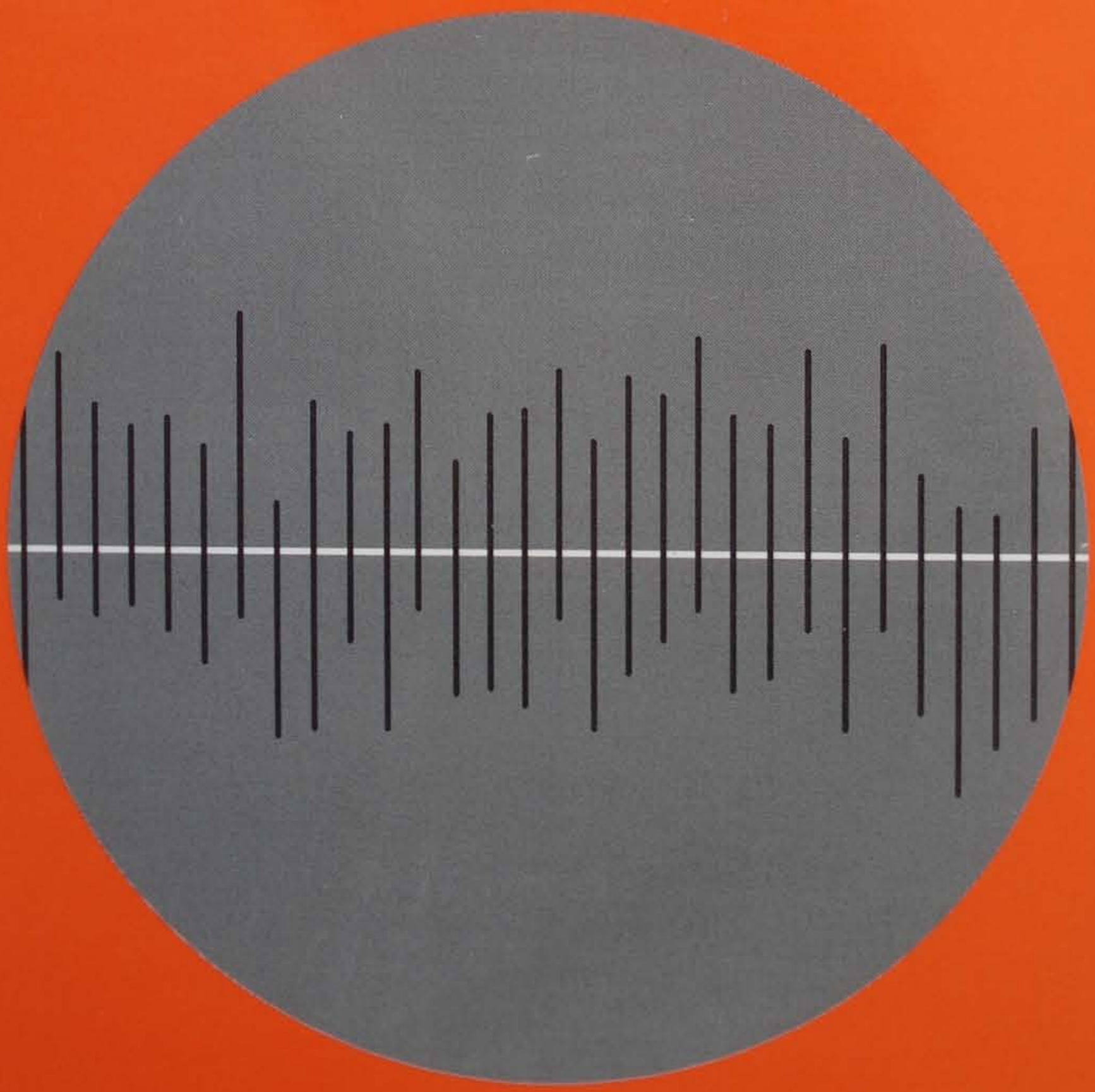
AAU

AMERICAN ANDRAGOGY UNIVERSITY

SERIE DE BIOLOGÍA FUNDAMENTAL

INTRODUCCIÓN A LA BIOESTADÍSTICA

SOKAL/ROHLF



editorial reverté, s.a.

Introducción a la

Bioestadística

Introducción a la

Bioestadística

Robert R. Sokal F. James Rohlf
STATE UNIVERSITY OF NEW YORK AT STONY BROOK

EDITORIAL REVERTÉ, S. A.
Barcelona - Bogotá - Buenos Aires - Caracas - México - Rio de Janeiro

Título de la obra original:
Introduction to Biostatistics

Edición original en lengua inglesa publicada por:
W. H. Freeman and Company, San Francisco

Versión española por:
Joaquina Gabarrón
Licenciada en Biología
Sección de Citogenética del Centro de Bioquímica Clínica de Espinardo (Murcia)

Propiedad de:
EDITORIAL REVERTÉ, S.A.
Loreto 13-15 Local B
08029 Barcelona, España
Tel.: 419-33-36 Fax: 419-51-89
E-Mail (Internet):
104164.23@compuserve.com

y

REVERTÉ EDICIONES, S.A. DE C.V.
Río Pánuco 141 Col. Cuauhtémoc
C.P. 06500 México, D.F.
Tel.: 55-33-56-58 al 60 Fax: 55-14-67-99
E-Mail (Internet):
101545.2361@compuserve.com

Edición en español

ISBN 968-6708-42-1 (México)
ISBN 84-291-1862-4 (España)

Impreso en México

Printed in Mexico

A Julie y Pat

Prólogo

La favorable acogida que nuestra extensa Biometría *ha recibido de profesores y estudiantes, y la sugerencia de numerosos compañeros, nos ha animado a escribir esta más breve* Introducción a la Bioestadística. *Este libro va dirigido a los estudiantes de estadística biológica que deberían poseer una base completa de la materia, requiriendo solamente una preparación elemental en matemáticas. Esperamos que el libro sea útil también en cursos breves de bioestadística como los que a menudo se imparten en facultades de medicina y otras escuelas profesionales. Pese a la necesaria brevedad, hemos conservado el estilo sencillo de nuestro más amplio volumen y confiamos en que las diversas características pedagógicas puestas de manifiesto en aquél, serán también apreciadas en éste.*

Muchos de los enfoques descritos en el prólogo de nuestro volumen anterior se repiten en éste; sin embargo, algunos han sido modificados en función de las diferentes características de los lectores a quienes nos dirigimos. Aunque suministramos esquemas de cálculo detallados para todos los métodos discutidos en el libro, hemos puesto menos énfasis en los aspectos de cálculos implicados en el tratamiento del material. Esto se ha hecho por dos razones: en muchos cursos para estudiantes universitarios, éstos tienen relativamente poca oportunidad y motivación para la realización de cálculos largos con material biológico de investigación; por otra parte, el desarrollo de calculadoras electrónicas ha revolucionado tanto las metodologías para cálculos estadísticos, que un tratamiento amplio de las diversas estrategias para cada tipo de calculadora disponible, caería fuera de los objetivos de este libro y además quedaría anticuado poco después de su publicación. Por tanto, confiaremos en que el profesor del curso aconseje a los estudiantes los mejores procedimientos de cálculo a seguir de acuerdo con los medios disponibles.

La materia está ordenada en capítulos y secciones, numerados según el sistema decimal convencional; el número que precede al punto indica el capítulo y el que sigue a éste, la sección. Los temas pueden encontrarse en el índice general y en el índice alfabético. Las tablas también están numeradas según el sistema decimal, indicando el primer número el

capítulo, y el segundo el número de la tabla dentro del capítulo. Ciertas tablas especiales son denominadas "cuadros" y están numerados como tales, también según el sistema decimal. Estos cuadros cumplen una doble función: ilustran sobre los métodos para resolver diversos tipos de problemas bioestadísticos y, por tanto, pueden ser utilizados por las personas que manejen el libro como modelos apropiados de cálculo. Normalmente contienen todos los pasos necesarios desde el planteamiento del problema hasta el resultado final; por esto, a los estudiantes familiarizados con el libro, pueden servir como resúmenes recordatorios del método. Un segundo uso importante de los cuadros tiene su origen en su utilidad como originales, los cuales se multicopiaban y entregaban a los estudiantes en los cursos de biometría de los autores. En el tiempo disponible de clases es imposible dar ni la mitad de la materia abarcada en el curso, si el contenido de estas hojas tiene que escribirse en la pizarra. De este modo, se puede remitir a los estudiantes a los cuadros y tablas del libro, con el consiguiente ahorro de tiempo y la posibilidad de dedicar su atención a la comprensión del contenido de estas hojas en lugar de copiarlas. Los profesores que utilicen este libro como texto, pueden servirse de los cuadros de forma similar.

Las figuras están numeradas por el mismo sistema que las tablas y cuadros, e igualmente los experimentos (de muestreo), expresiones (matemáticas) y ejercicios (de práctica). Los números de los apéndices van precedidos de la letra A.

Dado que el interés de este libro reside en las aplicaciones prácticas de la estadística a la biología, las discusiones de teoría estadística se atienden estrictamente al mínimo. Las demostraciones de algunas fórmulas aparecen en el apéndice A1 y deben ser estudiadas y reelaboradas por el estudiante.

Las tablas estadísticas necesarias para los métodos tratados en este libro se encuentran en el apéndice A2. Suponemos que los estudiantes son capaces de utilizar tablas matemáticas ordinarias que contengan logaritmos, raíces cuadradas y funciones trigonométricas. Las nuestras han sido extraídas de un volumen de tablas más amplio publicado separadamente (que incluye las tablas matemáticas ya mencionadas), titulado Statistical Tables *de Rholf y Sokal, W.H. Freeman and Company, 1969. Estamos agradecidos al editor de estas últimas Sir Ronald A. Fisher, F.R.S., al Dr. Frank Yates, F.R.S., y a Oliver & Boyd, de Edimburgo, por permitirnos copiar la Tabla III (nuestra Tabla III) de su libro* Statistical Tables for Biological Agricultural and Medical Research.

Al final de cada capítulo se dan ejercicios prácticos para los estudiantes que utilicen el libro en un curso de bioestadística o individualmente. De acuerdo con nuestras propias convicciones, éstos son en gran parte problemas de investigación reales. Algunos de ellos por lo tanto, requieren bastantes cálculos para su solución.

La mayor parte de lo expuesto en este texto, está sacada de nuestro más amplio libro de biometría. Por sugerencia de numerosos compañeros se ha añadido una nueva sección de probabilidad. Nuestro orden de exposición pasa, de forma convencional, de la estadística descriptiva a las distribuciones fundamentales y al contraste de hipótesis estadísticas elementales, para luego proceder inmediatamente al análisis de varianza. El familiar y tradicional test-t se trata simplemente como un caso especial de análisis de varianza y queda relegado a varias secciones de diversos capítulos apropiados del libro. Esto lo hemos hecho deliberadamente por dos razones: (1) Es urgente que los estudiantes se relacionen con el análisis de la varianza lo más pronto posible, ya que actualmente es

esencial para cada biólogo una buena base en el análisis de la varianza. (2) Si el análisis de la varianza se expone y comprende desde el principio, la necesidad de emplear la distribución-t queda muy reducida, excepto para establecer límites de confianza y otras pocas situaciones especiales. Todos los test-t pueden resolverse como análisis de varianza y muchos son más informativos cuando se realizan como tales. La cantidad de calculos es generalmente equivalente.

En otros temas nos hemos preocupado de introducir nuevas y mejoradas técnicas, haciendo más hincapié en éstas que en métodos anteriores que nos parecen menos adecuados. Ejemplos notables de tales innovaciones son: la adopción del procedimiento de tests de comparaciones múltiples a posteriori y el empleo del estadístico-G para el análisis de frecuencias, en lugar de los tradicionales tests chi-cuadrado que, por lo tanto, tratamos menos ampliamente.

Agradecemos a los Profesores K.J. Sonleitner, Theodore J. Crovello, y Albert J. Rowell sus amplias observaciones sobre la versión inicial del manuscrito. Los Profesores Arnold B. Larson y Gunter Schlager nos han proporcionado valiosas observaciones en el borrador de este libro. Estamos agradecidos también a Edwin Bryant, David Fisher, Koichi Fujü, John Kishpaugh y David Wool por la comprobación esmerada de la exactitud numérica de tablas y esquemas. Nuestras esposas, Julie y Pat, nos han ayudado mucho en el trabajo de redacción, y nuestra secretaria, Mrs. Ethel Savarese, ha sido imprescindible para conseguir la puesta en prensa del manuscrito.

Stony Brook, New York

Robert R. Sokal
F. James Rohlf

Indice analítico

Capítulo 1

Introducción

Este capítulo inicia el estudio de bioestadística. En principio se definirá este campo en sí (sección 1.1). Luego se revisará brevemente su desarrollo histórico (sección 1.2). La sección 1.3 concluye el capítulo con una discusión sobre las aportaciones que la persona adiestrada en estadística hace a la investigación biológica.

1.1 Definiciones

Se define la *bioestadística* como *la aplicación de métodos estadísticos a la solución de problemas biológicos*. También se le llama *estadística biológica* o *biometría*.

No se puede comprender bien la definición de bioestadística sin definir previamente la "estadística". Es una ciencia cuyo nombre resulta familiar incluso para el no profesional. El número de definiciones que se pueden encontrar está limitado solamente por el número de libros que se desee consultar. En su moderno sentido se puede definir como *el estudio científico de datos numéricos basados en fenómenos naturales*. Todas las partes de esta definición son importantes y merecen resaltarse.

Estudio científico: se considera de gran interés el comúnmente aceptado criterio de validez de evidencia científica. La objetividad en la presentación y evaluación de datos, y el código ético general de metodología científica deben tenerse en cuenta constantemente para no despertar el viejo bulo de que "los números nunca engañan, sólo los estadísticos lo hacen".

Datos: la estadística trata generalmente de poblaciones o grupos de individuos; por lo tanto, maneja cantidades de información, no un simple dato. Así, la medida de un solo animal o la respuesta de una sola prueba bioquímica generalmente no serán de interés.

Numéricos: si los datos de un estudio no pueden ser cuantificados, no serán tratables por análisis estadísticos. Los datos numéricos pueden ser,"medidas" como la longitud o anchura de una estructura o la cantidad de una sustancia química en un fluido corporal, o

"cuentas" como el número de cerdas o el número de dientes. Las diferentes clases de variables serán discutidas con detalle en el capítulo 2.

Fenómeno natural: se usa este término en sentido amplio, incluyendo todos aquellos eventos que ocurren en la naturaleza animada e inanimada sin el control del hombre y, además, aquellos evocados por el científico y parcialmente bajo su control, como en un experimento. Los diferentes científicos se interesan por diferentes categorías de fenómenos naturales, pero todos estarán de acuerdo en que el chirrido de los grillos, el número de guisantes en una vaina, y la edad a que madura un pollo, son fenómenos naturales. El latido de las ratas como respuesta a la adrenalina o la velocidad de mutación del maíz después de la irradiación pueden, no obstante, ser considerados naturales, aunque el hombre ha interferido el fenómeno mediante su experimentación. Sin embargo, los biólogos en general no considerarían como fenómeno natural el número de equipos estéreo alta fidelidad comprados por personas de diferente condición en un año, aunque los sociólogos y los ecólogos humanos también pueden considerarlo y creen que merece estudiarse. La condición "fenómeno natural" se incluye en la definición de estadística con el fin de remarcar que los fenómenos estudiados no son fenómenos arbitrarios sometidos a la voluntad y control del investigador, como lo está el número de animales empleados en un experimento.

El término "estadístico" se emplea también en otro sentido, aunque relacionado. Se refiere a cualquiera de las muchas cantidades estadísticas calculadas o estimadas, tales como la media, la desviación típica, o el coeficiente de correlación. Cada uno de éstos es un estadístico.

1.2 Desarrollo de la bioestadística

El origen de la estadística moderna se remonta al siglo XVII y deriva de dos fuentes; la primera de ellas se relaciona con las ciencias políticas y se presentó como una descripción cuantitativa de varios aspectos de los asuntos de estado (de ahí el término estadística). Esta materia también se denominó aritmética política. Las tasas y seguros motivaron que la gente llegara a interesarse en problemas de censos, longevidad y mortalidad. Tales estudios alcanzaron una importancia creciente especialmente en Inglaterra, país que prosperó durante el desarrollo de su imperio. John Graunt (1620-1674) y William Petty (1623-1687) fueron los pioneros de la estadística y otros siguieron su línea.

Prácticamente al mismo tiempo se desarrolló la segunda raíz de la estadística moderna: la teoría matemática de la probabilidad, nacida del interés por los juegos de azar entre las clases acomodadas de la época. A esta teoría hicieron aportaciones importantes los franceses Blaise Pascal (1623-1662) y Pierre de Fermat (1601-1665). Un suizo, Jacques Bernoulli (1654-1705), puso los cimientos de la moderna teoría de probabilidad en *Ars Conjectandi*, publicado después de su muerte. Abraham de Moivre (1667-1754), francés residente en Inglaterra, fue el primero en combinar la estadística de su época con la teoría de probabilidad, resolviendo problemas de anualidades. De Moivre fue el primero en aproximar la importante distribución normal por expansión de la binomial.

Un estímulo posterior para el desarrollo de la estadística surgió de la astronomía, en la cual muchas observaciones individuales tenían que hacerse encajar en una teoría coherente. Entre los líderes en este campo se cuentan astrónomos y matemáticos famosos del siglo XVIII, tales como Pierre Simon Laplace (1749-1827) en Francia y Karl Friedrich Gauss (1777-1855) en Alemania. La última aportación a la estadística es el método de mínimos cuadrados, que se tratará en posteriores capítulos de este libro.

Se cree que el primer personaje importante en bioestadística fue Adolphe Quetelet (1796-1874), astrónomo y matemático belga, que en su trabajo combinaba los métodos teóricos y prácticos de estadística y los aplicaba a problemas de biología, medicina, y sociología. A Francis Galton (1822-1911) primo de Charles Darwin, se le denominó padre de la bioestadística y eugenesia, dos materias que estudió interrelacionadamente. Lo imperfecto de las teorías genéticas de Darwin estimuló a Galton para intentar resolver los problemas de herencia. La mayor contribución de Galton a la biología es su aplicación de la metodología estadística al análisis de la variación biológica, así como el análisis de variabilidad y su estudio de regresión y correlación en medidas biológicas. Su esperanza de aclarar las leyes de la genética por medio de estos procedimientos fue en vano. Empezó con el material más difícil y con suposiciones erróneas. Sin embargo, su metodología fue el fundamento para la aplicación de la estadística a la biología. Karl Pearson (1857-1936), en el University College de Londres, se interesó por la aplicación de métodos estadísticos a la biología, particularmente en la demostración de la selección natural, por influencia de W.F.R. Weldon (1860-1906), zoólogo de la misma institución. A Weldon se le ha atribuido incidentalmente la creación del término biometría para el tipo de estudios a que se dedicaba. Pearson continuó en la tradición de Galton y sentó las bases para gran parte de la estadística descriptiva y de correlación. En este siglo la figura dominante en estadística y biometría ha sido Ronald A. Fisher (1890-1962). Sus muchas aportaciones a la teoría estadística serán obvias incluso para el que hojee por encima este libro.

Actualmente la estadística es un campo amplio y extremadamente activo cuyas aplicaciones conciernen a casi todas las ciencias e incluso a los estudios de humanidades. Constantemente se están encontrando otras y nadie puede predecir de qué rama de la estadística surgirán nuevas aplicaciones a la biología.

1.3 Punto de vista estadístico

La creciente importancia y aplicación de la estadística a los datos biológicos es evidente incluso al examinar de pasada cualquier revista de biología. ¿Por qué ha habido un incremento tan marcado en el uso de la estadística en biología? Aparentemente ha surgido por la comprobación de que en biología,la acción recíproca de variables de causa y respuesta obedece a leyes que no están en el modelo clásico de la física del Siglo XIX. En ese siglo, biólogos como Robert Mayer, Helmholtz, y otros, tratando de demostrar que los procesos biológicos no eran sino fenómenos fisicoquímicos, ayudaron a crear la impresión de que los métodos experimentales y la filosofía natural que habían producido un progreso tan espectacular de las ciencias físicas, deberían ser imitados plenamente en biología. Lamentablemente, la oposición a este punto de vista fue confundida con el movimiento vitalista, que condujo a teorías improductivas.

Así pues, muchos biólogos habían mantenido hasta entonces la tradición de conceptos de pensamiento estrictamente mecanicistas y deterministas, mientras los físicos, debido a

que sus ciencias eran más refinadas y trataban con partículas más "elementales", recurrieron a planteamientos estadísticos. En biología la mayoría de los fenómenos se ven afectados por muchos factores causales incontrolables en su variación y, a menudo no identificables. La estadística es necesaria para medir tales fenómenos variables con un error predecible y para descubrir la realidad de mínimas pero importantes diferencias.

Una mala interpretación de estos principios ha llevado a algunos biólogos a pensar que, si las diferencias inducidas por un experimento u observadas en la naturaleza no son tan grandes como para poder ser apreciadas por simple inspección (y por tanto, sin necesidad del análisis estadístico), no vale la pena investigarlas. Hay pocos campos auténticos de investigación en los que la estadística sea innecesaria debido a la naturaleza del fenómeno estudiado.

Debería subrayarse que el pensamiento estadístico no es realmente de diferente tipo que el pensamiento científico disciplinado ordinario, en el cual tratamos de cuantificar nuestras observaciones. En estadística expresamos nuestro grado de confianza o desconfianza como una probabilidad, más que como una vaga afirmación general. Por ejemplo, los biólogos hacen habitualmente afirmaciones de que las especies A son más grandes que las B, o que las hembras se encuentran más frecuentemente en el árbol M que en el N. Tales afirmaciones pueden y deberían expresarse más precisamente en forma cuantitativa. En muchos aspectos, la mente humana es una máquina estadística extraordinaria que absorbe muchos datos del mundo exterior, digiriéndolos y arrojándolos en forma resumida. Sabemos por experiencia que ciertos sucesos ocurren con frecuencia y otros raramente. "Un hombre que fuma" es frecuentemente observado, "un hombre que resbala en una piel de plátano" es raro. Por experiencia sabemos también que los japoneses son más bajos que los ingleses y que los egipcios son más morenos que los suecos. Asociamos trueno con relámpago casi siempre, moscas con basura frecuentemente en verano, pero es extremadamente raro asociar nevadas con el desierto meridional californiano. Tales conocimientos nos llegan como resultado de nuestra experiencia en la vida, tanto directa como indirectamente a través de otros, por comunicación directa o por medio de la lectura. Todos esos datos han sido procesados por el cerebro humano, extraordinario computador que proporciona un resumen. Este resumen se revisa constantemente, y aunque ocasionalmente defectuoso y equivocado, en conjunto es sorprendentemente bueno; es nuestro conocimiento del momento.

Aunque la estadística apareció para satisfacer las necesidades de la investigación científica, la evolución de su metodología afectó a las ciencias a que se aplicó. Así, por un proceso de retroalimentación positiva, la estadística, creada para servir las necesidades de las ciencias naturales, ha afectado a la filosofía de las ciencias biológicas. Para citar un ejemplo: el análisis de la varianza ha tenido gran efecto influenciando en los tipos de experimentos realizados por los investigadores; toda la genética cuantitativa, uno de cuyos problemas es la separación de efectos genéticos y ambientales, cuenta con el análisis de la varianza, y muchos conceptos de genética cuantitativa han sido elaborados directamente en torno al análisis de la varianza.

Capítulo 2
Los datos en biología

En la sección 2.1 expondremos el significado estadístico de los términos "muestra" y "población" que seguiremos utilizando a lo largo de este libro. A continuación entraremos en los tipos de observaciones que obtenemos del material de investigación biológica, con los cuales realizaremos los diversos cálculos en el resto del libro (sección 2.2). El grado de exactitud necesario para la toma de datos y el procedimiento para redondear los números se discutirán en la sección 2.3. Entonces ya estaremos preparados para considerar en la sección 2.4 ciertas clases de datos derivados, frecuentemente utilizados en biología, tales como razones e índices, que representan problemas peculiares con respecto a su exactitud y distribución. Es importante saber ordenar los datos en distribuciones de frecuencias, porque tales ordenaciones nos permiten sacar una impresión global de su aspecto general y presentarlos para procedimientos de cálculo posteriores. Las distribuciones de frecuencias así como la presentación de datos numéricos se discutirán en la próxima sección (2.5) de este capítulo. Por último, en la sección 2.6 describiremos brevemente el tratamiento de los datos para el computador.

2.1 Muestras y poblaciones

Vamos a definir ahora varios términos importantes necesarios para una comprensión de los datos biológicos. En bioestadística, generalmente los datos se basan en *observaciones individuales,* que son *observaciones o medidas tomadas de la mínima unidad de muestreo.* Con frecuencia, pero no necesariamente, estas mínimas unidades de muestreo son también individuos en el sentido biológico ordinario. Si pesamos 100 ratas, el peso de cada rata es una observación individual; los pesos de las 100 ratas juntos representan la *muestra de observaciones,* que se define como, *un conjunto de observaciones individuales seleccionadas por un procedimiento específico.* En este ejemplo, una observación individual está

basada en un individuo en sentido biológico, esto es, una rata; sin embargo, si hubiésemos estudiado el peso de una sola rata a través de un período de tiempo, la muestra de observaciones individuales estaría constituida por los pesos registrados en una sola rata en momentos sucesivos. Si deseamos medir la temperatura en un estudio de colonias de hormigas en el que cada colonia es una unidad básica de muestreo, la temperatura de cada colonia es una observación individual y la muestra de observaciones está formada por las temperaturas de todas las colonias consideradas. Si aceptamos que una estima del contenido en DNA de una célula espermática de mamífero es una observación individual, la muestra de observaciones puede estar constituida por las estimas del contenido en DNA de todas las células espermáticas estudiadas en un mamífero. Un sinónimo de observación individual es *item.*

Hasta el momento hemos evitado cuidadosamente especificar qué variable particular se estaba estudiando, porque los términos "observación individual" y "muestra de observaciones", tal como se han utilizado anteriormente, sólo definen la estructura pero no la naturaleza de los datos en un estudio. *La propiedad real medida por las observaciones individuales es el carácter o variable.* En estadística general el término más empleado es *variable;* sin embargo, en biología se utiliza frecuentemente como sinónimo la palabra *carácter.* En cada mínima unidad de muestreo puede medirse más de un carácter. Así, en un grupo de 25 ratones podemos medir el pH de la sangre y el número de eritrocitos. Cada uno de los 25 ratones (un individuo en sentido biológico) es la mínima unidad de muestreo; el pH de la sangre y el número de células rojas serían los dos caracteres estudiados; las lecturas de pH y los recuentos de células son observaciones individuales, dando lugar a dos muestras de 25 observaciones o una muestra bivariada de 25 observaciones, cada una de las cuales se refiere a una lectura de pH asociada con un recuento de eritrocitos.

A continuación vamos a definir *población.* La definición biológica de este término es bien conocida. Se refiere a todos los individuos de una especie determinada (o tal vez de una etapa del ciclo vital o de un sexo determinado) que se encuentran en un área limitada en un momento dado. En estadística, población se define como *la totalidad de observaciones individuales sobre las cuales se hacen inferencias, las cuales existen en cualquier parte del mundo o al menos dentro de un área de muestreo claramente especificada, limitada en espacio y tiempo.* Si se toman cinco hombres y se estudia el número de leucocitos en su sangre periférica,con la intención de sacar conclusiones sobre todos los hombres a partir de esta muestra de cinco, en este caso la población de la que se ha extraído la muestra representa los recuentos de leucocitos de todos los varones de la especie *Homo sapiens.* En cambio, si se restringe a muestra más estrechamente especificada, como por ejemplo cinco varones chinos de 20 años, limitando las conclusiones a este grupo particular, la población muestreada estará constituida por los números de leucocitos de todos los varones chinos de 20 años. En este sentido estadístico, la población se denomina a veces *universo.* Una población puede referirse a variables de un conjunto concreto de objetos o individuos como,por ejemplo,las longitudes de la cola de todos los ratones blancos del mundo, los recuentos de leucocitos de todos los varones chinos de 20 años, o el contenido en DNA de todas las células espermáticas de hamster; o bien puede referirse a resultados de experimentos, tales como las frecuencias de latidos cardíacos producidas en cobayas por inyecciones de adrenalina. En los primeros casos, la población

es generalmente finita; aunque en la práctica sería imposible obtener, contar y examinar todas las células espermáticas de hamster, todos los varones chinos de 20 años, o todos los ratones blancos del mundo, estas poblaciones son en realidad limitadas. Ciertas poblaciones más pequeñas tales como todas las grullas de una especie determinada de Norteamérica o todos los geómidos de una colonia determinada, pueden someterse perfectamente a un censo total. En cambio, un experimento puede repetirse infinitas veces (al menos en teoría). Así por ejemplo, la administración de adrenalina a cobayas podría repetirse mientras el experimentador pudiese obtener material y su salud y paciencia resistiesen. La muestra de experimentos realmente realizados es una muestra de número infinito de experimentos que *podrían* realizarse. Algunos de los métodos estadísticos que se van a desarrollar posteriormente distinguen entre muestreo de poblaciones finitas e infinitas. Sin embargo, aunque las poblaciones son teóricamente finitas en la mayor parte de las aplicaciones biológicas, generalmente son tan superiores a las muestras extraídas de ellas, que *de hecho* pueden considerarse como poblaciones infinitas.

2.2 Variables en biología

Cada disciplina biológica tiene su propia serie de variables que puede incluir medidas morfológicas convencionales, concentraciones de sustancias en fluidos corporales, velocidades de ciertos procesos biológicos, frecuencias de ciertos sucesos, como en genética y en biología de las radiaciones, lecturas físicas de maquinaria óptica o electrónica utilizada en investigación biológica, y otras muchas.

Ya hemos hecho referencia a variables biológicas de un modo general, pero aún no las hemos definido. Definiremos una *variable como una propiedad con respecto a la cual los individuos de una muestra difieren de algún modo verificable.* Si la propiedad no difiere dentro de una muestra que tenemos a mano, o al menos entre las muestras que se están estudiando, no puede ser de interés estadístico. Longitud, altura, peso, número de dientes, contenido en vitamina C y genotipos, son ejemplos de variables en grupos de organismos ordinarios, genética y fenotípicamente diferentes. No lo es, en cambio, la homeotermia en un grupo de mamíferos, puesto que son todos iguales a este respecto, pero sí sería naturalmente una variable la temperatura corporal de mamíferos.

Podemos dividir las variables biológicas como sigue:

Variables
Variables medibles
Variables continuas
Variables discontinuas
Variables clasificables en rangos
Atributos

Variables medibles son todas aquellas cuyos diferentes valores pueden expresarse de forma numéricamente ordenada. Pueden ser de dos clases. Las primeras, llamadas *varia-*

bles continuas, son las que al menos en teoría pueden tomar un número infinito de valores entre dos puntos determinados. Por ejemplo, entre las dos medidas de longitud 1,5 y 1,6 cm podrían medirse infinitas longitudes,siempre que se estuviese dispuesto a hacerlo y se dispusiese de un método de calibrado suficientemente preciso para obtener tales medidas. Cualquier valor de una variable continua tal como 1,57 mm de longitud es,por lo tanto, una aproximación al valor exacto, que en la práctica es imposible de conocer. Muchas de las variables estudiadas en biología son variables continuas, como por ejemplo las longitudes, áreas, volúmenes, pesos, ángulos, temperaturas, períodos de tiempo, porcentajes y velocidades.

En contraposición a las variables continuas están las *variables discontinuas,* conocidas también como *variables discretas o merísticas.* Estas, sólo tienen valores numéricos fijos, sin posibles valores intermedios. Así, el número de segmentos en un apéndice de un insecto puede ser 4, 5, o 6 pero nunca 5,5 o 4,3. Ejemplos de variables discontinuas son los números de ciertas estructuras (segmentos, cerdas, dientes o glándulas), el número de crías, el número de colonias de microorganismos o animales, o el número de plantas en un cuadrado determinado.

Algunas variables no pueden medirse pero al menos pueden ordenarse o alinearse por su magnitud. Así, en un experimento se puede registrar el orden de eclosión de diez pupas, sin especificar el momento exacto en que hizo eclosión cada una. En tales casos disponemos los datos como una *variable clasificable en rangos,* el orden de eclosión. Se han desarrollado métodos especiales para abordar tales variables, de los cuales se presentan varios en este libro. Al expresar una variable como una serie de rangos, 1, 2, 3, 4, 5, no implicamos que la diferencia en magnitud entre los rangos 1 y 2 sea idéntica, ni siquiera proporcional, a la diferencia entre 2 y 3.

Las variables que no pueden medirse sino que deben expresarse cualitativamente se llaman *atributos.* Todas ellas son propiedades tales como negro o blanco, grávida o ingrávida, muerto o vivo, macho o hembra. Cuando tales atributos se combinan con frecuencias pueden tratarse estadísticamente. De 80 ratones podemos afirmar, por ejemplo, que cuatro fueron negros, dos agutí, y el resto grises. Se llama *enumeración de datos* a la combinación de atributos con frecuencias en tablas apropiadas para el análisis estadístico. Así, la enumeración de datos para el color de los ratones ya mencionado, se haría como sigue:

Color	*Frecuencia*
Negros	4
Agutí	2
Grises	74
Número total de ratones	80

En ciertos casos, los atributos pueden transformarse en variables si se desea. Así, los colores pueden transformarse en longitudes de onda o valores del disco de colores que son variables medibles. Otros atributos susceptibles de ser alineados u ordenados pueden convertirse en variables clasificables. Por ejemplo, tres atributos referentes a una estructura como "pobremente desarrollado", "bien desarrollado", e "hipertrofiado" podrían ser convenientemente cifrados como 1, 2 y 3.

Un término no explicado aún es *variante.* En este libro lo utilizaremos para referirnos a una única lectura, recuento u observación de una variable determinada. Así, si tenemos medidas de la longitud de la cola de cinco ratones, la longitud de la cola será una variable continua y cada una de las cinco medidas de longitud será una variante. En este libro de texto identificamos las variables por letras mayúsculas, siendo el símbolo más común Y, que puede representar longitud de la cola de ratones. Una variante se referirá a una medida de longitud determinada; Y_i es la medida de la cola del ratón i e Y_4 es la del cuarto ratón de nuestra muestra.

2.3 Exactitud y precisión de los datos

"Exactitud" y "precisión" se utilizan como sinónimos en el lenguaje ordinario, pero en estadística los definimos más rigurosamente. *Exactitud* es la proximidad de un valor medido o calculado a su verdadero valor; *precisión* es la proximidad de medidas repetidas de la misma cantidad. Una balanza sesgada pero sensible puede dar pesos inexactos pero precisos. Una balanza poco sensible puede dar por azar una lectura exacta que, no obstante, será imprecisa, puesto que sería imposible que al repetir la medida diese un peso igualmente exacto. A menos que un instrumento de medida esté sesgado, la precisión conducirá a la exactitud. Por lo tanto, es necesario ocuparse principalmente de la primera.

De ordinario pero no necesariamente, las variantes precisas son números enteros. Así, cuando contamos cuatro huevos en un nido, si hemos contado correctamente no hay duda acerca del número exacto de huevos; es cuatro, ni tres ni cinco, y claramente no podría ser cuatro más o menos una parte fraccionaria. Las variables merísticas se miden generalmente como números exactos. Al parecer, las variables continuas que derivan de las merísticas, bajo ciertas condiciones pueden ser también números exactos. Por ejemplo, las razones entre números exactos son también exactas. Si en una colonia de animales hay 18 hembras y 12 machos, la razón de hembras a machos es 1,5, una variante continua y además un número exacto.

Sin embargo, la mayoría de las variables continuas son aproximadas; con esto queremos indicar que el valor exacto de la medida individual (la variante), es desconocido y probablemente imposible de conocer. La última cifra de la medida implicaría precisión, esto es, los límites entre los cuales creemos que se encuentra la verdadera medida. Así, una medida de 12,3 mm implica que la verdadera longitud de la estructura está entre 12,25 y 12,35 mm. Entre esos *límites implícitos,* no sabemos donde está exactamente la longitud real. Podemos preguntarnos donde caería una medida exacta de 12,25, ¿no sería igualmente probable que estuviese en cualquiera de las dos clases 12,2 o 12,3 lo cual no resolvería esta situación? Dicho argumento es correcto, pero cuando anotamos un número como 12,2 o 12,3 damos a entender que ya se ha tomado la decisión de ponerlo en la clase superior o inferior. Esta decisión no se ha tomado arbitrariamente sino que probablemente se ha basado en la mejor medida obtenible.

Si la escala de medida fuese tan precisa que se hubiese podido obtener claramente un valor de 12,25, entonces la medida debería haberse registrado originalmente con cuatro

cifras significativas. Por tanto, los límites implícitos siempre llevan una cifra más después de la última significativa medida por el observador.

De àquí se deduce que si anotamos la medida como 12,32, estamos implicando que el verdadero valor queda entre 12,315 y 12,325. Si no es éste nuestro propósito, no tendría objeto añadir la última cifra decimal a nuestras medidas originales. Si añadimos otra cifra, debemos implicar un aumento de precisión. Por consiguiente, vemos que exactitud y precisión en los números no es un concepto absoluto sino relativo. Suponiendo que no haya sesgo, un número es tanto más exacto a medida que somos capaces de escribirlo con más cifras significativas (aumenta su precisión). Para aclarar este concepto de relatividad de la exactitud, consideremos los tres números siguientes.

	Límites implícitos
193	192,5 -193,5
192,8	192,75 -192,85
192,76	192,755-192,765

Podemos imaginar que estos números son medidas registradas de la misma estructura. Vamos a suponer que tuviésemos el supremo conocimiento de que la verdadera longitud de dicha estructura fuese 192,758 unidades. Si eso fuese cierto, las tres medidas aumentarían en exactitud de arriba abajo. Se observará que los límites implícitos de la medida superior son más amplios que los de la siguiente, que a su vez lo son más que los de la inferior.

Las variantes merísticas, aunque ordinariamente son exactas, pueden registrarse de forma aproximada cuando se trata de números grandes. Así, cuando los recuentos se refieren al millar más próximo, un recuento de 36.000 insectos en un metro cúbico de suelo implica que el verdadero número varía de 35.500 a 36.500 insectos.

¿Con cuántas cifras significativas deberían anotarse las medidas? Si ordenamos la muestra por orden de magnitud del individuo más pequeño al más grande, una regla fácil de recordar es que *el número de clases consecutivas dispuestas en una serie desde la medida menor a la mayor, debería estar entre 30 y 300.* Así pues, si estamos midiendo una serie de conchas lo más aproximado al milímetro y la más grande es de 8 mm y la más pequeña de 4 mm de anchura, solamente hay cuatro clases consecutivas entre la medida mayor y la menor. Por tanto, deberíamos haber medido nuestras conchas con una cifra decimal más. En tal caso las dos medidas extremas podrían haber sido 8,2 mm y 4,1 mm, con 41 clases consecutivas entre ellas (contando la última cifra significativa como la unidad); éste habría sido un número adecuado de clases consecutivas. La razón de tal regla es que, un error de 1 en la última cifra significativa de una medida de 4 mm, constituiría un error inadmisible del 25 %, pero un error de 1 en la última cifra de 4,1 es menor que 2,5 %. Del mismo modo, si hubiésemos medido la altura de la más alta de una serie de plantas como 173,2 cm y la de la más baja como 26,6 cm, la diferencia entre estos límites comprendería 1466 clases consecutivas (de 0,1 cm), que son demasiadas. Por lo tanto, habría sido conveniente registrar la altura lo más aproximado al centímetro, es decir: 173 cm para la

más alta y 27 cm para la más baja, lo que daría 146 clases consecutivas. Utilizando la regla ya mencionada, para la mayoría de las medidas anotaremos dos o tres cifras.

La última cifra de un número aproximado debería ser siempre significativa; esto es, debería implicar para la verdadera medida un rango desde media clase consecutiva por debajo hasta media clase consecutiva por encima de la medida registrada, como se ha expuesto anteriormente. Esto se aplica a todas las cifras incluido el cero. Por lo tanto, los ceros no deberían escribirse al final de los números aproximados a la derecha de la coma decimal, excepto si se tiene intención de que sean significativos. Así, 7,80 debe implicar los límites de 7,795 a 7,805. Si está implícito de 7,75 a 7,85, la medida debería anotarse como 7,8.

Cuando se quiere reducir el número de cifras significativas, realizamos el proceso de *redondear* números. Las reglas para redondear son muy sencillas. Un número dígito que se quiere redondear no se cambia si va seguido por uno menor que 5. Si va seguido por uno mayor que 5 o por 5 seguido de otros números dígitos distintos de cero, se aumenta en uno. Cuando el número que se va a redondear va seguido por un 5 solo o seguido de ceros, se aumenta en uno si es impar pero no se cambia si es par. La razón de esta última regla es que cuando tales números se suman en una larga serie, por término medio habríamos elevado tantos números como habríamos rebajado; por lo tanto, estos cambios se equilibrarían. Poner en práctica las reglas anteriores para redondear los números siguientes con el número de cifras significativas que se indica.

Número	*Cifras significativas deseadas*	*Solución*
26,58	2	27
133,7137	5	133,71
0,03725	3	0,0372
0,03715	3	0,0372
18 316	2	18 000
17,3476	3	17,3

2.4 Variables derivadas

En trabajos biométricos, la mayoría de las variables son observaciones registradas como medidas directas o recuentos de material biológico, o como lecturas que son producto de diversos tipos de instrumentos. Sin embargo, en la investigación biológica hay una clase importante de variables que podemos llamar *variables derivadas* o *calculadas*, que generalmente están basadas en dos o más variables medidas independientemente cuyas relaciones se expresan de una determinada forma. Nos estamos refiriendo a razones, porcentajes, índices, velocidades y otras.

Una *razón* expresa como un solo valor la relación entre dos variables. En su forma más simple se expresa como, por ejemplo, 64:24, que puede representar el número de individuos tipo salvaje frente a mutantes o el número de machos frente a hembras, o la proporción de individuos parasitados frente a los no parasitados y así sucesivamente. Los

ejemplos anteriores implican razones basadas en recuentos; una razón basada en una variable continua podría igualmente expresarse como 1,2:1,8, que puede representar la relación de anchura a longitud en un esclerito de un insecto, o la relación entre las concentraciones de dos minerales contenidos en el agua o en el suelo. Las razones pueden expresarse también como fracciones; así las dos razones anteriores podrían expresarse como $\frac{6}{24}$ y $\frac{1,2}{1,8}$. Sin embargo, con fines de cálculo es más útil expresar la razón como un cociente. Por tanto las dos razones citadas anteriormente serían 2,666... y 0,666... respectivamente. Estos números son abstractos y no se expresan en unidades de medida de tipo alguno. Este es el modelo para las razones que consideraremos más adelante. Los *porcentajes* también son un tipo de razón. Razones y porcentajes son cantidades básicas en gran parte de la investigación biológica, ampliamente utilizadas y generalmente familiares.

Un *índice* es la razón de una variable anatómica dividida por otra mayor llamada estándar. Un ejemplo bien conocido de índice en este sentido es el índice cefálico en antropología médica. Expresado en sentido amplio, un índice podría ser el promedio de dos medidas, o en forma simple, tal como $\frac{1}{2}$ (longitud de A + longitud de B), o en forma ponderada, como $\frac{1}{3}$ [(2 × longitud de A) + longitud de B].

Las *velocidades* o *tasas* serán importantes en muchos campos experimentales de la biología. En esta categoría se incluirían la cantidad de una sustancia liberada por unidad de tiempo, las tasas reproductivas por unidad de población, tamaño y tiempo (tasas de nacimiento), y las tasas de defunción.

El uso de razones y porcentajes está profundamente arraigado en el pensamiento científico. Con frecuencia, las razones pueden ser la única vía significativa para interpretar y comprender ciertos tipos de problemas biológicos. Si el proceso biológico que se está investigando opera en razón de las variables estudiadas, debe examinarse esta razón para comprender el proceso. Así, Sinnott y Hammond (1935) encontraron que la herencia de la forma de las calabazas *Cucurbita pepo* podría ser interpretada por un índice de forma basado en una proporción longitud-anchura, pero no en términos de las dimensiones independientes de la forma. Igualmente, debe establecerse que la selección que afecta a las proporciones del cuerpo existe en la evolución de casi todos los organismos cuando se investiga adecuadamente.

Los inconvenientes del uso de razones son varios: en primer lugar, su relativa inexactitud. Vamos a volver a la razón $\frac{1,2}{1,8}$ mencionada más arriba, y recordemos de la sección anterior que una medida de 1,2 indica un verdadero rango de medida de la variable de 1,15 a 1,25; igualmente, una medida de 1,8 implica un rango de 1,75 a 1,85. Por lo tanto, vemos que la verdadera razón puede variar desde $\frac{1,15}{1,85}$ hasta $\frac{1,25}{1,75}$, ó 0,622 y 0,714 respectivamente. Observamos un error máximo posible de 4,2 % si 1,2 fuese una medida original: [(1,25 − 1,2)/1,2]; el error máximo correspondiente para la razón es 7,0 %: [(0,714 − 0,667)/0,667]. Además, el mejor estimador de una razón no es por lo general el punto medio entre sus posibles rangos. Así, en nuestro ejemplo el punto medio entre los límites implícitos es 0,668 y la verdadera razón es 0,666..., sólo una ligera diferencia que puede, no obstante, ser mayor en otros ejemplos.

Un segundo inconveniente de las razones y porcentajes es que sus distribuciones pueden ser un tanto raras y por tanto no estar distribuidas de forma más o menos normal (ver capítulo 5), como se requiere para muchos tests estadísticos. Con frecuencia esta dificultad puede superarse por transformación de la variable (como se discute en el capítulo 10).

Otro inconveniente de las razones es que no se suministra información de la relación entre las dos variables cuya razón se está considerando. A veces puede aprenderse más estudiando las variables de una en una y sus relaciones mutuas.

2.5 Distribuciones de frecuencias

Si fuésemos a extraer muestras, por ejemplo, de una población de pesos de niños recién nacidos, podríamos representar cada medida extraída por un punto a lo largo de un eje que indica magnitud de pesos de recién nacidos. Esto se representa en la Figura 2.1A, para una muestra de 25 pesos. Si extraemos muestras de la población repetidamente y obtenemos 100 pesos de recién nacidos, probablemente tendremos que colocar algunos de estos puntos sobre otros para registrarlos todos correctamente (figura 2.1B). Cuando continuamos muestreando cientos y miles de pesos de recién nacidos adicionales (figura 2.1C y D), el conjunto de puntos continuará aumentando en tamaño pero adoptará una forma claramente definida. La curva que traza el contorno de la nube de puntos aproxima la distribución de la variable. Recuérdese que una variable continua tal como el peso de recién nacidos puede adoptar una infinidad de valores entre dos puntos cualesquiera sobre la abscisa. La finura de nuestras medidas determinará cuán fino será el número de divisiones registradas entre dos puntos cualesquiera a lo largo del eje.

La distribución de una variable es de considerable interés biológico. Si encontramos que la distribución es asimétrica y máxima en una zona determinada, ello nos dice que tal vez haya selección a favor o contra los organismos que caen en uno de los extremos de la distribución, o que posiblemente la escala de medida elegida sea la que provoca una distorsión de la distribución. Si en una muestra de insectos inmaduros descubrimos que las medidas están bimodalmente distribuidas (con dos picos), esto indicaría que la población es dimórfica; pueden haberse entremezclado en la muestra especies o razas diferentes, o podría haber surgido el dimorfismo de la presencia de ambos sexos o de diferentes estadíos. Hay varios modelos característicos de distribuciones de frecuencias, la más común de las cuales es la distribución simétrica acampanada (aproximada por la última gráfica de la figura 2.1) o distribución normal, discutida en el capítulo 5. Hay también distribuciones sesgadas (más alargadas en un extremo que en otro), distribuciones en forma de ∪ como en la figura 2.2, distribuciones en forma de ∪, y otras, que prestan información significativa sobre ciertos tipos de parentescos. En los últimos capítulos y secciones tendremos más que decir sobre las implicaciones de diversos tipos de distribuciones.

Después de haberse obtenido los datos en un estudio determinado, deben ordenarse de forma adecuada para el cálculo e interpretación. Podemos suponer que inicialmente las variantes están dispuestas al azar o en el orden en que se han tomado las medidas. Una disposición sencilla sería una *ordenación* de los datos por orden de magnitud. Así, por ejemplo, las variantes 7, 6, 5, 7, 8, 9, 6, 7, 4, 6, 7, podrían disponerse por orden de magnitud decreciente como sigue: 9, 8, 7, 7, 7, 7, 6, 6, 6, 5, 4. Donde haya ciertas variantes del mismo valor, tales como el 6 y el 7 en este ejemplo ficticio, puede ocurrírsenos inmediatamente un artificio ahorrador de tiempo, a saber, anotar una frecuencia

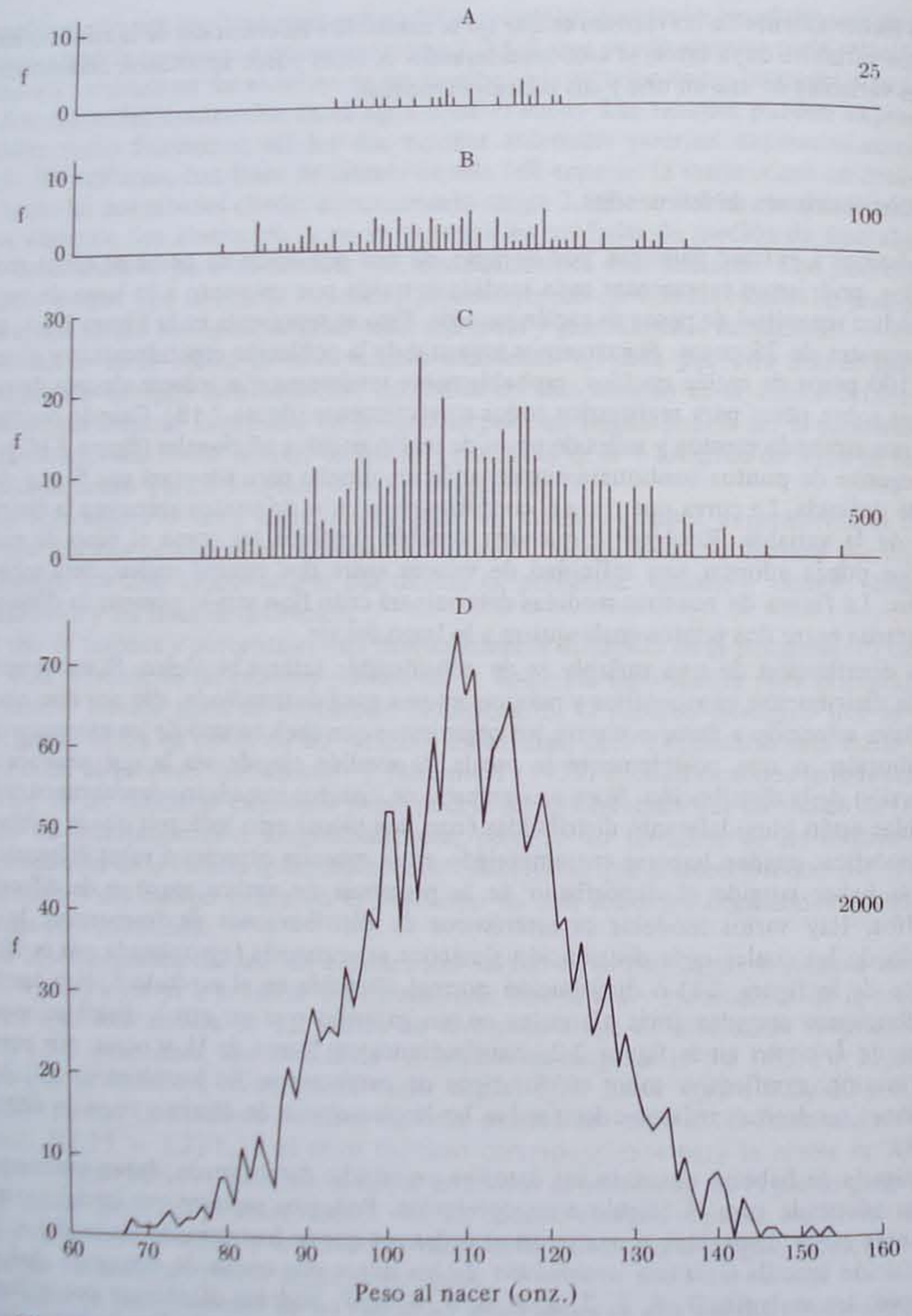


Fig. 2.1. Muestreo de una población de pesos de niños recién nacidos (una variable continua). A. Una muestra de 25. B. Una muestra de 100. C. Una muestra de 500. D. Una muestra de 2.000.

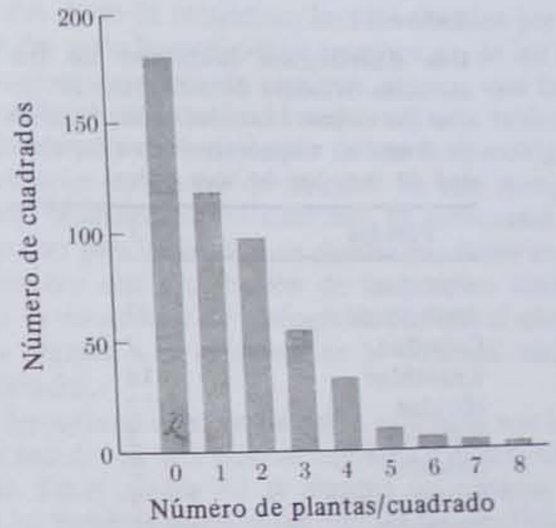


Fig. 2.2. Diagrama de barras. Frecuencias de la especie *Carex flacca* en 500 cuadrados. Datos de la tabla 2.2; original de Archibald (1950).

para cada variante recurrente, así 9, 8, 7(× 4), 6(× 3), 5, 4. Esta notación taquigráfica es una forma de representar una *distribución de frecuencias*, que es simplemente una ordenación de las clases de variantes con las frecuencias de cada clase indicada. Convencionalmente, una distribución de frecuencias se presenta en forma tabular, como sigue para el ejemplo anterior.

Variable	*Frecuencia*
Y	f
9	1
8	1
7	4
6	3
5	1
4	1

El anterior es un ejemplo de una *distribución de frecuencias cuantitativas*, ya que *Y* es claramente una variable medible. Sin embargo, las series y distribuciones de frecuencias no tienen que limitarse a tales variables. Podemos hacer distribuciones de frecuencias de atributos, llamadas *distribuciones de frecuencias cualitativas*. En éstas, las diversas clases se ponen en lista según cierto orden lógico o arbitrario. Por ejemplo, en genética podemos obtener una distribución de frecuencias cualitativas como sigue:

	f
A –	86
aa	32

TABLA 2.1

Una distribución cualitativa de frecuencias. Número de individuos del orden Heteróptera tabulados por familias en muestras completas de una comunidad de insectos de hoja estival.

Familia	*f*
Alydidos	2
Anthocoridae	37
Coreidae	2
Lygaeidae	318
Miridae	373
Nabidae	3
Neididae	5
Pentatomidae	25
Piesmidae	1
Reduviidae	3
Rhopalidae	2
Saldidae	1
Thyreocoridae	10
Tingidae	69
Heterópteros totales	851

Fuente: Datos de Whittaker (1952).

TABLA 2.2

Una distribución de frecuencias merística. Número de plantas de *Carex flacca* halladas en 500 cuadrados.

N.º de plantas por cuadrado *Y*	*Frecuencia observada* *f*
0	181
1	118
2	97
3	54
4	32
5	9
6	5
7	3
8	1
Total	500

Fuente: Datos de Archibald (1950).

Esto nos revela que hay dos clases de individuos, los identificados por el genotipo *A*, de los cuales se encontraron 86, y los homozigóticos recesivos *aa*, de los que se observaron 32 en la muestra. En ecología es bastante corriente obtener una lista de especies de habitantes en un área ecológica muestreada. El orden de tales tablas es usualmente el alfabético. En la tabla 2.1 se presenta un ejemplo de una lista ecológica que constituye una distribución de frecuencias cualitativa de familias. En este ejemplo, las familias de insectos mencionadas están ordenadas alfabéticamente; en otros casos la secuencia puede ser convencional, como ocurre para las familias de plantas con flores en botánica.

En la tabla 2.2 se muestra una distribución de frecuencias cuantitativa basada en variantes merísticas. Este es un ejemplo de ecología de plantas: el número de plantas por cuadrado muestreado se registra a la izquierda en la columna variable; la frecuencia observada se indica a la derecha.

Las distribuciones de frecuencias más comúnmente utilizadas son las distribuciones de frecuencias cuantitativas basadas en una variable continua y es conveniente familiarizarse completamente con ellas. En el cuadro 2.1 se muestra un ejemplo. Está basado en 25 longitudes del fémur de las hembras apomícticas (uno de los estadios del ciclo biológico) de una especie de áfidos. Las 25 lecturas (medidas en unidades de micrómetro codificadas) se presentan en la parte superior del cuadro 2.1 en el orden en que se obtuvieron las medidas. Podrían haberse ordenado según su magnitud. A continuación se disponen los datos en una distribución de frecuencias. Las variantes aumentan en magnitud por clases consecutivas de 0,1. La distribución de frecuencias se prepara introduciendo en la escala cada variante e indicando una cuenta por una señal convencional. Cuando se han encuadrado todos los elementos en las clases correspondientes, las marcas se convierten en números que indican las frecuencias en la próxima columna. Su suma se indica por Σf.

¿Qué hemos conseguido al resumir nuestros datos? Las 25 variantes originales están representadas ahora por 15 clases solamente. Observamos que las variantes 3,6, 3,8 y 4,3 tienen las frecuencias más altas. Sin embargo, observamos también que hay varias clases, tales como 3,4 o 3,7, que no están representadas por ningún áfido. Esto atribuye a la distribución de frecuencias completa una apariencia bastante estirada y dispersa. La razón de esto es que sólo tenemos 25 áfidos, demasiado pocos para introducirlos en una distribución de frecuencias con 15 clases. Para obtener una distribución más coherente y de aspecto uniforme debemos condensar nuestros datos en menos clases. Este proceso se conoce como *agrupamiento de clases* de las distribuciones de frecuencias; se representa en el cuadro 2.1 y se describe en los párrafos siguientes.

Deberíamos darnos cuenta de que lo que hacemos cuando agrupamos variantes individuales en clases de rango más amplio es sólo una prolongación del mismo proceso que se efectuó cuando obtuvimos la medida inicial. Así, como hemos visto en la sección 2.3, cuando medimos un áfido y anotamos la longitud de su fémur como 3,3 unidades, queremos decir con esto que la verdadera medida está entre 3,25 y 3,35 unidades, pero que fuímos incapaces de medir hasta la segunda cifra decimal. Al registrar inicialmente la medida como 3,3 unidades, estimamos que cae dentro de este rango. Si hubiésemos estimado que excedía el valor de 3,35, por ejemplo, le habríamos aplicado la marca superior próxima, 3,4. Por lo tanto, todas las medidas entre 3,25 y 3,35 fueron agrupadas en realidad en la clase identificada por la *marca de clase* 3,3. Nuestro *intervalo de clase* era

CUADRO 2.1

Preparación de una distribución de frecuencias y agrupamiento en menor número de clases con intervalos de clase más amplios.

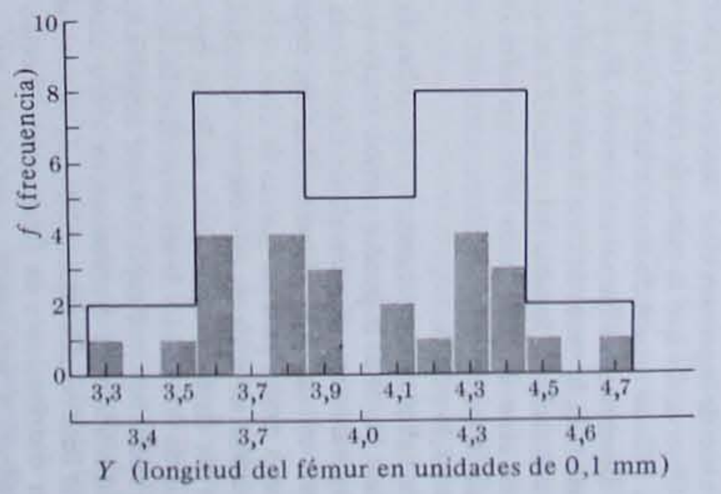
Veinticinco longitudes del fémur de hembras apomícticas del áfido *Pemphigus populi-transversus*. Las medidas están en mm $\times 10^{-1}$.

Medidas originales				
3,8	3,6	4,3	3,5	4,3
3,3	4,3	3,9	4,3	3,8
3,9	4,4	3,8	4,7	3,6
4,1	4,4	4,5	3,6	3,8
4,4	4,1	3,6	4,2	3,9

Distribución de frecuencias original				*Agrupamiento en 8 clases de intervalo 0,2*				*Agrupamiento en 5 clases de intervalo 0,3*			
Límites implícitos	*Y*	*Señales de anotación*	*f*	*Límites implícitos*	*Marca de clase*	*Señales de anotación*	*f*	*Límites implícitos*	*Marca de clase*	*Señales de anotación*	*f*
3,25-3,35	3,3	|	1	3,25-3,45	3,35	|	1	3,25-3,55	3,4	||	2
3,35-3,45	3,4		0								
3,45-3,55	3,5	|	1	3,45-3,65	3,55	𝍸	5				
3,55-3,65	3,6	||||	4					3,55-3,85	3,7	𝍸 |||	8
3,65-3,75	3,7		0	3,65-3,85	3,75	||||	4				
3,75-3,85	3,8	||||	4								
3,85-3,95	3,9	|||	3	3,85-4,05	3,95	|||	3	3,85-4,15	4,0	𝍸	5
3,95-4,05	4,0		0								
4,05-4,15	4,1	||	2	4,05-4,25	4,15	|||	3				
4,15-4,25	4,2	\	1					4,15-4,45	4,3	𝍸 |||	8
4,25-4,35	4,3	||||	4	4,25-4,45	4,35	𝍸 ||	7				
4,35-4,45	4,4	|||	3								
4,45-4,55	4,5	|	1	4,45-4,65	4,55	|	1	4,45-4,75	4,6	||	2
4,55-4,65	4,6		0								
4,65-4,75	4,7	|	1	4,65-4,85	4,75	|	1				
Σf			25				25				25

Fuente: Datos de R.R. Sokal

Histograma de la distribución de frecuencias original presentada anteriormente y de la distribución agrupada en 5 clases. El eje horizontal inferior muestra las marcas de clase para la distribución de frecuencias agrupada. Las barras sombreadas representan la distribución de frecuencias original; las barras en blanco representan la distribución agrupada.



de 0,1 unidades. Si ahora deseamos hacer intervalos de clase más amplios, no hacemos sino extender el rango dentro del cual se sitúan las medidas en una clase.

La referencia al cuadro 2.1 aclarará este proceso. A fin de inculcar al lector la flexibilidad del proceso, agrupamos los datos dos veces. En el primer ejemplo de agrupamiento, el intervalo de clase se ha duplicado en amplitud; esto es, se ha hecho de 0,2 unidades. Si empezamos por el final, los límites de clase implícitos serán ahora de 3,25 a 3,45, los de la próxima de 3,45 a 3,65 y así sucesivamente.

Nuestra próxima tarea es encontrar estas marcas de clase. Esto ha resultado bastante sencillo en la distribución de frecuencias mostrada en la parte izquierda del cuadro 2.1, en que las medidas originales se han utilizado como marcas de clase. Sin embargo, ahora utilizamos un intervalo de clase de doble amplitud que el anterior y las marcas de clase se calculan obteniendo el punto medio de los nuevos intervalos de clase. Así, para hallar la marca de clase de la primera clase tomamos el punto medio entre 3,25 y 3,45, que resulta ser 3,35. Observamos que la marca de clase tiene una cifra decimal más que las medidas originales. Ello no debe inducirnos a creer que repentinamente hemos conseguido mayor precisión. Siempre que elijamos un intervalo de clase cuya última cifra *significativa* sea par (en este caso 0,2), la marca de clase llevará una cifra decimal más que las medidas originales. En la parte derecha de la tabla del cuadro 2.1 se agrupan otra vez los datos, utilizando un intervalo de clase de 0,3. Como la última cifra significativa es impar, la marca de clase presenta ahora tantas cifras decimales como las variantes originales, siendo 3,4, el punto medio entre 3,25 y 3,55.

Una vez hallados correctamente los límites implícitos y la marca de clase para la primera clase, los otros se pueden escribir debajo sin ningún cálculo especial. Simplemente se suma repetidamente el intervalo de clase a cada uno de los valores. Así, comenzando con el límite inferior 3,25, al sumarle 0,2 obtenemos 3,45, 3,65, 3,85, y así sucesivamente; del mismo modo, para las marcas de clase obtenemos 3,35, 3,55, 3,75 y así sucesivamente. Debería ser evidente que cuanto más amplios sean los intervalos de clase más se condensan los datos, pero también se hacen menos precisos. Sin embargo, al mirar la distribución de frecuencias de las longitudes del fémur en el cuadro 2.1, advertimos que la estructura inicial bastante caótica se está simplificando por agrupamiento. Cuando agrupamos la distribución de frecuencias en cinco clases con un intervalo de clase de 0,3 unidades, se hace notablemente bimodal (esto es, posee dos picos de frecuencias).

Al hacer distribuciones de frecuencias deberían establecerse de 12 a 20 clases. No es necesario adherirse servilmente a esta regla, pero debería utilizarse con algo del sentido común que procede de la experiencia en el tratamiento estadístico de los datos. El número de clases depende en gran parte del tamaño de la muestra estudiada. Las muestras de menos de 40 ó 50 raramente deberían darse con tantas como 12 clases, ya que esto proporcionaría un número demasiado pequeño de frecuencias por clase. En cambio, las muestras de varios millares pueden agruparse provechosamente en más de 20 clases. Si fuese necesario agrupar los datos de áfidos del cuadro 2.1, probablemente no se agruparían en más de 6 clases.

Si los datos originales nos dan menos clases de las que pensamos que deberíamos tener, no puede hacerse nada si la variable es merística, ya que ésta es la naturaleza de los datos en cuestión. Sin embargo, con una variable continua, una escasez de clases indicaría que probablemente no hemos hecho nuestras medidas con suficiente precisión. Si hemos

seguido las reglas del número de cifras significativas establecidas en la sección 2.3, esto podría no haber ocurrido. Lamentablemente, con frecuencia ya se han obtenido las medidas antes de estudiar el informe estadístico. En tal caso no puede hacerse nada si hay pocas clases.

Cuando hay más clases de las deseadas, debería intentarse el agrupamiento. Cuando los datos son merísticos, los límites implícitos de variables continuas carecen de sentido. Sin embargo, en muchas variantes merísticas, tales como un número de cerdas que varía desde uno inferior de 13 a uno superior de 81, probablemente sería deseable agruparlos en clases, conteniendo varios recuentos cada una. Esto puede hacerse mejor utilizando como intervalo de clase un número impar, de modo que la marca de clase que representa los datos sea un número entero en vez de un fraccionario. Así, si agrupásemos en una clase los números de cerdas 13, 14, 15 y 16, la marca de clase tendría que ser 14,5, un valor sin sentido en términos de número de cerdas. Por lo tanto sería mejor agruparlos de 3 en 3 o de 5 en 5 lo cual daría como marca de clase los valores enteros 14 o 15.

¿Cuándo deberían agruparse los datos en distribuciones de frecuencias? Si se dispone de una calculadora y hay más de 100 ó 150 observaciones, probablemente valdrá la pena poner los datos en una distribución de frecuencias antes de realizar los cálculos estadísticos. Esta decisión está basada en el hecho de que el tiempo que se ahorra en el cálculo cuando se utiliza una distribución de frecuencias debe compararse con el tiempo que se gasta en establecer la distribución. Para frecuencias más bajas (esto es, < 100) llevaría más tiempo establecer la distribución que realizar los cálculos a partir de las observaciones sin tratar. Si es fácil disponer de un computador, la regla anterior también es naturalmente inválida. Puesto que un computador puede tratar cientos e incluso miles de observaciones en muy poco tiempo, generalmente no será necesario agruparlas a menos que estemos tratando con muestras extremadamente grandes, mucho más de 1.000 observaciones por muestra.

Si la forma de la distribución es un punto específico de interés, surge otra razón importante para obtener distribuciones de frecuencia. En una muestra de insectos inmaduros, el hallazgo de que ciertas medidas están bimodalmente distribuidas indicaría que la población es dimórfica. No obstante, en ciertos casos en los que se han tomado muestras repetidas de un proceso o fenómeno biológico particular y donde la forma de la distribución es bien conocida, no sería deseable hacer cada vez una distribución de frecuencias a menos que fuese necesario por razones de cálculo.

Si la forma de una distribución de frecuencias es de particular interés, muchas veces puede ser deseable presentar la distribución en forma gráfica cuando se discuten los resultados. Generalmente esto se hace por medio de diagramas de frecuencia, de los que hay dos tipos comunes. Para una distribución de datos merísticos utilizamos un *diagrama de barras*, como se muestra en la figura 2.2 para los datos de la tabla 2.2. La abscisa representa la variable (en nuestro caso el número de plantas por cuadrado), y la ordenada las frecuencias. La característica importante de este diagrama es que las barras no se tocan, lo que indica que la variable no es continua. En cambio, las variables continuas, tales como la distribución de frecuencias de las hembras apomícticas de áfidos, se representan gráficamente por un *histograma*, en el cual la amplitud de cada barra a lo largo de la abscisa representa un intervalo de clase de la distribución de frecuencias y las barras se tocan para mostrar que los límites reales de las clases son contiguos. El punto medio de la

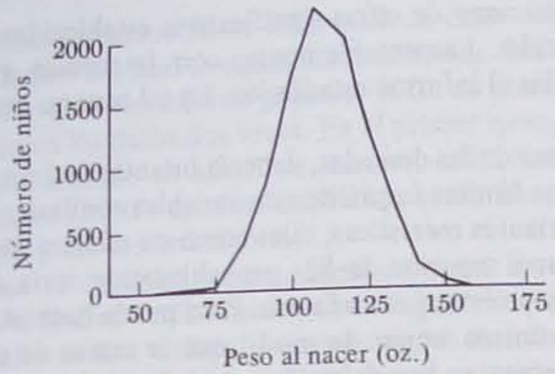


Fig. 2.3. Polígono de frecuencias. Pesos al nacer de: 9465 niños varones. Pacientes chinos, tercera clase, en Singapore, 1950 y 1951. Datos de Millis y Seng (1954).

barra corresponde a la marca de clase. En la parte inferior del cuadro 2.1 se presentan histogramas de la distribución de frecuencias de los datos de áfidos, no agrupados y agrupados. La altura de las barras representa la frecuencia de cada clase. Para demostrar que los histogramas son aproximaciones apropiadas para las distribuciones continuas que se encuentran en la naturaleza, podemos tomar un histograma y hacer más estrechos los intervalos de clase, produciendo más clases. Entonces el histograma claramente se ceñiría más a una distribución continua. Podemos continuar este proceso hasta que los intervalos de clase se aproximen al límite de amplitud infinitesimal. En este momento el histograma se convierte en la distribución continua de la variable. En ocasiones los intervalos de clase de una distribución de frecuencias continua agrupada son desiguales. Por ejemplo, en una distribución de frecuencias de edades podemos obtener más detalles en los diferentes estadios de individuos jóvenes e identificación menos exacta de las edades de individuos viejos. En estos casos, los intervalos de clase para los grupos más viejos serían más anchos y los de los grupos más jóvenes más estrechos. Al representar estos datos, las barras del histograma se dibujan con diferentes anchuras. La figura 2.3 muestra otra forma gráfica de representación de una distribución de frecuencias de una variable continua, el peso de niños recién nacidos. Como veremos más adelante, la forma de las distribuciones observadas en estos diagramas de frecuencias puede revelar mucho acerca de las situaciones biológicas que afectan a una variable determinada.

2.6 El tratamiento de los datos

El rápido y hábil tratamiento de los datos es esencial para la práctica con éxito de la estadística. Los lectores deben relacionarse con las diversas técnicas disponibles para efectuar los cálculos estadísticos. Aunque una calculadora de mesa es buena en su línea para considerarla una herramienta de investigación indispensable como un microscopio, puede uno encontrarse por una u otra circunstancia sin esta ayuda de cálculo. En tales casos, los cálculos estadísticos deben realizarse por los llamados métodos de papel y lápiz.

Como se verá en el próximo capítulo, deben transformarse los datos con el fin de procesarlos rápidamente, y el tipo de clave puede modificarse para que se ajuste a los métodos de papel y lápiz.

Pueden mencionarse aquí otros dos procesos para la simplificación de los cálculos. El primero es el desarrollo de tablas que proporcionan respuestas para ciertos problemas estadísticos típicos sin cálculo de ningún tipo. El segundo es el desarrollo de ciertas técnicas estadísticas más modernas, que por ser tan fáciles de realizar no necesitan ayudas mecánicas. Algunas de estas técnicas son inherentemente sencillas (por ejemplo, ciertos métodos no paramétricos tales como la prueba de los signos, sección 10.3); otras son aproximadas y, en consecuencia, son métodos menos eficientes que pueden servir como un primer resumen de los resultados deseados. Un ejemplo de un método que no es completamente eficiente es el empleo de la mediana (sección 3.3) en lugar de la media (sección 3.1). La exactitud en el cálculo dependerá del número de cifras manejadas durante el mismo. Los diferentes dispositivos de cálculo varían en su capacidad para retener el número de dígitos; por lo tanto, al resolver el mismo problema en una calculadora de mesa y en un computador, por ejemplo, los resultados obtenidos pueden ser diferentes. Incluso diferentes computadores no siempre darán exactamente las mismas respuestas.

En la mayor parte de los paises ya no se fabrican las calculadoras mecánicas. No obstante, aún se utilizan corrientemente y requieren una breve mención. Las *máquinas de sumar* resultan familiares a la mayoría de las personas por su amplia utilización en almacenes y oficinas. Están constituidas por un teclado que introduce los números en las máquinas y ciertas teclas operacionales para adición, entradas correctoras y otras diversas funciones. Los resultados se imprimen en un rollo de cinta de papel sujeto a la máquina. Una versión modificada de la máquina de sumar es la llamada *calculadora impresora.* Esta máquina es un descendiente lineal de la máquina de sumar eléctrica a la que se ha añadido la multiplicación y división automática de los números. Los resultados de estas operaciones y algunos pasos intermedios se imprimen en una cinta de papel.

La *calculadora de mesa* convencional (llamada también calculadora de mesa rotativa) no imprime los resultados de sus cálculos en cinta de papel pero los hace visibles al operador en visores situados en el carro de la máquina. Esta máquina realiza fundamentalmente cuatro operaciones: adición, sustracción, desplazamiento del carro con respecto al teclado y puesta a cero de los visores. La multiplicación se realiza por medio de adición repetida; la división se efectúa por sustracción repetida del dividendo del divisor. Tanto la multiplicación como la división se simplifican en gran parte moviendo el carro un número adecuado de lugares decimales. Algunos modelos poseen un teclado divisor (con un teclado multiplicador separado), otros llevan un teclado único.

Recientemente se han difundido las *calculadoras de mesa electrónicas.* Estas realizan los cálculos por medio de transistores y circuitos como en un computador electrónico. Estas máquinas combinan la conveniente accesibilidad y un precio relativamente bajo, con la rapidez instantánea de cálculo que se consigue por medios electrónicos en lugar de mecánicos. Los números de introducen por medio de una tecla y los resultados se exponen en un visor o imprimen en una cinta de papel. Muchos de los modelos son programables, permitiendo que se efectúen operaciones repetitivas automáticamente tan pronto como un número se introduce en la máquina. Algunos de estos modelos permiten programas bastante sofisticados (hasta varios cientos de instrucciones), y estos programas

se graban en cintas o fichas que pueden insertarse en la máquina preparándola inmediatamente para realizar una serie determinada de cálculos.

Desde mediados de los años 40 los computadores electrónicos han revolucionado el manejo y procesamiento de los datos. Un computador digital tiene las mismas posibilidades aritméticas básicas que las calculadoras de mesa mencionadas en la sección anterior (adición, sustracción, multiplicación y división), pero además tiene los siguientes rasgos distintivos. Se introducen los datos y los resultados del cálculo se presentan automáticamente bajo el control de una serie de instrucciones detalladas, previamente preparadas, que están almacenadas dentro del computador. Estas instrucciones controlan además la secuencia exacta de operaciones aritméticas que se van a efectuar. Los computadores tienen también la capacidad de ejecutar diferentes series de instrucciones dependiendo de que una determinada cantidad sea negativa, cero, o positiva. Así, es posible desarrollar *programas* sofisticados que pueden poner en juego diferentes órdenes en función de los resultados de cálculos previos.

Un computador digital típico consta de tres componentes principales: la *memoria*, que almacena los datos y las instrucciones, el *procesador*, que es el que realiza las operaciones aritméticas; y el *dispositivo periférico* que maneja la entrada, salida y almacenamiento intermedio de datos e instrucciones.

La materia que se presenta en este libro consta de cálculos estadísticos relativamente corrientes, la mayoría de los cuales probablemente hayan sido programados ya en cualquier centro de cálculo. En el libro más amplio de biometría de los autores (Sokal and Rohlf, 1969) se resaltan ciertos programas típicos que cubren materias de este libro y se escriben en el lenguaje de programador FORTRAN IV.

Ejercicios 2

2.1 Diferenciar entre los siguientes pares de términos y dar un ejemplo de cada uno: a) Poblaciones estadísticas y biológicas. b) Variante e individuo. c) Exactitud y precisión (repetibilidad). d) Intervalo de clase y marca de clase. e) Diagrama de barras e histograma. f) Abscisa y ordenada.

2.2 Redondear los números siguientes con tres cifras significativas: 106,55, 0,06819, 3,0495, 7815,01, 2,91,49 y 20,1500. ¿Cuáles son los límites implícitos antes y después de redondearlos? Redondear estos mismos números con una cifra decimal.

2.3 Dadas 200 medidas que varían desde 1,32 mm hasta 2,95 mm, ¿cómo se agruparían en una distribución de frecuencias? Dar límites de clase y marcas de clase.

2.4 Agrupar en una distribución de frecuencias las 40 medidas siguientes de anchura interorbital de una muestra de palomas y dibujar su histograma (datos de Olson y Miller, 1958). Las medidas están en mm.

12,2	12,9	11,8	11,9	11,6	11,1	12,3	12,2	11,8	11,8
10,7	11,5	11,3	11,2	11,6	11,9	13,3	11,2	10,5	11,1
12,1	11,9	10,4	10,7	10,8	11,0	11,9	10,2	10,9	11,6
10,8	11,6	10,4	10,7	12,0	12,4	11,7	11,8	11,3	11,1

2.5 ¿Con qué precisión se debería medir la longitud del ala de una especie de mosquitos en un estudio de variación geográfica, si el ejemplar más pequeño tiene una longitud de 2,8 mm aproximadamente y el más grande de 3,5 mm?

2.6 Buscar los logaritmos comunes de las 40 medidas del ejercicio 2.4 y hacer una distribución de frecuencias de estas variantes transformadas. Comentar el cambio resultante sobre el modelo de la distribución de frecuencias encontrada anteriormente.

Capítulo 3

Estadística descriptiva

Una etapa inicial y fundamental en cualquier ciencia es la etapa descriptiva. Hasta que los hechos puedan ser descritos exactamente tal como son, es prematuro un análisis de sus causas. La cuestión ¿qué? es antes que ¿cómo? A no ser que sepamos algo acerca de la distribución usual del contenido de azúcar de la sangre en una población de cobayas, así como sus fluctuaciones día a día y dentro de los días, seremos incapaces de averiguar el efecto de una dosis determinada de una droga sobre esta variable. En una muestra de tamaño normal sería claramente tedioso obtener información del material estudiando todas las observaciones individuales. Necesitamos algún modelo de resumen que nos permita abordar los datos en forma manejable y nos capacite para compartir con otros nuestras conclusiones en coloquios científicos y publicaciones. Un histograma o diagrama de barras de la distribución de frecuencias sería un tipo de resumen. Sin embargo, para la mayoría de los propósitos es necesario un resumen numérico que describa brevemente, aunque con exactitud, las propiedades de la distribución de frecuencias observadas. Las cantidades que proporcionan este resumen pueden llamarse *estadísticos descriptivos.* Este capítulo introducirá algunos de ellos y mostrará cómo se calculan.

En este capítulo se discutirán dos clases de estadísticos descriptivos: estadísticos de localización y estadísticos de dispersión. Los *estadísticos de localización* describen la posición de una muestra a lo largo de una dimensión determinada que representa una variable. Así, cuando medimos la longitud de ciertos animales nos gustaría saber si las medidas de la muestra están próximas a 2 cm. o 20 cm. Por lo tanto, un estadístico de localización debe dar un valor representativo para la muestra de observaciones. Sin embargo, este estadístico (conocido a veces también como medida de tendencia central) no describirá la forma de una distribución de frecuencias. Esta puede ser alargada o muy estrecha; puede ser gibosa o en forma de U; puede tener dos gibas o ser marcadamente asimétrica. Se necesitan medidas cuantitativas de tales aspectos de las distribuciones de frecuencias. Con este fin necesitamos definir y estudiar los *estadísticos de dispersión.*

La media aritmética descrita en la Sección 3.1 es indudablemente el estadístico de localización más importante, pero hay otros (la media geométrica, la media armónica, la mediana y la moda) que se mencionan brevemente en las secciones 3.2, 3.3, y 3.4. En la sección 3.5 se señala brevemente un estadístico sencillo de dispersión (el rango), y en la sección 3.6 se explica la desviación típica, el estadístico de dispersión más corriente. Nuestro primer caso de contrastes entre estadísticos de muestra y parámetros de población se presenta en la sección 3.7, junto con estadísticos de localización y dispersión. En la sección 3.8 hay una descripción de métodos de codificación de datos con el fin de simplificarlos para el cálculo a máquina de la media y desviación típica que se discute en la sección 3.9. El coeficiente de variación (estadístico que nos permite comparar el grado de dispersión relativa de diferentes muestras) se explica en la última sección (3.10).

Las técnicas disponibles después de haber dominado este capítulo no serán muy eficaces para resolver problemas biológicos pero serán herramientas indispensables para cualquier trabajo ulterior de bioestadística. Otros estadísticos descriptivos tanto de localización como de dispersión se abarcarán en capítulos posteriores.

Nota importante: En este capítulo nos encontraremos por primera vez con el uso de logaritmos. Para evitar confusión aquí y en capítulos subsiguientes, los logaritmos vulgares se han abreviado sistemáticamente como log y los logaritmos naturales como ln. Así, $\log x$ significa $\log_{10} x$ y $\ln x$ se refiere a $\log_e x$.

3.1 La media aritmética

El estadístico de localización más corriente resulta familiar a cualquiera. Se trata de la *media aritmética,* llamada habitualmente *media* o *promedio.* La media se calcula sumando todas las observaciones individuales o items de una muestra y dividiendo esta suma por el número de items de la muestra. Por ejemplo, como resultado de un análisis de gas en un respirómetro un investigador obtiene los cuatro porcentajes de oxígeno siguientes:

$$\begin{aligned} &14{,}9 \\ &10{,}8 \\ &12{,}3 \\ &\underline{23{,}3} \\ \text{Suma} = &\ 61{,}3 \end{aligned}$$

Calcula el porcentaje medio de oxígeno como la suma de los cuatro porcentajes (items) dividida por el número de items, en este caso por cuatro. Así, el porcentaje medio de oxígeno es

$$\text{Media} = \frac{61{,}3}{4} = 15{,}325\%$$

Al calcular una media se nos presenta la oportunidad de aprender simbolismo estadístico. Ya hemos visto (sección 2.2) que una observación individual se simboliza por Y_i, que

representa la observación i en la muestra. Cuatro observaciones podrían escribirse simbólicamente como sigue:

$$Y_1, Y_2, Y_3, Y_4$$

Definiremos el *tamaño de muestra n*, como el número de items en una muestra. En este ejemplo particular el *tamaño de muestra* es 4. Así, en una muestra grande podemos simbolizar la serie ordenada desde el primero al enésimo item como sigue:

$$Y_1, Y_2, \ldots, Y_n$$

Cuando deseamos sumar items, utilizamos la siguiente notación:

$$\sum_{i=1}^{i=n} Y_i = Y_1 + Y_2 + \cdots + Y_n$$

La letra mayúscula griega Σ, representa simplemente suma de los items indicados. La $i = 1$ significa que los items deberán sumarse comenzando por el primero y terminando por el enésimo como se indica por el $i = n$ encima del Σ. El subscrito y el sobrescrito son necesarios para indicar cuantos items deberán sumarse. La "$i =$" de la parte superior se omite usualmente por superflua. Por ejemplo, si hubiésemos querido sumar los tres primeros items solamente, habríamos escrito $\sum_{i=1}^{3} Y_i$. Por el contrario, si hubiésemos querido sumar todos ellos excepto el primero habríamos escrito $\sum_{i=2}^{n} Y_i$. Con algunas excepciones (que aparecerán en capítulos posteriores), es deseable omitir subscritos y sobrescritos que generalmente contribuyen a la complejidad de la fórmula y, cuando son innecesarios, distraen la atención del estudiante de las relaciones importantes expresadas por la fórmula. Más abajo se observan simplificaciones crecientes de la notación sumatoria del extremo izquierdo:

$$\sum_{i=1}^{i=n} Y_i = \sum_{i=1}^{n} Y_i = \sum_{i} Y_i = \sum^{n} Y = \sum Y$$

El tercer símbolo puede interpretarse como sigue: suma de las Y_i para todos los valores disponibles de i. Esta notación se utiliza frecuentemente, aunque no la utilizaremos en este libro. La siguiente, con n como sobrescrito, indica que deben sumarse n items de Y; nótese que el subíndice i de la Y se ha omitido por innecesario. Finalmente, a la derecha se presenta la notación más sencilla. Indica solamente suma de las Y. Esta será la forma que utilizaremos más frecuentemente y si un signo sumatorio precede a una variable se entenderá que es la suma de n items(todos los items de la muestra) a no ser que los subscritos o sobreescritos nos digan lo contrario.

Utilizaremos el símbolo $\bar{Y}$ para la media aritmética de la variable Y. Su fórmula se escribe como:

$$\bar{Y} = \frac{\sum Y}{n} = \frac{1}{n} \sum Y \tag{3.1}$$

En efecto, esta fórmula nos indica que se sumen todos los (n) items y se divida la suma por n. Aplazaremos hasta la sección 3.9 una discusión de la mecánica a seguir para calcular eficientemente una media.

La media de una muestra representa el centro de las observaciones. Si se dibuja el histograma de una distribución de frecuencias observadas en una hoja de cartulina, al recortar el histograma y dejarlo extendido contra una pizarra sosteniéndolo por debajo con un lápiz, existe la posibilidad de que esté desequilibrado, inclinándose a la izquierda o a la derecha. Al mover el lápiz y dirigirlo a una posición en la que se equilibre el histograma con exactitud, este punto de equilibrio será la media aritmética. En realidad, éste sería un método empírico para hallar la media aritmética de una distribución de frecuencias.

3.2 Otras medias

En los capítulos 10 y 11 veremos que las variables se transforman a veces en sus logaritmos o recíprocos. Si calculamos las medias de estas variables transformadas y después las cambiamos otra vez a la escala original, estas medias no coincidirán con las medias aritméticas calculadas a partir de las variables originales. Las medias resultantes reciben nombres especiales en estadística. La media de las variables transformadas logarítmicamente y transformada a su escala original, se llama *media geométrica*. Se calcula como:

$$G.M._Y = \text{antilog} \frac{1}{n} \sum \log Y \tag{3.2}$$

lo que indica que la media geométrica $G.M._Y$ es el antilogaritmo de la media de los logaritmos de la variable Y. Puesto que la suma de logaritmos es equivalente a la multiplicación de sus antilogaritmos, otra forma de representar esta cantidad es

$$G.M._Y = \sqrt[n]{Y_1 Y_2 Y_3 \cdots Y_n} \tag{3.3}$$

La media geométrica permite familiarizarnos con otro símbolo operador: pi mayúscula, Π, que puede leerse como producto. Del mismo modo que Σ simboliza la suma de los items que le siguen, Π simboliza la multiplicación de los mismos. Los subscritos y sobrescritos tienen exactamente el mismo significado que en el caso de la suma. Así, la expresión (3.3) para la media geométrica puede expresarse de forma reducida como sigue:

$$G.M._Y = \sqrt[n]{\prod_{i=1}^{n} Y_i} \tag{3.3a}$$

El cálculo de la media geométrica por la expresión (3.3a) es bastante pesado. En la práctica, la media geométrica debe calcularse transformando las variantes en logaritmos.

El recíproco de la media aritmética de los recíprocos se llama *media armónica*. Si lo simbolizamos por H_Y, la fórmula de la media armónica puede escribirse en forma concisa (sin subscritos ni sobrescritos) como

$$\frac{1}{H_Y} = \frac{1}{n}\sum \frac{1}{Y} \tag{3.4}$$

El lector puede demostrar que la media geométrica y la media armónica de los cuatro porcentajes de oxígeno son 14,65 % y 14,09 %, respectivamente. A no ser que los items individuales no varíen, la media geométrica es siempre menor que la media aritmética, y la media armónica es siempre menor que la media geométrica.

Algunos principiantes en estadística tienen dificultad para aceptar el hecho de que sean permisibles o incluso deseables otras medidas de localización o tendencia central distintas de la media aritmética. Piensan que la media aritmética es el promedio "lógico", y que cualquier otra media falsearía los datos. Este problema se relaciona con la escala de medida adecuada para representar los datos, esta escala no siempre es la escala lineal familiar a cualquiera, sino que a veces es preferible una escala logarítmica o recíproca. Si se tienen dudas acerca de esta cuestión, trataremos de mitigarlas en el capítulo 10, donde discutimos las razones para transformar variables.

3.3 La mediana

La *mediana M* es un estadístico de localización útil a veces en investigación biológica. Se define como el valor de la variable (en una serie ordenada) que tiene igual número de items a cada lado. Así, la mediana divide una distribución de frecuencias en dos mitades. En la siguiente muestra de cinco medidas,

14, 15, 16, 19, 23

$M = 16$, ya que la tercera observación tiene el mismo número de observaciones a ambos lados. Podemos hacer visible la mediana fácilmente si pensamos en una ordenación de mayor a menor, por ejemplo, una fila de hombres alineados por sus estaturas. El individuo mediano será pues la persona que tiene igual número de hombres a su derecha y a su izquierda. Su altura será la altura mediana de la muestra considerada. Esta cantidad se calcula fácilmente en una muestra ordenada de un número impar de individuos. Cuando el número es par, la mediana se calcula convencionalmente como el punto medio entre la variante que ocupa el lugar $n/2$ y la $[(n/2) + 1]$. Así, para la muestra de cuatro medidas

14, 15, 16, 19

la mediana sería el punto medio entre el segundo y el tercer item, o 15,5.

Siempre que un valor cualquiera de la variante aparezca más de una vez, pueden presentarse problemas para localizar la mediana. El cálculo del item mediana se complica más porque todos los miembros de una clase determinada en la que está situado el item mediana tendrán la misma marca de clase. La mediana es en ese caso la variante que ocupa el lugar $n/2$ en la distribución de frecuencias. Usualmente se calcula como el punto en que estaría localizado el individuo mediano entre los límites de clase de la clase mediana (suponiendo que los individuos estuviesen igualmente distribuidos en la clase).

La mediana es solamente un miembro de una familia de estadísticos que dividen una distribución de frecuencias en áreas iguales. Divide la distribución en dos mitades. Los tres *cuartiles* cortan la distribución en los puntos 25, 50 y 75 %, es decir, en puntos que dividen la distribución en primero, segundo, tercero y cuarto cuartos por área (y frecuencias). El segundo cuartil es naturalmente la mediana. (Hay además quintiles, deciles y percentiles, que dividen la distribución en 5, 10 y 100 porciones iguales, respectivamente.)

Desde el punto de vista de su aplicación en posteriores y más avanzados trabajos estadísticos, la mediana no es un estadístico útil (excepto para métodos "no paramétricos"; ver capítulo 10). Sin embargo, en ciertos casos especiales es una medida de localización más representativa que la media aritmética. Tales ejemplos incluyen casi siempre distribuciones asimétricas. Un ejemplo de economía frecuentemente citado sería una medida de localización adecuada para el salario "típico" de un empleado de una corporación. Los salarios muy altos de los pocos ejecutivos más antiguos elevarían la media aritmética, el centro de gravedad, hacia un valor completamente no representativo. En cambio la mediana estaría poco afectada por unos pocos salarios altos; ella representaría el punto particular en la escala de salarios sobre el cual quedan el 50 % de los salarios de la corporación, siendo la otra mitad inferiores a esta cifra.

Un ejemplo en que es preferible la aplicación de una mediana sobre una media aritmética en biología puede darse en poblaciones que presentan distribución asimétrica, tales como los pesos. Así, un peso mediano de varones americanos de 50 años puede ser un estadístico más significativo que el peso medio. La mediana es importante también en los casos en que puede resultar difícil o imposible obtener y medir todos los items de una muestra necesarios para obtener una media. Un ejemplo clarificará esta situación. Un etólogo animal está estudiando el tiempo necesario para que una muestra de animales realice una cierta fase del comportamiento. La variable que está midiendo es el tiempo desde el comienzo del experimento hasta que cada individuo lo ha realizado. Lo que él quiere obtener es un tiempo medio de realización. Sin embargo, este tiempo medio sólo podría calcularse después de haberse obtenido el dato de todos los individuos. Puede necesitarse mucho tiempo para que los animales más lentos completen su acción, más del que el observador desea gastar observándolos. Además, algunos de ellos pueden no responder nunca apropiadamente, haciendo imposible el cálculo de una media. Por lo tanto, un estadístico de localización conveniente para describir estos animales puede ser el tiempo mediano de realización, o un estadístico relacionado, tal como los percentiles 75 o 90. Así, tan pronto como el observador conozca cuál es el tamaño total de la muestra, no necesitará obtener medidas para el extremo derecho de su distribución. Ejemplos similares serían las respuestas a una droga o veneno en un grupo de individuos (la dosis letal mediana o dosis efectiva, LD_{50} o ED_{50}) o el tiempo mediano de aparición de una mutación en varias líneas de una especie.

3.4 La moda

La *moda* se refiere al valor de la variable "que más veces se obtiene" en una distribución de frecuencias, o el valor representado por el mayor número de individuos. Cuando se observa una distribución de frecuencias, la moda es el valor de la variable en el que la curva presenta un pico. En distribuciones de frecuencias agrupadas la moda como un punto no tiene mucho sentido. Habitualmente, basta identificar la clase modal. En biología, la moda no tiene muchas aplicaciones.

Las distribuciones que tienen dos picos (iguales o desiguales en altura) se llaman *bimodales;* las que tienen más de dos *multimodales.* En las distribuciones raras que tienen forma de U, llamamos *antimoda* al punto inferior del centro de la distribución.

Al valorar los méritos relativos de la media aritmética, la mediana y la moda, deben tenerse en cuenta varias consideraciones. En estadística, generalmente se prefiere la media ya que tiene un error estándar menor que otros estadísticos de localización (explicados en la sección 6.2), es más fácil calcularla matemáticamente, y tiene una propiedad adicional deseable (explicada en la sección 6.1): tenderá a distribuirse normalmente incluso si los datos originales no lo están. La media se ve marcadamente afectada por observaciones extrañas; la mediana y la moda no. Generalmente la media es más sensible a cambios en la

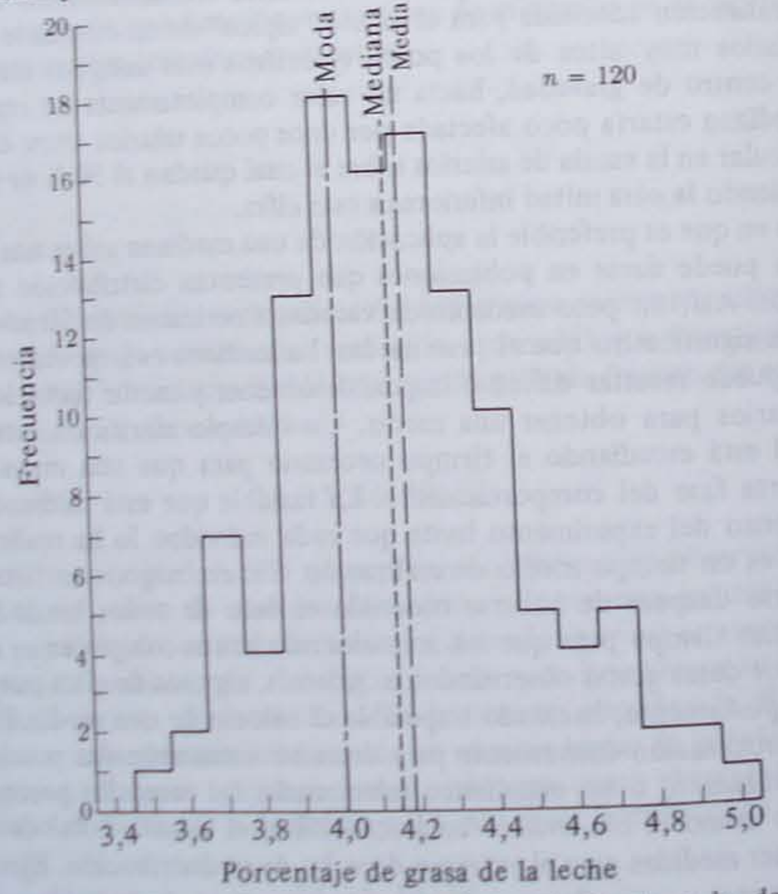


Fig. 3.1. Distribución de frecuencias asimétrica (estirada a la derecha) que muestra la situación de la media, mediana y moda. Porcentaje de grasa en 120 muestras de leche (del libro de registro de una ganadería canadiense).

forma de una distribución de frecuencias y si se quiere tener un estadístico que refleje tales cambios puede ser recomendable la media. En distribuciones simétricas unimodales, la media, la mediana, y la moda son idénticas. El mejor ejemplo de esto es la bien conocida distribución normal del capítulo 5. En una distribución asimétrica típica tal como se presenta en la figura 3.1, las posiciones relativas de la moda, mediana, y media son generalmente éstas: la media es la más próxima al extremo estirado de la distribución, la moda es la más lejana y la mediana está entre las dos. Una forma fácil de recordar esta secuencia es acordarse de que se presentan en orden alfabético desde el extremo más largo de la distribución.

3.5 El rango

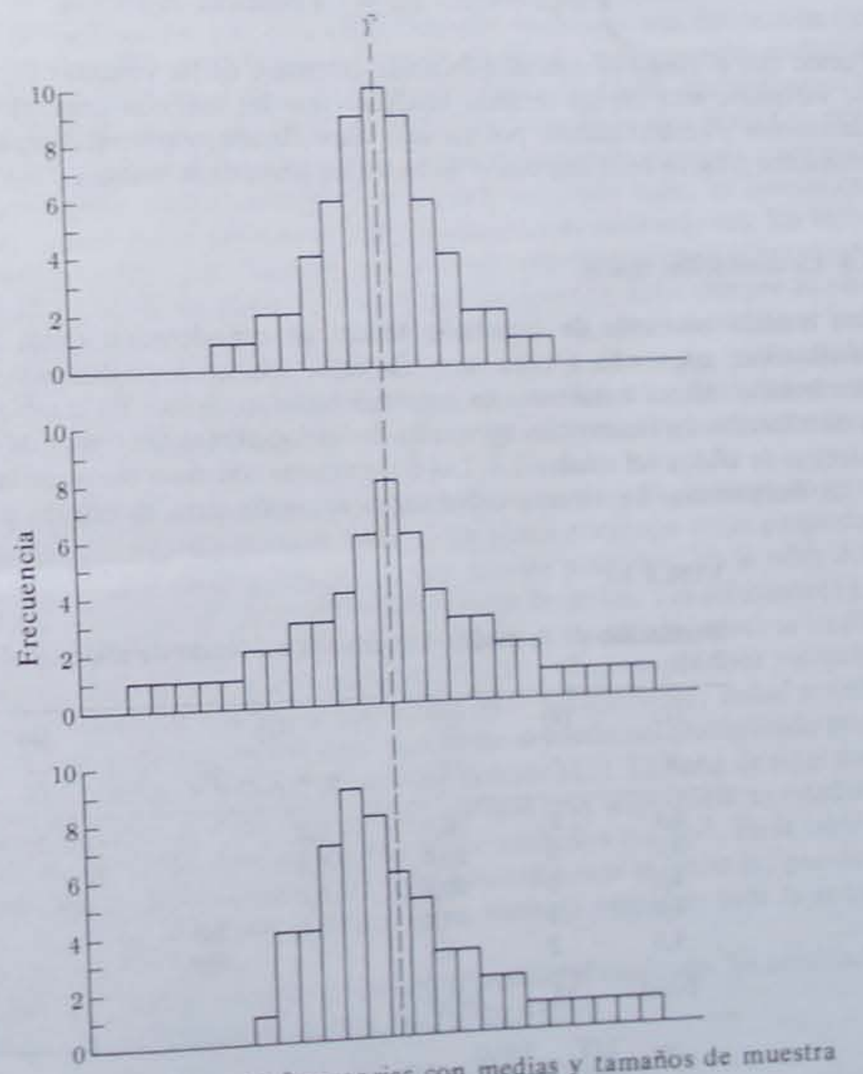


Fig. 3.2. Tres distribuciones de frecuencias con medias y tamaños de muestra idénticos, pero diferente patrón de dispersión.

Nos dirigimos ahora a medidas de dispersión. La figura 3.2 demuestra que distribuciones de aspecto completamente diferente pueden poseer idéntica media aritmética. Es obvio por lo tanto que deben encontrarse otras formas de caracterizar distribuciones.

Una medida sencilla de dispersión es el *rango*. Es la diferencia entre el mayor y el menor item en una muestra. Así, el rango de los cuatro porcentajes de oxígeno apuntados anteriormente (sección 3.1) es

$$\text{Rango} = 23{,}3 - 10{,}8 = 12{,}5\ \%$$

y el rango de las longitudes del fémur de áfidos (cuadro 2.1) es

$$\text{Rango} = 4{,}7 - 3{,}3 = 1{,}4 \text{ unidades de } 0{,}1 \text{ mm.}$$

Puesto que el rango es una medida de la extensión de las variantes a lo largo de la escala de variables, está en las mismas unidades que las medidas originales. El rango se ve claramente afectado incluso por un solo valor alejado, y por esta razón es solamente un estimador grosero de la dispersión de todos los items de la muestra.

3.6 La desviación típica

Una medida adecuada de dispersión tendrá en consideración todos los items de una distribución, asignando a cada item un valor relativo a su distancia del centro de la distribución. Ahora trataremos de construir este estadístico. En la tabla 3.1 presentamos la distribución de frecuencias agrupadas de las longitudes del fémur de las hembras apomícticas de áfidos del cuadro 2.1. Las dos primeras columnas muestran las marcas de clase y las frecuencias. La tercera columna es necesaria para el cálculo de la media de la

TABLA 3.1

Desviación de la media. Longitudes del fémur de áfidos (del cuadro 2.1).

(1)	(2)	(3)	(4)		(5)	
Y	f	fY	$y = Y - \bar{Y}$		fy	
3,4	2	6,8	−0,6		−1,2	
3,7	8	29,6	−0,3		−2,4	
4,0	5	20,0	0,0		0,0	
4,3	8	34,4		0,3		2,4
4,6	2	9,2		0,6		1,2
Total	25	100,0			−3,6	3,6

$$\bar{Y} = \frac{\Sigma fY}{\Sigma f} = \frac{100{,}0}{25} = 4{,}0$$

distribución de frecuencias (ver sección 3.9 posterior). Es la marca de clase Y multiplicada por la frecuencia f. El cálculo de la media aparece en la parte inferior de la tabla. La longitud media de fémur resulta ser 4,0 unidades.

La distancia de cada marca de clase a la media se calcula como la desviación siguiente:

$$y = Y - \bar{Y}$$

Por convenio, cada desviación individual (*"deviate"*) se calcula como la observación individual menos la media $Y - \bar{Y}$ en lugar de la inversa, $\bar{Y} - Y$. Las desviaciones individuales se simbolizan por letras minúsculas correspondientes a las letras mayúsculas de las variables. La columna (4) de la tabla 3.1 da las desviaciones individuales calculadas de este modo. Para comodidad en el cálculo, la columna se ha dividido en desviaciones positivas y negativas. Ahora las desviaciones han de multiplicarse por sus respectivas frecuencias para que aquellas que se presentan más frecuentemente contribuyan más a nuestra medida de dispersión que las que sólo se presentan rara vez. Por esta razón multiplicamos las desviaciones individuales por sus frecuencias f. Los resultados de estos cálculos se exponen en la columna (5), que conserva la separación entre desviaciones positivas y negativas.

Ahora nos proponemos calcular una desviación media sumando todas las desviaciones individuales y dividiendo por el número de dichas desviaciones en la muestra. Sin embargo, cuando sumamos nuestras desviaciones, observamos que las positivas y negativas se anulan, como se muestra en las sumas al final de la columna (5). Esto siempre es cierto para la suma de las desviaciones con respecto a la media, y está relacionado con el hecho de que la media es el centro de gravedad. Consecuentemente la desviación media también sería siempre igual a cero. Es recomendable estudiar el apéndice A1.1, el cual demuestra que la suma de las desviaciones en torno a la media de una muestra es siempre igual a cero.

Elevando al cuadrado las desviaciones individuales se evita el que la suma de las desviaciones en torno a la media sea siempre cero y da como resultado otras propiedades matemáticas deseables que consideraremos en una sección posterior. En la tabla 3.2 se presentan de nuevo los datos de las longitudes del fémur de áfidos. Las columnas (1), (2) y (3) son las marcas de clase, frecuencias y desviaciones, determinadas como se ha dicho previamente. Las desviaciones no están separadas ahora en columnas positivas y negativas. La columna (4) presenta sus cuadrados, los cuales son, naturalmente, todos positivos. Finalmente, la columna (5) presenta el cuadrado de las desviaciones multiplicado por sus frecuencias, es decir, columna (4) multiplicada por columna (2). La suma de estas desviaciones elevadas al cuadrado es 2,88. Esta es una cantidad muy importante en estadística, que para abreviar se denomina *suma de cuadrados* y se simboliza por Σy^2. En la tabla 3.2, la suma de cuadrados se simboliza por Σfy^2 pero habitualmente se omite la f puesto que Σ indica suma de todos los items posibles. Otro símbolo ordinario para la suma de cuadrados es *SS* (*sum of squares*).

El próximo paso es obtener la media de las n desviaciones al cuadrado. La cantidad que resulta se conoce como *varianza* o *desviación cuadrática media.*

$$\text{Varianza} = \frac{\Sigma y^2}{n} = \frac{2{,}88}{25} = 0{,}1152$$

La varianza es una medida de fundamental importancia en estadística y la utilizaremos a lo largo de este libro. De momento sólo necesitamos recordar que por haber elevado al cuadrado las desviaciones, la varianza se expresa en unidades al cuadrado. Para contrarrestar el efecto de elevar al cuadrado, extraemos ahora la raíz cuadrada positiva de la varianza y obtenemos la *desviación típica:*

$$\text{Desviación típica} = +\sqrt{\frac{\sum y^2}{n}} = 0{,}3394$$

Una desviación típica se expresa asimismo en las unidades de medida originales, puesto que es una raíz cuadrada de las unidades cuadráticas de la varianza.

Nota importante: No utilizar la técnica recién aprendida e ilustrada en la tabla 3.2 para el cálculo manual de una varianza y desviación típica. Esta técnica es excesivamente tediosa.

Es posible que el lector haya notado que hemos evitado asignar símbolo alguno a la varianza y desviación típica. En la próxima sección explicaremos porqué.

TABLA 3.2

La desviación típica. Método largo no recomendado para cálculos reales, pero presentado aquí para aclarar el significado de la desviación típica. Datos de la tabla 3.1.

(1)	(2)	(3)	(4)	(5)
Y	f	$y = Y - \overline{Y}$	y^2	fy^2
3,4	2	−0,6	0,36	0,72
3,7	8	−0,3	0,09	0,72
4,0	5	0,0	0,00	0,00
4,3	8	0,3	0,09	0,72
4,6	2	0,6	0,36	0,72
Total	25			2,88

$\overline{Y} = 4{,}0 \qquad \Sigma fy^2 = 2{,}88$

$\text{Varianza} = \frac{\Sigma fy^2}{n} = \frac{2{,}88}{25} = 0{,}1152 \qquad \text{Desviación típica} = \sqrt{0{,}1152} = 0{,}3394$

3.7 Estadísticos de muestra y parámetros

Hasta ahora hemos calculado estadísticos de muestras sin dar demasiada importancia a lo que representan estos estadísticos. Cuando se calculan correctamente, una media y desviación típica serán siempre medidas de localización y de dispersión absolutamente fieles para las muestras en las cuales están basadas. Así, la verdadera media de los cuatro porcentajes de oxígeno de la sección 3.1 es en realidad 15,325 %. La desviación típica de

las 25 longitudes del fémur es 0,3394 unidades cuando los items se agrupan como se ha expuesto. Sin embargo, en biología (o en estadística en general a este respecto), rara vez nos interesan medidas de localización y dispersión exclusivamente como resúmenes descriptivos de las muestras que hemos estudiado. Casi siempre nos interesan las *poblaciones* de las que se han extraído las muestras. Por lo tanto, lo que nos gustaría conocer no es la media de los cuatro porcentajes de oxígeno, sino el verdadero porcentaje de oxígeno del universo de lecturas del cual se han extraído estas cuatro. Igualmente, nos gustaría saber la verdadera longitud media del fémur de la población de hembras apomícticas de áfidos, no solamente la media de los 25 individuos que hemos medido. Cuando estudiamos dispersión, generalmente deseamos conocer las verdaderas desviaciones típicas de las poblaciones y no las de las muestras. Estos estadísticos de población, sin embargo, son desconocidos y (hablando en términos generales) imposibles de conocer. ¿Quién sería capaz de coger todas las hembras apomícticas de esta población de áfidos particular y medirlas? Así pues, necesitamos utilizar *estadísticos de muestra* como estimadores de *estadísticos de población* o *parámetros.*

En estadística es convencional utilizar letras griegas para parámetros de población y letras romanas para estadísticos de muestra. De este modo, la media de la muestra $\overline{Y}$ estima la media paramétrica de la población, μ. Igualmente, una varianza de muestra, simbolizada por s^2, estima una varianza paramétrica, simbolizada por σ^2. Estos estimadores deberían ser *insesgados.* Con esto queremos decir que las muestras (independientemente del tamaño de muestra) extraídas de una población con un parámetro conocido, deberían dar estadísticos de muestra que por término medio diesen el valor paramétrico. Un estimador que no cumple esto se dice que es *sesgado.* La media Y de una muestra es un estimador insesgado de la media paramétrica μ. En cambio, la varianza de una muestra calculada como hemos visto (en la sección 3.6) es sesgada. Por lo general subestimará la magnitud de la varianza de la población σ^2. Para superar este error, los estadísticos matemáticos han demostrado que al dividir las sumas de cuadrados por $n - 1$ en vez de por n, las varianzas de muestras resultantes serán estimadores insesgados de la varianza de la población. Por esta razón, es habitual calcular varianzas dividiendo la suma de cuadrados por $n - 1$. La fórmula de la desviación típica se da pues habitualmente como sigue:

$$s = +\sqrt{\frac{\sum y^2}{n-1}} \tag{3.5}$$

En los datos de hembras apomícticas de áfidos, la desviación típica se calcularía pues como:

$$s = \sqrt{\frac{2{,}88}{24}} = 0{,}3464$$

Observamos que este valor es sólo ligeramente superior que el estimado anteriormente de 0,3394. Es obvio que cuanto más grande sea el tamaño de muestra menor será la diferencia entre la división por n y por $n - 1$. No obstante, prescindiendo del tamaño de muestra, es buena costumbre dividir una suma de cuadrados por $n - 1$ cuando se calcula una varianza o una desviación típica. Puede suponerse que cuando se encuentra el símbolo s^2, se refiere a una varianza obtenida por división de la suma de cuadrados entre los *grados de libertad,* a los que generalmente se refiere la cantidad $n - 1$. El único caso en

que es apropiada la división de la suma de cuadrados por n es cuando el interés del investigador está verdaderamente limitado a la muestra que tiene a mano y a su varianza y desviación típica como estadísticos descriptivos de la muestra, y no como estimadores de los parámetros de población. Hay también casos en que el investigador posee datos de la población completa; en tales casos está perfectamente justificada la división por n ya que entonces no se está estimando un parámetro sino,en realidad,calculándolo. Así, la varianza de las longitudes del ala de todas las grullas adultas de una especie americana sería un valor paramétrico; igualmente, si se hubiese medido el C.I. de todos los ganadores del Premio Nobel de Física, su varianza sería un parámetro ya que está basada en la población completa.

3.8 Codificación de datos antes del cálculo

Antes de que podamos discutir cómo calcular medias y desviaciones típicas de una forma práctica en una calculadora de mesa, debe aprenderse un procedimiento más importante, la codificación de los datos originales. Por *codificación* entendemos la adición o sustracción de una constante a los datos originales y/o la multiplicación o división de estos datos por una constante. Frecuentemente los datos tienen que ser codificados porque se expresan originalmente con demasiados dígitos o son números muy grandes que pueden causar dificultades y errores durante el tratamiento de datos. La importancia de la correcta codificación de datos no puede sobrestimarse. El cálculo estadístico logrado, obteniendo el resultado correcto y no hundiéndose en una marisma de cifras, se facilita con frecuencia por codificación correcta de los datos originales.

Llamaremos *codificación aditiva* a la adición o sustracción de una constante (puesto que sustracción no es más que adición de un número negativo). Igualmente llamaremos *codificación multiplicativa* a la multiplicación o división por una constante (ya que división es multiplicación por el recíproco del divisor). Llamaremos *codificación de combinación* a la aplicación de ambas, aditiva y multiplicativa, a la misma serie de datos.

En el apéndice A1.2 examinamos las consecuencias de los tres tipos de codificación en el cálculo de medias, varianzas y desviaciones típicas. Los resultados para *medias* pueden resumirse como sigue: cuando se ha sumado una constante a cada variante, la media codificada $\overline{Y}_c$ se descodifica restándole la constante.

Cuando las variantes se han codificado multiplicándolas por una constante, $\overline{Y}_c$ puede descodificarse dividiéndola por la misma constante; igualmente, cuando las variantes se han codificado por división, la media se descodifica multiplicándola por esta misma constante.

Las medias codificadas por combinación pueden descodificarse efectuando la *operación inversa, en inversa secuencia*. Esto puede clarificarse por medio de un sencillo diagrama nemotécnico, que casualmente sirve igual para codificación simple multiplicativa o aditiva. Así, si hemos codificado una serie de variantes restándoles 5 y multiplicándolas por 10, descodificamos la media dividiéndola primero por 10 y sumándole luego 5:

$$\text{codificar} \xrightarrow[\times 10]{-5} \overline{Y}_c$$
$$\overline{Y} \xleftarrow[+10]{+5} \text{descodificar}$$

Igualmente, una media codificada basada en variantes que se han dividido por 100 y a las cuales se les ha sumado 1,2, podría descodificarse restando primero 1,2 y multiplicando después por 100:

$$\text{codificar} \xrightarrow[\div 100]{+1.2} \overline{Y}_c$$
$$\overline{Y} \xleftarrow[\times 100]{-1.2} \text{descodificar}$$

Los diagramas siguientes aclaran la codificación simple:

$$\text{codificar} \xrightarrow{+10} \qquad \text{o} \qquad \text{codificar} \xrightarrow{\times 4}$$
$$\xleftarrow[-10]{} \text{descodificar} \qquad\qquad \xleftarrow[\div 4]{} \text{descodificar}$$

Al considerar los efectos de la codificación de variantes en los valores de *varianzas* y *desviaciones típicas*, encontramos en primer lugar que las codificaciones no tienen efecto en las sumas de cuadrados, varianzas, ni desviaciones típicas. La demostración matemática se da en el apéndice A1.2, pero esto puede verse intuitivamente debido a que una codificación aditiva no afecta a la distancia de un item respecto de su media. La distancia de un item de 15 a su media de 10 sería 5. Si fuésemos a codificar las variantes restándoles la constante 10, el item sería ahora 5 y la media cero. La diferencia entre ellos continuaría siendo 5. Así, si se utiliza solamente la codificación aditiva, el único estadístico que necesita descodificación es la media. Pero la codificación multiplicativa sí afecta a las sumas de cuadrados, varianzas y desviaciones típicas. Las desviaciones típicas codificadas s_c tienen que dividirse por el código multiplicativo, lo mismo que debería hacerse para la media. Sin embargo, las sumas de cuadrados o varianzas tienen que dividirse por los códigos multiplicativos elevados al cuadrado, porque son términos cuadráticos y el factor multiplicativo se eleva al cuadrado durante las operaciones. En la codificación de combinación el código aditivo puede ignorarse. Los diagramas nemotécnicos siguientes pueden utilizarse para resumir la descodificación de desviaciones típicas:

$$\text{codificación} \xrightarrow[\times 10]{+5} s_c$$
$$s \xleftarrow[\div 10]{} \text{descodificar}$$

y para sumas de cuadrados o varianzas:

$$\text{codificación} \xrightarrow[+1.2]{+100} \sum y_c^2 \text{ o } s_c^2$$
$$\sum y^2 \text{ o } s^2 \xleftarrow[\times 100^2]{} \text{descodificar}$$

En los cuadros 3.1 y 3.2 que se discuten en la próxima sección, se presentan ejemplos de codificación y descodificación de datos.

3.9 Métodos prácticos para calcular la media y la desviación típica

El procedimiento utilizado para calcular la suma de cuadrados en la sección 3.6 puede expresarse por la fórmula siguiente:

$$\sum y^2 = \sum (Y - \overline{Y})^2 \tag{3.6}$$

Esta fórmula explica lo más claramente posible la naturaleza de la suma de cuadrados, pero excepto en computadores digitales es la más embarazosa. Para calcular Σy^2 por la expresión (3.6), primero tendríamos que calcular la media, después calcular la desviación de cada item respecto de la media. A continuación debería elevarse al cuadrado cada desviación, y finalmente sumarse estos cuadrados. La operación más pesada sería calcular y elevar al cuadrado las desviaciones, lo cual en la mayoría de las calculadoras de mesa no puede hacerse automáticamente si no es transcribiendo las desviaciones o al menos introduciéndolas manualmente. En nuestra experiencia, muchos estudiantes, una vez que han aprendido alguna técnica particular, tienden a repetirla por rutina. Por lo tanto, habiendo aprendido a calcular una suma de cuadrados como se muestra en la tabla 3.2 pudieran tender a repetir este procedimiento particular. No podemos pues insistir con la suficiente fuerza en que la tabla 3.2 se presenta exclusivamente por razones pedagógicas; ordinariamente no se calcula la suma de cuadrados como se muestra en ella. Vamos a desarrollar ahora una fórmula de cálculo para la desviación típica. En resumen, hay tres pasos necesarios para calcular este estadístico: (1) hallar la suma de cuadrados, Σy^2, (2) dividir por $n - 1$ para obtener la varianza, y (3) extraer la raíz cuadrada de la varianza para obtener la desviación típica. Los pasos (2) y (3) son operaciones sencillas y no es posible ningún mecanismo ahorrador de tiempo para ellas. Nuestros esfuerzos para simplificar el cálculo deben centrarse en el paso (1), el cálculo de la suma de cuadrados. La fórmula de cálculo habitual para esta cantidad es

$$\sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n} \tag{3.7}$$

Vamos a precisar lo que representa esta fórmula. El primer término del segundo miembro de la ecuación, ΣY^2 debería denominarse "suma de Y cuadrado" y debería distinguirse

$$\sum Y^2 = Y_1^2 + Y_2^2 + Y_3^2 + \cdots + Y_n^2$$

cuidadosamente de Σy^2, "la suma de cuadrados de Y". Estos nombres son inadecuados, pero están demasiado bien establecidos para pensar en rectificarlos. La otra cantidad de la expresión (3.7) es $(\Sigma Y)^2/n$. Con frecuencia se denomina *término de corrección*, CT. El numerador de este término es el cuadrado de la suma de las Y; esto es, primero se suman todos los valores de Y y luego se eleva al cuadrado esta suma. Por lo general, esta cantidad es diferente de ΣY^2, que primero eleva al cuadrado las Y y después las suma. Estos dos términos solamente son idénticos si todas las Y son iguales. Si no se está seguro de ello, se puede probar calculando estas cantidades para unos pocos números.

CUADRO 3.1

Cálculo de $\overline{Y}$ y s de datos no agrupados.

Datos de la longitud del fémur de áfidos, sin agrupar, como se muestran a la cabecera del cuadro 2.1.

Los datos se han codificado multiplicando por 10 para eliminar la coma decimal durante el cálculo. Las variables y estadísticos codificados se identifican por el subscrito c.

Cálculo	*Codificación y descodificación*
$n = 25$	Codificación: $Y_c = 10Y$
$\sum Y_c = 1001$	
$\overline{Y}_c = \frac{1}{n}\sum Y_c = 40{,}04$	Para descodificar $\overline{Y}_c$: $\overline{Y} = \frac{\overline{Y}_c}{10} = \frac{40{,}04}{10} = 4{,}004$
$\sum Y_c^2 = 40\,401$	
$\sum y_c^2 = \sum Y_c^2 - \frac{(\sum Y_c)^2}{n}$	
$= 40\,401 - \frac{(1001)^2}{25}$	
$= 320{,}960$	
$s_c^2 = \frac{\sum y_c^2}{n-1} = \frac{320{,}960}{24}$	
$= 13{,}37$	Para descodificar s_c^2: $s^2 = \frac{s_c^2}{10^2} = \frac{13{,}37}{100} = 0{,}1337$
$s_c = \sqrt{13{,}37} = 3{,}656$	Para descodificar s_c: $s = \frac{s_c}{10} = \frac{3{,}656}{10} = 0{,}3656$

¿Por qué la expresión (3.7) es idéntica a la expresión (3.6)? La demostración de esta identidad es muy sencilla y se da en el apéndice A1.3.

En las expresiones (3.1) y (3.7) vemos que para calcular la media y desviación típica de una muestra es necesario tener tres cantidades, n, ΣY, y ΣY^2. Las operaciones detalladas dependen de si los datos están sin agrupar o agrupados en una distribución de frecuencias. Hemos recomendado dejar los datos sin agrupar cuando el número de variantes es menor que 150 debido a que el tiempo necesario para establecer una distribución de frecuencias sería casi equivalente al tiempo ahorrado durante el cálculo si los datos estuviesen en forma de distribución de frecuencias. Cuando los datos están sin agrupar, el cálculo procede como en el cuadro 3.1 que está basado en los datos no agrupados de las longitudes del fémur de áfidos presentados en la cabecera del cuadro 2.1. Para eliminar la coma decimal durante el cálculo, los datos se han codificado multiplicando las variantes por 10.

Cuando los datos están agrupados en una distribución de frecuencias, los cálculos son considerablemente más sencillos. En el cuadro 3.2 se presenta un ejemplo. Estos son los

CUADRO 3.2

Cálculo de $\overline{Y}$, s, y CV de una distribución de frecuencias.

Pesos de varones chinos recién nacidos, en onzas

(1) *Marca de clase* Y	(2) f	(3) *Marca de clase codificada* Y_c
59,5	2	0
67,5	6	1
75,5	39	2
83,5	385	3
91,5	888	4
99,5	1729	5
107,5	2240	6
115,5	2007	7
123,5	1233	8
131,5	641	9
139,5	201	10
147,5	74	11
155,5	14	12
163,5	5	13
171,5	1	14
	$9465 = n$	

Fuente: Millis y Seng (1954).

Cálculo

$\sum fY_c = 59\,629$

$\overline{Y}_c = \sum f\overline{Y}_c / \sum f = 6{,}300$

$\sum fY_c^2 = 402\,987$

$CT = \dfrac{(\sum fY_c)^2}{n} = 375\,659{,}550$

$\sum fy_c^2 = \sum fY_c^2 - CT = 27\,327{,}450$

$s_c^2 = \dfrac{\sum fy_c^2}{n-1} = 2{,}888$

$s_c = 1{,}6991$

Codificación y descodificación

Codificación: $Y_c = \dfrac{Y - 59{,}5}{8}$

Para descodificar $\overline{Y}_c$: $\overline{Y} = 8\overline{Y}_c + 59{,}5$
$= 50{,}4 + 59{,}5$
$= 109{,}9$ oz.

Para descodificar s_c: $s = 8s_c = 13{,}593$ oz.

$$CV = \frac{s}{\overline{Y}} \times 100 = \frac{13{,}593}{109{,}9} \times 100 = 12{,}369\%$$

pesos de niños varones chinos recién nacidos, expuestos anteriormente en la figura 2.3. En primer lugar tienen que ser codificados para cambiar las marcas de clase un poco engorrosas. Esto se hace restando 59,5, la marca de clase inferior de la serie. Las marcas de clase que resultan son valores tales como 0, 8, 16, 24, 32, etc. Después se dividen por 8, lo que las transforma en 0, 1, 2, 3, 4, etc., que es la forma deseada. Ahora se efectúa el cálculo de una forma sencilla y elegante. Sumamos los productos de fY_c y fY_c^2 para obtener ΣfY_c y ΣfY_c^2. Al sumar las frecuencias se obtiene $\Sigma f = n$.

Una regla empírica común para estimar la media es promediar las observaciones mayor y menor para obtener el llamado *rango medio*. Para los datos de hembras apomícticas de áfidos del cuadro 2.1, este valor es $(4{,}7 + 3{,}3)/2 = 4{,}0$, que corresponde exactamente a la media de la muestra calculada. Las desviaciones típicas pueden estimarse a partir de los rangos por división adecuada de éstos.

Para muestras de	*Dividir el rango por*
10	3
30	4
100	5
500	6
1000	$6\frac{1}{2}$

El rango de los datos de áfidos es 1,4. Cuando este valor se divide por 4 se obtiene un estimador de la desviación típica de 0,35, que no equipara demasiado mal al valor de 0,3656 calculado en el cuadro 3.1. Estos métodos aproximados son muy útiles para detectar errores grandes en el cálculo.

3.10 El coeficiente de variación

Al haber presentado la desviación típica como una medida del grado de variación de los datos, nos podemos preguntar, "¿Ahora qué?". En esta etapa de nuestro aprendizaje de teoría estadística no se obtiene resultado últil alguno a partir de los cálculos que hemos efectuado, aunque los conocimientos prácticos recién adquiridos son básicos para todo el trabajo estadístico posterior. Hasta aquí, la única aplicación que podemos tener para la desviación típica es como una estimación del grado de variación en una población. Así, puede que queramos comparar las magnitudes de la desviación típica de poblaciones similares y ver si la población A es más o menos variable que la población B. No obstante, cuando las poblaciones difieren apreciablemente en cuanto a sus medias, la comparación de sus varianzas o desviaciones típicas sería bastante arriesgada. Por ejemplo, la desviación típica de las longitudes de la cola de elefantes es obviamente mucho mayor que la longitud total de la cola de un ratón. Con el fin de comparar el grado de variación en poblaciones que tienen diferentes medias, se ha desarrollado el *coeficiente de variación*.

Este es simplemente la desviación típica expresada como un porcentaje de la media. Su fórmula es

$$CV = \frac{s \times 100}{\bar{Y}} \tag{3.8}$$

Por ejemplo, el coeficiente de variación de los pesos de recién nacidos del cuadro 3.2 es 12,37 %, como se muestra al final de este cuadro. El coeficiente de variación es independiente de la unidad de medida y se expresa como un porcentaje.

Los coeficientes de variación son ampliamente utilizados cuando se quiere comparar la variación de dos poblaciones, independientemente de la magnitud de sus medias. Probablemente es de poco interés descubrir si los pesos de los niños chinos recién nacidos son más o menos variables que las longitudes del fémur de las hembras apomícticas de áfidos. Podemos calcular el último como 0,3656 × 100/4,004 = 9,13 %, lo cual sugiere que los pesos al nacer son más variables. Más frecuentemente desearemos comprobar si una muestra biológica determinada es más variable para un carácter que para otro. Así, para una muestra de ratas ¿es más variable el peso corporal que el contenido de azúcar de la sangre? Un segundo tipo de comparación frecuente, especialmente en sistemática, se da entre poblaciones diferentes para el mismo carácter. Así, es posible que hayamos medido la longitud del ala en muestras de pájaros de varias localidades. Deseamos conocer si una cualquiera de estas poblaciones es más variable que las otras. Examinando los coeficientes de variación de la longitud del ala en estas muestras, puede obtenerse una respuesta a esta cuestión.

Ejercicios 3

3.1 Hallar la media, desviación típica, y coeficiente de variación para los datos de palomas del ejercicio 2.4. Agrupar los datos en diez clases, calcular de nuevo $\bar{Y}$ y s, y compararlas con los resultados obtenidos a partir de los datos no agrupados. Calcular la mediana para los datos agrupados.

3.2 Hallar $\bar{Y}$, s, CV, y la mediana para los datos siguientes (mg de glicina por mg de creatinina en la orina de 37 chimpancés; de Gartler, Firschein, y Dobzhansky, 1956). SOLUCION. $\bar{Y} = 0{,}115$, $s = 0{,}10404$.

.008	.018	.056	.055	.135	.052	.077	.026	.440	.300
.025	.036	.043	.100	.120	.110	.100	.350	.100	.300
.011	.060	.070	.050	.080	.110	.110	.120	.133	.100
.100	.155	.370	.019	.100	.100	.116			

3.3 Los datos siguientes son porcentajes de grasa de la leche de 120 vacas Ayrshire de 3 años, seleccionados al azar de un libro de registro de ganado canadiense.

(a) Calcular $\bar{Y}$, s y CV directamente de los datos. Se encontrará ventajoso calcular ΣY y ΣY^2 para cada una de las cuatro columnas separadamente, ya que en caso de que se cometa un error sólo se tendrá que volver a calcular una columna en vez de las cuatro. Guardar los datos de cada columna. Se utilizarán posteriormente.

(b) Agrupar los datos en una distribución de frecuencias y calcular de nuevo $\bar{Y}$, s, y CV. Comparar los resultados con los de (a). ¿Cuánta precisión se ha perdido por el agrupamiento? Calcular también la mediana.

4.32	4.24	4.29	4.00
3.96	4.48	3.89	4.02
3.74	4.42	4.20	3.87
4.10	4.00	4.33	3.81
4.33	4.16	3.88	4.81
4.23	4.67	3.74	4.25
4.28	4.03	4.42	4.09
4.15	4.29	4.27	4.38
4.49	4.05	3.97	4.32
4.67	4.11	4.24	5.00
4.60	4.38	3.72	3.99
4.00	4.46	4.82	3.91
4.71	3.96	3.66	4.10
4.38	4.16	3.77	4.40
4.06	4.08	3.66	4.70
3.97	3.97	4.20	4.41
4.31	3.70	3.83	4.24
4.30	4.17	3.97	4.20
4.51	3.86	4.36	4.18
4.24	4.05	4.05	3.56
3.94	3.89	4.58	3.99
4.17	3.82	3.70	4.33
4.06	3.89	4.07	3.58
3.93	4.20	3.89	4.60
4.38	4.14	4.66	3.97
4.22	3.47	3.92	4.91
3.95	4.38	4.12	4.52
4.35	3.91	4.10	4.09
4.09	4.34	4.09	4.88
4.28	3.98	3.86	4.58

3.4 ¿Qué efecto tendría sobre los valores numéricos de los estadísticos siguientes $\bar{Y}$, s, CV, desviación media, mediana, moda, rango, sumar 5,2 a todas las observaciones? ¿Cuál sería el efecto de sumar 5,2 y después multiplicar las sumas por 8,0? Si primero multiplicásemos por 8,0 y después sumásemos 5,2, ¿supondría alguna diferencia en los estadísticos anteriores?

3.5 Demostrar que la ecuación de la varianza puede escribirse también como

$$s^2 = \frac{\sum Y^2 - n\bar{Y}^2}{n-1}$$

(Esta expresión no se utiliza mucho en la práctica porque puede llevar a errores graves de redondeo a menos que $\bar{Y}$ se calcule con muchas cifras significativas).

3.6 Estimar μ y σ utilizando el rango medio y el rango (ver sección 3.9) para los datos de los ejercicios 3.1, 3.2, y 3.3. ¿Hasta qué punto concuerdan estos valores con las estimaciones dadas por $\bar{Y}$ y s? SOLUCION. Los valores de μ y σ para el ejercicio 3.2 son 0,224 y 0,1014.

Capítulo 4

Introducción a las distribuciones de probabilidad: Binomial y de Poisson

En la sección 2.5 hemos encontrado por primera vez distribuciones de frecuencias. Por ejemplo, la tabla 2.2 muestra una distribución de una variable merística o discreta (discontinua), el número de plantas de chufa por cuadrado. Ejemplos de distribuciones para variables continuas son las longitudes del fémur de las hembras apomícticas de áfidos en el cuadro 2.1 o los pesos humanos de recién nacidos del cuadro 3.2. Cada una de estas distribuciones nos informa sobre la frecuencia absoluta de cualquier tipo de variable. Así, la mayor parte de los cuadrados contienen una, dos, o ninguna planta de chufa. En la clase 139,5 onzas de peso de recién nacidos, encontramos solamente 201 del total de 9 465 niños examinados; esto es, aproximadamente sólo el 2,1 % de los niños están en esa clase. Vemos claramente que estas distribuciones de frecuencia solamente son muestras de determinadas poblaciones. Los pesos de recién nacidos representan a una población de niños varones chinos de un área geográfica determinada. Pero si supiésemos que nuestra muestra era representativa de esa población, podríamos hacer toda clase de predicciones basadas en la distribución de frecuencias de la muestra. Por ejemplo, podríamos decir que aproximadamente el 2,1 % de los niños varones chinos nacidos en esta población, pesaría entre 135,5 y 143,5 onzas al nacer. Igualmente podríamos decir que la probabilidad de que un niño cualquiera de esta población pesara al nacer 139,5 onzas, es muy baja. Si a cada uno de los 9.465 pesos se les asignara un número, una vez mezclados los números en un sombrero, y sacado sólo uno, la probabilidad de que éste fuese uno de los 201 de la clase 139,5 onzas sería realmente muy baja, sólo 0,021. Sería mucho más probable que sacásemos un niño de 107,5 o 115,5 onzas, ya que los niños de estas clases están representados por frecuencias de 2 240 y 2 007, respectivamente. Finalmente, si tuviésemos que muestrear de una población desconocida de niños y encontrásemos que los primeros individuos extraídos tenían un peso al nacer de 170 onzas, probablemente rechazaríamos la hipótesis de que la población desconocida era la misma que la muestreada en el cuadro 3.2. A esta conclusión llegaríamos porque en la distribución del cuadro 3.2, sólo uno

entre casi 10 000 niños tenía un peso al nacer tan alto. Aunque fuese posible que hubiésemos muestreado de dicha población y obtenido un peso al nacer de 170 onzas, la probabilidad de que el primer individuo extraído tuviese tal valor, es realmente muy baja. Parece mucho más razonable suponer que la población desconocida de la que extraemos la muestra, tiene diferente media y posiblemente varianza, que la del cuadro 3.2.

Hemos utilizado esta distribución de frecuencias empírica para hacer ciertas predicciones (con qué frecuencia ocurrirá un suceso determinado), o para hacer juicios y tomar decisiones (¿es probable que un niño de determinado peso al nacer pertenezca a esta población?). Sin embargo, en muchos casos en biología, no haremos tales predicciones a partir de distribuciones empíricas, sino basándonos en consideraciones teóricas que a nuestro juicio son pertinentes. Podemos creer que los datos deberían estar distribuidos de una forma determinada por suposiciones básicas sobre la naturaleza de las fuerzas que actúan en el ejemplo que tenemos a mano. Si nuestros datos realmente observados no se ajustan a los valores esperados en base a estas suposiciones, tendremos serias dudas sobre ellas. Este es un uso común de las distribuciones de frecuencias en biología. Las suposiciones que se están comprobando llevan generalmente a una distribución de frecuencias teórica, conocida también como *distribución de probabilidad.* Esta puede ser una simple distribución de dos valores, tal como la razón 3:1 en un cruce mendeliano, o puede ser una función más complicada que trate de predecir el número de plantas por cuadrado. Si nos encontramos con que los datos observados no se ajustan a los esperados en base a la teoría, esto nos lleva con frecuencia al descubrimiento de algún mecanismo biológico que causa esta desviación de lo esperado. Los fenómenos de ligamiento en genética, de apareamiento preferente entre diferentes genotipos, en comportamiento animal, de agrupación de animales en ciertos lugares preferidos, o por el contrario su dispersión territorial, son casos a propósito. De este modo, haremos uso de la teoría de probabilidad para probar nuestras suposiciones sobre las leyes de incidencia de ciertos fenómenos biológicos. No obstante, deberíamos indicar al lector que la teoría de probabilidad es la base de toda la estructura de la estadística, ya que debido a la orientación no matemática de este libro, esto puede no ser completamente obvio.

En las secciones que siguen, presentaremos en primer lugar una breve discusión de probabilidad (sección 4.1), limitada a lo necesario para comprender las secciones que siguen al nivel propuesto de sofisticación matemática. A continuación nos ocuparemos de la distribución binominal (sección 4.2), que no sólo es importante en ciertos tipos de estudios,por ejemplo genéticos, sino que además es fundamental para una comprensión de las diversas clases de distribuciones de probabilidad que se discutirán en este libro.

La distribución de Poisson, sección 4.3, es de amplia aplicación en biología, especialmente para pruebas de contingencia sobre incidencia de ciertos sucesos. Esta y la binomial son distribuciones de probabilidad de variables discretas. La distribución continua de probabilidad más habitual es la distribución normal, discutida en el capítulo siguiente.

4.1 Probabilidad, muestreo al azar y contraste de hipótesis

Comenzaremos esta discusión con un ejemplo que no es biométrico ni biológico en sentido estricto. Con frecuencia hemos encontrado pedagógicamente eficaz el introducir

nuevos conceptos a través de situaciones completamente familiares al estudiante, aunque el ejemplo no sea relevante a la materia general de la bioestadística.

Vamos a recurrir a la Universidad de Matchless, una institución estatal situada entre los Apalaches y las Rocosas. Al mirar sus números de inscripción observamos la siguiente división de la clase estudiantil: 70 % de los estudiantes son americanos no graduados (AN) y 26 % americanos graduados (AG); el 4 % restante son extranjeros, de los cuales el 1 % son extranjeros no graduados (EN) y el 3 % extranjeros graduados (EG). En gran parte de nuestro trabajo utilizaremos proporciones en vez de porcentajes para mayor comodidad. De este modo, el registro de inscripción consta de 0,70 de AN, 0,26 de AG, 0,01 de EN y 0,03 de EG. La clase estudiantil total correspondiente al 100 % está representada por la cifra 1,0.

Si pudiésemos reunir a todos los estudiantes y muestrear 100 de ellos al azar, esperaríamos intuitivamente que por término medio 3 serían graduados extranjeros. El resultado real puede variar. Podría no haber ni un estudiante EG entre los 100 muestreados, o podría haber pocos más de 3. La razón del número de graduados extranjeros obtenido respecto del número total de estudiantes muestreados puede por lo tanto variar desde cero hasta un número considerablemente mayor que 0,03. Si aumentásemos nuestro tamaño de muestra a 500 o 1 000, sería menos probable que la razón fluctuase ampliamente alrededor de 0,03. Cuanto mayor sea la muestra extraída, la razón de estudiantes EG obtenidos respecto del total de estudiantes muestreados, se aproximará más estrechamente a 0,03. De hecho la *probabilidad* de muestrear un estudiante extranjero se define como el límite de la razón de estudiantes extranjeros respecto del número total de estudiantes extraídos cuando el tamaño de la muestra tiende a infinito. Así, podemos resumir la situación estableciendo que la probabilidad de que un estudiante de la Universidad de Matchless sea graduado extranjero es $P[EG] = 0{,}03$. Igualmente, la probabilidad de extraer un extranjero no graduado es $P[EN] = 0{,}01$, la de un americano no graduado es $P[AN] = 0{,}70$, y la de americanos graduados, $P[AG] = 0{,}26$.

Ahora vamos a imaginar el siguiente experimento: tratamos de sacar uno al azar entre los estudiantes de la Universidad de Matchless. Esto no es tan fácil como puede imaginarse. Si deseáramos realizar físicamente esta operación, tendríamos que establecer un lugar de reunión o trampa en cualquier parte del campus. Y para estar seguro de que la muestra es verdaderamente aleatoria con respecto a la población total de estudiantes, tendríamos que conocer muy a fondo la ecología de los estudiantes en el campus. Deberíamos tratar de situar nuestra trampa en algún lugar por el que cada estudiante tuviese la misma probabilidad de pasar. Pocos lugares de este tipo, si es que hay alguno, pueden encontrarse en una universidad. Los círculos de estudiantes son probablemente más frecuentados por estudiantes que viven solos y extranjeros,y menos por los que viven en casas de familias y habitaciones. En los clubs de estudiantes podrían encontrarse menos estudiantes extranjeros y graduados. Lógicamente no intentaríamos colocar nuestra trampa cerca de la Casa Internacional porque la probabilidad de muestrear un estudiante extranjero estaría enormemente aumentada. En las ventanillas de caja podríamos muestrear estudiantes que pagan enseñanza pero no becarios. No sabemos si la proporción de becarios entre estudiantes extranjeros o graduados es igual o diferente que la existente entre americanos o no graduados. Acontecimientos atléticos, reuniones políticas, bailes y tal atraerían un espectro diferencial de la clase estudiantil; verdaderamente, no parece vislumbrarse ninguna solución fácil. El momento del muestreo es igualmente importante, dependiente de la temporada, así como de la hora del día.

Aquellos lectores interesados en el muestreo de organismos de la naturaleza, ya habrán advertido problemas paralelos en su trabajo. Si fuésemos a muestrear solamente estudiantes que lleven turbantes o saris, la probabilidad de que fuesen extranjeros sería 1. Ya no podríamos hablar de una muestra al azar. En el ecosistema familiar de la universidad, estas violaciones del procedimiento correcto de muestreo son obvias a todos nosotros, pero no lo son tanto en ejemplos biológicos reales, en los que estamos poco familiarizados con la verdadera naturaleza del medio ambiente. ¿Cómo deberíamos proceder para obtener una muestra al azar de hojas de un árbol, de insectos de un campo, o de mutaciones en un cultivo? En el muestreo al azar, estamos tratando de permitir que las frecuencias de los diversos sucesos que ocurren en la naturaleza se reproduzcan inalteradamente en nuestros registros; esto es, esperamos que, por término medio, las frecuencias de estos sucesos en nuestra muestra sean las mismas que en la situación natural. Otra forma de decir esto es que en una muestra al azar, cada individuo de la población que se está muestreando tiene la misma probabilidad de estar incluido en la muestra.

Podríamos obtener una muestra al azar utilizando registros que representen al cuerpo estudiantil, tales como la guía de estudiantes, seleccionando al azar una de sus páginas y un nombre de la página. O bien podríamos asignar un número arbitrario a cada estudiante, escribir cada uno en una ficha o disco, ponerlos en una caja grande, mezclarlos bien y luego extraer un número.

Imaginemos ahora que extraemos un solo estudiante en la realidad por el método de la trampa, después de planificar detenidamente su colocación, de forma que el muestreo se haga al azar. ¿Cuáles son los resultados posibles? El estudiante podría ser un AN, AG, EN, o EG. Este *conjunto* de cuatro resultados posibles agota las posibilidades de este experimento. A este conjunto, que podemos representar por AN, AG, EN, EG se le llama *universo* o *espacio de muestreo*. Cualquier experimento singular conduciría a uno sólo de los cuatro posibles resultados (elementos) del conjunto. Dicho elemento de un espacio de muestreo se denomina *suceso simple*. Se distingue de *sucesos*, que es cualquier subconjunto del espacio de muestreo. Así, en el espacio de muestreo definido anteriormente, AN, AG, EN y FG, son sucesos simples. Algunos de los sucesos posibles son los siguientes: $\{AN, AG, EN\}$, $\{AN, AG, EG\}$, $\{AG, EG\}$, $\{AN, EG\}$, ... Según la definición de suceso, los sucesos simples así como el universo o espacio de muestreo total también son sucesos. Debería aclararse el significado de estos sucesos. Así, $\{AN, AG, EN\}$ implica que es un americano o un no graduado, o ambas cosas.

Dado el espacio de muestreo escrito más arriba, el suceso $\mathbf{A} = \{AN, AG\}$ incluye todas las posibilidades de sacar un estudiante americano. Igualmente el suceso $\mathbf{B} = \{AG, EG\}$ abarca las de obtener un estudiante graduado. La *intersección* de dos sucesos **A** y **B**, representada por $\mathbf{A} \cap \mathbf{B}$, comprende solamente los sucesos comunes a **A** y **B**. Como puede verse a continuación, sólo se limita a AG.

$$\{AN, AG\}$$
$$\{AG, EG\}$$

Así, $\mathbf{A} \cap \mathbf{B}$ es aquel suceso del espacio de muestreo que determina el muestreo de un estudiante graduado americano. Cuando la intersección de dos sucesos es el conjunto

vacío, como en $\mathbf{B} \cap \mathbf{C}$, siendo $\mathbf{C} = \{\text{AN, EN}\}$, se dice que los sucesos **B** y **C** son mutuamente excluyentes; no hay elementos comunes en estos dos sucesos del espacio de muestreo.

También podemos definir sucesos que son *unión* de otros dos sucesos del espacio de muestreo. Así, $\mathbf{A} \cup \mathbf{B}$ indica que ocurre **A** o **B** o ambos. En el ejemplo anterior, $\mathbf{A} \cup \mathbf{B}$ incluiría a todos los estudiantes que fuesen americanos, graduados, o las dos cosas, esto es, estudiantes americanos graduados.

¿Por qué el interés de definir espacios de muestreo y sucesos? Porque estos conceptos nos llevan a definiciones y operaciones útiles con respecto a la probabilidad de diversos resultados. Si podemos asignar un valor $0 \leqslant p \leqslant 1$ a cada suceso simple de un espacio de muestreo, de modo que la suma de todos los valores de p sea igual a la unidad, entonces este espacio se convierte en un *espacio de probabilidad* (finito). En nuestro ejemplo anterior los siguientes números se asociaron con los sucesos simples correspondientes del espacio de muestreo:

$$\{\text{AN, AG, EN, EG}\}$$
$$\{0{,}70,\ 0{,}26,\ 0{,}01,\ 0{,}03\}$$

Dado este espacio de probabilidad, ahora estamos capacitados para hacer afirmaciones con respecto a la probabilidad de determinados sucesos. Por ejemplo, ¿cuál es la probabilidad de que un estudiante muestreado al azar sea un americano graduado? Evidentemente es $P[\{\text{AG}\}] = 0{,}26$. ¿Cuál es la probabilidad de que sea americano o graduado? En términos de los sucesos definidos anteriormente, éste es un $P[\mathbf{A}] \cup P[\mathbf{B}] = P[\{\text{AN, AG}\}] + P[\{\text{AG, EG}\}] - P[\{\text{AG}\}] = 0{,}96 + 0{,}29 - 0{,}26 = 0{,}99$. Restamos $P[\{\text{AG}\}]$ porque de no hacerlo se incluiría dos veces, una vez en $P[\mathbf{A}]$ y otra en $P[\mathbf{B}]$, lo que llevaría al resultado absurdo de una probabilidad mayor que uno.

Vamos a suponer ahora que al muestrear un estudiante del cuerpo estudiantil de la Universidad de Matchless, éste resulta ser un graduado extranjero. ¿Qué conclusión podemos sacar de esto? Solamente por azar, este resultado ocurriría el 3 % de las veces, no muy frecuentemente. Probablemente debería rechazarse la hipótesis de muestreo al azar, ya que si la aceptásemos, este resultado del experimento sería improbable. Nótese que decimos *improbable*, no *imposible*. Es obvio que podríamos habernos encontrado primero con un EG; sin embargo, no es muy probable. La probabilidad de que un solo estudiante muestreado no sea un EG es 0,97. Si pudiéramos asegurarnos de que nuestro método de muestreo era al azar (como cuando se sacan números de estudiantes de un recipiente), tendríamos que decidir naturalmente que había ocurrido un suceso improbable. Todas las decisiones de este párrafo están basadas en nuestro conocimiento preciso de que la proporción de estudiantes de la Universidad de Matchless es verdaderamente la especificada por el espacio de probabilidad. Si dudásemos acerca de esto, el resultado de nuestro experimento de muestreo nos llevaría a suponer una mayor proporción de graduados extranjeros.

Ahora ampliaremos nuestro experimento al muestreo de dos estudiantes en vez de uno. ¿Cuáles son los posibles resultados? El nuevo espacio de muestreo puede representarse mejor por un diagrama (figura 4.1) que presenta el conjunto de los dieciséis sucesos simples posibles, como puntos en una red. Las combinaciones posibles, ignorando qué

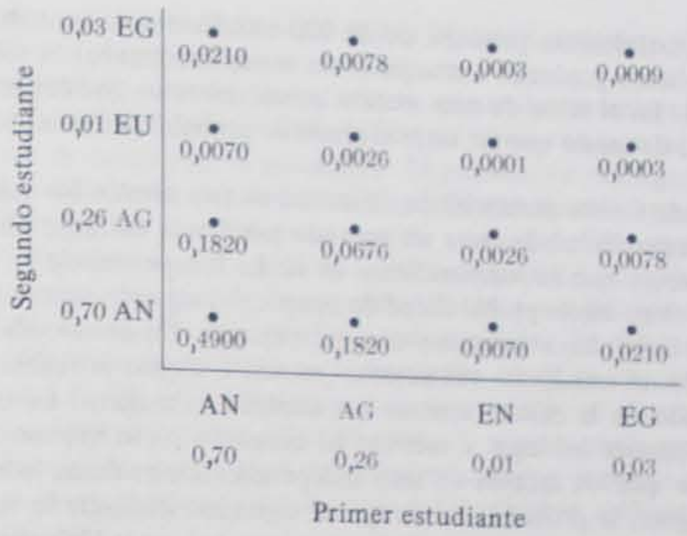


Fig. 4.1. Espacio de muestreo para el muestreo de dos estudiantes de la Universidad de Matchless. (Para más explicación ver texto.)

estudiante se sacó primero, son $\{\text{AN, AN}\}$, $\{\text{AN, AG}\}$, $\{\text{AN, EN}\}$, $\{\text{AN, EG}\}$, $\{\text{AG, AG}\}$, $\{\text{AG, EN}\}$, $\{\text{AG, EG}\}$, $\{\text{EN, EN}\}$, $\{\text{EN, EG}\}$, y $\{\text{EG, EG}\}$.

¿Cuáles serían las frecuencias esperadas de estos resultados? A partir del espacio de probabilidad anterior conocemos los resultados esperados para el muestreo de un estudiante, pero ¿cuál será el espacio probabilístico correspondiente al nuevo espacio de muestreo de 16 elementos? Ahora la naturaleza del procedimiento de muestreo resulta bastante importante. Podemos extraer con o sin *reemplazamiento*, es decir, podemos devolver a la población el primer estudiante sacado o podemos dejarlo fuera de ella. Si no lo devolvemos, la probabilidad de sacar un graduado extranjero ya no será exactamente 0,03. Esto se ve fácilmente. Vamos a suponer que la Universidad de Matchless tiene 10 000 estudiantes. En tal caso, puesto que hay 3 % de estudiantes extranjeros graduados, debe haber 300 EG en dicha Universidad. Después de sacar un graduado extranjero este número se reduce a 299 entre 9 999 estudiantes. Por consiguiente, la probabilidad de extraer un EG se convierte ahora en $299/9\,999 = 0{,}0299$, ligeramente menor que el valor 0,03 para el muestreo del primer estudiante. Por otra parte, si devolvemos el primer estudiante extranjero a la población de estudiantes y nos aseguramos de que dicha población vuelve a estar completamente al azar antes de extraer de nuevo (esto es, se le da la oportunidad de perderse entre la multitud del campus, o se mezclan bien los discos que llevan los números de estudiantes), la probabilidad de sacar un segundo estudiante EG es la misma que antes, 0,03. De hecho, si seguimos devolviendo los individuos extraídos a la población original, podemos muestrear de ella como si se tratase de una población infinita.

Las poblaciones biológicas son realmente finitas, pero con frecuencia son tan grandes que para casos de experimentos de muestreo podemos considerarlas efectivamente infinitas, tanto si devolvemos los individuos extraídos como si no. Al fin y al cabo, incluso en

esta población relativamente pequeña de 10 000 estudiantes, la probabilidad de extraer un segundo estudiante graduado extranjero (sin reemplazamiento) es sólo mínimamente diferente de 0,03. En el resto de esta sección consideraremos que los muestreos son con reemplazamiento, de modo que no varíe el grado de probabilidad de obtener un estudiante extranjero.

Hay una segunda fuente potencial de dificultad en este diseño. No solamente debemos suponer que la probabilidad de sacar un segundo estudiante extranjero es igual que la del primero sino también que es *independiente* de él. La independencia de sucesos significa que si un suceso tiene cierta probabilidad de ocurrir, un segundo suceso semejante tendrá la misma probabilidad haya ocurrido o no el primero. En el caso de los estudiantes, habiendo extraído un estudiante extranjero, ¿es más o menos probable que un segundo estudiante extraído de la misma manera sea también extranjero? La independencia de sucesos puede depender del lugar y método de muestreo. Si lo hacemos en el campus, es bastante probable que los sucesos no sean independientes; es decir, habiendo sacado un estudiante extranjero, la probabilidad de que el segundo estudiante lo sea aumenta, puesto que los estudiantes extranjeros tienden a reunirse. Así, en la Universidad de Matchless la probabilidad de que un estudiante que pasea con un graduado extranjero sea también un EG, sería mayor que 0,03.

En un espacio de muestreo, los sucesos **D** y **E** se definirán como independientes siempre que $P(\mathbf{D} \cap \mathbf{E}) = P(\mathbf{D}) \cdot P(\mathbf{E})$. Los valores de probabilidad asignados a los dieciséis puntos de la red del espacio de muestreo en la figura 4.1 han sido calculados para satisfacer la condición anterior. Así, suponiendo que $P[\mathbf{D}]$ sea igual a la probabilidad de que el primer estudiante sea un AN, esto es, $P[\{AN_1AN_2, AN_1AG_2, AN_1EN_2, AN_1EG_2\}]$ y $P[\mathbf{E}]$ igual a la probabilidad de que el segundo estudiante sea un EG, esto es, $P[\{AN_1EG_2, AG_1EG_2, EN_1EG_2, EG_1EG_2\}]$, observamos que la intersección $\mathbf{D} \cap \mathbf{E}$ es $\{AN_1EG_2\}$ cuyo valor en el espacio de probabilidad de la figura 4.1 es de 0,0210. Nos encontramos con que este valor es el producto de $P[\{AN\}] \cdot P[\{EG\}] = 0{,}70 \times 0{,}03 = 0{,}0210$. Estas relaciones mutuamente independientes han sido impuestas deliberadamente a todos los puntos del espacio de probabilidad. Por lo tanto, si las probabilidades de muestreo para el segundo estudiante son independientes del tipo de estudiante extraído primero, podemos calcular las probabilidades de los resultados sencillamente, como el producto de las probabilidades independientes. Así, la probabilidad de obtener dos estudiantes EG es $P[\{EG\}] \cdot P[\{EG\}] = 0{,}03 \times 0{,}03 = 0{,}0009$.

La probabilidad de obtener un estudiante AN y un EG en la muestra sería el producto $0{,}7 \times 0{,}03$. Sin embargo, es en realidad el doble. Es fácil ver la razón de esto. Solamente hay una forma de obtener dos estudiantes EG, esto es, sacando primero un EG y después otro EG. Igualmente, hay solamente una forma de sacar dos AN. Sin embargo, el muestreo de uno de cada tipo puede hacerse extrayendo primero un AN y después un EG, o bien, sacando primero un EG seguido de un AN. Así pues, la probabilidad es $2P[\{AN\}] \cdot P[\{EG\}] = 2 \times (0{,}70) \times (0{,}03) = 0{,}0420$.

Si realizamos tal experimento y obtenemos una muestra de dos estudiantes EG, llegaríamos a las conclusiones siguientes: Solamente 0,0009 muestras (el 9 % del 1 % o 9 entre los 10 000 casos) se esperaría que estuvieran formadas por dos estudiantes extranjeros graduados. Es bastante improbable que se obtenga tal resultado sólo por azar. Si $P(EG) = 0{,}03$ es una realidad, sospecharíamos por lo tanto que el muestreo no se hizo al azar o que

los sucesos no eran independientes (o que ambas hipótesis, muestreo al azar e independencia de sucesos, eran incorrectas).

A veces se confunde el muestreo al azar con el azar en la naturaleza. El primero es la representación fiel en la muestra de la distribución de sucesos en la naturaleza; el segundo es la independencia de sucesos en la naturaleza. El primero de éstos, generalmente está o debería estar bajo control del experimentador y está relacionado con la estrategia del buen muestreo. El segundo describe generalmente una propiedad innata de los objetos que se están extrayendo y es por tanto de mayor interés biológico. La confusión entre muestreo al azar e independencia de sucesos procede de que la falta de uno u otro puede producir frecuencias observadas que difieren de las esperadas. En nuestro ejemplo ilustrativo de la Universidad de Matchless, ya hemos visto cómo puede interpretarse desde ambos puntos de vista la falta de independencia en muestras de estudiantes extranjeros.

La explicación anterior de probabilidad es adecuada para nuestros propósitos actuales, pero demasiado incompleta para dar un conocimiento de este campo. Los lectores interesados en ampliar sus conocimientos de la materia deben dirigirse al Mosimann (1968) para una sencilla introducción.

4.2 La distribución binomial

Con vistas a la discusión que sigue simplificaremos nuestro espacio de muestreo para que conste de dos elementos solamente, estudiantes americanos y extranjeros, representados por $\{E, A\}$ e ignoraremos si son graduados o no. Vamos a representar el espacio de probabilidad por $\{p, g\}$, donde $p = P[E]$ es la probabilidad de que sea estudiante extranjero, y $q = P[A]$ es la probabilidad de que sea estudiante americano. Como antes, podemos calcular el espacio de probabilidad de muestras de dos estudiantes de la forma siguiente:

$$\{EE, EA, AA\}$$

$$\{p^2, 2pq, q^2\}$$

Si se trata de muestras de tres estudiantes, el espacio de probabilidad es como sigue:

$$\{EEE, EEA, EAA, AAA\}$$

$$\{p^3, 3p^2q, 3pq^2, q^3\}$$

De nuevo las muestras de tres estudiantes extranjeros o tres americanos sólo pueden obtenerse de una forma, y sus probabilidades son p^3 y q^3 respectivamente. Sin embargo, hay tres formas de obtener dos estudiantes de una clase y uno de la otra. Si A simboliza americano y E extranjero, la secuencia de muestreo para dos estudiantes extranjeros y un

americano podría ser AEE, EAE, y EEA. Así, la probabilidad de este resultado será $3p^2q$. Igualmente, la probabilidad de sacar dos estudiantes americanos y uno extranjero será $3pq^2$. Una forma adecuada de resumir estos resultados es por medio de la distribución binomial, que es aplicable a muestras de cualquier tamaño extraídas de poblaciones dicotómicas, es decir, que sus elementos pertenecen solamente a dos clases, por ejemplo, estudiantes que pueden ser extranjeros o americanos, individuos que pueden estar muertos o vivos, varones o hembras, negro o blanco, liso o rugoso, y así sucesivamente. Esto se realiza desarrollando el término binomial $(p+q)^k$, donde k es el tamaño de la muestra, p la probabilidad de que aparezca la primera clase, y q la de que aparezca la segunda. Por definición, $p + q = 1$; por lo tanto, q es una función de $p : q = 1 - p$. Desarrollaremos la expresión para muestras de $k = 1,2$ y 3:

Para muestras de 1, $(p + q)^1 = p + q$
Para muestras de 2, $(p + q)^2 = p^2 + 2pq + q^2$
Para muestras de 3, $(p + q)^3 = p^3 + 3p^2q + 3pq^2 + q^3$

Se observará que estas expresiones dan los mismos espacios de probabilidad discutidos previamente. Los coeficientes (números que preceden a las potencias de p y q) expresan el número de formas en que se obtiene un resultado particular. Un método fácil para calcular los coeficientes de los términos de la expresión binomial es por medio del triángulo de Pascal que se muestra a continuación.

k							
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	
...	.	.	.	.	.	.	.

El triángulo de Pascal proporciona los coeficientes de la distribución binomial, esto es, el número de posibles resultados de las diversas combinaciones de sucesos. Para $k = 1$, los coeficientes son 1, 1 respectivamente; para construir las líneas siguientes se escribe un 1 en los márgenes derecho e izquierdo de la línea, y cada valor intermedio se obtiene sumando los valores a su izquierda y derecha de la línea anterior. Este principio se sigue en todas las líneas. De esta forma, se pueden obtener los coeficientes para cualquier tamaño de muestra. La línea correspondiente a $k = 6$ estará formada por los coeficientes siguientes: 1, 6, 15, 20, 15, 6, 1. Los valores de p y q reciben potencias según un patrón determinado que sería fácil de reproducir para cualquier valor de k. Para $k = 4$:

$$p^4q^0 + p^3q^1 + p^2q^2 + p^1q^3 + p^0q^4$$

El exponente de p decrece desde 4 a 0 (k a 0 en caso general) al mismo tiempo que el de q aumenta de 0 a 4 (0 a k en caso general). Ya que cualquier número elevado a 0 es 1 y cualquier número elevado a 1 es él mismo, podemos simplificar esta expresión tal como se

muestra más abajo, y al mismo tiempo presentarla con los coeficientes del triángulo de Pascal para $k = 4$:

$$p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4$$

De este modo, podemos escribir casi a ojo el desarrollo del binomio para cualquier potencia razonable.

Supongamos que tenemos una población de insectos, de los cuales el 40 % exactamente se infectan por un determinado virus X. Si extraemos muestras de $k = 5$ insectos cada una y examinamos separadamente la presencia del virus en cada insecto, ¿qué distribución de muestras cabría esperar si la probabilidad de infección para cada insecto de una muestra fuese independiente de la de otros insectos de la misma? En este caso, $p = 0,4$ es la proporción de infectados, y $q = 0,6$ la de no infectados. Se supone que la población es tan grande que la cuestión de si el muestreo es con o sin reemplazamiento no es importante para fines prácticos. Las proporciones esperadas vendrían dadas por el desarrollo del binomio:

$$(p + q)^k = (0,4 + 0,6)^5$$

Con la ayuda del triángulo de Pascal obtendremos el espacio de probabilidad

$$\{p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5\}$$

o

$$(0,4)^5 + 5(0,4)^4(0,6) + 10(0,4)^3(0,6)^2 + 10(0,4)^2(0,6)^3 + 5(0,4)(0,6)^4 + (0,6)^5$$

que representa las proporciones esperadas de muestras de 5 insectos infectados, 4 infectados y 1 no infectado, 3 infectados y 2 no infectados, y así sucesivamente. Probablemente el lector ya habrá advertido que los términos de la distribución binomial realmente constituyen un tipo de distribución de frecuencias para los diferentes resultados. Asociada a cada resultado, tal como "cinco insectos infectados" hay una probabilidad de incidencia, en este caso $(0,4)^5 = 0,01024$. Esta es una distribución de frecuencias teórica o *distribución de probabilidad* de sucesos que pueden presentarse en dos clases. Describe la distribución esperada de resultados en muestras al azar de cinco insectos, el 40 % de los cuales son infectados. Esta distribución de probabilidad se conoce como *distribución binomial*, y el desarrollo del binomio da las frecuencias esperadas de las clases de dicha distribución.

En la tabla 4.1 se muestra un esquema apropiado para la presentación de una distribución binomial. En la primera columna aparece el número de insectos infectados por muestra, la segunda columna presenta potencias decrecientes de p desde p^5 hasta p^0, y la tercera, potencias crecientes de q desde q^0 hasta q^5. En la columna (4) se muestran los coeficientes binomiales del triángulo de Pascal. Las *frecuencias relativas esperadas*, que son las probabilidades de los diversos resultados, se presentan en la columna (5). Tales frecuencias esperadas las representamos por $\hat{f}_{rel}$ y son el producto de las columnas (2), (3) y (4). Su suma es igual a 1,0, ya que los sucesos alineados en la columna (1) agotan los

posibles resultados. En la columna (5) de la tabla 4.1 vemos que solamente alrededor del 1 % de las muestras cabe esperar que estén formadas por cinco insectos infectados. Comprobaremos si estas predicciones son válidas en un experimento real.

Experimento 4.1. Simular el muestreo de insectos infectados utilizando una tabla de números aleatorios tal como la tabla I. Estos son números de una cifra escogidos al azar, y cada número del 0 al 9 tiene la misma probabilidad de ser elegido. Por conveniencia, los números se agupan en lotes de veinte. Puesto que cada número tiene la misma probabilidad de aparecer, puedes mantener cuatro números cualesquiera (0, 1, 2, 3) para los insectos infectados y los restantes (4, 5, 6, 7, 8, 9) para los no infectados. La probabilidad de que cualquier número de la tabla seleccionado represente un insecto infectado (esto es, sea 0, 1, 2 o 3) es por lo tanto 40 % o 0,4, ya que éstos son cuatro de los diez números posibles. Asimismo, se supone que los valores de los sucesivos números son independientes de los de números previos. Por lo tanto, en este experimento deberían cumplirse las hipótesis de la distribución binomial. Entrar en la tabla de números al azar por un punto arbitrario (no siempre al principio) y observar sucesivos grupos de cinco números, anotando en cada grupo cuantos fueron 0, 1, 2, o 3. Elegir tantos grupos de cinco como el tiempo te permita, pero no menos de 100 grupos.

La columna (7) de la tabla 4.1 expone los resultados de un experimento de este tipo realizado por los alumnos de una clase de bioestadística durante un año. Se obtuvo un total de 2 423 muestras de cinco números de la tabla y en esta columna se presenta la distribución de los cuatro números que representan el porcentaje de infección. Las frecuencias observadas se designan por f. Para calcular las frecuencias esperadas en este ejemplo real, multiplicamos las frecuencias relativas esperadas $\hat{f}_{rel}$ de la columna (5) por $n = 2\,423$, el número de muestras extraídas. Esto da como resultado las *frecuencias absolutas esperadas,* designadas por $\hat{f}$, que aparecen en la columna (6). Al comparar las frecuencias observadas de la columna (7) con las frecuencias esperadas de la columna (6), observamos concordancia entre las dos columnas de números. En la figura 4.2 se representan las dos distribuciones. Si las frecuencias observadas no se ajustasen a las esperadas, podríamos pensar que la falta de ajuste se debía exclusivamente al azar. O podríamos pensar en rechazar una o más de las hipótesis siguientes: 1) que la verdadera proporción de números 0, 1, 2 y 3 es 0,4 (el rechazo de esta hipótesis normalmente no sería razonable, puesto que podemos confiar en el hecho de que la proporción de números 0, 1, 2 y 3 en una tabla de números al azar *es* 0,4 o muy próximo); 2) que el muestreo fue al azar; 3) que los sucesos son independientes.

TABLA 4.1

Frecuencias esperadas de insectos infectados en muestras de 5 insectos extraídos de una población infinitamente grande con un hipotético porcentaje de infección del 40 %.

(1) *Número de insectos infectados por muestra* Y	(2) *Potencias de* $p = 0.4$	(3) *Potencias de* $q = 0.6$	(4) *Coeficientes binomiales*	(5) *Frecuencias relativas esperadas* $\hat{f}_{rel}$	(6) *Frecuencias absolutas esperadas* $\hat{f}$	(7) *Frecuencias observadas* f
5	0.01024	1.00000	1	0.01024	24.8	29
4	0.02560	.60000	5	0.07680	186.1	197
3	0.06400	.36000	10	0.23040	558.3	535
2	0.16000	.21600	10	0.34560	837.4	817
1	0.40000	.12960	5	0.25920	628.0	643
0	1.00000	.07776	1	0.07776	188.4	202
			$\Sigma\hat{f}$ o $\Sigma f\,(=n)$	1.00000	2423.0	2423
			ΣY	2.00000	4846.1	4815
			Media	2.00000	2.00004	1.98721
			Desviación típica	1.09545	1.09543	1.11934

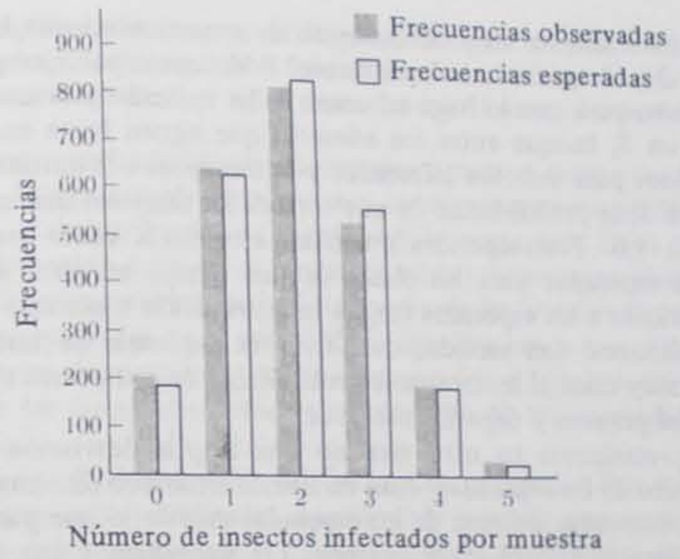


Fig. 4.2. Diagrama de barras de frecuencias observadas y esperadas dadas en la tabla 4.1.

Estas afirmaciones pueden reinterpretarse en términos del modelo de infección original con el cual comenzamos esta discusión. Si en lugar de un experimento de muestreo de números por un grupo de bioestadística, éste hubiese sido un experimento real de muestreo de insectos, concluiríamos que verdaderamente los insectos habían sido muestreados al azar y que no teníamos pruebas para rechazar la hipótesis de que la proporción de insectos infectados fuese 40 %. Si las frecuencias observadas no se hubiesen ajustado a las esperadas, la falta de ajuste podría atribuirse al azar, o a la conclusión de que la verdadera proporción de infección no es 0,4, o tendríamos que rechazar una o ambas de las hipótesis siguientes: 1) que el muestreo era al azar, y 2) que la incidencia de insectos infectados en estas muestras era independiente.

El experimento 4.1 se ha diseñado para dar muestras al azar y sucesos independientes.

¿Cómo podríamos simular un procedimiento de muestreo en que la incidencia de los números 0, 1, 2 y 3 no fuese independiente? Podríamos, por ejemplo, instruir al que realice el muestreo para que lo haga tal como se ha indicado previamente, pero cada vez que encuentra un 3, busque entre los números que siguen hasta encontrar otro de los cuatro establecidos para insectos infectados y lo incorpore a la muestra. Así, una vez que se encontrara un 3, la probabilidad de que otro de los números indicados se incluyera en la muestra sería 1,0. Tras repetidas muestras, esto daría como resultado frecuencias superiores a las esperadas para las clases de dos o más números de los indicados, y frecuencias inferiores a las esperadas (según la distribución binomial) para clases de uno solo. Podrían planearse una variedad de diferentes esquemas de muestreo de este tipo. Debería quedar muy claro al lector que la probabilidad de que ocurra el segundo suceso es diferente de la del primero y dependiente de él.

¿Cómo interpretaríamos en otro ejemplo una amplia desviación de las frecuencias observadas respecto de las esperadas? Aún no hemos estudiado técnicas para comprobar si las frecuencias observadas difieren de las esperadas más de lo que puede atribuirse solamente al azar. Esto se tratará en el capítulo 13. Supongamos que se ha realizado tal prueba y nos ha demostrado que nuestras frecuencias observadas son significativamente diferentes de las esperadas. Son probables dos tipos principales de desviación de lo esperado: 1) *agrupamiento* y 2) *repulsión*, expuestas en ejemplos simulados en la tabla 4.2. En ejemplos reales no tendríamos nociones a priori acerca de la magnitud de p (probabilidad de uno de los dos posibles resultados). En tales casos se acostumbra a obtener p de la muestra observada y se calculan las frecuencias esperadas utilizando la muestra p. Esto

TABLA 4.2

Distribuciones artificiales para mostrar agrupamiento y repulsión. Frecuencias esperadas de la tabla 4.1.

(1) *Número de insectos infectados por muestra* Y	(2) *Frecuencias absolutas esperadas* $\hat{f}$	(3) *Frecuencias agrupadas (contagiosas)* f	(4) *Desviación de lo esperado*	(5) *Frecuencias en repulsión* f	(6) *Desviación de lo esperado*
5	24,8	47	+	14	−
4	186,1	227	+	157	−
3	558,3	558	0	548	−
2	837,4	663	−	943	+
1	628,0	703	+	618	−
0	188,4	225	+	143	−
$\Sigma\hat{f}$ o n	2423,0	2423		2423,0	
ΣY	4846,1	4846		4846	
Media	2,00004	2,00000		2,00000	
Desviación típica	1,09543	1,20074		1,01435	

significaría que la hipótesis de que p es un determinado valor no puede ser contrastada, ya que por diseño, las frecuencias esperadas tendrán el mismo valor p que las observadas. Por lo tanto, las hipótesis se contrastan si las muestras son al azar y los sucesos independientes.

En la tabla 4.2 vemos que las frecuencias agrupadas tienen un exceso de observaciones en los extremos de la distribución de frecuencias, y consecuentemente un déficit de observaciones en el centro. Tal distribución se llama también *contagiosa*. Debe recordarse que el número total de elementos debe ser el mismo en ambas frecuencias, observadas y esperadas, a fin de hacerlas comparables. En la distribución de frecuencias en repulsión hay más observaciones de las esperadas en el centro de la distribución y menos en las colas. Estas discrepancias se ponen de manifiesto en las columnas (4) y (6) de la tabla 4.2, en la cual las desviaciones de las frecuencias observadas respecto de las esperadas aparecen como signos más o menos.

¿Qué implican estos fenómenos? En las frecuencias agrupadas, hay más muestras completamente (o en gran parte) infectadas, e igualmente más muestras completamente (o en gran parte) no infectadas, de las que se esperaría si las probabilidades de infección fuesen independientes. Esto podría deberse a mal diseño del muestreo. Por ejemplo, si el investigador al elegir sus muestras de cinco insectos tendiera siempre a escogerlos iguales, esto es, infectados o no infectados, probablemente aparecería tal resultado. Pero si el diseño del muestreo es bueno, los resultados son más interesantes. El agrupamiento significaría pues que las muestras de cinco están de algún modo relacionadas; así, si un insecto está infectado, es más probable que otros de la misma muestra lo estén. Esto podría ser cierto si procediesen de lugares adyacentes en una situación en que los vecinos fácilmente se infectan. O podrían ser miembros de la misma familia expuestos al mismo tiempo a la infección. O posiblemente la infección podría extenderse entre miembros de una muestra durante el tiempo que media desde que los insectos se recogen hasta que se examinan.

El fenómeno opuesto, repulsión, es más difícil de interpretar biológicamente. En tal distribución hay menos grupos homogéneos y más mezclados. Esto implica la idea de un fenómeno compensatorio: si se infectan algunos de los insectos de una muestra, es menos probable que se infecten los restantes. Tal situación podría presentarse lógicamente si los insectos infectados de la muestra pudieran de algún modo transmitir inmunidad a sus compañeros, pero es biológicamente improbable. Una interpretación más razonable de tal resultado es que para cada unidad de muestreo hubiese solamente un número limitado de patógenos disponibles, y una vez infectados algunos de los insectos, los restantes quedan libres de infección, simplemente porque no hay más agentes infecciosos. Esta situación es improbable en infecciones microbianas, pero en situaciones en que un número limitado de parásitos entra en el cuerpo del huésped, la repulsión puede ser más razonable.

A partir de las frecuencias esperadas y observadas de la tabla 4.1, podemos calcular la media y desviación típica del número de insectos infectados por muestra. Estos valores se dan al final de las columnas (5), (6) y (7) en la tabla 4.1. Observamos que en las columnas (5) y (6) las medias y desviaciones típicas son casi idénticas y sólo difieren trivialmente por errores de redondeo. En la columna (7), al ser una muestra de una población cuyos parámetros son los mismos que los de la distribución de frecuencias esperadas de las columnas (5) o (6), difieren algo. La media es ligeramente menor y la desviación típica ligeramente mayor que en las frecuencias esperadas. Si deseamos conocer la media y

desviación típica de distribuciones de frecuencias binomiales esperadas, no necesitamos realizar los cálculos presentados en la tabla 4.1. La media y desviación típica de una distribución binomial son respectivamente,

$$\mu = kp, \qquad \sigma = \sqrt{kpq}$$

Sustituyendo los valores $k = 5$, $p = 0{,}4$ y $q = 0{,}6$ del ejemplo anterior, obtenemos $\mu = 2{,}0$ y $\sigma = 1{,}09545$, que son idénticos a los valores calculados en la columna (5) de la tabla 4.1. Nótese que aquí utilizamos la notación paramétrica griega porque μ y σ son parámetros de una distribución de frecuencias esperadas y no estadísticos de muestreo como son la media y desviación típica de la columna (7). Las proporciones p y q son también valores paramétricos y en lenguaje estricto deberían distinguirse de las proporciones de muestreo. De hecho, en capítulos posteriores se utilizan $\hat{p}$ y $\hat{q}$ para proporciones paramétricas (en lugar de π, que convencionalmente se utiliza como la razón de la circunferencia respecto del diámetro de un círculo). Aquí, no obstante, preferimos conservar nuestra notación sencilla. Es interesante considerar las desviaciones típicas de las distribuciones de frecuencias en agrupamiento y en repulsión de la tabla 4.2. Observamos que la distribución en agrupamiento tiene una desviación típica mayor que la esperada y la de repulsión menor que la esperada. La comparación de desviaciones típicas de muestreo con sus valores esperados, es una medida de dispersión útil en tales ejemplos. Si deseamos expresar nuestra variable aleatoria como una proporción en vez de como un recuento, es decir, indicar incidencia media de infección en los insectos como 0,4, en lugar de 2 por muestra de 5, podemos utilizar otras fórmulas para la media y desviación típica en una distribución binomial:

$$\mu = p, \qquad \sigma = \sqrt{pq/k}$$

Ahora vamos a utilizar la distribución binomial para resolver un problema biológico. Basándonos en nuestro conocimiento de la citología y biología de la especie A, esperamos que la proporción de sexos entre sus descendientes sea 1 : 1. El estudio de una camada natural revela que de 17 crías, 14 fueron hembras y 3 machos. ¿Qué conclusiones podemos sacar de esta evidencia? Suponiendo que $p\,♀$ (probabilidad de que una cría sea hembra) sea 0,5 y que esta probabilidad sea independiente entre los miembros de la muestra, la distribución de probabilidad adecuada es la binomial para un tamaño de muestra $k = 17$. El desarrollo del binomio a la potencia 17 es una tarea terrible, que como veremos, afortunadamente no es necesario realizar en su totalidad. No obstante, debemos conocer los coeficientes binomiales que pueden obtenerse a partir de un desarrollo del triángulo de Pascal (bastante tedioso a menos que una vez obtenido se conserve para futura utilización), o bien calculando las frecuencias esperadas para cualquier clase de Y a partir de la fórmula general para cualquier término de la distribución binomial.

$$C(k, Y)p^{k-Y}q^{Y} \tag{4.1}$$

La expresión $C(k, Y)$ representa el número de combinaciones que pueden formarse a partir de k elementos tomados Y a Y. El valor numérico de esta expresión puede calcular-

se por la fórmula $k!/[Y!(k - Y)!]$, en que ! significa "factorial". En matemáticas factorial de k es el producto de todos los números enteros desde 1 hasta k incluido. Así, $5! = 1 \times 2 \times 3 \times 4 \times 5 = 120$. Por convenio, $0! = 1$. Nótese que al resolver fracciones que contienen factoriales, cualquier factorial se simplificará siempre frente a otro superior. Así, $5!/3! = (5 \times 4 \times 3!)/3! = 5 \times 4$. Por ejemplo, el coeficiente binomial para la frecuencia esperada en muestras de 5 items que contienen 2 insectos infectados es $C(5, 2) = 5!/2!3! = (5 \times 4)/2 = 10$.

TABLA 4.3

Algunas frecuencias esperadas de machos y hembras para muestras de 17 crías, suponiendo que la proporción de sexos sea 1 : 1 [$p♀ = 0{,}5, q♂ = 0{,}5; (p♀ + q♀)^k = (0{,}5 + 0{,}5)^{17}$].

(1) ♀♀	(2) ♂♂	(3) $p♀$	(4) $q♂$	(5) *Coeficientes binomiales*	(6) *Frecuencias relativas esperadas* $\hat{f}_{rel}$	
17	—	0,000 007 63	1	1	0,000 007 63	
16	1	0,000 015 26	,5	17	0,000 129 71	0,006 363 42
15	2	0,000 030 52	,25	136	0,001 037 68	
14	3	0,000 061 04	,125	680	0,005 188 40	
13	4	0,000 122 07	,0625	2380	0,018 157 91	

En la tabla 4.3 se presenta la realización del ejemplo. Se calculan potencias decrecientes de $p♀$ desde $p^{17}_{♀}$ hacia abajo, y potencias crecientes de $q♂$ (desde q^0 hasta q^4). Nótese que para los fines de nuestro problema no necesitamos continuar después del término correspondiente a 13 hembras y 4 machos. Al calcular las frecuencias esperadas relativas en la columna (6), observamos que la probabilidad de 14 hembras y 3 machos es 0,005 188 40, un valor muy pequeño. Si a este valor le sumamos todos los resultados "peores", es decir, todos los que son aún más improbables que 14 hembras y 3 machos basándonos en la hipótesis 1 : 1, obtenemos una probabilidad de 0,006 363 42, un valor todavía muy pequeño. En estadística, es una práctica corriente calcular la probabilidad de una desviación tan grande o mayor que un determinado valor. Basándose en estos resultados, son improbables una o más de las siguientes hipótesis: 1) que hayamos muestreado al azar en el sentido de obtener una muestra no sesgada; 2) que la verdadera proporción de sexos en la especie A sea 1 : 1; o 3) que los sexos de los descendientes sean independientes.

La falta de independencia de sucesos puede significar que aún cuando la proporción media de sexos sea 1 : 1, los miembros de cada familia o camada son fundamentalmente unisexuales, de modo que los descendientes de una determinada pareja tiendan a ser todos (o en gran parte) hembras, o bien todos (o en gran parte) machos. Para confirmar esta hipótesis necesitaríamos tener más muestras y además examinar la distribución de muestras por agrupamiento, que indicaría una tendencia hacia familias unisexuales.

Debemos ser muy precisos sobre las preguntas que hacemos acerca de nuestros datos. Realmente hay dos preguntas que podemos hacer sobre la proporción de sexos. Una, ¿son suficientemente desiguales las frecuencias de sexos como para que las hembras aparezcan más a menudo que los machos? Dos, ¿son desiguales las frecuencias de sexos? Podemos interesarnos solamente por la primera de estas cuestiones, ya que sabemos por experiencia que en este grupo particular de organismos los machos nunca son más frecuentes que las hembras. En tal caso, es apropiado el razonamiento seguido anteriormente. Sin embargo, si sabemos muy poco sobre este grupo de organismos y si nuestra pregunta es simplemente si la proporción de sexos de los descendientes es distinta, entonces tenemos que considerar ambos extremos de la distribución binomial; las desviaciones de la proporción 1 : 1 podrían ocurrir en ambas direcciones. En tal caso deberíamos considerar no solamente las probabilidades de muestras con 14 hembras y 3 machos (y todos los casos peores) sino también la probabilidad de muestras de 14 machos y 3 hembras (y todos los casos peores en esa dirección). Ya que esta distribución de probabilidad es simétrica (porque $p\,♀ = q\,♂ = 0{,}5$), simplemente duplicamos la probabilidad acumulativa de 0,006 363 42 obtenida previamente, lo que da por resultado 0,012 726 84. Este nuevo valor es todavía muy pequeño, lo que hace bastante improbable que la verdadera proporción de sexos sea 1 : 1. Esta es la primera experiencia con una de las aplicaciones más importantes de la estadística, el contraste de hipótesis. Una introducción formal a este campo se aplazará hasta la sección 6.8. Aquí podemos simplemente señalar que los dos planteamientos seguidos anteriormente se conocen como *prueba de una cola* y *prueba de dos colas,* respectivamente. A veces los estudiantes tienen dificultad para saber cual de los dos aplicar. En ejemplos posteriores trataremos de indicar en cada caso por qué se utiliza una prueba de una cola o una de dos colas.

Hemos dicho que una tendencia hacia familias unisexuales daría como resultado una distribución de frecuencias observadas en agrupamiento. Un caso real de esta índole es uno clásico en la literatura, los datos de proporción de sexos obtenidos por Geissler (1889) de los archivos de un hospital de Sajonia. La tabla 4.4 reproduce las proporciones de sexos en 6 115 familias de 12 hijos cada una procedente del más amplio estudio realizado por Geissler. Todas las columnas de la tabla deberían resultar ya familiares. Las frecuencias esperadas no se han calculado basándose en una hipótesis 1 : 1 puesto que se sabe que en poblaciones humanas la proporción de sexos al nacer no es 1 : 1. Como la proporción de sexos varía en diferentes poblaciones humanas, la mejor estima de ésta para la población de Sajonia se ha obtenido sencillamente utilizando la proporción media de varones en estos datos. Esta puede obtenerse calculando el promedio del número de varones por familia ($\bar{Y} = 6{,}23058$) para las 6 115 familias y convirtiéndolo en proporción. Este valor resulta ser 0,519 215. Por consiguiente, la proporción de hembras es de 0,480 785. En las desviaciones de las frecuencias observadas respecto de las esperadas, que se presentan en la columna (9) de la tabla 4.4, observamos considerable agrupamiento. Hay muchos más ejemplos de familias con todos los hijos varones o todos hembras (o casi todos) de los que indicarían probabilidades independientes. La base genética de esto no está clara, pero es evidente que hay ciertas familias que "tienden a tener niñas" y otras que "tienden a tener niños". También puede observarse evidencia de agrupamiento por el hecho de que s^2 es mucho mayor de lo que cabría esperar basándose en la distribución binomial [$\sigma^2 = kpq = 12 \times 0{,}519215 \times 0{,}480785 = 2{,}99557$].

Hay un contraste claro entre los datos de la tabla 4.1 y los de la tabla 4.4. En los datos de infección de insectos de la tabla 4.1 teníamos una proporción hipotética de infección basada en conocimientos exteriores. En los de proporción de sexos de la tabla 4.4 no teníamos tales conocimientos, utilizábamos un *valor empírico de p obtenido de los datos,* en lugar de un *valor hipotético exterior a los datos.* Esta es una diferencia cuya importancia se verá más adelante. En los datos de proporción de sexos de la tabla 4.3, como en gran cantidad de trabajos de genética mendeliana, se utiliza un valor hipotético de *p.*

TABLA 4.4

Proporción de sexos en 6 115 familias de doce hijos en Sajonia.

(1) ♂♂ Y	(2) ♀♀	(3) p♂	(4) q♀	(5) *Coeficientes binomiales*	(6) *Frecuencias relativas esperadas* $\hat{f}_{rel}$	(7) *Frecuencias absolutas esperadas* $\hat{f}$	(8) *Frecuencias observadas* f	(9) *Desviación de lo esperado* $(f - \hat{f})$
12	—	0,000 384	1	1	0,000 384	2,3	7	+
11	1	0,000 739	0,480 785	12	0,004 264	26,1	45	+
10	2	0,001 424	0,231 154	66	0,021 725	132,8	181	+
9	3	0,002 742	0,111 135	220	0,067 041	410,0	478	+
8	4	0,005 282	0,053 432	495	0,139 703	854,3	829	−
7	5	0,010 173	0,025 689	792	0,206 973	1265,6	1112	−
6	6	0,019 592	0,012 351	924	0,223 590	1367,3	1343	−
5	7	0,037 734	0,005 938	792	0,177 459	1085,2	1033	−
4	8	0,072 676	0,002 855	495	0,102 708	628,1	670	+
3	9	0,139 972	0,001 373	220	0,042 280	258,5	286	+
2	10	0,269 584	0,000 660	66	0,011 743	71,8	104	+
1	11	0,519 215	0,000 317	12	0,001 975	12,1	24	+
—	12	1	0,000 153	1	0,000 153	0,9	3	+
Total					0,999 998	6115,0	6115	
			$\bar{Y} = 6{,}23058$		$s^2 = 3{,}48985$			

Fuente: Geissler (1889).

4.3 La distribución de Poisson

En la aplicación típica de la binomial teníamos muestras relativamente pequeñas (2 estudiantes, 5 insectos, 17 crías, 12 hermanos) en las cuales aparecían dos estados alternativos con frecuencias variables (americano y extranjero, infectado y no infectado, macho y hembra). Sin embargo, muy frecuentemente estudiamos casos en los cuales el tamaño k de la muestra es muy grande. Estos presentarían un problema de cálculo considerable.

Hemos visto que el desarrollo del binomio $(p + q)^k$ es bastante tedioso cuando k es grande. Imagínese que se debe desarrollar la expresión $(0{,}001 + 0{,}999)^{1000}$. Obsérvese que en esta expresión no solamente es grande el tamaño de la muestra sino que además uno de los sucesos (representado por la probabilidad q) es mucho más frecuente que el otro (representado por la probabilidad p). El desarrollo de este binomio por los métodos aprendidos hasta ahora requiere una tabla de logaritmos muy exacta con 10 cifras decimales. Sin embargo, estas expresiones se encuentran corrientemente, y a veces son de gran importancia biológica. En tales casos, generalmente nos interesa sólo un extremo de la distribución. Este es el representado por los términos

$$p^0q^k,\ C(k,1)p^1q^{k-1},\ C(k,2)p^2q^{k-2},\ C(k,3)p^3q^{k-3},\ \ldots$$

El primer término representa 0 sucesos raros y k sucesos frecuentes en una muestra de k sucesos, el segundo término representa un suceso raro y $k - 1$ sucesos frecuentes, el tercero 2 sucesos raros y $k - 2$ frecuentes, y así sucesivamente. Las expresiones de la forma $C(k, i)$ son los coeficientes binomiales, representados por los términos combinatorios discutidos en la sección anterior. Aunque esta expresión permitiría el cálculo del extremo de la curva deseado, no obstante, sería un procedimiento muy complicado en vista de la magnitud de k. Para convencerse de esto, el lector podría intentar calcular uno o más términos en $(0{,}001 + 0{,}999)^{1000}$. Afortunadamente es mucho más fácil calcular otra distribución, la distribución de Poisson, que se aproxima estrictamente a los resultados deseados.

La *distribución de Poisson* es también una distribución de frecuencias discreta del número de veces que ocurre un suceso raro. Pero en contraste con la distribución binomial, el número de veces que un suceso no ocurre es infinitamente grande. Con miras a nuestro tratamiento actual se estudiará una variable de Poisson en una muestra espacial o en una temporal. Un ejemplo de la primera sería el número de plantas de musgo en un cuadrado de muestreo de una ladera, o el número de parásitos en un individuo huésped; un ejemplo de muestra temporal es el número de mutaciones que ocurren en un carácter genético en el período de un mes, o los casos de gripe registrados en una semana. La variable de Poisson, Y, será el número de sucesos por muestra. Puede tomar valores discretos desde 0 hacia arriba. Para seguir la distribución de Poisson, la variable debe tener dos propiedades: 1) Su media debe ser pequeña con respecto al máximo número posible de sucesos por unidad de muestreo. Así pues, el suceso debería ser "raro". Por ejemplo, un cuadrado en el que se cuentan plantas de musgo debería ser lo suficientemente grande para que un número sustancial de plantas de musgo pudieran hallarse allí si las condiciones biológicas fuesen favorables al desarrollo de numerosas plantas de musgo en el cuadrado. Un cuadrado de 1 cm^2 sería demasiado pequeño para que los musgos se distribuyesen según la ley de Poisson. Igualmente, un lapso de tiempo de 1 minuto sería irrealista para registrar nuevos casos de gripe en una ciudad, pero en una semana podrían presentarse un gran número de estos casos. 2) La incidencia del suceso debe ser independiente de incidencias previas dentro de la unidad de muestreo. Así, la presencia de una planta de musgo en un cuadrado no debe aumentar ni disminuir la probabilidad de que se desarrollen otras plantas de musgo en el cuadrado. Del mismo modo, el hecho de que se haya registrado un caso de gripe no debe afectar a la probabilidad de que se registren subsiguientes casos de

gripe. Los sucesos que reúnen estas condiciones ("sucesos raros y aleatorios") deberían distribuirse en forma de Poisson.

El propósito de adaptar una distribución de Poisson a varios sucesos raros en la naturaleza es probar si los sucesos ocurren con independencia mutua. Si es así, seguirán la distribución de Poisson. Si la ocurrencia de un suceso aumenta la probabilidad de un segundo suceso semejante, obtenemos una distribución en agrupamiento o contagiosa. Si la ocurrencia de un suceso impide la de un segundo suceso semejante en la unidad de muestreo, obtenemos una distribución en repulsión o espacialmente uniforme. La distribución de Poisson puede utilizarse como prueba de aleatoriedad o independencia de distribuciones no sólo con respecto al espacio sino también al tiempo.

Esta distribución debe su nombre al matemático francés Poisson, que la describió en 1837. Es una serie infinita cuyos términos se suman hasta alcanzar un valor igual a uno (como debe ser para cualquier distribución de probabilidad). La serie puede representarse como

$$\frac{1}{e^{\mu}},\ \frac{\mu}{1!e^{\mu}},\ \frac{\mu^2}{2!e^{\mu}},\ \frac{\mu^3}{3!e^{\mu}},\ \frac{\mu^4}{4!e^{\mu}},\ \ldots,\ \frac{\mu^r}{r!e^{\mu}},\ \ldots \tag{4.2}$$

que son las frecuencias relativas esperadas correspondientes a los siguientes recuentos del suceso raro Y:

$$0,\ 1,\ 2,\ 3,\ 4,\ \ldots,\ r,\ \ldots$$

Así, el primero de estos términos representa la frecuencia relativa esperada de muestras que no contienen suceso raro; el segundo término, un suceso raro; el tercero, dos sucesos raros; y así sucesivamente. El denominador de cada término contiene e^{μ}, donde e es la base de los logaritmos naturales o neperianos, una constante cuyo valor con cinco cifras decimales exactas es 2,71828. Reconocemos μ como la media paramétrica de la distribución; es constante para cualquier problema dado. El signo de admiración después del coeficiente en el denominador significa "factorial" y se ha explicado en la sección anterior.

Una forma de aprender más acerca de la distribución de Poisson es aplicarla a un caso real. En la parte superior del cuadro 4.1 se encuentra un resultado bien conocido desde la literatura estadística remota. Examina la distribución de células de levadura en 400 cuadrados de un hemocitómetro, una cámara de recuentos como la que se utiliza para hacer recuentos de células sanguíneas y otras estructuras microscópicas suspendidas en líquido. La columa (1) registra el número de células de levadura observadas en cada cuadrado y la columna (2) da las frecuencias observadas, el número de cuadrados que contienen un determinado número de células de levadura. Observamos que 75 cuadrados no contienen células de levadura, pero la mayoría de los cuadrados contienen 1 o 2 células. Solamente 17 cuadrados contienen 5 o más células de levadura.

¿Por qué deberíamos esperar que esta distribución de frecuencias se distribuyese en forma de Poisson? Tenemos aquí un suceso relativamente raro, la frecuencia de células de levadura por cuadrado del hemocitómetro, cuya medida se ha calculado y es 1'8. Es decir, hay 1'8 células por cuadrado, por término medio. Con respecto a la cantidad de espacio provisto en cada cuadrado y al número de células que podrían haber llegado a asentarse en un cuadrado cualquiera, el número real encontrado es verdaderamente bajo. Pudiéra-

CUADRO 4.1

Cálculo de frecuencias de Poisson esperadas.

Células de levadura en 400 cuadrados de un hemocitómetro: $\bar{Y} = 1{,}8$ células por cuadrado; $n = 400$ cuadrados muestreados.

(1) *Número de células por cuadrado* Y	(2) *Frecuencias observadas* f		(3) *Frecuencias absolutas esperadas* $\hat{f}$		(4) *Desviaciones de lo esperado* $f - \hat{f}$	
0	75		66,1		+	
1	103		119,0		−	
2	121		107,1		+	
3	54		64,3		−	
4	30		28,9		+	
5	13		10,4		+	
6	2		3,1		−	
7	1	17	0,8	14,5	+	+
8	0		0,2		−	
9	1		0,0		+	
	400		399,9			

Fuente: "Student" (1907).

Etapas del cálculo

El cálculo se basa en la expresión (4.3) multiplicada por n, ya que deseamos obtener frecuencias absolutas esperadas, $\hat{f}$.

1. Calcular $e^{\bar{Y}}$.

$$\log e^{\bar{Y}} = \bar{Y} \log e = \bar{Y}(0{,}43429) = (1{,}8)(0{,}43429)$$
$$= 0{,}78172$$
$$\text{antilog } (0{,}78172) = 6{,}0495$$

2. $\hat{f}_0 = \dfrac{n}{e^{\bar{Y}}} = ne^{-\bar{Y}} = \dfrac{400}{6{,}0495} = 400(0{,}16530) = 66{,}12$

3. $\hat{f}_1 = \left(\dfrac{n}{e^{\bar{Y}}}\right)\bar{Y} = 66{,}12(1{,}8) = 119{,}02$

4. $\hat{f}_2 = \left(\dfrac{n\bar{Y}}{e^{\bar{Y}}}\right)\dfrac{\bar{Y}}{2} = 119{,}02\left(\dfrac{1{,}8}{2}\right) = 107{,}12$

5. $\hat{f}_3 = \left(\dfrac{n\bar{Y}^2}{2e^{\bar{Y}}}\right)\dfrac{\bar{Y}}{3} = 107{,}12\left(\dfrac{1{,}8}{3}\right) = 64{,}27$

CUADRO 4.1 (continuación)

6. $\hat{f}_4 = \left(\dfrac{n\bar{Y}^3}{2 \times 3e^{\bar{Y}}}\right)\dfrac{\bar{Y}}{4} = 64{,}27\left(\dfrac{1{,}8}{4}\right) = 28{,}92$

7. $\hat{f}_5 = \left(\dfrac{n\bar{Y}^4}{2 \times 3 \times 4e^{\bar{Y}}}\right)\dfrac{\bar{Y}}{5} = 28{,}92\left(\dfrac{1{,}8}{5}\right) = 10{,}41$

8. $\hat{f}_6 = \left(\dfrac{n\bar{Y}^5}{2 \times 3 \times 4 \times 5e^{\bar{Y}}}\right)\dfrac{\bar{Y}}{6} = 10{,}41\left(\dfrac{1{,}8}{6}\right) = 3{,}12$

9. $\hat{f}_7 = \left(\dfrac{n\bar{Y}^6}{2 \times 3 \times 4 \times 5 \times 6e^{\bar{Y}}}\right)\dfrac{\bar{Y}}{7} = 3{,}12\left(\dfrac{1{,}8}{7}\right) = 0{,}80$

10. $\hat{f}_8 = \left(\dfrac{n\bar{Y}^7}{2 \times 3 \times 4 \times 5 \times 6 \times 7e^{\bar{Y}}}\right)\dfrac{\bar{Y}}{8} = 0{,}80\left(\dfrac{1{,}8}{8}\right) = 0{,}18$

Total	399,96
$\hat{f}_9$ y superiores	0,04

mos esperar que la presencia de determinadas células de levadura en un cuadrado fuese independiente de la presencia de otras células de levadura. Este es un tipo de aplicación de la distribución de Poisson que se encuentra corrientemente.

La única cantidad que necesitamos conocer para calcular las frecuencias relativas esperadas de una distribución de Poisson es la media del suceso raro. Puesto que no conocemos la media paramétrica de las células de levadura en este problema, utilizamos un estimador (la media de la muestra) y calculamos las frecuencias esperadas de una distribución de Poisson cuya μ es igual a la media de la muestra de la distribución de frecuencias observadas del cuadro 4.1. Para efectos de cálculo es conveniente volver a escribir la expresión (4.2) como

$$\frac{1}{e^{\bar{Y}}}, \quad \frac{1}{e^{\bar{Y}}}\left(\frac{\bar{Y}}{1}\right), \quad \frac{\bar{Y}}{e^{\bar{Y}}}\left(\frac{\bar{Y}}{2}\right), \quad \frac{\bar{Y}^2}{2e^{\bar{Y}}}\left(\frac{\bar{Y}}{3}\right), \quad \frac{\bar{Y}^3}{2 \times 3e^{\bar{Y}}}\left(\frac{\bar{Y}}{4}\right), \ldots . \tag{4.3}$$

Obsérvese en primer lugar que la media paramétrica μ se ha sustituido por la media de la muestra $\bar{Y}$. Cada término es exactamente el mismo matemáticamente que su correspondiente término en la expresión (4.2), pero se ha descompuesto en factores de forma apropiada para el cálculo. Después del primer término de la expresión (4.2), todos los siguientes constan del término anterior multiplicado por la media sobre un entero que aumenta en 1 para cada término subsiguiente. Así, sólo necesitamos calcular una vez la expresión $1/e^{\bar{Y}}$ para obtener la frecuencia del primer término, multiplicar ésta por $\bar{Y}/1$ para tener el segundo término, multiplicar el segundo por $\bar{Y}/2$ para el tercero, y así sucesivamente. Es importante no cometer error de cálculo, ya que en esta multiplicación en cadena la exactitud de cada término depende de la exactitud del término anterior a él.

La expresión (4.3) da las frecuencias relativas esperadas. Si como es más usual, se desean las frecuencias absolutas esperadas, basta multiplicar el primer término por n, el número de muestras, y luego continuar con los pasos computacionales como antes. Por este proceso de multiplicación en cadena n continúa como factor en cada término. En el cuadro 4.1 se representa el cálculo efectivo, y las frecuencias esperadas así obtenidas se alistan en la columna (3).

¿Qué hemos averiguado en este cálculo? Cuando comparamos las frecuencias observadas con las esperadas, observamos efectivamente un buen ajuste de nuestras frecuencias observadas a una distribución de Poisson de media 1,8, aunque todavía no hemos enseñado un test estadístico de bondad de ajuste (capítulo 13). No aparece un patrón claro de desviaciones de lo esperado. No podemos contrastar una hipótesis sobre la media porque la media de la distribución esperada se ha deducido de la media de la muestra de las variantes observadas. Como en la distribución binomial, agrupamiento o agregación significaría que la probabilidad de que una segunda célula de levadura se encontrara en un cuadrado no es independiente de la presencia de la primera, sino que es superior. Esto daría como resultado un agrupamiento de los ítems en las clases de los extremos de la distribución, de modo que habría algunos cuadrados con gran número de células.

La interpretación biológica del patrón de dispersión varía con el problema. Las células de levadura parecen estar aleatoriamente distribuidas en la cámara de recuento, lo que indica mezcla completa de la suspensión. Sin embargo, a no ser que se utilice el líquido de suspensión adecuado, las células rojas de la sangre con frecuencia se mantendrían juntas debido a una carga eléctrica. Este efecto denominado rouleaux estaría indicado por agrupamiento de las frecuencias observadas.

Obsérvese que en el cuadro 4.1, como en las tablas subsiguientes que dan ejemplos de aplicación de la distribución de Poisson, agrupamos las frecuencias bajas en un extremo de la curva, uniéndolas por medio de un corchete. Esto tiende a simplificar un poco los patrones de distribución. Sin embargo, el motivo principal de este agrupamiento está relacionado con el test G para bondad de ajuste (de frecuencias observadas a esperadas), que se discute en la sección 13.2. Para los fines de este test, ninguna frecuencia esperada $\hat{f}$ debería ser menor que 5.

Antes de pasar a otros ejemplos necesitamos mostrar algunos datos más sobre la distribución de Poisson. Probablemente se observará que para calcular las frecuencias esperadas sólo necesitamos conocer un parámetro, la media de la distribución. En cambio, en la binomial necesitamos dos parámetros, p y k. Así, la media define completamente la forma de una distribución de Poisson determinada. De esto se sigue que la varianza es una función de la media, y de hecho en una distribución de Poisson tenemos una interrelación muy simple entre las dos: $\mu = \sigma^2$, siendo la varianza igual a la media. La varianza del número de células de levadura por cuadrado basada en las frecuencias observadas en el cuadro 4.1 es igual a 1,965, no muy superior a la media de 1,8, indicando nuevamente que las células de levadura se distribuyen en forma de Poisson y por tanto aleatoriamente. Esta interrelación entre varianza y media sugiere una comprobación rápida de si una distribución de frecuencias observadas se distribuye en forma de Poisson incluso sin ajuste de las frecuencias esperadas a los datos. Calculamos simplemente un *coeficiente de dispersión*.

$$C.D. = \frac{s^2}{\bar{Y}}$$

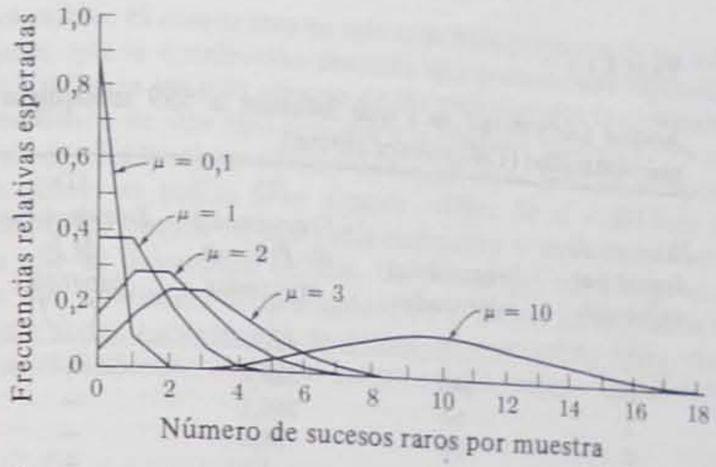


Fig. 4.3. Polígonos de frecuencias de la distribución de Poisson para diversos valores de la media.

Este valor será próximo a 1 en las distribuciones que son esencialmente de Poisson, será mayor que 1 en muestras agrupadas, y menor que 1 en casos de repulsión. En el ejemplo de las células de levadura, $C.D. = 1{,}092$.

En la figura 4.3 se presentan los patrones de cinco distribuciones de Poisson de medias diferentes como polígonos de frecuencias (línea que une los puntos medios de un diagrama de barras). Observamos que para el valor inferior de $\mu = 0{,}1$ el polígono de frecuencias es extremo en forma de J, pero al aumentar el valor de μ las distribuciones se hacen gibosas y finalmente casi simétricas.

Concluimos nuestro estudio de la distribución de Poisson con una consideración de dos ejemplos. El primer ejemplo muestra la distribución de un ácaro de agua en adultos de un mosquito quironómido (tabla 4.5). Este ejemplo es similar a la distribución de células de levadura, salvo que aquí la unidad de muestreo es un mosquito en lugar de un cuadrado del hemocitómetro. El suceso raro es la infestación de un mosquito por un ácaro. El coeficiente de dispersión es 2,225 y éste es reflejado claramente por las frecuencias observadas, que son mayores que las esperadas en los extremos y menores que las esperadas en el centro. Esta interrelación se ve fácilmente en la última columna de la distribución de frecuencias, que muestra los signos de las desviaciones (frecuencias observadas menos esperadas), y presenta un patrón de agrupamiento característico. Una posible explicación es que la densidad de ácaros en las diferentes charcas de las que emergieron los mosquitos quironómidos difiere considerablemente. Los quironómidos que salieron de charcas con muchos ácaros estarían parasitados por más de un ácaro, pero los de charcas en que eran escasos presentarían poca o ninguna infestación.

La segunda distribución (tabla 4.6) se ha obtenido de un estudio experimental de los efectos de diferentes densidades de padres del gorgojo de las judías Azuki. Estudiamos el número de gorgojos que salen por judía. Las larvas de estos gorgojos entran en las judías, dentro de ellas se nutren y se transforman en pupas, y luego salen por un orificio de eclosión. Así, el número de orificios por judía es una buena medida del número de

TABLA 4.5

Acaros (*Arrenurus sp.*) que infestan a 589 mosquitos quironómidos (*Calopsectra akrina*).

(1) *Número de ácaros por mosquito* Y	(2) *Frecuencias observadas* f		(3) *Frecuencias de Poisson esperadas* $\hat{f}$		(4) *Desviaciones de lo esperado* $f - \hat{f}$	
0	442		380,7		+	
1	91		166,1		−	
2	29		36,2		−	
3	14		5,3		+	
4	4		0,6		+	
5	6	27	0,1	5,7	+	+
6	2		0,0		+	
7	0		0,0		0	
8	1		0,0		+	
Total	589		589,0			

$\overline{Y} = 0{,}4363$ $\quad s^2 = 0{,}9709$ $\quad C.D. = 2{,}225$

Fuente: Datos de F. J. Rohlf.

TABLA 4.6

Gorgojos de las judías Azukis (*Callosobruchus chinensis*) que emergen de 112 judías azukis (*Phaseolus radiatus*).

(1) *Número de gorgojos que emergen por judía* Y	(2) *Frecuencias observadas* f		(3) *Frecuencias de Poission esperadas* $\hat{f}$		(4) *Desviación de lo esperado* $f - \hat{f}$	
0	61		70,4		−	
1	50		32,7		+	
2	1		7,6		−	
3	0	1	1,2	8,9	−	−
4	0		0,1		−	
Total	112		112,0			

$\overline{Y} = 0.4643$ $\quad s^2 = 0{,}269$ $\quad C.D. = 0{,}579$

Fuente: Utida (1943).

adultos que han salido. El suceso raro en este caso es la presencia de un solo gorgojo en la judía. Observamos que la distribución presenta una pronunciada repulsión. Hay muchas más judías que contienen un solo gorgojo de las previstas por la distribución de Poisson. Un hallazgo estadístico de este tipo conduce a la cuestión biológica "¿por qué?". En este caso se encontró que las hembras adultas de los gorgojos tendían a depositar sus huevos uniformemente sobre las judías a su alcance en vez de al azar. Esto impedía que se depositasen demasiados huevos en una judía cualquiera y excluía la fuerte competencia entre las larvas que se desarrollan en ella. Un factor que contribuye es la competencia entre las larvas que continúan alimentándose en la misma judía, dando como resultado que generalmente todas excepto una se destruyan o se echen fuera. Así se comprende fácilmente cómo el fenómeno biológico anterior daría origen a una distribución en repulsión.

Ejercicios 4

4.1 En el hombre la proporción de sexos en niños recién nacidos es aproximadamente 100 ♀♀: 105 ♂♂. Si extraemos 10 000 muestras al azar de 6 niños recién nacidos de la población total de tales niños durante un año, ¿cuál sería la frecuencia esperada de grupos de 6 varones, 5 varones, 4 varones, y así sucesivamente?

4.2 Las dos columnas que siguen dan la fertilidad de los huevos de la cepa CP de *Drosophila melanogaster* producida en 100 frascos de 10 huevos cada uno (datos de R.R. Sokal). Hallar las frecuencias esperadas suponiendo independencia de mortalidad para cada huevo de un frasco. Utilizar la media observada. Calcular la varianza esperada y compararla con la observada. Interpretar los resultados, sabiendo que los huevos de cada frasco son de la misma pareja y que los diferentes frascos contienen descendientes de diferentes parejas. SOLUCION: $\sigma^2 = 2{,}417$, $s^2 = 6{,}636$.

Número de huevos incubados Y	*Número de frascos* f
0	1
1	3
2	8
3	10
4	6
5	15
6	14
7	12
8	13
9	9
10	9

4.3 Calcular las frecuencias de Poisson esperadas para la distribución de frecuencias dada en la tabla 2.2 (número de plantas de la especie *Carex flacca* encontradas en 500 cuadrados).

4.4 El cuerpo médico del ejército está interesado sobre la enfermedad intestinal X. Por experiencia previa saben que los soldados que sufren la enfermedad contienen invariablemente el organismo patógeno en sus heces, y que para todos sus fines prácticos, cada muestra de heces del enfermo contiene estos organismos. Sin embargo, los organismos nunca son abundantes y así solamente el 20 % de todas las preparaciones hechas por el procedimiento estándar contendrán alguno (suponemos que si un organismo está presente en un portaobjetos se verá). ¿Cuántos portaobjetos por muestra de heces deberían pedir a los técnicos de laboratorio que preparasen y examinasen, para que en caso de que una muestra fuera positiva fuera falsamente diagnosticada como negativa en menos del 1 % de los casos (por término medio)? Basándose en la respuesta, ¿se recomendaría que tratasen de mejorar sus métodos de diagnóstico? SOLUCION: 21 portaobjetos.

4.5 Se hace un cruce en un experimento genético con Drosophila en el cual se espera que ¼ de la progenie tendrá ojos blancos y ½ tendrá el rasgo denominado "quetas chamuscadas". Supongamos que los dos loci genéticos se segregan independientemente. a) ¿Qué proporción de la progenia exhibiría ambos rasgos simultáneamente? b) Si se extraen 4 moscas al azar, ¿cuál es la probabilidad de que sean todas de ojos blancos? c) ¿Cuál es la probabilidad de que ninguna de las cuatro moscas tenga ojos blancos ni "quetas chamuscadas"? d) Si se extraen dos moscas ¿cuál es la probabilidad de que al menos una de ellas tenga ojos blancos o "quetas chamuscadas" o ambos rasgos?

4.6 Los lectores que han recibido uno o dos semestres de cálculo puede que deseen tratar de demostrar que la expresión (4.1) tiende a la expresión (4.2) cuando k tiende a infinito (y p se hace infinitesimal, para que $\mu = kp$ permanezca constante).

$$\left(1 - \frac{x}{n}\right)^n \longrightarrow e^{-x} \text{ cuando } n \longrightarrow \infty.$$

4.7 Resumir y comparar las hipótesis y parámetros en que se basan las distribuciones binomial y de Poisson.

4.8 Si la frecuencia del gen A es p y la frecuencia de a es q, ¿cuáles son las frecuencias esperadas de los cigotos AA, Aa, o aa (suponiendo que un cigoto diploide representa una muestra al azar de tamaño 2)? ¿Cuál sería la frecuencia esperada para un autotetraploide (para un locus próximo al centrómero, un cigoto puede considerarse como una muestra al azar de tamaño 4)?

4.9 Una población consta de tres tipos de individuos A_1, A_2 y A_3 con frecuencias relativas de 0,5, 0,2 y 0,3, respectivamente. a) ¿Cuál es la probabilidad de obtener solamente individuos del tipo A_1 en muestras de tamaño 1,2,3, . . . , n? b) ¿Cuáles serían las probabilidades de obtener solamente individuos que no fuesen de los tipos A_1 ni A_2 en una muestra de tamaño n? c) ¿Cuál es la probabilidad de obtener una muestra que contenga al menos una representación de cada tipo en muestras de tamaño 1,2,3,4,5 . . . , n?

Capítulo 5

La distribución de probabilidad normal

Las distribuciones de frecuencias teóricas del último capítulo eran todas discretas. Sus variables tomaban valores que variaban a intervalos enteros (variables merísticas). Así, el número de insectos infectados por muestra era 0, 1 ó 2 pero no podría tomar un valor intermedio entre éstos. Igualmente, el número de células de levadura por cuadrado de hemocitómetro es una variable merística y requiere una función de probabilidad discreta para describirla. Sin embargo, la mayoría de las variables encontradas en biología son continuas (como los pesos de niños recién nacidos o las longitudes del fémur de áfidos utilizados como ejemplos en los capítulos 2 y 3). Este capítulo trata más ampliamente de las distribuciones de variables continuas.

La primera sección (5.1) introduce distribuciones de frecuencias de variables continuas. En la sección 5.2 presentamos una forma de deducir la más común de estas distribuciones, la distribución de probabilidad normal, y en la sección 5.3 examinamos sus propiedades. En la sección 5.4 se presentan algunas aplicaciones de la distribución normal. En la sección 5.5 se cita una técnica gráfica para señalar desviaciones de la normalidad y para estimar la media y desviación típica en distribuciones aproximadamente normales, así como algunas de las razones de desviación de la normalidad en distribuciones de frecuencias observadas.

5.1 Distribuciones de frecuencias de variables continuas

Para variables continuas la distribución de probabilidad teórica o *función de densidad de probabilidad* puede representarse por una curva continua, como se indica en la figura 5.1. La altura de la curva da la densidad para un valor determinado de la variable. Por *densidad* entendemos la concentración relativa de variantes a lo largo del eje Y (como se indica en la figura 2.1). Nótese que el eje Y es la abscisa en la figura, siendo la ordenada la

frecuencia f o la densidad. Para comparar la distribución de frecuencias teórica con la observada es necesario dividir las dos en sus clases correspondientes, como se indica por las líneas verticales en la figura 5.1. Las funciones de densidad de probabilidad se definen de modo que la frecuencia esperada de observaciones entre dos límites de clase (líneas verticales) esté representada por el área bajo la curva entre estos límites. Por lo tanto, el área total bajo la curva es la suma de las frecuencias esperadas (1,0 ó n, dependiendo de si se han calculado frecuencias esperadas relativas o absolutas).

Cuando se hace una distribución de frecuencias de observaciones de una variable continua, la elección de límites de clase es arbitraria porque todos los valores de una variable son teóricamente posibles. En una distribución continua no se puede evaluar que la probabilidad de la variable sea exactamente igual a un determinado valor como 3 ó 3,5. Solamente puede estimarse la frecuencia de observaciones que están entre dos límites. Esto se debe a que el área de la curva correspondiente a cualquier punto a lo largo de ella es una infinitesimal. Así, para calcular las frecuencias esperadas para una distribución continua tendremos que calcular las proporciones del área bajo la curva entre los límites de clase. En las secciones 5.3 y 5.4 veremos cómo se hace esto en la distribución de frecuencias normal.

Las distribuciones de frecuencias continuas pueden comenzar y terminar en puntos finitos a lo largo del eje Y, como se presenta en la figura 5.1, o pueden extenderse indefinidamente hacia uno o ambos extremos de la curva, como en las figuras 6.11 y 5.3. La idea de un área bajo una curva cuando uno o los dos extremos tiende a infinito puede turbar a los no instruidos en cálculo. Sin embargo, afortunadamente esto no es un gran obstáculo conceptual, ya que en todos los casos que nos encontraremos, el extremo de la curva se aproximará al eje Y lo bastante rápidamente como para que la porción del área después de un cierto punto sea, para todos los efectos prácticos, cero y las frecuencias que representa sean infinitesimales.

Podemos ajustar distribuciones de frecuencias continuas a varias series de datos merísticos como, por ejemplo, el número de dientes en un organismo. En tales casos tenemos razón para creer que las variables biológicas subyacentes que causan diferencias en los

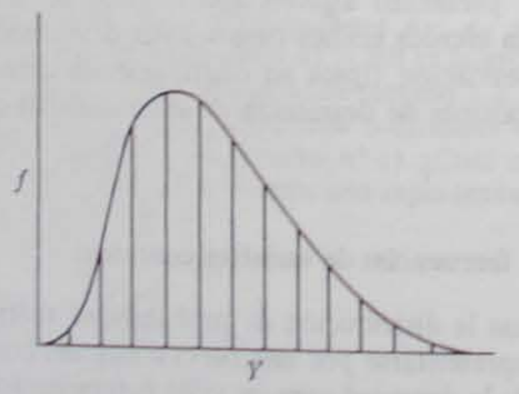


Fig. 5.1. Distribución de probabilidad de una variable continua. (Para explicación ver texto.)

números de la estructura son realmente continuas, aun cuando se expresen como una variable discreta.

Ahora pasaremos a discutir la función de densidad de probabilidad más importante en estadística, la distribución de frecuencias normal.

5.2 Deducción de la distribución normal

Hay varias maneras de deducir la distribución normal de frecuencias a partir de suposiciones elementales. La mayor parte de éstas requieren más matemáticas de las que esperamos de nuestros lectores. Por lo tanto, utilizaremos una aproximación ampliamente intuitiva que hemos encontrado de valor heurístico.

Vamos a considerar una distribución binomial de la forma habitual $(p + q)^k$ en la que k tiende a infinito. ¿Qué tipo de situación biológica podría dar origen a esta distribución binomial? Un ejemplo pudiera ser el caso en que muchos factores cooperan aditivamente en la producción de un efecto biológico. El siguiente caso hipotético posiblemente no esté demasiado alejado de la realidad. La intensidad de pigmentación de la piel en un animal será debida a la suma de muchos factores, algunos genéticos, otros ambientales. Como hipótesis simplificadora vamos a establecer que cada factor puede hallarse en dos estados solamente: presente o ausente. Cuando el factor está presente aporta una unidad de pigmentación al color de la piel, pero cuando está ausente no aporta nada a la pigmentación. Cada factor, independientemente de su naturaleza u origen, tiene idéntico efecto y los efectos son aditivos; si de los cinco posibles factores tres están presentes en un individuo, la intensidad de pigmentación sería tres unidades, siendo la suma de tres aportaciones de una unidad cada una. Una hipótesis final: cada factor tiene igual probabilidad de estar presente o ausente en un determinado individuo. Así, $p = P[F] = 0{,}5$ es la probabilidad de que el factor esté presente, mientras $q = P[f] = 0{,}5$, es la probabilidad de que el factor esté ausente.

Con un factor solo ($k = 1$), el desarrollo del binomio $(p + q)^1$ produciría dos clases de pigmentación entre los animales, es decir:

$\{F,\ \ f\ \}$	clases de pigmentación (espacio de probabilidad)
$\{0{,}5,\ \ 0{,}5\}$	frecuencias esperadas
$\{1,\ \ 0\ \}$	intensidad de pigmentación

La mitad de los animales tendría intensidad 1, la otra mitad 0. Con $k = 2$ factores presentes en la población (se supone que los factores se presentan independientemente uno de otro), la distribución de intensidades de pigmentación estaría representada por el desarrollo del binomio $(p + q)^2$:

$\{FF,\ \ Ff,\ \ ff\ \}$	clases de pigmentación (espacio de probabilidad)
$\{0{,}25,\ \ 0{,}50,\ \ 0{,}25\}$	frecuencias esperadas
$\{2,\ \ 1,\ \ 0\ \}$	intensidad de pigmentación

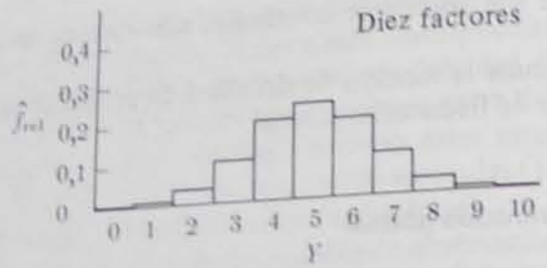


Fig. 5.2. Histograma basado en las frecuencias relativas esperadas que resultan del desarrollo del binomio $(0,5 + 0,5)^{10}$. El eje Y mide el número de factores de pigmentación F. (Para más explicación ver texto.)

Un cuarto de los individuos tendría intensidad de pigmentación 2, la mitad 1, y el cuarto restante 0.

El número de clases en el binomio aumenta con el número de factores. Las distribuciones de frecuencias son simétricas y las frecuencias esperadas en los extremos se hacen progresivamente menores al aumentar k. La distribución binomial para $k = 10$ se representa gráficamente como un histograma en la figura 5.2 (en vez de por un diagrama de barras como debería trazarse). Observamos que se aproxima al contorno acampanado habitual de la distribución normal (figuras 5.3 y 5.4). Si fuéramos a desarrollar la expresión para $k = 20$, nuestro histograma se aproximaría tanto a una distribución de frecuencias normal que no podríamos mostrar la diferencia entre ellos en un gráfico del tamaño de esta página. Al principio de este proceso hicimos varias suposiciones restrictivas severas por razón de simplicidad. ¿Qué ocurre cuando éstas se eliminan? Cuando $p \neq q$, la distribución también se aproxima a la normalidad cuando k tiende a infinito. Esto es difícil de ver intuitivamente porque cuando $p \neq q$ el histograma es al principio asimétrico. No obstante, puede demostrarse que cuando k, p y q son tales que $kpq \geqslant 3$, la distribución normal se aproximará estrechamente. En una situación más realista se permitirían factores que se hallan en más de dos estados, uno que hace una gran aportación, un segundo estado una menor aportación, y así sucesivamente. Sin embargo, también puede demostrarse que el polinomio $(p+q+r+\cdots+z)^k$ se aproxima a la distribución normal cuando k tiende a infinito. Pueden estar presentes diferentes factores con diferentes frecuencias y pueden tener diferentes efectos cuantitativos. Siempre que éstos sean aditivos e independientes, se alcanza la normalidad cuando k tiende a infinito.

La omisión de estas restricciones hace que las hipótesis conduzcan a una distribución normal compatible con innumerables situaciones biológicas. Por lo tanto, no es sorprendente que tantas variables biológicas se distribuyan aproximadamente según la ley normal.

Vamos a revisar las condiciones que tenderían a producir distribuciones de frecuencias normales: 1) si hay muchos factores que son simples o compuestos; 2) si estos factores se presentan con independencia; 3) si los factores son independientes en efecto, es decir, si sus efectos son aditivos; y 4) si contribuyen igual a la varianza. Todavía no estamos en situación de discutir la cuarta condición y ahora se menciona solamente para completar. Volveremos sobre esto en el capítulo 7.

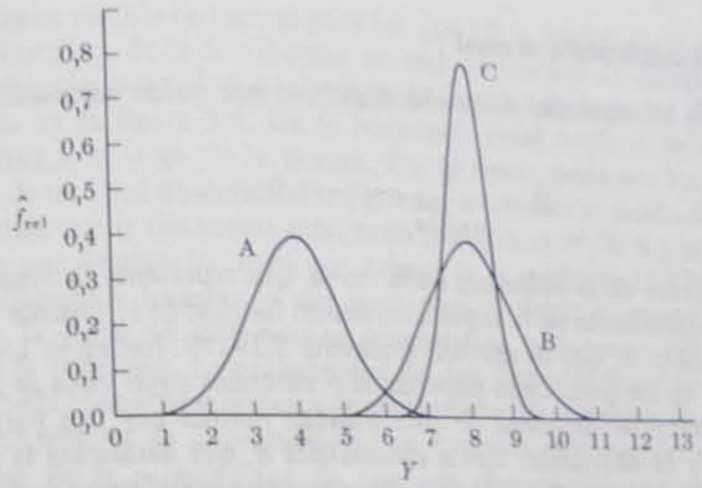


Fig. 5.3. Ilustración de cómo los cambios en los dos parámetros de la distribución normal afectan a la forma y posición de las curvas. A. $\mu = 4, \sigma = 1$; B. $\mu = 8, \sigma = 1$; C. $\mu = 8, \sigma = 0,5$.

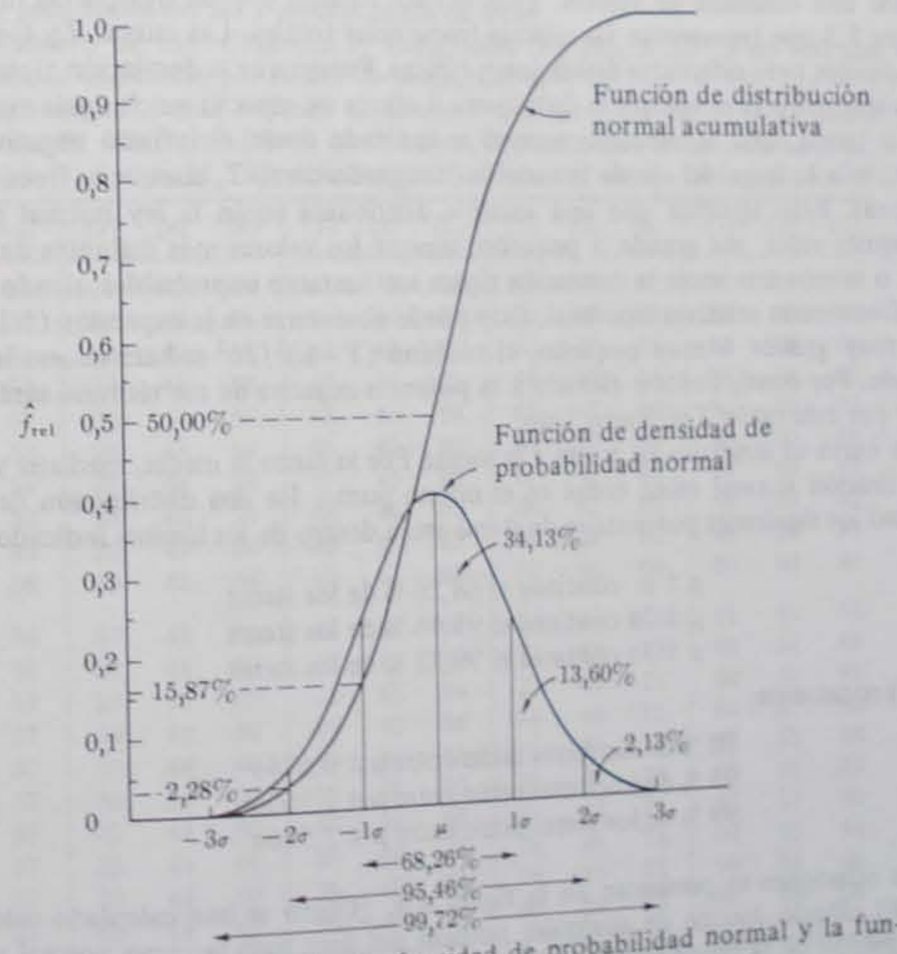


Fig. 5.4. Areas bajo la función de densidad de probabilidad normal y la función de distribución normal acumulativa.

5.3 Propiedades de la distribución normal

Realmente la *función de densidad de probabilidad normal* puede representarse por la expresión

$$Z = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(Y-\mu)^2/2\sigma^2} \qquad (5.1)$$

En ella Z indica la altura de la ordenada de la curva, que representa la densidad de los ítems. Es la variable dependiente en la expresión, siendo función de la variable Y. Hay dos constantes en la ecuación: π, que es aproximadamente 3,14159, haciendo $1/\sqrt{2\pi}$ igual a 0,39894, y e, la base de los logaritmos neperianos o naturales cuyo valor se aproxima a 2,71828. En una función de densidad de probabilidad normal hay dos parámetros, la media paramétrica μ y la desviación típica paramétrica σ, que determina la situación y forma de la distribución. Así, no hay sólo una distribución normal como pudiera parecer a los no iniciados que se encuentran la misma figura acampanada en los libros de texto elementales; más bien hay una infinidad de tales curvas, ya que estos parámetros pueden tomar una infinidad de valores. Esto se hace patente por las tres curvas normales de la figura 5.3 que representan las mismas frecuencias totales. Las curvas B y C tienen idénticas medias pero diferentes desviaciones típicas. Puesto que la desviación típica de la curva C es solamente la mitad que la de la curva B ofrece un aspecto mucho más estrecho.

En teoría, una distribución normal se extiende desde el infinito negativo al infinito positivo a lo largo del eje de la variable (designado como Y, aunque es frecuentemente la abscisa). Esto significa que una variable distribuida según la ley normal puede tomar cualquier valor, sea grande o pequeño, aunque los valores más distantes de la media en más o menos tres veces la desviación típica son bastante improbables, siendo muy escasas sus frecuencias relativas esperadas. Esto puede observarse en la expresión (5.1). Cuando Y sea muy grande o muy pequeño, el término $(Y-\mu)^2/2\sigma^2$ se hará necesariamente muy grande. Por consiguiente e elevado a la potencia negativa de ese término será muy pequeño y por esta razón Z será muy bajo.

La curva es simétrica en torno a la media. Por lo tanto la media, mediana y moda de la distribución normal están todas en el mismo punto. En una distribución de frecuencias normal los siguientes porcentajes de ítems están dentro de los límites indicados:

$\mu \pm \sigma$ contiene el 68,26 % de los ítems
$\mu \pm 2\sigma$ contiene el 95,46 % de los ítems
$\mu \pm 3\sigma$ contiene el 99,73 % de los ítems

Recíprocamente,

50 % de los ítems están entre $\mu \pm 0{,}674\sigma$
95 % de los ítems están entre $\mu \pm 1{,}960\sigma$
99 % de los ítems están entre $\mu \pm 2{,}576\sigma$

Estas relaciones se presentan en la figura 5.4. ¿Cómo se han calculado estos porcentajes? El cálculo directo de cualquier porción del área bajo la curva normal requiere una integración de la función presentada como expresión (5.1). Afortunadamente, para aquellos que no sepan cálculo (e incluso para los que sí) la integración ya se ha realizado en una forma alternativa de la distribución normal: la *función de distribución normal* (función de distribución *acumulativa* teórica de la función de densidad de probabilidad normal), indicada en la figura 5.4. Da la frecuencia total desde el infinito negativo hasta cualquier punto a lo largo de la abscisa. Por lo tanto, podemos buscar directamente la probabilidad de que una observación sea inferior a un valor indicado de Y. Por ejemplo, la figura 5.4 señala que la frecuencia total hasta la media es 50,00 % y la frecuencia hasta un punto igual a una desviación típica por debajo de la media es 15,87 %. Estas frecuencias se hallan gráficamente levantando una línea vertical desde un punto tal como $-\sigma$, hasta que corta a la curva de distribución acumulativa, y leyendo entonces la frecuencia (15,87 %) de la ordenada. La probabilidad de que una observación esté entre dos puntos arbitrarios, puede encontrarse restando la probabilidad de que dicha observación esté por debajo del punto inferior de la probabilidad de que una observación esté por debajo del punto inferior de la probabilidad de que una observación esté por debajo del punto

TABLA 5.1

Poblaciones de longitudes del ala y rendimientos de leche

Columna 1. Número de fila. *Columna 2.* Longitudes (en mm × 10^{-1}) de 100 alas de moscas domésticas dispuestas por orden de magnitud; $\mu = 45{,}5$, $\sigma^2 = 15{,}21$, $\sigma = 3{,}90$; distribución aproximadamente normal. *Columna 3.* Producción total de leche anual (en cientos de libras) de 100 vacas Jersey de dos años, dispuesta por orden de magnitud; $\mu = 66{,}61$, $\sigma^2 = 124{,}4779$, $\sigma = 11{,}597$; distribución fuertemente alejada de la normalidad.

(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
01	36	51	21	42	58	41	45	61	61	47	67	81	49	76
02	37	51	22	42	58	42	45	61	62	47	67	82	49	76
03	38	51	23	42	58	43	45	61	63	47	68	83	49	79
04	38	53	24	43	58	44	45	61	64	47	68	84	49	80
05	39	53	25	43	58	45	45	61	65	47	69	85	50	80
06	39	53	26	43	58	46	45	62	66	47	69	86	50	81
07	40	54	27	43	58	47	45	62	67	47	69	87	50	82
08	40	55	28	43	58	48	45	62	68	47	69	88	50	82
09	40	55	29	43	58	49	45	62	69	47	69	89	50	82
10	40	56	30	43	58	50	45	63	70	48	69	90	50	82
11	41	56	31	43	58	51	46	63	71	48	70	91	51	83
12	41	56	32	44	59	52	46	63	72	48	72	92	51	85
13	41	57	33	44	59	53	46	64	73	48	73	93	51	87
14	41	57	34	44	59	54	46	65	74	48	73	94	51	88
15	41	57	35	44	60	55	46	65	75	48	74	95	52	88
16	41	57	36	44	60	56	46	65	76	48	74	96	52	89
17	42	57	37	44	60	57	46	65	77	48	74	97	53	93
18	42	57	38	44	60	58	46	65	78	49	74	98	53	94
19	42	57	39	44	60	59	46	67	79	49	75	99	54	96
20	42	57	40	44	61	60	46	67	80	49	76	00	55	98

Fuente: Columna 2 - Datos adaptados de Sokal y Hunter (1955). Columna 3 - Datos de los registros del Gobierno canadiense.

superior. Por ejemplo, en la figura 5.4 podemos observar que la probabilidad de que una observación esté entre la media y un punto situado a una distancia de una desviación típica por debajo de la media es 0,5000 – 0,1587 = 0,3413.

La función de distribución normal está tabulada en la tabla **II**, Areas de la Curva Normal, y para comodidad en cálculos posteriores se ha restado 0,5 de todas las entradas. Por lo tanto, esta tabla registra la proporción del área entre la media y cualquier punto superior a ella en un número determinado de desviaciones típicas. Así por ejemplo, el área entre la media y el punto 0,5 desviaciones típicas superior a la media es 0,1915 del área total de la curva. Del mismo modo, el área entre la media y el punto 2,64 desviaciones típicas sobre la media es 0,4959 de la curva. Un punto a 4,0 desviaciones típicas de la media incluye 0,499968 del área entre él y la media. No obstante, puesto que la distribución normal se extiende desde el infinito negativo al positivo, es necesario atravesar una distancia infinita desde la media para alcanzar completamente la mitad del área bajo la curva. El manejo de la tabla de áreas de la curva normal se ilustra en la próxima sección.

Un experimento de muestreo dará una idea de la distribución de los ítems bajo una curva normal.

Experimento 5.1. Extraer muestras de dos poblaciones. La primera es una distribución de frecuencias aproximadamente normal de 100 longitudes del ala de moscas domésticas. La segunda población se desvía fuertemente de la normalidad. Es una distribución de frecuencias de la producción anual de leche total de 100 vacas Jersey. Ambas poblaciones se presentan en la tabla 5.1. Extraer muestras de ellas repetidamente para simular el muestreo de una población infinita. Obtener muestras de 35 ítems de cada una de las dos poblaciones. Esto se hace obteniendo dos juegos de 35 números al azar de dos cifras de la tabla de números aleatorios (tabla I) con la cual se familiarizó en el experimento 4.1. Anotar los números aleatorios en grupos de cinco y copiar a continuación de ellos el valor de Y (para cada longitud del ala o producción de leche) correspondiente al número aleatorio. Más abajo se presenta un ejemplo de tales grupos de cinco números y los cálculos necesarios, utilizando las longitudes del ala de moscas domésticas como ejemplo.

Estas muestras y los cálculos efectuados para cada una se utilizarán en capítulos subsiguientes. Por lo tanto, deben guardarse los datos.

Número aleatorio	*Longitud del ala* Y
16	41
59	46
99	54
36	44
21	42
	$\Sigma Y =$ 227
	$\Sigma Y^2 =$ 10 413
	$\bar{Y} =$ 45,4

TABLA 5.2

Tabla para registrar distribuciones de frecuencias de desviantes típicas $(Y_i - \mu)/\sigma$ en las muestras del experimento 5.1

	Longitudes del ala			*Rendimiento de leche*	
	Variantes que se hallan entre estos límites	*f*		*Variantes que se hallan entre estos límites*	*f*
$-\infty$			$-\infty$		
-3σ	———	———	-3σ	———	———
$-2\frac{1}{2}\sigma$	———	———	$-2\frac{1}{2}\sigma$	———	———
-2σ	36, 37	———	-2σ	———	———
$-1\frac{1}{2}\sigma$	38, 39	———	$-1\frac{1}{2}\sigma$	———	———
$-\sigma$	40, 41	———	$-\sigma$	51–55	———
$-\frac{1}{2}\sigma$	42, 43	———	$-\frac{1}{2}\sigma$	56–61	———
$\mu = 45{,}5$	44, 45	———	$\mu = 66{,}61$	62–66	———
$\frac{1}{2}\sigma$	46, 47	———	$\frac{1}{2}\sigma$	67–72	———
σ	48, 49	———	σ	73–77	———
$1\frac{1}{2}\sigma$	50, 51	———	$1\frac{1}{2}\sigma$	78–83	———
2σ	52, 53	———	2σ	84–88	———
$2\frac{1}{2}\sigma$	54, 55	———	$2\frac{1}{2}\sigma$	89–94	———
3σ	———	———	3σ	95–98	———
$+\infty$			$+\infty$		

En este experimento consideramos las 35 variantes para cada variable como una muestra única, sin separarlas en grupos de cinco. Puesto que la verdadera media y desviación típica (μ y σ) de las dos distribuciones son conocidas, se puede calcular la expresión $(Y_i - \mu)/\sigma$ para cada variante Y_i. Así, para la primera longitud del ala de una mosca doméstica muestreada anteriormente, se calcula

$$\frac{(41 - 45{,}5)}{3{,}90} = -1{,}1538$$

Esto significa que la primera longitud del ala es 1,1538 desviaciones típicas inferior a la verdadera media de la población. La desviación de la media medida en unidades de desviación típica se llama *desviación tipificada*. Los argumentos de la tabla **II** que expresan distancia de la media en unidades de σ se denominan *desviaciones típicas normales*. Agrupar las 35 variantes en una distribución de frecuencias, a continuación hacer lo mismo para los rendimientos de leche. Puesto que se conoce la media y desviación típica paramétricas, no es necesario calcular separadamente cada desviación sino que basta anotar los límites de clase en términos de la variable real así como en forma de desviación típica. Los límites de clase para tal distribución de frecuencias se presentan en la tabla 5.2. Combinar los resultados del propio muestreo con los de los compañeros de clase y estudiar el porcentaje de ítems de la distribución que se hallan a una, dos y tres desviaciones típicas a cada lado de la media. Observar las marcadas diferencias de distribución entre las longitudes del ala de moscas domésticas y los rendimientos de la leche.

5.4 Aplicaciones de la distribución normal

La distribución normal de frecuencias es la más ampliamente utilizada en estadística, y a veces recurriremos a ella en una variedad de situaciones. De momento podemos subdividir sus aplicaciones como sigue.

1. A veces hay que saber si una muestra determinada se distribuye según la ley normal antes de poder aplicarle una cierta prueba. Para comprobar si una muestra determinada sigue la ley normal, hay que calcular las frecuencias esperadas para una curva normal de la misma media y desviación típica, valiéndose de la tabla de áreas de la curva normal. En este libro solamente emplearemos métodos gráficos aproximados para probar la normalidad. Estos se destacan en la próxima sección.

2. Conociendo si una muestra se distribuye según la ley normal se pueden confirmar o rechazar ciertas hipótesis fundamentales acerca de la naturaleza de los factores que afectan al fenómeno estudiado. Esto está relacionado con las condiciones que contribuyen a la normalidad en una distribución de frecuencias, discutidas en la sección 5.2. Así, si encontramos que una determinada variable está normalmente distribuida, no tenemos motivo para rechazar la hipótesis de que los factores causales que afectan a la variable sean aditivos e independientes y de igual varianza. Por otra parte, cuando encontramos desviaciones de la normalidad, esto puede indicar que ciertas fuerzas, tales como la selección, afectan a la variable en estudio. Por ejemplo, la bimodalidad puede indicar una mezcla de muestras de dos poblaciones. La asimetría de los datos de rendimientos de leche puede reflejar el hecho de que éstos fuesen registros de vacas seleccionadas y las vacas de escasa producción no se incluyeran en el registro.

3. Si suponemos que una distribución dada es normal, podemos hacer predicciones y pruebas de determinadas hipótesis basadas en esta suposición. Más abajo se presenta un ejemplo de esta aplicación.

Se recordarán los datos de pesos de niños varones chinos recién nacidos ilustrados en el cuadro 3.2. La media de esta muestra de 9 465 pesos de recién nacidos es 109,9 onzas, y su desviación típica 13,593 onzas. Al extraer muestras al azar de las actas de nacimiento de esta población, ¿cuál es la probabilidad de obtener un peso de 151 onzas o superior? Este peso es considerablemente superior a la media de nuestra muestra, siendo la diferencia 151-109,9 = 41,1 onzas. Sin embargo, no podemos consultar la tabla de áreas de la curva normal con una diferencia en onzas. Debemos *tipificar* la diferencia, es decir, dividirla por la desviación típica para convertirla en una desviación tipificada. Al dividir la diferencia por la desviación típica obtenemos 41,1/13,593 = 3,02. Esto significa que un peso de nacimiento de 151 onzas es 3,02 desviaciones típicas mayor que la media. Suponiendo que los pesos de nacimiento sigan la ley normal, podemos consultar la tabla de áreas de la curva normal (tabla **II**), donde encontramos un valor de 0,4987 para 3,02 desviaciones típicas. Esto quiere decir que el 49,87 % del área de la curva se halla entre la media y un punto que dista de ella 3,02 desviaciones típicas. A la inversa, 0,0013 ó 0,13 % del área está más allá de 3,02 desviaciones típicas sobre la media. Así, suponiendo una distribución normal de pesos al nacer, solamente 0,13 % ó 13 de los 10 000 niños tendría un peso de 151 onzas *o más distante* de la media. Es bastante improbable que un solo ítem muestreado de esa población se desviara tanto de la media, y si se hubiese obtenido una muestra al azar de un peso de los registros de una población no especificada, pudiéramos por ello justificarnos al dudar si la observación proviene en realidad de la población que nosotros conocemos.

La probabilidad anterior se ha calculado de una cola de la distribución. Hemos hallado la probabilidad de que un individuo sea 3,02 desviaciones típicas *mayor* que la media. Si no tenemos conocimiento previo de que el individuo sea más pesado o más ligero que la media, sino que estamos interesados simplemente por lo que difiere de la media de la población, una cuestión apropiada sería: suponiendo que el individuo pertenezca a la población, ¿cuál es la probabilidad de observar un peso de nacimiento de un individuo tan alejado de la media en una u otra dirección? Esa probabilidad debe calcularse utilizando ambas colas de la distribución. La probabilidad anterior puede simplemente duplicarse puesto que la curva normal es simétrica. Así, 2 × 0,0013 = 0,0026. Esta también es tan pequeña que podríamos concluir que un peso al nacer tan alejado como 151 onzas es improbable que se haya originado de la población representada por nuestra muestra de niños varones chinos.

De este ejemplo podemos averiguar un detalle más importante. Nuestra hipótesis ha sido que los pesos al nacer siguen la ley normal. Sin embargo, el examen de la distribución de frecuencias del cuadro 3.2 muestra claramente que la distribución es asimétrica, estrechándose hacia la derecha. Aunque hay ocho clases sobre la clase media, solamente hay seis por debajo de ésta. A la vista de esta asimetría, las conclusiones acerca de una cola de la distribución no corresponderían necesariamente a la segunda cola. Hemos calculado que 0,13 % de los ítems se hallarían más allá de 3,02 desviaciones típicas sobre la media, lo cual corresponde a 151 onzas. En realidad nuestra muestra contiene 20 ítems (14 + 5 + 1) después de la clase 147,5 onzas cuyo límite superior es 151,5 onzas, casi el mismo que el peso al nacer para un solo individuo muestreado. Sin embargo, 20 ítems de los 9 465 de la muestra, es aproximadamente el 0,21 %, más del 0,13 % esperado de la distribución normal. Aunque a pesar de ello sería improbable encontrar un peso único tan grande como 151 onzas en la muestra, las conclusiones basadas en la hipótesis de normalidad pudieran ser erróneas si la probabilidad exacta fuese crítica para una prueba determinada. Nuestras conclusiones estadísticas solamente son tan válidas como nuestras hipótesis sobre los datos.

5.5 Desviaciones de la normalidad y métodos gráficos

En muchos casos una distribución de frecuencias observada se desviará evidentemente de la normalidad. Subrayaremos dos tipos de desviación de la normalidad. Uno es la *disimetría* que es otra forma de denominar la asimetría; disimetría significa que un extremo de la curva se alarga más que el otro. En tales curvas no coincidirán la media y la mediana. Las curvas se denominan torcidas a la derecha o a la izquierda dependiendo de que se alargue una u otra cola. El otro tipo de desviación de la normalidad es la *curtosis* o apuntamiento de una curva. Una curva *leptocúrtica* tiene más ítems cerca de la media y en las colas, con menos ítems en las regiones intermedias, que una distribución normal de la misma media y varianza. Una curva *platicúrtica* tiene menos ítems cerca de la media y en las colas que la curva normal pero tiene más ítems en las regiones intermedias. Una distribución bimodal es una distribución platicúrtica extrema.

Se han desarrollado métodos gráficos que examinan la forma de una distribución observada para desviaciones de la normalidad. Estos métodos permiten además estimaciones de la media y desviación típica de la distribución sin cálculo.

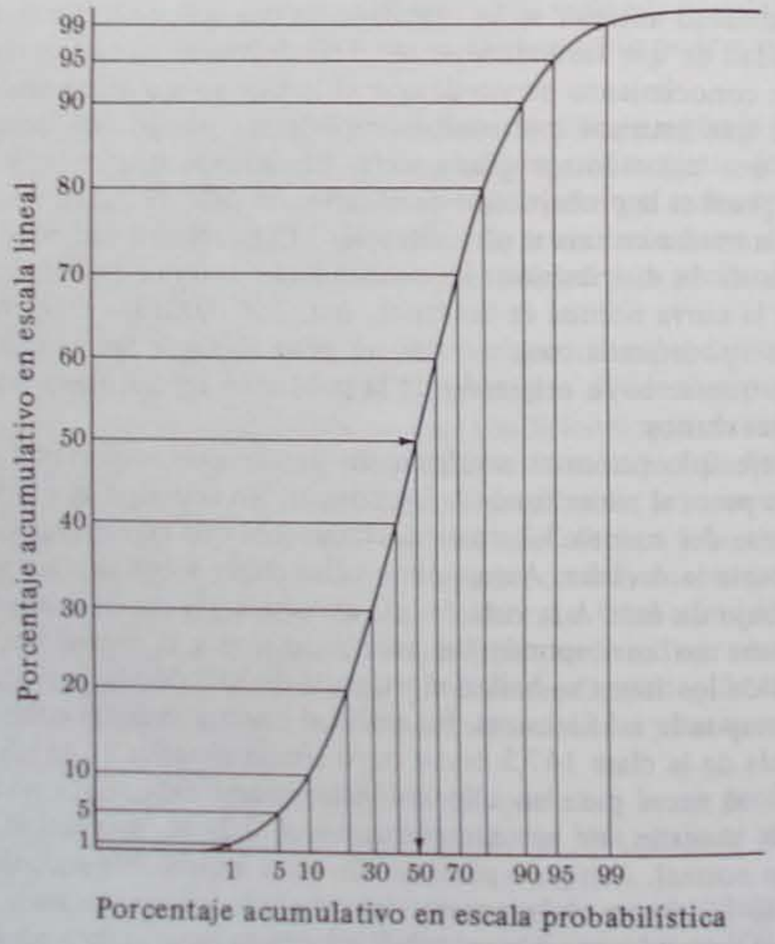


Fig. 5.5. Transformación de porcentajes acumulativos en escala de probabilidad normal.

Los métodos gráficos están basados en una distribución de frecuencias acumulativa. En la figura 5.4 vimos que una distribución normal representada gráficamente en forma acumulativa describe una curva en forma de **S** llamada curva sigmoidea. En la figura 5.5, la ordenada de la curva sigmoidea representa las frecuencias relativas expresadas como

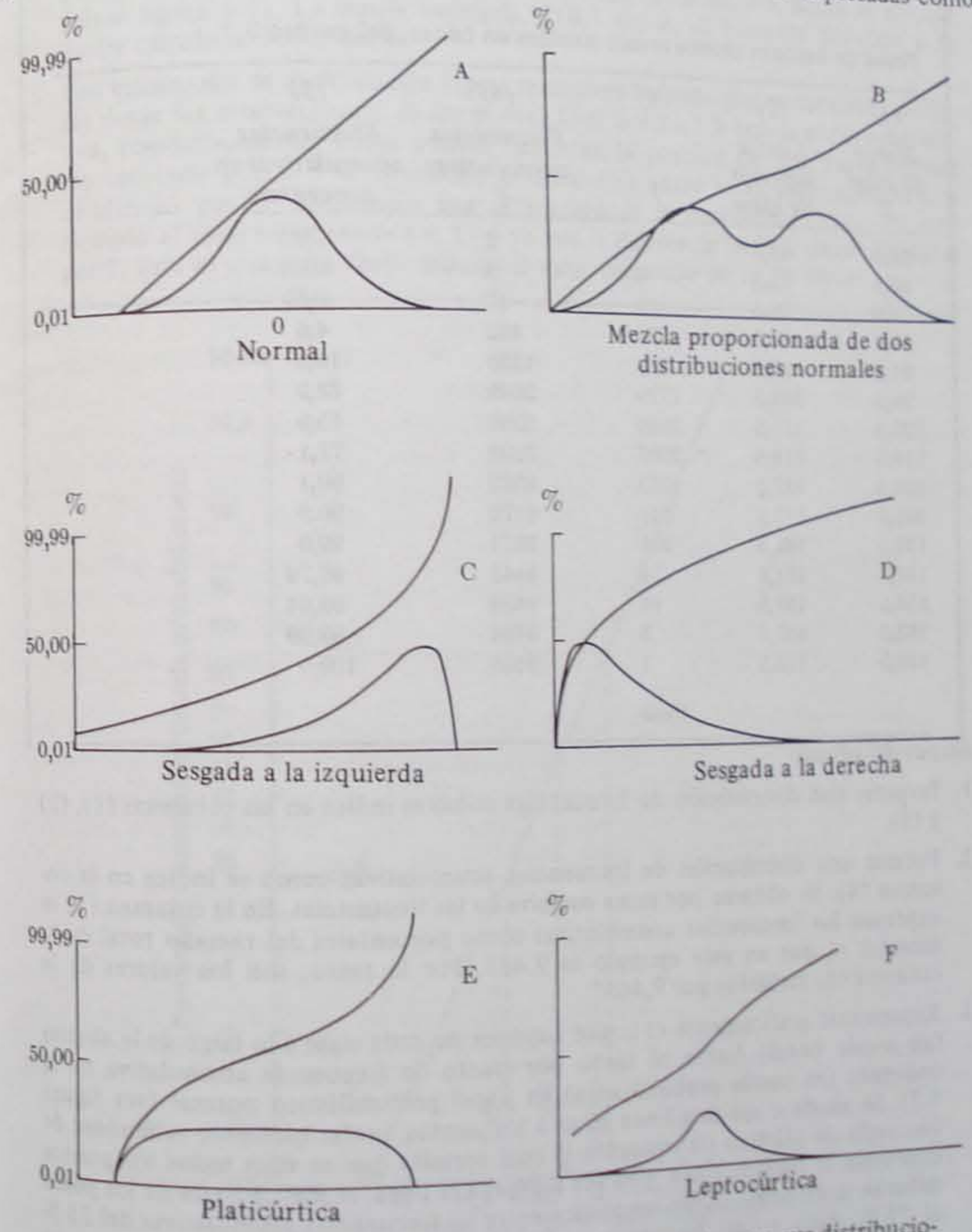


Fig. 5.6. Ejemplos de algunas distribuciones de frecuencias con sus distribuciones acumulativas representando la ordenada en escala probabilística normal. (Para explicación ver cuadro 5.1.)

CUADRO 5.1

Prueba gráfica para normalidad de una distribución de frecuencias y estimación de la media y desviación típica. Aplicación del papel probabilístico aritmético.

Pesos de varones chinos recién nacidos en onzas, del cuadro 3.2.

(1) *Marca de clase Y*	*(2)* *Límite superior de clase*	*(3)* *f*	*(4)* *Frecuencias acumulativas F*	*(5)* *Frecuencias acumulativas en porcentaje*
59,5	63,5	2	2	0,02
67,5	71,5	6	8	0,08
75,5	79,5	39	47	0,50
83,5	87,5	385	432	4,6
91,5	95,5	888	1320	13,9
99,5	103,5	1729	3049	32,2
107,5	111,5	2240	5289	55,9
115,5	119,5	2007	7296	77,1
123,5	127,5	1233	8529	90,1
131,5	135,5	641	9170	96,9
139,5	143,5	201	9371	99,0
147,5	151,5	74	9445	99,79
155,5	159,5	14	9459	99,94
163,5	167,5	5	9464	99,99
171,5	175,5	1	9465	100,0
		9465		

Etapas del cálculo

1. Preparar una distribución de frecuencias como se indica en las columnas (1), (2) y (3).
2. Formar una distribución de frecuencias acumulativas como se indica en la columna (4). Se obtiene por suma sucesiva de las frecuencias. En la columna (5) se expresan las frecuencias acumulativas como porcentajes del tamaño total de la muestra n, que en este ejemplo es 9.465. Por lo tanto, son los valores de la columna (4) divididos por 9.465.
3. Representar gráficamente el límite superior de cada clase a lo largo de la abscisa (en escala lineal) frente al tanto por ciento de frecuencia acumulativa en la ordenada (en escala probabilística) en papel probabilístico normal (ver figura 5.7). Se ajusta a ojo una línea recta a los puntos, preferiblemente valiéndose de una regla de plástico transparente la cual permite que se vean todos los puntos conforme se traza la línea. Una vez dibujada la línea, la mayor parte de los pesos deberán distribuirse sobre los puntos entre las frecuencias acumulativas del 25 % al 75 %. Esto es así porque una diferencia de un solo ítem puede suponer cambios apreciables en los porcentajes de los extremos. Observamos que las frecuencias superiores se desvían a la derecha de la línea recta. Esto es típico de los datos que son sesgados a la derecha (véase figura 5.6D).

4. Esta gráfica permite la estimación rápida de la media y desviación típica de una muestra. La media se aproxima por una estimación gráfica de la mediana. Cuanto más normal sea la distribución, más próxima a la mediana estará la media. La mediana se estima bajando una perpendicular desde la intersección del punto 50 % de la ordenada con la curva de frecuencia acumulativa, hasta la abscisa (véase figura 5.7). La media estimada, 110,7 onzas, es bastante próxima a la media calculada, 109,9 onzas.
5. Una estimación de la desviación típica se obtiene bajando perpendiculares similares desde las intersecciones de los puntos 15,9 % y 84,1 % con la curva acumulativa, respectivamente. Estos puntos encierran la porción de una curva normal representada por $\mu \pm \sigma$. Midiendo la diferencia entre estas perpendiculares y dividiendo por 2, obtenemos una estimación de la desviación típica. En este ejemplo el valor estimado es $s = 13{,}6$ ya que la diferencia es 27,2 onzas divididas por 2. Esta es una justa aproximación al valor calculado de 13,59 onzas.

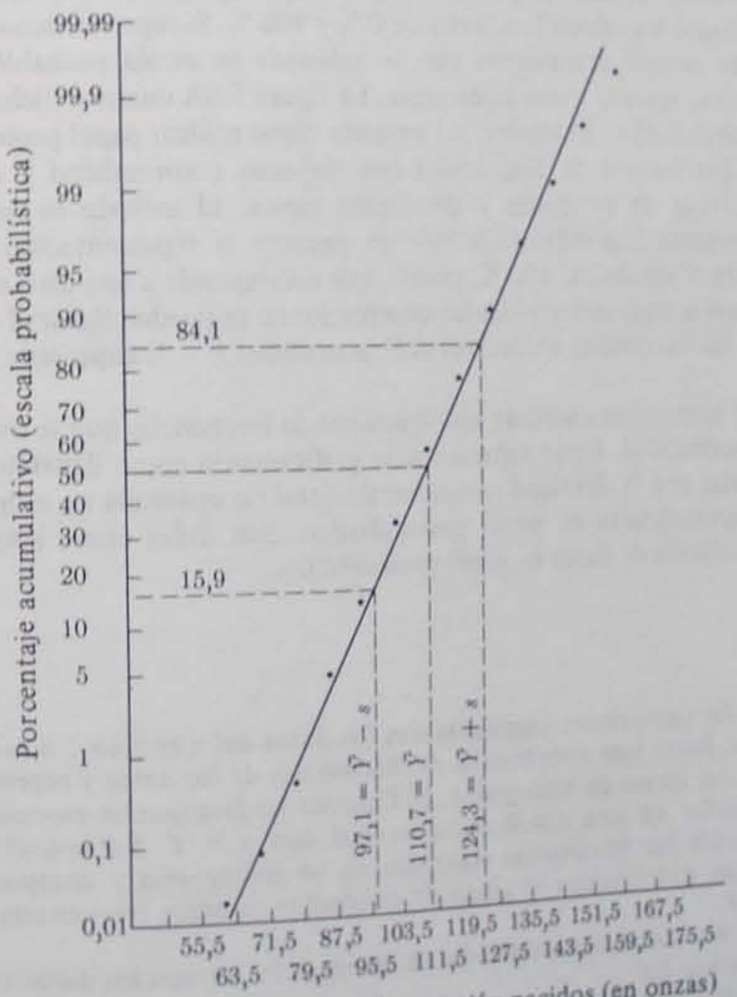


Fig. 5.7. Análisis gráficos de los datos del cuadro 5.1.

porcentajes. La pendiente de la curva acumulativa refleja los cambios en altura de la distribución de frecuencias en la que está basada. Así, la pendiente del segmento medio de la curva normal acumulativa corresponde a la altura relativamente mayor de la curva normal en torno a su media. En las figuras 5.4 y 5.5 la ordenada está en escala lineal. Otra escala posible es la *escala de probabilidad normal* (con frecuencia llamada simplemente *escala probabilística*), que puede originarse bajando perpendiculares desde la curva acumulativa normal, correspondientes a determinados porcentajes en la ordenada, hasta la abscisa (como se observa en la figura 5.5). La escala representada por la abscisa compensa la no linearidad de la curva acumulativa normal. Reduce la escala en torno a la media y la ensancha en los porcentajes acumulativos bajos y altos. Esta escala puede hallarse en *papel milimetrado probabilístico normal o aritmético* (o simplemente *papel probabilístico*), que generalmente se puede conseguir. Usualmente este papel tiene el margen largo graduado en escala probabilística, mientras el margen corto está en escala lineal. Nótese que no hay puntos 0 % ni 100 % en la ordenada. Esto no es posible ya que la distribución normal se extiende del infinito negativo al positivo, y por muy larga que hiciésemos nuestra línea nunca alcanzaríamos los valores limitantes de 0 % y 100 %. Si representamos gráficamente una distribución normal acumulativa con la ordenada en escala probabilística normal, corresponderá exactamente a una línea recta. La figura 5.6A muestra dicha gráfica trazada en papel probabilístico. El cuadro 5.1 muestra cómo utilizar papel probabilístico para examinar una distribución de frecuencias con respecto a normalidad y cómo obtener estimaciones gráficas de su media y desviación típica. El método es más eficaz para muestras medianamente grandes ($n > 50$); no permite la representación gráfica de la última frecuencia acumulativa, 100 %, puesto que corresponde a una distancia infinita de la media. Si interesa representar todas las observaciones, se pueden llevar al eje de ordenadas, en lugar de las frecuencias acumulativas F, la cantidad $F - \frac{1}{2}$ expresada como porcentaje de n.

La figura 5.6 muestra una serie de distribuciones de frecuencias que se desvían diferentemente de la normalidad. Están representadas gráficamente como distribuciones de frecuencias ordinarias con la densidad en una escala lineal (la ordenada no se indica) y como distribuciones acumulativas en papel probabilístico. Son útiles como pauta cuando se examina la distribución de datos en papel probabilístico.

Ejercicios 5

5.1 Realizar las operaciones siguientes con los datos del ejercicio 2.4. a) Si aún no se ha hecho, hacer una distribución de frecuencias de los datos y representarla gráficamente en forma de histograma. b) Calcular las frecuencias esperadas para cada clase basadas en una distribución normal con $\mu = \bar{Y}$ y $\sigma = s$. c) Representar gráficamente las frecuencias esperadas en un histograma y compararlas con las observadas. d) Comentar el grado de concordancia entre frecuencias observadas y esperadas.

5.2 Efectuar las operaciones indicadas en el ejercicio 5.1 con los datos transformados en el ejercicio 2.6.

5.3 Supóngase que la longitud del pétalo de una población de plantas de la especie X se distribuye normalmente con una media de $\mu = 3{,}2$ cm y una desviación típica de $\sigma = 1{,}8$. ¿Qué proporción de la población se esperaría que tuviese un pétalo de longitud a) mayor que 4,5 cm, b) mayor que 1,78 m, c) entre 2,9 y 3,6 cm? SOLUCION a) = 0,2353, b) = 0,7845 y c) = 0,154.

5.4 Hacer un análisis gráfico de los datos de la grasa de la leche dados en el ejercicio 3.3, utilizando papel probabilístico. Además, representar los datos en papel probabilístico con la abscisa en unidades logarítmicas. Comparar los resultados de los dos análisis.

5.5 Supóngase que los caracteres A y B son independientes y normalmente distribuidos con parámetros $\mu_A = 28{,}6$, $\sigma_A = 4{,}8$, $\mu_B = 16{,}2$ y $\sigma_B = 4{,}1$. Muestrear dos individuos al azar. ¿Cuál es la probabilidad de obtener muestras en las que ambos individuos sean menores que 20 para los dos caracteres? ¿Cuál es la probabilidad de que al menos uno de los individuos sea mayor que 30 para el carácter B?

5.6 Valiéndose de la información dada en el cuadro 3.2, ¿cuál es la probabilidad de obtener un individuo con un peso de nacimiento negativo? ¿Cuál es esta probabilidad si suponemos que los pesos de nacimiento se distribuyen según la ley normal?

Capítulo 6

Estimación y contraste de hipótesis

En este capítulo ofrecemos respuesta a dos cuestiones estadísticas fundamentales que cada biólogo debe responder repetidamente en el curso de su trabajo: 1) ¿hasta qué punto son fiables los resultados que hemos obtenido? y 2) ¿hasta qué punto es probable que las diferencias entre los resultados observados y los esperados en base a una hipótesis, se hayan producido solamente por azar? La primera cuestión, acerca de la fiabilidad, se soluciona por medio del establecimiento de límites de confianza para los estadísticos de muestra. La segunda cuestión introduce el contraste de hipótesis. Ambos temas pertenecen al campo de la inferencia estadística. La materia de este capítulo es fundamental para la comprensión de cualquiera de los siguientes.

En la sección 6.1 consideramos en primer lugar la forma de la distribución de medias y su varianza. En la sección 6.2 examinamos las distribuciones y varianzas de otros estadísticos. Esto nos lleva a la materia general de errores típicos, que son estadísticos que miden la fiabilidad de una estimación. Los límites de confianza acotan nuestras estimaciones de parámetros de población. En la sección 6.3 desarrollamos la idea de límite de confianza y mostramos su aplicación a muestras en que se conoce la verdadera desviación típica. Sin embargo, usualmente tratamos con muestras pequeñas distribuidas aproximadamente según la ley normal con desviaciones típicas desconocidas, en cuyo caso debe utilizarse la distribución *t*. En la sección 6.4 presentamos la distribución *t*. En la sección 6.5 se señala la aplicación de *t* al cálculo de límites de confianza para estadísticos de muestras pequeñas con desviaciones típicas de población desconocidas. En la sección 6.6 se explica otra distribución importante (ji-cuadrado) y a continuación se aplica a la fijación de límites de confianza para la varianza (sección 6.7). La teoría de contraste de hipótesis se presenta en la sección 6.8 y se aplica a varios casos que muestran las distribuciones *t* o normal (sección 6.9). Finalmente, la sección 6.10 ilustra sobre contraste de hipótesis para varianzas por medio de la distribución ji-cuadrado.

6.1 Distribución y varianza de medias

Comenzamos nuestro estudio de la distribución y varianza de medias con un experimento de muestreo.

Experimento 6.1. En el experimento 5.1 se pidió que se conservaran las medias de las siete muestras de 5 longitudes del ala de moscas domésticas y las siete medias similares de los rendimientos de leche. Podemos coger estas medias de cada estudiante de una clase, reunirlas con los resultados del muestreo de clases previas, y construir una distribución de frecuencias de estas medias. Para cada variable podemos obtener también la media de las siete medias, que es una media de una muestra de 35 ítems. Ahora haremos de nuevo una distribución de frecuencias de esta medias, aunque se requiere un número considerable de muestreadores para acumular un número suficiente de muestras de 35 items para una distribución de frecuencias significativas.

En la tabla 6.1 se presenta una distribución de frecuencias de 1 400 medias de muestras de 5 longitudes del ala de moscas domésticas. Por el momento consideremos solamente las

TABLA 6.1

Distribución de frecuencias de las medias de 1400 muestras al azar de 5 longitudes del ala de moscas domésticas. (Datos de la tabla 5.1.) Las marcas de clase se han elegido para que den intervalos de $\frac{1}{2}\,\sigma_{\bar{Y}}$ a cada lado de la media paramétrica μ.

	(1) *Marca de clase* Y *(en mm* $\times 10^{-1}$*)*	(2) *Marca de clase (en unidades de* $\sigma_{\bar{Y}}$*)*	(3) f
	39,832	$-3\frac{1}{4}$	1
	40,704	$-2\frac{3}{4}$	11
	41,576	$-2\frac{1}{4}$	19
	42,448	$-1\frac{3}{4}$	64
	43,320	$-1\frac{1}{4}$	128
	44,192	$-\frac{3}{4}$	247
	45,064	$-\frac{1}{4}$	226
$\mu = 45{,}5 \rightarrow$	45,936	$\frac{1}{4}$	259
	46,808	$\frac{3}{4}$	231
	47,680	$1\frac{1}{4}$	121
	48,552	$1\frac{3}{4}$	61
	49,424	$2\frac{1}{4}$	23
	50,296	$2\frac{3}{4}$	6
	51,168	$3\frac{1}{4}$	3
			1400
	$\bar{Y} = 45{,}480$	$s = 1{,}778$	$\sigma_Y = 1{,}744$

columnas (1) y (3). Realmente estas muestras no han sido obtenidas por alumnos de bioestadística sino por un computador digital, permitiéndonos reunir estos valores en muy poco tiempo. Al pie de la tabla se da su media y desviación típica. Estos valores se representan gráficamente en papel probabilístico en la figura 6.1. Nótese que la distribución aparece completamente normal al igual que la de las medias basadas en 200 muestras de 35 longitudes del ala señaladas en la figura 6.2. Esto ilustra un teorema importante. *Las medias de muestras de una población normalmente distribuida están por sí mismas normalmente distribuidas independientemente del tamaño n de la muestra.* Así pues observamos que las medias de las muestras de longitudes del ala se distribuyen normalmente tanto si están basadas en 5 lecturas individuales como en 35.

Igualmente las distribuciones de las medias de los rendimientos de leche (figuras 6.3 y 6.4), fuertemente sesgados, parecen acercarse a las distribuciones normales. Sin embargo, las medias basadas en cinco rendimientos de leche (figura 6.3) no concuerdan tanto con la distribución casi normal como las medias de 35 ítems (figura 6.4). Esto ilustra otro teorema de fundamental importancia en estadística. *Las medias de las muestras extraídas de una población de cualquier distribución se aproximarán a la normal al aumentar el tamaño de la muestra.* Este teorema, rigurosamente planteado (para el muestreo de poblaciones con varianza finita), se conoce como el *teorema central del límite.* La importancia de este teorema reside en que nos permite utilizar la distribución normal para hacer inferencias estadísticas sobre las medias de poblaciones en las cuales los ítems no están distribuidos normalmente en su totalidad.

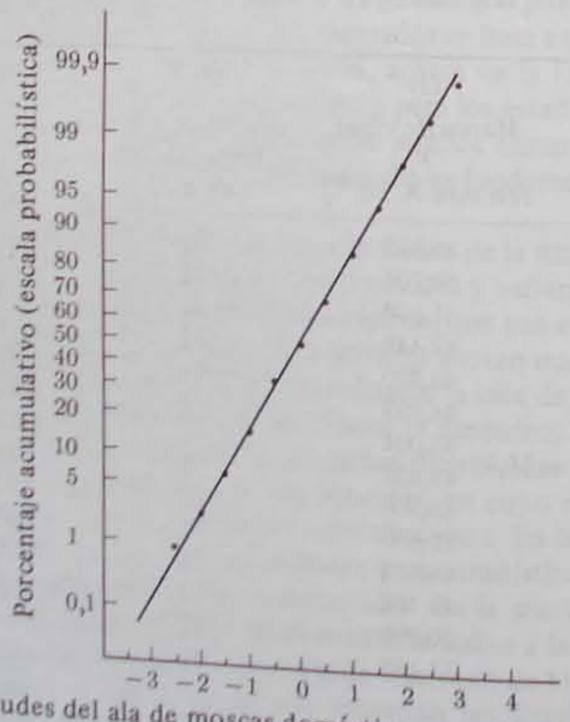


Fig. 6.1. Análisis gráfico de los datos de la tabla 6.1.

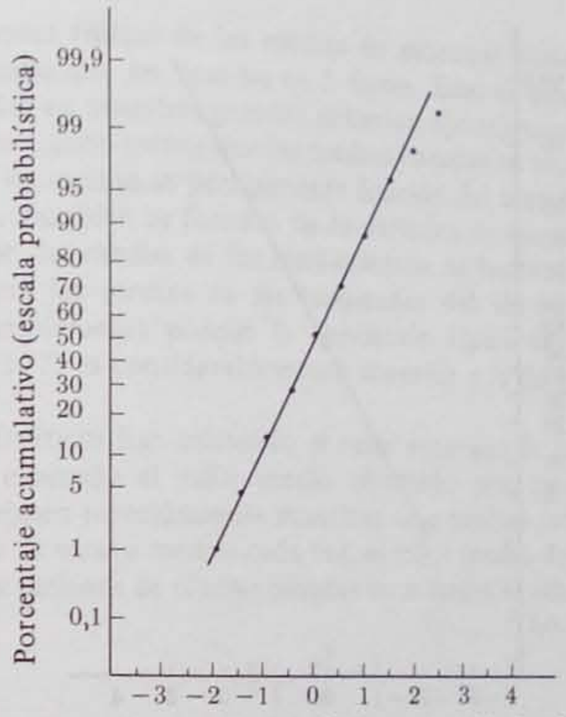


Fig. 6.2. Análisis gráfico de las medias de 200 muestras al azar de 35 longitudes del ala de moscas domésticas.

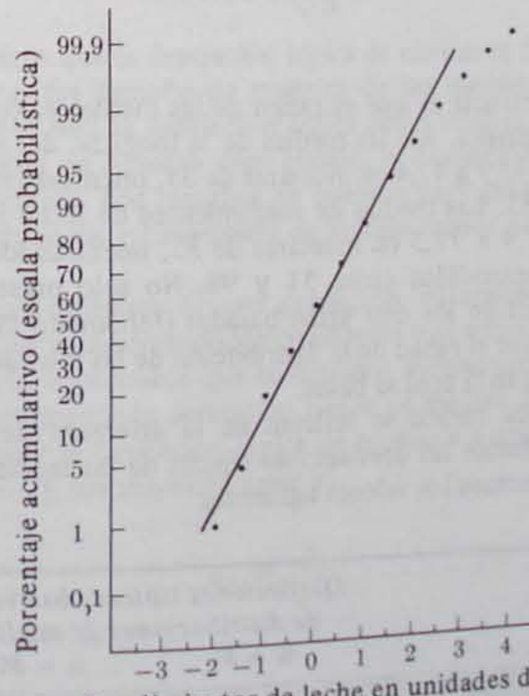


Fig. 6.3. Análisis gráfico de las medias de 1400 muestras al azar de 5 rendimientos de leche.

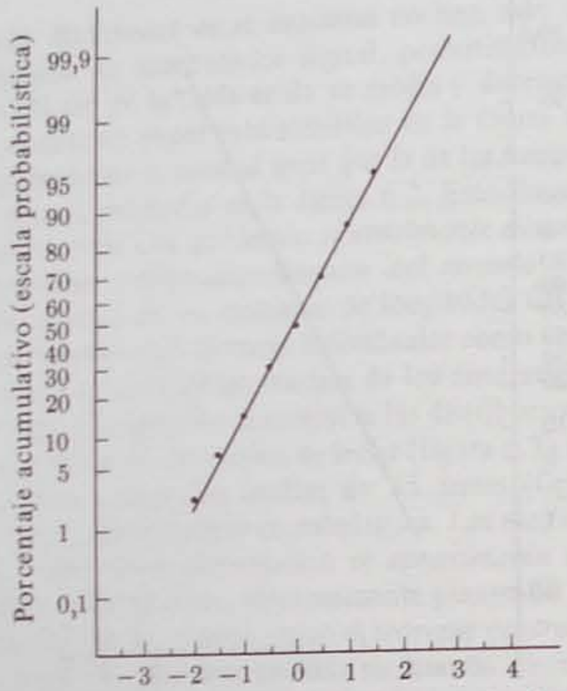


Fig. 6.4. Análisis gráfico de las medias de 200 muestras al azar de 35 rendimientos de leche.

Otro factor de importancia es que el rango de las medias es considerablemente menor que el de los ítems originales. Así las medias de la longitud del ala varían de 39,4 a 51,6 en muestras de 5 y de 43,9 a 47,4 en muestras de 35, mientras las longitudes individuales del ala varían de 36 a 55. Las medias de rendimientos de leche varían de 54,2 a 89,0 en muestras de 5 y de 61,9 a 71,3 en muestras de 35, mientras los rendimientos de leche individuales están comprendidos entre 51 y 98. No sólo presentan las medias menos dispersión que los ítems en los que están basadas (fenómeno fácil de comprender si se piensa un poco), sino que el rango de la distribución de las medias disminuye al aumentar el tamaño de la muestra en la cual se basan.

Las diferencias en los rangos se reflejan en la diferente desviación típica de estas distribuciones. Si calculamos las desviaciones típicas de las medias en las cuatro distribuciones en estudio, obtenemos los valores siguientes:

	Desviaciones típicas observadas de distribuciones de medias	
	$n = 5$	$n = 35$
Longitudes del ala	1,778	0,584
Rendimientos de leche	5,040	1,799

Nótese que las desviaciones típicas de las medias de muestras basadas en 35 ítems son considerablemente menores que las basadas en 5 ítems. Esto es también intuitivamente obvio. Las medias basadas en muestras grandes deberían aproximarse a la media paramétrica y por tanto variarán mucho menos que las medias basadas en muestras pequeñas. Por lo tanto, la varianza de las medias es parcialmente función del tamaño de las muestras en que se basan las medias. También es función de la varianza de los ítems en las muestras. Así, en la tabla anterior, las medias de los rendimientos de leche tienen una desviación típica mucho mayor que las medias de las longitudes del ala basadas en tamaño de muestra comparable, simplemente porque la desviación típica de los rendimientos de leche individuales (11,1597) es considerablemente superior a la de las longitudes del ala individuales (3,90).

Los estadísticos matemáticos han calculado el valor esperado de la varianza de medias. Se entiende por *valor esperado* el valor medio obtenido por muestreo infinitamente repetido. Así, si se extrajesen repetidamente muestras de a medias procedentes de n ítems y se calculase la varianza de estas a medias cada vez, el valor medio de estas varianzas sería el valor esperado. Para la varianza de medias basadas en n ítems el valor esperado es

$$\sigma^2_{\bar{Y}} = \frac{\sigma^2}{n} \tag{6.1}$$

Consecuentemente la desviación típica esperada de medias es

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} \tag{6.2}$$

En esta fórmula queda claro que la desviación típica de medias es función de la desviación típica de ítems así como del tamaño de muestra de las medias. Cuanto mayor sea el tamaño de muestra, menor será la desviación típica de las medias. De hecho, al aumentar el tamaño de muestra hasta un número muy grande, la desviación típica de las medias se anula. Esto es muy razonable. Los tamaños de muestra muy grandes, promediando muchas observaciones, producirían estimaciones de medias menos variables que las basadas en unos pocos ítems.

Cuando trabajamos con muestras de una población, naturalmente no conocemos su media paramétrica sino que solamente podemos obtener una estimación de muestreo s, de esta última. Además, sería improbable que tuviésemos numerosas muestras de tamaño n de las que calcular directamente la desviación típica de medias. Por lo tanto, ordinariamente tenemos que estimar la desviación típica de medias a partir de una muestra única utilizando la expresión (6.2), sustituyendo s por σ:

$$s_{\bar{Y}} = \frac{s}{\sqrt{n}} \tag{6.3}$$

Así, a partir de la desviación típica de una sola muestra, obtenemos un estimador de la desviación típica de medias que esperaríamos si fuésemos a obtener un conjunto de medias basadas en muestras de igual tamaño de n ítems de la misma población. Como

TABLA 6.2

Medias, desviaciones típicas y desviaciones típicas de medias (errores típicos) de cinco muestras al azar de 5 y 35 longitudes del ala de moscas domésticas y rendimientos de leche de vacas Jersey, respectivamente. (Datos de la tabla 5.1.) Los valores paramétricos de los estadísticos se dan en la sexta línea de cada categoría.

	(1) $\bar{Y}$	(2) s	(3) $s_{\bar{Y}}$
	Longitudes del ala		
	45,8	1,095	0,490
	45,6	3,209	1,435
$n = 5$	43,6	4,827	2,159
	44,8	4,764	2,131
	46,8	1,095	0,490
	$\mu = 45{,}5$	$\sigma = 3{,}90$	$\sigma_{\bar{Y}} = 1{,}744$
	45,37	3,812	0,644
	45,00	3,850	0,651
$n = 35$	45,74	3,576	0,604
	45,29	4,198	0,710
	45,91	3,958	0,669
	$\mu = 45{,}5$	$\sigma = 3{,}90$	$\sigma_{\bar{Y}} = 0{,}659$
	Rendimientos de leche		
	66,0	6,205	2,775
	61,6	4,278	1,913
$n = 5$	67,6	16,072	7,188
	65,0	14,195	6,348
	62,2	5,215	2,332
	$\mu = 66{,}61$	$\sigma = 11{,}160$	$\sigma_{\bar{Y}} = 4{,}991$
	65,429	11,003	1,860
	64,971	11,221	1,897
$n = 35$	66,543	9,978	1,687
	64,400	9,001	1,521
	68,914	12,415	2,099
	$\mu = 66{,}61$	$\sigma = 11{,}160$	$\sigma_{\bar{Y}} = 1{,}886$

veremos, este estimador de la desviación típica de una media es un estadístico muy importante y frecuentemente utilizado.

La tabla 6.2 presenta algunas estimaciones de las desviaciones típicas de medias que pudieran obtenerse de muestras al azar de las dos poblaciones que hemos estado discutiendo. Las medias de 5 muestras de longitudes de ala basadas en 5 individuos varían de 43,6 a 46,8, sus desviaciones típicas de 1,095 a 4,827 y la estimación de la desviación típica de las medias de 0,490 a 2,159.

Los rangos para las otras categorías de muestras de la tabla 6.2 incluyen igualmente los valores paramétricos de estos estadísticos. Las estimaciones de las desviaciones típicas de las medias de los rendimientos de leche se agrupan en torno al valor esperado ya que no son dependientes de la normalidad de las variantes. Sin embargo, en una muestra particular en la que por azar la desviación típica de muestreo es un estimador incorrecto de la desviación típica de población (como en la segunda muestra de 5 rendimientos de leche), el estimador de la desviación típica de medias está igualmente alejado del valor paramétrico.

Hay cierta diferencia entre la desviación típica de ítems y la de medias de muestreo. Si estimamos una desviación típica de población por medio de la desviación típica de una muestra, la magnitud de la estimación no cambiará al aumentar el tamaño de muestra. Podemos esperar que la estimación mejore y se aproxime a la desviación típica de la población. Sin embargo, su magnitud será la misma si la muestra está basada en 3, 30 ó 3 000 individuos. Esto puede verse claramente en la tabla 6.2. Los valores de s se aproximan más a σ en las muestras basadas en $n = 35$ que en muestras de $n = 5$. Pero la magnitud general es la misma en ambos ejemplos. Sin embargo, la desviación típica de medias disminuye al aumentar el tamaño de muestra como se deduce de la expresión (6.3). Así, las medias basadas en 3 000 ítems tendrán una desviación típica solamente la décima parte que la de medias basadas en 30 ítems. Esto es evidente según

$$\frac{s}{\sqrt{3000}} = \frac{s}{\sqrt{30} \times \sqrt{100}} = \frac{s}{\sqrt{30} \times 10}$$

Puesto que nuestras desviaciones típicas de muestra presentan considerable variación (véase tabla 6.2), las estimaciones de las desviaciones típicas de las medias varían también en consecuencia. Una estimación defectuosa de σ dará lugar a una mala estimación de $\sigma_{\bar{Y}}$.

6.2 Distribución y varianza de otros estadísticos

Lo mismo que hemos obtenido una media y una desviación típica de cada muestra de longitudes del ala y rendimientos de leche, podríamos haber obtenido también otros estadísticos de cada muestra, tales como una varianza, una mediana o un coeficiente de variación. Tras repetido muestreo y cálculo obtendríamos distribuciones de frecuencias para estos estadísticos y podríamos calcular sus desviaciones típicas lo mismo que hicimos para las distribuciones de frecuencias de las medias. En muchos casos los estadísticos se distribuyen normalmente como ocurría para las medias. En otros, los estadísticos sólo estarán normalmente distribuidos si están basados en muestras de una población normalmente distribuida, o en muestras grandes, o si se cumplen estas dos condiciones. En algunos casos, como en las varianzas, su distribución nunca es normal. Esto se hace patente en la figura 6.5, la cual muestra una distribución de frecuencias de las varianzas de las 1 400 muestras de longitudes del ala de moscas domésticas. Observamos que la distribución está fuertemente inclinada a la derecha, lo cual es característico de la distribución de varianzas.

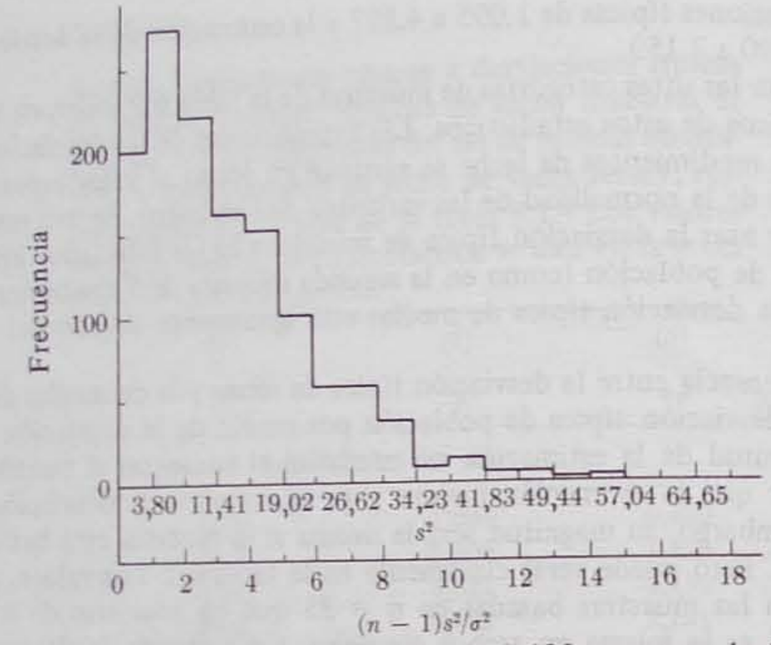


Fig. 6.5. Histograma de varianzas basadas en 1 400 muestras de 5 longitudes del ala de moscas domésticas de la tabla 5.1. La abscisa se da en términos de s^2 y $(n-1)s^2/\sigma^2$.

Las desviaciones típicas de diversos estadísticos se conocen generalmente como *errores típicos*. A veces los principiantes se desconciertan por una distinción imaginada entre desviaciones típicas y errores típicos. El error típico de un estadístico tal como la media (ó el *CV*), es la desviación típica de una distribución de medias (o de coeficientes de variación) para muestras de un determinado tamaño n. Así, los términos error típico y desviación típica se utilizan como sinónimos con la siguiente excepción: no es habitual utilizar error típico como sinónimo de desviación típica de los ítems en una muestra o población. El error típico o la desviación típica tienen que ser cualificados para hacer referencia a un estadístico determinado tal como la desviación típica del *CV* que es lo mismo que el error típico del *CV*. Convencionalmente, el término "error típico" utilizado sin ninguna limitación significa error típico de la media. "Desviación típica", usada sin restricción significa generalmente desviación típica de los ítems en una muestra o población. Así, cuando se lea que en una tabla se muestran medias, desviaciones típicas, errores típicos, y coeficientes de variación, esto significa que se presentan medias aritméticas, desviaciones típicas de ítems en muestras, desviaciones típicas de sus medias (= errores típicos de medias) así como coeficientes de variación. El siguiente resumen de términos puede ser útil:

$$\text{Desviación típica} = s = \sqrt{\sum y^2/(n-1)}$$
$$\text{Desviación típica de un estadístico } St = \text{error típico de un estadístico } St = s_{St}$$
$$\text{Error típico} = \text{error típico de una media} = \text{desviación típica de una media} = s_{\bar{Y}}$$

Los errores típicos no se obtienen usualmente de una distribución de frecuencias por muestreo repetido, sino que se estiman a partir de una sola muestra y representan la desviación típica esperada del estadístico en caso de que se hubiesen obtenido un gran número de tales muestras. Se recordará que en la sección previa hemos estimado de este modo el error típico de una distribución de medias a partir de una sola muestra.

El cuadro 6.1 presenta los errores típicos de cuatro estadísticos comunes. En la columna (1) se encuentra el estadístico cuyo error típico se describe; la columna (2) muestra la fórmula para el error típico estimado; la columna (3) da los grados de libertad en los que está basado el error típico (su aplicación se explica en la sección 6.5); la columna (4) ofrece observaciones sobre el rango de aplicación del error típico. Las aplicaciones de estos errores típicos se aclararán en las secciones siguientes.

CUADRO 6.1

Errores típicos de estadísticos usuales

(1)		(2)	(3)	(4)
	Estadístico	*Estimador del error típico*	*gl*	*Rangos de aplicación*
1	$\bar{Y}$	$s_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{s_Y}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}$	$n-1$	Válido para cualquier población con varianza finita.
2	Mediana	$s_{\text{med}} = (1{,}2533)s_{\bar{Y}}$	$n-1$	Muestras grandes de poblaciones normales.
3	s	$s_s = (0{,}7071068)\frac{s}{\sqrt{n}}$	$n-1$	Muestras de poblaciones normales. $(n > 15)$
4	CV	$s_{CV} = \frac{CV}{\sqrt{2n}}\sqrt{1+2\left(\frac{CV}{100}\right)^2}$	$n-1$	Muestras de poblaciones normales.
		$s_{CV} \approx \frac{CV}{\sqrt{2n}}$	$n-1$	Utilizado cuando $CV < 15$.

6.3 Introducción a límites de confianza

Los diversos estadísticos de muestra que hemos obtenido, tales como las medias o las desviaciones típicas, son estimadores de los parámetros de población μ o σ, respectivamente. Hasta ahora no hemos discutido la fiabilidad de estos estimadores. En primer lugar deseamos saber si los estadísticos de muestra son *estimadores no sesgados* de los parámetros de población, como se discutió en la sección 3.7. Pero saber por ejemplo que $\bar{Y}$ es un estimador no sesgado de μ no es suficiente. Nos gustaría hallar hasta qué punto es fiable una medida de μ. Los verdaderos valores de los parámetros casi siempre permanecen desconocidos y ordinariamente estimamos la fiabilidad de un estadístico muestral fijándole límites de confianza.

Para iniciar nuestra discusión de este asunto vamos a comenzar con el caso infrecuente de una población cuya media y desviación típica paramétricas son conocidas, μ y σ,

respectivamente. La media de una muestra de n ítems se simboliza por $\bar{Y}$. El error típico esperado de la media es $\sigma/\sqrt{n}$. Como hemos visto, las medias estarán normalmente distribuidas. Por lo tanto, según la sección 5.3, la región desde $1{,}96\sigma/\sqrt{n}$ por debajo de μ hasta $1{,}96\sigma/\sqrt{n}$ por encima de μ incluye el 95 % de las medias de muestreo de tamaño n. Otra forma de establecer esto es considerar la razón $(\bar{Y} - \mu)/(\sigma/\sqrt{n})$. Esta es la desviación típica de una media de muestreo respecto a la media paramétrica. Puesto que están normalmente distribuidas, el 95 % de tales desviaciones típicas se hallarán comprendidas entre $-1{,}96$ y $+1{,}96$. Podemos expresar simbólicamente este enunciado como sigue:

$$P\left\{-1{,}96 \leq \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq +1{,}96\right\} = 0{,}95$$

Esto significa que la probabilidad P de que las medias de muestreo $\bar{Y}$ difieran de la media paramétrica μ en no más de 1,96 errores típicos $\sigma/\sqrt{n}$ es igual a 0,95. La expresión entre corchetes es una desigualdad, en la cual todos sus términos pueden multiplicarse por $\sigma/\sqrt{n}$ para dar

$$\{-1{,}96\sigma/\sqrt{n} \leq (\bar{Y} - \mu) \leq +1{,}96\sigma/\sqrt{n}\}$$

Podemos volver a escribir esta expresión como

$$\{-1{,}96\sigma/\sqrt{n} \leq (\mu - \bar{Y}) \leq +1{,}96\sigma/\sqrt{n}\}$$

porque $-a \leqslant b \leqslant a$ implica que $a \geqslant -b \geqslant -a$, que puede escribirse como $-a \leqslant -b \leqslant a$. Y finalmente podemos pasar $-\bar{Y}$ al otro lado de los signos de desigualdad lo mismo que en una ecuación podría pasarse al otro lado del signo igual. Esto da la expresión final deseada:

$$P\left\{\bar{Y} - \frac{1{,}96\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + \frac{1{,}96\sigma}{\sqrt{n}}\right\} = 0{,}95 \tag{6.4}$$

o

$$P\{\bar{Y} - 1{,}96\sigma_{\bar{Y}} \leq \mu \leq \bar{Y} + 1{,}96\sigma_{\bar{Y}}\} = 0{,}95 \tag{6.4a}$$

Esto quiere decir que la probabilidad P de que el término $\bar{Y} - 1{,}96\sigma_{\bar{Y}}$ sea igual o menor que la media paramétrica μ y que el término $\bar{Y} + 1{,}96\sigma_{\bar{Y}}$ sea igual o superior que μ es 0,95. A los dos términos $\bar{Y} - 1{,}96\sigma_{\bar{Y}}$ y $\bar{Y} + 1{,}96\sigma_{\bar{Y}}$, L_1 y L_2 respectivamente, se les llama *límites de confianza* inferior y superior de la media al 95 %.

Otra forma de establecer la interrelación indicada por la expresión (6.4a) es que si obtuviésemos repetidamente muestras de un tamaño n de la población y construyésemos estos límites para cada una, podríamos esperar que el 95 % de los intervalos entre estos límites contendrían a la verdadera media, y solamente el 5 % de los intervalos no contendrían a μ. El intervalo desde L_1 a L_2 se llama *intervalo de confianza.*

Si no se está satisfecho al obtener el intervalo de confianza que contiene a la verdadera

media sólo 95 veces entre 100, se puede emplear 2,576 como coeficiente en lugar de 1,960. Recuerde que el 99 % del área de la curva normal está comprendida entre $\mu \pm 2{,}576\sigma$. Así, para calcular límites de confianza al 99 %, se calculan las dos cantidades $L_1 = \bar{Y} - 2{,}576\sigma/\sqrt{n}$ y $L_2 = \bar{Y} + 2{,}576\sigma/\sqrt{n}$ como límites de confianza inferior y superior, respectivamente. En este caso, 99 de cada 100 intervalos de confianza obtenidos por muestreo repetido contendrían a la verdadera media. El nuevo intervalo de confianza es más amplio que el del 95 % (ya que hemos multiplicado por un coeficiente mayor). Si todavía no se estuviese satisfecho con la fiabilidad de dicho límite de confianza, se podría aumentar multiplicando el error típico de la media por 3,291 para obtener los límites de confianza al 99,9 %. Este valor (ó 3,890 para límites del 99,99 %) podría hallarse por interpolación inversa en una tabla más amplia de áreas de la curva normal. El nuevo coeficiente haría aún más amplio el intervalo. Obsérvese que es posible construir intervalos de confianza que se espere contengan a μ un porcentaje de veces crecientemente superior. Primero se esperaría estar en lo cierto 95 veces de cada 100, luego 99 de cada 100 y finalmente 9 999 veces de cada 10 000. Pero conforme aumenta la confianza, la afirmación se hace más imprecisa ya que el intervalo de confianza se extiende. Vamos a examinar esto por medio de un ejemplo real.

Obtenemos una muestra de 35 longitudes del ala de moscas domésticas de la población de la tabla 5.1 de media y desviación típica conocidas ($\mu = 45{,}5$), ($\sigma = 3{,}90$). Vamos a suponer que la media de la muestra es 44,8. Podemos esperar que la desviación típica de medias basadas en muestras de 35 ítems sea $\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 3{,}90/\sqrt{35} = 0{,}6592$. Calculamos los límites de confianza como sigue:

el límite inferior es $L_1 = 44{,}8 - (1{,}960)(0{,}6592) = 43{,}51$
el límite superior es $L_2 = 44{,}8 + (1{,}960)(0{,}6592) = 46{,}09$

Recordemos que este es un caso infrecuente en el cual conocemos la verdadera media de la población ($\mu = 45{,}5$) y por ello sabemos que los límites de confianza incluyen a la media. Esperamos que el 95 % de los intervalos de confianza obtenidos en muestreo repetido incluyan a la media paramétrica. Podríamos incrementar la fiabilidad de estos límites recurriendo a intervalos de confianza al 99 %, reemplazando 1,960 por 2,576 en la expresión anterior y obteniendo $L_1 = 43{,}10$ y $L_2 = 46{,}50$. Podríamos conseguir una mayor confianza en que nuestro intervalo cubra a la media, pero estaríamos menos seguros acerca del verdadero valor de la media debido a la mayor amplitud de los límites. Aumentando todavía más el grado de confianza, es decir, hasta el 99,99 %, podríamos estar virtualmente seguros de que nuestros límites de confianza ($L_1 = 42{,}24$ $L_2 = 47{,}36$) contienen a la media de la población, pero los límites que incluyen a la media son ahora tan amplios que hacen nuestra predicción mucho menos útil que anteriormente.

Experimento 6.2. Para las siete muestras de 5 longitudes del ala de moscas domésticas y las siete muestras similares de rendimientos de leche manejadas últimamente en el experimento 6.1 (sección 6.1), calcular límites de confianza al 95 % de la media paramétrica para cada muestra y para la muestra total basada en 35 ítems. Basar los errores típicos de las medias en las desviaciones típicas de estas poblaciones (longitudes del ala de moscas domésticas $\sigma = 3{,}90$, rendimientos de leche $\sigma = 11{,}1597$). Indicar cuántos, en cada una de las cuatro clases de límites de confianza (longitudes del ala y rendimientos de leche, $n = 5$

y $n = 35$) eran correctos, es decir, contenían a la media paramétrica de la población. Reúne tus resultados con los de otros miembros de la clase.

Hemos ensayado el experimento con un computador para las 200 muestras de 35 longitudes del ala cada una, calculando límites de confianza de la media paramétrica, utilizando el error típico paramétrico de la media, $\sigma_{\bar{Y}} = 0{,}6592$. De las 200 paralelas a la ordenada trazadas en los intervalos de confianza, 194 (97,0 %) cortan a la media paramétrica de la población.

Para reducir la amplitud del intervalo de confianza tenemos que reducir el error típico de la media. Puesto que $\sigma_{\bar{Y}} = \sigma/\sqrt{n}$, esto solamente puede hacerse reduciendo la desviación típica de los ítems o aumentando el tamaño de muestra. La primera de estas alternativas frecuentemente no es accesible. Si muestreamos de una población en la naturaleza, ordinariamente no tenemos medio de reducir su desviación típica. Sin embargo, en muchos procedimientos experimentales podemos reducir la varianza de los datos. Por ejemplo, si estudiamos el peso del corazón en ratas y encontramos que su varianza es muy grande, es posible que fuésemos capaces de reducir esta varianza cogiendo ratas de una sola generación, en las cuales la variación de peso del corazón sería considerablemente menor. Así, controlando una de las variables del experimento se reduce la varianza de la variable respuesta, el peso del corazón. Del mismo modo, manteniendo constante la temperatura u otras variables ambientales en un procedimiento, frecuentemente podemos reducir la varianza de nuestra variable respuesta y por tanto obtener estimaciones más precisas de parámetros de población.

Una forma más habitual de reducir el error típico es aumentando el tamaño de muestra. Según la expresión 6.2 es evidente que el error típico disminuye al aumentar n; por consiguiente, cuando n tiende a infinito, el error típico y las longitudes de los intervalos de confianza tienden a cero. Esto concuerda con lo que hemos visto antes: en las muestras cuyo tamaño tiende a infinito, la media muestral tendería a la media paramétrica.

Debemos evitar un error habitual al expresar el significado de los límites de confianza de un estadístico. Cuando hemos establecido límites inferior y superior (L_1 y L_2, respectivamente) para un estadístico, queremos decir que la probabilidad de que este intervalo cubra a la media es 0,95, ó expresado de otro modo, que por término medio 95 de cada 100 intervalos de confianza obtenidos del mismo modo cubrirían a la media. *No podemos afirmar* que hay una probabilidad de 0,95 de que la verdadera media esté contenida dentro de un par determinado de límites de confianza, aunque esto pueda parecer que dice lo mismo. La última afirmación es incorrecta porque la verdadera media es un parámetro; de aquí que sea un valor fijo y por lo tanto esté dentro o fuera del intervalo. No puede estar dentro del intervalo dado el 95 % de las veces. Es importante pues aprender la exposición y el significado correcto de los límites de confianza.

Hasta ahora solamente hemos considerado medias basadas en muestras normalmente distribuidas con desviaciones típicas paramétricas conocidas. Sin embargo, podemos extender los métodos recién aprendidos a muestras de poblaciones con desviaciones típicas desconocidas pero en las cuales se sepa que la población sigue la ley normal y las muestras sean grandes, es decir $n \geq 100$. En tales casos utilizamos la desviación típica de muestreo para calcular el error típico de la media.

No obstante, cuando las muestras son pequeñas ($n < 100$) y no conocemos la desviación típica paramétrica, debemos considerar la fiabilidad de nuestra desviación típica de muestreo. Para hacer esto debemos valernos de la llamada distribución t o de Student. En la sección 6.5 aprenderemos cómo fijar límites de confianza utilizando la distribución t. Sin embargo, antes tendremos que familiarizarnos con esta distribución en la próxima sección.

6.4 Distribución t de Student

Las desviaciones $\bar{Y} - \mu$ de las medias de muestreo respecto de la media paramétrica de una distribución normal están normalmente distribuidas. Si estas desviaciones se dividen por la desviación típica paramétrica, $(\bar{Y} - \mu)/\sigma_{\bar{Y}}$ siguen estando normalmente distribuidas, con $\mu = 0$ y $\sigma = 1$. La sustracción de la constante μ a cada Y_i es simplemente una codificación aditiva (sección 3.8) y no alterará la forma de la distribución de medias, que es normal (sección 6.1). Dividiendo cada desviación por la constante $\sigma_{\bar{Y}}$ la varianza se reduce a la unidad, pero proporcionalmente lo mismo para la distribución total, de modo

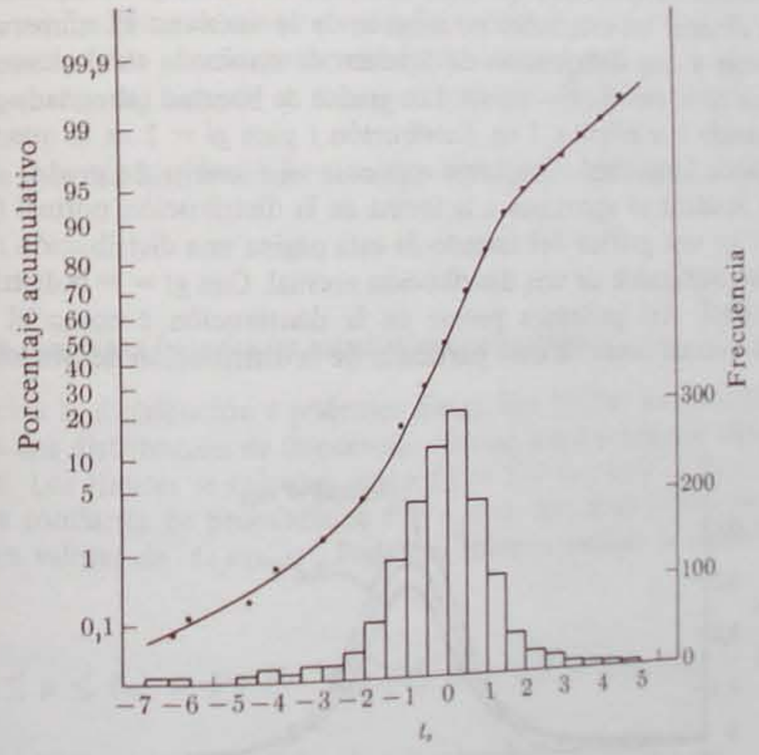


Fig. 6.6. Distribución a lo largo de la abscisa de la cantidad $t_s = (\bar{Y} - \mu)/s_{\bar{Y}}$ calculada para 1 400 muestras de 5 longitudes del ala de moscas domésticas representada como un histograma y como una distribución de frecuencias acumulativa. La ordenada de la derecha representa las frecuencias para el histograma, la de la izquierda la frecuencia acumulativa en escala probabilística.

que no se altera su forma y una distribución previamente normal continúa de igual modo.

Si, por otra parte, hubiésemos calculado la varianza s_i^2 de cada muestra y la desviación para cada media $\bar{Y}_i$ como $(\bar{Y}_i - \mu)/s_{\bar{Y}_i}$, donde $s_{\bar{Y}_i}$ representa el estimador del error típico de la media de la muestra i, habríamos encontrado que la distribución de las desviaciones es más abierta y aplanada que la distribución normal. Esto se representa en la figura 6.6, que muestra la razón $(\bar{Y}_i - \mu)/s_{\bar{Y}_i}$ para las 1 400 muestras de cinco longitudes del ala de moscas domésticas de la tabla 6.1. La nueva distribución fluctúa más ampliamente que la distribución normal correspondiente porque el denominador es el error típico de muestreo en lugar del paramétrico,y unas veces será menor y otras mayor que el esperado. Este aumento de variación se reflejará en la mayor varianza de la razón $(\bar{Y} - \mu)/s_{\bar{Y}}$. La distribución esperada de esta razón es la llamada distribución t conocida también como distribución "de Student" cuyo nombre se debe a W.S. Gosset que la describió por primera vez publicando bajo el seudónimo "Student". La distribución t es una función con una fórmula matemática complicada que no es necesario presentar aquí.

La distribución t comparte con la normal las propiedades de que es simétrica y se extiende del infinito negativo al positivo. Sin embargo, difiere de la normal en que adopta diferentes formas dependiendo del número de grados de libertad. Por grados de libertad nos referimos a la cantidad $n - 1$, donde n es el tamaño de muestra en el cual se ha basado una varianza. Se recordará que esta cantidad $n - 1$ es el divisor de una suma de cuadrados para obtener un estimador no sesgado de la varianza. El número de grados de libertad pertinente a una distribución de Student determinada es el mismo que el de la desviación típica en la razón $(\bar{Y} - \mu)/s_{\bar{Y}}$. Los grados de libertad (abreviado *gl* o a veces μ) pueden variar desde 1 a infinito. Una distribución t para $gl = 1$ es la que más marcadamente se desvía de la normal. Conforme aumenta el número de grados de libertad, la distribución de Student se aproxima a la forma de la distribución normal ($\mu = 0, \sigma = 1$) cada vez más, y en una gráfica del tamaño de esta página una distribución t de $gl = 30$ es esencialmente indistinguible de una distribución normal. Con $gl = \infty$ la distribución t es la distribución normal. Así podemos pensar en la distribución t como el caso general, considerando la normal como un caso particular de la distribución de Student con $gl = \infty$.

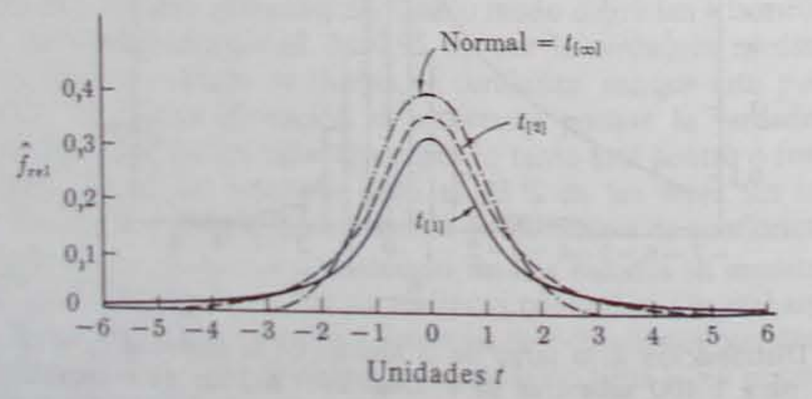


Fig. 6.7. Curvas de frecuencia de distribuciones t para 1 y 2 grados de libertad, comparadas con la distribución normal.

La figura 6.7 presenta distribuciones t para 1 y 2 grados de libertad comparadas con una distribución de frecuencias normal.

Hemos podido utilizar una sola tabla para las áreas de la curva normal codificando el argumento en unidades de desviación típica. Sin embargo, como las distribuciones t difieren en su forma para diferentes grados de libertad, será necesario tener una tabla t separada, correspondiente en estructura a la tabla de áreas de la curva normal, para cada valor de *gl*. Esto supondría juegos de tablas muy complicados y difíciles de manejar. Por tanto, las tablas t convencionales se preparan diferentemente. La tabla III presenta grados de libertad y probabilidad como argumentos, y los valores correspondientes de t como funciones. Las probabilidades indican el porcentaje del área en ambas colas de la curva (a la derecha e izquierda de la media) más allá del valor indicado de t. Así, buscando el *valor crítico* de t a una probabilidad $P = 0{,}05$ y $gl = 5$, encontramos $t = 2{,}571$ en la tabla III. Como ésta es una tabla de dos colas, la probabilidad de 0,05 significa que habrá 0,025 del área en cada cola a partir de un valor t de ± 2,571. Se recordará que el valor correspondiente para infinitos grados de libertad (para la curva normal) es 1,960. En la tabla III solo se presentan las probabilidades generalmente utilizadas. Se deben familiarizar completamente con la búsqueda de valores de t en esta tabla. Es una de las tablas de consulta más importantes. Un simbolismo completamente convencional es $t_{\alpha[\nu]}$, que representa el valor t de la tabla para ν grados de libertad y proporción α en las dos colas ($\alpha/2$ en cada una), el cual es equivalente al valor t para la probabilidad acumulativa de $1 - (\alpha/2)$. Para familiarizarse con la tabla procúrese buscar algunos de estos valores. Por ejemplo, asegúrese de que $t_{0,05[7]}$, $t_{0,01[3]}$, $t_{0,02[10]}$, y $t_{0,05[\infty]}$ corresponden a 2,365, 5,841, 2,764 y 1,960, respectivamente.

Utilizaremos ahora la distribución t para fijar límites de confianza a medias de muestras pequeñas.

6.5 Límites de confianza basados en estadísticos de muestreo

Una vez conocida la distribución t podemos ahora fijar límites de confianza a las medias de muestras de una distribución de frecuencias normal, cuya desviación típica paramétrica es desconocida. Los límites se calculan como $L_1 = \bar{Y} - t_{\alpha[n-1]}s_{\bar{Y}}$ y $L_2 = \bar{Y} + t_{\alpha[n-1]}s_{\bar{Y}}$ para límites de confianza de probabilidad $P = 1 - \alpha$. Así, para límites de confianza del 95 % utilizamos valores de $t_{0,05[n-1]}$. Podemos volver a escribir la expresión 6.4a como

$$P\{L_1 \leq \mu \leq L_2\} = P\{\bar{Y} - t_{\alpha[n-1]}s_{\bar{Y}} \leq \mu \leq \bar{Y} + t_{\alpha[n-1]}s_{\bar{Y}}\} = 1 - \alpha \qquad (6.5)$$

En el cuadro 6.2 se presenta un ejemplo de aplicación de esta expresión. Por medio de un experimento de muestreo podemos convencernos de la idoneidad de la distribución t para fijar límites de confianza a medias de muestras de una población normalmente distribuida con σ desconocida.

CUADRO 6.2

Límites de confianza para μ.

Longitudes del fémur de hembras apomícticas de áfidos de los cuadros 2.1 y 3.1: $\bar{Y} = 4{,}004$; $s = 0{,}366$; $n = 25$.

Valores para $t_{\alpha[n-1]}$ de una tabla t de dos colas (tabla **III**), donde $1 - \alpha$ es la proporción que expresa el grado de confianza y $n - 1$ son los grados de libertad:

$$t_{0{,}05[24]} = 2{,}064 \qquad t_{0{,}01[24]} = 2{,}797$$

Los límites de confianza del 95 % para la media de población, μ, vienen dados por las ecuaciones:

$$L_1 \text{ (límite inferior)} = \bar{Y} - t_{0{,}05[n-1]}\frac{s}{\sqrt{n}}$$

$$= 4{,}004 - \left(2{,}064\,\frac{0{,}366}{\sqrt{25}}\right) = 4{,}004 - 0{,}151$$

$$= 3{,}853$$

$$L_2 \text{ (límite superior)} = \bar{Y} + t_{0{,}05[n-1]}\frac{s}{\sqrt{n}}$$

$$= 4{,}004 + 0{,}151$$

$$= 4{,}155$$

Los límites de confianza del 99 % son:

$$L_1 = \bar{Y} - t_{0{,}01[24]}\frac{s}{\sqrt{n}}$$

$$= 4{,}004 - \left(2{,}797\,\frac{0{,}366}{\sqrt{25}}\right) = 4{,}004 - 0{,}205$$

$$= 3{,}799$$

$$L_2 = \bar{Y} + t_{0{,}01[24]}\frac{s}{\sqrt{n}}$$

$$= 4{,}004 + 0{,}205$$

$$= 4{,}209$$

Experimento 6.3. Repetir los cálculos y procedimientos del experimento 6.2 (sección 6.3), pero basar los errores típicos de las medias en las desviaciones típicas calculadas para cada muestra y utilizar el valor t apropiado en lugar de una desviación típica normal.

La figura 6.8 presenta límites de confianza del 95 % de 200 medias de muestreo de 35 longitudes del ala de moscas domésticas, calculados con t y $s_{\bar{Y}}$ en vez de con la curva

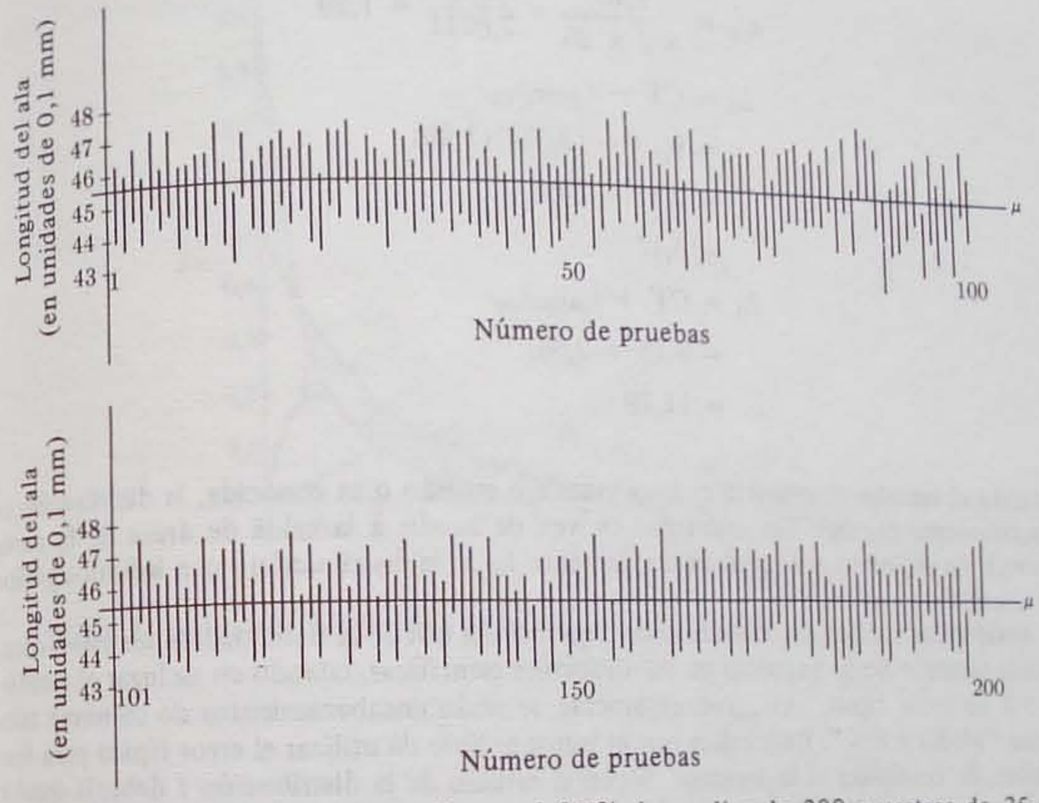


Fig. 6.8. Intervalos de confianza al 95 % de medias de 200 muestras de 35 longitudes del ala de moscas domésticas, basadas en errores típicos de muestreo $s_{\bar{Y}}$. La línea continua horizontal es la media paramétrica μ. La ordenada representa la variable.

normal y $\sigma_{\bar{Y}}$. Observamos que 191 (95,5 %) de los 200 intervalos de confianza cortan a la media paramétrica.

Podemos utilizar la misma técnica para fijar límites de confianza a cualquier estadístico siempre que siga la distribución normal. Esto se aplicará de manera aproximada a todos los estadísticos del cuadro 6.1. Así, por ejemplo, podemos fijar límites de confianza al coeficiente de variación de las longitudes del fémur de áfidos de los cuadros 2.1 y 3.1. Estos se calculan como

$$P\{CV - t_{\alpha[n-1]}s_{CV} \leq CV_P \leq CV + t_{\alpha[n-1]}s_{CV}\} = 1 - \alpha$$

donde CV_P simboliza el valor paramétrico del coeficiente de variación. Como el error típico del coeficiente de variación corresponde aproximadamente a $s_{CV} = CV/\sqrt{2n}$, procederemos como sigue:

$$CV = \frac{100s}{\bar{Y}} = \frac{100(0{,}3656)}{4{,}004} = 9{,}13$$

$$s_{CV} = \frac{9{,}13}{\sqrt{2 \times 25}} = \frac{9{,}13}{7{,}0711} = 1{,}29$$

$$\begin{aligned} L_1 &= CV - t_{{,}05[24]} s_{CV} \\ &= 9{,}13 - (2{,}064)(1{,}29) \\ &= 9{,}13 - 2{,}66 \\ &= 6{,}47 \\ L_2 &= CV + t_{{,}05[24]} s_{CV} \\ &= 9{,}13 + 2{,}66 \\ &= 11{,}79 \end{aligned}$$

Cuando el tamaño de muestra es muy grande o cuando σ es conocida, la distribución es efectivamente normal. Sin embargo, en vez de acudir a la tabla de áreas de la curva normal, de ordinario utilizaremos simplemente $t_{\alpha[\infty]}$, la distribución t con infinitos grados de libertad.

Aunque los límites de confianza son una medida útil de la fiabilidad de un estadístico, ordinariamente no se expresan en publicaciones científicas, citando en su lugar el estadístico ± su error típico. Así, frecuentemente se verán encabezamientos de columna tales como "Media ± E.S.". Esto indica que el lector es libre de utilizar el error típico para fijar límites de confianza si le interesa. Según el estudio de la distribución t debería quedar claro que no es posible fijar límites de confianza a un estadístico sin conocer el tamaño de muestra n en que está basado, que es necesario para calcular los grados de libertad. Así, es sumamente lamentable la citación ocasional de media y errores típicos sin expresar también el tamaño muestral n.

Es importante expresar un estadístico y su error típico con un número suficiente de cifras decimales. La siguiente regla empírica sirve de ayuda. Dividir el error típico por tres, señalar el lugar decimal del primer dígito distinto de cero del cociente; expresar el estadístico con ese número de lugares decimales significativos y poner un decimal más para el error típico. Esta regla es bastante sencilla como lo demuestra un ejemplo. Si la media y el error típico de una muestra se calculan como 2,354 ± 0,363, dividimos 0,363 por 3 lo que da 0,121. Por tanto la media debería expresarse con una cifra decimal y el error típico con dos. Así expresamos este resultado como 2,4 ± 0,36. En cambio, si la misma media tuviera un error típico de 0,243, al dividir éste por 3 habría dado 0,081 y el primer dígito distinto de cero habría estado en el segundo lugar decimal. Por lo tanto, la media debería haberse dado como 2,35 ± 0,243.

6.6 La distribución ji-cuadrado

Otra distribución continua de gran importancia en estadística es la distribución de χ^2 (léase *ji-cuadrado*). Necesitamos aprenderla ahora en relación con la distribución y límites de confianza de varianzas.

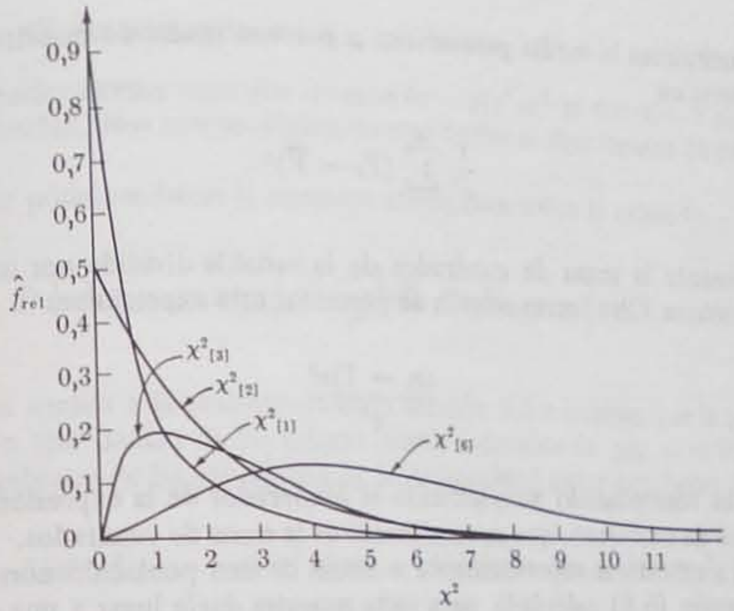


Fig. 6.9. Curvas de frecuencia de la distribución χ^2 para 1, 2, 3 y 6 grados de libertad.

La distribución ji-cuadrado es una función de densidad de probabilidad cuyos valores varían desde cero hasta el infinito positivo. Así, a diferencia de la distribución normal ó t, la función se aproxima asintóticamente al eje horizontal sólo en la cola derecha de la curva, no en ambas colas. La función que describe la distribución χ^2 es complicada y no se expondrá aquí. Como en t, no hay solamente una distribución χ^2 sino que hay una distribución para cada número de grados de libertad. Por lo tanto, χ^2 es función del número de grados de libertad ν. La figura 6.9 muestra funciones de densidad de probabilidad de las distribuciones χ^2 para 1, 2, 3 y 6 grados de libertad. Nótese que las curvas son marcadamente inclinadas hacia la derecha, en forma de ∪ al principio, pero más o menos acercándose a la simetría para grados de libertad superiores.

Podemos generar una distribución χ^2 de una población de desviaciones típicas normales. Se recordará que tipificamos una variable Y_i sometiéndola a la operación $(Y_i - \mu)/\sigma$. Vamos a simbolizar una variable tipificada por $Y'_i = (Y_i - \mu)/\sigma$. Imaginemos ahora muestras repetidas de n variantes Y_i de una población normal con media μ y desviación típica σ. Para cada muestra transformamos cada variante Y_i en Y'_i como hemos definido más arriba. Las cantidades $\sum^n Y'^2_i$ calculadas para cada muestra se distribuirán como una χ^2 con n grados de libertad. Utilizando la definición de Y'_i, podemos volver a escribir $\sum^n Y'^2_i$ como

$$\sum^n \frac{(Y_i - \mu)^2}{\sigma^2} = \frac{1}{\sigma^2} \sum^n (Y_i - \mu)^2 \qquad (6.6)$$

Cuando sustituimos la media paramétrica μ por una media de muestreo en esta expresión, se convierte en

$$\frac{1}{\sigma^2}\sum^{n}(Y_i - \overline{Y})^2 \tag{6.7}$$

que es simplemente la suma de cuadrados de la variable dividida por una constante, la varianza paramétrica. Otra forma común de presentar esta expresión es

$$\frac{(n-1)s^2}{\sigma^2} \tag{6.8}$$

en la cual se ha reemplazado simplemente el numerador de la expresión 6.7 por $n-1$ veces la varianza de muestreo, que naturalmente es la suma de cuadrados.

Si fuésemos a muestrear repetidamente n ítems de una población normalmente distribuida, la expresión (6.8) calculada para cada muestra daría lugar a una distribución χ^2 con $n-1$ grados de libertad. Nótese que aunque tenemos muestras de n ítems, hemos perdido un grado de libertad porque ahora estamos utilizando una media de muestreo en lugar de la media paramétrica. La figura 6.5, una distribución de muestreo de varianzas, tiene una segunda escala en la abscisa, que es la primera escala multiplicada por la constante $(n-1)/\sigma^2$. Esta escala convierte las varianzas de muestreo s^2 de la primera escala en la expresión (6.8). Puesto que la segunda escala es proporcional a s^2, la distribución de la varianza de muestreo servirá para representar una distribución de muestreo que aproxima χ^2. La distribución es marcadamente inclinada hacia la derecha como se esperaría en una distribución χ^2.

Las tablas χ^2 convencionales, como se demuestra en la tabla IV, dan los niveles de probabilidad ordinariamente requeridos y los grados de libertad como argumentos, y la χ^2 correspondiente a la probabilidad y a los *gl* como funciones. Cada ji-cuadrado en la tabla IV es el valor de χ^2 a partir del cual el área bajo la distribución χ^2 para ν grados de libertad representa la probabilidad indicada. Lo mismo que hemos utilizado subíndices para indicar la proporción acumulativa del área así como los grados de libertad representados por un valor determinado de t, los utilizaremos para χ^2 como sigue: $\chi^2_{\alpha[\nu]}$ indica el valor χ^2 a la derecha del cual se halla la proporción α del área bajo una distribución χ^2, para ν grados de libertad.

Vamos a aprender cómo se utiliza la tabla IV. Al observar la distribución de $\chi^2_{[2]}$ notamos que el 90 % de todos los valores de $\chi^2_{[2]}$ estaría a la derecha de 0,211, pero solamente el 5 % de los valores serían superiores a 5,991. Los estadísticos matemáticos han demostrado que el valor esperado de $\chi^2_{[\nu]}$ (la media de una distribución χ^2) es igual a sus grados de libertad ν. Así el valor esperado de una distribución $\chi^2_{[5]}$ es 5. Cuando examinamos valores del 50 % (las medianas) en la tabla χ^2, observamos que generalmente son inferiores al valor esperado (las medias). Así, para $\chi^2_{[5]}$ el punto 50 % es 4,351. Esto demuestra la asimetría de la distribución χ^2, estando la media a la derecha de la mediana. En la próxima sección se verá la primera aplicación de la distribución χ^2. Sin embargo, su más amplia utilidad se verá en relación con el capítulo 13.

6.7 Límites de confianza para varianzas

En la sección anterior hemos visto que la razón $(n-1)s^2/\sigma^2$ se distribuye como χ^2 con $n-1$ grados de libertad. Nos aprovechamos de este hecho al fijar límites de confianza a las varianzas.

En primer lugar podemos hacer la siguiente afirmación sobre la razón $(n-1)s^2/\sigma^2$:

$$P\left\{\chi^2_{(1-(\alpha/2))[n-1]} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{(\alpha/2)[n-1]}\right\} = 1-\alpha$$

Esta expresión es similar a la encontrada en la sección 6.3 e implica que la probabilidad P de que esta razón esté dentro de los valores límite indicados de $\chi^2_{[n-1]}$ es 0,95. La simple manipulación algebraica de las cantidades de la desigualdad entre corchetes da

$$P\{(n-1)s^2/\chi^2_{(\alpha/2)[n-1]} \leq \sigma^2 \leq (n-1)s^2/\chi^2_{(1-(\alpha/2))[n-1]}\} = 1-\alpha \tag{6.9}$$

Como $(n-1)s^2 = \sum y^2$, podemos simplificar la expresión (6.9) a

$$P\{\textstyle\sum y^2/\chi^2_{(\alpha/2)[n-1]} \leq \sigma^2 \leq \sum y^2/\chi^2_{(1-(\alpha/2))[n-1]}\} = 1-\alpha \tag{6.10}$$

Esta todavía parece una expresión complicada pero significa sencillamente que si dividimos la suma de cuadrados $\sum y^2$ por los dos valores de $\chi^2_{[n-1]}$ que limitan $1-\alpha$ del área de la distribución $\chi^2_{[n-1]}$, los dos cocientes encierran el verdadero valor de la varianza σ^2 con una probabilidad de $P = 1-\alpha$.

Un ejemplo numérico real aclarará esto. Supongamos que tenemos una muestra de 5 longitudes del ala de moscas domésticas con una varianza de muestreo de $s^2 = 13{,}52$. Si deseamos fijar límites de confianza del 95 % para la varianza paramétrica, hallamos el valor numérico de la expresión (6.10) para la varianza de muestreo s^2. Primero calculamos la suma de cuadrados para esta muestra: $4 \times 13{,}52 = 54{,}08$. Después buscamos los valores para $\chi^2_{0.025[4]}$ y $\chi^2_{0.975[4]}$. Como se piden límites de confianza del 95 %, en este caso α es igual a 0,05. Estos valores de χ^2 abarcan entre ellos el 95 % del área bajo la curva χ^2. Corresponden a 11,143 y 0,484, respectivamente, y los límites de la expresión (6.10) se convierten entonces en

$$L_1 = 54{,}08/11{,}143 \qquad \text{y} \qquad L_2 = 54{,}08/0{,}484$$

o

$$L_1 = 4{,}85 \qquad \text{y} \qquad L_2 = 111{,}74$$

Este intervalo de confianza es muy amplio pero no debemos olvidar que la varianza de muestreo está basada en 5 individuos solamente. Nótese además que el intervalo es asimétrico en torno a 13,52, la varianza de muestreo. Esto está en contraposición respecto a los intervalos de confianza encontrados anteriormente, los cuales eran simétricos en torno al estadístico de muestreo.

CUADRO 6.3

Límites de confianza para σ^2. Método de intervalos de confianza imparciales de mínima amplitud.

Longitudes del fémur de hembras apomícticas de áfidos de los cuadros 2.1 y 3.1: $n = 25$; $s^2 = 0{,}1337$.

Los factores de la tabla **VII** para $\nu = n - 1 = 24\ gl$ y coeficiente de confianza $(1 - \alpha) = 0{,}95$ son

$f_1 = 0{,}5943$ $\quad$ $f_2 = 1{,}8763$

y para un coeficiente de confianza de 0,99 son

$f_1 = 0{,}5139$ $\quad$ $f_2 = 2{,}3513$

Los límites de confianza del 95 % para la varianza de la población, vienen dados por las ecuaciones

L_1 = (límite inferior) $f_{18}{}^2 = 0{,}5943(0{,}1337) = 0{,}07946$

L_2 = (límite superior) $f_{28}{}^2 = 1{,}8763(0{,}1337) = 0{,}2509$

Los límites de confianza del 99 % son

$L_1 = f_{18}{}^2 = 0{,}5139(0{,}1337) = 0{,}06871$

$L_2 = f_{28}{}^2 = 2{,}3513(0{,}1337) = 0{,}3144$

El método descrito más arriba se denomina método de colas iguales porque en cada cola se sitúa el mismo valor de probabilidad (por ejemplo, 2 ½ %). Puede demostrarse que en vista de la asimetría de la distribución de varianzas, este método no produce los intervalos de confianza más cortos posibles. Puede desearse que el intervalo de confianza sea "el más corto" en el sentido de que la razón L_2/L_1 sea lo más pequeña posible. El cuadro 6.3 muestra cómo obtener estos intervalos de confianza imparciales más reducidos para σ^2 utilizando la tabla **VII**, basada en el método de Tate y Klett (1959). Esta tabla da $(n-1)/\chi^2_{[p,n-1]}$, donde p es un valor ajustado de $\alpha/2$ ó $1 - (\alpha/2)$ designada para producir los más cortos intervalos de confianza imparciales. El cálculo es muy simple.

6.8 Introducción al contraste de hipótesis

La aplicación más frecuente de la estadística en investigación biológica es probar ciertas hipótesis científicas. Los métodos estadísticos son importantes en biología porque ordinariamente los resultados de experimentos no están bien definidos, y por tanto se necesitan pruebas estadísticas para confirmar decisiones entre hipótesis alternativas. Una prueba estadística examina una serie de datos de muestreo, y sobre la base de una distribución esperada de los datos según una hipótesis determinada, lleva a la decisión de aceptar o rechazar dicha hipótesis y aceptar una alternativa. La naturaleza de las pruebas varía con los datos y las hipótesis, pero es común a todas ellas la misma filosofía general del

contraste de hipótesis, que se discutirá en esta sección. Estúdiese detenidamente la materia que se da a continuación porque es fundamental para comprender cada uno de los capítulos que siguen en este libro.

Nos gustaría refrescar la memoria sobre la muestra de 17 animales de la especie A, de los cuales 14 eran hembras y 3 machos. Estos datos fueron examinados con respecto a su ajuste a la distribución binomial presentada en la sección 4.2 y su análisis se muestra en la tabla 4.3. De esta tabla sacamos la conclusión de que si la proporción de sexos en la población fuese 1 : 1 ($p♀ = q♂ = 0{,}5$), la probabilidad de obtener una muestra con 14 machos y 3 hembras será 0,005188, lo que hace muy improbable que tal resultado pudiera obtenerse sólo por azar. Vimos que es convencional incluir todos los resultados "peores", es decir, todos los que se alejan aún más del resultado esperado según la hipótesis $p♀ = q♂ = 0{,}5$. Incluyendo todos los resultados peores la probabilidad es 0,006363, un valor todavía muy pequeño. El cálculo anterior está basado en la idea de una prueba de una cola, en el cual sólo estamos interesados en las desviaciones de la proporción de sexos 1 : 1 que muestran una preponderancia de hembras. Si no tenemos prejuicio sobre la dirección de las desviaciones de lo esperado, debemos calcular la probabilidad de obtener una muestra tan divergente de lo esperado como 14 hembras y 3 machos *en una y otra dirección*. Esto requiere la probabilidad de obtener una muestra de 3 hembras y 14 machos (y todas las muestras peores) o de obtener 14 hembras y 3 machos (y todas las muestras peores). Esta prueba es de dos colas y como la distribución es simétrica duplicamos la probabilidad discutida previamente para dar 0,012726.

¿Qué significa esta probabilidad? Nuestra hipótesis es que $p♀ = q♂ = 0{,}5$. Vamos a llamar a esta hipótesis H_0, la *hipótesis nula*, que es la que se contrasta. Se le llama hipótesis nula porque supone que no hay diferencia real entre el verdadero valor de p en la población de la cual hemos muestreado y el valor hipotético de $\hat{p} = 0{,}5$; por ejemplo, en el caso presente pensamos que la única razón de que nuestra muestra no ofrezca una proporción de sexos 1 : 1 es por causa de error de muestreo. Si la hipótesis nula $p♀ = q♂ = 0{,}5$ es cierta, entonces aproximadamente 13 muestras entre 1000 serán tan desviantes ó más que ésta en una u otra dirección *sólo por azar*. Así, es completamente *posible* que hayamos logrado por azar una muestra de 14 hembras y 3 machos, pero no es muy *probable* ya que un suceso tan desviante solamente ocurriría alrededor de 13 veces de cada 1 000 o el 1,3 % de las veces. Si efectivamente obtenemos tal muestra, podemos tomar una de dos decisiones. Podemos decidir que en realidad la hipótesis nula es cierta (es decir la proporción de sexos es 1 : 1) y que la muestra obtenida resulta ser precisamente una de las de la cola de la distribución, ó podemos decidir que una muestra tan desviante es un suceso demasiado improbable para justificar la aceptación de la hipótesis nula. Por lo tanto, podemos decidir que la hipótesis de que la proporción de sexos es 1 : 1 no es cierta. Una u otra de estas decisiones puede ser correcta dependiendo de la veracidad de la cuestión. Si de hecho la hipótesis 1 : 1 es correcta, la primera decisión (aceptar la hipótesis nula) será correcta. Si decidimos rechazar la hipótesis bajo estas circunstancias, cometemos un error. *El rechazo de una hipótesis nula cierta se denomina error de tipo I.* Por otra parte, si en realidad la verdadera proporción de sexos de la población es distinta de 1 : 1, la primera decisión (aceptar la hipótesis 1 : 1) es un error, denominado *error de tipo II*, que es *la aceptación de una hipótesis nula falsa.* Finalmente, si la hipótesis 1 : 1 no es cierta y decidimos rechazarla, tomamos nuevamente la decisión correcta. Así, hay dos

tipos de decisiones correctas, aceptar una hipótesis nula cierta y rechazar una hipótesis nula falsa, y dos tipos de errores, tipo I, rechazar una hipótesis nula cierta, y tipo II, aceptar una hipótesis nula falsa.

Antes de realizar una prueba tenemos que decidir qué magnitud de error de tipo I (rechazo de una hipótesis cierta) vamos a permitir. Incluso cuando extraemos muestras de una población de parámetros conocidos, siempre habrá algunas muestras que por casualidad sean muy desviantes. Las más desviantes de éstas es probable que nos induzcan a error haciéndonos creer que nuestra hipótesis H_0 es falsa. Si permitimos que un 5 % de las muestras nos lleven a un error del tipo I, entonces rechazaremos 5 de cada 100 muestras de la población, decidiendo que éstas no son de la población dada. En la distribución que se estudia, esto significa que rechazaríamos todas las muestras de 17 animales que tuviesen 13 de un sexo y 4 del otro. Esto puede verse recurriendo a la columna (3) de la tabla 6.3, donde se muestran las frecuencias esperadas de los diversos resultados según la hipótesis $p♀ = q♂ = 0{,}5$. Esta tabla es una ampliación de la tabla 4.3 anterior, que presentaba solamente una cola de esta distribución. Efectivamente, obtendríamos un error tipo I ligeramente menor del 5 % si sumásemos las frecuencias relativas esperadas de ambas colas, empezando por la clase de 13 de un sexo y 4 del otro. Según la tabla 6.3 puede verse que la frecuencia relativa esperada en las dos colas será $2 \times 0{,}0245209 = 0{,}0490418$. En una distribución de frecuencias discreta, tal como la binomial, no podemos calcular exactamente errores del 5 % como en una distribución continua, en la cual podemos medir exactamente el 5 % del área. Si nos decidimos por un error aproximado del 1 %, rechazaríamos la hipótesis $p♀ = q♂$ para todas las muestras de 17 animales que tuviesen 14 ó más de un sexo (en la tabla 6.3 vemos que la $\hat{f}_{rel.}$ en las colas es igual a $2 \times 0{,}0063629 = 0{,}0127258$). Así, cuanto más pequeño sea el error de tipo I que estamos dispuestos a tolerar, más desviante tiene que ser una muestra para que rechacemos la hipótesis nula H_0. De modo natural puede que se tienda a tener un error lo más pequeño posible. Puede decidirse trabajar con un error de tipo I sumamente pequeño, tal como 0,1 % o incluso 0,01 %, aceptando la hipótesis nula a no ser que la muestra sea extremadamente desviante. La dificultad de esta aproximación es que a pesar de precaver contra un error del primer tipo, se pudiera caer en un error del segundo tipo (tipo II) aceptando la hipótesis nula cuando en realidad no es cierta y sí lo es una hipótesis alternativa H_1. Luego veremos cómo sucede esto.

Primero vamos a aprender alguna terminología más. El error de tipo I se expresa más frecuentemente como una probabilidad y se simboliza por α. Cuando se expresa como un porcentaje se conoce también como *nivel de significación*. Así, un error de tipo I de $\alpha = 0{,}05$ corresponde a un nivel de significación del 5 % para una prueba determinada. Cuando en una distribución de frecuencias separamos áreas proporcionales a α, el error de tipo I, la porción de la abscisa bajo el área que se ha separado se llama *región de rechazo o región crítica* de una prueba, y la porción de la abscisa que llevaría a la aceptación de la hipótesis nula se denomina *región de aceptación*. La figura 6.10A es un diagrama de barras que indica la distribución esperada de resultados en el ejemplo de proporción de sexos, dada H_0. Las líneas discontinuas separan aproximadamente las regiones de rechazo del 1 %, de la región de aceptación del 99 %.

Ahora vamos a echar una ojeada más detenida al error de tipo II. Este es la probabilidad de aceptar la hipótesis nula cuando en realidad es falsa. Si se intenta evaluar la

TABLA 6.3

Frecuencias relativas esperadas para muestras de 17 animales según dos hipótesis. Distribución binomial.

(1) ♀♀	(2) ♂♂	(3) $H_0: p♀ = q♂ = \frac{1}{2}$ $\hat{f}_{rel}$	(4) $H_1: p♀ = 2q♂ = \frac{1}{3}$ $\hat{f}_{rel}$
17	0	0,0000076	0,0010150
16	1	0,0001297	0,0086272
15	2	0,0010376	0,0345086
14	3	0,0051880	0,0862715
13	4	0,0181580	0,1509752
12	5	0,0472107	0,1962677
11	6	0,0944214	0,1962677
10	7	0,1483765	0,1542104
9	8	0,1854706	0,0963815
8	9	0,1854706	0,0481907
7	10	0,1483765	0,0192763
6	11	0,0944214	0,0061334
5	12	0,0472107	0,0015333
4	13	0,0181580	0,0002949
3	14	0,0051880	0,0000421
2	15	0,0010376	0,0000042
1	16	0,0001297	0,0000002
0	17	0,0000076	0,0000000
	Total	1,0000002	0,9999999

probabilidad de error de tipo II, inmediatamente aparece un problema. Si la hipótesis nula H_0 es falsa, alguna otra hipótesis H_1 debe ser cierta. Pero a no ser que se pueda especificar H_1, no se está en condiciones de calcular el error de tipo II. Un ejemplo aclarará esto inmediatamente. Supongamos que en nuestro caso de proporción de sexos tenemos solamente dos posibilidades razonables, 1) nuestra hipótesis $H_0: p♀ = q♂$, ó 2) una hipótesis alternativa $H_1: p♀ = 2q♂$, la cual formula que la proporción de sexos es 2 : 1 a favor de hembras, de modo que $p♀ = \frac{2}{3}$ y $q♂ = \frac{1}{3}$. Ahora tenemos que calcular frecuencias esperadas para la distribución binomial $(p♀ + q♂)^N = (\frac{2}{3} + \frac{1}{3})^{17}$ para hallar las probabilidades de los diversos resultados según esta hipótesis. Estas se representan gráficamente en la figura 6.10B y en la tabla 6.3 se tabulan y comparan con las frecuencias esperadas de la distribución anterior.

Supongamos que nos hemos decidido por un error de tipo I $\alpha \approx 0{,}01$ ($\approx$ significa "aproximadamente igual a") como se muestra en la figura 6.10A. A este nivel de significación aceptaríamos la H_0 para todas las muestras de 17 que contengan 13 o menos animales de un sexo. Aproximadamente el 99 % de todas las muestras caerán en esta categoría. Pero ¿qué ocurre si H_0 no es cierta y H_1 sí lo es? Claramente, de la población representada por la hipótesis H_1 también podríamos obtener resultados en que un sexo estuviese

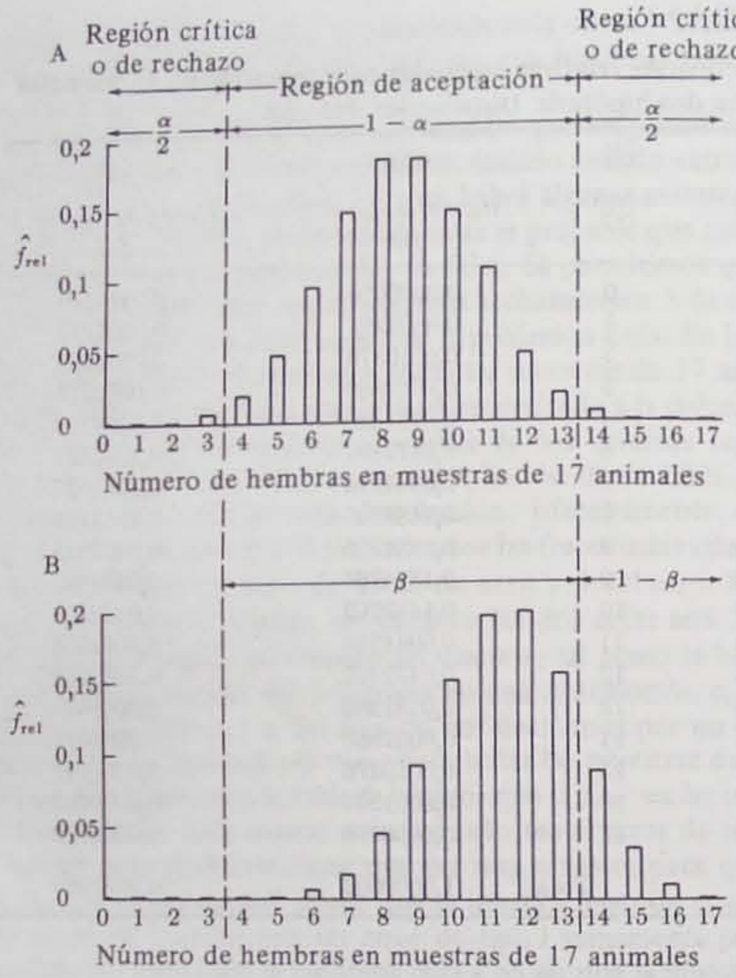


Fig. 6.10. Distribuciones esperadas de resultados cuando se extraen muestras de dos poblaciones hipotéticas. A. H_0: $p♀ = q♂ = \frac{1}{2}$. B. H_1: $p♀ = 2q♂ = \frac{2}{3}$. Las líneas discontinuas separan las regiones críticas de la región de aceptación en la distribución de la figura A. El error del tipo I α es aproximadamente igual a 0,01.

representado 13 veces o menos en muestras de 17. Tenemos que calcular qué proporción de la curva que representa la hipótesis H_1 coincidirá en parte con la región de aceptación de la distribución que representa la hipótesis H_0. En este caso hallamos que 0,8695 de la distribución que representa H_1 se superpone con la región de aceptación de H_0 (véase figura 6.10B). Así, si realmente H_1 es cierta (y correspondientemente H_0 falsa), aceptaríamos erróneamente la hipótesis nula 86,95 % de las veces. Este porcentaje corresponde a la proporción de muestras de H_1 que está entre los límites de las regiones de aceptación de H_0. Esta proporción se llama β, el error de tipo II expresado como proporción. En este ejemplo β es bastante grande. Naturalmente, una muestra de 17 animales es poco satisfactoria para discriminar entre las dos hipótesis. Aunque el 99 % de las muestras según H_0

caerían en la región de aceptación, el 87 % lo haría según H_1. Una sola muestra que caiga en la región de aceptación no nos permitiría llegar a una decisión entre las hipótesis con un alto grado de fiabilidad. Si la muestra tuviese 14 hembras o más, concluiríamos que H_1 era correcta. Si tuviese 3 hembras o menos, podríamos concluir que ni H_0 ni H_1 eran ciertas. Al aproximarse H_1 a H_0 (como en H_1: $p♀ = 0{,}55$, por ejemplo), las dos distribuciones se superpondrían cada vez más y la magnitud de β aumentaría, haciendo menos posible aún la discriminación entre las hipótesis. Por el contrario, si H_1 representase $p♀ = 0{,}9$, las distribuciones estarían mucho más separadas y se reduciría el error de tipo II. Luego claramente la magnitud de β depende, entre otras cosas, de los parámetros de la hipótesis alternativa H_1 y no puede especificarse si no se conocen estos últimos.

Cuando se establece la hipótesis alternativa como en el ejemplo anterior (H_1: $p♀ = 2q♂$), la magnitud del error de tipo I, α, que estamos dispuestos a tolerar, determinará la magnitud del error de tipo II, β. Cuanto más pequeña sea la región crítica α en la distribución según H_0, mayor será la región de aceptación $1 - \alpha$ en esta distribución. Sin embargo, cuanto mayor sea $1 - \alpha$, mayor será su solapamiento con la distribución que representa H_1, y por tanto mayor será β. Convéncete de esto en la figura 6.10. Moviendo hacia fuera las líneas de trazos reducimos las regiones críticas que representan el error de tipo I, α, en el diagrama A. Pero al hacer esto, una parte mayor de la distribución de H_1 en el diagrama B quedará dentro de la región de aceptación de la hipótesis nula. Así, al reducir α estamos aumentando β y en cierto sentido malogrando nuestras propias intenciones. En la mayor parte de las aplicaciones, los científicos desearían mantener pequeños los dos errores, puesto que no desean rechazar una hipótesis nula cuando es cierta ni aceptarla cuando otra hipótesis es correcta. Más adelante veremos qué medidas pueden tomarse para reducir β al mismo tiempo que α se mantiene constante a un nivel preestablecido.

Aunque los niveles de significación α pueden variarse a voluntad, los investigadores se encuentran frecuentemente limitados porque para muchas pruebas no se han tabulado las probabilidades acumulativas de las distribuciones apropiadas, y por tanto deben valerse de los niveles de probabilidad publicados. Estos son ordinariamente 0,05, 0,01, y 0,001, aunque a veces se encuentran otros diferentes. Cuando se ha rechazado una hipótesis nula a un nivel indicado de α, decimos que la muestra es *significativamente diferente* de la población paramétrica o hipotética con probabilidad $P \leqslant \alpha$. Generalmente, valores de α superiores a 0,05 no se consideran *estadísticamente significativos.* Un nivel de significación del 5 % ($P = 0{,}05$) corresponde a un error de tipo I en 20 pruebas, un nivel del 1 % ($P = 0{,}01$) a un error en 100 pruebas. Los niveles de significación menores que el 1 % ($P \leqslant 0{,}01$) casi siempre se juzgan significativos; los situados entre el 5 % y el 1 % pueden considerarse significativos al arbitrio del investigador. Puesto que significación estadística tiene un sentido técnico particular (H_0 rechazada a $P \leqslant \alpha$), utilizaremos el adjetivo significativo solamente en este sentido; su utilización en artículos y comunicaciones científicas debería impedirse a no ser que esté claramente implícito este significado técnico. Para fines descriptivos generales, sinónimos tales como importante, intencionado, marcado, notable y otros pueden servir para subrayar diferencias y efectos.

Ahora se hace un breve comentario de la hipótesis nula representada por distribuciones de probabilidad asimétricas. Supongamos que nuestra hipótesis nula en el caso de proporción de sexos hubiera sido H_0: $p♀ = \frac{2}{3}$, como se ha discutido anteriormente. En la figura

6.10B se presenta la distribución de muestras de 17 crías de esta población. Es claramente asimétrica y por esta razón las regiones críticas tienen que definirse independientemente. Para una determinada prueba de dos colas podemos, bien sea duplicar la probabilidad P de una desviación en la dirección del extremo más próximo y comparar $2P$ con α, el nivel de significación convencional, o bien comparar P con $\alpha/2$, la mitad del nivel de significación convencional. En este último caso el máximo valor de P que se considera convencionalmente significativo es 0,025.

Revisaremos lo que hemos aprendido por medio de un segundo ejemplo, incluyendo esta vez una distribución de frecuencias continua, las longitudes del ala de moscas domésticas distribuidas normalmente, de media paramétrica $\mu = 45{,}5$ y varianza $\sigma^2 = 15{,}21$. Las medias basadas en muestras de 5 ítems extraídas de éstas, también estarán normalmente distribuidas como se ha demostrado en la tabla 6.1 y en la figura 6.1. Vamos a suponer que alguien se presenta con una sola muestra de 5 longitudes del ala de moscas domésticas y se desea probar si podrían pertenecer a la población indicada. La hipótesis nula será $H_0{:}\mu = 45{,}5$ ó $H_0{:}\mu = \mu_0$, donde μ es la verdadera media de la población de la que se ha muestreado y μ_0 representa la media paramétrica hipotética de 45,5. De momento supondremos que no tenemos evidencia de que la varianza de nuestra muestra sea muy superior ó inferior a la varianza paramétrica de las longitudes del ala de moscas domésticas. Si fuese así, no sería lógico suponer que nuestra muestra procede de la población indicada. Hay una prueba crítica de hipótesis sobre la varianza de muestreo de la que nos ocuparemos más adelante. En la figura 6.11, la curva del centro representa la distribución esperada de medias de muestras de 5 longitudes del ala de moscas domésticas de la población indicada. A lo largo de la abscisa se delimitan las regiones de aceptación y crítica para un error de tipo I, $\alpha = 0{,}05$. Los límites de las regiones críticas se calculan como sigue (recuérdese que $t_{[\infty]}$ es equivalente a la distribución normal):

$$L_1 = \mu_0 - t_{0{,}05[\infty]}\sigma_{\bar{Y}} = 45{,}5 - (1{,}96)(1{,}744) = 42{,}08$$

y

$$L_2 = \mu_0 + t_{0{,}05[\infty]}\sigma_{\bar{Y}} = 45{,}5 + (1{,}96)(1{,}744) = 48{,}92$$

Así, para medias menores que 42,08 o mayores que 48,92 consideraríamos improbable que se hubiesen muestreado de esta población. Por lo tanto, para estas medias rechazaríamos la hipótesis nula. La prueba que presentamos es de dos colas porque no tenemos nociones a priori sobre las posibles alternativas a nuestra hipótesis nula. Si pudiésemos suponer que la verdadera media de la población de la cual se ha tomado la muestra solamente puede ser igual o mayor que 45,5, la prueba sería de una cola.

Ahora vamos a examinar diversas hipótesis alternativas. Una hipótesis alternativa pudiera ser que la verdadera media de la población de la que proviene nuestra muestra sea 54,0, pero la varianza sea la misma que antes. Podemos expresar esta hipótesis como $H_1{:}\mu = 54{,}0$ ó $H_1{:}\mu = \mu_1$, donde μ_1 representa la media paramétrica alternativa 54,0. A partir de la tabla de áreas de la curva normal y nuestro conocimiento de la varianza de las medias, podemos calcular la proporción de la distribución denotada por H_1 que coincidiría en parte con la región de aceptación denotada por H_0. Encontramos que 54,0 está a 5,08 unidades de medida de 48,92, el límite superior de la región de aceptación de H_0. Esto

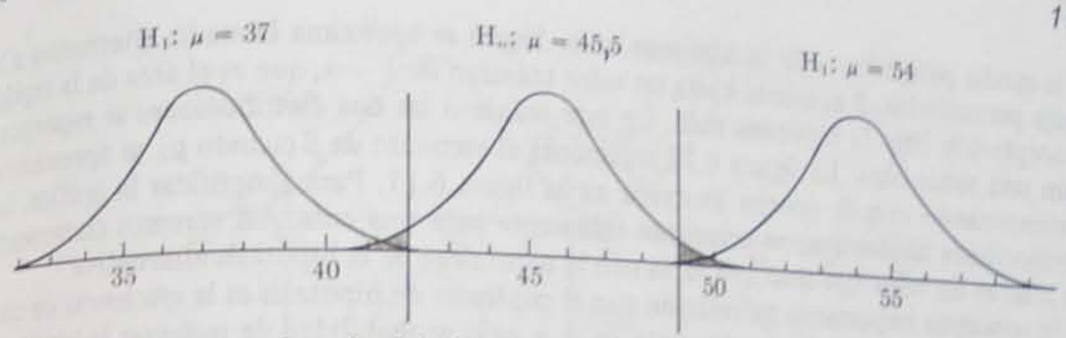


Fig. 6.11. Distribución esperada de medias de muestras de 5 longitudes del ala de moscas domésticas de poblaciones normales indicadas por μ, como se muestra en las curvas anteriores, y $\sigma_{\bar{Y}} = 1{,}744$. La curva del centro representa la hipótesis nula, $H_0{:}\mu = 45{,}5$, las curvas de los lados representan hipótesis alternativas, $\mu = 37$ ó $\mu = 54$. Las líneas verticales delimitan regiones críticas del 5 % para la hipótesis nula (2 ½ % en cada cola, sombreadas).

corresponde a $5{,}08/1{,}744 = 2{,}91\sigma_{\bar{Y}}$ unidades. En la tabla de áreas de la curva normal (tabla II) encontramos que 0,0018 del área estará más allá de $2{,}91\sigma$ en una cola de la curva. Así, según esta hipótesis alternativa, 0,0018 de la distribución de H_1 coincidirá en parte con la región de aceptación de H_0. Este es β, el error de tipo II según esta hipótesis alternativa. Realmente esto no es del todo correcto. Ya que la cola izquierda de la distribución H_1 continúa hasta el infinito negativo, se sale de la región de aceptación y pasa a la región crítica de la parte izquierda de H_0. No obstante, esto representa solamente una cantidad infinitesimal del área de H_1 (el límite crítico inferior de H_0, 42,08, dista $6{,}83\sigma_{\bar{Y}}$ unidades de $\mu_1 = 54$) y puede ignorarse.

Nuestra hipótesis alternativa H_1 indicaba que μ_1 es 8,5 unidades mayor que μ_0. Sin embargo, como se ha dicho anteriormente, puede que no tengamos fundamento a priori para creer que la verdadera media de nuestra muestra sea mayor o menor que μ. Por esta razón podemos suponer simplemente que se aparta 8,5 unidades de medida de 45,5. En este caso debemos calcular de igual manera β para la hipótesis alternativa de que $\mu_1 = \mu_0 - 8{,}5$. Así, las hipótesis alternativas se convierten en $H_1{:}\mu = 54{,}0$ ó $37{,}0$, ó $H_1{:}\mu = \mu_1$, donde μ_1 representa 54 ó 37, las medias paramétricas alternativas. Como las distribuciones son simétricas, β es el mismo para las dos hipótesis alternativas. Por lo tanto, el error de tipo II para la hipótesis H_1 es 0,0018, independientemente de cual de las dos hipótesis alternativas sea correcta. Si H_1 es realmente cierta, 18 de cada 10 000 muestras llevarían a una aceptación incorrecta de H_0, una proporción de error muy baja. En la figura 6.11 se presentan estas relaciones.

Nos podemos preguntar, justamente, qué razón tenemos para creer que el valor paramétrico alternativo de la media es 8,5 unidades de medida a cada lado de $\mu_0 = 45{,}5$. Sería realmente extraordinario si tuviésemos cualquier justificación para esta opinión. De hecho, la verdadera media lo mismo puede estar a 7,5 que a 6 que a cualquier número de unidades a cada lado de μ_0. Si trazamos curvas para $H_1{:}\mu = \mu_0 \pm 7{,}5$, encontramos que β ha aumentado considerablemente estando ahora más próximas las curvas para H_0 y H_1. Así, la magnitud de β dependerá de lo distante que esté la media paramétrica alternativa

de la media paramétrica de la hipótesis nula. Según se aproxima la media alternativa a la media paramétrica, β aumenta hasta un valor máximo de $1 - \alpha$, que es el área de la región de aceptación bajo la hipótesis nula. En este máximo las dos distribuciones se superpondrían una sobre otra. La figura 6.12 representa el aumento de β cuando μ_1 se aproxima a μ, comenzando con la prueba ilustrada en la figura 6.11. Para simplificar la gráfica, las distribuciones alternativas se presentan solamente para una cola. Así veremos claramente que β no es un valor fijo sino que varía con la naturaleza de la hipótesis alternativa.

Un concepto importante en relación con el contraste de hipótesis es la *eficiencia* de una prueba. Esta es $1 - \beta$, el complemento de β, y es la probabilidad de rechazar la hipótesis nula cuando en realidad es falsa y la hipótesis alternativa es correcta. Evidentemente, para cualquier prueba dada nos gustaría que $1 - \beta$ fuese lo más grande posible y β lo más pequeño posible. Como generalmente no podemos especificar una determinada hipótesis alternativa, tenemos que describir β ó $1 - \beta$ para una continuidad de valores alternativos. Cuando $1 - \beta$ se representa gráficamente de este modo, el resultado se denomina *curva de eficiencia* para la prueba que se considera. La figura 6.13 muestra la curva de eficiencia para el ejemplo ya discutido de la longitud del ala de moscas domésticas. Esta puede compararse con la figura 6.12 de la cual se deriva directamente. La figura 6.12 pone de relieve el error de tipo II, β, y la figura 6.13 representa gráficamente el complemento de este valor, $1 - \beta$. Observamos que la eficiencia de la prueba disminuye claramente al aproximarse la hipótesis alternativa a la hipótesis nula. El sentido común confirma estas conclusiones: podemos tomar decisiones claras y firmes de si nuestra muestra procede de una población de media 45,5 ó 60,0. La eficiencia es esencialmente 1. Pero si la hipótesis alternativa es que $\mu_1 = 45{,}6$, que difiere solamente en 0,1 del valor supuesto según la hipótesis nula, será difícil decidir cuál de estas hipótesis es correcta y la eficiencia será muy baja.

Para mejorar la eficiencia de una determinada prueba (o disminución de β) mientras se mantiene constante α para una hipótesis nula establecida, debemos incrementar el tamaño de muestra. Si en lugar de muestrear 5 longitudes del ala hubiésemos muestreado 35, la distribución de medias sería mucho más estrecha. Así las regiones críticas para idéntico error de tipo I comenzarían ahora en 44,21 y 46,79. Aunque las regiones de aceptación y de rechazo han continuado proporcionalmente iguales, la región de aceptación se ha reducido mucho más en valor absoluto. Previamente no podíamos, con confianza, rechazar la hipótesis nula para una media de muestreo de 48.0. Ahora, cuando se basa en 35 individuos, una media tan alejada como 48,0 solamente aparecería 15 veces entre 100 000 y por lo tanto, se rechazaría la hipótesis. ¿Qué ha ocurrido al error tipo II? Puesto que las curvas de distribución no son tan abiertas como antes, hay menos solapamiento entre ellas; si la hipótesis alternativa H_1: $\mu = 54{,}0$ ó $37{,}0$ es correcta, la probabilidad de que la hipótesis nula pudiera ser aceptada por error (error de tipo II) es infinitesimalmente pequeña. Si permitimos que μ_1 se aproxime a μ_0, aumentará β, naturalmente, pero siempre será menor que el correspondiente valor para un tamaño de muestra de $n = 5$. Esta comparación se presenta en la figura 6.13 donde la eficiencia para la prueba con $n = 35$ es mucho mayor que para $n = 5$. Si aumentásemos nuestro tamaño de muestra hasta 100 ó 1 000, la eficiencia se incrementaría todavía más. Así llegamos a una conclusión importante: si una determinada prueba no es suficientemente sensible podemos incrementar su sensibilidad (eficiencia) aumentando el tamaño de muestra.

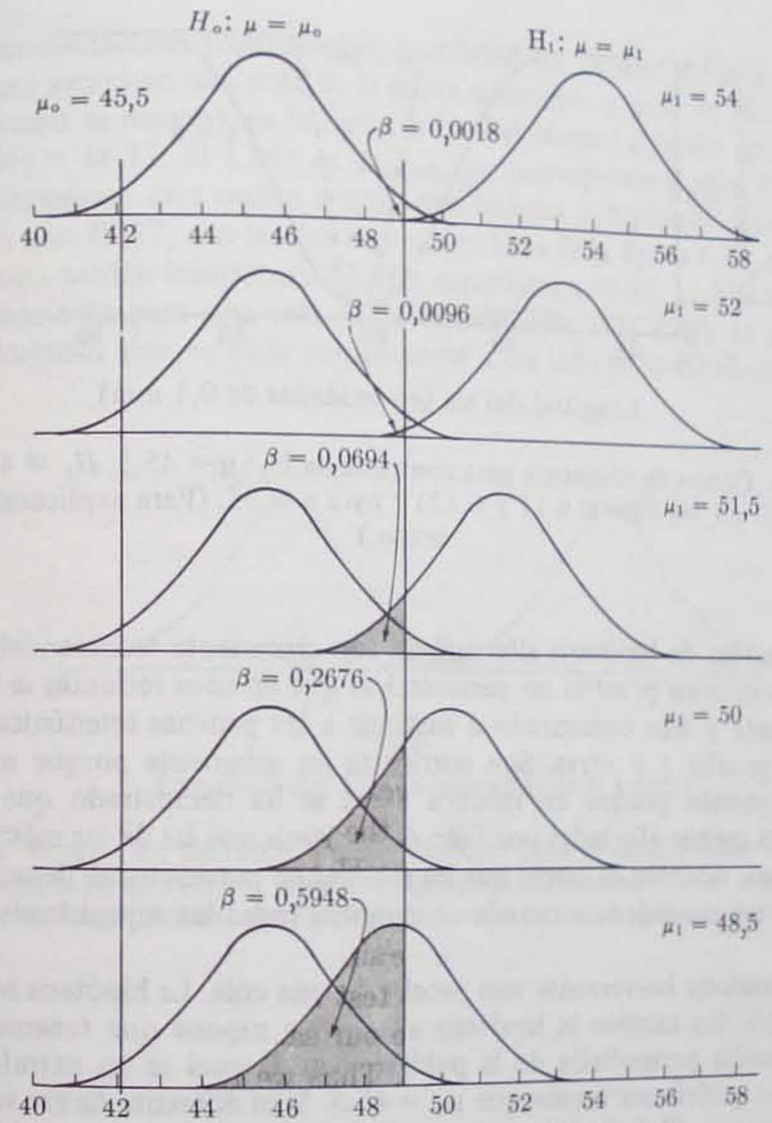


Fig. 6.12. Diagrama para ilustrar incrementos en el error de tipo II, β, según se aproxima la hipótesis alternativa H_1 a la hipótesis nula H_0, es decir μ_1 se aproxima a μ. El sombreado representa β. Las líneas verticales señalan los límites de las regiones críticas del 5 % (2 ½ % en cada cola) para la hipótesis nula. Para simplificar la gráfica las distribuciones alternativas se representan solamente para una cola. Datos idénticos a los de la figura 6.11.

Hay otra forma más de aumentar la eficiencia de una prueba. Si no podemos aumentar el tamaño de muestra, la eficiencia puede elevarse cambiando la naturaleza de la prueba. Las diferentes técnicas estadísticas que contrastan aproximadamente la misma hipótesis pueden diferir sustancialmente tanto en la magnitud real como en las pendientes de sus curvas de eficiencia. Las pruebas que mantienen niveles de eficiencia superiores sobre

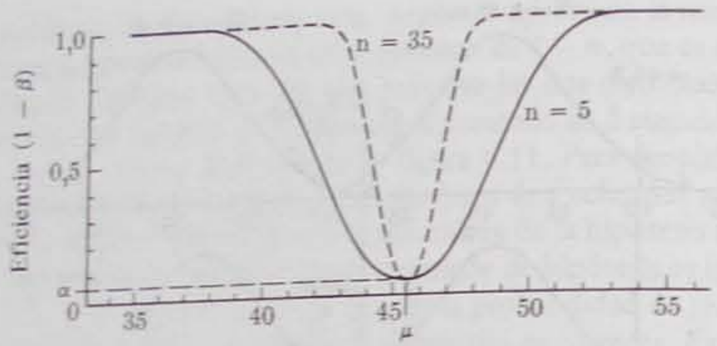


Fig. 6.13. Curvas de eficiencia para contraste de $H_0: \mu = 45{,}5$, $H_1 \neq 45{,}5$ para $n = 5$ (como en las figuras 6.11 y 6.12) y para $n = 35$. (Para explicación véase el texto.)

rangos considerables de hipótesis alternativas, son claramente recomendables. Ya hemos mencionado las diversas pruebas no paramétricas que en años recientes se han generalizado crecientemente y han comenzado a sustituir a las pruebas estadísticas tradicionales tales como la prueba t y otras. Son corrientes no solamente porque son sencillas de realizar, sino también porque en muchos casos se ha demostrado que sus curvas de eficiencia se ven menos afectadas por fallo de hipótesis que las de los métodos paramétricos. Sin embargo, también es cierto que las pruebas no paramétricas tienen una eficiencia total inferior a las paramétricas cuando se cumplen todas las suposiciones de las pruebas paramétricas.

Vamos a considerar brevemente una prueba de una cola. La hipótesis nula es $H_0: \mu_0 = 45{,}5$ como antes. En cambio la hipótesis alternativa supone que tenemos motivo para creer que la media paramétrica de la población de la cual se ha extraído la muestra, posiblemente no podría ser menor que $\mu_0 = 45{,}5$. Si es diferente de ese valor, solamente podría ser superior a 45,5. Pudiéramos tener dos fundamentos para esta hipótesis. Primero, pudiéramos tener alguna causa biológica para tal opinión. Puede que nuestras moscas fuesen una población diminuta y cualquier otra población de la cual podría haber procedido nuestra muestra debe ser más grande. Una segunda razón pudiera ser que estuviésemos interesados solamente en una desviación de diferencia. Por ejemplo, podemos probar el efecto de una sustancia química en el alimento larval, destinada a aumentar la magnitud de la muestra de moscas. Por lo tanto, esperaríamos que $\mu_1 \geqslant \mu_0$, y no estamos interesados en contrastar para ninguna μ_1 que sea menor que μ_0 porque este efecto es exactamente lo contrario de lo que esperamos. Del mismo modo, si estamos investigando el efecto de cierta droga como remedio para el cáncer, pudiéramos querer comparar la población no tratada que tiene una tasa media de mortalidad θ (de cáncer) con la población tratada cuya tasa es θ_1. Nuestra hipótesis alternativa será $H_1: \theta_1 < \theta$. Es decir, no nos interesa ningún θ_1 que sea mayor que θ porque si nuestra droga incrementa la mortalidad de cáncer, es una perspectiva de tratamiento de poco valor.

Cuando se realiza esta prueba de una cola, la región de rechazo sobre la abscisa está solamente bajo una cola de la curva que representa la hipótesis nula. Así, para nuestros

datos de moscas domésticas (distribución de medias de tamaño de muestra $n = 5$), la región de rechazo estará en una cola de la curva solamente y para un error de tipo I del 5 % aparecerá como se muestra en la figura 6.14. Calculamos el límite crítico como $45{,}5 + (1{,}645)(1{,}744) = 48{,}37$. El 1,645 es $t_{0,10[\infty]}$, que corresponde al valor 5 % para un test de una cola. Compárese esta región crítica, que rechaza la hipótesis nula para todas las medias mayores que 48,37, con las dos regiones críticas de la figura 6.12, que rechazan la hipótesis nula para medias inferiores a 42,08 y superiores a 48,92. La hipótesis alternativa solamente se considera para una cola de la distribución, y la curva de eficiencia de la prueba no es simétrica sino estirada con respecto a un lado de la distribución solamente.

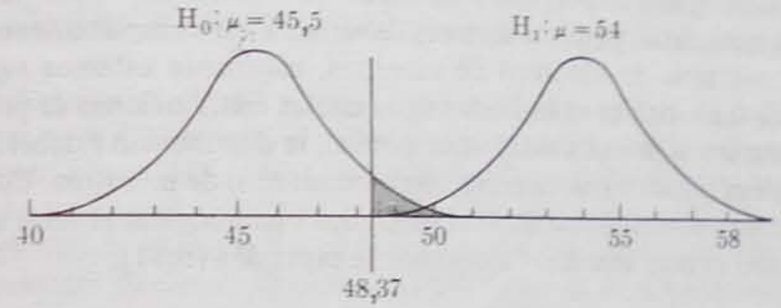


Fig. 6.14. Prueba de significación de una cola para la distribución de la figura 6.11. Ahora la línea vertical separa la región de rechazo al 5 % en una cola de la distribución (el área correspondiente de la curva ha sido sombreada).

6.9 Pruebas de hipótesis simples que utilizan la distribución t

Pasaremos a aplicar nuestro conocimiento recientemente adquirido del contraste de hipótesis a un sencillo ejemplo que incluye a la distribución t.

Los reglamentos del Estado ordenan que la dosis patrón de un cierto preparado biológico debería ser de 600 unidades de actividad por centímetro cúbico. Preparamos 10 muestras de este preparado y probamos la potencia de cada una. Encontramos que el número medio de unidades de actividad por muestra es 592,5 unidades por cm³ y la desviación típica de las muestras es 11,2. ¿Cumple nuestra muestra la norma estatal? Establecida con más precisión, nuestra hipótesis nula es $H_0: \mu = \mu_0$. La hipótesis alternativa es que la dosis no es igual a 600, ó $H_1: \mu \neq \mu_0$. Pasamos a calcular el significado de la desviación $\bar{Y} - \mu_0$ expresada en unidades de desviación típica. La desviación típica pertinente es la de medias (el error típico de la media), *no* la desviación típica de los ítems porque se trata de la desviación de una media de muestreo en torno a una media paramétrica. Por lo tanto calculamos $s_{\bar{Y}} = s/\sqrt{n} = 11{,}2/\sqrt{10} = 3{,}542$. A continuación examinamos la desviación $(\bar{Y} - \mu_0)/s_{\bar{Y}}$. Hemos visto anteriormente en la sección 6.4 que

una desviación dividida por una desviación típica estimada estará distribuida según la distribución t con $n-1$ grados de libertad. Por lo tanto escribimos

$$t_s = \frac{\overline{Y} - \mu_0}{s_{\overline{Y}}} \tag{6.11}$$

Esto indica que esperaríamos que esta desviación se distribuyese como una variante t. Obsérvese que en la expresión (6.11) hemos escrito t_s. En la mayor parte de los libros de texto se encontrará esta razón identificada simplemente como t, pero en realidad la distribución t es una distribución paramétrica y teórica que generalmente sólo es aproximada pero nunca igualada por datos de muestreo observados. Esto puede parecer una característica secundaria, pero los lectores deberían estar completamente seguros de que en cualquier contraste de hipótesis de muestras, solamente estamos *suponiendo* que la distribución de las variables examinadas sigue ciertas distribuciones de probabilidad teóricas. Para ajustarse a la práctica estadística general, la distribución t debería tener realmente una letra griega (como τ), sirviendo t como estadístico de muestreo. Como esto violaría la práctica antigua, preferimos utilizar el subíndice s para indicar el valor de muestreo.

La prueba real es muy sencilla. Calculamos la expresión (6.11),

$$t_s = \frac{592{,}5 - 600}{3{,}542} = \frac{-7{,}5}{3{,}542} = -2{,}12, \qquad gl = n - 1 = 9,$$

y la comparamos con los valores esperados de t para 9 grados de libertad. Como la distribución t es simétrica, ignoraremos el signo de t_s y siempre lo introduciremos en la tabla III bajo su valor positivo. Los dos valores a cada lado de t_s son $t_{0{,}05[9]} = 2{,}26$ y $t_{0{,}10[9]} = 1{,}83$. Estos son valores de t para pruebas de dos colas, apropiados en este ejemplo porque la hipótesis alternativa es que $\mu \neq 600$; es decir, puede ser menor o mayor. Parece que el nivel de significación de nuestro valor de t_s está entre el 5 % y 10 %; si la hipótesis nula es realmente cierta, la probabilidad de obtener una desviación tan grande o mayor que 7,5 está entre 0,05 y 0,10. A niveles de significación habituales esto es insuficiente para declarar la media de muestreo significativamente diferente del patrón. Por consiguiente, aceptamos la hipótesis nula. En lenguaje convencional presentaremos los resultados del análisis estadístico como sigue. "La media de muestreo no es significativamente diferente del patrón aceptado." En una comunicación científica esta afirmación debería respaldarse siempre por un valor de probabilidad, y la forma apropiada de presentar esto es escribir $0{,}10 > P > 0{,}05$, que significa que la probabilidad de esta desviación está entre 0,05 y 0,10. Otra forma de decir esto es que el valor de t_s *no es significativo* (frecuentemente abreviado como *ns*).

Una costumbre frecuentemente encontrada es el uso de asteriscos después del valor calculado de la prueba de significación, como en $t_s = 2{,}86^{**}$. Generalmente los símbolos representan los siguientes rangos de probabilidad:

$$^{*} = 0{,}05 > P > 0{,}01, \qquad ^{**} = 0{,}01 > P > 0{,}001, \qquad ^{***} = P < 0{,}001$$

Sin embargo, como algunos autores a veces denotan otros rangos por estos asteriscos, el significado de los símbolos debe indicarse en cada comunicación científica.

Pudiera argumentarse que en un preparado biológico el interés del investigador no debería residir en si la muestra difiere significativamente de un patrón, sino en si está significativamente por debajo del patrón. Este puede ser uno de esos preparados biológicos en los cuales un exceso del componente activo no es perjudicial, pero un déficit haría la preparación inefectiva a la dosis convencional. En este caso la prueba es de una cola y se realiza exactamente de la misma manera salvo que los valores críticos de t para una prueba de una cola están en la mitad de las probabilidades de las pruebas de dos colas. Así $t_{0{,}025[9]} = 2{,}26$, el valor 0,05 anterior, y $t_{0{,}05[9]} = 1{,}83$, el valor 0,10 anterior, haciendo nuestro valor de t_s observado, 2,12, "significativo al nivel 5 %" ó, expresado con más precisión, significativo a $0{,}05 > P > 0{,}025$. Si estuviésemos dispuestos a aceptar un nivel de significación del 5 %, consideraríamos el preparado significativamente por debajo del patrón.

Puede sorprender que el mismo ejemplo, utilizando los mismos datos y pruebas de significación, lleve a dos conclusiones diferentes, y puede comenzarse a preguntar si algunas de las cosas que se han oído sobre estadística y estadísticos no son, después de todo, correctas. La explicación reside en el hecho de que los dos resultados son respuestas a diferentes preguntas. Si examinamos si nuestra muestra es significativamente diferente del patrón en cualquier dirección, debemos concluir que no es suficientemente diferente para que rechacemos la hipótesis nula. Si, por otra parte, no consideramos el hecho de que la verdadera media de muestreo μ podría ser mayor que la standard establecida μ_0, la diferencia encontrada por nosotros es claramente significativa. En este ejemplo queda claro que en cualquier prueba estadística debe expresarse claramente si se ha realizado una prueba de una cola o de dos colas, en el caso de que la naturaleza del ejemplo fuese tal que hubiese cualquier duda sobre la cuestión. Deberíamos señalar también que esta diferencia en los resultados no es necesariamente típica. Se debe solamente a que el resultado en este caso está en un área limítrofe entre clara significación y no significación. Si la diferencia entre muestra y patrón hubiera sido de 10,5 unidades de actividad, la muestra habría sido sin duda significativamente diferente del patrón para la prueba de una cola o de dos colas.

La promulgación de una media patrón es generalmente insuficiente para el establecimiento de un patrón riguroso para un producto. Si la varianza entre las muestras es suficientemente grande, nunca será posible establecer una diferencia significativa entre la media de muestreo y la patrón. Este es un punto importante que debería quedar completamente claro. Recuérdese que el error típico puede aumentar de dos maneras, disminuyendo el tamaño de muestra o aumentando la desviación típica de los elementos repetitivos. Ambos son aspectos indeseables de cualquier procedimiento experimental.

La prueba descrita más arriba para el preparado biológico nos lleva a una prueba general para la significación de un estadístico, es decir, para la significación de una desviación de cualquier estadístico de un parámetro, el cual se expone a grandes rasgos en el cuadro 6.4. Esta prueba se aplica siempre que se espere que los estadísticos sigan la ley de distribución normal. Cuando el error típico se estima a partir de la muestra, se utiliza la distribución t. Sin embargo, cuando la distribución normal es sólo un caso especial $t_{[\infty]}$ de la distribución t, la mayoría de los estadísticos aplican la distribución t con los grados

CUADRO 6.4

Contraste de la significación de un estadístico, es decir, la significación de una desviación respecto de un parámetro. Para estadísticos normalmente distribuidos.

Etapas del cálculo

1. Calcular t_s como la siguiente razón,

$$t_s = \frac{St - St_p}{s_{St}}$$

donde St es un estadístico de muestreo, St_p es el valor paramétrico frente al cual se va a contrastar el estadístico de muestreo, y s_{St} es su error típico estimado obtenido en el cuadro 6.1, o en otra parte de este libro.

2. Las hipótesis pertinentes son

$H_0: St = St_p$ $\quad$ $H_1: St \neq St_p$

para una prueba de dos colas.

$H_0: St = St_p$ $\quad$ $H_1: St > St_p$

o

$H_0: St = St_p$ $\quad$ $H_1: St < St_p$

para una prueba de una cola.

3. En la prueba de dos colas, buscar el valor crítico de $t_{\alpha[\nu]}$, donde α es el error de tipo I pertinente y ν los grados de libertad correspondiente al error típico utilizado (véase Cuadro 6.1). En la prueba de una cola buscar el valor crítico de $t_{2\alpha[\nu]}$ para un nivel de significación de α.

4. Aceptar o rechazar la hipótesis apropiada en 2 basándose en el valor de t_s en 1 comparado con los valores críticos de t en 3.

de libertad pertinentes desde 1 hasta infinito. Un ejemplo de ello es la prueba t para la significación de un coeficiente de regresión que se presenta en la etapa 2 del cuadro 11.4.

6.10 Contraste de la hipótesis $H_0: \sigma^2 = \sigma_0^2$

El método del cuadro 6.4 solamente puede utilizarse si el estadístico está normalmente distribuido. En la varianza esto no es así. Como hemos visto en la sección 6.6, las sumas de cuadrados divididas por σ^2 siguen la distribución χ^2. Por lo tanto, para contrastar la hipótesis de que una varianza de muestreo es diferente de una varianza paramétrica debemos emplear la distribución χ^2.

Vamos a utilizar como ejemplo el preparado biológico de la sección anterior. Se nos ha dicho que la desviación típica era 11,2 basada en 10 muestras. Por lo tanto, la varianza debe haber sido 125,44. Supongamos que el gobierno postula que la varianza de las muestras del preparado no debería ser superior a 100,0. ¿Es nuestra varianza de muestreo

significativamente superior a 100? Recordando de la expresión (6.8) que $(n-1)s^2/\sigma^2$ se distribuye como $\chi^2_{[n-1]}$, procederemos como sigue. En primer lugar calculamos

$$\begin{aligned} X^2 &= (n-1)s^2/\sigma^2 \\ &= (9)125{,}44/100 \\ &= 11{,}290 \end{aligned}$$

Nótese que calculamos la cantidad X^2 en lugar de χ^2 para hacer hincapié de nuevo en que estamos obteniendo un estadístico de muestreo que compararemos con la distribución paramétrica. El uso de X^2 para designar el estadístico de muestreo que se ajusta a una distribución χ^2 está ampliamente establecido. Siguiendo el esquema general del cuadro 6.4, a continuación establecemos nuestras hipótesis nula y alternativa que son $H_0: \sigma^2 \leqslant \sigma_0^2$ y $H_1: \sigma^2 > \sigma_0^2$; es decir, vamos a realizar una prueba de una cola. Después se halla el valor crítico de χ^2 como $\chi^2_{\alpha[\nu]}$, donde α indica el error de tipo I y ν los grados de libertad pertinentes. La cantidad α representa la proporción de la distribución χ^2 a la derecha del valor dado, como se ha descrito en la sección 6.6, y ahora se ve por qué hemos utilizado el símbolo α para esa porción de la curva; corresponde al error de tipo I. Para 9 grados de libertad encontramos en la tabla IV que

$$\chi^2_{0,05[9]} = 16{,}919, \qquad \chi^2_{0,10[9]} = 14{,}684, \qquad \chi^2_{0,50[9]} = 8{,}343$$

Observamos que la probabilidad de obtener un χ^2 tan grande como 11,290 es pues superior que 0,10 pero inferior que 0,50, suponiendo que la hipótesis nula sea cierta. Así X^2 no es significativa al nivel 5 %; no tenemos fundamento para rechazar la hipótesis nula, y debemos concluir que la varianza de las 10 muestras del preparado biológico no puede ser superior a la standard permitido por el gobierno. Si hubiésemos decidido probar si la varianza es diferente de la standard, permitiéndole desviarse en cualquier dirección, la hipótesis para esta prueba de dos colas habría sido $H_0: \sigma^2 = \sigma_0^2$ y $H_1: \sigma^2 \neq \sigma_0^2$, y un error de tipo I del 5 % habría dado para la prueba de dos colas los valores críticos siguientes:

$$\chi^2_{0,975[9]} = 2{,}700, \qquad \chi^2_{0,025[9]} = 19{,}023$$

Estos valores representan ji-cuadrados en los puntos que delimitan regiones críticas del 2 ½ % en cada cola de la distribución χ^2. Un valor de X^2 menor que 2,700 o mayor que 19,023 habría sido prueba de que la varianza de muestreo no pertenecía a esta población. Nuestro valor de $X^2 = 11{,}290$ habría llevado nuevamente a una aceptación de la hipótesis nula.

En el capítulo próximo veremos que hay otra prueba de significación válida para contrastar la hipótesis sobre varianzas de la presente sección. Esta es la prueba F, matemáticamente equivalente, pero que es sin embargo, una prueba más general, permitiéndonos probar la hipótesis de que dos varianzas de muestreo proceden de poblaciones con varianzas iguales.

Ejercicios 6

6.1 Diferenciar entre errores de tipo I y de tipo II. ¿Qué significa la potencia de una prueba estadística?

6.2 Puesto que es posible contrastar una hipótesis estadística con cualquier tamaño de muestra, ¿por qué se prefieren tamaños mayores?

6.3 Los límites de confianza al 95 % para μ obtenidos en una determinada muestra han sido 4,91 y 5,67 g. ¿Es correcto decir que 95 veces de cada 100 la media de la población, μ, se halla dentro del intervalo desde 4,91 hasta 5,67 g? Si no, ¿cuál sería la afirmación correcta?

6.4 Fijar límites de confianza del 99 % para la media, mediana, coeficiente de variación, y varianza de los datos de pesos de nacimiento dados en el cuadro 3.2. SOLUCION. Los límites inferiores son 109,540, 109,060, 12,136 y 178,698, respectivamente.

6.5 Fijar límites de confianza del 95 % a las medias de la tabla 6.2. ¿Son todos estos límites correctos? (Es decir, ¿contienen a μ?)

6.6 En un estudio de llamadas de apareamiento en el sapo arbóreo *Hyla ewingi*, Littlejohn (1965) encontró que la duración de la señal de llamada en una muestra de 39 observaciones de Tasmania tenía una media de 189 milisegundos y una desviación típica de 32 milisegundos. Establecer intervalos de confianza del 95 % para la media y la varianza.

6.7 En la sección 4.3 se dio el coeficiente de dispersión como un índice de si los datos se ajustan o no a una distribución de Poisson. Como en una verdadera distribución de Poisson, la media, μ, es igual a la varianza paramétrica, σ^2, el coeficiente de dispersión es análogo a la expresión (6.8). Utilizando los datos de ácaros del cuadro 4.5, contrastar la hipótesis de que la verdadera varianza es igual a la media de muestreo, en otras palabras, que hemos muestreado de una distribución de Poisson (en la cual el coeficiente de dispersión sería igual a la unidad). Nótese que en estos ejemplos la tabla ji-cuadrado no es adecuada, de modo que deben calcularse valores críticos aproximados utilizando el método dado con la tabla **IV**. En la sección 7.3 se presentará una prueba de significación alternativa que evita este problema. SOLUCION. $(n-1) \times \text{CD} = 1308{,}30$, $X^2_{0{,}05[588]} \approx 645{,}708$.

6.8 Utilizando el método descrito en el ejercicio 6.7, examinar el ajuste de la distribución observada a una distribución de Poisson, contrastando la hipótesis de que el verdadero coeficiente de dispersión para los datos de la tabla 4.6 es igual a la unidad.

6.9 En el comportamiento clinocinético directo, relativo a la temperatura, los animales tienden más a menudo hacia el extremo caliente del gradiente que hacia el frío; la dirección de la tendencia es al azar. En una simulación de este comportamiento, en un ordenador, se obtuvieron los siguientes resultados. La posición media de un gradiente de temperatura apareció en $-1{,}352$. La desviación típica fue 12,267 y $n = 500$ individuos. El gradiente fue dividido en unidades: el cero en el medio del gradiente, punto de partida de los animales; el signo menos corresponde al extremo frío y el signo más al extremo más caliente. Comprobar la hipótesis de que el comportamiento clinocinético directo de los animales no resulta de una tendencia gregaria hacia el extremo frío o hacia el caliente, es decir comprobar la hipótesis de que μ, la posición media en el gradiente, era cero.

6.10 En un estudio de medidas de los picos de los papamoscas vespertinos, Johnson (1966) encontró que la longitud del pico de los machos tenía una media de 8,14 ± 0,021 y un coeficiente de variación de 4,67 %. Partiendo de estos datos, deducir cuantos individuos se utilizaron. SOLUCION: $n = 328$.

Capítulo 7

Introducción al análisis de la varianza

Ahora pasamos al estudio del análisis de la varianza. Este método, desarrollado por R. A. Fisher, es fundamental para toda la aplicación de la estadística a la biología y especialmente al diseño experimental. Una manera de enfocar el análisis de la varianza es considerarlo una prueba de si dos o más medias de muestreo podrían haberse obtenido de poblaciones con la misma media paramétrica, con respecto a una variable determinada. Alternativamente, podríamos concluir que estas medias difieren entre sí hasta tal punto que debemos suponer que se han muestreado de poblaciones diferentes. En los casos en que solamente se trata de dos muestras, la distribución t ha sido utilizada tradicionalmente para probar diferencias significativas entre medias. Sin embargo, el análisis de la varianza es una prueba más general que permite contrastar dos o muchas muestras, y por esta razón lo presentamos al principio para equipar al lector con el arma más potente para su arsenal estadístico. En la sección 8.4 discutiremos la prueba t para dos muestras como un caso particular.

En la sección 7.1 abordaremos el tema en terreno familiar, el experimento de muestreo de las longitudes del ala de moscas domésticas. De estas muestras obtendremos dos estimadores independientes de la varianza de población. En la sección 7.2 nos desviaremos para presentar otra distribución continua más, la distribución F, necesaria para la prueba de significación en el análisis de la varianza. La sección 7.3 es otra disgresión en la que mostramos cómo puede utilizarse la recién aprendida distribución F para probar si dos muestras tienen la misma varianza. Ahora estamos preparados para la sección 7.4, en la que examinamos los efectos de someter las muestras a diferentes tratamientos. La próxima sección (7.5) describe la descomposición de sumas de cuadrados y de grados de libertad, el *análisis de varianza* propiamente dicho. Las dos últimas secciones (7.6 y 7.7) se ocupan de un modo más ordenado de los dos modelos científicos para los que es conveniente el análisis de la varianza, el llamado modelo de efectos del tratamiento (Modelo I) y el modelo del componente de la varianza (Modelo II).

Excepto la sección 7.3, todo el capítulo es ampliamente teórico. En el capítulo 8 consideraremos los detalles prácticos del cálculo. No obstante, es necesario un conocimiento completo de la materia del capítulo 7 para resolver ejemplos reales de análisis de la varianza en el capítulo 8.

7.1 Las varianzas de muestreo y sus medias

Abordamos el análisis de la varianza por medio del experimento de muestreo familiar de las longitudes del ala de moscas domésticas (experimento 5.1 y tabla 5.1), en el cual combinábamos siete muestras de 5 longitudes del ala para formar muestras de 35. En la tabla 7.1 hemos reproducido una de estas muestras. Las siete muestras de 5, aquí llamadas grupos, se alistan verticalmente en la mitad superior de la tabla. Antes de pasar a explicar más ampliamente la tabla 7.1 debemos familiarizarnos con la terminología y simbolismo añadidos para abordar este tipo de problema. Denominamos a nuestras muestras *grupos*; a veces se denominan *clases* y por otros términos más que veremos después. En cualquier problema de análisis de varianza tendremos dos o más muestras o grupos semejantes y utilizaremos el símbolo a para el número de grupos. Así, en el ejemplo actual $a = 7$. Cada grupo o muestra está basado en n ítems, como antes; en la tabla 7.1, $n = 5$. El número total de ítems de la tabla es a veces n, que en este caso corresponde a 7×5 ó 35.

Las sumas de los ítems de cada grupo se presentan en la fila bajo la línea horizontal que separa. En un análisis de varianza, los signos sumatorios ya no pueden ser tan sencillos como anteriormente. Podemos sumar los ítems de un grupo solamente o los ítems de la tabla entera. Por lo tanto, tenemos que utilizar superíndices con el símbolo sumatorio. De acuerdo con nuestra costumbre de utilizar la notación más sencilla posible, cuando no sea probable que esto nos induzca a error, utilizaremos $\sum^{n} Y$ para indicar la suma de los ítems de un grupo y $\sum^{an} Y$ para indicar la suma de todos los ítems de la tabla. La suma de los ítems de cada grupo se expone en la primera fila bajo la línea horizontal. La media de cada grupo, simbolizada por $\bar{Y}$, se halla en la próxima fila y se calcula sencillamente como $\sum^{n} Y/n$. Las dos filas restantes en esa parte de la tabla 7.1 presentan $\sum^{n} Y^2$ y $\sum^{n} y^2$, para cada grupo por separado. Estas son las cantidades ya familiares, suma de los Y cuadrado y suma de cuadrados de Y.

De la suma de cuadrados de cada grupo podemos obtener una estimación de la varianza de la población de longitudes del ala de moscas domésticas. Así, en el primer grupo $\sum^{n} y^2 = 29{,}2$. Por lo tanto, nuestra estimación de la varianza de la población es

$$s^2 = \frac{\sum^{n} y^2}{(n-1)} = 29{,}2/4 = 7{,}3$$

una estimación bastante baja. Como tenemos una suma de cuadrados para cada grupo, podríamos obtener una estimación de la varianza de población de cada una de ellas. No obstante, es lógico que obtengamos una mejor estimación si de algún modo hacemos un promedio de estas diferentes estimaciones de varianzas. Esto se hace calculando una *media ponderada*. Realmente en este ejemplo bastaría una media simple, ya que cada estimación de la varianza está basada en muestras del mismo tamaño. Sin embargo,

TABLA 7.1

Siete muestras (grupos) de 5 longitudes del ala de moscas domésticas seleccionadas al azar. (Datos del experimento 5.1 y de la tabla 5.2). Media paramétrica, $\mu = 45{,}5$; varianza, $\sigma^2 = 15{,}21$.

	a grupos (a = 7)							*Cálculo de la suma de cuadrados de las medias*	*Cálculo de la suma de cuadrados total*
	1	*2*	*3*	*4*	*5*	*6*	*7*		
n individuos por grupo ($n = 5$)	41	48	40	40	49	40	41		
	44	49	50	39	41	48	46		
	48	49	44	46	50	51	54		
	43	49	48	46	39	47	44		
	42	45	50	41	42	51	42		
$\sum^n Y$	218	240	232	212	221	237	227	$\sum^a \overline{Y} = 317{,}4$	$\sum^{an} Y = 1587$
$\overline{Y}$	43,6	48,0	46,4	42,4	44,2	47,4	45,4	$\overline{\overline{Y}} = 45{,}34$	$\overline{Y} = 45{,}34$
$\sum^n Y^2$	9534	11 532	10 840	9034	9867	11 315	10 413	$\sum^a \overline{Y}^2 = 14\,417{,}24$	$\sum^{an} Y^2 = 72535$
$\sum^n y^2$	29,2	12,0	75,2	45,2	98,8	81,2	107,2	$\sum^a (\overline{Y} - \overline{\overline{Y}})^2 = 25{,}417$	$\sum^{an} y^2 = 575{,}886$

preferimos dar la fórmula general, la cual da resultados igualmente satisfactorios tanto para este caso como para casos de tamaños de muestra diferentes, en los que es necesaria la media ponderada. Una fórmula general para calcular una media ponderada de cualquier estadístico *St* es la siguiente:

$$\overline{St} = \frac{\sum_{i=1}^{i=m} w_i St_i}{\sum_{i=1}^{i=m} w_i} \tag{7.1}$$

donde se están promediando m estadísticos, St_i, cada uno ponderado por el factor w_i. En este caso cada varianza de muestreo s_i^2 se multiplica por sus grados de libertad, $w_i = n_i - 1$, dando como resultado una suma de cuadrados $(\sum y^2)_i$, ya que $(n_i - 1)s_i^2 = \sum y_i^2$. Así, el numerador de la expresión (7.1) es la suma de las sumas de cuadrados. El denominador es $\sum^a (n_i - 1) = 7 \times 4$, la suma de los grados de libertad de cada grupo. Por consiguiente, la varianza media es

$$s^2 = \frac{29{,}2 + 12{,}0 + 75{,}2 + 45{,}2 + 98{,}8 + 81{,}2 + 107{,}2}{28} = \frac{448{,}8}{28} = 16{,}029$$

Esta cantidad es una estimación de 15,21, la varianza paramétrica de las longitudes del ala de moscas domésticas. Esta estimación, basada en 7 estimaciones independientes de varianzas de grupos, se denomina *varianza media intragrupos* o simplemente *varianza intragrupos*. Nótese que utilizamos la expresión *intra*grupos, aunque en capítulos previos utilizábamos el término varianza *de* grupos. La razón de que hagamos esto es que hasta ahora todas las estimaciones de varianza utilizadas para calcular la varianza media han resultado de sumas de cuadrados que miden la variación dentro de una columna. Como veremos más abajo, se pueden calcular también varianzas entre grupos, separando por grupos limítrofes.

Para obtener una segunda estimación de la varianza de la población tratamos las medias de los siete grupos, $\overline{Y}$, como si fuesen una muestra de siete observaciones. Los estadísticos que resultan se exponen en la parte superior derecha de la tabla 7.1, titulada Cálculo de la suma de cuadrados de las medias. En este ejemplo, hay siete medias; en el caso general habrá a medias. Calculamos primero $\sum^a \overline{Y}$, la suma de las medias. Nótese que este simbolismo es un tanto chapucero. Para ser completamente correcto deberíamos identificar esta cantidad como $\sum_{i=1}^{i=a} \overline{Y}_i$, sumando las medias desde el grupo 1 hasta el grupo a. La cantidad calculada a continuación es $\overline{\overline{Y}}$, la media total de las medias de grupos, calculada como $\overline{\overline{Y}} = \sum^a \overline{Y}/a$. La suma de las siete medias es $\sum^a \overline{Y} = 317{,}4$ y la media total es $\overline{\overline{Y}} = 45{,}34$, una aproximación bastante cercana a la media paramétrica $\mu = 45{,}5$. La suma de cuadrados representa las desviaciones de las medias de grupo respecto de la media total, $\sum^a (\overline{Y} - \overline{\overline{Y}})^2$. Para esto necesitamos en primer lugar la cantidad $\sum^a \overline{Y}^2$, que es igual a 14 417,24. La fórmula de cálculo habitual para la suma de cuadrados aplicada a estas medias es $\sum^a \overline{Y}^2 - [(\sum^a \overline{Y})^2/a] = 25{,}417$. De la suma de cuadrados de las medias obtenemos una *varianza entre las medias* de la forma convencional como sigue: $\sum^a (\overline{Y} - \overline{\overline{Y}})^2/(a - 1)$. Dividimos

por $a - 1$ en lugar de $n - 1$ porque la suma de cuadrados se basa en a ítems (medias). Así, la varianza de las medias $s^2_{\bar{Y}} = 25{,}417/6 = 4{,}2362$. Hemos visto en el capítulo 6, expresión (6.1), que cuando se muestrea al azar de una sola población

$$\sigma^2_{\bar{Y}} = \frac{\sigma^2}{n}$$

y por tanto

$$\sigma^2 = n\sigma^2_{\bar{Y}}$$

Así, podemos estimar una varianza de ítems multiplicando la varianza de las medias por el tamaño de muestra en que están basadas las medias (suponiendo que hayamos muestreado de una sola población). Cuando hacemos esto para nuestro ejemplo actual obtenemos $s^2 = 5 \times 4{,}2362 = 21{,}181$. Esta es una segunda estimación de la varianza paramétrica 15,21. No es tan próxima al valor correcto como la estimación previa basada en la varianza media intragrupos, pero esto es de esperar ya que está basada en 7 "observaciones" solamente. Necesitamos un nombre que describa esta varianza para distinguirla tanto de la varianza de las medias de la cual se ha calculado como de la varianza intragrupos con la cual se comparará. La denominaremos *varianza entre grupos*; es n veces la varianza de las medias y es un estimador independiente de la varianza paramétrica σ^2 de las longitudes del ala de moscas domésticas. En esta etapa puede no quedar claro por qué las dos estimaciones de σ^2 que hemos obtenido, la varianza intragrupos y la varianza entre grupos, son independientes. Rogamos se admita que en verdad lo son. Aunque esto no es de ningún modo una prueba de independencia, obsérvese que en este ejemplo las dos estimaciones son realmente diferentes, siendo 16,029 para la varianza intragrupos y 21,181 para la varianza entre grupos.

TABLA 7.2

Datos ordenados para el análisis de la varianza simple de clasificación sencilla, de forma completamente aleatoria.

		a grupos									
		1		*2*		*3*	...	*i*	...	*a*	
n ítems	1	Y_{11}		Y_{21}		Y_{31}	...	Y_{i1}	...	Y_{a1}	
	2	Y_{12}		Y_{22}		Y_{32}	...	Y_{i2}	...	Y_{a2}	
	3	Y_{13}		Y_{23}		Y_{33}	...	Y_{i3}	...	Y_{a3}	
	⋮	⋮		⋮		⋮		⋮		⋮	
	j	Y_{1j}		Y_{2j}		Y_{3j}	...	Y_{ij}	...	Y_{aj}	
	⋮	⋮		⋮		⋮		⋮		⋮	
	n	Y_{1n}		Y_{2n}		Y_{3n}	...	Y_{in}	...	Y_{an}	
Sumas		$\sum^n Y$	$\sum^n Y_1$		$\sum^n Y_2$		$\sum^n Y_3$	...	$\sum^n Y_i$	...	$\sum^n Y_a$
Medias		$\bar{Y}$	$\bar{Y}_1$		$\bar{Y}_2$		$\bar{Y}_3$	...	$\bar{Y}_i$	...	$\bar{Y}_a$

Vamos a revisar lo que hemos hecho hasta ahora, expresándolo de un modo más ordenado. La tabla 7.2 representa una tabla generalizada para datos como los de las muestras de longitudes del ala de moscas domésticas. Cada longitud del ala individual se representa por Y, con subíndices que indican su posición en la tabla de datos. La longitud del ala de la mosca j de la muestra o grupo i se expresa por Y_{ij}. Así se observará que el primer subíndice cambia con cada columna representando un grupo de la tabla, y el segundo subíndice cambia con cada fila representando un ítem individual. Utilizando esta notación podemos calcular la varianza de la muestra 1 como

$$\frac{1}{n-1}\sum_{j=1}^{j=n}(Y_{1j} - \bar{Y}_1)^2$$

La varianza intragrupos, que es la varianza media de las muestras, se calcula como

$$\frac{1}{a(n-1)}\sum_{i=1}^{i=a}\sum_{j=1}^{j=n}(Y_{ij} - \bar{Y}_i)^2$$

Obsérvese la doble suma. Significa que ponemos en primer lugar $i = 1$ (siendo i el índice del Σ exterior), cuando comenzamos por el primer grupo. Sumamos las desviaciones cuadráticas de todos los ítems respecto de la media del primer grupo, variando el índice j del Σ interior desde 1 hasta n en el proceso. Después volvemos a la suma exterior, ponemos $i = 2$, sumamos las desviaciones cuadráticas para el grupo 2 desde $j = 1$ hasta $j = n$. Este proceso se continúa hasta que i, el índice del Σ exterior, se lleva hasta a. En otras palabras, sumamos en primer lugar todas las desviaciones cuadráticas dentro de un grupo y añadimos esta suma a las sumas similares de todos los demás grupos. La varianza entre grupos se calcula como

$$\frac{n}{a-1}\sum_{i=1}^{i=a}(\bar{Y}_i - \bar{\bar{Y}})^2$$

Ahora que tenemos dos estimadores independientes de la varianza de la población, ¿qué haremos con ellos? Podríamos querer hallar si en realidad estiman el mismo parámetro. Para probar esta hipótesis necesitamos una prueba estadística que evalúe la probabilidad de que las dos varianzas de muestra sean de la misma población. Esta prueba utiliza la distribución F, que se considera a continuación.

7.2 La distribución F

Vamos a planear otro experimento de muestreo. Este es bastante pesado, por lo cual no pediremos que se realice. Supóngase que se está muestreando al azar de una población normalmente distribuida, tal como las longitudes del ala de moscas domésticas de media μ

y varianza σ^2. El procedimiento de muestreo consiste en muestrear primero n_1 ítems y calcular su varianza s_1^2, seguido por el muestreo de n_2 ítems y cálculo de su varianza s_2^2. Los tamaños de muestra n_1 y n_2 pueden ser iguales o no entre sí pero son fijos para cualquier experimento de muestreo. Así, por ejemplo, podríamos muestrear siempre 8 longitudes del ala para la primera muestra (n_1) y 6 para la segunda (n_2). Una vez que se ha obtenido cada par de valores, calculamos

$$F_s = \frac{s_1^2}{s_2^2}$$

Esta razón será próxima a 1 porque estas varianzas son estimaciones de la misma cantidad. Su valor real dependerá de las magnitudes relativas de las varianzas s_1^2 y s_2^2. Si tomamos repetidamente muestras de tamaños n_1 y n_2, calculando las razones F_s de sus varianzas, de hecho la media de estas razones se aproximará a $(n_2 - 1)/(n_2 - 3)$, que es próximo a 1,0, cuando n_2, es grande. La distribución esperada de este estadístico se denomina *distribución F* en honor de R. A. Fisher. Esta es otra distribución descrita por una función matemática complicada de la que no es necesario que nos ocupemos ahora. A diferencia de las distribuciones t y χ^2, la forma de la distribución F está determinada por *dos* valores para ν_1 y ν_2 grados de libertad. Así, para cada combinación posible de valores ν_1, ν_2, variando cada ν desde 1 hasta infinito, existe una distribución F diferente. Recuérdese que la distribución F es una distribución de probabilidad teórica lo mismo que las distribuciones t y χ^2. Las razones de varianzas basadas en las varianzas de muestreo, s_1^2/s_2^2, son estadísticos de muestreo que pueden seguir o no la distribución F. Por lo tanto, hemos distinguido la razón de varianzas de muestreo llamándola F_s, de acuerdo con nuestra costumbre comúnmente aceptada de escoger símbolos diferentes para estadísticos de muestreo y distribuciones de probabilidad (tales como t_s y X^2 para t y χ^2).

Hemos discutido la generación de una distribución F tomando repetidamente dos muestras de la misma distribución normal. También podríamos haberla originado por muestreo de dos distribuciones normales distintas que difieren en su media, pero idénticas en sus varianzas paramétricas, es decir, con $\mu_1 \neq \mu_2$ pero $\sigma_1^2 = \sigma_2^2$. Así obtenemos una distribución F tanto si las muestras proceden de la misma población normal como de diferentes, siempre que sus varianzas sean idénticas.

La figura 7.1 presenta varias distribuciones F típicas. Para muy pocos grados de libertad la distribución tiene forma de ∪ pero se hace gibosa y fuertemente inclinada hacia la derecha al aumentar ambos grados de libertad. La tabla V presenta la distribución de probabilidad acumulativa de F para tres valores de probabilidad seleccionados. Los valores de la tabla representan $F_{\alpha[n_1, n_2]}$, donde α es la proporción de la distribución F a la derecha del valor dado de F (en una cola) y ν_1, ν_2 son los grados de libertad pertenecientes al numerador y al denominador de la razón de varianzas, respectivamente. La tabla está ordenada de modo que por la parte superior se lee ν_1, los grados de libertad pertenecientes a la varianza superior (numerador) y a lo largo del margen izquierdo se lee ν_2, los grados de libertad pertenecientes a la varianza inferior (denominador). En cada intersección de valores de grados de libertad registramos tres valores de F, decrecientes en magnitud, de α. Por ejemplo, una distribución F con $\nu_1 = 6$, $\nu_2 = 24$ es 2,51 para $\alpha = 0{,}005$. Con esto queremos decir que 0,05 del área bajo la curva se halla a la derecha de $F = 2{,}51$.

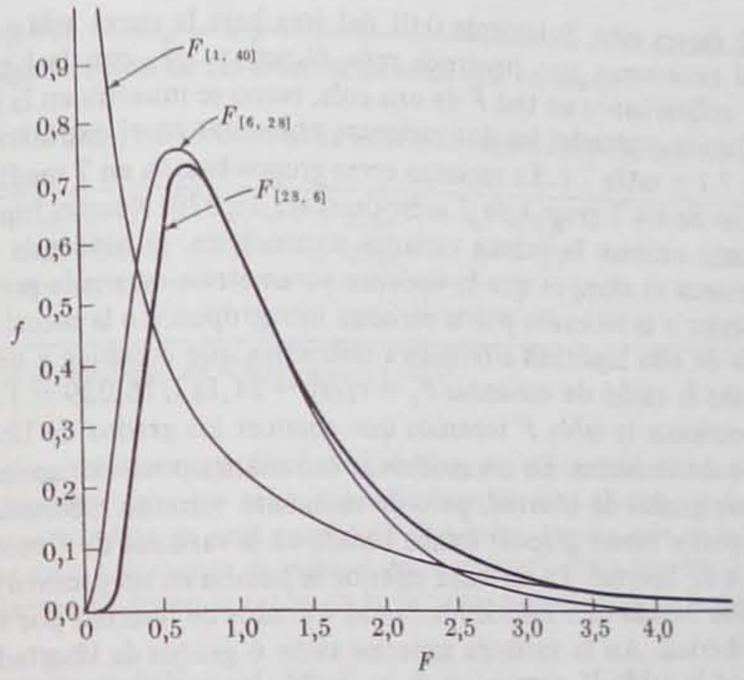


Fig. 7.1. Tres distribuciones F típicas.

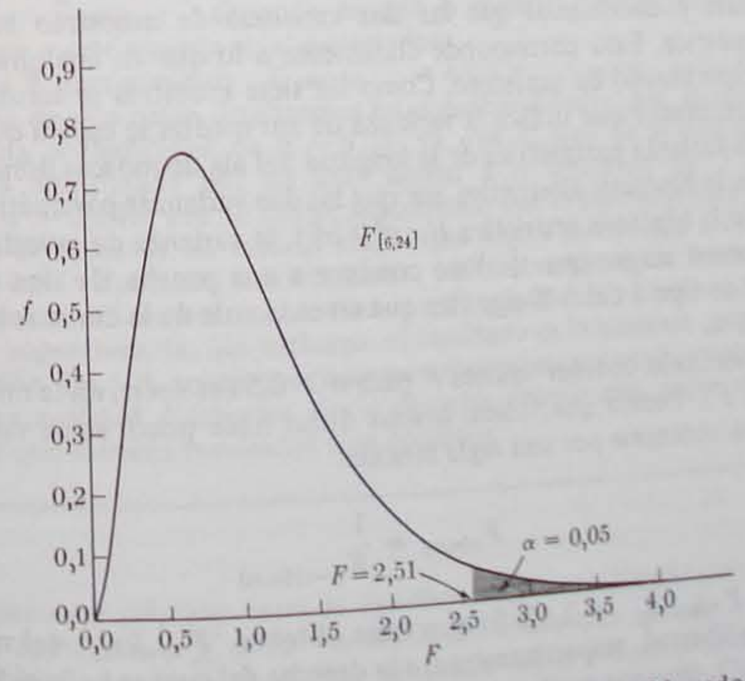


Fig. 7.2. Curva de frecuencia de la distribución F para 6 y 24 grados de libertad, respectivamente. En $F = 2{,}51$ se separa una región crítica del 5 % en una cola.

La figura 7.2 ilustra esto. Solamente 0,01 del área bajo la curva está a la derecha de $F = 3{,}67$. Así, si tuviésemos una hipótesis nula $H_0: \sigma_1^2 = \sigma_2^2$, con la hipótesis alternativa $H_1: \sigma_1^2 > \sigma_2^2$, utilizaríamos un test F de una cola, como se muestra en la figura 7.2.

Ahora podemos contrastar las dos varianzas obtenidas en el experimento de muestreo de la sección 7.1 y tabla 7.1. La varianza entre grupos basada en 7 medias era 21,180 y la varianza dentro de los 7 grupos de 5 individuos era 16,029. Nuestra hipótesis nula es que las dos varianzas estiman la misma varianza paramétrica, la hipótesis alternativa en un análisis de varianza es siempre que la varianza paramétrica estimada por la varianza entre grupos es superior a la estimada por la varianza intragrupos. En la sección 7.4 se explicará el fundamento de esta hipótesis alternativa restrictiva que conduce a una prueba de una cola. Calculamos la razón de varianzas $F_s = s_1^2/s_2^2 = 21{,}181/16{,}029 = 1{,}32$. Antes de que podamos inspeccionar la tabla F tenemos que conocer los grados de libertad pertinentes para esta razón de varianzas. En un análisis de la varianza posterior aprenderemos fórmulas sencillas para grados de libertad, pero de momento vamos a razonarlo entre nosotros. La varianza superior (entre grupos) estaba basada en la varianza de 7 medias; por lo tanto habría 6 grados de libertad. La varianza inferior se basaba en un promedio de 7 varianzas, cada una de ellas basada en 5 individuos dando 4 grados de libertad por varianza; $7 \times 4 = 28$ grados de libertad. Así la varianza superior tiene 6 grados de libertad y la inferior 28. Si confrontamos la tabla V para $\nu_1 = 6$, $\nu_2 = 24$, los argumentos más próximos en la tabla, encontramos que $F_{0{,}05[6,24]} = 2{,}51$. Para $F = 1{,}32$, que corresponde al valor F_s realmente obtenido, α es claramente mayor que 0,05. Así, podemos esperar que más del 5 % de todas las razones de varianza de muestras basadas en 6 y 28 grados de libertad, respectivamente, tengan valores F_s mayores que 1,32. No tenemos pruebas para rechazar la hipótesis nula y concluimos que las dos varianzas de muestreo estiman la misma varianza paramétrica. Esto corresponde claramente a lo que en cualquier caso sabíamos por nuestro experimento de muestreo. Como las siete muestras se tomaron de la misma población, el estimador que utiliza la varianza de sus medias se espera que produzca otro estimador de la varianza paramétrica de la longitud del ala de moscas domésticas.

Siempre que la hipótesis alternativa sea que las dos varianzas paramétricas son desiguales (en lugar de la hipótesis restrictiva $H_1: \sigma_1^2 > \sigma_2^2$), la varianza de muestreo s_1^2 podría ser tanto menor como mayor que s_2^2. Esto conduce a una prueba de dos colas y en estos casos un error de tipo I del 5 % significa que en cada cola de la curva se hallarán regiones críticas del 2 ½ %.

A veces es necesario obtener valores F para $\alpha > 0{,}5$ (es decir, en la mitad izquierda de la distribución F). Puesto que, como hemos dicho hace poco, estos valores rara vez se tabulan, pueden obtenerse por una regla sencilla:

$$F_{\alpha[\nu_1,\nu_2]} = \frac{1}{F_{(1-\alpha)[\nu_2,\nu_1]}} \tag{7.2}$$

Por ejemplo, $F_{0{,}05[5,24]} = 2{,}62$. Si queremos obtener $F_{0{,}95[5,24]}$ (el valor de F para 5 y 24 grados de libertad, respectivamente, a la derecha del cual se halla el 95 % del área de la distribución F), tenemos que buscar primero $F_{0{,}05[24,5]} = 4{,}53$. Entonces $F_{0{,}95[5,24]}$ es el recíproco de 4,53, que corresponde a 0,221. Así, el 95 % de una distribución F para 5 y 24 grados de libertad está a la derecha de 0,221.

CUADRO 7.1

Contraste de significación de las diferencias entre dos varianzas.

Supervivencia, en días, de la cucaracha *Blattella vaga* cuando se mantiene sin comida ni agua.

Hembras	$n_1 = 10$	$\bar{Y}_1 = 8{,}5$ días	$s_1^2 = 3{,}6$
Machos	$n_2 = 10$	$\bar{Y}_2 = 4{,}8$ días	$s_2^2 = 0{,}9$

$H_0: \sigma_1^2 = \sigma_2^2$ $\quad$ $H_1: \sigma_1^2 \neq \sigma_2^2$

Fuente: Datos modificados de Willis y Lewis (1957).

La hipótesis alternativa es que las dos varianzas son diferentes. No tenemos fundamento para suponer que un sexo sea más variable que el otro. A la vista de la hipótesis alternativa ésta es una prueba de dos colas. Puesto que en la tabla V y en muchas otras tablas solamente se expone ampliamente la cola derecha de la distribución F, calculamos F_s como la razón de la varianza mayor sobre la menor:

$$F_s = \frac{s_1^2}{s_2^2} = \frac{3{,}6}{0{,}9} = 4{,}00$$

Debido a que la prueba es de dos colas, buscamos el valor crítico de $F_{\alpha/2[\nu_1,\nu_2]}$, donde α es el error de tipo I aceptado, $\nu_1 = n_1 - 1$ y $\nu_2 = n_2 - 1$, los grados de libertad para la varianza superior e inferior, respectivamente. El que busquemos $F_{\alpha/2[\nu_1,\nu_2]}$ o $F_{\alpha/2[\nu_2,\nu_1]}$ depende de que la muestra 1 o la muestra 2 tenga la varianza mayor y esté situada en el numerador.

En la tabla V encontramos $F_{.025[9,9]} = 4{,}03$ y $F_{.05[9,9]} = 3{,}18$. Como ésta es una prueba de dos colas, duplicamos estas probabilidades. Así, el valor F de 4,03 representa una probabilidad de $\alpha = 0{,}05$, ya que el área de la cola derecha de $\alpha = 0{,}025$ está equilibrada por un área similar a la izquierda de $F_{.975[9,9]} = 1/F_{.025[9,9]} = 0{,}248$. Por lo tanto, suponiendo que la hipótesis nula sea cierta, la probabilidad de observar un valor F mayor que 4,00 y menor que $1/4{,}00 = 0{,}25$ es $0{,}10 > P > 0{,}05$. Hablando estrictamente, las dos varianzas de muestreo, no son significativamente diferentes, los dos sexos eran igualmente variables en cuanto a su duración de supervivencia. Sin embargo, el resultado es lo bastante próximo al nivel de significación del 5 % como para hacernos sospechar que posiblemente las varianzas fuesen en realidad diferentes. Sería deseable repetir este experimento con la esperanza de que saliesen resultados más decisivos.

Hay una importante relación entre la distribución F y la distribución χ^2. Se puede recordar que la razón $X^2 = \sum y^2/\sigma^2$ se distribuía como una χ^2 con $n - 1$ grados de libertad. Si se divide el numerador de esta expresión por $n - 1$, se obtiene la razón $F_s = s^2/\sigma^2$, que es una razón de varianzas con una distribución esperada de $F_{[n-1,\infty]}$. Los grados de libertad del numerador son $n - 1$, los grados de libertad de la suma de cuadrados o varianza de muestreo; los grados de libertad del denominador se consideran infinitos

porque solamente basándonos en un número infinito de ítems podemos obtener la verdadera varianza paramétrica de una población. Por lo tanto al dividir un valor de X^2 por $n - 1$ grados de libertad, obtenemos un valor F_s con $n - 1$ e ∞ *g.l.*, respectivamente. En general, $\chi^2_{[\nu]}/\nu = F_{[\nu,\infty]}$. Podemos convencernos de esto examinando las tablas F y χ^2. En la tabla χ^2 (tabla **IV**) encontramos que $\chi^2_{0,05[10]} = 18{,}307$. Dividiendo este valor por 10 *g.l.* obtenemos 1,8307. En la tabla F (tabla **V**) encontramos, para $\nu_1 = 10$, $\nu_2 = \infty$, que $F_{0,05[10,\infty]} = 1{,}83$. Así, los dos estadísticos de significación están estrechamente relacionados y, al carecer de una tabla χ^2, podríamos arreglarnos con una tabla F solamente, utilizando los valores de $\nu F_{[\nu,\infty]}$ en lugar de $\chi^2_{[\nu]}$.

Antes de volver al análisis de la varianza, primero aplicaremos nuestro recién adquirido conocimiento de la distribución F para contrastar una hipótesis acerca de dos varianzas muestrales.

7.3 La hipótesis $H_0: \sigma_1^2 = \sigma_2^2$

En la tabla 7.1 se presenta un contraste de la hipótesis nula de que dos poblaciones normales representadas por dos muestras tienen la misma varianza. Como se verá más adelante, la aceptación de esta hipótesis nula es un prerrequisito para algunas pruebas que llevan a una decisión de si las medias de dos muestras vienen de la misma población. No obstante, esta prueba es de interés por derecho propio. Repetidas veces será necesario probar si dos muestras tienen la misma varianza. En genética podemos encontrarnos con la necesidad de conocer si una generación filial es más variable para un carácter que la generación paterna. En sistemática nos gustaría descubrir si dos poblaciones locales son igualmente variables. En biología experimental querríamos demostrar bajo cuál de dos procedimientos experimentales serán más variables las lecturas. En general sería preferible el procedimiento menos variable; si ambos fuesen igualmente variables, el investigador seguiría el que fuese más sencillo o menos costoso de emprender.

7.4 Heterogeneidad entre medias de muestreo

Ahora modificaremos los datos de la tabla 7.1, discutidos en la sección 7.1. Supongamos que estos siete grupos de moscas domésticas no representaban muestras al azar de la misma población, sino que resultaban del siguiente experimento. Cada muestra se había cultivado en un recipiente de cultivo distinto y el medio de cada uno de los recipientes se había preparado de diferente modo. Unos tenían más agua, otros más azúcar, otros más materia sólida. Vamos a suponer que la muestra 7 representa el medio estándar frente al cual nos proponemos comparar las otras muestras. Los diversos cambios en el medio afectan a los tamaños de las moscas que salen de él; éste a su vez afecta a las longitudes del ala que hemos medido.

Supongamos los efectos siguientes resultantes del tratamiento del medio:

Medio 1 – disminuye la longitud media del ala de una muestra en 5 unidades.
2 – disminuye la longitud media del ala de una muestra en 2 unidades.
3 – no cambia la longitud media del ala de una muestra.
4 – aumenta la longitud media del ala de una muestra en 1 unidad.
5 – aumenta la longitud media del ala de una muestra en 1 unidad.
6 – aumenta la longitud media del ala de una muestra en 5 unidades.
7 – (control) no cambia la longitud media del ala de una muestra.

Simbolizaremos el efecto de tratamiento i como α_i. (Nótese que este empleo de α no está relacionado con su uso como símbolo para error de tipo I). Lamentablemente hay más símbolos necesarios en estadística que letras en los alfabetos griego y romano. Así α_i toma los valores siguientes para los efectos de tratamiento anteriores:

$$\alpha_1 = -5 \qquad \alpha_4 = 1$$
$$\alpha_2 = -2 \qquad \alpha_5 = 1$$
$$\alpha_3 = 0 \qquad \alpha_6 = 5$$
$$\alpha_7 = 0$$

Nótese que $\sum^a \alpha_i = 0$; esto es, la suma de los efectos se anula. Esta es una propiedad conveniente que generalmente se da como cierta, pero es innecesaria para nuestro argumento. Ahora podemos modificar la tabla 7.1 sumando los valores apropiados de α_i a cada muestra. En la muestra 1 el valor de α_1 es -5; por tanto la primera longitud del ala, que era 41 (véase tabla 7.1), se convierte ahora en 36, la segunda longitud del ala, anteriormente 44, se convierte en 39, y así sucesivamente. Para la segunda muestra α_2 es -2, transformando la primera longitud del ala de 48 a 46. Donde α_i sea 0, las longitudes del ala no varían; donde α_i sea positivo, se incrementan en la magnitud indicada. Los valores transformados pueden examinarse en la tabla 7.3, la cual se dispone de forma idéntica a la tabla 7.1.

Ahora repetimos nuestros cálculos previos. Primero calculamos la suma de cuadrados de la primera muestra y encontramos que es 29,2. Si comparamos este valor con la suma de cuadrados de la primera muestra en la tabla 7.1, se encuentra que los dos valores son idénticos. Igualmente todos los demás valores de $\sum y^2$, la suma de cuadrados de cada grupo, son idénticos a sus valores previos. ¿Por qué ocurre esto? El efecto de sumar α_i a cada grupo es simplemente el de una codificación aditiva, ya que α_i es constante para cualquier grupo. En el apéndice A1.2 vimos que las codificaciones aditivas no afectan a las sumas de cuadrados ni varianzas. Por consiguiente, no solamente cada suma de cuadrados diferente es la misma que antes sino que la varianza media intragrupos continúa siendo 16,029. Ahora vamos a calcular la varianza de las medias. Esta es $100{,}617/6 = 16{,}770$, que es un valor mucho mayor que el de la varianza de medias encontrada antes, 4,236. Cuando multiplicamos por $n = 5$ para obtener una estimación de σ^2 obtenemos la varianza de los grupos, que ahora es 83,848 y ya no es ni siquiera próxima a una

TABLA 7.3

Datos de la tabla 7.1 con efectos de tratamiento fijo α_i o efectos aleatorios A_i añadidos a cada muestra.

	a grupos (a = 7)							*Cálculo de la suma de cuadrados de medias*	*Cálculo de la suma de cuadrados total*
	1	*2*	*3*	*4*	*5*	*6*	*7*		
α_i o A_i	−5	−2	0	+1	+1	+5	0		
	36	46	40	41	50	45	41		
	39	47	50	40	42	53	46		
n individuos por grupo ($n = 5$)	43	47	44	47	51	56	54		
	38	47	48	47	40	52	44		
	37	43	50	42	43	56	42		
$\sum^{n} Y$	193	230	232	217	226	262	227	$\sum^{a} \bar{Y} = 317{,}4$	$\sum^{an} Y = 1587$
$\bar{Y}$	38,6	46,0	46,4	43,4	45,2	52,4	45,4	$\bar{Y} = 45{,}34$	$\bar{Y} = 45{,}34$
	$\binom{43,6}{-5}$	$\binom{48,0}{-2}$	$\binom{46,4}{+0}$	$\binom{42,4}{+1}$	$\binom{44,2}{+1}$	$\binom{47,4}{+5}$	$\binom{45,4}{+0}$	$\left(45{,}34 + \frac{-5-2+1+1+5}{7}\right)$	
$\sum^{n} Y^2$	7479	10 592	10 840	9463	10 314	13 810	10 413	$\sum^{a} \bar{Y}^2 = 14\,492{,}44$	$\sum^{an} Y^2 = 72\,911$
$\sum^{n} y^2$	29,2	12,0	75,2	45,2	98,8	81,2	107,2	$\sum^{a} (\bar{Y} - \bar{\bar{Y}})^2 = 100{,}617$	$\sum^{an} y^2 = 951{,}886$

TABLA 7.4

Datos de la tabla 7.3 ordenados del mismo modo que en la tabla 7.2.

	a grupos					
	1	*2*	*3*	... *i* ...		*a*
n ítems 1	$Y_{11} + \alpha_1$	$Y_{21} + \alpha_2$	$Y_{31} + \alpha_3$	... $Y_{i1} + \alpha_i$	...	$Y_{a1} + \alpha_a$
2	$Y_{12} + \alpha_1$	$Y_{22} + \alpha_2$	$Y_{32} + \alpha_3$	... $Y_{i2} + \alpha_i$	...	$Y_{a2} + \alpha_a$
3	$Y_{13} + \alpha_1$	$Y_{23} + \alpha_2$	$Y_{33} + \alpha_3$	... $Y_{i3} + \alpha_i$	...	$Y_{a3} + \alpha_a$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
j	$Y_{1j} + \alpha_1$	$Y_{2j} + \alpha_2$	$Y_{3j} + \alpha_3$	... $Y_{ij} + \alpha_i$	...	$Y_{aj} + \alpha_a$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	$Y_{1n} + \alpha_1$	$Y_{2n} + \alpha_2$	$Y_{3n} + \alpha_3$	... $Y_{in} + \alpha_i$	...	$Y_{an} + \alpha_a$
Sumas	$\sum^{n} Y_1 + n\alpha_1$	$\sum^{n} Y_2 + n\alpha_2$	$\sum^{n} Y_3 + n\alpha_3$	... $\sum^{n} Y_i + n\alpha_i$	...	$\sum^{n} Y_a + n\alpha_a$
Medias	$\bar{Y}_1 + \alpha_1$	$\bar{Y}_2 + \alpha_2$	$\bar{Y}_3 + \alpha_3$	... $\bar{Y}_i + \alpha_i$	...	$\bar{Y}_a + \alpha_a$

estimación de σ^2. Repetimos la prueba F con las nuevas varianzas y encontramos que $F_s = 83{,}848/16{,}029 = 5{,}23$, que es mucho mayor que el más próximo valor crítico de $F_{0{,}01[6,24]} = 2{,}51$. De hecho, el F_s observado es mayor que $F_{0{,}05[6,24]} = 3{,}67$. Claramente, la varianza superior, que representa la varianza entre grupos, se ha hecho significativamente mayor. Es sumamente improbable que las dos varianzas representen la misma varianza paramétrica.

¿Qué ha ocurrido? Podemos explicarlo fácilmente por medio de la tabla 7.4, que representa simbólicamente a la tabla 7.3 del mismo modo que la tabla 7.2 representa la tabla 7.1. Observamos que a cada grupo se le ha añadido una constante α_i, y que esta constante cambia las sumas de los grupos en $n\alpha_i$ y las medias de estos grupos en α_i. En la sección 7.1 calculábamos la varianza intragrupos como

$$\frac{1}{a(n-1)} \sum_{i=1}^{i=a} \sum_{j=1}^{j=n} (Y_{ij} - \bar{Y}_i)^2$$

Si intentamos repetir esto, nuestra fórmula se complica más porque a cada Y_{ij} y a cada $\bar{Y}_i$ se le ha sumado ahora α_i. Por lo tanto escribimos

$$\frac{1}{a(n-1)} \sum_{i=1}^{i=a} \sum_{j=1}^{j=n} [(Y_{ij} + \alpha_i) - (\bar{Y}_i + \alpha_i)]^2$$

Cuando desarrollamos los paréntesis dentro de los corchetes al cuadrado, la segunda α_i cambia de signo y las α_i se anulan, dejando la expresión exactamente como antes, lo que justifica nuestra observación anterior de que la varianza intragrupos no varía pese a los efectos del tratamiento.

La varianza de las medias se ha calculado previamente por la fórmula

$$\frac{1}{a-1}\sum_{i=1}^{i=a}(\overline{Y}_i-\overline{\overline{Y}})^2$$

Sin embargo, en la tabla 7.4 vemos que la nueva media total es igual a

$$\frac{1}{a}\sum_{i=1}^{i=a}(\overline{Y}_i+\alpha_i)=\frac{1}{a}\sum_{i=1}^{i=a}\overline{Y}_i+\frac{1}{a}\sum_{i=1}^{i=a}\alpha_i=\overline{\overline{Y}}+\bar{\alpha}$$

Cuando sustituimos los nuevos valores para las medias de grupo y la media total, la fórmula aparece como

$$\frac{1}{a-1}\sum_{i=1}^{i=a}[(\overline{Y}_i+\alpha_i)-(\overline{\overline{Y}}+\bar{\alpha})]^2$$

que a su vez da

$$\frac{1}{a-1}\sum_{i=1}^{i=a}[(\overline{Y}_i-\overline{\overline{Y}})+(\alpha_i-\bar{\alpha})]^2$$

Elevando al cuadrado la expresión entre corchetes, obtenemos los términos

$$\frac{1}{a-1}\sum_{i=1}^{i=a}(\overline{Y}_i-\overline{\overline{Y}})^2+\frac{1}{a-1}\sum_{i=1}^{i=a}(\alpha_i-\bar{\alpha})^2+\frac{2}{a-1}\sum_{i=1}^{i=a}(\overline{Y}_i-\overline{\overline{Y}})(\alpha_i-\bar{\alpha})$$

El primero de estos términos lo reconocemos inmediatamente como la anterior varianza de las medias, $s^2_{\overline{Y}}$. El segundo es una cantidad nueva pero es familiar por su aspecto general. Es sin duda una varianza o al menos una cantidad análoga a una varianza. La tercera expresión es nueva. Es la llamada covarianza que todavía no nos hemos encontrado. Ahora no nos ocuparemos de ella excepto para decir que en casos tales como el presente, donde la magnitud de los efectos de tratamiento α_i es independiente de la $\overline{Y}_i$ a la cual se suman, el valor esperado de esta cantidad es cero; por lo tanto no contribuiría a la nueva varianza de las medias.

La independencia de los efectos de tratamiento y las medias de muestreo es un concepto importante que debemos comprender claramente. Si no hubiésemos aplicado diferentes tratamientos a las botellas de medio sino simplemente tratado todos los recipientes como controles, habríamos obtenido, no obstante, diferencias entre las medias de la longitud del ala. Esas son las diferencias encontradas en la tabla 7.1, con muestreo al azar de la misma población. Por casualidad algunas de estas medias son mayores y algunas menores. En nuestra planificación del experimento no podíamos predecir qué medias de muestreo serían pequeñas y cuales serían grandes. Por consiguiente, al planificar nuestros tratamientos no teníamos modo de emparejar grandes efectos de tratamiento, como el del medio 6.

con la media que por azar sería la mayor, como la de la muestra 2. Asimismo, la media de muestreo más pequeña (muestra 4) no está asociada con el efecto de tratamiento más pequeño. Solamente si la magnitud de los efectos de tratamiento se correlacionase deliberadamente con las medias de muestreo (esto sería difícil de hacer en el experimento diseñado aquí), el tercer término de la expresión, la covarianza, tendría un valor esperado distinto de cero.

El segundo término de la expresión anterior es claramente añadido como resultado de los efectos de tratamiento. Es análogo a una varianza pero no puede llamarse así, ya que no está basado en una variable aleatoria sino más bien en tratamientos deliberadamente elegidos en gran parte bajo nuestro control. Variando la magnitud y naturaleza de los tratamientos podemos modificar más o menos a voluntad la cantidad análoga a la varianza. Lo llamaremos por tanto, el *componente aditivo debido a los efectos de tratamiento.* Como los valores de α_i se disponen de modo que $\bar{\alpha}=0$, podemos volver a escribir el término central como

$$\frac{1}{a-1}\sum_{i=1}^{i=a}(\alpha_i-\bar{\alpha})^2=\frac{1}{a-1}\sum_{i=1}^{i=a}\alpha_i^2=\frac{1}{a-1}\sum^{a}\alpha^2$$

En el análisis de la varianza multiplicamos la varianza de las medias por n para estimar la varianza paramétrica de los ítems. Como se sabe, a la cantidad así obtenida la llamaremos varianza de los grupos. Cuando hacemos esto para el caso en que están presentes efectos de tratamiento obtenemos

$$n\left(s^2_{\overline{Y}}+\frac{1}{a-1}\sum^{a}\alpha^2\right)=s^2+\frac{n}{a-1}\sum^{a}\alpha^2$$

Así vemos que el estimador de la varianza paramétrica de la población se incrementa en la cantidad

$$\frac{n}{a-1}\sum^{a}\alpha^2$$

que es n veces el componente aditivo debido a los efectos de tratamiento. Encontramos que la razón de varianzas F_s es significativamente mayor de lo permitido para ser compatible con la hipótesis nula. Ahora es evidente el porqué de esto. Estábamos contrastando la razón de varianzas esperando hallar F aproximadamente igual a $\sigma^2/\sigma^2=1$. Sin embargo tenemos en realidad

$$F\approx\frac{\sigma^2+\dfrac{n}{a-1}\sum^{a}\alpha^2}{\sigma^2}$$

En esta fórmula deliberadamente presentada en forma asimétrica, queda claro que la prueba F es sensible a la presencia del componente aditivo debido a los efectos de tratamiento.

En este momento se debería tener la primera visión clara de la utilidad del análisis de la varianza. Nos permite probar si hay efectos de tratamiento aditivos, es decir, si un grupo de medias puede considerarse simplemente muestras al azar de la misma población o si los tratamientos que han afectado a cada grupo por separado han conducido a un cambio suficiente de estas medias como para que ya no puedan considerarse muestras de la misma población. Si eso es cierto, estará presente un componente aditivo debido a los efectos de tratamiento y puede detectarse por una prueba F en la prueba de significación del análisis de la varianza. En este estudio generalmente no nos interesa la magnitud de

$$\frac{n}{a-1}\sum^{a}\alpha^2$$

sino que estamos interesados en la magnitud de los diferentes valores de α_i. En nuestro ejemplo estos son los efectos de diferentes formulaciones del medio en la longitud del ala. Si en lugar de longitud del ala de moscas domésticas estuviésemos midiendo presión sanguínea en muestras de ratas, y los diferentes grupos se hubiesen sometido a diferentes drogas o diferentes dosis de la misma droga, las cantidades representarían los efectos de las drogas en la presión sanguínea, que es claramente el tema de interés para el investigador. También podemos estar interesados en el estudio de diferencias del tipo $\alpha_1 - \alpha_2$, que nos lleva a la cuestión de la significación de las diferencias entre los efectos de dos tipos cualesquiera de medio o dos drogas cualesquiera. Pero vamos un poco por delante de nuestro argumento.

Cuando el análisis de la varianza implica efectos de tratamiento del tipo recién estudiado, lo denominamos análisis de la varianza modelo I. Más adelante en este capítulo (sección 7.6), se definirá claramente el modelo I. Hay otro modelo denomindo análisis de la varianza modelo II en el cual los efectos aditivos para cada grupo no son tratamientos fijos sino que son efectos al azar. Con esto queremos decir que no hemos planeado ni fijado deliberadamente el tratamiento para ningún grupo sino que los efectos reales en cada grupo son al azar y sólo parcialmente bajo nuestro control. Supongamos que las siete muestras de moscas domésticas de la tabla 7.3 representasen la descendencia de siete hembras seleccionadas al azar de una población, cultivadas en un medio uniforme. Habría diferencias genéticas entre estas hembras y sus siete camadas reflejarían esto. La naturaleza exacta de estas diferencias es dudosa e imprevisible. Antes de que las midamos realmente no hay modo de conocer si la camada 1 tendría alas más largas que la 2, ni hay modo de controlar este experimento para que la camada 1 desarrolle de hecho alas más largas. Como hasta cierto punto podemos asegurar, los factores genéticos para la longitud del ala se distribuyen de una manera desconocida en la población de moscas domésticas (pudiéramos esperar que estuviesen normalmente distribuidos) y nuestra muestra de siete es una muestra al azar de estos factores. Otro ejemplo para un modelo II podría ser que en lugar de preparar nuestros siete cultivos de un solo lote de medio, hubiésemos preparado siete lotes por separado, uno inmediatamente después de otro y ahora estuviésemos analizando la variación entre los lotes. No nos interesarían las diferencias exactas de un lote a otro. Aun cuando éstas se midiesen no estaríamos en condiciones de interpretarlas. No habiendo modificado deliberadamente el lote 3, no tenemos idea de porqué debería producir alas más largas que el lote 2, por ejemplo. Sin embargo nos interesaríamos por la magnitud de la varianza de los efectos aditivos. Así, si utilizásemos 7 botellas de medio derivadas de un lote, podríamos esperar que la varianza de las medias de las botellas fuese $\sigma^2/5$, puesto que había 5 moscas por botella. Pero cuando se basan en diferentes lotes de medio podría esperarse que la varianza fuese mayor, porque entrarían en juego todos los imponderables accidentes de formulación y diferencias ambientales durante la preparación del medio, que hacen que un lote de medio sea diferente de otro. El interés se centraría en el componente aditivo de varianza procedente de diferencias entre lotes. Igualmente, en el otro ejemplo nos interesaríamos por el componente aditivo de la varianza que procede de las diferencias genéticas entre las hembras.

Ahora echaremos una ojeada rápida a la formulación algebraica del análisis de varianza en el caso del modelo II. En la tabla 7.3 la segunda fila a la cabeza de las columnas de datos muestra no solamente α_i sino también A_i, que es el símbolo que utilizaremos para un efecto de grupo aleatorio. Utilizaremos una letra mayúscula para indicar que el efecto es una variable. El álgebra para calcular las dos estimaciones de la varianza de población es la misma que en el modelo I, excepto que imaginamos que α_i es sustituido por A_i en la tabla 7.4. La estimación de la varianza entre medias representa ahora la cantidad

$$\frac{1}{a-1}\sum_{i=1}^{i=a}(\overline{Y}_i-\overline{Y})^2+\frac{1}{a-1}\sum_{i=1}^{i=a}(A_i-\overline{A})^2+\frac{2}{a-1}\sum_{i=1}^{i=a}(\overline{Y}_i-\overline{Y})(A_i-\overline{A})$$

El primer término es la varianza de las medias $s^2_{\overline{Y}}$ como antes, y el último término es la covarianza entre las medias de grupo y los efectos aleatorios A_i, cuyo valor esperado es cero (como anteriormente) porque los efectos aleatorios son independientes de la magnitud de las medias. El término intermedio es una verdadera varianza ya que A_i es una variable aleatoria. Lo simbolizamos por s^2_A y lo denominamos *componente aditivo de la varianza entre grupos*. Representaría el componente aditivo de la varianza entre hembras o entre lotes de medio, dependiendo de cual de los diseños antes discutidos recordemos. La existencia de este componente aditivo de la varianza se demuestra por la prueba F. Si los grupos son muestras al azar podemos esperar que F se aproxime a $\sigma^2/\sigma^2 = 1$, pero con un componente aditivo de varianza la razón esperada es

$$F \approx \frac{\sigma^2 + n\sigma^2_A}{\sigma^2}$$

Nótese que σ^2_A, el valor paramétrico de s^2_A, está multiplicado por n, ya que tenemos que multiplicar por n la varianza de las medias para obtener un estimador independiente de la varianza de la población. En un modelo II no nos interesamos por la magnitud de A_i ni por diferencias tales como $A_1 - A_2$, sino que nos interesa la magnitud de σ^2_A y su magnitud relativa con respecto a σ^2, que generalmente se expresa como el porcentaje $100s^2_A/(s^2 + s^2_A)$. Puesto que la varianza entre grupos estima $\sigma^2 + n\sigma^2_A$, podemos calcular s^2_A como

$$\frac{1}{n}\text{(varianza entre grupos} - \text{varianza intragrupos)} = \frac{1}{n}[(s^2 + ns^2_A) - s^2] = \frac{1}{n}(ns^2_A) = s^2_A$$

Para el ejemplo que nos ocupa $s_A^2 = \frac{1}{8}(83{,}848 - 16{,}029) = 13{,}56$. Este componente aditivo de la varianza entre grupos es

$$\frac{100 \times 13{,}56}{16{,}029 + 13{,}56} = \frac{1356}{29{,}589} = 45{,}83\%$$

de la suma de las varianzas entre e intragrupos. El modelo II se discutirá formalmente al final de este capítulo (sección 7.7); los métodos para estimar componentes de varianza se tratarán con detalle en el próximo capítulo.

7.5 Descomposición de la suma de cuadrados total y los grados de libertad

Hasta ahora hemos ignorado otra varianza que puede calcularse a partir de los datos de la tabla 7.1. Si eliminamos la clasificación en grupos, podemos considerar que los datos de moscas domésticas son una muestra única de $an = 35$ longitudes del ala y calcular la media y varianza de estos ítems de manera convencional. Las diversas cantidades necesarias para este cálculo se exponen en la última columna de la derecha en las tablas 7.1 y 7.3, encabezada como Cálculo de la suma de cuadrados total. Obtenemos una media de $\bar{Y} = 45{,}34$ para la muestra de la tabla 7.1, que es por supuesto, la misma que la cantidad $\bar{\bar{Y}}$ calculada previamente a partir de las siete medias de grupo. La suma de cuadrados de los 35 ítems es 575,886, que da una varianza de 16,938 cuando se divide por 34 grados de libertad. Repitiendo estos cálculos para los datos de la tabla 7.3 obtenemos $\bar{Y} = 45{,}34$ (la misma que en la tabla 7.1 porque $\sum^a \alpha_i = 0$) y $s^2 = 27{,}997$, que es considerablemente mayor que la correspondiente varianza de la tabla 7.1. La varianza total calculada a partir de los *an* ítems es otro estimador de σ^2. En el primer caso es un buen estimador, pero en la segunda muestra (tabla 7.3), en que están presentes componentes aditivos debidos a los efectos del tratamiento o componentes aditivos de varianza, es un estimador pobre de la varianza de la población.

No obstante, el objeto de calcular la varianza total en un análisis de la varianza no es para utilizarla como otro estimador más de σ^2 sino para facilitar el cálculo. Esto se ve mejor cuando disponemos nuestros resultados en una *tabla de análisis de la varianza* convencional, como se muestra en la tabla 7.5. Esta tabla está dividida en cuatro columnas. La primera identifica el origen de la variación como entre grupos, intragrupos y total (grupos reunidos para formar una sola muestra). La columna encabezada por *g.l.* da los grados de libertad por los que deben dividirse las sumas de cuadrados pertinentes a esa fuente de variación para obtener la varianza. El número de grados de libertad para la variación entre grupos es $a - 1$, el de la variación intragrupos es $a(n - 1)$, y el de la variación total es $an - 1$. Las dos columnas siguientes muestran las sumas de cuadrados y varianzas, respectivamente. Nótese que las sumas de cuadrados introducidas en la tabla de análisis de varianza son la suma de cuadrados entre grupos, la suma de cuadrados intragrupos, y la suma de cuadrados de la muestra total de *an* ítems. Se observa que las varianzas no se denominan como tales en el análisis de la varianza, sino que generalmente se llaman *medias cuadráticas*, ya que en un modelo I no estiman una varianza de población. Estas cantidades no son verdaderas medias de cuadrados, porque las sumas de cuadrados se

TABLA 7.5

Tabla de análisis de la varianza para los datos de la tabla 7.1.

(1)		(2)	(3)	(4)
	Origen de la variación	*g.l.*	*Sumas de cuadrados S.C.*	*Medias cuadráticas M.C.*
$\bar{Y} - \bar{\bar{Y}}$	Entre grupos	6	127,086	21,181
$Y - \bar{Y}$	Intragrupos	28	448,800	16,029
$Y - \bar{\bar{Y}}$	Total	34	575,886	16,938

dividen por los grados de libertad y no por el tamaño de la muestra. Las sumas de cuadrados y medias cuadráticas se abrevian frecuentemente como *S.C.* y *M.C.* respectivamente.

Las sumas de cuadrados y medias cuadráticas de la tabla 7.5 son las mismas que las obtenidas previamente, salvo mínimos errores de redondeo. Hay que señalar, no obstante, una propiedad importante de las sumas de cuadrados. Se han obtenido una independientemente de otra, pero cuando sumamos la *S.C.* entre grupos y la *S.C.* intragrupos obtenemos la *S.C.* total. Las sumas de cuadrados son aditivas. Otra forma de decir esto es que podemos descomponer la suma de cuadrados total en una parte debida a variación entre grupos y otra debida a variación intragrupos. Obsérvese que los grados de libertad también son aditivos y que el total de 34 *g.l.* puede descomponerse en 6 *g.l.* entre grupos y 28 *g.l.* intragrupos. Por tanto, si conocemos dos de las sumas de cuadrados cualesquiera (y sus grados de libertad correspondientes) podemos calcular la tercera y completar nuestro análisis de la varianza. Hay que hacer notar que las medias cuadráticas no son aditivas. Esto es obvio ya que generalmente $(a + b)/(c + d) \neq a/c + b/d$.

Utilizaremos la fórmula de cálculo de la suma de cuadrados (expresión 3.7) para demostrar porqué estas sumas de cuadrados son aditivas. Aunque es una desviación algebraica, se pone aquí en vez de en el apéndice porque estas fórmulas nos llevarán también a las fórmulas de cálculo propiamente dichas del análisis de la varianza. Como puede haberse sospechado, el método que hemos utilizado para obtener las sumas de cuadrados no es el procedimiento de cálculo más rápido; en la práctica empleamos un método mucho más sencillo. La suma de cuadrados de las medias en notación simplificada es

$$S.C._{\text{medias}} = \sum^a (\bar{Y} - \bar{\bar{Y}})^2 = \sum^a \bar{Y}^2 - \frac{\left(\sum^a \bar{Y}\right)^2}{a}$$

$$= \sum^a \left(\frac{1}{n}\sum^n Y\right)^2 - \frac{1}{a}\left[\sum^a \left(\frac{1}{n}\sum^n Y\right)\right]^2$$

$$= \frac{1}{n^2}\sum^a \left(\sum^n Y\right)^2 - \frac{1}{an^2}\left(\sum^a \sum^n Y\right)^2$$

Nótese que la desviación de las medias respecto de la media total, primero se reordena para adaptarse a la fórmula de cálculo [expresión (3.7)] y después se escribe cada media en términos de sus variantes constituyentes. La reunión de denominadores fuera de los signos sumatorios, da la fórmula final deseada. Para obtener la suma de cuadrados de los grupos, multiplicamos la $S.C._{medias}$ por n, como anteriormente se ha hecho. Esto da

$$S.C._{\text{grupos}} = n \times S.C._{\text{medias}} = \frac{1}{n}\sum^{a}\left(\sum^{n} Y\right)^2 - \frac{1}{an}\left(\sum^{a}\sum^{n} Y\right)^2$$

A continuación calculamos la suma de cuadrados intragrupos:

$$S.C._{\text{intragrupos}} = \sum^{a}\sum^{n}(Y - \overline{Y})^2 = \sum^{a}\left[\sum^{n} Y^2 - \frac{1}{n}\left(\sum^{n} Y\right)^2\right]$$

$$= \sum^{a}\sum^{n} Y^2 - \frac{1}{n}\sum^{a}\left(\sum^{n} Y\right)^2$$

La suma de cuadrados total representa

$$S.C._{\text{total}} = \sum^{a}\sum^{n}(Y - \overline{\overline{Y}})^2$$

$$= \sum^{a}\sum^{n} Y^2 - \frac{1}{an}\left(\sum^{a}\sum^{n} Y\right)^2$$

Ahora copiamos las fórmulas para estas sumas de cuadrados, ligeramente reordenadas como sigue:

$$S.C._{\text{grupos}} = \frac{1}{n}\sum^{a}\left(\sum^{n} Y\right)^2 - \frac{1}{an}\left(\sum^{a}\sum^{n} Y\right)^2$$

$$S.C._{\text{intragrupos}} = -\frac{1}{n}\sum^{a}\left(\sum^{n} Y\right)^2 + \sum^{a}\sum^{n} Y^2$$

$$S.C._{\text{total}} = \sum^{a}\sum^{n} Y^2 - \frac{1}{an}\left(\sum^{a}\sum^{n} Y\right)^2$$

Sumando la expresión de la $S.C._{grupos}$ a la de la $S.C._{intragrupos}$ obtenemos una cantidad que es idéntica a la recién desarrollada como $S.C._{total}$. Esta demostración explica porqué las sumas de cuadrados son aditivas.

No vamos a realizar ninguna demostración sino simplemente establecer que los grados de libertad correspondientes a las sumas de cuadrados son también aditivos. El total de grados de libertad se descompone en los grados de libertad correspondientes a la variación entre grupos y los de la variación de los ítems dentro de los grupos.

Antes de continuar vamos a revisar el significado de las tres medias cuadráticas en el

TABLA 7.6

Tabla de análisis de la varianza para los datos de la tabla 7.3.

(1)		(2)	(3)	(4)
Origen de la variación		*g.l.*	*Sumas de cuadrados S.C.*	*Medias cuadráticas M.C.*
$\overline{Y} - \overline{\overline{Y}}$	Entre grupos	6	503,086	83,848
$Y - \overline{Y}$	Intragrupos	28	448,800	16,029
$Y - \overline{\overline{Y}}$	Total	34	951,886	27,997

análisis de la varianza. La *M.C.* total es un estadístico de dispersión de los 35 (an) ítems en torno a su media, la media total 45,34. Describe la varianza de la muestra total debida a todas y a cada una de las causas y estima σ^2 cuando no hay efectos de tratamiento aditivos ni componentes aditivos de varianza entre grupos. La *M.C.* intragrupos, conocida también como la *media cuadrática de los errores* o *individual*, da la dispersión media de los 5 (n) ítems de cada grupo en torno a las medias de grupo. Si los a grupos fuesen muestras al azar de una población homogénea ordinaria, la *M.C.* intragrupos estimaría σ^2. La *M.C.* entre grupos está basada en la varianza de las medias de grupo, la cual describe la dispersión de las 7 (a) medias de grupo en torno a la media total. Si los grupos son muestras al azar de una población homogénea, la varianza esperada de su media será σ^2/n. Por lo tanto, a fin de obtener las tres varianzas del mismo orden de magnitud, multiplicamos la varianza de las medias por n para obtener la varianza entre grupos. Si no hay efectos aditivos de tratamiento ni componentes aditivos de varianza, la *M.C.* entre grupos es un estimador de σ^2. De lo contrario es un estimador de

$$\sigma^2 + \frac{n}{a-1}\sum^{a}\alpha^2 \qquad \text{o} \qquad \sigma^2 + n\sigma_A^2$$

dependiendo de que el análisis de la varianza que tenemos a mano sea modelo I o II.

Las relaciones de aditividad que acabamos de aprender son independientes de la presencia de efectos aditivos de tratamiento o efectos aleatorios. Podríamos demostrar esto algebraicamente, pero es más sencillo examinar la tabla 7.6, que resume el análisis de varianza de la tabla 7.3, en la cual se han sumado a cada muestra α_i o A_i. La relación de aditividad sigue conservándose, aunque los valores para *S.C.* entre grupos y *S.C.* total son diferentes de los de la tabla 7.5.

Otra forma de considerar la descomposición de la variación es estudiar la desviación de las medias en un caso particular. Refiriéndonos a la tabla 7.1 podemos considerar la longitud del ala del primer individuo del 7º, grupo, que es 41. Su desviación respecto de su media de grupo es

$$Y_{71} - \overline{Y}_7 = 41 - 45{,}4 = -4.4$$

La desviación de la media de grupo respecto de la media total es

$$\bar{Y}_7 - \bar{\bar{Y}} = 45{,}4 - 45{,}34 = 0{,}06$$

y la desviación de la longitud del ala individual respecto de la media total es

$$Y_{71} - \bar{\bar{Y}} = 41 - 45{,}34 = -4{,}34$$

Nótese que estas desviaciones son aditivas. La desviación del ítem respecto de la media de grupo y la de la media de grupo respecto de la media total, suman la desviación total del ítem respecto de la media total. Estas desviaciones se expresan algebraicamente como $(Y - \bar{Y}) + (\bar{Y} - \bar{\bar{Y}}) = (Y - \bar{\bar{Y}})$. Elevando al cuadrado y sumando estas desviaciones para *an* ítems resultará

$$\sum^{a}\sum^{n}(Y - \bar{Y})^2 + n\sum^{a}(\bar{Y} - \bar{\bar{Y}})^2 = \sum^{an}(Y - \bar{\bar{Y}})^2$$

Antes de elevar al cuadrado, las desviaciones estaban en la relación $a + b = c$. Después de elevar al cuadrado esperaríamos que estuviesen en la forma $a^2 + b^2 + 2ab = c^2$. ¿Qué ha ocurrido con el término doble producto correspondiente a $2ab$? Este es

$$2\sum^{an}(Y - \bar{Y})(\bar{Y} - \bar{\bar{Y}}) = 2\sum^{a}[(\bar{Y} - \bar{\bar{Y}})\sum^{n}(Y - \bar{Y})]$$

un término de tipo covarianza que siempre es cero, ya que $\sum^{n}(Y - \bar{Y}) = 0$ para cada uno de los *a* grupos (demostración en el apéndice A1.1).

En los márgenes izquierdos de las tablas que dan los resultados del análisis de varianza (tabla 7.5 y 7.6) identificamos las desviaciones representadas por cada nivel de variación. Nótese que las desviaciones se suman correctamente: la desviación entre grupos más la desviación intragrupos es igual a la desviación total de los ítems en el análisis de la varianza, $(\bar{Y} - \bar{\bar{Y}}) + (Y - \bar{Y}) = (Y - \bar{\bar{Y}})$.

7.6 Análisis de la varianza modelo I

Un punto importante a tener en cuenta es que la disposición básica de los datos, así como el cálculo propiamente dicho y la prueba de significación, en la mayor parte de los casos es igual para los dos modelos. Los fines del análisis de la varianza difieren para los dos modelos así como algunas de las pruebas suplementarias y cálculos que siguen a la prueba de significación inicial.

Ahora vamos a intentar determinar la variación encontrada en un caso de análisis de la

varianza. Esto no solamente nos llevará a una interpretación más formal del análisis de la varianza, sino que además nos dará un mayor conocimiento de la naturaleza de la variación en sí. Con miras a la discusión recurriremos nuevamente a las longitudes del ala de moscas domésticas de la tabla 7.3. Hacemos la pregunta "¿qué es lo que determina que una determinada longitud del ala adquiera el valor que presenta? La tercera longitud del ala de la primera muestra de moscas se registra como 43 unidades. ¿Cómo podemos explicar esta lectura?

En principio, no sabiendo nada más acerca de esta mosca particular, nuestra mejor suposición acerca de la longitud de su ala, es la media total de la población, la cual sabemos que es $\mu = 45{,}5$. Sin embargo, tenemos información adicional con respecto a esta mosca. Es un miembro del grupo 1, el cual se ha sometido a un tratamiento que disminuye la media del grupo en 5 unidades. Por lo tanto, $\alpha_1 = -5$ y esperaríamos que nuestro individuo Y_{13} (el tercer individuo del grupo 1) midiese $45{,}5 - 5 = 40{,}5$ unidades. Pero en realidad mide 43 unidades, que es 2,5 unidades superior al valor esperado. ¿A qué podemos atribuir esta desviación? Es la variación individual de las moscas dentro de un grupo debida a la varianza de los individuos en la población ($\sigma^2 = 15{,}21$). Todos los efectos genéticos y ambientales que hacen diferente una mosca doméstica de otra entran en juego para producir esta varianza. Por medio de experimentos cuidadosamente planeados podríamos averiguar algo acerca de la causa de esta varianza y atribuirla a ciertos factores genéticos o ambientales específicos. Además podríamos ser capaces de eliminar parte de la varianza. Por ejemplo, utilizando solamente parientes carnales (hermanos y hermanas) en una botella de cultivo cualquiera, reduciríamos la variación genética en los individuos, e indudablemente la varianza intragrupos sería más pequeña. Sin embargo, es imposible tratar de eliminar completamente toda la varianza. Aunque pudiésemos eliminar toda la varianza genética, habría, no obstante, varianza ambiental, y aún en el caso más improbable en que pudiésemos hacer desaparecer ambas, quedaría el error de medida, de modo que nunca obtendríamos exactamente la misma lectura ni siquiera en la misma mosca particular. La *M.C.* intragrupos queda siempre como una parte residual de la naturaleza de las cosas, diferente de un experimento a otro. Es por esto que la varianza intragrupos se denomina también varianza de los errores o media cuadrática de los errores. No es un error en el sentido de que se haya cometido una equivocación, sino en el sentido de que proporciona una medida de la variación con la que se tiene que contar al tratar de estimar diferencias significativas entre grupos. La varianza de los errores está compuesta de desviaciones individuales para cada individuo, simbolizada por ϵ_{ij}, el componente aleatorio de la variante individual que ocupa el lugar *j* en el grupo que ocupa el lugar *i*. En nuestro caso, $\epsilon_{13} = 2{,}5$, puesto que el valor real observado es 2,5 unidades superior a su valor esperado 40,5.

Ahora expresaremos más formalmente esta relación. En un análisis de varianza modelo I suponemos que las diferencias entre las medias de grupo, si las hay, son debidas a los efectos del tratamiento fijo determinado por el investigador. El objeto del análisis de la varianza es estimar las verdaderas diferencias entre las medias de grupo. Cada variante individual puede descomponerse como sigue:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{7.3}$$

donde $i = 1, \ldots, a, j = 1, \ldots, n$, ϵ_{ij} representa una variable independiente, normalmente distribuida con media $\bar{\epsilon}_{ij} = 0$, y varianza $\sigma_\epsilon^2 = \sigma^2$. Por lo tanto, una determinada lectura está formada por la media total de la población, μ, una desviación fija α_i de la media del grupo i respecto de la media total μ, y una desviación aleatoria ϵ_{ij} del individuo que ocupa el lugar j del grupo i respecto de su valor esperado, que es $(\mu + \alpha_i)$. Recuérdese que tanto α_i como ϵ_{ij} pueden ser positivos o negativos. El valor esperado (media) de los valores de ϵ_{ij} es cero y su varianza es la varianza paramétrica de la población, σ^2. Para que se verifiquen todos los supuestos del análisis de la varianza, la distribución de ϵ_{ij} debe ser normal.

En un análisis de la varianza modelo I examinamos diferencias del tipo $\alpha_1 - \alpha_2$ entre las medias de grupo, probando la presencia de un componente aditivo debido a los tratamientos. Si encontramos que este componente está presente, rechazamos la hipótesis nula de que los grupos proceden de la misma población y aceptamos la hipótesis alternativa de que al menos parte de las medias de grupo son diferentes entre sí, lo que indica que al menos algunos de los valores de α_i son diferentes en magnitud. A continuación, generalmente queremos comprobar cuáles de los valores de α_i son diferentes entre sí. Esto se hace por medio de pruebas de significación, con hipótesis alternativas tales como $H_1: \alpha_1 > \alpha_2$ o $H_1: (\alpha_1 + \alpha_2) > \alpha_3$. Es decir, estos comprueban si la media del grupo 1 es significativamente mayor que la media del grupo 2, o bien si la media del grupo 3 es menor que el promedio de las medias de los grupos 1 y 2.

A continuación siguen algunos ejemplos de análisis de la varianza modelo I en diversas disciplinas biológicas. Un experimento en el que ensayamos los efectos de diferentes drogas en lotes de animales conduce a un análisis de varianza modelo I. Estamos interesados en los resultados de los tratamientos y las diferencias entre ellos. Los tratamientos son fijos y determinados por el investigador. Esto se verifica también cuando examinamos los efectos de diferentes dosis de un determinado factor, una sustancia química, o la cantidad de luz a que ha sido expuesta una planta o las temperaturas a que se han mantenido botellas de cultivo de insectos. El tratamiento no tiene que ser completamente conocido y manipulado por el investigador; con tal de que sea fijo y repetible, se aplicará el modelo I. Si hubiésemos querido comparar los pesos de nacimiento de los niños chinos en el hospital de Malaya con los pesos de niños chinos nacidos en un hospital del continente chino, también habría sido un análisis de varianza modelo I. Los efectos del tratamiento serían en este caso "continente respecto a Malaya", que resumen toda una serie de factores diferentes, genéticos y ambientales, algunos conocidos por nosotros, pero la mayoría de ellos desconocidos. No obstante, éste es un tratamiento definido que podemos describir y además repetir; es decir, si lo deseamos, podemos muestrear de nuevo pesos de nacimiento de niños tanto en Malaya como en el continente chino.

Otro ejemplo de análisis de la varianza modelo I sería un estudio de pesos corporales para grupos de animales de diferente edad. Los tratamientos serían las edades, que son fijas. Si hallamos que hay una diferencia significativa en peso entre las edades, podríamos completar esto con la cuestión de si hay una diferencia de la edad 2 a la 3 o solamente de la edad 1 a la 2. En su mayor parte los análisis de varianza modelo I son el resultado de un experimento y de manipulación deliberada de factores por el investigador. No obstante, el estudio de diferencias tales como la comparación de pesos de nacimiento de dos países, si bien no es un experimento propiamente dicho también cae dentro de esta categoría.

7.7 Análisis de la varianza, modelo II

La estructura de la variación en un análisis de varianza modelo II es bastante similar a la del modelo I:

$$Y_{ij} = \mu + A_i + \epsilon_{ij} \tag{7.4}$$

donde $i = 1, \ldots, a, j = n$, ϵ_{ij} representa una variable independiente normalmente distribuida con media $\bar{\epsilon}_{ij} = 0$ y varianza $\sigma_\epsilon^2 = \sigma^2$, y A_i representa una variable normalmente distribuida, independiente de todos los valores de ϵ, con media $\bar{A}_i = 0$ y varianza σ_A^2. La diferencia esencial es que en lugar de efectos de tratamiento fijos α_i, ahora consideramos efectos aleatorios A_i, que difieren de un grupo a otro. Como los efectos son aleatorios, es inútil estimar la magnitud de dichos efectos aleatorios para un grupo cualquiera, o las diferencias de un grupo a otro, pero podemos estimar su varianza, el componente aditivo de la varianza entre grupos σ_A^2. Comprobamos su presencia y estimamos su magnitud s_A^2, así como su porcentaje de contribución a la variación en un análisis de la varianza modelo II.

Algunos ejemplos aclararán las aplicaciones del análisis de varianza modelo II. Supongamos que queremos determinar el contenido en *DNA* de células hepáticas de rata. Cogemos cinco ratas y hacemos tres preparaciones de cada uno de los cinco hígados obtenidos. La muestra de lecturas será de $a = 5$ grupos con $n = 3$ lecturas por grupo. Probablemente las cinco ratas se han extraído al azar de la colonia disponible al investigador. Deben ser diferentes en varios aspectos, genética y ambientalmente; pero no tenemos información precisa acerca de la naturaleza de estas diferencias. Por tanto, si averiguamos que la rata 2 tiene ligeramente más *DNA* en sus células hepáticas que la rata 3, poco podemos hacer con esta información porque es improbable que tengamos ninguna base para llevar hasta el fin este problema. Sin embargo, estaremos interesados en estimar la varianza de las tres réplicas dentro de un hígado cualquiera y la varianza entre las cinco ratas; es decir, ¿hay varianza σ_A^2 entre ratas además de la varianza σ^2 esperada basándose en las tres réplicas? La varianza entre las tres réplicas probablemente surja solo de diferencias en la técnica y posiblemente de diferencias en contenido de *DNA* en diferentes partes del hígado (improbable en un homogenado). La varianza aditiva entre ratas, si existiera, pudiera deberse a diferencias en ploidía o fenómenos relacionados. El grado relativo de variación entre ratas e "intrarratas" (= entre preparaciones) nos conduciría a planear más estudios de este tipo. Si hubiese poca varianza entre las preparaciones y relativamente más variación entre las ratas, necesitaríamos menos preparaciones y más ratas. Por otra parte, si la varianza entre ratas es proporcionalmente menor, utilizaríamos menos ratas y más preparaciones por rata.

En un estudio del grado de variación del pigmento de la piel en poblaciones humanas podríamos querer estudiar diferentes familias dentro de un grupo racial homogéneo y hermanos y hermanas dentro de cada familia. La varianza intrafamilias sería la media cuadrática del error y probaríamos un componente aditivo de varianza entre familias. Esperaríamos un componente aditivo de varianza σ_A^2 porque hay diferencias genéticas entre familias que determinan el grado de pigmentación de la piel. Estaríamos especialmente interesados en las proporciones relativas de las dos varianzas σ^2 y σ_A^2 porque nos proporcionarían importante información genética. Según nuestros conocimientos de teo-

ría genética esperaríamos que la varianza entre familias fuese mayor que la varianza entre hermanos y hermanas dentro de una familia.

Los ejemplos anteriores ilustran los dos tipos de problemas que comprende el análisis de varianza modelo II, que es más probable que surjan en trabajos biológicos. Uno se ocupa del problema general del diseño de un experimento y la magnitud del error experimental a diferentes niveles de replicación, tales como el error entre réplicas dentro de hígados de rata, error entre lotes, experimentos, y así sucesivamente. Los otros se refieren a la variación entre e intrafamilias, entre e intrahembras, entre e intrapoblaciones, y así sucesivamente, ocupándose del problema general de la relación entre variación genética y fenotípica.

Ejercicios 7

7.1 En un estudio que compara la composición química de la orina de chimpancés y gorilas (Gastler, Firschein, y Dobzhansky, 1956) se obtuvieron los siguientes resultados. Para 37 chimpancés la varianza de la cantidad de ácido glutámico en miligramos por miligramos de creatinina fue 0,01069. Un estudio similar basado en seis gorilas dio una varianza de 0,12442. ¿Hay una diferencia significativa entre la variabilidad en chimpancés y gorilas? SOLUCION. $F_s = 11{,}639$, $F_{0{,}025[5,36]} \approx 2{,}90$.

7.2 Los datos siguientes proceden de un experimento realizado por Sewal Wright. Cruzó conejos gigantes polacos y flamencos y obtuvo 27 conejos F_1. Se cruzaron estos y se obtuvieron 112 conejos F_2. Hemos obtenido los siguientes datos de longitud del fémur de estos conejos

	n	$\bar{Y}$	s
F_1	27	83,39	1,65
F_2	112	80,5	3,81

¿Hay un grado de variabilidad significativamente mayor en las longitudes del fémur entre los conejos de la F_2 que entre los conejos de la F_1? ¿Qué fenómeno genético bien conocido viene ilustrado por estos datos?

7.3 Demuestra que es posible representar el valor de una variante individual como sigue: $Y_{ij} = (\bar{\bar{Y}}) + (\bar{Y}_i - \bar{\bar{Y}}) + (Y_{ij} - \bar{Y}_i)$. ¿Qué estima cada uno de los términos entre paréntesis en un análisis de la varianza modelo I y en un modelo II?

7.4 Para los datos de la tabla 7.3, háganse tablas que representen la descomposición del valor de cada variante en sus tres componentes, $\bar{\bar{Y}}$, $(\bar{Y}_i - \bar{\bar{Y}})$, $(Y_{ij} - \bar{Y}_i)$. La primera tabla constaría pues de 35 valores, todos iguales a la media total. En la segunda tabla todas las entradas de una determinada columna serían iguales a la diferencia entre la media de esa columna y la media total. Y la última tabla constará de las desviaciones de cada variante individual respecto de su media de columna. Estas tablas representan estimaciones de uno de los componentes individuales de la expresión (7.3). Calcular la media y sumas de cuadrados para cada tabla.

Capítulo 8

Análisis de la varianza de clasificación simple

Ya estamos preparados para estudiar casos reales de análisis de varianza en diversas aplicaciones y modelos. El presente capítulo trata del tipo más sencillo de análisis de varianza, el *análisis de varianza de clasificación simple*. Esto significa que los grupos de muestras se clasifican por un solo criterio. Las dos interpretaciones de las siete muestras de longitudes del ala de moscas domésticas estudiadas en el capítulo anterior, diferentes formulaciones del medio (modelo I), y progenies de diferentes hembras (modelo II), representarían un criterio único de clasificación. Otros ejemplos serían diferentes temperaturas a las que se han criado grupos de animales o diferentes suelos en los que se han cultivado muestras de plantas.

En la sección 8.1 comenzaremos por establecer las fórmulas básicas de cálculo para el análisis de la varianza, basadas en los tópicos incluidos en el capítulo anterior. La sección 8.2 da un ejemplo del caso ordinario con tamaños de muestra iguales. Ilustraremos este caso por medio de un análisis de la varianza modelo I. Puesto que los cálculos básicos para el análisis de la varianza, son los mismos en los dos modelos, no es necesario repetir la ilustración con un modelo II. Este último modelo se destaca en la sección 8.3, que expone las complicaciones de cálculo secundarias que resultan de tamaños de muestreo diferentes, ya que todos los grupos en el análisis de la varianza no han de tener necesariamente el mismo tamaño de muestreo. Se exponen además algunos cálculos especiales para un modelo II, la estimación de componentes de la varianza. Las fórmulas resultan particularmente sencillas para el caso de dos muestras (sección 8.4). En el modelo I de este caso puede aplicarse también la prueba t matemáticamente equivalente.

Cuando un análisis de la varianza modelo I se ha encontrado que es significativo, llevando a la conclusión de que las medias no son de la misma población, es deseable contrastar las medias de diversas maneras para descubrir qué pares de medias son diferentes entre sí, y las medias pueden separarse en grupos que sean significativamente diferentes entre sí. Las llamadas pruebas de comparaciones múltiples se incluyen en las

secciones 8.5 y 8.6. La primera trata de las llamadas comparaciones planificadas, diseñadas antes de hacer la prueba; la segunda sección de pruebas a posteriori que se proponen al investigador como resultado de su análisis.

8.1 Fórmulas para el cálculo

En la sección 7.5 vimos que la suma de cuadrados y los grados de libertad totales pueden descomponerse aditivamente en los que pertenecen a variación entre grupos y los que pertenecen a variación intragrupos. Lo más sencillo es calcular la suma de cuadrados entre grupos, dejando la suma de cuadrados intragrupos para obtenerla por la sustracción $S.C._{\text{total}} - S.C._{\text{grupos}}$. Esta regla no se aplica a los computadores digitales; en ellos la molestia de calcular sumas de cuadrados intragrupos no es de importancia, pero la exactitud sí. En la sección 7.5 llegábamos a las siguientes fórmulas de cálculo para estas sumas de cuadrados:

$$S.C._{\text{total}} = \sum^{a}\sum^{n} Y^2 - \frac{1}{an}\left(\sum^{a}\sum^{n} Y\right)^2$$

$$S.C._{\text{grupos}} = \frac{1}{n}\sum^{a}\left(\sum^{n} Y\right)^2 - \frac{1}{an}\left(\sum^{a}\sum^{n} Y\right)^2$$

Estas fórmulas suponen el mismo tamaño de muestreo n para cada grupo y en la sección 8.3 se modificarán para tamaños de muestras diferentes. No obstante, en su forma actual bastan para aclarar algunos puntos generales acerca de los procedimientos de cálculo del análisis de la varianza.

En primer lugar observamos que el segundo término restado en cada suma de cuadrados es idéntico. Este término representa la suma de todas las variantes en el análisis de la varianza (la suma total), elevada al cuadrado y dividida por el número total de variantes. Es comparable al segundo término de la suma de cuadrados ordinaria [expresión (3.7)]. Este término se denomina a veces *término de corrección* (abreviado *T.C.*).

El primer término para la suma de cuadrados total es simple. Es la suma de todos los cuadrados de las variantes de la tabla de análisis de varianza. Así, la suma de cuadrados total, que describe la variación de una sola muestra no estructurada de *an* ítems, es simplemente la fórmula familiar de la suma de cuadrados de la expresión (3.7).

El primer término de la suma de cuadrados entre grupos se obtiene elevando al cuadrado la suma de los ítems de cada grupo, dividiendo cada cuadrado por su tamaño de muestreo y sumando los cocientes de esta operación para cada grupo. Como el tamaño de muestreo de cada grupo es el mismo en las fórmulas anteriores, podemos sumar primero todos los cuadrados de las sumas de grupo y luego dividir su suma por la constante n.

De la fórmula para la suma de cuadrados entre grupos surge una importante regla de cálculo para el análisis de la varianza. *Para hallar la suma de cuadrados entre cualquier serie de grupos, elevar al cuadrado la suma de cada grupo y dividirla por su tamaño de muestreo; sumar los cocientes de estas operaciones y restar de esta suma un término de corrección. Para hallar este término de corrección, sumar todos los ítems de la serie, elevar al cuadrado esta suma, y dividirla por el número de ítems en que se basa.*

8.2 Igual *n*

Ilustraremos un análisis de la varianza de clasificación simple, con tamaños de muestras iguales, por medio de un ejemplo de modelo I. El cálculo hasta la primera prueba de significación incluyendo ésta, es idéntico para los dos modelos. Así, el cálculo del cuadro 8.1 podría servir también para un análisis de la varianza modelo II con igual tamaño de muestreo.

Los datos son de un experimento de fisiología vegetal. Registran la longitud, en unidades codificadas, de secciones de guisante desarrolladas en cultivo de tejidos con auxina presente. El fin del experimento era contrastar los efectos de la adición de diversos azúcares en el crecimiento, medidos por la longitud. Se utilizaron cuatro grupos experimentales, representando tres azúcares diferentes y una mezcla de azúcares, más un control sin azúcar. Para cada tratamiento se hicieron diez observaciones (repeticiones). El término "tratamiento" ya implica un análisis de varianza modelo I. Está claro que los cinco grupos no representan muestras al azar de todas las condiciones experimentales posibles, sino que han sido deliberadamente planeados para probar los efectos de ciertos azúcares en la velocidad de crecimiento. Estamos interesados en el efecto de los azúcares en la longitud, y nuestra hipótesis nula será que no hay componente aditivo debido a los efectos del tratamiento entre los cinco grupos; es decir, se supone que las medias de población son todas iguales.

El cálculo se representa en el cuadro 8.1. Después de haberse calculado las cantidades de la 1 hasta la 7 se introducen en una tabla de análisis de la varianza, como se muestra en el cuadro. Primero se presentan las fórmulas generales para esta tabla, seguidas por una tabla ocupada por el ejemplo específico. Habiendo cinco tratamientos observamos 4 grados de libertad entre grupos y 45 *g.l.* intragrupos, que representan 5 veces (10-1) grados de libertad. Encontramos que la media cuadrática entre grupos es considerablemente mayor que la media cuadrática del error, provocando la sospecha de que esté presente un componente aditivo debido a los efectos del tratamiento. Si la $M.C._{\text{grupos}}$ es igual o menor que la $M.C._{\text{intra}}$, no nos molestamos en continuar con el análisis, puesto que no tendríamos pruebas de la presencia de un componente aditivo de la varianza. Nos podemos preguntar cómo sería posible que la $M.C._{\text{grupos}}$ fuese menor que la $M.C._{\text{intra}}$. Debe recordarse que los dos son estimadores independientes. Si no hay componente aditivo de la varianza entre grupos, es tan probable que el estimador de la varianza entre grupos sea menor como que sea mayor que la varianza intragrupos.

Las expresiones para los valores esperados de las medias cuadráticas se exponen también en la primera tabla de análisis de la varianza del cuadro 8.1. Son las expresiones que se han aprendido en el capítulo anterior para un modelo I.

Puede parecer que llevamos un número innecesario de cifras en los cálculos del cuadro 8.1. Esto es necesario a veces para asegurar que la suma de cuadrados del error, cantidad 7, tiene suficiente exactitud.

Como v_2 es relativamente grande, los valores críticos de F se han calculado por interpolación armónica en la tabla V (véase nota al pie de la tabla III en interpolación armónica).

CUADRO 8.1

Análisis de la varianza de clasificación simple, con tamaños de muestra iguales.

El efecto de la adición de diferentes azúcares en la longitud, en unidades del micrómetro ocular (× 0,114 = mm), de secciones de guisantes crecidas en cultivo de tejidos y en presencia de auxina: $n = 10$ (réplicas por grupo). Este es un análisis de la varianza modelo I.

	Tratamientos ($a = 5$)				
Observaciones, por ejemplo, réplicas	*Control*	*Adición de glucosa al 2 %*	*Adición de fructosa al 2 %*	*Adición de glucosa al 1 % + fructosa al 1 %*	*Adición de sacarosa al 2 %*
1	75	57	58	58	62
2	67	58	61	59	66
3	70	60	56	58	65
4	75	59	58	61	63
5	65	62	57	57	64
6	71	60	56	56	62
7	67	60	61	58	65
8	67	57	60	57	65
9	76	59	57	57	62
10	68	61	58	59	67
$\sum Y$	701	593	582	580	641
$\bar{Y}$	70.1	59.3	58.2	58.0	64.1

Fuente: Datos de W. Purves.

Cálculos preliminares

1. Suma total $= \sum^{a}\sum^{n} Y = 701 + 593 + \cdots + 641 = 3097$

2. Suma de los cuadrados de las observaciones $= \sum^{a}\sum^{n} Y^2$
$= 75^2 + 67^2 + \cdots + 68^2 + 57^2 + \cdots + 67^2 = 193\,151$

3. Suma de los cuadrados de las sumas de grupo dividida por $n = \frac{1}{n}\sum^{a}\left(\sum^{n} Y\right)^2$
$= \frac{1}{10}[701^2 + 593^2 + \cdots + 641^2] = \frac{1}{10}[1\,929\,055] = 192\,905{,}50$

4. Suma total elevada al cuadrado y dividida por el tamaño de muestreo total = término de corrección
$$TC = \frac{1}{an}\left(\sum^{a}\sum^{n} Y\right)^2 = \frac{(3097)^2}{5 \times 10} = \frac{9\,591\,409}{50} = 191\,828{,}18$$

5. $S.C._{\text{total}} = \sum^{a}\sum^{n} Y^2 - CT$
= cantidad 2 − cantidad 4 = 193 151 − 191 828,18 = 1322,82

CUADRO 8.1 (continuación)

6. $S.C._{\text{grupos}} = \frac{1}{n}\sum^{a}\left(\sum^{n} Y\right)^2 - CT$
= cantidad 3 − cantidad 4 = 192 905,50 − 191 828,18 = 1 077,32

7. $S.C._{\text{intragrupos}} = S.C._{\text{total}} - S.C._{\text{grupos}} =$
= cantidad 5 − cantidad 6 = 1 322,82 − 1 077,32 = 245,50

La tabla de análisis de la varianza se construye como sigue.

Origen de la variación	*g.l.*	*S.C.*	*M.C.*	F_s	*M.C. esperada*
$\bar{Y} - \bar{\bar{Y}}$ Entre grupos	$a - 1$	6	$\frac{6}{(a-1)}$	$\frac{M.C._{\text{grupos}}}{M.C._{\text{intragrupos}}}$	$\sigma^2 + \frac{n}{a-1}\sum^{a}\alpha^2$
$Y - \bar{Y}$ Intragrupos	$a(n-1)$	7	$\frac{7}{a(n-1)}$		σ^2
$Y - \bar{\bar{Y}}$ Total	$an - 1$	5			

Sustituyendo los valores calculados en la tabla anterior obtendremos la siguiente.

Tabla de análisis de la varianza

Origen de la variación	*g.l.*	*S.C.*	*M.C.*	*F.*
$\bar{Y} - \bar{\bar{Y}}$ entre grupos (entre tratamientos)	4	1077,32	269,33	49,33
$Y - \bar{Y}$ intragrupos (error, réplicas)	45	245,50	5,46	
$Y - \bar{\bar{Y}}$ Total	49	1322,82		
$F_{0,05[4,45]} = 2{,}58$		$F_{0,01[4,45]} = 3{,}77$	$F_{0,001[4,45]} = 5{,}57$	

Conclusiones. – Hay un componente aditivo altamente significativo ($P \ll 0{,}01$), debido a los efectos del tratamiento en la media cuadrática entre grupos (tratamientos). Los diferentes tratamientos de azúcares tienen sin duda un efecto significativo en el crecimiento de las secciones de guisante.

Véase secciones 8.5 y 8.6 para la realización de un análisis de la varianza modelo I: es decir, el método para determinar qué medias son significativamente diferentes entre sí.

Los valores críticos se han dado aquí solamente para presentar un registro completo del análisis. Ordinariamente, al enfrentarnos con este ejemplo, no nos preocuparía el cálculo de estos valores de F. La comparación de la razón de varianzas observada $F_s = 49{,}33$ con $F_{0{,}01[4,40]} = 3{,}83$, el valor crítico conservativo (la F tabulada inmediata con menos grados de libertad) nos convencería para rechazar la hipótesis nula. La probabilidad de que los cinco grupos difieran tanto como lo hacen, por azar es casi infinitesimalmente pequeña. Sin duda los azúcares producen un efecto aditivo de tratamiento, inhibiendo aparentemente el crecimiento y reduciendo en consecuencia la longitud de las secciones de guisantes.

En esta etapa no estamos en condiciones de decir si cada tratamiento es diferente de otro, o si los azúcares son diferentes del control, pero no diferentes entre sí. Estas pruebas son necesarias para completar un análisis modelo I, pero aplazaremos su discusión hasta las secciones 8.5 y 8.6.

8.3 Diferente *n*

Esta vez utilizaremos como ejemplo un análisis de varianza modelo II. Recuérdese que, hasta la prueba de significación F incluida, los cálculos son exactamente los mismos tanto si el análisis de la varianza es modelo I como si es modelo II. Indicaremos la etapa del cálculo en la que habría una divergencia de operaciones dependiendo del modelo.

En la tabla 8.1 se presenta el ejemplo. Se refiere a una serie de medidas morfológicas de la anchura del escudo (placa dorsal) de muestras de larvas de garrapata, obtenidas de cuatro diferentes individuos huéspedes del conejo norteamericano de cola de algodón. Estos cuatro huéspedes se obtuvieron al azar de una localidad. No sabemos nada acerca de sus orígenes ni su constitución genética. Representan una muestra al azar de la población de individuos huéspedes de la localidad dada. No estaríamos en condiciones de interpretar diferencias entre larvas de ningún huésped, ya que no sabemos nada de los orígenes de los conejos. No obstante, los biólogos de poblaciones se interesan por estos análisis porque ofrecen una respuesta a las siguientes cuestiones. ¿Son las varianzas de las medias de los caracteres larvarios entre huéspedes mayores que las esperadas, basándose en las varianzas de los caracteres intrahuéspedes? Podemos calcular la varianza media de la anchura del escudo larval en un huésped. Este será nuestro término "error" en el análisis de la varianza. Después contrastamos la media cuadrática entre grupos observada y vemos si contiene un componente aditivo de la varianza. ¿Qué representaría tal componente aditivo de la varianza? La media cuadrática intrahuéspedes (es decir, de las larvas en un huésped cualquiera) representa diferencias genéticas entre las larvas y diferencias en las experiencias ambientales de estas larvas. La varianza aditiva entre huéspedes demuestra diferenciación significativa entre las larvas, debida posiblemente a diferencias entre los huéspedes que afectan a la larva. También puede deberse a diferencias genéticas entre las larvas, si cada huésped llevase una familia de garrapatas, o al menos una población cuyos individuos están más relacionados entre sí de lo que lo están con las larvas de garrapata de otros huéspedes. En este ejemplo el interés está en las magnitudes de las varianzas. En vista de la elección al azar de los huéspedes éste es un caso claro de análisis de la varianza modelo II.

TABLA 8.1

Datos y tabla para un análisis de la varianza de clasificación simple, con tamaños de muestra diferentes. Anchura del escudo (placa dorsal) de larvas de la garrapata *Haemaphysalis leporispalustris* en muestras de 4 conejos americanos de cola de algodón. Medidas en micrómetros. Este es un análisis de la varianza modelo II.

Huéspedes ($a = 4$)			
1	*2*	*3*	*4*
380	350	354	376
376	356	360	344
360	358	362	342
368	376	352	372
372	338	366	374
366	342	372	360
374	366	362	
382	350	344	
	344	342	
	364	358	
		351	
		348	
		348	
$\sum_{i=1}^{n_i} Y$ 2978	3544	4619	2168
n_i 8	10	13	6
$\sum_{i=1}^{n_i} Y^2$ 1 108 940	1 257 272	1 642 121	784 536

Fuente: Datos de P. A. Thomas.

Tabla de análisis de la varianza

Objeto de la variación	*g.l.*	*S.C.*	*M.C.*	F_s
$\bar{Y} - \bar{\bar{Y}}$ Entre grupos (entre huéspedes)	3	1808	602,7	5,26**
$Y - \bar{Y}$ Intragrupos (error; entre larvas en un huésped)	33	3778	114,5	
$Y - \bar{\bar{Y}}$ Total	36	5586		

$F_{.05[3,33]} = 2.89$ $F_{.01[3,33]} = 4,44$

** = $P < 0.01$

Conclusión. – Hay un componente aditivo de la varianza significativo ($P < 0{,}01$) entre huéspedes para la anchura del escudo en larvas de garrapata.

El cálculo sigue el esquema provisto en el cuadro 8.1, excepto que el símbolo $\sum^{n}$ tiene que escribirse ahora $\sum^{n_i}$ puesto que los tamaños de muestreo difieren para cada grupo. Los pasos 1, 2 y del 4 hasta el 7 se realizan como antes. Solamente el paso 3 tiene que modificarse apreciablemente. Este es

3. Suma de los cuadrados de las sumas de grupo, cada una dividida por su tamaño de muestreo

$$= \sum^{a} \frac{\left(\sum^{n_i} Y\right)^2}{n_i} = \frac{(2978)^2}{8} + \frac{(3544)^2}{10} + \cdots + \frac{(2168)^2}{6} = 4\,789\,091$$

Los valores críticos de F al 5 % y 1 % se muestran debajo de la tabla de análisis de la varianza en la tabla 8.1 (2,89 y 4,44, respectivamente). Debería confirmarse por uno mismo en la tabla V. Obsérvese que no se da el argumento $\nu_2 = 33$. Por lo tanto, se debe interpolar entre los argumentos que representan 30 y 40 grados de libertad, respectivamente. Los valores presentados se han calculado utilizando interpolación armónica. Sin embargo, una vez más, no ha sido necesario efectuar tal interpolación. El valor conservativo de F, $F_{\alpha[3,30]}$, es 2,92 y 4,51 para $\alpha = 0{,}05$ y $\alpha = 0{,}01$, respectivamente. El valor observado de F_s es 5,26, considerablemente superior al valor conservativo de $F_{0,01}$. Por consiguiente rechazamos la hipótesis nula ($H_0: \sigma_A^2 = 0$) de que no hay componente aditivo de la varianza entre grupos y que las dos medias cuadráticas estimen la misma varianza. Aceptamos en cambio la hipótesis alternativa de la existencia de un componente aditivo de varianza σ_A^2.

¿Cuál es el significado biológico de esta conclusión? Por una u otra razón las garrapatas de diferentes huéspedes difieren más entre sí que las garrapatas de un huésped cualquiera. Esto puede deberse a cierta influencia modificadora de los huéspedes sobre las garrapatas (diferencias biológicas en la sangre, diferencias en la piel, diferencias en el medio ambiente del huésped; todas ellas bastante improbables en este caso) o puede deberse a diferencias genéticas entre las garrapatas. Posiblemente las de cada huésped representen los descendientes de una sola pareja de padres y las diferencias de las garrapatas entre huéspedes representen diferencias genéticas entre familias; o la selección haya actuado diferentemente en las poblaciones de garrapatas en cada huésped, o los huéspedes hayan emigrado al lugar de reunión de diferentes áreas geográficas en las que las garrapatas difieren en la anchura del escudo. De estas diversas posibilidades, a la vista de la biología del organismo, las diferencias genéticas entre familias parece la más razonable.

Hasta este punto los cálculos habrían sido idénticos en un análisis de varianza modelo I. Si éste hubiese sido modelo I, la conclusión habría sido que hay un efecto de tratamiento significativo en vez de un componente aditivo de la varianza. Sin embargo, ahora debemos completar los cálculos propios de un análisis de varianza modelo II. Estos incluirán la estimación del componente aditivo de varianza y el cálculo del porcentaje de variación a los dos niveles.

Como el tamaño de muestreo n_i difiere entre grupos en este ejemplo, no podemos escribir $\sigma^2 + n\sigma_A^2$ para la $M.C._{\text{grupos}}$ esperada. Es obvio que ningún valor individual de n

sería apropiado en la fórmula. Por lo tanto utilizamos un n medio; no obstante, éste no es simplemente $\bar{n}$, la media aritmética de los valores de n, sino que es

$$n_0 = \frac{1}{a-1}\left(\sum^{a} n_i - \frac{\sum^{a} n_i^2}{\sum^{a} n_i}\right) \tag{8.1}$$

que es un promedio ordinariamente próximo a $\bar{n}$ pero siempre menor, a menos que los tamaños de muestreo sean iguales, cuando $n_0 = \bar{n}$. En este ejemplo

$$n_0 = \frac{1}{4-1}\left([8 + 10 + 13 + 6] - \frac{8^2 + 10^2 + 13^2 + 6^2}{8 + 10 + 13 + 6}\right) = 9{,}009$$

En las expresiones para las medias cuadráticas esperadas presentadas en la tabla de análisis de la varianza, está claro cómo se obtienen el componente de varianza entre grupos σ_A^2 y la varianza del error σ^2. Naturalmente los valores que obtenemos son simplemente estimaciones y por lo tanto se escriben como s_A^2 y s^2. El componente aditivo de varianza s_A^2 se estima como $(M.C._{\text{grupos}} - M.C._{\text{intra}})/n$. Siempre que los tamaños de muestreo sean diferentes, el denominador se convierte en n_0. En este ejemplo $(602{,}7 - 114{,}5)/9{,}009 = 54{,}190$. Frecuentemente no nos interesan mucho los valores absolutos de estos componentes de la varianza, sino sus magnitudes relativas. Con este fin los sumamos y expresamos cada uno como un porcentaje de esta suma. Así $s^2 + s_A^2 = 114{,}5 + 54{,}190 = 168{,}690$ y s^2 y s_A^2 son 67,9 % y 32,1 % de esta suma, respectivamente. Se encuentra relativamente más variación intragrupos (larvas en un huésped) que entre grupos (huéspedes).

8.4 Dos grupos

Una prueba frecuente en estadística es establecer *la significación de la diferencia entre dos medias.* Esto puede hacerse fácilmente por medio de un *análisis de la varianza para dos grupos.* El cuadro 8.2 muestra este procedimiento para un análisis de la varianza modelo I. el caso ordinario.

El ejemplo del cuadro 8.2 se refiere al inicio de la madurez reproductiva en la pulga de agua, *Daphnia longispina.* Este se mide como la edad media (en días) al comienzo de la reproducción. Cada variante de la tabla es en realidad un promedio y un posible fallo en el análisis podría ser que estos promedios no estuviesen basados en tamaños de muestra iguales. Sin embargo, no se nos ha dado esta información y tenemos que proceder en el supuesto de que cada lectura de la tabla sea una variante igualmente fiable. Las dos series representan diferentes cruzamientos genéticos y las siete réplicas de cada serie son clonos derivados del mismo cruzamiento genético. Este ejemplo es claramente un modelo I, ya que la cuestión a responder es si la serie I difiere de la serie II en la edad media al comienzo de la reproducción. El examen de los datos revela que la edad media al inicio de la reproducción es muy similar para las dos series. Nos sorprendería por tanto encontrar que son significativamente diferentes. No obstante, de todos modos realizaremos una

CUADRO 8.2

Prueba de la diferencia de medias entre dos grupos.

Edad media (en días) al comienzo de la reproducción en *Daphnia longispina* (cada variante es una media basada en números de hembras aproximadamente iguales). Se comparan dos series derivadas de cruzamientos genéticos diferentes y conteniendo siete clones cada una; $n = 7$ clones por serie. Este es un análisis de la varianza modelo I.

	Series (a = 2)	
	I	*II*
	7,2	8,8
	7,1	7,5
	9,1	7,7
	7,2	7,6
	7,3	7,4
	7,2	6,7
	7,5	7,2
$\sum^{n} Y$	52,6	52,9
$\bar{Y}$	7,5143	7,5571
$\sum^{n} Y^2$	398,28	402,23

Fuente: Datos de Ordway desde Banta (1939).

Análisis de la varianza de clasificación simple, para dos grupos con iguales tamaños de muestra.

Tabla de análisis de la varianza

Origen de la variación	*g.l.*	*S.C.*	*M.C.*	F_s
$\bar{Y} - \bar{\bar{Y}}$ Entre grupos (series)	1	0,00643	0,00643	0,0141
$Y - \bar{Y}$ Intragrupos (error; clones dentro de una serie)	12	5,48571	0,45714	
$Y - \bar{\bar{Y}}$ Total	13	5,49214		

$F_{0.05[1,12]} = 4,75$

Conclusión. – Como $F_s \ll F_{0.05[1,12]}$, se acepta la hipótesis nula. Las medias de las dos series no son significativamente diferentes; es decir, las dos series no difieren en la edad media al comienzo de la reproducción.

Prueba t de la hipótesis de que dos medias de muestreo proceden de una población con igual μ: límites de confianza de la diferencia entre dos medias.

Esta prueba supone que las varianzas de las poblaciones de las que se han extraído las dos muestras son idénticas. En caso de duda acerca de esta hipótesis, comprobar por el método del cuadro 7.1, sección 7.3.

CUADRO 8.2 (continuación)

La fórmula apropiada para t_s es una de las siguientes:

Expresión (8.2), cuando los tamaños de muestra son diferentes y n_1 ó n_2 ó ambos son pequeños (< 30) : $g.l. = n_1 + n_2 - 2$.

Expresión (8.3), cuando los tamaños de muestra son idénticos (independientemente del tamaño): $g.l. = 2(n - 1)$.

Expresión (8.4), cuando ambos n_1 y n_2 son diferentes pero grandes (> 30): $g.l. = n_1 + n_2 - 2$.

Para los datos presentes, puesto que los tamaños de muestra son iguales, elegimos la expresión (8.3):

$$t_s = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n}(s_1^2 + s_2^2)}}$$

No hemos calculado las varianzas, pero podemos transformar la fórmula en una basada en sumas de cuadrados, que se obtienen en un paso del cálculo anterior a las cantidades provistas.

$$t_s = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sum^{n} y_1^2 + \sum^{n} y_2^2}{n(n-1)}}}$$

Contrastamos la hipótesis nula de que $\mu_1 - \mu_2 = 0$. Por lo tanto sustituimos esta cantidad por cero en este ejemplo y obtenemos $y_1^2 = 3,029$ y $y_2^2 = 2,457$. Luego

$$t_s = \frac{7,5143 - 7,5571}{\sqrt{(3,029 + 2,457)/7 \times 6}} = \frac{-0,0428}{\sqrt{5,486/42}} = \frac{-0,0428}{0,3614} = -0,1184$$

Los grados de libertad para este ejemplo son $2(n - 1) = 2 \times 6 = 12$. El valor crítico de $t_{0,05[12]} = 2,179$. Como el valor absoluto de nuestro t_s observado es menor que el valor crítico t, las medias no son significativamente diferentes, que es el mismo resultado que se ha obtenido por el análisis de la varianza.

Límites de confianza de la diferencia entre dos medias

$$L_1 = (\bar{Y}_1 - \bar{Y}_2) - t_{\alpha[\nu]} s_{\bar{Y}_1 - \bar{Y}_2}$$

$$L_2 = (\bar{Y}_1 - \bar{Y}_2) + t_{\alpha[\nu]} s_{\bar{Y}_1 - \bar{Y}_2}$$

En este caso $\bar{Y}_1 - \bar{Y}_2 = -0,0428$, $t_{0,05[12]} = 2,18$ y $s_{\bar{Y}_1 - \bar{Y}_2} = 0,3614$ tal como se ha calculado anteriormente como el denominador de la prueba t. Por consiguiente

$$L_1 = -0,0428 - (2,18)(0,3614) = -0,8307$$

$$L_2 = -0,0428 + (2,18)(0,3614) = 0,7451$$

Los límites de confianza del 95 % contienen el punto cero (ninguna diferencia) como era de esperar, ya que la diferencia $\bar{Y}_1 - \bar{Y}_2$ se ha encontrado que no era significativa.

prueba. Como puede verse por ahora, no se puede decir si es significativa la magnitud absoluta de una diferencia. Esto depende de la magnitud de la media cuadrática del error, que representa la varianza intraseries.

Los cálculos para el análisis de la varianza no se exponen. Serían los mismos que en el cuadro 8.1. Con iguales tamaños de muestra y solamente dos grupos hay otro procedimiento de cálculo más directo que el ordinario. La cantidad 6, $S.C._{\text{grupos}}$, puede calcularse directamente por la siguiente fórmula:

$$S.C._{\text{grupos}} = \frac{\left(\sum^{n} Y_1 - \sum^{n} Y_2\right)^2}{2n} = \frac{(52{,}6 - 52{,}9)^2}{14} = 0{,}00643$$

Hay solamente 1 grado de libertad entre los dos grupos. El valor crítico $F_{0{,}05[1,12]}$ se da debajo de la tabla de análisis de la varianza, pero no es realmente necesario consultarlo. El examen de las medias cuadráticas en el análisis de la varianza muestra que la $M.C._{\text{grupos}}$ es mucho menor que la $M.C._{\text{intra}}$; por lo tanto el valor de F_s está muy por debajo de la unidad y de ningún modo puede existir un componente aditivo debido a los efectos del tratamiento entre las series. En los casos en que $M.C._{\text{grupos}} \leqslant M.C._{\text{intra}}$, usualmente no nos molestamos en calcular F_s porque el análisis de varianza no podría ser de manera alguna significativo.

Hay otro método para resolver un análisis de la varianza modelo I para dos muestras. Este es una prueba *t de las diferencias entre dos medias*. Esta prueba t es el método tradicional para resolver dicho problema y ya puede resultar familiar como una relación previa con el trabajo estadístico. No tiene ventaja positiva en cuanto a facilidad de cálculo ni comprensión, y como se verá más adelante es matemáticamente equivalente al análisis de la varianza del cuadro 8.2. Se presenta aquí principalmente con el fin de completar. Parecería demasiada ruptura con la tradición no presentar la "la prueba t" en un texto de bioestadística.

En la sección 6.4 nos hemos informado sobre la distribución t y hemos visto que una distribución t de $n - 1$ *g.l.* podría obtenerse a partir de una distribución del término $(\bar{Y}_i - \mu)/s_{\bar{Y}_i}$, en que $s_{\bar{Y}_i}$ tienen $n - 1$ grados de libertad e $\bar{Y}$ está normalmente distribuida. El numerador de este término representa una desviación de una media de muestreo respecto de una media paramétrica y el denominador un error típico para tal desviación. Ahora vemos que la expresión

$$t_s = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right]\left(\frac{n_1 + n_2}{n_1 n_2}\right)}} \tag{8.2}$$

también se distribuye como t. La expresión (8.2) parece complicada, pero realmente tiene la misma estructura que el término más simple para t. El numerador es una desviación, esta vez no entre una sola media de muestreo y la media paramétrica, sino entre una simple diferencia de dos medias de muestreo, $\bar{Y}_1$ e $\bar{Y}_2$, y la verdadera diferencia entre las medias de las poblaciones representadas por estas medias. En una prueba de este tipo nuestra hipótesis nula es que las dos muestras vienen de la misma población; es decir, deben tener la misma media paramétrica. Así la diferencia $\mu_1 - \mu_2$ se espera que sea cero.

Por lo tanto, contrastamos la desviación de la diferencia $\bar{Y}_1 - \bar{Y}_2$ respecto de cero. El denominador de la expresión (8.2) es un error típico, el error típico de la diferencia entre dos medias $s_{\bar{Y}_1 - \bar{Y}_2}$. La parte izquierda de la expresión, que está entre corchetes, es una media ponderada de las varianzas de las dos muestras, s_1^2 y s_2^2, calculada a la manera de la sección 7.1. El término de la derecha del error típico es la fórmula de cálculo más fácil de $(1/n_1) + (1/n_2)$, que es el factor por el que debe multiplicarse la varianza media intragrupos para convertirla en una varianza de la diferencia de medias. La analogía con la multiplicación de una varianza de muestreo s^2 por $1/n$ para transformarla en una varianza de una media $s_{\bar{Y}}^2$ debería ser obvia.

La prueba esbozada aquí supone iguales varianzas en las dos poblaciones muestreadas. Este es también un supuesto de los análisis de la varianza realizados hasta ahora, aunque no hemos hecho hincapié en esto. Para dos varianzas solamente, la igualdad puede probarse por el procedimiento del cuadro 7.1.

Cuando los tamaños de muestra son iguales en una prueba de dos muestras, la expresión (8.2) se simplifica hasta

$$t_s = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n}(s_1^2 + s_2^2)}} \tag{8.3}$$

que es la fórmula aplicada en el ejemplo del cuadro 8.2. Cuando los tamaños de muestra son diferentes pero muy grandes, de modo que las diferencias entre n_i y $n_i - 1$ son relativamente insignificantes, la expresión (8.2) se reduce a la más sencilla

$$t_s = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_2} + \frac{s_2^2}{n_1}}} \tag{8.4}$$

En el apéndice A1.4 se expone la simplificación de la expresión (8.2) hasta las expresiones (8.3) y (8.4). Los grados de libertad correspondientes a las expresiones (8.2) y (8.4) son $n_1 + n_2 - 2$; y para la expresión (8.3) *g.l.* es $2(n - 1)$.

En el cuadro 8.2 se muestra la prueba de significación para las diferencias entre medias, utilizando la prueba t. Esta prueba es de dos colas porque nuestra hipótesis alternativa es $H_1 : \mu_1 \neq \mu_2$. Los resultados de esta prueba son idénticos a los del análisis de la varianza del mismo cuadro: las dos medias no son significativamente diferentes. Hemos mencionado anteriormente que estos dos resultados son en realidad matemáticamente idénticos. Podemos demostrar esto del modo más sencillo posible elevando al cuadrado el valor obtenido para t_s, que debería ser idéntico al valor F_s del análisis de la varianza correspondiente. Puesto que $t_s = -0{,}1184$ en el cuadro 8.2, $t_s^2 = 0{,}0140$. Dentro del error de redondeo éste es igual al F_s obtenido en el análisis de la varianza ($F_s = 0{,}0141$). ¿Por qué sucede así? Expondremos dos razones de esto. Hemos visto que $t_{[\nu]} = (\bar{Y} - \mu)/s_{\bar{Y}}$, donde ν es el número de grados de libertad de la varianza de la media $s_{\bar{Y}}^2$; por tanto $t_{[\nu]}^2 = (\bar{Y} - \mu)^2/s_{\bar{Y}}^2$. Sin embargo, esta expresión puede considerarse como una razón de varianzas. El denominador es claramente una varianza con ν grados de libertad. El numerador es

también una varianza. Es una desviación simple al cuadrado, que representa una suma de cuadrados que posee 1 en lugar de cero grados de libertad (ya que es una desviación de la verdadera media μ y no de una media de muestreo). Una suma de cuadrados basada en 1 grado de libertad es a la vez una varianza. Así t^2 es una razón de varianzas; específicamente, $t^2_{[\nu]} = F_{[1,\nu]}$, como hemos visto más arriba. En el apéndice A1.5 demostramos algebraicamente que los valores de t_s^2 y F_s obtenidos en el cuadro 8.2 son cantidades idénticas.

También podemos demostrar la relación entre t y F en las tablas estadísticas. En la tabla V encontramos que $F_{.05[1,10]} = 4{,}96$. Este valor se supone que es igual a $t^2_{.05[10]}$. La raíz cuadrada de 4,96 es 2,227 y cuando buscamos $t_{.05[10]}$ en la tabla III encontramos que es 2,228, que está dentro del error de redondeo aceptable del valor anterior. Como t se aproxima a la distribución normal al tender a infinito sus grados de libertad, $F_{\alpha[1,\nu]}$ se acerca a la distribución del cuadrado de la desviante normal cuando $\nu \to \infty$.

La relación entre t^2 y F conduce a otra relación. Acabamos de averiguar que cuando $\nu_1 = 1$, $F_{[\nu_1,\nu_2]} = t^2_{[\nu_2]}$. De la sección 7.2 sabemos que $\chi^2_{[\nu_1]}/\nu_1 = F_{[\nu_1,\infty]}$. Por lo tanto, cuando $\nu_1 = 1$ y $\nu_2 = \infty$, $\chi^2_{[1]} = F_{[1,\infty]} = t^2_{[\infty]}$. Esto puede demostrarse por las tablas IV, V y III, respectivamente.

$$\chi^2_{.05[1]} = 3.841$$
$$F_{.05[1,\infty]} = 3.84$$
$$t_{.05[\infty]} = 1.960 \qquad t^2_{.05[\infty]} = 3.8416$$

La prueba t para las diferencias entre dos medias es útil cuando queremos fijar límites de confianza a esta diferencia. El cuadro 8.2 muestra como calcular límites de confianza del 95 % para la diferencia entre las medias de serie en el ejemplo de *Daphnia*. El error típico y grados de libertad apropiados dependen de que se elija para t_s la expresión (8.2), (8.3) u (8.4). No nos sorprende descubrir que en este caso los límites de confianza de la diferencia incluyen el valor de cero, variando desde $-0{,}8307$ hasta $+0{,}7451$. Esto debe ser así cuando se encuentra que una diferencia no es significativamente diferente de cero. Podemos interpretar esto diciendo que no podemos excluir el cero como el verdadero valor de la diferencia entre las medias de las dos series.

8.5 Comparaciones entre medias: tests a priori

Hemos visto que después del test de significación inicial, un análisis de la varianza modelo II se completa con la estimación de los componentes aditivos de la varianza. Ordinariamente completamos un análisis de la varianza modelo I, de más de dos grupos, examinando los datos con más detalle, contrastando qué medias son diferentes entre sí o qué grupos de medias son diferentes de otros grupos semejantes o de medias individuales. Vamos a referirnos otra vez a los análisis de la varianza modelo I tratados hasta ahora en este capítulo. Podemos disponer inmediatamente del caso de dos muestras del cuadro 8.2, la edad media de las pulgas de agua al comienzo de la reproducción. Como se recordará,

no había diferencia significativa en edad entre las dos series genéticas. Pero aunque hubiese habido tal diferencia, no serían posibles otras pruebas. Sin embargo, los datos de longitud de secciones de guisante dados en el cuadro 8.1 muestran una diferencia significativa entre los cinco tratamientos (basados en 4 grados de libertad). Aunque sabemos que las medias no son todas iguales, no sabemos cuáles difieren entre sí. Esto nos conduce al tema de prueba entre pares y grupos de medias. Así por ejemplo, podríamos probar el control frente a los 4 tratamientos experimentales que representan los azúcares añadidos. La cuestión a contrastar sería, ¿afecta la adición de azúcares a la longitud de las secciones de guisantes? Podríamos probar también las diferencias entre los tratamientos de azúcares. Una prueba lógica podría ser azúcares puros (glucosa, fructosa y sacarosa) frente a azúcares mezclados (glucosa al 1,%, fructosa al 1 %).

Un punto importante sobre estas pruebas es que se diseñan y eligen independientemente de los resultados del experimento. Deberían planearse *antes* de que el experimento se haya realizado de hecho y obtenido los resultados. Estas comparaciones se denominan *planificadas* o *comparaciones a priori*. Estas pruebas se aplican con independencia de los resultados del análisis completo preliminar de la varianza. Por el contrario, después de haberse realizado el experimento podríamos querer comparar ciertas medias que observamos que son notablemente diferentes. Por ejemplo, la sacarosa con una media de 64,1 parece haber tenido menor efecto de inhibición del crecimiento que la fructosa con una media de 58,2. Pudiéramos por tanto querer comprobar si hay en realidad una diferencia significativa entre los efectos de la fructosa y la sacarosa. Estas comparaciones, que se proponen como resultado del experimento efectuado, se denominan *no planificadas* o *comparaciones a posteriori*. Estas pruebas solamente se realizan si el análisis de la varianza global preliminar es significativo. Incluyen pruebas de comparación entre todos los pares de medias posibles. Cuando hay a medias, puede naturalmente haber $a(a-1)/2$ posibles comparaciones entre pares de medias. La razón de que hagamos esta distinción entre comparaciones a priori y a posteriori, es que las pruebas de significación apropiadas para las dos comparaciones son diferentes. Un sencillo ejemplo análogo explicará por qué esto es así.

Vamos a suponer que hemos muestreado de una población de hombres de altura aproximadamente normal. Hemos calculado su media y desviación típica. Si de esta población muestreamos dos hombres a la vez, podemos predecir la diferencia entre ellos basándonos en la teoría estadística ordinaria. Algunos hombres serán muy similares y otros relativamente diferentes. Sus diferencias estarán normalmente distribuidas con una media de 0 y una varianza esperada de $2\sigma^2$, por razones que se descubrirán en un capítulo posterior (sección 12.2). Así, si obtenemos una gran diferencia entre una pareja de hombres extraídos al azar, tendrá que ser un número suficiente de desviaciones típicas mayor que cero, para que rechacemos nuestra hipótesis nula de que los dos hombres provienen de la población indicada. Por otra parte, si observásemos las alturas de las personas antes de muestrearlas y luego cogiésemos parejas que pareciesen ser muy diferentes, es obvio que obtendríamos repetidamente diferencias entre parejas de hombres que diferirían en varias desviaciones típicas. Estas diferencias estarían separadas del resto en la distribución de frecuencias esperada de las diferencias, y una y otra vez rechazaríamos nuestra hipótesis nula cuando de hecho es cierta. Los hombres se han extraído de la misma población, pero debido a que no se han extraído al azar sino que se han examinado antes de

extraerlos, la distribución de probabilidad en la cual se basa nuestro contraste de hipótesis ya no es válida. Es obvio que en una muestra grande de una distribución normal las colas diferirán en 5 a 7 desviaciones típicas, y que si tomamos deliberadamente individuos de cada cola y los comparamos, parecerán ser en gran manera significativamente diferentes entre sí, aunque pertenezcan a la misma población.

Cuando comparamos medias que difieren mucho entre sí como resultado de algún tratamiento en el análisis de la varianza, hacemos exactamente lo mismo que al coger el hombre más alto y el más bajo de la distribución de frecuencias de alturas. Si queremos saber si éstas son significativamente diferentes entre sí, no podemos utilizar la distribución ordinaria de probabilidad en la que se basa el análisis de la varianza, sino que tenemos que utilizar pruebas de significación especiales. Estas pruebas a posteriori se discutirán en la próxima sección. Esta se ocupa de la realización de pruebas a priori, aquellas comparaciones planificadas antes de la realización del experimento.

La regla general para hacer una comparación planificada es sumamente sencilla y vuelve a estar relacionada con la regla de obtención de la suma de cuadrados para cualquier serie de grupos (discutida al final de la sección 8.1). Para comparar k grupos de tamaño cualquiera n_i, se eleva al cuadrado la suma de cada grupo, se divide por su tamaño de muestreo n_i, y se suman los k cocientes así obtenidos. De la suma de estos cocientes se resta un término de corrección, que será la suma total de todos los grupos de esta comparación, elevada al cuadrado y dividida por el número de ítems de esta suma total. Si la comparación incluye a todos los grupos en el análisis de la varianza, el término de corrección será el *T.C.* general del estudio. En cambio si la comparación incluye solamente algunos de los grupos del análisis de la varianza, el *T.C.* será diferente, restringiéndose solamente a estos grupos.

TABLA 8.2

Medias, sumas de grupo, y tamaños de muestreo, de los datos del cuadro 8.1. Longitud de secciones de guisantes desarrollados en cultivo de tejidos (en unidades del micrómetro ocular).

	Control	*Glucosa al 2 %*	*Fructosa al 2 %*	*Glucosa al 1 % + fructosa al 1 %*	*Sacarosa al 2 %*	Σ
$\bar{Y}$	70,1	59,3	58,2	58,0	64,1	$(61,94 = \bar{\bar{Y}})$
ΣY	701	593	582	580	641	3097
n	10	10	10	10	10	50

Estas reglas pueden aprenderse mejor por medio de un ejemplo. La tabla 8.2 registra las medias, sumas de grupo y tamaños de muestra del experimento de las secciones de guisante del cuadro 8.1. Se recordará que había diferencias altamente significativas entre los grupos. Ahora queremos probar si los controles son diferentes en los cuatro tratamientos que representan adición de azúcar. Habrá por lo tanto dos grupos, el grupo control y el grupo "azúcares", con una suma de 2 396 y un tamaño de muestreo de 40. Por consiguiente calculamos

S.C. (control respecto a azúcares)

$$= \frac{(701)^2}{10} + \frac{(593 + 582 + 580 + 641)^2}{40} - \frac{(701 + 593 + 582 + 580 + 641)^2}{50}$$

$$= \frac{(701)^2}{10} + \frac{(2396)^2}{40} - \frac{(3097)^2}{50} = 832{,}32$$

En este caso el término de corrección es el mismo que para el análisis de la varianza porque incluye a todos los grupos del estudio. El resultado es una suma de cuadrados para la comparación entre estos dos grupos. Puesto que una comparación entre dos grupos tiene solamente un grado de libertad, la suma de cuadrados es al mismo tiempo una media cuadrática. Esta media cuadrática se contrasta con la media cuadrática del error del análisis de la varianza, para dar la siguiente comparación:

$$F_s = \frac{M.C.\ (\text{control frente a azúcares})}{M.C._{\text{intra}}} = \frac{832{,}32}{5{,}46} = 152{,}44$$

$$F_{0{,}05[1.45]} = 4.05, \qquad F_{0{,}01[1.45]} = 7{,}23,$$

Esta comparación es altamente significativa, mostrando que la adición de azúcares retarda significativamente el crecimiento de las secciones de guisante.

A continuación probamos si la mezcla de azúcares es significativamente diferente de los azúcares puros. Utilizando la misma técnica calculamos

S.C. (azúcares mezclados respecto a azúcares puros)

$$= \frac{(580)^2}{10} + \frac{(593 + 582 + 641)^2}{30} - \frac{(593 + 582 + 580 + 641)^2}{40}$$

$$= \frac{(580)^2}{10} + \frac{(1816)^2}{30} - \frac{(2396)^2}{40} = 48{,}13$$

Aquí el *T.C.* es diferente, basándose en la suma de los azúcares solamente. La prueba estadística apropiada es

$$F_s = \frac{M.C.\ (\text{azúcares mezclados respecto a azúcares puros})}{M.C._{\text{intra}}} = \frac{48{,}13}{5{,}46} = 8{,}82$$

Este es significativo a la vista de los valores críticos de $F_{\alpha[1.45]}$ dados en el párrafo anterior.

Una última prueba final se hace entre los tres azúcares. Esta media cuadrática tiene 2 grados de libertad, ya que está basada en tres medias. Así calculamos

$$S.C.\text{ (entre azúcares puros)} = \frac{(593)^2}{10} + \frac{(582)^2}{10} + \frac{(641)^2}{10} - \frac{(1816)^2}{30} = 196{,}87$$

$$M.C.\text{ (entre azúcares puros)} = \frac{S.C.\text{ (entre azúcares puros)}}{g.l.} = \frac{196{,}87}{2} = 98{,}435$$

$$F_s = \frac{M.C.\text{ (entre azúcares puros)}}{M.C._{\text{intra}}} = \frac{98{,}435}{5{,}46} = 18{,}06$$

Este F_s es altamente significativo puesto que incluso $F_{0{,}01[2,40]} = 5{,}18$.

Concluimos que la adición de los tres azúcares retarda el crecimiento de las secciones de guisante, que los azúcares mezclados afectan a las secciones de forma diferente respecto de los azúcares puros, y que los azúcares puros son significativamente diferentes entre sí, probablemente debido a que la sacarosa tiene una media mucho mayor. No podemos probar la sacarosa frente a los otros dos, porque ésta sería una prueba a posteriori que se nos propondría después de haber visto los resultados. Para realizar esta prueba necesitamos los métodos de la próxima sección.

Sin embargo, nuestras pruebas a priori podrían haber sido muy diferentes, dependiendo por completo de nuestras hipótesis iniciales. Así, podríamos haber probado inicialmente el control respecto a los azúcares, seguido por disacáridos (sacarosa) respecto a monosacáridos (glucosa, fructosa, glucosa + fructosa), seguido por mezclados frente a monosacáridos puros y finalmente por glucosa frente a fructosa.

El tipo y número de pruebas a priori vienen determinados por las hipótesis sobre los datos. No obstante, hay ciertas restricciones. Naturalmente no sería correcto decidir a priori que uno desearía comparar cada media frente a cada otra media [$a(a-1)/2$ comparaciones]. Para a grupos la suma de los grados de libertad de las diferentes pruebas a priori no debería exceder de $a - 1$. Además, es deseable estructurar las pruebas de tal manera que cada una de ellas pruebe una relación independiente entre las medias (como se ha hecho en el ejemplo anterior). Por ejemplo, si ya hubiésemos hallado que la media uno difería de la tres, preferiríamos no probar si las medias uno, dos y tres diferían puesto que la significación de la segunda implica significación de la primera.

Puesto que estas pruebas son independientes, las tres sumas de cuadrados que hemos obtenido hasta ahora, basadas en 1,1 y 2 *g.l.* respectivamente, en conjunto totalizan la suma de cuadrados entre tratamientos del análisis de la varianza primitivo basado en 4 grados de libertad. Así:

		g.l.
S.C. (control respecto a azúcares)	= 832,32	1
S.C. (azúcares mezclados respecto a azúcares puros)	= 48,13	1
S.C. (entre azúcares puros)	= 196,87	2
S.C. (entre tratamientos)	= 1077,32	4

Esto demuestra nuevamente la elegancia del análisis de la varianza. Las sumas de cuadrados entre tratamientos pueden descomponerse en diferentes partes que son sumas de cuadrados por sí mismas, con sus grados de libertad correspondientes. Una suma de cuadrados mide la diferencia entre los controles y los azúcares, la segunda entre los azúcares mezclados y los azúcares puros, y la tercera, la varianza residual entre los tres azúcares. Podemos presentar todos estos resultados como una tabla de análisis de la varianza, tal como se expone en la tabla 8.3.

TABLA 8.3

Tabla de análisis de la varianza del cuadro 8.1, con la suma de cuadrados entre tratamientos, descompuesta en comparaciones planificadas.

Origen de la variación	*S.C.*	*g.l.*	*M.C.*	F_s
Tratamientos	1077,32	4	269,33	49,33**
Control respecto a azúcares	832,32	1	832,32	152,44**
Mezclados respecto a azúcares puros	48,13	1	48,13	8,82**
Entre azúcares puros	196,87	2	98,44	18,03**
Intra	245,50	45	5,46	
Total	1322,82	49		

8.6 Comparaciones entre medias: pruebas a posteriori

Si un análisis de la varianza de clasificación simple es significativo, quiere decir naturalmente que

$$\frac{M.C._{\text{grupos}}}{M.C._{\text{intra}}} \geq F_{\alpha[a-1,a(n-1)]} \tag{8.5}$$

Como $M.C._{\text{grupos}}/M.C._{\text{intra}} = S.C._{\text{grupos}}/[(a-1)M.C._{\text{intra}}]$, podemos volver a escribir la expresión (8.5) como

$$S.C._{\text{grupos}} \geq (a-1)\, M.C._{\text{intra}}\; F_{\alpha[a-1,a(n-1)]} \tag{8.6}$$

Por ejemplo, en el cuadro 8.1, en el que el análisis de la varianza es significativo, $S.C._{\text{grupos}} = 1077{,}32$. Sustituyendo en la expresión (8.6) obtenemos

$$1077{,}32 > (5-1)\,(5{,}46)\,(2{,}58) = 56{,}35 \text{ para } \alpha = 0{,}05$$

Por lo tanto, es posible calcular un valor crítico de *S.C.* para una prueba de significación de un análisis de la varianza. Así, otro modo de calcular la significación global sería ver si