieren a cantidades crecientes del factor. El análisis de regresión de los caps. 10 al 13 se puede aplicar en el caso presente. Sin embargo,

cuando se presentan incrementos iguales entre niveles sucesivos de un factor, puede usarse un codificador sencillo para llevar a cabo el análisis con menor esfuerzo. Además, se han desarrollado métodos simples para tratar la regresión polinomial y la homogeneidad de las varias componentes de la regresión. Nos referimos, en particular, al uso de los valores de *polinomios ortogonales* o *coeficientes para comparaciones ortogonales* en regresión. Los polinomios ortogonales son ecuaciones tales que cada uno está asociado con una potencia de la variable independiente, por ejemplo con  $X$ ,  $X^2$  o  $X^3$ , y todos están no correlacionados por pares o son ortogonales. Esto permite un cálculo independiente de cualquier contribución de acuerdo con el grado de la variable independiente y una prueba independiente de la contribución. Cada suma de cuadrados corresponde a la reducción adicional debida al ajuste de una curva con un grado más. O sea que el ajuste es secuencial. En la sec. 19.5 se da una exposición general de la construcción de polinomios ortogonales.

Los polinomios ortogonales para  $X$  igualmente espaciados se definen mediante

$$\begin{aligned}\xi_0 &= 1, \text{ todos los } X; \quad \xi_1 = \frac{X_i - \bar{X}}{d}; \quad \xi_2 = \left\{ \left( \frac{X_i - \bar{X}}{d} \right)^2 - \frac{n^2 - 1}{12} \right\}; \\ &\cdots; \quad \xi_{k+1} = \xi_1 \xi_k - \frac{k^2(n^2 - k^2)}{4(4k^2 - 1)} \xi_{k-1} \quad (15.13)\end{aligned}$$

donde  $d$  es el espaciamiento entre  $X$  consecutivos,  $k$  es el grado del polinomio y  $n$  es el número de niveles del factor.

El primer polinomio es de grado cero y claramente se refiere a la media. El segundo es de primer grado, mide cada  $X$  a partir de la media en unidades del espaciamiento de  $X$ , y se refiere a la regresión lineal; si hay tres  $X$ ,  $\xi_1$ , toma los valores  $-1, 0, +1$ ; si hay cuatro  $X$ ,  $\xi_1$ , toma los valores  $-3/2, -1/2, +1/2, +3/2$ ; y así sucesivamente. Los valores de los polinomios, hallados por substitución de los valores de  $X$ , se multiplican por un coeficiente  $\lambda$  para obtener enteros; luego se tabulan éstos.

La tabla 15.11 contiene los valores de hasta seis tratamientos igualmente espaciados de los polinomios multiplicados por los valores de  $\lambda$ , sumas de cuadrados de esos enteros y valores  $\lambda$ . Los coeficientes y los divisores hasta  $n = 75$  y  $n = 104$  tratamientos los dan Fisher y Yates (15.6) y Anderson y Houseman (15.1), respectivamente. Estos van hasta polinomios de grado quinto. Los polinomios ortogonales no son aplicables a tratamientos desigualmente espaciados. Para este caso, ver la sec. 19.7 y Robson (19.8).

Con tres niveles de un factor, hay dos grados de libertad que pueden particionarse en uno, asociados con la respuesta lineal y otro con la cuadrática o de segundo grado. Como una de segundo grado o mayor siempre pasa por 3 puntos, puede ser más apropiado referirse a la última como asociada a una respuesta no lineal o con *falta de ajuste* a una respuesta lineal. Para cuatro niveles, se dispone de un grado de libertad más para estimar la respuesta cúbica; nuevamente podría ser mejor referirse a la última como desviaciones de la respuesta cuadrática o *falta de ajuste*; y así sucesivamente. Si se tienen más de cuatro niveles, comúnmente son sólo de interés las respuestas lineal y cuadrática y a veces la cúbica. Puede ser deseable calcular la suma de cuadrados para cada comparación del conjunto con el objeto de revisar el trabajo.

Como las comparaciones individuales son ortogonales, la suma de sus sumas de cuadrados es igual a la suma de cuadrados del factor correspondiente. Cada una de las sumas

**Tabla 15.11 Coeficientes y divisores para comparaciones ortogonales en la regresión: tratamientos igualmente espaciados**

Número de tratamientos	Grado del polinomio	Totales de tratamiento						Divisor = $\sum c_i^2$	$\lambda$
		$T_1$	$T_2$	$T_3$	$T_4$	$T_5$	$T_6$		
2	1	-1	+1					2	2
3	1	-1	0	+1				2	1
	2	+1	-2	+1				6	3
4	1	-3	-1	+1	+3			20	2
	2	+1	-1	-1	+1			4	1
	3	-1	+3	-3	+1			20	10/3
5	1	-2	-1	0	+1	+2		10	1
	2	+2	-1	-2	-1	+2		14	1
	3	-1	+2	0	-2	+1		10	5/6
	4	+1	-4	+6	-4	+1		70	35/12
6	1	-5	-3	-1	+1	+3	+5	70	2
	2	+5	-1	-4	-4	-1	+5	84	3/2
	3	-5	+7	+4	-4	-7	+5	180	5/3
	4	+1	-3	+2	+2	-3	+1	28	7/12
	5	-1	+5	-10	+10	-5	+1	252	21/10

de cuadrados se comprueba mediante el término de error, siendo la hipótesis nula la de que la media de la población para la comparación es cero o que  $\beta$  para el polinomio ortogonal particular es cero. Si solamente el efecto lineal es significante, concluimos que el aumento en la respuesta entre niveles sucesivos del factor es constante dentro de la variación aleatoria del orden del error experimental. La respuesta puede ser positiva o negativa, según que aumente o disminuya con los incrementos adicionales del factor. Un efecto cuadrático significante indica que una parábola se ajusta mejor a los datos, esto es, que explica significativamente más variación entre medias de tratamientos que una recta; es decir, el aumento o disminución por cada incremento adicional no es constante sino que varía progresivamente. Al planear experimentos en los cuales los niveles de uno o más factores suponen aumentos igualmente espaciados, es aconsejable tener el nivel más alto por encima del valor para el cual se espera la máxima respuesta.

Se ilustrará el uso de los polinomios ortogonales con un experimento sobre soya efectuado por Lambert. Se estudió el efecto sobre la producción de semillas del espaciamiento de cinco surcos que se diferencian por incrementos de 6 pulgadas, con soyas Ottawa Mandarin en seis bloques de un diseño de bloques completos al azar. Los datos, análisis de la varianza y aplicaciones de comparaciones de regresión ortogonales se dan en la tabla 15.12.

La última parte de la tabla 15.12 ilustra el uso de polinomios ortogonales para partitionar la suma de cuadrados de tratamientos (espaciamiento entre surcos) en componentes lineal, cuadrática, cúbica y cuártica. En términos del cap. 14, la SC de las columnas de  $SC(X | X^0)$ ,  $SC(X^2 | X^0, X)$ ,  $SC(X^3 | X^0, X, X^2)$  y  $SC(X^4 | X^0, X, X^2, X^3)$ , el procedi-

**Tabla 15.12 Rendimientos de soya Ottawa Mandarin cultivadas en Rosemount, Minnesota, 1951, en bushels por acre**

Bloque	Espaciamientos de los surcos, en					Totales de bloques
	18	24	30	36	42	
1	33.6	31.1	33.0	28.4	31.4	157.5
2	37.1	34.5	29.5	29.9	28.3	159.3
3	34.1	30.5	29.2	31.6	28.9	154.3
4	34.6	32.7	30.7	32.3	28.6	158.9
5	35.4	30.7	30.7	28.1	29.6	154.5
6	36.1	30.3	27.9	26.9	33.4	154.6
Totales de tratamiento	210.9	189.8	181.0	177.2	180.2	939.1
Medias	35.15	31.63	30.17	29.53	30.03	31.30

\* Valor estimado. Ver también los grados de libertad para el error y el total.

#### Análisis de la varianza

Fuente	gl	SC	CM
Bloques	5	5.41	
Espacamiento de surcos	4	125.66	31.42**
Error	19	73.92	3.89
Total	28	204.99	

Participación de la SC de espaciamiento de los surcos mediante el uso de polinomios ortogonales.

Efecto	Espaciamiento de surcos en pulgadas rendimientos en bushels por acre					Q	$r \sum c_i^2$	SC	F
	18	24	30	36	42				
Lineal	-2	-1	0	+1	+2	-74.0	6(10)	91.27	23.46**
Cuadrática	+2	-1	-2	-1	+2	53.2	6(14)	33.69	8.66**
Cúbica	-1	+2	0	-2	+1	-5.5	6(10)	0.50	< 1
Cuártica	+1	-4	+6	-4	+1	9.1	6(70)	0.20	< 1
Total							125.66		

*Fuente:* Datos usados con el permiso de J.W. Lambert, Universidad de Minnesota, St. Paul, Minnesota.

miento es una aplicación particular de comparaciones ortogonales con un solo grado de libertad tal como se vio en la sec. 8.3. Así para la componente lineal, la ec. (8.3) da

$$Q = -2(210.9) - 1(189.8) + 0(181.0) + 1(177.2) + 2(180.2) = -74.0$$

La suma de cuadrados para esta comparación es, según la ec. (8.6),

$$\text{SC}(X | X^0) = \frac{Q^2}{r \sum c_i^2} = \frac{(-74.0)^2}{6[(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2]} = 91.27$$

Los valores de  $\sum c_i^2$  se dan en la tabla 15.11 para cada comparación. Cada suma de cuadrados tiene un grado de libertad, así que también es un cuadrado medio; los valores  $F$  se obtienen dividiendo cada cuadrado medio por el cuadrado medio del error.

El análisis indica efectos lineal y cuadrático altamente significantes para los tratamientos de espaciamiento de surcos. En promedio, el rendimiento disminuye a medida que la distancia entre los surcos aumenta. La componente lineal es la porción de la suma de cuadrados atribuible a la regresión lineal del rendimiento con respecto al espaciamiento. La componente cuadrática mide la mejora adicional debida al ajuste de un polinomio de segundo orden. Indica que la disminución en el rendimiento resulta menor para cada incremento o aumento en el espaciamiento. La relación entre espaciamiento entre surcos y rendimiento promedio en bushels por acre se representa en la fig. 15.4.

Con los resultados anteriores en mente, se podría decidir ajustar una curva de segundo grado a través de los promedios. Esto puede hacerse mediante técnicas de regresión múltiple, pero los polinomios ortogonales ofrecen otra posibilidad cómoda. Necesitamos la ec. (15.14).

$$\hat{Y} = \bar{Y} + b_1 \lambda_1 \xi_1 + b_2 \lambda_2 \xi_2 \quad (15.14)$$

Para  $b_1$  y  $b_2$ , necesitamos las cantidades apropiadas  $\sum (X - \bar{X})(Y - \hat{Y}) / \sum (X - \bar{X})^2$ . Los numeradores son los  $Q$  de la tabla 15.12 y los denominadores las correspondientes ( $r \sum c_i^2$ ).  $\lambda_1$  y  $\lambda_2$  se encuentran en la tabla 15.11 para cinco tratamientos y polinomios de

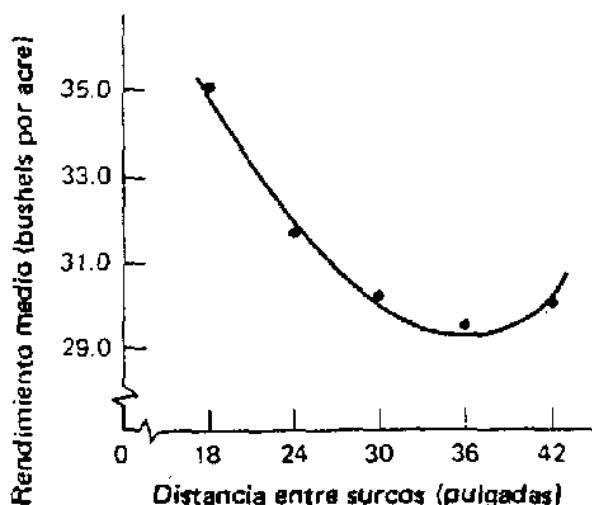


Figura 15.4 Relación entre rendimiento de semillas y espaciamiento entre surcos de soya Ottawa Mandarin.

primero y segundo grados. El espaciamiento,  $d$ , es 6 pulgadas y  $n = 5$ , es el número de niveles. La ec. (15.14) se convierte en

$$\hat{Y} = 31.30 + \frac{(-74.0)}{10(6)}(1)\left(\frac{X - 30}{6}\right) + \frac{53.2}{14(6)}(1)\left[\left(\frac{X - 30}{6}\right)^2 - \frac{5^2 - 1}{12}\right]$$

Esta es una forma cómoda cuando se usa la ecuación para hacer estimaciones. También podemos presentarla en la forma polinomial usual.

$$\hat{Y} = 52.03333 - 1.26111X + .01759X^2$$

En esta última forma, los errores de redondeo pueden ser un problema.

Ahora se muestra otra aplicación de los polinomios ortogonales para los datos de la tabla 15.13. Estos son raíces cuadradas del número de retoños de grama por pie cuadrado, 52 días después de una aspersión con hidrácido maleico. El experimento fue efectuado por Zick (15.19) en Madison, Wisconsin, e intervenían dos factores, o sea, hidrácido maleico en dosis de 0, 4 y 8 lb/acre, llamadas tasas  $R$  y días de retardo en el cultivo después de la aspersión, llamados días  $D$ . Los datos se usan para ilustrar cómo los valores de los polinomios ortogonales pueden usarse en la partición de las sumas de cuadrados para tasas y tasas por días en componentes lineales y cuadráticos. En estos datos, como la suma de cuadrados de la interacción es inferior al error, no hay entonces para que particionarla por razones diferentes a la pura ilustración.

La lógica de base indica calcular efectos simples como observaciones nuevas y considerar luego su suma o media como el efecto principal y su varianza o diferencias como falta de homogeneidad o interacción. El procedimiento de cálculo no siempre parece lógico.

Obsérvese al pie de la tabla 15.13. Para  $D$ , hay tres factores simples, esto es,  $64.9 - 61.5 = 3.4$ ,  $49.3 - 48.7 = 0.6$  y  $39.5 - 37.5 = 2.0$  en términos de totales. Estas son diferencias entre sumas de observaciones; la varianza de cada una es  $4(2)\sigma^2$ . La suma de los tres efectos simples, un múltiplo del efecto principal  $D$ , es 6.0 con varianza  $4(3)2\sigma^2$ . Se indica en forma conveniente como comparación 1. Con base por-observación,  $SC(D) = 6.0^2/4(3)2 = 1.5$ .

La varianza, una medida de la homogeneidad o heterogeneidad, entre estos efectos simples es  $3.4^2 + 0.6^2 + 2.0^2 - 6.0^2/3 = 3.92$ . Con base por observación, esto llega a ser  $3.92/4(2) = 0.49$  con dos grados de libertad. Esta es la interacción tasas  $\times$  días. Recuérdese que esta interacción es el no llevar los efectos simples a ser análogos. Antes de esto, hemos encontrado la interacción como un residuo en una tabla de dos vías, pero cuando un factor sólo tiene dos niveles, una varianza ofrece otra posibilidad.

Si  $W$  representa una diferencia, por ejemplo un efecto simple, entonces hemos usando la ecuación siguiente:  $\sum W_i^2 = (\sum W)^2/n + \sum (W_i - \bar{W})^2$  ecuación que solemos ver con el término  $(\sum W)^2/n$  en el primer miembro de la igualdad y que así iguala una definición y una fórmula de cálculo de una varianza.

Las comparaciones 2 y 3 son esencialmente las de regresiones lineales de la respuesta respecto de la tasa  $R$  dentro de cada una de las demoras de tres y diez días. Llamémoslas  $R_L|(d_1 = 3)$  y  $R_L|(d_2 = 10)$ . Los divisores las convierten en coeficientes de regresión con base por observación, pero para una unidad de 4 lb/acre en vez de una libra. Ahora

**Tabla 15.13 Raíz cuadrada del número de retoños de grama por pie cuadrado 52 días después de aspersión con hidrácido maleico**

Días de retardo en el cultivo	Cantidad de hidrácido maleico, lb, $R$	Bloques				Total
		1	2	3	4	
3	0	15.7	14.6	16.5	14.7	61.5
	4	9.8	14.6	11.9	12.4	48.7
	8	7.9	10.3	9.7	9.6	37.5
10	0	18.0	17.4	15.1	14.4	64.9
	4	13.6	10.6	11.8	13.3	49.3
	8	8.8	8.2	11.3	11.2	39.5
Totales		73.8	75.7	76.3	75.6	301.4

#### Análisis de la varianza

Fuente	gl	SC	Cuadrado medio
Bloques	3	0.58	0.19
Tasas, $R$	2	153.66	76.83**
Días, $D$	1	1.50	1.50
Tasas X Días ( $R \times D$ )	2	0.49	0.25
Error	15	39.38	2.63
Total	23		

No. Comparación	3 días			10 días			Suma	Divisor	SC
	0	4	8	0	4	8			
1    ** $D$	-1	-1	-1	+1	+1	+1	6.0	4(6)	1.5
2    * $R_L  (d_1 = 3)$	-1	0	+1				-24.0	4(2)	72.0
3    * $R_L  (d_2 = 10)$				-1	0	+1	-25.4	4(2)	80.645
4    ** $R_L$	-1	0	+1	-1	0	+1	-49.4	4(4)	152.5225
5 $R_L D$	+1	0	-1	-1	0	+1	-1.4	4(4)	.1225
6    * $R_Q  (d_1 = 3)$	+1	-2	+1				1.6	4(6)	.1067
7    * $R_Q  (d_2 = 10)$				+1	-2	+1	5.8	4(6)	1.4017
8    ** $R_Q$	+1	-2	+1	+1	-2	+1	7.4	4(12)	1.1408
9 $R_Q D$	-1	+2	-1	+1	-2	+1	4.2	4(12)	.3675

\* Efectos simples

\*\* Efectos principales

ya, hemos ajustado un modelo con dos coeficientes de regresión. Se ha visto que son negativos, como era de esperar para un herbicida. Las sumas de cuadrados calculadas son reducciones atribuibles a la regresión dentro de cada valor de  $D$ .

Las comparaciones 2 y 3 pueden considerarse como extensiones de la idea de efectos simples. La combinación da un efecto principal, en este caso, el efecto principal  $R_L$ , la regresión lineal total respecto de las tasas combinadas sobre la demora en días. La comparación 4 es el resultado.

La combinación tácitamente supone homogeneidad. Para medir la falta de homogeneidad de  $R_L$  o la interacción con  $D$ , examíñese la diferencia entre los coeficientes de regresión. Se ha visto que la comparación 5 hace esto. Esta da la interacción *tasas*  $\times$  *días*,  $R_L D$ , y es de observar que los coeficientes de la comparación pueden obtenerse multiplicando los de  $R_L$  por  $D$ . Este es el método preferible.

Las comparaciones 6 y 7 son similares a las 2 y 3, pero son efectos cuadráticos en vez de lineales. Los contrastes o sumas divididas por el número apropiado en la columna del divisor y multiplicado por el valor  $\lambda$  de la tabla 15.11 dan coeficientes de regresión. Estos son para el polinomio cuadrático definido por la ec. (15.13). Las sumas de cuadrados calculadas implican que se han calculado dos coeficientes, uno para cada nivel de  $D$ .

O bien, sumamos la información para obtener el efecto principal  $R_Q$ , comparación 8, y tomamos la diferencia para medir la interacción  $R_Q D$ , comparación 9.

Nótese que la suma de cuadrados del efecto principal  $R_L$  y la interacción  $R_L D$  y las sumas de cuadrados para los efectos simples  $R_L$  son otras particiones de una parte de la suma de cuadrados de tratamientos;  $152.5225 + 0.1225 = 72.0 + 80.645$ . Una relación similar se verifica para el efecto cuadrático:  $1.14083 + 0.3675 = 0.10667 + 1.40167$ .

En general, la partición del efecto principal y de la interacción da más información. Establecemos si hay homogeneidad en los coeficientes de regresión que intervienen y se obtiene un valor combinado. Los coeficientes de comparación indican que toda la información se usa en cada comparación. Si los datos resultan heterogéneos, podemos calcular coeficientes separados usando subconjuntos apropiados de los datos. Este último procedimiento estaría, naturalmente, orientado por los resultados como consecuencia de una prueba de significancia.

Como no hay interacción significante, el interés se centra únicamente en los efectos principales. La suma de cuadrados para el efecto lineal de las tasas es altamente significante mientras que el efecto cuadrático es menor que el error. Los días no parecen ser una fuente de variación. Puede concluirse que la disminución en la raíz cuadrada del número de retoños de grama es la misma para un incremento en el hidrácido maleico, así la demora en el cultivo sea de 3 a 10 días. Si hubiese resultado significante la componente lineal de interacción, ello indicaría que la disminución en la raíz cuadrada del número de retoños por cada incremento en la tasa diferiría para los dos días; o sea que dos coeficientes de regresión de la respuesta con respecto a tasas serían diferentes. Entonces, sería necesario examinar los efectos simples, esto es, las componentes lineales o coeficientes de regresión lineal dentro de cada uno de los dos días.

Con dos o más variables independientes medidas, se puede construir una superficie de respuesta que pase por las medias de los tratamientos. Esta es una extensión de la idea de polinomio, tal como ocurrió con los datos de la soya. También se dispone de las técnicas de regresión múltiples, pero lo ilustraremos con polinomios ortogonales. Así mismo, dejamos de lado el hecho de que el único efecto de tratamientos que es significante es  $R_L$ ,

e incluimos todas las posibles componentes. En este caso, nuestra superficie se ajustará perfectamente a las medias, dentro de los errores de redondeo.

La ecuación deseada es

$$\hat{Y} = \bar{Y} + b_1 \lambda_1 \xi_1 + b_2 \lambda_2 \xi_2 + b_3 \lambda_3 \xi_3 + b_4 \lambda_1 \xi_1 \lambda_3 \xi_3 + b_5 \lambda_2 \xi_2 \lambda_3 \xi_3 \quad (15.15)$$

donde  $\xi_1$  y  $\xi_2$  son polinomios lineales y cuadráticos para  $R$ , y  $\xi_3$  es el polinomio lineal y único para  $D$ . Nótese que  $\xi_1 \xi_3$  y  $\xi_2 \xi_3$  se refieren a las interacciones  $R_L D$  y  $R_Q D$ . Estos son uno de los dos términos cuadráticos y el término cúbico en la ecuación.

A partir de la tabla 15.13,  $b_1 = -49.4/4(14)$ ,  $b_2 = 7.4/12(4)$ ,  $b_3 = 6.0/6(4)$ ,  $b_4 = -1.4/4(4)$  y  $b_5 = 4.2/12(4)$ . Los polinomios ortogonales y los valores de  $\lambda$  se obtienen de la ec. (15.13) y de la tabla 15.11. La ec. (15.15) se torna en

$$\begin{aligned}\hat{Y} = & 12.5583 + \left(\frac{-49.4}{16}\right)1\left(\frac{X_1 - 4}{4}\right) + \frac{7.4}{48}(3)\left(\left(\frac{X_1 - 4}{4}\right)^2 - \frac{3^2 - 1}{12}\right) \\ & + \frac{6.0}{24}(2)\left(\frac{X_2 - 6.5}{7}\right) + \frac{(-1.4)}{16}\left(\frac{X_1 - 4}{4}\right)2\left(\frac{X_2 - 6.5}{7}\right) \\ & + \frac{4.2}{48}(3)\left(\left(\frac{X_1 - 4}{4}\right)^2 - \frac{3^2 - 1}{12}\right)2\left(\frac{X_2 - 6.5}{7}\right)\end{aligned}$$

Esta es una forma cómoda para sustituir los valores ( $X_1$ ,  $X_2$ ). La forma polinomial es como sigue

$$\begin{aligned}\hat{Y} = & 15.0107114 - .71875X_1 - .0015625X_1^2 \\ & + .1214286X_2 - .04375X_1 X_2 + .0046875X_1^2 X_2\end{aligned}$$

Cuando el rendimiento muestra una respuesta lineal a un factor  $A$  y cuando esta respuesta no es homogénea para los niveles de otro factor  $B$ , entonces tenemos una interacción  $A$  lineal por  $B$  o  $A_L B$ . A veces se desea examinar la naturaleza de esta interacción mediante la determinación de si los coeficientes de regresión lineal presentan una respuesta lineal a  $B$  y así sucesivamente. Esta componente de interacción se denominaría interacción  $A$  lineal por  $B$  lineal o  $A_L B_L$ . Tal como ocurre con otras interacciones, ésta es simétrica en  $A_L$  y  $B_L$ ; esto es, podemos considerar esta interacción como la correspondiente a los coeficientes de regresión lineal de la respuesta a  $B$  y su linealidad sobre los niveles de  $A$ . Análogamente, podríamos considerar  $A_Q B_L$ ,  $A_L B_Q$ ,  $A_Q B_Q$ , y otras interacciones como éstas.

El cálculo de sumas de cuadrados para tales interacciones es directo. Para la comparación de la interacción  $A_L B_L$ , simplemente multiplicamos los coeficientes de las dos comparaciones  $A_L$  y  $B_L$ . Estos se aplican a los totales de tratamientos y la cantidad resultante se eleva al cuadrado y se divide por el número de observaciones en cada total multiplicado por la suma de cuadrados de los coeficientes que se acaban de encontrar. En la siguiente sección, tratamos un ejemplo de tal interacción donde el espaciamiento no es igual y la interacción tiene un significado especial.

**Ejercicio 15.7.1** A partir de la tabla 15.12, obtener medias de tratamiento y espaciamiento de los surcos. Con esta información, calcular el coeficiente de regresión del rendimiento respecto al espaciamiento de surcos, con los métodos de los caps. 10 ó 13. Obsérvese que el valor hallado es el mismo que el calculado brevemente según la ec. (15.14).

**Ejercicio 15.7.2** Usese una de las formas de la ec. (15.14) para encontrar valores ajustados para las cinco medias del problema de soya. Hallar las cinco desviaciones  $\bar{Y} - \hat{Y}$ , elevarlas al cuadrado y sumarlas. Esto medirá la falta de ajuste a una ecuación cuadrática. Con base por observación, deberá ser igual a la suma de las sumas cúbica y cuártica de cuadrados. La varianza de una media de seis observaciones es  $\sigma^2/6$ , así que 6 es el multiplicador.

¿Es su suma de las desviaciones satisfactoriamente próxima a cero, tal como debiera ser?  
¿Es  $6[SC(\text{desviaciones})] = SC(\text{cúbica}) + SC(\text{cuártica})$ ?

**Ejercicio 15.7.3** Usese una de las formas de la ec. (15.15) para encontrar los valores ajustados de las seis medias de grama. Obsérvense las desviaciones de las medias observadas con respecto a los valores ajustados. Todas deben ser cero.

Usar la última forma y aproximar todos los coeficientes de regresión, incluso  $b_0$ , hasta cuatro cifras decimales. Repetir el ejercicio que se acaba de hacer.

¿Sería satisfactorio un proceso de cálculo que sólo diera cuatro cifras decimales?

**Ejercicio 15.7.4** Para los datos del ejercicio 9.3.5, tal como se han analizado en el ejercicio 15.4.2, examinar las componentes lineales y de falta de ajuste para la salinidad. ¿Son éstas homogéneas para los niveles de nitrógeno? ¿Para los niveles de aireación?

**Ejercicio 15.7.5** Con los datos del ejercicio 9.8.1 tal como se han analizado en el ejercicio 15.4.3, particionar la respuesta a la edad en componentes lineal y no lineal. ¿Son éstas homogéneas para las categorías estado físico?

## 15.8 Un solo grado de libertad para no aditividad

Tukey (15.13) da un método para aislar una suma de cuadrados del error con el propósito de probar la no aditividad; ésta tiene un grado de libertad. El método originalmente propuesto es aplicable a la clasificación de dos vías o diseño de bloques completos aleatorizados y se ilustra más abajo. También se ha ideado un método para diseños cuadrado latino (15.14).

Ahora calculamos una suma de cuadrados para no aditividad con un grado de libertad para los datos de la tabla 9.2. Los datos y los cálculos se presentan en la tabla 15.14. Lo primero es medir desviaciones de medias de tratamientos y bloque respecto de la media total. Luego calculamos los valores  $Q_j$ , definidos por la ec. (15.16), al pie de la primera parte de la tabla. Así

$$Q_j = \sum_i (Y_{i\cdot} - \bar{Y}_{..}) Y_{ij} \quad j = 1, \dots, b (= 4) \quad (15.16)$$

Esto es, multiplicar cada desviación de una media de tratamiento respecto de la media total por el correspondiente valor en el bloque 1 y sumar. Nótese que la ec. (15.16) es la misma ec. (8.3) en que los datos proporcionan los  $c_i$  como  $c_i = (\bar{Y}_{i\cdot} - \bar{Y}_{..})$ ;  $Y_{ij}$  es un total de una observación. Repetir para todos los bloques y obtener así los  $Q_j$ . Nótese que cada cálculo es equivalente a calcular el numerador de un coeficiente de regresión; calculamos

**Tabla 15.14 Un grado de libertad para no aditividad: un ejemplo**  
 (Ver también la tabla 9.2)

Tratamientos (estado en que se ha inoculado)	Bloques				Medias de tratamiento descodificadas $\bar{Y}_i - \bar{Y}_.$	$\bar{Y}_i - \bar{Y}_.$
	1	2	3	4		
Plántulas	4.4	5.9	6.0	4.1	35.10	-0.43
Floración temprana	3.3	1.9	4.9	7.1	34.30	-1.23
Flotación completa	4.4	4.0	4.5	3.1	34.00	-1.53
Cloración completa(1/100)	6.8	6.6	7.0	6.4	36.70	1.17
Madurez	6.3	4.9	5.9	7.1	36.05	0.52
Sin inoculación	6.4	7.3	7.7	6.7	37.03	1.50
Medias bloques descodificadas	35.27	35.10	36.00	35.75	35.53	
$\bar{Y}_j - \bar{Y}_.$	-0.26	-0.43	0.47	0.22		0.00
$Q_j = \sum_i (\bar{Y}_i - \bar{Y}_.) Y_{ij}$	8.149	10.226	7.316	5.991		

$$\begin{aligned}
 Q &= \sum_j (\bar{Y}_j - \bar{Y}_.) Q_j = \sum_j (\bar{Y}_j - \bar{Y}_.) \sum_i (\bar{Y}_i - \bar{Y}_.) Y_{ij} \\
 &= \sum_{i,j} (\bar{Y}_j - \bar{Y}_.) (\bar{Y}_i - \bar{Y}_.) Y_{ij} \\
 &= -1.759
 \end{aligned}$$

$$\begin{aligned}
 \frac{Q^2}{r \sum c_i^2} &= \frac{Q^2}{\sum_i (\bar{Y}_i - \bar{Y}_.)^2 \sum_j (\bar{Y}_j - \bar{Y}_.)^2} \quad \text{con } r = 1 \\
 &= \frac{(-1.759)^2}{7.928(.5218)} = .7483 \quad 1 \text{ gl como}
 \end{aligned}$$

#### Análisis de la varianza

Fuente de variación	gl	SC	CM	F
Bloques	$r - 1 = 3$	3.14	1.05	
Tratamientos	$t - 1 = 5$	31.65	6.33	4.83**
Error	$(r - 1)(t - 1) = 15$	19.72	1.31	
Aditividad	1	.75	.75	<1
Residuo	14	18.97	1.36	
Total	$rt - 1 = 23$	54.51		

la regresión de la respuesta de los individuos en cualquier bloque con respecto a las desviaciones de las medias de tratamientos respecto de la media global. Sólo hay que elevar al cuadrado cada  $Q_j$  y dividir por

$$\sum_i c_i^2 = \sum_i (\bar{Y}_i - \bar{Y}_.)^2$$

para obtener cuatro sumas de cuadrados tratamiento lineal  $T_L$ :

El siguiente cálculo comprende medir la homogeneidad de los coeficientes de regresión; en particular, vemos si presentan una tendencia lineal. El resultado se puede considerar

como una interacción de tratamiento lineal por bloque lineal o  $T_L B_L$ . Como coeficientes, usamos las desviaciones de las medias de los bloques respecto de la media global, esto es, los  $(\bar{Y}_{ij} - \bar{Y}_{..})$ . El cálculo es

$$Q = \sum_j (\bar{Y}_{ij} - \bar{Y}_{..}) Q_j \quad (15.17)$$

Finalmente,

$$Q = \sum_{i,j} (\bar{Y}_{i.} - \bar{Y}_{..})(\bar{Y}_{.j} - \bar{Y}_{..}) Y_{ij} \quad (15.18)$$

Se ve que la ecuación es simétrica en  $i$  y  $j$ .

Ahora calculamos las sumas de cuadrados atribuibles a la no aditividad mediante

$$\frac{Q^2}{\sum c_i^2} = \frac{Q^2}{\sum (\bar{Y}_{i.} - \bar{Y}_{..})^2 \sum (\bar{Y}_{.j} - \bar{Y}_{..})^2} \quad \text{con 1 gl} \quad (15.19)$$

Hemos visto que esto puede interpretarse como una interacción  $T_L B_L$ , donde los datos proporcionan los coeficientes; ordinariamente, en un experimento de dos factores, dispondríamos factores igualmente espaciados y apelamos a una tabla de valores de polinomios ortogonales.

La naturaleza de la no aditividad que se investiga puede verse escribiendo de nuevo la ec. (15.18) en una forma equivalente, esto es, como

$$Q = \sum_{i,j} (\bar{Y}_{i.} - \bar{Y}_{..})(\bar{Y}_{.j} - \bar{Y}_{..})(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}) \quad (15.20)$$

Ahora  $\bar{Y}_{i.} - \bar{Y}_{..}$  y  $\bar{Y}_{.j} - \bar{Y}_{..}$  son estimaciones de  $\tau_i$  y  $\beta_j$  respectivamente. Por tanto,  $(\bar{Y}_{i.} - \bar{Y}_{..})(\bar{Y}_{.j} - \bar{Y}_{..})$  es una estimación de la contribución de tratamiento de bloque que se espera en la celda  $i, j$ -ésima si los efectos de bloque y tratamiento, es decir, los efectos principales, son multiplicativos en vez de aditivos. También,  $Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}$  una estimación de la componente de error en la celda  $i, j$ -ésima cuando el supuesto de un modelo lineal aditivo es válido. Así, pues,  $Q$  es el numerador del coeficiente muestral de regresión del error de un modelo aditivo respecto al producto de los efectos. Ahora

$$\sum_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 \sum_j (\bar{Y}_{.j} - \bar{Y}_{..})^2 = \sum_{i,j} [(\bar{Y}_{i.} - \bar{Y}_{..})(\bar{Y}_{.j} - \bar{Y}_{..})]^2$$

así que finalmente la suma de cuadrados para no aditividad es la parte de la suma residual de cuadrados acostumbrada que puede atribuirse a esta regresión. Cuando el cuadrado medio de no aditividad es significante y no se debe a unas cuantas observaciones aberrantes, es necesaria una transformación. Harter y Lum (15.7) presentan la idea de regresión y no aditividad para un experimento de dos factores.

**Ejercicio 15.8.1** Los datos de la tabla 15.13 son datos transformados. Calcular la suma de cuadrados atribuible a no aditividad con un grado de libertad. La transformación, ¿ha resultado en datos para los cuales no se aplica todavía el modelo aditivo?

**Ejercicio 15.8.2** En el ejercicio 9.16.1 había que transformar los datos antes del análisis. Comprobar la no aditividad de los datos. Comprobar la no aditividad de los datos transformados. ¿Cuál fue el objeto de la formación?

**Ejercicio 15.8.3** En el ejercicio 9.16.2, había que transformar los datos antes del análisis. Comprobar los datos en cuanto a no aditividad antes y después de la transformación. ¿Cuál era el objeto de la transformación?

## Referencias

- 15.1. Anderson, R. L., y E. E. Houseman: "Tables of orthogonal polynomial values extended to  $N = 104$ ," *Iowa Agr. Exp. Sta. Res. Bull.*, 297, 1942.
- 15.2. Box, G. E. P., y J. S. Hunter: "Experimental designs for exploring response surfaces," in V. Chew (ed.), *Experimental Designs in Industry*, págs. 138-190, Wiley, Nueva York, 1958.
- 15.3. Cochran, W. G.: "Testing a linear relation among variances," *Biom.*, 7:17-32 (1951).
- 15.4. Crump, S. L.: "The estimation of variance components in analysis of variance," *Biom. Bull.*, 2:7-11 (1946).
- 15.5. Federer, W. T.: *Experimental Design*, Macmillan, Nueva York, 1955.
- 15.6. Fisher, R. A., y F. Yates: *Statistical Tables for Biological, Agricultural and Medical Research*, 5a. ed., Hafner, Nueva York, 1957.
- 15.7. Harter, H. L., y M. D. Lum: "A note on Tukey's one degree of freedom for non-additivity," *Abstract 474, Biom.*, 14:136-137 (1958).
- 15.8. Henderson, C. R.: "Estimation of variance and covariance components," *Biom.*, 9:226-252 (1953).
- 15.9. Satterthwaite, F. E.: "An approximate distribution of estimates of variance components," *Biom. Bull.*, 2:110-114 (1946).
- 15.10. Scheffé, H.: "Statistical methods for evaluation of several sets of constants and several sources of variability," *Chem. Eng. Progr.*, 50:200-205 (1950).
- 15.11. Schultz, E. F., Jr.: "Rules of thumb for determining expectations of mean squares in analysis of variance," *Biom.*, 11:123-135 (1955).
- 15.12. Snedecor, G. W.: *Statistical Methods*, 5a. ed., Iowa State College Press, Ames, Iowa, 1956.
- 15.13. Tukey, J. W.: "One degree of freedom for non-additivity," *Biom.*, 5:232-242 (1949).
- 15.14. Tukey, J. W.: "Reply to 'Query 113,'" *Biom.*, 11:111-113 (1955).
- 15.15. Wagner, R. E.: "Effects of depth of planting and type of soil on the emergence of small-seeded grasses and legumes," M.Sc. thesis, University of Wisconsin, Madison, 1943.
- 15.16. Wilk, M. B., y O. Kempthorne, "Fixed, mixed, and random models," *J. Amer. Statist. Ass.*, 50: 1144-1167 (1955).
- 15.17. Wilkinson, W. S.: "Influence of diethylstilbestrol on feed digestibility and on blood and liver composition of lambs," Tesis doctoral, University of Wisconsin, Madison, 1954.
- 15.18. Yates, F.: "The principles of orthogonality and confounding in replicated experiments," *J. Agr. Sci.*, 23:108-145 (1933).
- 15.19. Zick, W.: "The influence of various factors upon the effectiveness of maleic hydrazide in controlling quack grass, *Agropyron repens*," Tesis doctoral, University of Wisconsin, Madison, 1956.
- 15.20. Gaylor, D. W., y F. N. Hopper: "Estimating the degrees of freedom for linear combinations of mean squares by Satterthwaite's formula," *Technometrics*, 11:691-706 (1969).
- 15.21. Rawls, R. F.: "Training for increased comprehension with accelerated word rates in auditory reading media (compressed speech)," Tesis doctoral, North Carolina State University, Raleigh, N.C., 1970.
- 15.22. Searle, S. R.: *Linear Models*, Wiley, Nueva York, 1976.

- 15.23. Anderson, V. L.: "Restriction errors for linear models (an aid to develop models for designed experiments)," *Biom.*, 26: 255-268 (1970).
- 15.24. Anderson, V. L., y R. A. McLean: *Design of Experiments: A Realistic Approach*, Marcel Dekker, Nueva York, 1974.

---

## CAPITULO DIECISEIS

---

### ANALISIS DE LA VARIANZA IV: DISEÑO Y ANALISIS DE PARCELAS DIVIDIDAS

#### 16.1 Introducción

En nuestra exposición anterior sobre experimentos factoriales (cap. 15), se suponía que el conjunto de todas las combinaciones de tratamientos se aplicaba a las unidades experimentales de acuerdo con el proceso de aleatorización apropiado para el diseño completamente aleatorio, de bloques completos aleatorizados o de cuadrado latino. Pero son posibles otros procesos de aleatorización. Una de las aleatorizaciones alternas da lugar al diseño de *parcelas divididas*, que es una clase especial de diseño de bloques incompletos. El de parcelas divididas y algunas de sus aplicaciones es el tema de este capítulo.

#### 16.2 Diseños de parcelas divididas

Los diseños de parcelas divididas se usan frecuentemente en experimentos factoriales. Tales diseños pueden incorporar uno o más de los diseños completamente aleatorio, de bloques completos aleatorizados o de cuadrado latino. El principio básico es éste : las *parcelas completas* o *unidades completas*, a las cuales se les aplican niveles de uno o más factores, se dividen en *subparcelas* o *subunidades* a las cuales se les aplican niveles de uno o más factores adicionales. De este modo, cada unidad completa se convierte en un bloque para los tratamientos de subunidades. Por ejemplo, considérese un experimento para probar el factor *A* a cuatro niveles en tres bloques de un diseño de bloques completos al azar. Un segundo factor *B*, a dos niveles, puede superponerse mediante división de cada unidad del factor *A* en dos subunidades, y asignando los dos tratamientos *B* a esas subunidades. Aquí las unidades *A* son las unidades completas y las unidades *B* son las subunidades.

Después de la aleatorización, el diagrama de distribución, puede ser como sigue

Bloque 1	Bloque 2			Bloque 3		
$a_4 b_2$	$a_1 b_2$	$a_2 b_1$	$a_3 b_2$	$a_2 b_1$	$a_1 b_2$	$a_4 b_1$
$a_4 b_1$	$a_1 b_1$	$a_2 b_2$	$a_3 b_1$	$a_2 b_2$	$a_1 b_1$	$a_3 b_2$

Nótese que la aleatorización es en *dos etapas*. Primero aleatorizamos niveles del factor *A* en las unidades completas; luego aleatorizamos niveles del factor *B* en las subunidades, dos por unidad completa. Cada parcela dividida puede considerarse en lo que concierne al factor *B*, pero solo con un bloque incompleto en lo que atañe al conjunto completo de tratamientos. Por esta razón, los diseños de parcelas divididas pueden llamarse diseños de *bloques incompletos*.

El diseño de parcelas divididas es deseable en las siguientes situaciones:

1. Puede usarse cuando los tratamientos relacionados con los niveles de uno o más de los factores necesitan mayores cantidades de material experimental en una unidad experimental que los tratamientos de otros factores. Esto es común en experimentación sobre el campo, el laboratorio, industrial y social. Por ejemplo, en un experimento sobre el campo uno de los factores puede ser métodos de preparación del suelo o aplicación de un fertilizante, factores que necesitan ambos, por lo general, parcelas o unidades experimentales grandes. El otro factor puede ser variedades, las cuales se pueden comparar usando parcelas más pequeñas. Otro ejemplo es el experimento diseñado para comparar las cualidades de conservación de la crema de helado hecha a partir de diferentes fórmulas y almacenada a diferentes temperaturas. Para una sola replicación, el procedimiento sería la elaboración de un gran lote mediante cada fórmula, unidades completas, y luego dividir cada lote en almacenamientos separados a las diferentes temperaturas, las subunidades.
2. El diseño puede usarse si va a incorporarse en un experimento un factor adicional para aumentar su alcance. Por ejemplo, supongamos que el objeto principal de un experimento es comparar los efectos de varios fungicidas como protectores contra infección por una enfermedad. Para aumentar el alcance del experimento, se incluyen varias variedades de las cuales se sabe que difieren en su resistencia a la enfermedad. Aquí las variedades se podrían organizar en unidades completas y los protectores de semillas en subunidades.
3. A partir de la información anterior, se puede saber que pueden esperarse diferencias mayores entre los niveles de ciertos factores que entre los niveles de otros. En este caso, las combinaciones de los tratamientos para los factores donde se esperan diferencias grandes pueden asignarse aleatoriamente a las unidades completas, simplemente por comodidad.
4. El diseño se usa cuando se desea mayor precisión para comparaciones entre ciertos factores, que para otras. Esta es esencialmente la misma que la tercera situación, pero las razones pueden ser diferentes.

En resumen, como se espera que la variación entre las subunidades sea menor que entre las unidades, en los experimentos con parcelas divididas, los factores que necesitan

menores cantidades de material experimental, o que sean de mayor importancia, de los que se espera que presenten menores diferencias, o para los cuales por alguna razón se deseé mayor precisión, se asignan a las subunidades.

Ahora se expone la forma del análisis de la varianza para un experimento de dos factores en parcelas divididas para un diseño en bloques completos al azar. Sea  $r$  el número de bloques,  $a$  el número de niveles de  $A$  o unidades completas por bloque y  $b$  el número de niveles de  $B$  o subunidades por unidad completa. Supóngase que  $r = 3$ ,  $a = 4$  y  $b = 2$ . Las unidades enteras constituyen un grupo  $ra = 12$  unidades. Los 11 grados de libertad entre *unidades completas* se partitionan en 2 grados de libertad para bloques, 3 grados de libertad para el efecto principal  $A$  y 6 grados de libertad para un error experimental para las comparaciones de las unidades completas. Dentro de cada unidad completa hay un grado de libertad asociado con la variación entre subunidades dentro de una unidad completa, lo que da un total de 12 grados de libertad *dentro de unidades completas* para el experimento. Estos 12 grados de libertad están partitionados en un grado de libertad para el efecto principal  $B$ , 3 grados de libertad para la interacción  $AB$ , y 8 grados de libertad para un error experimental para las comparaciones entre las subunidades.

La partición de los grados de libertad para un diseño de parcelas divididas en la cual las unidades completas están dispuestas completamente al azar, en bloques completos aleatorizados y en un cuadrado latino, se presenta en la tabla 16.1. El factor  $A$ , aplicado a las unidades completas, tiene  $a$  niveles; y el factor  $B$ , aplicado a las subunidades, tiene  $b$  niveles. El factor  $B$  puede aplicarse a las subunidades en arreglos diferentes a los que aquí se usan.

Como cada unidad entera se debe dividir para acomodar todos los niveles de  $B$ , una unidad completa es un bloque en cuanto se refiere a  $B$ . En consecuencia, se aplica un análisis de dos vías con las fuentes: bloques,  $B$ , y el residuo o bloques  $\times B$ . Si se partitionan los bloques, como sucede para dar el análisis de la unidad completa, entonces el residuo puede partitionarse también en cada componente multiplicado por  $B$ . Es claro que, desearemos examinar la interacción  $AB$ . Por otro lado, estamos preparados para incluir bloques  $\times B$  en el error ( $b$ ) para el diseño de bloques completos al azar, y filas  $\times B$  y columnas  $\times B$  en el error ( $b$ ) para el cuadrado latino. Todo esto implica que para el diseño de bloques completos al azar, los bloques no interactúan con el factor  $B$ ; y que para el diseño de cuadrado latino, ni las filas ni las columnas interactúan con el factor  $B$ . Si existe alguna razón para dudar de este supuesto, debe partitionarse aún más el error ( $b$ ) en componentes de acuerdo con un modelo más completo. Cualquier duda respecto al modelo debe examinarse a la luz de la naturaleza del material experimental y a la experiencia anterior con el mismo. Obsérvese que el error ( $b$ ) tiene el mismo número de grados de libertad.

El error de la unidad completa, convenientemente designado  $E_a$ , usualmente es mayor que el error de las subunidades, designado  $E_b$ . Esto se debe a que las observaciones en las subunidades de la misma unidad completa tienden a correlacionarse positivamente y así reaccionar de modo más semejante que las subunidades de diferentes unidades enteras.  $E_a$  no puede ser menor que  $E_b$ , excepto por azar; y si esto sucede, es apropiado considerar ambos  $E_a$  y  $E_b$  como estimaciones de la misma  $\sigma^2$  y, en consecuencia, las dos sumas de cuadrados pueden combinarse y luego dividirse por los grados de libertad combinados para obtener una estimación de  $\sigma^2$ .

Ocasionalmente,  $E_a$  puede ser muy inferior a  $E_b$ . En experimentos de campo, las unidades completas pueden encontrarse sobre un solo eje perpendicular a un gradiente de

**Tabla 16.1** Participación de los grados de libertad para un diseño de parcelas divididas con diferentes organizaciones de la unidad completa

Completemtamente aleatorio (r replicaciones)			Bloques completos al azar (r = replicaciones = bloques)			Cuadrado latino (r = a replicaciones = lado del cuadro)		
Fuente	gl	Fuente	gl	Fuente	gl	Filas	Columnas	
Análisis de unidad completa								
A	$a - 1$	Bloques	$r - 1$			$a - 1$	$a - 1$	
Error (a)	$a(r - 1)$	A	$a - 1$			$a - 1$	$a - 1$	
Total unidad completa	$ar - 1$	Error (a)	$(a - 1)(r - 1)$			$(a - 1)(a - 2)$	$(a - 1)(a - 2)$	
		Total Unidad completa	$ar - 1$			Total unidad completa	$a^2 - 1$	
Análisis de la subunidad								
B	$b - 1$	B	$b - 1$			$b - 1$	$b - 1$	
AB	$(a - 1)(b - 1)$	AB	$(a - 1)(b - 1)$			$(a - 1)(b - 1)$	$(a - 1)(b - 1)$	
Error (b)	$a(r - 1)(b - 1)$	Error (b)	$a(r - 1)(b - 1)$			Error (b)	Error (b)	
Subtotal	$ab(b - 1)$	Subtotal	$ab(b - 1)$			Subtotal	Subtotal	
Total	$abr - 1$	Total	$abr - 1$			Total	Total	
						$a^2b - 1$	$a^2b - 1$	

fertilidad y así ser muy similares, mientras que las subunidades pueden ser muestras a lo largo de un gradiente de fertilidad y resultar así muy diferentes. Por lo tanto, la variación entre subunidades ha sido maximizada a expensas de la variación entre unidades completas, al contrario de lo que se anticipó para el experimento.

Si el experimento factorial no está planeado en un diseño de parcelas divididas, entonces el diseño que se usa tiene una cierta precisión general aplicable a las medias de los tratamientos. Respecto de este experimento, el diseño de parcelas divididas deberá dar más precisión para las comparaciones de subunidades, pero a costa de menos precisión para las comparaciones de unidad completa, ya que no es probable que la precisión total se modifique. Las desviaciones estándar o errores estándar apropiados para las comparaciones entre diferentes medias se dan en la tabla 16.2. En los primeros tres casos, los divisores con los números de subunidades en una media; en el último caso,  $r$  es el número de subunidades en una media de tratamiento, pero el divisor correcto es  $rb$ .

Las comparaciones de dos medias  $A$ , al mismo nivel o a diferentes niveles de  $B$ , comprenden tanto el efecto principal de  $A$  como la interacción  $AB$ . Esto es evidente al observar los símbolos de tratamiento. Ellos son tanto comparaciones de unidades completas como de subunidades; es apropiado usar un promedio ponderado de  $E_a$  y  $E_b$  tal como se da en la tabla 16.2. Las ponderaciones son  $(b - 1)$  y 1, su suma es  $b$ , así  $b$  aparece en el divisor. Para tales comparaciones, la relación de la diferencia de tratamiento al error estándar no sigue la distribución  $t$  de Student. La aproximación de la sec. 5.9 puede adaptarse para obtener un valor para todo nivel de significancia. Sean  $t_a$  y  $t_b$  los valores  $t$  tabulados al nivel de significancia elegido para los correspondientes grados de libertad de  $E_a$  y  $E_b$ . Entonces

$$t' = \frac{(b - 1)E_b t_b + E_a t_a}{(b - 1)E_b + E_a} \quad (16.1)$$

Tabla 16.2 Errores estándar en un diseño de parcelas divididas

Diferencia entre	Medida como*	Ejemplo	Error estándar de la diferencia
Dos medias $A$	$a_i - a_j$	$a_1 - a_2$	$\sqrt{\frac{2E_a}{rb}}$
Dos medias $B$	$b_i - b_j$	$b_1 - b_2$	$\sqrt{\frac{2E_b}{ra}}$
Dos medias $B$ al mismo nivel de $A$	$a_i b_j - a_i b_k$	$a_1 b_1 - a_1 b_2$	$\sqrt{\frac{2E_b}{r}}$
Dos medias $A$ al 1. Mismo nivel de $B$ o 2. Diferentes niveles de $B$ (dos medias de tratamiento).	$a_i b_j - a_k b_j$	$a_1 b_1 - a_2 b_1$	$\sqrt{\frac{2[(b - 1)E_b + E_a]}{rb}}$
	$a_i b_j - a_k b_l$	$a_1 b_2 - a_2 b_1$	

\* Todas las medias se miden con base en una subunidad. Esto está implicado en los procedimientos de cálculo.

es el valor, al nivel de significancia elegido, con el cual comparamos nuestro  $t$  muestral. Así,  $t'$  corresponde a un  $t$  tabulado. Caerá entre  $t_a$  y  $t_b$ .

Muchas variantes del diseño de parcelas divididas son de uso común. Una de éstas supone la división de cada subunidad en  $c$  sub-subunidades para la inclusión de un tercer factor  $C$  a  $c$  niveles. Los niveles del factor  $C$  se asignan aleatoriamente a las sub-subunidades. La partición de los grados de libertad es exactamente como aparece en la tabla 16.1 añadiendo un análisis de sub-subunidades. Esto es

Fuente	gl
$C$	$c - 1$
$AC$	$(a - 1)(c - 1)$
$BC$	$(b - 1)(c - 1)$
$ABC$	$(a - 1)(b - 1)(c - 1)$
Error ( $c$ )	$ab(r - 1)(c - 1)$
Subtotal	$abr(c - 1)$
Total (sub-subunidades)	$abcr - 1$

Los cálculos de sumas de cuadrados se hacen con base en la sub-subunidad. Los divisores son el número de sub-subunidades en cualquier total elevado al cuadrado. La suma de cuadrados para el error ( $c$ ),  $E_c$ , se obtiene restando de la suma de cuadrados total para las sub-subunidades, la suma de todas las otras sumas de cuadrados. Este diseño se conoce como de parcelas subdivididas. Para otras variantes del diseño de parcelas divididas, se remite al lector a Cochran y Cox (16.2) y Federer (16.3).

No es necesario tener una división adicional por cada factor. Si se tienen tres factores las combinaciones  $AB$ , pueden asignarse a las unidades completas y los niveles del factor  $C$  a las subunidades, o los niveles de  $A$  a las unidades completas y las combinaciones  $BC$  a las subunidades.

**Ejercicio 16.2.1** Considere un experimento factorial con  $A$  aleatorizado sobre parcelas completas y  $B$  dentro de éstas. Supóngase un diseño de bloques completos al azar con dos bloques.

Elaborar una tabla con ocho columnas que representen los cuatro símbolos de tratamientos para cada uno de los bloques de tal forma que cada encabezamiento de columna represente una observación. Asegúrese de que las observaciones en la misma parcela completa sean adyacentes.

En la tabla, escribir los coeficientes para comparaciones con un grado de libertad y que representan: (1) bloques, (2)  $A$ , (3) bloques  $\times A$  = error ( $a$ ), (4)  $B$ , (5)  $AB$ , (6) bloques  $\times B$ , y (7) bloques  $\times AB$ . Obsérvese que los contrastes (6) y (7) en conjunto constituyen el error ( $b$ ).

Obsérvese que cada una de las comparaciones (1), (2) y (3) tienen el mismo coeficiente en cualquier parcela completa. Las diferencias que implican cambios de signo, sólo se presentan entre las parcelas enteras. No hay comparaciones de parcelas divididas en esta parte del análisis.

Obsérvese también que las comparaciones restantes tienen un +1 y un -1 dentro de cada parcela completa o en parcelas divididas.

Ahora las razones para análisis de parcelas divididas debe resultar más obvio. También podemos decir que el efecto principal  $A$  se confunde con parcelas completas.

### 16.3 Un ejemplo de parcelas divididas

En un experimento llevado a cabo por D. C. Arny en la Universidad de Wisconsin, se compararon los rendimientos de cuatro lotes de avenas para tres tratamientos químicos de las semillas y un control sin tratamiento. Dos de los lotes de semillas eran Vicland, designados Vicland (1) cuando estaban infectados por *H. Victoriae* y Vicland (2) cuando no estaban infectados. Los otros dos lotes de semillas eran muestras de avena Clinton y Branch que son resistentes a *H. Victoriae*. Los lotes de semillas, factor *A*, se distribuyeron aleatoriamente a las parcelas completas dentro de cada bloque; los protectantes de las semillas, factor *B*, se asignaron aleatoriamente a las subparcelas dentro de cada parcela completa. El diseño de parcelas completas era un diseño de bloques completos al azar de cuatro bloques. Los rendimientos en bushels por acre se dan en la tabla 16.3.

El análisis de la varianza de los datos se calcula con base en subunidades, la unidad en la cual se mide la respuesta. El tratamiento es como sigue. Sea  $Y_{ijk}$  el rendimiento en el bloque  $i$ -ésimo de la subunidad que recibe el nivel  $j$ -ésimo del factor *A* y el nivel  $k$ -ésimo del factor *B*. Entonces  $Y_{i..}$  es el total para el bloque  $i$ -ésimo, o sea, la suma de  $ab$  observaciones, en las subunidades;  $Y_{..j}$  es el total para todas las subunidades que reciben el factor *A* al nivel  $j$ -ésimo, la suma de  $rb$  observaciones;  $Y_{..k}$  es el total de todas las subunidades que reciben el factor *B* al nivel  $k$ -ésimo, la suma de  $ra$  observaciones;  $Y_{ij.}$  es el total de una unidad completa, la suma de  $b$  observaciones, etc.

*Paso 1* Calcular el factor de corrección y la suma de cuadrados total.

$$\text{Factor de corrección } \frac{Y_{...}^2}{rab} = \frac{3,379.8^2}{64} = 178,485.13$$

$$\begin{aligned} \text{SC(total)} &= \sum_{i, j, k} Y_{ijk}^2 - FC \\ &= 42.9^2 + \cdots + 47.4^2 - FC = 7,797.39 \end{aligned}$$

*Paso 2* Completar el análisis de las unidades completas,

$$\begin{aligned} \text{SC(unidades completas)} &= \frac{\sum_{i, j} Y_{ij.}^2}{b} - FC \\ &= \frac{190.6^2 + \cdots + 209.6^2}{4} - FC = 6,309.19 \end{aligned}$$

$$\begin{aligned} \text{SC(bloques)} &= \frac{\sum_i Y_{i..}^2}{ab} - FC \\ &= \frac{965.3^2 + \cdots + 743.9^2}{4(4)} - FC = 2,842.87 \end{aligned}$$

$$\text{SC}(A = \text{lotes de semillas}) = \frac{\sum_j Y_{..j}^2}{ra} - FC$$

Tabla 16.3 Rendimientos de avenas, en bushels por acre

Lote de semillas, A	Bloque	Treatment, B				Totales
		Control	Ceresan	M	Panogen	
Vicland (1)	1	42.9	53.8	49.5	44.4	190.6
	2	41.6	58.5	53.8	41.8	195.7
	3	28.9	43.9	40.7	28.3	141.8
	4	30.8	46.3	39.4	34.7	151.2
Totales		144.2	202.5	183.4	149.2	679.3
Vicland (2)	1	53.3	57.6	59.8	64.1	234.8
	2	69.6	69.6	65.8	57.4	262.4
	3	45.4	42.4	41.4	44.1	173.3
	4	35.1	51.9	45.4	51.6	184.0
Totales		203.4	221.5	212.4	217.2	854.5
Clinton	1	62.3	63.4	64.5	63.6	253.8
	2	58.5	50.4	46.1	56.1	211.1
	3	44.6	45.0	62.6	52.7	204.9
	4	50.3	46.7	50.3	51.8	199.1
Totales		215.7	205.5	223.5	224.2	868.9
Branch	1	75.4	70.3	68.8	71.6	286.1
	2	65.6	67.3	65.3	69.4	267.6
	3	54.0	57.6	45.6	56.6	213.8
	4	52.7	58.5	51.0	47.4	209.6
Totales		247.7	253.7	230.7	245.0	977.1
Totales de tratamiento		811.0	883.2	850.0	835.6	3,379.8
Bloques	1	2	3	4		
Tratamientos	965.3	936.8	733.8	743.9		

Fuente: Datos usados con permiso de D.C. Arny, Universidad de Wisconsin, Madison, Wisconsin.

$$= \frac{679.3^2 + \dots + 977.1^2}{4(4)} - FC = 2,848.02$$

$$\begin{aligned} SC[\text{error } (a)] &= SC(\text{unidades completas}) - SC(\text{bloques}) - SC(A) \\ &= 6,309.19 - (2,842.87 + 2,848.02) = 618.30 \end{aligned}$$

Paso 3 Completar el análisis de las subunidades.

$$\begin{aligned} SC(B = \text{tratamiento de semillas}) &= \frac{\sum_{k=1}^k Y_{..k}^2}{ra} - FC \\ &= \frac{811.0^2 + \dots + 835.6^2}{4(4)} - FC = 170.53 \end{aligned}$$

$$\begin{aligned}
 SC(AB) &= \frac{\sum_{j,k} Y_{jk}^2}{r} - FC - SC(A) - SC(B) \\
 &= \frac{144.2^2 + \dots + 245.0^2}{4} - FC - (2,848.02 + 170.53) \\
 &= 586.47
 \end{aligned}$$

$$\begin{aligned}
 SC[\text{error } (b)] &= SC(\text{total}) - SC(\text{unidades completas}) - SC(B) - SC(AB) \\
 &= 7,797.39 - 6,309.19 - 170.53 - 586.47 = 731.20
 \end{aligned}$$

Las sumas de cuadrados se llevan a una tabla de análisis de la varianza, como la tabla 16.4, que es la que suministra los datos en estudio. Se dan dos coeficientes de variación. Para parcelas completas, CV se define como  $= (\sqrt{E_a/b}/\bar{Y}_{..})100$ . Esto equivale a omitir la división en subparcelas y a hacer sólo el análisis de parcelas completas. Para el modelo fijo, el valor de  $F$  para los lotes de semillas exige que  $E_a$  esté en el denominador; los para tratamientos de semillas e interacciones exigen  $E_b$ . El valor  $F$  para los lotes de semillas es altamente significante, el para los tratamientos de las semillas es casi significante al nivel del 5 por ciento, y el para la interacción es altamente significante. Para el modelo aleatorio, la elección del denominador para los diversos contrastes  $F$  no es tan directa. Se remite al lector a la tabla 16.9.

Como la interacción es significante, las diferencias en las respuestas entre los lotes de semillas varían para los tratamientos de las semillas en una forma que el azar y la hipótesis nula no pueden explicar fácilmente; es importante examinar los efectos simples. Los efectos simples de mayor interés están entre los cuatro tratamientos de semillas dentro de cada lote de semillas. Para las comparaciones deseadas, las medias de tratamiento y los errores estándar, al igual que en la tabla 16.2, se dan en la tabla 16.5.

**Tabla 16.4 Análisis de la varianza para los datos de la tabla 16.3  
(Con base en subunidades)**

Fuente de variación	gl	SC	Cuadrado medio	F
Bloques	3	2,842.87	947.62	
Lotes de semillas, factor A	3	2,848.02	949.34	13.82**
Error (a)	9	618.30	68.70	
Tratamientos de las semillas, factor B	3	170.53	56.84	2.80
Interacción AB	9	586.47	65.16	3.21**
Error (b)	36	731.20	20.31	
Total	63	7,797.39		

$$\text{Coeficiente de variación: } CV(a) = \frac{\sqrt{68.70/4}}{52.8} (100) = 7.8\%;$$

$$CV(b) = \frac{\sqrt{20.31}}{52.8} (100) = 8.5\%$$

**Tabla 16.5 Rendimientos medios y errores estándar**

Rendimientos medios de avenas, en bushels por acre, con los datos de la tabla 16.3

Lotes de semillas	Tratamientos de semillas				Medias de lotes de semillas
	Control	Ceresan M	Panogen	Agrox	
Vicland (1)	36.1	50.6	45.9	37.3	42.5
Vicland (2)	50.9	55.4	53.1	54.3	53.4
Clinton	53.9	51.4	55.9	56.1	54.3
Branch	61.9	63.4	57.7	61.3	61.1
Medias de tratamientos de semillas	50.7	55.2	53.1	52.2	52.8

## Errores estándar

Diferencia entre	Error estándar como bushels por acre	gl
Medias de dos lotes de semillas, factor A	$\sqrt{\frac{2(68.70)}{4(4)}} = 2.93$	9
Medias de dos tratamientos de semillas, factor B	$\sqrt{\frac{2(20.31)}{4(4)}} = 1.59$	36
Medias de dos tratamientos en el mismo lote de semilla	$\sqrt{\frac{2(20.31)}{4}} = 3.19$	36
Medias de dos lotes para el mismo tratamiento de semilla	$\sqrt{\frac{2[(3)20.31 + 68.70]}{4(4)}} = 4.02$	—

Para calcular un valor  $t$ , correspondiente a un  $t_{0.025}$  tabulado para comparar los promedios de dos lotes de semillas con el mismo tratamiento para las mismas, usamos la ec. (16.1). Los  $t$  tabulados para 9 y 36 grados de libertad son 2.262 y 2.028, respectivamente. Así,  $t_{0.025}$  para la comparación es

$$t' = \frac{3(20.31)2.028 + 68.70(2.262)}{3(20.31) + 68.70} = 2.152$$

Tal valor  $t'$  siempre se encuentra entre los dos valores de  $t$  usados en su cálculo, ya que es una media ponderada de tales  $t$  tabulados. Este hecho a menudo obviará la necesidad del cálculo de  $t'$ . Esta prueba se hace frente a alternativas bilaterales con  $\alpha = 0.05$ .

Para comparaciones en que entre el control con cada uno de los protectantes de semillas dentro de los lotes, el método de Dunnett da

$$t_{.05}(36 \text{ gl}) s_{\bar{y}_c - \bar{y}_i} = 2.14(3.19) = 6.8 \text{ bushels por acre} \quad \text{Prueba unilateral}$$

$$t_{.01}(36 \text{ gl}) s_{\bar{y}_c - \bar{y}_i} = 2.84(3.19) = 9.1 \text{ bushels por acre} \quad \text{Prueba unilateral}$$

donde  $t$  procede de las tablas de Dunnett, tabla A.9. Aquí se usa una prueba unilateral porque se espera que un protectante sea beneficioso. La probabilidad se aplica al conjunto de enunciados, esto es, un conjunto de tres, en vez de para cada enunciado por separado.

Concluimos que con Vicland (1) el aumento en rendimiento comparado con el control es altamente significante para Ceresan M y Panogen, pero no para Agrox. Para los otros lotes de semillas no se encontraron diferencias significativas.

**Ejercicio 16.3.1** En 1951 J.W. Lambert, de la Universidad de Minnesota, comparó el efecto del espaciamiento de cinco surcos sobre el rendimiento de dos variedades de soya. El diseño era de parcelas divididas con variedades como tratamientos de parcelas completas en un diseño de bloques completos al azar; los espaciamientos entre surcos se aplicaron a subparcelas. El rendimiento en bushels por acre para seis bloques se dan en la tabla siguiente.

Variedad	Espaciamiento entre surcos, pulgadas	Bloque					
		1	2	3	4	5	6
OM*	18	33.6	37.1	34.1	34.6	35.4	36.1
	24	31.1	34.5	30.5	32.7	30.7	30.3
	30	33.0	29.5	29.2	30.7	30.7	27.9
	36	28.4	29.9	31.6	32.3	28.1	26.9
	42	31.4	28.3	28.9	28.6	18.5	33.4
B	18	28.0	25.5	28.3	29.4	27.3	28.3
	24	23.7	26.2	27.0	25.8	26.8	23.8
	30	23.5	26.8	24.9	23.3	21.4	22.0
	36	25.0	25.3	25.6	26.4	24.6	24.5
	42	25.7	23.2	23.4	25.6	24.5	22.9

\* OM = Ottawa Mandarin; B = Blackhawk.

Escribir el análisis de la varianza. Hacer los datos. Calcular el C.V para unidades enteras y subunidades.

**Ejercicio 16.3.2** Particionar la suma de cuadrados para los espaciamientos de los surcos y la interacción de variedad por espaciamiento entre surcos en lineal, cuadrática y desviaciones respecto de la cuadrática = componentes de falta de ajuste. Usar valores del polinomio ortogonal donde sea posible. Interpretar los datos detalladamente.

**Ejercicio 16.3.5** Las partidas en la columna de "Totales" de la tabla 16.3 son rendimientos en parcelas completas. Analíicense esos 16 valores como si fueran los resultados del experimento completo. ¿Cómo se comparan con el coeficiente de variación?

**Ejercicio 16.3.4** Medidas respectivas tales como las de los datos de antes y después de tratamiento a veces se analizan como si fueran de un experimento de parcelas divididas.

Consideremos que  $3 \times 14 = 42$  individuos se asignan aleatoriamente en igual número a tres tratamientos. Ahora se toman dos observaciones en el transcurso del tiempo. Aun cuando no podemos claramente aleatorizar el pre y post como tratamientos, analizaremos los datos como si fuese un experimento de parcelas divididas. Consideremos los datos del ejercicio 7.3.3.

El análisis corresponde al primero de los presentados en la tabla 16.1. El error ( $a$ ) puede describirse como "entre sujetos dentro de tratamientos" y el error ( $b$ ) como "(pre frente a post)

X sujetos dentro de tratamientos". El error ( $a$ ) se calcula a partir de las respuestas combinadas para las dos pruebas; el error ( $b$ ) considera la homogeneidad de este tipo de variabilidad en las dos pruebas.

Completar el análisis de la varianza. ¿Dónde se busca una prueba de una respuesta global a los tratamientos? ¿Dónde está la comprobación que busca el problema de determinar si existe una respuesta diferencial a los tratamientos?

Los ejercicios 7.3.3 y 7.3.4 dieron análisis de los mismos datos. Si los presentes análisis hubiesen combinado todos los términos de error previamente calculados, se tendrían ahora  $6 \times 13 = 78$  gl para estimar el error. ¿En qué se ha convertido esta información? (Sugerencia: Releer el párrafo anterior).

**Ejercicio 16.3.5** Los datos pre y post del ejercicio 15.3.3 puede analizarse como se sugirió en el ejercicio 16.3.4. El análisis de las parcelas completas supone un factorial de  $2 \times 2$ , mientras que las subparcelas introducirán un tercer factor a dos niveles.

Analizar el conjunto completo de datos. ¿Dónde está la información relativa a las diferencias entre los métodos de instrucción?

#### 16.4 Datos faltantes en diseños de parcelas divididas

Las fórmulas para estimar observaciones faltantes en el diseño de parcelas divididas las da Anderson (16.1). Considérese el caso en que falta una sola subunidad y el tratamiento es  $a_j b_k$ . Sea  $Y$  la observación de la subunidad faltante, y sea  $W$  el total de subunidades observadas en la unidad entera de donde falta la observación,  $(a_j b_k)$  el total de subunidades observadas que recibieron el mismo tratamiento  $a_j b_k$ , y sea  $(a_j)$  el total de subunidades observadas que recibieron el nivel  $j$ -ésimo de  $A$ . Entonces la estimación del valor faltante está dada por la ec. (16.2). Nótese que ésta es la ec. (9.8), en lo que los bloques son las parcelas completas, donde  $A$  se encuentra en el nivel  $j$ -ésimo.

$$Y = \frac{rW + b(a_j b_k) - (a_j)}{(r-1)(b-1)} \quad (16.2)$$

Por ejemplo, supóngase que falta en la tabla 16.3 el valor del control en el bloque 1 para Vicland (1), es decir, 42.9. Entonces

$$W = 190.6 - 42.9 = 147.7$$

$$(a_j b_k) = 144.2 - 42.9 = 101.3$$

$$(a_j) = 679.3 - 42.9 = 636.4$$

y

$$Y = \frac{4(147.7) + 4(101.3) - 636.4}{3(3)} = \frac{359.6}{9} = 40.0$$

Si faltan valores, cada uno en tratamientos diferentes de unidades completas, estimar los valores faltantes dentro de cada tratamiento en unidad completa tal como se des-

Tabla 16.6 Errores estándar para el diseño de parcelas divididas con datos faltantes

Comparación	Medida como	Error estándar de la diferencia
Diferencia entre dos medias $A$	$a_i - a_j$	$\sqrt{\frac{2(E_a + fE_b)}{rb}}$
Diferencia entre dos medias $B$	$b_i - b_j$	$\sqrt{\frac{2E_b(1 + fb/a)}{ra}}$
Diferencia entre dos medias $B$ al mismo nivel de $A$	$a_i b_j - a_k b_l$	$\sqrt{\frac{2E_b(1 + fb/a)}{r}}$
Diferencia entre dos medias $A$		
1. Al mismo nivel de $B$	$a_i b_j - a_k b_j$	$\sqrt{\frac{2E_a + 2E_b[(b-1) + fb^2]}{rb}}$
2. A diferentes niveles de $B$	$a_i b_j - a_k b_l$	

cribió anteriormente. Si falta más de una subunidad en un tratamiento de una unidad completa, aplicar repetidamente la ecuación antes vista.

El cálculo de la suma de cuadrados para el análisis de la varianza se lleva a cabo en la forma usual una vez que se han insertado en los cálculos el valor o valores faltantes. Por cada valor faltante de subunidad se resta un grado de libertad al error ( $b$ ). La estimación de  $E_b$  es insesgada; sin embargo, los cuadrados medios para tratamientos y para  $E_a$  son sesgados hacia arriba. Si sólo faltan pocos valores, los sesgos pueden omitirse. Anderson (16.1) da procedimientos para obtener estimaciones insesgadas, igual que los procedimientos para estimar una unidad completa faltante. El lector también puede consultar a Khargonkar (16.4).

Cochran y Cox (16.2) dan fórmulas para estimar los errores estándar de las diferencias entre dos medias en las que entran valores faltantes. Tales fórmulas se reproducen en la tabla 16.6.

Cuando sólo falta un valor, el factor  $f$  de la tabla 16.6 es  $1/[2(r-1)(b-1)]$  para comparaciones en las que entra un promedio con el valor faltante y otra media. Sin embargo, si falta más de una observación,  $f$  depende de la ubicación de las subunidades faltantes. La siguiente aproximación es correcta para ciertos casos, pero tiende a ser ligeramente grande para otros casos.

$$f = \frac{k}{2(r-d)(b-k+c-1)}$$

donde  $k$ ,  $c$  y  $d$  sólo se refieren a observaciones faltantes para las dos medias que se comparan; en particular

$k$  = número de observaciones faltantes

$c$  = número de bloques que contienen observaciones faltantes

$d$  = número de observaciones en el tratamiento de subunidad  $a_i b_k$  más afectado

**Ejercicio 16.4.1** Para los datos del ejercicio 16.3.1, supóngase que faltan los valores para la variedad OM, espaciamiento entre surcos de 18, bloque 1, para la variedad B, espaciamiento entre surcos de 30, bloque 5 y para la variedad OM, espaciamiento entre surcos de 42, bloque 5. Calcular los valores faltantes para estas observaciones. Construir una tabla como la tabla 16.6, colocando valores numéricos excepto para  $E_a$  y  $E_b$ , los cuales deben obtenerse mediante un nuevo análisis.

## 16.5 Diseño de bloques divididos

Los diseños de parcelas divididas antes estudiados se conocen como de parcelas divididas en el espacio, ya que cada unidad completa se subdivide en subunidades distintas. En algunos experimentos, se hacen observaciones sucesivas en la misma unidad completa durante un período de tiempo. Por ejemplo, en un cultivo forrajero como la alfalfa, los datos sobre rendimiento en forraje usualmente se obtienen dos o más veces por año en un período de varios años. Tales datos son de alguna manera análogos a los provenientes de un diseño de parcelas divididas en muchos aspectos, y sus análisis a veces se efectúan como tales y se les conoce como parcelas divididas en el tiempo.

Sin embargo, un aspecto de interés de tales experimentos es que no hay parcelas divididas dentro de las parcelas completas en el sentido físico del estudio e ilustración iniciales. Cada corte es de toda la parcela completa. Esto implica una cierta simetría de los dos factores, así que el uno parece elección tan probable como el otro para denominarse el tratamiento de la parcela completa. Tales diseños, incluyendo aquéllos donde las parcelas de cada factor se cruzan físicamente entre sí, se llaman *diseños de bloques divididos* o *diseños con ambos factores en franjas*.

Cuando las parcelas se cruzan físicamente unas con otras, un diagrama aleatorizado puede ser como sigue:

Bloque 1			Bloque 2			Bloque 3			
	$a_3$	$a_1$		$a_2$	$a_1$	$a_3$		$a_1$	
$b_1$				$b_2$				$b_2$	
$b_2$				$b_1$				$b_1$	

Con una aleatorización en mente, se usa a menudo el análisis alterno que sigue. Se dan tres términos de error. Se espera que éstos puedan ser relativamente altos para efectos principales, pero relativamente pequeños para interacciones. Este análisis parece particularmente apropiado en cuanto la experiencia nos dice qué bloques por A, bloques por B, y los bloques por AB son a menudo no homogéneos. Esto no es muy difícil de prever. Por ejemplo, sean los bloques áreas diferentes sobre un suelo inclinado. Toda diferencia en respuesta a cultivares, A, puede ser bastante homogénea en bloques, de tal forma que bloques por A sea de magnitud razonable. Por otra parte, las diferencias entre cortes, B, pueden depender del bloque en que uno con buena humedad de diferencias notorias, mientras que otro con inadecuada humedad entre cortes da lugar a un segundo crecimiento bajo y por tanto origina diferencias mayores. Por consiguiente, bloques por B como fuente de

variación es de diferente orden de magnitud que bloques por  $A$ . Finalmente, esta última modalidad de diferencias puede depender de los cultivares, ya que probablemente difieren en su capacidad para resistir a la sequía. La componente bloques por  $AB$  puede ser aún de diferente orden de magnitud. Esta partición equivale al reconocimiento de fuentes posibles de variación e incluir un término en el modelo para cada una de ellas y efectuar el análisis de acuerdo con ello.

Supóngase que los rendimientos se miden en cada parcela para los  $b$  cortes de  $a$  cultívaras de alfalfa en un diseño de bloques completos al azar con  $r$  bloques. Sea  $Y_{ijk}$  la observación en el bloque  $i$ -ésimo de la variedad  $j$ -ésima donde se hizo el corte  $k$ -ésimo.

*Paso 1* Hacer un análisis de la varianza para cada corte, esto es, un análisis para los  $Y_{ij1}$ , para los  $Y_{ij2}$ , y así sucesivamente. (Este paso sólo se aplica a experimentos donde cosechas múltiples dan lugar a los datos).

*Paso 2* Preparar una tabla de 2 vías de los totales de bloques por variedad para todos los cortes, esto es, una tabla de  $Y_{ij..}$  (ver tabla 16.7). Estos corresponden a totales de unidades completas en el diseño de parcelas divididas.

*Paso 3* A partir de los totales de la tabla, calcular el análisis para las unidades enteras con base en las subunidades como en la tabla 16.8, es decir, usar divisores basados en el número de subunidades.

*Paso 4* Preparar una tabla parecida de dos vías para los totales de bloques por cortes para todas las variedades; o sea, una tabla de  $Y_{i..k}$ . Calcular un análisis de dos vías con base en las subunidades, tal como en la segunda parte de la tabla 16.8. Como SC(bloques) ya se encuentra en la tabla del análisis de la varianza, no se repite.

*Paso 5* Completar el análisis de subunidades tal como en la tabla 16.8. Se necesitan una tabla  $AB$  y una SC total con base por subunidad.

Los totales necesarios para el cálculo de las sumas de cuadrados de  $B$ ,  $AB$  y  $RB$  se encuentran en los análisis individuales de los cortes. Así, SC( $B$ ) requiere de los totales glo-

**Tabla 16.7** Totales de las parcelas completas para el análisis de unas parcelas divididas en el tiempo

Bloque	Variedad					Totales de bloques
	1	...	$j$	...	$a$	
1	$Y_{11..}$	...	$Y_{1j..}$	...	$Y_{1a..}$	$Y_{1..}$
$i$	$Y_{i1..}$	...	$Y_{ij..}$	...	$Y_{ia..}$	$Y_{i..}$
$r$	$Y_{r1..}$	...	$Y_{rf..}$	...	$Y_{ra..}$	$Y_{r..}$
Totales de variedad	$\overline{Y_{1..}}$	...	$\overline{Y_{f..}}$	...	$\overline{Y_{a..}}$	$\overline{Y_{..}}$

Tabla 16.8 Análisis de la varianza para un diseño de bloques divididos

Fuente	gl	SC
Bloques, $R$	$r - 1$	$\sum_i Y_{i..}^2/ab - FC$
Variedades, $A$	$a - 1$	$\sum_j Y_{.j.}^2/rb - FC$
Error ( $a$ ), $RA$	$(r - 1)(a - 1)$	$\sum_{i,j} Y_{ij.}^2/b - FC - SC(R) - SC(A)$
Unidades completas para $A$	$ra - 1$	$\sum_{i,j} Y_{ij.}^2/b - FC$
Cortes, $B$	$b - 1$	$\sum_k Y_{.k.}^2/ra - FC$
Errores ( $b$ ), $RB$	$(r - 1)(b - 1)$	$\sum_{i,k} Y_{i.k.}^2/a - FC - SC(R) - SC(B)$
Unidades completas para $B$	$rb - 1$	$\sum_{i,k} Y_{i.k.}^2/a - FC$
Variedades $\times$ cortes, $AB$	$(a - 1)(b - 1)$	$\sum_{j,k} Y_{jk.}^2/r - FC - SC(A) - SC(B)$
Error ( $c$ ), $RAB$	$(r - 1)(a - 1)(b - 1)$	$\sum_{i,j,k} Y_{ijk.}^2 - FC - SC(\text{unidades completas para } A)$ $- SC(B) - SC[\text{error } (b)] - SC(AB)$
Total (subunidades)	$rab - 1$	$\sum_{i,j,k} Y_{ijk.}^2 - FC$

bales de los análisis individuales de cortes,  $SC(AB)$  requiere los totales de variedades a partir de los análisis individuales de cortes, y la  $SC(RB)$  necesita de los totales de los bloques de los análisis individuales de los cortes.

Existe cierta relación entre los análisis de los cortes individuales y los análisis combinados. Estos sirven como comprobación de los cálculos o como procedimientos de cálculo. La  $SC(B)$  es la suma de los términos de corrección individuales menos  $C$ , el término de corrección total. Las sumas de los grados de libertad y las sumas de cuadrados de  $A$  y  $AB$  son iguales a las correspondientes de  $A$  en los cortes individuales. Esto implica que la  $SC(AB)$  puede obtenerse restando  $SC(A)$ , en los análisis combinados de la suma, de las  $SC(A)$  en los análisis de los cortes individuales. En forma similar, las sumas de los grados de libertad y las sumas de cuadrados para  $R$  y  $RB$ , en el análisis combinado, son iguales a los de  $R$  en los cortes individuales. También, la  $SC$  total es igual a la suma de las  $SC$  totales para los análisis individuales más  $SC(B)$ .

Si deben analizarse los resultados de varios años con varios cortes cada año para un número dado de variedades, es muy posible que se presente el problema de heterogeneidad y correlación de la varianza del error. Se han propuesto otros métodos de análisis. Steel (16.5) ha propuesto un análisis multivariable, procedimiento que también ha sido usado por Tukey (16.6) al combinar los resultados de un grupo de experimentos.

### 16.6 Modelos de parcelas divididas y de bloques divididos

Sea

$$Y_{ijk} = \mu + \rho_i + \alpha_j + \gamma_{ij} + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk} \quad (16.3)$$

la observación en el bloque  $i$ -ésimo de un diseño de bloques completos al azar, bajo el tratamiento  $j$ -ésimo de la unidad completa con el  $k$ -ésimo tratamiento de la subunidad. Sean  $i = 1, \dots, r$  bloques,  $j = 1, \dots, a$  tratamientos de las unidades completas y  $k = 1, \dots, b$  tratamientos de subunidades. Sean  $\gamma_{ij}$  y  $\varepsilon_{ijk}$  que están distribuidos normal e independiente en torno a medias cero, con  $\sigma^2_\gamma$  como varianza común de los  $\gamma$ , es decir de los componentes aleatorios de unidades enteras, y con  $\sigma^2_\varepsilon$  como varianza común de los  $\varepsilon$ , es decir de los componentes aleatorios de subunidades. Esta representación sirve para el análisis de la varianza del diseño; o de bloques completos de la tabla 16.1 y para el ejemplo de la sec. 16.3.

El modelo puede ser fijo, aleatorio o mixto. Si cualquiera de los  $\alpha$  o los  $\beta$  es aleatorio, entonces los  $(\alpha\beta)$  son aleatorios. Los valores esperados de los cuadrados medios se dan en la tabla 16.9. Observando la tabla se ve claramente cuál es la hipótesis nula que puede probarse usando los errores ( $a$ ) y ( $b$ ). Para los modelos aleatorios y mixtos donde las interacciones son reales, es necesario sintetizar un error en varios casos. Para el procedimiento, ver la sec. 15.5.

En la sección 16.5, se representó una observación mediante

$$Y_{ijk} = \mu + \rho_i + \alpha_j + \beta_k + \theta_{ik} + (\alpha\beta)_{jk} + \varepsilon_{ijk} \quad (16.4)$$

Se ha incluido una componente para bloques por cortes. Supongamos que los  $\rho$  y los  $\alpha$  son aleatorios y que los  $\beta$  son fijos. Este es un modelo mixto. Los valores esperados de los cuadrados medios se dan en la tabla 16.10. Para probar los cortes es necesario sintetizar un error.

**Ejercicio 16.6.1** Para el experimento que dio lugar a la tabla 16.10, supóngase que  $A$  y  $B$  son fijos. Usar las reglas del cap. 15 para obtener los valores esperados para los cuadrados medios en este caso. ( Nótese que  $AB$  también será fijo.)

**Ejercicio 16.6.2** Para un experimento similar al que dio lugar a la tabla 16.8, suponer  $A$  fijo y  $B$  aleatorio. ¿Cuáles serán los valores esperados para los cuadrados medios en este experimento?

### 16.7 Parcelas divididas en espacio y tiempo

Un ejemplo de parcelas divididas en espacio y tiempo es un experimento con cultivo perenne con un diseño de parcelas divididas. Considerar un experimento con forraje para evaluar tratamientos con diferentes fertilizantes, (factor  $A$ ), dispuesto en unidades completas con varias prácticas de manejo (factor  $B$ ), como subunidades y llevado a cabo sin realeatorización para un número de años (factor  $C$ ). Los datos de tales estudios se analizan corrientemente cada año tal como se descubrió para las parcelas divididas en la sec. 16.2. Un análisis combinado de los datos puede efectuarse para determinar las respuestas

Tabla 16.9 Valores esperados de los cuadrados medios para un modelo de parcelas divididas en un diseño de bloques completos al azar

Fuente de variación	gl	Modelo I Efectos fijos	Modelo II Efectos aleatorios
Bloques	$r - 1$	$\sigma_e^2 + b\sigma_y^2 + ab\sigma_p^2$	$\sigma_e^2 + b\sigma_y^2 + ab\sigma_p^2$
$A$	$a - 1$	$\sigma_e^2 + b\sigma_y^2 + rb \frac{\sum \alpha_j^2}{a - 1}$	$\sigma_e^2 + b\sigma_y^2 + r\sigma_{ay}^2 + rba\sigma_p^2$
Error ( $a$ )	$(r - 1)(a - 1)$	$\sigma_e^2 + b\sigma_y^2$	$\sigma_e^2 + b\sigma_y^2$
$B$	$b - 1$	$\sigma_e^2 + ra \frac{\sum \beta_k^2}{b - 1}$	$\sigma_e^2 + r\sigma_{ay}^2 + rao\sigma_p^2$
$AB$	$(a - 1)(b - 1)$	$\sigma_e^2 + r \frac{\sum (\alpha\beta)_{jk}^2}{(a - 1)(b - 1)}$	$\sigma_e^2 + r\sigma_{ay}^2$
Error ( $b$ )	$a(b - 1)(r - 1)$	$\sigma_e^2$	$\sigma_e^2$
Modelo mixto			
$A$ aleatorio, $B$ fijo		$A$ fijo, $B$ aleatorio	
$\sigma_e^2 + b\sigma_y^2 + ab\sigma_p^2$		$\sigma_e^2 + b\sigma_y^2 + ab\sigma_p^2$	
$\sigma_e^2 + b\sigma_y^2 + rba\sigma_p^2$		$\sigma_e^2 + b\sigma_y^2 + r \frac{a}{a - 1} \sigma_{ay}^2 + rb \frac{\sum \alpha^2}{a - 1}$	
$\sigma_e^2 + b\sigma_y^2$		$\sigma_e^2 + b\sigma_y^2$	
$\sigma_e^2 + r \frac{b}{b - 1} \sigma_{ay}^2 + ra \frac{\sum \beta^2}{b - 1}$		$\sigma_e^2 + rao\sigma_p^2$	
$\sigma_e^2 + r \frac{b}{b - 1} \sigma_{ay}^2$		$\sigma_e^2 + r \frac{a}{a - 1} \sigma_{ay}^2$	
$\sigma_e^2$		$\sigma_e^2$	

promedio de los tratamientos para todos los años y ver si estas respuestas son coherentes de año en año. La forma del análisis se presenta en la tabla 16.11, donde  $Y_{ijkm}$  representa la observación hecha en el bloque  $i$ -ésimo de la subunidad que recibe el tratamiento consistente en el nivel  $j$ -ésimo del factor  $A$  y el nivel  $k$ -ésimo del factor  $B$  en el año  $m$ -ésimo, factor  $C$ . Entonces  $Y_{...}$  es un total de un bloque para todas las subunidades y años; es un total de  $abc$  observaciones. También  $Y_{jk.}$  es el total para la combinación de tratamientos  $j$ ,  $k$ -ésimo para todos los bloques y años; es un total de  $rc$  observaciones. En forma similar se dan otros totales.

Las partes I y II de la tabla 16.11 dan la partición de los grados de libertad y las sumas de cuadrados para la porción del análisis correspondiente al promedio de las respuestas a los tratamientos para todos los años. El procedimiento es idéntico al que se des-

**Tabla 16.10 Valores esperados de los cuadrados medios para un modelo posible para los datos de la tabla 16.8**

Modelo: los  $\rho$  y los  $\alpha$  aleatorios, los  $\beta$  fijos

Fuente de variación	gl	Valores esperados
Bloques	$r - 1$	$\sigma_e^2 + a\sigma_\theta^2 + b\sigma_\gamma^2 + ab\sigma_\rho^2$
Variables, $A$	$a - 1$	$\sigma_e^2 + b\sigma_\gamma^2 + rba\sigma_\alpha^2$
Error ( $a$ ), $RA$	$(r - 1)(a - 1)$	$\sigma_e^2 + b\sigma_\gamma^2$
Cortes, ( $B$ )	$b - 1$	$\sigma_e^2 + a\sigma_\theta^2 + r \frac{b}{b - 1} \sigma_{\alpha\theta}^2 + ra \frac{\sum \beta_k^2}{b - 1}$
Error ( $b$ ), $RB$	$(r - 1)(b - 1)$	$\sigma_e^2 + a\sigma_\theta^2$
$AB$	$(a - 1)(b - 1)$	$\sigma_e^2 + r \frac{b}{b - 1} \sigma_{\alpha\theta}^2$
Error ( $c$ ), $RAB$	$(r - 1)(a - 1)(b - 1)$	$\sigma_e^2$

cribe para el diseño de unas parcelas divididas en las secc. 16.2 y 16.3, excepto por la adición de la constante  $c$  en los divisores. Los totales usados en los cálculos son para  $c$  años. Las partes III y V son extensiones de los análisis presentados en la sec. 16.5.

Los errores estándar para las comparaciones entre medias de tratamientos para todos los años son los mismos que los de la tabla 16.2, excepto que  $c$  se incluye en el divisor ya que las medias a compararse cubren  $c$  años. Las mismas precauciones que se discutieron en la sec. 16.5 se aplican a las comparaciones entre tratamientos dentro y entre años individuales. Como el análisis completo y la interpretación de los experimentos en que entran unidades divididas en espacio y en tiempo puede ser complejo, se sugiere al lector buscar asesoría al respecto.

### 16.8 Series de experimentos semejantes

Muchos experimentos agrícolas se llevan a cabo en varias localidades y durante varios años. Esta práctica es muy común en evaluaciones de variedades de diferentes cultivos. El propósito es obtener información que permite hacer recomendaciones para años futuros en una zona amplia. Tanto localidades como años pueden considerarse como tipos amplios de replicación, donde las localidades son replicaciones o muestras del área para la cual se desea la información y los años son replicaciones o muestras de años futuros.

Una serie de experimentos similares puede efectuarse para determinar el efecto de diferentes condiciones ambientales, tales como diferentes longitudes del día en una respuesta, tal como el crecimiento de las plantas o el número de yodo en lino; o para determinar si diferentes llantas se comportan en forma parecida bajo diferentes condiciones de conducción. La repetición de experimentos semejantes bajo diferentes condiciones es esencial para proveer variación en las condiciones externas en estudio. Así mismo, puede llevarse a cabo un experimento en varios laboratorios para evaluar un producto por un

• Tabla 16.11 Análisis de la varianza para parcelas divididas en espacio y tiempo

Referencia No.	Fuente	gl	SC
I	Bloques, $R$	$r - 1$	$\sum_i Y_{i..}^2 / abc - FC$
	fertilizantes, $A$	$a - 1$	$\sum_j Y_{j..}^2 / rbc - FC$
	Error ( $a$ ), $RA$	$(r - 1)(a - 1)$	$\sum_{i,j} Y_{ij..}^2 / bc - FC - SC(R) - SC(A)$
	Subtotal I	$ra - 1$	$\sum_{i,j} Y_{ij..}^2 / bc - FC$
	Prácticas de manejo, $B$	$b - 1$	$\sum_k Y_{k..}^2 / rac - FC$
II	$AB$	$(a - 1)(b - 1)$	$\sum_{j,k} Y_{jk..}^2 / rc - FC - SC(A) - SC(B)$
	Error ( $b$ ), $RB + RAB$	$(r - 1)a(b - 1)$	$\sum_{i,j,k} Y_{ijk..}^2 / c - FC - SC(I) - SC(B) - SC(AB)$
	Subtotal I + II	$rab - 1$	$\sum_{i,j,k} Y_{ijk..}^2 / c - FC$

	Años, C	$c - 1$	$\sum_{i,m} Y_{i..m}^2 / rab - FC$
III	$E(c), RC$	$(r - 1)(c - 1)$	$\sum_{i,m} Y_{i..m}^2 / ab - FC - SC(R) - SC(C)$
	Subtotal III	$rc - 1$	$\sum_{i,m} Y_{i..m}^2 / ab - FC$
IV	$AC$	$(a - 1)(c - 1)$	$\sum_{j,m} Y_{j..m}^2 / rb - FC - SC(A) - SC(C)$
	$E(d), RAC$	$(r - 1)(a - 1)(c - 1)$	$\sum_{i,j,m} Y_{ij..m}^2 / b - FC - SC(I) - SC(C) - SC(E_t) - SC(AC)$
	Subtotal I + III + IV	$rac - 1$	$\sum_{i,j,m} Y_{ij..m}^2 / b - FC$
V	$BC$	$(b - 1)(c - 1)$	$\sum_{k,m} Y_{...km}^2 / ra - FC - SC(B) - SC(C)$
	$ABC$	$(a - 1)(b - 1)(c - 1)$	$\sum_{j,k,m} Y_{j..km}^2 / r - FC - SC(A) - SC(B)$
	$E(e), RBC + RABC$	$(r - 1)a(b - 1)(c - 1)$	$-SC(C) - SC(AB) - SC(AC) - SC(BC)$ $\sum_{i,j,k,m} Y_{ijk..m}^2 - FC - SC(I + II + III + IV)$ $-SC(BC) - SC(ABC)$
	Gran total	$rabc - 1$	$\sum_{i,j,k,m} Y_{ijk..m}^2 - FC$

ensayo biológico o un procedimiento de determinaciones químicas; el objeto puede ser el asegurarse de la exactitud en los distintos laboratorios y si se llega o no a las mismas conclusiones.

El procedimiento para el análisis de tales datos varía con los objetivos. Los análisis preliminares son generalmente iguales. Sin embargo, los análisis finales difieren y suelen ser muy complejos, los detalles rebasan el campo de este texto. El lector puede consultar a Cochran y Cox (16.2), Federer (16.3), Steel (16.5) y Tukey (16.6), donde a su vez pueden encontrarse otras referencias. Aquí sólo se exponen algunos de los puntos más importantes que tienen que ver con el análisis de una serie de experimentos agrícolas.

Los efectos de los tratamientos, en experimentos agrícolas repetidos en varias localidades por un período de años, suelen ser fijos, es decir, no son conjuntos aleatorios de tratamientos, sino que los selecciona con anterioridad el experimentador como los que tienen mayores posibilidades de éxito. Los efectos de localidades y años se consideran corrientemente como aleatorios, donde las localidades corresponden a un conjunto aleatorio de las posibles localidades y los años un conjunto aleatorio de años futuros. En la práctica, esos supuestos rara vez, si acaso, se cumplen. En la mayoría de las situaciones, las localidades usadas no se seleccionan aleatoriamente sino que son estaciones experimentales o campos de ubicación permanente en el área para la cual se desea la información. Se supone que tales localidades son como mínimo representativas de tipos especiales de suelo o áreas. Las mismas localidades se usan año a año mientras dure el experimento. Con cultivos anuales, generalmente se usan diferentes campos en las estaciones experimentales seleccionadas. Finalmente, una muestra de varios años sucesivos no siempre será representativa de años futuros.

Las varianzas del error a menudo difieren considerablemente de una prueba a otra. Cuando se usa en el denominador de  $F$  el término de error combinado, tal heterogeneidad tiende a invalidar la prueba  $F$  para comparaciones en las que entran la interacción de tratamientos con localidades y con años. El resultado es que la prueba  $F$  puede producir demasiados resultados significantes, y en un experimento dado, el nivel de probabilidad establecido puede ser bastante incorrecto si la heterogeneidad es extrema. Un enfoque prudente al probar interacciones en que intervengan tratamientos consiste en comparar el  $F$  calculado con el  $F$  tabulado para  $(t - 1)$  y  $n$  grados de libertad, donde  $t$  es el número de tratamientos, y  $n$  el de grados de libertad para el error en una sola prueba, en vez de compararlo con el  $F$  tabulado para los grados de libertad de la interacción y el error combinado. Los verdaderos números de grados de libertad para la distribución de tal razón que dan en algún punto entre los dos extremos. Si esta prueba resulta significante hay pocas dudas; la dificultad está en la decisión respecto a los valores  $F$  que se encuentran entre los extremos.

La heterogeniedad de la interacción frecuentemente complica la interpretación de los datos de experimentos semejantes. Esto se produce cuando ciertos tratamientos difieren apreciablemente de ensayo a ensayo mientras otros no. Tal heterogeneidad invalida las pruebas  $F$  de tratamientos frente a interacción, de una forma similar a la descrita para probar la interacción con el error combinado. En algunos casos, puede ser útil una subdivisión de las comparaciones de tratamientos e interacciones en componentes de interés especial.

En la interpretación de los datos de ensayos de variedades llevados a cabo durante un período de años, es muy útil estudiar cuidadosamente las medias de las variedades en

las diferentes localidades. O sea, los análisis deben llevarse a cabo para cada estación, para cada año y para cada estación durante los varios años. Usualmente se espera una interacción de variedades por localidades, especialmente si las localidades o variedades difieren grandemente. Tal interacción no quiere decir que todas las variedades reaccionen en forma diferente para todas las estaciones, sino que sólo algunas lo hacen en ciertas estaciones. Aun así, el cambio puede ser sólo de la magnitud de las diferencias, no que varíen el orden o rango. Una determinada variedad puede ubicarse en el primer puesto en todas o casi todas las localidades a pesar de una interacción muy grande de variedades por localidad. En este caso, siendo todo lo demás igual, la sola variedad podría recomendarse para toda el área cubierta por la prueba. Sin embargo, si una variedad se comporta bien en algunas localidades pero no en otras, entonces podrían hacerse recomendaciones en diferentes partes del área.

Otro punto tiene que ver con la variabilidad existente dentro de las variedades de una prueba a otra. Dos variedades pueden dar aproximadamente el mismo resultado en cada una de las diversas localidades. Sin embargo, una puede variar considerablemente de año en año mientras que otra puede ser bastante constante. Si pudieramos hacer predicciones, anticipándonos un año y con razonable exactitud, sobre las condiciones climáticas que probablemente se encontrarian, entonces se podrían hacer recomendaciones específicas respecto a qué variedad sembrar. Como esto no es del todo posible actualmente, entonces se le ha de recomendar al agricultor interesado en un rendimiento razonablemente uniforme en cada estación que plante la variedad que es constante.

**Ejercicio 16.8.1** Se compararon doce cepas de soya en un experimento de bloques completos al azar, con 3 bloques en 3 localidades en el estado de Carolina del Norte. Los rendimientos se dan en gramos por parcela.

Variedad	Plymouth			Clayton			Clinton		
	1	2	3	1	2	3	1	2	3
Tracy	1,307	1,365	1,542	1,178	1,089	960	1,583	1,841	1,464
Centennial	1,425	1,475	1,276	1,187	1,180	1,235	1,713	1,684	1,378
N72-137	1,289	1,671	1,420	1,451	1,177	1,723	1,369	1,608	1,647
N72-3058	1,250	1,202	1,407	1,318	1,012	990	1,547	1,647	1,603
N72-3148	1,546	1,489	1,724	1,345	1,335	1,303	1,622	1,801	1,929
R73-81	1,344	1,197	1,319	1,175	1,064	1,158	1,800	1,787	1,520
D74-7741	1,280	1,260	1,605	1,111	1,111	1,099	1,820	1,521	1,851
N73-693	1,583	1,503	1,303	1,388	1,214	1,222	1,464	1,607	1,642
N73-877	1,656	1,371	1,107	1,254	1,249	1,135	1,775	1,513	1,570
N73-882	1,398	1,497	1,583	1,179	1,247	1,096	1,673	1,507	1,390
N73-1102	1,586	1,423	1,524	1,345	1,265	1,178	1,894	1,547	1,751
R75-12	911	1,202	1,012	1,136	1,161	1,004	1,422	1,393	1,342

**Fuente:** Datos cortesía de C.A. Brim, Universidad del Estado de Carolina del Norte, Raleigh, Carolina del Norte.

Elaborar un análisis de la varianza para cada localidad.

Combinar los análisis. Se tendrán ahora SC(localidades) y SC (bloques de las localidades) que explican la variación entre los nueve bloques. SC(localidades) es nueva. Las tres SC(sepas) dentro de localidades deberá particionarse de modo más informativo en SC(cepas) con 11 grados de libertad y SC(localidades X cepas) con 22 gl. Las tres componentes de SC(error) se combinan lo mismo que sus grados de libertad.

¿Qué conclusiones útiles se pueden sacar de análisis combinado?

**Ejercicio 16.8.2** Supóngase que los datos del ejercicio 7.3.3 se han presentado desde el siguiente punto de vista: Cada conjunto de datos pre y post corresponde a un experimento individual que tiene que ver con la evaluación de una sola experiencia aislada.

Analizar cada conjunto de datos.

Combinar los 3 análisis en uno solo. ¿De qué información suplementaria se dispone ahora que no se pudo encontrar en los análisis originales.

Comparar el análisis combinado con el del ejercicio 16.3.4.

## Referencias

- 16.1. Anderson, R. L.: "Missing-plot techniques," *Biom. Bull.*, 2:41-47 (1946).
- 16.2. Cochran, W. G., y G. M. Cox: *Experimental Designs*, 2a. ed., Wiley, Nueva York, 1957.
- 16.3. Federer, W. T.: *Experimental Design*. Macmillan, Nueva York, 1955.
- 16.4. Khargonkar, S. A.: "The estimation of missing plot values in split-plot and strip trials," *J. Ind. Soc. Agr. Statist.*, 1:147-161 (1948).
- 16.5. Steel, R. G. D.: "An analysis of perennial crop data," *Biom.*, 11:201-212 (1955).
- 16.6. Tukey, J. W.: "Diadic anova," *Hum. Biol.*, 21:65-110 (1949).

## ANALISIS DE LA COVARIANZA

### 17.1 Introducción

El análisis de la covarianza trata de dos o más variables medidas y donde cualquier variable independiente medible no se encuentra a niveles predeterminados, como en un experimento factorial. Hace uso de conceptos tanto del análisis de varianza como de la regresión. Este capítulo trata la covarianza lineal. A menudo, una relación lineal es una aproximación razonablemente buena para una relación no lineal con tal que los valores de las variables independientes no cubran un intervalo muy amplio.

### 17.2 Usos del análisis de la covarianza

Los usos más importantes del análisis de la covarianza son:

1. Controlar el error y aumentar la precisión.
2. Ajustar medias de tratamientos de la variable dependiente a las diferencias en conjuntos de valores de variables independientes correspondientes.
3. Ayudar en la interpretación de datos, especialmente en lo concerniente a la naturaleza de los efectos de los tratamientos.
4. Particionar una covarianza total o suma de productos cruzados en componentes.
5. Estimar datos faltantes.

Estos usos se exponen ahora con más detalle.

1. *Control del error* La varianza de una media de tratamiento es  $\sigma_{\bar{Y}}^2 = \sigma^2/n$ . Así, para disminuir esta varianza, sólo tenemos dos enfoques: el aumento del tamaño de la muestra o el control de la varianza en una población muestreada.

El control de  $\sigma^2$  se logra mediante el diseño experimental o mediante el uso de una o más covariables. Ambos métodos pueden usarse simultáneamente. Cuando se usa la covarianza como método para reducir el error, esto es, de controlar  $\sigma^2$ , se hace reconociendo el hecho de que la variación observada de la variable dependiente  $Y$  es parcialmente atribuible a la variación de la variable independiente  $X$ . El uso de la covariable exige el uso de las técnicas de regresión de los caps. 10, 13 y 14.

El uso de la covarianza para controlar el error es un medio de aumentar la precisión con la cual los efectos de los tratamientos pueden medirse eliminando, por regresión, ciertos efectos reconocidos que no pueden ser o no han sido controlados efectivamente por el diseño experimental. Por ejemplo, en un experimento de nutrición animal para comparar el efecto de varias raciones en el momento de peso, los animales asignados a un bloque varían en peso inicial. Ahora, si el peso inicial está correlacionado con la ganancia de peso, una porción del error experimental en la ganancia puede deberse a diferencias en el peso inicial. Mediante el análisis de la covarianza, esta porción, una contribución que puede atribuirse a diferencias en el peso inicial puede calcularse y eliminarse del error experimental para ganancia.

Forester (17.6) ilustra la aplicación de la covarianza al control del error en datos en campos donde la variabilidad ha sido aumentada mediante experimento previo. Federer y Schlottfeldt (17.4) usan la covarianza como un substituto del uso de bloques para controlar gradientes en el material experimental. Outhwaite y Rutherford (17.9) generalizaron los resultados de Federer y Schlottfeldt.

*2. Ajuste de medias de tratamientos* Cuando la variación observada en  $Y$  puede atribuirse parcialmente a la variación en  $X$ , la variación entre las  $\bar{Y}$  de los tratamientos también debe afectarse por las  $\bar{X}$  de los tratamientos. Para que sean comparables, las  $\bar{Y}$  de los tratamientos deberán ajustarse para hacer de ellas las mejores estimaciones de lo que hubieran sido si todas las  $\bar{X}$  de los tratamientos hubiesen sido iguales. Si el objeto principal de la covarianza es ajustar las  $\bar{Y}$  de tratamiento, es también en reconocimiento de una situación de regresión que se exige el correspondiente ajuste del error. En todo caso, es necesario medir la regresión apropiada, independientemente de otras fuentes de variación que pueden invocarse en el modelo.

La idea general es evidente en la fig. 17.1 para dos tratamientos. Para cada tratamiento, se ve que la variación de  $X$  contribuye a la variación de  $Y$ . Así pues, se ve la necesidad de controlar la varianza del error mediante el uso de la covariable. Al mismo tiempo, la distancia entre  $\bar{X}_1$  y  $\bar{X}_2$  puede contribuir grandemente a la diferencia entre  $\bar{Y}_1$  y  $\bar{Y}_2$ . Si las  $Y$  de los tratamientos se han observado a partir de una  $\bar{X}$  común, digamos  $X_0$ , entonces serán comparables. Así pues, ajustar las medias de los tratamientos es evidente.

Como ilustración consideremos las arvejas de enlatar. Este cultivo aumenta rápidamente el rendimiento con el aumento en madurez. En un ensayo para evaluar los rendimientos de diferentes variedades, es difícil cosecharlas todas en el mismo estado de madurez. Un análisis de rendimientos no ajustados a las diferencias en madurez puede tener poco valor. Sin embargo, la madurez puede usarse como una covariable cuando se mide mediante un instrumento mecánico, el tenderómetro, que mide la presión necesaria para punzar las arvejas. Una comparación de rendimientos ajustados a diferencias en madurez tendría más sentido que una comparación entre rendimientos no ajustados.

En experimentos de campo, los rendimientos pueden ajustarse a las diferencias en la productividad de las parcelas determinada mediante ensayos de uniformidad. Un ensayo

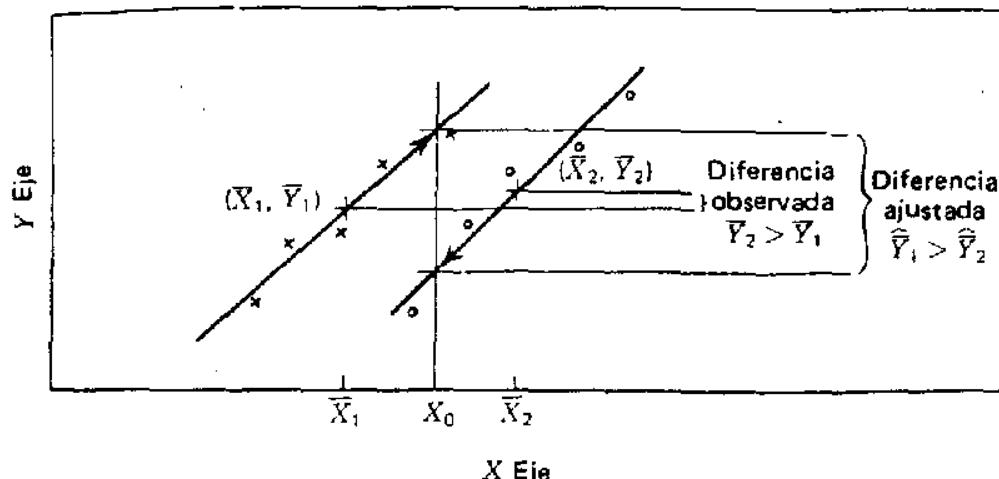


Figura 17.1 Control del error y ajuste de las medias de tratamientos mediante la covarianza.

de esta naturaleza mide los rendimientos de las parcelas tratadas uniformemente antes del resultado del experimento principal. Love (17.7) concluye que con los cultivos anuales la mayor precisión que se logra con el uso de datos de uniformidad rara vez vale la pena; sin embargo, con cultivos perennes, tales como cultivos de árbol, hay mucho que ganar.

En experimentos de nutrición animal, las diferencias entre medias de tratamientos no ajustadas puede obedecer a las diferencias en el valor nutritivo de las raciones, a diferencias en las cantidades consumidas, o a ambas. Si las diferencias entre ganancias medias de peso para las diferentes raciones se ajustan a un consumo común de alimento, las medias ajustadas indican si las raciones difieren o no en valor nutritivo. Aquí al proveer información de base sobre la forma como los tratamientos producen los efectos, la covarianza toca los principios fundamentales de los resultados de la investigación.

*3. Interpretación de datos* Todo procedimiento aritmético y técnica estadística asociada se proponen contribuir a la interpretación de datos. Así, los usos 1 y 2 definitivamente tienen que ver con interpretación de datos. Sin embargo, se piensa que el uso de 3 sea más específico en cuanto que el análisis de la covarianza a menudo ayuda al experimentador a entender los principios que fundamentan los resultados de una investigación. Por ejemplo, puede ser bien sabido que ciertos tratamientos producen efectos tanto en la variable dependiente como en las independientes. La covarianza, como medio de controlar el error y ajustar medias de tratamientos, se usa primordialmente cuando la variable independiente mide efectos ambientales y en ella no influyen los tratamientos. Pero si ocurre así, la interpretación de los datos cambia. Esto es así porque las medias de tratamientos para la variable independiente son las mismas. El ajuste elimina parte de los efectos de tratamientos cuando las medias de la variable independiente están afectadas por los tratamientos. La covarianza debe usarse con precaución.

En un ensayo de fertilización en remolacha azucarera, los tratamientos pueden producir diferencias en densidad. Cuando la densidad, la variable independiente, está afectada por los tratamientos, el análisis del rendimiento ajustado a las diferencias elimina parte del efecto de tratamiento y entonces el experimentador puede desorientarse en la interpretación de los datos. Aun así, el análisis de la covarianza puede proporcionar información

útil. El rendimiento total es función del peso promedio por remolacha y de la densidad. Ahora, si en la densidad influye los tratamientos, el análisis de la covarianza del rendimiento ajustado a la cosecha mide esencialmente los efectos de los tratamientos sobre el peso promedio de las remolachas. Un valor significante de  $F$  para los rendimientos ajustados a las diferencias de densidad indicaría que los tratamientos afectan en promedio los pesos individuales de las remolachas.

Para cultivos como la remolacha azucarera, en los que la variación en número de plantas por parcela es aleatoria y hay correlación entre el número y el rendimiento, el error experimental del rendimiento aumentará con la variación aleatoria de la covariable. Para compensar este aumento y ajustar las medias de rendimiento, a veces se hace un ajuste en proporción al número de plantas. Este procedimiento no es recomendable ya que produce por lo general corrección excesiva en las parcelas con la densidad más baja puesto que los rendimientos raramente son proporcionales al número de plantas por parcela. O sea que la verdadera regresión del rendimiento respecto de la densidad, para cualquier tratamiento, rara vez pasa por el origen. El análisis de covarianza proporciona un método más apropiado y satisfactorio de ajuste de los datos experimentales.

En situaciones en las que se presentan diferencias reales entre los tratamientos para la variable independiente, pero que no son el efecto directo de los tratamientos, se justifica el ajuste. Por ejemplo, considérese un ensayo varietal para el cual se ha producido semilla de diferentes variedades en diferentes áreas. Tal semilla puede tener una germinación muy diferente, no debido a causas intrínsecas de las mismas, sino como resultado de las condiciones ambientales en las cuales crecieron. En consecuencia, las diferencias en densidad pueden ocurrir aun cuando se controle la tasa de siembra. En esta situación, está justificado el uso de la covarianza tanto para control del error como para ajuste de rendimiento.

El problema de si la covarianza es o no aplicable, es algo que el experimentador debe juzgar con cuidado. Además, debe tenerse extremo cuidado en la interpretación de las diferencias entre tratamientos ajustados. De Lury (17.3) presenta un estudio de un caso en el que una variable independiente está influida por los tratamientos.

*4. La partición de una covarianza total* Tal como ocurre en el análisis de la varianza, se partitiona una suma de productos. Si bien este uso es el título del capítulo, a menudo es una parte incidental, aunque necesaria, de un análisis de la covarianza.

Una covarianza proveniente de un experimento replicado se partitiona cuando queremos determinar la relación entre dos o más variables medidas cuando en la relación no influyen otras fuentes de variación. Por ejemplo, considérense los datos de un diseño en bloques completos al azar con cuatro bloques de 25 líneas aleatorias de soya para los cuales se desea determinar la relación entre el contenido de aceite y el de proteína. La suma total de productos de las 100 observaciones puede partitionarse en componentes de acuerdo con las fuentes de variación, o sea bloques, líneas y error. Las componentes tienen 3, 24 y 72 grados de libertad, respectivamente. Si los distintos coeficientes de regresión y correlación correspondientes a estas fuentes difieren significativamente, entonces la regresión y correlación totales son heterogéneas y no interpretables. Para este experimento, el interés radicaría en la relación para las medias de las líneas y para el residuo, donde éste mide la relación promedio entre las dos variables observadas, dentro de las líneas, tras eliminar los efectos de bloques globales. A menudo hay buenas razones para suponer que estas dos regresiones serán diferentes.

*5. Estimación de datos faltantes* Las fórmulas dadas anteriormente para estimar datos faltantes dan lugar a una suma de cuadrados residual mínima. Pero la suma de cuadrados de tratamientos presenta un sesgo hacia arriba. El uso de la covarianza para estimar las parcelas faltantes lleva a una suma residual de cuadrados mínima más una de cuadrados de tratamientos no sesgada. El procedimiento de la covarianza es sencillo, pero más difícil de describir que los procedimientos anteriores que exigían poco más que una fórmula.

### 17.3 El modelo y los supuestos para la covarianza

Los supuestos para la covarianza son una combinación de los supuestos para el análisis de la varianza y de la regresión lineal. El modelo aditivo lineal para un diseño dado es el correspondiente al análisis de la varianza más un término adicional para la variable concomitante o independiente. Así, para el diseño de bloques al azar con una observación por celda, la descripción matemática está dada por

$$Y_{ij} = \mu + \tau_i + \rho_j + \beta(X_{ij} - \bar{X}_{..}) + \varepsilon_{ij} \quad (17.1)$$

La variable dependiente, que se está analizando, generalmente se denota  $Y$ , mientras que la variable usada en el control del error y el ajuste de medias, la variable independiente o covariante, se denota  $X$ .

Es interesante reescribir esta expresión en las siguientes formas:

$$\begin{aligned} Y_{ij} - \beta(X_{ij} - \bar{X}_{..}) &= \mu + \tau_i + \rho_j + \varepsilon_{ij} \\ Y_{ij} - \tau_i - \rho_j &= \mu + \beta(X_{ij} - \bar{X}_{..}) + \varepsilon_{ij} \end{aligned}$$

En el primer caso, hacemos énfasis en los aspectos del diseño experimental del problema. Deseamos efectuar un análisis de la varianza de valores que han sido ajustados a la regresión respecto de una variable independiente. Se está destacando el uso 2 (sec. 17.2), aunque obviamente tenemos en mente los usos 1 y 3.

En el segundo caso, se acentúa el enfoque de la regresión. Deseamos medir la regresión de  $Y$  con respecto a  $X$  sin la interferencia de efectos de bloques y tratamientos. Ahora nos interesa más el uso 4. Observar que si  $X$  no se midiera, entonces  $\beta(X_{ij} - \bar{X}_{..})$  no se podría determinar y quedaría pues incluida en el término residual o de error.

Los supuestos necesarios para el uso válido de la covarianza son :

1. Los  $X$  son fijos, medidos sin error, e independientes de los tratamientos.
2. La regresión de  $Y$  respecto a  $X$ , después de eliminar diferencias de bloques y tratamientos, es lineal e independiente de tratamientos y bloques.
3. Los residuos se distribuyen normal e independientemente con media cero y varianza común.

El supuesto 1 establece que los  $X$  son fijos. Esto quiere decir que, en la obtención de los valores sesgados, se repite el mismo conjunto de  $X$ . A su turno, las inferencias sólo se aplican al conjunto de los  $X$  realmente observados. Mientras que los  $X$  no se seleccionen

exactamente o se visualicen como idénticos en muestreo real repetido, entonces las inferencias se harán para valores interpolados en vez de extrapolados. También, los  $X$  deben medirse exactamente de tal forma que las medias de población se identifiquen con propiedad. En realidad, la medición del error va a ser simplemente trivial en relación con la variación observada. El supuesto 1 también aplica, al igual que lo establece, que para el uso normal de la covarianza, los tratamientos no afectarán a los valores  $X$ , porque al fijarlos, pueden escogerse o restringirse por razones de comodidad. Ya se ha indicado que la covarianza puede usarse cuando los valores  $X$  están así afectados, pero debe usarse con prudencia.

El supuesto 2 establece que el efecto de  $X$  sobre  $Y$  es aumentar o disminuir todo  $Y$ , en promedio, en un múltiplo constante de la desviación del correspondiente  $X$  respecto de la  $\bar{X}_{..}$  para todo el experimento, es decir, en  $\beta(X_{ij} - \bar{X}_{..})$ . Se supone que la regresión es estable u homogénea. Así, no se requiere subíndice en  $\beta$  para relacionarlo con bloques o tratamientos. Un caso de tal relación se expone en la sec. 17.9.

El supuesto 3 es aquél del cual depende la validez de las pruebas usuales,  $t$  y  $F$ . Un análisis, tal como lo determina el modelo, da una estimación válida de la varianza común cuando se han aleatorizado los tratamientos dentro de los bloques. El supuesto de normalidad no es necesario para estimar las componentes de la varianza en  $Y$ ; pero es necesaria la aleatorización.

La varianza residual se estima con base en los estimadores mínimos cuadrados de  $\mu$ , los  $\tau_i$ , los  $\rho_j$  y  $\beta$  indicados con acentos circunflejos. La ec. (17.2) cumple

$$\sum_{i,j} [Y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\rho}_j - \hat{\beta}(X_{ij} - \bar{X}_{..})]^2 = \text{mínimo} \quad (17.2)$$

Aquí ajustamos el modelo completo, al cual se apela cuando todas las hipótesis alternas aplicadas son verdaderas. Para las estimaciones por mínimos cuadrados, es correcta la ec. (17.3)

$$\sum_{i,j} [Y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\rho}_j - \hat{\beta}(X_{ij} - \bar{X}_{..})] = 0 \quad (17.3)$$

La suma de todas las desviaciones es cero. No es necesario obtener y usar las estimaciones de los parámetros en la ec. (17.2) como tampoco para el análisis de bloques completos al azar sin covarianza. Sin embargo, las ecs. (17.4) a (17.6) definen las estimaciones y dan la varianza residual

$$\hat{\mu} = \bar{Y} \quad (17.4)$$

$$\hat{\tau}_i = t_i = \bar{Y}_i - \bar{Y}_{..} - b(\bar{X}_i - \bar{X}_{..}) \quad (17.4)$$

$$\hat{\rho}_j = r_j = \bar{Y}_j - \bar{Y}_{..} - b(\bar{X}_{.j} - \bar{X}_{..})$$

$$\hat{\beta} = b = \frac{E_{XY}}{E_{XX}} \quad (17.5)$$

$$\hat{\sigma}_{Y_{..} X}^2 = s_{Y_{..} X}^2 = \frac{E_{YY} - (E_{XY})^2/E_{XX}}{f_e} \quad (17.6)$$

donde  $E_{xx}$ ,  $E_{xy}$  y  $E_{yy}$  son sumas de productos ajustadas al error; por ejemplo  $E_{xx}$  es la suma de cuadrados del error para  $X$  y  $f_e$  los grados de libertad para el error. Puede verse en la segunda de las ecs. (17.4) que para estimar el efecto de tratamiento  $\tau_i$ , la desviación de toda media de tratamiento respecto de la media general debe ajustarse en la cantidad  $b(\bar{X}_i - \bar{X}_{..})$ . Este ajuste elimina todo efecto atribuible a la variable  $X$ . Son las medias de tratamientos ajustadas las que son comparables. Todas las estimaciones son coeficientes de regresión parcial como en cap. 14. Se deberán introducir variables ficticias para  $\hat{\mu}$ , los  $\hat{\tau}$  y los  $\hat{\rho}$  para que la analogía sea completa. La ec. (17.6) es la fórmula de cálculo para lo que corresponde a la fórmula de definición, dada como ec. (17.2).

#### 17.4 Prueba de medias de tratamientos ajustadas

La tabla 17.1 da el análisis de la covarianza para un diseño de bloques completos al azar y, al mismo tiempo, ilustra el procedimiento general. Obsérvese la nueva notación con letras mayúsculas y subíndices pareados para indicar sumas de productos; un cuadrado es un tipo de producto particular.

La lógica del procedimiento depende del ajuste de modelos mediante técnicas de regresión múltiple. Una analogía estricta exige la inclusión de  $r - 1$  y  $t - 1$  variables ficticias para efectos de bloques y tratamientos, respectivamente, lo mismo que todas las covariables medidas. Nuestro interés se centra en el aspecto de la regresión donde  $SC(\text{total}, \text{ajustada})$  se partitiona en componentes atribuibles a la regresión y al error o residuo. Esto debe hacerse para el *modelo completo*, el que está dentro de  $H_1$ , y de nuevo para el *modelo reducido*, que está dentro de  $H_0$ . La reducción adicional debido a la introducción, en el modelo, del conjunto de parámetros que se prueban mediante

$$\begin{aligned} SC(\text{regresión}|H_1) - SC(\text{regresión}|H_0) \\ = SC(\text{residuos}|H_0) - SC(\text{residuos}|H_1) \end{aligned}$$

recuérdese que

$$SC(\text{regresión}|H_1) = SC(\{\tau_i\}, \{\rho_j\}, \beta)$$

La "reducción adicional" en la forma de cuadrado medio se prueba con  $CM(\text{residuos}|H_1)$

La tabla 17.1 esquematiza el proceso. Primero se ajusta el modelo completo. Se emplea un proceso secuencial, tal como se expuso en la sec. 14.7 con las salidas de computador. La secuencia seguida ajusta efectos de bloques y tratamientos en primer lugar. Esto es cómodo debido a la ortogonalidad. La columna  $Y$ ,  $Y$  se calcula con el residuo

$$E_{yy} = SC(\text{Total, ajustado}) - SC(\{\tau_i\}, \{\rho_j\})$$

aún sin ajustar por la regresión con respecto a  $X$ . Para esto también se necesitan los residuos  $E_{xx}$  y  $E_{xy}$ . La idea se discutió en la sec. 14.7 cuando se consideró la segunda parte de la tabla 14.5. Ahora úsese la línea del "Error" para calcular la contribución atribuible a la regresión ajustada para bloques y tratamientos, es decir,

$$\frac{E_{xy}^2}{E_{xx}} = SC(\beta | \{\tau_i\}, \{\rho_j\}).$$

**Tabla 17.1 Prueba de medias de tratamiento ajustadas**  
**Análisis de la covarianza para el diseño de bloques completos al azar**

Fuente	gl	Sumas de productos de			$\Sigma(Y - \bar{Y})^2$	CM
		X, X	Y, Y	g!		
Total	$rt - 1$	$\sum(X - \bar{X})^2$	$\sum(X - \bar{X})(Y - \bar{Y})$	$\sum(Y - \bar{Y})^2$		
Bloques	$r - 1$	$R_{xx}$	$R_{yy}$	$R_{xy}$		
Tratamientos	$t - 1$	$T_{xx}$	$T_{yy}$	$T_{xy}$		
Error	$(r - 1)(t - 1)$	$E_{xx}$	$E_{yy}$	$E_{xy}$	$(r - 1)(t - 1) - 1$	$E_{yy} - \frac{(E_{xy})^2}{E_{xx}}$
Tratamientos	$r(t - 1)$ + error	$S_{xx}$	$S_{yy}$	$S_{xy}$	$r(t - 1) - 1$	$S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$
Tratamientos ajustados	$t - 1$				$\left[ S_{yy} - \frac{(S_{xy})^2}{S_{xx}} \right]$ $- \left[ E_{yy} - \frac{(E_{xy})^2}{E_{xx}} \right]$	CM( $T$ , ajustados)

Compárese con la ec. (10.18). La suma de cuadrados del error o residuo final es

$$\sum (Y - \hat{Y})^2 = SC(\text{total, ajustada}) - SC(\{\tau_i\}, \{\rho_j\}, \beta)$$

donde

$$SC(\{\tau_i\}, \{\rho_j\}, \beta) = T_{YY} + R_{YY} + \frac{E_{XY}^2}{E_{XX}}.$$

Finalmente, tenemos  $CME = s_{Y \cdot X}^2$  para el modelo completo. El coeficiente de regresión parcial de  $Y$  con respecto a  $X$  está dado por  $b = E_{XY}/E_{XX}$ ; estima el  $\beta$  de la ec. (17.1).

Segundo, se ajusta el modelo reducido. Aquí, no hay efectos de tratamiento; en el modelo sólo hay efectos de bloques y de regresión con respecto a  $X$ . Comiéncese por ajustar por bloques, pero tratamientos no, ya que los últimos ya no se encuentran en el modelo. Se necesitan  $R_{YY} = SC(\{\rho_j\})$  y una nueva  $SC(\text{residuos}) = \sum (Y - \bar{Y})^2 - R_{YY}$ . Como hemos visto,  $R_{XX}$ ,  $R_{XY}$  y el correspondiente  $SC(\text{residuos})$  nuevo será necesario cuando se ajusta  $\beta$ . Claramente  $SC(\text{residuos})$  para  $Y$  es  $T_{YY} + E_{YY} = S_{YY}$  y lo será independientemente del diseño experimental usado. Los otros residuos son similares. Ahora, ajúste el  $\beta$  y obténgase  $SC(\beta | \{\rho_j\}) = S_{XY}^2/S_{XX}$ . A su turno, calcúlese

$$\sum (Y - \hat{Y})^2 = SC(\text{total ajustado}) - SC(\{\rho_j\}, \beta) = S_{YY} - \frac{S_{XY}^2}{S_{XX}},$$

donde

$$SC(\{\rho_j\}, \beta) = R_{XX} + \frac{S_{XY}^2}{S_{XX}},$$

Finalmente, la diferencia entre las sumas residuales de cuadrados para los dos modelos es la cantidad atribuible a la inclusión de los efectos de los tratamientos como el último conjunto de componentes en el modelo, o sea, luego de los efectos de los bloques y el término de la regresión. Es  $SC(\{\tau_i\} | \{\rho_j\}, \beta)$ . Hemos hecho lo equivalente a encontrar una suma de cuadrados atribuible al ajuste de coeficientes de regresión parcial para los efectos de los tratamientos.

Para probar el cuadrado medio de tratamientos ajustado, el cuadrado medio del error apropiado es  $s_{Y \cdot X}^2$ . Cuando se desea hacer varias pruebas como cuando se partitiona una suma de cuadrados de tratamientos en componentes, las pruebas exactas exigen un cálculo separado de tratamientos más error y tratamientos ajustados para cada comparación. En la sec. 17.10 se ilustra una aproximación para abreviar los cálculos.

**Ejercicio 17.4.1** Matrone et al (17.8) compararon el efecto de tres factores, cada uno a dos niveles sobre ganancias de peso en corderos. Los factores fueron heno de soya  $A$ , a partir de parcelas fertilizadas bien sea con KCa o PkCa, suplemento de urea  $B$ , y suplemento dicálcico  $C$ . Se usó un diseño en bloques completos al azar con dos repeticiones, determinándose el consumo de alimento para cada cordero. En la tabla anexa se dan los datos respectivos.

Ignorar la naturaleza factorial del experimento y efectuar los cálculos tal como se indica en la tabla 17.1. Calcular  $b$  y  $s_{Y \cdot X}^2$ . (En la siguiente sección se dan los procedimientos de cálculo). En término del análisis estadístico de los datos, ¿qué tan valiosas son las cifras decimales?

Ración	Bloque 1 Corderos hembras		Bloque 2 Corderos machos	
	X = consumo	Y = ganancia	X = consumo	Y = ganancia
000	209.3	11.2	286.9	27.2
100	252.4	26.1	302.1	30.6
010	241.5	13.2	246.8	15.4
001	259.1	24.4	273.2	24.0
110	201.1	18.8	274.6	20.1
101	287.5	31.0	276.3	24.4
011	286.6	27.9	270.7	29.9
111	255.7	20.8	253.0	20.8

**Ejercicio 17.4.2** Los datos del ejercicio 7.3.3 se han analizado como dos experimentos completamente al azar y, en el ejercicio 16.3.4, como un experimento de parcelas divididas. En el último caso, el modelo exige un incremento fijo atribuible a cada tratamiento, que se ha de probar mediante pre-post, y una heterogeneidad sobre las componentes de tratamiento, que se ha de probar mediante la interacción de (pre-post) X tratamientos.

Sin embargo, los aumentos pueden relacionarse linealmente con los registros antes de la prueba, de tal forma que la ganancia es mayor (o menor) según dónde se comienza la escala. Esto implica que las medidas antes pueden ser una covariante satisfactoria.

Efectuar los cálculos tal como sugiere la tabla 17.1 y la exposición que la acompaña, ¿qué modificaciones deben hacerse en la tabla 17.1? Calcular  $b$  y  $s_{Y \cdot X}^2$ .

**Ejercicio 17.4.3** Los datos del ejercicio 7.3.3 son incompletos y deberán incluir las observaciones del ejercicio 7.4.2. Incluir esas observaciones y considerar cómo analizar el conjunto completo de datos. ¿Es más complicado que lo que se usó en el ejercicio 17.4.2? Hacer los cálculos del caso. Calcular  $b$  y  $s_{Y \cdot X}^2$ .

**Ejercicio 17.4.4** Considérense los datos del ejercicio 15.3.3. Sean los registros de antes de la prueba los valores de una covariante y sean los datos después de la prueba los valores de una variable dependiente. Dejar de lado la naturaleza factorial de los tratamientos y hacer los cálculos necesarios para probar medias de tratamiento ajustadas.

## 17.5 La covarianza en el diseño de bloques completos al azar

La forma del análisis de la covarianza en un diseño de bloques completos al azar se presenta en la tabla 17.1. El proceso se ilustra usando los datos de la tabla 17.2 procedentes de un experimento realizado por J. L. Bowers y L. B. Permenter, Escuela Superior del Estado de Misisipi, en el cual se compararon 11 variedades de habas en cuanto al contenido de ácido ascórbico. Por experiencias previas, se sabe que el aumento de madurez producía una disminución en el contenido de vitamina C. Como no todas las variedades tenían la misma madurez en el momento de la cosecha y como todas las parcelas de una variedad dada no alcanzaban el mismo nivel de madurez en el mismo día, no era posible cosechar todas las parcelas en el mismo estado de madurez. Así que se observó el porcentaje de materia seca basado en 100 g de habas acabadas de cosechar como un índice de la madurez y se usó como covariante.

El cálculo de las sumas de productos para el diseño de bloques completos al azar se da básicamente en las ecs. (9.1) a (9.5) y (10.2). Para los datos de la tabla 17.2 se efectúan los siguientes cálculos. Las sumas de productos para el total son

$$\sum (X_{ij} - \bar{X}_{..})^2 = \sum X_{ij}^2 - \frac{\bar{X}_{..}^2}{rt} = 2,916.22$$

$$\sum (Y_{ij} - \bar{Y}_{..})^2 = 61,934.42$$

$$\sum (X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}) = \sum X_{ij} Y_{ij} - \frac{\bar{X}_{..} \bar{Y}_{..}}{rt} = -12,226.14$$

Las sumas de los productos para bloques son

$$R_{XX} = \frac{\sum_j X_{.j}^2}{t} - \frac{\bar{X}_{..}^2}{rt} = 367.85$$

$$R_{YY} = 4,968.94$$

$$R_{XY} = \frac{\sum_j X_{.j} Y_{.j}}{t} - \frac{\bar{X}_{..} \bar{Y}_{..}}{rt} = -1,246.66$$

Tabla 17.2 Y Contenido de ácido ascórbico † y X porcentaje de materia seca ‡ en habas

	Replicación												5 Totales de variedades	
	1		2		3		4		5		6			
	X	Y	X	Y	X	Y	X	Y	X	Y	X <sub>r</sub>	Y <sub>r</sub>		
1	34.0	93.0	33.4	94.8	34.7	91.7	38.9	80.8	36.1	80.2	177.1	440.5		
2	39.6	47.3	39.8	51.5	51.2	33.3	52.0	27.2	56.2	20.6	238.8	179.9		
3	31.7	81.4	30.1	109.0	33.8	71.6	39.6	57.5	47.8	30.1	183.0	349.6		
4	37.7	66.9	38.2	74.1	40.3	64.7	39.4	69.3	41.3	63.2	196.9	338.2		
5	24.9	119.5	24.0	128.5	24.9	125.6	23.5	129.0	25.1	126.2	122.4	628.8		
6	30.3	106.6	29.1	111.4	31.7	99.0	28.3	126.1	34.2	93.6	153.6	538.7		
7	32.7	106.1	33.8	107.2	34.8	97.5	35.4	86.0	37.8	88.8	174.5	485.6		
8	34.5	61.5	31.5	83.4	31.1	93.9	36.1	69.0	38.5	46.9	121.7	354.7		
9	31.4	80.5	30.5	106.5	34.6	76.7	30.9	91.8	36.8	68.2	164.2	423.7		
10	21.2	149.2	25.3	151.6	23.5	170.1	24.8	155.2	24.6	146.1	119.4	772.2		
11	30.8	78.7	26.4	116.9	33.2	71.8	33.5	70.3	43.8	40.9	167.7	378.6		
Total bloques	348.8	990.7	342.1	1,134.9	373.8	995.9	382.4	962.2	422.2	806.8	1,869.3	4,890.5		
	X <sub>r</sub>	Y <sub>r</sub>												

† Miligramos por 100 g de peso seco

‡ Basado en 100 g de haba recién cosechados.

La suma de productos para tratamientos (variedades) son

$$T_{xx} = \frac{\sum_i X_{ii}^2}{r} - \frac{X_{..}^2}{rt} = 2,166.71$$

$$T_{yy} = 51,018.18$$

$$T_{xy} = \frac{\sum_i X_{ij} Y_{ij}}{r} - \frac{X_{..} Y_{..}}{rt} = -9,784.14$$

Las sumas de los productos del error encontrados por diferencia son

$$\begin{aligned} E_{xx} &= \sum (X_{ij} - \bar{X}_{..})^2 - R_{xx} - T_{xx} \\ &= 2,916.22 - 367.85 - 2,166.71 = 381.66 \\ E_{yy} &= \sum (Y_{ij} - \bar{Y}_{..})^2 - R_{yy} - T_{yy} = 5,947.30 \\ E_{xy} &= \sum (X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}) - R_{xy} - T_{xy} \\ &= -12,226.14 - (-1,246.66) - (-9,784.14) \\ &= -1,195.34 \end{aligned}$$

Estos resultados se llevan a la tabla 17.3.

Estas sumas de productos contienen el material para el análisis de la varianza de  $X$  y  $Y$ , lo mismo que para el análisis de la covarianza. Así, para probar la hipótesis de que no hay diferencias entre las medias de variedades no ajustadas para el contenido de ácido ascórbico  $Y$ ,

$$F = \frac{T_{yy}/(t-1)}{E_{yy}/(r-1)(t-1)} = \frac{51,018.18/10}{5,947.30/40} = 34.31^{**} \quad 10 \text{ y } 40 \text{ gl}$$

Concluimos que hay diferencias reales entre medias de variedades para el porcentaje de materia seca en las habas en el momento de la cosecha.

Para probar la hipótesis de que no existen diferencias entre las medias de las variedades por porcentaje de materia seca  $X$  en el momento de la cosecha

$$F = \frac{T_{xx}/(t-1)}{E_{xx}/(r-1)(t-1)} = \frac{2,166.71/10}{381.66/40} = 22.71^{**} \quad 10 \text{ y } 40 \text{ gl}$$

Concluimos que existen diferencias verdaderas entre las medias de las variedades por porcentaje de materia seca en las habas en el momento de la cosecha.

Tabla 17.3 Análisis de la covarianza de los datos de la tabla 17.2

Fuente de variación	gl	Sumas de productos			Y ajustados por X			F
		X, X	X, Y	Y, Y	gl	SC	CM	
Total	54	2,916.22	-12,226.14	61,934.42				
Bloques	4	367.85	-1,246.66	4,968.94				
Tratamientos	10	2,166.71	-9,784.14	51,018.18				
Error	40	381.66	-1,195.34	5,947.30	39	2,203.45	56.50	
Tratamientos + error	50	2,548.37	-10,979.48	56,965.48	49	9,661.13		
Tratamientos ajustados					10	7,457.62	745.76	13.20**

El último *F* significante ilustra la situación mencionada en la sec. 17.2, esto es, que pueden existir diferencias significantes entre medias de tratamientos de la variable dependiente. Las diferencias en madurez que se miden por el porcentaje de materia seca no son necesariamente efectos de tratamiento. Esto ocurre en parte porque no todas las variedades fueron cosechadas en el mismo estado de madurez. En todo caso, se van a ajustar las medias de ácido ascórbico de las variedades al mismo porcentaje de materia seca o madurez ya que ésta es la etapa lógica para comparar variedades.

Para probar la hipótesis de que no existen diferencias entre medias de tratamientos ajustadas, es necesario calcular la suma de cuadrados del error y la suma de cuadrados de tratamiento más error de *Y* ajustadas por sus respectivas regresiones con respecto a la co-variable *X*. La suma de cuadrados para probar la hipótesis de que no hay diferencias entre medias de tratamientos ajustadas es la diferencia entre estas sumas de cuadrados ajustadas. El procedimiento se esquematiza en la tabla 17.1 y se ilustra en la tabla 17.3.

Para el error, el coeficiente de regresión está dado por la ec. (17.5) así:

$$b_{YX} = \frac{E_{XY}}{E_{XX}} = \frac{-1,195.34}{381.66} \\ = -3.13 \text{ mg de ácido ascórbico por 1% de materia seca}$$

La suma de cuadrados de *Y* atribuible a la regresión de *Y* con respecto a *X* es

$$b_{YX} E_{XY} = \frac{(E_{XY})^2}{E_{XX}} = \frac{(-1,195.34)^2}{381.66} = 3,743.85 \quad \text{con 1 gl}$$

La suma de cuadrados ajustadas está implicada en la ec. (17.6) como

$$E_{YY} - \frac{(E_{XY})^2}{E_{XX}} = 5,947.30 - 3,743.74 \\ = 2,203.45 \quad \text{con } (r-1)(t-1)-1 = 39 \text{ gl}$$

y la varianza residual es

$$s_{Y \cdot X}^2 = \frac{2,203.45}{39} = 56.50$$

Para tratamientos más error, la suma de cuadrados ajustada es

$$\begin{aligned} S_{YY} - \frac{(S_{XY})^2}{S_{XX}} &= 56,965.48 - \frac{(10,979.48)^2}{2,548.37} \\ &= 9,661.13 \quad \text{con } r(t-1) - 1 = 49 \text{ gl} \end{aligned}$$

Para medias de tratamientos ajustadas, la suma de cuadrados es la diferencia entre la suma de cuadrados de tratamientos más error y la suma del error, es decir,

$$(ajustados) T_{YY} = 9,661.13 - 2,203.45 = 7,457.62 \quad \text{con } t-1 = 10 \text{ gl}$$

La suma de cuadrados ajustados para el error se usa para estimar la varianza dentro de todas y cada una de las poblaciones de valores  $Y$ . Se supone que la varianza es homogénea. La suma de cuadrados ajustadas para tratamientos más error es una estimación de la suma de cuadrados esperada para las dos fuentes combinadas si todas las  $\bar{X}$  de tratamientos fuesen iguales.

Para que los ajustes dados sean válidos y útiles, es necesario que la regresión de  $Y$  con respecto a  $X$  sea homogénea para todos los tratamientos ajustados para las diferencias de bloques. La homogeneidad de la regresión es, entonces, un supuesto, como lo es el de la homogeneidad de la varianza del error luego del ajuste por regresión. No existe una forma simple de probar los supuestos de homogeneidad.

Para probar las hipótesis de que no hay diferencias entre medias de tratamientos para  $Y$  ajustada por la regresión de  $Y$  con respecto a  $X$ ,

$$\begin{aligned} F &= \frac{\text{CM}(\text{medias ajustadas de tratamientos})}{s_{Y \cdot X}^2} \\ &= \frac{745.76}{56.50} = 13.20^{**} \quad \text{con 10 y 39 gl} \end{aligned}$$

La  $F$  altamente significante es prueba de que existen diferencias reales entre las medias de los tratamientos para  $Y$  cuando se ajustan por  $X$ . Si las medias de tratamientos no ajustadas hubiesen sido significantes pero no las ajustadas, ello indicaría que las diferencias entre las medias no ajustadas sólo reflejaría diferencias en madurez y no en contenido de ácido ascórbico a una madurez común.

**Ejercicio 17.5.1** Probar medias de tratamiento ajustadas para el ejercicio 17.4.1. Probar la hipótesis nula de que  $\beta = 0$ .

Ejercicio 17.5.2 Probar medias ajustadas de tratamientos para el ejercicio 17.4.2. Probar  $H_0: \beta = 0$ .

Ejercicio 17.5.3 Probar medias ajustadas de tratamiento para el ejercicio 17.4.3. Probar  $H_0: \beta = 0$ .

Ejercicio 17.5.4 Probar medias ajustadas de tratamiento para el ejercicio 17.4.4. Probar  $H_0: \beta = 0$ .

## 17.6 Ajuste de las medias de tratamiento

La fórmula para ajustar medias de tratamiento se da esencialmente en la ec. (17.4), y es una aplicación del procedimiento dado en la ec. (10.9). La idea básica también se ilustra en la fig. 17.1. En notación familiar, la ecuación para una media de tratamiento ajustada es

$$\hat{Y}_{i\cdot} = \bar{Y}_{i\cdot} - b_{YX}(\bar{X}_{i\cdot} - \bar{X}_{..}) \quad (17.7)$$

$b_{YX} = b$  es el coeficiente de regresión del error. Las medias de tratamiento ajustadas son estimaciones de lo que serían las medias de tratamiento si todas las  $\bar{X}_{i\cdot}$  coincidieran con  $\bar{X}_{..}$ . Las medias de tratamiento ajustadas se dan en la tabla 17.4 para los datos de la tabla 17.2.

El error estándar de una media de tratamiento ajustada es simplemente una modificación de la ec. (10.13). Esto es,

$$s_{\hat{Y}_{i\cdot}} = s_{Y \cdot X} \sqrt{\frac{1}{r} + \frac{(\bar{X}_{i\cdot} - \bar{X}_{..})^2}{E_{XX}}}$$

Tabla 17.4 Ajuste de medias de tratamientos; datos de la tabla 17.2

Variedad No.	Media porcentual materia seca $\bar{X}_{i\cdot}$	Desviación $\bar{X}_{i\cdot} - \bar{X}_{..}$	Ajuste $b_{YX}(\bar{X}_{i\cdot} - \bar{X}_{..})$	Media observada del contenido de ácido ascórbico $\bar{Y}_{i\cdot}$	Media ajustada del contenido de ácido ascórbico $\hat{Y}_{i\cdot} = \bar{Y}_{i\cdot} - b_{YX} \times (\bar{X}_{i\cdot} - \bar{X}_{..})$
1	35.42	1.43	-4.48	88.10 (5)†	92.59 (5)
2	47.76	13.77	-43.10	35.98(11)	79.12 (8)
3	36.60	2.61	-8.17	69.92 (9)	78.10 (9)
4	39.38	5.39	-16.87	67.64(10)	84.53 (6)
5	24.48	-9.51	29.77	125.76 (2)	95.98 (4)
6	30.72	-3.27	10.24	107.74 (3)	97.51 (3)
7	34.90	0.91	-2.85	97.12 (4)	99.98 (2)
8	34.34	0.35	-1.10	70.94 (8)	72.04(11)
9	32.84	-1.15	3.60	84.74 (6)	81.15 (7)
10	23.88	-10.11	31.64	154.44 (1)	122.78 (1)
11	33.54	-0.45	1.41	75.72 (7)	74.32(10)
$\bar{X} = 33.99$		$\Sigma = -0.03‡$	$\Sigma = +0.09‡$	$\bar{Y}_{..} = 88.92‡$	Media = 88.92‡

† Rangos.

‡ Teóricamente  $\sum (\bar{X}_{i\cdot} - \bar{X}_{..}) = 0$  Luego  $\sum b_{YX}(\bar{X}_{i\cdot} - \bar{X}_{..}) = 0$  y  $\sum \bar{Y}_{i\cdot} = \sum \hat{Y}_{i\cdot}$ . En la práctica, pueden aparecer errores por aproximación de cifras.

El error estándar de la diferencia entre dos medias de tratamiento ajustadas lo da Wishart (17.12) así:

$$s_{\bar{Y}_t - \bar{Y}_{t'}} = \sqrt{s_{Y \cdot X}^2 \left[ \frac{2}{r} + \frac{(\bar{X}_t - \bar{X}_{t'})^2}{E_{XX}} \right]} \quad (17.8)$$

Por ejemplo, para comparar las variedades 7 y 10 (suponiendo una razón legítima), se requiere

$$s_{\bar{Y}_7 - \bar{Y}_{10}} = \sqrt{56.50 \left[ \frac{2}{5} + \frac{(34.90 - 23.88)^2}{381.66} \right]} = 6.38 \text{ mg de ácido ascórbico}$$

Es la comparación de las medias ajustadas  $\bar{Y}_7 = 90.97$  y  $\bar{Y}_{10} = 122.78$  y evidentemente la diferencia resulta significante.

La ecuación (17.8) exige un cálculo separado para cada comparación. Esto probablemente se justifica si el experimento demandó una cantidad apreciable de tiempo o fue algo más que un experimento preliminar. Sin embargo, Finney (17.5) sugiere la siguiente aproximación  $s_{\bar{Y}_t - \bar{Y}_{t'}}$ , la cual utiliza un promedio en vez de las  $(\bar{X}_t - \bar{X}_{t'})$  por separado requeridas por la ec. (17.8). La fórmula es

$$s_{\bar{Y}_t - \bar{Y}_{t'}} = \sqrt{\frac{2s_{Y \cdot X}^2}{r} \left[ 1 + \frac{T_{XX}}{(t-1)E_{XX}} \right]} \quad (17.9)$$

Por la tabla 17.3

$$s_{\bar{Y}_7 - \bar{Y}_{10}} = \sqrt{\frac{2(56.50)}{5} \left[ 1 + \frac{2,166.71}{10(381.66)} \right]} = 5.95 \text{ mg de ácido ascórbico}$$

para nuestro ejemplo. Para medias de tratamiento con desigual número remplácese  $2/r$  en las ecs. (17.8) y (17.9) por  $(1/r_1 + 1/r_2)$ . Cochran y Cox (17.2) aseguran que la aproximación de Finney es, por lo general, bastante cercana si los grados de libertad para el error son más de 20, ya que es pequeña la contribución de los errores de muestreo en  $b_{YX}$ , el factor de ajuste.

Si la variación entre las  $\bar{X}_t$  es significante, por ejemplo, debido a los tratamientos, la fórmula aproximada puede conducir a serios errores y no debe usarse. Esto deberá aplicarse aquí también, ya que las diferencias entre medias de tratamientos para la materia seca  $X$  fueron significantes.

**Ejercicio 17.6.1** Calcular medias ajustadas de tratamiento para los 8 tratamientos del ejercicio 17.4.1. Calcular la desviación estándar de la media para la ración 111. Calcular la desviación estándar para la diferencia entre medias para los tratamientos 000 y 111.

Calcular las medias ajustadas de tratamientos para las dos medias utilizadas para medir el efecto principal  $A$ . Calcular la desviación estándar de la diferencia entre estas medias. Probar la hipótesis nula de que no existe efecto principal  $A$ .

Usar la ecuación (17.9) para calcular una desviación estándar aproximada entre medias  $A$ . Compararla con la desviación estándar exacta. ¿Se presentaron diferencias significantes entre las medias de tratamiento para la covariante?

**Ejercicio 17.6.2** Calcular las tres medias ajustadas de tratamiento para el ejercicio 17.4.2. Encontrar las desviaciones estándar para las tres diferencias.

Utilizar la propuesta de Finney, ec. (17.9), para encontrar una sola desviación estándar que sea aproximada a las que se acaban de calcular. ¿Fueron significantes las diferencias entre las medias de tratamiento para la covariante?

**Ejercicio 17.6.3** Considérese el ejercicio 17.4.3. ¿Qué fórmula se usaría para comparar dos medias ajustadas de tratamiento basadas en diferente número de observaciones?

**Ejercicio 17.6.4** Calcular las medias ajustadas de tratamientos para los cuatro tratamientos del ejercicio 17.4.4.

Calcular las dos medias ajustadas para braille e imprenta. Para los tratamientos 1 y 2. Calcular la desviación estándar para la diferencia entre medias para braille e imprenta. Entre medias de los tratamientos 1 y 2. Probar la hipótesis nula de que no hay diferencias entre medias de población ajustadas para braille e imprenta. Para los tratamientos 1 y 2.

## 17.7 Aumento de precisión debido a la covarianza

Para probar la efectividad de la covarianza como medio de controlar el error, se hace una comparación de la varianza de una media de tratamiento antes y después del ajuste para la variable independiente  $X$ . El cuadrado medio del error antes del ajuste es  $5.947.30/40 = 148.68$  con 40 grados de libertad y después del ajuste es 56.50 con 39 grados de libertad. Es necesario ajustar el último valor para dejar margen al error de muestreo en el coeficiente de regresión usado en el ajuste. El cuadrado medio del error efectivo luego del ajuste por  $X$  es

$$s_{Y \cdot X}^2 \left[ 1 + \frac{T_{XX}}{(t-1)E_{XX}} \right] = 56.50 \left[ 1 + \frac{2.166.71}{10(381.66)} \right] = 88.58 \quad (17.10)$$

Una estimación de la precisión relativa es  $(148.68/88.58)/100 = 168$  por ciento. Esto indica que 100 replicaciones con covarianza son tan efectivas como 168 sin ello, una razón de aproximadamente 3:5.

**Ejercicio 17.7.1** Calcular el cuadrado medio efectivo para el ejercicio 17.4.1. A partir de la precisión relativa, concluir cuántos bloques se necesitarían para obtener la misma precisión sin usar la covariante.

**Ejercicio 17.7.2** Calcular el cuadrado medio efectivo para el ejercicio 17.4.2. A partir de la precisión relativa, concluir cuántas replicaciones se necesitarían para lograr la misma precisión sin el uso de la covariante.

**Ejercicio 17.7.3** Calcular el cuadrado medio efectivo para el ejercicio 17.4.4. A partir de la precisión relativa, concluir cuántas replicaciones se necesitarían para obtener la misma precisión sin el uso de la covariante.

**Tabla 17.5 Partición de una covarianza; datos de la tabla 17.2**

Fuente de variación	gl	Sumas de productos	$b_{yx}$	$r$
Total	54	-12,226.14	-4.19	-0.910
Bloques	4	-1,246.66	-3.39	-0.922
Tratamientos	10	-9,784.14	-4.52	-0.931
Error	40	-1,195.34	-3.13	-0.793

### 17.8 Partición de la covarianza

El término análisis de la covarianza implica un uso al que generalmente no se le da relieve, esto es, la partición efectuada en la columna de productos cruzados. Esta se amplía en la tabla 17.5 para incluir los coeficientes de regresión y correlación.

Cuando la hipótesis nula de que no hay diferencias de tratamientos entre las medias de tratamientos ajustados es verdadera, los tratamientos y el error proporcionan estimaciones independientes de coeficientes de regresión y correlación comunes. Si una prueba de medias de tratamientos ajustadas presenta significancia, entonces puede presumirse que las regresiones de tratamiento y error difieren. Si la prueba no presenta significancia, entonces puede haber o no una regresión de tratamientos. Si hay regresión de tratamiento, entonces es la misma que la del error. En casos raros, puede interesar probar la homogeneidad de los coeficientes de regresión. La suma de cuadrados apropiada para el numerador está dada por

SC de regresión de tratamiento frente a la de error

$$\begin{aligned}
 &= \frac{T_{xy}^2}{T_{xx}} + \frac{E_{xy}^2}{E_{xx}} - \frac{(T_{xy} + E_{xy})^2}{T_{xx} + E_{xx}} \\
 &= \frac{(-9,784.14)^2}{2,166.71} \\
 &\quad + \frac{(-1,195.34)^2}{381.66} - \frac{(-10,979.48)^2}{2,548.37} \\
 &= 621.31
 \end{aligned} \tag{17.11}$$

Obsérvese que cada término es una suma de cuadrados; la primera es para una regresión en la que se usan medias de tratamiento; la segunda, para una regresión en la que se usan residuos, y la tercera, la de una regresión combinada. En conjunto constituye una varianza entre dos coeficientes de regresión con las ponderaciones apropiadas, como se estudió en la sec. 10.13.

La varianza apropiada para probar la hipótesis nula de homogeneidad generalmente se considera como la varianza del error. [Se ha sugerido el uso de desviaciones a partir de la regresión de tratamientos para estimar la varianza del coeficiente de regresión de trata-

miento y la varianza del error para el coeficiente de regresión del error. Estas varianzas a menudo difieren y requieren de un procedimiento de prueba como el de la sec. 5.9; los divisores para las ecs. (5.15) serán  $T_{xx}$  y  $E_{xx}$  en vez de  $n_1$  y  $n_2$ .] Para nuestro ejemplo

$$F = \frac{621.31}{56.50} = 11.00^{**} \quad \text{con } 1 \text{ y } 39 \text{ gl.}$$

Esta prueba es equivalente a usar la prueba de  $t$  de la ec. (10.19).

La prueba de homogeneidad de las regresiones de tratamientos y error a veces se presenta en forma alterna. Puede usarse el siguiente argumento para explicar y justificar el procedimiento.

1. Calcular la suma de cuadrados para las medias de tratamiento ajustadas. Este puede considerarse como la suma de cuadrados de desviaciones de medias de tratamiento respecto de una recta de regresión común, un promedio móvil. Como la pendiente de la recta de regresión se obtiene a partir del error, no se partitiona un grado de libertad de los disponibles para medias de tratamientos.
2. Calcular la suma de cuadrados de desviaciones de medias de tratamientos respecto de su propia regresión. Esta tendrá  $t - 2$  grados de libertad ya que las medias proporcionan la estimación del coeficiente de regresión.
3. Restar la última suma de cuadrados de la primera. Si las regresiones son las mismas, el resultado no deberá ser mayor que el esperado del muestreo aleatorio. Si el resultado no puede atribuirse razonablemente al azar y a la hipótesis nula, entonces concluimos que las regresiones no son homogéneas.

Para nuestro ejemplo

$$\begin{aligned} \text{Las desviaciones respecto de la regresión de tratamiento} &= T_{yy} - \frac{T_{xy}^2}{T_{xx}} \\ &= 51,018.08 - \frac{(-9,784.14)^2}{2,166.71} = 6,836.27 \quad \text{con } t - 2 = 9 \text{ gl} \end{aligned}$$

SC de regresión de tratamiento frente a error

$$\begin{aligned} &= T_{yy} - \text{ajustados} - \text{desviaciones respecto de la regresión de tratamiento} \\ &= 7,457.57 \text{ (10 gl)} - 6,836.27 \text{ (9 gl)} = 621.30 \quad \text{con 1 gl} \end{aligned}$$

De nuevo,  $F = 621.30 / 56.50 = 11.00^{**}$  con 1 y 39 gl. Puesto que las diferencias entre medias de tratamiento ajustadas fueron significantes, este resultado era de esperarse y normalmente no se hubiera hecho la prueba.

Una comparación de los coeficientes de correlación puede hacerse mediante el uso del procedimiento apropiado de la sec. 11.5. Este procedimiento indica que no existe diferencia entre los coeficientes de correlación. Que esto puede ocurrir, es decir, que los coeficientes de regresión pueden ser muy diferentes aunque los de correlación sean iguales, puede verse por la siguiente relación

$$r = b_{yx} \frac{s_x}{s_y}$$

**Tabla 17.6** Homogeneidad de regresiones dentro de tratamientos para un diseño completamente al azar

Tratamientos	gl	$\sum (X - \bar{X})^2$	$\sum (X - \bar{X})(Y - \bar{Y})$	$\sum (Y - \bar{Y})^2$	gl	SC residual
1	$n_1 - 1$	$E_{xx}(1)$	$E_{yy}(1)$	$E_{rr}(1)$	$n_1 - 2$	SC <sub>1</sub> (residuales)
2	$n_2 - 1$	$E_{xx}(2)$	$E_{yy}(2)$	$E_{rr}(2)$	$n_2 - 2$	SC <sub>2</sub> (residuales)
⋮	⋮	⋮	⋮	⋮	⋮	⋮
t	$n_t - 1$	$E_{xx}(t)$	$E_{yy}(t)$	$E_{rr}(t)$	$n_t - 2$	SC <sub>t</sub> (residuales)
Residuos de regresiones individuales					$\sum n_i - 2t$	$\sum SC_i$ (residuales) = error = A
Total de regresiones simples	$\sum n_i - t$	$\sum E_{xx}(i)$	$\sum E_{yy}(i)$	$\sum E_{rr}(i)$	$\sum n_i - t - 1$	$\sum E_{rr}(i) - \frac{(\sum E_{yy}(i))^2}{\sum E_{xx}(i)} = B$
Diferencias para homogeneidad de regresiones					$t - 1$	$B - A$

Para probar la homogeneidad de la regresión,  $F = [(B - A)/(t - 1)]/[A/(\sum n_i - 2t)]$ , con  $t - 1$  y  $\sum n_i - 2t$  gl.

Dos  $r$  pueden ser muy similares a pesar de  $b_{YX}$  muy diferentes con tal que la razón  $s_X/s_Y$  compense esas diferencias.

### 17.9 Homogeneidad de coeficientes de regresión

Cuando el diseño experimental es completamente al azar, puede calcularse la regresión de  $Y$  con respecto a  $X$  para cada tratamiento. En este caso, el supuesto usual de homogeneidad de los coeficientes de regresión puede plantearse como una hipótesis nula y probarse mediante una prueba de  $F$  apropiada en un análisis de la covarianza. (Para otros diseños, no estamos al tanto de otros métodos actualmente disponibles para probar la homogeneidad de los coeficientes de regresión).

El procedimiento se presenta a continuación y se resume en la tabla 17.6

#### 1. Calcular

$$\sum_j (X_{ij} - \bar{X}_{i\cdot})^2 \quad \sum_j (X_{ij} - \bar{X}_{i\cdot})(Y_{ij} - \bar{Y}_{i\cdot}) \quad \text{y} \quad \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$$

para cada tratamiento.

Para el  $i$ -ésimo tratamiento, designarlos  $E_{XX}(i)$ ,  $E_{XY}(i)$ , y  $E_{YY}(i)$ , respectivamente.

#### 2. A partir de lo anterior y para cada tratamiento, calcular

$$SC_i(\text{regresión}) = \frac{E_{XY}^2(i)}{E_{XX}(i)}$$

(no se presentan en la tabla 17.6) y la suma residual de cuadrados.

$$SC_i(\text{residuos}) = E_{YY}(i) - \frac{E_{XY}^2(i)}{E_{XX}(i)}, \quad i = 1, \dots, t.$$

Ver la tabla 17.6. Ahora, quedan ajustadas  $t$  medias y  $t$  coeficientes de regresión.

3. Totalizar las sumas de cuadrados de los residuos para dar la suma residual de cuadrados de las desviaciones respecto de las rectas de regresión individuales. Este es el residuo luego del ajuste en 2 y se denota  $A$  en la tabla 17.6. Tiene  $\sum (n_i - 2) = \sum n_i - 2t$  grados de libertad.
4. Totalizar las  $E_{XX}(i)$ ,  $E_{XY}(i)$  y  $E_{YY}(i)$  para todos los tratamientos.
5. A partir de los resultados en el paso 4, calcular la suma de cuadrados de  $Y$  atribuible a la regresión con respecto a  $X$  y a la suma residual de cuadrados. Esta suma residual de cuadrados es la de las desviaciones respecto de las rectas de regresión individuales, cada una de las cuales pasa por los propios  $(\bar{X}, \bar{Y})$  pero todas con un coeficiente de regresión común. Aquí se han ajustado  $t$  medias y una regresión. El residuo se denota  $B$  en la tabla 17.6 y tiene  $\sum (n_i - 1) - 1 = \sum n_i - t - 1$  grados de libertad.
6. Calcular  $B - A$ . Esta es la cantidad de la suma de cuadrados total para  $Y$  que puede atribuirse a las diferencias entre los coeficientes de regresión. Es necesariamente positiva, ya que siempre podremos mejorarla mediante el ajuste de varios coeficientes en vez de una.

La cantidad  $B - A$  también puede obtenerse mediante el cálculo de  $\sum SC_i$  (regresión con  $t$  gl y restando, de este total,  $[\sum Exr(i)]^2 / \sum Exx(i)$  con un gl). La diferencia  $B - A$  tiene  $t - 1$  grados de libertad. El término de error apropiado es el cuadrado medio de las desviaciones con respecto a las regresiones individuales.

**Ejercicio 17.9.1** Probar la hipótesis nula de homogeneidad de los coeficientes de regresión para los datos del ejercicio 10.8.1. Comparar el  $F$  resultante con el valor  $t^2$ , donde  $t$  es el que se obtuvo en el ejercicio 10.8.1.

**Ejercicio 17.9.2** En el ejercicio 17.4.2 se analizaron los datos del ejercicio 7.3.3 como un problema de covarianza. ¿Qué nos dice lo comprobado respecto a la homogeneidad de los tres coeficientes de regresión que pueden calcularse?

**Ejercicio 17.9.3** Repítase el ejercicio 17.9.2 refiriéndose a los ejercicios 7.4.2 y 17.4.3.

**Ejercicio 17.9.4** En el ejercicio 17.4.4, se analizaron los datos del ejercicio 15.3.3 como un problema de covarianza. Probar la homogeneidad de los cuatro coeficientes de regresión que se supusieron homogéneos en esa oportunidad. ¿De cuántos grados de libertad se dispone para estimar cada coeficiente de regresión?

## 17.10 La covarianza cuando se partitiona la suma de cuadrados de tratamiento

La aplicación del análisis de la covarianza a un experimento factorial se ilustra con los datos en la tabla 17.7, tomada de Wishart (17.13). El experimento fue un factorial  $3 \times 2$

**Tabla 17.7** Pesos iniciales  $X$  y ganancias de peso  $Y$  en libras en cerdos para tocineta en un ensayo de nutrición

Jaulas (bloques)	Sexo	Raciones						Totales	
		$a_1$		$a_2$		$a_3$		$X$	$Y$
		$X$	$Y$	$X$	$Y$	$X$	$Y$		
1	$M = b_1$	38	9.52	39	8.51	48	9.11	269	56.83
	$F = b_2$	48	9.94	48	10.00	48	9.75		
2	$M$	35	8.21	38	9.95	37	8.50	202	54.04
	$F$	32	9.48	32	9.24	28	8.66		
3	$M$	41	9.32	46	8.43	42	8.90	238	52.94
	$F$	35	9.32	41	9.34	33	7.63		
4	$M$	48	10.56	40	8.86	42	9.51	272	59.88
	$F$	46	10.90	46	9.68	50	10.37		
5	$M$	43	10.42	40	9.20	40	8.76	222	55.44
	$F$	32	8.82	37	9.67	30	8.57		
Totales	$M$	205	48.03	203	44.95	209	44.78	617	137.76
	$F$	193	48.46	204	47.93	189	44.98	586	141.37
	$M + F$	398	96.49	407	92.88	398	89.76	1,203	279.13

Tabla 17.8 Análisis de la covarianza para los datos de la tabla 17.7

Fuente de variación	gl	Sumas de productos			Y ajustado por X			
		X, X	X, Y	Y, Y	gl	SC	CM	F
Total	29	1,108.70	78.507	16.3453				
Bloques (jaulas)	4	605.87	39.905	4.8518				
Raciones	2	5.40	-0.147	2.2686				
Sexo	1	32.03	-3.730	0.4344				
Ración × sexo	2	22.47	3.112	0.4761				
Error	20	442.93	39.367	8.3144	19	4.8156	0.2535	
Ración + error	22	448.33	39.220	10.5830	21	7.1520		
Diferencias para probar medias de ración ajustadas.					2	2.3366	1.1683	4.61*
Sexo + error	21	474.96	35.637	8.7488	20	6.0749		
Diferencias para probar medias de sexo ajustadas.					1	1.2594	1.2594	4.97*
Ración × sexo + error	22	465.40	42.479	8.7905	21	4.9133		
Diferencias para probar interacción sexo × ración ajustada					2	0.0977	0.0489	0.19

en un diseño de bloques completos al azar con cinco repeticiones. Los tratamientos fueron tres raciones, factor  $A$ , y dos sexos, factor  $B$ . El procedimiento es aplicable siempre que se desee particionar una suma de cuadrados de tratamientos.

El análisis se presenta en la tabla 17.8. El procedimiento de cálculo es el mismo que el de la sec. 17.5, excepto que la suma de cuadrados de tratamientos se partitiona en componentes de efectos principales y de interacciones, como se vio en el cap. 15. Para una referencia fácil, escribase  $T_{XY} = A_{XY} + B_{XY} + AB_{XY}$  para denotar la partición de la suma de productos cruzados de tratamientos en componentes para las raciones  $A$ , sexo  $B$  e interacción  $AB$ . Una notación similar se usa con  $T_{XX}$  y  $T_{YY}$ . Para partitionar la suma de productos cruzados de tratamientos, cálculese

$$A_{XY} = \frac{\sum_i (a_i)_X (a_i)_Y}{rb} - \frac{(\sum X_{ij})(\sum Y_{ij})}{rt}$$

$$= \frac{398(96.49) + 407(92.88) + 398(89.76)}{5(2)} - \frac{1,203(279.13)}{5(6)}$$

$$= -0.147$$

$$B_{XY} = \frac{\sum (b_j)_X (b_j)_Y}{ra} - \frac{(\sum X_{ij})(\sum Y_{ij})}{rt}$$

$$= \frac{617(137.76) + 586(141.37)}{5(3)} - \frac{1,203(279.13)}{5(6)} = -3.730$$

$$\begin{aligned} AB_{XY} &= T_{XY} - A_{XY} - B_{XY} \\ &= \frac{205(48.03) + \dots + 189(44.98)}{5} - \frac{1,203(279.13)}{5(6)} - (-0.147) \\ &\quad - (-3.730) = 3.112 \end{aligned}$$

donde  $(a_i)_X$  es el total de los valores de  $X$  para el nivel  $i$  del factor  $A$ , y así sucesivamente, con  $t = ab$ . Las sumas de productos calculadas directamente se llevan a la tabla 17.8 y los residuos se obtienen restando del total de la suma de productos las componentes de bloques, sexo, raciones y raciones  $\times$  sexo. Para probar las hipótesis nulas de que no hay diferencias entre los niveles del factor  $A$ , los niveles del factor  $B$  y de que no hay interacción  $AB$ , luego del ajuste por la variable concomitante, el procedimiento es básicamente el que se indica en las tablas 17.1 y 17.3. Sin embargo, cada hipótesis se prueba separadamente, tal como se ve en la tabla 17.8. Por lo tanto, la suma de cuadrados ajustada para probar la hipótesis nula acerca de las raciones es la diferencia entre la suma de cuadrados ajustada para las raciones más el error y la suma de cuadrados ajustada para el error. Por ejemplo, la suma de cuadrados ajustada para el error es

$$\begin{aligned} E_{YY} - \frac{(E_{XY})^2}{E_{XX}} &= 8.3144 - \frac{(39.367)^2}{442.93} \\ &= 4.8156 \quad \text{con } (r-1)(t-1)-1 = 19 \text{ gl} \end{aligned}$$

La suma de cuadrados para raciones más error es

$$\begin{aligned} (A_{YY} + E_{YY}) - \frac{(A_{XY} + E_{XY})^2}{A_{XX} + E_{XX}} \\ &= 2.2686 + 8.3144 - \frac{[(-0.147) + (39.367)]^2}{[5.40 + 442.93]} \\ &= 7.1520 \quad \text{con } (a-1) + (r-1)(t-1)-1 = 21 \text{ gl} \end{aligned}$$

La diferencia

$$\begin{aligned} \left[ A_{YY} + E_{YY} - \frac{(A_{XY} + E_{XY})^2}{A_{XX} + E_{XX}} \right] - \left[ E_{YY} - \frac{(E_{XY})^2}{E_{XX}} \right] &= 7.1520 - 4.8155 \\ &= 2.3366 \quad \text{con } a-1 = 2 \text{ gl} \end{aligned}$$

es la suma de cuadrados ajustada para raciones.

Las sumas de cuadrados ajustadas para sexo y raciones por sexo se obtienen de manera análoga.

**Tabla 17.9 Análisis de la varianza de  $[Y - b(X - \bar{X})]$  para los datos de las tablas 17.7 y 17.8**

Fuente de variación	gl	Suma de cuadrados	Cuadrado medio	F	F*
Raciones	2	2.3374	1.1687	4.61*	4.61*
Sexo	1	1.3507	1.3507	5.33*	4.97*
Ración X sexo	2	0.1004	0.0502	0.20	0.19
Error	19	4.8156	0.2535		

\*F de la tabla 17.8

Las pruebas  $F$  para las tres hipótesis nulas se efectúan con el cuadrado medio de error ajustado como denominador. Por ejemplo, para probar la hipótesis nula de que no hay diferencias entre medias de raciones después del ajuste a una  $\bar{X}$  común,  $F = 1.1683/0.2534$  con 2 y 19 grados de libertad.

En un experimento en el que se partitiona la suma de cuadrados de tratamientos para varias hipótesis nulas, el procedimiento anterior supone mucho tiempo, pero se justifica para la mayoría de los experimentos. Cochran y Cox (17.2) dan la siguiente aproximación que permite economizar tiempo. Preparar un análisis de las  $Y$  ajustadas, esto es, de los  $[Y - b(X - \bar{X})]$  donde  $b = E_{XY}/E_{XX}$ , a partir del análisis de covarianza preliminar. Esto se ilustra en la tabla 17.9, utilizando el análisis de la covarianza de la tabla 17.8. El análisis de  $Y - b(X - \bar{X})$  se efectúa calculando  $E_{YY} - 2bE_{XY} + b^2E_{XX}$  para el error y las correspondientes cantidades para cada fuente de variación para las cuales se desea probar una hipótesis nula; en todos los casos,  $b = E_{XY}/E_{XX} = 39.367/442.93 = 0.0889$  lb de ganancia por libra de peso inicial. Por ejemplo, para sexo,

$$0.4344 - 2(0.0889)(-3.730) + (0.0889)^2(32.03) = 1.3507$$

Este método aproximado da el cuadrado medio residual correcto dentro de los errores de aproximación; las sumas de cuadrados para efectos principales e interacciones, y los correspondientes valores  $F$ , son mayores que los obtenidos mediante el método exacto de la tabla 17.8. La sobreestimación de  $F$  rara vez es considerable cuando la variación en  $X$  es aleatoria y sólo se debe comprobar por el método exacto cuando apenas sea significante. El procedimiento aproximado no deberá usarse cuando  $X$  está influida por los tratamientos o bien es mayor que la debido al azar, ya que los valores  $F$  pueden ser considerablemente erróneos en tales casos.

Las medias ajustadas de tratamiento se obtienen como en la tabla 17.4.

**Ejercicio 17.10.1** Calcular pruebas exactas de medias ajustadas de tratamiento para cada uno de los efectos principales e interacciones para los datos del ejercicio 17.4.1. Decir cuál es la hipótesis nula, en cada caso. Comparar los resultados de la prueba para el efecto principal  $A$  con la prueba  $t$  de la misma hipótesis del ejercicio 17.6.1.

**Ejercicio 17.10.2** Calcular las pruebas exactas de efectos principales e interacciones usando el método aproximado de Cochran y Cox. Comparar los valores de  $F$  resultante con los que se encontraron para el ejercicio 17.10.1.

**Tabla 17.10 Contenido medio de ácido ascórbico de tres muestras de 2 g de raíces de mostaza, en miligramos por 100 g de peso seco**

Bloque (día)	1	2	3	4	5	Total
<b>Tratamiento:</b>						
A		887	897	850	975	3,609
B	857	1,189	918	968	909	4,841
C	917	1,072	975	930	954	4,848
Total	1,774	3,148	2,790	2,748	2,838	13,298

**Ejercicio 17.10.3** Calcular las pruebas exactas de medias ajustadas de tratamiento para los dos efectos principales y la interacción del ejercicio 17.4.4. Comparar los resultados con los correspondientes del ejercicio 17.6.4.

### 17.11 Estimación de observaciones faltantes mediante la covarianza

Para ilustrar este uso de la covarianza, considérense los datos de la tabla 17.10, tomados de Tucker et al. (17.11). Aunque se dispone de todas las observaciones, supóngase que falta la observación en la esquina superior izquierda.

El procedimiento, que da una estimación insesgada de las sumas de cuadrados de tratamiento y de error, es:

1. Hacer  $Y = 0$  para la parcela faltante.
2. Definir una covariante como  $X = 0$  para un  $Y$  observado,  $X = +1$  ( $o - 1$ ) para  $Y = 0$ .
3. Efectuar el análisis de la covarianza (ver tabla 17.11).
4. Calcular  $b = E_{XY}/E_{XX}$  y cambiar de signo para estimar el valor faltante.

El valor de la parcela,  $-b$ , es esencialmente, un ajuste de la llamada observación  $Y = 0$ , para dar una estimación de los  $Y$  que se hubieran obtenido si  $X$  hubiese sido 0 en vez de 1.

La ecuación (9.8) de la parcela faltante da un valor de 800 para un cuadrado medio residual de 5,186 (el mismo resultado obtenido por el procedimiento de covarianza) y una suma ajustada de cuadrados de tratamiento de 25,293, que es sesgada. La ec. (9.9) puede usarse para calcular el sesgo cuando se usa la ec. (9.8). El análisis del procedimiento de covarianza lleva directamente a una prueba no sesgada de medias de tratamiento ajustadas.

En la práctica, se ve que el método de covarianza para estimar parcelas faltantes es cómodo y sencillo. La técnica puede ampliarse para trazar varias parcelas faltantes, introduciendo una nueva variable independiente para cada parcela faltante y utilizando covarianza múltiple, como se ilustra en la siguiente sección. Bartlett (17.1) da un ejemplo sobre el uso de la covarianza para calcular parcelas faltantes así como del tipo usual de covariante.

**Ejercicio 17.11.1** Aplique la técnica de la covarianza al problema de la parcela faltante del ejercicio 9.6.1. Obtener el valor de la parcela faltante, el cuadrado medio del error y el cuadrado medio de tratamiento. Comparar estos resultados con los del ejercicio 9.6.1.

**Tabla 17.11 Análisis de la covarianza como alternativa de la ecuación de la parcela faltante para los datos de la tabla 17.10.**

Fuente	gl	Sumas de productos			gl	Ajustadas	
		X, X	X, Y	Y, Y		SC(Y)	CM
Total	14	$1 - \frac{1}{15} = \frac{14}{15} = \frac{gl}{n}$	-886.53	945.296			
Bloques	4	$\frac{1}{3} - \frac{1}{15} = \frac{4}{15} = \frac{gl}{n}$	-295.20	359.823			
Tratamiento	2	$\frac{1}{5} - \frac{1}{15} = \frac{2}{15} = \frac{gl}{n}$	-164.73	203.533			
Residuo	8	Por sustracción = $\frac{8}{15} = \frac{gl}{n}$	-426.60	381.940	7	40.713	5.816
Tratamientos + residuo	10	$\frac{10}{15} = \frac{gl}{n}$	-591.33	585.473	9	60.966	
Tratamientos ajustados					2	20.253	10.126
		Parcela faltante = $-b = \frac{426.60}{8/15} = 800$ (aprox)					

**Ejercicio 17.11.2** Si hubiera habido observaciones faltantes en los datos del ejercicio 17.4.2 y 17.4.4, ¿hubiera sido necesario alguna técnica de observación faltante? ¿Por qué?

### 17.12 Covarianza con dos variables independientes

A veces una variable dependiente se ve afectada por dos o más variables independientes. Un método para analizar tales datos es el análisis de la covarianza múltiple. En la tabla 17.12 se da un ejemplo de tales datos, con dos variables  $X_1$  y  $X_2$  que miden peso inicial y forraje consumido (trébol Ladino) por conejillos de indias y una variable dependiente,  $Y$ , que mide el aumento de peso. {Puede demostrarse [ver Steel (17.10)] que cuando el peso inicial es una covariante, el peso final como alternativa del aumento de peso lleva al mismo análisis de covarianza, en el sentido de que las sumas de cuadrados ajustadas para los dos análisis son idénticas}.

El peso inicial no se ve afectado por los tratamientos, pero se incluye para controlar el error y ajustar medias de tratamiento. El peso inicial fue en gran medida la base para distribuir los animales en los tres bloques, pero aún quedó suficiente variación como para recomendar su medición. La cantidad de forraje consumido está afectado por los tratamientos. Tal variable se introduce para contribuir a la interpretación de los datos. Las diferencias entre medias de tratamiento ajustadas por aumento de peso puede deberse a diferencias en lo apetitoso del sabor y, por lo tanto, en consumo de forraje, o pudo deberse a diferencias en el valor nutritivo o a ambas cosas. Así, la comparación entre medias de tratamientos para aumentos ajustados por consumo de información sobre los valores nutritivos de los tratamientos y por lo tanto ayuda en la interpretación de los datos.

Tabla 17.12 Peso inicial  $X_1$ , forraje consumido  $X_2$  y aumento de peso  $Y$ , todos en gramos, en un ensayo de nutrición con conejillos de indias.

Tratamiento del suelo		No fertilizados			Fertilizados		
Bloque		$X_1$	$X_2$	$Y$	$X_1$	$X_2$	$Y$
Franco limoso de Miami							
1		220	1,155	224	222	1,326	237
2		246	1,423	289	268	1,559	265
3		262	1,576	280	314	1,528	256
Total		728	4,154	793	804	4,413	758
Media		242.7	1,384.7	264.3	268.0	1,471.0	252.7
Arena fina de Plainfield							
1		198	1,092	118	205	1,154	82
2		266	1,703	191	236	1,250	117
3		335	1,546	115	268	1,667	117
Total		799	4,341	424	709	4,071	316
Media		266.3	1,447.0	141.3	236.3	1,357.0	105.3
Franco limoso de Almena							
1		213	1,573	242	188	1,381	184
2		236	1,730	270	259	1,363	129
3		288	1,593	198	300	1,564	212
Total		737	4,896	710	747	4,308	525
Media		245.7	1,632.0	236.7	249.0	1,436.0	175.0
Turba de Carlisle							
1		256	1,532	241	202	1,375	239
2		278	1,220	185	216	1,170	207
3		283	1,232	185	225	1,273	227
Total		817	3,984	611	643	3,818	673
Media		272.3	1,328.0	203.7	214.3	1,272.7	224.3
$\sum X_1 = 5,984$		$\sum X_2 = 33,985$		$\sum Y = 4,810$			
$\bar{X}_1 = 249.3$		$\bar{X}_2 = 1,416.0$		$\bar{Y} = 200.4$			
$\sum X_1^2 = 1,526,422$		$\sum X_2^2 = 48,971,371$		$\sum Y^2 = 1,045,898$			
$\sum X_1 X_2 = 8,555,357$		$\sum X_1 Y = 1,199,664$		$\sum X_2 Y = 6,904,945$			

Fuente: Datos obtenidos por cortesía de W. Wedin, anteriormente en la Universidad de Wisconsin, Madison, Wisconsin.

**Tabla 17.13 Suma de productos para el análisis de la covarianza múltiple de los datos de la tabla 17.12**

Fuente de variación	gl	Sumas de productos					
		$X_1, X_1$	$X_2, X_2$	$X_1, X_2$	$X_1, Y$	$X_2, Y$	$Y, Y$
Total	23	34,411	847,195	81,764	371	93,785	81,894
Bloques	2	20,397	122,438	49,815	917	2,835	496
Tipos de suelo	3	480	165,833	3,006	612	5,008	57,176
Fertilizante	1	1,320	24,384	5,674	1,973	8,479	2,948
Interacción	3	6,055	61,163	5,491	-2,902	11,284	5,545
Error	14	6,159	473,377	17,778	-229	66,179	15,729

El procedimiento es como sigue:

1. Obtener la suma de productos como aparece en las secc. 17.5 y 17.10. (Este es un experimento factorial con fertilizantes y tipos de suelo como factores). Los resultados se dan en la tabla 17.13.
2. Calcular los conjuntos necesarios de coeficientes de regresión parcial y las correspondientes sumas de cuadrados atribuibles a regresión. Por definición, los parciales han sido ajustados por todas las demás componentes en el modelo. Así, para el modelo completo, ahora necesitamos resolver la ec. (17.12), donde los  $E_{ij}$  son las sumas de cuadrados o productos de residuos una vez eliminados los efectos de bloque y tratamiento. Comparar con la línea del error de la tabla 17.1. Los subíndices 1, 2 y  $Y$  se refieren a las variables  $X_1$ ,  $X_2$  y  $Y$ .

$$\begin{aligned} E_{11}b_{Y1 \cdot 2} + E_{12}b_{Y2 \cdot 1} &= E_{1Y} \\ E_{12}b_{Y1 \cdot 2} + E_{22}b_{Y2 \cdot 1} &= E_{2Y} \end{aligned} \quad (17.12)$$

Relacionar estas ecuaciones con la ecuación matricial (14.10) y con la numérica inmediatamente después de la ec. (14.11)

Para cada modelo asociado con una hipótesis nula, los correspondientes  $T_{ij}$  de tratamiento deben sumarse a los  $E_{ij}$  de los errores para obtener los  $S_{ij}$  tal como en la tabla 17.1. Así tenemos

- 2a. Para el *error* las ecuaciones son

$$6,159b_{Y1 \cdot 2} + 17,778b_{Y2 \cdot 1} = -229$$

$$17,778b_{Y1 \cdot 2} + 473,377b_{Y2 \cdot 1} = 66,179$$

Su solución es  $b_{Y1 \cdot 2} = -0.4944$  y  $b_{Y2 \cdot 1} = 0.1584$  para el modelo completo.

- 2b. Por *tipos de suelo más error*, las ecuaciones son

$$6,639b_{Y1 \cdot 2} + 20,784b_{Y2 \cdot 1} = 383$$

$$20,784b_{Y1 \cdot 2} + 639,210b_{Y2 \cdot 1} = 71,187$$

**Tabla 17.14 Cálculo de las sumas de cuadrados de tratamientos ajustadas para el análisis de la covarianza múltiple comenzado en la tabla 17.13**

Fuente de variación	gl	Sumas de cuadrados			gl	Y ajustados por $X_1$ y $X_2$		
		$S_{Yr}^2$	$b_{Y1 \cdot 2} S_{1r}^2$	$b_{Y2 \cdot 1} S_{2r}^2$		SS	CM	F
Error	14	15,729	113.2	10,482.8	12	5,133.0	427.8	
Tipos de suelo	17	72,905	-124.1	8,677.7	15	64,351.4		
+ error								
Diferencia para probar SC tipos de suelo ajustada					3	59,218.4	19,739.5	46.14**
Fertilizante + error	15	18,677	-485.2	12,176.7	13	6,985.5		
Diferencias para probar SC fertilizantes ajustada.					1	1,852.5	1,852.5	4.33
Interacción + error	17	21,274	1,817.9	13,184.2	15	6,271.9		
Diferencia para probar SC interacción ajustada					3	1,138.9	379.6	0.89

<sup>†</sup> Esta notación  $S_{ij}$  es para incluir  $E_{ij}$ .

Su solución es  $b_{Y1 \cdot 2} = -0.3240$  y  $b_{Y2 \cdot 1} = 0.1219$  para este modelo reducido, y así sucesivamente.

3. Calcular la suma de cuadrados de  $Y$  ajustada por  $X_1$  y  $X_2$  a partir de la fórmula (17.13). Esta es la suma de cuadrados residual o de error.

$$E_{YY} - b_{Y1 \cdot 2} E_{1Y} - b_{Y2 \cdot 1} E_{2Y} \quad (17.13)$$

Obsérvese que los dos términos con signos negativos constituyen la SC(regresión) =  $\hat{\beta}_A' X_A' Y_A$  tenemos

$$15,729 - (-0.4944)(.229) - 0.1584(66,179) = 5,133.0$$

con  $14 - 2 = 12$  gl

Las sumas de cuadrados ajustadas para tipos de suelo más error, fertilizantes más error e interacción más error se calculan en forma similar usando  $S_{ij}$ . Estos valores se presentan en la tabla 17.14.

4. Calcular la suma de cuadrados para medias de tratamientos ajustadas como diferencias entre la suma de cuadrados ajustada para el error calculado en el paso 3 y el error apropiado más la suma de cuadrados de tratamiento calculado en el mismo paso.

Por ejemplo, la suma de cuadrados ajustada para tipo de suelo es

$$64,351.4 - 5,133.0 = 59,218.4 \quad \text{con } 15 - 12 = 3 \text{ gl.}$$

Tabla 17.15 Medias de tratamiento ajustadas para los datos de la tabla 15.12

Tratamiento	F	No F	$\bar{Y}_i$	$b_{Y1 \cdot 2}(\bar{X}_{1i} - \bar{X}_{1..})$	$b_{Y2 \cdot 1}(\bar{X}_{2i} - \bar{X}_{2..})$	Aumento medio diario ††		
						$\hat{P}_i$	Ajustado	No ajustado
Franco limoso de Miami	F	252.7	-9.2	+8.7	253.2	4.60	4.59	
	No F	264.3	+3.3	-5.0	266.0	4.84	4.81	
Arena fina de Plainfield	F	105.3	+6.4	-9.3	108.2	1.97	1.92	
	No F	141.3	-8.4	+4.9	144.8	2.63	2.57	
Franco limoso de Almena	F	175.0	+0.1	+3.2	171.7	3.12	3.18	
	No F	236.7	+1.8	+34.2	200.7	3.65	4.30	
Turba de Carlisle	F	224.3	+17.3	-22.7	229.7	4.18	4.08	
	No F	203.7	-11.4	-13.9	229.0	4.16	3.70	
Total		1,603.3	0.1	0.1	1,603.3			

†  $\hat{P}_i = \bar{Y}_i - b_{Y1 \cdot 2}(\bar{X}_{1i} - \bar{X}_{1..}) - b_{Y2 \cdot 1}(\bar{X}_{2i} - \bar{X}_{2..})$ .

†† Los aumentos diarios ajustados y no ajustados se obtienen dividiendo las  $\hat{P}_i$  y  $P_i$  por 55, el número de días en que los animales están bajo ensayo.

Las otras sumas de cuadrados ajustadas para probar las varias hipótesis nulas también se dan en la tabla 17.14.

5. Calcular los valores  $F$  tal como se necesitan. Para probar medias ajustadas de tipos de suelos,

$$F = \frac{19,739.5}{427.8} = 46.14^{**} \quad \text{con } 3 \text{ y } 12 \text{ gl}$$

Los otros valores de  $F$  también se presentan en la tabla 17.14.

Las medias de tratamiento ajustadas para aumento de peso se calculan a partir de

$$\hat{Y}_i = \bar{Y}_i - b_{Y1 \cdot 2}(\bar{X}_{1i} - \bar{X}_{1..}) - b_{Y2 \cdot 1}(\bar{X}_{2i} - \bar{X}_{2..}) \quad (17.14)$$

donde  $i$  se refiere al tratamiento para los  $4(2) = 8$  combinaciones de tratamientos, el subíndice numérico se refiere a la correspondiente variable independiente y  $\bar{X}_{1..}$  y  $\bar{X}_{2..}$  se refieren a medias generales. Por ejemplo, la media ajustada para aumento de peso de animales alimentados con forraje proveniente de un franco limoso de Miami fertilizado es

$$252.7 - (-0.4944)(268.0 - 249.3) - (0.1584)(1,471 - 1,416) = 253.2 \text{ g}$$

Todas las medias de tratamiento ajustadas por aumento de peso se dan en la tabla 17.15. El ensayo de nutrición duró 55 días y el promedio de aumento diario tanto para medias ajustadas también se dan en la misma tabla.

El error estándar de una diferencia entre dos medias de tratamiento ajustada se da en la ec. (17.15) para los tratamientos  $i$  e  $i'$

$$s_{\hat{Y}_{ij} - \hat{Y}_{i'j'}} = \sqrt{s_{Y_{\cdot \cdot 12}}^2 \left[ \frac{2}{r} + \frac{(X_{1ij} - \bar{X}_{1i'})^2 E_{22} - 2(\bar{X}_{1ij} - \bar{X}_{1i'}) \times (\bar{X}_{2ij} - \bar{X}_{2i'}) E_{12} + (\bar{X}_{2ij} - \bar{X}_{2i'})^2 E_{11}}{E_{11} E_{22} - E_{12}^2} \right]} \quad (17.15)$$

La ecuación (17.15) es una aplicación de la ec. (13.13). El  $2/r$  aparece después de la varianza de  $\bar{Y}$  en la ec. (17.14). Los términos restantes,  $E_{22}$ ,  $-E_{12}$  y  $E_{11}$ , son su divisor común,  $E_{11} E_{22} - E_{12}^2$  son elementos de la matriz de varianza-covarianza de los  $b$ .

Una fórmula aproximada correspondiente a la ec. (17.9) para una sola covariable está dada por

$$s_{\hat{Y}_{ij} - \hat{Y}_{i'j'}} = \sqrt{\frac{2}{r} s_{Y_{\cdot \cdot 12}}^2 \left[ 1 + \frac{T_{11} E_{22} - 2T_{12} E_{12} + T_{22} E_{11}}{(t-1)(E_{11} E_{22} - E_{12}^2)} \right]} \quad (17.16)$$

Como ilustración de la ec. (17.15), la desviación estándar de la diferencia en respuesta al forraje fertilizado y no fertilizado cultivado en arena de Plainfield es

$$\begin{aligned} s_{\hat{Y}_{ij} - \hat{Y}_{i'j'}} &= \\ &\sqrt{427.8 \left[ \frac{2}{3} + \frac{(236.3 - 266.3)^2 473,377 - 2(236.3 - 266.3)(1,357.0) - 1,447.0)17,778 + (1,357.0 - 1,447.0)^2 6,159}{6,159(473,377) - (17,778)^2} \right]} \\ &= 18.65 \text{ g} \end{aligned}$$

Mediante la ec. (17.16) obtenemos una desviación estándar aproximada de

$$\begin{aligned} s_{\hat{Y}_{ij} - \hat{Y}_{i'j'}} &= \\ &\sqrt{\frac{2(427.7)}{3} \left[ 1 + \frac{7,855(473,377) - 2(14,171)(17,778) + 251,380(6,159)}{(8-1)[(6,159)(473,377) - (17,778)^2]} \right]} \\ &= 18.97 \text{ g} \end{aligned}$$

La ecuación (17.15) da una desviación estándar diferente para cada comparación, mientras que (17.16) da la misma, así que produce un valor demasiado grande para ciertas comparaciones y muy pequeño para otras. Cuando la variación entre medias de tratamientos para uno o más de las variables independientes es mayor que lo que normalmente es atribuible al azar, como cuando están influidas por tratamientos, puede presentarse un serio error de la aplicación de la ec. (17.16).

**Ejercicio 17.12.1** La tabla acompañante, obtenida por cortesía de C.R. Weber, Escuela Superior del Estado de Iowa, corresponde a un experimento de bloques completos al azar con 4 repeticiones. Se sembraron once cepas de soya. Los datos y la definición de las variables son:

$X_1$  = madurez, medida en días más tarde que la variedad Hawkeye.

$X_2$  = voleamiento, medido en una escala de 1 a 5.

$Y$  = infección por cáncer del tallo, medida como un porcentaje de los tallos infectados.

Cepa	Bloque 1			Bloque 2			Bloque 3			Bloque 4		
	$X_1$	$X_2$	$Y$									
Lincoln	9	3.0	19.3	10	2.0	29.2	12	3.0	1.0	9	2.5	6.4
A7-6102	10	3.0	10.1	10	2.0	34.7	9	2.0	14.0	9	3.0	5.6
A7-6323	10	2.5	13.1	9	1.5	59.3	12	2.5	1.1	10	2.5	8.1
A7-6520	8	2.0	15.6	5	2.0	49.0	8	2.0	17.4	6	2.0	11.7
A7-6905	12	2.5	4.3	11	1.0	48.2	13	3.0	6.3	10	2.5	6.7
C-739	4	2.0	25.2	2	1.5	36.5	2	2.0	23.4	1	2.0	12.9
C-776	3	1.5	67.6	4	1.0	79.3	6	2.0	13.6	2	1.5	39.4
H-6150	7	2.0	35.1	8	2.0	40.0	7	2.0	24.7	7	2.0	4.8
L6-8477	8	2.0	14.0	8	1.5	30.2	10	1.5	7.2	7	2.0	8.9
L7-1287	9	2.5	3.3	9	2.0	35.8	13	3.0	1.1	9	3.0	2.0
Bav. Sp.	10	3.5	3.1	10	3.0	9.6	11	3.0	1.0	10	3.5	0.1

El principal objetivo fue averiguar si la madurez o el voleamiento están relacionados con la técnica más estrechamente. Esto se determinará a partir de la regresión múltiple del error. A propósito, probar la hipótesis de que no hay diferencias entre medias ajustadas por variedades.

¿Cuál es la prueba apropiada para lograr el objetivo principal?

Ejercicio 17.12.2 ¿Cuál es la ecuación apropiada para calcular medias ajustadas de tratamiento?

Ejercicio 17.12.3 Calcular una desviación estándar aproximada para la diferencia entre cualquier par de medias de tratamiento ajustadas. ¿Existe alguna razón para pensar que esta desviación estándar no se puede aplicar en este caso?

### 17.13 Cálculos de alta velocidad y salidas de computador

No es muy probable que muchos investigadores efectúan los cálculos de sus propios análisis de la covarianza con dos o más covariables. Los computadores de alta velocidad hacen esto proporcionando varias salidas impresas. La tabla 17.16 corresponde a una salida de SAS (14.13).

La tabla 17.16 da un análisis de la covarianza para los datos de la tabla 17.12. La primera partición de la suma de cuadrados total ajustadas es para MODELO y ERROR; el primero incluye la media. Recuérdese que el R-CUADRADO 100 mide el porcentaje de la SC(TOTAL CORREGIDO) atribuible al modelo. La SC(error) y el CM(error) son aproximadamente las mismas que aparecen en la tabla 17.14, elaborada en un computador de escritorio. Probablemente los cálculos de la tabla 17.16 son más exactos.

Las sumas de cuadrados tipo I se calculan secuencialmente. Estas son ortogonales para bloques y tratamientos como consecuencia del diseño experimental y las mismas que en la columna  $Y$ ,  $Y$  de la tabla 17.13. Para  $X_1$  y  $X_2$ , la suma es  $8.49 + 10,585.05 = 10,593.54$ , básicamente lo mismo que  $113.2 + 10,482.8 = 10,596.0$  según la tabla 17.14.

**Tabla 17.16 Análisis de la covarianza múltiple  
PROCEDIMIENTO DEL MODELO LINEAL GENERAL**

VARIABLE DEPENDIENTE: Y	GL	SUMA DE CUADRADOS	CUADRADO MEDIO	VALOR F	PR > F
FUENTE					
MODELO	11	76758.29697647	6978.02699786	16.31	0.0001
ERROR	12	5135.51635686	427.96136307		
TOTAL CORREGIDO	23	81893.83333333			
R-CUADRADO	C.V.	DESV EST	MEDIA Y		
0.937290	10.3221	20.68722705	200.41666667		
FUENTE	GL	TIPO I SC	VALOR F	PR > F	
BLOQUE	2	495.58333333	0.58	0.5754	
SUELO	3	57176.16666667	44.53	0.0001	
FERTILIZANTE	1	2948.16666667	6.89	0.0222	
SUELO*FERTILIZANTE	3	5544.83333333	4.32	0.0278	
X1	1	8.49366072	0.02	0.8903	
X2	1	10585.05331575	24.73	0.0003	
FUENTE	GL	TIPO IV SC	VALOR F	PR > F	
BLOQUE	2	395.40120649	0.46	0.6408	
SUELO	3	59216.44313232	46.12	0.0001	
FERTILIZANTE	1	1850.62488503	4.32	0.0597	
SUELO*FERTILIZANTE	3	1136.58247353	0.89	0.4764	
X1	1	1341.58895597	3.13	0.1020	
X2	1	10585.05331575	24.73	0.0003	

**Tabla 17.17 Análisis de la covarianza múltiple  
PROCEDIMIENTO DE MODELO LINEAL GENERAL**

MEDIAS POR MINIMOS CUADRADOS

SUELO	Y	ERROR EST	PROB >  T
	MEDIA MC	MEDIA MC	HO: MEDIA MC = 0
1	259.598426	8.594161	0.0001
2	126.545658	8.485458	0.0001
3	186.164174	9.334271	0.0001
4	229.368408	9.145719	0.0001

NOTA: PARA ASEGURAR EL NIVEL DE PROTECCION TOTAL, SOLO SE DEBERAN USAR PROBABILIDADES ASOCIADAS CON COMPARACIONES PRE-PLANEADAS.

FERTILIZANTE	Y	ERROR EST	PROB >  T	PROB >  T  HO:
	MEDIA MC	MEDIA MC	HO: MEDIA MC = 0	MEDIA 1 MC = MEDIA 2 MC
1	210.118178	6.292618	0.0001	0.0597
2	190.715155	6.292618	0.0001	

SUELO	FERTILIZANTE	Y	ERROR EST	PROB >  T
		MEDIA MC	MEDIA MC	HO: MEDIA MC = 0
1	1	266.006700	12.078556	0.0001
1	2	253.190151	12.917782	0.0001
2	1	144.833755	12.769879	0.0001
2	2	108.257562	12.444568	0.0001
3	1	200.653832	13.986703	0.0001
3	2	171.674517	11.962663	0.0001
4	1	228.978425	14.268837	0.0001
4	2	229.738391	15.152531	0.0001

En la tabla 17.16, las sumas secuenciales de cuadrados pueden interpretarse individualmente, mientras que los términos individuales de regresión  $b_{Y_1 \cdot 2} E_{1r}$  y  $b_{Y_2 \cdot 1} E_{2r}$  no tienen significado y sólo son parte de un todo.

Cada suma de cuadrados tipo IV de la tabla 17.16 es una suma de cuadrados ajustada por otras componentes del modelo. Cada una corresponde a la "diferencia para probar una SC ajustada" de la tabla 17.14. No se dan los cuadrados medios. Presumiblemente, el interés primordial en este análisis estriba en los efectos principales y la interacción. Obsérvese que podemos probar las hipótesis nulas respecto a los coeficientes de regresión parcial separados. Estas proveen información respecto al valor de  $X_1$  y  $X_2$  en el control del error.

La tabla 17.17 consta de partes de otras salidas impresas. Da medias por mínimos cuadrados o medias de tratamiento ajustadas para cada nivel de ambos efectos principales y para todas las combinaciones de tratamientos. La codificación para suelos es para la secuencia en la tabla 17.12. Para fertilizantes, 1 corresponde a no fertilizado y 2 para fertilizado. Se da el error estándar para cada media ajustada; éstos son desiguales debido a los ajustes.

Las hipótesis individuales de que cada media de población es cero no son realistas en este caso, aunque se prueban. La prueba de la diferencia entre medias de fertilizantes

ya fue hecha, con  $F$ , en la tabla 17.16; obsérvese que la probabilidad de encontrar un valor mayor del criterio de prueba es la misma para  $F$  que para  $|t|$ . Obsérvese también la llamada de atención respecto a las comparaciones no planeadas. Esto es aquí importante ya que el programa exige comparaciones por pares de medias de todas las medias (en esta salida impresa se suprime las comparaciones).

**Ejercicio 17.13.1** Preparar los datos del ejercicio 17.12.1 para cálculos a alta velocidad en un computador. Comparar los resultados de salida impresa con la de los cálculos anteriores.

Comparar las salidas impresas con las que se dan en las tablas 17.16 y 17.17.

## Referencias

- 17.1. Bartlett, M. S.: "Some examples of statistical methods of research in agriculture and applied biology," *J. Roy. Statist. Soc. Suppl.*, 4:137-183 (1937).
- 17.2. Cochran, W. G., y G. M. Cox: *Experimental Designs*, 2a. ed., Wiley, Nueva York, 1957.
- 17.3. DeLury, D. B.: "The analysis of covariance," *Biom.*, 4:153-170 (1948).
- 17.4. Federer, W. T., y C. S. Schlottfeldt: "The use of covariance to control gradients in experiments," *Biom.*, 10:282-290 (1954).
- 17.5. Finney, D. J.: "Standard errors of yields adjusted for regression on an independent measurement," *Biom. Bull.*, 2:53-55 (1946).
- 17.6. Forester, H. C.: "Design of agronomic experiments for plots differentiated in fertility by past treatments," *Iowa Agr. Exp. Sta. Res. Bull.* 226, 1937.
- 17.7. Love, H. H.: "Are uniformity trials useful?" *J. Amer. Soc. Agron.*, 28:234-245 (1936).
- 17.8. Matrone, G., F. H. Smith, V. B. Weldon, W. W. Woodhouse, Jr., W. J. Peterson, y K. C. Beeson: "Effects of phosphate fertilization on the nutritive value of soybean forage for sheep and rabbits," *U.S. Dep. Agr. Tech. Bull.* 1086, 1954.
- 17.9. Outhwaite, A. D. y A. Rutherford: "Covariance analysis as alternative to stratification in the control of gradients," *Biom.*, 11:431-440 (1955).
- 17.10. Steel, R. G. D.: "Which dependent variable?  $Y$  or  $Y - X$ ?" *Mimeo Series BU-54-M*, Biometrics Unit, Cornell University, Ithaca, N.Y., 1954.
- 17.11. Tucker, H. P., J. T. Wakeley, y F. D. Cochran: "Effect of washing and removing excess moisture by wiping or by air current on the ascorbic acid content of turnip greens," *S. Coop. Ser. Bull.*, 10, 54-56 (1951).
- 17.12. Wishart, J.: "Tests of significance in the analysis of covariance," *J. Roy. Statist. Soc. Suppl.*, 3:79-82 (1936).
- 17.13. Wishart, J.: "Growth-rate determinations in nutrition studies with the bacon pig and their analysis," *Biometrika*, 30:16-28 (1938).
- 17.14 *Biom.*, 13:261-405, No. 3 (1957). (Este número consta de siete trabajos dedicados a la covarianza).

**ANALISIS DE LA VARIANZA V:  
NUMERO DESIGUAL DE SUBCLASES****18.1 Introducción**

En el capítulo 9 se trató del diseño de bloques completos al azar. Los datos provenientes de tal diseño normalmente se presentan bajo una clasificación de dos vías; bloques y tratamientos. En algunos casos, se puede tomar más de una observación por tratamiento dentro de un bloque. Estas pueden tomarse en partes separadas de una sola unidad experimental; esto corresponde al muestreo y su análisis apropiado se dio en la tabla 9.10. O bien, estas observaciones pueden tomarse en unidades experimentales separadas; este caso también se vio en la sec. 9.8 como un diseño de bloques generalizado. En este capítulo estudiaremos más detalladamente el diseño de bloques generalizado con números de subclases proporcionales, pero dedicaremos más tiempo al análisis de datos con un número no proporcionado de subclases.

**18.2 Observaciones múltiples dentro de subclases**

Los datos a menudo se presentan en una clasificación de dos vías, donde las *celdas* o *subclases* contienen observaciones sobre más de una unidad experimental. Este sería el caso para el diseño de bloques aleatorizado generalizado expuesto en la sec. 9.8. Aquí, cada uno de los tratamientos  $t$  aparece  $n_i$  veces en el bloque  $i$ -ésimo y los tratamientos se aleatorizan en  $n_i t$  unidades experimentales que forman el bloque.

En algunos casos, como el del ejercicio 9.8:1, los bloques pueden corresponder a un sistema de clasificación diferente al de bloques puramente físicos de las unidades experimentales. En la práctica, la aleatorización puede ser un supuesto, con tal que sea razonable. Finalmente, el número de observaciones por celda no tiene que ser igual ni proporcional.

Una gran cantidad de datos caen en esta categoría general. Una fuente son los experimentos con animales domésticos y salvajes y con pájaros. Por ejemplo, un investigador

puede interesarse por el tamaño de nidadas de faisanes, de hembras salvajes y de hembras soltadas en la primavera en diferentes localidades. Es claro que los números que pueden observarse en estado salvaje no se pueden controlar, y las posibilidades de muestras pueden ser limitadas; el número de las que se han soltado debe depender de la disponibilidad de habitat. En tales datos, casi con seguridad, entra un número no proporcionado de subclases.

Cuando las observaciones se hacen sobre unidades experimentales en vez de en unidades de muestreo, la variación entre unidades dentro de una celda mide el verdadero error experimental o error puro. En consecuencia, la descripción matemática de una observación es

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad (18.1)$$

Hemos usado  $\alpha$  y  $\beta$  en vez de  $\tau$  y  $\rho$  para destacar el hecho de que para datos con número no proporcionado de subclases, los llamados bloques a menudo corresponden a un sistema de clasificación. En este caso, la posibilidad de una interacción de bloques por tratamientos debe presentarse como una hipótesis. Los supuestos del modelo con que los  $\varepsilon$  tienen distribución normal e independiente con media cero y varianza común; los  $\alpha$  y  $\beta$  pueden ser fijos o aleatorios; los  $(\alpha\beta)$  son componentes de interacción. Este modelo contrasta con aquel en el cual las observaciones se hacen en unidades de muestreo.

### 18.3 Análisis de un número proporcionado de subclases

En la tabla 18.1 se da una ilustración de números proporcionales de subclases. El análisis no presenta dificultades particulares. La suma de cuadrados del error es la suma de las sumas de cuadrados dentro de las celdas. La suma de cuadrados de un efecto principal se calcula a partir de totales marginales apropiados, como para un experimento de un solo factor con números desiguales; ésta es una suma ponderada de cuadrados de desviaciones. La suma de cuadrados de la interacción se obtiene restando las sumas de cuadrados de los efectos principales de la suma ponderada de cuadrados entre totales de las celdas, esto es,

$$\sum_{i,j} \frac{Y_{ij}^2}{n_{ij}} - \frac{Y_{..}^2}{n_{..}} = SC(A) - SC(B) \quad (18.2)$$

Tabla 18.1 Análisis de números de subclases proporcionales

Modelo:  $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$

Factor	B	Número de observaciones			Análisis de la varianza	
		Totales A			Fuente	gl
A	Nivel	$b_1$	$b_2$	$b_3$	$A$	1
	$a_1$	3	4	6	$B$	2
	$a_2$	6	8	12	$AB$	2
	Total B	9	12	18	Error	33
	Total A	13	26	39	Total	38

donde  $Y_{ij}$  representa el total de una celda con  $n_{ij}$  observaciones, o como

$$\frac{(a_1 b_1)^2}{3} + \cdots + \frac{(a_2 b_3)^2}{12} - \frac{Y^2}{39} = SC(A) - SC(B)$$

para la ilustración.

Las pruebas de hipótesis dependen de los supuestos de las componentes del modelo.

En el desarrollo de un análisis de un conjunto de datos, se necesita estimar los diferentes parámetros desconocidos del modelo. Recuérdese que  $SC(\text{modelo}) = \hat{\beta}' X' Y$ . Esto es cierto tanto para datos balanceados, tales como los que provienen de un experimento de bloques completos al azar, como para los datos no balanceados, aquellos con número no proporcionado de subclases. (Una vez se ha desarrollado el análisis, nunca podemos usar a sabiendas las estimaciones). Con el objeto de obtener un conjunto único de estimaciones, generalmente es necesario imponer restricciones al modelo o limitaciones a la solución; esto último con el único propósito de obtener una solución, como se expuso en la sec. 7.5. Supóngase que se imponen las restricciones. Para el diseño de bloques completos al azar con una observación por celda, la expresión matemática para una observación es  $Y_{ij} = \mu + \tau_i + \rho_j + \varepsilon_{ij}$ . En la población,  $\sum \tau_i = 0$  (efectos fijos) o  $\mu_\tau = 0$  (efectos aleatorios), así que necesitamos que  $\sum \hat{\tau}_i = 0$ , donde  $\hat{\tau}_i$  es nuestra estimación de  $\tau_i$ , ya que eso se deduce naturalmente del modelo. También se exige que  $\sum \hat{\rho}_j = 0$ .

Para la ilustración de la tabla 18.1, donde se supone la existencia de la interacción, el análisis bosquejado impone las restricciones de que las sumas ponderadas de  $\hat{\alpha}$ , de  $\hat{\beta}$  de  $(\hat{\alpha}\hat{\beta})$  sea cero, donde las ponderaciones son esencialmente las proporciones. Aquí

$$\begin{aligned}\hat{\alpha}_1 + 2\hat{\alpha}_2 &\approx 0 \\ 3\hat{\beta}_1 + 4\hat{\beta}_2 + 6\hat{\beta}_3 &= 0 \\ (\hat{\alpha}\hat{\beta})_{11} + 2(\hat{\alpha}\hat{\beta})_{21} &= 0 \\ (\hat{\alpha}\hat{\beta})_{12} + 2(\hat{\alpha}\hat{\beta})_{22} &= 0 \\ (\hat{\alpha}\hat{\beta})_{13} + 2(\hat{\alpha}\hat{\beta})_{23} &= 0 \\ 3(\hat{\alpha}\hat{\beta})_{11} + 4(\hat{\alpha}\hat{\beta})_{12} + 6(\hat{\alpha}\hat{\beta})_{13} &= 0\end{aligned}$$

y

$$3(\hat{\alpha}\hat{\beta})_{21} + 4(\hat{\alpha}\hat{\beta})_{22} + 6(\hat{\alpha}\hat{\beta})_{23} = 0$$

Para estas restricciones, la media experimental  $\bar{Y}_{..}$ , una media ponderada de las medias de las celdas, es una estimación insesgada de  $\mu$ . El examen de los totales y medias de tratamiento de la tabla 18.2 muestra que esto constituye un conjunto razonable de restricciones desde el punto de vista puramente de los cálculos. Por ejemplo, para estimar la varianza entre los  $\beta$ , o la suma de los cuadrados de los  $\beta$ , debemos eliminar los  $(\alpha\beta)$  de las medias de  $B$ , ya que difieren de una media a otra y, en consecuencia, contribuirían a la varianza entre las medias  $B$ . Por otra parte, las restricciones implican, como es de esperar, las fre-

Tabla 18.2 Totales y medias de tratamiento para la ilustración de la tabla 18.1

Factor	B		Totales A		Medias A	
	Nivel $b_1$	$b_2$	$b_3$			
$A$	$3\mu + 3x_1 + 3\beta_1 + 3(\alpha\beta)_{11} + \epsilon_{11}$	$4\mu + 4x_1 + 4\beta_2 + 4(\alpha\beta)_{12} + \epsilon_{12}$	$6\mu + 6x_1 + 6\beta_3 + 6(\alpha\beta)_{13} + \epsilon_{13}$	$13\mu + 13x_1 + 3\beta_1 + 4\beta_2 + 4(\alpha\beta)_{11} + 6\beta_3 + 3(\alpha\beta)_{12} + 6(\alpha\beta)_{13} + \epsilon_{1..}$	$\mu + x_1 + \frac{1}{13}\beta_1 + \frac{4}{13}\beta_2 + \frac{6}{13}\beta_3 + \frac{1}{13}(x\beta)_{11} + \frac{4}{13}(x\beta)_{12} + \frac{6}{13}(x\beta)_{13} + \epsilon_{1..}$	
	$6\mu + 6x_2 + 6\beta_1 + 6(\alpha\beta)_{21} + \epsilon_{21}$	$8\mu + 8x_2 + 8\beta_2 + 8(\alpha\beta)_{22} + \epsilon_{22}$	$12\mu + 12x_2 + 12\beta_3 + 12(\alpha\beta)_{23} + \epsilon_{23}$	$26\mu + 26x_2 + 6\beta_1 + 8\beta_2 + 12\beta_3 + 6(x\beta)_{21} + 8(x\beta)_{22} + 12(x\beta)_{23} + \epsilon_{2..}$	$\mu + x_2 + \frac{1}{13}\beta_1 + \frac{4}{13}\beta_2 + \frac{6}{13}\beta_3 + \frac{1}{13}(x\beta)_{21} + \frac{4}{13}(x\beta)_{22} + \frac{6}{13}(x\beta)_{23} + \epsilon_{2..}$	
Totales B	$9\mu + 3x_1 + 6x_2 + 9\beta_1 + 3(\alpha\beta)_{11} + 6(\alpha\beta)_{21} + \epsilon_{1..}$	$12\mu + 4x_1 + 8x_2 + 12\beta_2 + 4(\alpha\beta)_{12} + 8(\alpha\beta)_{22} + \epsilon_{2..}$	$18\mu + 6x_1 + 12x_2 + 18\beta_3 + 6(\alpha\beta)_{13} + 8(x\beta)_{23} + \epsilon_{3..}$	$39\mu + 13x_1 + 26x_2 + 18\beta_1 + 18\beta_2 + 18\beta_3 + 3(\alpha\beta)_{11} + 4(\alpha\beta)_{12} + 6(\alpha\beta)_{13} + 8(x\beta)_{12} + 12(x\beta)_{13} + \epsilon_{1..}$	$\mu + \frac{1}{13}x_1 + \frac{4}{13}x_2 + \frac{1}{13}\beta_1 + \frac{4}{13}\beta_2 + \frac{6}{13}\beta_3 + \frac{1}{13}(x\beta)_{11} + \frac{4}{13}(x\beta)_{12} + \frac{6}{13}(x\beta)_{13} + \epsilon_{1..}$	
	$\mu + \frac{1}{3}x_1 + \frac{2}{3}x_2 + \beta_1 + \frac{1}{3}(\alpha\beta)_{11} + \frac{2}{3}(\alpha\beta)_{21} + \epsilon_{1..}$	$\mu + \frac{1}{3}x_1 + \frac{2}{3}x_2 + \beta_2 + \frac{1}{3}(\alpha\beta)_{12} + \frac{2}{3}(\alpha\beta)_{22} + \epsilon_{2..}$	$\mu + \frac{1}{3}x_1 + \frac{2}{3}x_2 + \beta_3 + \frac{1}{3}(\alpha\beta)_{13} + \frac{2}{3}(\alpha\beta)_{23} + \epsilon_{3..}$	$\mu + \frac{1}{3}x_1 + \frac{2}{3}x_2 + \beta_1 + \frac{1}{3}\beta_2 + \frac{2}{3}\beta_3 + \frac{1}{3}(\alpha\beta)_{11} + \frac{4}{3}(\alpha\beta)_{12} + \frac{5}{3}(\alpha\beta)_{13} + \frac{1}{3}(\alpha\beta)_{21} + \frac{2}{3}(\alpha\beta)_{22} + \frac{5}{3}(\alpha\beta)_{23} + \tilde{\epsilon}_{1..}$	$\mu + \frac{1}{3}x_1 + \frac{2}{3}x_2 + \beta_1 + \frac{1}{3}\beta_2 + \frac{2}{3}\beta_3 + \frac{1}{3}(\alpha\beta)_{11} + \frac{4}{3}(\alpha\beta)_{12} + \frac{5}{3}(\alpha\beta)_{13} + \frac{1}{3}(\alpha\beta)_{21} + \frac{2}{3}(\alpha\beta)_{22} + \frac{5}{3}(\alpha\beta)_{23} + \tilde{\epsilon}_{1..}$	
Medias B				$\text{Media no ponderada} = \frac{1}{6}[(\mu + x_1 + \beta_1 + (\alpha\beta)_{11} + \epsilon_{11}) + (\mu + x_1 + \beta_1 + (\alpha\beta)_{12} + \epsilon_{12}) + (\mu + x_1 + \beta_3 + (\alpha\beta)_{13} + \epsilon_{13}) + (\mu + x_2 + \beta_2 + (\alpha\beta)_{22} + \epsilon_{22}) + (\mu + x_2 + \beta_3 + (\alpha\beta)_{23} + \epsilon_{23})] = \mu + [3 \sum x_i + 2 \sum \beta_j + \Sigma (\alpha\beta)_{ij} + (\sum \tilde{\epsilon}_{ij})]/6$		

cuencias relativas con las cuales los varios  $\alpha$ ,  $\beta$ , y  $(\alpha\beta)$  deben estar presentes en sus poblaciones respectivas. Esta es información respecto al modelo y determina el proceso de cálculo.

Se pueden presentar otras restricciones en un análisis de una forma diferente. Considerese el siguiente conjunto diferente de restricciones

$$\sum \hat{\alpha}_i = 0 \quad \sum \hat{\beta}_j = 0 \quad \sum_i (\hat{\alpha}\hat{\beta})_{ij} = 0 \quad \text{para todo } j$$

$$\sum_j (\hat{\alpha}\hat{\beta})_{ij} = 0 \quad \text{para todo } i$$

Ahora bien, el promedio *no ponderado* de las medias de las celdas es una estimación no sesgada de  $\mu$ , como puede verse en la tabla 18.2. Para estas restricciones, el análisis es diferente y más engorroso que el descrito al principio de esta sección.

Ahora supóngase que el modelo realmente no incluye una interacción. Al eliminar los  $(\alpha\beta)$  en la tabla 18.2, vemos que todas las medias  $B$  contienen la misma suma ponderada de las  $\alpha$ . Así que no contribuyen en nada a la varianza entre las medias  $B$ . Cualquier tipo de restricciones razonable sobre las  $\alpha$  no afectaría nuestra estimación de la varianza o de la suma de cuadrados entre las  $\beta$ .

Así, resultará claro que la elección de un conjunto de restricciones, sobre las estimaciones de los diversos efectos, tienen un efecto real sobre el análisis de la varianza cuando hay interacción, y esta elección deberá hacerse con base en supuestos que son parte del modelo.

**Ejercicio 18.3.1** En el ejercicio 15.3.3, se da un conjunto de datos balanceados. A este conjunto agréguese las siguientes observaciones para estudiantes que prefirieron imprenta a braille.

Datos	Prueba previa		Prueba posterior		
	Tratamiento:	1	2	1	2
	98,64	87,89	93,80	87,88	

Analizar todos los datos de la prueba previa usando el método de esta sección.

Analizar todos los datos de la prueba posterior usando el método de esta sección.

#### 18.4 Análisis de un número no proporcionado de subclases

Ante todo, el modelo y el análisis para la clasificación de una vía con desigual número de observaciones por tratamiento se expuso ya en el cap. 7. La suma de cuadrados del error es una SC(dentro de tratamientos) y la SC(tratamientos) es una suma ponderada de cuadrados entre medias con ponderaciones que dependen de la varianza de cada media.

Para el análisis de datos en una clasificación de dos vías con un número de subclases no proporcionado, primero obsérvese la tabla 18.3. Es claro que no podemos calcular una

**Tabla 18.3 Número de observaciones y medias de filas y columnas para ilustrar un caso de un número no proporcionado de subclases.**

Modelo:  $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$ , sin interacción.

		Números				Medias A
Números	$a_i$	$b_1$	$b_2$	$b_3$	Total	
		3	5	12	20	$\mu + \alpha_1 + \frac{1}{10}\beta_1 + \frac{1}{10}\beta_2 + \frac{12}{10}\beta_3 + \epsilon_{1..}$
	$a_2$	6	7	8	21	$\mu + \alpha_2 + \frac{6}{11}\beta_1 + \frac{7}{11}\beta_2 + \frac{8}{11}\beta_3 + \epsilon_{2..}$
Total		9	12	20	41	
Medias B			$\mu + \frac{1}{2}\alpha_1 + \frac{5}{8}\alpha_2$	$\mu + \frac{1}{2}\alpha_1 + \frac{1}{12}\alpha_2$	$\mu + \frac{11}{10}\alpha_1 + \frac{5}{8}\alpha_2$	$+ \beta_1 + \epsilon_{1..}$
			$+ \beta_2 + \epsilon_{2..}$	$+ \beta_3 + \epsilon_{3..}$		

varianza entre las medias ponderadas  $A$  exenta del efecto  $B$ . No hay restricción sobre las  $\beta$  que pueda eliminar esta dificultad ya que la razón del coeficiente de  $\beta_1$  al coeficiente de  $\beta_2$  varía con el nivel  $A$ . Una posibilidad es obtener las medias de filas y columnas como promedios no ponderados de medias de celdas y usar la restricción de que  $\sum \hat{\beta}_j = 0$ , con la condición de que exista como mínimo un registro en cada celda. Esto plantea el problema de la eficiencia, ya que las medias tienen diferentes varianzas. Una dificultad similar resulta a propósito de la suma de cuadrados para el factor  $B$ .

La imposibilidad de obtener una suma de cuadrados para  $A$  o para  $B$  que esté libre del otro factor se resume el decir que los datos *no son ortogonales*. Como resultado nos atenemos al método básico de mínimos cuadrados. Este método se le llamó alguna vez *método de ajuste de constantes* y se recomendaba cuando el modelo no incluía interacción, aunque podría efectuarse una prueba de interacción con poco más esfuerzo. En ese tiempo, los métodos de cálculo eran menos perfeccionados y el ajuste de las otras constantes de interacción era difícil.

El método se aplica cuando algunas celdas no tienen observaciones. Puede usarse con covariables. Para las técnicas, nos apoyamos en la exposición sobre regresión dada en los caps. 13 y 14.

Hacemos la ilustración estableciendo primeramente un problema con  $n_{11} = 3$  y  $n_{12} = 1$ ,  $n_{21} = 4$  y  $n_{22} = 2$ . Teniendo en cuenta la ec. (18.1) el conjunto de observaciones se escribe de la siguiente manera

$$\begin{array}{l}
 \begin{array}{ccccccccc}
 \mu & \alpha_1 & \alpha_2 & \beta_1 & \beta_2 & (\alpha\beta)_{11} & (\alpha\beta)_{12} & (\alpha\beta)_{21} & (\alpha\beta)_{22} \\
 \left( \begin{array}{c} Y_{111} \\ Y_{112} \\ Y_{113} \\ Y_{121} \\ Y_{211} \\ Y_{212} \\ Y_{213} \\ Y_{214} \\ Y_{221} \\ Y_{222} \end{array} \right) = \left( \begin{array}{c} 1 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \\ 1 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \\ 1 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \\ 1 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \\ 1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \\ 1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \\ 1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \\ 1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \\ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 \\ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 \end{array} \right) + \left( \begin{array}{c} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \\ (\alpha\beta)_{21} \\ (\alpha\beta)_{22} \end{array} \right) + \left( \begin{array}{c} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{113} \\ \epsilon_{121} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{213} \\ \epsilon_{214} \\ \epsilon_{221} \\ \epsilon_{222} \end{array} \right) \quad (18.3)
 \end{array}
 \end{array}$$

La ecuación (18.3) es simplemente  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ . Los parámetros que se colocan sobre la matriz  $\mathbf{X}$  son una repetición conveniente de los elementos del vector  $\boldsymbol{\beta}$ . Todos los nueve  $X$  son variables ficticias o indicadoras que registran la presencia (1) o ausencia (0) del parámetro. Así, la primera ecuación dice que  $Y_{111} = \mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11} + \varepsilon_{111}$ .

Las ecuaciones por mínimos cuadrados son  $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$  donde

$$\mathbf{X}'\mathbf{X} = \left( \begin{array}{c|cc|cc|ccccc} 10 & 4 & 6 & 7 & 3 & 3 & 1 & 4 & 2 \\ \hline 4 & 4 & 0 & 3 & 1 & 3 & 1 & 0 & 0 \\ 6 & 0 & 6 & 4 & 2 & 0 & 0 & 4 & 2 \\ \hline 7 & 3 & 4 & 7 & 0 & 3 & 0 & 4 & 0 \\ 3 & 1 & 2 & 0 & 3 & 0 & 1 & 0 & 2 \\ \hline 3 & 3 & 0 & 3 & 0 & 3 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 4 & 0 & 4 & 4 & 0 & 0 & 0 & 4 & 0 \\ 2 & 0 & 2 & 0 & 2 & 0 & 0 & 0 & 2 \end{array} \right) \quad (18.4)$$

$\mathbf{Y}$

$$\mathbf{X}'\mathbf{Y} = \left( \begin{array}{c} Y_{...} \\ Y_{1..} \\ Y_{2..} \\ Y_{.1} \\ Y_{.2} \\ Y_{11..} \\ Y_{12..} \\ Y_{21..} \\ Y_{22..} \end{array} \right)$$

La solución de estas ecuaciones depende de  $(\mathbf{X}'\mathbf{X})^{-1}$  si es que existe. Sin embargo, notamos que la suma de los coeficientes de las filas 2 y 3 es igual a la de los coeficientes de la fila 1, que la suma de los coeficientes de las filas 4 y 5 es igual a la de los de la fila 1, que la suma de los de las filas 6 y 7 es igual a la de los de la fila 2, que la suma de los de las filas 6 y 8 es igual a la de los de la fila 4, que la suma de los de las filas 7 y 9 es igual a la de los de la fila 5, y que la suma de los de las filas 8 y 9 es igual a la de los de la fila 3. El modelo está sobreparametrizado y  $\mathbf{X}'\mathbf{X}$  es singular.

Cuando  $\mathbf{X}'\mathbf{X}$  es singular, las ecuaciones normales tienen muchas soluciones. Para obtener una solución única, se necesita de restricciones sobre el modelo o de limitaciones a la solución. Suponemos que se emplean computadores de alta velocidad con sus correspondientes salidas impresas.

Eugene J. Eisen efectuó un experimento con ratones en la Universidad del Estado de Carolina del Norte. Entre las características estudiadas estaba el tamaño de la camada. Esta es una variable discreta, pero los números van desde 1 hasta 25. Aquí se analiza esta variable como una variable continua, para la cual los supuestos de normalidad e independencia y de homogeneidad del error son válidos.

El tamaño de la camada puede verse afectado por la historia de la selección de la población y por el tamaño de la camada en la cual se crio la madre. Estas son las fuentes de variación aquí estudiadas. Los datos provienen de cinco líneas.

11. Seleccionados por tamaño grande de la camada  $L$
12. Seleccionados por peso elevado del cuerpo a las seis semanas  $W$
14. Seleccionados por un índice que combine  $L$  bajo y  $W$  elevado
15. Un control no seleccionado
16. Seleccionados por un índice que combine  $L$  alto y  $W$  bajo

El número de crianza, NREAR en las tablas de las salidas impresas, se refiere al tamaño de la camada en la cual se criaron las hembras. Estas se han ajustado artificialmente según se necesite en el momento del nacimiento a 8, 12 ó 16. En la madurez, las hembras han producido, a su turno, camadas cuyo tamaño corresponde a la observación considerada.

Los números de las celdas iban de 44 a 50. Esta desproporción está lejos de ser seria, pero los números no son aún ni iguales ni proporcionales.

La salida del SAS(14.13) disponible proporciona, para cada celda,  $n_{ij}$ ,  $Y_{ij}$ ,  $\bar{Y}_{ij}$ ,  $\min Y_{ijk}$ ,  $\max Y_{ijk}$ ,  $s^2_{ij}$ ,  $s_{ij}/\sqrt{n_{ij}}$  y el coeficiente de variación. Todo esto es información útil. La tabla 18.4 da solamente los  $\bar{Y}_{ij}$ , y los  $n_{ij}$ .

La tabla 18.5 es una salida impresa del análisis de la varianza. Había 717 observaciones, así que hay 716 gl para la suma total de cuadrados corregida o ajustada. La SC(modelo) es

$$\hat{\beta}' \mathbf{X}' \mathbf{Y} - \bar{Y}_{..}^2/n_{..} = \sum \left( \frac{Y_{ij}^2}{n_{ij}} \right) - \frac{\bar{Y}_{..}^2}{n_{..}} = \text{SC(entre celdas)}$$

**Tabla 18.4 Tamaño medio de camada y número de camadas observadas,  $\bar{Y}_{ij}(n_{ij})$ , para poblaciones seleccionadas de acuerdo con el tamaño de la camada en la cual fue criada la hembra.**

		Hembra criada en tamaño de la camada		
		8	12	16
Población = línea	11	18.23(48)	17.55(47)	16.48(48)
	12	14.53(49)	14.00(49)	13.74(47)
	14	10.87(47)	11.17(47)	10.29(48)
	15	12.16(49)	11.84(49)	11.96(50)
	16	14.40(50)	14.80(44)	13.67(45)

*Fuente:* Datos obtenidos por cortesía de Eugene J. Eisen, Universidad del Estado de Carolina del Norte, Raleigh, Carolina del Norte.

Tabla 18.5 Análisis de la varianza de los datos de los ratones

PROCEDIMIENTO GENERAL PARA MODELOS LINEALES								
Modelo: $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$								
VARIABLE DEPENDIENTE: Y		GL	SUMA DE CUADRADOS CUADRADO MEDIO	VALOR F	PR>F	R-CUADRADO C.V.		
FUENTE								
MODELO	14	3840.25238981	274.30374213	29.06	0.0001	0.366936 22.4195		
ERROR	702	6625.47145956	9.43799353		DESV EST	MEDIA		
TOTAL CORREGIDO	716	10465.73384937			3.07213176	13.70292887		
FUENTE	GL	TIPO I SC	VALOR F	PR>F	GL	TIPO III SC	VALOR	PR > F
POP	4	3698.99948468	97.98	0.0001	4	3695.93018099	97.90	0.0001
NREAR	2	85.66916501	4.54	0.0110	2	87.53206237	4.64	0.0100
POR*NREAR	8	55.58374012	0.74	0.6596	8	55.58374012	0.74	0.6596

La SC(modelo) se partitiona más todavía en sumas de cuadrados de efecto principal y de interacción. Las entradas SC TIPO I se han calculado en forma secuencial, así que sólo la interacción ha sido ajustada por todos los demás elementos en el modelo. Las entradas SC TIPO III son sumas de cuadrados para reducciones atribuibles a la inclusión de elementos correspondientes a la fuente nombrada, última en el modelo. Las hipótesis nulas que se prueban están, en términos de medias de celdas, descritas por el modelo completo:

$$\begin{aligned} H_0(\text{poblaciones}) : \mu_{11} + \mu_{12} + \mu_{13} &= \mu_{21} + \mu_{22} + \mu_{23} = \mu_{31} + \mu_{32} + \mu_{33} \\ &= \mu_{41} + \mu_{42} + \mu_{43} = \mu_{51} + \mu_{52} + \mu_{53} \end{aligned}$$

$$\begin{aligned} H_0(\text{número de crianza}) : \mu_{11} + \mu_{21} + \mu_{31} + \mu_{41} + \mu_{51} &= \mu_{12} + \mu_{22} + \mu_{32} + \mu_{42} + \mu_{52} \\ &= \mu_{13} + \mu_{23} + \mu_{33} + \mu_{43} + \mu_{53} \end{aligned}$$

Para estos datos, la interacción no es significante. Interprétese esto como algo que dice que los conjuntos de las medias de la línea 5 (POP) se comportan casi paralelamente para los tres tamaños de camadas, en las cuales se criaron las hembras (NREAR), o que las diferencias entre las medias de líneas no son dependientes del número de crianza.

La tabla 18.6 es otra parte de la salida impresa de SAS. Da las medias por mínimos cuadrados para las líneas y los tamaños de camadas en las cuales se criaron las hembras. Estas se calculan a partir de estimaciones por mínimos cuadrados de los parámetros así ( $\hat{\mu} + \hat{\alpha}_i$ ) y ( $\hat{\mu} + \hat{\beta}_j$ ). Se dan los errores estándar. Se prueban las hipótesis nulas,  $H_0: \mu = 0$ , para cada media dada. Esta parte del programa no tiene sentido aquí. Las medias dentro

**Tabla 18.6 Medias por mínimos cuadrados para efectos principales**

Procedimiento general para modelos lineales

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

POP	ERROR EST		PROB >  T  HO: MC MEDIA = 0	PROB >  T  HO: MC MEDIA(I) = MC MEDIA(J)	MC MEDIA(I) - MC MEDIA(J)				
	MC MEDIA	MC MEDIA			1	2	3	4	5
11	17.4205083	0.2569172	0.0001	1 .	0.0001	0.0001	0.0001	0.0001	0.0001
12	14.0917644	0.2551759	0.0001	2 0.0001	. .	0.0001	0.0001	0.5922	
14	10.7780733	0.2578203	0.0001	3 0.0001	0.0001	. .	0.0009	0.0001	
15	11.9866667	0.2525391	0.0001	4 0.0001	0.0001	0.0009	. .	0.0001	
16	14.2873737	0.2609795	0.0001	5 0.0001	0.5922	0.0001	0.0001	. .	

NOTA: PARA ASEGURAR EL NIVEL DE PROTECCION TOTAL, SOLO USAR PROBABILIDADES ASOCIADAS CON COMPARACIONES PRE-PLANEADAS

NREAR	ERROR EST		PROB >  T  HO: MC MEDIA = 0	PROB >  T  HO: MC MEDIA(I) = MC MEDIA(J)	MC MEDIA(I) - MC MEDIA(J)		
	MC MEDIA	MC MEDIA			1	2	3
8	14.0390769	0.1971210	0.0001	1 .	0.5501	0.0039	
12	13.8711187	0.2001340	0.0001	2 0.5501	. .	0.0232	
16	13.2284362	0.1992537	0.0001	3 0.0039	0.0232	. .	

NOTA: PARA ASEGURAR EL NIVEL DE PROTECCION TOTAL, SOLO USAR PROBABILIDADES ASOCIADAS CON COMPARACIONES PRE-PLANEADAS

**Tabla 18.7 Análisis de la varianza para los datos de los ratones**  
 Procedimiento general para modelos lineales

$$\text{Modelo: } Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

VARIABLE DEPENDIENTE: Y		FUENTE	GL	SUMA DE CUADRADOS	CUADRADO MÉDIO	VALOR F	PR > F	R-CUADRADO	C.V.
MODELO		6	3784.6684969	630.77810828	67.03	0.0001	0.361625	22.3862	
ERROR		710	6681.05519968	9.409931690			DESV EST	MEDIA	
TOTAL CORREGIDO		716	10465.72384937				3.06756205	13.70292887	
FUENTE		GL	TIPO I SC	VALOR F	PR > F	GL	TIPO III SC	VALOR F	PR > F
POP		4	3698.99948668	98.27	0.0001	4	3693.04715859	98.12	0.0001
NREAR		2	85.66916501	4.55	0.0109	2	85.66916501	4.55	0.0109

#### MEDIAS CUADRADAS MINIMAS

POP	MC MEDIA	STD ERR	PROB >  T	PROB >  T  HO: MC MEDIA = LSMFAN(j)				
				HO: MC MEDIA = 0	T	1	2	3
11	17.4206343	0.2565249	0.0001	1		0.0001	0.0001	0.0001
12	14.0899613	0.2534750	0.0001	2	0.0001		0.0001	0.0001
14	10.7780127	0.2574266	0.0001	3	0.0001	0.0001		0.6075
15	11.9897149	0.2521544	0.0001	4	0.0001	0.0001	0.0008	0.0001
16	14.2770880	0.2602676	0.0001	5	0.0001	0.0001	0.0001	0.0001

NOTA: PARA ASSEGURAREL NIVEL DE PROTECCION TOTAL, SOLO USAR PROBABILIDADES ASOCIADAS CON COMPARACIONES PRE-PLANEADAS

NREAR	MC MEDIA	MC MEDIA HO: MC MEDIA = 0	PROB >  T	PROB >  T  HO: MC MEDIA (1) = MC MEDIA (j)		
				1	2	3
8	14.0381786	0.1967991	0.0001	1		0.5295
12	13.8617923	0.1997314	0.0001	2	0.5295	
16	13.2332758	0.1988791	0.0001	3	0.0041	0.0260

NOTA: PARA ASSEGURAREL NIVEL DE PROTECCION TOTAL, SOLO USAR PROBABILIDADES ASOCIADAS CON COMPARACIONES PRE-PLANEADAS

de cada conjunto se comparan por pares. Si estas fuesen comparaciones razonables planeadas, deberán preferirse. Obsérvese la advertencia contra la aceptación ciega de estos resultados. También se dan y prueban las quince medias por mínimos cuadrados. Estas son medias de celdas cuando se incluye interacción en el modelo, y ya aparecen en la tabla 18.4.

Como la interacción es aproximadamente del mismo orden de magnitud del error, no hay necesidad de proponer un modelo con interacción. Este es el caso para el cual se propuso originalmente el método de ajuste de constantes o por mínimos cuadrados. Supóngase que proponemos un modelo semejante.

Para la ilustración, en la matriz  $\mathbf{X}$  de la ec. (18.3) ya no se necesitan las cuatro últimas columnas. La matriz  $5 \times 5$  de la esquina izquierda de  $\mathbf{X}'\mathbf{X}$  en la ec. (18.4) se convierte en la matriz  $\mathbf{X}'\mathbf{X}$  que se necesita. No se necesitan los  $Y_{ij}$  en  $\mathbf{X}'\mathbf{Y}$ . Hay 4 ecuaciones normales menos, y lo que al principio se llamó interacción, ahora es parte del error. Sin embargo, la nueva matriz  $\mathbf{X}'\mathbf{X}$  es aun singular.

Para los datos de los ratones,  $\mathbf{X}$  sólo tiene  $1 + 5 + 3 = 9$  columnas. Las últimas  $5 \times 3 = 15$  columnas y 15 filas de la  $\mathbf{X}'\mathbf{X}$  original ya no se necesitan más, y la matriz  $9 \times 9$  de la esquina superior izquierda es la nueva matriz  $\mathbf{X}'\mathbf{X}$ . La tabla 18.7 da el nuevo análisis.

Las sumas de cuadrados tipo I son idénticas a las de la tabla 18.5, excepto que no hay término de interacción. Se sigue la misma secuencia en el ajuste de parámetros en tanto sean exigidos por el modelo. Las sumas de cuadrados tipo III son diferentes ya que se ajustan menos parámetros. De nuevo, ambos efectos principales resultan significantes; se descartan ambas hipótesis nulas.

Ya que no hay interacción en el modelo, el único interés está en los efectos principales. Las medias apropiadas son las de las líneas y los números de crianza. Las medias por mínimos cuadrados se incluyen en la tabla 18.7. Estas no son las mismas de la tabla 18.6, ya que se ha resuelto un conjunto distinto de ecuaciones. Como la interacción era lo suficientemente pequeña como para dejarla por fuera del modelo y la desproporción no es grande, las diferencias son pequeñas. Obsérvese, además, la advertencia sobre el uso de las pruebas de programas.

**Ejercicio 18.4.1** En la sec. 2.15, se hizo una recomendación acerca de los tamaños relativos de la desviación estándar y de la unidad de la medida.

¿Cuál fue esa recomendación y cómo se aplica aquí?

**Ejercicio 18.4.2** ¿Cuál pudo haber sido una transformación adecuada de los datos antes del análisis de la varianza? ¿Por qué se escogió esa transformación?

**Ejercicio 18.4.3** Sugírase otro enfoque para analizar los datos originales de los ratones de modo que se puedan aplicar los caps. 9 y 15. (*Sugerencia:* la desproporción no es grande).

**Ejercicio 18.4.4** Los datos de Rawes (15.21) dados en el ejercicio 15.3.3 están incompletos. Había más observaciones sobre el mismo tratamiento y dos tratamientos más. Los últimos eran:

3. Cinco lecciones sobre técnicas de audición eficientes antes de adiestramiento en prácticas de audición y
4. Un grupo control.

Los datos adicionales son como siguen:

Datos antes de la prueba

	Tratamiento 1	Tratamiento 2
Braille	87	
Imprenta	98, 64	87, 89, 82, 91

Datos después de la prueba

Braille	92	
Imprenta	93, 80	87, 88, 90, 86

Datos antes de la prueba

	Tratamiento 3	Tratamiento 4
Braille	97, 86, 73	92, 91
Imprenta	94, 77, 97, 90, 64, 86, 83, 85	80, 66, 80, 86, 95, 75, 96, 86, 91

Datos después de la prueba

Braille	96, 79, 75	93, 97
Imprenta	90, 68, 98, 95, 73, 78, 78, 92	73, 48, 86, 78, 93, 85, 95, 88, 95

Utilizando todos los datos después de la prueba, obtener un análisis en que braille e imprenta sea las categorías de una dirección de clasificación, y los tratamientos, los de la otra.

Ejercicio 18.4.5 Analizar los datos de después de la prueba, pero utilizando los datos de antes de la prueba como valores de una covariante.

## 18.5 Otras técnicas analíticas

Yates (18.5) propuso análisis basados en medias de celdas. Estos son el *método de medias no ponderadas* y el *método de los cuadrados de medias ponderadas*. El primer método es aproximado en el sentido de que se procede como si las medias tuvieran varianzas iguales. Como  $\sigma_y^2 = \sigma^2/n_{ij}$ , las varianzas son desiguales cuando  $n_{ij}$  no es constante, que es la base de este capítulo.

El método de los cuadrados de medias ponderados utiliza funciones lineales de medias y, en el análisis, se ponderan éstas de acuerdo con la varianza de la función lineal. Para este método, las hipótesis probadas son las que se han dado en la sec. 18.4 para el método de mínimos cuadrados.

Estos métodos y otros se consideran en Gosslee y Lucas (18.2). Se refieren a efectos sobre niveles de significancia tabulados y al cálculo de la potencia.

Searle (18.3, cap. 8) también expone estos métodos y las hipótesis que se prueban.

Steel y Torrie (18.4, cap. 13) ilustran métodos para tablas de  $r \times 2$  y  $2 \times 2$ . Alguna vez, estos métodos fueron populares, pero los computadores modernos los han dejado casi anticuados.

**Referencias**

- 18.1. Addelman, S.: "The generalized randomized block design," *Amer. Statist.*, 23(4): 35-36 (1969).
- 18.2. Gosslee, D. G., y H. L. Lucas: "Analysis of variance of disproportionate data when interaction is present," *Biom.*, 21:115-133 (1965).
- 18.3. Searle, S. R.: *Linear Models*, 1a. ed., Wiley, Nueva York, 1971.
- 18.4. Steel, R. G. D., y J. H. Torrie: *Principles and Procedures of Statistics*, 1a. ed., McGraw-Hill, Nueva York, 1960.
- 18.5. Yates, F.: "The analysis of multiple classifications with unequal numbers in the different subclases," *J. Amer. Statist. Ass.*, 29:51-66 (1934).

**AJUSTE DE CURVAS****19.1 Introducción**

El tipo más común y sencillo de ajuste de curvas es de la línea recta. Sin embargo, cuando se representan pares de observaciones, éstas suelen quedar sobre una línea curva; las teorías biológicas y de algún otro tipo hasta pueden exigir una curva de forma especificada. Este capítulo considera tal regresión. Además, hay una breve exposición de la construcción y uso de polinomios ortogonales.

**19.2 Regresión no lineal**

Una relación entre dos variables puede ser aproximadamente lineal cuando se estudia en un intervalo limitado, pero puede ser marcadamente curvilínea si se amplía el intervalo. Por ejemplo, la relación entre madurez y rendimiento en arvejas para enlatar usualmente es lineal sobre el intervalo de madurez aceptable para la industria de enlatado. Pero, el aumentar la madurez, la tasa de aumento del rendimiento se disminuye, esto es, se vuelve curvilínea. Análogamente, la tasa de aumento en rendimiento tiende a mejorar en las etapas de inmadurez. Así, pues, para describir la relación en todo el intervalo es inadecuada una ecuación lineal.

Además de usar una curva apropiadamente descriptiva, es procedimiento acertado quitar del error toda componente que mida la regresión curvilínea. Así, si una observación se describe apropiadamente con  $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$ , y usamos  $Y = \beta_0 + \beta_1 X + \varepsilon$  como modelo, entonces se asigna a la medida del error una parte de la variación entre medias de población, o sea, la asociada con  $\beta_2 X^2$ . Es claro que ello exagera nuestra medida del error.

La selección de la forma de la ecuación de regresión que mejor expresa una relación curvilínea no siempre es problema simple. Prácticamente no hay límite en cuanto al número de tipos de curvas que pueden expresarse por ecuaciones matemáticas. Entre las

ecuaciones posibles, puede haber muchas igualmente buenas para minimizar la SC(residuos). Por tanto, al escoger la forma de la curva, es deseable tener alguna teoría dada por especialistas que trabajen en el campo de la materia del tema. Además, también puede ser bueno considerar la labor que entra en el ajuste de la regresión y si se cumplen los supuestos acostumbrados necesarios para la validez de la estimación y los procedimientos de prueba.

Tales consideraciones nos llevan a clasificar las relaciones curvilíneas en dos tipos: lineales y no lineales en los parámetros. Los modelos que son *lineales en los parámetros* son aquellos para los cuales se dispone de técnicas de regresión múltiple, entre ellos los modelos polinomiales. Los modelos que *no son lineales en los parámetros* son *intrínsecamente lineales* si los hace lineales una transformación. Ejemplos típicos de esta situación son las curvas logarítmica y exponencial. Modelos que no se pueden linealizar mediante una transformación son intrínsecamente no lineales y los análisis correspondientes se llaman *regresiones no lineales*. Este problema no se estudia aquí, pero en Draper y Smith (19.3, cap. 10) se da una introducción y Gallant (19.6) presenta un artículo expositivo sobre el tema.

Las transformaciones tienen por objeto proporcionar un procedimiento más fácil de ajuste y/o procedimientos válidos de estimación y prueba. Por ejemplo, podemos convenir en que la ecuación  $E(Y) = \beta_0 X^{\beta_1}$  se basa en un sólido razonamiento biológico. Entonces  $\log E(Y) = \log \beta_0 + \beta_1 \log X$  es una ecuación lineal si el par de observaciones se considera como  $(\log Y, \log X)$ . Los procedimientos de los caps. 10 y 17 son aplicables. Con datos como esos, no es infrecuente encontrar que los supuestos que se refieren a la normalidad casi sean más apropiados en escala transformada que en la original.

Ahora consideramos dos tipos generales de curvas: polinomiales y exponencial o logarítmica. He aquí algunos ejemplos, para los cuales se muestran en la fig. 19.1 las formas generales. Para las ecuaciones exponenciales,  $e$  puede reemplazarse por cualquiera otra constante sin que se afecte la forma de la curva que se ajusta. Para la ecuación  $E(Y) = \beta_0 X^{\beta_1}$  los valores enteros del exponente  $\beta_1$  dan casos especiales de polinomios. Sin embargo, es más probable que se use este tipo de curva cuando se desea un experimento fraccionado como ocurre a menudo en el campo de la economía.

Polinomial	Exponencial	Logarítmica
<b>Lineal</b> $E(Y) = \beta_0 + \beta_1 X$	$e^{E(Y)} = \beta_0 X^{\beta_1}$	$E(Y) = \beta_0' + \beta_1 \log X$
<b>Cuadrática</b> $E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2$	$E(Y) = \beta_0 \beta_1^X$	$\log E(Y) = \beta_0' + \beta_1' X$
<b>Cúbica</b> $E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$	$E(Y) = \beta_0 X^{\beta_1}$	$\log E(Y) = \beta_0' + \beta_1 \log X$

Los polinomios pueden tener picos y depresiones cuyo número a lo más es uno menos que el exponente más alto. Por ejemplo, la ilustración de la parte inferior izquierda de la fig. 19.1 tiene un pico y una depresión, o sea dos de tales puntos en una curva en que el exponente más elevado es 3. A los picos se les llama *máximos* y a las depresiones se les llama *mínimos*. Al ajustar curvas polinomiales, el investigador se interesa usualmente en un segmento dado del intervalo total representado por la ecuación.

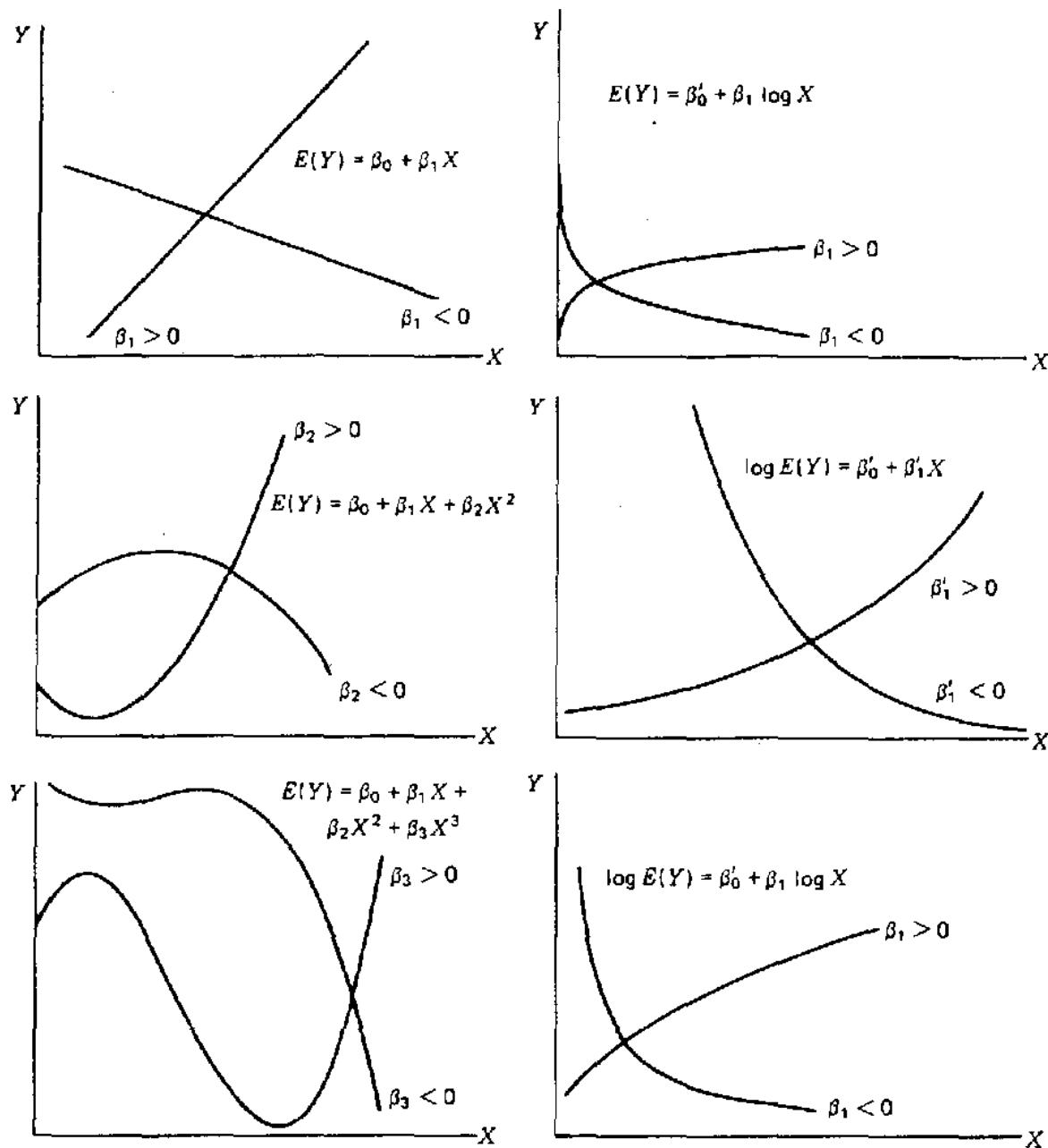


Figura 19.1 Tipos generales de curvas

Las curvas exponenciales o logarítmicas, excepto las de la forma  $\log E(Y) = \beta'_0 + \beta'_1 \log X$  se caracterizan por un aplanamiento hacia un extremo del intervalo. Por ejemplo, la curva  $E(Y) = \log X$  se aproxima más y más a  $X = 0$  a medida que  $Y$  toma valores negativos numéricamente más y más grandes; pero esta curva nunca cruza la recta vertical  $X = 0$ . Los números negativos no tienen logaritmos reales.

### 19.3 Curvas logarítmicas o exponenciales

Las curvas logarítmicas, o simplemente log, son lineales cuando se representan en papel logarítmico, apropiado. Refiriéndonos a la fig. 19.1 (lado derecho, de arriba a abajo) tenemos:

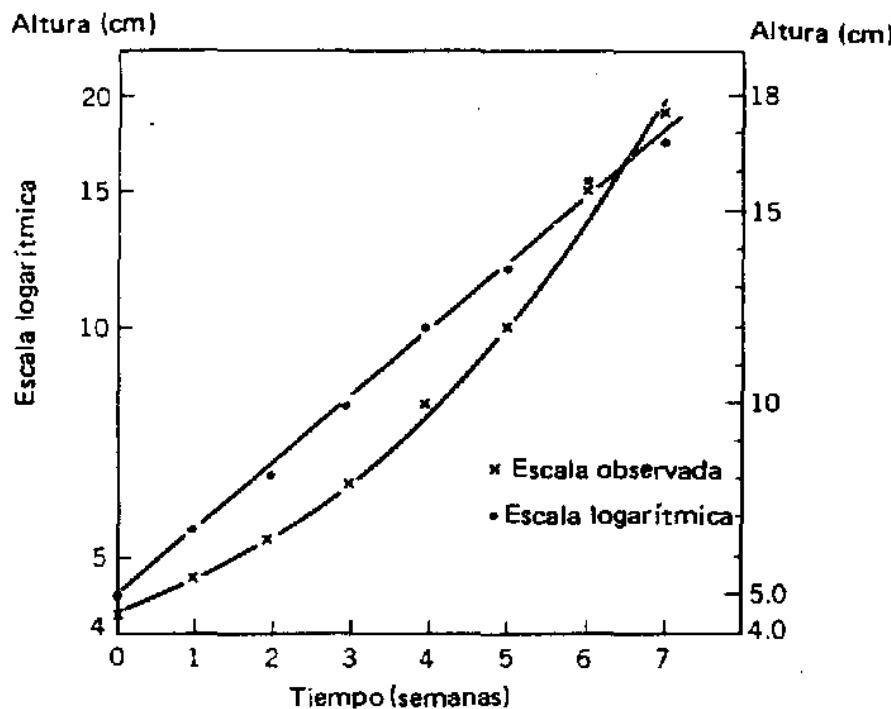


Figura 19.2 Puntos observados representados en escalas de intervalo fijo y logarítmica.

1.  $e^Y = \beta_0 X^{\beta_1}$  o  $Y = \beta'_0 + \beta'_1 \log X$ . Representando en papel semilogarítmico, los puntos  $(X, Y)$  dan lugar a una recta, en la que  $Y$  se representa en la escala de intervalos iguales, y  $X$ , en la escala logarítmica. En esencia, el papel semilogarítmico encuentra y representa los logs de  $X$ .
2.  $Y = \beta_0 \beta'_1 X$  o  $\log Y = \beta'_0 + \beta'_1 X$ . Los puntos  $(X, Y)$  dan lugar a una recta al representarlos en papel semilogarítmico, en la que  $Y$  va en escala logarítmica y  $X$  en la escala de intervalos iguales (ver también fig. 19.2).
3.  $Y = \beta_0 X^{\beta_1}$  o  $\log Y = \beta'_0 + \beta'_1 \log X$ . Esta se comporta como una recta cuando se representa en doblemente logarítmico. En el que ambas escalas son logarítmicas. (Ver también fig. 19.3).

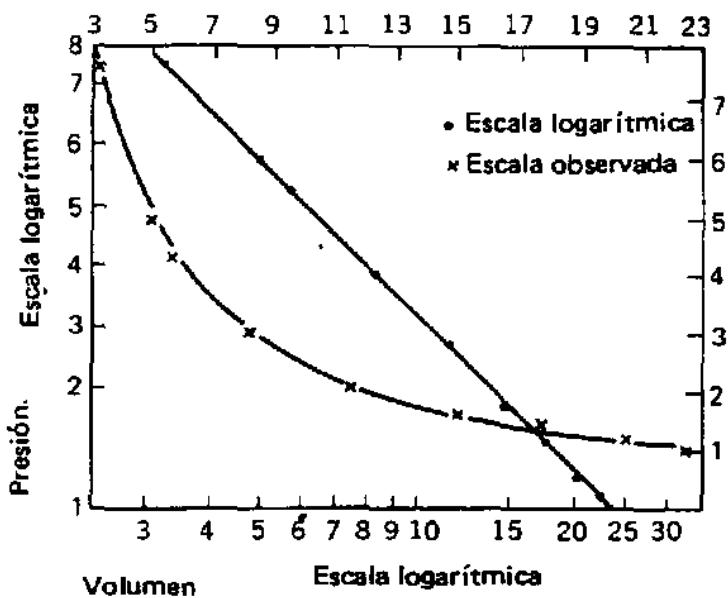


Figura 19.3 Puntos observados representados en escalas de intervalo fijo y logarítmica.

Para determinar si una curva logarítmica puede describir datos, suele ser suficiente con representar los datos en papel logarítmico. Una vez tomada la decisión respecto al tipo de curva logarítmica, se transforman los valores observados de  $X$  o  $Y$  o de ambos a logaritmos antes de realizar los cálculos. Los datos transformados se tratan luego por los métodos de los caps. 10 y 17. Los supuestos usuales se aplican a los datos transformados en lugar de a los originales.

Para ilustrar lo anterior, W. J. Drapala, Escuela Superior del Estado de Misisipi, ha proporcionado dos conjuntos de datos. En la tabla 19.1 se presentan los datos de la altura en centímetros por encima de los cotiledones, tomadas semanalmente, de repollos Golden Acre. Los datos aparecen representados en la fig. 19.2, con  $X$  en la escala de intervalo fijo y  $Y$  en la escala logarítmica, y también en la escala de intervalo fijo. Obsérvese como el conjunto de puntos toma la forma lineal cuando se representan en papel semilogarítmico.

Ahora procedemos con los cálculos para los datos transformados, como se hizo en los caps. 10 y 13. O bien un programa de computador provee información, tal como en la tabla 19.2.

La ecuación de regresión es

$$\widehat{\log Y} = 0.6497 + 0.0866X$$

La media de los valores de  $\log Y$  es 0.9528.

Los datos de la tabla 19.3 son presiones, en atmósferas de gas oxígeno a 25°C al ocupar varios volúmenes, medidos en litros. Los datos se comportan en forma curvilínea cuando se representan en una escala de intervalos iguales, y en forma casi lineal en la escala logarítmica doble (ver fig. 19.3). Se ajusta la relación  $\log Y = b_0 + b_1 \log X$ . También son apropiados los procedimientos de los caps. 10 y 13. La tabla 19.4 es parte de una salida impresa del computador del análisis de datos. La ecuación de regresión es

$$\widehat{\log Y} = 1.3790 - 1.0022 \log X$$

**Tabla 19.1 Altura por encima de los cotiledones de repollo Golden Acre media a intervalos semanales**

Semanas después de la primera observación $X$	Altura cm $Y$	Logaritmo decimal de altura $Y$
0	4.5	0.653
1	5.5	0.740
2	6.5	0.813
3	8.0	0.903
4	10.0	1.000
5	12.0	1.079
6	15.5	1.190
7	17.5	1.243

**Tabla 19.2** Análisis de los datos de cotiledones mediante SAS  
PROCEDIMIENTO GENERAL DE MODELOS LINEALES

VARIABLE DEPENDIENTE: LOG Y	GL	SUMA DE CUADRADOS	CUADRADO MEDIO	VALOR F	PR > F	R. CUADRADO	C.V.
FUENTE							
MODELO	1	0.31497610	0.31497610	2347.62	0.0001	0.997451	1.2157
ERROR	6	0.00080501	0.00013417		DESV EST		LOG Y MEDIA
TOTAL CORREGIDO	7	0.31578111			0.01158310		0.95276619
FUENTE	GL	TIPO I SC	VALOR F	PR > F	GL	TIPO IV SC	VALOR F
X	1	0.31497610	2347.62	0.0001	1	0.31497610	2347.62
PARAMETRO	ESTIMACION	T PARA HO:	PR >  T		ERRORESTANDAR		
INTERCEPTO	0	PARAMETRO = 0			DE LA ESTIMACION		
X	0.64966880	86.89	0.0001			0.00747686	
	0.08659926	48.45	0.0001			0.00178731	

**Tabla 19.3 Presión de oxígeno a 25°C al ocupar varios volúmenes**

Volumen, litros, <i>X</i>	Presión, atm, <i>Y</i>	Logaritmo decimal	
		Volumen <i>X</i>	Presión <i>Y</i>
3.25	7.34	0.512	0.866
5.00	4.77	0.699	0.679
5.71	4.18	0.757	0.621
8.27	2.88	0.918	0.459
11.50	2.07	1.061	0.316
14.95	1.59	1.175	0.201
17.49	1.36	1.243	0.134
20.35	1.17	1.309	0.068
22.40	1.06	1.350	0.025

Obsérvese que el valor de *F* deberá comenzar con 4. Pero el número es demasiado grande para el espacio disponible, así, todos los 9 se imprimen por convención de programación.

**Ejercicio 19.3.1** Brockington et al. (19.2) recolectaron los datos que se presentan en la tabla adjunta, sobre la relación entre contenido de humedad y la humedad relativa intersticial de semillas de maíz entero.

Muestra No.	Contenido de humedad		Humedad relativa equilibrio, porcentaje
	Brown-Duval	Horno de dos etapas	
1	7.0	9.4	40.0
2	7.5	9.9	40.0
3	11.6	12.9	59.0
4	11.8	12.6	63.5
5	12.9	14.1	71.5
6	13.2	14.7	71.0
7	14.0	15.2	76.5
8	14.2	14.6	75.5
9	14.6	15.2	79.0
10	14.8	15.8	79.0
11	15.7	15.8	82.0
12	17.3	17.2	85.5
13	17.4	17.0	85.0
14	17.8	18.2	87.5
15	18.0	18.5	86.5
16	18.8	18.2	88.0
17	18.9	19.4	90.0
18	20.0	20.3	90.5
19	20.7	19.9	88.5
20	22.4	19.5	89.5
21	22.5	19.8	91.0
22	26.8	22.6	92.0

**Tabla 19.4 Análisis de los datos de presión mediante el SAS  
PROCEDIMIENTO GENERAL DE MODELOS LINEALES**

VARIABLE DEPENDIENTE: LOG Y	GL	SUMA DE CUADRADOS	CUADRADO MEDIO	VALOR F	PR > F	R-CUADRADO	C.V.
FUENTE	1	0.70897285	0.70897285	99999.99	0.999998	0.0000	0.1054
MODELO						DESV EST	LOG Y MEDIA
ERROR	7	0.00000109	0.00000016			0.00039463	0.3745348
TOTAL CORREGIDO	8	0.70897394					
FUENTE	GL	TIPO SC	VALOR F	PR > F	GL	TIPO IV SC	VALOR F
LOG X	1	0.70897285	99999.99	0.0000	1	0.70897285	99999.99
							PR > F 0.0000
PARAMETRO	ESTIMACION	T PARA HO:	PR >  T		ERROR ESTANDAR		
INTERCEPTO	1.37899089	2820.73	0.0001		DE LA ESTIMACION		
LOG X	-1.00219470	-2133.68	0.0001			0.00048888	
						0.00046970	

Representar  $Y$  = humedad relativa de equilibrio respecto de  $X$  = contenido de humedad (cualquiera de las dos medidas) con  $X$  en escala ordinaria y en escala logarítmica. ¿Cuál parece ser la escala apropiada para obtener una recta? Calcular la regresión lineal de la humedad relativa de equilibrio respecto al log del contenido de humedad. ¿Qué porcentaje de la SC(total) queda aplicado por la regresión lineal?

**Ejercicio 19.3.2** En un estudio sobre el tamaño y forma óptimos de parcela, Weber y Horner (19.9) consideraron la longitud de la parcela y la varianza de las medias de parcelas por unidad básica de rendimiento en gramos y el porcentaje de proteína (entre otras formas y características de parcelas). Obtuvieron los datos que se indican en la tabla adjunta.

Varianza			
Forma	Número de unidades	Rendimiento, g	Proteína, porcentaje
8 x 1	1	949	.116
16 x 1	2	669	.080
24 x 1	3	540	.053
32 x 1	4	477	.048

Representar los dos conjuntos de varianzas con respecto al número de unidades después de transformar las tres variables a escala logarítmica. ¿Los datos resultantes guardan alguna relación razonablemente lineal?

#### 19.4 El polinomio de segundo grado

En la sección 14.8 se trató en forma breve de los modelos polinomiales. Aquí hacemos una ilustración sobre el polinomio de grado 2.

Sea la ecuación (19.1) la descripción matemática de una observación. Es claro que la linealidad se aplica a los parámetros que deben estimarse y no a los observables.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \quad (19.1)$$

En esta ecuación,  $\beta_1$  y  $\beta_2$  son coeficientes de regresión parcial, tal como se expuso en el cap. 14, pero no pueden interpretarse sino en espacio que no sea de más de dos dimensiones.

Para estimaciones por mínimos cuadrados de los parámetros, ha de cumplirse la ec. (19.2).

$$\sum (Y - \hat{Y})^2 = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 X_i^2)^2 = \min \quad (19.2)$$

Las ecuaciones normales o pdr mínimos cuadrados son

$$X'X\hat{\beta} = X'Y \quad (19.3)$$

Tabla 19.5 Rendimiento  $Y$ , en libras por parcelas, y lectura del tenderómetro  $X$ , de arvejas Alaska, cultivadas en Madison, Wisconsin, 1953

$\bar{Y}$	Rendimiento, libras por parcela:	24.0 22.0 26.5 22.0 25.0 37.5 36.0 39.5 32.0 26.5 55.5 49.5 56.0 55.5	
$\bar{X}$	Lectura del tenderómetro:	76.2 76.8 77.3 79.2 80.0 87.8 93.2 93.5 94.3 96.8 97.5 99.5 104.2 106.3	
$\bar{Y}$	Rendimiento, libras por parcela:	58.0 61.5 69.0 71.5 73.0 76.5 78.5 74.0 71.5 77.0 85.5	
$\bar{X}$	Lectura del tenderómetro:	106.7 119.0 119.7 119.8 119.8 123.5 141.0 142.3 145.5 149.0 150.0	
$\sum X = 2,698.9$	$\sum X^2 = 305,148.35$	$\sum X^3 = 36,045,287.22$	$\sum X^4 = 4,429,289,685.23$
$\sum Y = 1,303.5$	$\sum Y^2 = 78,797.25$	$\sum XY = 152,129.55$	$\sum X^2Y = 18,424,791.12$

donde

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum X_i & \sum X_i^2 \\ \sum X_i & \sum X_i^2 & \sum X_i^3 \\ \sum X_i^2 & \sum X_i^3 & \sum X_i^4 \end{pmatrix} \quad \text{y} \quad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \\ \sum X_i^2 Y_i \end{pmatrix}$$

Como ilustración se usan los datos de la tabla 19.5. Las entradas numéricas de  $\mathbf{X}'\mathbf{X}$  y  $\mathbf{X}'\mathbf{Y}$  se dan allí. La tabla 19.6 proporciona un análisis de la varianza, pruebas de hipótesis y estimaciones de los parámetros necesarios. Aproximadamente el 91 por ciento ( $= 100R^2$ ) de la variación, medida por la suma de cuadrados total, puede explicarse por una ecuación cuadrática en  $X$ . La ecuación de regresión muestral es

$$\hat{Y} = -138.5068 + 2.7133X - 0.0084X^2$$

La figura 19.4 da tanto la ecuación lineal como la cuadrática. Los cálculos para la regresión lineal proceden de la salida impresa del computador, pero aquí no se presentan.

Ejercicio 19.4.1 A la tabla 10.5 agregar el par de observaciones (1,949, 1,976). Calcular la regresión cuadrática del número de caballos por año. ¿Sería deseable incluir  $X^2$  en la ecuación de regresión? ¿Por qué?

## 19.5 Polinomios ortogonales

Cuando los aumentos entre niveles sucesivos de  $X$  son iguales y los valores de  $Y$  tienen una varianza común, pueden usarse tablas de valores de polinomios ortogonales en los cálculos que llevan a las pruebas de hipótesis respecto a la bondad de ajuste de polinomios de diversos grados. En la sec. 15.7 se ilustran ambos procedimientos.

Los polinomios ortogonales pueden usarse en situaciones en que las  $X$  no están igualmente espaciadas o las observaciones tienen varianzas desiguales pero conocidas. Los poli-

**Tabla 19.6 Análisis de los datos de arveja Alaska mediante SAS  
PROCEDIMIENTO GENERAL DE MODELOS LINEALES**

VARIABLE DEPENDIENTE: Y	GL	SUMA DE CUADRADOS CUADRADO MEDIO	VALOR F	PR > F	R CUADRADO	C.V.
FUENTE						
MODELO	2	9888.85463514	4944.42731757	115.24	0.0001	0.912866
ERROR	22	943.90536486	42.90478931		DESV EST	12.5627
TOTAL CORREGIDO	24	10832.76000000		6.55017475	Y MEDIA	52.14000000
FUENTE						
X	1	9441.75392063	220.06	0.0001	1	917.28607392
X•X	1	447.10071451	10.42	0.0039	1	447.10071451
						21.38    0.0001
						10.42    0.0039
PARAMETRO	ESTIMACION	T PARA HO:	PR >  T	ERROR ESTANDAR		
INTERCEPTO		PARAMETRO = 0		DE LA ESTIMACION		
X	-138.50680942	-4.33	0.0003	31.98150177		
X•X	2.71528314	4.62	0.0001	0.586880743		
	-0.00837858	-3.23	0.0039	0.00259550		

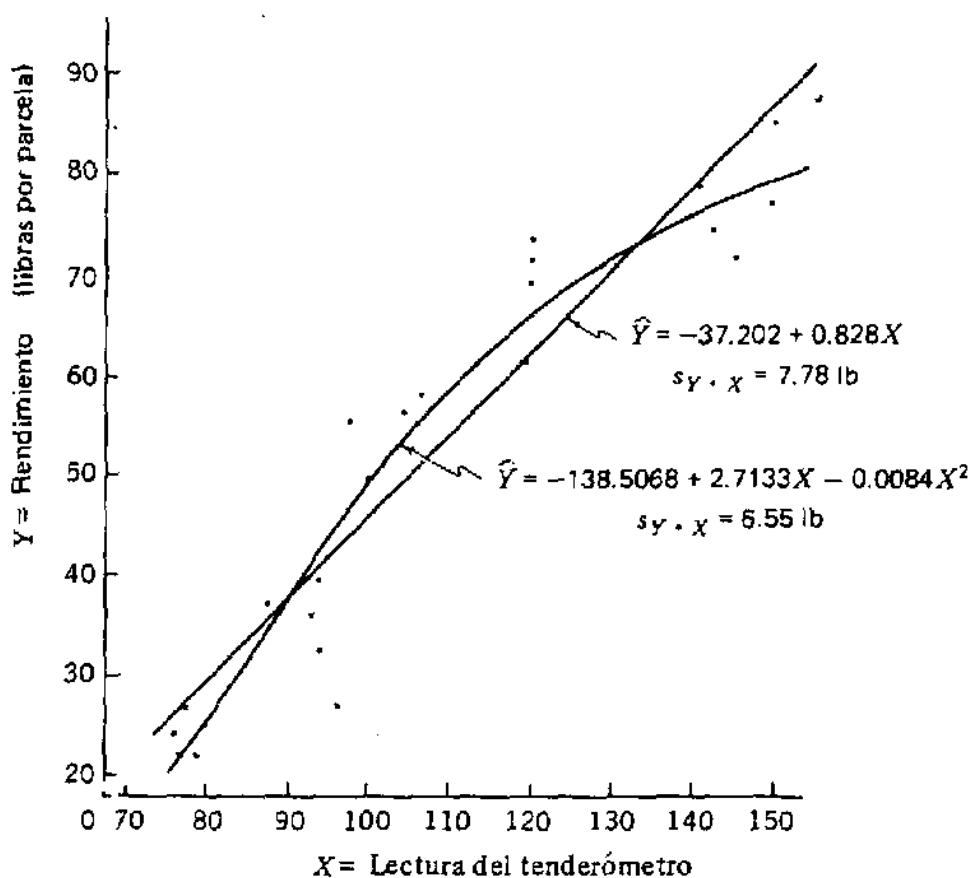


Figura 19.4 Relación entre rendimiento y lectura del tenderómetro para los datos de la tabla 19.5

nomios sucesivos son independientes unos de otros y nos permiten calcular otras sumas de cuadrados atribuibles a las varias potencias de  $X$ .

Para  $X$  desigualmente espaciadas o medias  $Y$  basadas en números desiguales, o ambas, Robson (19.8) usa lo que esencialmente es una razón de Fisher (19.4, 19.5) para obtener una fórmula recurrente para calcular coeficientes. (Una fórmula recurrente es la que se aplica una y otra vez, utilizando en cada nueva aplicación la aplicación anterior). Robson procede como sigue:

$$\hat{Y}_i = b_0 f_0(X_i) + b_1 f_1(X_i) + \cdots + b_r f_r(X_i) \quad (19.4)$$

$i = 1, \dots, n > r$ , donde  $\hat{Y}_i$  es la media de población estimada de las  $Y$  cuando  $X = X_i$ . El polinomio  $f_j(X_i)$  es de grado  $j$  y proporcionará los valores de los polinomios ortogonales, uno para cada  $X$ , necesarios para determinar los coeficientes de regresión. Los coeficientes son

$$b_j = \sum_i Y_i f_j(X_i) \quad (19.5)$$

Cada coeficiente de regresión representa una comparación, tal como se define en la sec. 8.3. Estas comparaciones son ortogonales y dan otras sumas de cuadrados atribuibles a la inclusión de  $X$  a la potencia  $j$ -ésima en la ecuación de regresión. Así,  $b_0$  es el grado

cero o efecto medio (no el principal),  $b_1$  el primer grado o efecto lineal, y así sucesivamente. Finalmente

$$\sum_i Y_i^2 = b_0^2 + \cdots + b_{n-1}^2$$

si efectuamos de  $r$  a  $n - 1$  de tal manera que todos los  $g_l = n$  hayan sido tenidos en cuenta individualmente.

La fórmula de recurrencia dada por Robson es la ec. (19.6) con  $c_h$  definido mediante la ec. (19.7).

$$f_h(X_i) = \frac{1}{c_h} \left[ X_i^h - \sum_{j=0}^{h-1} f_j(X_i) \sum_{g=1}^n X_g^h f_j(X_g) \right] \quad (19.6)$$

$$c_h^2 = \sum_i \left[ X_i^h - \sum_{j=0}^{h-1} f_j(X_i) \sum_{g=1}^n X_g^h f_j(X_g) \right]^2 \quad (19.7)$$

Obsérvese que  $c_h^2$  es la suma de cuadrados de cantidades, tales como las entre corchetes de la ec. (19.6).

Ahora vamos a ilustrar el uso de la fórmula de recurrencia. Supóngase que tenemos medias del porcentaje de digestión de celulosa a 6, 12, 18, 24, 36, 48 y 72 h. Por la aplicación de las ecs. (19.6) y (19.7) se tiene

$$f_0(X_i) = \frac{1}{c_0} (X_i^0) = \frac{1}{c_0}$$

$$c_0^2 = X_1^0 + \cdots + X_n^0 = n$$

(Una cantidad elevada a la potencia cero es igual a uno). Ahora bien,

$$f_0(X_1) = \frac{1}{\sqrt{n}} = f_0(X_2) = \cdots = f_0(X_n)$$

De la ecuación (19.5),

$$b_0 = \sum_i Y_i \frac{1}{\sqrt{n}}$$

y la reducción atribuible a este polinomio es

$$b_0^2 = \frac{(\sum Y)^2}{n}$$

Este es el término de corrección, tal como pudimos haberlo esperado. Los restantes  $n - 1$  coeficientes de regresión tienen que ver con la partición de los  $n - 1$  grados de libertad para medias de tratamientos.

En seguida

$$\begin{aligned}
 f_1(X_i) &= \frac{1}{c_1} \left[ X_i - f_0(X_i) \sum_{g=1}^n X_g f_0(X_g) \right] \\
 &= \frac{1}{c_1} \left( X_i - \frac{1}{\sqrt{n}} \sum_{g=1}^n X_g \frac{1}{\sqrt{n}} \right) \\
 &= \frac{1}{c_1} (X_i - \bar{X}) \\
 c_1^2 &= \sum (X_i - \bar{X})^2
 \end{aligned}$$

Ahora

$$\bar{X} = 216/7 = 30.86 \quad y \quad \sum (X_i - \bar{X})^2 = 3,198.86$$

$$f_1(X_1) = \frac{(6 - 30.86)}{\sqrt{3,198.86}}, \dots, f_1(X_n) = \frac{(72 - 30.86)}{\sqrt{3,198.86}}$$

Finalmente

$$b_1 = \sum Y_i \frac{(X_i - \bar{X})}{\sqrt{\sum (X_i - \bar{X})^2}}$$

y la reducción atribuible a regresión lineal es

$$b_1^2 = \frac{\sum Y_i (X_i - \bar{X})]^2}{\sum (X_i - \bar{X})^2}$$

También se esperaba este resultado. La cantidad entre corchetes del numerador usualmente se escribe en la forma  $\sum (Y_i - \bar{Y})(X_i - \bar{X})$ , pero las dos formas son equivalentes. Obsérvese que para  $r = 1$ , la ec. (19.4) se convierte en

$$\begin{aligned}
 \hat{Y} &= \frac{(\sum Y_i)}{\sqrt{n}} \frac{1}{\sqrt{n}} + \frac{\sum Y_i (X_i - \bar{X})}{\sqrt{\sum (X_i - \bar{X})^2}} \frac{(X - \bar{X})}{\sqrt{\sum (X_i - \bar{X})^2}} \\
 &= \bar{Y} + \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} (X - \bar{X})
 \end{aligned}$$

Para la ecuación de segundo grado,

$$f_2(X_i) = \frac{1}{c_2} \left[ X_i^2 - \frac{1}{\sqrt{n}} \left( \sum_{g=1}^n X_g^2 \frac{1}{\sqrt{n}} \right) \right]$$

$$\begin{aligned}
 & - \frac{X_i - \bar{X}}{\sqrt{\sum_i (X_i - \bar{X})^2}} \sum_g X_g^2 \frac{X_g - \bar{X}}{\sqrt{\sum_i (X_i - \bar{X})^2}} \Big] \\
 & = \frac{1}{c_2} \left[ X_i^2 - \frac{1}{n} \sum_g X_g^2 - \frac{(X_i - \bar{X}) \sum_g X_g^2 (X_g - \bar{X})}{\sum_i (X_i - \bar{X})^2} \right]
 \end{aligned}$$

De nuevo, debemos encontrar  $c_2^2$  mediante la ec. (19.7). Obsérvese que  $f_2(X_i)$  es un polinomio de grado dos; tiene un  $X_i^2$  un  $X_i$  como un  $X_i - \bar{X}$  multiplicado por una constante, y un término constante, que es,  $-\sum X_g^2/n$ . El hecho de que exista un  $X_i$  indica que la ecuación cuadrática completa tendrá un coeficiente de  $X$  que difiere del coeficiente de  $X$  en la ecuación lineal; el coeficiente de  $X_i$  en  $f_2(X_i)$  provee el ajuste.

Obsérvese que hemos calculado las funciones ortogonales para el caso general. Entonces, para nuestro ejemplo, hemos usado las  $X$  del experimento para obtener los coeficientes de las  $Y$  para la función lineal de las  $Y$  que da la reducción adicional atribuible a la más alta potencia de  $X$  introducida.

**Ejercicio 19.5.1** A. van Tienhoven, Universidad de Cornell, llevó a cabo un experimento factorial  $5 \times 5$  para estudiar los efectos de la hormona tiroide (HT) y una hormona estimulante de la tiroide (HET) sobre las alturas del epitelio folicular (entre otras respuestas) en pollos. La TH se mide en unidades  $\gamma$ , HET en unidades Junkmann-Schoeller, y la respuesta en unidades de micrómetro. En la tabla adjunta se presentan los totales de tratamiento. Cada total proviene de cinco observaciones.

HET					
	.00	.03	.09	.27	.81
HT	.00	3.42	6.21	11.21	14.40
	.04	5.64	5.85	9.16	18.30
	.16	5.13	8.39	12.74	15.20
	.64	5.37	5.24	9.14	17.66
	2.56	4.54	6.49	8.37	14.23
					16.90

*Fuente:* Datos no publicados y usados con permiso de A. van Tienhoven, Universidad de Cornell, Ithaca, Nueva York.

Se supone que la respuesta medida en alturas del epitelio sigue una curva logarítmica. Los primeros niveles de HT y HET no se encuentran en secuencias logarítmicas igualmente espaciadas con los otros niveles. Pero se quería obtener información para este tratamiento particular.

Un análisis preliminar es como sigue:

Omitir los tratamientos  $HET = 0.00$  y  $HT = 0.00$ . Representar la respuesta a HT, para cada nivel de HET, en papel logarítmico con tratamiento en la escala logarítmica. ¿En esta escala aparece lineal la respuesta? Repítase lo anterior para la respuesta a HET.

Emplear polinomios ortogonales para calcular sumas de cuadrados para las respuestas lineal, cuadrática y cúbica globales a la HET para los cuatro niveles diferentes de cero. Tomar la escala logarítmica para los tratamientos de modo que se pueda usar la tabla de coeficientes directamente. Hacer una prueba de significancia para cada una de las respuestas.

Considerar cómo se probaría la homogeneidad de estas varias respuestas para los niveles de HT. Probar la homogeneidad de las respuestas cúbicas, ya que miden desviaciones con respecto a las respuestas cuadráticas y pueden considerarse como candidatos para un término de error.

Calcular las sumas de cuadrados para las formas lineal, cuadrática y cúbica globales de las respuestas a la HT.

¿Cómo se encontrarían coeficientes para medir la HET(lineal) por HT(lineal), HET(lineal) por HT(cuadrática), y HET(cuadrática) por HT(lineal)? Hallar y probar estas componentes.

### Análisis preliminar

Fuente	gl	SC
Total	123	1,378.85
TSH	4	540.33
TH	4	48.09
TSH × TH	16	154.67
Error	99†	635.76

\* Faltó una observación.

Ejercicio 19.5.2 Usar la ec. (19.6) para calcular el polinomio ortogonal cúbico.

Ejercicio 19.5.3 A partir de los resultados del texto y del ejercicio 19.5.2, calcular una tabla de coeficientes de polinomios ortogonales para cuatro y cinco tratamientos, igualmente espaciados e igualmente replicados. Comentar.

### Referencias

- 19.1. Anderson, R. L. y E. E. Houseman: "Tables of orthogonal polynomial values extended to  $n = 104$ ," *Iowa Agr. Exp. Sta. Res. Bull.* 297, 1942.
- 19.2. Brockington, S. F., H. C. Dorin, y H. K. Howerton: *Hygroscopic equilibria of whole kernel corn*, "Cereal Chem.", 26:166-173 (1949).
- 19.3. Draper, N. R., y H. Smith, Jr.: *Applied Regression Analysis*, 1a. ed., Wiley, Nueva York, 1966.
- 19.4. Fisher, R.A.: "The influence of rainfall on the yield of wheat at Rothamsted," *Phil. Trans. Roy. Soc.*, B, 213:89-142 (1925).
- 19.5. Fisher, R. A.: *Statistical Methods for Research Workers*, 11a. ed., rev., Hafner, Nueva York, 1950.
- 19.6. Gallant, A. R.: "Nonlinear regression," *Amer. Statist.*, 29:73-81 (1975).
- 19.7. Pearson, E. S., y H. O. Hartley (eds): *Biometrika Tables for Statisticians*, vol 1, Cambridge, Nueva York, 1954.
- 19.8. Robson, D. S.: "A simple method for constructing orthogonal polynomials when the independent variable is unequally spaced," *Biom.*, 15:187-191 (1959).
- 19.9. Weber, C. R., y T. W. Horner: "Estimation of cost and optimum plot size and shape for measuring yield and chemical characters in soybeans," *Agron. J.*, 49:444-449 (1957).

---

**CAPITULO  
VEINTE**

---

## **ALGUNOS USOS DE JI-CUADRADO**

### **20.1 Introducción**

El criterio de prueba ji cuadrado generalmente se asocia más con datos enumerativos. Sin embargo, la distribución ji cuadrado, es una distribución continua basada en la distribución normal. A esta altura, introduciremos un capítulo sobre la distribución ji cuadrado para destacar, en cierto grado, su verdadera naturaleza asociándola con datos provenientes de distribuciones continuas en varias situaciones útiles. En los dos capítulos subsiguientes pasamos entonces a ilustrar su uso con datos enumerativos.

La ji cuadrado se definió en la sec. 3.9 como una suma de cuadrados de variables independientes normalmente distribuidas, con media cero y varianza uno, como lo ilustran las ecs. (3.13) y (3.14). En este capítulo, mostraremos cómo calcular un intervalo de confianza para  $\sigma^2$  usando la distribución  $\chi^2$ . Este es un procedimiento exacto. El criterio de prueba ji cuadrado también se utiliza cuando es una aproximación obvia; por ejemplo, en la sec. 11.5, al probar la homogeneidad de las varianzas y de la bondad de ajuste de datos continuos observados a distribuciones teóricas.

### **20.2 Intervalo de confianza para $\sigma^2$**

Consideremos la expresión

$$P(\chi_0^2 \leq \chi^2 \leq \chi_1^2) = .95$$

Este es un enunciado acerca de la variable  $\chi^2$ . Una elección obvia para  $\chi_0^2$  y  $\chi_1^2$  serían los valores  $\chi_{0.975}^2$  y  $\chi_{0.025}^2$ , tales que

$$P(\chi^2 \leq \chi_{0.975}^2) = .025 \quad \text{y} \quad P(\chi^2 \geq \chi_{0.025}^2) = .025$$

Es cómodo utilizar estos dos valores cuando se calcula un intervalo de confianza para  $\sigma^2$ . Sin embargo, no dan el intervalo de confianza más corto posible.

El procedimiento acostumbrado para calcular un intervalo de confianza del 95 por ciento para  $\sigma^2$  parte de la combinación del par de enunciados probabilísticos anteriores. Así, obtenemos

$$P(\chi^2_{.975} \leq \chi^2 \leq \chi^2_{.025}) = .95$$

donde  $\chi^2 = (n - 1)s^2/\sigma^2$ , por definición. (Para un intervalo de confianza del 99 por ciento, usar  $\chi^2_{.995}$  y  $\chi^2_{.005}$  y así sucesivamente). Esta fórmula conduce a

$$P\left[\frac{(n - 1)s^2}{\chi^2_{.025}} \leq \sigma^2 \leq \frac{(n - 1)s^2}{\chi^2_{.975}}\right] = .95 \quad (20.1)$$

La expresión anterior es aún una ecuación acerca de la variable aleatoria  $\chi^2$  o respecto a  $s^2$ , aunque ahora aparezca como una ecuación para  $\sigma^2$ . A partir de una muestra particular calculamos  $(n - 1)s^2/\chi^2_{.025}$  y  $(n - 1)s^2/\chi^2_{.975}$ , como los extremos del intervalo de confianza del 95 por ciento para  $\sigma^2$ .

Por ejemplo, para la muestra 1 de la tabla 4.3, se tiene  $SC(n - 1)s^2 = 2,376.40$  para  $n = 10$  (sabemos que  $\sigma^2 = 144$ ). Para un intervalo de confianza de  $\sigma^2$ , calculamos

$$\frac{2,376.40}{\chi^2_{.025}} = \frac{2,376.40}{19.0} = 125.07$$

$$\text{y} \quad \frac{2,376.40}{\chi^2_{.975}} = \frac{2,376.40}{2.70} = 880.15$$

Decimos ahora que  $\sigma^2$  está entre 125.07 y 880.15, a menos que se tratara de una muestra no usual. En este caso, sabemos que  $\sigma^2$  cae en el intervalo de confianza porque el muestreo se ha hecho en una población conocida.

**Ejercicio 20.2.1** Calcular un intervalo de confianza de 90 por ciento para  $\sigma^2$ , para la segunda muestra de la tabla 4.3. Calcular un intervalo de confianza del 95 por ciento para la tercera muestra y un intervalo de confianza del 99 por ciento para la cuarta muestra.

**Ejercicio 20.2.2** Calcular intervalos de confianza del 95 por ciento para  $\sigma^2$ , para cada tipo de suelo muestreado en la tabla 5.6. ¿Se traslanan los intervalos? Comparar este resultado con el de la prueba F de  $H_0: \sigma_1^2 = \sigma_2^2$  frente a  $H_1: \sigma_1^2 \neq \sigma_2^2$ . Comentar.

**Ejercicio 20.2.3** Partir de la expresión  $P(\chi^2 \leq \chi^2_{.05}) = 0.95$  y demostrar que conduce a una cota inferior para el tamaño de  $\sigma^2$ .

**Ejercicio 20.2.4** Partir de la expresión  $P(\chi^2 \geq \chi^2_{.995}) = 0.95$  y demostrar que conduce a una cota superior para el tamaño de  $\sigma^2$ .

**Nota:** Los ejercicios 20.2.3 y 20.2.4 suministran técnicas para construir intervalos de confianza unilaterales.

**Ejercicio 20.2.5** Construir un intervalo de confianza unilateral del 90 por ciento para  $\sigma^2$  en la segunda muestra de la tabla 4.3. Dejarlo abierto a la derecha.

**Ejercicio 20.2.6** Construir un intervalo de confianza unilateral, del 95 por ciento de  $\sigma^2$ , para la tercera muestra de la tabla 4.3. Igualar a cero el extremo de la izquierda.

### 20.3 Homogeneidad de la varianza

En un estudio sobre la herencia del tamaño de la semilla en lino, Myers (20.8), obtuvo las varianzas entre pesos de grupos de 50 semillas provenientes de plantas individuales de padres y de la generación  $F_1$ . Los datos aparecen en la tabla 20.1.

Sería deseable probar la homogeneidad de tales varianzas por las mismas razones por las que se prueba homogeneidad de medias o para establecer si se cumple o no un supuesto de homogeneidad, requerido para un análisis válido de la varianza de los datos. Un estudio de tales supuestos se dio en la sec. 7.10. Si se descarta la hipótesis de homogeneidad de la varianza, entonces lo probable es que se haga un análisis de la varianza sobre datos transformados. Las transformaciones se vieron en la sec. 9.16.

El procedimiento de prueba, que es aproximado, se debe a Bartlett (20.1, 20.2) y es una modificación de la prueba de razón de verosimilitud de Neyman-Pearson. Se efectúa en la tabla 20.1. Si se usan logaritmos naturales (base  $e$ ), no se requiere el factor 2.3026; éste sólo se usa cuando se emplean logaritmos decimales.

Tabla 20.1 Prueba de la homogeneidad de  $k$  varianzas

Clase	g1	$\sum (Y - \bar{Y})^2$	$s^2$	$\log s^2$	$(n - 1) \log s^2$	$1/(n - 1)$
Redwing	81	4,744.17	58.57	1.76768	143.18208	.01235
Ottawa 770 B	44	3,380.96	76.84	1.88559	82.96596	.02273
$F_1$	13	1,035.71	79.67	1.90129	24.71677	.07692
TOTALES	138	9,160.84			250.86481	.11200
COMBINACIONES			66.38	1.82204	-251.44152	

$$\chi^2 = 2.3026 \left[ \sum (n_i - 1) \right] \log \bar{s}^2 - \sum (n_i - 1) \log s_i^2 \}$$

$$= 2.3026(251.44152 - 250.86481) = 1.3279 \quad \text{con } 2 \text{ g1}$$

$$\text{Factor de corrección} = 1 + \frac{1}{3(k-1)} \left[ \sum \frac{1}{n_i - 1} - \frac{1}{\sum (n_i - 1)} \right]$$

$$= 1 + \frac{1}{3(2)} \left[ .11200 - \frac{1}{138} \right] = 1.01746$$

$$\text{corregido } \chi^2 = \frac{1.3279}{1.01746} = 1.305 \quad \text{no es significante}$$

El factor de corrección deberá mejorar la aproximación a  $\chi^2$  cuando los tamaños de muestra son pequeños. Es siempre mayor que 1 y su uso disminuye el valor bruto de  $\chi^2$ . Así, generalmente calculamos el  $\chi^2$  corregido sólo cuando el  $\chi^2$  bruto es significante, pero está cercano al valor crítico.

Para usar apropiadamente esta prueba, suponemos que las distribuciones de base son normales. Cuando esto no ocurre, la prueba puede detectar la no normalidad en vez de la heterogeneidad en las varianzas. Esta prueba es más sensible a la no normalidad que el uso de  $F$  en el análisis de la varianza.

Los valores independientes de  $\chi^2$  son aditivos, o sea que la distribución de una suma semejante también se distribuye como  $\chi^2$  con un número de grados de libertad igual a la suma de los grados de libertad de los sumandos. Esto es válido sólo para los  $\chi^2$  sin corregir, pero no para los corregidos. Así que si se tuvieran varianzas de varios años, podría ser informativo hacer comparaciones entre y dentro de años. También si se dispone de varianzas de varias generaciones en segregación, podemos hacer comparaciones entre padres y  $F_1$  (como en el ejemplo), entre generaciones en segregación y entre los dos grupos. El  $\chi^2$  total puede usarse, al menos, como comprobación aritmética.

**Ejercicio 20.3.1** Las varianzas de los datos de la tabla 7.1 varían desde 1.28 hasta 33.64. Si bien este intervalo es bien amplio, cada una tiene sólo 4 grados de libertad. Probar la homogeneidad de estas varianzas. ¿Puede observarse cómo se reducirían los cálculos cuando todas las varianzas se basan en el mismo número de grados de libertad?

*Nota:* Pearson y Hartley (20.9) han dado tablas de valores críticos del criterio de prueba  $s_{\max}^2/s_{\min}^2$  y de la prueba de Cochran (20.3)  $s_{\max}^2/\sum s_i^2$ . También hay tablas para la primera prueba en Rohlf y Sokal (20.12) y para la segunda en Dixon y Massey (20.11). Estas pruebas exigen tamaños de muestra iguales. Como la prueba de Bartlett, también son sensibles a la normalidad.

**Ejercicio 20.3.2** Verificar la homogeneidad de las varianzas en el ejercicio 7.3.1.

**Ejercicio 20.3.3** Las varianzas en el ejercicio 7.4.1 se basan en números diferentes de grados de libertad. Verificar la homogeneidad.

## 20.4 Bondad de ajuste para distribuciones continuas

A menudo es deseable saber si un conjunto de datos se aproxima o no a una distribución dada, tal como la normal o la ji cuadrado. Como ejemplo, considérense los datos de la tabla 20.2 presentados antes en la tabla 2.5. Probemos la distribución observada con la distribución normal.

Para comparar una distribución observada con una distribución normal, se requieren las frecuencias esperadas por celda. El rendimiento de 3 g incluye rendimientos hasta 5.5 g, el de 8 g incluye rendimientos mayores de 5.5 y hasta 10.5 g, etc. Para calcular las frecuencias esperadas, es necesario conocer las probabilidades asociadas con cada intervalo. Estas se hallan en la tabla A.4. Deben estimarse valores de  $\mu$  y  $\sigma$  a partir de nuestros datos. Encontramos  $\bar{Y} = 31.93$  y  $s = 12.80$  y los consideramos como  $\mu$  y  $\sigma$ . Ahora bien

$$P(Y < 5.5) = P\left(Z < \frac{5.5 - 31.93}{12.80}\right)$$

Tabla 20.2 Valores de rendimiento observados y esperados, en gramos, de 229 plantas espaciadas de soya Richland

$$\begin{aligned}
 &= P(Z < -2.065) = .0194 \\
 P(5.5 < Y < 10.5) &= P\left(\frac{5.5 - 31.93}{12.80} < Z < \frac{10.5 - 31.93}{12.80}\right) \\
 &= P(-2.065 < Z < -1.674) = .0471 - .0194 \\
 &= .0277
 \end{aligned}$$

y así sucesivamente. Cada probabilidad multiplicada por la frecuencia total 229 da una frecuencia esperada. La probabilidad asociada con la última celda es la probabilidad de un valor mayor que 65.5 y no es la probabilidad de un valor mayor que 65.5, pero no mayor que 70.5. Los cálculos se muestran en la tabla 20.2. La suma de las probabilidades difiere de 1 tan sólo por errores de aproximación. La suma de las frecuencias esperadas diferirá de 229 sólo por errores de aproximación.

Calculemos el del criterio de prueba

$$\begin{aligned}
 \chi^2 &= \sum \frac{(\text{observado} - \text{esperado})^2}{\text{esperado}} \quad (20.2) \\
 &= 10.62 \quad \text{con } (14 - 1 - 2) = 11 \quad \text{gl no significante.}
 \end{aligned}$$

En general, el número de grados de libertad para  $\chi^2$  es el número de celdas menos el número de restricciones impuestas al muestreo y el número de parámetros independientes estimados. Aquí, tenemos 14 celdas restringidas a probar muestras de tamaño 229, y debemos estimar los 2 parámetros de la distribución normal, o sea,  $\mu$  y  $\sigma^2$ . El criterio de prueba se distribuye aproximadamente como  $\chi^2$ , siempre que las frecuencias esperadas no sean demasiado pequeñas.

Las frecuencias esperadas empiezan en 1.0. Este es un valor esperado pequeño. Algunos autores han sugerido de cinco a diez como mínimo y recomiendan agrupar las primeras dos celdas como una nueva primera celda, y las últimas tres o cuatro, como una nueva última celda. Esto mejoraría la aproximación a  $\chi^2$ . Sin embargo, puesto que las colas de una distribución ofrecen, a menudo, la mejor fuente de comprobación para distinguir entre distribuciones hipotéticas, el mejoramiento se da a expensas de la potencia de la prueba. Cochran (20.4 a 20.6) ha demostrado que hay poca perturbación a un nivel del 5 por ciento cuando un solo valor esperado es tan bajo como 0.5 y dos valores esperados pueden ser tan bajos como 1.0 para menos grados de libertad que los de nuestro ejemplo. El nivel del 1 por ciento muestra mayor perturbación que el nivel del 5 por ciento. Para la tabla 20.2, el agrupamiento parece innecesario. No hay evidencia que indique que la distribución normal no proporciona un ajuste adecuado.

**Ejercicio 20.4.1** Ajustar una distribución a los datos de grasa de leche de la tabla 4.1 suponiendo  $\mu = 40$  y  $\sigma = 12$  lb. Probar la bondad de ajuste.

**Ejercicio 20.4.2\*** Ajustar una distribución normal a las 500 medias de la tabla 4.4, suponiendo que no se conocen la media y la varianza de la población. Probar la bondad del ajuste.

## 20.5 Combinaciones de probabilidades de pruebas de significancia

Fisher (20.7) ha demostrado que  $-2 \log P$ , donde los logaritmos son en base  $e$  y  $P$  es la probabilidad de obtener un valor del criterio de prueba tan extremo o más que el obtenido en una prueba particular, se distribuye como  $\chi^2$  con dos grados de libertad. Tales valores pueden sumarse. En la práctica, es costumbre usar logaritmos en base 10 y multiplicar el resultado por 2.3026 para obtener logaritmos en base  $e$ . Es conveniente hacer la multiplicación después de hecha la suma en lugar de hacerla por valores separados.

Esta relación entre  $P$  y  $\chi^2$  puede usarse cuando se desea agrupar información disponible de datos no susceptibles de combinación. Tales datos deben proceder de ensayos independientes.

Por ejemplo, podemos tener información sobre la diferencia en respuesta a un control o tratamiento patrón y un tratamiento experimental particular provenientes de varios experimentos bastante diferentes. Estos tratamientos pueden ser los únicos comunes a los experimentos, que pueden incluir diferentes diseños experimentales. Cada experimento puede mostrar que la probabilidad de un valor mayor del criterio de prueba, no necesariamente el mismo para todas las pruebas, está entre 0.15 y 0.05 para la comparación particular. Las probabilidades se calculan sobre el supuesto de que la hipótesis nula es cierta. Ningún valor puede ser significante, aunque puede sugerir la posibilidad de una diferencia real. La prueba se aplicaría a las probabilidades combinadas, y la combinación se efectúa agregando los valores de  $-2 \log P$ .

El uso de este procedimiento de agrupación requiere que las tablas del criterio de prueba original sean razonablemente completas con respecto a los niveles de probabilidad; por tanto, debe usarse interpolación. En seguida se remite la suma de los valores  $-2 \log P$  a la tabla  $\chi^2$ , tabla A.5, donde se obtiene el valor de  $P$  para la información conjunta. Este conjunto  $\chi^2$  tiene  $2k$  grados de libertad, cuando se combinan  $k$  probabilidades.

Wallis (20.10) señala que esta prueba rara vez o nunca es ideal pero que es probable que sea muy satisfactorio en la práctica.

## Referencias

- 20.1. Bartlett, M. S.: "Properties of sufficiency and statistical tests," *Proc. Roy. Soc.*, A160:268-282 (1937).
- 20.2. Bartlett, M. S.: "Some examples of statistical methods of research in agriculture and applied biology," *J. Roy. Statist. Soc. Suppl.*, 4:137-183 (1937).
- 20.3. Cochran, W. G.: "The distribution of the largest of a set of estimated variances as a fraction of their total," *Ann. Eugen.*, 11:47-52 (1941).
- 20.4. Cochran, W. G.: "The  $\chi^2$  correction for continuity," *Iowa State Coll. J. Sci.*, 16:421-436 (1942).
- 20.5. Cochran, W. G.: "The  $\chi^2$  test of goodness of fit," *Ann. Math. Statist.*, 23:315-345 (1952).
- 20.6. Cochran, W. G.: "Some methods for strengthening the common  $\chi^2$  tests," *Biom.*, 10:417-451 (1954).
- 20.7. Fisher, R. A.: *Statistical Methods for Research Workers*, 11 ed., rev., Hafner, Nueva York, 1950.
- 20.8. Myers, W. M.: "A correlated study of the inheritance of seed size and botanical characters in the flax cross, Redwing x Ottawa 770B," *J. Amer. Soc. Agron.*, 28:623-625 (1936).
- 20.9. Pearson, E. S., y H. O. Hartley (eds.): *Biometrika Tables for Statisticians*, 3a. ed. Cambridge, Nueva York, 1966.

- 20.10. Wallis, W. A.: "Compounding probabilities from independent significance tests," *Econ.*, 10:229-248 (1942).
- 20.11. Dixon, W. J., y F. J. Massey, Jr.: *Introduction to Statistical Analysis*, 3a. ed., McGraw-Hill, Nueva York, 1969.
- 20.12. Rohlf, F. J., y R. R. Sokal: *Statistical Tables*. Freeman, San Francisco, 1969.

## DATOS ENUMERATIVOS I: CLASIFICACIONES DE UNA VÍA

### 21.1 Introducción

Este capítulo se refiere a *datos enumerativos* clasificados de acuerdo con un criterio único.

En los datos enumerativos entra en general una variable discreta, o sea, una característica cualitativa, en vez de una cuantitativa; por lo tanto, son números de individuos pertenecientes a clases bien definidas. Por ejemplo, se muestrea una población y se observa en la muestra el número de machos y hembras, o se cuenta el número de respuestas afirmativas, negativas o indecisas a una pregunta de un cuestionario, y así sucesivamente.

### 21.2 El criterio de prueba $\chi^2$

En la sección 3.9, ec. (3.13) se define el estadígrafo  $\chi^2$  cuadrado con  $n$  grados de libertad como la suma de los cuadrados de  $n$  variables independientes, distribuidas normalmente con medias cero y varianzas uno. Es decir

$$\chi^2 = \sum_i \frac{(Y_i - \mu_i)^2}{\sigma_i^2} \quad (21.1)$$

donde  $Y_i$  son independientes.

Excepto en el capítulo 20 la  $\chi^2$  se ha mencionado con poco frecuencia y sólo cuando intervenía en otras distribuciones; por ejemplo,  $F$  es una razón de dos  $\chi^2$  divididas por sus grados de libertad. En este capítulo los criterios de prueba llamados  $\chi^2$  se usan con frecuencia. Sin embargo, ahora tratamos con datos discretos, de modo que los criterios de prueba son cantidades  $\chi^2$  como se definen en la ec. (21.1), aun cuando la hipótesis nula sea cierta. Esto significa que el criterio de prueba puede distribuirse sólo aproximadamente como cantidades  $\chi^2$ .

Cuando está asociada con datos discretos, la distribución  $\chi^2$  suele estar en conjunción con una prueba de *bondad de ajuste*. El criterio de prueba es

$$\chi^2 = \sum \frac{(\text{observados} - \text{esperados})^2}{\text{esperados}} \quad (21.2)$$

La suma se toma sobre todas las celdas en el sistema de clasificación. *Observados* se refiere a los números observados en las celdas; *esperados* a los números promedio o valores esperados cuando la hipótesis es cierta, o sea, a los valores teóricos. La suma de dichas desviaciones, es decir, valores de  $(\text{observados} - \text{esperados})$ , será igual a cero dentro de los errores de aproximación. El número de grados de libertad que interviene se estudiará para varias situaciones a medida que se presenten.

### 21.3 Tablas de dos celdas, límites de confianza para una proporción o porcentaje

Muchas situaciones de muestreo dan margen para solo dos sucesos posibles, por ejemplo, los números de síes y noes en respuesta a una pregunta o los números de individuos que muestran la presencia o ausencia de una característica cualitativa. A menudo se requiere la estimación de una proporción poblacional o una prueba de hipótesis acerca de una proporción. Considérese el siguiente ejemplo. En el estudio de *Dentaria* en una zona cercana a Ithaca, Nueva York, una clase de taxonomía y ecología observó el número de plantas que florecen o no florecen. Esto se hizo con varias muestras. En una muestra hubo 42 plantas con flor y 337 sin flor. Establecemos un intervalo de confianza para la proporción de plantas con flor en la población.

Supóngase que el muestreo ha sido aleatorio y que procede de una población estable. Como sólo hay dos sucesos posibles asociados con un individuo, tenemos una *población binomial*, como se la llama según se vio en el cap. 3. El parámetro que va a estimarse por un intervalo de confianza es la proporción de plantas que florecen en la población de plantas o la probabilidad de que una planta tomada al azar sea planta que florece. El parámetro se denota generalmente mediante  $p$  y su estimación, la proporción observada, por  $\hat{p}$ . El valor esperado o media de todos los valores posibles  $\hat{p}$  es  $p$ , es decir  $\hat{p}$  es un estimador insesgado de  $p$ , la media de la población. La varianza del estadígrafo  $\hat{p}$  es  $p(1-p)/n$ , donde  $n$  es el número total de observaciones en la muestra; se estima por  $\hat{p}(1-\hat{p})/n$  cuando es necesario. Obsérvese que la varianza se calcula a partir de la media. Con frecuencia  $1-p$  y  $1-\hat{p}$  se escriben  $q$  y  $\hat{q}$ , respectivamente.

Ahora procedemos a usar la distribución normal como una aproximación a la binomial, sabiendo que un supuesto básico es falso.

*1. La aproximación normal* Como aproximación, se dice que  $\hat{p}$  está distribuido normalmente con media  $p$  y varianza  $p(1-p)/n$ , estimado por  $\hat{p}(1-\hat{p})/n$ . Por lo tanto, el intervalo de confianza del 95 por ciento para  $p$  es

**Tabla 21.1 Tamaño de muestra binomial para que sea aplicable la aproximación normal**

$\hat{p}$	$n\hat{p}$ = número observado en la clase más pequeña	$n$ = tamaño de muestra
0.5	15	30
0.4	20	50
0.3	24	80
0.2	40	200
0.1	60	600
0.05	70	1,400

Fuente: Reimpreso con permiso de W.G.Cochran,  
"Sampling Techniques", tabla 3.3, página 41, John Wiley  
& Sons, Nueva York, 1953.

$$\begin{aligned} \hat{p} \pm Z_{.05}(\text{normal}) \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \\ = \frac{42}{379} \pm 1.96 \sqrt{\frac{42}{379} \frac{337}{379} \frac{1}{379}} = .111 \pm .032 = (.079, .143) \end{aligned} \quad (21.3)$$

Esta aproximación no es la única posible basada en la distribución normal, pero es cómoda y ha de diferir poco de cualquiera otra basada en un tamaño de muestra razonable. La tabla 21.1 da los límites inferiores para tamaños de muestras, que se deben a Cochran (21.2). La tabla sugiere que nuestra aproximación puede ser insuficiente ya que deberíamos estar cerca de  $n = 600$ .

**2. La distribución exacta** Cuando se decide que una aproximación no es adecuada, se requiere tablas de la distribución binomial. Hay muchas de esas tablas, por ejemplo, *Tablas de Distribución Binomial de Probabilidades Acumuladas* por The Computation Laboratory, Harvard University (21.7), contienen sumas de las probabilidades de  $k$  términos para  $k = 1, 2, \dots, n + 1$  para  $n = 1, 2, \dots, 1,000$  observaciones en la muestra, y  $p = 0.01[0.01]0.50$ . Así mismo, The National Bureau of Standards Applied Mathematics Series 6(21.11) da probabilidades para términos individuales. Los intervalos de confianza calculados por estas tablas raramente serán simétricos con respecto a  $\hat{p}$ , ya que asocian como la mitad de la probabilidad de un error Tipo I con cada cola de la distribución.

Esencialmente elegimos dos valores de  $p$ , por ejemplo  $p_1$  y  $p_2$ , con la propiedad de que sean los valores mínimo y máximo de  $p$  que puedan ser hipotetizados y hallados aceptables con base en los datos observados. En la práctica, es imposible, por lo general, encontrar probabilidades exactamente iguales a la mitad de la probabilidad de un error Tipo I. En algunos casos puede ser necesario colocar la probabilidad completa en una cola, por ejemplo, si todas o casi todas las observaciones caen en una de las dos clases. En este caso un límite de confianza será 0.00 ó 1.00.

Las tablas A.14 dan intervalos de confianza del 95 y 99 por ciento para la distribución binomial que se han obtenido de tablas más extensas por Mainland et al. (21.3). Para ilustrar su uso, se lanzó una moneda 30 veces y se observaron 13 caras y 17 sellos. En las tablas, los intervalos de confianza del 95 y 99 por ciento para la probabilidad  $p$  de obtener cara en un solo lanzamiento están dados como  $0.2546 < p < 0.6256$  y  $0.2107 < p < 0.6772$ , respectivamente. Los límites de confianza para la probabilidad de una cola pueden encontrarse en la tabla o restando de 1.00 cada uno de los límites anteriores.

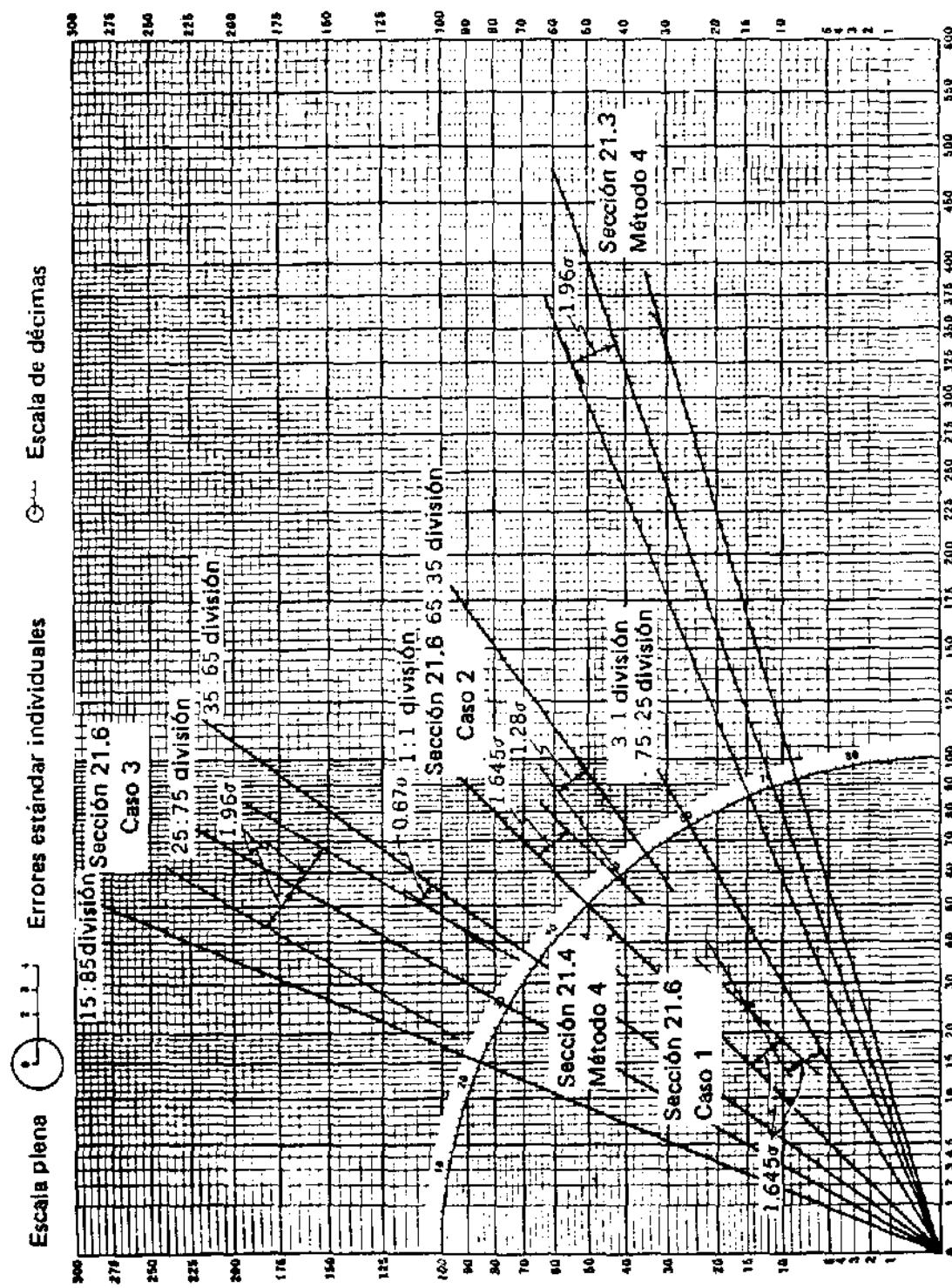
Si queremos usar la tabla A.14B para los datos de *Dentaria*, tenemos que interpolar entre  $n = 300$  y  $n = 500$ . Para  $\hat{p} = 0.11$ , el límite inferior sobre el intervalo de confianza del 95 por ciento está entre 0.0771 y 0.0841 y el límite superior entre 0.1508 y 0.1406. Una primera aproximación sitúa los puntos extremos a mitad de camino entre los pares de valores 0.0806 y 0.1457. El intervalo no difiere mucho de la aproximación normal, especialmente si consideramos que es probable que se deseen más de dos cifras decimales.

**3. Diagrama de Clopper y Pearson** Como alternativa para calcular un intervalo de confianza, se puede usar un diagrama llamado *Regiones de confianza para p*, como el que dan Clopper y Pearson (21.1), o un diagrama similar. La tabla A.15 es uno de estos diagramas. Para los datos de *Dentaria*, calculamos  $\hat{p} = 42/379 = 0.111$  y se localiza en la escala  $\hat{p} = Y/n$ ; verticalmente sobre este eje encontramos dos rectas marcadas para el tamaño de muestra cercano a 379 (400 es el valor más cercano); ahora nos movemos horizontalmente sobre la escala  $p$ , donde encontramos los valores para el intervalo de confianza. Para el intervalo de confianza del 95 por ciento, obtenemos (0.08, 0.15) sin intentar hacer ninguna interpolación real para el tamaño de muestra en 379.

Los diagramas como el de la tabla A.15 se calculan por un método aproximado de interpolación, que da un resultado más preciso que la aproximación normal. Los intervalos de confianza no serán simétricos con respecto a  $\hat{p}$ .

**4. Papel de probabilidad binomial** Otro procedimiento es el uso del papel de probabilidad binomial, diseñado por Mosteller y Tukey (21.4). Este papel y los resultados obtenidos con él se basan en la transformación de los datos a una escala sobre la cual los números resultantes tengan una distribución aproximadamente normal con varianza casi constante; lo que implica que la media y la varianza son aproximadamente independientes.

La tabla A.16 es un papel binomial (ver también fig. 21.1, donde se opera con un ejemplo). Para los datos de *Dentaria* representese el punto (42, 337) o (337, 42). Debido a las dimensiones del papel, se representa el último. Una recta trazada por este punto hasta el origen corta el cuarto círculo en el punto cuyas coordenadas son los dos porcentajes del problema. Leemos como proporciones 0.89 y 0.11. En la obtención del intervalo de confianza interviene un triángulo con coordenadas (37, 43), (337, 42) y (338, 42), cuyo lado mayor o hipotenusa está más lejos del origen. Obsérvese que un número de los pares observados se incrementa en uno para dar la primera coordenada, luego el otro para la tercera, solamente se incrementa un número para dar cada una de las nuevas coordenadas. Nuestro triángulo no se puede distinguir de un punto. Alrededor de este punto trazamos un círculo de radio 1.96 desviaciones estándar a plena escala. Estas están dadas en el papel. Las tangentes a este círculo trazadas a partir del origen corta en el cuarto de círculo en puntos cuyas coordenadas son los límites de confianza. Leemos 0.08 y 0.14. Si



**Figura 21.1** Algunos ejemplos del uso del papel de probabilidad binomial

se deseara un intervalo de confianza para la proporción de plantas que no florecen, deberíamos leer las dos abscisas. Si el punto se hubiera podido distinguir como triángulo, se habrían medido dos desviaciones estándar hacia arriba a partir del ángulo agudo superior y hacia abajo a partir del ángulo agudo inferior antes de trazar tangentes al cuarto círculo por el origen.

Vemos que todos nuestros métodos prácticamente concuerdan para este problema. Esto depende tanto del tamaño de la muestra como del valor de  $p$ . El cambio a porcentajes implica una multiplicación por 100.

**Ejercicio 21.3.1** Otras muestras de *Dentaria* de la misma área produjeron 6 plantas con flores frente a 20 sin flores, y 29 con flores frente a 485 sin flores. Obtenga en cada caso intervalos de confianza del 95 por ciento para la proporción de plantas con flores en la población. Comentar sobre la validez de cada procedimiento.

**Ejercicio 21.3.2** En una sección de  $30 \times 10$  m de un área sembrada de *Dentaria*, se enumeraron todas las plantas y se obtuvieron 296 plantas con flores y 987 sin flor. Suponiendo que estas plantas son una muestra aleatoria de la población de *Dentaria*, obtener un intervalo de confianza del 95 por ciento para la proporción de plantas con flor en la población. Usar papel de probabilidad binomial para este propósito. [Habrá que cambiar las escalas representado (296, 987) como (29.6, 98.7) y usar la escala 1:10 de errores estándar].

## 21.4 Tablas de dos celdas, pruebas de hipótesis

Un procedimiento para intervalo de confianza, como el de la sec. 21.3 incluye una prueba de hipótesis entre alternativas bilaterales. Pero, si se desea probar una hipótesis, puede ser más cómodo calcular un criterio de prueba en vez de calcular un intervalo de confianza. En general, el modelo binomial dice que  $P(\text{de que una observación aleatoria esté en la } i\text{-ésima celda}) = p_{i0}$ ,  $i = 1, 2$  con  $p_{10} + p_{20} = 1$ .

Consideremos una generación  $F_1$  de *Drosophila melanogaster* con 35 machos y 46 hembras. Se requiere probar la hipótesis de una razón 1:1 entre sexos, es decir,  $H_0: p_{10} = 0.50$ .

*1. La aproximación normal* Usando la aproximación normal, se prueba la hipótesis  $\mu = p = 0.5$ , la razón 1:1, usando varianza  $\sigma_p^2 = \sigma^2/n = p(1-p)/n = (0.5)(0.5)/81$ , donde 81 = 35 + 46. Cálculos

$$Z = \frac{35/81 - .5}{\sqrt{(.5)(.5)/81}} = -1.22$$

De la tabla A.4, se necesita un valor de  $Z \pm 1.96$  para un nivel de significancia del 5 por ciento. De la misma tabla,  $P(|Z| \geq 1.22) = 2(0.1112) = 0.2224$ . No hay evidencia para negar la hipótesis nula.

Puede hacerse una prueba con alternativas unilaterales. Por ejemplo, el punto al 5 por ciento de significancia está en  $Z = 1.645$  si buscamos un  $p$  mayor de 0.5; y es  $Z = -1.645$  si deseamos un  $p$  menor de 0.5. Para otros niveles de significancia, ver sec. 3.6, casos 1, 1a y 1b.

2. El criterio  $\chi^2$  Aplicando la ec. (21.2) al ejemplo de la *Drosophila*, calculamos

$$\begin{aligned}\chi^2 &= \sum \frac{(\text{observados} - \text{esperados})^2}{\text{esperados}} \\ &= \frac{(35 - 40.5)^2}{40.5} + \frac{(46 - 40.5)^2}{40.5} = 1.49 \quad \text{con 1 gl}\end{aligned}$$

Aquí hemos usado los valores observados y esperados en lugar de una proporción como en  $Z$ . En la tabla A.5, encontramos  $0.25 > p(\chi^2 \text{ mayor por azar si la razón es } 1:1) > 0.10$ . Puesto que sólo entra un grado de libertad, tenemos el cuadrado de una sola desviación normal. Justo éste se calculó como  $Z = 1.22$  y  $P(|Z| \geq 1.22) = 0.2224$ . Ahora bien  $Z^2 = 1.49 = \chi^2$ . Esta relación entre  $Z$  y  $\chi^2$  se cumple solamente para un solo grado de libertad. La equivalencia de estas pruebas se demuestra fácilmente por álgebra elemental. La prueba  $\chi^2$  se efectúa con alternativas bilaterales. La interpretación de la comprobación muestral no cambia.

Para tablas de dos celdas, las desviaciones son siempre iguales en magnitud pero de signo opuesto. Esto nos permite escribir la ec. (21.2) con un gl como

$$\chi^2 = \frac{(Y - \mu)^2}{np(1-p)} \quad (21.4)$$

donde  $Y$  es uno de los números de celda observados,  $n_1$  o  $n_2$ , y  $\mu$  es el valor esperado para esa celda, o sea,  $np$  o  $n(1-p)$ . Algunas veces la ec (21.4) se escribe a veces con  $p$  y  $1-p$  remplazados por sus estimaciones muestrales; esto da  $\chi^2 = n(Y - \mu)^2/n_1 n_2$ . Los valores calculados por los dos criterios no serán iguales por lo general.

También podemos usar

$$\chi^2 = \frac{(r_2 n_1 - r_1 n_2)^2}{r_1 r_2 (n_1 + n_2)} \quad (21.5)$$

para una razón  $r_1 : r_2$  con números observados  $n_1$  y  $n_2$ . Si la razón se expresa como  $r : 1$  con  $r \geq 1$ , entonces

$$\chi^2 = \frac{(n_1 - rn_2)^2}{r(n_1 + n_2)} \quad (21.6)$$

donde  $n_2$  es el número observado en la celda con el valor esperado más pequeño. Con las ecs. (21.5) y (21.6) se obtiene el mismo resultado de la ec. (21.2).

Para mejorar la aproximación a la distribución  $\chi^2$  y así poder obtener un valor de probabilidad más exacto de la tabla  $\chi^2$ , Yates (21.9) ha propuesto una *corrección de continuidad*, aplicable cuando el criterio tiene un solo grado de libertad. Esta corrección se propone hacer que la distribución real del criterio, tal como se ha calculado a partir de

datos discretos, se acerque más a la distribución  $\chi^2$  basada en desviaciones normales. La aproximación exige que el valor absoluto de cada desviación sea disminuido en  $\frac{1}{2}$ . Así,

$$\begin{aligned}\chi^2 \text{ ajustado} &= \sum \frac{(|\text{observados} - \text{esperados}| - 0.5)^2}{\text{esperado}} & (21.7) \\ &= \frac{(|35 - 40.5| - .5)^2}{40.5} + \frac{(|46 - 40.5| - .5)^2}{40.5} \\ &= 1.23 \quad \text{para los datos de } Drosophila\end{aligned}$$

Este ajuste conduce a un valor menor de  $\chi^2$ . Por lo tanto, al probar hipótesis, el ajuste se justifica sólo cuando el valor de  $\chi^2$  no ajustado es mayor que el valor de  $\chi^2$  tabulado al nivel de probabilidad deseado.

**3. La distribución exacta** Cuando el tamaño de la muestra es pequeño, es recomendable usar la distribución exacta, es decir, la binomial. "Pequeño" puede definirse como un número menor que el número apropiado de la tabla 21.1. Pueden usarse las tablas de la distribución binomial para probar hipótesis acerca de  $p$  sin calcular límites de confianza. En vista del trabajo tan dispendioso, puede ser más ventajoso el cálculo de tales límites o el uso de las tablas de límites de confianza existentes. Usar las tablas A.14 o de Mainland (21.3) para alternativas bilaterales. Si las alternativas son unilaterales, obténgase el intervalo de confianza para dos veces la probabilidad de un error de Tipo I aceptable para fines de pruebas, o bien, póngase toda la probabilidad en la cola apropiada.

**4. Papel de probabilidad binomial** El papel de probabilidad binomial también puede usarse para probar la hipótesis nula. Para los datos de *Drosophila*, representar el triángulo (35, 46), (35, 47), (36, 46). Este triángulo es distingible de un punto. La probabilidad hipotetizada se obtiene como una división, aquí 1:1 ó 50:50, y así sucesivamente, que se representa como una recta por el origen y cualquier par de valores, que describen la división (ver fig. 21.1). La distancia más corta del triángulo a la división, en este caso a partir de (36, 46), se compara con la escala de error estándar y se ve que es un poco mayor que una desviación estándar (compárese con  $Z = 1.22$ ).

**Ejercicio 21.4.1** Woodward y Rasmussen (21.8) estudiaron el desarrollo de la cubierta y la arista en la cebada. Concluyeron que este desarrollo está determinado por dos pares de genes. Cuatro de los nueve posibles genotipos  $F_2$ , se deben segregar en la generación  $F_3$ . Las razones esperadas son 3:1 para los caracteres citados abajo con los datos. (La clasificación correcta de arista fue difícil para el tercer conjunto de datos).

2,376 cubiertas para 814 de aristas cortas

1,927 cubiertas para 642 de aristas largas

1,685 aristas largas para 636 aristas cortas

623 aristas largas para 195 aristas cortas

Probar  $H_0: p = \frac{1}{2}$  para el carácter apropiado en cada uno de los cuatro casos. Comentar sobre lo apropiado de la(s) prueba(s) que se utilicen.

## 21.5 Pruebas de hipótesis para un conjunto limitado de alternativas

Cuando se han obtenido los datos y se ha probado una hipótesis, el valor de criterio de prueba puede conducirnos a aceptar la hipótesis. La aceptación de la hipótesis nula significa que ella y el azar ofrecen una explicación razonable de los datos existentes. Sin embargo, el azar y todo valor de  $p$  dentro del intervalo de confianza no nos conducirán a negar la hipótesis nula. Así, la aceptación de la hipótesis nula, precisamente como se la ha planteado, sería un enunciado muy fuerte. De otro lado, el tipo de hipótesis alternativa hasta ahora considerada, ha sido un conjunto de alternativas. Por esto, la aceptación de la hipótesis alterna es un enunciado general, aunque un tanto vago. Es fuerte sólo con respecto a la hipótesis nula, puesto que rechaza esta hipótesis.

En ciertos problemas genéticos, puede existir un número limitado de hipótesis nulas posibles para elegir y ninguna de ellas puede ser una elección obvia para hipótesis nula. Por ejemplo, el experimentador puede tener para elegir entre razones 1:1 y 3:1. ¿Qué razón debería probarse como hipótesis nula, y cómo se prueba con una sola alternativa en oposición a un conjunto de alternativas? Si se prueban ambas hipótesis, se puede concluir que una u otra es satisfactoria. En este caso es claro que el tamaño de la muestra ha sido insuficiente para distinguir entre las razones al nivel de significancia elegido. Cuando son posibles más de dos razones, los resultados de probar cada razón como una hipótesis nula pueden ser más confusos.

Problemas como los del párrafo precedente son una clase especial de problemas y requieren una solución especial. Así, en el análisis de la varianza, una prueba  $F$  es una solución general correcta. Sin embargo, si el experimentador desea contrastar diferencias entre todos los posibles pares de medias, se recomiendan procedimientos como los indicados en el cap. 8; si los tratamientos son un conjunto factorial, entonces es aconsejable una prueba de efectos principales e interacciones. Aquí también, donde hay un número finito de razones posibles, se aconseja una alternativa a los métodos presentados en este capítulo.

Una solución parcial al problema de elegir entre dos o más razones binomiales se da en las tablas A.17, construidas por NaNagara (21.5). Estas tablas dan regiones de aceptación para las diversas razones, junto con las probabilidades de tomar una decisión errónea de acuerdo con cuál hipótesis sea cierta. Esta solución parte de la premisa de que no siempre es deseable fijar previamente la probabilidad de un error de Tipo I a uno de los niveles acostumbrados y dejar de lado la probabilidad de un error de Tipo II. Esta premisa es particularmente válida cuando ninguna hipótesis es obviamente nula. Aquí, es mejor fijar con anticipación las probabilidades de los posibles errores, y el tamaño de la muestra se elige de tal forma que no se excedan las probabilidades.

El método de solución, llamado *minimax*, minimiza la probabilidad máxima de un error de cualquier tipo. Por ejemplo, supóngase que un individuo pertenece a una de dos clases distintas y que tenemos sólo dos hipótesis acerca de la razón que entra en juego, a saber 1:1 ó 3:1. Supóngase que el tamaño de la muestra es 20. Si observamos 0.20, entonces lógicamente debemos aceptar la hipótesis 1:1. Igualmente si observamos las razones 1:19, 2:18, ..., 10:10. Pasemos ahora a 15:5; aquí lógicamente debemos aceptar la hipótesis 3:1. Lo mismo sucede si observamos 16:4, ..., 20:0. Las decisiones más difíciles serán para los valores observados 11:9, ..., 14:6.

Consideremos ahora cada uno de estos casos posibles, asociemos una regla para decidir entre las dos razones en cada caso, examinemos el resultado de cada regla y finalmente decidamos cuál es la mejor.

*Regla propuesta 1* Si observamos 0, 1, ..., 11 en la primera clase (el primer 1 de 1:1 o el 3 de 3:1), aceptamos la hipótesis 1:1 y rechazamos la hipótesis 3:1; si observamos 12, ..., 20 en la primera clase, rechazamos la hipótesis 1:1 y aceptamos la hipótesis 3:1.

*Resultado de la regla propuesta 1* Para juzgar esta regla, ya sea por su propio mérito o en relación con otras reglas, hay que saber qué hacer si la razón verdadera es 1:1 y qué hacer si la razón verdadera es 3:1.

*Resultado cuando la razón verdadera es 1:1* Con esta regla y una razón 1:1 cierta, tomamos una decisión errónea si observamos 12, ..., 20 en la primera clase, porque la regla exige que se acepte la hipótesis 3:1. Calculemos la probabilidad de una decisión errónea. Para ello sumamos las probabilidades asociadas con 12, ..., 20 en la primera clase cuando la razón verdadera es 1:1. Esto requiere una tabla de la distribución binomial para  $n = 20$  y  $p = 0.5$ . La probabilidad es 0.2517.

*Resultado cuando la razón verdadera es 3:1* Con la misma regla, pero con una razón verdadera 3:1, tomamos una decisión errónea si observamos 0, ..., 11 en la primera clase, porque la regla exige que se acepte la hipótesis 1:1. La probabilidad de una decisión errónea se encuentra sumando las probabilidades asociadas con 0, ..., 11 en la primera clase cuando la razón 3:1 es la razón verdadera. Para estas probabilidades necesitamos una tabla de probabilidad binomial para  $n = 20$  y  $p = 0.75$  (o  $p = 0.25$ ). La probabilidad es 0.0409.

Si ahora examinamos el resultado de la regla propuesta, vemos que no es demasiado satisfactorio si la razón verdadera es 1:1, pero que lo es si la razón verdadera es 3:1. Ciertamente sería bueno tener un mejor balance entre las dos probabilidades. Por lo tanto, consideremos otras reglas.

*Regla propuesta 2* Si observamos 0, ..., 12 en la primera clase, aceptamos la hipótesis 1:1 y rechazamos la 3:1; si observamos 13, ..., 20 en la primera clase, rechazamos la hipótesis 1:1 y aceptamos la 3:1.

*Resultado de la regla propuesta 2 si 1:1 es la razón verdadera* Ahora tomamos una decisión errónea si observamos 13, ..., 20 en la primera clase, porque la regla requiere la aceptación de la hipótesis 3:1. Nuevamente nos referimos a la tabla de probabilidad binomial para  $n = 20$  y  $p = 0.5$  y encontramos la suma de las probabilidades asociadas con 13, ..., 20 en la primera clase. Esta es 0.1316.

*Resultado de la regla propuesta 2 si 3:1 es la razón verdadera* Ahora tomamos una decisión errónea si observamos 0, ..., 12 en la primera clase; nos referimos a la tabla de probabilidad binomial para  $n = 20$  y  $p = 0.75$  (o  $p = 0.25$ ) y hallamos la suma de las probabilidades asociadas con 0, ..., 12 en la primera clase. La suma es 0.1018.

Esta regla da un mejor resultado que la anterior si tratamos de tener las probabilidades de una decisión errónea aproximadamente iguales, independientemente de cuál hipótesis sea cierta. Podemos considerar que ambas probabilidades son demasiado altas.

*Otras reglas propuestas* Las otras reglas que se van a considerar son: (3) aceptar la hipótesis 1:1 y rechazar la hipótesis 3:1 si observamos 0, ..., 13 en la primera clase; rechazar la hipótesis 1:1 y aceptar la hipótesis 3:1 si observamos 14, ..., 20 en la primera clase; y (4) aceptar la hipótesis 1:1 y rechazar la hipótesis 3:1 si observamos 0, ..., 14 en la primera clase; rechazar la hipótesis 1:1 y aceptar la hipótesis 3:1 si observamos 15, ..., 20 en la primera clase.

Las probabilidades de tomar decisiones erróneas se dan en la tabla 21.2 justamente para tres de las cuatro reglas que se acaban de exponer. Los números bajo el encabezamiento "regiones de aceptación para razones" se refieren a los números observados en la primera clase.

*Cómo decidir entre las posibles reglas* Ahora debemos decidir cuál es la mejor regla. Nuestro criterio de la "mejor" es que podamos mantener dentro de límites razonables la peor situación en que podamos vernos. En la tabla 21.2 observamos que la regla 1 puede llevarnos a decisiones erróneas cerca del 25 por ciento de las veces; la regla 2, a un 13 por ciento, y la regla 3, alrededor del 25 por ciento de las veces. Estos valores, subrayados en la tabla 21.2, corresponden al grupo que contiene la peor situación posible. El más pequeño de estos valores, el mínimo de las probabilidades máximas de tomar una decisión errónea, es 0.1316. Esto es lo mejor que podemos hacer y la mejor regla es la regla correspondiente, o sea la regla 2.

Teniendo, pues, en cuenta la tabla 21.2: para una muestra de tamaño 20, supóngase que la razón 1:1 es aceptada cuando hay 11 o menos en el grupo potencialmente mayor, y que se acepta la razón 3:1 cuando el número es 12 o más. Entonces se toma la decisión errónea con probabilidades 0.0409 si la razón 3:1 es la verdadera y con probabilidad 0.2517 si la razón 1:1 es la verdadera. Esto es esencialmente lo que sucede si se toma 3:1 como razón para la hipótesis nula, si la hipótesis alternativa es apropiadamente unilateral, y si el error de Tipo I se fija en el 5 por ciento; 4.09 por ciento es lo más cercano que podemos llegar al 5 por ciento. Por otra parte, si la región de aceptación para la razón 1:1 requiere 13 o menos en la clase potencialmente mayor y para la razón 3:1 14 o más en esa clase, la probabilidad de tomar una decisión errónea es 0.0577 si la verdadera es la razón 1:1 y 0.2142 si la verdadera es la razón 3:1. Esto es aproximadamente equivalente a probar la hipótesis nula de una razón 1:1 con hipótesis alternativas apropiadamente unilaterales y con un error de Tipo I fijo en 0.05, el porcentaje más cercano a 5 que podemos

**Tabla 21.2 Probabilidades de tomar decisiones erróneas para razones 1:1 y 3:1,  $n = 20$**

Regla	Regiones de aceptación para razones		Probabilidades de tomar una decisión errónea cuando la razón verdadera es	
	1 : 1	3 : 1	1 : 1	3 : 1
1	0-11	12-20	.2517	.0409
2	0-12	13-20	.1316	.1018
3	0-13	14-20	.0577	.2142

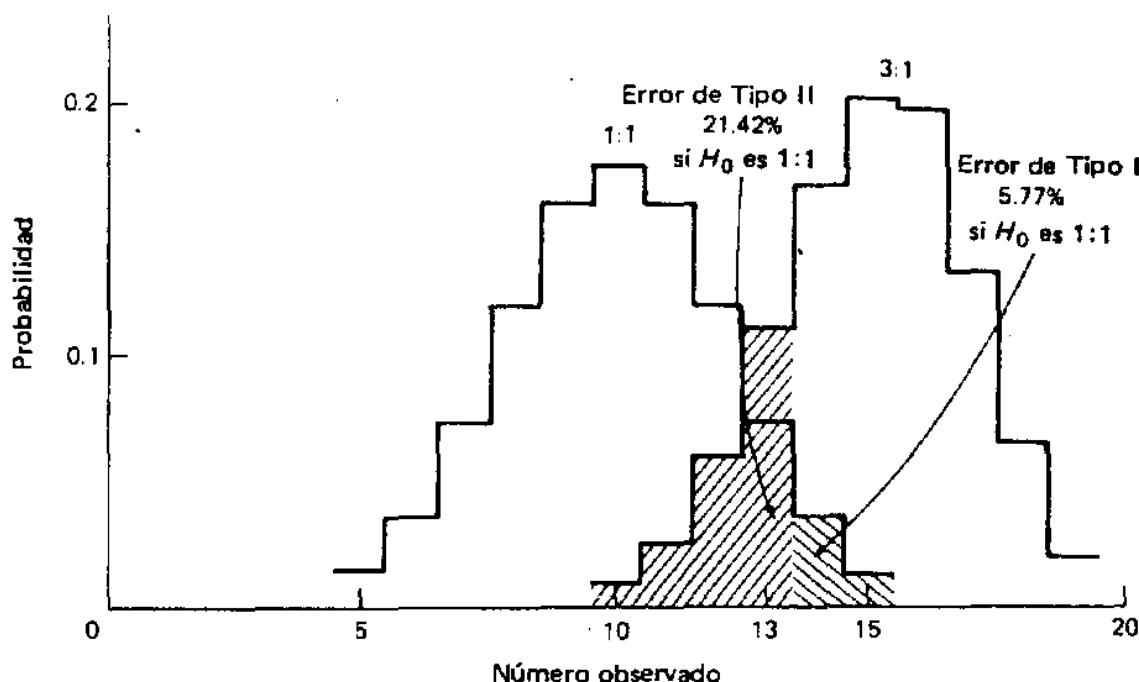


Figura 21.2 Distribuciones de probabilidad binomial para razones 1:1 y 3:1 y  $n = 20$

obtener es 5.77. Con el punto de división entre 12 y 13, las probabilidades de una decisión errónea son más aproximadamente iguales independientemente de cuál sea la hipótesis verdadera. Para las tres situaciones, las probabilidades máximas de error son 0.2517, 0.1316 y 0.2142. Al elegir 0–12 para la primera clase, elegimos la solución cuya probabilidad máxima de decisión errónea es un mínimo, o sea, 0.1316.

En la figura 21.2 se visualiza la situación existente cuando la región de aceptación es 0–13 para la razón 1:1. Supóngase que la razón verdadera es 1:1. Si obtenemos 14 o más individuos en la primera clase, aceptamos erróneamente la razón 3:1. La probabilidad asociada con este error es la suma de las probabilidades de observar exactamente 14, exactamente 15, ..., exactamente 20, o sea  $P = 0.0577$ , calculada para la razón 1:1. Esta es la probabilidad de un error de Tipo I si la hipótesis nula es que la razón es 1:1. Supóngase ahora que la verdadera razón es 3:1. Si obtenemos 13 o menos individuos en la primera clase, aceptamos erróneamente la razón 1:1. La probabilidad de este tipo de error es la suma de las probabilidades de observar exactamente 13, exactamente 12, ..., exactamente 0, o sea  $P = 0.2142$ , calculada para la razón 3:1. Esta es la probabilidad de un error de Tipo II para hipótesis nula de que la razón es 1:1. Para mayor claridad, las probabilidades de la fig. 21.2 se muestran como columnas llenas en vez de rectas verticales.

Las tablas A.17 contienen únicamente soluciones mínimas. Para ilustrar su uso, consideremos la tabla A.17A para un tamaño de muestra 44. Aceptamos la hipótesis 1:1 si las observaciones 0–27 pertenecen al grupo potencialmente mayor; si este grupo contiene 28 a 44 observaciones, aceptamos la hipótesis 3:1. En la peor situación, estaríamos en error 4.81 por ciento de las veces en promedio; éste sería el caso si los datos procedieran siempre de una distribución con una razón 1:1. Si nunca se presentaran datos de distribuciones diferentes a una distribución con una razón 3:1, entonces estaríamos en error 3.18 por ciento de las veces en promedio.

Puesto que se usó la distribución binomial para obtener las probabilidades tabuladas, es virtualmente imposible encontrar las probabilidades exactas de 0.05 y 0.01 en estas tablas. Las tablas restantes de este grupo dan reglas para aceptar varias hipótesis alternativas para tamaños de muestras dados, y las probabilidades de tomar decisiones erróneas.

Mather (21.10) expone un criterio algo diferente, llamado la *razón ambigua*, para distinguir entre dos razones. Primero, expresa estas como  $l_1 : 1$  y  $l_2 : 1$  donde  $l_i \geq 1$ . Entonces la razón ambigua es  $\sqrt{l_1 l_2} : 1$ , un valor que determina el punto de división entre regiones. Esto conduce a la *segregación ambigua* tal que  $\chi^2$  será el mismo valor, independientemente de qué razón que se propone como hipótesis nula.

Supongamos que se necesita distinguir entre las razones 1:1 y 3:1 con base en un tamaño muestral de 20. La razón ambigua es  $\sqrt{1(3)} : 1 = 1.732 : 1$ . El punto de división es entonces  $[1.732/(1 + 1.732)]20 = 0.63397(20) = 12.7$ . En consecuencia, aceptamos la razón 1:1 si la clase potencialmente mayor tiene 12 o menos individuos y la razón 3:1 si se tienen 13 o más individuos.

Las dos regiones son las mismas para el criterio minimax, aunque las razones 0.63397 y 0.63091 son ligeramente diferentes.

**Ejercicio 21.5.1** En un problema de genética, se necesitan suficientes datos para distinguir entre razones de prueba 3:1 y 7:1. ¿Cuántos hijos deberán observarse si el experimentador se considera satisfecho con un 10 por ciento por error independientemente de qué razón sea la verdadera?

**Ejercicio 21.5.2** Un experimentador tiene para clasificar 727 plantas  $F_2$  y sabe que tendrá que distinguir entre las razones 9:7, 13:3 y 15:1. ¿Cuál debería ser la regla para distinguir entre estas razones si desea tener la protección máxima cuando la naturaleza le presente en la peor situación?

**Ejercicio 21.5.3** Si tratamos de usar la técnica de razón ambigua en el problema de distinguir entre tres razones, ¿cuántos puntos de división, pueden encontrarse? ¿Cuál sería su solución?

## 21.6 Tamaño de la muestra

Supóngase que deseamos probar la hipótesis nula que el parámetro desconocido  $p$  de una población binomial es un valor específico. Al hacer la prueba, será necesario poder detectar cierta alternativa con frecuencia razonable. En la sección anterior se expuso el procedimiento minimax para elegir las regiones de aceptación y de rechazo cuando nos enfrentamos a un conjunto limitado de posibles valores  $p$ . Las tablas A.17 asociadas pueden usarse cuando se desea ver de qué clase de protección se dispone para las reglas allí dadas; también pueden usarse hasta cierto punto para elegir el tamaño de la muestra.

Consideremos el uso del papel de probabilidad binomial para elegir el tamaño de la muestra en varios casos.

**Caso 1** Supóngase que sólo dos razones se consideran posibles. Se desea conocer el tamaño de muestra necesario para distinguir entre las dos razones.

Por ejemplo, sean las razones 1:1 y 3:1 las únicas posibles. Además, independientemente de cuál razón sea la verdadera, la gravedad de una decisión errónea es casi la misma. Es decir, que deseamos tener una probabilidad aproximadamente igual de tomar una decisión errónea sea cual fuere la razón verdadera, fijemos esta probabilidad en 0.05.

El procedimiento que se presenta se basa en el uso del papel de probabilidad binomial. Primero, tracemos las divisiones 1:1 y 3:1 en el papel (ver fig. 21.1). Sobre el lado más cercano de una división a la otra, trazamos una paralela a 1.645 unidades en la escala plena (alrededor de  $1.645 \times 5 = 8.2$  mm en el papel gráfico comercial) de la primera división. El valor 1.645 es tal que  $P(Z \geq 1.645) = 0.05$  y se toma porque buscamos un procedimiento de prueba esencialmente unilateral. Independientemente de cual hipótesis sea la verdadera. Estas dos rectas, cada una paralela a cada división, se cortan aproximadamente en el punto (25, 15). Este punto está situado en el vértice derecho inferior del triángulo rectángulo cuyo ángulo recto está en (24, 15) y en el vértice izquierdo superior del ángulo derecho en (25, 14). O sea que el triángulo con ángulo recto en (25, 15) tiene sus otros ángulos fuera de los límites de confianza unilaterales al 95 por ciento. En esencia da un valor crítico para rechazar  $H_0$  a  $\alpha = 0.05$ , independientemente de si se ha especificado una razón 1:1 o una razón 3:1. Por lo tanto, el número de observaciones requeridas en nuestra muestra es  $25 + 15 - 1 = 39$  observaciones. Un tamaño de muestra de 39 nos permitirá tomar una decisión en todos los casos; un tamaño de muestra nos dará mayor protección que la deseada. Para el tamaño de muestra mayor, tendremos que alterar nuestras probabilidades para obtener una regla de decisión entre alternativas que cubra todos los posibles resultados muestrales.

Si podemos comparar este resultado con el obtenido por las tablas A.17, encontramos una ligera diferencia. La tabla apropiada indica que el número requerido para garantizar la protección deseada es 44, pero que en realidad es más de lo que se necesita. La tabla también indica que 40 observaciones casi dan la protección deseada. Recordemos que el papel de probabilidad binomial se basa en una transformación que da una variable con distribución casi normal y que sólo puede darnos un tamaño de muestra aproximado. Esto explica la discrepancia.

*Caso 2* Supóngase que tenemos sólo un candidato para la hipótesis nula, pero que se desea tener una seguridad razonable de poder detectar un valor alterno de un tamaño especificado.

Por ejemplo, supóngase que estamos haciendo un muestreo de opinión pública con una pregunta que exige un sí o un no como respuesta. Deseamos probar  $H_0: p = 0.5$  pero queremos detectar  $p = 0.65$  si está en una dirección especificada. Decidimos probar  $H_0$  al nivel del 5 por ciento usando una prueba unilateral y detectar  $p = 0.65$  si es la razón verdadera el 90 por ciento de las veces.

Trácense las divisiones 1:1 y 65:35 (ver fig. 21.1). Trazar una paralela a 1.645 desviaciones estándar en la escala plena a la división 1:1. Esta recta estará entre las dos divisiones y representa una prueba de una cola de  $H_0: p = 0.5$  al nivel del 5 por ciento. También, entre las dos divisiones, trazar una paralela a 65:35 a 1.28 desviaciones estándar de ella en la escala plena (alrededor de  $1.28 \times 5 = 6.4$  mm). Esta recta corresponde a una prueba de una cola de  $H_0: p = 0.65$  al nivel del 10 por ciento. Leemos el punto (52, 38) en la intersección de las paralelas. El tamaño de muestra apropiado será aproximadamente 90. En ciertos casos, tales como éste, resulta difícil leer exactamente la escala y el papel. Si se desea más exactitud, hay que valerse de que  $\sigma = 5.080$  mm en la escala plena del papel comercial.

*Caso 3* Supóngase que tenemos un único candidato para la hipótesis nula, pero que tenemos una probabilidad razonable de detectar una alternativa si difiere hasta en un 10 por ciento.

Por ejemplo, supóngase que  $H_0: p = 0.25$ . Vamos a probar esta hipótesis nula usando una prueba de dos colas al nivel de significancia del 5 por ciento. Queremos determinar una proporción de 15 ó 35 por ciento, si tal proporción es la verdadera alrededor del 75 por ciento de las veces.

Represéntese la división 25:75. Trácense, además, las divisiones 35:65 y 15:85 (ver fig. 21.1). A cada lado de la división 25:75 se traza una paralela a 1.96 desviaciones estándar de ella en la escala plena. Estas corresponden al procedimiento bilateral de prueba. Si se miran simplemente las sumas de las coordenadas de los puntos donde se cortan las paralelas a las divisiones 35:65 y 15:85, se verificará que no hay prueba que nos garantice con precisión una probabilidad prefijada de detección para cada alternativa. (Los puntos referidos son tales que nos dan 50 por ciento de detección). Seremos cautelosos y pediremos el 75 por ciento de protección en el caso menos favorable, o sea, cuando  $p = 0.35$ . Ahora trácese una paralela a la división 35:65 a 0.67 unidades de ella. [ $P(Z \geq 0.67) = 0.25$  para la distribución normal]. Las dos rectas se cortan aproximadamente en (46, 98). Por tanto, el tamaño requerido de la muestra es 144 aproximadamente.

**Ejercicio 21.6.1** Usar el papel de probabilidad binomial para determinar el tamaño de la muestra necesario para distinguir entre las siguientes razones: 3:1 y 7:1; 9:7 y 15:1; 27:37 y 3:1. Suponer que estamos preparados para rechazar falsamente una razón alrededor del 5 por ciento de las veces, pero no con mayor frecuencia.

**Ejercicio 21.6.2** Para un problema de muestreo con datos discretos, se ha decidido probar  $H_0: p = 0.40$ . Se usará una prueba unilateral al nivel del 1 por ciento. Se requiere detectar  $H_1: p = 0.65$ , si es la hipótesis verdadera, alrededor del 80 por ciento de las veces. ¿Cuál es el tamaño de muestra necesario?

**Ejercicio 21.6.3** Supóngase que hay que distinguir entre las razones 1:1, 3:1 y 7:1. Quisiéramos fijar la probabilidad de rechazar cada hipótesis, siendo verdadero, alrededor del 0.05 pero obviamente tendríamos que arreglar las cosas de tal modo que ésta sea la probabilidad máxima de una decisión errónea. ¿Se puede elaborar un procedimiento para encontrar el tamaño de muestra necesario usando papel de distribución binomial? Comparar el resultado con el obtenido por la tabla A.17.

## 21.7 Tablas de una vía con $n$ celdas

Robertson (21.6) proporciona datos que incluyen la progenie  $F_2$  de un cruce de cebada. Los caracteres observados son no en dos filas frente a en dos filas y normal o verde frente a planta clorótica. Estos datos están registrados en la tabla 21.3. Se desea probar la hipótesis de una segregación normal de híbrida en razón 9:3 : 3:1. Los valores esperados están dados por  $np$  donde  $n = 1,898$  y  $p = 9/16, 3/16, 3/16$ ; y  $1/16$ , respectivamente.

Calcular  $\chi^2$  como

$$\chi^2 = \sum \frac{(\text{observados} - \text{esperados})^2}{\text{esperados}}$$

$$= 54.36^{**} \quad \text{con } 3 \text{ gl}$$

Tabla 21.3 Valores observados y esperados para un progenie  $F_2$ 

	Verde		Clorótico		
	No en dos filas	En dos filas	No en dos filas	En dos filas	Totales
Observado	1,178	291	273	156	1,898
Esperado	1,067.6	355.9	355.9	118.6	1,898
(Dif) <sup>2</sup> /esperado	11.416	11.835	19.310	11.794	54.36

Los grados de libertad son el número de celdas disminuido en uno debido a que sólo el tamaño de la muestra es fijo y no hay parámetros estimados. La evidencia está en contra de la razón teórica de 9:3:3:1. El genetista probablemente concluirá que la asociación fue causa de un ajuste tan insuficiente. (Estos datos se estudian más detalladamente en el cap. 22).

¿Cuál era nuestra hipótesis alternativa? Simplemente que la razón era diferencia de 9:3:3:1. Cuando se especifica una hipótesis alternativa, resulta deseable un método más efectivo para decidir entre las alternativas posibles. Aunque la distribución multinomial es lo indicado como distribución exacta, es difícil definir regiones de aceptación y rechazo. La tabla A.17 se utiliza para elegir entre unos cuantos conjuntos de razones de tres resultados.

**Ejercicio 21.7.1** Woodward y Rasmussen (21.8) han planteado la siguiente hipótesis: 9 cubiertas a 3 de aristas largas a 3 aristas cortas como razón en la generación  $F_2$ . Los datos observados fueron 348:115:157. Probar sus hipótesis. ¿Cuántos grados de libertad hay para el criterio  $\chi^2$ ?

## Referencias

- 21.1. Clopper, C. J. y E. S. Pearson: "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika* 26:404 (1934).
- 21.2. Cochran, W. G.: *Sampling Techniques*, Wiley, Nueva York, 1953.
- 21.3. Mainland, D., L. Herrera, y M. I. Sutcliffe: *Tables for Use with Binomial Samples*, published by the authors, Nueva York, 1956.
- 21.4. Mosteller, F., y J. W. Tukey: "The uses and usefulness of binomial probability paper," *J. Amer. Statist. Ass.*, 44:174-212 (1949).
- 21.5. NaNagara, P.: "Testing Mendelian ratios," M.S. Thesis, Cornell University, Ithaca, N.Y., 1953.
- 21.6. Robertson, D. W.: "Maternal inheritance in barley," *Genetics* 22:104-113 (1937).
- 21.7. *Tables of the Cumulative Binomial Probability Distribution*, Ann. Computation Lab. of Harvard Univ., vol. 35, Harvard University Press, Cambridge, Mass., 1955.
- 21.8. Woodward, R. W., y D. C. Rasmussen: "Hood and awn development in barley determined by two gene pairs," *Agron. J.*, 49:92-94 (1957).
- 21.9. Yates, F.: "Contingency tables involving small numbers and the  $\chi^2$  test," *J. Roy. Statist. Soc. Soc. Suppl.*, 1:217-235 (1934).
- 21.10. Mather, K., *Measurement of Linkage in Heredity*, 2a. ed. Methuen, Londres, 1951.
- 21.11. *Tables of the Binomial Probability Distribution*, National Bureau of Standards Applied Mathematics Series 6, 1949.

---

**DATOS ENUMERATIVOS II:  
TABLAS DE CONTINGENCIA**


---

### **22.1 Introducción**

Con frecuencia los individuos se clasifican de acuerdo con varias variables. Por ejemplo, una persona puede clasificarse como fumadora o no fumadora, y al mismo tiempo como un individuo con o sin enfermedad coronaria; una mosca de la fruta puede clasificarse como macho o hembra y de acuerdo con su ascendencia, etc. Hay dos variables de clasificación en cada caso.

En otros casos, los individuos pueden asignarse a grupos y luego clasificarse dentro de cada grupo con respecto a alguna variable. Por ejemplo, un individuo puede asignarse a un grupo de tratamiento o a un grupo de control y posteriormente clasificarse según la respuesta a un estímulo.

Para ambas ilustraciones, los datos se registran en una tabla de doble entrada en forma conveniente. Este capítulo se ocupa del análisis de datos enumerativos en tales tablas, llamadas a menudo tablas de contingencia.

### **22.2 El modelo de muestreo aleatorio**

Weir (22.25) proporcionó los datos de la tabla 22.1, una tabla de contingencia de  $r \times c = 6 \times 6$ . Aquí,  $n_{..} = 4,396$  individuos de una población de cebada que ha sido clasificada simultáneamente de acuerdo con un lugar geométrico de  $A$  y  $B$  esterasa. Se empleó una técnica electroforética para determinar el genotipo o la composición genética de cada individuo. El total  $n_{..} = 4,396$  se consideró fijo. No hay totales marginales fijos, sino que pueden variar de una muestra a otra en 4,396 observaciones.

Un modelo de probabilidad no complicado dice que la probabilidad de que un individuo seleccionado al azar sea clasificado en la  $i, j$ -ésima celda es  $p_{ij}$ ,  $i, j = 1, \dots, 6$  con

$$\sum_{i,j} p_{ij} = 1$$

Tabla 22.1 Una tabla  $6 \times 6$  de frecuencias genéticas

	$A_1 A_1$	$A_2 A_2$	$A_3 A_3$	$A_1 A_2$	$A_1 A_3$	$A_2 A_3$	
$B_1 B_1$	311	5	205	2	103	9	$635 = n_1.$
$B_2 B_2$	63	954	2,172	2	3	175	$3,369 = n_2.$
$B_3 B_3$	13	275	7	1	0	4	$300 = n_3.$
$B_1 B_2$	1	34	12	1	1	15	$64 = n_4.$
$B_1 B_3$	0	1	0	0	1	1	$3 = n_5.$
$B_2 B_3$	2	6	1	5	1	10	$25 = n_6.$
	390	1,275	2,397	11	109	214	4,396
	$= n_{..1}$	$= n_{..2}$	$= n_{..3}$	$= n_{..4}$	$= n_{..5}$	$= n_{..6}$	$= n_{..}$

Fuente: Datos obtenidos por cortesía de Bruce Weir, Universidad del Estado de Carolina del Norte, Raleigh, North Carolina. Ver también Ref. 22.25.

Un modelo más restringido requiere que la probabilidad de un genotipo particular  $A$  sea independiente de la del genotipo  $B$ . Esto significa que si se conocen las probabilidades para las clases  $A$  y  $B$ , entonces sus productos dan las probabilidades de celda. Este enunciado describe el modelo simétrico en  $A$  y  $B$ . La *independencia* y no independencia o *interacción* se ilustran en la tabla 22.2.

Para tales datos, la hipótesis nula usual es la de independencia, o sea,

$$H_0: p_{ij} = p_{i\cdot} p_{\cdot j} \quad (22.1)$$

En este caso,  $p_{ij}$  es la probabilidad de que un individuo al azar sea clasificado en la  $i, j$ -ésima celda. Las  $p_{i\cdot}$  y  $p_{\cdot j}$  son las probabilidades de fila y columna respectivamente, con

$$\sum_i p_{i\cdot} = 1 = \sum_j p_{\cdot j}$$

La hipótesis alternativa es

$$H_1: p_{ij} \neq p_{i\cdot} p_{\cdot j}$$

El criterio de prueba es

$$\chi^2 = \sum \frac{(\text{observado} - \text{esperado})^2}{\text{esperado}} \quad (r-1)(c-1) \text{ gl} \quad (22.2)$$

Este criterio de prueba es llamado también  $X^2$ , debido a que sólo aproximadamente está distribuido como  $\chi^2$ . Para una buena aproximación, los valores esperados no deberían ser demasiado pequeños.

Tabla 22.2 Probabilidades en tablas de contingencia de dos vías

Independiente			Dependiente				
	$A_1$	$A_2$	Suma		$A_1$	$A_2$	Suma
$B_1$	$\frac{3}{4} \left( \frac{3}{4} \right) = \frac{9}{16}$	$\frac{3}{4} \left( \frac{1}{4} \right) = \frac{3}{16}$	$\frac{3}{4}$	$B_1$	$\frac{5}{8}$	$\frac{1}{8}$	$\frac{3}{4}$
$B_2$	$\frac{3}{4} \left( \frac{1}{4} \right) = \frac{3}{16}$	$\frac{1}{4} \left( \frac{1}{4} \right) = \frac{1}{16}$	$\frac{1}{4}$	$B_2$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$
Suma	$\frac{3}{4}$	$\frac{1}{4}$	1	Suma	$\frac{3}{4}$	$\frac{1}{4}$	1

Probabilidades marginales dadas;  
la independencia permite el cálculo  
de las probabilidades de celda

Probabilidades marginales dadas;  
la dependencia no permite el cálculo  
de probabilidades de celda

*Nota:* Las probabilidades de las celdas deben sumar las probabilidades marginales por filas y columnas.

Los valores esperados se calculan en el supuesto de que la hipótesis nula es verdadera. Puesto que las  $p_{i\cdot}$  y las  $p_{\cdot j}$  no están dadas, deben estimarse. Para éstas, usamos los totales marginales y así obtener proporciones muestrales apropiadas.

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n_{..}} \quad i = 1, \dots, r \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n_{..}} \quad j = 1, \dots, c$$

Ahora bien,  $\hat{p}_{ij} = \hat{p}_{i\cdot} \hat{p}_{\cdot j}$  y los valores esperados están dados por

$$E_{ij} = \hat{p}_{ij} n_{..} = \frac{n_{i\cdot} n_{\cdot j}}{n_{..}} \quad (22.3)$$

Por ejemplo,

$$E_{54} = \frac{3(11)}{4,396} = .0075$$

El valor 0.0075 está muy por debajo del mínimo aceptable indicado por Cochran (sec. 20.4). En consecuencia, reducimos la tabla combinando  $A_1$  y  $A_2$ . En términos de la tabla,  $A_1 A_1$ ,  $A_2 A_2$ , y  $A_1 A_2$  se convierten en  $A_1^* A_1^*$ , mientras que  $A_1 A_3$  y  $A_2 A_3$  se convierten en  $A_1^* A_3$ . Combinar en forma análoga  $B$ . La tabla 22.3 es el resultado.

Así mismo, se calculan los valores esperados. El valor más pequeño es 2.06 y no ha de causar dificultad. Las sumas por filas y columnas de los valores esperados deben ser iguales a las sumas de observaciones correspondientes. Las sumas análogas de desviaciones

Tabla 22.3 Cálculo de  $\chi^2$  para una tabla  $r \times c$ 

		$A_1 A_1^*$	$A_1 A_3$	$A_1^* A_3$	Totales
$B_1^* B_1^*$	Observado	1373	2389	306	4,068
	Esperado	1,550.95	2,218.15	298.90	4,068.00
	Desviación	-177.95	170.85	7.10	.00
	$\chi^2$	20.42	13.16	.17	
$B_3 B_3$	Observado	289	7	4	300
	Esperado	114.38	163.58	22.04	300.00
	Desviación	174.62	-156.58	-18.04	.00
	$\chi^2$	266.60	149.88	14.77	
$B_1^* B_3$	Observado	14	1	13	28
	Esperado	10.68	15.27	2.06	28.01
	Desviación	3.32	-14.27	10.94	-.01
	$\chi^2$	1.04	13.33	58.13	
Totales	Observado	1,676	2,397	323	4,396
	Esperado	1,676.01	2,397.00	323.00	4,396.01
	Desviación	-.01	.00	.00	-.01

deben ser iguales a cero. Las pequeñas discrepancias de estas sumas pueden atribuirse a errores de aproximación.

Los valores  $\chi^2$  cuadrado de la tabla 22.3 realmente son contribuciones al valor total  $\chi^2$ . Obsérvese que cada una se mide con relación a la celda  $E_{ij}$  como  $(n_{ij} - E_{ij})^2/E_{ij}$ . De este modo, una desviación grande donde  $E_{ij}$  es grande y una pequeña donde  $E_{ij}$  es pequeño, pueden contribuir más o menos igualmente a  $\chi^2$ . Cuando el  $\chi^2$  total es significante, las magnitudes de las contribuciones pueden ser útiles en la interpretación de los datos.

Los grados de libertad para esta tabla  $r \times c$  son  $(r - 1)(c - 1)$ . Aquí,  $gl = 2(2) = 4$ . Un argumento razonable es que  $n = 4,396$  es una restricción del muestreo y que hemos tenido que estimar probabilidades de dos filas independientes y de dos columnas independientes, ya que no se hicieron suposiciones acerca de sus valores. Recuérdese que

$$\sum \hat{p}_{i \cdot} = 1 = \sum_j \hat{p}_{\cdot j}$$

En consecuencia,  $gl = 3(3) - 1 - 2 - 2 = 4$ .

Aquí,  $\chi^2_4 = 539.50$ , donde  $\chi^2_{0.01} = 13.3$ . Los datos y el azar no respaldan la hipótesis nula. Concluimos que las probabilidades de fila y columna no son independientes y que hay interacción. No es posible aplicar un conjunto único de probabilidades a todas y cada una de las filas; lo mismo ocurre con las columnas. Usese  $\hat{p}_{ij} = n_{ij}/n$ . Si bien es necesario ahora estimar probabilidades de celda.

Un valor grande  $\chi^2$  indica falta de independencia de las variables de clasificación, pero de poca información acerca del grado de independencia. Una medida de la dependencia es la dada por:

$$\frac{\chi^2}{n(t - 1)} \quad (22.4)$$

donde  $t$  es el menor valor de  $r$  y  $c$ , y  $n$  es el número total de observaciones en la tabla. Este criterio está entre 0 y 1. Su distribución está obviamente relacionada con  $\chi^2$  por un simple cambio de variable. Aquí, el valor es  $539.50/4,396(2) = 0.06$ , que no es muy grande debido, en parte, al gran tamaño de la muestra.

Existen otros coeficientes de contingencia.

**Ejercicio 22.2.1** Weir proporciona otros datos de frecuencias para dos lugares geométricos de esterasa en una población de cebada. Los datos son los siguientes:

		<i>B</i> Genotipo		
		1, 1	2, 2	1, 2
C Genotipo	1, 1	2,172	1,019	178
	2, 2	212	608	118
1, 2		13	49	27

*Fuente:* Datos utilizados con autorización de Bruce Weir, Universidad del Estado de Carolina del Norte, Raleigh, North Carolina. Ver también Ref. 22.25.

Usar  $\chi^2$  para probar la hipótesis nula de que la probabilidad de ser un genotipo *B* particular es independiente de la probabilidad de ser un genotipo *C* particular.

**Ejercicio 22.2.2** Varios árboles de pino blanco fueron clasificados de acuerdo con clases de edad y reacción injertos de roya de los pinos. Los datos son los siguientes

		Edad en años del árbol padre			
		4	10	20	$\geq 40$
Reacción	Sanos	7	6	11	15
	Enfermos	14	11	5	8

Usar  $\chi^2$  para probar la hipótesis nula de que la probabilidad de una reacción sana es independiente de la probabilidad de seleccionarse de una clase particular de edad. (Estos datos son analizados más detalladamente en la sec. 22.10).

### 22.3 El modelo de muestreo aleatorio estratificado

Consideremos los datos de Di Raimondo (22.2) para ratones de la tabla 22.4. Los ratones no fueron tratados con penicilina, sino inyectados con un inóculo bacteriano (*staphylococcus aureus*) cultivado en caldo enriquecido con las vitaminas niacinamida (NA), ácido

Tabla 22.4 Valores observados y valores esperados para los datos de ratones.

Inóculo	Vivos = A		Muertos		Totales	
	Observado	Esperado	Observado	Esperado	Observado	Esperado
NA	10	9.65	30	30.35	40	40
AF	9	9.65	31	30.35	40	40
Paba	9	12.06	41	37.94	50	50
B <sub>6</sub>	13	9.65	27	30.35	40	40
Totales	41	41.01	129	128.99	170	170

fólico (AF), ácido *p*-aminobenzoico (Paba), y B<sub>6</sub> como piridoxina, cada uno en concentración de 10 µg/ml.

Es claro que se especificó el tamaño de cada grupo de tratamiento. Estas son *muestras independientes*, donde la asignación de los ratones a grupos fue aleatoria. Las únicas probabilidades que intervienen son las de  $P(A)$  y  $P(\text{no } A) = P(\bar{A})$ ; estas pueden variar con el inóculo.

La hipótesis nula usual expresa que el parámetro  $p_A$ , que es parámetro binomial, es constante para todo inóculo, como lo es  $p_A$ , o sea,

$$H_0: \begin{cases} p_{Ai} = p_A \\ p_{\bar{A}i} = 1 - p_A \end{cases} \text{ para todo } i, \text{ con } p_{Ai} + p_{\bar{A}i} = 1$$

Esta es una hipótesis de *homogeneidad*.

La hipótesis alternativa, o de heterogeneidad, es

$$H_1: p_{Ai} \neq p_{A'i} \quad \text{algún } i, i'$$

Nuevamente el criterio de prueba es la ec. (22.2).

Los valores esperados se calculan en el supuesto de que la hipótesis nula es verdadera.  $p_A$  debe estimarse puesto que no está dado. Para esto, se combinan todos los datos, lo cual da

$$\hat{p}_A = \frac{n_{..1}}{n_{..}} \quad \text{y} \quad \hat{p}_{\bar{A}} = \frac{n_{..2}}{n_{..}}$$

Para hallar los valores esperados, multiplíquense estas probabilidades por cada tamaño de muestra como sigue

$$\begin{aligned} E_{ij} &= \hat{p}_A n_{i..} \quad (\text{o } \hat{p}_{\bar{A}} n_{i..}) \\ &= \frac{n_{i..} n_{..j}}{n_{..}} \end{aligned} \tag{22.5}$$

Esta es la misma ec. (22.3).

Un argumento razonable para justificar los grados de libertad dice que cada uno de los totales  $r$  de fila ha sido fijado, pero que sólo se han tenido que estimar  $c - 1$  probabilidades independientes; aquí  $c - 1 = 1$ . En consecuencia,  $gl = rc - r - (c - 1) = (r - 1)(c - 1)$ .

Aquí,  $\chi^2_3 = 2.63$ , donde  $\chi^2_{.05} = 7.81$ . En efecto,  $0.50 > P(\chi^2_3 > 2.63) > 0.25$ . Con base en esta muestra, concluimos que la razón de ratones vivos a muertos sólo varía al azar de un inóculo a otro. Si hay diferencias reales en las razones de población, entonces nuestra muestra no ha sido lo suficientemente grande para detectar estas diferencias.

Los términos independencia y homogeneidad tienden a intercambiarse en su uso. Lo mismo ocurre para falta de independencia, dependencia, interacción y heterogeneidad.

Se dispone de otras técnicas de cálculo; Steel y Torrie (22.21), dan algunas.

**Ejercicio 22.3.1** Green (22.5) obtuvo tres distribuciones de frecuencia en un estudio de la herencia de habilidad combinatoria en el maíz. Lo que se comprobó estuvo a favor de una varianza común. Entonces se tornó interesante ver si las distribuciones tenían la misma forma independiente de su localización. La tabla siguiente de distribuciones de frecuencia usadas para comparar las formas.

Muestra $F_2$	Centro de clase: Errores estándar por encima o por debajo de la media							Total
	-3	-2	-1	0	1	2	3	
Alto x alto	1	5	28	19	22	8	—	83
Alto x bajo	2	5	19	28	23	6	—	83
Bajo x bajo	3	9	18	24	18	9	2	83

Probar la hipótesis de homogeneidad de forma. (Recuérdese que gran parte de la información referente a la forma de una distribución está asociada con las colas, como se ha visto en la sec. 20.4. Por tanto, combínense sólo las clases  $+2$  y  $+3$ ).

¿Sería apropiado considerar el modelo estratificado como correcto para estos datos? Explicar.

**Ejercicio 22.3.2** Rinke (22.18) observó los datos de la siguiente tabla en un estudio de la herencia de la cumarina en trébol dulce.

Clases para hábito de crecimiento	Clases de cumarina en 1/100 de 1 por ciento			Total
	0-14	15-29	Superior a 30	
Estándar	8	24	7	39
Estándar intermedio	11	28	17	56
Alfa intermedia	9	31	11	51
Alfa	1	14	11	26
Total	29	97	46	172

¿Se comprueba independencia entre hábito de crecimiento y contenido de cumarina? ¿Qué modelo de muestreo sería apropiado para estos datos? Explicar.

**Ejercicio 22.3.3** Calcular el coeficiente de contingencia para los datos de los ejercicios 22.3.1 y 22.3.2. Comentar sobre las aplicaciones a estos datos.

#### 22.4 Tabla cuádruple o de $2 \times 2$

Los datos de la tabla 21.3 entran naturalmente en una tabla de dos factores, o sea la tabla 22.5. Aquí el modelo apropiado es el de muestreo aleatorio simple. Probemos la hipótesis de independencia sin ningún supuesto acerca de las razones verdaderas. Mientras que la ec. (22.2) es apropiada, la ec. (22.6) da una alternativa conveniente para  $\chi^2_1$ , sea cual fuere el modelo aplicado.

$$\begin{aligned}\chi^2_1 &= \frac{(n_{11}n_{22} - n_{12}n_{21})^2 n_{..}}{n_{1..}n_{2..}n_{.1}n_{.2}} \\ &= 50.54^{**} \quad \text{para estos datos}\end{aligned}\tag{22.6}$$

Estos datos nos llevan a rechazar la hipótesis nula.

Grizzle (22.6) consideró la corrección de continuidad (ver sec. 22.5) para tablas  $2 \times 2$  generadas con  $H_0$ , calculando tanto las  $\chi^2$  corregidas como las no corregidas, y comparando los resultados empíricos con la distribución  $\chi^2_1$  a los niveles del 5 y del 1 por ciento. Sus conclusiones fueron que la  $\chi^2$  sin corregir es la que se comporta mejor, y que la  $\chi^2$  corregida es demasiado prudente, es decir, que descarta la  $H_0$  demasiado pocas veces. Así mismo, para esperanzas menores de 5, la  $\chi^2$  sin corregir es conservadora y la  $\chi^2$  corregida es aún más prudente.

**Aproximación normal** La ecuación (22.6) es equivalente a la aproximación normal dada por

$$Z = \chi = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})(1/n_{1..} + 1/n_{2..})}}\tag{22.7}$$

donde  $\hat{p}_i = n_{i1}/n_{..}$  estima el parámetro de población en la  $i$ -ésima muestra, que es de tamaño  $n_{..}$ , y  $\hat{p}$  se obtiene combinando los datos.

Es claro que esta ecuación compara dos proporciones muestrales, de modo que se refiere al modelo estratificado para muestras independientes.

**Tabla 22.5** Datos de cebada en una tabla  $2 \times 2$

	No en dos filas	En dos filas	Totales
Verde	1,178	291	1,469
Clorótica	273	156	429
Totales	1,451	447	1,898

La varianza de  $\hat{p}_1 - \hat{p}_2$  también puede estimarse por  $\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2$ . El interés reciente sobre cuál varianza es mejor en el denominador de  $Z$  ha conducido a Eberhardt y Fligner (22.22), Petkau (22.23), y Conoyer (22.24) a hacer algunas recomendaciones. Parece que la ec. (22.7) es en general satisfactoria, especialmente para tamaños de muestras más pequeños.

**Papel binomial** Usando el papel de probabilidad binomial, pueden compararse dos probabilidades muestrales, tabla A.16 (Ref. 22.15). Para ello,

1. Representar cada par de conteo y el par de conteo total.
2. Sumar las distancias medias, es decir, desde el centro de cada hipotenusa a partir de los pares de conteo hasta la recta que pasa por el par de conteo total.
3. Comparar esta suma con  $1.96\sqrt{2} = 2.77$  ó  $2.57\sqrt{2} = 3.64$  unidades en la escala plena para los niveles del 5 y del 1 por ciento, respectivamente.

**Ejercicio 22.4.1** Di Raimondo (22.2) utilizó dos caldos como control en el experimento mencionado en la sec. 22.3. Eran (1) un caldo simple considerado como patrón y (2) uno con 0.15U de penicilina por milímetro. La tabla siguiente muestra los datos combinados de los experimentos con ratones no tratados.

Tratamiento	Vivos	Muertos	Totales
Patrón	8	12	20
Penicilina	48	62	110
Totales	56	74	130

¿Justifica la comprobación que la hipótesis  $P(\text{vivo})$  es la misma para los dos tratamientos?

¿Qué criterio de prueba se ha empleado y cuál es la probabilidad de encontrar un valor mayor que el observado? Pruébese ahora con el otro criterio. Hallar también la probabilidad de un valor mayor que el observado. ¿Se halló  $\chi^2 = Z^2$ ? ¿Fueron iguales las dos probabilidades de un valor mayor extremo?

**Ejercicio 22.4.2** C.A. Perrotta, de la Universidad de Wisconsin, observó la frecuencia de las visitas de ratones a trampas (pequeñas piezas de madera) previamente tratadas con orina de ratones en comparación con otras que estaban limpias. Los siguientes resultados se obtuvieron en 1954.

La trampa fue	Con orina	Limpia
Visitada	17	3
No visitada.	9	23

Probar la hipótesis nula de que la frecuencia de visita no se afecta con el tratamiento. Para ello, calcular la interacción  $\chi^2$ . Compárese el resultado con el que se ha leído en el papel de probabilidad binomial.

¿Cuál de los modelos propuestos es más apropiado para estos datos?

**Ejercicio 22.4.3** Smith y Nielsen (22.19) han estudiado aislamientos de pasto azul de Kentucky y de una variedad de pasturas. Sus observaciones incluyen los datos que se presentan en la tabla adjunta. ¿Hay evidencia de heterogeneidad en la respuesta al moho?

No.	Características de la pastura	Con moho	Sin moho	Total
1	Tierra baja buena, moderadamente pastoreada	107	289	396
2	Tierra baja buena, moderadamente pastoreada	291	81	372

¿Cuál de los modelos propuestos se ajusta más a los datos?

**Ejercicio 22.4.4** Smith y Nielsen (22.19) observaron los mismos campos para roya y obtuvieron los datos que se dan en la tabla. ¿Hay alguna evidencia de heterogeniedad en la respuesta a la roya?

Campo	Con roya	Sin roya	Total
1	372	24	396
5	330	48	378

¿Cuál de los modelos propuestos se ajusta más a los datos?

## 22.5 "Prueba exacta" de Fisher

Ocasionalmente sólo es posible obtener cantidades limitadas de datos si, por ejemplo, se necesita destruir unidades experimentales costosas o de difícil de consecución para obtener los datos. Cuando en una tabla  $2 \times 2$  los números son pequeños (todos los totales de filas y columnas son inferiores a 15), puede ser mejor calcular probabilidades exactas en vez de confiar en una aproximación. Esta prueba usualmente se lleva a cabo contra alternativas unilaterales para cualquiera de los modelos probabilísticos discutidos, o cuando son fijados todos los totales marginales. Este último tipo puede llamarse *modelo de muestreo por cuotas*. En el cálculo de probabilidades, las únicas tablas consideradas son aquellas con totales marginales, lo mismo que los observados.

Considerar los siguientes datos:

	Tienen	No tiene	Total
Estándar	5	2	7
Tratamiento	3	3	6
Total	8	5	13

Cuando se requiere la prueba de homogeneidad, calculamos la probabilidad de obtener la distribución observada o una distribución más extrema, las más extremas son:

6	1	7
2	4	6
8	5	13

y

7	0	7
1	5	6
8	5	13

Por tanto, requerimos de la suma de las probabilidades asociada con las tres distribuciones dadas. Los totales marginales son los mismos para las tres tablas. La suma de las probabilidades se usará para juzgar la significancia.

La probabilidad asociada con la distribución

$n_{11}$	$n_{12}$	$n_{1\cdot}$
$n_{21}$	$n_{22}$	$n_{2\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot \cdot}$

es:

$$P = \frac{n_{1\cdot}! n_{2\cdot}! n_{\cdot 1}! n_{\cdot 2}!}{n_{11}! n_{12}! n_{21}! n_{22}! n_{\cdot \cdot}!} \quad (22.8)$$

donde  $n_{ij}!$  es definido por

$$n! = n(n - 1) \cdots 1 \quad y \quad 0! = 1 \quad (22.9)$$

Léase  $n!$  como  $n$  factorial.

Las probabilidades para nuestras tres tablas son

$$P = \frac{7! 6! 8! 5!}{5! 2! 3! 3! 13!} = .3263$$

$$P = \frac{7! 6! 8! 5!}{6! 1! 2! 4! 13!} = .0816$$

$$P = \frac{7! 6! 8! 5!}{7! 0! 1! 5! 13!} = .0047$$

La suma de las probabilidades es 0.426. Es claro que para responder la pregunta de significancia era suficiente el cálculo de la primera o segunda probabilidades solamente. En la

práctica, se usa esta aproximación calculando primero la mayor probabilidad individual, y así sucesivamente. Obsérvese que las probabilidades calculadas se refieren a sucesos más extremos en una dirección. Posiblemente éste es el tipo de prueba que se necesita en una comparación de un patrón y un tratamiento, o sea, una prueba unilateral. Si las alternativas son bilaterales, los niveles de significancia del 5 y del 1 por ciento exigen probabilidades de 0.025 y 0.005.

Pearson y Hartley (22.17) hacen observar que esta prueba es casi siempre conservadora, y cuando la prueba es de dos colas, las probabilidades quedan más exactamente aproximadas con  $\chi^2$  corregida. Para calcular esto, remplácese  $(n_{11}n_{22} - n_{12}n_{21})$  por  $(|n_{11}n_{22} - n_{12}n_{21}| - n_{..}/2)$  en el numerador de la ec. (22.6).

Aunque calcular las probabilidades es relativamente simple, Mainland (22.12, 22.13) suministra cómodas tablas que eliminan la necesidad de cálculo, la primera referencia incluye probabilidades exactas para cada término en tamaños de muestras iguales o diferentes hasta de 20 observaciones y una tabulación de las tablas 2 X 2 que indican significancia junto con el nivel exacto de significancia. La segunda referencia incluye los contrastes mínimos requeridos para significancia en una selección de tamaños de muestra hasta de 500. Pearson y Hartley (22.17, tabla 38) también suministran tablas de significancia.

**Ejercicio 22.5.1** Para los datos de esta sección, establecer las posibles tablas restantes para totales marginales fijos. Calcular la probabilidad exacta para cada tabla restante y demostrar que la suma de las probabilidades para todas las tablas posibles es uno. ¿Son simétricas las probabilidades en el conjunto?

**Ejercicio 22.5.2** Probar la hipótesis de homogeneidad para los datos del ejercicio 22.4.2. Usese el procedimiento exacto de la sec. 22.5 y supóngase que las alternativas son bilaterales. Comparar el resultado con el obtenido previamente.

**Ejercicio 22.5.3** Calcular  $\chi^2$  corregida para los datos de esta sección y del ejercicio 22.5.2. Con la tabla A.5, hallar la probabilidad aproximada de obtener un valor  $\chi^2$  mayor que el observado. ¿Cómo se compara esta probabilidad con la prueba exacta frente a alternativas bilaterales?

## 22.6 Muestras no independientes en tablas 2 × 2

Supóngase que se obtienen varios pares de individuos, pareados como mellizos, para evaluar tratamientos. Por ejemplo, podríamos querer remedios para el dolor de cabeza o preventivos para el mareo. Tiene sentido asignar un tratamiento particular a un individuo seleccionado al azar en cada par y al individuo restante, el otro tratamiento. El experimento resultante es semejante a un diseño en bloques al azar. A veces es posible utilizar el mismo individuo para ambos tratamientos, pero en diferentes tiempos. El orden de asignación será aleatorio en este caso.

Aquí, las muestras no son independientes, así que los procedimientos analíticos previos de este capítulo no son aplicables. Sin embargo, si se consideran las observaciones pareadas como observaciones bivariadas, vectores de dos componentes, entonces el problema es fácil de resolver. Los vectores posibles son (1, 1), (1, 0), (0, 1) y (0, 0), donde 1 representa éxito y 0 fracaso, y la primera componente es la observación en el individuo

**Tabla 22.6 Presentación de datos para  $n_{ij}$ , observaciones apareadas dependientes**

		El segundo tratamiento fue		Totales
		E <sup>†</sup>	F	
El primer tratamiento fue	S	1, 1	1, 0	$n_{1\cdot}$
	F	0, 1	0, 0	
Totals		$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{..}$

<sup>†</sup>E = éxito F = fracaso

que recibe el primer tratamiento, y la segunda es en el que recibe el segundo tratamiento. Puede calcularse ahora una covarianza a partir de estas observaciones bivariadas incluyéndola en el denominador de la ec. (22.7), o puede usarse la tabla 22.6 en la forma sencilla que se presenta a continuación.

En el modelo, las probabilidades son las de una observación bivariada aleatoria perteneciente a la  $i, j$ -ésima celda;

$$\sum_{i,j} p_{ij} = 1$$

La hipótesis nula es  $P(\text{éxito para el primer tratamiento}) = P(\text{éxito para el segundo tratamiento})$ , o

$$H_0: p_{1\cdot} = p_{\cdot 1} \quad \text{o} \quad p_{11} + p_{12} = p_{11} + p_{21}$$

Esto se reduce a

$$H_0: p_{12} = p_{21}$$

$$H_1: p_{12} \neq p_{21}$$

En consecuencia, sólo hay que tratar de las entradas  $n_{12}$  y  $n_{21}$ , preguntándose si son iguales dentro del muestreo aleatorio. La ec. (21.5) es un criterio de prueba apropiado con  $r = 1$ , es decir,

$$\chi^2_1 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \quad 1 \text{ gl} \quad (22.10)$$

La distribución binomial exacta puede usarse como alternativa con  $p = 0.5$ . Calcúlense las probabilidades para la división observada  $n_{12} : n_{21}$  y para las divisiones más extremas con  $n = n_{12} + n_{21}$  fijo, y, finalmente, duplíquese esta probabilidad si la prueba es bilateral.

**Ejercicio 22.6.1** Supóngase que se realiza un experimento con 150 parejas de individuos. En un estudio de analgésicos se compara un supuesto analgésico con un placebo, o píldora ficticia, que a menudo se hace de lactosa. Ningún paciente sabe qué clase de píldora recibe.

Supóngase que los resultados muestran que el 40 por ciento obtiene alivio con el placebo y el 70 por ciento usando el analgésico. Estos porcentajes fijan todos los totales marginales. Completar con un valor el número de pares en que ambos individuos se benefician con su tratamiento. ¿Cuántos valores más se pueden completar a elección para estos totales marginales fijos?

Probar  $H_0: p_{11} = p_{22}$  para la tabla que se emplee.

## 22.7 Homogeneidad de muestras de dos celdas

La tabla de dos celdas es probablemente la presentación tabular más común de datos discretos. Es decir, la mayoría de los datos discretos están relacionados con la presencia o ausencia de una característica, o sea, con una *dicotomía*. A menudo se dispone de muchas muestras con información similar y se desea combinarlas para obtener una mejor estimación de la proporción en la población. La combinación es apropiada cuando las muestras son homogéneas. Es importante entonces probar la hipótesis de homogeneidad.

Si bien la ec. (22.2) es apropiada, la ec. (22.11) da una visión diferente.

$$\chi^2 = \frac{\sum \hat{p}_i n_{i1} - \hat{p} n_{..}}{\hat{p}(1 - \hat{p})} \quad (22.11)$$

(El subíndice 1 puede ser remplazado por el subíndice 2.) En este caso,  $\hat{p}_i$  es la estimación de  $p$  para la  $i$ -ésima muestra y  $\hat{p}$  lo es para los datos combinados. Por tanto,  $\hat{p}_i = n_{i1}/n_{..}$ . Al usar esta estimación obtenemos la alternativa

$$\chi^2 = \frac{\sum (n_{ij}^2/n_{..}) - n_j^2/n_{..}}{n_{..} n_{.2}/n_{..}^2} \quad j = 1 \text{ o } 2 \quad (22.12)$$

Ahora podemos observar que el denominador es una suma ponderada de cuadrados de los  $\hat{p}$ , puesto que  $n_{ij}^2/n_{..} = n_{..} (n_{ij}/n_{..})^2$ . Esto relaciona el cálculo con un análisis ponderado de la varianza.

Consideremos el siguiente ejemplo. Smith (22.20) observó en trébol dulce hábitos de crecimiento anual frente a los de crecimiento bienal y su herencia. Examinó 38 progenies segregantes, de las cuales los resultados para las seis primeras figuran en la tabla 22.7. Aplicando la prueba de homogeneidad dada por la ec. (22.12), obtenemos

$$\begin{aligned} \chi^2 &= \frac{\sum (n_{ij}^2/n_{..}) - n_j^2/n_{..}}{n_{..} n_{.2}/n_{..}^2} = \frac{8.711885 - 8.404545}{.157252} \\ &= 1.95 \quad \text{con 5 gl, ns} \end{aligned}$$

Tablas 22.7 Progenies segregantes de trébol dulce

Cultivo	Valores observados			Valores (3:1) esperados			$\chi^2$ (1 gl)
	Anual	Bienal	Total	Anual	Bienal		
4-3	18	6	24	18.00	6.00	0.0000	
4-11	33	7	40	30.00	10.00	1.2000	
4-14	38	12	50	37.50	12.50	0.0267	
4-15	19	5	24	18.00	6.00	0.2222	
4-16	39	7	46	34.50	11.50	2.3478	
4-21	30	6	36	27.00	9.00	1.3333	
Totales	177	43	220	165.00	55.00	5.1300 (6 gl)	

No hay pruebas para concluir que las progenies difieran en la proporción de hábito anual o bienal. La proporción común de plantas de hábito bienal se estimaría como  $\hat{p} = 43/220 = 0.195$  para las progenies segregantes de estos seis cultivos.

En algunos casos, hay bases biológicas serias o de otra índole para formular como hipótesis la razón común. En este caso particular, se plantea la hipótesis nula de una razón 3:1 para hábito anual a bienal.

Cuando se conoce o formula la hipótesis de la proporción de población, la ec. (22.12) se modifica para dar

$$\chi^2 = \frac{\sum_i (n_{ij}^2/n_{i\cdot}) - n_{\cdot j}^2/n_{..}}{p(1-p)} \quad j = 1 \text{ o } 2 \quad (22.13)$$

Aquí  $p$  es la proporción de población, y  $p = \frac{1}{4}$  o  $\frac{3}{4}$ . Para nuestro ejemplo,

$$\chi^2 = \frac{8.711885 - 8.404545}{.25(.75)} = 1.64 \quad \text{con 5 gl}$$

Las ecuaciones (22.12) y (22.13) tienen el mismo denominador. La ec. (22.13) da un  $\chi^2$  menor que la ec. (22.12) siempre que la proporción de la población esté más cerca de 0.5 que es la proporción observada.

**Ejercicio 22.7.1** Mendel (22.14), en su estudio genético clásico, observó la variación de planta a planta. En un serie de experimentos sobre la forma de la semilla, las primeras diez plantas dieron los siguientes resultados:

Plantas	1	2	3	4	5	6	7	8	9	10
Semilla redonda	45	27	24	19	32	26	88	22	28	25
Semilla angular	12	8	7	10	11	6	24	10	6	7

Probar la hipótesis nula  $H_0$ : 3 redonda; 1 angular, usando totales. Probar la homogeneidad de la razón 3:1 para las 10 plantas, en el supuesto de que la razón 3:1 es la verdadera. ¿Cuál es el valor de la homogeneidad  $\chi^2$  cuando no se supone la razón 3:1?

**Ejercicio 22.7.2** Mendel (22.14) también observó la variación de planta a planta en un experimento sobre el color del albumen en las semillas. Los resultados para las primeras diez plantas fueron:

Planta	1	2	3	4	5	6	7	8	9	10
Albumen amarillo	25	32	14	70	24	20	32	44	50	44
Albumen verde	11	7	5	27	13	6	13	9	14	18

Repetir el ejercicio 22.7.1 para este conjunto de datos.

*Nota:* Eisenhart (22.3) ofrece un comentario breve pero interesante sobre Mendel, con una presentación de sus descubrimientos y la acogida que tuvieron.

## 22.8 Aditividad de $\chi^2$

La información de datos de tablas como la 22.7 no queda completamente agotada por una prueba de homogeneidad  $\chi^2$  cuando el investigador tiene una hipótesis acerca de la proporción de población. Por ejemplo, puede calcularse un  $\chi^2$  para probar la hipótesis nula para cada cultivo (fila). En la tabla 22.7, éstos se muestran en la última columna. Ningún  $\chi^2$  individual es significante.

Los  $\chi^2$  independientes pueden sumarse. Obtenemos un  $\chi^2$  total basado en 6 grados de libertad. El valor es 5.1300 y no significante. (Sólo pueden sumarse valores de  $\chi^2$  sin ajustar).

También podemos calcular un  $\chi^2$  con base en los totales 177 y 43. En este caso,  $\chi^2 = 3.4909$  con 1 grado de libertad y no es significante, aunque está próximo al punto 5 por ciento;  $\chi_{0.05}^2 = 3.841$  para un grado de libertad.

Finalmente, la diferencia entre la suma de los  $\chi^2$  y el  $\chi^2$  global mide la interacción. Por tanto,  $5.1300 - 3.4909 = 1.6391$  con 5 grados de libertad en comparación con 1.64 por cálculo directo. Esto está lejos de ser significante; alrededor de los dos tercios de la suma de los  $\chi^2$ , están asociados con una prueba de la razón de las observaciones totales.

Ahora consideremos los  $\chi^2$  calculados para este ejemplo.

**1.  $\chi^2$  individuales** Cada uno da información acerca de una muestra particular. Cuando cada muestra es pequeña, será difícil detectar una desviación respecto de la hipótesis nula a menos que sea considerable. También, si hay muchas muestras y la hipótesis nula es verdadera, esperamos el valor ocasional, como 1 en 20, de presentar significancia falsamente.

**2. La suma de los  $\chi^2$  individuales** Aquí tenemos una información combinada de las muestras. Por ejemplo, supóngase que la población verdadera fuera tal que  $\chi^2 = 2.000$ , en promedio; encontramos  $0.20 > P > 0.10$ . Los  $\chi^2$  individuales estarían distribuidos en torno a este valor, algunos mayores y otros más pequeños. Algunos  $\chi^2$  serían significantes, otros no. Esto hace difícil evaluar la información, a menos que sumemos los  $\chi^2$ . Si tenemos 20

muestras, entonces la suma debería ser alrededor de  $20(2,0000) = 40.000$  con 20 grados de libertad; esto es altamente significante.

Una dificultad obvia se presenta en la interpretación de una suma de  $\chi^2$ . Supóngase que las razones de población difieren de una muestra a otra, respecto de la hipótesis nula. Esto tiende a hacer que todos los  $\chi^2$  individuales sean demasiado grandes. Una suma significante puede atribuirse a más de una hipótesis alternativa, es decir, muestras heterogéneas.

**3.  $\chi^2$  sobre totales** En este caso combinamos las muestras y calculamos  $\chi^2$  con un grado de libertad. Si las muestras son homogéneas pero la hipótesis nula es falsa, entonces tenemos una muestra grande y estamos en capacidad de detectar una desviación pequeña respecto de la hipótesis nula. Por ejemplo, supóngase que se muestrea de una población en la cual la hipótesis es 3:1 y que observamos individuos en razón 17:3. Aquí  $\chi^2 = 1.07$  y no es significante. Sin embargo, si tenemos cuatro veces tantos individuos como los indicados y observamos exactamente la misma razón, es decir, si observamos  $(17 \times 4) : (3 \times 4) = 68:12$ ; entonces  $\chi^2 = 4.27$  y es significante.

Si las muestras son homogéneas, algunas con razones mayores y otras con razones menores que la de la hipótesis, entonces la combinación de muestras puede conducir a un  $\chi^2$  bajo y a la aceptación de la hipótesis nula. Por ejemplo, si 3:1 es la razón de la hipótesis nula, entonces una muestra de 67:13 da  $\chi^2 = 3.27$ , cerca al 5 por ciento tabulado. También, una muestra de 53:27 tiene  $\chi^2 = 3.27$ . Esta muestra se aparta de la hipótesis nula en otra dirección. Al combinar las muestras, tenemos una razón observada de 120:40, que se ajusta perfectamente a la hipótesis nula.

**4.  $\chi^2$  de homogeneidad** Ahora se pone de manifiesto la importancia de la homogeneidad; es una medida de la semejanza o desemejanza de las muestras. Es independiente de la razón de la hipótesis nula ya que las muestras pueden diferir de esta razón sin que aumente la heterogeneidad  $\chi^2$ , siempre y cuando difieran en la misma dirección y aproximadamente en el mismo grado.

En resumen,  $\chi^2$  en totales y  $\chi^2$  de homogeneidad constituyen un análisis razonablemente adecuado de los datos. Si la homogeneidad  $\chi^2$  es significante, entonces puede ser prudente considerar los  $\chi^2$  individuales. La suma de los  $\chi^2$  individuales puede ser difícil de interpretar en sí misma.

**Ejercicio 22.8.1** Para los datos del ejercicio 22.7.1, calcular  $\chi^2$  de cada planta para probar  $H_0$ : la razón 3:1. Súmense estos  $\chi^2$  y demuéstrese que la suma es igual a la suma de los dos  $\chi^2$  calculados en el ejercicio 22.7.1. Obsérvese también que los grados de libertad son aditivos.

**Ejercicio 22.8.2** Repítase el ejercicio 22.8.1 con los datos del ejercicio 22.7.2.

**Ejercicio 22.8.3** Discutir los resultados obtenidos en los ejercicios 22.7.1 y 22.8.1; 22.7.2 y 22.8.2.

## 22.9 Más sobre la aditividad de $\chi^2$

En la tabla 22.4 los datos son homogéneos;  $\chi^2$  (3 gl) = 2.63. Los datos del ejercicio 22.4.1 también son homogéneos;  $\chi^2$  (1 gl) = 0.0913. En consecuencia, parece apropiado combi-

**Tabla 22.8 Datos combinados de la tabla 22.4 y del ejercicio 22.4.1**

Tratamiento	Vivos	Muertos	Total
Vitaminas	41	129	170
Controles	56	74	130
Total	97	203	300
$\chi^2 = \frac{[129(56) - 41(74)]^2}{97(203)170(130)} = 12.10^{**} \text{ con } 1 \text{ gl}$			

nar los datos en cada tabla y verificar si los datos del tratamiento y control son homogéneos. En la tabla 22.8 se tienen los datos combinados.

La prueba de la tabla 22.8 es contraria a una probabilidad común de muerte de ratones no tratados inyectados con *Staphylococcus aureus* cultivados en los dos conjuntos de condiciones.

El manejo de los datos de Di Raimondo hasta aquí debe sugerir aditividad de  $\chi^2$ . Así que preguntamos: “¿Es  $\chi^2$  (en vitaminas) +  $\chi^2$  (en controles) +  $\chi^2$  (vitaminas frente a controles) =  $\chi^2$  (vitaminas y controles)?” La respuesta es no. En este caso, la suma de los  $\chi^2$  es  $2.63 + 0.09 + 12.10 = 14.82$  con  $3 + 1 + 1 = 5$  gl;  $\chi^2$  (vitaminas y controles) = 14.41 con 5 grados de libertad (ver tabla 22.9). La diferencia entre los dos  $\chi^2$ , cada uno con 5 grados de libertad, es pequeña, pero no se debe a errores de aproximación.

Si las tres comparaciones, en vitaminas, en controles, entre vitaminas y controles, se hubieran hecho en un análisis de la varianza, entonces habrían sido claramente independientes y la suma de sus sumas de cuadrados habría sido la misma que la entre tratamientos. Comprendidos vitaminas y controles. No sucede lo mismo con datos discretos y valores de  $\chi^2$ .

Probablemente pocos objetarían las tres comparaciones no independientes hechas si el investigador las considerara con sentido. Sin embargo, pueden hacerse comparaciones independientes, que vamos a ilustrar a continuación. Se indicarán situaciones en las que tales comparaciones tienen sentido y, debido a su independencia, se han de preferir. La ilustración que damos parecerá algo artificial.

Consideremos los datos de la tabla 22.9, para los cuales ahora se harán comparaciones independientes; éstas y sus cálculos se deben a Irwin (22.7), Lancaster (22.10), (22.11) y Kimball (22.9).  $\chi^2 = 14.41^{**}$  con 5 grados de libertad se obtuvo a partir de la ec. (22.2). Este valor será repartido en  $\chi^2$  independientes con un solo grado de libertad cada una. Empezamos aplicando la ec. (22.14) muy similar a la ec. (22.6), pero, si se la examina cuidadosamente, se encontrará que difiere tanto en el numerador como en el denominador. Las ecuaciones restantes son modificaciones obvias de la ec. (22.14). Su aplicación da:

$$\begin{aligned}\chi_1^2 &= \frac{n_{\cdot\cdot}(n_{11}n_{22} - n_{12}n_{21})^2}{n_{\cdot 1}n_{\cdot 2}n_{1\cdot}n_{2\cdot}(n_{1\cdot} + n_{2\cdot})} \\ &= \frac{300^2[8(62) - 12(48)]^2}{97(203)20(110)130} = .1023, ns\end{aligned}\quad (22.14)$$

Tabla 22.9 Datos de la tabla 22.4 y el ejercicio 22.4.1

Tratamiento	Vivos	Muertos	Total
Patrón	8	12	20
Penicilina	48	62	110
NA	10	30	40
AF	9	31	40
Paba	9	41	50
B <sub>6</sub>	13	27	40
Total	97	203	300

$\chi^2 = 14.41^{**}$  con 5 gl

$$\chi_2^2 = \frac{n_{\cdot\cdot}^2[(n_{11} + n_{21})n_{32} - (n_{12} + n_{22})n_{31}]^2}{n_{\cdot 1}n_{\cdot 2}n_{\cdot 3}(n_{1\cdot} + n_{2\cdot})(n_{1\cdot} + n_{2\cdot} + n_{3\cdot})} \quad (22.15)$$

$$= \frac{300^2[56(30) - 74(10)]^2}{97(203)40(130)170} = 4.5686^*$$

$$\chi_3^2 = \frac{n_{\cdot\cdot}^2[(n_{11} + n_{21} + n_{31})n_{42} - (n_{12} + n_{22} + n_{32})n_{41}]^2}{n_{\cdot 1}n_{\cdot 2}n_{\cdot 4}(n_{1\cdot} + n_{2\cdot} + n_{3\cdot})(n_{1\cdot} + n_{2\cdot} + n_{3\cdot} + n_{4\cdot})} \quad (22.16)$$

$$= \frac{300^2[66(31) - 104(9)]^2}{97(203)40(170)210} = 3.9436^*$$

$$\chi_4^2 = \frac{n_{\cdot\cdot}^2[(n_{11} + n_{21} + n_{31} + n_{41})n_{52} - (n_{12} + n_{22} + n_{32} + n_{42})n_{51}]^2}{n_{\cdot 1}n_{\cdot 2}n_{\cdot 5}(n_{1\cdot} + n_{2\cdot} + n_{3\cdot} + n_{4\cdot})(n_{1\cdot} + \dots + n_{5\cdot})} \quad (22.17)$$

$$= \frac{300^2[75(41) - 135(9)]^2}{97(203)50(210)260} = 5.7921^*$$

$$\chi_5^2 = \frac{n_{\cdot\cdot}^2[(n_{11} + \dots + n_{51})n_{62} - (n_{12} + \dots + n_{52})n_{61}]^2}{n_{\cdot 1}n_{\cdot 2}n_{\cdot 6}(n_{1\cdot} + \dots + n_{5\cdot})n_{\cdot\cdot}} \quad (22.18)$$

$$= \frac{300^2[84(27) - 176(13)]^2}{97(203)40(260)300} = .0006$$

La suma de estos  $\chi^2$  es 14.41 con 5 grados de libertad, igual que para la tabla total.

Las fórmulas generales para  $\chi^2$  independientes en toda tabla  $r = 2$  son bien claras de acuerdo con los casos particulares presentados.

Al revisar esta ilustración, vemos que  $\chi^2_1$  tiene sentido por cuanto compara dos patrones. También,  $\chi^2_2$  tiene sentido por cuanto comparamos el promedio de las respuestas a dos patrones, para los cuales no hay prueba de una respuesta diferente, con la respuesta a un tratamiento. Como la significancia se halla a este nivel, tiene poco sentido promediar las respuestas de tres tratamientos, uno de los cuales es ciertamente diferente de los otros dos, para comparación con la respuesta a un cuarto tratamiento. Por esta razón, nos hemos referido a nuestra ilustración como artificial.

El uso de  $\chi^2$  independientes, como los calculados aquí, parece tener aplicación real en el caso de dos tratamientos y un patrón o dos patrones y un tratamiento. Si los dos tratamientos o los controles parecen diferir, es dudosa la utilidad de promediarlos y compararlos con las restantes entradas. Ya encontrada la significancia, se podría proceder a efectuar las pruebas de no independencia que parezcan más provechosas.

**Ejercicio 22.9.1** ¿En qué forma difieren las ecs. (22.6) y (22.14)?

**Ejercicio 22.9.2** A los datos de moho del ejercicio 22.4.3, añádanse los siguientes:

No.	Tipo de paso	Con moho	Sin moho	Total
9	Buena tierra alta, moderadamente pastoreada	280	144	424

A los datos de roya del ejercicio 22.4.4, añádanse:

Campo	Con roya	Sin roya	Total
9	371	54	425

Si se considera que los métodos de esta sección se aplican a una tabla de contingencia  $3 \times 2$ , procédase a aplicarlos. En cualquier caso, justifíquese la posición.

## 22.10 Regresión lineal, tablas $r \times 2$

Cochran (22.1) da un método para determinar la regresión lineal en tablas  $r \times 2$ , donde las filas están en orden natural. Se asignan puntuaciones  $Z_i$  a las filas para tratar de convertirlas a valores en una escala continua. El coeficiente de regresión  $b$  de la estimación de la probabilidad  $\hat{p}_i$  en regresión de fila respecto de  $Z_i$  es una regresión ponderada dada por la ec. (22.19); a  $\hat{p}_i$  se asignan  $n_{i\cdot}/\hat{p}(1 - \hat{p})$  ponderados,

donde  $\hat{p} = \sum_i n_{i2} / \sum_i n_{i\cdot}$ .

$$b = \frac{\sum_i n_{i\cdot}(\hat{p}_i - \hat{p})(Z_i - \bar{Z}_w)}{\sum_i n_{i\cdot}(Z_i - \bar{Z}_w)^2} \quad \text{definición de la fórmula} \quad (22.19)$$

$\bar{Z}_w$  es la media ponderada de  $Z_i$  definida por

$$\bar{Z}_w = \sum_i n_i Z_i / \sum_i n_i$$

El criterio para probar  $b$  está dado por la ec. (22.20), si bien se distribuye sólo aproximadamente como  $\chi^2$ , y tiene un grado de libertad.

$$\chi^2 = \frac{\left[ \sum_i n_i (\hat{p}_i - \bar{p})(Z_i - \bar{Z}_w) \right]^2}{\bar{p}(1 - \bar{p}) \sum_i n_i (Z_i - \bar{Z}_w)^2} \quad (22.20)$$

En un experimento realizado por R.F. Patton de la Universidad de Wisconsin con pino blanco, se comparó el efecto de la edad de árboles padres, de los cuales habían tomado cortes, sobre la susceptibilidad de injertos a la roya, como se muestra en la tabla 22.10. Todos los injertos fueron hechos en transplantes de 4 años de edad.

Calcular  $\chi^2$  por la ec. (22.21), fórmula de operación de la ec. (22.20)

$$\begin{aligned} \chi^2 &= \frac{\left[ \sum_i n_{i2} Z_i - n_{..} (\sum_i n_i Z_i / n_{..}) \right]^2}{\left[ \sum_i n_i Z_i^2 - \left( \sum_i n_i Z_i \right)^2 / n_{..} \right] (\bar{p})(1 - \bar{p})} \\ &= \frac{[14(1) + \dots + 8(8) - (38)(303)/77]^2}{[21(1) + \dots + 184(8) - 303^2/77](.4935)(.5065)} \\ &= \frac{29.53^2}{156.17} = 5.58^* \quad \text{con 1 gl} \end{aligned} \quad (22.21)$$

Tabla 22.10 Edad de árbol padre y reacción de injertos a la roya

Edad del árbol padre, en años	Escala $Z_i$	Sano $n_{i1}$	Enfermo $n_{i2}$	Total $n_i$	$\hat{p}_i = n_{i2}/n_i$ porcentaje	$n_i Z_i$
4	1	7	14	21	67 33%	21
10	2	6	11	17	65 38%	34
20	4	11	5	16	31 19%	64
40 y más	8	15	8	23	35 24%	184
Total		39	38	77	0.4935	303
		$= n_{..1}$	$= n_{..2}$	$= n_{..}$	$= \bar{p}$	

Fuente: Datos usados con permiso de R.F. Patton, Universidad de Wisconsin, Madison, Wisconsin.

El  $\chi^2$  total, calculado por la ec. (22.12), puede ser particionado como sigue:

	gl	$\chi^2$
Regresión de $\hat{p}_i$ respecto de $Z_i$	1	5.58*
Desviaciones respecto de la regresión	2	2.58
Total	3	8.16*

## 22.11 Tamaño de la muestra en tablas $2 \times 2$

El problema del tamaño de la muestra en tablas de dos celdas se estudia en las secs. 21.5 y 21.6. Ahora volvemos nuestra atención al problema del tamaño de la muestra necesario para detectar diferencias en tablas  $2 \times 2$ .

Considérense los datos de la tabla 22.8. El tratamiento "vitaminas" se compone de un conjunto homogéneo de tratamientos; el tratamiento "controles" se compone de un par homogéneo. Es concebible que un investigador desee experimentar además con uno de los tratamientos y uno de los patrones, o continuar con un experimento diferente pero similar donde la única información relacionada sea proporcionada por este experimento. Como los controles dan una mayor proporción de "vivos" es razonable suponer que éste continuará siendo el caso y probar sólo para un conjunto de alternativas unilaterales.

La ecuación (22.22) ha sido dada por Paulson y Wallis (22.16) para determinar el tamaño de la muestra.

$$n = 1,641.6 \left( \frac{Z_\alpha + Z_\beta}{\text{arcsen} \sqrt{p_S} - \text{arcsen} \sqrt{p_E}} \right)^2 \quad (22.22)$$

En esta ecuación, los ángulos (arcsen significa "el ángulo cuyo seno es") están dados en grados como en la tabla A.10,  $n$  es el número de observaciones en cada muestra,  $Z_\alpha$  es la desviación normal tal que  $P(Z \geq Z_\alpha) = \alpha$ ,  $Z_\beta$  es la desviación normal tal que  $P(Z \geq Z_\beta) = \beta$ ,  $p_S$  es la probabilidad de población o proporción asociada con el tratamiento patrón o de control, y  $p_E$  lo es para el tratamiento experimental. Las cantidades  $Z_\alpha$  y  $Z_\beta$  se obtienen de la tabla A.4. Puesto que  $p_S$  y  $p_E$  nunca están disponibles, deben ser calculadas, esto se aplica particularmente a  $p_S$ , mientras que  $p_E$  puede ser elegido en forma bastante arbitraria. Una estimación razonablemente buena no afectaría seriamente el resultado, a menos que por lo menos una de las  $p$  fuera muy pequeña o muy grande. El procedimiento de prueba que supone la ec. (22.22) es  $\chi^2$  frente a alternativas unilaterales y ajustada por continuidad; la hipótesis nula es la de la de homogeneidad. Si la hipótesis nula es verdadera, entonces será rechazada con probabilidad  $\alpha$ ; si es verdadera la alternativa que  $p_S$  es tan grande como lo establecido, entonces será rechazada con probabilidad  $\beta$ . Si la verdadera  $p_S$  se acerca más a  $p_E$  que lo establecido, entonces dejaríamos de detectar este hecho más del  $100\beta$  por ciento de las veces; si la verdadera  $p_S$  difiere de  $p_E$  más que lo establecido, entonces dejaríamos de detectar este factor menos del  $100\beta$  por ciento de las

veces, es decir, lo detectamos con probabilidad mayor que  $1 - \beta$ . (Si en lugar de grados se usan radianes para los ángulos, remplácese 1,641.6 por 0.5).

Ahora ilustramos el uso de la ec. (22.22). Supóngase que planeamos efectuar un experimento muy semejante al resumido en la tabla 22.8, pero con solo un patrón o control y un tratamiento. Si la hipótesis nula de homogeneidad es verdadera, estamos dispuestos a rechazarla sólo el 1 por ciento de las veces; por tanto  $\alpha = 0.01$ . No conocemos  $p_S$  pero puede estimarse por la tabla 22.8 como  $\hat{p}_S = 56/130 = 0.43$ . Deseamos detectar  $p_E$  cuando se acerca a la magnitud observada en la tabla 22.8, o sea,  $p_E = 41/170 = 0.24$ , con probabilidad 0.75. Es decir que  $1 - \beta = 0.75$  y  $\beta = 0.25$ .

Usando la ec. (22.22), obtenemos

$$n = 1,641.6 \left( \frac{2.327 + 0.675}{40.98 - 29.33} \right)^2 = 109 \text{ ratones por tratamiento}$$

Si entra un solo método en el experimento, es decir, si debemos decidir cuál de dos razones teóricas aplicar para una muestra única, entonces la ec. (22.22) puede usarse, pero obtenemos el doble del tamaño de muestra requerido.

En la referencia (22.16) hay un monograma basado en la ec. (22.22) que puede usarse para determinar el tamaño de la muestra. Lo único que se necesita además es una regla. Kermack y McKendrick (22.8) también dan solución para este problema, incluso una tabla de valores de  $n$ .

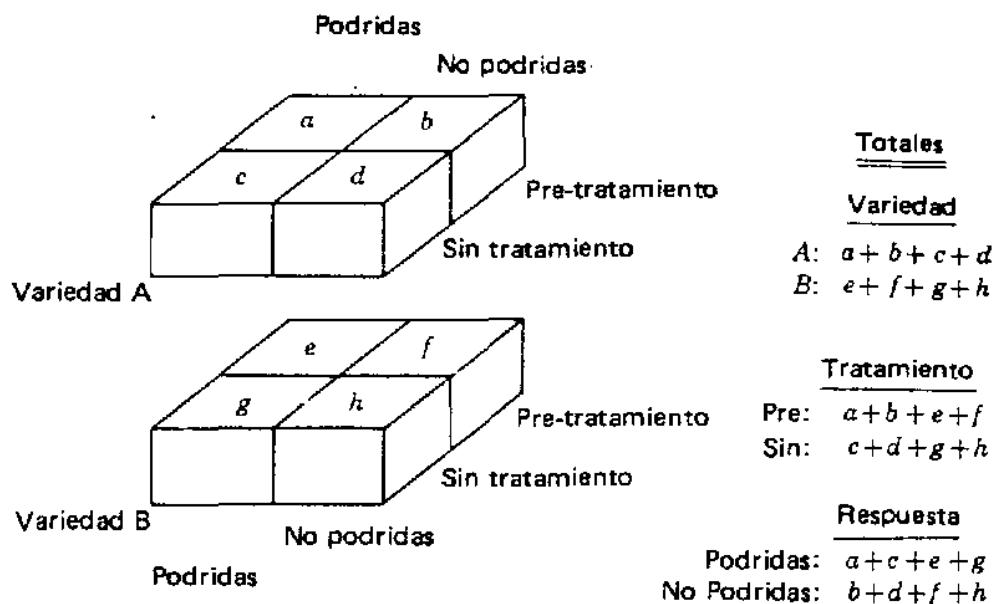
**Ejercicio 22.11.1** Supóngase que un procedimiento patrón da una respuesta de cerca del 65 por ciento de exterminio en una situación binomial, pero que la respuesta es lo suficientemente variable para que se incluya el patrón en un experimento. Se decidió considerar un tratamiento propuesto con el cual se sostiene que hay una mortalidad del 80 por ciento.  $Z(\text{o } \chi^2)$  va a ser el criterio de prueba unilateral con un nivel de significancia del 5 por ciento; se desea detectar en la alternativa un 80 por ciento de mortalidad con una probabilidad de 0.90. ¿Cuál es el tamaño de muestra necesario para cada tratamiento?

## 22.12 Clasificación de $n$ vías

En tablas de más de dos dimensiones los datos binomiales presentan problemas estadísticos de interpretación. En la sec. 23.5, se expone el uso de una transformación para algunos de tales problemas. Por ahora, consideremos solamente la tabla  $2 \times 2 \times 2$

En la figura 22.1, las letras podrían ser números de semillas germinadas que se puden por el pie o que no se pudren (la respuesta) según variedad y tratamiento. Podemos formular como hipótesis que cualquiera que sea la interacción de tratamiento por respuesta, difiere de una variedad a otra en no más de la variación aleatoria. Al probar esta hipótesis, se prueba una *interacción de tres factores o de segundo orden*.

Sean  $p_1, \dots, p_8$  las verdaderas frecuencias relativas o probabilidades, generalmente desconocidas, de que una observación esté en las celdas con entradas  $a, \dots, h$ , respectivamente. Si no hay interacción de tratamientos por respuesta para la variedad  $A$ , entonces  $p_1/p_2 = p_3/p_4$  y  $p_1 p_4 = p_2 p_3$ . Así mismo, para la variedad  $B$ ,  $p_5/p_6 = p_7/p_8$  y  $p_5 p_8 = p_6 p_7$ . Si hay interacción, las igualdades no se cumplen. En cualquier caso, las razones  $p_1 p_4 / p_2 p_3$  y  $p_5 p_8 / p_6 p_7$  están directamente relacionadas con la interacción en cada tabla.

Figura 22.1 Representación de una tabla  $2 \times 2 \times 2$ 

La hipótesis nula para una interacción de tres factores es que la interacción de tratamiento por respuesta, se presente o no, es igual para las dos variaciones. Esta es equivalente a la hipótesis  $p_1 p_4 / p_2 p_3 = p_5 p_8 / p_6 p_7$ , o

$$H_0: p_1 p_4 p_6 p_7 = p_2 p_3 p_5 p_8 \quad (22.23)$$

Esta hipótesis supone un solo grado de libertad y, por tanto, una sola desviación. Esta puede obtenerse despejando la ecuación cúbica

$$(a + X)(d + X)(f + X)(g + X) = (b - X)(c - X)(e - X)(h - X) \quad (22.24)$$

El criterio de prueba para  $H_0$  está dado por

$$\chi^2 = X^2 \sum_i \frac{1}{E_i} \quad (22.25)$$

$E_i$  es el valor esperado para la  $i$ -ésima celda.

*Ejemplo* Galton obtuvo los datos de la tabla 22.11 a partir de 78 familias. Los hijos fueron clasificados como de ojos claros o no, según que tuvieran un progenitor o no de ojos claros y de acuerdo a si tenían o no un abuelo de ojos claros.

La ecuación cúbica es

$$(1,928 + X)(395 + X)(508 + X)(225 + X) \\ = (552 - X)(303 - X)(596 - X)(501 - X)$$

**Tabla 22.11 Número de hijos de acuerdo con una característica de ojos claros en hijo, padre y abuelo**

Abuelo		Claros		No	
Padre		Claros	No.	Claros	No.
Hijo	Claros	1,928	552	596	508
	No.	303	395	225	501

que se transforma en

$$5,008X^3 + 1,174,832X^2 + 1,262,521,792X + 37,104,335,424 = 0$$

La solución para esta ecuación puede obtenerse por tanteo. Recomendamos como valores iniciales de tanteo múltiplos de 10. En este caso, deben ser evidentemente múltiplos negativos. Una vez obtenidos dos valores tales que el primer miembro de la ecuación cambie de signo, la localización de la solución queda razonablemente bien establecida. La solución final supondrá muy pocos valores más de tanteo. Para nuestra ecuación, la solución es  $X = -30.1$ . De la ec. (22.25),

$$\chi^2 = (-30.1)^2 \left( \frac{1}{1,928 - 30.1} + \cdots + \frac{1}{501 + 30.1} \right) = 16.93 \quad \text{con 1 gl}$$

Los datos dejan poca duda acerca de una interacción de segundo orden. O bien, las dos interacciones de primer orden no son homogéneas. La prueba es simétrica en cuanto una comparación de todo par de interacciones de primer orden, es decir, las dos variedades por respuesta o las dos variedades por tratamiento o las dos respuestas por tratamiento conducen a la misma hipótesis nula, o sea, la ec. (22.23).

## Referencias

- 22.1. Cochran, W. G.: "Some methods for strengthening the common  $\chi^2$  tests," *Biom.*, 10:417-451 (1954).
- 22.2. Di Raimondo, F.: "In vitro and in vivo antagonism between vitamins and antibiotics," *Int. Rev. Vitamin Res.*, 23:1-12 (1951).
- 22.3. Eisenhart, C.: "Anniversaries in 1965 of interest to statisticians," *Amer. Statist.*, 19:21-29 (1965).
- 22.4. Federer, W. T.: "Variability of certain seed, seedling, and young-plant characters of guayule," *U.S. Dept. Agr. Tech. Bull.*, 919, 1946.
- 22.5. Green, J. M.: "Inheritance of combining ability in maize hybrids," *J. Amer. Soc. Agron.*, 40: 58-63 (1948).

- 22.6. Grizzle, J.: "Continuity correction in the  $\chi^2$ -test for  $2 \times 2$  tables," *The Amer. Statist.*, 21(4): 28-32 (1967).
- 22.7. Irwin, J. O.: "A note on the subdivision of  $\chi^2$  into components," *Biometrika*, 36:130-134 (1949).
- 22.8. Kermack, W. O., y A. G. McKendrick: "The design and interpretation of experiments based on a four-fold table: The statistical assessment of the effects of treatment," *Proc. Roy. Soc. Edinburgh*, 60:362-375 (1940).
- 22.9. Kimball, A. W.: "Short-cut formulas for the exact partition of  $\chi^2$  in contingency tables," *Biom.*, 10:452-458 (1954).
- 22.10. Lancaster, H.O.: "The derivation and partition of  $\chi^2$  in certain discrete distributions," *Biometrika*, 36:117-129 (1949).
- 22.11. Lancaster, H. O.: "The exact partition of  $\chi^2$  and its application to the problem of polling of small expectations," *Biometrika*, 37:267-270 (1950).
- 22.12. Mainland, D.: "Statistical methods in medical research. I. Qualitative Statistics (Enumeration Data)," *Can. J. Res., sect. E. Med. Sci.*, 26:1-166 (1948).
- 22.13. Mainland, D. L. Herrera, y M. I. Sutcliffe: *Tables for Use with Binomial Samples*, published by the authors, Nueva York, 1956.
- 22.14. Mendel, G.: *Versuche über Pflanzen Hybriden*, 1866. Versión del inglés de Harvard University Press, Cambridge, Mass, 1948.
- 22.15. Mosteller, F., y J. W. Tukey: "The uses and usefulness of binomial probability paper," *J. Amer. Statist. Ass.*, 44:174-212 (1949).
- 22.16. Paulson, E., y W. A. Wallis: Chap. 7 en C. Eisenhart, M. W. Hastay, y W. A. Wallis (eds.), *Techniques of Statistical Analysis*, McGraw-Hill, Nueva York, 1947.
- 22.17. Pearson, E. S., y H. O. Hartley (eds.): *Biometrika Tables for Statisticians*, vol. 1, Cambridge University Press, Nueva York, 1954.
- 22.18. Rinke, E. H.: "Inheritance of coumarin in sweet clover," *J. Amer. Soc. Agron.*, 37:635-642 (1945).
- 22.19. Smith, D. C., y E. L. Nielsen: "Comparisons of clonal isolations of *Poa pratensis L.*, etc." *Agron. J.*, 43:214-218 (1951).
- 22.20. Smith, H. B.: "Annual versus biennial growth habit and its inheritance in *Melilotus alba*," *Amer. J. Bot.*, 14:129-146 (1927).
- 22.21. Steel, R. G. D., y J. H. Torrie: *Principles and Procedures of Statistics*, 1st ed., McGraw-Hill, Nueva York, 1960.
- 22.22. Eberhardt, K. R., y M. A. Fligner: "A comparison of two tests for equality of two proportions," *Amer. Statist.*, 31(4):151-155 (1977).
- 22.23. Petkau, J.: "A fundamental question of practical statistics," carta al editor, *Amer. Statist.*, 32(3):114 (1978).
- 22.24. Conover, W. J': "A fundamental question of practical statistics," carta al editor, *Amer. Statist.*, 32(3):114 (1978).
- 22.25. Weir, B. S., R. W. Allard, y A. L. Kahler: "Analysis of complex allozyme polymorphisms in a barley population," *Genetics*, 72:505-523 (1972).

---

## CAPITULO VEINTITRES

---

### ALGUNAS DISTRIBUCIONES DISCRETAS

#### 23.1 Introducción

En los capítulos 21 y 22 tratamos datos enumerativos, que surgen de distribuciones discretas. Sin embargo, al hablar de ellos, pocas veces comentamos el tipo de distribución discreta, a menos que hubiéramos dicho específicamente que la base era una distribución binomial o multinomial y deseado probar una hipótesis nula relativa a razones de población.

En este capítulo, se hace una breve exposición sobre varias distribuciones discretas útiles. La exposición incluye algunos de los usos y pruebas asociadas con las distribuciones.

#### 23.2 La distribución hipergeométrica

Supóngase que tenemos una población finita de 25 pájaros amenazada de extinción. De éstos, 20 son adultos y 5 polluelos. Si sacamos uno al azar, entonces la probabilidad de que sea adulto es  $20/25$  y  $5/25$  de que sea polluelo. Si sacamos un adulto, entonces las probabilidades al sacar los próximos son  $19/24$  y  $5/24$ ; si sacamos un polluelo, las probabilidades son  $20/24$  y  $4/24$ . En cualquiera de los dos casos, ya no son  $20/25$  ni  $5/25$  porque el muestreo ha sido *sin reemplazo*. Este es completamente diferente de un problema binomial donde la probabilidad asociada a un evento específico es constante de un ensayo a otro. En este caso, la ocurrencia de un evento no afecta la ocurrencia de ningún otro. En el lanzamiento de monedas, se aplica la distribución binomial, así que se define esta propiedad de independencia de eventos diciendo que una moneda no tiene memoria.

El problema de la población de pájaros sirve para introducir la *distribución hipergeométrica*. Un problema hipergeométrico más general puede plantearse como sigue: dados  $N$  elementos entre ellos  $N_1$  con propiedad  $A$  y  $N - N_1$  con propiedad no  $A$ , ¿cuál es la probabilidad de que una muestra de  $n$  elementos extraída sin remplazo, contenga  $n_1$

elementos con la propiedad  $A$  y  $n - n_1$  con la propiedad no  $A$ ? La probabilidad asociada con este evento es

$$P(n_1) = \frac{\binom{N_1}{n_1} \binom{N - N_1}{n - n_1}}{\binom{N}{n}} \quad (23.1)$$

donde

$$\binom{N_1}{n_1} = \frac{N_1!}{n_1!(N_1 - n_1)!} \quad (23.2)$$

para  $N_1!$  definido por la ec. (22.9). El sistema de probabilidades se llama *distribución hipergeométrica*. Obsérvese que hay que decir que el muestreo es sin reemplazo, es decir, que una vez muestreado un elemento, éste es retirado permanentemente de la población. Si el muestreo fuera con reemplazo, su retorno a la población resultaría en probabilidades independientes asociadas con cada extracción, y tendríamos una población binomial.

Con la distribución hipergeométrica están relacionados dos problemas estadísticos comunes: (1) inspección de calidad o muestreo de aceptación y (2) problemas de recaptura de identificaciones o de recaptura de marcas.

**Inspección de calidad o muestreo de aceptación** Aquí el problema es estimar la población de defectuosos. Un defectuoso puede ser un recipiente de levadura seca con agujereado que debería ser hermético, una mano de bridge sin una figura, o un loro con fiebre en el zoológico. Aquí el tamaño de la población es conocido, y hay que estimar el número de defectuosos. Si una muestra de  $n$  observaciones contiene  $Y$  defectuosos, entonces  $N_1$  se puede estimar por

$$\hat{N}_1 = \frac{Y}{n} N \quad (23.3)$$

La varianza de la estimación,  $\hat{N}_1$ , es estimada por

$$S_{\hat{N}_1}^2 = \frac{N(N - n)}{n} \frac{Y}{n} \left(1 - \frac{Y}{n}\right) \frac{n}{n - 1} \quad (23.4)$$

**Problemas de recaptura de identificaciones** Aquí tenemos un problema bastante común en las poblaciones de vida silvestre. Se captura un número conocido de pájaros, peces o animales, se identifican o se marcan de alguna manera y se les vuelve a la población. Se captura otra muestra después y se observa el número de individuos recapturados. En este caso, el problema es estimar el tamaño de la población. Una estimación la da

$$\hat{N} = \frac{N_1 n}{Y} \quad (23.5)$$

donde  $N_1$  es el número de individuos marcados en la población, y  $Y$  es el número de individuos marcados en la muestra de tamaño  $n$ .

Las distribuciones hipergeométricas pueden generalizarse además para incluir más de dos clases de elementos.

**Ejercicio 23.2.1** Si se tiene alguna facilidad con el álgebra, es fácil transformar la ec. (22.8) y demostrar que es una probabilidad de una distribución hipergeométrica.

### 23.3 La distribución binomial

El muestreo de una distribución hipergeométrica es como la extracción de frijoles de un saco sin reemplazo, mientras que el muestreo de una distribución binomial es semejante a la extracción de frijoles, de un saco con reemplazo. O sea que la probabilidad asociada con un evento, tal como extraer un frijol marcado especialmente es constante de un ensayo a otro. Los eventos se dicen *mutuamente independientes*, o simplemente *independientes*. La distribución binomial se estudió en el cap. 3, donde se dio el siguiente cálculo de probabilidades para eventos múltiples :

$$P(Y = n_1 | n) = \binom{n}{n_1} p^{n_1} (1 - p)^{n - n_1} \quad Y = 0, 1, \dots, n \quad (23.6)$$

En los capítulos 21 y 22, se incluyeron pruebas de hipótesis nulas acerca de parámetros binomiales. En este capítulo, nos interesaremos más por el supuesto de una distribución binomial y la haremos parte de nuestra hipótesis, y, por tanto, estará sujeta a prueba.

La distribución binomial puede generalizarse para obtener la distribución multinomial en la cual hay más de dos resultados posibles para cada ensayo. De los posibles resultados sólo ocurrirá uno en cada ensayo.

### 23.4 Ajuste de una distribución binomial

¿Cuál es la probabilidad de que en una familia de 6 hijos, todos sean niñas? Supóngase que se aplica la distribución binomial y que  $P(\text{niña}) = 1/2$  antes de cada nacimiento. Entonces  $P(6 \text{ niñas}) = (1/2)^6 = 0.015625$ . Evidentemente este es un evento poco usual. Pero hay muchas familias de seis hijos. Si pensamos en 1,000 familias semejantes, el valor esperado es  $np = 1,000(0.015625) = 15.625$ . Es decir, si observamos muchas familias de seis hijos, estaremos casi seguros de observar familias de seis niñas. Es de interés saber si ocurren con más o menos frecuencia de la que debiera ocurrir si las probabilidades son binomiales. Es decir: ¿tenemos una distribución binomial con la independencia de eventos requerida y una probabilidad constante de un evento a otro?

Para determinar si tenemos una distribución binomial, debemos observar muchas familias de tamaño seis. Los resultados de nuestras observaciones serán registrados como seis niñas y ningún niño, cinco niñas y un niño, ..., ninguna niña y seis niños –en total siete grupos posibles de resultados. Como cada grupo de resultados proporciona información relativa a la distribución verdadera, debemos evitar combinar grupos si es que es posible.

Tabla 23.1 Ajuste de una distribución binomial

Y	f	Datos		Probabilidad		Cálculos			
		Frecuencia,		$P(Y = n_1   5)$	Coeficiente	$.4^y$	$.6^{5-y}$	P	Frecuencia esperada
0	13	$\binom{5}{0}$	.4 <sup>0</sup>	.6 <sup>5</sup>	1	1	1	.07776	.07776 6.2208
1	18	$\binom{5}{1}$	.4	.6 <sup>4</sup>	5	.4	.1296	.25920	20.7360
2	20	$\binom{5}{2}$	.4 <sup>2</sup>	.6 <sup>3</sup>	10	.16	.216	.34560	27.6480
3	18	$\binom{5}{3}$	.4 <sup>3</sup>	.6 <sup>2</sup>	10	.064	.36	.23040	18.4320
4	6	$\binom{5}{4}$	.4 <sup>4</sup>	.6	5	.0256	.6	.07680	6.1440
5	5	$\binom{5}{5}$	.4 <sup>5</sup>	.6 <sup>0</sup>	1	.01024	1	.01024	.8192
Totales	80		1.00					1.00	80.0000

Para todos los datos:  $p = .4$

$$\hat{p} = \sum Yf(Y)/5(80) = 161/400 = .4025$$

$$\hat{\sigma}_p^2 = p(1 - p)/n = .4(.6)/400 = .0006$$

$$\sigma_p^2 = \hat{p}(1 - \hat{p})/n = .4025(.5975)/400 = .0006$$

Conjunto de cinco muestras:  $p = .4$

$$\hat{p} = (80 \text{ Valores desde } 0 [2] 1)$$

$$\sigma_p^2 = .4(.6)/5 = .048$$

$$\mu_Y = np = 5(.4) = 2, \sigma_Y^2 = np(1 - p) = 5(.4)(.6) = 1.2$$

$$\bar{Y} = \sum Yf(Y)/\sum f(Y) = 161/80 = 2.0125$$

En este tipo común de problema puede entrar en juego el número de animales con cierto atributo genético en carandas de igual tamaño, o el número que responde a cierto estímulo si se utiliza el mismo número de plantas o animales de ensayo en ensayo. La respuesta puede referirse al efecto del medio ambiente sobre una característica genética, a la habilidad de obtener una respuesta constante (dentro de la variación en el muestreo binomial) a una técnica, o simplemente la habilidad para determinar si el análisis de un conjunto de datos se basa en el supuesto de una distribución binomial.

Para ilustrar el procedimiento de ajuste de una distribución binomial, consideremos los datos de la tabla 23.1, que se obtuvieron de la siguiente manera: en una clase de 80 estudiantes se asignó a cada uno el problema de extraer 10 muestras aleatorias de 10 observaciones de la tabla 4.1. Las 10 muestras fueron pareadas al azar y se calcularon las diferencias. Se calculó  $t = \bar{D}/s_D$  para 5 muestras de 10 diferencias aleatorias. Los resultados debieran seguir la distribución  $t$ , puesto que la verdadera media de las diferencias es cero y la población original es normal. Pero como las calculadoras y la estadística eran cosa nueva para todos los estudiantes, ello introducía la posibilidad de error.

Puesto que primero se desea ilustrar la distribución binomial para  $p$  conocido, elijamos arbitrariamente  $t_{0.20}(9 \text{ gl}) = 0.883$  y observemos el número de valores  $t$  mayores en valor absoluto que 0.883 por cada uno de los 80 conjuntos de cinco muestras. Sea este número  $Y$ . Ahora debemos ver qué tan bien se ajusta una distribución binomial con  $p = 0.4$  en vista de que  $P(|t| > t_{0.20}) = 0.40$  a nuestros datos. Por la ec. (23.6),

$$P(Y = n_1 | 5) = \binom{5}{n_1} .4^{n_1} .6^{5-n_1} \quad \text{para } n_1 = 0, 1, \dots, 5$$

Cuando cada probabilidad se multiplica por 80 de la frecuencia esperada. Estas frecuencias y los cálculos necesarios se indican en la tabla 23.1. Obsérvese que las frecuencias observadas no son necesarias para los cálculos. Más información sobre la población y la muestra se da al pie de la tabla.

Si bien los coeficientes son fáciles de obtener con la ec. (23.2), también se pueden hallar mediante el triángulo de Pascal. La construcción del triángulo es simple si se observa que cada número es la suma de los dos números de la línea anterior que están inmediatamente encima a derecha e izquierda; el primero y último números son siempre 1.

Número observado = $n$	Coeficiente binomiales						
1			1	1			
2			1	2	1		
3		1	3	3	1		
4	1	4	6	4	1		
5	1	5	10	10	5	1	
:			⋮				

Ahora estamos en condiciones de probar la hipótesis nula de que tenemos una distribución binomial con  $p = 0.4$ . Es una prueba de bondad de ajuste, como se expuso en el cap. 20; el criterio de prueba es  $\chi^2$ . Los cálculos se realizan en la tabla 23.2 y se obtiene  $\chi^2 = 31.215^{**}$  con 5 grados de libertad. La hipótesis nula de una distribución binomial

Tabla 23.2 Prueba de la distribución binomial con  $p = 0.4$

$Y$	Frecuencia observada	Frecuencia esperada	$(\text{Desviación})^2$	
			Desviación	Frecuencia esperada
0	13	6.2208	6.7792	7.388
1	18	20.7360	-2.7360	.361
2	20	27.6480	-7.6480	2.116
3	18	18.4320	-.4320	.010
4	6	6.1440	-.1440	.003
5	5	.8192	4.1808	21.337
Totales	80	80	0.0	31.215

con  $p = 0.4$  no es aceptable. Como no hubo que estimar  $p$ , sólo perdemos un grado de libertad, el cual está asociado con la restricción de que el número de observaciones es  $n = 80$ .

Cuando los datos no respaldan la hipótesis nula, podemos buscar la causa en el valor escogido para  $p$ , en la naturaleza de la distribución independientemente del valor de  $p$ , o en ambos. La proporción observada no difiere apreciablemente de la que se supuso en la hipótesis, así que concluimos que los datos no están distribuidos binomialmente. Las celdas con  $Y = 0$  y  $Y = 5$  son las únicas con desviaciones positivas; también son las mayores contribuyentes a  $\chi^2$ . Parece aconsejable comprobar los procedimientos de cálculo de los estudiantes que obtienen valores constantemente elevados o constantemente bajos de  $t$ .

Comparemos ahora el procedimiento de ajuste de la distribución binomial con el resumen de procedimientos, sec. 22.8, usados para los datos de la tabla 22.7. Estos procedimientos, aplicados a los presentes datos, requerirían una tabla  $80 \times 2$ . Los  $\chi^2$  individuales serían de poca utilidad mientras no se averigüe si los resultados de los 80 ensayos fueron homogéneos o no. La suma de los  $\chi^2$  individuales normalmente debería particionarse para probar una hipótesis acerca de  $p$  y una de homogeneidad. El valor  $\chi^2$  para totales sería una prueba de  $H_0: p = 0.4$ , y aceptaríamos la hipótesis nula ( $\hat{p} = 0.4025$ ). El  $\chi^2$  de homogeneidad prueba la hipótesis de una  $p$  constante de un ensayo a otro. Esperamos que muestre significancia. Para este ejemplo, los dos procedimientos bien podrían conducir a las mismas conclusiones.

¿Pueden los procedimientos llevar a diferentes conclusiones? Sí. Supóngase que de los 5 valores en cada uno de nuestros 80 ensayos, tres son menores que 0.883 y dos son mayores. La razón observada es la de la hipótesis y los datos son homogéneos. Es decir, podemos tener menos variación que la binomial y no estar en capacidad de determinar esto por los métodos del cap. 22. Los métodos de este capítulo están diseñados expresamente para detectar divergencias respecto de la distribución binomial.

Surge otro problema en el ajuste de una distribución binomial. Es el problema de ajustar una distribución binomial cuando no hay un valor conocido o supuesto en la hipótesis para  $p$ . *Para un  $p$  desconocido*, la prueba del ajuste de una distribución binomial prueba la naturaleza binomial de la variación, incluyendo la constancia de  $p$  de un ensayo a otro. Es necesario estimar  $p$ , que da cuenta de un grado de libertad. Así que con los datos de la tabla 23.1, estimamos  $p$  como  $\hat{p} = 0.4025$ , concluimos el proceso de ajuste de la tabla 23.1 usando 0.4025 como el valor de  $p$ , y continuamos con el procedimiento de prueba de la tabla 23.2. El  $\chi^2$  resultante tiene  $6 - 2 = 4$  grados de libertad en lugar de 5.

**Ejercicio 23.4.1** Completar el triángulo de Pascal hasta  $n = 10$ . ¿Cuántos coeficientes hay para  $n = 7$ ? ¿Cuántos hay en general?

**Ejercicio 23.4.2** Con los datos de la tabla 23.1 probar la hipótesis nula, de que  $p = 0.4$  suponiendo una distribución binomial.

**Ejercicio 23.4.3** Calcular la media teórica y la varianza del número de éxitos en cinco ensayos, usando las columnas encabezadas  $Y$  y la frecuencia esperada. Obsérvese que son iguales a  $5(0.4)$  y  $5(0.4)(0.6)$  calculadas por fórmula. (Ver sec. 2.17; casi no es necesario codificar).

**Ejercicio 23.4.4** Para cada uno de los 80 conjuntos de cinco muestras, también se ha observado el número de  $t$  mayores que 1.833 en valor absoluto. Los datos son los siguientes:

Número de $ t $ menores que 1.833:	5	4	3	2	1	0
Número de muestras:	51	18	6	2	2	1

¿Cuál es el valor teórico de  $p$  en esta distribución binomial? Ajustar una distribución binomial para el valor teórico de  $p$ . Al probar la bondad de ajuste de los datos, ¿sería deseable combinar algunos de los resultados? Probar la bondad de ajuste, asegurándose de establecer el número de grados de libertad para el criterio de prueba.

**Ejercicio 23.4.5** Los mismos 80 grupos de 5 valores de  $t$  fueron observados con respecto a otro valor tabulado de  $t$ . Si bien los resultados están disponibles y se dan a continuación, el valor de  $|t|$  no lo está.

Número de $ t $ menores que el valor no establecido:	5	4	3	2	1	0
Número de muestras:	23	30	16	5	4	2

Ajustar una distribución binomial a estos datos y probar la bondad de ajuste. ¿Cuántos grados de libertad tiene el criterio de prueba? Usando todos los datos, estimar la  $p$  binomial, por medio de un intervalo de confianza del 95 por ciento. A la luz de la prueba de bondad de ajuste, ¿es el intervalo de confianza demasiado amplio, demasiado estrecho, justo?

### 23.5 Transformación para la distribución binomial

En la sección 9.16, la transformación arcosen  $\sqrt{Y}$  se recomendó para datos binomiales. La tabla A.10 usada para este propósito y los ángulos resultantes se dan en grados. La varianza de una observación es aproximadamente  $821/n$ .

Cuando un investigador confía en que la variación en los datos es netamente binomial, esta transformación y la varianza teórica pueden ser de mucha utilidad. Por ejemplo, muchas *tablas de contingencia r × c × 2* pueden presentarse como tablas de proporciones  $r \times c$ . Si estas proporciones son transformadas de acuerdo con la transformación arcosen, se pueden emplear procedimientos corrientes de análisis de la varianza para obtener sumas de cuadrados de efecto principal y de interacción. Las proporciones transformadas dan una tabla de dos vías con solo una observación por celda. Debemos usar la varianza teórica.

Para probar una hipótesis nula, el procedimiento es el siguiente: calcular la suma de cuadrados apropiada usando las proporciones transformadas y procedimientos corrientes de análisis de la varianza. Si los denominadores de las proporciones son los mismos, entonces los cálculos son tales que todas las sumas de cuadrados están basadas por observación, donde por observación se refiere al común denominador. Por tanto, la varianza que ha de usarse para cada suma de cuadrados es  $821/n$ , donde  $n$  es el común denominador. Si los denominadores son desproporcionados, los cálculos se efectúan por métodos implícitos en la sec. 22.7, con las ponderaciones basadas en estos denominadores. Ahora, por observación se refiere al ensayo binomial único y la varianza que ha de usarse para cada suma de cuadrados es  $821/n$ , donde  $n = 1$ .

En vista de que  $\chi^2$  es igual a una suma de cuadrados dividida por  $\sigma^2$ , los cuadrados medios no se calculan. En lugar de ello, cada suma de cuadrados que se va a probar se divide por  $821/n$  para obtener un  $\chi^2$  basado en el número de grados de libertad asociado con la suma de cuadrados en el numerador. Los  $\chi^2$  resultantes se registran en la tabla A.5 para juzgar su significancia. En el caso de denominadores desiguales, los valores de  $\chi^2$  no serán aditivos.

**Ejercicio 23.5.1** Revisar el ejercicio 9.16.2 a la luz del nuevo conocimiento de la transformación arcsen. ¿Era razonable el término de error utilizado entonces? ¿Son aditivos los  $\chi^2$  en este ejercicio? ¿Por qué?

### 23.6 La distribución de Poisson

Esta distribución discreta se relaciona algunas veces con la distribución binomial con  $p$  pequeño y  $n$  grande. Con todo, es una distribución por derecho propio, y el muestreo aleatorio de organismos en un medio, el recuento de insectos en parcelas de cultivos, semillas de malezas nocivas en muestras de semillas, el número de algunos tipos de partículas radioactivas emitidas, pueden producir datos que siguen una distribución de Poisson.

Las probabilidades para una distribución Poisson están dadas por

$$P(Y = k) = \frac{e^{-\mu} \mu^k}{k!} \quad (23.7)$$

Esto se lee como "la probabilidad que la variable aleatoria  $Y$  tome el valor de  $k$  es igual a...". El valor de  $k$  puede ser 0, 1, ...; no hay punto final. La media de la distribución es  $\mu$ ; la varianza es también  $\mu$ . Es usual tomar varias observaciones y la media de éstas,  $\bar{Y}$ , da una estimación de  $\mu$ , y  $\sigma^2$ .

Un procedimiento exacto para obtener un intervalo de confianza para la media de la distribución Poisson (incluso una tabla) es el que dan Fisher y Yates (23.3). Blischke también estudia el problema (23.1).

Para ilustrar el *ajuste de una distribución de Poisson*, hemos elegido una muestra que da la distribución verdadera de células de levadura sobre 400 cuadrados de un hemaciótmetro. Estos datos fueron obtenidos por Student (23.6) y se presentan en la tabla 23.3.

Puesto que el parámetro  $\mu$  es desconocido, se estimará a partir de los datos. Para los cálculos se procede como en la tabla 23.3. Hemos vuelto a escribir la ec. (23.7) como una fórmula recursiva dada en la ec. (23.8) y utilizamos esta fórmula para obtener probabilidades.

$$P(Y = k) = \frac{\mu}{k} P(Y = k - 1) \quad (23.8)$$

El primer paso es encontrar  $P_{0,1}$  como se indica al pie de la tabla. Esto requiere que encontremos el antilogaritmo de 1.703594 – la única vez que se emplea la tabla de logaritmos. Si se dispone de una tabla de  $e^x$  hasta este único uso de la tabla de logaritmos se vuelve innecesario ( $e^{-\mu} = 1/e^\mu$ ). De aquí en adelante, el cálculo se realiza en forma secuencial, con las probabilidades registradas como aparecen en el computador. El paso final para el ajuste es multiplicar cada probabilidad por la frecuencia total; en este caso, 400.

Para probar la bondad del ajuste  $\chi^2$  es un criterio de prueba apropiado. El método se presenta en la tabla 23.4; las frecuencias esperadas provienen de la tabla 23.3. Para esta prueba tenemos cinco clases, estando las dos últimas combinadas, ya que una de las frecuencias esperadas es menor que 1. Además, fue necesario estimar  $\mu$ . Por tanto, los grados

Tabla 23.3 Ajuste de una distribución de Poisson

$\gamma$	Frecuencia observada	Probabilidad Ecuación (23.7)	Cálculos†	Probabilidad	Frecuencia esperada
0	213	$P_0 = e^{-\mu}$	.5054	.5054	202.16
1	128	$P_1 = \mu P_0$	.6825(.5054)	.3449	137.96
2	37	$P_2 = \frac{\mu}{2} P_1$	.34125 $P_1$	.1177	47.08
3	18	$P_3 = \frac{\mu}{3} P_2$	.2275 $P_2$	.0268	10.72
4	3	$P_4 = \frac{\mu}{4} P_3$	.170625 $P_3$	.0046	1.84
5	1	$P_5 = \frac{\mu}{5} P_4$	.1365 $P_4$	.0006	.24
> 5	0	$1 - \sum_{i=0}^5 P_i$	1-1.0000	.0000	.00
$\bar{Y} = 273/400 = 0.6825$				1.0000	400.00

†  $\log P_0 = -\mu \log e = -0.6825(0.434295) = -0.296406 = \bar{Y} = 0.6825; P_0 = 0.5054$

de libertad son  $5 - 1 - 1 = 3$  gl. Queda comprobado que estos datos no se ajustan a una distribución Poisson. Si existiera una buena razón para creer que los datos surgen con probabilidades de Poisson, entonces tenemos una muestra no usual o bien la media de la distribución no es estable.

Fisher (23.2) ha propuesto una medida alternativa de  $\chi^2$  para probar la bondad del ajuste de una distribución de Poisson. Esta prueba se emplea más especialmente para distribuciones observadas cuando las frecuencias esperadas son pequeñas. Rao y Chakravarti (23.4) han considerado el problema más ampliamente.

Tabla 23.4 Prueba de la bondad de ajuste de una distribución de Poisson

$\gamma$	Frecuencia observada	Frecuencia esperada	Desviación	$\frac{(O - E)^2}{E}$
0	213	202.16	10.84	.581
1	128	137.96	-9.96	.719
2	37	47.08	-10.08	2.158
3	18	10.72	7.28	4.944
4	3	1.84	2.08	1.772
5	1	.24		
Totales	400	400.00	0.00	10.174

**Ejercicio 23.6.1** Student observó las siguientes distribuciones. Ajustar una distribución de Poisson a una o más de estas muestras.

Muestra	Y												
	0	1	2	3	4	5	6	7	8	9	10	11	12
2	103	143	98	42	8	4	2						
3	75	103	121	54	30	13	2	1	0	1			
4	0	20	43	53	86	70	54	37	18	10	5	2	2

Probar la bondad de ajuste.

### 23.7 Otras pruebas con distribuciones de Poisson

El hecho de que la media y la varianza de una distribución de Poisson sean iguales sugiere que su razón pudiera proporcionar una prueba de significancia. La razón usualmente utilizada es

$$\chi^2_{n-1} = \frac{\sum (Y_i - \bar{Y})^2}{\bar{Y}} = \frac{(n-1)s^2}{\bar{Y}} \quad (23.9)$$

Obsérvese que es la razón de la suma de cuadrados, y no de la varianza, a la media. Es la ecuación  $\chi^2 = \sum [(O - E)^2/E]$  expresada en términos diferentes. Descartar  $H_0: \mu = \sigma^2$  equivale a descartar la hipótesis de que la distribución subyacente es de Poisson, o de que  $\mu$  es estable.

Aplicaremos este criterio de prueba a la muestra 2, en el ejercicio 23.6.1. Obtenemos

$$\chi^2_{399} = \frac{513.40}{1.3225} = 388.20$$

Obsérvese que los grados de libertad son  $400 - 1 = 399$  gl. Los valores  $Y$  son 0, 1, ..., 6. Como la tabla A.5 no da valores de  $\chi^2$  para 399 grados de libertad, confiamos en que  $\sqrt{2\chi^2} = \sqrt{2n - 1}$  tiene aproximadamente distribución normal con media cero y varianza uno. Por lo tanto, podemos usar la tabla A.4. Pero una ojeada a la tabla A.5 indica que el valor de  $\chi^2$  para la mediana ( $P = 0.5$ ) es aproximadamente igual a los grados de libertad. Parece poco útil completar los cálculos.

Este criterio de prueba puede usarse en situaciones experimentales como la de Student o en situaciones más generales. Así, podemos tener  $t$  tratamientos y hacer  $n$  observaciones en cada uno. El diseño experimental puede ser completamente al azar con igual número de repeticiones, un diseño de bloques completos al azar, o un cuadrado latino. La suma de variables Poisson también es una variable Poisson. Por tanto, si la hipótesis nula de que no hay diferencias entre tratamientos es válida, los totales de tratamientos seguirán

una distribución de Poisson, aun cuando haya diferencia real entre bloques. La ec. (23.9) con  $t - 1$  grados de libertad es apropiada para probar la hipótesis nula de que no hay diferencias entre medias de tratamientos.

Una situación un poco especial se presenta cuando hay probabilidades demasiado pequeñas, por ejemplo, cuando se observa el número de mutantes. Si las muestras son grandes y los números son pequeños, podemos suponer una distribución de Poisson. Por ejemplo, se observaron dos líneas de maíz y resultaron los siguientes datos:

	No. mutantes	Mutantes	Total (aprox.)
A	$5 \times 10^5$	10	$5 \times 10^5$
B	$6 \times 10^5$	4	$6 \times 10^5$

Quisiéramos probar la hipótesis nula de que la probabilidad de una mutación es igual para cada línea. Una prueba puede basarse en determinar si la división 10:4 es improbable al muestrear una población donde las proporciones verdaderas son  $5 \times 10^5 : 6 \times 10^5$  y nos detenemos después de observar 14 mutantes. Este problema supone una distribución condicional de Poisson que puede relacionarse con una distribución binomial.

Si usamos la distribución binomial, el valor esperado de  $p$  es  $5(10^5)/[5(10^5) + 6(10^5)] = \frac{5}{11}$ , que está entre 0.45 y 0.46. Usamos ahora la ec. (23.6) y obtenemos

$$P(Y \geq 10 | n = 14) = \sum_{n_1=10}^{14} \binom{14}{n_1} p^{n_1} (1-p)^{14-n_1}$$

Esto es igual a 0.0426 para  $p = 0.45$  y 0.0500 para  $p = 0.46$ , como se ve en la referencia 21.7.

Puesto que la probabilidad de obtener 10:4 o una división más extrema, dentro de la hipótesis nula, es pequeña (muy cerca de 0.05), se rechaza la hipótesis nula de que la probabilidad de mutación es la misma para cada línea. Otros criterios de prueba para este ejemplo son los expuestos por Steel (23.5).

**Ejercicio 23.7.1** Usar la ec. (23.9) para probar la hipótesis nula de que  $\mu$  es estable para los datos de la tabla 23.3. Hágase lo mismo para una muestra que pudo haberse empleado para completar el ejercicio 23.6.1. Comparénse los resultados obtenidos para los dos criterios.

**Ejercicio 23.7.2** Visualizar una situación en que los procedimientos de las secs. 23.6 y 23.7 llevarían a diferentes conclusiones. Explíquese la razón por la cual esto ocurre en términos de los supuestos, hipótesis nulas e hipótesis alternativas.

**Ejercicio 23.7.3** Aplicar el criterio de prueba  $\chi^2$  a los datos utilizados en esta sección. Comparar el resultado con el que se da en el texto.

### Referencias

- 23.1. Blischke, W. R.: "A comparison of equal-tailed, shortest, and unbiased confidence intervals for the chi-square distribution," tesis de maestría, Cornell University, Ithaca, N.Y., 1958.
- 23.2. Fisher, R. A.: "The significance of deviations from expectation in a Poisson series," *Biom.*, 6:17-24 ( 950).
- 23.3. Fisher, R. A., y F. Yates: *Statistical Tables for Biological, Agricultural and Medical Research*, 5a. ed., Hafner, Nueva York, 1957.
- 23.4. Rao, C. R., y I. M. Chakravarti: "Some small sample tests of significance for a Poisson distribution," *Biom.*, 12:264-282 (1956).
- 23.5. Steel, R. G. D.: "A problem involving minuscule probabilities," *Mimeo Series BU-81-M*, Biometrics Unit, Cornell University, Ithaca, N.Y., 1957.
- 23.6. Student: "On the error of counting with a haemacytometer," *Biometrika*, 5:351-360 (1907).

# VEINTICUATRO

---

## ESTADISTICA NO PARAMETRICA

### 24.1 Introducción

En las técnicas estudiadas hasta ahora, especialmente las que implican distribuciones continuas, se han puesto de relieve los supuestos fundamentales para los cuales son válidas las técnicas. Estas técnicas se emplean para estimación de parámetros y para probar las hipótesis relativas a ellos. Se llaman estadígrafos *paramétricos*. Los supuestos generalmente especifican la forma de la distribución y los caps. 1 a 17 tratan más que todo de datos en los que la distribución original es normal.

En una gran cantidad de datos recolectados, no es fácil especificar la distribución original. Para operar con tales datos, necesitamos estadística *de distribución libre*, es decir, procedimientos que no dependan de una distribución original específica. Si no especificamos la naturaleza de la distribución original, entonces ordinariamente no trataremos con parámetros. La *estadística no paramétrica* compara distribuciones y no parámetros. Estas estadísticas pueden ser sensibles a cambios en localización, dispersión o en ambos. No trataremos de establecer una distinción entre estadística de distribución libre y no paramétrica, sino que más bien a todos ellos los llamaremos estadígrafos no paramétricos.

La estadística no paramétrica tiene varias ventajas.

1. Cuando es posible hacer solamente supuestos débiles acerca de la naturaleza de las distribuciones que fundamentan los datos, entonces los estadígrafos no paramétricos son los apropiados. Los estadígrafos no paramétricos se aplican a una amplia clase de distribuciones más que a una distribución única o a todas las distribuciones posibles.
2. A veces, solo será posible poco más que categorizar los datos por falta de una escala de medición adecuada. En este caso, lo ideal es hacer una prueba no paramétrica. Otras veces, la categorización puede ser una forma de colectar datos con prontitud

—tantos datos que la potencia de una prueba no paramétrica sea adecuada para las necesidades del investigador.

3. Cuando es posible asignar rangos a los datos, se dispone de procedimientos no paramétricos. Por ejemplo, puede asignarse rangos por textura o sabor a productos alimenticios; o bien clasificar plantas por una infección de virus o parcelas por una infestación de insectos; en un ensayo de variedades que implique muchas localizaciones, las varianzas pueden ser heterogéneas violando los supuestos usuales para un análisis de la varianza válido, y los rangos pueden ser la mejor medida para el análisis.
4. Como la estadística no paramétrica usa recuentos, rangos o los signos de diferencias en observaciones pareadas, suelen ser, aunque no siempre, rápida y fácil de aplicar y de aprender.

Los procedimientos no paramétricos también tienen desventajas. Si se sabe que la forma de la población original es razonablemente cercana a una distribución para la cual hay una teoría normal, o si los datos pueden transformarse de modo que éste sea el caso, entonces los procedimientos no paramétricos extraen menos información de la que hay disponible en los datos. Si todos los experimentos de un investigador son tales que la hipótesis nula es verdadera, entonces los procedimientos no paramétricos son tan buenos como cualesquiera otros, puesto que el investigador fija la tasa de error. Sin embargo, si la hipótesis nula es falsa, entonces el problema usual es detectar diferencias entre medias. Los procedimientos no paramétricos son tan efectivos para este propósito como los procedimientos clásicos, siempre que sean válidos los supuestos acerca de la distribución original. En particular, la eficiencia de los procedimientos no paramétricos es bastante alta para muestras pequeñas, por ejemplo, para  $n \leq 10$ ; decrece a medida que  $n$  crece. Por otra parte, la eficiencia puede no ser importante para muestras muy grandes.

## 24.2 Prueba $\chi^2$ de bondad de ajuste

Frecuentemente, lo que deseamos no es saber algo acerca de los parámetros de una distribución supuesta, sino acerca de su forma. Es decir, deseamos probar la hipótesis de que los datos muestrales provienen de una distribución especificada. El criterio de prueba  $\chi^2$  está definido por las ecs. (20.2) y (21.2). Es apropiado para datos que caen en categorías. No se requiere escala para definir categorías, aunque puede existir y usarse una escala. Las probabilidades se requieren para calcular valores esperados para cada categoría o celda. Cochran (20.4 - 20.6) propone que ninguno de estos sea menor que 1. Las probabilidades pueden ser dadas enteramente por teoría o, en parte, estimarse a partir de los datos.

El uso de  $\chi^2$  ya se ha ilustrado. En la sec. 21.7, se efectuó una prueba de bondad de ajuste de una distribución 9:3:3:1 usando  $\chi^2$ . Las categorías eran puramente nominales y no se usaron los datos para determinar la razón y, en consecuencia, las probabilidades. En la sec. 20.4, se probó la hipótesis nula de que entraba en juego una distribución normal. Se midieron las observaciones originales y se usó la escala para determinar las categorías. Los parámetros normales tuvieron que ser estimados antes de poder calcular las probabilidades por celda. En las secs. 23.4 y 23.6, las distribuciones binomial y de Poisson fueron los modelos para las pruebas de bondad de ajuste. Los datos eran recuentos.

### 24.3 Prueba de Kolmogorov-Smirnov con una muestra

La prueba de bondad de ajuste fue desarrollada por Kolmogorov (24.10) para probar hipótesis acerca de distribuciones continuas con parámetros dados. Es considerada como conservadora, es decir,  $P(\text{rechazar } H_0 | H_0 \text{ verdadera}) < \alpha$ , tabulado, cuando se estiman los parámetros. También se le emplea para probar hipótesis acerca de distribuciones discretas.

En la sección 24.6, se presenta la prueba de Kolmogorov-Smirnov con dos muestras. Fue desarrollada por Smirnov (24.11). Las semejanzas en las pruebas ha llevado a la asociación de ambos nombres con ambas pruebas.

Supóngase que se desea probar la hipótesis nula de que los datos originales de la tabla 2.4 siguen una distribución normal. Puesto que no hay razón para especificar valores de los parámetros para esta distribución, los estimaremos mediante  $\bar{Y} = 75.943$  y  $s = 1.227$ .

Para la prueba, se necesitan la distribución muestral acumulado y la distribución acumulada hipotética. El estadígrafo de prueba frente a alternativas bilaterales es

$$D = \sup_Y |F_n(Y) - F_0(Y)| \quad (24.1)$$

donde  $F_n(Y)$  es la distribución muestral acumulada y  $F_0(Y)$  es la distribución acumulada en la hipótesis  $H_0$ . Hay que "tomar el supremo (sup), sobre todo  $Y$ , del valor absoluto de la diferencia".

La tabla 24.1 y la fig. 24.1 ilustran cómo se calcula  $D$ . En la tabla 24.1 se registran las observaciones ordenadas por rangos y las frecuencias acumuladas. Obsérvese que 76.0 ocurre tres veces mientras que todos los demás valores ocurren solo una vez.  $F_n(Y)$  es la frecuencia relativa acumulada o la distribución muestral acumulada. En la fig. 24.1, obsérvese que  $F_n(Y) = 0$  hasta  $Y = 73.9$ ; aquí, salta a  $F_n(Y) = \frac{1}{14} = 0.0714$  y permanece constante hasta que el siguiente  $Y$  es observado, en  $Y = 74.2$ , y así sucesivamente. En  $Y = 77.7$ ,  $F_n(Y) = 1$  y continua así para  $Y > 77.7$ . Se calculan valores  $Z$  para cada  $Y$  observado. Estos se llevan a la tabla A.4 para calcular  $F_0(Y) = P(Y \leq Y_i)$ . Se empleó la inter-

Tabla 24.1 Cálculos de  $F_n(Y)$  y  $F_0(Y)$

$Y$	Frecuencia acumulada	$F_n(Y)$	$Z = \frac{Y - 75.943}{1.227}$	$F_0(Y)$	$ F_n(Y_i) - F_0(Y_i) $	$ F_n(Y_{i-1}) - F_0(Y_i) $
73.9	1	.0714	-1.6648	.0480	.0234	.0480
74.2	2	.1429	-1.4203	.0778	.0651	.0064
74.6	3	.2143	-1.0944	.1369	.0774	.0060
74.7	4	.2857	-1.0129	.1555	.1302	.0588
75.4	5	.3571	-.4424	.3291	.0280	.0434
76.0	8	.5714	.0466	.5186	.0528	.1615
76.5	9	.6429	.4540	.6750	.0321	.1036
76.6	10	.7143	.5355	.7038	.0105	.0609
76.9	11	.7857	.7800	.7823	.0034	.0680
77.3	12	.8571	1.1060	.8656	.0085	.0799
77.4	13	.9286	1.1875	.8825	.0461	.0254
77.7	14	1.0000	1.4320	.9239	.0761	.1047

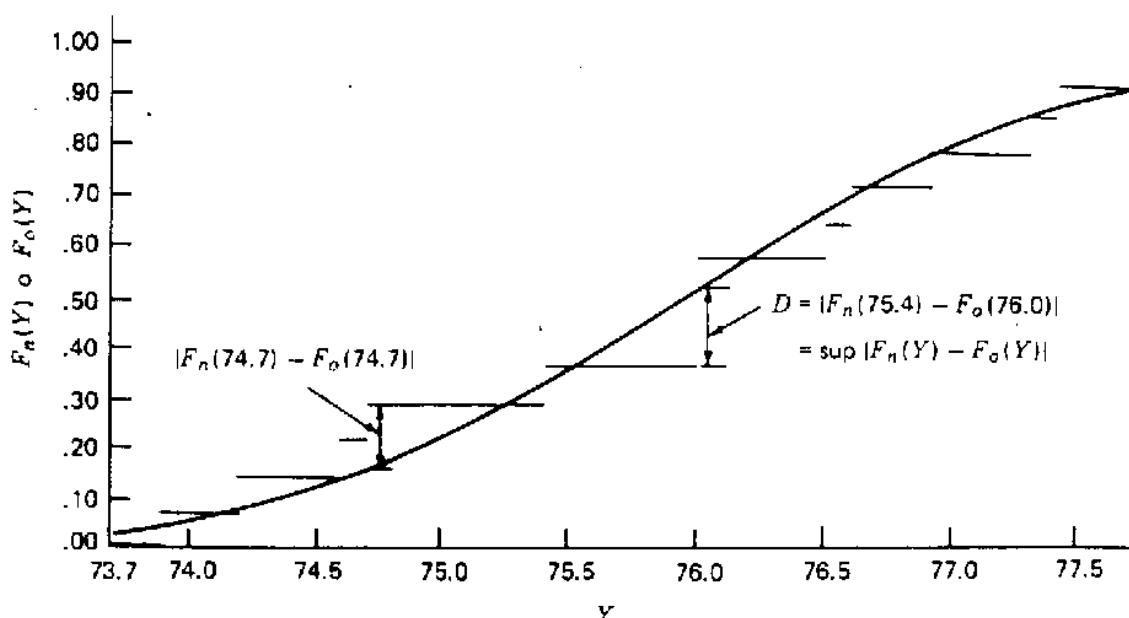


Figura 24.1 Gráfica para la prueba Kolmogorov-Smirnov de una muestra

polación lineal. Se representaron los valores  $F_0(Y)$  y se trazó la curva lisa en forma de S,  $F_0(Y)$ .

Para encontrar  $D$ , búsquese la distancia vertical más amplia entre  $F_n$  y  $F_0$ . Esto no ocurre necesariamente en un  $Y$  observado. En particular, no es suficiente observar  $|F_n(Y) - F_0(Y)|$  en la tabla 24.1. En la fig. 24.1, se ve que  $D$  ocurre en el punto extremo derecho del intervalo  $(75.4, 76.0)$ , donde  $D = |F_n(75.4) - F_0(76.0)| = 0.1615$ . Se podría decir que necesitamos  $F_n(Y)$  y  $F_0(Y)$  justamente antes que  $Y$  alcance el valor de 76.0. Así, además de observar todo  $|F_n(Y_i) - F_0(Y_i)|$  es necesario también observar todo  $|F_n(Y_{i-1}) - F_0(Y_i)|$ . Estos valores se dan también en la tabla 24.1. El valor  $|F_n(74.7) - F_0(74.7)| = 0.1302$  también era un posible  $D$ ; en este caso, la distancia se calcula en el extremo izquierdo del intervalo  $(74.7, 75.4)$ .

Para probar la hipótesis nula  $H_0: F(Y) = F_0(Y)$ , para todo  $Y$ , donde  $F_0$  es la distribución normal acumulada con  $\mu = 75.943$  y  $\sigma = 1.227$  contra la alternativa  $H_1: F(Y) \neq F_0(Y)$ , para al menos un  $Y$ , referirse con  $D = 0.1615$  a la tabla A.22. Para  $\alpha = 0.05$  y  $n = 14$ , el valor crítico es 0.349. No hay pruebas para rechazar la hipótesis nula.

También pueden hacerse pruebas con alternativas unilaterales. En este caso, pruébese  $H_0: F(Y) \geq F_0(Y)$  contra  $H_1: F(Y) < F_0(Y)$  o  $H_0: F(Y) \leq F_0(Y)$  con  $H_1: F(Y) > F_0(Y)$ . El criterio de prueba es

$$D^+ = \sup_Y [F_0(Y) - F_n(Y)] \quad \text{o} \quad D^- = \sup_Y [F_n(Y) - F_0(Y)]$$

O sea que para encontrar  $D^+$  verifíquese que  $F_0(Y)$  esté arriba de  $F_n(Y)$  cuando la alternativa es  $F(Y) < F_0(Y)$ ; y para encontrar  $D^-$ , compruébese que  $F_0(Y)$  esté abajo de  $F_n(Y)$  cuando la alternativa es  $F(Y) > F_0(Y)$ .

Esta prueba se usa también para probar la bondad de ajuste en datos discretos y de datos continuos agrupados. Para datos discretos,  $H_0: p_i = p_{i0}$  para todo  $i$ , es decir, la hi-

pótesis nula da la probabilidad de que una observación aleatoria caiga en la  $i$ -ésima celda, para todo  $i$ . Aquí también se permite la agrupación de celdas. Para datos continuos agrupados,  $F_n(Y)$  en la ec. (24.1) es el histograma acumulado el extremo de cada intervalo y  $F_0(Y)$  es la probabilidad en la  $H_0$ , acumulada en forma similar.

Para datos discretos y datos continuos agrupados, esta prueba se considera conservadora, esto es, los valores de  $D$  tabulados son mayores de lo necesario. O sea, el  $\alpha$  verdadero es más pequeño que el valor tabulado. Para valores críticos más exactos, ver Pettit y Stephens (24.13).

**Ejercicio 24.3.1** Los datos de la tabla 2.3 fueron analizados en el supuesto de que seguían una distribución normal. Sin embargo, hubo una desviación muy grande. ¿Era justificado el supuesto de normalidad?

**Ejercicio 24.3.2** En la tabla 5.4, se considerarán dos muestras de aumento de peso. Para cada muestra, ¿era justificado el supuesto de una distribución normal original subyacente?

**Ejercicio 24.3.3** En la tabla 5.5, se supuso que las diez diferencias estaban normalmente distribuidas. Probar la hipótesis nula de que esto es cierto.

**Ejercicio 24.3.4** La tabla 2.1 contiene datos agrupados de una distribución continua. Probar la hipótesis nula de que la distribución es normal. (Ver también la sec. 20.4).

**Ejercicio 24.3.5** En la sec. 2.5 se incluyen datos correspondientes a la distribución de un conjunto de cifras. Probar la hipótesis nula de que estos siguen una distribución uniforme, es decir, probar  $H_0$ :  $P(\text{una observación aleatoria} = i) = 1/10, i = 0[1]9$ .

**Ejercicio 24.3.6** Probar la hipótesis nula de que los datos de la tabla 2.2 siguen una distribución binomial con  $p = 0.5$ .

**Ejercicio 24.3.7** Probar la hipótesis nula de que los datos de las dos primeras columnas de la tabla 23.1 siguen una distribución binomial con  $p = 0.4$ .

**Ejercicio 24.3.8** Probar la hipótesis nula de que los datos de las dos primeras columnas de la tabla 23.3 siguen una distribución de Poisson.

#### 24.4 La prueba de signos

En esta prueba, consideramos medianas en vez de medias. La mediana es el valor tal que una mitad de la probabilidad queda de cada lado. Es claro que la media y la mediana serán las mismas en distribuciones simétricas.

La prueba de signos se basa en los signos de las diferencias entre valores pareados. Esto significa que pueden usarse también cuando las observaciones pareadas son simplemente ordenadas por rangos.

Para ilustrar el procedimiento de la prueba, consideremos los datos del ejercicio 5.7.1. Estos datos son constantes de enfriamiento de ratones recién sacrificados y de los mismos ratones llevados de nuevo a temperatura corporal. Las diferencias, recién sacrificados menos calentados de nuevo, son :+ 92, + 139, - 6, + 10, + 81, - 11, + 45, - 25, - 4, + 22, + 2, + 41, + 13, + 8, + 33, + 45, - 33, - 45, y - 12. Hay 12 más (+) y 7

menos (-). Estos números sirven para probar la hipótesis nula de que cada diferencia tiene una mediana cero —o sea, que los más (+) y los menos (-) ocurren con igual probabilidad.

Para probar la hipótesis nula de que cada diferencia pertenece a una distribución de probabilidad con mediana 0, puede aplicarse cualquiera de los criterios de prueba de la sec. 21.4. La ec. (24.2) es apropiada para probar  $H_0: p = 0.5$ , como es este el caso.

$$\chi^2 = \frac{(n_+ - n_-)^2}{n_+ + n_-} \quad (24.2)$$

Los valores  $n_+$  y  $n_-$  son los números de signos más y de signos menos. Para nuestro ejemplo,  $\chi^2 = (12 - 7)^2/19 = 25/19 = 1.32$  y no es significante, claro está.

Puede hacerse un ajuste de continuidad modificando el numerador de la ec. (24.2) en  $(|n_+ - n_-| - 1)^2$  pero esto hace poca diferencia cerca de los valores críticos para  $\alpha = 0.05$  y  $0.01$ . Dixon y Massey (24.14) dan valores críticos con base en la distribución binomial.

Esta prueba es fácil de aplicar y si planeamos usarla, la recolección de datos puede simplificarse. No necesitamos homogeneidad de varianzas en el sentido usual, puesto que cada diferencia puede provenir de una distribución continua diferente, siempre y cuando todas las distribuciones tengan como mediana cero. Por supuesto, las diferencias deben ser independientes. Además, la prueba no es sensible a errores grandes de registro. Esto puede ser de cierta importancia cuando se revisan trabajos de otros años y de otros investigadores.

La prueba tiene la desventaja de eliminar mucha información en la magnitud de las diferencias. Así, es imposible detectar una divergencia respecto de la hipótesis nula con menos de seis pares de observaciones. Se hace más útil con 20 o más pares de observaciones. Cuando hay empates, como sucede en la práctica, pueden asignarse en igual número para las categorías más y menos, o simplemente descartarse junto con la información que contienen.

La prueba de signos puede modificarse para tratar otras varias situaciones. Por ejemplo,

1. Para una muestra de una población única, podemos probar la hipótesis nula de que la mediana es un valor especificado. Obsérvese el número de observaciones que están por encima o por debajo del valor formulado como hipótesis y utilizar la ec. (24.2) como criterio de prueba.
2. Para observaciones pareadas, podemos preguntar si el tratamiento  $A$  da una respuesta mayor en  $C$  unidades que la del  $B$ . Este es el modelo lineal usual sin las restricciones acostumbradas. Obsérvense los signos de las diferencias  $Y_{1i} - (Y_{2i} + C)$  y aplíquese la prueba de signos.
3. Para observaciones pareadas, podemos preguntar si el tratamiento  $A$  da una respuesta que sea el  $k$  por ciento mejor que la que da  $B$ . Este es el modelo no aditivo que requeriría de una transformación si se usaran los procedimientos de los capítulos del análisis de la varianza. Obsérvense los signos de las diferencias  $Y_{1i} - (1 + K)Y_{2i}$ , donde  $K$  es la fracción decimal correspondiente al  $k$  por ciento, y úsese la prueba de signos.

Ejercicio 24.4.1 Aplicar la prueba de signos para los datos de la tabla 5.5. ¿Se llega a la misma conclusión que antes?

Ejercicio 24.4.2 Aplicar la prueba de signos para los datos del ejercicio 5.7.1. Del ejercicio 5.7.2. Del ejercicio 5.7.3. Del ejercicio 5.7.4. ¿Cuál sería la hipótesis nula que se prueba en cada caso?

## 24.5 Prueba de rangos signados de Wilcoxon

Esta prueba (refs. 21.7 y 21.9) es un mejoramiento de la prueba de signos en cuanto a detectar diferencias reales con tratamientos pareados. El mejoramiento es atribuible al uso de las magnitudes de las diferencias.

Los pasos en el procedimiento son:

1. Asignar rangos a las diferencias entre valores pareados en forma ascendente sin considerar el signo.
2. Asignar a los rangos los signos de las diferencias originales.
3. Calcular la suma de los rangos positivos  $T_+$  y la de los rangos negativos  $T_-$ . Estas están relacionadas por la ecuación  $T_+ + T_- = n(n + 1)/2$ . Elijase el menor numéricamente entre  $T_+$  y  $T_-$  llamándolo  $T$ . Sólo hay que calcular la suma menor, si se ve claro cuál de ellas va a ser.
4. Comparar la suma obtenida en el paso 3 con el valor crítico.

Aplicando el procedimiento anterior a los datos de los ratones de la sec. 24.4, obtenemos lo siguiente:

Diferencia:	+2,	-4,	-6,	+8,	+10,	-11,	-12,	+13,	+22,	-25
Rango signado:	+1,	-2,	-3,	+4,	+5,	-6,	-7,	+8,	+9,	-10
Diferencia:	-33,	+33,	+41,	-45,	+45,	+45,	+81,	+92,	+139,	
Rango signado:	-11½,	+11½,	+13,	-15,	+15,	+15,	+17,	+18,	+19	

La suma de los rangos negativos es  $T_- = T = 54.5$ .

Este valor, sin tener en cuenta el signo, se refiere a la tabla A.18 para juzgar la significancia. El valor crítico al nivel del 5 por ciento es 46, por lo tanto, concluimos que lo comprobado no es suficiente para negar la hipótesis nula. Nótese que los valores pequeños de  $T$  son significantes. Son los valores grandes de  $Z, t, \chi^2$ , y  $F$ , los que generalmente aportan la prueba contra la hipótesis nula.

Fuera del alcance de la tabla A.18, pueden usarse  $Z$  y la tabla A.4 para probar significancia. Para  $Z = (T - \mu_T)/\sigma_T$ ,  $\mu_T$  y  $\sigma_T$  los dan las ecs. (24.3) para  $n$  igual al número de pares.

$$\mu_T = \frac{n(n + 1)}{4} \quad \text{y} \quad \sigma_T = \sqrt{\frac{n(n + 1)(2n + 1)}{24}} \quad (24.3)$$

Obsérvese cómo se tratan los empates en este ejemplo. Se da el valor promedio a cada rango. Esto es necesario cuando los rangos empatados incluyen ambos signos, pero

no en otro caso. Al mejorar la prueba de signos para tener la prueba de rangos signados, ha sido necesario introducir un supuesto. El supuesto es que cada diferencia proviene de una distribución simétrica. De este modo, si el supuesto es válido, una inferencia acerca de la mediana también se aplica a la media. No es necesario que cada diferencia tenga la misma distribución. Esta prueba también puede usarse para una muestra única, en la que se desee probar una hipótesis nula acerca de la mediana. Aquí el valor hipotético se resta de cada observación y los valores resultantes se tratan como en las instrucciones precedentes.

**Ejercicio 24.5.1** Aplicar la prueba de rangos signados de Wilcoxon a los datos de la tabla 5.5. Comparar la conclusión con las obtenidas en el texto y en el ejercicio 24.4.1.

**Ejercicio 24.5.2** Aplicar la prueba de rangos signados de Wilcoxon a los datos del ejercicio 5.7.1. Del ejercicio 5.7.2. Del ejercicio 5.7.3. Del ejercicio 5.7.4. Comparar las conclusiones actuales con las obtenidas en las secc. 5.7 y 24.4. ¿Qué hipótesis nula se prueba en cada caso?

#### 24.6 Prueba de Kolmogorov-Smirnov de dos muestras

La prueba requiere dos muestras independientes y prueba la hipótesis nula de que éstas proceden de distribuciones idénticas. Si las muestras son  $Y_{11}, \dots, Y_{1n_1}$  y  $Y_{21}, \dots, Y_{2n_2}$ , entonces tenemos  $H_0: F_1(Y) = F_2(Y)$ , donde  $F_i$  es la verdadera función de distribución acumulada pero no especificada. El criterio de prueba exige que se comparan las dos funciones de distribución muestral. En particular, observamos la diferencia numérica máxima entre ellas. Birnbaum y Hall (24.15) dan tablas de valores críticos para  $n_1 = n_2$  y Massey (24.16) para  $n_1 \neq n_2$ .

Ilustramos el procedimiento con los datos de la tabla 5.2. La tabla 24.2 ayudará a aclarar las instrucciones.

1. Asignar rangos a todas las observaciones juntas.
2. Determinar las funciones de distribución acumulada muestral,  $F_n(Y_1)$  y  $F_n(Y_2)$ .
3. Calcular  $|F_n(Y_1) - F_n(Y_2)|$  en cada uno de los  $n_1 + n_2$  valores de  $Y$ .
4. Encontrar  $D$  y comparar con el valor crítico de las tablas A.23A y A.23B para sacar una conclusión.

En el ejemplo  $D = 6/7 = 36/42$  y en la tabla A.23B el valor crítico para  $\alpha = 0.01$  es  $5/6 = 35/42$ . Los datos no respaldan la hipótesis nula, de modo que se la rechaza. Concluimos que las dos muestras son de poblaciones diferentes. El orden de rangos en la tabla 24.2 sugiere ciertamente que las distribuciones difieren en localización. Sin embargo, la prueba es sensible también a diferencias en varianzas, ya que es una prueba de la igualdad de distribuciones más que de parámetros específicos.

Es fácil hacer pruebas contra alternativas unilaterales. Si  $H_1: F_1(Y) > F_2(Y)$ , entonces el criterio es  $D^+ = |F_n(Y_1) - F_n(Y_2)|$  para  $F_n(Y_1) > F_n(Y_2)$ . Si  $H_1: F_1(Y) < F_2(Y)$ , entonces el criterio es  $D^- = |F_n(Y_1) - F_n(Y_2)|$  para  $F_n(Y_1) < F_n(Y_2)$ .

La prueba también se usa para probar la igualdad de distribuciones discretas. Aquí no es una prueba exacta, pero es conservadora como se definió previamente.

**Tabla 24.2 Cálculos para una prueba de Kolmogorov-Smirnov de dos muestras**

$Y_1$	$F_n(Y_1)$	$Y_2$	$F_n(Y_2)$	$ F_n(Y_1) - F_n(Y_2) $
53.2	1/7			$ 1/7 - 0  = 1/7$
53.6	2/7			$ 2/7 - 0  = 2/7$
54.4	3/7			$ 3/7 - 0  = 3/7$
56.2	4/7			$ 4/7 - 0  = 4/7$
56.4	5/7			$ 5/7 - 0  = 5/7$
57.8	6/7			$ 6/7 - 0  = 6/7 = D$
		58.7	1/6	$ 6/7 - 1/6  = 29/42$
		59.2	2/6	$ 6/7 - 2/6  = 22/42$
		59.8	3/6	$ 6/7 - 3/6  = 15/42$
61.9	7/7			$ 1 - 3/6  = 1/2$
		62.5	4/6	$ 1 - 4/6  = 1/3$
		63.1	5/6	$ 1 - 5/6  = 1/6$
		64.2	6/6	$ 1 - 1  = 0$

**Ejercicio 24.6.1** Los datos de la tabla 5.6 presentaron alguna dificultad puesto que era discutible el supuesto de varianzas homogéneas. Probar la hipótesis nula de que las dos poblaciones originales son idénticas.

**Ejercicio 24.6.2** Aplicar la prueba de Kolmogorov-Smirnov a los datos del ejercicio 5.5.1. ¿Exactamente cuál es la hipótesis nula? ¿Qué hipótesis alternativa sería la más apropiada? ¿Es esta la alternativa con que se probó? ¿Qué criterio de prueba se emplearía para detectar esta alternativa? ¿Cuál es la conclusión?

**Ejercicio 24.6.3** Repetir el ejercicio 24.6.2 con los datos del ejercicio 5.5.2.

**Ejercicio 24.6.4** Aplicar la prueba de Kolmogorov-Smirnov a los datos del ejercicio 5.5.5. Utilizar una hipótesis alternativa bilateral.

**Ejercicio 24.6.5** Repetir el ejercicio 24.6.2 con los datos del ejercicio 5.5.6. (Los faisanes machos serían en promedio, más pesados que las hembras.)

**Ejercicio 24.6.6** Repetir el ejercicio 24.6.2 con los datos de la tabla 5.4.

## 24.7 Prueba de Wilcoxon-Mann-Whitney con dos muestras

Wilcoxon (24.7) desarrolló esta prueba de localización para dos muestras independientes de igual tamaño. La prueba fue ampliada por Mann y Whitney (24.3) para muestras de desigual tamaño.

La prueba para observaciones no pareadas es como sigue, donde  $n_1 \leq n_2$ .

1. Asignar rangos a las observaciones de ambas muestras en orden ascendente.
2. Totalizar los rangos de la muestra más pequeña. Llamar esto  $T$ .
3. Calcular  $T' = n_1(n_1 + n_2 + 1) - T$ , el valor que se obtendría para la muestra más

pequeña si las observaciones se hubieran ordenado de mayor a menor. (No es la suma de los rangos para la otra muestra.)

4. Comparar la suma de menor rango con los valores tabulados.

Ahora aplicamos la prueba a los datos de la tabla 5.2 sobre coeficientes de digestibilidad en ovejas (O) y novillos (N). Las observaciones se ordenan y se les asignan rangos así: 53.2(O),1; 53.6(O),2; 54.4(O),3; 56.2(O),4; 56.4(O),5; 57.8(O),6; 58.7(N),7; 59.2(N),8; 59.8(N),9; 61.9(N),10; 62.5(N),11; 63.1(N),12; 64.2(N),13. Cuando ocurren observaciones empatadas, se asigna el rango promedio. Las sumas de los rangos para novillos (N), la muestra más pequeña, es  $T = 6(6 + 7 + 1) - 60 = 24$ .

El valor observado del criterio de prueba se compara con los valores de la tabla A.19, elaborada por White (24.6). White también da una tabla para  $p = 0.001$ . De nuevo, obsérvese que los valores pequeños de este criterio de prueba llevan a rechazar la hipótesis nula. Dado que el valor, al 5 por ciento, de la menor suma rasqueada es 27, rechazamos la hipótesis nula. El análisis de la varianza dió un valor de  $F$  que está justamente por encima del nivel del 1 por ciento.

La diferencia entre las dos conclusiones puede obedecer a una o más causas. Primera, si los supuestos básicos del análisis de la varianza son verdaderos, entonces esperaríamos estar en capacidad mejor de detectar diferencias reales respecto de la hipótesis nula. Segunda, si los supuestos básicos son falsos, podemos estar detectando los supuestos falsos en vez de diferencias reales. La prueba de aditividad de Tukey y la prueba  $F$  (bilateral) de homogeneidad de varianzas ( $F$  de dos colas) pueden usarse para examinar la validez de los supuestos. En nuestro ejemplo, la diferencia entre las conclusiones parece trivial.

Si las tablas no son adecuadas, podemos usar la media y la desviación estándar de  $T$  dadas por

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} \quad \text{y} \quad \sigma_T = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \quad (24.4)$$

Con estos valores y  $T$ , podemos calcular la cantidad  $Z = (T - \mu_T)/\sigma_T$  que tiene distribución aproximadamente normal. La tabla A.4 puede usarse para juzgar la significancia.

**Ejercicio 24.7.1** Aplicar la prueba de Wilcoxon-Mann-Whitney a los datos de la tabla 5.4. ¿Qué hipótesis nula se está probando? ¿Cuál era la hipótesis alternativa? ¿Es la conclusión diferente de la del texto?

**Ejercicio 24.7.2** Aplicar la prueba de Wilcoxon-Mann-Whitney a los datos de la tabla 5.6. ¿Es apropiada la prueba en este caso?

**Ejercicio 24.7.3** Aplicar la prueba de Wilcoxon-Mann-Whitney a los datos del ejercicio 5.5.1. ¿Cuál es la hipótesis nula? ¿Qué alternativa se emplearía? ¿Cuál es la conclusión?

**Ejercicio 24.7.4** Repetir el ejercicio 24.7.3 con los datos del ejercicio 5.5.2.

**Ejercicio 24.7.5** Repetir el ejercicio 24.7.3 con los datos del ejercicio 5.5.5.

**Ejercicio 24.7.6** Repetir el ejercicio 24.7.3 con los datos del ejercicio 5.5.6.

## 24.8 Prueba de la mediana

Esta prueba fue dada por Mood (24.4) para usarse con dos muestras independientes. Prueba la hipótesis nula de que dos distribuciones continuas muestreadas tienen una mediana común. Es fácil de usar cuando las alternativas son bilaterales, caso considerado aquí. El procedimiento es como sigue.

1. Ordenar las dos muestras como una sola, de menor a mayor. Para los empates, se da el promedio de los rangos que se hubieran asignado.
2. Hallar la mediana.
3. Para cada muestra, observar el número de observaciones mayores que la mediana.
4. Utilizar estos dos números y los dos tamaños de muestra para completar una tabla de contingencia  $2 \times 2$
5. Probar la significancia por  $\chi^2$  con un grado de libertad si ambos tamaños de muestra exceden de 10; en otro caso, la ec. (22.8) es la apropiada, especialmente si la suma de los dos tamaños de muestra es pequeña. Esta es un término de la distribución hipergeométrica.

Para los coeficientes de digestibilidad, la mediana es 58.7. Hay 1 O mayor que la mediana, y 5 N. La tabla  $2 \times 2$  es la siguiente :

	O	N
Mayores	1	5
$n_i - \text{mayores}$	6	1
	7	6
		13

El total de la fila “ $n_i - \text{mayores}$ ” será igual a  $(n_1 + n_2 + 1)/2$  si  $n_1 + n_2$  es impar, y a  $(n_1 + n_2)/2$  si  $n_1 + n_2$  es par.

**Ejercicio 24.8.1** Probar la hipótesis nula de que las diferencias muestreadas que dan la tabla anterior tienen la misma mediana. Calcular la probabilidad de una tabla tan extrema o más que la observada usando la ec. (22.8). Comparar el resultado con los del análisis de la varianza y con los de la prueba de Wilcoxon-Mann-Whitney. ¿Se ha usado la misma hipótesis alternativa cada vez?

**Ejercicio 24.8.2** Aplicar la prueba de la mediana a los datos de la tabla 5.4. De la tabla 5.6. Del Ejercicio 5.5.1. Del ejercicio 5.5.2. Del ejercicio 5.5.5. Del ejercicio 5.5.6.

**Ejercicio 24.8.3** Resumir los resultados de la aplicación de los procedimientos de prueba paramétricas y no paramétricas a los datos a que se refiere el ejercicio 24.8.2. ¿Han sido las hipótesis alternativas las mismas siempre?

## 24.9 Prueba de Kruskal-Wallis con $k$ muestras

Kruskal y Wallis (24.2) han elaborado un criterio de prueba basado en rangos, el cual es apropiado para el diseño completamente al azar. Para  $k = 2$ , es equivalente a la prueba de

Wilcoxon-Mann-Whitney. Como en las otras pruebas de rangos, se supone que todas las poblaciones muestradas son continuas e idénticas, excepto posiblemente en la localización. La hipótesis nula es que todas las poblaciones tienen la misma localización.

El procedimiento para aplicar la prueba es el siguiente:

1. Asignar rangos a todas las observaciones de menor a mayor.
2. Sumar los rangos para cada muestra.
3. Calcular el criterio de prueba y comparar con los valores tabulados.

El criterio de prueba es

$$H = \frac{12}{n(n+1)} \sum_i \frac{R_i^2}{n_i} - 3(n+1) \quad (24.5)$$

Aquí  $n_i$  es el número de observaciones en la  $i$ -ésima muestra,  $i = 1, \dots, k$ ,  $n = \sum n_i$  y  $R_i$  es la suma de los rangos para la  $i$ -ésima muestra.  $H$  se distribuye como  $\chi^2$  con  $k - 1$  grados de libertad si los  $n_i$  son demasiado pequeños. Para  $k = 2$ , úsese la prueba de Wilcoxon-Mann-Whitney y los valores críticos dados en la tabla A.18. Para  $k = 3$  y todas las combinaciones de los  $n_i$  hasta 5, 5, 5, Kruskal y Wallis dan una tabla de probabilidades exactas (24.2). A los empates se da el rango promedio, y cuando son de tratamientos diferentes, puede hacerse una corrección en  $H$ . Esta corrección no cambia por lo general el valor de  $H$  en forma apreciable. La corrección es

$$\text{Divisor} = 1 - \frac{\sum T}{(n-1)n(n+1)} \quad (24.6)$$

donde  $T = (t-1)t(t+1)$  para cada grupo de empates y  $t$  es el número de observaciones empatadas en el grupo. Para obtener un  $H$  corregido, se utiliza este número como divisor.

Ahora aplicamos el procedimiento a los datos de la tabla 7.1. Las observaciones y sus rangos son (los números o letras entre paréntesis se refieren al tratamiento): 9.1(4), 1; 11.6(13), 2; 11.8(13), 3; 11.9(4), 4; 14.2(13), 5; 14.3(13), 6; 14.4(13), 7; 15.8(4), 8; 16.9(C), 9; 17.0(4), 10; 17.3(C), 11; 17.7(5), 12; 18.6(7), 13; 18.8(7), 14; 19.1(C), 15; 19.4(4), 17; 19.4(1), 17; 19.4(C), 17; 20.5(7), 19; 20.7(7), 20; 20.8(C), 21; 21.0(7), 22; 24.3(5), 23; 24.8(5), 24; 25.2(5), 25; 27.0(1), 26; 27.9(5), 27; 32.1(1), 28; 32.6(1), 29; 33.0(1), 30. Las sumas de los rangos para cada muestra son:  $R(1) = 130$ ,  $R(5) = 111$ ,  $R(4) = 40$ ,  $R(7) = 88$ ,  $R(13) = 23$ ,  $R(C) = 73$ . Ahora

$$H = \frac{12}{30(31)} \frac{130^2 + \dots + 73^2}{5} - 3(31) = 21.64 \quad \text{con } 6 - 1 = 5 \text{ gl}$$

Puesto que 21.64 está por encima de la probabilidad 0.005, se rechaza la hipótesis nula. Se llegó a la misma conclusión con el análisis de la varianza.

**Ejercicio 24.9.1** Aplicar la prueba de Kruskal-Wallis a los datos del ejercicio 7.3.1. ¿Qué hipótesis nula se está probando? ¿Contra qué alternativa? ¿Qué supuestos se han hecho?

**Ejercicio 24.9.2** Repetir el ejercicio anterior con los datos del ejercicio 7.3.3 usando los datos de antes y después, pero por separado.

**Ejercicio 24.9.3** Repetir el ejercicio 24.9.1 con los datos del ejercicio 7.4.1.

**Ejercicio 24.9.4** Repetir el ejercicio 24.9.1 con los datos del ejercicio 7.4.2.

## 24.10 Prueba de la mediana para $k$ muestras

La *prueba de la mediana* puede aplicarse también a datos procedentes de un diseño completamente al azar. El procedimiento es como sigue.

1. Asignar rangos a todas las observaciones de menor a mayor.
2. Hallar la mediana.
3. Hallar el número de observaciones superiores a la mediana para cada tratamiento.
4. Completar una tabla  $2 \times k$  con los números obtenidos en el paso 3 y las diferencias entre los  $n_i$  y estos números. (Ver la prueba de la mediana para dos muestras de la sec. 24.8).
5. Calcular  $\chi^2$  con  $k - 1$  grados de libertad para la tabla de contingencia obtenida en el paso 4.

**Ejercicio 24.10.1** Aplicar la prueba de la mediana para los datos del ejercicio 7.3.1. ¿Qué hipótesis nula se está probando? ¿Con qué alternativa? ¿Cuáles son los supuestos que fundamentan el procedimiento de prueba?

**Ejercicio 24.10.2** Repetir el ejercicio 24.10.1 con los datos del ejercicio 7.3.3. Del ejercicio 7.4.1. Del ejercicio 7.4.2.

**Ejercicio 24.10.3** Resumir los resultados del análisis de la varianza, de la prueba de Kruskal-Wallis, de la prueba de la mediana, para los conjuntos de datos en que se emplearon todas las tres pruebas

## 24.11 Prueba de Friedman para la clasificación de dos vías

Probablemente el diseño experimental más común es el diseño de bloques completos al azar con más de dos tratamientos. Friedman (24.1) ha propuesto la prueba siguiente para tales diseños:

1. Asignar rangos a los tratamientos dentro de cada bloque del más bajo al más alto.
2. Obtener la suma de los rangos para cada tratamiento.
3. Probar la hipótesis nula de que las poblaciones de un bloque son idénticas con la alternativa de que al menos un tratamiento proviene de poblaciones con una localización diferente en una dirección. El criterio de prueba es

$$\chi_r^2 = \frac{12}{bt(t+1)} \sum_i r_i^2 - 3b(t+1) \quad (24.7)$$

Tabla 24.3 Contenido de aceite en semillas de lino Redwing†

Bloque	Tratamiento					
	S	EB	FB	FB(1/100)	R	U
1	4.4(−)	3.3(−)	4.4(−)	6.8(+)	6.3(+)	6.4(+)
	2.5	1	2.5	6	4	5
2	5.9(+)	1.9(−)	4.0(−)	6.6(+)	4.9(−)	7.3(+)
	4	1	2	5	3	6
3	6.0(+)	4.9(−)	4.5(−)	7.0(+)	5.9(−)	7.7(+)
	4	2	1	5	3	6
4	4.1(−)	7.1(+)	3.1(−)	6.4(−)	7.1(+)	6.7(+)
	2	5.5	1	3	5.5	4
Total de rangos	12.5	9.5	6.5	19	15.5	21

$\chi^2_r = \frac{12}{4(6)7} (12.5^2 + \dots + 21^2) - 3(4)7 = 11.07$  con 5 gl;  $\chi^2_{0.05}(5 \text{ gl}) = 11.1$

† Ver tabla 9.2

con  $t - 1$  grados de libertad, donde  $t$  es el número de tratamientos,  $b$  el número de bloques, y  $r_i$ , la suma de los rangos para el  $i$ -ésimo tratamiento. Obsérvese que 12 y 3 son constantes y que no dependen del tamaño del experimento. La definición de  $r_i$  implica que  $r_{ij}$  es el rango del  $i$ -ésimo tratamiento en el  $j$ -ésimo bloque. Este criterio de prueba mide la homogeneidad de las  $t$  sumas y se distribuye aproximadamente como  $\chi^2$ . La aproximación es más diferente para valores pequeños de  $t$  y  $b$ . Friedman ha elaborado tablas de la distribución exacta de  $\chi^2_r$  de algunos pares de valores de  $t$  y  $b$  pequeños.

Aplicaremos ahora el procedimiento de Friedman a los datos de la tabla 9.2, presentados nuevamente en la tabla 24.3. Las respuestas a los tratamientos se dan frente a los números de bloque, y los rangos debajo. A los tratamientos empatados se les atribuye el rango promedio dentro del bloque donde ocurren. El valor del criterio de prueba es correcto al 5 por ciento para el ejemplo; el valor  $F$  estaba más allá del punto 1 por ciento.

El estadígrafo de prueba puede ajustarse para empates dividiendo por el valor

$$\text{Divisor} = \frac{1 - \sum_{i=1}^b T_i}{bt(t^2 - 1)} \quad (24.8)$$

donde  $T_i = \sum_h t_{ih}^3 - \sum_h t_{ih}$  y  $t_{ih}$  = el número de observaciones empatadas para un rango dado en el  $i$ -ésimo bloque. Aquí  $h$  es el índice de la sumatoria para los conjuntos de empates en el bloque. Por ejemplo, si hay tres observaciones de 8 empatadas para los rangos 4, 5, y 6 en el  $i$ -ésimo bloque y no hay más empates, entonces solamente hay un  $t_i$ , o sea 3; así que  $h = 1$  solamente y  $T_i = 3^3 - 3 = 24$ .

**Ejercicio 24.11.1** Aplicar el procedimiento de Friedman a los datos del ejercicio 9.3.1. ¿Qué hipótesis nula se está probando? ; Con qué alternativas? ; Qué supuestos se han hecho?

Ejercicio 24.11.2 Repetir el ejercicio 24.11.1 usando los datos del ejercicio 9.3.2.

Ejercicio 24.11.3 Repetir el ejercicio 24.11.1 usando los datos del ejercicio 9.3.5.

## 24.12 Una prueba de la mediana para la clasificación de dos vías

Mood (24.4) da otro procedimiento. Esta prueba supone que las distribuciones son continuas e idénticas excepto en localización. Obsérvese que esto requiere homogeneidad de la varianza. Se prueba si las contribuciones de tratamiento a las medianas de celda son todas cero. Las contribuciones son tales que si son reales, su mediana es cero. Para el modelo aleatorio o mixto, ningún supuesto es necesario acerca de interacción; para el modelo fijo, se supone cero para la interacción.

El procedimiento de prueba para tratamientos es como sigue :

1. Hallar la mediana de las observaciones en cada bloque.
2. Reemplazar cada observación por un signo más o menos, según que esté por encima o por debajo de la mediana de las observaciones en el bloque. (Ver tabla 24.3).
3. Registrar los números de signos más y menos, por tratamiento en una tabla  $2 \times t$ .
4. Probar el resultado como si se tratara de una tabla de contingencia ordinaria.

Ahora aplicamos el procedimiento para los datos de la tabla 24.3. No es necesario en general, calcular medianas. Por ejemplo, en el bloque 1 la mediana está entre 4.4 y 6.3; esto es suficiente para asignar signos más (+) y menos (-). Los signos dados en la tabla 24.3 son para esta prueba. La tabla de contingencia de dos vías se presenta abajo. Ahora sólo es necesario calcular  $\chi^2$  y comparar con los valores tabulados para  $t - 1 = 5$  grados de libertad.

Tratamiento	S	EB	FB	FB(1/100)	R	U
Por encima	2	1	0	3	2	4
No por encima	2	3	4	1	2	0

Ejercicio 24.12.1 Completar la prueba empezada arriba. ¿Cuál es la conclusión?

Ejercicio 24.12.2 Aplicar la prueba de la mediana a los datos del ejercicio 9.3.1. Del ejercicio 9.3.2. Del ejercicio 9.3.5.

Ejercicio 24.12.3 Resumir los resultados de los procedimientos de prueba y las conclusiones para la prueba F, la prueba de Friedman y la prueba de la mediana.

## 24.13 Desigualdad de Chebyshev

Esta desigualdad dice que

$$P(|Y - \mu| > k\sigma) \leq \frac{1}{k^2} \quad (24.9)$$

Como  $Y$  simplemente es una variable aleatoria, podemos sustituir por  $\bar{Y}$  siempre que remplacemos  $\sigma$  por  $\sigma_{\bar{Y}}$ . También es posible hacer otras sustituciones puesto que  $\mu$  y  $\sigma$  forman parte del enunciado de la probabilidad, se puede escribir mejor como independiente de la distribución que como no paramétrico. La desigualdad es válida para una distribución con varianza finita.

Apliquemos la desigualdad para el problema de determinar si las líneas de maíz de la sec. 23.7 difieren en la tasa de mutación. La distribución puede ser binomial; con el fin de estimar una varianza para la diferencia entre las proporciones, supondremos que este será el caso. La desviación estándar es

$$\hat{\sigma} = \sqrt{\frac{14}{11(10^5)} - \frac{11(10^5) - 14}{11(10^5)}} \left[ \frac{1}{5(10^5)} + \frac{1}{6(10^5)} \right] = \sqrt{\frac{14}{11(10^5)} \frac{11}{30(10^5)}} \text{ (aprox.)}$$

La hipótesis nula es que  $\mu$ , la diferencia entre proporciones, es cero. El valor observado de  $Y$  para utilizarlo en la ec. (24.9) es

$$Y = \frac{10}{5(10^5)} - \frac{4}{6(10^5)} = \frac{40}{30(10^5)}$$

Ahora, por la parte entre paréntesis de la ec. (24.9), se tiene

$$\frac{40}{30(10^5)} > k \sqrt{\frac{14}{11(10^5)} \frac{11}{30(10^5)}}$$

Si se despeja  $k$ , obtenemos  $k = 1.95$  y  $1/k^2 = 0.26$ . Por tanto, la probabilidad de obtener una diferencia, entre proporciones mayor que la obtenida es menor que 0.26.

Si hubiéramos estado preparados para suponer una distribución normal, entonces  $k$  habría sido una desviación normal y este procedimiento de dos colas habría exigido una probabilidad correspondiente de 0.05, bien lejos de 0.26. Vemos la importancia de usar un procedimiento no paramétrico cuando es imposible tener mucha confianza en un supuesto acerca de una distribución subyacente, y, simultáneamente, la importancia de aprovechar todo supuesto razonable. Además, parece que frecuentemente tendremos que confiar en una estimación de  $\sigma$ .

La desigualdad de Chebyshev también puede usarse para determinar el tamaño de la muestras se necesitan si deseamos obtener este resultado con una probabilidad de 0.90? por unidad de área de jamón almacenado en condiciones normales durante cierto tiempo dado. Deseamos que nuestra estimación esté dentro de  $\sigma/4$  de la verdadera  $\mu$ . ¿Cuántas muestras se necesitan si deseamos obtener este resultado con una probabilidad de 0.90?

En términos de un enunciado probabilístico, deseamos que

$$P\left(|\bar{Y} - \mu| > \frac{\sigma}{4}\right) \leq .10$$

Relacionando esto con la ec. (24.9) tenemos

$$P\left(|\bar{Y} - \mu| > \frac{\sqrt{n}}{4} \frac{\sigma}{\sqrt{n}}\right) \leq .10 = \frac{4^2}{n}$$

(Evidentemente  $k = \sqrt{n}/4$  y  $1/k^2 = 4^2/n$ .) Despejando  $n$  en la segunda igualdad, se tiene

$$n = \frac{4^2}{.10} = 160$$

El tamaño de muestra necesario es 160. Aquí no tuvimos que especificar  $\sigma$ .

#### 24.14 Coeficiente de correlación de rangos de Spearman

El coeficiente de correlación  $r$  se aplica a la distribución normal bivariante, distribución que no es muy común. Se han propuesto algunos coeficientes que no exigen suponer una distribución normal bivariante.

El coeficiente de correlación de rangos de Spearman se aplica a datos en forma de rangos. Pueden recolectarse como rangos o se les pueden asignar rangos según observaciones en alguna otra escala. Mide la correspondencia entre rangos, de tal manera que no es necesariamente una medida de correlación lineal. El procedimiento que se sigue es:

1. Asignar rangos a las observaciones para cada variable.
2. Hallar la diferencia en rangos para las observaciones pareadas. Sea  $d_i$  = la diferencia para el  $i$ -ésimo par.
3. Estimar  $\rho$  por la ec. (24.10).
4. Si el número de pares es grande, la estimación se puede probar usando el criterio dado en la ec. (24.11).

A continuación se dan las ecs. (24.10) y (24.11).

$$r_s = 1 - \frac{6 \sum_i d_i^2}{(n-1)n(n+1)} \quad (24.10)$$

donde  $r_s$  es el coeficiente de correlación de rangos de Spearman y  $n$  es el número de  $d$ . El criterio

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}} \quad (24.11)$$

se distribuye como una  $t$  de Student con  $n - 2$  grados de libertad.

Ahora aplicamos el procedimiento a los datos del ejercicio 11.2.1 para los caracteres  $T$  = longitud de tubo y  $L$  = longitud del limbo. A los empates se asigna el rango promedio.

$T:$	49, 44, 32, 42, 32, 53, 36, 39, 37, 45, 41, 48, 45, 39, 40, 34, 37, 35
Rango:	17, 13, 1.5, 12, 1.5, 18, 5, 8.5, 6.5, 14.5, 11, 16, 14.5, 8.5, 10, 3, 6.5, 4
$L:$	27, 24, 12, 22, 13, 29, 14, 20, 16, 21, 22, 25, 23, 13, 20, 15, 20, 13
Rango:	17, 15, 1, 12.5, 2.5, 18, 4, 9, 6, 11, 12.5, 16, 14, 7, 9, 5, 9, 2.5
Diferencia:	0, -2, .5, -5, -1, 0, 1, -5, .6, 3.5, -1.5, 0, .5, 1.5, 1, -2, -2.5, 1.5

Según la ec. (24.10),

$$r_s = 1 - \frac{6(37.50)}{17(18)19} = .9613$$

Como comprobación,  $\sum d_i = 0$ . También  $r_s$  debe estar entre -1 y +1.

Obviamente este valor de  $r_s$  es altamente significante, claro está.

También se dispone de otra aplicación. En ciertos casos, un conjunto de objetos puede tener un orden "verdadero". Por ejemplo, un conjunto de fichas pintadas puede tener un incremento en algún color o un conjunto de muestras de sabores puede tener un incremento en la cantidad de un aderezo. Si solicitamos a un jurado de colores o a un degustador asignar rangos al conjunto, entonces disponemos de un verdadero patrón de comparación de rangos. La correlación de rangos de Spearman es válida para la competencia de clasificación.

Ejercicio 24.14.1 Calcular  $t$  por la ec. (24.11). ¿Qué hipótesis nula se probaría cuando se compara este  $t$  con los  $t$  tabulados en la tabla A.3?

Ejercicio 24.14.2 Calcular los valores de  $r_s$  y  $t$  para  $T$  y  $N$ , y  $L$  y  $N$  para los datos del ejercicio 11.2.1. Comparar la  $t$  muestral con los valores correspondientes de  $t$  tabulados. ¿Qué conclusiones se sacan?

Ejercicio 24.14.3 Calcular  $r_s$  y  $t$  para los datos del ejercicio 11.2.2. ¿Qué conclusiones se obtienen?

## 24.15 Prueba de asociación del cuadrante de Olmstead-Tukey

Olmstead y Tukey (24.5) han desarrollado la siguiente prueba no paramétrica para la asociación de dos variables continuas, a la que llamaron prueba de *suma del cuadrante*. Los valores extremos son frecuentemente los mejores indicadores de una asociación entre variables y esta prueba les da una ponderación especial. Se calcula como sigue:

1. Representar las observaciones pareadas.
2. Obtener las medianas para cada variable.
3. Empezando por la parte superior, contar hacia abajo el número de observaciones

(usando el eje  $Y$ ) que aparezcan, hasta donde sea preciso cruzar la mediana vertical. Registrar este número junto con el signo del cuadrante.

4. Repetir lo mismo del paso 3 empezando por la derecha, usando la mediana horizontal.
5. Repetir lo mismo partiendo de la parte inferior y de la izquierda.
6. Calcular la suma del cuadrante y comparar con los valores tabulados.

La tabla A.20 es apropiada para juzgar la significancia de toda suma de cuadrante.

Un *número par de parejas* no plantea problemas en obtener las medianas y aplicar la prueba. Para un *número impar de parejas*, esta mediana pasa por un punto presumiblemente diferente. Sean estos puntos  $(X_m, Y)$  y  $(X, Y_m)$ . Con el fin de calcular la suma del cuadrante, remplácese estos dos pares por el único par  $(X, Y)$ . Esto deja un número par de parejas.

Ahora aplicamos el procedimiento para los datos del ejercicio 11.2.2. Estos datos están representados en la fig. 24.2, donde se calculó también la suma del cuadrante. Hemos usado flechas en la figura para mostrar cada punto de llegada como resultado del primer cruce de una media cuando se cuenten las cuatro direcciones. La suma del cuadrante es 17 y altamente significante.

Los empates ocurren en situaciones como la siguiente : si contamos hacia abajo hasta la primera observación que nos hace cruzar la mediana vertical, este punto puede tener un valor  $Y$  que es común a los puntos del otro lado. Para este valor  $Y$ , tenemos pues puntos que son favorables para la inclusión en la suma del cuadrante y puntos que no lo son. Olmstead y Tukey (24.5) sugieren que tales grupos empatados sean tratados como el número de puntos antes de cruzar la mediana fuese

$$\frac{\text{Número favorable para inclusión}}{1 + \text{Número desfavorable}}$$

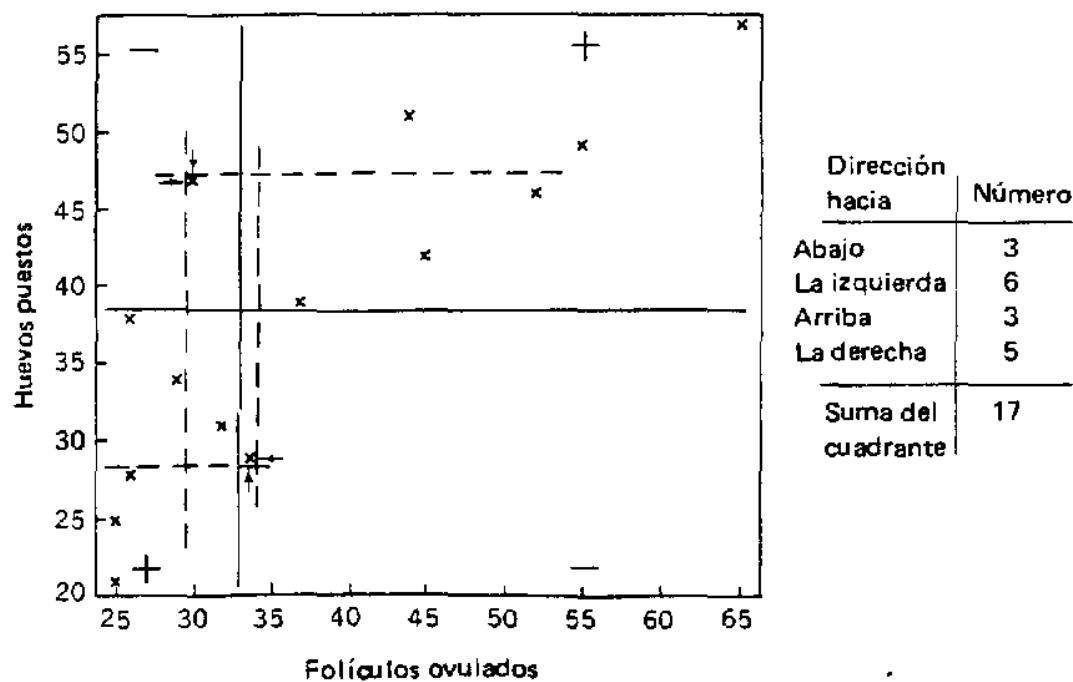


Figura 24.2 Prueba de asociación del cuadrante

Se ve ahora que la prueba de la suma del cuadrante es fácil de aplicar y da ponderación especial a valores extremos de las variables.

**Ejercicio 24.15.1** Aplicar la prueba de Olmstead-Tukey a los datos usados para ilustración en la sec. 24.14.

**Ejercicio 24.15.2** Aplicar la prueba de Olmstead-Tukey para buscar una asociación entre  $T$  y  $N$ , y entre  $L$  y  $N$ , en los datos del ejercicio 11.2.1.

**Ejercicio 24.15.3** Aplicar la prueba de Olmstead-Tukey para buscar una asociación entre las variables en los datos del ejercicio 11.2.2.

## 24.16 Prueba de aleatorización para regresión

Si planteamos como hipótesis que  $\beta = 0$  en un problema de regresión lineal, entonces esto es equivalente a decir que las parejas son aleatorias. A su vez, todos los posibles pareamientos son igualmente probables. Supóngase que construimos todos los  $n!$  apareamientos posibles y calculamos un valor de  $b$  para cada uno. Ahora tenemos una distribución empírica de  $b$  para la cual podemos obtener percentiles y valores críticos.

Los datos como se observan originalmente dieron un valor específico de  $b$ . Este puede compararse con la distribución empírica para concluir si se acepta o no la hipótesis nula.

Para obtener los pareamientos es conveniente ordenar el conjunto de valores de una variable y aleatorizar el otro conjunto de valores. El número  $n!$  aumenta muy rápido, de modo que puede ser suficiente muestrear todo el conjunto de aleatorizaciones. En todo caso, es esencial un equipo computador de alta velocidad.

## Referencias

- 24.1. Friedman, M.: "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Statist. Ass.*, 32:675-701 (1937).
- 24.2. Kruskal, W. H., y W. A. Wallis: "Use of ranks in one-criterion variance analysis," *J. Amer. Statist. Ass.*, 47:583-621 (1952).
- 24.3. Mann, H. B., y D. R. Whitney: "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, 18:50-60 (1947).
- 24.4. Mood, A. M.: *Introduction to the Theory of Statistics*, McGraw-Hill, Nueva York, 1950.
- 24.5. Olmstead, P. S., y J. W. Tukey: "A corner test for association," *Ann. Math. Statist.*, 18:495-513 (1947).
- 24.6. White, C.: "The use of ranks in a test of significance for comparing two treatments," *Biom.*, 8:33-41 (1952).
- 24.7. Wilcoxon, F.: "Individual comparisons by ranking methods," *Biom. Bull.*, 1:80-83 (1945).
- 24.8. Wilcoxon, F.: "Probability tables for individual comparisons by ranking methods," *Biom. Bull.*, 3:119-122 (1947).
- 24.9. Wilcoxon, F.: *Some Rapid Approximate Statistical Procedures*, American Cyanamid Company, Stamford, Conn., 1949.
- 24.10. Kolmogorov, A. N.: "Sulla determinazione empirica di una legge di distribuzione," *Giorn. Ist. Ital. Attuari*, 4:83-91 (1933).
- 24.11. Smirnov, N. V.: "Estimate of deviation between empirical distribution functions in two independent samples" (in Russian), *Bull. Moscow Univ.*, 2:3-16 (1939).

- 24.12. Miller, L. H.: "Table of percentage points of Kolmogorov statistics," *J. Amer. Statist. Ass.*, 51: 111-121 (1956).
- 24.13. Pettit, A. N., y M. A. Stephens: "The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data," *Technometrics*, 19: 205-210 (1977).
- 24.14. Dixon, W. J., y F. J. Massey, Jr.: *Introduction to Statistical Analysis*, 3a. ed., McGraw-Hill, Nueva York, 1969.
- 24.15. Birnbaum, Z. W., y R. A. Hall. "Small sample distribution for multi-sample statistics of the Smirnov type," *Ann. Math. Statist.*, 31:710-720 (1960).
- 24.16. Massey, F. J., Jr.: "Distribution table for the deviation between two sample cumulatives," *Ann. Math. Statist.*, 23:435-441 (1952).
- 24.17. Davis, L. S., "Table Errata 266," *Math. Comput.*, 12:262-263 (1958).

---

CAPITULO  
VEINTICINCO

---

## MUESTREO DE POBLACIONES FINITAS

### 25.1 Introducción

El muestreo estudiado hasta ahora se ha referido a experimentos en los cuales las unidades experimentales no se obtuvieron generalmente por procedimientos aleatorios. La aleatorización se usó para asignar tratamientos a unidades y las poblaciones eran hipotéticas en tanto que las unidades diferían por errores aleatorios. Ahora dirigimos nuestra atención a poblaciones que ya no son teóricas sino a poblaciones cuyas unidades experimentales pueden enumerarse y, en consecuencia, se pueden muestrear aleatoriamente. Por ejemplo, los silos se muestrean para determinar los residuos de insecticidas, los suelos para análisis químicos, las poblaciones de plantas, con propósitos taxanómicos, las frutas, para determinar calidad, los campos de trigo, para estimaciones de rendimientos y calidad antes de la cosecha, las razas primitivas, para analizar muchas características, las personas, para conocer sus opiniones, y así sucesivamente. En todos estos ejemplos, las poblaciones que interesan son infinitas.

Un problema nuevo se presenta en el muestreo de poblaciones finitas. Por ejemplo, si deseamos obtener información de muestra de distribuidores de semillas al por mayor en un estado, estaremos muestreando una población finita. Si la muestra es grande, como digamos 25 por ciento de los distribuidores de semillas al por mayor, hay que saber si las técnicas existentes se pueden aplicar o si será necesario desarrollar otras nuevas.

Al muestrear *poblaciones finitas* hay tres maneras bien distintas de hacer la selección. Estas son

1. Muestreo aleatorio
2. Muestreo sistemático
3. Muestreo autoritario

*El muestreo aleatorio* será nuestra mayor preocupación. La aleatorización puede introducirse en el procedimiento de muestreo de varias maneras que nos dan diversos dise-

ños de muestras. Gracias a la aleatorización pueden obtenerse estimaciones válidas del error. Se puede aplicar la teoría de la probabilidad y se pueden sacar conclusiones válidas.

*El muestreo sistemático* se usa cuando cada  $k$ -ésimo individuo de la población se incluye en la muestra. Tal procedimiento es siempre muy fácil, pero evidentemente insatisfactorio si en la población se presentan tendencias o ciclos no reconocidos aún. Dado que las poblaciones se deben enumerar antes del muestreo, pueden introducirse en forma inconsciente ciertas relaciones entre una o más de las características investigadas y orden de enumeración. En general, no es seguro suponer que no existe tal relación.

El muestreo sistemático puede efectuarse en forma tal, que puede obtenerse una estimación no sesgada del error de muestreo. Esto requiere de más de una muestra sistemática. Para una sola muestra sistemática, las fórmulas de que se dispone para estimar la varianza de una media suponen el conocimiento de la forma de la población.

*El muestreo autoritario* exige que una persona, bien familiarizada con el material que va a muestrearse, extraiga la muestra sin tener en cuenta la aleatorización. Tal procedimiento depende completamente del conocimiento y pericia de la persona que hace el muestreo. Puede producir buenos resultados en algunos casos, pero rara vez se recomienda.

## 25.2 Organización del estudio

Un considerable esfuerzo debe dedicarse a la planeación y ejecución de un estudio muestral, además de al muestreo efectivo. Vamos a hacer arbitrariamente una lista de cinco etapas para efectuar un muestreo.

1. Aclaración de objetivos
2. Definición de la unidad de muestreo y de la población
3. Selección de la muestra
4. Realización del estudio
5. Análisis de los datos

Se exponen estas etapas en forma breve.

*1. Aclaración de los objetivos* Esto consiste, ante todo, en establecer objetivos tan concisamente como sea posible, en que cada objetivo se enuncia como una hipótesis que va a probarse, un intervalo de confianza que ha de calcularse, o una decisión que debe tomarse.

Con este objetivo primario en mente, consideremos qué datos deben recolectarse. Cuando se tienen varios objetivos en mente, tal vez haya que modificar nuestras ideas respecto a qué datos deben recolectarse con el propósito de lograr todos los objetivos. Hasta podemos modificar nuestros objetivos de modo que el estudio no se vuelva demasiado complejo y costoso.

Usualmente, el grupo de estudio contará con un presupuesto fijo y deseará maximizar la cantidad de información por cada peso gastado. O bien, los objetivos incluirán una declaración sobre la cantidad de información deseada, generalmente en la forma del tamaño de un intervalo de confianza y tendremos que minimizar el costo.

*2. Definición de la unidad de muestreo y de la población* Hasta cierto punto esto se ha hecho en la etapa 1. Para el muestreo, la población debe dividirse en *unidades de mues-*

treo que, en conjunto, constituyen la población. Pueden existir varias posibilidades de escoger las unidades de muestreo. La elección final puede ser un tanto arbitraria, pero debe ser utilizable. Si estamos muestreando personas, podemos escoger el individuo, la familia, o los ocupantes de alguna vivienda específica como unidad de muestreo. Cualquiera que sea la selección hay que localizar e identificar la unidad sobre el terreno.

Para el muestreo aleatorio, debemos poder *enumerar* todas las unidades de muestreo, y tener una lista de todas las unidades. Puede ser necesario revisar las unidades existentes, por ejemplo, listas de niños escolares, granjeros, etc., o bien hacer nuevas listas, lo que parezca más factible y económico. Se pueden juxtaponer cuadriculas sobre los mapas del terreno, o bosques u otras zonas de las cuales se necesite obtener muestras de cultivos o información respecto a la cobertura de la vida silvestre. Aquí, puede ser necesario tener inventiva y hasta arbitrariedad, especialmente si tenemos que hacer el muestreo en un área de forma irregular.

Mientras se decide sobre la unidad de muestreo, es necesario considerar qué va a medirse y qué métodos de medida deben usarse. ¿Estamos midiendo estatura, peso u opinión? Si así es, ¿cómo lo vamos a hacer? ¿Puede usarse un cuestionario para medir estrés emocional? Si así es, ¿pueden emplearse como entrevistadores estudiantes universitarios que buscan un trabajo de tiempo parcial? ¿Se les puede adiestrar en unos días? ¿Puede ser de igual utilidad un candidato a ingeniero que un estudiante de premedicina?

*3. Selección de la muestra* Las formas en que pueden extraerse una muestra se llaman *diseños muestrales*. De ellos se hablará en secciones posteriores de este capítulo.

La selección del tamaño de la muestra se relaciona, en parte, con los recursos disponibles; si son inadecuados para obtener una muestra lo suficientemente grande para lograr los objetivos propuestos, deben revisarse los objetivos o retardar el estudio hasta cuando se tengan los fondos suficientes.

El diseño y el tamaño de la muestra darán una buena idea respecto a la extensión y naturaleza de las tablas y cálculos necesarios.

*4. Realización del estudio* Probablemente, será necesario adiestrar a parte del personal con el objeto de lograr uniformidad en la localización o identificación de las unidades de muestreo y en el registro de las respuestas a cuestionarios u otros datos. Será necesario un cronograma de actividades. Generalmente, se requiere de un esquema para una verificación temprana de la validez de los datos registrados en los diversos formatos. Debe preverse qué hacer en caso de que haya que tomar decisiones rápidas frente a hechos inesperados.

*5. Análisis de los datos* Primero, será necesario corregir los datos en cuanto a errores de registro e invalidez. Finalmente, el estudio muestral se deberá revisar en cuanto a maneras posibles de mejorar estudios futuros.

### 25.3 Muestreo probabilístico

Supóngase que nuestra población se ha definido claramente y que se ha hecho un listado de las unidades de muestreo. Ahora también podemos hacer una lista de todas las posibles muestras. Usamos el término *muestreo probabilístico* cuando

1. Cada unidad de muestreo tiene, o se le ha asignado, una probabilidad conocida de estar en la muestra.
2. Hay selección aleatoria en alguna etapa del procedimiento de muestreo y está directamente relacionado con probabilidades conocidas. La selección aleatoria supone un procedimiento mecánico para seleccionar las unidades que deben incluirse en la muestra.
3. El método de cálculo de una estimación de una media se establece claramente y llevará a un solo valor de la estimación. Esto es parte del análisis de los datos. Al estimar una media, usamos las probabilidades de selección asignadas a las unidades muestrales. Estas suministrarán ponderaciones, cada una de las cuales será cierto múltiplo constante del inverso de la probabilidad.

Cuando se cumplen estos criterios, puede asignarse una probabilidad de selección a cada muestra y a cada estimación. Por tanto, podemos construir una distribución de probabilidades de las estimaciones dadas por nuestro plan de muestreo. De esta manera, podemos evaluar el valor de nuestro plan y compararlo con otros planes de muestreo probabilístico. La evaluación consiste en medir la exactitud de toda estimación por la magnitud de su desviación estándar.

Cuando las probabilidades asignadas a cada unidad de muestreo son iguales, entonces los pesos que han de usarse en el cálculo de las estimaciones de las medias son todas iguales. Realmente no necesitamos pensar concretamente en las ponderaciones, ya que la muestra es *autoponderada*. Aunque tales muestras son fáciles de analizar, carecen de ciertas ventajas que poseen otros planes de muestreo probabilístico, ventajas tales como facilidad y bajo costo de administración por unidad de información y la capacidad de obtener estimaciones para estratos individuales (ver sec. 25.5).

Una muestra probabilística no garantiza que todas nuestras estimaciones sean no sesgadas. Ya hemos visto que en el muestreo aleatorio a partir de una población normal, optamos por utilizar  $s = \sqrt{\sum (Y - \bar{Y})^2/(n - 1)}$  si bien es una estimación sesgada de  $\sigma$ . También se usan estimaciones sesgadas en estudios muestrales. Deben usarse, naturalmente, con precaución, ya que pueden introducir distorsiones en las unidades probabilísticas. En particular, cuando se promedian estimaciones sesgadas (no necesariamente aritméticamente), el efecto sobre el promedio y su uso posterior puede no ser claro. Ahora se exponen varios tipos de muestreo probabilístico.

#### 25.4 Muestreo aleatorio simple

Para el muestreo aleatorio, se hace un listado de la población y se fija el plan y tamaño de la muestra. Para el *muestreo aleatorio simple*, cada muestra posible tiene la misma probabilidad de ser seleccionada. Este es el criterio importante.

En el proceso efectivo de selección de las unidades muestrales en una población finita, se usa una tabla de números aleatorios y el muestreo se hace sin *reemplazo*. Aparte de esto, las unidades de muestreo se extraen independientemente.

*Notación y definiciones* Como ahora estamos tratando principalmente con poblaciones finitas, se requieren nueva notación y definiciones. La notación y las definiciones, como

se verá, no son del todo coherentes en la literatura del muestreo. Trataremos de usar letras mayúsculas para cantidades de población y letras minúsculas para cantidades muestrales; también se usa  $\sigma^2$ .

Para comenzar, sea  $Y_i$  la observación  $i$ -ésima en la población. También usamos  $Y_i$  para describir la  $i$ -ésima observación muestral cuando no hay confusión posible.

Tamaño de la población:  $N$

Tamaño de la muestra:  $n$

Media de la población:

$$\bar{Y} = \frac{\sum Y_i}{N} = \frac{Y}{N}, \quad \text{variable continua}$$

$$P = \frac{A}{N} \quad \text{proporción}$$

Para una proporción  $Y_i = 0$  ó  $1$ ;  $\sum Y_i/N$  es la proporción de individuos que poseen una característica específica, así que puede servir también como una definición de la media de la población. Es común remplazar  $\sum_i Y_i$  por  $A$ . Para un porcentaje la media apropiada es  $100 P$ .

Media muestral:

$$\hat{Y} = \hat{y} = \frac{\sum Y_i}{n} = \frac{y}{n}, \quad \text{variable continua}$$

$$\hat{P} = p = \frac{a}{n}, \quad \text{proporción.}$$

Para una proporción,  $a$  remplaza a  $\sum_i Y_i$ . Como los totales de población y sus estimaciones son a menudo de interés, son bien corrientes las cantidades  $Y$ ,  $A$ ,  $\bar{Y}$ ,  $a$ .

Varianza de la población:

$$\begin{aligned} \sigma^2 &= \frac{\sum_i (Y_i - \bar{Y})^2}{N} \\ S^2 &= \frac{\sum_i (Y_i - \bar{Y})^2}{N - 1} \end{aligned} \tag{25.1}$$

Usamos la ec. (25.1) para definir la varianza de la población, ya que nuestra definición de  $s^2$ , ec. (25.4), da una estimación no sesgada de  $S^2$ .

Varianza de la población de una media:

$$S_{\bar{y}}^2 = \frac{S^2}{n} \left( \frac{N - n}{N} \right) \tag{25.2}$$

$$S_p^2 = \frac{PQ}{n} \left( \frac{N-n}{N-1} \right) \quad \text{donde } Q = 1 - P \quad (25.3)$$

Varianza muestral:

$$s^2 = \frac{\sum_i (Y_i - \bar{y})^2}{n-1} \quad \text{una estimación insesgada de } S^2 \quad (25.4)$$

El numerador se calcula como  $\sum Y_i^2 - (\sum Y_i)^2/n$ .

Varianza muestral de una media:

$$s_{\bar{y}}^2 = \frac{s^2}{n} \left( \frac{N-n}{N} \right) \quad (25.5)$$

$$s_p^2 = \frac{pq}{n-1} \frac{N-n}{N} \quad \text{donde } q = 1 - p \quad (25.6)$$

La ecuación (25.6) da una estimación insesgada de  $S_p^2$ , pero generalmente no se usa cuando se calculan intervalos de confianza. La forma más familiar está implícita en la ec. (25.8).

La cantidad  $(N-n)/N$  se conoce como *corrección de población finita* o cpf. También puede escribir  $1 - n/N$  y a  $n/N$  se le llama *fracción de muestreo*. Si la fracción de muestreo es pequeña, digamos menos del 5 por ciento, puede omitirse. Es de interés observar que  $qp/(n-1)$  es una estimación no sesgada de la varianza de la población independientemente de que la población sea finita o no, o sea que usamos una estimación insesgada de la varianza de la población en los caps. 20 a 23 cuando usamos  $\hat{p}(1-\hat{p})/n$ . (Recuérdese que en los caps. 20 a 23  $p$  se usa como parámetro y  $\hat{p}$  como estimación).

El intervalo de confianza para una media está dado por la ec. (25.7). Obsérvese que se hace uso de la cpf.

$$\text{IC} = \bar{y} \pm t \left[ \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N}} \right] \quad (25.7)$$

Obviamente estamos suponiendo que  $\bar{y}$  se distribuye normalmente, sabiendo que la población de los  $Y$  no es normal, ya que es una población finita. Además, se muestrea sin remplazo.

El intervalo de confianza para una proporción exige una distribución hipergeométrica, sec. 23.2, si ha de ser completamente válido. El investigador interesado puede remitirse a los gráficos de Chung y DeLury (25.3). Una aproximación corriente es

$$\text{IC} = p \pm t \sqrt{\frac{pq}{n} \frac{N-n}{N-1}} \quad (25.8)$$

Obsérvese que la desviación estándar estimada no es la que da la ec. (25.6), pero es comparable con la cantidad poblacional dada por la ec. (25.3). La estimación usada en la ec. (25.8) es la más común. La tabla 21.1 puede usarse para juzgar lo apropiado de la ec. (25.8).

El muestreo aleatorio simple se usa cuando se sabe que la población no es muy variable cuando la verdadera población cae entre el 20 y el 80 por ciento. Cuando hay variación considerable, las unidades de muestreo deberán agruparse en estratos de tal modo que puede esperarse que la variación dentro de los estratos sea inferior a la variación entre estratos. Esto lleva al muestreo estratificado. Idea muy parecida es la que lleva a un análisis de la varianza entre grupos y dentro de grupos.

**Ejercicio 25.4.1** Considérese la población finita que consiste en los números 1, 2, ..., 6. Calcular la media y la varianza. Considérense todas las posibles muestras de dos observaciones cuando se muestrea esta población sin remplazo. Hacer una tabla de medias y varianzas muestrales y la frecuencia de ocurrencia de cada uno de los valores. Demostrar que la media y varianza muestrales de la ec. (25.5) son estimaciones insesgadas de la media y varianza poblacionales de la ec. (25.1). Si el ejercicio hubiese dicho "muestra con remplazo" ¿qué cambios se requerirían en los cálculos?

**Ejercicio 25.4.2** Se va a muestrear una población consistente en 6,000 compradores de muebles, mediante un cuestionario enviado por correo en relación con su preferencia por determinado mueble. Se toma una muestra aleatoria de 250 personas y se envían los cuestionarios. Como la preferencia se refiere a un accesorio poco costoso, se prevén sólo respuestas sí o no. Todos los cuestionarios fueron devueltos con 187 respuestas sí. Estimar la verdadera proporción de respuestas sí en la población mediante un intervalo de confianza del 95 por ciento.

**Ejercicio 25.4.3** Con la misma población del ejercicio 25.4.2, se envió un cuestionario más largo a una muestra aleatoria de 750 compradores. Sólo fueron devueltos 469 cuestionarios. Estimar la verdadera proporción de los que *respondieron* en la población mediante un intervalo de confianza del 90 por ciento.

## 25.5 Muestreo estratificado

La varianza estimada de una media de población está dada por la ec. (25.5) y, para una proporción, por la ec. (25.6) o la alternativa más frecuente que implica la ec. (25.8). Para disminuir la longitud del intervalo de confianza que estima la media de población, podemos aumentar  $n$  o disminuir la varianza de la población. Obviamente, ambas posibilidades deben considerarse.

La forma obvia de disminuir una varianza de población es construir *estratos* con las unidades de muestreo, así que la variación total se partitiona de tal manera que la mayor parte posible se asigne a diferencia entre estratos. Así que la variación dentro de los estratos se mantiene baja. La variación entre medias de estratos en la población no contribuye al error de muestreo de la estimación de la media de la población. Ver la ec. (25.15).

La reducción en la variación de la estimación de la media de población es una razón muy importante para la estratificación. Pero en muchos estudios entran varias variables y una buena estratificación para una variable puede no serlo para otra. Así vemos que los estratos a menudo se construyen con base puramente geográfica. En general, esto da resultado, y así encontramos municipios, condados y zonas de recursos de tierra usados como estratos. Este tipo de estratificación a menudo es conveniente por razones administrativas,

ya que es posible obtener la cooperación de organismos, municipios, condados u otros organismos convenientemente localizados.

Además de aumentar la precisión con la cual se miden las medias, la estratificación permite un trabajo eficiente de asignación de recursos ya que podemos usar cualquier método para decidir cuántas unidades de muestreo se han de tomar de cada estrato. Se supone que se hará muestreo en cada estrato. A menudo se desean estimaciones de medias de estratos y, en tales casos, la estratificación es esencial.

*Notación y definiciones* La notación y definiciones para el muestreo aleatorio estratificado están obviamente relacionadas y son extensiones de las ya dadas en la sec. 25.4 bajo el mismo encabezamiento.

Sea  $Y_{ki}$  la observación  $i$ -ésima en el estrato  $k$ -ésimo,  $k = 1, \dots, s$ . Los tamaños de los estratos, medias y varianzas se designarán mediante  $N_k$ ,  $\bar{Y}_k$ , o  $P_k$  y  $S_k^2$  con los valores muestrales correspondientes  $n_k$ ,  $\bar{y}_k$  o  $p_k$  y  $s_k^2$ .

La media y la varianza de los estratos son:

$$\bar{Y}_k = \frac{\sum_{i=1}^{N_k} Y_{ki}}{N_k} = \frac{Y_k}{N_k}$$

$$P_k = \frac{A_k}{N_k}$$

$$S_k^2 = \frac{\sum_{i=1}^{N_k} (Y_{ki} - \bar{Y}_k)^2}{N_k - 1}$$

según la ec. (25.1)

La media y la varianza muestral para el estrato  $k$ -ésimo :

$$\hat{Y}_k = \bar{y}_k = \frac{\sum_{i=1}^{n_k} Y_{ki}}{n_k} = \frac{y_k}{n_k}$$

$$\hat{P}_k = p_k = \frac{a_k}{n_k}$$

$$s_k^2 = \frac{\sum_{i=1}^{n_k} (Y_{ki} - \bar{y}_k)^2}{n_k - 1}$$

según la ec. (25.4)

También se necesitarán parámetros y estadígrafos para la población completa.

Sean

$$N = \sum_k N_k \quad \text{y} \quad n = \sum_k n_k$$

La razón  $N_k/N$  se presenta con bastante frecuencia y se la representa con  $W_k$ , esto es  $W_k = N_k/N$  donde  $W$  corresponde a la ponderación.

Media de población (est significa estratificado):

$$\bar{Y}_{\text{est}} = \frac{\sum_k N_k \bar{Y}_k}{N} = \sum_k W_k \bar{Y}_k \quad (25.9)$$

$$P_{\text{est}} = \frac{\sum_k N_k P_k}{N} = \sum_k W_k P_k \quad (25.10)$$

Estimación de la media de población

$$\hat{Y}_{\text{est}} = \bar{y}_{\text{est}} = \frac{\sum_k N_k \bar{y}_k}{N} = \sum_k W_k \bar{y}_k \quad (25.11)$$

$$\hat{P}_{\text{est}} = p_{\text{est}} = \frac{\sum_k N_k p_k}{N} = \sum_k W_k p_k \quad (25.12)$$

(Las medias muestrales son  $\bar{y} = \sum n_k \bar{y}_k/n$  y  $p = \sum n_k p_k/n$ .)

La varianza de la estimación de la media de población:

$$\begin{aligned} \sigma^2(\bar{y}_{\text{est}}) &= \sum_k \left( \frac{N_k}{N} \right)^2 \frac{S_k^2}{n_k} \frac{N_k - n_k}{N_k} \\ &= \frac{1}{N^2} \sum_k N_k (N_k - n_k) \frac{S_k^2}{n_k} \end{aligned} \quad (25.13)$$

Compárese la ec. (25.5) con la primera expresión para  $\sigma^2(\bar{y}_{\text{est}})$ .

$$\sigma^2(p_{\text{est}}) = \sum_k \frac{N_k^2}{N^2} \frac{P_k Q_k}{n_k} \frac{N_k - n_k}{N_k - 1} \quad (25.14)$$

Compárese la varianza dada en la ec. (25.3) con  $\sigma^2(p_{\text{est}})$ .

La varianza muestral de la estimación de la media de población:

$$\begin{aligned} s^2(\bar{y}_{\text{est}}) &= \frac{1}{N^2} \sum_k N_k (N_k - n_k) \frac{s_k^2}{n_k} \\ &= \sum_k W_k^2 \frac{s_k^2}{n_k} - \frac{1}{N} \sum_k W_k s_k^2 \end{aligned} \quad (25.15)$$

La primera forma se obtiene de la ec. (25.13) sustituyendo parámetros por estimadores. Es una estimación no sesgada de  $\sigma^2(\bar{y}_{est})$ . La segunda forma puede usarse para cálculos.

$$s^2(p_{est}) = \sum_k W_k^2 \frac{p_k q_k}{n_k} \frac{N_k - n_k}{N_k - 1} \quad (25.16)$$

Esta ecuación se obtiene de la ec. (25.14). Da una estimación sesgada de  $\sigma^2(p_{est})$ , pero es de uso común.

Si la corrección de población finita es pequeña, se omite cuando se calculan los intervalos de confianza.

Al estimar la media de la población, se usan ponderaciones. Por esta razón, la estimación de la media de la población y de la media muestral,  $\bar{y}_{est}$  y  $\bar{y}$ , respectivamente, no tienen que ser las mismas. Sin embargo,  $n_1/N_1 = \dots = n_s/N_s = n/N$ , entonces  $\bar{y}_{est} = \bar{y}$ . A ésto se le llama *asignación proporcional* y se dice que la muestra es *autoponderada*.

Cuando se usa la asignación proporcional y las varianzas *dentro de los estratos* son homogéneas, los resultados relativos a las varianzas se pueden resumir en una tabla de análisis de la varianza con las fuentes de variación para el total, entre y dentro de estratos. El valor de la estratificación particular puede estimarse comparando la desviación estándar de  $\bar{y}_{est}$ , calculada a partir del cuadrado medio dentro de estratos, con la de  $\bar{y}$ , calculada a partir del cuadrado medio total. Cochran (25.2) y Hansen y otros (25.4) dan procedimientos exactos.

**Ejercicio 25.5.1** El problema de *los que no responden* es general, y de él padecen los usuarios de cuestionarios enviados por correo y aún los entrevistadores. Supóngase que se escogen al azar 900 compradores de una población de 6,000. Se reciben respuestas de 250 y, de éstos, 195 están a favor de una sugerencia. Estimar la proporción de "los que están a favor" en la población de los que responden; utilizar un intervalo de confianza del 95 por ciento.

De los que no responden conocidos, se extrae una muestra aleatoria de 50 y se entrevistan. De éstos, "a favor" hay 30 en la muestra de la población de los que no responden. Estimar mediante un intervalo de confianza del 95 por ciento, la proporción de los que responden "a favor" en esta población.

A veces la respuesta y no respuesta a los cuestionarios enviados por correo se usan como criterio de estratificación. Supóngase que los resultados anteriores se consideran como provenientes de tales estratos. Estimar la proporción de la población de los que están "a favor" y la desviación estándar de la estimación. Dar un juicio crítico práctico y dos teóricos sobre el procedimiento.

¿Cómo se podrían usar los resultados de la muestra para contrastar si los dos estratos difieren o no en respuesta a la sugerencia?

## 25.6 Asignación óptima

La estratificación generalmente produce una disminución en la varianza de la estimación de la media de la población. Sin embargo, la asignación proporcional no siempre es una *asignación óptima* y, por esto, puede necesitarse una *fracción muestral variable*.

**Costo fijo** En algunos experimentos en muestreo, el costo de obtener una observación a partir de una unidad de muestreo no varía nada de un estrato a otro, y se puede omitir al determinar las fracciones muestrales para los diferentes estratos. El problema está en mi-

nimizar  $\sigma^2(\bar{y}_{est})$  tal como aparece en la ec. (25.13) o  $\sigma^2(p_{est})$  dada por la ec. (25.14). Las fracciones muestrales se determinan por el tamaño del estrato y su variabilidad, y es bien claro que un estrato mayor implica un número de observaciones mayor, como ocurriría con un estrato con alta variabilidad. Se ha demostrado que la asignación óptima se obtiene cuando el número de observaciones tomadas en un estrato se determina mediante

$$n_k = n \frac{N_k S_k}{\sum_k N_k S_k} \quad (25.17)$$

Obsérvese que el denominador es la suma extendida a todos los estratos.

Aunque la aplicación de esta fórmula necesita de los parámetros  $S_k$ ,  $k = 1, \dots, s$ , a menudo es necesario usar estimaciones. Así, se desea información muestral para los años entre censos, se pueden usar desviaciones estándar con respecto al censo precedente más cercano. En otros casos, puede ser necesario hacer uso de información de otros estudios muestrales relacionados. Cuando no se dispone de estimaciones de  $S_k$ , se recomienda la asignación proporcional.

A veces, la ecuación (25.17), dará uno o más valores de  $n_k$  mayores que sus correspondientes  $N_k$ . En tales casos, el 100 por ciento del muestreo se hace en estos estratos y las restantes fracciones muestrales se ajustan de modo que la muestra total sea del tamaño planeado originalmente. Por ejemplo, si  $n_s > N_s$  al aplicar la ec. (25.17), entonces se hace  $n_s = N_s$  para propósitos del muestreo y recalcularmos los restantes  $n_k$  con la ecuación

$$n_k = \frac{(n - N_s)N_k S_k}{\sum_{k=1}^{s-1} N_k S_k} \quad k = 1, \dots, s-1$$

Cuando el muestreo estratificado es para proporciones, el  $n_k$  se puede determinar mediante

$$n_k = n \frac{N_k \sqrt{P_k Q_k}}{\sum_k N_k \sqrt{P_k Q_k}} \quad (25.18)$$

Esta ecuación es una aproximación; es similar a la ec. (25.17), si bien ésta no es una aproximación.

Lo que se gana en precisión como resultado del uso de una asignación óptima en vez de una proporcional, probablemente no es tanto para estimar proporciones como para estimar medias de variables continuas. Como la asignación proporcional ofrece la comodidad de las muestras autoponderadas, frecuentemente se recomienda cuando se han de estimar proporciones.

**Costo variable** Cuando el costo para obtener una observación varía de estrato en estrato, se necesita cierta función de costo que dé el costo total. Una sencilla función de costo es la dada por

$$\text{Costo} = C = a + \sum c_k n_k \quad (25.19)$$

donde  $a$  es un costo fijo, independientemente del tipo de asignación del muestreo a los estratos, y  $c_k$  representa el costo por observación en el estrato  $k$ . Para esta función de costo, la  $\sigma^2(\bar{y}_{est})$  mínima se obtiene si tomamos una muestra grande en un estrato grande, una muestra grande cuando la varianza del estrato es alta, y una muestra pequeña cuando el costo del estrato es alto. Es decir, el tamaño de la muestra para todo estrato es proporcional a  $N_k S_k / \sqrt{c_k}$ .

En un estudio efectivo tal vez tengamos que operar con un presupuesto fijo o bien haya que estimar la varianza de la media de población con una precisión especificada. El último requisito determina el tamaño de la muestra y, a su turno, el presupuesto.

Para un *presupuesto fijo*, el tamaño de muestra óptimo para cada estrato es

$$n_k = \frac{N_k S_k / \sqrt{c_k} (C - a)}{\sum_k N_k S_k \sqrt{c_k}} \quad (25.20)$$

Para una *varianza fija*, la ec. (25.21) da el tamaño óptimo de la muestra para cada estrato. En este caso, minimizamos el costo para una varianza fija o predeterminada.

$$n_k = \frac{N_k S_k}{\sqrt{c_k}} \frac{\sum_k N_k S_k \sqrt{c_k}}{N^2 \sigma^2(\bar{y}_{st}) + \sum_k N_k S_k^2} \quad (25.21)$$

$$= \frac{W_k S_k}{\sqrt{c_k}} \frac{\sum_k W_k S_k \sqrt{c_k}}{\sigma^2(\bar{y}_{st}) + \sum_k W_k S_k^2 / N}$$

donde  $W_k = N_k / N$ .

Las estimaciones aproximadas de los  $S_k^2$  y  $c_k$  suelen ser bastante adecuadas para estimar tamaños de muestra óptimos para los estratos. Cuando el muestreo es para estimar *proporciones*,  $S_k$  puede remplazarse por  $\sqrt{P_k Q_k}$  en las ecs. (25.20) y (25.21) para tener  $n_k$  aproximadamente óptimos.

**Ejercicio 25.6.1** Demuéstrese que las ecs. (25.17) y (25.20) dan los mismos resultados que la asignación proporcional cuando las varianzas de los estratos son homogéneos y el costo  $c_k$  no varía.

**Ejercicio 25.6.2** Los siguientes datos provienen de R. J. Jessen (25.5). Los estratos son tipos de áreas de cultivo,  $N_k$ , el número de granjas rurales,  $S_k^2(1)$  es una varianza para el número de porcinos y  $S_k^2(2)$  es una varianza para el número de ovejas. Los  $N_k$  corresponden a datos del censo de 1939, mientras que los  $S_k^2$  se han obtenido de una muestra tomada en 1939 y son solamente estimaciones.

Para cada conjunto de datos, calcular los tamaños de muestras utilizando asignaciones proporcional y óptima para un tamaño de muestra total de 800. Comparar los resultados.

Estrato	1	2	3	4	5	Estado
$N_k$	39,574	38,412	44,017	36,935	41,832	200,770
$S_k^2(1)$	1,926	2,352	2,767	1,967	2,235	2,303
$S_k^2(2)$	764	20	618	209	87	235

## 25.7 Muestreo multietápico o por conglomerados

En ciertos esquemas de muestreo, las unidades de muestreo están en grupos de igual o desigual tamaño y los grupos, en vez de unidades, son los que se muestran aleatoriamente. A tales grupos se les llama *unidades primarias de muestreo* o upm. Las observaciones pueden obtenerse sobre todas las unidades elementales o éstas, a su vez, pueden ser muestreadas. Por ejemplo, podemos estar interesados en individuos, las unidades elementales, y podemos obtenerlas extrayendo una muestra aleatoria de familias, las unidades primarias de muestreo, y observar dentro de ellas todas las unidades. Esto sería un plan de *muestreo por conglomerados simple* o un plan de *muestreo de una etapa*. Al muestrear el suelo en un terreno para llevar a cabo un experimento, podemos dividir el terreno en parcelas experimentales, colocar una cuadrícula encima de cada parcela para definir las unidades de muestreo, y luego obtener varias observaciones de cada parcela. Esto sería un muestreo en *dos etapas o submuestreo*, en el que la primera etapa era esencialmente un censo.

Obviamente pueden idearse muchos tipos de muestreo por conglomerados. La mayoría tendrá ciertas ventajas obvias en relación con el costo y la aplicabilidad, ya que el costo de pasar de una upm a otra es probablemente mayor que el de pasar de una subunidad a otra, porque la identificación de los upm puede ser más simple que identificar la subunidad. Cuando un conglomerado se define por asociación con un área, tenemos muestreo de área. Por ejemplo, podemos muestrear secciones de a cuarto de milla cuadrada de área, el conglomerado o upm, y enumerar todas las granjas en la upm.

Supóngase que una población consiste en  $N$  upm, de la cual sacamos una muestra aleatoria de tamaño  $n$ ; cada upm consiste en  $M$  subunidades, de las cuales extraemos una muestra de tamaño  $m$  para cada una de las  $n$  upm. (Las letras  $M$  y  $N$ , y  $m$  y  $n$  se intercambian a menudo en la literatura de muestreo). Ahora hay  $MN$  elementos en la población y  $mn$  en la muestra. El plan es de muestreo en dos etapas o de submuestreo.

Los cálculos usualmente se efectúan por elementos, tal como se acostumbra en el análisis de la varianza. Una observación se denota mediante  $Y_{ij}$ , donde  $j$  se refiere al elemento e  $i$  a la upm. La media de todos los elementos en una upm se designa  $\bar{Y}_i$ , o simplemente  $\bar{Y}_i$ , y la media de población por  $\bar{Y}_..$  o simplemente  $\bar{Y}$ . Las medias muestrales correspondientes están dadas por los símbolos  $\bar{y}_{ij}$ , o  $\bar{y}_i$  y  $\bar{y}_{..}$  o  $\bar{y}$ .

Supongamos ahora que  $N$  y  $M$  son infinitos y definimos un elemento por

$$Y_{ij} = \bar{Y} + \delta_i + \varepsilon_{ij} \quad (25.22)$$

Este es un modelo lineal tal como se expuso en las secc. 7.6 y 7.7 con notación algo diferente. Si denotamos la varianza de los  $\delta$  por  $S_a^2$  y las de los  $\varepsilon$  por  $S_w^2$ , entonces los cuadros medios muestrales definidos como en la tabla 25.1 son estimaciones de las cantidades

en la columna de valor esperado. Obsérvese que no se intenta que  $s_a^2$  sea una estimación de  $S_a^2$ .

Las sumas de cuadrados se calculan usualmente a partir de las siguientes fórmulas de cálculo y no por las fórmulas de definición de la tabla 25.1.

$$\text{Entre las upm: } (n - 1)s_a^2 = \frac{\sum_i Y_i^2}{m} - \frac{Y_{..}^2}{nm}$$

$$\text{Dentro de las upm: } n(m - 1)s_w^2 = \sum_i \left( \sum_{j=1}^m Y_{ij}^2 - \frac{Y_{i..}^2}{m} \right)$$

$$= \text{SC(total)} - \text{SC(dentro de las upm)}$$

$$\text{SC(total)} = \sum_{i,j} Y_{ij}^2 - \frac{Y_{..}^2}{nm}$$

Si relacionamos el análisis de la varianza dado en la tabla 25.1 con las secs. 7.6 y 7.8, vemos que el cuadrado medio dentro de las upm puede llamarse *error muestral* y que el cuadrado medio entre upm puede llamarse *error experimental*.

El error experimental en lugar del error muestral es el apropiado en cuanto a la estimación de  $\bar{Y}$  mediante un intervalo de confianza. El error experimental se basa en la unidad escogida al azar en la primera etapa de muestreo y corresponde a la parcela a la cual se aplica un tratamiento en forma aleatoria en un experimento en el terreno o de laboratorio. En el curso corriente de los sucesos, sería de esperar que el error muestral fuese menor que el error experimental ya que esperaríamos más homogeneidad dentro de las upm que entre las upm. Así, el error muestral no sería apropiado para calcular un intervalo de confianza para  $\bar{Y}$ .

La varianza de la media muestral,  $s^2(\bar{y})$ , se estima mediante  $s_a^2/nm$ , estimación no sesgada de la verdadera varianza. Ahora podemos estimar tanto  $S_w^2$  como  $S_a^2$  y construir

**Tabla 25.1 Análisis de la varianza y valores esperados en muestreo en dos etapas**

Fuente de variación	gl	Cuadrado medio	Valores esperados del cuadrado medio
Entre las upm	$n - 1$	$s_a^2 = \frac{n \sum_i (\bar{y}_i - \bar{y})^2}{n - 1}$	$S_w^2 + mS_a^2$
Dentro las upm	$n(m - 1)$	$s_w^2 = \frac{\sum_i \sum_j (Y_{ij} - \bar{y}_i)^2}{n(m - 1)}$	$S_w^2$
Total	$nm - 1$	$\frac{\sum_{i,j} (Y_{ij} - \bar{y})^2}{nm - 1}$	

estimaciones de las varianzas de las medias de tratamientos para las diferentes asignaciones de nuestros esfuerzos. Así, para el esquema presente,

$$\sigma^2(\bar{y}) = \frac{S_w^2}{nm} + \frac{S_a^2}{n}$$

No se disminuye  $S_a^2/n$  tomando más submuestras, pero  $S_a^2$  probablemente sea lo que más contribuya a  $\sigma^2(\bar{y})$ . Si fuésemos a muestrear  $n$ , entonces disminuiríamos ambas contribuciones. Así, en teoría, la mejor asignación a nuestro esfuerzo es tomar tantas upm como sea posible y tomar muy pocos elementos dentro de cada upm si ello implica un esfuerzo considerable; naturalmente, necesitamos dos de tales elementos de cada upm si tenemos que estimar bien sea  $S_w^2$  o  $S_a^2$  conservando facilidad de cálculo.

También se dispone de la teoría para poblaciones finitas y cuando las upm difieren en el número de elementos que contienen. El lector interesado puede consultar Cochran (25.2) y a Hansen et al. (25.4).

Al muestrear para proporciones con  $N$  conglomerados y  $M$  elementos por conglomerado, extráiganse  $n$  conglomerados y enumérense completamente. Entonces una proporción observada es una proporción verdadera  $P_i$  para el conglomerado  $i$ -ésimo y no está sujeta a variación muestral. Estimamos la proporción de población por

$$\hat{P} = p_{nM} = \frac{\sum_i P_i}{n}$$

y su varianza por

$$s^2(p_{nM}) = \frac{N - n}{N} \frac{1}{n} \frac{\sum_i (P_i - p_{nM})^2}{n - 1}$$

Cuando el esquema de muestreo supone tomar sólo  $m$  de los  $M$  elementos en un conglomerado, entonces sólo se estima  $P_i$  y debe introducirse un término para la variación muestral dentro de conglomerados en la varianza de la estimación de la proporción de población  $P$ . Ahora tenemos

$$\hat{P} = p_{nm} = \bar{p} = \frac{\sum_i p_i}{n}$$

y

$$s^2(\bar{p}) = \frac{M - m}{M - 1} \frac{m}{m - 1} \frac{1}{Nnm} \sum_i p_i q_i + \frac{N - n}{N} \frac{1}{n} \frac{\sum_i (p_i - \bar{p})^2}{n - 1}$$

Ladell (25.6) y Cochran (25.1) describen un interesante experimento de muestreo donde se impone un control local, en forma de una restricción sobre el submuestreo. En

**Tabla 25.2 Análisis de la varianza de los datos de cienpiés**

Fuente	gl	Suma de cuadrados	Cuadrado medio
Filas	4	515.44	128.86
Columnas	4	523.44	130.86
Error experimental	16	712.16	44.51
Entre mitades de parcelas	25	2.269.00	90.76
Error muestral	100	3.844.00	38.44
Totales	149	7.864.04	

el ejemplo particular intervino la superimposición de un diseño de cuadrado latino, inicialmente sin tratamientos, sobre un área experimental y la obtención de seis muestras de suelos de cada parcela. En estas muestras de suelos se hicieron recuentos de cienpiés. Se dispuso de manera que se tomaron tres muestras de cada una de las mitades norte y sur de la parcela. En consecuencia, las diferencias no aleatorias en el número de cienpiés entre las mitades no influyen en las comparaciones de tratamientos o en el error experimental. Los resultados se presentan en la tabla 25.2; aquí combinados el “error experimental” usual y “tratamientos” ya que no hubo verdaderos tratamientos. Es claro que “el control local” aumentó apreciablemente la precisión del experimento.

**Ejercicio 25.7.1** Suponga que deseamos hacer una estimación antes de la cosecha del rendimiento de trigo en un estado donde se cultiva trigo. El área sembrada de trigo se divide, para los fines del muestreo, en parcelas de un acre. Se toma una muestra aleatoria de 250 parcelas y se obtienen dos submuestras de cada una de las 250 parcelas. Cada submuestra es de 2 pies cuadrados, aproximadamente 1/10,000 de acre, así que no se necesita aplicar la teoría del muestreo finito.

El análisis de la varianza da un error muestral de 20 (dentro de las ump) y un error experimental de 70 (entre las ump). (Los rendimientos fueron convertidos a bushels por acre). Efectuar el análisis de la varianza. Estimar las componentes de la varianza. Calcular la varianza de una media de tratamiento. Estimar la varianza de una media de tratamiento suponiendo que la tasa de submuestreo se dobla (cuatro submuestreos, en vez de dos). ¿Da esto una apreciable ganancia en precisión? (Expresar la varianza estimada como un porcentaje de la observada).

**Ejercicio 25.7.2.** Utilizar los datos de la tabla 25.2 para calcular el error experimental como si no se hubiese superimpuesto un diseño a las parcelas. (Usar una media ponderada de cuadrados medios de filas, columnas y error experimental). ¿Cuál habría sido el error de muestreo si no se hubiera usado ningún control local? (Usar una media ponderada de los cuadrados medios de las mitades de las parcelas y el error de muestreo). ¿Cuál habría sido el error experimental sin el diseño y sin el control local? (Sumar los valores que se acaban de calcular para los errores experimental y muestral. Este incluye el error muestral sin control local dos veces de modo que el error muestral con control total debe restarse ahora). Calcular la desviación estándar de cada una de las tres varianzas que se acaba de calcular.

**Ejercicio 25.7.3** ¿Cuál es el mínimo número de submuestras por media parcela de control local si se necesita estimar el error de muestreo y, al mismo tiempo, no perder la facilidad de cálculo?

### Referencias

- 25.1. Cochran, W. G.: "The information supplied by the sampling results," *Ann. Appl. Biol.*, 25: 383-389 (1938).
- 25.2. Cochran, W. G.: *Sampling Techniques*, Wiley, Nueva York, 1953.
- 25.3. Chung, J. H., y D. B. DeLury: *Confidence Limits for the Hypergeometric Distribution*, University of Toronto Press, Toronto, Ontario, 1950.
- 25.4. Hansen, M. H., W. N. Hurwitz, y W. G. Madow: *Sample Survey Methods and Theory*, 2 vols., Wiley, Nueva York, 1953.
- 25.5. Jessen, R. J.: "Statistical investigation of a sample survey for obtaining farm facts," *Iowa Agr. Exp. Sta. Res. Bull.* 304, 1942.
- 25.6. Ladell, W. R. S.: "Field experiments on the control of wireworms," *Ann. Appl. Biol.*, 25: 341-382 (1938).

## TABLAS

- A.1 Diez mil dígitos aleatorizados
  - A.2 Valores de la razón, intervalo dividida por la desviación estándar  $\sigma$  para tamaños de muestras desde 20 hasta 1000.
  - A.3 Valores de  $t$
  - A.4 Probabilidad de hallar un valor al azar de  $Z = (Y - \mu)/\sigma$  mayor que los valores tabulados en los márgenes.
  - A.5 Valores de  $\chi^2$
  - A.6 Valores de  $F$
  - A.7 Amplitudes studentizadas significativas para el 5 y 1 por ciento de la nueva prueba de amplitud múltiple.
  - A.8 Puntos porcentuales superiores de la amplitud studentizada,  $q_\alpha = (\bar{Y}_{\max} - \bar{Y}_{\min})/s_y$
  - A.9 Tabla de  $t$  para comparaciones de una y dos colas entre  $p$  medias de tratamientos y un control para un coeficiente de confianza conjunto de  $P = 0.95$  y  $P = 0.99$
  - A.10 Transformación de la  $\sqrt{\text{porcentaje arco seno}}$
  - A.11 Franja de confianza para el coeficiente de correlación  $\rho$ :  $P = 0.95$  y  $P = 0.99$
  - A.12 Transformación de  $r$  a  $Z$
  - A.13 Valores significativos de  $r$  y  $R$
  - A.14 Límites de confianza binomiales
  - A.15 Franjas de confianza para proporciones: Coeficientes de confianza de 0.95 y 0.99
  - A.16 Papel de probabilidad binomial de Mosteller-Tukey
  - A.17 Tamaño de muestra y probabilidad de tomar decisiones erróneas para un conjunto limitado de alternativas.
  - A.18 Prueba de rangos signados de Wilcoxon
  - A.19 Puntos críticos de sumas de rangos
  - A.20 Niveles significativos de trabajo para magnitudes de sumas de cuadrantes.
  - A.21 Valores de  $t$  de riesgo-promedio-mínimo de Duncan-Waller
  - A.22 Valores críticos para la prueba de una muestra de Kolmogorov-Smirnov
  - A.23 Valores críticos para la prueba de dos muestras de Kolmogorov-Smirnov
  - A.24 Alfabeto griego
- Índice

Tabla A.1 Diez mil dígitos aleatorizados

	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
00	88758	66605	33843	43623	62774	25517	09560	41880	85126	60755
01	35661	42832	16240	77410	20686	26656	59698	86241	13152	49187
02	26335	03771	46115	88133	40721	06787	95962	60841	91788	86386
03	60826	74718	56527	29508	91975	13695	25215	72237	06337	73439
04	95044	99896	13763	31764	93970	60987	14692	71039	34165	21297
05	83746	47694	06143	42741	38338	97694	69300	99864	19641	15083
06	27998	42562	63402	10056	81668	48744	08400	83124	19896	18805
07	82685	32323	74625	14510	85927	28017	80588	14756	54937	76379
08	18386	13862	10988	04197	18770	72757	71418	81133	69503	44037
09	21717	13141	22707	68165	58440	19187	08421	23872	03036	34208
10	18446	83052	31842	08634	11887	86070	08464	20565	74390	36541
11	66027	75177	47398	66423	70160	16232	67343	36205	50036	59411
12	51420	96779	54309	87456	78967	79638	68869	49062	02196	55109
13	27045	62626	73159	91149	96509	44204	92237	29969	49315	11804
14	13094	17725	14103	00067	68843	63565	93578	24756	10814	15185
15	92382	62518	17752	53163	63852	44840	02592	88572	03107	90169
16	16215	50809	49326	77232	90155	69955	93892	70445	00906	57002
17	09342	14528	64727	71403	84156	34083	35613	35670	10549	07468
18	38148	79001	03509	79424	39625	73315	18811	86230	99682	82896
19	23689	19997	72382	15247	80205	58090	43804	94548	82693	22799
20	25407	37726	73099	51057	68733	75768	77991	72641	95386	70138
21	25349	69456	19693	85568	93876	18661	69018	10332	83137	88257
22	02322	77491	56095	03055	37738	18216	81781	32245	84081	18436
23	15072	33261	99219	43307	39239	79712	94753	41450	30944	53912
24	27002	31036	85278	74547	84809	36252	09373	69471	15606	77209
25	66181	83316	40386	54316	29505	86032	34563	93204	72973	90760
26	09779	01822	45537	13128	51128	82703	75350	25179	86104	40638
27	10791	07706	87481	26107	24857	27805	42710	63471	08804	23455
28	74833	55767	31312	76611	67389	04691	39687	13596	88730	86850
29	17583	24038	83701	28570	63561	00098	60784	76098	84217	34997
30	45601	46977	39325	09286	41133	34031	94867	11849	75171	57682
31	60683	33112	65995	64203	18070	65437	13624	90896	80945	71987
32	29956	81169	18877	15296	94368	16317	34239	03643	66081	12242
33	91713	84235	75296	69875	82414	05197	66596	13083	46278	73498
34	85704	86588	82837	67822	95963	83021	90732	32661	64751	83903
35	17921	26111	35373	86494	48266	01888	65735	05315	79328	13367
36	13929	71341	80488	89827	48277	07229	71953	16128	65074	28782
37	03248	18880	21667	01311	61806	80201	47889	83052	31029	06023
38	50583	17972	12690	00452	93766	16414	01212	27964	02766	28786
39	10636	46975	09449	45986	34672	46916	63881	83117	53947	95218
40	43896	41278	42205	10425	66560	59967	90139	73563	29875	79033
41	76714	80963	74907	16890	15492	27489	06067	22287	19760	13056
42	22393	46719	02083	62428	45177	57562	49243	31748	64278	05731
43	70942	92042	22776	47761	13503	16037	30875	80754	47491	96012
44	92011	60326	86346	26738	01983	04186	41388	03848	78354	14964
45	66456	00126	45685	67607	70796	04889	98128	13599	93710	23974
46	96292	44348	20898	02227	76512	53185	03057	61375	10760	26889
47	19680	07146	53951	10935	23333	76233	13706	20502	60405	09745
48	67347	51442	24536	60151	05498	64678	87569	65066	17790	55413
49	95888	59255	06898	99137	50871	81265	42223	83303	48694	81953

Tabla A.1 Diez mil dígitos aleatorizados (continuación)

	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
00	70896	44520	64720	49898	78088	76740	47460	83150	78905	59870
01	56809	42909	25853	47624	29486	14196	75841	00393	42390	24847
02	66109	84775	07515	49949	61482	91836	48126	80778	21302	24975
03	18071	36263	14053	52526	44347	04923	68100	57805	19521	15345
04	98732	15120	91754	12657	74675	78500	01247	49719	47635	55514
05	36075	83967	22268	77971	31169	68584	21336	72541	66959	39708
06	04110	45061	78062	18911	27855	09419	56459	00695	70323	04538
07	75658	58509	24479	10202	13150	95946	55087	38398	18718	95561
08	87403	19142	27208	35149	34889	27003	14181	44813	17784	41036
09	00005	52142	65021	64438	69610	12154	98422	65320	79996	01935
10	43674	47103	48614	70823	78252	82403	93424	05236	54588	27757
11	68597	68874	35567	98463	99671	05634	81533	47406	17228	44455
12	91874	70208	06308	40719	02772	69589	79936	07514	44950	35190
13	73854	19470	53014	29375	62256	77488	74388	53949	49607	19816
14	65926	34117	55344	68155	38099	56009	03513	05926	35584	42328
15	40005	35246	49440	40295	44390	83043	26090	80201	02934	49260
16	46686	29890	14821	69783	34733	11803	64845	32065	14527	38702
17	02717	61518	39583	72863	50707	96115	07416	05041	36756	61065
18	17048	22281	35573	28944	96889	51823	57268	03866	27658	91950
19	75304	53248	42151	93928	17343	88322	28683	11252	10355	65175
20	97844	62947	62230	30500	92816	85232	27222	91701	11057	83257
21	07611	71163	82212	20653	21499	51496	40715	78952	33029	64207
22	47744	04603	44522	62783	39347	72310	41460	31052	40814	94297
23	54293	43576	88116	67416	34908	15238	40561	73940	56850	31078
24	67556	93979	73363	00300	11217	74405	18937	79000	68834	48307
25	86581	73041	95809	73986	49408	53316	90841	73808	53421	82315
26	28020	86282	83365	76600	11261	74354	20968	60770	12141	09539
27	42578	32471	37840	30872	75074	79027	57813	62831	54715	26693
28	47290	15997	86163	10571	81911	92124	92971	80860	41012	58666
29	24856	63911	13221	77028	06573	33667	30732	47280	12926	67276
30	16352	24836	60799	76281	83402	44709	78930	82969	84468	36910
31	89060	79852	97854	28324	39638	86936	06702	74304	39873	19496
32	07637	30412	04921	26471	09605	07355	20466	49793	40539	21077
33	37711	47786	37468	31963	16908	50283	80884	08252	72655	58926
34	82994	53232	58202	73318	62471	49650	15888	73370	98748	69181
35	31722	67288	12110	04776	15168	68862	92347	90789	66961	04162
36	93819	78050	19364	38037	25706	90879	05215	00260	14426	88207
37	65557	24496	04713	23688	26623	41356	47049	60676	72236	01214
38	88001	91382	05129	36041	10257	55558	89979	58061	28957	10701
39	96648	70303	18191	62404	26558	92804	15415	02865	52449	78509
40	04118	51573	59356	02426	35010	37104	98316	44602	96478	08433
41	19317	27753	39431	26996	04465	69695	61374	06317	42225	62025
42	37182	91221	17307	68507	85725	81898	22588	22241	80337	89033
43	82990	03607	29560	60413	59743	75000	03806	13741	79671	25416
44	97294	21991	11217	98087	79124	52275	31088	32085	23089	21498
45	86771	69504	13345	42544	59616	07867	78717	82840	74669	21515
46	26046	55559	12200	95106	56496	76662	44880	89457	84209	01332
47	39689	05999	92290	79024	70271	93352	90272	94495	26842	54477
48	83265	89573	01437	43786	52986	49041	17952	35035	88985	84671
49	15128	35791	11296	45319	06330	82027	90808	54351	43091	30387

Tabla A.1 Diez mil dígitos aleatorizados (*continuación*)

	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
50	54441	64681	93190	00993	62130	44484	46293	60717	50239	76319
51	08573	52937	84274	95106	89117	65849	41356	65549	78787	50442
52	81067	68052	14270	19718	88499	63303	13533	91882	51136	60828
53	39737	58891	75278	98046	52284	40164	72442	77824	72900	14886
54	34958	76090	08827	61623	31114	86952	83645	91786	29633	78294
55	61417	72424	92626	71952	69709	81259	58472	43409	84454	88648
56	99187	14149	57474	32268	85424	90378	34682	47606	89295	02420
57	13130	13064	36485	48133	35319	05720	76317	70953	50823	06793
58	65563	11831	82402	46929	91446	72037	17205	89600	59084	55718
59	28737	49502	06060	52100	43704	50839	22538	56768	83467	19313
60	50353	74022	59767	49927	45882	74099	18758	57510	58560	07050
61	65208	96466	29917	22862	69972	35178	32911	08172	06277	62795
62	21323	38148	26696	81741	25131	20087	67452	19670	35898	50636
63	67875	29831	59330	46570	69768	36671	01031	95995	68417	68665
64	82631	26260	86554	31881	70512	37899	38851	40568	54284	24056
65	91989	39633	59039	12526	37730	68848	71399	28513	69018	10289
66	12950	31418	93425	69756	34036	55097	97241	92480	49745	42461
67	00328	27427	95474	97217	05034	26676	49629	13594	50525	13485
68	63986	16698	82804	04524	39919	32381	67488	05223	89537	59490
69	55775	75005	57912	20977	35722	51931	89565	77579	93085	06467
70	24761	56877	56357	78809	40748	69727	56652	12462	40528	75269
71	43820	80926	26795	57553	28319	25376	51795	26123	51102	89853
72	66669	02880	02987	33615	54206	20013	75872	88678	17726	60640
73	49944	66725	19779	50416	42800	71733	82052	28504	15593	51799
74	71003	87598	61296	95019	21568	86134	66096	65403	47166	78638
75	52715	04593	69484	93411	38046	13000	04293	60830	03914	75357
76	21998	31729	89963	11573	49442	69467	40265	56066	36024	25705
77	58970	96827	18377	31564	23555	86338	79250	43168	96929	97732
78	67592	59149	42554	42719	13553	48560	81167	10747	92552	19867
79	18298	18429	09357	96436	11237	88039	81020	00428	75731	37779
80	88420	28841	42628	84647	59024	52032	31251	72017	43875	48320
81	07627	88424	23381	29680	14027	75905	27037	22113	77873	78711
82	37917	93581	04979	21041	95252	62450	05937	81670	44894	47262
83	14783	95119	68464	08726	74818	91700	05961	23554	74649	50540
84	05378	32640	64562	15303	13168	23189	88198	63617	58566	56047
85	19640	96709	22047	07825	40583	99500	39989	96593	32254	37158
86	20514	11081	51131	56469	33947	77703	35679	45774	06776	67062
87	96763	56249	81243	62416	84451	14696	38195	70435	45948	67690
88	49439	61075	31558	59740	52759	55323	95226	01385	20158	54054
89	16294	50548	71317	32168	86071	47314	65393	56367	46910	51269
90	31381	94301	79273	32843	05862	36211	93960	00671	67631	23952
91	98032	87203	03227	66021	99666	98368	39222	36056	81992	20121
92	40700	31826	94774	11366	81391	33602	69608	84119	93204	26825
93	68692	66849	29366	77540	14978	06508	10824	65416	23629	63029
94	19047	10784	19607	20296	31804	72984	60060	50353	23260	58909
95	82867	69266	50733	62630	00956	61500	89913	30049	82321	62367
96	26528	28928	52600	72997	80943	04084	86662	90025	14360	64867
97	51166	00607	49962	30724	81707	14548	25844	47336	57492	02207
98	97245	15440	55182	15368	85136	98869	33712	95152	50973	98658
99	54998	88830	95639	45104	72676	28220	82576	57381	34438	24565

Fuente: Preparado por Fred Gruenberger, Numerical Analysis Laboratory, University of Wisconsin, Madison, Wisconsin, 1952.

Tabla A.1 Diez mil dígitos aleatorizados (continuación)

	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
50	58649	85086	16502	97541	76611	94229	34987	86718	87208	05426
51	97306	52449	55596	66739	36525	97563	29469	31235	79276	10831
52	09942	79344	78160	11015	55777	22047	57615	15717	86239	36578
53	83842	28631	74893	47911	92170	38181	30416	54860	44120	73031
54	73778	30395	20163	76111	13712	33449	99224	18206	51418	70006
55	88381	56550	47467	59663	61117	39716	32927	06168	06217	45477
56	31044	21404	15968	21357	30772	81482	38807	67231	84283	63552
57	00909	63837	91328	81106	11740	50193	86806	21931	18054	49601
58	69882	37028	41732	37425	80832	03320	20690	32653	90145	03029
59	26059	78324	22501	73825	16927	31545	15695	74216	98372	28547
60	38573	98078	38982	33078	93524	45606	53463	20391	81637	37269
61	70624	00063	81455	16924	12848	23801	55481	78978	26795	10553
62	49806	23976	05640	29804	38988	25024	76951	02341	63219	75864
63	05461	67523	48316	14613	08541	35231	38312	14969	67279	50502
64	76582	62153	53801	51219	30424	32599	49099	83959	68408	20147
65	16660	80470	75062	75588	24384	27874	20018	11428	32265	07692
66	60166	42424	97470	88451	81270	80070	72959	26220	59939	31127
67	28953	03272	31460	41691	57736	72052	22762	96323	27616	53123
68	47536	86439	95210	96386	38704	15484	07426	70675	06888	81203
69	73457	26657	36983	72410	30244	97711	25652	09373	66218	64077
70	11190	66193	66287	09116	48140	37669	02932	50799	17255	06181
71	57062	78964	44455	14036	36098	40773	11688	33150	07459	36127
72	99624	67254	67302	18991	97687	54099	94884	42283	63258	50651
73	97521	83669	85968	16135	30133	51312	17831	75016	80278	68953
74	40273	04838	13661	64757	17461	78085	60094	27010	80945	66439
75	57260	06176	49963	29760	69546	61336	39429	41985	18572	98128
76	03451	47098	63495	71227	79304	29753	99131	18419	71791	81515
77	62331	20492	15393	84270	24396	32962	21632	92965	38670	44923
78	32290	51079	06512	38806	93327	80086	19088	59887	98416	24918
79	28014	80428	92853	31333	32648	16734	43418	90124	15086	48444
80	18950	16091	29543	65817	07002	73115	94115	20271	50250	25061
81	17403	69503	01866	13049	07263	13039	83844	80143	39048	62654
82	27999	50489	66613	21843	71746	65868	16208	46781	93402	12323
83	87076	53174	12165	84495	47947	60706	64034	31635	65169	93070
84	89044	45974	14524	46906	26052	51851	84197	61694	57429	63395
85	98048	64400	24705	75711	36232	57624	41424	77366	52790	84705
86	09345	12956	49770	80311	32319	48238	16952	92088	51222	82865
87	07086	77628	76195	47584	62411	40397	71857	54823	26536	56792
88	93128	25657	46872	11206	06831	87944	97914	64670	45760	34353
89	85137	70964	29947	27795	25547	37682	96105	26848	09389	64326
90	32798	39024	13814	98546	46585	84108	74603	94812	73968	68766
91	62496	26371	89880	52078	47781	95260	83464	65942	91761	53727
92	62707	81825	40987	97656	89714	52177	23778	07482	91678	40128
93	05500	28982	86124	19554	80818	94935	61924	31828	79369	23507
94	79476	31445	59498	85132	24582	26024	24002	63718	79164	43556
95	10653	29954	97568	91541	33139	84525	72271	02546	64818	14381
96	30524	06495	00886	40666	68574	49574	19705	16429	90981	08103
97	69050	22019	74066	14500	14506	06423	38332	34191	82663	85323
98	27908	78802	63446	07674	98871	63831	72449	42705	26513	19883
99	64520	16618	47409	19574	78136	46047	01277	79146	95759	36781

**Tabla A.2** Valores de la razón, amplitud dividido por la desviación estándar  $\sigma$ , para tamaños de muestras desde 20 hasta 1 000

Número en la muestra	Amplitud/ $\sigma$	Número en la muestra	Amplitud/ $\sigma$
20	3.7	200	5.5
30	4.1	300	5.8
50	4.5	400	5.9
70	4.8	500	6.1
100	5.0	700	6.3
150	5.3	1,000	6.5

*Fuente:* Resumido con autorización de los fideicomisarios y de los editores de *Biometrika*, tomada de E. S. Pearson y H. O. Hartley, *Biometrika Tables for Statisticians*, vol. 1, Cambridge University Press, 1954. Tabla original de L. H. C. Tippett, "On the extreme individuals and the range of samples taken from a normal population" *Biometrika*, 17: págs., 364-387, (1925).

Tabla A.3 Valores de  $t$ 

$gl$	Probabilidad de un valor más alto de $t$								
	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	.679	.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
$\infty$	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291
$gl$	0.25	0.2	0.15	0.1	0.05	0.025	0.01	0.005	0.0005
	Probabilidad de un valor positivo más alto de $t$								

Fuente: Esta tabla es un resumen de la Tabla III de Fisher y Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, publicada por Oliver y Boyd Ltd, Edinburgh, con autorización de los autores y editores.

**Tabla A.4 Probabilidad de hallar un valor al azar de  $Z = (Y - \mu)/\sigma$  mayor que los valores tabulados en los márgenes**

Tabla A.5 Valores de  $\chi^2$ 

$g^1$	Probabilidad de un valor más alto de $\chi^2$												
	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005
1	0.393	0.157	0.082	0.0393	0.0168	.102	.455	.132	.271	.384	.502	6.63	7.88
2	0.100	0.201	0.606	.103	.211	.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	0.0717	0.115	.216	.352	.584	1.21	2.37	4.11	6.25	7.81	9.36	11.3	12.8
4	0.0445	0.077	.484	.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9
5	0.027	0.042	.554	.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	16.7
6	0.012	0.018	.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8
7	0.009	0.014	1.24	1.60	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5
8	0.007	0.011	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1
9	0.006	0.009	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	21.7
10	0.005	0.007	2.16	2.56	3.26	3.94	4.87	6.74	9.34	12.6	16.0	18.3	23.6
11	0.004	0.006	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	22.0
12	0.003	0.005	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3
13	0.002	0.004	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7
14	0.002	0.003	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	27.7
15	0.001	0.002	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	29.1
16	0.001	0.002	5.14	5.81	6.91	7.96	9.34	11.9	16.3	19.4	23.5	26.3	30.6
17	0.001	0.001	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	32.8
18	0.001	0.001	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	34.8
19	0.001	0.001	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	37.2
20	0.001	0.001	7.43	8.26	9.60	10.9	12.4	15.5	19.3	23.8	28.4	31.4	38.6
21	0.001	0.001	8.03	8.90	10.3	11.6	13.1	16.3	20.3	24.9	29.6	32.7	37.6
22	0.001	0.001	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	37.7
23	0.001	0.001	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	41.6
24	0.001	0.001	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	44.2
25	0.001	0.001	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	46.9
26	0.001	0.001	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.4
27	0.001	0.001	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2
28	0.001	0.001	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5
29	0.001	0.001	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7
30	0.001	0.001	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	49.6
31	0.001	0.001	14.5	15.8	17.6	19.3	21.4	25.3	30.2	35.6	41.1	45.0	52.3
32	0.001	0.001	15.2	16.6	18.4	20.1	22.2	26.1	31.0	36.2	41.9	46.0	53.7
33	0.001	0.001	16.0	17.4	19.2	20.9	23.0	26.9	31.8	37.0	42.8	47.0	56.9
34	0.001	0.001	16.8	18.2	20.0	21.7	23.8	27.7	32.6	38.4	44.3	49.2	57.6
35	0.001	0.001	17.5	19.0	20.8	22.5	24.6	28.5	33.4	39.2	45.0	50.8	59.5
36	0.001	0.001	18.2	19.7	21.4	23.1	25.2	29.1	34.0	39.8	45.6	51.4	60.9
37	0.001	0.001	18.9	20.4	22.1	23.8	25.9	29.8	34.7	40.5	46.3	52.1	61.8
38	0.001	0.001	19.6	21.1	22.8	24.5	26.6	30.5	35.4	41.2	47.0	52.8	62.5
39	0.001	0.001	20.3	21.8	23.5	25.2	27.3	31.2	36.1	41.9	47.7	53.5	63.2
40	0.001	0.001	21.0	22.2	24.4	26.5	29.1	33.7	39.3	45.6	51.8	57.8	66.8
41	0.001	0.001	21.7	22.9	25.1	27.2	29.7	34.3	40.9	47.1	53.3	59.3	67.7
42	0.001	0.001	22.4	23.6	25.8	27.9	30.4	35.0	41.6	47.8	54.0	60.5	69.2
43	0.001	0.001	23.1	24.3	26.5	28.6	31.1	35.7	42.3	48.5	55.2	61.9	70.9
44	0.001	0.001	23.8	25.0	27.2	29.3	31.8	36.4	43.0	49.2	56.0	62.7	72.5
45	0.001	0.001	24.5	25.7	27.9	29.9	32.4	37.0	43.6	50.0	56.7	63.4	74.2
46	0.001	0.001	25.2	26.4	28.6	30.7	33.2	37.7	44.3	50.7	57.4	64.1	75.9
47	0.001	0.001	25.9	27.1	29.3	31.4	33.9	38.4	45.0	51.4	58.1	64.8	76.5
48	0.001	0.001	26.6	27.8	29.9	32.1	34.6	39.1	45.7	52.1	58.8	65.5	77.2
49	0.001	0.001	27.3	28.5	30.6	32.8	35.1	39.6	46.3	52.7	59.4	66.1	77.9
50	0.001	0.001	28.0	29.2	31.3	33.4	35.8	40.3	47.0	53.4	60.1	66.8	78.6
51	0.001	0.001	28.7	29.9	32.0	34.1	36.5	41.0	47.7	54.1	60.8	67.5	79.3
52	0.001	0.001	29.4	30.6	32.7	34.8	37.1	41.6	48.3	54.7	61.4	68.1	79.0
53	0.001	0.001	30.1	31.3	33.4	35.5	37.8	42.5	49.2	55.6	62.3	69.0	80.7
54	0.001	0.001	30.8	32.0	34.1	36.2	38.5	43.2	50.0	56.3	63.0	69.7	81.4
55	0.001	0.001	31.5	32.7	34.8	36.9	39.1	43.9	50.7	57.0	63.7	70.4	82.1
56	0.001	0.001	32.2	33.4	35.5	37.6	39.8	44.6	51.4	57.7	64.4	71.1	82.8
57	0.001	0.001	32.9	34.1	36.2	38.3	40.5	45.3	52.1	58.4	65.1	71.8	83.5
58	0.001	0.001	33.6	34.8	36.9	39.0	41.2	46.0	52.8	59.1	65.8	72.5	84.2
59	0.001	0.001	34.3	35.5	37.6	39.7	41.9	46.7	53.5	59.8	66.5	73.2	84.9
60	0.001	0.001	35.0	36.2	38.3	40.4	42.6	47.4	54.2	60.5	67.2	74.0	85.6

Fuente: Esta tabla es un extracto de "Table of percentage points of the  $\chi^2$  distribution", Biometrika, 32: 188-189 (1941), de Catherine M. Thompson. Se publica aquí con la gentil autorización de la autora y del editor de Biometrika.

Tabla A.6 Valores de  $F$ 

$g_f$ del denominador	Probabilidad de un valor más alto de $F$	$g_f$ del numerador								
		1	2	3	4	5	6	7	8	9
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
	.025	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3
	.010	4052	4999.5	5403	5625	5764	5859	5928	5982	6022
	.005	16211	20000	21615	22500	23056	23437	23715	23925	24091
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.005	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	.005	53.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88
4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	.005	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14
5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	.005	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77
6	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
	.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
	.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
	.005	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39
7	.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
	.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	.005	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51
8	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	.005	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34
9	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	.005	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54
10	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
	.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
	.005	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97
11	.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
	.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
	.010	9.63	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
	.005	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54
12	.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	.005	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20
13	.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
	.050	4.57	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
	.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31
	.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
	.005	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94
14	.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
	.050	4.80	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
	.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21
	.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	.005	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72

Tabla A.6 Valores de  $F$  (continuación)

gl del numerador												
10	12	15	20	24	30	40	60	120	$\infty$	P	gl	
60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33	.100	1	
241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3	.050		
968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018	.025		
6056	6106	6157	6209	6235	6261	6287	6313	6339	6366	.010		
24224	24426	24630	24836	24940	25044	25148	25253	25359	25465	.005		
9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49	.100	2	
19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50	.050		
39.40	39.41	39.43	39.45	39.46	39.47	39.48	39.49	39.50	39.50	.025		
99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50	.010		
199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5	199.5	.005		
5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13	.100	3	
8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53	.050		
14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90	.025		
27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13	.010		
43.69	43.39	43.08	42.78	42.62	42.47	42.31	41.99	41.83	41.83	.005		
3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76	.100	4	
5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63	.050		
8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26	.025		
14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46	.010		
20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.32	.005		
3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10	.100	5	
4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36	.050		
6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02	.025		
10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02	.010		
13.62	13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.14	.005		
2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72	.100	6	
4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67	.050		
5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85	.025		
7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88	.010		
10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.88	.005		
2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47	.100	7	
3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23	.050		
4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14	.025		
6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65	.010		
8.38	8.18	7.97	7.75	7.65	7.53	7.42	7.31	7.19	7.08	.005		
2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29	.100	8	
3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93	.050		
4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67	.025		
5.61	5.57	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86	.010		
7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	5.95	.005		
2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16	.100	9	
3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71	.050		
3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33	.025		
5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31	.010		
6.42	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.19	.005		
2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06	.100	10	
2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54	.050		
3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08	.025		
4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91	.010		
5.85	5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	4.64	.005		
2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97	.100	11	
2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40	.050		
3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88	.025		
4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60	.010		
5.42	5.24	5.05	4.86	4.76	4.65	4.55	4.44	4.34	4.23	.005		
2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90	.100	12	
2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30	.050		
3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72	.025		
4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36	.010		
5.09	4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.90	.005		
2.14	2.10	2.05	2.01	1.96	1.96	1.93	1.90	1.88	1.85	.100	13	
2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21	.050		
3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60	.025		
4.10	3.96	3.82	3.66	3.59	3.31	3.43	3.34	3.25	3.17	.010		
4.82	4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	3.65	.005		
2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80	.100	14	
2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13	.050		
3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49	.025		
3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00	.010		
4.60	4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	3.44	.005		

Tabla A.6 Valores de  $F$  (continuación)

gl del denominador	Probabilidad de un valor más alto de $F$	gl del numerador								
		1	2	3	4	5	6	7	8	9
15	.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
	.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12
	.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
	.005	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54
16	.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
	.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
	.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05
	.010	8.53	6.23	5.20	4.77	4.44	4.20	4.03	3.89	3.78
	.005	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38
17	.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
	.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
	.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98
	.010	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
	.005	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25
18	.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
	.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	.005	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14
19	.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
	.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
	.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
	.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	.005	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04
20	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	.005	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96
21	.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
	.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	.005	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88
22	.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
	.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	.005	9.73	6.81	5.55	5.02	4.61	4.32	4.11	3.94	3.81
23	.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
	.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
	.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	.005	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75
24	.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
	.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
	.005	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69
25	.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
	.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
	.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
	.005	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64
26	.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
	.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
	.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
	.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
	.005	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60
27	.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
	.050	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
	.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
	.005	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56
28	.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
	.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
	.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61
	.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
	.005	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52

Tabla A.6 Valores de  $F$  (continuación)

$g_f$ del numerador												
10	12	15	20	24	30	40	60	120	$\infty$	P	gl	
2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76	.100	15	
2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07	.050		
3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40	.025		
3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87	.010		
4.42	4.23	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.26	.005		
2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72	.100	16	
2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01	.050		
2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32	.025		
3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75	.010		
4.27	4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22	3.11	.005		
2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69	.100	17	
2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96	.050		
2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25	.025		
3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65	.010		
4.14	3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	2.98	.005		
1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66	.100	18	
2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92	.050		
2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19	.025		
3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57	.010		
4.03	3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99	2.87	.005		
1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63	.100	19	
2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88	.050		
2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13	.025		
3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49	.010		
3.93	3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	2.78	.005		
1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61	.100	20	
2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84	.050		
2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09	.025		
3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42	.010		
3.85	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.69	.005		
1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59	.100	21	
2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81	.050		
2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04	.025		
3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36	.010		
3.77	3.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73	2.61	.005		
1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57	.100	22	
2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78	.050		
2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00	.025		
3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31	.010		
3.70	3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66	2.55	.005		
1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55	.100	23	
2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76	.050		
2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97	.025		
3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26	.010		
3.64	3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	2.48	.005		
1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53	.100	24	
2.25	2.19	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73	.050		
2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94	.025		
3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21	.010		
3.59	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.43	.005		
1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52	.100	25	
2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71	.050		
2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91	.025		
3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17	.010		
3.54	3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	2.38	.005		
1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50	.100	26	
2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69	.050		
2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88	.025		
3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13	.010		
3.49	3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	2.33	.005		
1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49	.100	27	
2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67	.050		
2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85	.025		
3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10	.010		
3.45	3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.29	.005		
1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48	.100	28	
2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65	.050		
2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83	.025		
3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06	.010		
3.41	3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.25	.005		

Tabla A.6 Valores de  $F$  (continuación)

$g_f$ del denominador	Probabilidad de un valor más alto de $F$	$g_f$ del numerador								
		1	2	3	4	5	6	7	8	9
29	.100	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
	.050	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
	.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59
	.010	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
	.005	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48
30	.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
	.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
	.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
	.005	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45
40	.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
	.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45
	.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
	.005	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22
60	.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
	.050	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
	.010	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
	.005	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01
120	.100	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68
	.050	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
	.025	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22
	.010	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
	.005	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81
30	.100	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63
	.050	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88
	.025	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11
	.010	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41
	.005	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62

Fuente: Una parte de "Tables of percentage points of the inverted beta ( $F$ ) distribution", *Biometrika*, vol. 33 (1943) por M. Merrington y C. M. Thompson y de la Tabla 18 de *Biometrika Tables for Statisticians*, vol. 1, Cambridge University Press, 1954, editado por E. S. Pearson y H. O. Hartley. Reproducido con permiso de los autores, editores, y de los fideicomisarios de *Biometrika*.

Tabla A.6 Valores de  $F$  (continuación)

gl del numerador												
10	12	15	20	24	30	40	60	120	120	$\infty$	P	gl
1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47	1.64	.050	20
2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64	1.81	.025	
2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81	2.03	.010	
3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.21	2.33	.005	
3.38	3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	2.21	2.31	.001	
1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46	1.62	.050	30
2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.79	1.79	.025	
2.51	2.41	2.31	2.21	2.14	2.07	2.01	1.94	1.87	2.01	2.01	.010	
2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.11	2.11	.005	
3.34	3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.18	2.18	.003	
1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38	1.51	.050	40
2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51	1.64	.025	
2.39	2.29	2.18	2.07	2.01	1.96	1.88	1.80	1.72	1.72	1.80	.010	
2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.92	1.93	.005	
3.12	2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	2.06	1.93	.003	
1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29	1.39	.050	60
1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.47	1.56	.025	
2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.67	1.67	1.69	.010	
2.63	2.50	2.35	2.20	2.12	2.05	1.94	1.84	1.73	1.73	1.69	.005	
2.90	2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	1.83	1.83	.003	
1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.26	1.25	.050	120
1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.45	1.45	1.45	1.45	.025	
2.16	2.05	1.95	1.82	1.76	1.69	1.64	1.53	1.53	1.53	1.53	.010	
2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.66	1.66	1.66	.005	
2.71	2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.75	1.75	1.75	.003	
1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.24	1.24	1.24	.050	80
1.83	1.75	1.67	1.57	1.52	1.46	1.40	1.34	1.34	1.34	1.34	.025	
2.05	1.94	1.83	1.71	1.64	1.57	1.51	1.43	1.43	1.43	1.43	.010	
2.32	2.18	2.04	1.88	1.79	1.70	1.64	1.57	1.57	1.57	1.57	.005	
2.52	2.36	2.19	2.00	1.90	1.87	1.79	1.72	1.72	1.72	1.72	.003	

Tabla A.7 Amplitudes studentizadas significativas para 5 y 1 por ciento de la nueva prueba de amplitud múltiple

Tabla A.7 Amplitudes studentizadas significativas para 5 y 1 por ciento de la nueva prueba de amplitud múltiple (Continuación)

$g^f$ del error	Nivel significativo	$\rho$ = número de medias para la amplitud a probarse																		
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
16	.05	3.00	3.15	3.23	3.30	3.34	3.37	3.39	3.41	3.43	3.45	3.46	3.47	3.47	3.47	3.47	3.47	3.47	3.47	3.47
	.01	4.13	4.34	4.45	4.54	4.60	4.67	4.72	4.76	4.79	4.84	4.88	4.91	4.93	4.94	4.94	4.94	4.94	4.94	4.94
17	.05	2.98	3.13	3.22	3.28	3.33	3.36	3.38	3.40	3.42	3.44	3.45	3.46	3.47	3.47	3.47	3.47	3.47	3.47	3.47
	.01	4.10	4.30	4.41	4.50	4.56	4.63	4.68	4.72	4.75	4.80	4.83	4.86	4.88	4.88	4.88	4.88	4.88	4.88	4.88
18	.05	2.97	3.12	3.21	3.27	3.32	3.35	3.37	3.39	3.41	3.43	3.45	3.46	3.47	3.47	3.47	3.47	3.47	3.47	3.47
	.01	4.07	4.27	4.38	4.46	4.53	4.59	4.64	4.68	4.71	4.76	4.79	4.82	4.84	4.85	4.85	4.85	4.85	4.85	4.85
19	.05	2.96	3.11	3.19	3.26	3.31	3.35	3.37	3.39	3.41	3.43	3.44	3.45	3.46	3.46	3.46	3.46	3.46	3.46	3.46
	.01	4.05	4.24	4.35	4.43	4.50	4.56	4.61	4.64	4.67	4.72	4.76	4.79	4.81	4.81	4.82	4.82	4.82	4.82	4.82
20	.05	2.95	3.10	3.18	3.25	3.30	3.34	3.36	3.38	3.40	3.43	3.44	3.45	3.46	3.46	3.46	3.46	3.46	3.46	3.46
	.01	4.02	4.22	4.33	4.40	4.47	4.53	4.58	4.61	4.65	4.69	4.73	4.76	4.78	4.79	4.79	4.79	4.79	4.79	4.79
22	.05	2.93	3.08	3.17	3.24	3.29	3.32	3.35	3.37	3.39	3.42	3.44	3.45	3.46	3.46	3.46	3.46	3.46	3.46	3.46
	.01	3.99	4.17	4.26	4.36	4.42	4.48	4.53	4.57	4.60	4.65	4.68	4.71	4.74	4.75	4.75	4.75	4.75	4.75	4.75
24	.05	2.92	3.07	3.15	3.22	3.28	3.31	3.34	3.37	3.38	3.41	3.44	3.45	3.46	3.46	3.46	3.46	3.46	3.46	3.46
	.01	3.96	4.14	4.24	4.33	4.39	4.44	4.49	4.53	4.57	4.62	4.64	4.67	4.70	4.72	4.72	4.72	4.72	4.72	4.72
26	.05	2.91	3.06	3.14	3.21	3.27	3.30	3.34	3.36	3.38	3.41	3.43	3.45	3.46	3.46	3.46	3.46	3.46	3.46	3.46
	.01	3.93	4.11	4.21	4.30	4.36	4.41	4.46	4.50	4.53	4.58	4.62	4.65	4.67	4.69	4.69	4.69	4.69	4.69	4.69
28	.05	2.90	3.04	3.13	3.20	3.26	3.30	3.33	3.35	3.37	3.40	3.43	3.45	3.46	3.46	3.46	3.46	3.46	3.46	3.46
	.01	3.91	4.08	4.18	4.28	4.34	4.39	4.43	4.47	4.51	4.56	4.60	4.62	4.65	4.67	4.69	4.69	4.69	4.69	4.69
30	.05	2.89	3.04	3.12	3.20	3.25	3.29	3.32	3.35	3.37	3.40	3.43	3.44	3.46	3.46	3.46	3.46	3.46	3.46	3.46
	.01	3.89	4.06	4.16	4.22	4.32	4.36	4.41	4.45	4.48	4.54	4.58	4.61	4.63	4.65	4.65	4.65	4.65	4.65	4.65
40	.05	2.86	3.01	3.10	3.17	3.22	3.27	3.30	3.33	3.35	3.39	3.42	3.44	3.46	3.46	3.46	3.46	3.46	3.46	3.46
	.01	3.82	3.99	4.10	4.17	4.24	4.30	4.34	4.37	4.41	4.46	4.51	4.54	4.57	4.59	4.59	4.59	4.59	4.59	4.59
60	.05	2.83	2.98	3.08	3.14	3.20	3.24	3.28	3.31	3.33	3.37	3.40	3.43	3.45	3.45	3.45	3.45	3.45	3.45	3.45
	.01	3.76	3.92	4.03	4.12	4.17	4.23	4.27	4.31	4.34	4.39	4.44	4.47	4.50	4.53	4.53	4.53	4.53	4.53	4.53
100	.05	2.80	2.95	3.05	3.12	3.18	3.22	3.26	3.29	3.32	3.36	3.40	3.42	3.45	3.45	3.45	3.45	3.45	3.45	3.45
	.01	3.71	3.86	3.98	4.06	4.11	4.17	4.21	4.25	4.29	4.35	4.38	4.42	4.45	4.48	4.48	4.48	4.48	4.48	4.48
400	.05	2.77	2.92	3.02	3.09	3.15	3.23	3.26	3.29	3.34	3.38	3.41	3.44	3.47	3.47	3.47	3.47	3.47	3.47	3.47
	.01	3.64	3.80	3.90	3.98	4.04	4.09	4.14	4.17	4.20	4.26	4.31	4.34	4.37	4.41	4.41	4.41	4.41	4.41	4.41

Fuente: Resumida de la publicación D. B. Duncan, "Multiple range and multiple  $F$  tests", *Biometrics*, 11: 1-42 (1955), con el permiso del editor y el autor.

Tabla A.8 Puntos porcentuales superiores de la amplitud studentizada,

$$q_\alpha = (\bar{Y}_{\max} - \bar{Y}_{\min})/s_{\bar{Y}}$$

gl del error	$\alpha$	$p = \text{número de}$									
		2	3	4	5	6	7	8	9	10	11
5	.05	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17
	.01	5.70	6.97	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48
6	.05	3.46	4.34	4.90	5.31	5.63	5.89	6.12	6.32	6.49	6.65
	.01	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30
7	.05	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30
	.01	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55
8	.05	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05
	.01	4.74	5.63	6.20	6.63	6.96	7.24	7.47	7.68	7.87	8.03
9	.05	3.20	3.95	4.42	4.76	5.02	5.24	5.43	5.60	5.74	5.87
	.01	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.32	7.49	7.65
10	.05	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72
	.01	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36
11	.05	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61
	.01	4.39	5.14	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13
12	.05	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.40	5.51
	.01	4.32	5.04	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94
13	.05	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43
	.01	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79
14	.05	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36
	.01	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66
15	.05	3.01	3.67	4.08	4.37	4.60	4.78	4.94	5.08	5.20	5.31
	.01	4.17	4.83	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55
16	.05	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26
	.01	4.13	4.78	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46
17	.05	2.98	3.63	4.02	4.30	4.52	4.71	4.86	4.99	5.11	5.21
	.01	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38
18	.05	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17
	.01	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31
19	.05	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14
	.01	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25
20	.05	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11
	.01	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19
24	.05	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01
	.01	3.96	4.54	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02
30	.05	2.89	3.49	3.84	4.10	4.30	4.46	4.60	4.72	4.83	4.92
	.01	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85
40	.05	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.74	4.82
	.01	3.82	4.37	4.70	4.93	5.11	5.27	5.39	5.50	5.60	5.69
60	.05	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73
	.01	3.76	4.28	4.60	4.82	4.99	5.13	5.25	5.36	5.45	5.53
120	.05	2.80	3.36	3.69	3.92	4.10	4.24	4.36	4.48	4.56	4.64
	.01	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.38
$\infty$	.05	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55
	.01	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23

Fuente: Esta tabla es un extracto de la Tabla 29, *Biometrika Tables for Statisticians*, vol. I, Cambridge University Press, 1954. Se produce con gentil autorización de los fideicomisarios y los editores de *Biometrika*, E. S. Pearson y H. O. Hartley. El trabajo original apareció en una publicación de J. M. May, "Extended and corrected tables of upper percentage points of 'Studentized' range", en *Biometrika*, 39: 192-193 (1952).

Tabla A.8 Puntos porcentuales superiores de la amplitud studentizada

 $q_\alpha = (\bar{Y}_{\max} - \bar{Y}_{\min})/s_y$  (continuación)

medias de tratamiento									$\alpha$	gl del error
12	13	14	15	16	17	18	19	20		
7.32	7.47	7.60	7.72	7.83	7.93	8.03	8.12	8.21	.05	5
10.70	10.89	11.08	11.24	11.40	11.55	11.68	11.81	11.93	.01	
6.79	6.92	7.03	7.14	7.24	7.34	7.43	7.51	7.59	.05	6
9.49	9.65	9.81	9.95	10.08	10.21	10.32	10.43	10.54	.01	
6.43	6.55	6.66	6.76	6.85	6.94	7.02	7.09	7.17	.05	7
8.71	8.86	9.00	9.12	9.24	9.35	9.46	9.55	9.65	.01	
6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.80	6.87	.05	8
8.18	8.31	8.44	8.55	8.66	8.76	8.85	8.94	9.03	.01	
5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.64	.05	9
7.78	7.91	8.03	8.13	8.23	8.32	8.41	8.49	8.57	.01	
5.83	5.93	6.03	6.11	6.20	6.27	6.34	6.40	6.47	.05	10
7.48	7.60	7.71	7.81	7.91	7.99	8.07	8.15	8.22	.01	
5.71	5.81	5.90	5.99	6.06	6.14	6.20	6.26	6.33	.05	11
7.25	7.36	7.46	7.56	7.65	7.73	7.81	7.88	7.95	.01	
5.62	5.71	5.80	5.88	5.95	6.03	6.09	6.15	6.21	.05	12
7.06	7.17	7.26	7.36	7.44	7.52	7.59	7.66	7.73	.01	
5.53	5.63	5.71	5.79	5.86	5.93	6.00	6.05	6.11	.05	13
6.90	7.01	7.10	7.19	7.27	7.34	7.42	7.48	7.55	.01	
5.46	5.55	5.64	5.72	5.79	5.85	5.92	5.97	6.03	.05	14
6.77	6.87	6.96	7.05	7.12	7.20	7.27	7.33	7.39	.01	
5.40	5.49	5.58	5.65	5.72	5.79	5.85	5.90	5.96	.05	15
6.66	6.76	6.84	6.93	7.00	7.07	7.14	7.20	7.26	.01	
5.35	5.44	5.52	5.59	5.66	5.72	5.79	5.84	5.90	.05	16
6.56	6.66	6.74	6.82	6.90	6.97	7.03	7.09	7.15	.01	
5.31	5.39	5.47	5.55	5.61	5.68	5.74	5.79	5.84	.05	17
6.48	6.57	6.66	6.73	6.80	6.87	6.94	7.00	7.05	.01	
5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79	.05	18
6.41	6.50	6.58	6.65	6.72	6.79	6.85	6.91	6.96	.01	
5.23	5.32	5.39	5.46	5.53	5.59	5.65	5.70	5.75	.05	19
6.34	6.43	6.51	6.58	6.65	6.72	6.78	6.84	6.89	.01	
5.20	5.28	5.36	5.43	5.49	5.55	5.61	5.66	5.71	.05	20
6.29	6.37	6.45	6.52	6.59	6.65	6.71	6.76	6.82	.01	
5.10	5.18	5.25	5.32	5.38	5.44	5.50	5.54	5.59	.05	24
6.11	6.19	6.26	6.33	6.39	6.45	6.51	6.56	6.61	.01	
5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.48	.05	30
5.93	6.01	6.08	6.14	6.20	6.26	6.31	6.36	6.41	.01	
4.91	4.98	5.05	5.11	5.16	5.22	5.27	5.31	5.36	.05	40
5.77	5.84	5.90	5.96	6.02	6.07	6.12	6.17	6.21	.01	
4.81	4.88	4.94	5.00	5.06	5.11	5.16	5.20	5.24	.05	60
5.60	5.67	5.73	5.79	5.84	5.89	5.93	5.98	6.02	.01	
4.72	4.78	4.84	4.90	4.95	5.00	5.05	5.09	5.13	.05	120
5.44	5.51	5.56	5.61	5.66	5.71	5.75	5.79	5.83	.01	
4.62	4.68	4.74	4.80	4.85	4.89	4.93	4.97	5.01	.05	
5.29	5.35	5.40	5.45	5.49	5.54	5.57	5.61	5.65	.01	$\infty$

**Tabla A.9A Tabla de  $t$  para comparaciones de una cola entre  $p$  medias de tratamiento y un control para un coeficiente de confianza conjunto de  $P = 0.95$  y  $P = 0.99$**

$gl$ del error	$P$	$p =$ número de medias de tratamiento, sin incluir el control								
		1	2	3	4	5	6	7	8	9
5	.95	2.02	2.44	2.68	2.85	2.98	3.08	3.16	3.24	3.30
	.99	3.37	3.90	4.21	4.43	4.60	4.73	4.85	4.94	5.03
6	.95	1.94	2.34	2.56	2.71	2.83	2.92	3.00	3.07	3.12
	.99	3.14	3.61	3.88	4.07	4.21	4.33	4.43	4.51	4.59
7	.95	1.89	2.27	2.48	2.62	2.73	2.82	2.89	2.95	3.01
	.99	3.00	3.42	3.66	3.83	3.96	4.07	4.15	4.23	4.30
8	.95	1.86	2.22	2.42	2.55	2.66	2.74	2.81	2.87	2.92
	.99	2.90	3.29	3.51	3.67	3.79	3.88	3.96	4.03	4.09
9	.95	1.83	2.18	2.37	2.50	2.60	2.68	2.75	2.81	2.86
	.99	2.82	3.19	3.40	3.55	3.66	3.75	3.82	3.89	3.94
10	.95	1.81	2.15	2.34	2.47	2.56	2.64	2.70	2.76	2.81
	.99	2.76	3.11	3.31	3.45	3.56	3.64	3.71	3.78	3.83
11	.95	1.80	2.13	2.31	2.44	2.53	2.60	2.67	2.72	2.77
	.99	2.72	3.06	3.25	3.38	3.48	3.56	3.63	3.69	3.74
12	.95	1.78	2.11	2.29	2.41	2.50	2.58	2.64	2.69	2.74
	.99	2.68	3.01	3.19	3.32	3.42	3.50	3.56	3.62	3.67
13	.95	1.77	2.09	2.27	2.39	2.48	2.55	2.61	2.66	2.71
	.99	2.65	2.97	3.15	3.27	3.37	3.44	3.51	3.56	3.61
14	.95	1.76	2.08	2.25	2.37	2.46	2.53	2.59	2.64	2.69
	.99	2.62	2.94	3.11	3.23	3.32	3.40	3.46	3.51	3.56
15	.95	1.75	2.07	2.24	2.36	2.44	2.51	2.57	2.62	2.67
	.99	2.60	2.91	3.08	3.20	3.29	3.36	3.42	3.47	3.52
16	.95	1.75	2.06	2.23	2.34	2.43	2.50	2.56	2.61	2.65
	.99	2.58	2.88	3.05	3.17	3.26	3.33	3.39	3.44	3.48
17	.95	1.74	2.05	2.22	2.33	2.42	2.49	2.54	2.59	2.64
	.99	2.57	2.86	3.03	3.14	3.23	3.30	3.36	3.41	3.45
18	.95	1.73	2.04	2.21	2.32	2.41	2.48	2.53	2.58	2.62
	.99	2.55	2.84	3.01	3.12	3.21	3.27	3.33	3.38	3.42
19	.95	1.73	2.03	2.20	2.31	2.40	2.47	2.52	2.57	2.61
	.99	2.54	2.83	2.99	3.10	3.18	3.25	3.31	3.36	3.40
20	.95	1.72	2.03	2.19	2.30	2.39	2.46	2.51	2.56	2.60
	.99	2.53	2.81	2.97	3.08	3.17	3.23	3.29	3.34	3.38
24	.95	1.71	2.01	2.17	2.28	2.36	2.43	2.48	2.53	2.57
	.99	2.49	2.77	2.92	3.03	3.11	3.17	3.22	3.27	3.31
30	.95	1.70	1.99	2.15	2.25	2.33	2.40	2.45	2.50	2.54
	.99	2.46	2.72	2.87	2.97	3.05	3.11	3.16	3.21	3.24
40	.95	1.68	1.97	2.13	2.23	2.31	2.37	2.42	2.47	2.51
	.99	2.42	2.68	2.82	2.92	2.99	3.05	3.10	3.14	3.18
60	.95	1.67	1.95	2.10	2.21	2.28	2.35	2.39	2.44	2.48
	.99	2.39	2.64	2.78	2.87	2.94	3.00	3.04	3.08	3.12
120	.95	1.66	1.93	2.08	2.18	2.26	2.32	2.37	2.41	2.45
	.99	2.36	2.60	2.73	2.82	2.89	2.94	2.99	3.03	3.06
$\infty$	.95	1.64	1.92	2.06	2.16	2.23	2.29	2.34	2.38	2.42
	.99	2.33	2.56	2.68	2.77	2.84	2.89	2.93	2.97	3.00

Fuente: Esta tabla se reproduce de "A multiple comparison procedure for comparing several treatments with a control", J. Am. Stat. Assn., 50: 1096-1121 (1955), con autorización del autor, C. W. Dunnett, y del editor.

**Tabla A.9B Tabla de  $t$  para comparaciones de dos colas entre  $p$  tratamientos y un control para un coeficiente de confianza conjunto de  $P = 0.95$  y  $P = 0.99$**

$gl$ del error	$P$	$p$ = número de medias de tratamiento, sin incluir el control								
		1	2	3	4	5	6	7	8	9
5	.95	2.57	3.03	3.39	3.66	3.88	4.06	4.22	4.36	4.49
	.99	4.03	4.63	5.09	5.44	5.73	5.97	6.18	6.36	6.53
6	.95	2.45	2.86	3.18	3.41	3.60	3.75	3.88	4.00	4.11
	.99	3.71	4.22	4.60	4.88	5.11	5.30	5.47	5.61	5.74
7	.95	2.36	2.75	3.04	3.24	3.41	3.54	3.66	3.76	3.86
	.99	3.50	3.95	4.28	4.52	4.71	4.87	5.01	5.13	5.24
8	.95	2.31	2.67	2.94	3.13	3.28	3.40	3.51	3.60	3.68
	.99	3.36	3.77	4.06	4.27	4.44	4.58	4.70	4.81	4.90
9	.95	2.26	2.61	2.86	3.04	3.18	3.29	3.39	3.48	3.55
	.99	3.25	3.63	3.90	4.09	4.24	4.37	4.48	4.57	4.65
10	.95	2.23	2.57	2.81	2.97	3.11	3.21	3.31	3.39	3.46
	.99	3.17	3.53	3.78	3.95	4.10	4.21	4.31	4.40	4.47
11	.95	2.20	2.53	2.76	2.92	3.05	3.15	3.24	3.31	3.38
	.99	3.11	3.45	3.68	3.85	3.98	4.09	4.18	4.26	4.33
12	.95	2.18	2.50	2.72	2.88	3.00	3.10	3.18	3.25	3.32
	.99	3.05	3.39	3.61	3.76	3.89	3.99	4.08	4.15	4.22
13	.95	2.16	2.48	2.69	2.84	2.96	3.06	3.14	3.21	3.27
	.99	3.01	3.33	3.54	3.69	3.81	3.91	3.99	4.06	4.13
14	.95	2.14	2.46	2.67	2.81	2.93	3.02	3.10	3.17	3.23
	.99	2.98	3.29	3.49	3.64	3.75	3.84	3.92	3.99	4.05
15	.95	2.13	2.44	2.64	2.79	2.90	2.99	3.07	3.13	3.19
	.99	2.95	3.25	3.45	3.59	3.70	3.79	3.86	3.93	3.99
16	.95	2.12	2.42	2.63	2.77	2.88	2.96	3.04	3.10	3.16
	.99	2.92	3.22	3.41	3.55	3.65	3.74	3.82	3.88	3.93
17	.95	2.11	2.41	2.61	2.75	2.85	2.94	3.01	3.08	3.13
	.99	2.90	3.19	3.38	3.51	3.62	3.70	3.77	3.83	3.89
18	.95	2.10	2.40	2.59	2.73	2.84	2.92	2.99	3.05	3.11
	.99	2.88	3.17	3.35	3.48	3.58	3.67	3.74	3.80	3.85
19	.95	2.09	2.39	2.58	2.72	2.82	2.90	2.97	3.04	3.09
	.99	2.86	3.15	3.33	3.46	3.55	3.64	3.70	3.76	3.81
20	.95	2.09	2.38	2.57	2.70	2.81	2.89	2.96	3.02	3.07
	.99	2.85	3.13	3.31	3.43	3.53	3.61	3.67	3.73	3.78
24	.95	2.06	2.35	2.53	2.66	2.76	2.84	2.91	2.96	3.01
	.99	2.80	3.07	3.24	3.36	3.45	3.52	3.58	3.64	3.69
30	.95	2.04	2.32	2.50	2.62	2.72	2.79	2.86	2.91	2.96
	.99	2.75	3.01	3.17	3.28	3.37	3.44	3.50	3.55	3.59
40	.95	2.02	2.29	2.47	2.58	2.67	2.75	2.81	2.86	2.90
	.99	2.70	2.95	3.10	3.21	3.29	3.36	3.41	3.46	3.50
60	.95	2.00	2.27	2.43	2.55	2.63	2.70	2.76	2.81	2.85
	.99	2.66	2.90	3.04	3.14	3.22	3.28	3.33	3.38	3.42
120	.95	1.98	2.24	2.40	2.51	2.59	2.66	2.71	2.76	2.80
	.99	2.62	2.84	2.98	3.08	3.15	3.21	3.25	3.30	3.33
$\infty$	.95	1.96	2.21	2.37	2.47	2.55	2.62	2.67	2.71	2.75
	.99	2.58	2.79	2.92	3.01	3.08	3.14	3.18	3.22	3.25

Fuente: Esta tabla se reproduce de "A multiple comparison procedure for comparing several treatments with a control", *J. Am. Stat. Assn.*, 50: 1096-1121 (1955), con autorización del autor, C. W. Dunnett, y del editor.

Tabla A.10 Transformación de la  $\sqrt{\text{porcentaje arco seno}}$ 

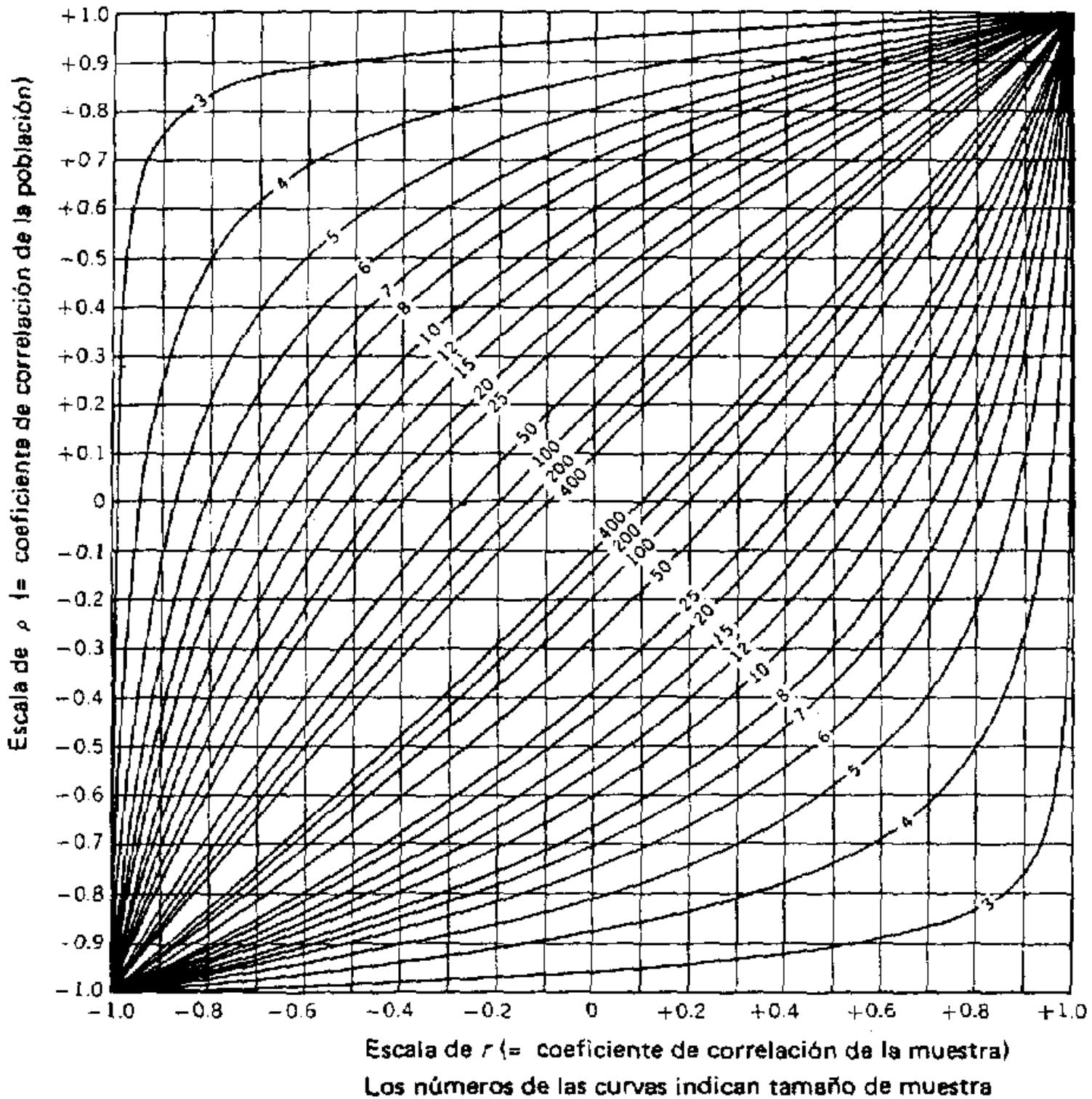
Transformación de porcentajes binomiales, en los márgenes, a ángulos de igual información en grados. Los signos + y - que siguen a los valores angulares que terminan en 5 son la guía en la aproximación a un decimal.

%	0	1	2	3	4	5	6	7	8	9
0.0	0	0.57	0.81	0.99	1.15-	1.28	1.40	1.52	1.62	1.72
0.1	1.81	1.90	1.99	2.07	2.14	2.22	2.29	2.36	2.43	2.50
0.2	2.56	2.53	2.69	2.75-	2.81	2.87	2.92	2.98	3.03	3.09
0.3	3.14	3.19	3.24	3.29	3.34	3.39	3.44	3.49	3.53	3.58
0.4	3.63	3.67	3.72	3.76	3.80	3.85-	3.89	3.93	3.97	4.01
0.5	4.05+	4.09	4.13	4.17	4.21	4.25+	4.29	4.33	4.37	4.40
0.6	4.44	4.48	4.52	4.55+	4.59	4.62	4.66	4.69	4.73	4.76
0.7	4.80	4.83	4.87	4.90	4.93	4.97	5.00	5.03	5.07	5.10
0.8	5.13	5.16	5.20	5.23	5.26	5.29	5.32	5.35+	5.38	5.41
0.9	5.44	5.47	5.50	5.53	5.56	5.59	5.62	5.65+	5.68	5.71
1	5.74	6.02	6.29	6.55-	6.80	7.04	7.27	7.49	7.71	7.92
2	8.13	8.33	8.53	8.72	8.91	9.10	9.28	9.46	9.63	9.81
3	9.98	10.14	10.31	10.47	10.63	10.78	10.94	11.09	11.24	11.39
4	11.54	11.68	11.83	11.97	12.11	12.25-	12.39	12.52	12.66	12.79
5	12.92	13.05+	13.18	13.31	13.44	13.56	13.69	13.81	13.94	14.06
6	14.18	14.30	14.42	14.54	14.65+	14.77	14.89	15.00	15.12	15.23
7	15.34	15.45+	15.56	15.68	15.79	15.89	16.00	16.11	16.22	16.32
8	16.43	16.54	16.64	16.74	16.85-	16.95+	17.05+	17.16	17.26	17.35
9	17.46	17.56	17.66	17.76	17.85+	17.95+	18.05-	18.15-	18.24	18.34
10	18.44	18.53	18.63	18.72	18.81	18.91	19.00	19.09	19.19	19.28
11	19.37	19.46	19.55+	19.64	19.73	19.82	19.91	20.00	20.09	20.18
12	20.27	20.36	20.44	20.53	20.62	20.70	20.79	20.88	20.96	21.05-
13	21.13	21.22	21.30	21.39	21.47	21.56	21.64	21.72	21.81	21.89
14	21.97	22.06	22.14	22.22	22.30	22.38	22.46	22.55-	22.63	22.71
15	22.79	22.87	22.95-	23.03	23.11	23.19	23.26	23.34	23.42	23.50
16	23.58	23.66	23.73	23.81	23.89	23.97	24.04	24.12	24.20	24.27
17	24.35+	24.43	24.50	24.58	24.65+	24.73	24.80	24.88	24.95+	25.03
18	25.10	25.18	25.25+	25.33	25.40	25.48	25.55-	25.62	25.70	25.77
19	25.84	25.92	25.99	26.06	26.13	26.21	26.28	26.35-	26.42	26.49
20	26.56	26.64	26.71	26.78	26.85+	26.92	26.99	27.06	27.13	27.20
21	27.28	27.35-	27.42	27.49	27.56	27.63	27.69	27.76	27.83	27.90
22	27.97	28.04	28.11	28.18	28.25-	28.32	28.38	28.45+	28.52	28.59
23	28.66	28.73	28.79	28.86	28.93	29.00	29.06	29.13	29.20	29.27
24	29.33	29.40	29.47	29.53	29.60	29.67	29.73	29.80	29.87	29.93
25	30.00	30.07	30.13	30.20	30.26	30.33	30.40	30.46	30.53	30.59
26	30.66	30.72	30.79	30.85+	30.92	30.98	31.05-	31.11	31.18	31.24
27	31.31	31.37	31.44	31.50	31.56	31.63	31.69	31.76	31.82	31.88
28	31.95-	32.01	32.08	32.14	32.20	32.27	32.33	32.39	32.46	32.52
29	32.58	32.65-	32.71	32.77	32.83	32.90	32.96	33.02	33.09	33.15-
30	33.21	33.27	33.34	33.40	33.46	33.52	33.58	33.65-	33.71	33.77
31	33.83	33.89	33.96	34.02	34.08	34.14	34.20	34.27	34.33	34.39
32	34.45-	34.51	34.57	34.63	34.70	34.76	34.82	34.88	34.94	35.00
33	35.06	35.12	35.18	35.24	35.30	35.37	35.43	35.49	35.55-	35.61
34	35.67	35.73	35.79	35.85-	35.91	35.97	36.03	36.09	36.15+	36.21
35	36.27	36.33	36.39	36.45+	36.51	36.57	36.63	36.69	36.75+	36.81
36	36.87	36.93	36.99	37.05-	37.11	37.17	37.23	37.29	37.35-	37.41
37	37.47	37.52	37.58	37.64	37.70	37.76	37.82	37.88	37.94	38.00
38	38.06	38.12	38.17	38.23	38.29	38.35+	38.41	38.47	38.53	38.59
39	38.65-	38.70	38.76	38.82	38.88	38.94	39.00	39.06	39.11	39.17
40	39.23	39.29	39.35-	39.41	39.47	39.52	39.58	39.64	39.70	39.76
41	39.82	39.87	39.93	39.99	40.05-	40.11	40.16	40.22	40.28	40.34
42	40.40	40.46	40.51	40.57	40.63	40.69	40.74	40.80	40.86	40.92
43	40.98	41.03	41.09	41.15-	41.21	41.27	41.32	41.38	41.44	41.50
44	41.55+	41.61	41.67	41.73	41.78	41.84	41.90	41.96	42.02	42.07
45	42.13	42.19	42.25-	42.30	42.36	42.42	42.48	42.53	42.59	42.65-
46	42.71	42.76	42.82	42.88	42.94	42.99	43.05-	43.11	43.17	43.22
47	43.28	43.34	43.39	43.45+	43.51	43.57	43.62	43.68	43.74	43.80
48	43.85+	43.91	43.97	44.03	44.08	44.14	44.20	44.25+	44.31	44.37
49	44.43	44.48	44.54	44.60	44.66	44.71	44.77	44.83	44.89	44.94

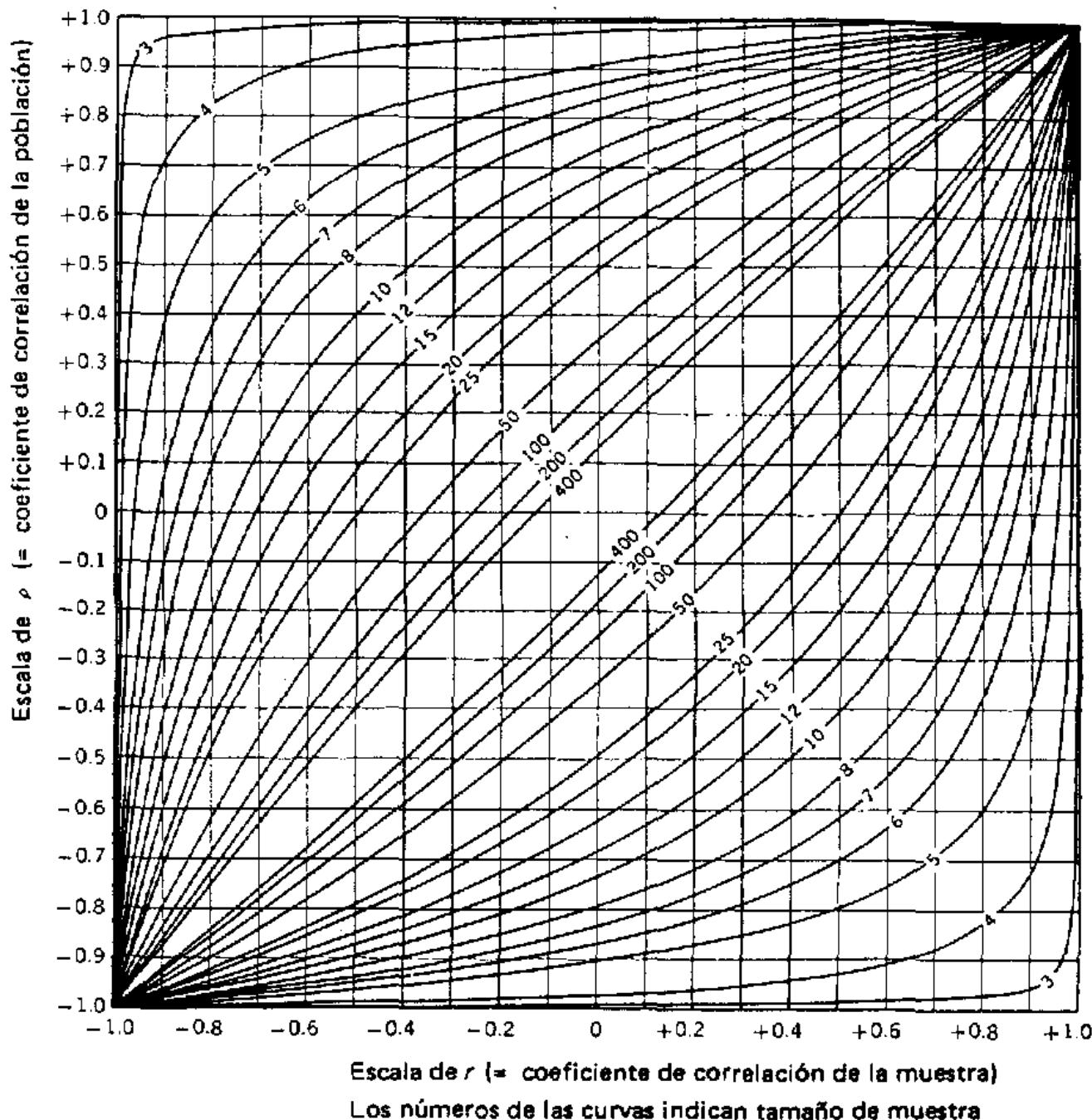
Fuente: Esta tabla apareció en *Plant Protection* (Leningrad), 12: 67 (1937), y se reproduce con autorización del autor, C. I. Bliss.

Tabla A.10 Transformación de la  $\sqrt{\text{porcentaje arco seno}}$  (continuación)

**Tabla A.11A Franjas de confianza para el coeficiente de correlación  $\rho$ :  $P = 0.95$**



Fuente: Esta tabla se reproduce con la autorización de E. S. Pearson de F. N. David, *Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples*, Cambridge University Press para los fideicomisarios de *Biometrika*, 1938.

**Tabla A.11B Franjas de confianza para el coeficiente de correlación  $\rho$ :  $P = .99$** 

Escala de  $\rho$  (= coeficiente de correlación de la población)

Los números de las curvas indican tamaño de muestra

Fuente: Esta tabla se reproduce con la autorización de E. S. Pearson de F. N. David, *Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples*, Cambridge University Press para los fideicomisarios de *Biometrika*, 1938.

**Tabla A.12** Transformación de  $r$  a  $Z$ 

Valores de  $Z = 0.5 \ln(1+r)/(1-r) = \tanh^{-1} r$  aparecen en el cuerpo de la tabla para valores correspondientes de  $r$ , los coeficientes de correlación, están en los márgenes.

$r$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0	.00000	.01000	.02000	.03001	.04002	.05004	.06007	.07012	.08017	.09024
.1	.10034	.11045	.12058	.13074	.14093	.15114	.16139	.17167	.18198	.19234
.2	.20273	.21347	.22366	.23419	.24477	.25541	.26611	.27686	.28768	.29857
.3	.30952	.32055	.33165	.34283	.35409	.36544	.37689	.38842	.40006	.41180
.4	.42385	.43561	.44769	.45990	.47223	.48470	.49731	.51007	.52298	.53606
.5	.54931	.56273	.57634	.59014	.60415	.61838	.63283	.64752	.66246	.67767
.6	.69315	.70892	.72500	.74142	.75817	.77530	.79281	.81074	.82911	.84795
.7	.86730	.88718	.90764	.92873	.95048	.97295	.99621	.1.02033	.1.04537	.1.07143
.8	1.09861	1.12703	1.15682	1.18813	1.22117	1.25615	1.29334	1.33308	1.37577	1.42192
.9	1.47222	1.52752	1.58002	1.65339	1.73805	1.83178	1.94591	2.09229	2.29756	2.64665

Fuente: Esta tabla es un resumen de la Tabla XII de *Standard Four-figure Mathematical Tables*, 1931, de L. M. Milne-Thompson y L. J. Comrie, reproducida con autorización de los autores y editores, MacMillan and Company, Londres.

Tabla A.13 Valores significativos de  $r$  y  $R$ 

$g_f$ del error	$P$	Variables independientes				$g_f$ del error	$P$	Variables independientes			
		1	2	3	4			1	2	3	4
1	.05	.997	.999	.999	.999	24	.05	.388	.470	.523	.562
	.01	1.000	1.000	1.000	1.000		.01	.496	.565	.609	.642
2	.05	.950	.975	.983	.987	25	.05	.381	.462	.514	.553
	.01	.990	.995	.997	.998		.01	.487	.555	.600	.633
3	.05	.878	.930	.950	.961	26	.05	.374	.454	.506	.545
	.01	.959	.976	.983	.987		.01	.478	.546	.590	.624
4	.05	.811	.881	.912	.930	27	.05	.367	.446	.498	.536
	.01	.917	.949	.962	.970		.01	.470	.538	.582	.615
5	.05	.754	.836	.874	.898	28	.05	.361	.439	.490	.529
	.01	.874	.917	.937	.949		.01	.463	.530	.573	.606
6	.05	.707	.795	.839	.867	29	.05	.355	.432	.482	.521
	.01	.834	.886	.911	.927		.01	.456	.522	.565	.598
7	.05	.666	.758	.807	.838	30	.05	.349	.426	.476	.514
	.01	.798	.855	.885	.904		.01	.449	.514	.558	.591
8	.05	.632	.726	.777	.811	35	.05	.325	.397	.445	.482
	.01	.765	.827	.860	.882		.01	.418	.481	.523	.556
9	.05	.602	.697	.750	.786	40	.05	.304	.373	.419	.455
	.01	.735	.800	.836	.861		.01	.393	.454	.494	.526
10	.05	.576	.671	.726	.763	45	.05	.288	.353	.397	.432
	.01	.708	.776	.814	.840		.01	.372	.430	.470	.501
11	.05	.553	.648	.703	.741	50	.05	.273	.336	.379	.412
	.01	.684	.753	.793	.821		.01	.354	.410	.449	.479
12	.05	.532	.627	.683	.722	60	.05	.250	.308	.348	.380
	.01	.661	.732	.773	.802		.01	.325	.377	.414	.442
13	.05	.514	.608	.664	.703	70	.05	.232	.286	.324	.354
	.01	.641	.712	.755	.785		.01	.302	.351	.386	.413
14	.05	.497	.590	.646	.686	80	.05	.217	.269	.304	.332
	.01	.623	.694	.737	.768		.01	.283	.330	.362	.389
15	.05	.482	.574	.630	.670	90	.05	.205	.254	.288	.315
	.01	.606	.677	.721	.752		.01	.267	.312	.343	.368
16	.05	.468	.559	.615	.655	100	.05	.195	.241	.274	.300
	.01	.590	.662	.706	.738		.01	.254	.297	.327	.351
17	.05	.456	.545	.601	.641	125	.05	.174	.216	.246	.269
	.01	.575	.647	.691	.724		.01	.228	.266	.294	.316
18	.05	.444	.532	.587	.628	150	.05	.159	.198	.225	.247
	.01	.561	.633	.678	.710		.01	.208	.244	.270	.290
19	.05	.433	.520	.575	.615	200	.05	.138	.172	.196	.215
	.01	.549	.620	.665	.698		.01	.181	.212	.234	.253
20	.05	.423	.509	.563	.604	300	.05	.113	.141	.160	.176
	.01	.537	.608	.652	.685		.01	.148	.174	.192	.208
21	.05	.413	.498	.522	.592	400	.05	.098	.122	.139	.153
	.01	.526	.596	.641	.674		.01	.128	.151	.167	.180
22	.05	.404	.488	.542	.582	500	.05	.088	.109	.124	.137
	.01	.515	.585	.630	.663		.01	.115	.135	.150	.162
23	.05	.396	.479	.532	.572	1,000	.05	.062	.077	.088	.097
	.01	.505	.574	.619	.652		.01	.081	.096	.106	.115

Fuente: Reproducido del libro, de G. W. Snedecor, *Statistical Methods*, cuarta edición, The Iowa State College Press, Ames, Iowa, 1946, con autorización del autor y del editor.

Tabla A.14A Límites de confianza binomiales †

Número con características	<i>P</i>	Tamaño de muestra			
		10	15	20	25
0	.95	.0000-.3085	.0000-.2180	.0000-.1685	.0000-.1372
	.99	.0000-.4113	.0000-.2976	.0000-.2327	.0000-.1910
1	.95	.0025-.4450	.0017-.3200	.0013-.2485	.0010-.2036
	.99	.0005-.5440	.0003-.4027	.0002-.3170	.0002-.2624
2	.95	.0252-.5560	.0166-.4049	.0124-.3170	.0098-.2605
	.99	.0108-.6480	.0071-.4871	.0053-.3870	.0042-.3208
3	.95	.0667-.6520	.0433-.4807	.0321-.3793	.0255-.3124
	.99	.0370-.7350	.0239-.5607	.0177-.4505	.0140-.3748
4	.95	.1220-.7380	.0780-.5514	.0575-.4365	.0455-.3610
	.99	.0768-.8091	.0488-.6278	.0358-.5065	.0283-.4241
5	.95	.1870-.8130	.1185-.6162	.0868-.4913	.0684-.4072
	.99	.1280-.8720	.0803-.6889	.0585-.5605	.0460-.4700
6	.95		.1633-.6774	.1190-.5430	.0935-.4514
	.99		.1167-.7440	.0845-.6095	.0662-.5138
7	.95		.2129-.7338	.1538-.5920	.1206-.4938
	.99		.1587-.7954	.1140-.6570	.0890-.5556
8	.95			.1910-.6395	.1496-.5350
	.99			.1460-.7010	.1136-.5954
9	.95			.2305-.6848	.1797-.5748
	.99			.1808-.7430	.1401-.6336
10	.95			.2720-.7280	.2112-.6132
	.99			.2175-.7825	.1680-.6704
11	.95				.2441-.6506
	.99				.1975-.7055
12	.95				.2781-.6869
	.99				.2284-.7393
13	.95				
	.99				
14	.95				
	.99				

† Los intervalos de confianza hallados en estas tablas son tales que las probabilidades de aproximadamente 0.025 y 0.005 están relacionadas con eventos raros en cada extremo. Así, las probabilidades de confianza son como mínimo 0.95 y 0.99 y son 0.975 y 0.995 cuando el número con la característica es cero.

Tabla A.144 Límites de confianza binomiales (continuación)

Tamaño de muestra					P	Número con características
30	50	100	500	1,000		
.0000-.1157	.0000-.0711	.0000-.0362	.0000-.0074	.0000-.0037	.95	0
.0000-.1619	.0000-.1005	.0000-.0516	.0000-.0105	.0000-.0053	.99	
.0009-.1779	.0005-.1066	.0002-.0545	.0001-.0111	.0000-.0056	.95	1
.0002-.2233	.0001-.1398	.0000-.0721	.0000-.0148	.0000-.0074	.99	
.0082-.2209	.0049-.1372	.0024-.0704	.0005-.0144	.0002-.0072	.95	2
.0035-.2735	.0021-.1721	.0010-.0894	.0002-.0184	.0001-.0092	.99	
.0211-.2653	.0126-.1657	.0062-.0853	.0012-.0174	.0006-.0087	.95	3
.0116-.3203	.0069-.2032	.0034-.1057	.0007-.0218	.0003-.0109	.99	
.0377-.3074	.0223-.1925	.0110-.0993	.0022-.0204	.0011-.0102	.95	4
.0234-.3639	.0138-.2313	.0068-.1208	.0013-.0250	.0007-.0125	.99	
.0564-.3474	.0332-.2182	.0164-.1129	.0032-.0232	.0016-.0116	.95	5
.0379-.4044	.0222-.2580	.0110-.1353	.0022-.0281	.0011-.0141	.99	
.0770-.3856	.0454-.2431	.0224-.1260	.0044-.0259	.0022-.0130	.95	6
.0543-.4426	.0318-.2842	.0156-.1493	.0031-.0310	.0015-.0156	.99	
.0992-.4229	.0582-.2675	.0286-.1390	.0056-.0286	.0028-.0144	.95	7
.0729-.4801	.0425-.3092	.0208-.1628	.0041-.0339	.0020-.0170	.99	
.1229-.4589	.0717-.2912	.0351-.1516	.0069-.0313	.0035-.0157	.95	8
.0930-.5158	.0540-.3336	.0263-.1761	.0052-.0368	.0026-.0185	.99	
.1473-.4940	.0858-.3144	.0420-.1640	.0083-.0339	.0041-.0170	.95	9
.1143-.5500	.0660-.3573	.0321-.1892	.0063-.0396	.0031-.0199	.99	
.1729-.5280	.1004-.3372	.0490-.1762	.0096-.0365	.0048-.0183	.95	10
.1369-.5835	.0786-.3804	.0382-.2020	.0075-.0423	.0037-.0213	.99	
.1993-.5613	.1154-.3595	.0562-.1883	.0110-.0390	.0050-.0196	.95	11
.1606-.6157	.0920-.4032	.0445-.2145	.0087-.0450	.0043-.0226	.99	
.2266-.5939	.1307-.3817	.0636-.2002	.0125-.0416	.0062-.0209	.95	12
.1850-.6469	.1056-.4256	.0510-.2269	.0099-.0477	.0050-.0240	.99	
.2546-.6256	.1463-.4034	.0711-.2120	.0139-.0441	.0069-.0221	.95	13
.2107-.6772	.1198-.4473	.0577-.2392	.0112-.0504	.0056-.0253	.99	
.2835-.6566	.1623-.4248	.0787-.2237	.0154-.0465	.0077-.0234	.95	14
.2373-.7066	.1342-.4688	.0646-.2513	.0126-.0530	.0063-.0267	.99	

Para interpolar en esta tabla, usar la fórmula  $IC(T)n(T)/n$ , donde  $IC(T)$  y  $n(T)$  son los intervalos de confianza tabulados y el tamaño de muestra inmediatamente inferior al tamaño de muestra observado,  $n$ . Por ejemplo, si tres individuos de 40 tienen la característica, entonces el valor más bajo del intervalo de confianza de 99% se calcula así  $0.0116(30)/40 = 0.0087$ .

Fuente: Resumido de *Statistical Tables for Use with Binomial Samples*, publicados por D. Mainland, L. Herrera, y M. J. Sutcliffe, Nueva York, 1956, con autorización de los autores.

Tabla A.14B Límites de confianza binomiales

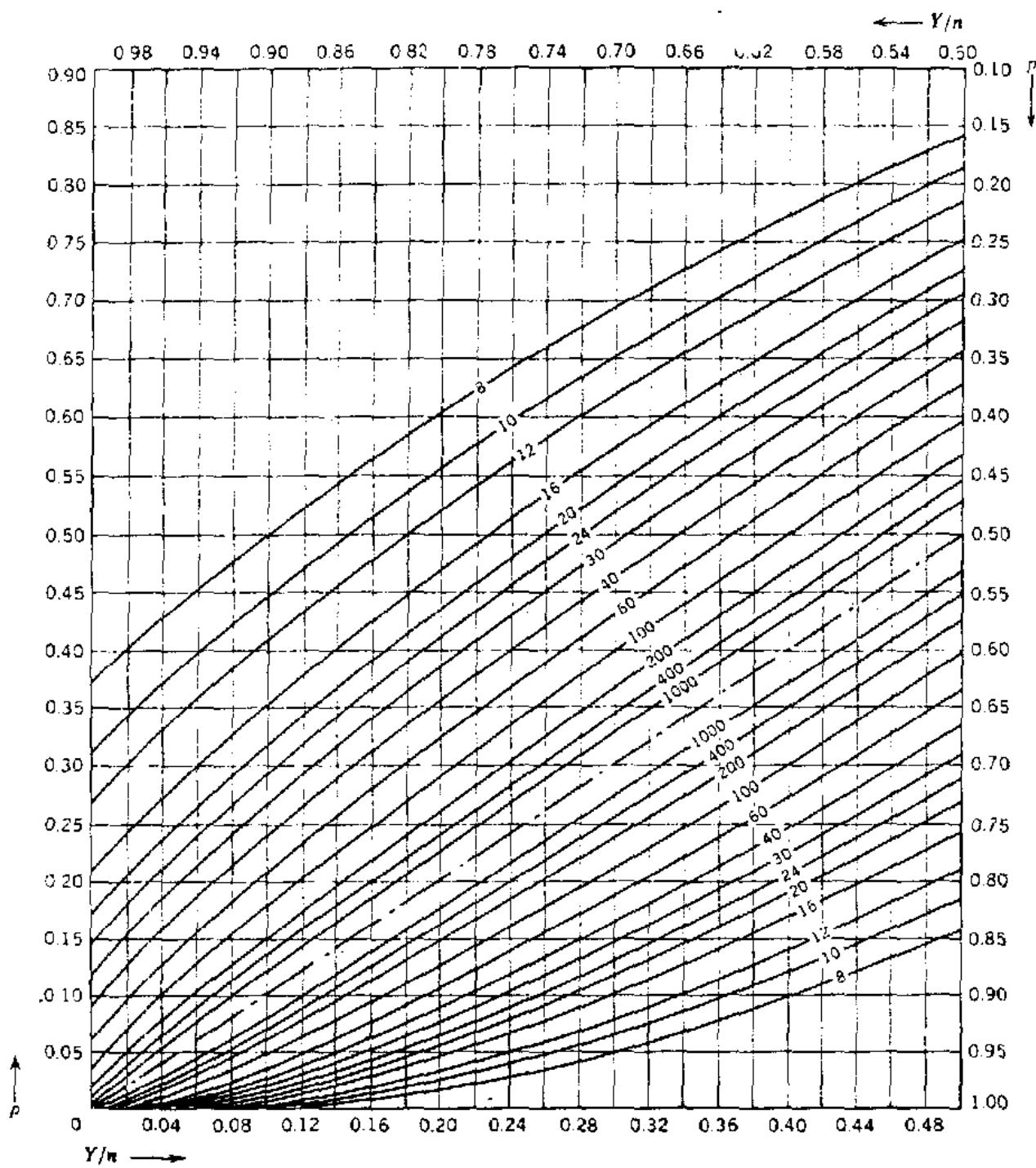
Fracción observada	P	Tamaño de muestra					
		50	75	150	300	500	1,000
.01	.95				.0021-.0289	.0032-.0232	.0048-.0183
	.99				.0011-.0361	.0022-.0280	.0037-.0213
.02	.95				.0086-.0420	.0106-.0356	.0129-.0301
	.99				.0067-.0500	.0087-.0412	.0113-.0336
.03	.95				.0152-.0550	.0179-.0481	.0211-.0419
	.99				.0122-.0640	.0152-.0544	.0188-.0459
.04	.95				.0217-.0681	.0253-.0605	.0292-.0536
	.99				.0177-.0779	.0217-.0675	.0264-.0582
.05	.95			.0211-.0981	.0283-.0811	.0326-.0729	.0373-.0654
	.99			.0156-.1150	.0232-.0918	.0283-.0807	.0339-.0705
.06	.95			.0283-.1104	.0363-.0928	.0411-.0843	.0463-.0764
	.99			.0219-.1279	.0307-.1040	.0363-.0924	.0425-.0818
.07	.95			.0355-.1227	.0444-.1045	.0496-.0956	.0552-.0873
	.99			.0282-.1408	.0381-.1162	.0443-.1042	.0512-.0931
.08	.95			.0427-.1350	.0524-.1162	.0581-.1070	.0642-.0983
	.99			.0345-.1537	.0455-.1284	.0523-.1160	.0598-.1043
.09	.95			.0499-.1473	.0605-.1280	.0666-.1183	.0732-.1093
	.99			.0408-.1666	.0529-.1406	.0604-.1277	.0684-.1156
.10	.95			.0571-.1595	.0685-.1397	.0751-.1297	.0821-.1203
	.99			.0471-.1796	.0604-.1528	.0684-.1395	.0770-.1269
.11	.95			.0651-.1711	.0771-.1508	.0841-.1406	.0914-.1310
	.99			.0544-.1915	.0685-.1643	.0770-.1507	.0860-.1378
.12	.95			.0730-.1827	.0857-.1620	.0930-.1516	.1006-.1416
	.99			.0617-.2035	.0767-.1758	.0856-.1619	.0951-.1486
.13	.95			.0810-.1942	.0943-.1732	.1020-.1625	.1099-.1523
	.99			.0690-.2155	.0848-.1873	.0942-.1731	.1041-.1595
.14	.95			.0890-.2058	.1030-.1843	.1109-.1734	.1192-.1630
	.99			.0764-.2274	.0930-.1988	.1028-.1843	.1131-.1704
.15	.95			.0780-.2512	.0970-.2174	.1116-.1955	.1198-.1844
	.99			.0628-.2844	.0837-.2394	.1012-.2103	.1114-.1955
.16	.95			.0857-.2626	.1054-.2285	.1205-.2064	.1290-.1950
	.99			.0690-.2961	.0916-.2508	.1097-.2214	.1203-.2063
.17	.95			.0934-.2741	.1139-.2396	.1294-.2172	.1382-.2057
	.99			.0759-.3078	.0995-.2622	.1183-.2325	.1292-.2172
.18	.95			.1011-.2855	.1223-.2508	.1384-.2281	.1474-.2164
	.99			.0829-.3195	.1074-.2736	.1269-.2436	.1381-.2281
.19	.95			.1088-.2969	.1307-.2619	.1473-.2390	.1566-.2271
	.99			.0898-.3312	.1153-.2850	.1355-.2547	.1471-.2390
.20	.95			.1163-.3084	.1392-.2731	.1562-.2498	.1658-.2378
	.99			.0967-.3429	.1232-.2964	.1440-.2657	.1560-.2499
.21	.95			.1248-.3194	.1479-.2839	.1654-.2604	.1752-.2483
	.99			.1042-.3541	.1316-.3075	.1529-.2765	.1651-.2605
.22	.95			.1327-.3304	.1567-.2947	.1745-.2711	.1845-.2588
	.99			.1116-.3652	.1399-.3185	.1618-.2874	.1743-.2712
.23	.95			.1409-.3414	.1654-.3055	.1837-.2817	.1939-.2693
	.99			.1191-.3764	.1482-.3295	.1706-.2982	.1834-.2818
.24	.95			.1490-.3524	.1742-.3164	.1929-.2924	.2033-.2799
	.99			.1265-.3876	.1585-.3405	.1795-.3090	.1926-.2925
.25	.95	.1384-.3927	.1572-.3634	.1830-.3272	.2020-.3030	.2126-.2904	.2234-.2781
	.99	.1125-.4365	.1340-.3988	.1648-.3516	.1884-.3198	.2017-.3031	.2155-.2869

Tabla A.14B Límites de confianza binomiales (*continuación*)

Fracción observada	<i>p</i>	Tamaño de muestra					
		50	75	150	300	500	1,000
.26	.95	.1465-.4034	.1656-.3741	.1920-.3378	.2113-.3135	.2221-.3008	.2331-.2883
	.99	.1198-.4472	.1419-.4095	.1734-.3623	.1975-.3303	.2110-.3136	.2250-.2973
.27	.95	.1545-.4140	.1741-.3848	.2010-.3484	.2207-.3239	.2316-.3111	.2427-.2986
	.99	.1271-.4579	.1498-.4203	.1821-.3729	.2066-.3409	.2204-.3241	.2346-.3076
.28	.95	.1626-.4247	.1826-.3955	.2100-.3590	.2300-.3344	.2411-.3215	.2524-.3089
	.99	.1344-.4686	.1577-.4310	.1907-.3836	.2157-.3515	.2297-.3346	.2441-.3180
.29	.95	.1706-.4354	.1911-.4061	.2190-.3695	.2393-.3449	.2506-.3319	.2621-.3192
	.99	.1408-.4792	.1656-.4418	.1994-.3943	.2247-.3621	.2390-.3451	.2537-.3284
.30	.95	.1787-.4461	.1996-.4168	.2280-.3801	.2487-.3553	.2601-.3423	.2717-.3295
	.99	.1491-.4899	.1735-.4525	.2080-.4050	.2338-.3726	.2483-.3555	.2632-.3387
.31	.95	.1871-.4565	.2083-.4272	.2372-.3905	.2582-.3656	.2697-.3525	.2815-.3397
	.99	.1568-.5002	.1818-.4629	.2169-.4155	.2431-.3830	.2578-.3659	.2729-.3490
.32	.95	.1955-.4668	.2171-.4376	.2464-.4009	.2676-.3760	.2793-.3628	.2912-.3499
	.99	.1646-.5105	.1901-.4733	.2250-.4259	.2524-.3934	.2673-.3762	.2825-.3592
.33	.95	.2038-.4772	.2259-.4481	.2556-.4113	.2771-.3863	.2890-.3731	.3009-.3601
	.99	.1723-.5208	.1984-.4838	.2347-.4364	.2617-.4038	.2768-.3865	.2922-.3695
.34	.95	.2122-.4876	.2346-.4585	.2648-.4217	.2866-.3966	.2986-.3833	.3107-.3703
	.99	.1801-.5311	.2067-.4942	.2436-.4468	.2710-.4142	.2862-.3969	.3018-.3797
.35	.95	.2206-.4980	.2434-.4689	.2740-.4320	.2961-.4069	.3082-.3936	.3204-.3805
	.99	.1878-.5414	.2150-.5046	.2525-.4572	.2803-.4246	.2957-.4072	.3114-.3900
.36	.95	.2293-.5080	.2524-.4790	.2834-.4422	.3057-.4171	.3179-.4038	.3302-.3905
	.99	.1960-.5513	.2296-.5147	.2617-.4674	.2897-.4348	.3053-.4174	.3212-.4002
.37	.95	.2380-.5181	.2615-.4892	.2928-.4524	.3153-.4273	.3276-.4139	.3400-.4007
	.99	.2042-.5612	.2322-.5217	.2700-.4776	.2992-.4450	.3149-.4276	.3309-.4103
.38	.95	.2467-.5281	.2705-.4993	.3022-.4626	.3249-.4375	.3373-.4241	.3498-.4109
	.99	.2123-.5710	.2409-.5348	.2800-.4879	.3087-.4552	.3245-.4378	.3407-.4205
.39	.95	.2554-.5382	.2795-.5095	.3116-.4728	.3345-.4477	.3470-.4343	.3597-.4210
	.99	.2205-.5809	.2495-.5449	.2891-.4981	.3181-.4655	.3342-.4480	.3504-.4306
.40	.95	.2641-.5482	.2885-.5196	.3210-.4830	.3441-.4579	.3568-.4444	.3695-.4311
	.99	.2167-.5908	.2581-.5549	.2983-.5083	.3276-.4757	.3438-.4582	.3602-.4408
.41	.95	.2731-.5580	.2978-.5296	.3305-.4931	.3539-.4679	.3666-.4545	.3793-.4412
	.99	.2372-.6004	.2670-.5647	.3076-.5182	.3372-.4857	.3535-.4683	.3700-.4509
.42	.95	.2821-.5678	.3070-.5395	.3401-.5031	.3636-.4780	.3764-.4646	.3892-.4512
	.99	.2457-.6099	.2759-.5745	.3170-.5282	.3468-.4958	.3632-.4783	.3798-.4610
.43	.95	.2910-.5776	.3163-.5494	.3496-.5132	.3733-.4881	.3862-.4746	.3991-.4613
	.99	.2542-.6195	.2849-.5843	.3264-.5382	.3564-.5059	.3729-.4884	.3896-.4710
.44	.95	.3000-.5874	.3256-.5593	.3592-.5232	.3830-.4981	.3960-.4847	.4090-.4714
	.99	.2627-.6290	.2938-.5941	.3337-.5482	.3660-.5159	.3827-.4985	.3995-.4811
.45	.95	.3090-.5971	.3348-.5693	.3687-.5333	.3928-.5082	.4058-.4948	.4189-.4814
	.99	.2712-.6386	.3027-.6038	.3431-.5582	.3756-.5260	.3924-.5086	.4093-.4912
.46	.95	.3183-.6067	.3443-.5790	.3785-.5431	.4026-.5182	.4157-.5048	.4288-.4914
	.99	.2800-.6478	.3119-.6133	.3547-.5680	.3834-.5359	.4022-.5185	.4192-.5012
.47	.95	.3275-.6162	.3538-.5886	.3882-.5530	.4125-.5281	.4256-.5148	.4387-.5014
	.99	.2889-.6569	.3211-.6228	.3642-.5777	.3952-.5458	.4121-.5285	.4291-.5112
.48	.95	.3368-.6257	.3633-.5983	.3979-.5625	.4223-.5381	.4355-.5247	.4487-.5114
	.99	.2978-.6661	.3304-.6323	.3738-.5875	.4049-.5557	.4219-.5385	.4390-.5212
.49	.95	.3461-.6352	.3728-.6080	.4076-.5728	.4321-.5481	.4454-.5347	.4586-.5214
	.99	.3067-.6753	.3396-.6417	.3834-.5973	.4147-.5656	.4318-.5484	.4489-.5312
.50	.95	.3553-.6447	.3823-.6177	.4173-.5827	.4420-.5580	.4553-.5447	.4685-.5315
	.99	.3155-.6845	.3488-.6512	.3929-.6071	.4245-.5755	.4416-.5584	.4589-.5411

Fuente: Resumido de *Statistical Tables for Use with Binomial Samples*, publicado por D. Mainland, L. Herrera y M. I. Sutcliffe, Nueva York, 1956, con autorización de los autores.

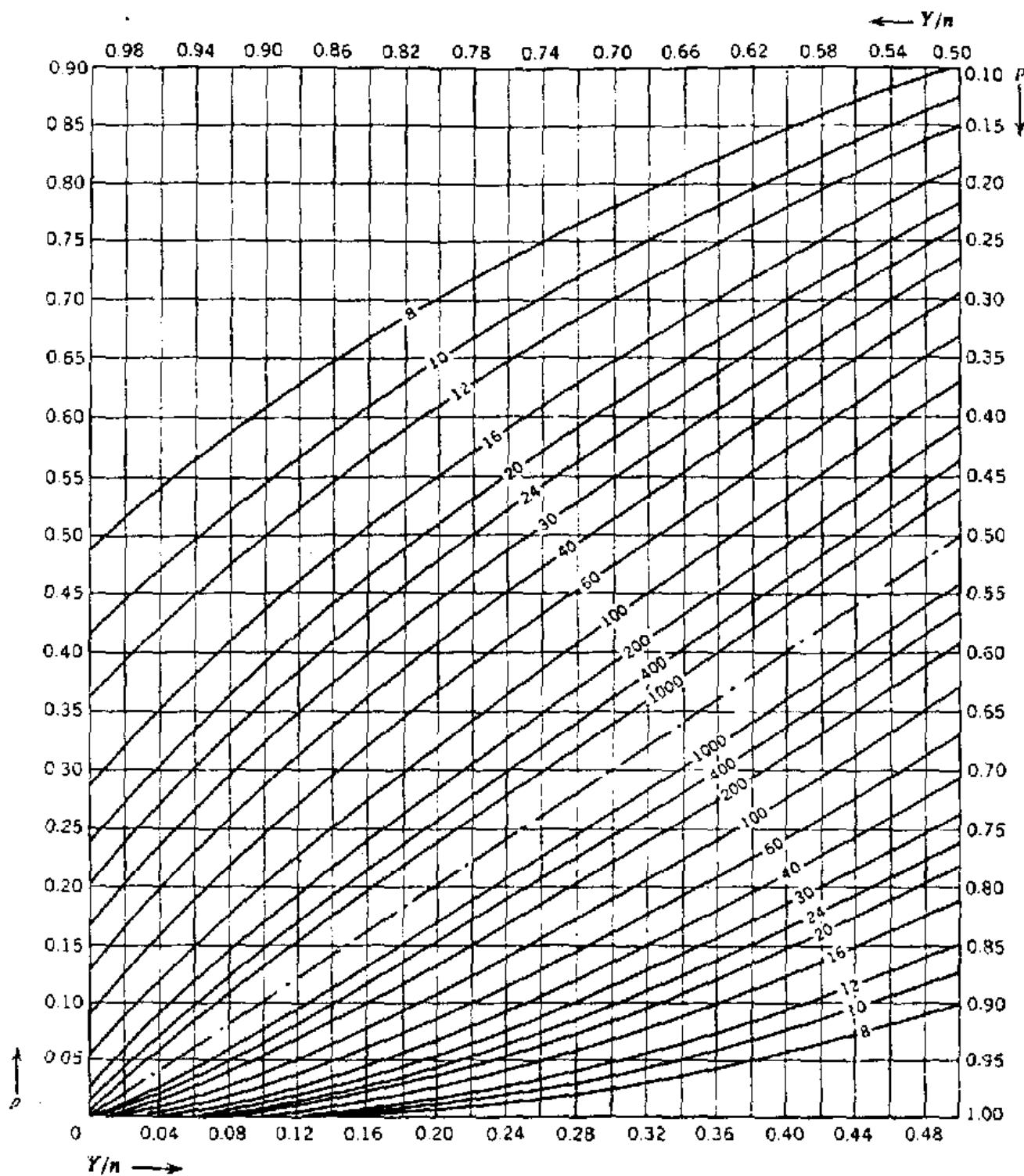
**Tabla A.15A Franjas de confianza para proporciones: coeficiente de confianza de 0.95**



Los números sobre las curvas indican el tamaño de muestra,  $n$ . Para un valor dado de la abscisa  $Y/n$ ,  $p_A$  y  $p_B$  son ordenadas leídas en (o interpoladas entre) las curvas superior e inferior apropiadas, y  $\Pr(p_A \leq p \leq p_B) \geq 1 - 2\alpha$ .

*Fuente:* Reproducida con autorización de los fideicomisarios y de los editores de *Biometrika*, de E. S. Pearson y H. O. Hartley, *Biometrika Tables for Statisticians*, vol. 1, Cambridge University Press, 1954.

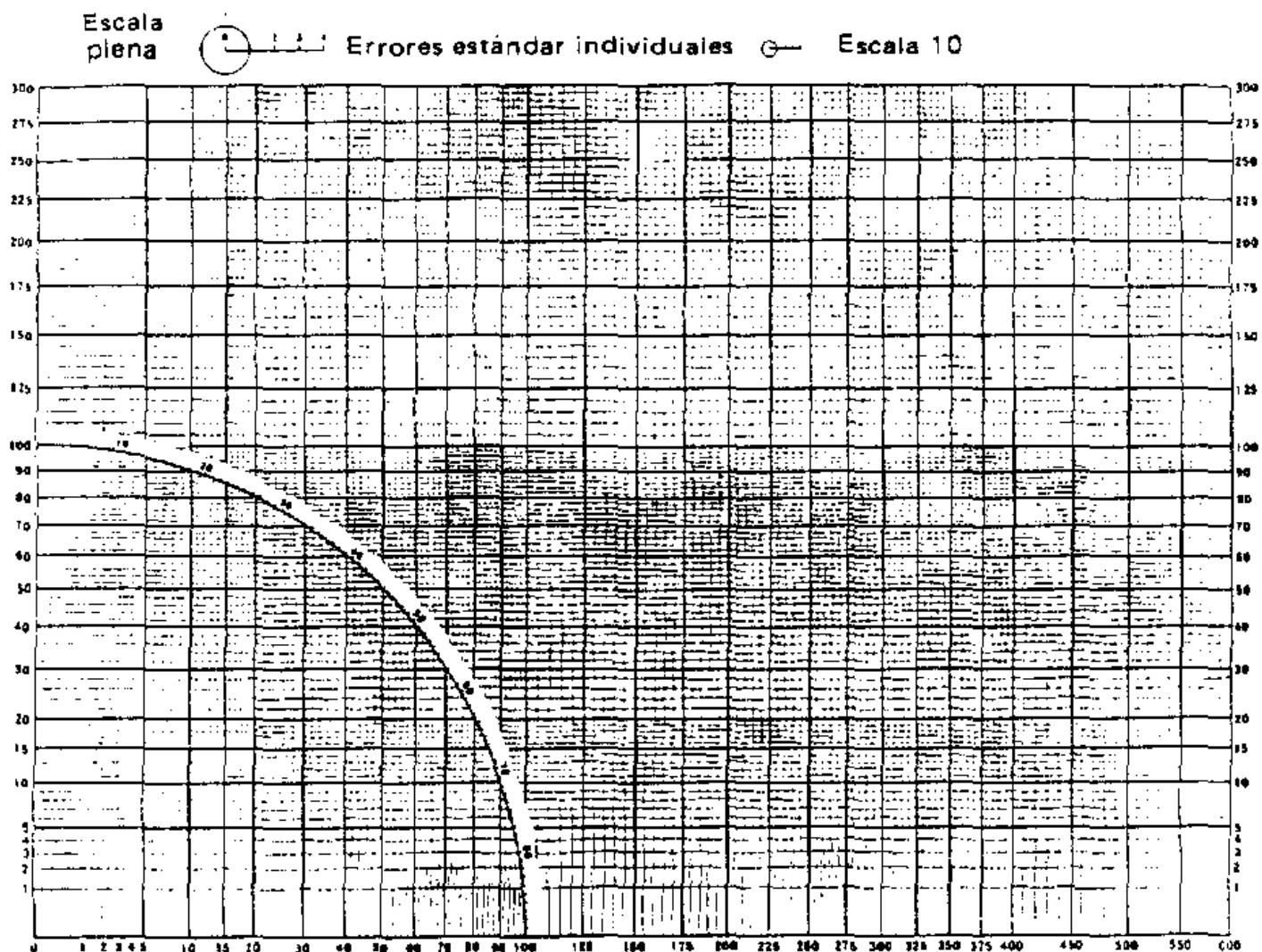
**Tabla A.15B Franjas de confianza para proporciones: coeficiente de confianza de 0.99**



Los números sobre las curvas indican el tamaño de las muestras,  $n$ . Nota: El proceso de lectura de la curva puede simplificarse con la ayuda de la esquina en ángulo recto de una hoja de papel o una tarjeta, a lo largo de los bordes marcados en las escalas que aparecen en la esquina superior izquierda del diagrama.

Fuente: Reproducido con el permiso de los fideicomisarios y de los editores de *Biometrika*, de E. S. Pearson y H. O. Hartley, *Biometrika Tables for Statisticians*, vol. 1, Cambridge University Press, 1954.

Tabla A.16 Papel probabilístico binomial Mosteller-Tukey



*Fuente:* Este diagrama apareció originalmente en F. Mosteller y J. W. Tukey, "The uses and usefulness of binomial probability paper", *J. Am. Stat. Assn.*, 44: 174-212 (1949). Se reproduce con autorización de los autores, editor y Codex Book Company, Inc.

**Tabla A.17A** Tamaño de muestra y probabilidad de tomar una decisión errónea entre las pruebas de cruzamiento de proporciones 1:1 y 3:1

Tamaño de muestra $n$	Proporción aceptada	Clases en regiones de aceptación	Probabilidad de tomar una decisión errónea
20	1:1	0-12	.1316
	3:1	13-20	.1018
30	1:1	0-18	.1002
	3:1	19-30	.0507
40	1:1	0-25	.0403
	3:1	26-40	.0544
44	1:1	0-27	.0481
	3:1	28-44	.0318
50	1:1	0-31	.0325
	3:1	32-50	.0287
60	1:1	0-37	.0259
	3:1	38-60	.0154
70	1:1	0-44	.0112
	3:1	45-70	.0163
80	1:1	0-50	.0092
	3:1	51-80	.0089

$$\left. \begin{array}{l} 1:1 = .5: .5 \\ 3:1 = .75: .25 \end{array} \right\} \text{La línea divisoria } R/n = 0.63091$$

Fuente: Reproducido de Prasert NaNagara, *Testing Mendelian Ratios*, Tesis de Maestría 1953, Cornell University, Ithaca, N. Y.

**Tabla A.17B Tamaño de muestra y probabilidad de tomar una decisión errónea entre pruebas de cruzamiento de proporciones 3:1 y 7:1†**

Tamaño de muestra $n$	Proporción aceptada	Clases en regiones de aceptación	Probabilidad de tomar una decisión errónea
20	3:1	0-16	.2252
	7:1	17-20	.2347
30	3:1	0-24	.2026
	7:1	25-30	.1644
40	3:1	0-32	.1820
	7:1	33-40	.1190
50	3:1	0-40	.1637
	7:1	41-50	.0879
60	3:1	0-49	.0859
	7:1	50-60	.1231
70	3:1	0-57	.08
	7:1	58-70	.09
80	3:1	0-65	.08
	7:1	66-80	.06
90	3:1	0-73	.07
	7:1	74-90	.05
100	3:1	0-82	.04
	7:1	83-100	.07
110	3:1	0-90	.04
	7:1	91-110	.05
200	3:1	0-164	.01
	7:1	165-200	.01+
210	3:1	0-171	.011
	7:1	172-210	.006

$$\left. \begin{array}{l} 3:1 = .75: .25 \\ 7:1 = .875: .125 \end{array} \right\} \text{La línea divisoria } R/n = 0.81786$$

† Ver Tabla A.17A

**Tabla A.17C Tamaño de muestra y probabilidad de tomar una decisión errónea entre pruebas de cruzamiento con proporciones 1:1, 3:1 y 7:1 †**

Tamaño de muestra $n$	Proporción aceptada	Clases en regiones de aceptación	Probabilidad de tomar una decisión errónea
20	1:1	0-11	.2517
	3:1	12-16	.2661*
	7:1	17-20	.2347
30	1:1	0-17	.1808
	3:1	18-24	.2242*
	7:1	25-30	.1644
40	1:1	0-23	.1341
	3:1	24-32	.1936*
	7:1	33-40	.1190
50	1:1	0-31	.0325
	3:1	32-41	.1203
	7:1	42-50	.1660*
60	1:1	0-37	.0259
	3:1	38-49	.1013
	7:1	50-60	.1231*
80	1:1	0-50	.01
	3:1	51-65	.08*
	7:1	66-80	.06
100	1:1	0-63	.00
	3:1	64-82	.04
	7:1	83-100	.07*
110	1:1	0-69	.00
	3:1	70-90	.04
	7:1	91-110	.05*
200	1:1	0-126	.00
	3:1	127-164	.01
	7:1	165-200	.01+*
210	1:1	0-132	.000
	3:1	133-171	.011*
	7:1	172-210	.006

$$1:1 = .5: .5$$

$$3:1 = .75: .25$$

$$7:1 = .875: .125$$

} La línea divisoria  $R/n \doteq 0.63091$   
 } La línea divisoria  $R^1/n \doteq 0.81786$

† Ver Tabla A.17A

**Tabla A.17D Tamaño de muestra y probabilidad de tomar una decisión errónea entre las  $F_2$  para las proporciones 9:7, 13:3 y 15:1 †**

Tamaño de muestra $n$	Proporción aceptada	Clases en regiones de aceptación	Probabilidad de tomar una decisión errónea
20	9:7	0-12	.29
	13:3	13-17	.34*
	15:1	18-20	.15
30	9:7	0-19	.17
	13:3	20-26	.18*
	15:1	27-30	.12
50	9:7	0-34	.03
	13:3	35-44	.09*
	15:1	45-50	.09
75	9:7	0-51	.020
	13:3	52-66	.048*
	15:1	67-75	.044
150	9:7	0-104	.000
	13:3	105-132	.010*
	15:1	133-150	.006

$$9:7 = .5625; .4375$$

$$13:3 = .8125; .1875$$

$$15:1 = .9375; .0625$$

La línea divisoria  $R/n \doteq 0.69736$

La línea divisoria  $R^1/n \doteq 0.88478$

† Ver Tabla A.17 A

Tabla A.17E Tamaño de muestra y probabilidad de tomar una decisión errónea entre las  $F_2$  para las proporciones 27:37, 55:9 y 63:1†

Tamaño de muestra $n$	Proporción aceptada	Clases en regiones de aceptación	Probabilidad de tomar una decisión errónea
20	27:37	0-12	.03
	55:9	13-18	.21*
	63:1	19-20	.04
30	27:37	0-19	.01
	55:9	20-28	.06
	63:1	29-30	.08*
40	27:37	0-26	.00
	55:9	27-37	.07*
	63:1	38-40	.03
50	27:37	0-33	.00
	55:9	34-47	.02
	63:1	48-50	.04*
75	27:37	0-49	.000
	55:9	50-70	.014*
	63:1	71-75	.007
90	27:37	0-59	.000
	55:9	60-84	.008*
	63:1	85-90	.003
95	27:37	0-63	.000
	55:9	64-89	.005*
	63:1	90-95	.004
100	27:37	0-66	.000
	55:9	67-94	.003
	63:1	95-100	.005*

$$27:37 = .421875; .578125$$

$$55:9 = .859375; .140625$$

$$63:1 = .984375; .015625$$

La línea divisoria  $R/n \doteq 0.665214$

La línea divisoria  $R^1/n \doteq 0.94178$

† Ver Tabla A.17A

Tabla A.17F Tamaño de muestra y probabilidad de tomar una decisión errónea entre las  $F_2$  de las proporciones 27:37, 9:7, 3:1, 13:3, 55:9, 15:1 y 63:1†

Tamaño de muestra $n$	Proporción aceptada	Clases en regiones de aceptación	Probabilidad de tomar una decisión errónea
50	27:37	0-22	.342
	9:7	23-29	.405
	3:1	30-37	.517
	13:3	38-41	.521*
	55:9	42-45	.415
	15:1	46-48	.372
75	63:1	49-50	.177
	27:37	0-33	.330
	9:7	34-44	.319
	3:1	45-57	.379
	13:3	58-63	.383
	55:9	64-69	.400*
100	15:1	70-73	.375
	63:1	74-75	.328
	27:37	0-49	.059
	9:7	50-66	.096
	3:1	67-77	.314
	13:3	78-84	.372*
500	55:9	85-92	.352
	15:1	93-97	.334
	63:1	98-100	.206
	27:37	0-246	.00
	9:7	247-330	.00
	3:1	331-387	.10
800	13:3	388-418	.10*
	55:9	419-451	.09
	15:1	452-483	.01
	63:1	484-500	.00
	27:37	0-393	.00
	9:7	394-528	.00
1375	3:1	529-625	.02
	13:3	626-670	.04+*
	55:9	671-722	.04
	15:1	723-772	.00
	63:1	773-800	.00
	27:37	0-676	.00
27:37 } 9:7 } 3:1 } 13:3 } 55:9 } 15:1 } 63:1 }	677-908	.00	
	27:37 } 9:7 } 3:1 } 13:3 } 55:9 } 15:1 } 63:1 }	909-1075	.00
	27:37 } 9:7 } 3:1 } 13:3 } 55:9 } 15:1 } 63:1 }	1076-1151	.01
	27:37 } 9:7 } 3:1 } 13:3 } 55:9 } 15:1 } 63:1 }	1152-1241	.01
	27:37 } 9:7 } 3:1 } 13:3 } 55:9 } 15:1 } 63:1 }	1242-1328	.00
	27:37 } 9:7 } 3:1 } 13:3 } 55:9 } 15:1 } 63:1 }	1329-1375	.00
<p>La línea divisoria <math>R/n \doteq 0.4921</math>      La línea divisoria <math>R/n \doteq 0.6605</math>      La línea divisoria <math>R/n \doteq 0.7823</math>      La línea divisoria <math>R/n \doteq 0.8368</math>      La línea divisoria <math>R/n \doteq 0.9031</math>      La línea divisoria <math>R/n \doteq 0.9660</math></p>			

† Ver Tabla A.17A

**Tabla A.17G Tamaño de muestra y probabilidad de tomar una decisión errónea entre pruebas de cruzamiento de las proporciones 2:1:1, 1:2:1 y 1:1:2†**

Tamaño de muestra $n$	Probabilidad de tomar una decisión errónea
20	.1890
40	.0645
45	.0501
50	.0394
70	.0147
75	.0113
80	.0099

Aceptar 2:1:1 cuando el primer grupo es más grande que los otros dos; aceptar 1:2:1 cuando el segundo grupo es más grande que los otros dos; y aceptar 1:1:2 cuando el tercer grupo es el grupo más grande.

† Ver Tabla A.17A

**Tabla A.17H Tamaño de muestra y probabilidad de tomar una decisión errónea entre las pruebas de cruzamiento 1:1:2, 1:1:4 y 1:1:6†**

Tamaño de muestra $n$	Proporción aceptada	Clases en regiones de aceptación	Probabilidad de tomar una decisión errónea
50	1:1:2	0-27	.240
	1:1:4	28-35	.305*
	1:1:6	36-50	.252
100	1:1:2	0-54	.18
	1:1:4	55-71	.16
	1:1:6	72-100	.21*
200	1:1:2	0-117	.01
	1:1:4	118-142	.10
	1:1:6	143-200	.11*
330	1:1:2	0-193	.00
	1:1:4	194-234	.05*
	1:1:6	235-330	.04
646	1:1:2	0-377	.000
	1:1:4	378-458	.008
	1:1:6	459-646	.010*

Aceptar 1:1:2 cuando el número de individuos en el tercer grupo  $z$  es menor que  $0.5850 n$ ; aceptar 1:1:4 cuando  $z$  esté entre  $0.5850 n$  y  $0.7095 n$ ; aceptar 1:1:6 cuando  $z$  sea mayor que  $0.7095 n$ .

† Ver Tabla A.17A

**Tabla A.171** Tamaño de muestra y probabilidad de tomar una decisión errónea entre las  $F_2$  de las proporciones 9:6:1, 9:3:4 y 12:3:1†

Tamaño de muestra $n$	Proporción aceptada	Clases en regiones de aceptación	Probabilidad de tomar una decisión errónea
20	9:6:1	$y \geq 6$ $x < 14$	.213
	9:3:4	$y < 6$ $x < 14$	.312*
	12:3:1	$y \geq 6$ $x \geq 14$	.214
50	9:6:1	$y \geq 14$ $x < 34$	.084
	9:3:4	$y < 14$ $x < 34$	.134*
	12:3:1	$y \geq 14$ $x \geq 34$	.098
75	9:6:1	$y \geq 21$ $x < 50$	.054
	9:3:4	$y < 21$ $x < 50$	.075*
	12:3:1	$y \geq 21$ $x \geq 50$	.039
90	9:6:1	$y \geq 25$ $x < 60$	.035
	9:3:4	$y < 25$ $x < 60$	.051*
	12:3:1	$y \geq 25$ $x \geq 60$	.028
150	9:6:1	$y \geq 42$ $x < 100$	.008
	9:3:4	$y < 42$ $x < 100$	.009*
	12:3:1	$y \geq 42$ $x \geq 100$	.008

$x$  y  $y$  son los números de individuos en los grupos primero y segundo de la muestra, respectivamente.

Aceptar 9:6:1 cuando  $y \geq 0.27457n$  y  $x < 0.6605n$

Aceptar 9:3:4 cuando  $y < 0.27457n$  y  $x < 0.6605n$

Aceptar 12:3:1 cuando  $x \geq 0.6605n$

† Ver Tabla A.17A

**Tabla A.17J Tamaño de muestra y probabilidad de tomar una decisión errónea entre las  $F_2$  de las proporciones 27:9:28 y 81:27:148†**

Tamaño de muestra $n$	Proporción aceptada	Clases de $z$ en regiones de aceptación	Probabilidad de tomar una decisión errónea
20	27:9:28	0-10	.2144
	81:27:148	11-20	.3125
40	27:9:28	0-20	.1694
	81:27:148	21-40	.1998
60	27:9:28	0-30	.1345
	81:27:148	31-60	.1370
100	27:9:28	0-50	.09
	81:27:148	51-100	.07
135	27:9:28	0-68	.05+
	81:27:148	69-135	.05-
200	27:9:28	0-101	.02
	81:27:148	102-200	.02
269	27:9:28	0-136	.010
	81:27:148	137-269	.009

$z$  es el número de individuos en el tercer grupo de la muestra. La línea divisoria  $R/n \doteq 0.50788$ .

† Ver Tabla A.17A

**Tabla A.18 Prueba de rangos signados de Wilcoxon**

Valores tabulados de  $T$  son tales que valores más pequeños ocurren por azar con una probabilidad dada†

Pares $n$	Probabilidad			Pares $n$	Probabilidad		
	.05	.02	.01		.05	.02	.01
6	0	—	—	16	30	24	20
7	2	0	—	17	35	28	23
8	4	2	0	18	40	33	28
9	6	3	2	19	46	38	32
10	8	5	3	20	52	43	38
11	11	7	5	21	59	49	43
12	14	10	7	22	66	56	49
13	17	13	10	23	73	62	55
14	21	16	13	24	81	69	61
15	25	20	16	25	89	77	68

† Las probabilidades son para pruebas de dos colas. Para pruebas de una cola, las probabilidades de la tabla vienen a ser 0.025, 0.01 y 0.005.

Fuente: Reproducida de F. Wilcoxon, *Some Rapid Approximate Statistical Procedures*, American Cyanamid Company, Stamford, Conn., 1949, con autorización del autor y la Cyanamid Company. Los valores de esta tabla fueron obtenidos por aproximación de los valores dados por Tukey en Memorandum Rept. 17, "The simplest signed rank tests", Grupo de Investigación Estadística, Princeton, 1949.

Tabla A.19 Puntos críticos de sumas ranqueadas

(alternativas de dos colas)

$n_2 = n$ más grande	$P$	$n_1 = n$ más pequeño												
		2	3	4	5	6	7	8	9	10	11	12	13	14
4	.05			10										
	.01			—										
5	.05		6	11	17									
	.01		—	—	15									
6	.05		7	12	18	26								
	.01		—	10	16	23								
7	.05		7	13	20	27	36							
	.01		—	10	17	24	32							
8	.05	3	8	14	21	29	38	49						
	.01	—	—	11	17	25	34	43						
9	.05	3	8	15	22	31	40	51	63					
	.01	—	6	11	18	26	35	45	56					
10	.05	3	9	15	23	32	42	53	65	78				
	.01	—	6	12	19	27	37	47	58	71				
11	.05	4	9	16	24	34	44	55	68	81	96			
	.01	—	6	12	20	28	38	49	61	74	87			
12	.05	4	10	17	26	35	46	58	71	85	99	115		
	.01	—	7	13	21	30	40	51	63	76	90	106		
13	.05	4	10	18	27	37	48	60	73	88	103	119	137	
	.01	—	7	14	22	31	41	53	65	79	93	109	125	
14	.05	4	11	19	28	38	50	63	76	91	106	123	141	160
	.01	—	7	14	22	32	43	54	67	81	96	112	129	147
15	.05	4	11	20	29	40	52	65	79	94	110	127	145	164
	.01	—	8	15	23	33	44	56	70	84	99	115	133	151
16	.05	4	12	21	31	42	54	67	82	97	114	131	150	169
	.01	—	8	15	24	34	46	58	72	86	102	119	137	155
17	.05	5	12	21	32	43	56	70	84	100	117	135	154	
	.01	—	8	16	25	36	47	60	74	89	105	122	140	
18	.05	5	13	22	33	45	58	72	87	103	121	139		
	.01	—	8	16	26	37	49	62	76	92	108	125		
19	.05	5	13	23	34	46	60	74	90	107	124			
	.01	3	9	17	27	38	50	64	78	94	111			
20	.05	5	14	24	35	48	62	77	93	110				
	.01	3	9	18	28	39	52	66	81	97				
21	.05	6	14	25	37	50	64	79	95					
	.01	3	9	18	29	40	53	68	83					
22	.05	6	15	26	38	51	66	82						
	.01	3	10	19	29	42	55	70						
23	.05	6	15	27	39	53	68							
	.01	3	10	19	30	43	57							
24	.05	6	16	28	40	55								
	.01	3	10	20	31	44								
25	.05	6	16	28	42									
	.01	3	11	20	32									
26	.05	7	17	29										
	.01	3	11	21										
27	.05	7	17											
	.01	4	11											
28	.05	7												
	.01	4												

Fuente: Reproducido de Colin White, "The use of ranks in a test of significance for comparing two treatments", *Biometrics*, 8: 33-41 (1950), con autorización del editor y el autor.

Tabla A.20 Niveles significativos de trabajo para magnitudes de sumas de cuadrantes

<i>Nivel significativo</i>	<i>Magnitud de suma de cuadrante †</i>
.10	9
.05	11
.02	13
.01	14-15
.005	15-17
.002	17-19
.001	18-21

† La magnitud más pequeña se aplica a muestras de tamaño grande, la magnitud más grande se aplica a muestra de tamaño pequeño. Magnitud mayor o igual a dos veces el tamaño de la muestra menos 6 no deberá usarse.

Fuente: Reproducido de P. S. Olmstead y J. W. Tukey, "A corner test for association", *Annals Math Stat.*, 18: 495-513 (1947), con autorización de los autores, el editor y gentileza de Bell Telephone Laboratories, Inc.

Tabla A.21 Valores de  $t$  de riesgo-promedio-mínimo Durcan-Waller ( $k = 100$ )

$q$	4	6	8	10	12	14	16	18	20	24	30	40	60	120	$\infty$
$\dagger F = 1.2 (a = .913, b = 2.449)$															
2-6	*	*	*	*	*	*	*	*	*	*	*	*	*	*	2.85
8	2.85	2.91	2.94	2.96	2.97	2.98	2.99	2.99	2.99	3.00	3.00	3.00	3.00	3.00	3.00
10	2.85	2.93	2.98	3.01	3.04	3.05	3.06	3.07	3.08	3.09	3.10	3.11	3.12	3.12	3.12
12	2.85	2.95	3.01	3.05	3.08	3.10	3.12	3.13	3.14	3.16	3.17	3.19	3.20	3.21	3.22
14	2.85	2.96	3.03	3.08	3.12	3.14	3.16	3.18	3.19	3.21	3.23	3.25	3.27	3.29	3.31
16	2.85	2.97	3.05	3.11	3.15	3.18	3.20	3.22	3.24	3.26	3.28	3.31	3.33	3.36	3.38
20	2.85	2.99	3.08	3.14	3.19	3.23	3.26	3.28	3.30	3.33	3.37	3.40	3.44	3.47	3.50
40	2.85	3.02	3.13	3.22	3.29	3.35	3.39	3.43	3.47	3.52	3.58	3.64	3.72	3.79	3.87
100	2.85	3.04	3.17	3.28	3.36	3.44	3.50	3.55	3.59	3.67	3.76	3.86	3.98	4.11	4.23
$\infty$	2.85	3.05	3.20	3.32	3.42	3.50	3.58	3.64	3.70	3.80	3.91	4.06	4.24	4.45	4.22
$\dagger F = 1.4 (a = .845, b = 1.871)$															
2-4	*	*	*	*	*	*	*	*	*	*	*	*	*	*	2.57
6	2.85	2.85	2.84	2.83	2.82	2.81	2.80	2.80	2.80	2.79	2.78	2.77	2.75	2.74	2.72
8	2.85	2.88	2.89	2.90	2.90	2.89	2.89	2.89	2.89	2.88	2.88	2.87	2.86	2.85	2.83
10	2.85	2.90	2.93	2.94	2.95	2.95	2.96	2.96	2.96	2.96	2.96	2.95	2.94	2.93	2.92
12	2.85	2.92	2.95	2.98	2.99	3.00	3.00	3.01	3.01	3.01	3.01	3.01	3.00	2.99	2.98
14	2.85	2.93	2.97	3.00	3.02	3.03	3.04	3.04	3.04	3.05	3.05	3.06	3.06	3.05	3.03
16	2.85	2.94	2.99	3.02	3.04	3.06	3.07	3.08	3.08	3.09	3.09	3.10	3.10	3.09	3.08
20	2.85	2.95	3.01	3.05	3.08	3.10	3.11	3.12	3.13	3.14	3.15	3.16	3.16	3.16	3.14
40	2.85	2.98	3.06	3.12	3.16	3.19	3.22	3.24	3.25	3.28	3.30	3.31	3.32	3.32	3.28
100	2.85	2.99	3.09	3.16	3.22	3.26	3.29	3.32	3.34	3.41	3.43	3.45	3.42	3.31	
$\infty$	2.85	3.01	3.12	3.20	3.26	3.31	3.35	3.39	3.42	3.50	3.53	3.54	3.46	3.46	3.22

 $\dagger$  Condiciones límites para la Tabla A.21Para  $F \leq 2.4$ , interpolar sobre  $a = 1/F^{1/2}$  a menos que ambos  $q > 100$  y  $f < 10$ , entonces usar  $b = [F/(F - 1)]^{1/2}$ .Para  $F > 2.4$ , interpolar sobre  $b$ , a menos que  $q \leq 20$  y  $f \leq 20$ , entonces usar  $a$ .

Fuente: Comunicación personal de David B. Duncan.

		2.21									
		* * * * *					* * * * *				
		* * * * *					* * * * *				
2		•	•	•	•	•	•	•	•	•	•
4	*	2.74	2.67	2.63	2.59	2.56	2.54	2.52	2.51	2.49	2.46
6	*	2.79	2.74	2.70	2.67	2.64	2.62	2.60	2.59	2.57	2.54
8	*	2.81	2.77	2.74	2.71	2.69	2.67	2.65	2.64	2.62	2.59
10	*	2.85	2.83	2.80	2.77	2.74	2.72	2.70	2.69	2.67	2.65
12	*	2.85	2.84	2.82	2.79	2.77	2.75	2.73	2.71	2.70	2.67
14	*	2.85	2.85	2.83	2.81	2.79	2.77	2.75	2.73	2.72	2.69
16	*	2.85	2.85	2.84	2.82	2.80	2.78	2.76	2.74	2.73	2.70
20	*	2.85	2.86	2.85	2.84	2.82	2.80	2.78	2.77	2.75	2.72
40	*	2.85	2.88	2.89	2.88	2.86	2.85	2.83	2.81	2.80	2.77
100	*	2.85	2.89	2.91	2.90	2.89	2.88	2.86	2.84	2.82	2.79
1000	*	2.85	2.90	2.92	2.92	2.91	2.90	2.89	2.88	2.86	2.85
∞	*	2.85	2.90	2.92	2.92	2.91	2.90	2.89	2.88	2.86	2.85

\* Todas las diferencias no significativas.

<sup>†</sup> Ver condiciones límite en la página 615.

ESTADOS UNIDOS DE COLOMBIA

Tabla A.21 Valores de  $t$  de riesgo-promedio-mínimo Duncan-Waller ( $k = 100$ ) (continuación)

$q$	4	6	8	10	12	14	16	18	20	24	30	40	60	120	$\infty$
$f$															
2	*	*	*	*	*	*	*	*	*	*	*	*	*	*	2.16
4	2.85	2.71	2.63	2.57	2.53	2.49	2.47	2.44	2.43	2.40	2.37	2.34	2.31	2.28	2.25
6	2.85	2.75	2.68	2.63	2.58	2.55	2.52	2.50	2.48	2.46	2.42	2.39	2.36	2.32	2.28
8	2.85	2.77	2.71	2.66	2.62	2.59	2.56	2.54	2.52	2.49	2.45	2.42	2.38	2.34	2.29
10	2.85	2.79	2.73	2.68	2.64	2.61	2.58	2.56	2.54	2.50	2.47	2.43	2.39	2.34	2.30
12	2.85	2.79	2.74	2.70	2.66	2.62	2.60	2.57	2.55	2.52	2.48	2.44	2.39	2.35	2.29
14	2.85	2.80	2.75	2.71	2.67	2.64	2.61	2.58	2.56	2.53	2.49	2.44	2.40	2.35	2.29
16	2.85	2.81	2.76	2.72	2.68	2.65	2.62	2.59	2.57	2.53	2.49	2.45	2.40	2.34	2.29
20	2.85	2.82	2.77	2.73	2.69	2.66	2.63	2.60	2.58	2.54	2.50	2.45	2.40	2.34	2.28
40	2.85	2.83	2.80	2.76	2.72	2.69	2.66	2.63	2.60	2.56	2.51	2.46	2.39	2.33	2.27
100	2.85	2.84	2.81	2.78	2.74	2.71	2.67	2.64	2.62	2.57	2.51	2.45	2.39	2.32	2.26
$\infty$	2.85	2.85	2.83	2.79	2.76	2.72	2.68	2.65	2.62	2.57	2.51	2.45	2.38	2.31	2.25
$f$															
2	*	*	*	*	*	*	*	*	*	*	*	*	*	*	2.09
4	2.85	2.68	2.57	2.50	2.45	2.41	2.38	2.35	2.33	2.30	2.27	2.24	2.20	2.17	2.13
6	2.85	2.71	2.61	2.54	2.49	2.44	2.41	2.39	2.36	2.33	2.29	2.26	2.22	2.18	2.14
8	2.85	2.72	2.63	2.56	2.51	2.47	2.43	2.40	2.38	2.34	2.31	2.27	2.22	2.18	2.14
10	2.85	2.74	2.65	2.58	2.52	2.48	2.44	2.41	2.39	2.35	2.31	2.27	2.22	2.18	2.13
12	2.85	2.74	2.66	2.59	2.53	2.49	2.45	2.42	2.40	2.36	2.31	2.27	2.22	2.18	2.13
14	2.85	2.75	2.66	2.60	2.54	2.49	2.46	2.43	2.40	2.36	2.32	2.27	2.22	2.17	2.13
16	2.85	2.75	2.67	2.60	2.55	2.50	2.46	2.43	2.40	2.36	2.32	2.27	2.22	2.17	2.12
20	2.85	2.76	2.68	2.61	2.55	2.51	2.47	2.43	2.41	2.36	2.32	2.27	2.22	2.17	2.12
40	2.85	2.77	2.70	2.63	2.57	2.52	2.48	2.44	2.41	2.37	2.32	2.26	2.21	2.16	2.11
100	2.85	2.78	2.71	2.64	2.58	2.53	2.49	2.45	2.42	2.37	2.31	2.26	2.21	2.16	2.11
$\infty$	2.85	2.79	2.71	2.65	2.59	2.53	2.49	2.45	2.42	2.37	2.31	2.26	2.20	2.15	2.11

 $\dagger F = 2.4 (a = .645, b = 1.309)$  $\dagger F = 3.0 (a = .577, b = 1.225)$

$\dagger F = 4.0 (a = .500, b = 1.155)$ 

2	2.58	2.44	2.35	2.29	2.25	2.22	2.20	2.18	2.15	2.12	2.09	2.06	2.03	2.00	
4	2.85	2.63	2.50	2.41	2.35	2.30	2.27	2.24	2.22	2.18	2.15	2.12	2.08	2.05	2.01
6	2.85	2.65	2.52	2.43	2.37	2.32	2.28	2.25	2.23	2.19	2.16	2.12	2.08	2.04	2.01
10	2.85	2.67	2.55	2.46	2.39	2.34	2.30	2.26	2.24	2.20	2.16	2.12	2.08	2.04	2.00
20	2.85	2.69	2.57	2.47	2.40	2.35	2.30	2.27	2.24	2.20	2.15	2.11	2.07	2.03	1.99
$\infty$	2.85	2.71	2.59	2.49	2.42	2.36	2.31	2.27	2.24	2.19	2.15	2.11	2.06	2.02	1.99

 $\dagger F = 6.0 (a = .408, b = 1.095)$ 

2	2.85	2.53	2.37	2.27	2.21	2.16	2.13	2.10	2.08	2.04	2.02	1.99	1.96	1.93	1.90
4	2.85	2.56	2.40	2.30	2.23	2.18	2.14	2.12	2.09	2.06	2.02	1.99	1.96	1.93	1.90
6	2.85	2.58	2.42	2.31	2.24	2.19	2.15	2.12	2.09	2.06	2.02	1.99	1.95	1.92	1.89
10	2.85	2.59	2.43	2.32	2.24	2.19	2.15	2.12	2.09	2.06	2.02	1.99	1.95	1.92	1.89
20	2.85	2.60	2.44	2.32	2.25	2.19	2.15	2.12	2.09	2.05	2.02	1.98	1.95	1.92	1.89
$\infty$	2.85	2.61	2.44	2.33	2.25	2.19	2.15	2.12	2.09	2.05	2.02	1.98	1.95	1.92	1.89

 $\dagger F = 10.0 (a = .316, b = 1.054)$ 

2	2.85	2.48	2.30	2.19	2.12	2.07	2.04	2.01	1.99	1.96	1.93	1.90	1.87	1.85	1.82
4	2.85	2.49	2.31	2.20	2.13	2.08	2.04	2.01	1.99	1.96	1.93	1.90	1.87	1.84	1.82
6	2.85	2.50	2.31	2.20	2.13	2.08	2.04	2.01	1.99	1.96	1.93	1.90	1.87	1.84	1.82
$10-\infty$	2.85	2.51	2.32	2.20	2.13	2.08	2.04	2.01	1.99	1.96	1.93	1.90	1.87	1.84	1.82

 $\dagger F = 25.0 (a = .200, b = 1.021)$ 

2-4	2.85	2.40	2.20	2.10	2.03	1.99	1.95	1.93	1.91	1.88	1.86	1.83	1.80	1.78	1.76
$10-\infty$	2.85	2.41	2.21	2.10	2.03	1.99	1.95	1.93	1.91	1.88	1.86	1.83	1.80	1.78	1.76

 $\dagger F = \infty (a = 0, b = 1)$ 

2- $\infty$	2.85	2.33	2.13	2.03	1.97	1.93	1.90	1.88	1.86	1.84	1.81	1.79	1.76	1.74	1.72
-------------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------

• Todas las diferencias no significativas.

† Ver condiciones límite en la página 615.

Tabla A.22 Valores críticos para la prueba de una muestra de Kolmogorov-Smirnov

Prueba unilateral, $\alpha =$	.10	.05	.025	.01	.005
Prueba bilateral, $\alpha =$	.20	.10	.05	.02	.01
n = 1	.900	.950	.975	.990	.995
2	.684	.776	.842	.900	.929
3	.565	.636	.708	.785	.829
4	.493	.565	.624	.689	.734
5	.447	.509	.563	.627	.669
6	.410	.468	.519	.577	.617
7	.381	.436	.483	.538	.576
8	.358	.410	.454	.507	.542
9	.339	.387	.430	.480	.513
10	.323	.369	.409	.457	.489
11	.308	.352	.391	.437	.468
12	.296	.338	.375	.419	.449
13	.285	.325	.361	.404	.432
14	.275	.314	.349	.390	.418
15	.266	.304	.338	.377	.404
16	.258	.295	.327	.366	.392
17	.250	.286	.318	.355	.381
18	.244	.279	.309	.346	.371
19	.237	.271	.301	.337	.361
20	.232	.265	.294	.329	.352
21	.226	.259	.287	.321	.344
22	.221	.253	.281	.314	.337
23	.216	.247	.275	.307	.330
24	.212	.242	.269	.301	.323
25	.208	.238	.264	.295	.317
26	.204	.233	.259	.290	.311
27	.200	.229	.254	.284	.305
28	.197	.225	.250	.279	.300
29	.193	.221	.246	.275	.295
30	.190	.218	.242	.270	.290
31	.187	.214	.238	.266	.285
32	.184	.211	.234	.262	.281
33	.182	.208	.231	.258	.277
34	.179	.205	.227	.254	.273
35	.177	.202	.224	.251	.269
36	.174	.199	.221	.247	.265
37	.172	.196	.218	.244	.262
38	.170	.194	.215	.241	.258
39	.168	.191	.213	.238	.255
40	.165	.189	.210	.235	.252
Aproximación para					
$n > 40:$		1.0730	1.2239	1.3581	1.5174
		$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$

Fuente: Esta tabla es un extracto de "Table of percentage points of Kolmogorov Statistics", J. Amer. Statist. Assoc., 51: 111-121 (1956), con autorización del autor, L. H. Miller, y el editor.

**Tabla A.23A Valores críticos para la prueba dos muestras de Kolmogorov-Smirnov,  $n_1 = n_2$**

Prueba unilateral	$1 - \alpha = 0.90$	0.95	0.975	0.99	0.995
Prueba bilateral	$1 - \alpha = 0.80$	0.90	0.95	0.98	0.99
$n = 3$	2/3	2/3			
4	3/4	3/4	3/4		
5	3/5	3/5	4/5	4/5	4/5
6	3/6	4/6	4/6	5/6	5/6
7	4/7	4/7	5/7	5/7	5/7
8	4/8	4/8	5/8	5/8	6/8
9	4/9	5/9	5/9	6/9	6/9
10	4/10	5/10	6/10	6/10	7/10
11	5/11	5/11	6/11	7/11	7/11
12	5/12	5/12	6/12	7/12	7/12
13	5/13	6/13	6/13	7/13	8/13
14	5/14	6/14	7/14	7/14	8/14
15	5/15	6/15	7/15	8/15	8/15
16	6/16	6/16	7/16	8/16	9/16
17	6/17	7/17	7/17	8/17	9/17
18	6/18	7/18	8/18	9/18	9/18
19	6/19	7/19	8/19	9/19	9/19
20	6/20	7/20	8/20	9/20	10/20
21	6/21	7/21	8/21	9/21	10/21
22	7/22	8/22	8/22	10/22	10/22
23	7/23	8/23	9/23	10/23	10/23
24	7/24	8/24	9/24	10/24	11/24
25	7/25	8/25	9/25	10/25	11/25
26	7/26	8/26	9/26	10/26	11/26
27	7/27	8/27	9/27	11/27	11/27
28	8/28	9/28	10/28	11/28	12/28
29	8/29	9/29	10/29	11/29	12/29
30	8/30	9/30	10/30	11/30	12/30
31	8/31	9/31	10/31	11/31	12/31
32	8/32	9/32	10/32	12/32	12/32
33	8/33	9/33	11/33	12/33	13/33
34	8/34	10/34	11/34	12/34	13/34
35	8/35	10/35	11/35	12/35	13/35
36	9/36	10/36	11/36	12/36	13/36
37	9/37	10/37	11/37	13/37	13/37
38	9/38	10/38	11/38	13/38	14/38
39	9/39	10/39	11/39	13/39	14/39
40	9/40	10/40	12/40	13/40	14/40
Aproximación para $n > 40$ :	1.5174 $\sqrt{n}$	1.7308 $\sqrt{n}$	1.9206 $\sqrt{n}$	2.1460 $\sqrt{n}$	2.3018 $\sqrt{n}$

Fuente: Esta tabla es un extracto de "Small-sample distribution for multi-sample statistics of the Smirnov type", *Ann. Math. Statist.*, 31: 710-720 (1960), con autorización de los autores, Z. W. Birnbaum y R. A. Hall, y el editor.

**Tabla A.23B** Valores críticos para la prueba de dos muestras de Kolmogorov-Smirnov,  $n_1 \neq n_2$ 

Prueba unilateral		$1 - \alpha = 0.90$	0.95	0.975	0.99	0.995
Prueba bilateral		$1 - \alpha = 0.80$	0.90	0.95	0.98	0.99
$n_1 = 3$	$n_2 = 4$	3/4	3/4			
	5	2/3	4/5	4/5		
	6	2/3	2/3	5/6		
	7	2/3	5/7	6/7	6/7	
	8	5/8	3/4	3/4	7/8	
	9	2/3	2/3	7/9	8/9	8/9
	10	3/5	7/10	4/5	9/10	9/10
	12	7/12	2/3	3/4	5/6	11/12
$n_1 = 4$	$n_2 = 5$	3/5	3/4	4/5	4/5	
	6	7/12	2/3	3/4	5/6	5/6
	7	17/28	5/7	3/4	6/7	6/7
	8	5/8	5/8	3/4	7/8	7/8
	9	5/9	2/3	3/4	7/9	8/9
	10	11/20	13/20	7/10	4/5	4/5
	12	7/12	2/3	2/3	3/4	5/6
	16	9/16	5/8	11/16	3/4	13/16
$n_1 = 5$	$n_2 = 6$	3/5	2/3	5/6	5/6	
	7	4/7	23/35	5/7	29/35	6/7
	8	11/20	5/8	27/40	4/5	4/5
	9	5/9	3/5	31/45	7/9	4/5
	10	1/2	3/5	7/10	7/10	4/5
	15	8/15	3/5	2/3	11/15	11/15
	20	1/2	11/20	3/5	7/10	3/4
$n_1 = 6$	$n_2 = 7$	23/42	4/7	29/42	5/7	5/6
	8	1/2	7/12	2/3	3/4	3/4
	9	1/2	5/9	2/3	13/18	7/9
	10	1/2	17/30	19/30	7/10	11/15
	12	1/2	7/12	7/12	2/3	3/4
	18	4/9	5/9	11/18	2/3	13/18
	24	11/24	1/2	7/12	5/8	2/3
$n_1 = 7$	$n_2 = 8$	1/2	33/56	5/8	41/56	3/4
	9	31/63	5/9	40/63	5/7	47/63
	10	33/70	39/70	43/70	7/10	53/70
	14	3/7	1/2	4/7	9/14	5/7
	28	3/7	13/28	15/28	17/28	9/14
$n_1 = 8$	$n_2 = 9$	4/9	13/24	5/8	2/3	3/4
	10	19/40	21/40	23/40	27/40	7/10
	12	11/24	1/2	7/12	5/8	2/3
	16	7/16	1/2	9/16	5/8	5/8
	32	13/32	7/16	1/2	9/16	19/32

**Tabla A.23B Valores críticos para la prueba de dos muestras de Kolmogorov-Smirnov,  $n_1 \neq n_2$  (continuación)**

Prueba unilateral		$1 - \alpha = 0.90$	$0.95$	$0.975$	$0.99$	$0.995$
Prueba bilateral		$1 - \alpha = 0.80$	$0.90$	$0.95$	$0.98$	$0.99$
$n_1 = 9$	$n_2 = 10$	7·15	1/2	26/45	2/3	31/45
	12	4/9	1/2	5/9	11/18	2/3
	15	19/45	22/45	8/15	3/5	29/45
	18	7/18	4/9	1/2	5/9	11/18
	36	13/36	5/12	17/36	19/36	5/9
$n_1 = 10$	$n_2 = 15$	2·5	7/15	1/2	17/30	19/30
	20	2/5	9/20	1/2	11/20	3/5
	40	7/20	2/5	9/20	1/2	
$n_1 = 12$	$n_2 = 15$	23/60	9/20	1/2	11/20	7/12
	16	3/8	7/16	23/48	13/24	7/12
	18	13/36	5/12	17/36	19/36	5/9
	20	11/30	5/12	7/15	31/60	17/30
$n_1 = 15$	$n_2 = 20$	7/20	2/5	13/30	29/60	31/60
$n_1 = 16$	$n_2 = 20$	27/80	31/80	17/40	19/40	41/80
Aproximación para						
muestras grandes: $\sqrt{\frac{n_1 + n_2}{n_1 n_2}} \times$						
		1.0730	1.2239	1.3581	1.5174	1.6276

Fuente: Esta tabla es un extracto de "Distribution table for the deviation between two sample cumulatives", *Ann. Math. Statist.*, 23: 435-441 (1952), con autorización del autor, F. J. Massey, Jr., y el editor.

**Tabla A.24 Alfabeto griego**

(Letra y nombre)

A $\alpha$	Alfa	H $\eta$	Eta	N $\nu$	Nu	T $\tau$	Tau
B $\beta$	Beta	$\Theta \theta$	Theta	$\Xi \xi$	Xi	$\Upsilon \upsilon$	Upsilon
$\Gamma \gamma$	Gamma	I $\iota$	Iota	O $\circ$	Omicron	$\Phi \phi$	Phi
$\Delta \delta$	Delta	K $\kappa$	Kappa	$\Pi \pi$	Pi	X $\chi$	Ji
E $\epsilon$	Epsilon	$\Lambda \lambda$	Lambda	P $\rho$	Rho	$\Psi \psi$	Psi
Z $\zeta$	Zeta	M $\mu$	Mu	$\Sigma \sigma$	Sigma	$\Omega \omega$	Omega

## DISTRIBUCION DE PROBABILIDAD NORMAL ESTANDAR

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
-3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
-2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3065	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

# INDICE

- Abscisa, 232
- Aditividad:  
con transformaciones, 226  
de efectos e interacciones principales, 332-334  
de  $\chi^2$  cuadrado ( $\chi^2$ ), 498-501  
de efectos, 162, 164  
de sumas de cuadrados, 136, 196
- Agrupación:  
doble (*ver Cuadrado latino*)  
efecto de, 34
- Ajuste:  
de continuidad (*ver Corrección de continuidad*)  
de la media, 22, 135  
de medias de tratamientos, 393-394, 406, 422  
prueba de, 398-400, 404-406, 422-423
- Aleatoriedad, 9-10, 539
- Aleatorización: 129-130, 132-133, 163, 189-190  
en dos etapas, 369  
para el cuadrado latino, 215  
para el diseño de bloques, 132-133  
completamente aleatorio, 132-133, 189-190  
prueba de, 539
- Amplitud, 23  
mínima significante, 181  
múltiple, 180  
studentizada, 179, 181.
- Análisis con un número proporcionado de subclases, 429, 432-439, 441
- Análisis de covarianza, 392  
aumento de precisión mediante el, 408  
ecuación del modelo, 396-398  
estimaciones de parámetros, 397-398  
múltiple, 418  
para factoriales, 413-415, 420-422  
para bloques aleatorizados, 391, 395  
supuestos, 396-398  
usos, 392
- Análisis de varianza: 134, 495-499  
para bloques aleatorizados, 190-193  
con muestreo, 148-154, 156-159  
para el cuadrado latino, 215-218  
para bloques divididos, 381-383  
para parcelas divididas, 374-376  
en espacio y tiempo, 384-386  
para un diseño completamente aleatorio, 134, 138  
con muestreo, 208-210  
con número desigual de repeticiones, 140-141  
ponderado, 498  
supuestos, 162-164
- Análisis probit, 239
- Argumento, 231
- Asignación:  
óptima, 161, 550-552  
proporcional, 550
- Banda de confianza, 246-247
- Base por observación, 136-137, 152, 171, 185, 201, 514

- Bloques:**
- completos, 126, 127, 190-191, 193
  - divididos, 381
  - incompletos, 127, 369
  - variación entre, 193
- Captura-recaptura, 509-510**
- Clasificación:**
- anidada (*ver también* Diseño anidado)
  - de dos vías, 189
  - de una vía (*ver* Diseño completamente aleatorio)
  - jerárquica, 150, 154
  - (*ver también* Diseño anidado)
- Codificación:**
- en la tabla de frecuencias, 32-34
  - (*ver también* Polinomios ortogonales)
- Coeficiente:**
- de alienación, 267
  - de confianza, 62, 65
  - de contingencia, 480
  - de correlación (*ver* Coeficiente de correlación)
  - de determinación, 258, 267
    - múltiple, 319
  - de indeterminación, 267
  - de regresión, 233, 245-246, 258
  - de variabilidad, 25-27, 30, 154
  - de variación, 25-26, 30, 154, 376
- Coeficiente de correlación, 263-267, 409, 410**
- distribución del, 269-271
  - homogeneidad del, 271-273
  - intervalo de confianza para el, 269-270
  - momento-producto, 263
  - múltiple, 307, 316, 318
  - parcial, 307, 316, 317
  - prueba del, 269-271
  - sesgo en el, 272-273
  - (*ver también* Correlación)
- Coeficiente(s) de regresión, 233, 245-246, 258, 324**
- de tratamiento contra error, 409-410
  - diferencia entre, 250
  - estándar, 225-226
  - homogeneidad del, 412
  - para polinomios ortogonales, 362
  - parcial, 307, 313, 398
- Cofactor, 286, 317, 318**
- Comparaciones:**
- independientes, 499-501
  - múltiples, 168
  - no planeadas, 167-169, 177, 427
  - planeadas, 167-169, 171, 178-180, 334
  - pareadas, 169, 179, 185-186
  - significantes, 168-169, 333-334
- (ver también Contrastes)
- Complemento, 41**
- Componentes de la varianza, 159-161**
- (*ver también* Valor esperado de cuadrados medios)
- Conformes:**
- para la adición, 279
  - para la multiplicación, 280
- Conjunto exhaustivo de resultados (eventos), 38**
- Contrastes, 174**
- con media, repetida de manera desigual, 185
  - de datos, 433
  - no ortogonales, 174, 433
  - ortogonales (*ver* Polinomios ortogonales)
- Control del error (*ver también* Error experimental)**
- Corrección de continuidad:**
- con prueba de signos, 526-527
  - en tablas de dos celdas, 472
  - en tablas de  $2 \times 2$ , 489, 491
- Corrección de población finita, 546, 550**
- Correlación, 263-264, 268-269**
- de interclases, 274
  - de intraclasses, 273, 275
  - de rangos de Spearman, 536-537
  - lineal, 263
  - momento-producto, 274
  - múltiple, 304, 307
  - parcial, 303, 307
  - simple, 303, 307
  - total, 303, 307
- (ver también Coeficiente de correlación)
- Costo:**
- en el muestreo, 550-551
  - en la planeación de experimentos, 159-160
- Covarianza, 106-108, 234**
- entre  $b_0$  y  $b_1$ , 294
  - entre los coeficientes de regresión parcial, 313
  - entre  $Y$  y  $b$ , 246
- (ver también Análisis de covarianza)
- cpf (corrección de población finita), 546, 550**
- Criterio de la prueba de Ji cuadrado, 458, 466, 483**
- aditividad del, 498-499
  - ajustado (*ver* Corrección por continuidad)
  - bondad del ajuste (*ver* Prueba de bondad del ajuste)
  - con muestras dependientes, 497
  - distribución de Poisson, 517-518
  - muestras independientes, 483, 485
  - para la regresión lineal en tablas de  $r \times 2$ , 501-503
  - tablas de contingencia, 482-486
  - tablas de dos celdas, 470-471
  - tablas de  $2 \times 2$ , 493-495

- tablas de  $2 \times 1 \times 2$ , 504-506
- tablas de una vía, 480
- Criterio de prueba (*ver Prueba de significación*)
- Cuadrado latino, 213-215
  - aleatorización, 215
  - análisis del, 215-218
  - aumentado, 221
  - datos faltantes, 219-220
  - eficiencia, 221-223
  - incompleto, 220-221
  - modelos, 223
  - valores esperados de los cuadrados medios, 223
- Cuadrado medio, 20
  - residual (*ver Error experimental*)
    - (*ver también Cuadrado latino, valor esperado de los cuadrados medios*)
- Cuartiles, 17
- Cuasi-razón, 348
- Curva:
  - de poder, 109-110
  - exponencial, 443, 444-450
  - gaussiana, 46
  - laplaciana, 46
  - logarítmica, 443, 444-450
  - normal, 3, 46, 48
  - historia de la, 48
- Darwin, Charles, 3
- Datos, 7, 9-10
  - binomiales (*ver Distribución binomial*)
  - continuos, 14
  - cualitativos, 12
  - cuantitativos, 13
  - discretos, 14
  - faltantes (*ver Parcela faltante*)
  - por categoría (*ver Datos enumerativos*)
  - presentación de, 12-14
- Datos con un número no proporcionado de subclases, 429
  - análisis de, 429, 432-439
  - en porcentajes, 228
  - modelo de, 429
- Datos enumerativos, 12, 466, 508
  - en tablas de contingencia 482, 485-486
- Deciles, 17
- De Méré, 2
- De Moivre, A., 2, 48
- Dependencia, 488
  - lineal, 284
- Desigualdad de Chebyshév, 534-536
- Desviación, 57
  - de la muestra, 16, 20
  - media, 23
  - media absoluta, 23
- media cuadrática, 21
- Desviación estándar, 20-21
  - a partir de la tabla de frecuencia, 32
  - con codificación, 30-31
  - de diferencias, 76-78
  - de la media, 24-25, 72-73
  - de la regresión parcial, 322
  - de una estimación no sesgada, 71-72
  - de  $Y$  para  $X$  fija, 244
    - (*ver también Error estándar*)
- Determinante, 285-286
- Diagonal principal, 278, 294
- Diagrama:
  - de barras, 12
  - de dispersión, 264-266
- Diagramas, 12, 14
- Dicotomía, 495
- Diferencia mínima significante, de Fisher (protegida), 170
- Discrepancia, 136
- Diseño anidado, 150
- Diseño completamente aleatorizado, 132-134
  - aleatorización 132, 207-208
  - análisis de, 134
    - no paramétrico, 535-536
    - con muestreo, 148-154
    - modelos de, 144-147, 154-155, 274
- Diseño de bloques completos al azar, 126, 190
  - análisis de, 190-193
  - análisis no paramétrico de, 532-535
  - con covarianza, 401-405, 418-423
  - con datos faltantes, 202-206
  - eficiencia del, 207-208
  - generalizado, 190, 208, 428
  - modelos para, 211-213, 396-398
  - término de error, 196-201
- Diseño de bloques divididos, 381-383
  - modelos para, 384-386
- Diseños de bloques generalizado, 428
- Diseño de bloques incompletos, 126, 369
  - balanceado, 127
- Diseños muestrales, 543
- Diseño de parcelas divididas, 126, 368-373
  - análisis del, 374-378
  - con datos faltantes, 379-380
  - en espacio y tiempo, 384-386
  - modelos, 384
- Diseño experimental,
  - (*ver también los diseños específicos*)
- Diseño sistemático, 130, 163
- Diseños aumentados, 221
- Dispersion, 19-20, 23
- Distribución, 8-9
  - acumulada, 9, 46

- binomial, 40-43, 46-47, 467, 508, 510, 511-513
- bivariante, 256-258, 263
- de diferencias, 76-79
- de Ji cuadrado, 56-58, 466-467
  - para combinación de probabilidades, 468
  - para intervalos de confianza en  $\sigma^2$ , 458-459
- del coeficiente de correlación, 268
- de medias, 55-56, 68-70
- de Poisson, 510-517
  - ajuste de la 515-517
  - prueba de, 517-518
  - transformación para la, 228, 514
- de probabilidad, acumulada, 41-43
- derivada, 27, 30, 55
- F, 136-137
- hipergeométrica, 508-510, 546
- multinomial, 41, 481
- normal, 46-48, 65, 68, 70-71
  - probabilidades para la, 48-54
- principal, 24, 55, 68, 70
- sesgada, 72
- $t$ , 58, 73-75, 103
  - de Student, 58, 73, 79, 103
  - uniforme, 45
  - $\chi^2$  (Ji cuadrado), 56
- DMS (*ver Prueba de Waller-Duncan*)
- dms (diferencia mínima significante), 185
  
- Ecuación de regresión (*ver Recta de regresión*)
  - estándar, 325
  - parcial, 307, 311-312, 325
- Ecuaciones de mínimos cuadrados, 290-291, 434
  - (*ver también Ecuaciones normales*)
- Ecuaciones inconsistentes, 284
- Ecuaciones normales, 290, 309
  - con variables indicadoras, 298-299
  - para la regresión lineal, 291
  - para la regresión múltiple, 309-310
  - para la regresión polinomial, 450-451
    - (*ver también Ecuaciones de mínimos cuadrados*)
- Efecto principal (*ver Experimento factorial*)
- Efecto simple (*ver Experimento factorial*)
- Efectos:
  - aditivos (*ver Modelo lineal aditivo*)
  - residuales, 221
- Eficiencia (*ver Eficiencia relativa*)
- Eficiencia relativa, 60, 125
  - a partir de la asignación óptima, 552
  - a partir de la estratificación, 547
  - con covarianza, 408
  - para el cuadrado latino, 221-222
  
- para el diseño de bloques, 193, 207-208
- para procedimientos no paramétricos, 520-521
- Eliminación gaussiana, 320
- Empates, 526, 530-531
- Encuesta, 541-542
  - muestral, 541-542
- Ensayo:
  - binomial, 40
  - biológico, 256
  - de Bernoulli, 40
  - de uniformidad, 127, 393
- Ensayos independientes, 41-42, 485, 510
- Error:
  - combinado, 136, 162-163
  - cuadrático medio, 408
  - de clase dos (*ver Error tipo II*)
  - de clase uno (*ver Error tipo I*)
  - de muestreo, 27, 148, 150, 152-153, 158, 210, 547, 554
  - generalizado, 136, 163-164
  - puro, 429
- Error estándar, 24-25, 30, 71-72
  - de la diferencia, 193
    - con desigual número de repeticiones, 407
    - con valores faltantes, 202-206, 219-221, 380
    - entre medias ajustadas, 407
  - de la media ajustada, 396, 398, 421-423
  - de la regresión lineal, 322-323
  - de una estimación, 244
  - para parcelas divididas, 372
- Error experimental, 121-122, 125-127, 129, 153-159, 160-162, 166, 209-210, 392-393, 554
  - control del, 125, 126, 392, 426
  - mediante covarianza, 392-393, 426
  - naturaleza del, 196-198
  - partición del, 198-200
- Error tipo I, 85, 109-114, 166, 178, 476
- Error tipo II, 85, 109-114, 476
- Errores, 25-30, 162
  - correlacionados, (*ver también Residual*)
- Escalar, 279, 285
  - multiplicación por un, 280
- Espacio muestral, 380
- Esperanza (*ver Valor esperado*)
- Estadística:
  - definición, 1, 2
  - de prueba (*ver Prueba de significación*)
  - historia de la, 2-4
  - no paramétrica, 520, 521
  - paramétrica, 520
    - (*ver también Estimación*)
- Estimación, 59-62, 296-297
  - conjunta, 324

- (ver también Estimaciones por mínimos cuadrados)
- Estimaciones por mínimos cuadrados, 28, 196, 202, 290-291, 397
- Estimador sesgado, 61
- Estimadores lineales, 59-61
- Estratificación, 547-548
- Estrato, 547, 548, 552
- Experimento, 118-120
- aleatorio, 39
  - factorial, 328-334
    - con covarianza, 413-416, 417-423
    - de 2 x 2, 334
    - de 3 x 3 x 2, 340-343
    - modelos lineales para, 346-351
    - superficies de respuesta, 352-354
    - con polinomios ortogonales, 358-359
      - (ver también Diseño de bloques divididos; diseño de parcelas divididas)
- Extrapolación, 254-255, 324
- Factor (ver Experimento factorial)
- de eficiencia, 207, 222
  - de mejoramiento, 267
  - de precisión, 207, 222-223
- Falta de ajuste, 355
- Familia de hipótesis, 176, 179
- Fermat, 2
- Fiducial ( $F$ ), 62
- Fisher, R. A., 3, 88, 129
- Función, 103
  - de densidad, 9, 47
  - de densidad de probabilidades, 8-9, 44-46
  - de densidad normal, 48
  - de distribución acumulada, 9, 46
  - lineal, 60
    - varianza de la, 105-106
  - (ver también Contrastes)
- Fracción de muestreo, 546, 550-551
  - ajuste de la, 551
  - variable, 550-551
- Gauss, K.F., 2
- Gosset, W. S., 3, 58, 88
- Grados de libertad, 23, 155-158, 349
  - efectivos, 102, 103, 157, 349
- Grupo de resultados, 189
- Heterogeneidad (ver Homogeneidad)
- Hipótesis:
- alternativa, 85, 86 481
  - nula (ver Prueba de significación)
- Histograma, 13-14, 44
- Homogeneidad:
- de componentes del error, 199-201
- de Ji cuadrado, 498
- de la regresión, 250, 409-412
  - tratamiento contra error, 409-410
- de la varianza, 108, 461
- de muestras independientes, 487-488, 495-496
- Impresión, 320-322, 424, 426, 437
- Improbables, 37
- Independencia:
- de errores, 163
  - lineal, 284
    - (ver también Interacción)
- Índice:
- de columna, 277
  - de fila, 277
  - de sumatoria, 16
- Inferencia, 2, 4, 5, 59-62
  - (ver también Estimaciones de mínimos cuadrados)
- Información, cantidad de (ver Precisión; Eficiencia relativa)
- Inseguimiento, 59, 72, 546
  - de la estimación de  $\sigma^2$ ,
  - de  $s^2$ , 72, 417
  - de  $S^2$ , 546
  - de  $2S^2$ , 78-79
  - de SC (tratamientos) con datos estimados, 417
    - (ver también Sesgo)
- Inspección de calidad, 509
- Interacción, 210, 325, 332, 337, 342-344, 359, 360, 376
  - en tablas de contingencia, 483-486
  - heterogeneidad de la, 389
    - (ver también Experimento factorial)
- Intercorrelación, 323
- Intervalo,
- de clase, 32
  - de la media,
  - de tolerancia, 75
  - studentizado,
- Intervalo de confianza, 61-63, 75, 83-84, 121
  - con corrección de población finita, 550
  - de longitud fija, 116-117, 224-226
  - para proporciones, 467-471, 551-552
  - para una predicción, 63, 253-255
  - simultáneo, 177-178, 183, 315
    - (ver también Banda de confianza)
- Intervalos:
- de confianza conjuntos (ver Intervalo de confianza simultáneo)
  - mínimos significantes, 181
  - múltiples,
- Intrínsecamente lineal, 443
- Inversa, 283-287

- por la derecha, 284
- por la izquierda, 284
- Ji cuadrado:
  - corregida (*ver* Corrección por continuidad)
  - independiente (*ver* aditividad de Ji cuadrado)
- Laplace, P. S. de, 2
- Látices parcialmente balanceados, 126
- Ley estadística, 233
- Límites de confianza (*ver* Intervalo de confianza)
- Lineal en los parámetros, 443
- Lyell, Charles, 2
- Matriz (ces), 277
  - adicción de, 279
  - álgebra, 278
  - cuadrada, 278, 281
  - de correlación, 317
  - de diseño, 289
  - dimensión de una, 278
  - ecuación matricial, 289-291, 420
  - identidad, 281
  - inversa, 283, 285, 286, 317
  - multiplicación de, 280
  - no singular, 285, 291
  - simétrica (*idéntica*), 279
  - singular, 285, 298, 434
  - sustracción de, 279
  - varianza-covarianza, 294
- Máximos, 443
- Media:
  - aritmética, 15, 17, 19, 30-31, 35, 544
  - armónica, 17-18
  - de Ji cuadrado, 57
  - de la muestra, 25, 28, 30
  - de una función lineal, 103-104
  - de una tabla de frecuencias, 34-35
  - de una variable binomial, 43, 551
  - distribución de la, 55
  - geométrica, 17
  - no ponderada, 432-433
  - ponderada, 17, 93, 141, 146, 261, 272, 430, 433, 502
- Mediana, 17
- Medias de tratamientos ajustadas, 393-394
  - pruebas de, 398-400
- Mendel, G., 3
- Menor, 285, 317
  - con signo, 317
- Método:
  - científico, 4-5
  - de ajustes de constantes, 433, 439
  - de cuadrados de medias ponderadas, 440
- de Doolittle, 320
- de medias no ponderadas, 440
- de mínimos cuadrados, 433, 437
- de prueba de hipótesis, 437
- empírico, 65
- minimax, 474, 478
- Mínimos, 443
- Moda, 17
- Modelo aditivo (*ver* Modelo lineal aditivo)
- Modelo completo (lineal), 311, 397, 400, 420
- Modelo de efectos aleatorios (*ver* Modelo II)
- Modelo de efectos fijos (*ver* Modelo I)
- Modelo de regresión, 236-240
  - lineal (*ver* Modelo lineal aditivo)
- Modelo lineal aditivo, 27-28
  - con correlación de intraclasses, 274-275
  - con covarianza, 396-398
  - para la regresión, 236-238
  - para muestras independientes, 96-97
  - para observaciones pareadas, 101-102
  - para polinomios, 324-325
  - para regresión múltiple, 308
- Modelo mixto:
  - para bloques aleatorizados, 211-213
  - para el cuadrado latino, 223
  - para factoriales, 347-351
  - para la prueba de la mediana, 534
  - para parcelas divididas y bloques divididos, 384
- Modelo reducido, 311, 312, 398
- Modelo I:
  - para bloques aleatorizados, 211, 213
  - para el cuadrado latino, 223
  - para el diseño completamente aleatorio, 144-146
  - para factoriales, 347-348
  - para la prueba de la mediana, 534
  - para la regresión, 236-238
    - en la predicción de  $X$ , 256
  - para parcelas divididas y bloques divididos, 384
- Modelo II:
  - con correlación de intraclasses, 274
  - para bloques aleatorizados, 211-213
  - para el cuadrado latino, 223
  - para el diseño completamente aleatorio, 144-146
  - para factoriales, 347-351
  - para la prueba de la mediana, 534
  - para la regresión, 236-240, 256-258
  - para parcelas divididas y bloques divididos, 384
- Muestra, 9-10, 65
  - autoponderada, 544-550
  - completamente aleatoria, 11

- desviación estándar de la (*ver* Desviación estándar)
  - en dos etapas, de Stein, 116-117
- Muestras:
  - aleatorias, 10, 68, 541-542
  - independientes, 93, 102-103, 487, 489, 499, 501
  - no independientes, 493, 499-501
- Muestreo:
  - con reemplazo, 11, 510
  - sin reemplazo, 508, 510, 544
  - sistemático, 542
    - (*ver también* tipos específicos de muestreo, a continuación)
- Muestreo aleatorio estratificado, 547-550
- Muestreo aleatorio simple, 544-547
- Muestreo autoritario, 542
- Muestreo de aceptación, 509
- Muestreo de área, 553, 555-559
- Muestreo de población finita, 541
- Muestreo probabilístico, 543-544
- Muestreo de una etapa, 553
- Muestreo multietápico, 553-556
- Muestreo por conglomerados, 553-556
- Muestreo por cuotas, 491
- Muestreo por conglomerado simple, 553
- Multicolinealidad, 324
- Neyman, J., 3
- Nivel (*ver* Experimento factorial)
  - de protección, 181
  - de significancia, 62, 84-85, 87
- No aditividad:
  - prueba de Tukey para la, 529
    - (*ver también* Aditividad)
- No independencia (*ver* Interacción)
- No linealidad de los parámetros, 443
- Nonios ("no usual"), 55-56
- "No responden", 550
- Notación de puntos, 134-135, 150-151, 190
- Número efectivo de repeticiones:
  - en el cuadrado latino, 220-221
  - en el diseño de bloques, 206, 220-221
- Observaciones concomitantes, uso de, 127, 231, 248-249
- Observaciones pareadas:
  - aleatoriamente, 76
  - en la prueba del signo, 526-527
  - significativas, 98-99, 107
- Orden:
  - de una matriz, 277
  - de un modelo, 324
- Ordenada, 232
  - al origen, 232
- Ortogonalidad, 174
  - de bloques y tratamientos, 189, 203
- Papel para gráficas de probabilidad binomial, 469
  - para intervalo de confianza, 469-470
  - para prueba de hipótesis, 474, 490
  - para tamaño de la muestra, 478-480
- Parámetro, 16, 59
  - de incomodidad, 103
  - observable, 43, 237
- Parcela experimental (*ver* Unidad experimental)
- Parcela faltante:
  - en bloques aleatorios, 202-207
  - en el cuadrado latino, 219-221
  - mediante covarianza, 417-418
- Parcelas completas, 368, 369-370
- Parcamiento:
  - aleatorio, 76
  - significativo (*ver* Observaciones pareadas significativas)
- Pascal, B., 2
- Pearson, E. S., 3
- Pearson, Karl, 2
- Pendiente, 233
  - (*ver también* Coeficiente de regresión)
- Percentiles, 17
- Permutación, 133, 189
- Peso del error, 184
- Plano de regresión, 304-306
- Población, 9-10
  - finita, 43, 541
    - (*ver también* Distribución)
- Poder, 85, 109, 113, 121
- Polygono de frecuencias, 13-14
- Polinomio (al)
  - de segundo grado, 450-451
  - modelos, 324
    - (*ver también* Polinomios ortogonales)
- Polinomios ortogonales, 355-362, 451-456
- Ponderaciones, 147
- Postmultiplicación, 279
- Precisión, 119, 122, 123-125, 127, 139, 261, 273, 393
  - (*ver también* Eficiencia relativa)
- Predicción, 63, 253, 256, 296-297, 324
  - de  $X$ , 256
- Premultiplicación, 279
- Probabilidad, 37-40, 41, 43
  - de confianza, 65
- Probabilidades independientes (*ver* Ensayos independientes)
- Promedio (*ver* Media)
- Prueba:
  - de efectos sugerida por los datos, 175-176,

153037

- de significancia, 83-88  
 de las dos colas, 87, 108  
 de una cola, 87, 108  
 para no aditividad, 363-365  
 para un conjunto limitado de alternativas, 474-478  
 poder de una (*ver Poder*)  
 secuencial, 320  
 simultánea (*ver Comparaciones múltiples*)
- Prueba de Bartlett**, 228, 462-463
- Prueba de bondad del ajuste**, 467  
 de Kolmogorov-Smirnov, 522, 524, 527  
 para datos binomiales (binómicos), 512-513  
 para datos con distribución de Poisson, 515-516  
 para datos en categorías, 521  
 para polinomios, 451  
 para una distribución continua, 461-463, 467
- Prueba de Duncan**, 181-184, 186
- Prueba de Dunnett**, 182-183, 186
- Prueba dms de Fisher** (protegida), 170
- Prueba de Friedman**, 532-533
- Prueba exacta de Fisher**, 491-493
- Prueba F**, 92, 93, 108-109, 136-138, 147
- Prueba de Kolmogorov-Smirnov**:  
 con dos muestras, 527  
 con una muestra, 522-524
- Prueba de Kruskal-Wallis**, 530-531
- Prueba de la mediana**, 524-525, 530  
 para  $k$  muestras, 532  
 para clasificaciones de dos vías, 534
- Prueba de rangos**, 526, 528-529  
 signados, de Wilcoxon, 527
- Prueba de signos**, 524-525  
 (*ver también Prueba de la mediana*)  
 prueba de Olmstead-Tukey, de asociación del cuadrante, 537-538  
 prueba de Scheffé, 177-178, 179, 185  
 prueba de Student-Newman-Keuls, 180, 181, 186  
 prueba de Tukey, 179-180, 186, 363-365, 529  
 prueba de Waller-Duncan, 184
- Prueba de Wilcoxon-Mann-Whitney**, 528-529
- Prueba nueva de amplitud múltiple**, 181-182, 186
- Pruebas de una cola** (*ver Prueba de significación*)
- Pruebas orientadas por resultados** (*ver Prueba de significación*)
- Punto muestral**, 38
- R-Cuadrado**, 322
- Rango**:  
 de columna, 285  
 de una fila, 285  
 de una matriz, 285
- Rangos, 526, 530-533  
 empatados, 531, 532  
 "Rastreo de datos", 177
- Razón**:  
 ambigua, 478  
 F sintetizada, 158  
 mínima significativa, 227
- Razones genéticas**:  
 alternativas limitadas, 474-478  
 tablas  $r \times c$ , 482  
 tamaño de la muestra, 478-480
- Recaptura de la marca**, 509
- Recta**:  
 ajustada, 233  
 de regresión, banda de confianza para la, 243-245
- Redes parcialmente balanceadas** (látices), 126
- Región**:  
 crítica (*ver Región de rechazo*)  
 de aceptación, 87, 110-114  
 para razones alternativas, 474-478  
 de confianza, 315, 324  
 de rechazo, 87, 110-114  
 elección de la, 92
- Regresión**:  
 a través del origen, 258-260  
 curvilinea, 442  
 fuentes de variación en la, 244  
 lineal múltiple (*ver Regresión múltiple*), 501  
 modelos de, 236-240  
 múltiple, ecuación de, 236-237, 304, 306-307, 320  
 no lineal, 443  
 parcial, 306-307  
 ponderada, 238, 261-262, 501-502  
 suma de cuadrados, 242, 246, 257, 259  
 varianza respecto de la, 242
- Regresiones independientes** (*ver Homogeneidad*)
- Relación funcional**, 233, 238
- Repetición**, 122-124, 126, 130, 190
- Replicación oculta**, 331
- Réplica** (*ver Repetición*)
- Residual**, 241, 290  
 gráfica, 324  
 (*ver también Errores*)
- Restricciones**, 145, 430-432  
 (*ver también Restricciones sobre el modelo*)
- Restricciones sobre el modelo**, 145, 196, 430-432, 433, 434
- Resultado que no se puede descomponer**, 38
- Resultados independientes** (eventos), 41, 510
- Resultados mutuamente excluyentes** (eventos), 38-40

- A.S. (*ver Impresión*)
- SC PARCIAL, 322-323
- SC SECUENCIAL, 324
- Segregación ambigua, 478
- Sensibilidad (*ver Precisión*)
- Seriedad del error (*ver Peso del error*)
- Sesgo, 58
  - en la correlación transformada, 272-273
  - en la suma de cuadrados de tratamientos con datos estimados, 202, 219-220, 380
  - (*ver también* Insesgamiento)
- Significancia estadística, 124
  - (*ver también* Significante)
- Significante, 87, 169
  - altamente, 88, 376
  - no, 169
- Singularidad, 285
- S-N-K, prueba de (Student-Newman-Keuls), 180-181, 186
- Sobréparametrizado, 144, 434
- "Student" (W.S. Gossett), 3
- Submuestra, 148
- Submuestreo, 159, 553
- Subparcelas, 368, 372-373, 374, 376
- Subunidades, 368
- Suma:
  - de cuadrados, dentro de grupos, 21, 135-136, 241-242
  - no ajustada, 23, 241
  - del cuadrante, 537-539
  - de productos, 234
  - de rangos, 526
  - mínima de cuadrados (*ver Estimaciones mínimas de cuadrados*)
  - ponderada, 146, 430, 432, 499
    - de cuadrados, 142, 147, 262, 430, 433
    - de desviaciones, 240, 262
- Superficie:
  - de regresión, región de confianza de la, 354
  - de respuesta, 353-354, 362
- Supremo, 522
- Supuestos débiles, 520
- Tabla:
  - cuádruple (de  $2 \times 2$ ), 489-490
    - prueba exacta para la, 491-493
    - prueba para muestras no independientes, 493-495
  - de frecuencia, 14, 32-35, 44
  - de números aleatorios, 10, 68, 133
  - tablas de contingencia (*ver también* Criterio de  $\chi^2$ ; prueba de  $\chi^2$  cuadrado)
    - de la muestra:
      - de  $2 \times 2$ , 503-504
- óptimo, 552
- Tasa de error, 62, 84-85
  - por comparación, 168, 176, 246
  - por experimento, 176, 246
  - por familia, 176-177
  - por punto, 246, 254
- Tendencia central, 14, 15, 19
- Término (factor) de corrección, 22, 135
- Transformación, 163, 226-228, 443
  - angular, 228
  - arcosen, 228, 514
  - logarítmica, 163, 227-228
  - para datos binomiales, 228, 514
  - para datos con distribución de Poisson, 227
  - para datos en porcentajes, 227-228
  - para el coeficiente de correlación, 270-272
  - raíz cuadrada,  $\sqrt{Y}$ , 227-228
  - seno inverso, 228
- Transformaciones:
  - sen  $Y - 1/\sqrt{Y}$ , 228
- Transpuesta, 277
- Tratamientos, 120
  - combinación de, 329
  - comparaciones de (*ver Contrastes: Positivos ortogonales*)
  - selección de, 128
  - suma de cuadrados, 135-136
- Tratamientos ficticios, 76
- Triángulo de Pascal, 512-514
- Unidad experimental, 120, 132, 133
  - agrupación de la, 126
  - tamaño y forma de la, 127-128, 208
- Unidad de muestreo, 120, 148, 542-543
  - número de, 159-161
  - primaria (upm), 553
- Unidades completas, 368, 369, 370
- Universo, 9
- Valor:
  - absoluto, 23
  - ajustado, 290
  - crítico, 87, 10
  - de regresión, 242, 290, 307-308, 539
  - promedio (*ver* Valor esperado)
- Valor esperado, 105, 211
  - de cuadrados medios para el diseño completamente aleatorio, 14
  - de funciones lineales, para bloques 104-106, 211-213
  - para cuadrados latinos, 223
  - para factoriales, 349-354
  - para muestreo en dos etapas
  - para parcelas divididas, 384,

- Víncres:**  
 ajustados, 242, 243, 290, 307-308  
 de clase, 13, 32  
 exentos, 341, 539
- Variable:** 7-9  
 aleatoria, 7, 39  
 binaria, 298, 302, 434  
 concomitante, 231  
 continua, 8, 44  
 cualitativa, 8  
 cuantitativa, 8
- X** dependiente, 231, 264  
 de probabilidad, 7-39  
 discontinua, 8  
 discreta, 8, 39  
 estandarizada, 54, 251, 264  
 ficticia, 289, 398, 434
- Y** independiente, 251  
 indicadora, 298, 302, 434  
 normal estandarizada, 24
- Variación (ver también Varianza)**
- Variancia:** 23, 24, 71, 2303, 105, 111, 345, 346  
 de diferencias, 76-78  
 error, 24-25, 346, 547, 549, 549  
 en la media, 24-25, 346, 547, 549, 549  
 en  $\bar{x}$ , 302 ..  
 de  $\bar{x}$ , 545, 548, 555  
 de  $\bar{x} - \bar{y}_j$ , 467, 495-496  
 de  $y^1, 57$   
 infelina, 60  
 residual, 357, 405 (ver también error experimental)  
 respecto a la regresión, 242, 244
- Vector:**  
 uno, 285 ..  
 columna, 278
- V**  $\Sigma$ , 278
- W**  $\oplus$ , 285, 289
- Wald, A**, 34
- Z** (ver Distribución normal)

## ALGUNOS COMENTARIOS A CERCA DEL RESULTADO ANTERIOR

La prueba de F detecta diferencias altamente significativas ( $\alpha=0.01$ ) entre promedios de tratamientos. En consecuencia, rechazamos la hipótesis nula  $H_0: \tau_i = 0$  para  $i=1,\dots,6$ . Otras estadísticas para evaluar el ajuste (o precisión) del modelo son el coeficiente de determinación ( $R^2=SC_{\text{modelo}}/SC_{\text{total}}$ ) y el coeficiente de variación ( $cv = 100 * \sqrt{cme / Y..}$ ).

El valor  $R^2$  (0.7496) indica que, aproximadamente, el 75% de la variación total de la variable NITROGEN es explicada por el efecto de tratamientos.

El coeficiente de variación, es una medida adimensional e indica el grado de precisión del experimento.

### Pruebas de Comparación Múltiples

Cuando se comparan más de dos promedios, la prueba de  $F$  en ANOVA nos dice si las medias son significativamente diferentes, unas de otras, pero no dice entre cuales. Los métodos de comparación múltiples (o de separación de medias) dan información más detallada a cerca de tales diferencias. Existe una variedad de pruebas de comparación múltiple disponibles en los procedimientos ANOVA y GLM del SAS, como opciones de la proposición MEANS).

Decidir sobre qué tipo de prueba utilizar en un caso específico depende de varios factores, entre los cuales se puede mencionar:

- Si los tratamientos tienen estructura (ver arreglos factoriales) o no.
- Si las comparaciones son planeadas(a priori) o no planeadas(a posteriori).
- Tipo de tasa de error.
- Si la prueba de  $F$  resulta significativa o no.
- Si se desea controlar el error Tipo I, o el error Tipo II, o ambos. Una prueba de significancia liberal es la que tiende a una mayor tasa de Error Tipo I; esto es, se declaran diferencias significativas cuando no hay diferencias. Una prueba conservadora tiende a una mayor tasa de error Tipo II, o sea, declara no diferencias cuando pueden existir.
- Si el número de observaciones (repeticiones) por tratamiento es igual o no.
- Si se requiere intervalos de confianza para las comparaciones en prueba.

### Tipos de tasas de error en pruebas de comparación múltiples.

Cuando comparamos dos medias y usamos, por ejemplo una prueba de  $t$ , con un nivel de significancia,  $\alpha= 0.05$ ; pero al comparar 10 promedios con pruebas de  $t$ , estamos efectuando  $10(10-1)/2=45$  pruebas, cada una con una probabilidad de Error tipo I igual a 0.05 (un falso rechazo de la hipótesis nula). El chance de hacer al menos un Error Tipo I es mucho mayor que 0.05. Es difícil calcular la probabilidad exacta pero suponiendo independencia entre las comparaciones (aproximación pesimista), una probabilidad límite superior de hacer al menos un error Tipo I (*tasa de error por experimento*) da igual a  $1-(1-0.05)^{45}=0.90$ .

La probabilidad real es un poco menor que 0.90, pero a medida que se incrementa el número de medias, el chance de hacer al menos un error Tipo I se acerca a 1.

Si usted decide controlar la tasa de error Tipo I para cada comparación, usted está controlando la *tasa de error por comparación*. De otro lado, si usted desea controlar la tasa de error Tipo I global para todas las comparaciones, usted está controlando la *tasa de error por experimento*. Es decisión del experimentador controlar una u otra tasa de error, pero en muchas circunstancias es conveniente mantener baja la tasa de error por experimento. Los métodos Estadísticos para hacer dos o más inferencias mientras se controla la probabilidad de hacer al menos un error Tipo I, se llaman Métodos de Inferencia Simultáneos (ver referencias bibliográficas en el volumen 2 del manual SAS/Stat Users's Guide, versión 6).

Se ha sugerido que la tasa de error por experimento se puede mantener a un nivel  $\alpha$  ejecutando una prueba de F global en el ANOVA a un nivel  $\alpha$  y haciendo comparaciones posteriores solo si la prueba de F es significante, como una prueba LSD protegida de Fisher. Esta aserción es falsa si hay mas de tres medias. Suponga que para 10 medias, una media poblacional difiere de las demás por una cantidad suficientemente grande y que la potencia (probabilidad de rechazar correctamente la hipótesis nula) de la prueba de F es cercana a 1 pero los demás promedios son iguales entre sí. Tendremos  $9(9-1)/2=36$  pruebas de  $t$  con hipótesis nulas ciertas, con un límite superior de 0.84 sobre la probabilidad de al menos un error Tipo I. Así, debemos distinguir entre *tasa de error por experimento bajo la hipótesis nula completa*, en la cual todas las medias de población son iguales y *tasa de error por experimento bajo hipótesis nula parcial*, en la cual algunas medias son iguales pero otras difieren.

SAS usa en la discusión del tema las abreviaturas siguientes:

**CER** Tasa de error por comparación

**EERC** Tasa de error por experimento bajo la hipótesis nula completa

**EERP** Tasa de error por experimento bajo hipótesis nula parcial

**MEER** Tasa de error máxima por experimento bajo cualquiera hipótesis nula, parcial o completa.

### Recomendaciones

Si usted desea controlar la CER se recomiendan los métodos pruebas de T repetidas o LSD no protegida de Fisher (opciones T o LSD). Si se quiere controlar la MEER, cuando no se requieran intervalos de confianza y se tenga igual tamaño de celdas se recomiendan las pruebas REGWF y REGWQ. Si se quiere controlar la MEER, se requieren intervalos de confianza y tenemos desigual tamaño de celdas se recomiendan las pruebas de Tukey o Tukey-Kramer (opción TUKEY). Si usted está de acuerdo con la aproximación Bayesiana (minimizar el riesgo de Bayes bajo pérdidas aditivas en lugar de controlar tasas de error Tipo I), use la prueba de Waller-Duncan (opción Waller).

En el ejercicio anterior las proposiciones:

**MEANS TRAT/LSD DUNCAN TUKEY SCHEFFE SNK WALLER REGWQ SMM SIDAK BON lines;** y,

**MEANS TRAT/DUNNETT('COMPOST') DUNNETTU('COMPOST')**

**DUNNETTL('COMPOST')**, producen las pruebas de comparación múltiples para los  $6(5)/2=15$  pares de medias de tratamientos involucrados en el experimento.