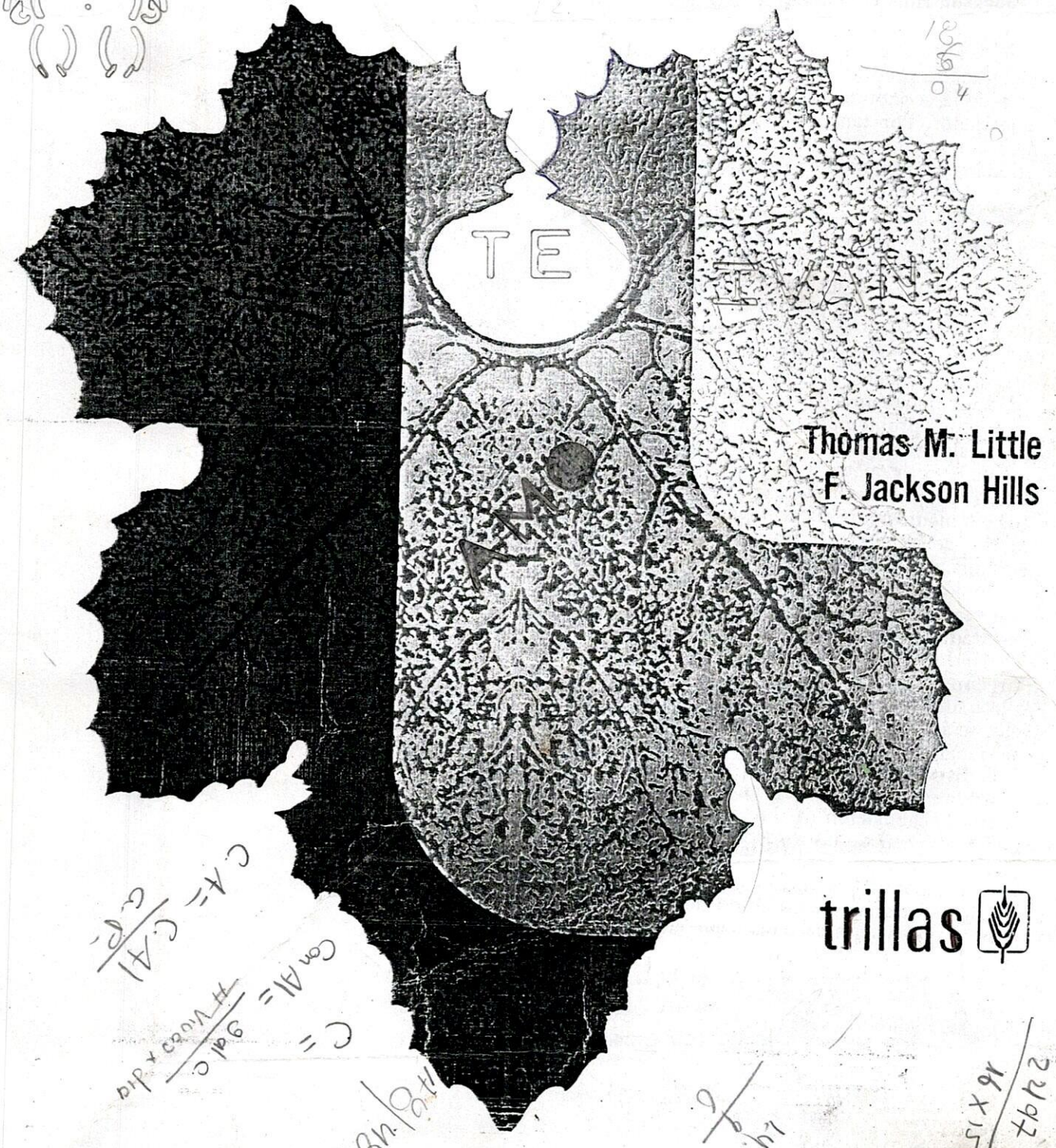
Métodos estadísticos para la investigación en la agricultura



Thomas M. Little
F. Jackson Hills

trillas

MÉTODOS ESTADÍSTICOS PARA LA INVESTIGACIÓN EN LA AGRICULTURA

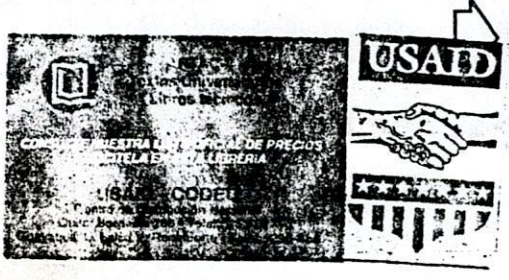
Thomas M. Little
F. Jackson Hills

La agricultura constituye actualmente la base de la economía en los países latinoamericanos. Por tanto, el presente volumen constituye un instrumento de gran utilidad, pues expone un estudio profundo acerca de las aplicaciones que tiene la estadística en la investigación para el mejoramiento de las técnicas agropecuarias. A través de este libro se ofrece una explicación sencilla y clara de lo que es en sí la ciencia estadística, cómo utilizarla correctamente y de qué maneras se puede aplicar en el diseño, la ejecución y el análisis de experimentos agrícolas. Asimismo contiene planteamientos atinentes a cada uno de los problemas que se tratan, y las razones por las cuales el conocimiento de la estadística es necesario.

Con el propósito de dar a conocer al lector los medios más eficaces para la resolución de problemas relativos al sector agrícola, se ha incluido en la obra el desarrollo detallado de diversos experimentos con plantas y animales, y sus respectivos resultados.

Teniendo en cuenta la necesidad de contar con un método para la realización adecuada de investigaciones agropecuarias, se abarcan no sólo las formas lógicas de la investigación, como el razonamiento inductivo y el deductivo, sino también los procedimientos que se llevan a cabo mediante la tabulación de datos que se obtienen a partir de las leyes de la causalidad exacta.

El texto contiene también gran número de figuras y gráficas, así como de tablas para la ordenación de datos, que contri-



Métodos estadísticos para la investigación en la agricultura

Thomas M. Little

Biómetra de extensión,
profesor emérito de la
University of California,
Riverside

F. Jackson Hills

Agrónomo de extensión,
University of California,
Davis

EDITORIAL TRILLAS

México, Argentina, España,
Colombia, Puerto Rico, Venezuela

Catalogación en la fuente

Little, Thomas M.
Métodos estadísticos para la investigación en la agricultura. – 2a ed. – México : Trillas, 1989 (reimp. 1991).
270 p. : il. ; 28 cm.
Traducción de: Statiscal methods in agricultural research
Bibliografía: p. 270
ISBN 968-24-3629-X

1. Agricultura - Métodos estadísticos. I. t.

LC- S540.A3.S7'L5.5 D- 630.20112'L218m 625

Título de esta obra en inglés:
Statiscal Methods in Agricultural Research

Versión autorizada en español de la segunda reimpresión publicada en ingles por
© T.M. Little y F. J. Hills,
California, E. U. A.

Primera edición en español, 1976 (ISBN 968-24-0528-9)
Reimpresiones, 1978, 1979, 1981, 1983, 1984, 1985 y 1987
Segunda edición en español, 1989 (ISBN 968-24-3629-X)
Reimpresión, 1990

Segunda reimpresión, octubre 1991

Impreso en México
Printed in Mexico

Prólogo

Desde 1957 hemos entrenado agentes de extensión en métodos estadísticos para mejorar la creciente participación del Servicio de Extensión Agrícola de la Universidad de California, en la resolución de problemas de investigación. El precursor de este libro, *Métodos experimentales para extensionistas* y su folleto asociado, *Correlación y regresión*, fueron diseñados como parte de ese esfuerzo. Nuestro primer curso intensivo, "Diseños para la investigación y análisis", fue impartido en 1965, y desde entonces el curso ha sido repetido en veinte ocasiones, entrenando a más de 250 agentes de extensión y miembros del cuerpo universitario. *Métodos estadísticos para la investigación en la agricultura* ha sido diseñado como un texto para este curso, en un intento por explicar, en los términos más sencillos posibles, de qué trata la ciencia estadística, cómo es utilizada en el proyecto, la ejecución y el análisis de experimentos, y por qué su utilización es necesaria.

La mayor parte de lo que hemos escrito aparece más detallado en varios de los excelentes textos disponibles acerca de proyectos estadísticos y análisis. Nuestro propósito es brindar al experimentador un minucioso conocimiento de importantes instrumentos estadísticos, en el menor tiempo posible. Los procedimientos son presentados con el fin de facilitar una comprensión del *cómo* de cada técnica mediante un intento de aplicar muchos de los *porqué*, a través de un enfoque intuitivo.

Los lectores interesados en poseer mayores detalles acerca de ciertos temas, deben consultar el apéndice A.13, donde encontrarán una lista de lecturas complementarias.

T.M. Little y F. J. Hills

Índice

Simbología

Convencional	
a	Intercepto o punto donde una recta o curva cruza el eje de las *Y*.
b	Pendiente de una recta, también denominada coeficiente de regresión.
c	Coeficiente de X^2 en cualquier ecuación de polinomios.
c_i	Coeficientes para las ecuaciones de polinomios apropiadas a los datos provenientes de incrementos a intervalos iguales de la variable independiente.
C	Término de corrección para calcular una suma de cuadrados.
LC	Límites de confianza de una media o de una diferencia de medias.
CV	Coeficiente de variación.
d	Diferencia entre dos valores de la variable.
$\bar{d}$	Diferencia entre dos medias.
gl	Grados de libertad.
F	Razón entre dos varianzas o cuadrados medios.
K	Factor utilizado en el cálculo abreviado de ecuaciones de polinomios de regresión.
L	Diferencia común entre los valores de *X* a intervalos iguales.
DSM	Diferencia significativa mínima.
M	Número de medias en una población de medias.
CM	Cuadrado medio. Generalmente se usa una tercera letra para indicar un cuadrado medio específico; por ejemplo, *CME* es el cuadrado medio del error.
N	Número de elementos en una población.
n	Número de elementos en una muestra.
P	Suma de los productos de una serie de coeficientes y sus correspondientes valores de *Y*, $P = \Sigma(c_i Y)$.
Q	Número de diferencias de medias en una población de diferencias de medias o en un factor calculado en la prueba de aditividad de Tukey.
r	Coeficiente de correlación lineal.
r^2	Coeficiente de determinación lineal.
R^2	Coeficiente de determinación de una correlación simple no lineal o de una correlación múltiple.
SC	Suma de cuadrados. En el análisis de varianza, *SC* indica la suma total de cuadrados; la letra mayúscula siguiente indica alguna otra suma específica de cuadrados; por ejemplo, *SCT* es la suma de cuadrados para tratamientos.
DSM	Diferencia significativa mínima en la prueba de rango múltiple de Duncan.

Convencional

s, s^2	Desviación estándar y varianza, respectivamente, de una población de elementos individuales estimadas a partir de una muestra.
$s_{\bar{x}}$, $s^2_{\bar{x}}$	Error estándar y varianza, respectivamente, de una población de medias estimadas a partir de una muestra.
$s_{\bar{d}}$, $s^2_{\bar{d}}$	Error estándar y varianza, respectivamente, de una población de diferencias de medias estimadas a partir de dos muestras.
t	Una diferencia expresada en unidades de error estándar.
U, V	Coeficientes en curvas periódicas apropiadas y sumas en partición de cuadrados para datos tomados a intervalos iguales de tiempo durante un ciclo completo.
X	Un elemento. Valor de una unidad experimental individual.
$\bar{X}$	Media aritmética de todos los elementos de una muestra.
x	Desviación de un elemento con respecto a la media, $X-\bar{X}$.
Y	Un elemento. El valor de una observación en particular. La *Y* es utilizada en lugar de la *X*, para indicar que la variable dependiente está en correlación y regresión.
$\bar{Y}$	Media aritmética de una muestra de valores de Y.
y	Desviación de un valor de *Y* con respecto a la media, $Y-\bar{Y}$.
$\hat{Y}$	Valor estimado de *Y* a partir de una ecuación de regresión.
z	Diferencia expresada en unidades de error estándar, cuando el error estándar de la población es conocido o es estimado a partir de una muestra grande.

Letras griegas

μ	Media de una población de elementos individuales.
$\mu_{\bar{x}}$	Media de una población de medias.
$\mu_{\bar{d}}$	Media de una población de diferencias de medias.
Σ	Símbolo de sumatoria. La suma de diversas unidades.
$\sigma, \sigma_Y, \sigma_X$	Desviación estándar de una población de elementos individuales. Y indica una población de *Y* elementos; y X, una población de *X* elementos.
$\sigma^2, \sigma^2_Y, \sigma^2_X$	Varianza de una población de elementos individuales.
$\sigma_{\bar{x}}, \sigma^2_{\bar{x}}$	Desviación estándar y varianza, respectivamente, de una población de medias.
$\sigma_{\bar{d}}, \sigma^2_{\bar{d}}$	Desviación estándar y varianza, respectivamente, de una población de diferencias de medias.
χ^2	Ji cuadrada. Valor calculado para comprobar la efectividad o idoneidad, utilizado con datos basados en cómputos.

Capítulo 1

Lógica, investigación y experimento

"El propósito de la ciencia estadística es suministrar una base objetiva para el análisis de problemas en los que los datos se apartan de las leyes de la causalidad exacta. Se ha ideado un sistema lógico general de razonamiento inductivo, aplicable a datos de esta naturaleza, y en la actualidad se utiliza ampliamente en la investigación científica. Es importante comprender sus principios, tanto para los investigadores científicos como para aquellos cuyos intereses residen en la aplicación de avances tecnológicos resultantes de dichas investigaciones. Esto es especialmente cierto para las ciencias agrícolas y biológicas."

D. J. Finney, Introducción a la ciencia estadística en la agricultura.

La cita anterior es una exposición concisa de la importancia de la ciencia estadística en la agricultura. Para entender más ampliamente qué es un "sistema lógico de razonamiento inductivo", debemos repasar algunos conceptos elementales de lógica. Cuando clasificamos los problemas de acuerdo con el sistema de razonamiento empleado en su resolución, encontramos precisamente dos clases de problemas.

RAZONAMIENTO DEDUCTIVO

En primer lugar, existe el tipo de problema en el cual se nos da algún principio o conjunto de principios generales y se nos pide que determinemos qué sucedería bajo un conjunto específico de condiciones. El tipo de razonamiento empleado, de lo general a lo particular, recibe el nombre de **razonamiento deductivo.**

Unos cuantos ejemplos servirán para aclarar este concepto.

Dada la fórmula general para el área de un círculo, $A = \pi r^2$, ¿cuál es el área de un círculo cuyo radio es de 15 centímetros?

Dada una clave y descripciones de las malas hierbas de California, ¿a qué especie debe pertenecer cierta mala hierba?

Dadas las leyes de Boyle y Charles, ¿cómo debemos esperar que cierto volumen de gas cambie al ser sometido a cambios de presión y temperatura específicos?

Dados algunos principios generales de control de enfermedades, ¿qué respuesta cabe esperar al aplicar determinada dosis de un fungicida a un acre de un cultivo en particular?

Dada una moneda neutral cuya probabilidad de que caiga en cara al ser lanzada al aire es de un medio, ¿qué sucederá cuando dicha moneda sea lanzada al aire diez veces?

El lector deberá pensar en otros ejemplos basados en su propia experiencia.

Casi todos los problemas encontrados durante nuestra educación formal fueron de este tipo, donde la solución requirió el razonamiento deductivo. Frecuentemente se dice que un agricultor debería ser "versado en los fundamentos básicos". Esto implica que debería tener bajo su control un gran acervo de principios generales y las habilidades del razonamiento deductivo para aplicarlos a casos específicos.

RAZONAMIENTO INDUCTIVO (2)

El segundo tipo de problema es opuesto al primero. Se nos dan algunos casos específicos y se nos pide que lleguemos a algunos principios generales que serán aplicados a todos los miembros de la clase representada por dichos casos. El razonamiento empleado, de lo específico a lo general, se denomina **razonamiento inductivo.**

Los siguientes ejemplos de problemas, que requieren del razonamiento inductivo, son análogos a aquellos que se han presentado anteriormente para ilustrar el tipo de problemas deductivos:

Dadas las áreas y los radios de diversos círculos, ¿qué fórmula general podemos dar, expresando la relación entre las áreas y los radios de todos los círculos?

Dados diversos tipos de una especie de mala hierba no descrita, ¿cómo podríamos describir a la especie como un todo y expresar su relación con otras especies en una clave?

Dada una serie de observaciones sobre el volumen de un gas bajo diferentes condiciones de presión y temperatura, ¿qué leyes generales explicarán tales observaciones?

Dados los resultados de una serie de pruebas de control de enfermedades, ¿qué recomendaciones generales se pueden hacer respecto de la utilización de los métodos de control?

Dados los resultados de lanzar al aire una moneda diez veces, ¿a qué conclusiones podemos llegar respecto del sesgo o el insesgamiento de la moneda?

Nótese que todos los problemas de este tipo tienen algo en común: todos empiezan con un **grupo de observaciones**. En algunos casos, como en la descripción de una nueva especie, las observaciones de fenómenos son hechas simplemente en la medida en que éstos tienen lugar en la naturaleza; sin embargo, por regla general, las observaciones se realizan bajo condiciones controladas. Los factores objeto de estudio se hacen variar en alguna forma sistemática, mediante la aplicación de **tratamientos**. Otros factores que pueden ejercer influencia sobre las observaciones son minimizados hasta el punto en que la práctica lo permita. Tenemos, entonces, un **experimento**.

EL PROBLEMA DEL INVESTIGADOR (3)

Hemos dicho que casi todos los problemas encontrados en nuestra enseñanza formal son del tipo que requiere del razonamiento deductivo. Podemos afirmar también que casi todos los problemas que afronta un agrónomo son aquellos que requieren del razonamiento inductivo.

¿Cuál es el problema típico que afrontan los investigadores agrícolas? Este podría establecerse en los siguientes términos generales: ¿afectará la utilización de una práctica nueva o diferente el resultado de algún segmento en particular de la empresa agrícola?; si así fuese, ¿en qué extensión lo haría? Puesto que estas interrogantes nunca pueden ser contestadas con un 100% de seguridad, debemos tener en cuenta también el riesgo y el costo de una toma de decisión incorrecta. Esto resultará más evidente a medida que avancemos en el estudio de este volumen.

Para resolver tal problema, por regla general se requiere un experimento. En el experimento más simple, debe haber sólo dos tratamientos: la nueva práctica y la vieja. Un experimento más complicado puede incluir diversas proporciones o métodos de aplicación de la nueva práctica. Aún más complejos son aquellos experimentos en los que los efectos de diversas prácticas se estudian simultáneamente.

Cualquiera que sea el diseño del experimento, su propósito es suministrar un medio de realizar observaciones (muestreo probabilístico) que pueda utilizarse para formular generalizaciones plausibles acerca de la práctica, objeto de estudio. Llegar a tales generalizaciones es un problema típico del razonamiento inductivo.

El lector no debe tener la equivocada impresión de que el razonamiento inductivo contempla una línea independiente de pensamiento, distinta del razonamiento deductivo. Las conclusiones inductivas deben ser siempre comprobadas mediante precisos métodos deductivos.

EL ELEMENTO ALEATORIO

Otra frase que aparece en la cita al inicio de este capítulo requiere alguna aclaración. ¿Qué se entiende por "problemas en los que los datos se apartan de la leyes de la causalidad exacta"?

Fijándonos en los ejemplos de problemas presentados en párrafos anteriores, notamos que existen algunas diferencias importantes entre los mismos. En el problema de encontrar el área de un círculo, no hay incertidumbre en cuanto a la respuesta. Para cualquier radio dado, sólo puede existir una respuesta.

El problema de la moneda lanzada al aire constituye algo diferente. El supuesto general es que la moneda es neutral; pero incluso para una sola oportunidad, estamos inseguros en cuanto al resultado. Puede obtenerse sólo uno de los dos posibles, siendo ambos igualmente probables. La pregunta de qué sucederá cuando la moneda sea lanzada diez veces al aire, tiene una respuesta aún más insegura, existiendo para la misma once resultados posibles según el número de veces en que ésta caiga en cara; estos resultados difieren en cuanto a su probabilidad de registrarse. Obviamente ocurrirán casos fortuitos en esta ocasión, para los cuales no existe una relación simple uno a uno entre causa y efecto.

Tal situación es casi universal en el campo de la agricultura. No importa cuánto sepa un científico sobre nutrición y fisiología: no podrá predecir con exactitud cuál será el aumento de peso de una res o la cosecha de una parcela de papas bajo un conjunto de condiciones dado. Variaciones aleatorias, debidas a una multitud de causas, siempre harán variar los resultados, sin importar la cantidad de esfuerzo desplegado para controlar todos los factores conocidos.

El término aleatorio resulta difícil de definir; pero aun sin una clara definición, todos sabemos lo que significa, en grado suficiente como para considerar su importancia en lo que a efectos sobre los resultados biológicos se refiere.

Cuando el elemento aleatorio forma parte de un problema, se introducen dificultades reales. Estas resultan mucho más serias en el campo del razonamiento inductivo que en el del razonamiento deductivo.

Consideremos el problema deductivo que consiste en lanzar al aire una moneda neutral diez veces. Mediante métodos deductivos podemos enumerar todos los once posibles resultados y calcular fácilmente la

probabilidad de cada uno de ellos; por ejemplo, supóngase que preguntamos: "¿Cuál es la probabilidad de obtener cinco caras y cinco cruces?" Esta respuesta puede encontrarse mediante el cálculo del valor de

$$(10 \times 9 \times 8 \times 7 \times 6)/(2 \times 3 \times 4 \times 5 \times 2^{10})$$

cuyo resultado es 63/256 o 24.6%. A medida que el número de lanzamientos se incrementa, o los supuestos iniciales se modifican para incluir ciertos grados de sesgo en la moneda, los cálculos se vuelven más trabajosos aunque siguen siendo directos, y los resultados son sencillos y definidos. Afortunadamente, la teoría de probabilidades ha sido desarrollada por matemáticos, de modo que disponemos de métodos abreviados y tablas que reducen en gran medida los cálculos necesarios en casos complicados.

Considérese ahora el problema inverso o inductivo. Si una moneda es lanzada al aire diez veces y cae cinco veces en cara y cinco en cruz, ¿qué podemos decir acerca del sesgo o insesgamiento de la misma? Todo lo que podemos decir con seguridad es que la moneda no tenía dos caras ni dos cruces. Si ésta fuese neutral, cabría esperar este resultado en aproximadamente 25% de las veces en que el intento se repita. Podemos afirmar, con elevado grado de probabilidad de estar en lo cierto, que la moneda no se halla fuertemente inclinada a favor de la cara o de la cruz. Debemos recordar que nunca podemos hacer tal aseveración con absoluta seguridad. Incluso con una moneda fuertemente sesgada (una que cae en cara el 90% de las veces), el resultado observado de cinco caras y cinco cruces tal vez sería **posible**, pero no muy **probable**.

La otra aseveración que podemos formular acerca de la moneda es que nos sentimos bastante confiados de que su grado de inclinación se encontraba en algún punto entre una ligera inclinación a favor de la cruz y una ligera inclinación a favor de la cara. Nótese que existen infinitas posibilidades en ese intervalo, y que cero inclinación es una de ellas. Esto es muy importante para entender que sin más conocimientos acerca de la moneda que los resultados de estos diez lanzamientos, no hay justificación para concluir que la moneda era neutral. Con un mayor número de lanzamientos sería posible estrechar el intervalo de inclinaciones que podrían ser razonablemente esperadas para producir el resultado observado; pero nunca seremos capaces de establecer con seguridad que la moneda era neutral.

Premeditadamente, hemos evitado definir los términos sesgo **fuerte** y **leve**, en bien de la simplicidad; sin embargo, resulta posible, mediante métodos estadísticos, determinar qué escala de sesgos aceptaremos o rechazaremos de acuerdo al grado de confiabilidad que deseemos tener en nuestras conclusiones.

Ahora podemos ver que la respuesta a la pregunta: "¿Qué podemos decir acerca del sesgo de la moneda?" fue bastante vaga. El lector que está acostumbrado sólo a las respuestas precisas de las matemáticas deductivas, puede estar desilusionado con la vaguedad de la respuesta; sin embargo, a pesar de lo insatisfactoria que ésta puede parecer, la verdadera naturaleza del razonamiento inductivo es tal que esta respuesta es la mejor que podemos dar. Como dijo el gran filósofo matemático Alfred North Whitehead: "La teoría de la inducción es la desesperanza de la filosofía —y aun todas nuestras actividades están basadas en la misma."

El investigador no debe desesperanzarse en sus intentos por responder preguntas a través de observaciones y experimentos; no obstante, deberá darse cuenta de que sus respuestas no serán nunca absolutas. Deberá hacer generalizaciones con precaución y sólo después de efectuar cuidadosas observaciones y de ejercitar los mejores sistemas de razonamiento bajo su control.

NECESIDAD DE LA EVALUACIÓN ESTADÍSTICA

La mayoría de los agrónomos ven rápidamente la necesidad del análisis estadístico para sentar una base objetiva de evaluación; algunos ejemplos pueden resultar útiles. Si cosechamos dos áreas iguales de trigo en un campo, el grano producido en estas áreas, si son siembras por varas de longitud o mitades del campo entero, rara vez será igual; el peso de los frutos de árboles adyacentes en un huerto, difícilmente es el mismo; la proporción de aumento de peso de dos animales cualesquiera de la misma especie y raza, casi siempre

difiere. Las diferencias de este tipo entre cultivos o animales son debidas a diferencias genéticas y ambientales más allá del control razonable de un experimentador. No hay errores en el sentido de estar equivocados; éstas representan la variabilidad entre unidades experimentales, denominada **error experimental.**

Una vez que reconocemos la existencia de esta variabilidad, entendemos la dificultad para evaluar una nueva práctica, mediante su aplicación a una unidad experimental única y su comparación con otra unidad que es similar, pero **no tratada.** El efecto de la nueva práctica se confunde con la variabilidad no determinada. Así, un experimento con una sola réplica suministra una medición incompleta del efecto del tratamiento; además, puesto que no existen dos unidades experimentales igualmente tratadas, éste no suministra mediciones del error experimental. La ciencia estadística supera estas dificultades, requiriendo la recolección de datos experimentales que permitirán una estimación imparcial de los efectos del tratamiento y la evaluación de las diferencias del tratamiento a través de pruebas de significación basadas en mediciones del **error experimental.**

Los efectos del tratamiento son estimados mediante la aplicación de tratamientos a por lo menos dos unidades experimentales, por regla general a más de dos, y promediando los resultados para cada tratamiento. Las pruebas de significación determinan la probabilidad de que las diferencias de tratamiento pudieran haber ocurrido solamente por casualidad.

Existen tres importantes principios, inherentes a todos los proyectos experimentales que son esenciales para los objetivos de la ciencia estadística:

1. **Repetición.** La repetición significa que un tratamiento se efectúa dos o más veces. Su función es suministrar una estimación del error experimental y brindar una medición más precisa de los efectos del tratamiento. El número de repeticiones que se requerirán en un experimento particular, depende de la magnitud de las diferencias que deseamos detectar y de la variabilidad de los datos con los que estamos trabajando. Considerando estos dos aspectos al inicio de un experimento, evitaremos muchas equivocaciones.

2. **Muestreo aleatorio.** El muestreo aleatorio es la asignación de tratamientos a unidades experimentales, de modo que todas las unidades consideradas tengan iguales probabilidades de recibir un tratamiento. Su función es asegurar estimaciones imparciales de medias de tratamientos y del error experimental.

3. **Control local.** Este principio de diseño experimental permite ciertas restricciones sobre la selección aleatoria para reducir el error experimental; por ejemplo, en el diseño de bloques completos al azar, los tratamientos son agrupados en bloques que se espera tengan un desempeño diferente, en el cual cada uno de ellos presenta un **efecto de bloque** que se puede separar de la variación total del experimento.

EL MÉTODO CIENTÍFICO

La investigación puede definirse en forma amplia como el estudio sistemático de un sujeto con el fin de descubrir nuevos hechos o principios. El procedimiento para la investigación se conoce generalmente como **método científico,** el cual, aunque difícil de definir con precisión, usualmente contiene los siguientes elementos:

1. **Hechos observados.** Se dice que la ciencia empieza con la observación, a través de la cual se establecen diversos factores.

2. **Hipótesis.** La consideración del conjunto de hechos acerca de un sujeto conduce al establecimiento de una hipótesis —una idea provisoria de cómo los hechos han de ser interpretados y explicados.

3. **Experimento.** El experimento es un ensayo destinado a probar la validez de la hipótesis propuesta.

4. **Resultados y su interpretación.** Los resultados del experimento establecen hechos adicionales. La interpretación de estos nuevos hechos a la luz de lo ya conocido, conduce al apoyo, rechazo o alteración de la hipótesis y de ese modo volvemos nuevamente a través de la misma serie de pasos.

El **experimento** es un instrumento de investigación utilizado para descubrir algo desconocido o para probar un principio o una hipótesis. Es un importante paso del método científico, y las preguntas que éste aspira a contestar serán fundamentales para el apoyo o rechazo de una hipótesis.

CARACTERÍSTICAS DE UN EXPERIMENTO BIEN PLANEADO

1. **Simplicidad.** La selección de tratamientos y la disposición experimental deberán hacerse del modo más simple posible y deberán ser consistentes con los objetivos del experimento.

2. **Grado de precisión.** Deberá haber una gran probabilidad de que el experimento sea capaz de medir diferencias de tratamientos con los grados de precisión deseados por el experimentador. Esto implica un diseño apropiado y un número suficiente de repeticiones.

3. **Ausencia de error sistemático.** Debe planearse el experimento para asegurar que las unidades experimentales que reciban un tratamiento no difieran sistemáticamente de aquellas que reciben otro tratamiento, de modo que pueda obtenerse una estimación imparcial de cada efecto de tratamiento.

4. **Rango de validez de las conclusiones.** Las conclusiones deben tener un rango de validez tan amplio como sea posible. Un experimento replicado en tiempo y espacio incrementaría el rango de validez de las conclusiones que podrían sacarse del mismo. Un conjunto factorial de tratamientos es otro medio para incrementar el rango de validez de un experimento. En un experimento factorial, los efectos de un factor son evaluados bajo niveles variantes de un segundo factor.

5. **Cálculo del grado de incertidumbre.** En cualquier experimento existe siempre algún grado de incertidumbre en cuanto a la validez de las conclusiones. El experimento deberá ser concebido de modo que resulte posible calcular la probabilidad de obtener los resultados observados debido únicamente al azar.

PROCEDIMIENTO PARA LA EXPERIMENTACIÓN

En la planeación y la conducción de un experimento hay un gran número de consideraciones que deben ponderarse cuidadosamente si el experimento ha de ser exitoso. Algunos de los pasos más importantes a dar son:

1. **Definición del problema.** El primer paso en la resolución de un problema consiste en establecer clara y concisamente el problema con que estamos tratando. Si èl problema no puede definirse, hay muy pocas probabilidades de que éste sea resuelto alguna vez. Cuando el problema se ha comprendido, debemos ser capaces de formular preguntas que, una vez contestadas, conduzcan a la solución.

2. **Determinación de los objetivos.** Estos pueden ser la forma en que las preguntas serán contestadas, la hipótesis que se va a comprobar o los efectos que se desea estimar. Los objetivos deberán redactarse en términos precisos. Dado este paso, el experimentador está capacitado para planear más efectivamente sus procedimientos experimentales. Cuando hay más de un objetivo, debemos ordenarlos de acuerdo con su importancia, como si tuvieran un lugar en el diseño experimental. Al establecer los objetivos debemos evitar la vaguedad o exceso de ambigüedad de los planteamientos.

3. **Análisis crítico del problema y de los objetivos.** La racionalidad y utilidad de las metas del experimento deberán considerarse cuidadosamente a la luz del estatus actual de conocimiento acerca

del problema. ¿Son los objetivos del experimento realmente importantes para la solución del problema?

La selección de un procedimiento para la investigación depende, en gran medida, del objeto de estudio en que la investigación está siendo conducida, así como de los objetivos de la investigación. La investigación puede ser descriptiva y contemplar una encuesta por muestreo, o puede contemplar un experimento o una serie de experimentos controlados. Algunas de las siguientes consideraciones, o todas estarán incluidas en el procedimiento para la mayoría de las áreas de investigación.

4. **Selección de tratamientos.** El éxito del experimento reside en la cuidadosa selección de tratamientos, cuya evaluación responderá a las preguntas que tengamos.

5. **Selección del material experimental.** En la selección del material experimental, considérense los objetivos del experimento, así como la población sobre la cual se harán las inferencias. El material utilizado deberá ser representativo de la población sobre la cual deseamos probar nuestros tratamientos.

6. **Selección del diseño experimental.** Nuevamente, aquí es importante hacer una consideración de los objetivos; pero una regla general podría ser elegir el diseño más simple que parece brindar la precisión requerida por nosotros.

7. **Selección de la unidad de observación y del número de repeticiones.** Para experimentos de campo, con plantas, estos medios determinan el tamaño y la forma de las parcelas de campo. Para experimentos con animales, estos medios determinan el número de animales que han de ser consíderados como una unidad experimental. La experiencia de otros experimentos similares es de incalculable valor para la toma de tales decisiones. El tamaño de la parcela y el número de repeticiones deberán ser elegidos para obtener la precisión requerida en la estimación de los tratamientos.

8. **Control de los efectos entre unidades adyacentes.** Esto suele llevarse a cabo a través de la utilización de callejones de demarcación y mediante la selección aleatoria de tratamientos.

9. **Consideración acerca de los datos que se van a recabar.** Los datos recabados deberán evaluar apropiadamente los efectos del tratamiento, de acuerdo con los objetivos del experimento; además, se deberá brindar atención a la recolección de datos que explicarán el desempeño de los tratamientos.

10. **Esbozo del análisis estadístico y del resumen de los resultados.** En el análisis de varianza, anótense las fuentes de variación y los grados de libertad asociados. Inclúyanse las diversas pruebas F que se planearon. Considérese cómo pueden utilizarse los resultados y prepárense posibles tablas de resumen o gráficas que muestren los efectos esperados. Compárense estos resultados esperados con los objetivos del experimento, a fin de verificar si el mismo suministrará las respuestas buscadas.

    A estas alturas es conveniente que todos estos puntos sean revisados por un estadístico, así como por uno o más colegas. La revisión realizada por otros puede revelar aspectos que pasaron inadvertidos. Ciertas alteraciones o ajustes pueden enriquecer grandemente el experimento, así como posibilitar un aprovechamiento más completo del trabajo que se va a realizar.

11. **Conducción del experimento.** En la conducción del experimento, aplíquense procedimientos libres de sesgos personales o favoritismos. Apliquese el diseño experimental para recabar datos, de modo que las diferencias entre individuos o las diferencias asociadas con el orden de recolección puedan ser removidas del error experimental. Evítese la fatiga en el acopio de datos. Vuélvanse a comprobar inmediatamente las observaciones que parecen fuera de lugar. Organícese la recolección de datos para facilitar el análisis y para evitar errores al recopiarlos. Si es necesario copiar los datos, compruébense inmediatamente los números copiados con los originales.

12. **Análisis de los datos e interpretación de los resultados.** Todos los datos deberán analizarse tal como fueron planeados; los resultados se deberán interpretar a la luz de las condiciones experimentales; se comprobará la hipótesis y deberá definirse la relación de los resultados con los hechos previamente establecidos. Recuérdese que la estadística no demuestra nada y que existe siempre una probabilidad de que las conclusiones puedan ser erróneas. Por tanto, considérense las consecuencias de tomar una decisión incorrecta. Evítese llegar a una conclusión, aun cuando ésta sea **estadísticamente significativa**, si la misma aparece fuera de lugar con respecto a hechos previamente establecidos. En este caso, debe investigarse exhaustivamente el asunto.

13. **Elaboración de un completo, legible y correcto informe de la investigación. No existen resultados negativos.** Si la hipótesis nula no se rechaza es una evidencia positiva de que pueden no existir verdaderas diferencias entre los tratamientos sometidos a prueba. Nuevamente recúrrase a los colegas y sométanse las conclusiones al tamiz de sus opiniones.

Nótese que la mayoría de los pasos anteriores no son estadísticos; sin embargo, el análisis estadístico constituye una parte importante de la experimentación. La ciencia estadística ayuda al investigador a concebir su experimento y a evaluar objetivamente los datos numéricos resultantes. Como experimentadores, muchos de nosotros tendremos el tiempo o la tendencia a transformarnos en biometristas competentes; pero todos podemos aprender y practicar las reglas básicas de la experimentación:

1. Repetición. Esta es la única forma en que seremos capaces de medir la validez de nuestras conclusiones en un experimento.
2. Selección aleatoria. El análisis estadístico depende de la asignación de tratamientos a las parcelas en una forma aleatoria, puramente objetiva.
3. Cooperación. Búsquese ayuda cuando existan dudas acerca del diseño, la ejecución o el análisis de un experimento. No se espera que seamos expertos estadísticos, pero debemos saber lo suficiente para entender los importantes principios de la experimentación científica, para estar alertas a los engaños más comunes y solicitar cooperación cuando la necesitemos.

RESUMEN

El razonamiento que parte de un principio general hacia una conclusión específica es un **proceso deductivo.** El razonamiento **inductivo** llega a un principio general a partir de una conclusión particular. Los **experimentos** son conducidos para suministrar hechos específicos a partir de los cuales se establecen las conclusiones generales o principios, contemplando así el razonamiento **inductivo.**

La variabilidad es una característica del material biológico y plantea el problema de decidir si las diferencias entre unidades experimentales se deben a la variabilidad no ponderada o a los efectos reales del tratamiento. La ciencia estadística ayuda a superar esta dificultad, requiriendo el acopio de datos para suministrar estimaciones imparciales de los efectos del tratamiento y para evaluar las diferencias de tratamiento mediante pruebas de significación basadas en mediciones de la variabilidad no ponderada.

Tres importantes principios del diseño experimental son: la **repetición,** la **selección aleatoria** y el **control local.**

El **Método científico** contempla un proceso en flujo desde los hechos observados hasta la hipótesis para la experimentación, la cual suministra más hechos que anularán, ampliarán o alterarán la hipótesis.

Un experimento bien concebido y diseñado deberá ser lo más simple posible, tener grandes posibilidades de alcanzar su objetivo y evitar los errores tendenciosos y sistemáticos. Sus conclusiones deberán poseer un amplio rango de validez, y los datos recabados a partir del mismo deben estar sujetos al análisis a través de procedimientos estadísticos válidos.

El procedimiento para la experimentación contempla: definir un problema, establecer los objetivos, analizar el problema y los objetivos, seleccionar los tratamientos, el material experimental, el diseño experimental, las unidades experimentales y el número de repeticiones, controlar los efectos entre unidades adyacentes, recabar datos y analizar, interpretar e informar sobre los resultados.

Capítulo 2

Algunos conceptos básicos

ALGUNAS DEFINICIONES

En un experimento, la unidad a la que se aplican los tratamientos recibe el nombre de **unidad experimental** o **parcela.** La unidad experimental puede constar de una sola hoja, un árbol completo, un área de terreno que contenga diversas plantas, un solo animal, diversos animales o todo un rebaño.

Una característica medible de una unidad experimental se denomina **variable.** Una variable puede ser **discreta** (discontinua), y tomar sólo valores específicos (por ejemplo, el número de plantas enfermas en cada parcela), o **continua,** en cuyo caso toma cualquier valor entre ciertos límites (por ejemplo, la producción de grano de una parcela de cebada). Las mediciones individuales de una variable reciben el nombre de **elemento.**

En el lenguaje estadístico, una **población** es un conjunto de mediciones o cómputos de una única variable, tomado sobre todos los individuos que se ha especificado pertenecen a la población.

Una población puede ser relativamente pequeña, como la producción de granos por acre de todos los campos de cebada de un área específica en determinado año; o grande, como las estaturas de todos los hombres mayores de 20 años en los Estados Unidos, o las cosechas que resultarían de todas las parcelas posibles de una forma dada que podrían disponerse en un área experimental. Incluso una **pequeña** población suele implicar la medición de un gran número de individuos o unidades experimentales. Podemos tener la población de una variable a partir de unidades experimentales individuales; una población de medias de muestras de la variable, o una población de diferencias entre pares de medias de la muestra.

Una **muestra** es un conjunto de mediciones que constituye parte de una población. A partir de la muestra obtenemos información y hacemos inferencias acerca de una población. Por esta razón, es importante que la muestra sea representativa de la población. Para obtener una muestra representativa utilizamos las técnicas aleatorias de muestreo. Una **muestra aleatoria** es aquella en que cualquier medición individual tiene tantas probabilidades de ser incluida como cualquier otra.

DISTRIBUCIONES DE FRECUENCIA

Las poblaciones se describen mediante características denominadas **parámetros.** Los parámetros son valores fijos; por ejemplo, la media aritmética de todos los elementos de una población es un parámetro. Este sólo

tiene un valor, aunque raramente sepamos cuál es. Las muestras son descritas por las mismas características, pero cuando éstas se aplican a las muestras reciben el nombre de **estadísticos.** La media de una muestra es un estadístico. Calculamos los estadísticos de las muestras para estimar los parámetros de la población. Obviamente, los estadísticos varían de muestra a muestra.

Diferentes valores de una variable presentan distintas **frecuencias** de incidencia en la población. Para describir (caracterizar) convenientemente una población se organizan los datos provenientes de una muestra grande, mediante la construcción de una **tabla de frecuencia**, un **histograma de frecuencia** y un **polígono de frecuencia.** En una tabla de frecuencia (tabla 2.1), los elementos son clasificados de acuerdo con diversos intervalos de clase, en los cuales aquellos encajan. Los totales pueden entonces ser marcados como frecuencias de ocurrencia para cada intervalo de clase y puede construirse un histograma de frecuencia (véase figura 2.1). Conectando los puntos medios de los intervalos de clase obtendremos un polígono de frecuencia.

Si fuésemos a marcar la frecuencia de las cosechas de grano de diversas varas cuadradas de centeno, el porcentaje de crema de la leche de muchas vacas, el aumento de peso de diversos grupos de ovejas, el número de lesiones de escara por papa en un millar de papas o las lecturas refractométricas de diversas cebollas, los gráficos resultantes mostrarían diversas características importantes en común. Todas las curvas podrían ser acampanadas con su punto más alto cercano al medio, representando la clase más común. Estas podrían desviarse bastante simétricamente sobre cualquiera de sus lados hacia las clases raras, excepcionales, en sus dos extremos.

A partir de las distribuciones de frecuencia, las probabilidades pueden calcularse para la ocurrencia de un elemento de cualquier tamaño o rango de tamaño especificado. Quizá la distribución de frecuencia más importante en la teoría y práctica de la estadística sea la **distribución normal.**

LA DISTRIBUCIÓN NORMAL

La mayoría de datos biológicos (y, de hecho, de diversos campos de aplicación), al ser graficados en una curva de frecuencia, se asemejan bastante a una ecuación matemáticamente definida, denominada **curva de frecuencia normal.** En la figura 2.1, una curva de frecuencia normal ha sido superpuesta a un histograma y a un polígono de frecuencia de lecturas refractométricas de cebollas.

Las curvas de distribución normal pueden diferir en cuanto a la posición del punto medio (el punto de mayor frecuencia) y a la dispersión de los datos, pero todas pueden ser descritas mediante sólo dos parámetros: la **media** y la **desviación estándar.** La media es una medida de tendencia central, es decir, describe el punto central alrededor del cual se sitúan los valores de la variable. La desviación estándar es una medida de dispersión (o amplitud o variación). Antes de analizar la distribución normal más ampliamente, consideraremos las medidas de tendencia central y de dispersión.

Medidas de tendencia central

La más común y usualmente la mejor medida de tendencia central es la **media aritmética.** Se utilizarán dos símbolos para representar a la media aritmética (a partir de aquí, la llamaremos solamente **media**): la letra griega μ (Mu) para la media de una población y $\overline{X}$ para la media de una muestra. Mu (μ) es un parámetro (una característica fija que rara vez conocemos) y $\overline{X}$ una variable; ésta varía de muestra a muestra de una misma población. La media es un estadístico.

Tabla 2.1. Tabla de Frecuencia. Lecturas refractométricas en 10 000 cebollas

Intervalos de clase	Punto medio	Tabulación	Frecuencia
6.8– 7.2	7.0	𝍸 𝍸	10
7.3– 7.7	7.5	𝍸 𝍸 𝍸 1111	19
7.8– 8.2	8.0		60
⋮			
10.8–11.2	11.0		1600
11.3–11.7	11.5		1700
⋮			
14.3–14.7	14.5		65
14.8–15.2	15.0		50
15.3–15.7	15.5	𝍸 𝍸 𝍸 𝍸 𝍸	25
15.8–16.2	16.0	𝍸 𝍸 𝍸 𝍸	20
16.3–16.7	16.5	𝍸 𝍸 11	12

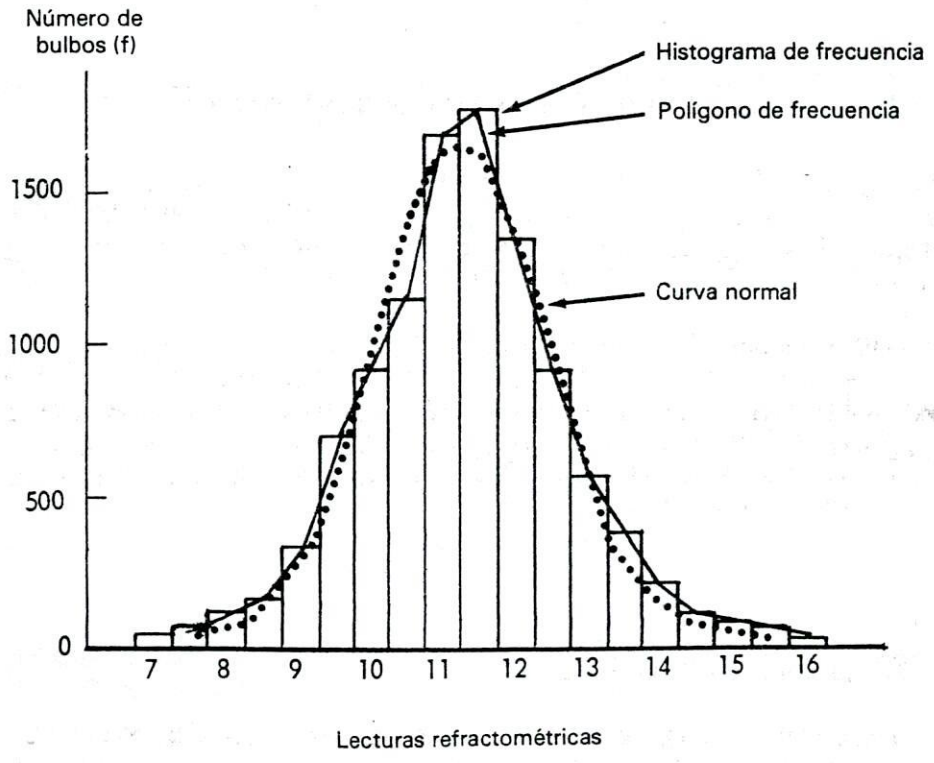


Figura 2.1. Distribución de frecuencia de lecturas refractométricas de 10 000 cebollas, y la curva teórica de la distribución normal.

La media de una población se define como:

$$\mu = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N}$$

donde X_1, X_2, etc., son los elementos de la población y N es el número de elementos en la población. Así X_N es el elemento N-ésimo de la población.

La media (μ) puede definirse mediante una notación abreviada, denominada **notación de sumatoria.**

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

En esta forma abreviada, la letra mayúscula griega Σ (sigma) indica que debemos sumar todos los valores de X_i. Los índices de la sumatoria $i = 1 \ldots N$, indican que los valores de X_i van desde el valor de X_1 hasta el de X_N.

Puesto que rara vez conocemos el valor de μ lo estimamos a partir de la media de una muestra $\bar{X}$, la cual se define como sigue:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

donde la n del denominador representa el número de elementos en la muestra. Cuando resulta evidente qué valores de X han de sumarse, la notación suele reducirse a ΣX_i o incluso a ΣX.

Frecuentemente deseamos representar la diferencia entre un elemento (X) y una media ($\bar{X}$). Tales desviaciones se representan a menudo con la letra minúscula en tipo cursivo x, o y. $x_i = X_i - \bar{X}$ y $y = Y_i - \bar{Y}$. También se utilizan Y y $\bar{Y}$ para representar un elemento y la media de una muestra, respectivamente.

Una propiedad de la media consiste en que las sumas de sus desviaciones son iguales a cero; por ejemplo, en la tabla 2.2:

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{(3 + 4 + \ldots + 1)}{5} = \frac{15}{5} = 3 \text{ g/planta, y}$$

$$\Sigma (X_i - \bar{X}) = \Sigma x_i = (3 - 3) + (4 - 3) + \ldots + (1 - 3) = 0.$$

Por otro lado, la suma de los cuadrados de las desviaciones de la media es mínima; es decir, la suma de los cuadrados de las desviaciones de cualquier otro valor resultará una larga suma de cuadrados.

Otras medidas de tendencia central que no utilizaremos en esta breve obra son: la mediana, o sea el valor situado en el centro de los elementos cuando éstos son agrupados en orden de magnitud —si el número de elementos es par, la mediana será el promedio de los valores centrales— y la moda, el valor de la ocurrencia más frecuente. En una distribución normal, la media, la mediana y la moda son iguales.

Tabla 2.2. Peso seco de cinco plantas.

g./planta X	$(X - \bar{X})$	$(X - \bar{X})^2$
3	0	0
4	1	1
5	2	4
2	−1	1
1	−2	4
$\Sigma X = 15,$	$\bar{X} = 3,$	$\Sigma(X - \bar{X})^2 = 10$

$\bar{X} = 3$
$S^2 = 2,5$
$S = 1.581$
$CV = 52.704$
$S_{\bar{x}} = \sqrt{\frac{2,5}{5}} = 0.707$

$Lc = \bar{x} \pm t\, s_{\bar{x}}$

$Lc = 0.5 = 3 + 2.776\ (0.707) =$
$Lc = 4.962 =$

$Lc_{0.5} = 3 - 2.776\ (0.707)$
$Lc = 1.037.$

Medidas de dispersión

La medida de dispersión más común, y la mejor para la mayoría de los propósitos, es la **varianza** o su raíz cuadrada, la **desviación estándar**. También aquí utilizaremos dos símbolos para representarlas: σ^2 para la varianza de una población y s^2 para la mejor estimación de σ^2 que puede ser obtenida a partir de una muestra. Las raíces cuadradas respectivas, σ y s, representan la desviación estándar de la población y su estimador.

La varianza de la población se define como:

$$\sigma^2 = \frac{\Sigma (X_i - \mu)^2}{N},$$

donde N es el número de elementos en la población. La mejor estimación de σ^2 a partir de una muestra pequeña (donde n es menor que 60), se define como sigue:

$$s^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1},$$ donde n es el número de elementos en la muestra.

Si conocemos el valor de μ, la mejor estimación de σ^2 a partir de una muestra es:

$$s^2 = \frac{\Sigma(X_i - \mu)^2}{n},$$ siendo n el número de elementos en la muestra.

Sin embargo, rara vez, si es que ello es posible, conocemos el valor de μ, de modo que en el numerador la sustituimos por su estimador $\bar{X}$. Ahora bien, mientras que $\bar{X}$ es el promedio igual a μ, varía de muestra a muestra y rara vez es exactamente igual a μ. Vimos en párrafos anteriores que $\Sigma(X_i - \bar{X})^2$ es menor que la suma de cuadrados de las desviaciones de cualquier otro valor diferente de $\bar{X}$. Por tanto, si $\bar{X}$ no es exactamente igual a μ, $\Sigma(X_i - \bar{X})^2$ es menor que $\Sigma(X_i - \mu)^2$.

Esto significa que $\frac{\Sigma(X_i - \bar{X})^2}{n}$ dará una estimación demasiado pequeña de σ^2. De ahí resulta que la corrección apropiada puede ser hecha mediante la utilización en el denominador de $n-1$, en vez de n. Es decir, en promedio:

$$\frac{\Sigma(X_i - \bar{X})^2}{n-1} = \frac{\Sigma(X_i - \mu)^2}{n} \cong \sigma^2.$$

El numerador, $\Sigma(X_i - \bar{X})^2$, es una **suma de cuadrados;** en este caso la suma de los cuadrados de las desviaciones de elementos individuales de sus medias.

Utilizaremos la pequeña muestra de la tabla 2.2 para ilustrar el cálculo de s^2 y s:

$$s^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1} = \frac{(3-3)^2 + (4-3)^2 + (5-3)^2 + (2-3)^2 + (1-3)^2}{5-1}$$

$$= \frac{(0)^2 + (1)^2 + (2)^2 + (-1)^2 + (-2)^2}{4} = \frac{0+1+4+1+4}{4} = \frac{10}{4} = 2.5$$

$$s = \sqrt{2.5} = 1.58 \text{ g / planta.}$$

Para muestras pequeñas sin decimales, en las que la media suele ser un número entero, s^2 y s pueden calcularse fácilmente a partir de la fórmula de la definición; pero para muestras grandes se cuenta con un método abreviado que resulta de más sencilla aplicación, especialmente si se utiliza una calculadora. Es posible demostrar que

$\Sigma(X - \bar{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$; por tanto, una fórmula de trabajo conveniente para s^2 es:

$$s^2 = \frac{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}}{n-1}.$$

El término del extremo derecho en el numerador recibe el nombre de **término de corrección o factor de corrección,** y lo representaremos en esta obra como C. $C = \frac{(\Sigma X_i)^2}{n}$. A la expresión del denominador, (n--1), se le llama grados de libertad (denotados gl) sobre los cuales está basada la varianza; en este caso, uno menos que el número de elementos en la muestra.

Aplicando esta fórmula a los datos de la tabla 2.2, obtenemos:

$$s^2 = \frac{(3)^2 + (4)^2 + (5)^2 + (2)^2 + (1)^2 - \frac{(3+4+5+2+1)^2}{5}}{5-1} = \frac{55 - \frac{(15)^2}{5}}{4}$$

$$= \frac{55-45}{4} = \frac{10}{4} = 2.5, \text{ como antes.}$$

Otras medidas de dispersión son el rango y la desviación media; pero éstas no serán estudiadas aquí, dada la utilidad bastante mayor de s^2 y s.

La variabilidad entre las unidades experimentales de experimentos que involucran diferentes unidades de medida y tamaños de parcelas, puede ser comparada a través de **coeficientes de variabilidad.** El coeficiente de variabilidad expresa la desviación estándar por unidad experimental, como un porcentaje de la media general del experimento.

$CV = \frac{s}{\bar{X}}(100)$. Por ejemplo: si en un experimento con remolachas, la media de todas las parcelas fue de 30.5 toneladas por acre y la desviación estándar por parcela fue de 1.18 toneladas por acre,

$$CV = \frac{1.18}{30.5}(100) = 3.9\%.$$

En un experimento que involucra tratamientos de semillas de frijol de media luna, la media general es de 82.7 retoños por parcela, y $s = 5.8$ retoños por parcela; entonces

$$CV = \frac{5.8}{82.7}(100) = 7.0\%.$$

La comparación entre los dos CV muestra que existió 1.8 veces $\left(\frac{7.0}{3.9}\right)$ más variabilidad entre las parcelas dentro de un tratamiento del experimento con el frijol de media luna.

Características de la distribución normal

Las distribuciones normales sólo varían entre sí con respecto a la media y/o la desviación estándar. La media determina la posición de una curva sobre el eje horizontal (abscisa). La desviación estándar determina el grado de amplitud o dispersión entre los elementos. La figura 2.2a muestra dos distribuciones normales con idénticas desviaciones estándar, pero con medias distintas. Las dos distribuciones normales de la figura 2.2b contienen idénticas medias, pero tienen diferentes desviaciones estándar.

Las áreas bajo las curvas, limitadas por cualquier rango dado de valores sobre el eje de las X, corresponden al porcentaje de elementos de la población que caen dentro del rango designado de valores de X; por ejemplo: el rango de valores de la media, más o menos una unidad de desviación estándar, contiene 68.27% de todos los elementos de la población. Un intervalo de $\mu \pm 1.96\sigma$ contiene el 95% de los elementos, y $\mu \pm 2.58\sigma$ contiene el 99% de todos los elementos. Así pues, suponiendo que la curva normal de la figura 2.1 sea la verdadera distribución de los datos relativos a las cebollas, que $\mu = 11.25$ y $\sigma = 1.40$ si podemos tomar aleatoriamente una cebolla de dicha población habría un 95% de probabilidades de que su lectura refractométrica se encontrara entre $\mu \pm 1.96$ (1.40), es decir, entre 8.51 y 13.99. Recíprocamente, habría un 5% de probabilidades de que su valor fuera menor que 8.51 o mayor que 13.99.

Afortunadamente, para determinar las probabilidades no es necesario construir una curva de frecuencia normal para cada conjunto de datos. Cualquier curva normal puede ser convertida en una curva de **probabilidad estándar normal** (figura 2.3), cambiando el eje de las Y en eje de probabilidades, expresando cada frecuencia como una fracción decimal de N, el número total de observaciones, y cambiando el eje de las X en eje de las verdaderas unidades de medida para la desviación estándar denominada z. Un valor z de cualquier valor de X se calcula mediante la sustracción de la media (μ) y la división del resultado entre la desviación estándar (σ). Por tanto, $z = \frac{(X - \mu)}{\sigma}$. Cuando $X = \mu$, $z = 0$; y cuando $X - \mu = \sigma$, $z = 1$. Así pues, el eje de las X de la curva de probabilidad estándar normal se encuentra en términos de unidades z con $\mu = 0$ y $\sigma = 1$. El área total bajo la curva es igual a 1 y el área bajo la curva entre dos valores dados de z es igual al porcentaje de la población que se encuentra dentro de los valores prescritos de z. En la elección aleatoria de un elemento de una distribución normal y el cálculo de su valor z, podríamos esperar obtener un valor z de

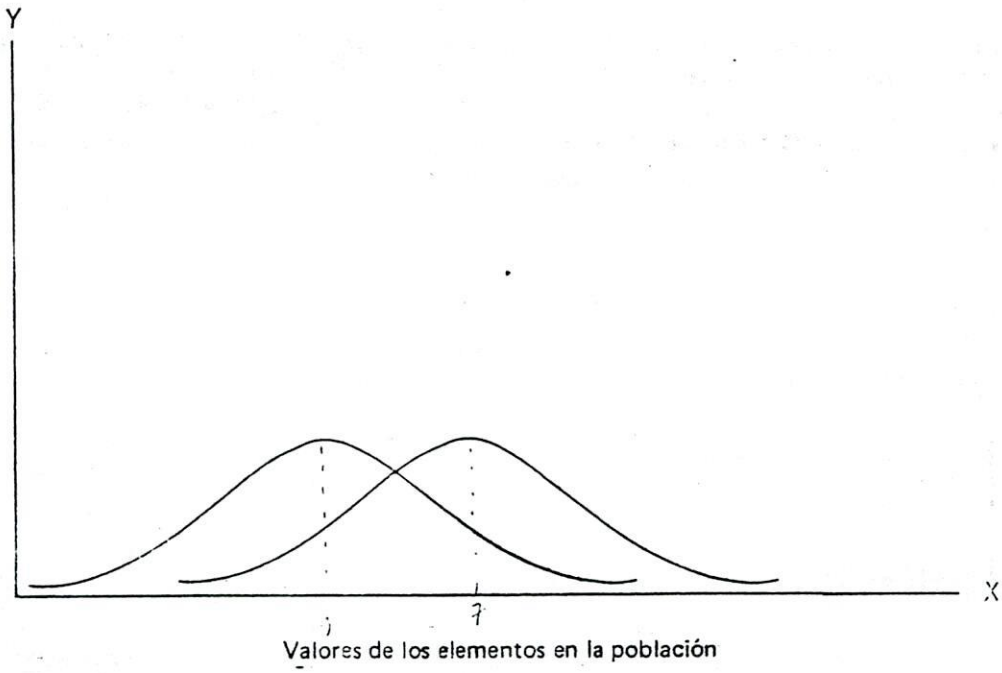


Figura 2.2a. Distribuciones normales — desviaciones estándar iguales, medias diferentes.

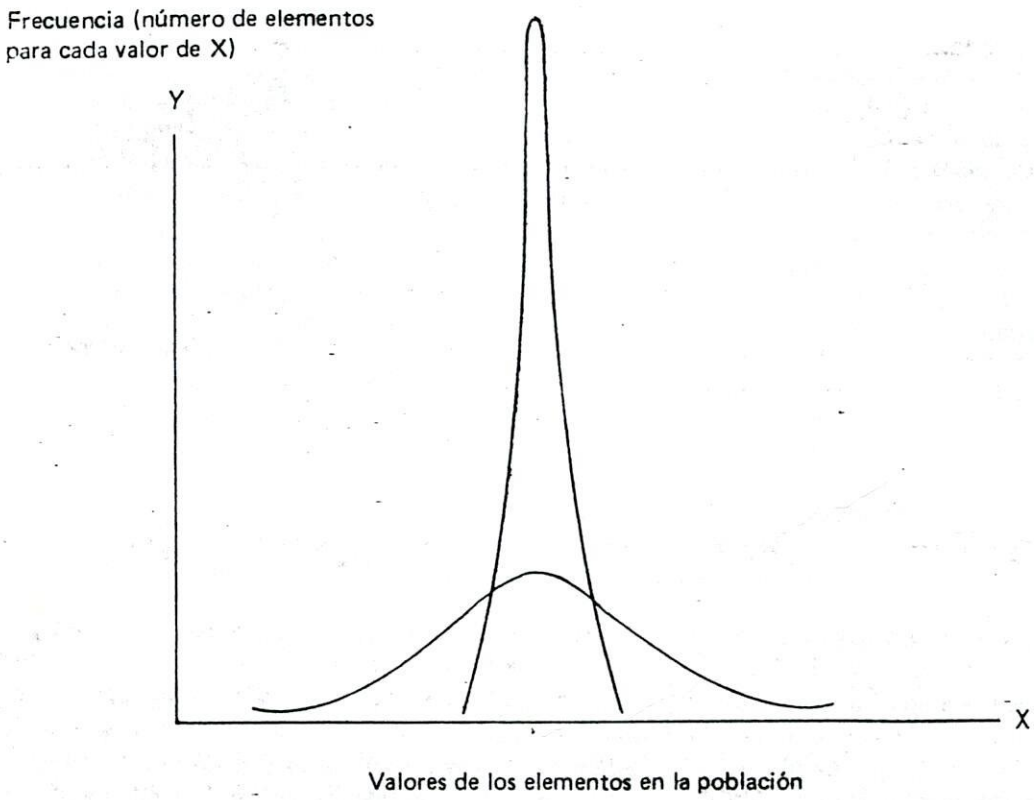


Figura 2.2b. Distribuciones normales — medias iguales, desviaciones estándar diferentes.

1.96 o mayor en sólo el 2.5% de las veces, y un valor de − 1.96 o menor en sólo el 2.5% de los casos. En el 95% de las veces en que un elemento es sacado, su valor z se encontrará dentro del intervalo ± 1.96 (figura 2.3)

En la mayoría de los libros destinados al estudio de la estadística se incluye una tabla de áreas bajo la curva normal de probabilidad, correspondientes a valores de z. Dicha tabla no se incluye en este texto debido a que, como se verá un poco más adelante, otra distribución, la t, resulta más apropiada para las pequeñas muestras generalmente utilizadas en la investigación agrícola.

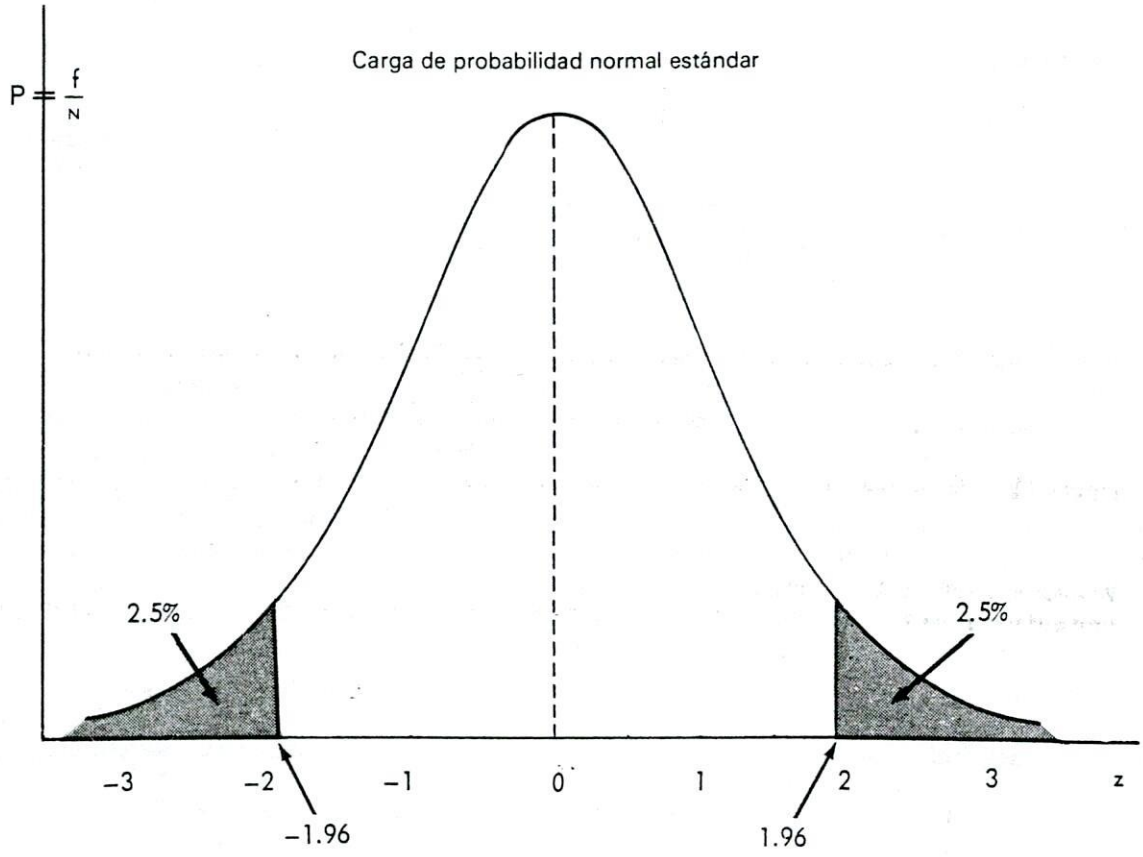
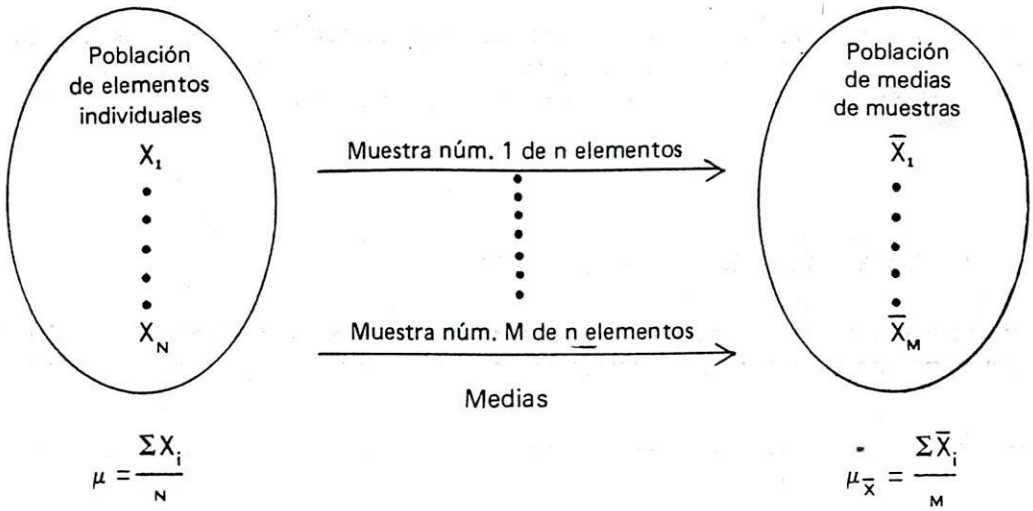


Figura 2.3. La curva de probabilidad normal estándar con valores z, que incluyen el 95% del área. Los elementos de cualquier distribución normal se convierten en valores z mediante: $z = \frac{x - \mu}{\sigma}$. La media de los valores z es igual a cero y $\sigma = 1$.

MUESTREO A PARTIR DE UNA DISTRIBUCIÓN NORMAL

El tipo de muestra con la cual usualmente trabajaremos consiste en una serie de parcelas o animales sometidos a cierto tratamiento. El efecto del tratamiento se estima mediante el cálculo de la media de la muestra ($\bar{X}$). Sabemos que las repeticiones del experimento (en efecto, obteniendo otras muestras) arrojarán una serie de medias diferentes. Esto plantea entonces el problema de saber qué tan bien representado es el verdadero efecto del tratamiento por una única media. Un enfoque a este problema consiste en calcular **límites de confianza**, un rango de valores, dentro de los cuales estará situada la verdadera media del efecto



Medias

$$\mu = \frac{\Sigma X_i}{N} \qquad \mu_{\bar{X}} = \frac{\Sigma \bar{X}_i}{M}$$

$$\mu = \mu_{\bar{X}}$$

Varianzas

$$\sigma^2 = \frac{\Sigma (X_i - \mu)^2}{N} \qquad \sigma^2_{\bar{X}} = \frac{\Sigma (\bar{X}_i - \mu)^2}{M}$$

$$\sigma^2 = n\,\sigma^2_{\bar{X}}$$

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

Figura 2.4. Una población de elementos individuales y una población de medias generadas por muestreos sucesivos, así como sus medidas, varianzas y relaciones importantes.

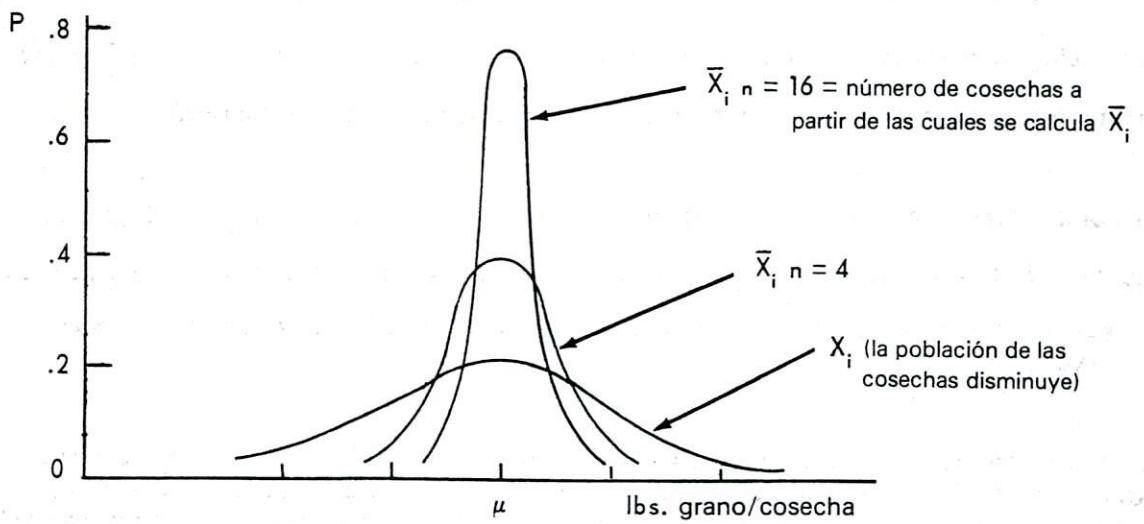


Figura 2.5. Distribuciónes de frecuencia de poblaciones de medias, con variaciones en el tamaño de las muestras, generadas por muestreos sucesivos tomados de la misma población normalmente distribuida de la producción de cosechas de grano. Las distribuciones (todas normales) se van haciendo más angostas y altas conforme aumenta el tamaño de la muestra de acuerdo con la relación $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$.

Si deseamos tener un 95% de confianza en que los límites incluirán a μ, utilizamos $z = \pm 1.96$, puesto que un valor z de esta magnitud o de mayor tamaño sólo sería obtenido en un 5% de las veces en que una muestra fuese sacada.

La distribución t y los límites de confianza

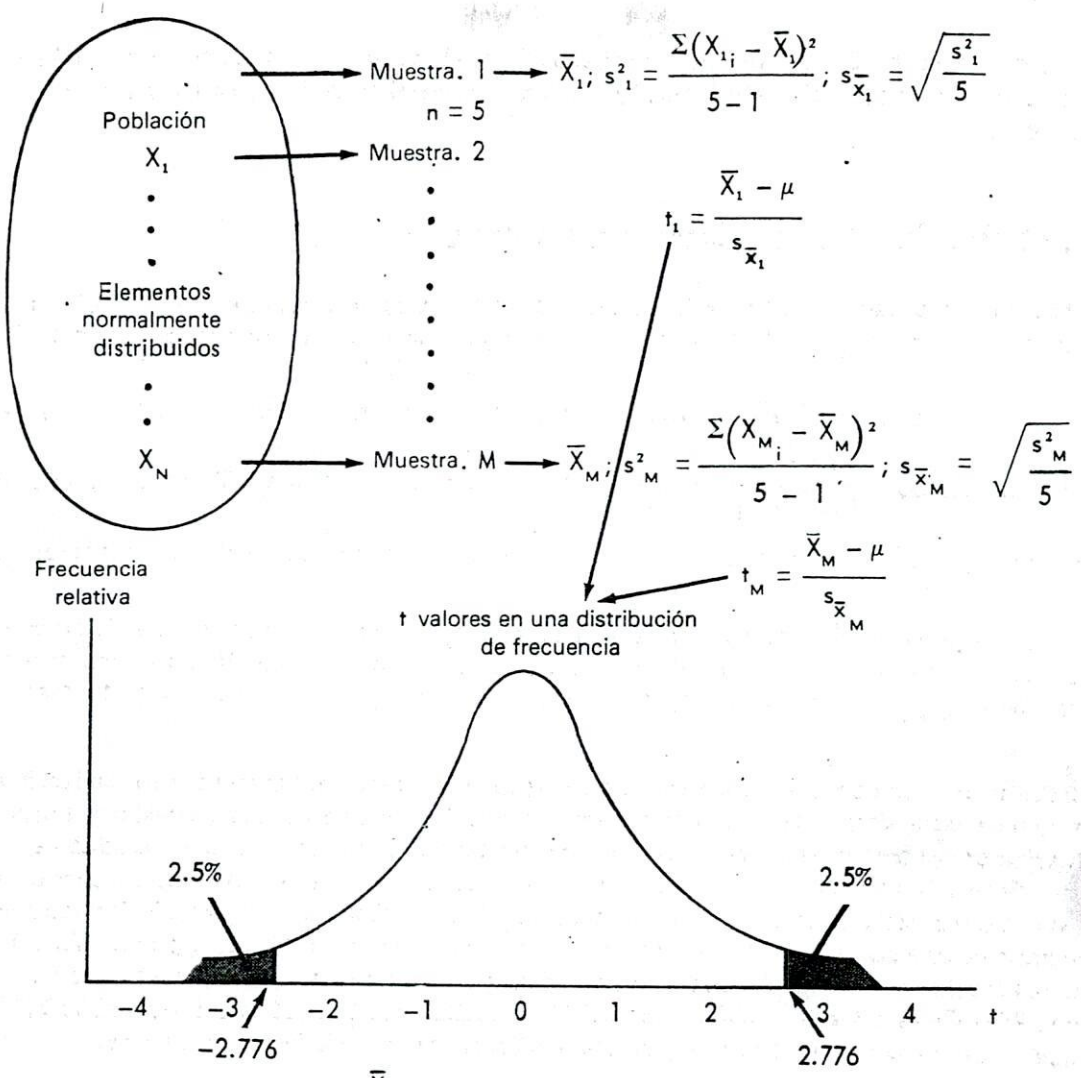
En el muestreo esbozado en la figura 2.4, estábamos interesados en la distribución de medias de muestras que contenían una media $(\mu_{\bar{x}})$ igual a μ y una varianza $(\sigma^2_{\bar{x}})$ igual a $\frac{\sigma^2}{n}$. Finalmente, convertimos dichas medias en valores z a través de la relación $z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}}$. Considérese ahora otra extracción repetida de muestras de tamaño n de una población de elementos normalmente distribuidos (figura 2.6), y el cálculo de un nuevo estadístico: $t = \frac{\bar{X} - \mu}{s_{\bar{x}}}$. Nótese que éste difiere de z sólo en el denominador. Para encontrar los valores z, dividimos la expresión del numerador entre el parámetro $\sigma_{\bar{x}}$ de la población; pero para determinar los valores t, la dividimos entre $s_{\bar{x}}$ una estimación de $\sigma_{\bar{x}}$ calculada para cada muestra. Para cada muestra de la figura 2.6, calcúlese $\bar{X}$; s^2; $s_{\bar{x}}$; y t. Luego, organícense los valores t en una distribución de frecuencia.

Puesto que $\sigma_{\bar{x}}$ es una constante, la variación de los valores z depende solamente de la variación de $\bar{X}$ de muestra a muestra, y se encuentra normalmente distribuida. Por otro lado, la variación en t depende tanto de la variación de $\bar{X}$ como de la variación de $s_{\bar{x}}$ de muestra a muestra. Por ende, el estadístico t es más variable que z, y tiene una distribución que no es normal, pero presenta pocos valores cerca del centro y un mayor número de éstos hacia los extremos de la distribución. Esto se ilustra en la figura 2.7. Debido a esta diferencia en la forma de una distribución t en comparación con una distribución normal, el punto más allá del cual el 2.5% de los valores de t estará situado, se encuentra más apartado de la media que en una distribución normal. En la figura 2.6, donde n = 5, el 5% de los valores de t serán iguales o mayores que ± 2.776. Para cualquier muestra aleatoria, se pueden calcular límites de confianza (LC) dentro de los cuales μ caerá con una confianza específica. Esto se hace mediante la solución de $\pm t = \frac{\bar{X} - \mu}{s_{\bar{x}}}$ para μ y denominando a los dos valores resultantes límites de confianza: $LC = \bar{X} \pm t\, s_{\bar{x}}$. Si deseamos tener un 95% de confianza en que LC contendrá a μ multiplicamos $s_{\bar{x}}$ por un valor tabular de t, dependiendo de los grados de libertad n – 1 y del nivel de probabilidad de 5% (tabla A 2). Para una muestra donde n = 5, $s_{\bar{x}}$ se multiplica por 2.776.

A modo de ilustración, considérese la muestra de la tabla 2.2, donde $n = 5$, $\bar{X} = 3$ y $s^2 = 2.5$. Entonces

$$s_{\bar{x}} = \sqrt{\frac{2.5}{5}} = 0.707 \text{ y } LC_{95} = 3 \pm 2.776(0.707) = 4.96 \text{ para 1.04 g/planta.}$$

En consecuencia, podemos decir con un 95% de confianza, que μ se encuentra dentro de este rango. Es incorrecto afirmar que la **probabilidad** de que μ se halle dentro de estos límites de confianza es del 95%, puesto que, basado en los estadísticos de la muestra en particular, μ **se** encontrará o **no se** encontrará en el intervalo calculado. Podríamos haber extraído una muestra cuyos $\bar{X}$ y/o s^2 se desviaran suficientemente de μ y/o σ^2 de modo que LC_{95} no contuviera a μ; sin embargo, la probabilidad de extraer tal muestra es de sólo un 5%.



Límites de confianza. Resuélvase $\pm t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$ para los límites de confianza de μ : $LC = \bar{X} \pm t\, s_{\bar{X}}$.

Ejemplo: Tabla 2.2 donde $n = 5$, $\bar{X} = 3$, $s^2 = 2.5$.

$$s_{\bar{X}} = \sqrt{\frac{2.5}{5}} = 0.707 \text{ g/planta}$$

$$LC_{95} = \bar{X} \pm 2.776\,(0.707) = 4.96 \text{ a } 1.04 \text{ g/planta.}$$

Figura 2.6. Generación de la distribución t para una muestra de medida 5 y cálculo de los límites de confianza (véase texto).

Para cada tamaño de muestra existe una distribución t única. Cuanto mayor es el tamaño de la muestra, mayor será la aproximación de t al valor z normalmente distribuido. En el último renglón de la mayoría de las tablas de valores de t para grados infinitos de libertad, $t = z$ (tabla A 2).

Si la muestra que nos ocupa es grande, o sea, $n > 60$, entonces $s_{\bar{X}}$ suministra una estimación suficientemente buena de $\sigma_{\bar{X}}$ de modo que $\frac{\bar{X} - \mu}{s_{\bar{X}}}$ se aproxima a una distribución normal y brinda una estimación apropiada de z. Entonces $LC = \bar{X} \pm z\, s_{\bar{X}}$. Puesto que en la mayoría de las investigaciones agrícolas las repeticiones (n) son, por regla general, inferiores a 60, los valores z rara vez se utilizan.

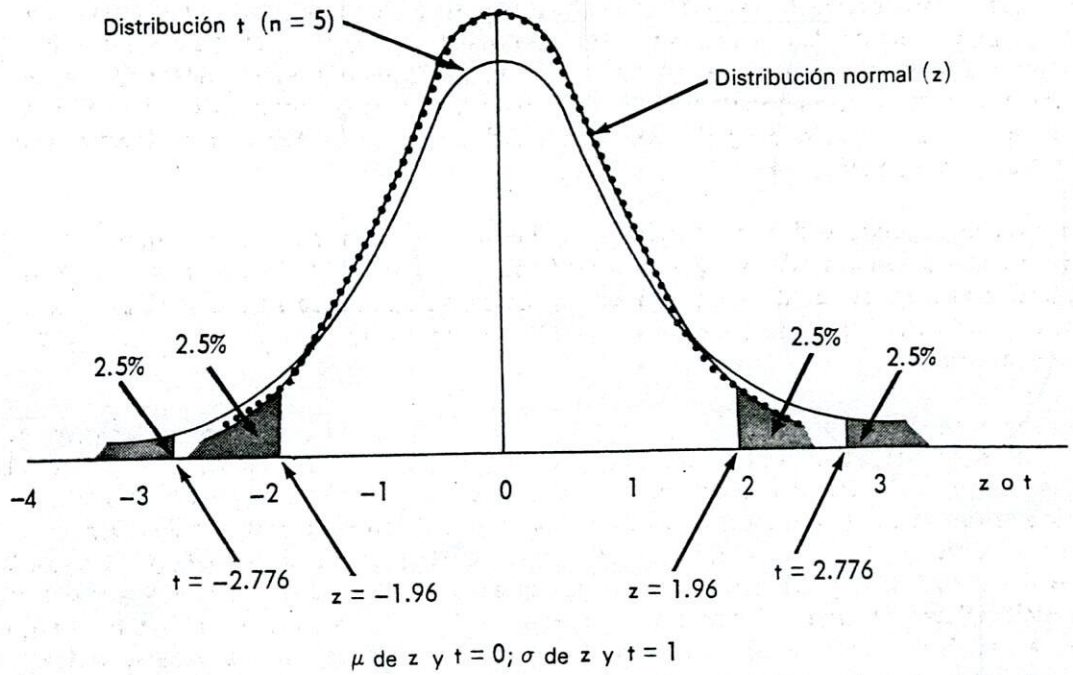


Figura 2.7. Distribución de z comparada con la distribución t con base en una muestra de medida 5. Conforme aumenta la medida de la muestra la distribución t se aproxima a la distribución normal z. (Valores de t y z los cuales excluyen el 5% del área indicada debajo de cada curva.)

HIPÓTESIS ESTADÍSTICAS Y PRUEBAS DE SIGNIFICACIÓN

Una hipótesis estadística es un supuesto referente a algún parámetro. La hipótesis nula se utiliza frecuentemente. Para determinar si un tratamiento tuvo algún efecto, consideraremos a la **hipótesis nula** como la ausencia de efectos. Procedemos entonces a verificar la probabilidad de que las medias, tan divergentes como aquellas de nuestras muestras, pudieran ocurrir sólo por casualidad, si éstas fueran en realidad muestras aleatorias de poblaciones normalmente distribuidas con medias y varianzas iguales. Si nuestro análisis conduce a la conclusión de que podríamos esperar tales diferencias de medias bastante frecuentemente por casualidad, no rechazamos la hipótesis nula y concluimos que no tenemos buenas evidencias de un efecto real del tratamiento. Si el análisis indica que las diferencias observadas rara vez podrían ocurrir en muestras aleatorias extraídas de poblaciones con medias y varianzas iguales, rechazamos la hipótesis nula y concluimos que al menos un tratamiento tiene efectos reales. Se dice que al menos una de las medias es **significativamente** diferente de las otras.

Si la probabilidad de que la variación observada entre medias, que pudiera ocurrir por casualidad, es de un 5% o menor, decimos que las medias son **significativamente diferentes.** Si la probabilidad de que la variación observada entre medias la cual pudiese esperarse que ocurriera por casualidad, es de 1% o menor, decimos que las diferencias son **altamente significativas.**

El hecho de que la hipótesis nula no sea rechazada y que concluyamos que no existen diferencias significativas entre las medias, no prueba que alguno de los tratamientos no produjo efectos. Siempre hay una probabilidad definida de que existió un efecto real, pero que el experimento fue demasiado insensible para detectar la diferencia en el nivel de probabilidad deseado.

A estas alturas debemos entender que nada existe de mágico acerca del nivel de significación del 5%. Las conclusiones que hagamos respecto a un experimento son nuestras propias conclusiones, no las del estadístico, y deben estar basadas en algo más que en la evidencia estadística. La lógica de las conclusiones debe considerarse a la luz de lo que ya se conoce sobre el sujeto. Debemos evitar aceptar de inmediato un resultado **significativo** si éste no tiene sentido a la luz de otros hechos conocidos. Siempre existe la posibilidad de que nuestro resultado significativo haya ocurrido por casualidad y de que hayamos cometido un error al rechazar la hipótesis nula.

Considérense las consecuencias de una equivocación. Si éstas son serias, como el estar equivocados al recomendar un cambio que requeriría un gasto considerable para una utilidad incrementada relativamente pequeña, podemos vacilar en cuanto al rechazo de la hipótesis nula con base en una única prueba, aun cuando los resultados sean significativos en el nivel del 5%. En tal situación, pruebas adicionales son claramente necesarias.

Por otro lado, si las consecuencias de una equivocación no son serias, podemos rechazar la hipótesis nula, aun cuando el análisis estadístico indicara que podríamos esperar tal resultado por casualidad, con una frecuencia de una entre 15 o incluso de una entre 10 oportunidades; por ejemplo, considérese la verificación de un nuevo y barato tratamiento de semillas, donde el análisis combinado de diversos experimentos de campo resulta justamente insuficiente para ser significativo en el nivel del 5%. Además, supóngase que los resultados de diversos experimentos en invernaderos indicaron que el nuevo tratamiento suministró significativamente una mejor protección contra los agentes patógenos más importantes que atacan a los retoños del cultivo en cuestión. En tal situación, estaremos justificados para rechazar la hipótesis nula, incluso hasta el punto de recomendar la práctica a los agricultores, mientras procedemos a verificar más ampliamente nuestras conclusiones en experimentos de campo adicionales.

La distribución F

Una prueba F es una razón entre dos varianzas y se utiliza para determinar si dos estimaciones de varianza independientes pueden ser admitidas como estimaciones de la misma varianza. Esta razón fue denominada F por George W. Snedecor, en honor del fallecido Ronald A. Fisher, pionero de la utilización de las estadísticas matemáticas en la agricultura. En el análisis de varianza, la prueba F se utiliza para verificar la igualdad de medias, o sea, para responder a la pregunta: ¿es razonablemente posible admitir que las medias del tratamiento resultaron del muestreo de poblaciones con medias iguales? Esto puede ilustrarse con una descripción de cómo una porción de la tabla de valores de F podría ser determinada. Considérese lo siguiente: de una población normalmente distribuida (figura 2.8), extráiganse 5 muestras ($m = 5$) de un número específico de elementos; 9 por ejemplo ($n = 9$). Calcúlense las medias de estas 5 muestras ($\overline{X}_1 \ldots \overline{X}_5$). Estímese σ^2 mediante el cálculo de s^2 **para cada muestra:**

$s^2_1 = \sum_{i=1}^{9} (X_{1i} - \overline{X}_1)^2/(9-1)$; etc., para $s^2_2 \ldots s^2_5$. Súmense estas estimaciones de σ^2 para obtener una estimación promedio (combinada): $s^2 = (s^2_1 + \ldots s^2_5)/5$. Estímese ahora la varianza de las medias $(\sigma^2_{\overline{x}})$ de las **medias** de las 5 muestras:

$$s^2_{\overline{x}} = \sum_{i=1}^{5} \left(\overline{X}_i - \overline{\overline{X}}\right)^2/(5-1).$$

A partir de $s^2_{\overline{x}}$ estímese nuevamente σ^2, utilizando la relación $s^2 = n\, s^2_{\overline{x}}$.

Calcúlese la razón de varianza F, donde

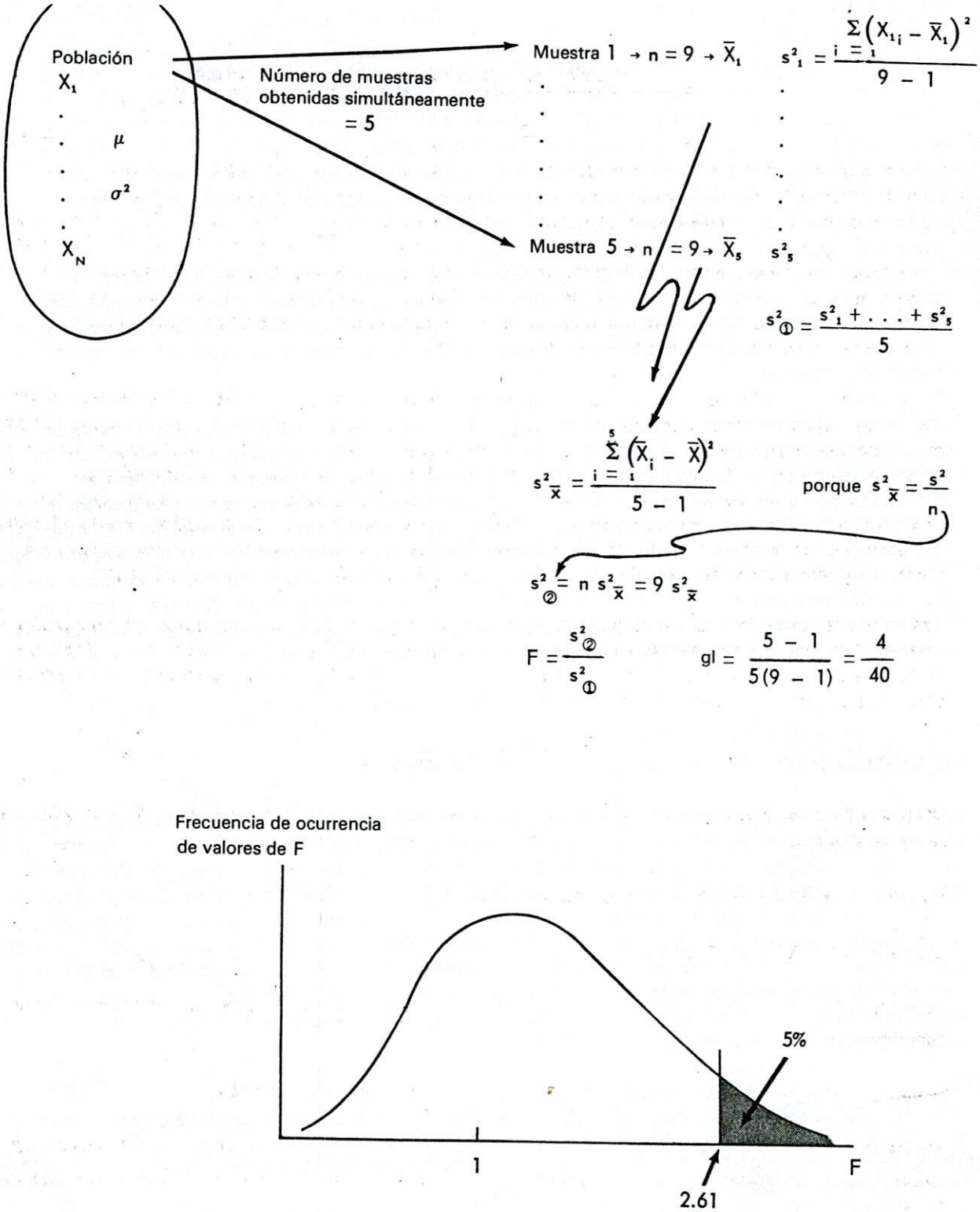


Figura 2.8. La obtención sucesiva de 5 muestras de n = 9, a partir de una población de elementos ($X_1 \ldots X_n$) normalmente distribuida, genera una distribución F. Cinco por ciento de los valores de F serán 2.61 o mayores (véase texto).

$$F = \frac{s^2, \text{ calculado a partir de las medias de la muestra}}{s^2, \text{ calculado por la combinación de las varianzas de la muestra}}.$$

Los grados de libertad para el numerador son $m - 1 = 4$ (donde m es el número de muestras) y para el denominador $m(n - 1) = 5(8) = 40$ (donde n es el número de elementos en cada muestra). Supóngase ahora que este procedimiento de muestreo se repite hasta que todos los conjuntos posibles de muestras hayan sido extraídos y registrados, que las frecuencias para obtener valores de F de diversos tamaños se hayan registrado y que la curva de frecuencia se haya trazado. F 2.61 es el valor más allá del cual se encuentra el 5% de los valores calculados. Este es el valor para el nivel de 5% encontrado en una tabla de F para los grados de libertad 4 y 40 (tabla A3). Análogamente, los valores F pueden determinarse para otros tamaños de muestra, números de muestras y para otros niveles de probabilidad (2.5%, 1%, etc).

Puesto que ambas varianzas de la razón F son estimaciones de la misma varianza (σ^2), ésta se acercará a 1, a menos que se haya extraído un conjunto poco usual de muestras. La distribución F para el tamaño de muestra que estamos considerando ($m = 5$, $n = 9$) se parecerá al gráfico de la figura 2.8. El área bajo la curva representa la frecuencia de obtención de cualquier valor dado de F. Para cualquier extracción específica de un conjunto de muestras de $m = 5$ y $n = 9$ las probabilidades de que el valor calculado de F sea igual o mayor que 2.61 son de un 5%. Por otro lado, existe un 95% de probabilidades de que cualquier extracción dada de tal conjunto de muestras producirá un valor F menor que 2.61. Nótese que la prueba F es una prueba que persigue la unidad; o sea, no estamos interesados en la probabilidad de que F sea igual a algún valor menor que 1.

Los anteriores experimentos hipotéticos de muestreo están destinados a mostrar cómo las distribuciones t y F pueden obtenerse mediante el muestreo de una población de elementos normalmente distribuidos. Las tablas de valores de t y F no se determinan mediante estos laboriosos procedimientos de muestreo, sino que se calculan a partir de precisas y complicadas relaciones matemáticas.

RESUMEN

Unidad experimental (o parcela, para un área de terreno en el campo). Unidad de material experimental sobre la cual se aplica un tratamiento.

Variable. Característica medible de una unidad experimental.

Elemento. Medición específica de una variable.

Población. Conjunto de mediciones de una variable, tomadas sobre todos los individuos que se encuentran en la población.

Muestra. Conjunto de mediciones (elementos) que constituye una parte de la población.

Parámetro. Una característica de la población (la media, por ejemplo). Un parámetro es un valor fijo que rara vez conocemos. Los parámetros son estimados a partir de las muestras. Por regla general, se representan con letras griegas (μ, σ), etc.

Estadístico. Una característica de una muestra; se utiliza con frecuencia para estimar un parámetro. Generalmente se representa con las letras ($\bar{X}$, s), etc.

Distribución normal. Curva matemáticamente definida, en forma de campana, resultante de la marcación de las frecuencias de ocurrencia de valores de un elemento (eje de las Y) contra el rango de los valores del elemento (eje de las X). Una distribución normal se describe únicamente por su media y su desviación estándar.

La media de una población de elementos individuales, μ.

$$\mu = \frac{\Sigma X_i}{N}, \text{ donde N es el número de individuos en la población.}$$

La estimación de μ a partir de una muestra, $\bar{X}$.

$$\bar{X} = \frac{\Sigma X_i}{n}, \text{ donde n es el número de individuos en la muestra.}$$

La varianza de una población de elementos individuales, σ^2.

$$\sigma^2 = \frac{\Sigma(X_i - \mu)^2}{N}$$

La desviación estándar de una población de elementos individuales, σ.

$$\sigma = \sqrt{\sigma^2}$$

La estimación de σ^2 a partir de una muestra, s^2.

$$s^2 = \frac{\Sigma(X_i - \bar{X})^2}{n - 1} \text{ (fórmula de definición)} \qquad s^2 = \frac{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}}{n - 1} \text{ (fórmula de trabajo)}$$

Término de corrección utilizado en la fórmula de trabajo C.

$$F.C = \frac{(\Sigma X_i)^2}{n}$$

Estimación de σ a partir de una muestra s.

$$s = \sqrt{s^2}$$

Coeficiente de variacion, CV.

$$CV = \frac{s}{\bar{X}}(100)$$

Curva normal estándar. Curva de frecuencia normal en la cual las frecuencias sobre el eje de las Y se encuentran en términos de proporciones para el número total de observaciones, y la escala sobre el eje de las X está dada en términos de las desviaciones a partir de la media en número de desviaciones estándar denominada escala z. $z = \frac{X - \mu}{\sigma}$

Una población de medias. Población de todas las medias posibles $(\bar{X})$ de un tamaño de muestra específico (n), extraída de una población de individuos.

La media de una población de medias $\mu_{\bar{x}}$.

$$\mu_{\bar{x}} = \frac{\Sigma \bar{X}_i}{M} = \mu, \text{ donde M es el número de medias de la muestra.}$$

La varianza de una población de medias, $\sigma^2_{\bar{x}}$.

$$\sigma^2_{\bar{x}} = \frac{\Sigma\left(\bar{X}_i - \mu\right)^2}{M}$$

La desviación estándar de una población de medias, o error estándar, $\sigma_{\bar{x}}$.

$$\sigma_{\bar{x}} = \sqrt{\sigma^2_{\bar{x}}}$$

La relación entre σ^2 y $\sigma^2_{\bar{x}}$.

$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$, donde n es el número de elementos en cada media de la muestra (tamaño de la muestra).

La estimacion de $\sigma^2_{\bar{x}}$ a partir de m muestras, $s^2_{\bar{x}}$.

$$s^2_{\bar{x}} = \frac{\Sigma\left(\bar{X}_i - \bar{\bar{X}}\right)^2}{m - 1}$$

La estimación de $\sigma^2_{\bar{x}}$ a partir de una única muestra de tamaño n.

$$s^2_{\bar{x}} = \frac{\Sigma\left(X_i - \bar{X}\right)^2}{n(n-1)} = \frac{s^2}{n}$$

La estimación de σ^2 cuando $s^2_{\bar{x}}$ es conocido.

$\sigma^2 \cong s^2 = n\, s^2_{\bar{x}}$, donde n es el número de elementos en cada muestra.

Límites de confianza de μ, muestras grandes (n > 60).

$LC = \bar{X} + z\,\sigma_{\bar{x}}$, donde $\bar{X}$ es la media de la muestra grande, z es un valor tabular basado en el nivel de probabilidad deseado y $\sigma_{\bar{x}}$ se estima mediante $s_{\bar{x}}$, calculado a partir de la muestra.

Límites de confianza de μ, muestras pequeñas.

$$LC = \bar{X} \pm t\, s_{\bar{x}}$$

$$S_{\bar{x}} = \sqrt{\frac{S^2}{n}}$$

F, la razón entre dos estimaciones de σ^2.

$$F = \frac{s^2, \text{ cálculo a partir de las medias de la muestra}}{s^2, \text{ calculado por la combinación de las varianzas de la muestra}}$$

Capítulo 3

Análisis de varianza y pruebas de significación

Disponemos ahora de los conceptos estadísticos para entender el análisis de varianza; sin embargo, antes de examinar experimentos complicados resultará ilustrativo ver cómo podemos utilizar los instrumentos que hemos adquirido para analizar el caso más sencillo de dos muestras (tratamientos).

EXPERIMENTO CON DOS MUESTRAS

Emplearemos los datos de la tabla 3.1 para ilustrar el procedimiento en el análisis de varianza.

Tabla 3.1. Cosechas (10^2 lb/acre) de las variedades de trigo A y B en parcelas a las cuales las variedades fueron asignadas aleatoriamente.

Tratamientos	Repeticiones					Total	$\bar{X}$
	I	II	III	IV	V		
A	19	14	15	17	20	85	$17 = \bar{X}_A$
B	23	19	19	21	18	100	$20 = \bar{X}_B$
						185	$18.5 = \bar{X}$

1. Estímese la varianza de la población a partir de la cual cada muestra fue extraída.

$$s^2_A = \frac{\Sigma\left(X_{A_i} - \bar{X}_A\right)^2}{n_A - 1} = \frac{(19 - 17)^2 + \ldots + (20 - 17)^2}{5 - 1} = \frac{(2)^2 + \ldots + (3)^2}{4} = \frac{26}{4} = 6.5$$

$$s^2_B = \frac{\Sigma\left(X_{B_i} - \bar{X}_B\right)^2}{n_B - 1} = \frac{(23 - 20)^2 + \ldots + (18 - 20)^2}{5 - 1} = \frac{(3)^2 + \ldots + (-2)^2}{4} = \frac{16}{4} = 4.0$$

2. Suponiendo que s^2_A y s^2_B estiman una varianza común (σ^2), obténgase la mejor estimación de dicha varianza (s^2) combinando $s^2_A + s^2_B$. Luego estimaremos σ^2 en otra forma y así designaremos la estimación de σ^2 basada en la variabilidad **dentro** de las muestras como $s^2_1 = (s^2_A + s^2_B)/2 = (6.5 + 4.0)/2 = 5.25$.

variabilidad dentro de las muestras

3. Asumiendo la hipótesis nula de que estas dos muestras son muestras aleatorias extraídas de la misma población y que, por tanto, $\bar{X}_A$ y $\bar{X}_B$, estiman la media de la población (μ), estímese la varianza de medias ($\sigma^2_{\bar{X}}$) a partir de las medias de las muestras A y B.

media de 4 y 8 general

$$s^2_{\bar{X}} = \frac{\Sigma(\bar{X}_i - \bar{X})^2}{m - 1} = \frac{(17 - 18.5)^2 + (20 - 18.5)^2}{2 - 1} = \frac{(-1.5)^2 + (1.5)^2}{1} = 4.5$$

Varianza de medias

$m-1$

4. Nuevamente estímese σ^2 utilizando la relación $s^2_{\bar{X}} = \frac{s^2}{n}$ y resolviéndola para s^2. Recuérdese que n es el número de elementos sobre los cuales está basada cada media de la muestra. Denotaremos esta estimación de σ^2 como s^2_2. $s^2_2 = n\, s^2_{\bar{X}} = 5(4.5) = 22.5$.

$S^2_{\bar{x}} = \frac{S^2}{n}$ variabilidad entre tratamientos $S^2_2 = n \cdot S^2_{\bar{x}}$

Contamos ahora con dos estimaciones de σ^2: s^2_1 basada en la variabilidad **dentro** de cada muestra, y s^2_2 basada en la variabilidad **entre** las muestras. Suponiendo que la hipótesis nula es verdadera, podríamos esperar que s^2_1 y s^2_2 fuesen casi iguales, puesto que ambas son estimaciones de la misma varianza (σ^2). Podemos determinar la probabilidad de obtener estimaciones divergentes de σ^2 calculando una razón F y confrontando el resultado en una tabla de valores F. Para calcular esta razón F, siempre colocamos la varianza estimada a partir de las medias de la muestra (tratamiento) (s^2_2) en el numerador y la varianza estimada a partir de elementos individuales, en el denominador. Entonces: $F = \frac{s^2_2}{s^2_1}$.

F calculado.

Si los dos tratamientos (muestras) provienen de poblaciones que tienen medias diferentes, s^2_2 contendrá un componente que refleje esta diferencia y será mayor que s^2_1. Para nuestro experimento, $F = \frac{22.5}{5.25} = 4.29$.

El numerador, s^2_2, está basado en un grado de libertad, puesto que existen dos medias de la muestra. El denominador, s^2_1, está basado en la combinación de los grados de libertad **dentro** de cada muestra. Cada muestra tiene 5 elementos; por tanto, 4 gl, de modo que los grados de libertad para s^2_1 son 4 + 4 = 8.

A partir de una tabla de valores F, obtenemos los valores de F que podríamos esperar con una probabilidad especificada si la hipótesis nula fuese verdadera y nuestras medias de la muestra difirieran sólo por casualidad. Para 1 y 8 grados de libertad (numerador y denominador), podríamos esperar un valor de F igual a 4.29 o mayor, con una probabilidad de aproximadamente un 7%. En otras palabras... si la verdadera diferencia entre las medias es igual a cero ($\mu_A - \mu_B = \mu_{\bar{d}} = 0$), la probabilidad de obtener una estimación de $\mu_{\bar{d}} = 3$ quintales por acre es de aproximadamente 7%. Por regla general, no estamos dispuestos a apostar a que este acontecimiento (el cual tiene un 7% de probabilidades de ocurrir) no ocurre; por tanto, sería imprudente rechazar la hipótesis nula y concluir que la media de la variedad A es en realidad distinta de la media de la variedad B. Por otro lado, una diferencia entre medias de variedades de 3 gl por acre, si es real, representa una considerable ganancia económica. En consecuencia, debemos decidirnos por la evaluación de las dos variedades en experimentos adicionales.

UNA POBLACIÓN DE DIFERENCIAS DE MEDIAS Y LA PRUEBA t DE SIGNIFICACIÓN ESTADÍSTICA

Podemos también utilizar una prueba t para evaluar las posibilidades de que dos medias sean significativamente diferentes. En primer lugar, necesitamos ver cómo una población de diferencias de medias es generada a partir de una población de elementos normalmente distribuidos; en particular, necesitamos saber cómo los parámetros de esta nueva población están relacionados con los parámetros de las poblaciones originales y con las poblaciones de medias también originadas en la obtención de la población de diferencias de medias.

Si a partir de dos poblaciones normalmente distribuidas, $X_1, X_2, \ldots X_N$ y $Y_1, Y_2, \ldots Y_N$, extraemos todas las posibles muestras de un tamaño dado y calculamos sus medias, tendremos dos poblaciones adicionales, $\bar{X}_1, \bar{X}_2 \ldots X_M$ y $\bar{Y}_1, \bar{Y}_2, \ldots \bar{Y}_M$. Si tomamos ahora todos los posibles pares de medias y sustraemos luego $\bar{X}_1 - \bar{Y}_1, \bar{X}_1 - \bar{Y}_2, \ldots \bar{X}_1 - \bar{Y}_M, \bar{X}_2 - \bar{Y}_1, \ldots \bar{X}_2 - \bar{Y}_M, \ldots \bar{X}_M - \bar{Y}_M$, tendremos una quinta población de **diferencias de medias** (véase la figura 3.1). Las siguientes relaciones entre las medias y las desviaciones estándar de dichas poblaciones pueden demostrarse matemáticamente, pero aquí sólo se establecerán. La media de las diferencias de medias es igual a la diferencia entre las medias de las medias de la muestra a partir de las poblaciones X y Y: $\mu_{\bar{d}} = \mu_{\bar{X}} - \mu_{\bar{Y}}$. Esta diferencia es también igual a la diferencia entre la media de la población X y la población Y:

$$\mu_{\bar{d}} = \mu_{\bar{X}} - \mu_{\bar{Y}} = \mu_X - \mu_Y. \text{ Si } \mu_X = \mu_Y, \text{ entonces } \mu_{\bar{d}} = 0.$$

La varianza de la población de diferencias de medias es:

$$\sigma^2_{\bar{d}} = \frac{\Sigma\left(\bar{d}_i - \mu_{\bar{d}}\right)^2}{Q}$$ que es igual a la suma de las varianzas de las respectivas medias. Por tanto:

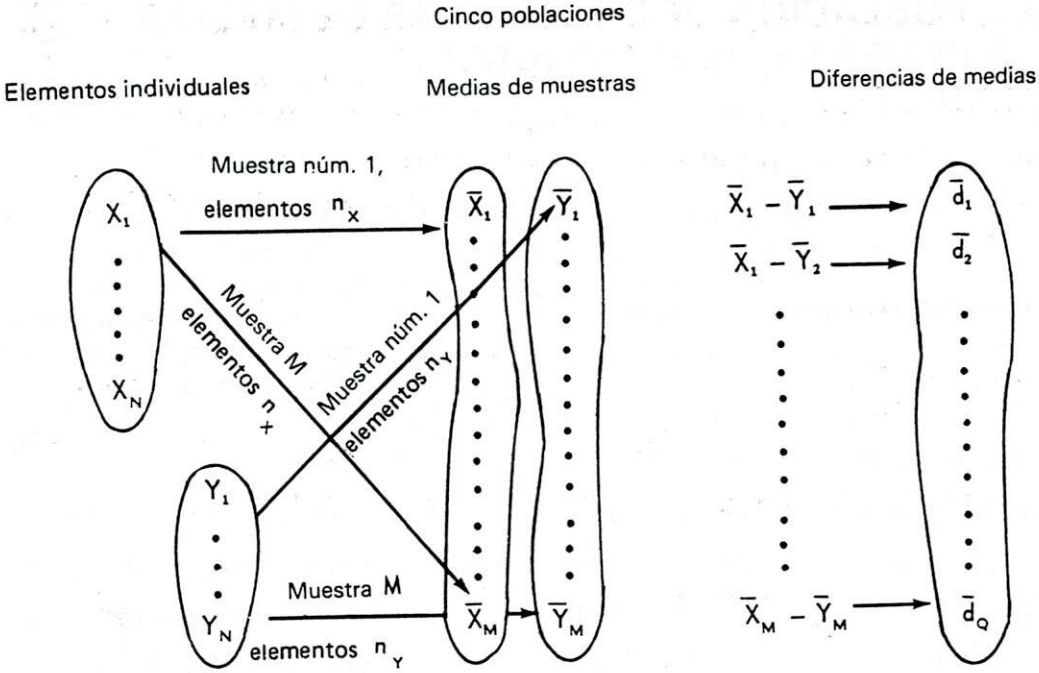
$\sigma^2_{\bar{d}} = \sigma^2_{\bar{X}} + \sigma^2_{\bar{Y}}$. A partir de dos muestras, $\sigma^2_{\bar{d}}$ se estima mediante $s^2_{\bar{d}}$ de las varianzas de las medias de la muestra: $s^2_{\bar{d}} = s^2_{\bar{X}} + s^2_{\bar{Y}}$. Puesto que $s^2_{\bar{X}} = \frac{s^2_X}{n_X}$ y $s^2_{\bar{Y}} = \frac{s^2_Y}{n_Y}$, $\quad s^2_{\bar{d}} = \frac{s^2_X}{n_X} + \frac{s^2_Y}{n_Y}$.

La raíz cuadrada de la varianza de las diferencias de medias suele denominarse **error estándar de una diferencia**. Comúnmente, en el análisis estadístico, una varianza se calcula a partir de otra.

Importantes relaciones entre varianzas que utilizaremos a menudo son:

$$s^2_{\bar{X}} = \frac{s^2_X}{n}; \quad s^2_{\bar{d}} = s^2_{\bar{X}} + s^2_{\bar{Y}}; \; s^2_{\bar{d}} = \frac{s^2_X}{n_X} + \frac{s^2_Y}{n_Y}; \quad \text{y cuando } n_X = n_Y \text{ y } s^2_X = s^2_Y,$$

hacemos $n_X = n_Y = n$ y $s^2_X = s^2_Y = s^2$; por tanto: $s^2_{\bar{d}} = \frac{2s^2}{n}$.



Medias

$$\mu_X \quad \mu_Y \qquad \mu_{\bar{X}} \quad \mu_{\bar{Y}} \qquad \mu_{\bar{d}}$$

$$\mu_X - \mu_Y = \mu_{\bar{X}} - \mu_{\bar{Y}} = \mu_{\bar{d}}$$

$$(\text{Si } \mu_X = \mu_Y \longrightarrow \mu_{\bar{d}} = 0)$$

Varianzas

$$\sigma^2_X \quad \sigma^2_Y \qquad \sigma^2_{\bar{X}} \quad \sigma^2_{\bar{Y}} \qquad \sigma^2_{\bar{d}} = \frac{\Sigma (\bar{d}_i - \mu_{\bar{d}})^2}{Q}$$

$$\sigma^2_{\bar{d}} = \sigma^2_{\bar{X}} + \sigma^2_{\bar{Y}} = \frac{\sigma^2_X}{n_X} + \frac{\sigma^2_Y}{n_Y}$$

Cuando $\sigma^2_X = \sigma^2_Y = \sigma^2$ y $n_X = n_Y = n$, entonces: $\sigma^2_{\bar{d}} = 2\frac{\sigma^2}{n}$

Figura 3.1. Generación de poblaciones de medias y diferencias de medias, a partir de dos poblaciones de elementos individuales y relaciones entre parámetros (véase texto).

Tratándose de una población de **diferencias de medias**, la fórmula para determinar t es: $t = \frac{\bar{d} - \mu_{\bar{d}}}{s_{\bar{d}}}$. Para el experimento de la tabla 3.1, deseamos conocer la probabilidad de que las muestras X_A y X_B pudiesen haber provenido de poblaciones con medias idénticas $(\mu_A = \mu_B)$. Este planteamiento es análogo al del análisis anterior, donde nos referimos a las poblaciones X y Y; sólo que ahora pasamos a denominarlas X_A y X_B. La diferencia de medias de nuestras medias de la muestra es $\bar{d} = \bar{X}_A - \bar{X}_B = 17 - 20 = 3(10^2)$ libras por acre.

El error estándar de la diferencia es:

$$s_{\bar{d}} = \sqrt{s^2_{\bar{X}_A} + s^2_{\bar{X}_B}}\ , = \sqrt{\frac{s^2_A}{n_A} + \frac{s^2_B}{n_B}} = \sqrt{\frac{6.5}{5} + \frac{4}{5}} = \sqrt{\frac{10.5}{5}} = \sqrt{2.10} = 1.449.$$

Asumiendo la hipótesis nula de que $\mu_A = \mu_B$, $\left(\mu_{\bar{d}} = 0\right)$, t se calcula como sigue:

$$t = \frac{\bar{d} - \mu_{\bar{d}}}{s_{\bar{d}}} = \frac{3 - 0}{1.449} = 2.07.$$

A partir de la Tabla A.2, podemos determinar el valor absoluto mínimo de t, que tiene un 5% de probabilidades de ocurrir. Si suponemos que $\sigma^2_A = \sigma^2_B$ buscaremos t basado en los grados de libertad combinados dentro de las muestras, en este caso 4 + 4 = 8. El valor esperado de t para el nivel de probabilidad de 5% es 2.306; por tanto, nuestro tratamiento nuevamente se juzga como no significativo. Nótese que $t^2 = F$, $(2.07)^2 = 4.285$. Considerando los errores circundantes, éste es igual a nuestro F de 4.29, previamente calculado.

Un punto que debe enfatizarse aquí es que el análisis del procedimiento de varianza y el cálculo de un valor F conducen a las mismas conclusiones que la prueba t. Los investigadores frecuentemente expresan la idea de que existe algo único y más poderoso respecto de la prueba t, en comparación con la prueba F del análisis de varianza. Las pruebas son equivalentes, a la vez que el análisis del procedimiento de varianza es, por regla general, más fácil de llevar a cabo.

Un punto adicional debe establecerse con respecto a la utilización de una prueba t: esta prueba es apropiada cuando $\sigma_A \neq \sigma_B$. En este caso, la prueba F del análisis de varianza no es válida. Cuando $\sigma_A \neq \sigma_B$ y $n_A = n_B = n$, el valor de t requerido para la significación es para $n - 1$ grados de libertad. En nuestro ejemplo, $n = 5$, y el valor de t requerido en el nivel de 5% sería el valor tabular para 4 grados de libertad (g.l.) o 2.776. Cuando $n_A \neq n_B$ el valor de t requerido debe ser calculado como si estuviese en algún punto entre el valor tabular de t para $n_A - 1$ y $n_B - 1$ grados de libertad. Cuando $\sigma_A \neq \sigma_B$ y $n_A \neq n_B$ el valor requerido de t es:

$$t = \frac{t_A s^2_{\bar{X}_A} + t_B s^2_{\bar{X}_B}}{s^2_{\bar{X}_A} + s^2_{\bar{X}_B}}$$

donde t_A y t_B son valores tabulares de t para $n_A - 1$ y $n_B - 1$ grados de libertad, respectivamente.

Tenemos una estimación de la diferencia de medias de la población, a saber: $\bar{X}_A - \bar{X}_B = 3(10^2)$ libras por acre, y podemos desear calcular un intervalo de confianza dentro del cual se hallará la verdadera diferencia de medias de la población, a menos que las muestras que hayamos extraído sean muy poco comunes. Con una

confianza del 95%, podemos afirmar que $\mu_{\bar{d}}$ se encuentra dentro del intervalo $d \pm t_{.95}\, s_{\bar{d}}$. Por tanto, los límites de confianza de 95% son

$$LC_{95} = 3 \pm 2.306\ (1.449) = 3 \pm 3.34 = -0.34 \text{ para } 6.34(10^2) \text{ lb por acre.}$$

Nótese que este intervalo de confianza incluye el cero, lo cual es otra forma de mostrar que las medias de los tratamientos A y B no son significativamente diferentes.

Diferencia significativa mínima

La diferencia significativa mínima (DSM) se estudiará con mayor amplitud en el capítulo 6, en la sección titulada Separación de medias, pero debe mencionarse aquí, puesto que es una forma de la prueba t, objeto de nuestra consideración. La fórmula para calcular la DSM entre dos medias es: $DSM = t\, s_{\bar{d}}$. En el caso de experimentos que sólo contemplan dos tratamientos, no es necesario calcular la DSM, puesto que solamente hay una diferencia de medias a considerar y una prueba F o t revelará si la diferencia es significativa.

MÉTODOS PARA INCREMENTAR LA PRECISIÓN

Dos expresiones, **precisión y exactitud**, suelen emplearse como sinónimos, aunque en el lenguaje estadístico tengan significados ligeramente distintos. Exactitud se refiere a la proximidad con que puede hacerse una medición en particular, mientras que precisión se refiere a la magnitud de la diferencia entre dos tratamientos que un experimento es capaz de detectar. Antes de estudiar las técnicas para lograr la precisión de los experimentos, consideraremos brevemente la exactitud en la recolección y recopilación de datos.

Siempre que sea posible, los registros originales deben recabarse de modo que no sea necesario volver a copiarlos. Si se está utilizando un equipo de procesamiento electrónico, el acopio de datos puede organizarse de manera que las cifras originales se empleen para perforar la tarjeta de datos. Esto evita posibles errores en el proceso de recopiado. Si las cifras deben ser trasladadas, habrán de comprobarse inmediatamente.

En el momento en que los datos son recabados, deben examinarse en cuanto a las cifras fuera de lugar y reexaminarse en su totalidad, para evitar posibles errores. Existe ya suficiente variación inherente a los datos biológicos, sin permitir mayores deslices a consecuencia de errores humanos.

¿Con qué aproximación deben tomarse las medidas?,¿se pesa con una aproximación de centésimas, décimas o para todo un kilogramo? La exactitud espuria debe evitarse. Resulta innecesario registrar cifras no garantizadas por la precisión de un experimento. Cochran y Cox (A. 13) dan la siguiente guía general para el número de cifras significativas que deben ser retenidas: si el coeficiente de variabilidad se encuentra entre 0.4 y 4%, regístrense los datos originales para cuatro cifras significativas; si el CV se encuentra entre 4 y 40%, reténganse tres cifras; y si el CV es mayor que un 40%, regístrense sólo dos cifras. Una regla más precisa es la de que el intervalo completo no debe exceder de un cuarto de la desviación estándar por parcela. Si s es estimado en 12.50, entonces ¼ (12.5) = 3.1; por tanto, regístrense los datos para la unidad entera más cercana. Si s = 2.5, ¼ (2.5) = 0.6, regístrese para una aproximación de 0.1. Si s = 0.20, ¼ (0.20) = 0.05, regístrese para una aproximación de 0.01.

El instrumento utilizado para la ponderación o medición no necesita ser más exacto de lo requerido por la precisión del experimento; por ejemplo, si se ha de medir una serie de pesos y ha de redondearse a la cantidad más cercana en kilogramos, las escalas utilizadas pueden expresarse en unidades completas de kilogramo en vez de hacerlo en divisiones de tal unidad.

La llamada **regla de ingenieros** es un buen ejemplo a seguir para el mecanismo de redondear números. Esta regla establece que el dígito retenido debe redondearse hacia arriba si el dígito que ha de excluirse es mayor que 5; sin embargo, si el dígito siguiente al que ha de ser retenido es 5, el dígito precedente debe redondearse para hacerlo par si éste es impar y mantenerse como está si el mismo es par; o sea, para redondear 21.55 con una aproximación de 0.1, obtendríamos 21.6; pero para redondear 21.45, obtendríamos 21.4.

En el análisis de varianza es mejor retener el número completo de cifras obtenidas a partir de la suma total de cuadrados no corregida. Si los datos originales contienen una cifra decimal, la suma de cuadrados registrará dos lugares decimales.

Evítense los dígitos superfluos en la presentación final de los resultados. Redondéense las medias de los tratamientos a una décima del error estándar estimado de una media. Si $s_{\bar{x}} = 2.56$, $0.1\ (2.56) = 0.2$, por tanto, redondéense los tratamientos con una aproximación de 0.1. Entonces, una media de 15.12 se presenta como 15.1.

Cuanto mayor sea la variabilidad entre parcelas igualmente tratadas, mayor será el error asociado a la diferencia entre dos medias, y menor será la precisión del experimento para detectar diferencias debidas a los tratamientos. El error estándar de la diferencia entre dos medias disminuye cuando s decrece y n aumenta. $s_{\bar{d}} = \sqrt{\frac{2s^2}{n}}$ (donde n es el número de réplicas). Por tanto, los métodos para incrementar la precisión de un experimento son diseñados para disminuir la variabilidad no determinada por parcela o para aumentar el número efectivo de réplicas.

La precisión puede mejorarse mediante: a) el incremento de las repeticiones, b) la cuidadosa selección de los tratamientos, c) el refinamiento de la técnica, d) la selección del material experimental, e) la selección de la unidad experimental, f) la toma de medidas adicionales, y g) el agrupamiento planeado de unidades experimentales.

Incremento de repeticiones

La precisión de un experimento siempre puede aumentarse mediante repeticiones adicionales, aunque el grado de mejoramiento decrece rápidamente cuando el número de repeticiones aumenta; por ejemplo, en comparación con un experimento con cuatro repeticiones, para duplicar el grado de precisión con el cual dos medias pueden separarse se requieren 16 repeticiones. Esto se sigue del efecto que ejerce el número de repeticiones (n) sobre la diferencia requerida para separar dos medias a un nivel de significación dado, $NSD = t\sqrt{\frac{2s^2}{n}}$. Esto no es exactamente así por que cuando n se incrementa, t se vuelve ligeramente menor, aunque se encuentra lo suficientemente cercano para ser utilizado como una regla práctica.

Por regla general, en la investigación de campo y de cosechas de vegetales, se requieren entre cuatro y ocho repeticiones para un grado de precisión razonable. En la planeación de un experimento debemos estar razonablemente seguros de que seremos capaces de detectar una diferencia verdadera de la magnitud en la cual estamos interesados. Si la probabilidad de que podamos alcanzar nuestro objetivo con el número de repeticiones que estamos dispuestos a emplear es reducida, y no existen otros medios razonables para mejorar la precisión, estaríamos bien advertidos de no realizar el experimento; —o al menos para suspenderlo, hasta que tengamos recursos suficientes para llevar a cabo el experimento en una forma que tenga buenas probabilidades de alcanzar nuestro objetivo. La tabla 2.1, diseñada por Cochran y Cox (A. 13), es conveniente para estimar el número de repeticiones requeridas para detectar una diferencia específica.

Selección de tratamientos

Una cuidadosa selección de tratamientos no sólo es importante para alcanzar los objetivos del experimentador, sino que puede también incrementar la precisión del experimento. Por ejemplo, en el estudio de los efectos de un herbicida, fungicida, fertilizante o insecticida, resulta más útil determinar cómo las unidades experimentales responden a dosis cada vez mayores de nuestro material experimental, que decidir si dos dosis sucesivas son o no son significativamente diferentes. Por tanto, una serie de dosis apropiadas hará posible la planificación de pruebas de significación que sean más sensibles que la mera comparación de medias adyacentes en un conjunto. Se comentará esto más detalladamente en los capítulos 6 y 15.

La utilización de experimentos factoriales, en los que dos o más tipos de tratamientos son sometidos a prueba simultáneamente, puede resultar en un considerable mejoramiento de la precisión. Los experimentos factoriales se analizarán en la próxima sección, y con gran amplitud en el capítulo 6.

Refinamiento de la técnica

Una técnica defectuosa puede incrementar el error experimental y el sesgo sobre los efectos del tratamiento. A su vez, una buena técnica debe perseguir los siguientes objetivos: a) tratamientos uniformemente aplicados, b) trazado adecuado y mediciones insesgadas de los efectos del tratamiento, c) prevención de errores sistemáticos, y d) control de las influencias externas, de modo que todos los tratamientos sean comparablemente afectados.

Selección del material experimental

Para ciertos tipos de estudios, resulta deseable un material uniforme y cuidadosamente seleccionado; sin embargo, en la selección del material experimental debemos tener en mente la población acerca de la cual deseamos hacer inferencias. Es por ello que en la mayoría de las investigaciones aplicadas a la agricultura, resulta importante utilizar los tipos de material experimental que se emplearán en la producción real.

Selección de la unidad experimental

El tamaño y la forma de la parcela afecta la precisión. Por regla general, la variabilidad disminuye con el incremento del tamaño de la parcela, aunque una vez que cierto tamaño ha sido alcanzado, el incremento de la precisión decrece rápidamente para tamaños mayores. En la determinación de la producción, usualmente se registra una pequeña ganancia en precisión mediante el empleo de parcelas mayores que 0.1 acres; para la mayoría de los cultivos, las áreas cosechadas de 0.01 a 0.02 acres registran una buena precisión. LeClerg y colaboradores (A. 13) han estudiado el tamaño y la forma de las parcelas de campo para diversos cultivos y citan muchas referencias de utilidad. Las parcelas rectangulares son más eficientes en la superación de la heterogeneidad del suelo cuando sus ejes de longitud están dirigidos hacia la mayor variación del suelo.

El número creciente de animales por unidad experimental también aumenta la precisión; sin embargo, si los animales pueden ser manipulados individualmente, la precisión se verá más incrementada mediante la utilización de individuos como unidades experimentales teniendo un mayor número de repeticiones que empleando el mismo número de animales con más de uno por unidad experimental.

Toma de medidas adicionales

Una técnica conocida como análisis de covarianza puede utilizarse algunas veces para eliminar una importante fuente de variación entre las unidades experimentales. Los pesos iniciales del ganado empleado en un experimento de alimentación pueden utilizarse para eliminar dicho efecto sobre la tasa de aumento durante cierto periodo de alimentación. El análisis de covarianza contempla considerables cálculos; sin embargo no se examinará en este libro. Si el lector está interesado en esta técnica, puede recurrir a la obra de Snedecor (A.13), Steele y Torry (A.13) o a otro texto completo sobre técnicas estadísticas.

Agrupamiento planeado de las unidades experimentales

El agrupamiento planeado contempla la aplicación del principio de diseño experimental llamado **control local.** Mediante ciertas restricciones sobre la elección aleatoria de los tratamientos para unidades experimentales, es posible eliminar ciertas fuentes de variación, tales como cambios en la fertilidad del suelo a lo largo de un área experimental, o diferencias en la capacidad de aumentar de peso asociadas con la edad y el peso de los animales. La agrupación en diversas formas de las unidades experimentales da lugar a varios diseños experimentales. Las ventajas y desventajas de cada una de estas formas se analizarán para los diseños que se presentarán en los capítulos subsecuentes.

EXPERIMENTOS FACTORIALES

En un experimento factorial, los efectos de dos o más factores se investigan en forma simultánea. Si se sospecha que la conducta de un factor varía con los cambios de otro, dicha conducta puede probarse mediante un conjunto factorial de tratamientos, planificado en un diseño experimental adecuado.

Cuando dos o más factores (cada uno debe estar en dos o más niveles) se prueban en todas las combinaciones posibles, se dice que los tratamientos resultantes son factoriales. Los efectos diferenciales de un factor sobre otro reciben el nombre de interacción. El descubrimiento de las interacciones amplía las conclusiones de un experimento. El rango de validez del experimento se incrementa, lo cual es una característica deseable de un experimento bien planeado. Incluso si no se detectan interacciones, en los experimentos factoriales los resultados son más ampliamente aplicables, puesto que se ha demostrado que los efectos principales de los tratamientos se mantienen para un rango más grande de condiciones.

Ejemplos de combinaciones de factores en un experimento son: probar varios niveles de diferentes fertilizantes del suelo; evaluar el efecto de una hormona sobre la capacidad de aumento de peso de la oveja macho y la de la oveja hembra.

Un conjunto factorial de tratamientos se muestra en la tabla 3.2. Los nueve **tratamientos** son todas las combinaciones posibles de tres niveles de dosificación de un insecticida y tres niveles de dosificación de un fungicida, utilizados como tratamientos de semillas del frijol de media luna.

Tabla 3.2. Tratamientos de la semilla del frijol de media luna. Combinación factorial de tres niveles de dosificación de un fungicida con tres niveles de dosificación de un insecticida

Dosis de fungicida	Dosis de insecticida		
	I_0 (ninguna)	I_1	I_2
F_0 (ninguna)	F_0I_0	F_0I_1	F_0I_2
F_1	F_1I_0	F_1I_1	F_1I_2
F_2	F_2I_0	F_2I_1	F_2I_2

Este conjunto de tratamientos permite evaluar la contribución relativa del fungicida y del insecticida para el surgimiento de los retoños de frijol de media luna. Véase la tabla 3.3 para los promedios del tratamiento, y la figura 3.2 para una presentación gráfica de los resultados que muestran el significado de la interacción.

Tabla 3.3. Efecto de los niveles de fungicida e insecticida en el tratamiento de la semilla sobre el surgimiento de los retoños de frijol de media luna. Los valores presentados son retoños por cada 100 semillas

Fungicida (onza por 100 lb de semillas)	Insecticida (onza por 100 lb de semillas) 0 (I_0)	⅙ (I_1)	⅓ (I_2)	Efecto promedio del fungicida
	Insecticida x Fungicida (Medias)			
0 (F_0)	68	58	48	59
1 ⅓ (F_1)	94	93	90	92
2 ⅔ (F_2)	89	92	92	91

% Surgimiento

En la figura 3.2, nótese la disminución del surgimiento al incrementar la dosis de insecticida cuando éste fue utilizado sin el fungicida. Dicha disminución no se registra cuando un fungicida se añade al tratamiento de la semilla. El efecto diferencial del insecticida, dependiendo de si un fungicida fue o no utilizado, es lo que denominamos **interacción**. Si una interacción no se registra, la disposición factorial multiplica el número de repeticiones para probar los efectos promedio totales de los componentes del tratamiento. Nótese que no existe un apreciable efecto diferencial del insecticida en las dosis F_1 y F_2 del fungicida. En otras palabras, no hay interacción I × F con respecto a las dosis F_1 y F_2 del fungicida. En este caso, la mejor estimación de los

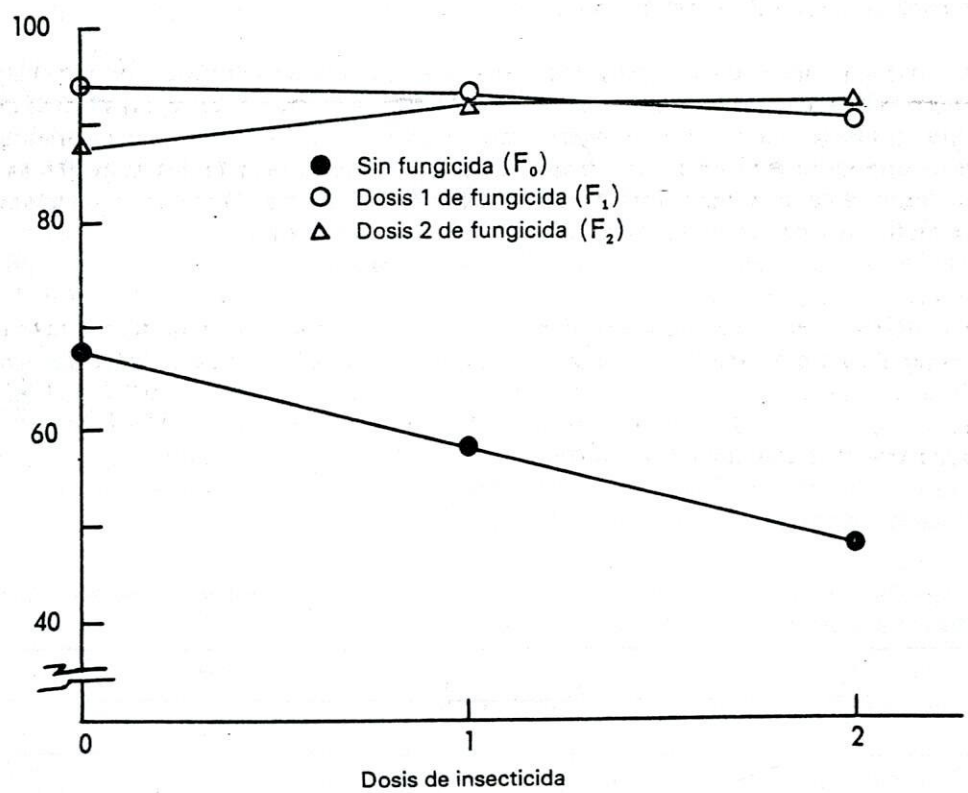


Figura 3.2. Representación gráfica de los promedios de los tratamientos de la tabla 3.3.

efectos de tales dosis la constituyen los promedios de las mismas sobre todos los niveles del insecticida. Los promedios resultantes, F_1 = 92% y F_2 = 91% (tabla 3.3), están basados en 3 × el número de repeticiones de un tratamiento en particular. No se indica la superioridad de la dosis más elevada del fungicida.

Ocasionalmente, el lector habrá leído acerca de **diseños factoriales**. Esta terminología no es estrictamente correcta; lo factorial es la combinación de tratamientos, no el diseño.

EL ANÁLISIS DE VARIANZA Y EL DISEÑO EXPERIMENTAL

La diferencia principal entre los diseños experimentales radica en la forma en que se agrupan o clasifican las unidades experimentales. En todos los diseños las unidades experimentales se clasifican por tratamientos; pero en algunos, éstos se clasifican preferentemente en bloques, filas, parcelas principales y otras modalidades. El análisis de varianza utiliza las medias de dichos agrupamientos, denominadas fuente de variación, para estimar varianzas o más precisamente cuadrados medios. Un cuadrado medio que estima la dispersión entre mediciones de parcelas debidas a causas aleatorias también se calcula; ésta se denomina **error experimental**. En ausencia de diferencias reales debidas a medias de los tratamientos, bloques u otras fuentes de variación, dichos cuadrados medios serán, en promedio, iguales. Sólo esporádicamente un cuadrado medio se desviará de otro de manera considerable, exclusivamente por casualidad. Cuando una prueba F indica que el cuadrado medio de una de las fuentes de variación es significativamente mayor que el cuadrado medio debido a efectos aleatorios, decimos que existen diferencias reales entre las medias de aquella fuente particular de variación; empero, recuérdese: siempre existe una probabilidad definida de que estemos equivocados en semejante conclusión. Está en manos del experimentador seleccionar las probabilidades para las cuales se encuentra dispuesto a concluir que existen efectos reales.

Es frecuente describir los resultados que cabría esperar con una probabilidad del 5% o menor como **significativos** y a aquellos esperados con un 1% o menor como **altamente significativos**. Cuando un experimentador aplica la frase "los tratamientos son significativamente diferentes", realmente está diciendo que si la hipótesis nula es verdadera, las probabilidades de obtener tales diferencias de medias del tratamiento son sólo de un 5%. Está afirmando que no hubo tal probabilidad de ocurrencia en su experimento y que, por tanto, el resultado significativo se debió a un efecto real del tratamiento.

En los capítulos siguientes se explicarán las características principales de los diseños experimentales comúnmente utilizados en la investigación de campo, se proporcionará un ejemplo de cada uno y se dará a conocer el procedimiento a seguir en el análisis de los datos. Se utiliza el mismo conjunto de datos para los primeros diseños: el diseño completamente aleatorio y el diseño de bloques completos al azar. Esto muestra las posibles ventajas de un diseño sobre el otro, manteniendo la sencillez de los cálculos, de modo que podamos concentrarnos en lo que se está haciendo y por qué.

RESUMEN

La significación estadística de una **diferencia** entre dos medias de la muestra puede verificarse mediante la razón F en un **análisis de varianza**, o por medio de una **prueba t**. Ambas pruebas son estadísticamente equivalentes, $t^2 = F$. El análisis de varianza y la prueba F suelen ser más fáciles de calcular.

Capítulo 4

Diseño completamente aleatorio

Este diseño es el más sencillo y se origina por la asignación aleatoria de tratamientos a un conjunto de unidades experimentales previamente determinado. Por regla general, éste no es el diseño más eficiente para ensayos de campo con plantas, pero puede constituir la disposición más factible para verificar ciertos tipos de tratamientos en animales.

En este diseño puede probarse cualquier número de tratamientos. Resulta deseable, aunque no esencial, asignar el mismo número de unidades experimentales a cada tratamiento.

Las principales ventajas del diseño son la sencillez y la flexibilidad. Una de sus desventajas consiste en que algún otro diseño suele ser capaz de estimar el error estándar por unidad experimental (error experimental) con un mayor grado de precisión.

MUESTREO ALEATORIO

Se puede asignar un número arbitrariamente a cada una de las unidades requeridas de parcelas de campo o de animales que han de ser utilizadas en el experimento. El número de unidades experimentales será igual al número de tratamientos × el número de repeticiones. Una tabla de números aleatorios resulta conveniente para elegir a las unidades experimentales que recibirá cada tratamiento. Si cada tratamiento ha de repetirse cuatro veces, los primeros cuatro números aleatorios obtenidos se asignarán al tratamiento A, los siguientes cuatro números aleatorios al tratamiento B, y así sucesivamente.

Por ejemplo, deseamos probar tres tipos diferentes de hormonas, cada una en una dosis única, para determinar sus efectos sobre la capacidad de aumento de peso de las ovejas. Entonces, incluyendo el control, tenemos cuatro tratamientos. Los 16 grupos de ovejas asignados al experimento se numeran del 1 al 16. Aplicando la tabla de números aleatorios (tabla A.1), empezamos arbitrariamente hacia abajo en las columnas 5 y 6, y registramos los primeros cuatro números que encontramos entre el 1 y el 16. Estos son: 14, 13, 9 y 8. Estos grupos de ovejas se asignan al tratamiento A. Continuando hacia abajo por las columnas 5 y 6, los próximos cuatro grupos destinados a recibir el tratamiento B son: 12, 11, 6 (ahora nos desplazamos a las columnas 6 y 7, y realizamos la lectura) y 5. Después de que cuatro grupos hayan sido idénticamente asignados al tratamiento C, los cuatro restantes se asignarán al tratamiento D.

ANÁLISIS DE VARIANZA

Los datos de este experimento están organizados por tratamiento en la tabla 4.1.

Tabla 4.1. Aumento de peso de ovejas agrupadas por tratamientos (libras por animal, por 100 días).

Tratamiento	Repeticiones I	II	III	IV	Tratamiento Total (T_t)	Media ($\bar{X}_t$)
A (control) Criolla	47	52	62	51	212	53
B Corriedale	50	54	67	57	228	57
C Chevriot	57	53	69	57	236	59
D Merinos	54	65	74	59	252	63
				Total principal = 928		Media principal ($\bar{X}$) = 58

$FC = \frac{(\Sigma x_i)^2}{n} = \frac{(928)^2}{16} = 53'824.$

Fuentes de variación y grados de libertad

Se construye una tabla de análisis de varianza (tabla 4.2) y se completan las dos primeras columnas. Existen sólo dos fuentes de variación en el diseño completamente aleatorio: entre unidades experimentales **dentro** de un tratamiento, la cual denominamos **error experimental**, y aquella **entre** medias del tratamiento.

Tabla 4.2. Análisis de varianza.

Fuente de variación	Grados de libertad (gl)	Suma de cuadrados (sc)	Cuadrado medio (cm)	F Observado	F requerido 5%	F requerido 1%
Total	15	854	—	—	—	—
Tratamientos	3	208	69.3	1.29	3.49	5.95
Error	12	646	53.8			

$C.M_{trat} = \frac{SC_{trat}}{Gdl\ para\ trat}$ $C.M_{error} = \frac{SCerror}{Gde\ l\ de\ error}$ $F_{cal} = \frac{S^2_{trat}}{S^2_{error}}$ $\frac{69.33}{53.83} =$

Los grados de libertad son uno menos que el número de observaciones para cada fuente de variación: existen cuatro tratamientos y, por tanto, 3 gl; hay cuatro unidades experimentales por tratamiento, en consecuencia, 3 gl para cada tratamiento × 4 tratamientos, arroja un total de 12 gl para el error. Los grados de libertad asociados con la variabilidad total del experimento son uno menos que el número total de unidades experimentales: 16 – 1 = 15 gl. Nótese que los grados de libertad asociados con las fuentes de variación son aditivos. Esto facilita la determinación de los gl para el error, mediante la sustracción del gl para el total: 15 – 3 = 12.

Para facilitar el cálculo de los gl y las sumas de cuadrados para el **error**, situamos la variación **total** en primer lugar en la tabla de análisis de varianza, pero calculamos su suma de cuadrados después de que se han determinado todas las demás sumas de cuadrados, excepto el **error**.

Término de corrección (C)

$$C = \frac{(\Sigma X)^2}{rn} = \frac{(928)^2}{4(4)} = 53\,824$$

Nota: r = número de repeticiones, n = número de tratamientos. Si el número de repeticiones no es el mismo para todos los tratamientos, el divisor correcto es Σr_i. Si, por ejemplo, la primera repetición del tratamiento A de la tabla 4.1 estuviese ausente, el término de corrección sería:

$$C = \frac{(\Sigma X)^2}{3+4+4+4} = \frac{(881)^2}{15}.$$

Sumas de cuadrados y cuadrados medios

Tratamiento. SCT y CMT. $SCT = \frac{\Sigma(T_t)^2}{r} - C$, donde T_t = totales del tratamiento y r = número de repeticiones en cada tratamiento. Cuando el número de repeticiones no es el mismo para todos los tratamientos, el cuadrado de un total debe ser dividido entre el número de repeticiones, obteniéndose dicho total antes de efectuar la suma. Entonces, $SCT = \Sigma\frac{(T_t)^2}{r} - C$

$$SCT = \frac{(212)^2 + (228)^2 + \ldots + (252)^2}{4} - C = 54\,032 - 53\,824 = 208$$

La SCT se anota en la tabla 4.2. El cuadrado medio para el tratamiento (CMT) se obtiene dividiendo la SCT entre el gl para el tratamiento:

CMT = SCT/gl (T) = 208/3 = 69.3 lo cual se anota en la tabla 4.2

Total. SC. No necesitamos un CM para el total, puesto que éste contiene varianzas para todas las fuentes de variación:

$$SC = \Sigma(X)^2 - C = (47)^2 + (50)^2 + \ldots + (59)^2 - C = 54\,678 - 53\,824 = 854.$$

Error. SCE, CME

$$SCE = SC - SCT = 854 - 208 = 646$$

$$SME = SCE/gl\,(E) = 646/12 = 53.8$$

Nota: cualquier suma de cuadrados puede calcularse a partir de los totales mediante la siguiente fórmula:

Suma de cuadrados $= \Sigma\frac{(T_i)^2}{r_i} - \frac{(\Sigma X_i)^2}{\Sigma r_i}$ donde las T_i representan una serie de totales de elementos; las r_i el **número de elementos que completan cada total**, y las X_i son los elementos que completan los totales. Si r es el mismo para todos los totales, entonces $SC = \frac{\Sigma T_i^2}{r} - \frac{(\Sigma X_i)^2}{rn}$ donde n = número de totales. La mayoría de los estudiantes suelen equivocarse al decidir sobre el divisor para el primer término a continuación del signo =. Si un experimento tiene seis tratamientos, cada uno con cinco repeticiones, y ha de calcularse una suma de cuadrados para los **tratamientos**, existirán seis totales para elevar al cuadrado en el numerador del primer término, pero el **divisor, r, será igual a cinco, puesto que cada total está constituido por cinco elementos.**

Tabla 7.3. Separación ortogonal de los tratamientos de la tabla 7.2.

Fuente de variación	gl	SC	CM	F observado	F requerido 5%	F requerido 1%
Tratamientos	5	185.77	37.154	24.56	2.71	4.10
Sin N y con N	1	180.200	180.200	119.10	4.35	8.10
N orgánico y N inorgánico	1	3.816	3.816	2.52		
N de amonio y N de nitrato	1	0.202	0.202	0.13		
$(NH_4)_2SO_4$ vs NH_4NO_3	1	1.334	1.334	0.88		
$NaNO_3$ vs $CaNO_3$	1	0.213	0.213	0.14		
Error	20	30.25	1.513			

Tabla 7.4. Coeficientes de tratamiento para verificar la ortogonalidad de las comparaciones.

	Tratamientos y totales del tratamiento					
Comparación	No N 148.6	$(NH_4)_2SO_4$ 186.1	NH_4NO_3 182.1	$CO(NH_2)_2$ 188.9	$Ca(NO_3)_2$ 183.8	$NaNO_3$ 182.2
Sin N y con N	+5	−1	−1	−1	−1	−1
N orgánico y N inorgánico	0	−1	−1	+4	−1	−1
NH_4–N vs NO_3–N	0	+1	+1	0	−1	−1
$(NH_4)_2SO_4$ vs NH_4NO_3	0	+1	−1	0	0	0
$Ca(NO_3)_2$ vs $NaNO_3$	0	0	0	0	+1	−1

Las sumas de cuadrados pueden calcularse a partir de los totales del tratamiento como sigue:

$$SC \text{ Sin N y con N} = \frac{(148.6)^2}{6} + \frac{(186.1 + \ldots + 182.2)^2}{30} - \frac{(1\,071.7)^2}{36}$$

$$= 3680.327 + 28\,403.787 - 31\,903.914 = 180.200$$

Cuando la comparación involucra un solo grado de libertad, el método más corto para calcularlas, utilizando los polinomios ortogonales de la tabla 7.4, es: $SC = \frac{[\Sigma(c_i T_i)]^2}{r(\Sigma c_i^2)}$.

Por tanto:

$$SC \text{ Sin N y con N} = \frac{[5(148.6) - 186.1 - 182.1 - 188.9 - 183.8 - 182.2]^2}{6(30)} = \frac{(-180.1)^2}{180} = 180.200$$

Columnas $SCC = \frac{\Sigma(T_c)^2}{n} - C$ donde n = número de parcelas en cada columna.

$$= \frac{(183.6)^2 + \ldots + (178.8)^2}{6} - 31\,903.91 = 33.67$$

$$CMC = SCC/gl(C) = 33.67/5 = 6.734$$

Tratamientos. $SCT = \frac{\Sigma(T_t)^2}{r} - C$, donde r = número de repeticiones de cada tratamiento.

$$= \frac{(148.6)^2 + \ldots + (182.2)^2}{6} - 31\,903.91 = 185.77$$

$$CMT = SCT/gl(T) = 185.77/5 = 37.154$$

Total. $SC = \Sigma(X)^2 - C = [(28.2)^2 + (32.1)^2 + \ldots + (27.4)^2 + (29.1)^2] - 31\,903.91$

$$= 32\,185.79 - 31\,903.91 = 281.88$$

Error. $SCE = SC - SCH - SCC - SCT = 281.88 - 32.19 - 33.67 - 185.77 = 30.25$

$$CME = SCE/gl(E) = 30.25/20 = 1.513$$

VALORES F

$$F(\text{hileras}) = CMH/CME = 6.438/1.513 = 4.26$$

$$F(\text{columnas}) = CMC/CME = 6.734/1.513 = 4.45$$

$$F(\text{tratamientos}) = CMT/CME = 37.154/1.513 = 24.56$$

Las tres razones F están basadas en 5 y 20 grados de libertad. Los valores requeridos para la significación estadística se obtienen a partir de la tabla A.3 y se anotan en la tabla de análisis de varianza. Las tres fuentes de variación se clasifican como altamente significativas. Por ello concluimos que existen diferencias reales tanto entre hileras y columnas, como entre tratamientos.

SEPARACIÓN DE MEDIAS

Al planificar el experimento con remolacha de azúcar, diseñado para evaluar los efectos de diferentes fuentes de nitrógeno, el investigador formuló diversas preguntas que podrán ser contestadas mediante la separación de la suma de cuadrados de tratamientos en el conjunto ortogonal de comparaciones indicado en la tabla 7.3.

Los procedimientos para probar la ortogonalidad de las comparaciones y para completar la tabla 7.3 se muestran en la tabla 7.4.

Nótese que la suma de todas las hileras es igual a cero, que la suma de los productos de los coeficientes correspondientes de dos comparaciones cualesquiera es igual a cero y que, por tanto, las comparaciones de tratamientos son ortogonales.

Tabla 7.1. Producciones agrupadas por tratamiento e hilera.

Tratamiento[1]		Hilera I	II	III	IV	V	VI	Total del tratamiento (T_t)	Media del tratamiento
Sin nitrógeno	F	28.2	24.8	21.7	26.7	25.8	21.4	148.6	24.8
$(NH_4)_2SO_4$	A	32.1	30.6	31.9	30.4	30.3	30.8	186.1	31.0
NH_4NO_3	B	33.1	29.5	30.1	28.8	29.9	30.7	182.1	30.4
$CO(NH_2)_2$	C	32.4	29.4	30.8	33.1	33.5	29.7	188.9	31.5
$Ca(NO_3)_2$	D	29.1	33.0	30.6	31.4	32.3	27.4	183.8	30.6
$NaNO_3$	E	31.1	31.0	28.8	31.9	30.3	29.1	182.2	30.4
Total de la hilera (T_r)		186.0	178.3	173.9	182.3	182.1	169.1	1071.7(ΣX)	$\bar{X}$=29.8

[1] Cada uno de los materiales se aplica para suministrar 100 libras de N por acre.

Tabla 7.2. Análisis de varianza, ensayo con una fuente de nitrógeno para la remolacha de azúcar.

Fuente de variación	Grados de libertad gl	Suma de cuadrados SC	Cuadrado medio CM	F observado	F requerido 5%	F requerido 1%
Total	35	281.88				
Hileras	5	32.19	6.438	4.26	2.71	4.10
Columnas	5	33.67	6.734	4.45		
Tratamientos	5	185.77	37.154	24.56		
Error	20	30.25	1.513			

Término de corrección

$$C = (\Sigma X)^2/r^2 = (1071.7)^2/6(6) = 31\,903.91$$

Sumas de cuadrados y cuadrados medios

Hileras. $SCH = \dfrac{\Sigma(T_r)^2}{n} - C = \dfrac{(186.0)^2 + \ldots + (169.1)^2}{6} - 31\,903.91 = 32.19$, donde n = número de parcelas en cada hilera.

$$CMH = SCH/gl\,(R) = 32.19/5 = 6.438$$

MUESTREO ALEATORIO

Empecemos con un cuadro latino cualquiera (sistemático o aleatorio) con el número de tratamientos requerido para nuestro experimento; por ejemplo, deseamos asignar aleatoriamente seis tratamientos, A, B, C, D, E y F. Empezamos con el cuadro latino que aparece a continuación (figura 7.4); recurrimos a una tabla de números aleatorios (tabla A.1), escogemos un lugar arbitrario para comenzar (hilera 5) y continuamos a lo largo y de regreso (sobre la hilera 6), asignando los números 1, 3, 5, 4, 2 y 6 a las hileras 1 a la 6. Continuando a lo largo de la hilera 6 de la tabla de números aleatorios y de regreso (de derecha a izquierda) sobre la hilera 7, asignamos los números 4, 2, 5, 1, 3, 6 a las columnas.

	Columnas					
Hilera	4	2	5	1	3	6
1	B	D	E	F	A	C
3	C	E	A	D	F	B
5	A	F	C	B	E	D
4	D	A	F	C	B	E
2	F	B	D	E	C	A
6	E	C	B	A	D	F

Figura 7.4. Procedimiento para la reordenación aleatoria de un cuadro latino de 6 x 6. Las hileras y columnas han de ser reordenadas aleatoriamente en el orden indicado por una tabla de números aleatorios..

El nuevo cuadro latino se completa ahora como en la figura 7.1 mediante la reordenación de las hileras y columnas del cuadro original (figura 7.4), como se indicó con los números aleatorios.

ANÁLISIS DE VARIANZA

Analizaremos aquí los datos de la remolacha de azúcar contenidos en la figura 7.1. Los tratamientos fueron las cuatro fuentes de nitrógeno y un control, como se muestra en la tabla 7.1, donde las producciones de las parcelas se organizan en tratamientos e hileras. Los totales de las columnas se obtienen mediante la suma de las columnas, en la figura 7.1. Aparecen en la parte inferior de la tabla 7.1.

Los totales de las columnas (suma de las columnas de la figura 7.1) son: 183.6, 173.3, 169.7, 179.5, 186.9, 178.7.

El análisis de varianza se presenta en la tabla 7.2, y el procedimiento para completar el análisis aparece a continuación de la misma.

Fuentes de variación y grados de libertad

Los datos se clasifican en tres formas: hileras, columnas y tratamientos. La variación debida a cada uno de estos componentes es medida y sustraída del total para obtener el error experimental. Los grados de libertad son, como es usual, el número de observaciones asociadas con cada fuente de variación, menos uno. Los grados de libertad para el error pueden obtenerse por sustracción ($35 - 5 - 5 - 5 = 20$) o multiplicando $(n - 1)(n - 2)$ donde n = número de tratamientos, $(6 - 1)(6 - 2) = 5 \times 4 = 20$.

Hilera	Columna					
	I	II	III	IV	V	VI
I	F 28.2	D 29.1	A 32.1	B 33.1	E 31.1	C 32.4
II	E 31.0	B 29.5	C 29.4	F 24.8	D 33.0	A 30.6
III	D 30.6	E 28.8	F 21.7	C 30.8	A 31.9	B 30.1
IV	C 33.1	A 30.4	B 28.8	D 31.4	F 26.7	E 31.9
V	B 29.9	F 25.8	E 30.3	A 30.3	C 33.5	D 32.3
VI	A 30.8	C 29.7	D 27.4	E 29.1	B 30.7	F 21.4

Figura 7.1. Cuadro latino de seis por seis. Las letras indican los tratamientos de la tabla 7.1. Las variables representan cosechas de remolacha de azúcar, en toneladas por acre.

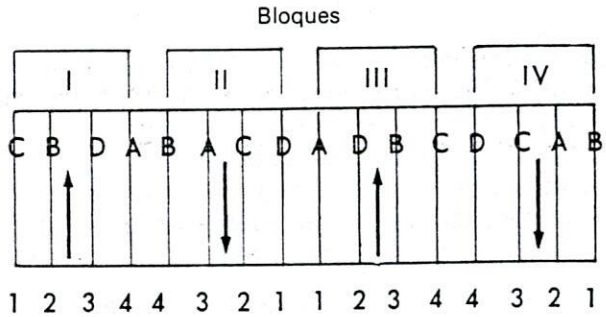


Figura 7.2. Cuadro latino con cuatro tratamientos de semilla (A, B, C y D) asignados a las unidades sembradoras 1, 2, 3 y 4. Las flechas indican la dirección del movimiento de la plantadora. Las fuentes de variación y los grados de libertad son: bloques = 3; unidades plantadoras = 3; tratamiento de semilla = 3; error = 6.

Hilera (periodos)	Columna (operadores)		
	I	II	III
I	B	A	C
II	C	B	A
III	A	C	B
IV	B	C	A
V	C	A	B
VI	A	B	C

Figura 7.3. Tres tratamientos en un cuadro latino doble. Las fuentes de variación y los grados de libertad son: hileras = 5; columnas = 2; tratamientos = 2; error = 8. Los tratamientos (A, B y C) son tres diferentes calculadoras de mesa.

Capítulo 7

Diseño de cuadro latino

En este diseño, la distribución aleatoria de los tratamientos se restringe más ampliamente mediante la agrupación de los mismos, tanto en columnas como en hileras (bloques). Así pues resulta posible eliminar la variabilidad del error experimental asociada con ambos efectos. Cada tratamiento ocurre el mismo número de veces (usualmente una vez) en cada hilera y columna. Este diseño proporcionará una comparación más precisa de los efectos del tratamiento que la suministrada por el diseño de bloques al azar sólo en caso de que exista una variación apreciable asociada con las columnas.

Las hileras y columnas pueden referirse a la distribución espacial de las unidades experimentales o al orden en el cual los trata nientos se realizan; por ejemplo, en la figura 7.1 las hileras y columnas se refieren a la disposición física de las parcelas de remolacha de azúcar. En la figura 7.3, los tratamientos A, B y C son tres diferentes modelos de calculadoras de mesa por probarse; las **columnas** son tres operadores diferentes y las **hileras** son las seis distintas oportunidades en que los tres operadores prueban la máquina. Cada operador prueba cada máquina en dos oportunidades, y las tres máquinas en su totalidad son probadas en cada periodo. De ahí que los efectos del periodo y de los operadores sean fuentes de variación medibles, independientes de las máquinas, y que puedan eliminarse de la variabilidad total del experimento, reduciendo el error experimental.

Un cuadro latino requiere al menos tantas repeticiones como tratamientos existan; por tanto no resulta práctico para experimentos con un gran número de tratamientos. Los cuadros latinos más comúnmente utilizados son aquellos que tienen entre cuatro y ocho tratamientos, con una sola unidad experimental por tratamiento en cada columna e hilera. La figura 7.1 ilustra una disposición de cuadro latino de parcelas de remolacha de azúcar, para recibir seis diferentes tratamientos con fertilizantes.

Hay ocasiones en que un cuadro latino puede ser ventajoso, incluso cuando las parcelas forman una línea continua. Considérese, por ejemplo, un experimento diseñado para probar cuatro tratamientos de semillas donde las parcelas individuales han de ser hileras particulares a lo largo del área experimental. Debe utilizarse una sembradora con cuatro unidades plantadoras. Las unidades plantadoras pueden diferir en cuanto a la proporción de siembra. Para eliminar el efecto de la plantadora, cada tratamiento de semilla puede asignarse a una unidad sembradora diferente en cada uno de los cuatro bloques, de modo que cada tratamiento es sembrado el mismo número de veces por cada unidad de siembra, como en la figura 7.2.

Cuando el número de tratamientos se reduce y existen buenas razones para creer que se registrará un apreciable efecto de columna, la variación puede eliminarse en dos direcciones mediante la utilización de dos cuadros latinos (con asignaciones aleatorias independientes), como en la figura 7.3.

RESUMEN

El problema de decidir qué medias del tratamiento son significativamente diferentes recibe el nombre de separación de medias. Existen tres enfoques generales para la separación de medias: la aplicación de las diferencias significativas mínimas; la utilización de las pruebas de rango múltiple, y las pruebas F planeadas.

La diferencia significativa mínima se calcula como sigue:

$$DSM = t\sqrt{\frac{2(CME)}{r}}$$ donde t es un valor tabulado elegido para los grados de libertad del error y el nivel de significación deseado, CME = cuadrado medio para el error, y r = número de elementos sobre los cuales se basan las medias que se van a separar. Para separar dos medias basadas en un número desigual de observaciones, aplicamos la fórmula siguiente:

$$DSM = t\sqrt{\frac{CME}{r_1} + \frac{CME}{r_2}}$$

La prueba de rango múltiple de Duncan es la más popular entre un grupo de pruebas de rango disponibles; se calcula como DSMn = R (DSM), donde R = valor tabular para grados de libertad del error, nivel de significación y distancia de separación entre dos medias en un arreglo de medias de tratamientos. DSM es la diferencia significativa mínima.

Por regla general, las pruebas F planeadas ofrecen el procedimiento más preciso para la separación de medias. Estas pruebas pueden formular y responder a tantas preguntas independientes como grados de libertad existan para los tratamientos. Las preguntas deben planearse antes de que se realice el experimento.

La suma de cuadrados para un solo grado de libertad puede calcularse a partir de un conjunto de coeficientes cuya suma es igual a cero, mediante la fórmula:

$$SC = \frac{[\Sigma(c_i T_i)]^2}{r(\Sigma c_i^2)}$$, donde c_i es el conjunto de coeficientes, T_i es el conjunto de totales del tratamiento, y r es el número de elementos que integran cada T_i.

Dos comparaciones son ortogonales (independientes) si los productos de los coeficientes correspondientes arrojan, al ser sumados, un total igual a cero.

Si se efectúan tantas comparaciones ortogonales como grados de libertad existan para los tratamientos, sus sumas de cuadrados se adicionarán a la suma de cuadrados para los tratamientos.

Los coeficientes para medir tendencias pueden obtenerse de la tabla A. 11, si los niveles de tratamiento están igualmente espaciados.

Las comparaciones indican una disminución lineal altamente significativa en la germinación a medida que aumenta la dosis de insecticida en el caso en que no son tratadas con el fungicida; no hubo una respuesta de tendencia significativa cuando se empleó fungicida.

El paso final consiste en calcular una función lineal dosis-respuesta de cuadrados mínimos para el efecto de los tratamientos de semilla con insecticida, en ausencia del tratamiento con fungicida. Siguiendo el método del capítulo 15, $\hat{Y}_L = \overline{Y} + (K_2 P_1) X'$, donde $\hat{Y}_L$ = germinación estimada/100 semillas; $\overline{Y}$ = germinación media a partir de las dosis de insecticida F_0I_0, F_0I_1, F_0I_2; K_2 proviene de la tabla A. 11; $P_1 = \Sigma(c_i Y_i)$; y X' son valores codificados de las dosis de insecticida: $-1, 0, +1$. Se deja al lector la tarea de verificar que $\hat{Y}_L = 58.3 - 9.7X'$ y que, en términos de los niveles efectivos de dosificación del insecticida (es decir, $X = 0, \frac{1}{6}, \frac{1}{3}$), $\hat{Y}_L = 68.0 - 58.2X$.

Tabla 6.5. Coeficientes de comparación destinados a determinar funciones de respuesta del surgimiento de retoños del frijol de media luna para niveles de dosificación de insecticida.

	Tratamientos y totales del tratamiento					
Comparación	F_0I_0 341	F_0I_1 290	F_0I_2 244	F_2I_0 446	F_2I_1 459	F_2I_2 460
Insecticida lineal: F_0	+1	0	−1	0	0	0
Insecticida residual: F_0	+1	−2	+1	0	0	0
Insecticida lineal: F_2	0	0	0	+1	0	−1
Insecticida residual: F_2	0	0	0	+1	−2	+1

A continuación se calculan las sumas de cuadrados para dichas comparaciones. Puesto que cada comparación involucra un solo grado de libertad, nuevamente podemos utilizar la siguiente ecuación:

$$SC = \frac{[\Sigma(c_i T_i)]^2}{r(\Sigma c_i^2)}.$$ Por tanto:

$$SC(IL:F_0) = \frac{(341 - 244)^2}{5(2)} = 940.9$$

$$SC(IR:F_0) = \frac{[341 - 2(290) + 244]^2}{5(6)} = 0.83$$

$$SC(IL:F_2) = \frac{(446 - 460)^2}{5(2)} = 19.6$$

$$SC(IR:F_2) = \frac{[446 - 2(459) + 460]^2}{5(6)} = 4.8$$

Los cuadrados medios son iguales a las sumas de cuadrados, puesto que cada uno está basado en un solo grado de libertad, y los valores F se han calculado mediante la división de cada uno de los cuadrados medios entre el CME, como en la tabla 6.6

Tabla 6.6. Cuadrados medios y valores F para probar la significación de las respuestas a los niveles de dosificación de insecticida.

Fuente de variación	gl	CM	F observado	F requerido 5%	F requerido 1%
Insecticida lineal: F_0	1	940.9	24.43	4.35	8.10
Insecticida residual: F_0	1	0.8	0.02		
Insecticida lineal: F_2	1	19.6	0.51		
Insecticida residual: F_2	1	4.8	0.12		
Error	20	38.51			

Tabla 6.3. Retoños de frijol de media luna surgidos a partir de 100 semillas sembradas por parcela. F_0 y F_2 son iguales a cero y 2⅔ onzas de fungicida/100 lb de semillas, respectivamente. I_0, I_1, e I_2 son iguales a 0, ⅙, y ⅓ onzas de insecticida/100 lb de semillas, respectivamente.

Tratamiento		Bloques I	II	III	IV	V	T_t	$\bar{X}_t$
F_0	I_0	55	69	71	78	68	341	68.2
F_0	I_1	65	47	55	64	59	290	58.0
F_0	I_2	47	37	58	48	54	244	48.8
							T_{F_0} = 875	58.3 = $\bar{X}_{F_0}$
F_2	I_0	91	76	92	92	95	446	89.2
F_2	I_1	85	93	97	88	96	459	91.8
F_2	I_2	84	94	94	96	92	460	92.0
							T_{F_2} = 1 365	91.0 = $\bar{X}_{F_2}$
T_b		427	416	467	466	464	2 240	74.7 = $\bar{X}$

	Niveles de insecticida I_0	I_1	I_2
Totales (T_I)	787	749	704
Medias ($\bar{X}_I$)	78.7	74.9	70.4

$C = (2\,240)^2/30 = 167\,253.33$

$\Sigma X^2 = 177\,394$

que fueron tratadas con el fungicida. Estas cuatro preguntas son las comparaciones de la primera columna de la tabla 6.5. Los coeficientes para calcular las sumas de cuadrados para las comparaciones se han tomado de la tabla A.11, para n = 3 y los tres niveles de dosificación de insecticidas.

Tabla 6.4. Análisis de varianza del surgimiento de retoños de frijol de media luna.

Fuente de variación	gl	SC	CM	F observado	F requerido 5%	F requerido 1%
Total	29	10 140.67				
Bloques	4	401.00	100.25			
Tratamiento	5	8 969.47	1 793.89			
Fungicida	1	8 003.33	8 003.33	207.82	4.35	8.10
Insecticida	2	345.27	172.64	4.48	3.49	5.85
F x I	2	620.87	310.44	8.05		
Error	20	770.20	38.51			

Suma de Cuadrados para I × S

$$SC(I \times S) = SCT - SCI - SCS = 208 - 144 - 64 = 0, \text{ o, mediante coeficientes,}$$

$$SC(I \times S) = \frac{[1(212) - 1(236) - 1(228) + 1(252)]^2}{4(4)} = \frac{(0)^2}{16} = 0$$

Nótese que si aplicamos el método de los coeficientes los cálculos resultan mucho más sencillos que aquellos que implica el procedimiento del término de corrección. En cada caso, sólo necesitamos elevar al cuadrado un número, en vez de sumar los cuadrados de dos números grandes; no se requiere un término de corrección. Nótese también que la suma de las tres sumas de cuadrados componentes, es exactamente igual a la suma de cuadrados de tratamientos calculada en la forma usual, lo que suministra un medio de control sobre los cálculos.

Puesto que cada suma de cuadrados sólo tiene un grado de libertad, el cuadrado medio en cada caso es el mismo que la suma de cuadrados.

Las pruebas F se efectúan mediante la división de cada cuadrado medio entre el CME. Nótese que a través de la utilización de estas pruebas F más sensibles, hemos aprendido algo que ni la prueba de la DSM ni la de Duncan nos revelaron. Contamos ahora con buena evidencia de que las ovejas machos aumentaron más rápidamente de peso que las ovejas hembras.

Para analizar otro ejemplo de separación de la suma de cuadrados de tratamientos en componentes previamente planeados para pruebas F, véase la sección sobre Separación de Medias, en el capítulo 7.

Comparaciones de tendencia

Frecuentemente resulta deseable estudiar un factor a diversos niveles; por ejemplo, incrementos de un fertilizante, fechas de cosecha, dosis de un pesticida o de un herbicida, etc. En estos casos, el experimentador está interesado en la naturaleza de la respuesta de las unidades experimentales para niveles variables de un tratamiento. El análisis estadístico debe diseñarse para evaluar la tendencia de la respuesta.

Siempre que sea posible, resulta aconsejable aplicar los valores de un factor en una progresión aritmética, ya que esto permite utilizar métodos abreviados de cálculo (estudiados ampliamente en el capítulo 15) para estimar funciones de regresión apropiadas; por ejemplo, en el experimento de tratamiento de semillas de frijol de media luna resumido en la figura 3.2, uno de los objetivos consistió en determinar la naturaleza de la respuesta a las dosis de insecticida con y sin un tratamiento fungicida. Una parte de los datos provenientes de dicho experimento se presenta en la tabla 6.3, mientras que el análisis de varianza se resume en la tabla 6.4. Nótese que las dosis del insecticida forman una serie lineal: 0, $\frac{1}{6}$, $\frac{1}{3}$. La interacción significativa (F × I) indica que la respuesta a los niveles de dosificación del insecticida depende de que las semillas hayan sido tratadas o no con el fungicida.

La existencia de la interacción justifica el examen de la naturaleza de dichas respuestas. El primer paso consiste en elaborar una tabla de coeficientes de comparación (tabla 6.5) para simplificar el cálculo de las sumas de cuadrados deseadas. Nótese (tabla 6.4) que existen cuatro grados de libertad asociados con el insecticida y con el F × I. Entonces, podemos formular cuatro preguntas independientes (ortogonales) para indicar la naturaleza de la interacción. Con semillas no tratadas con fungicida, ¿existe evidencia para una disminución lineal de la germinación para dosis cada vez mayores de insecticida?; ¿existe evidencia para otra función que describa mejor la respuesta? Estas mismas preguntas pueden también formularse para semillas

Tabla 6.2. Partición ortogonal de los tratamientos del experimento de implantación de hormonas en ovejas.

Fuente de variación	gl	SC	CM	F observado	F requerido 5%	F requerido 1%
Tratamiento	3	208	69.33	8.91	3.86	6.99
Implantaciones	1	144	144	18.51	5.12	10.56
Sexo	1	64	64	8.23		
I × S	1	0	0	0		
Error	9	70	7.78			

La suma de cuadrados para tratamientos se separa como en la tabla 6.2. A fin de calcular dichas sumas de cuadrados y los valores F para contestar cada una de nuestras preguntas, procedemos como sigue:

Para calcular las sumas de cuadrados de los componentes de tratamientos, utilizaremos primero el procedimiento del término de corrección y luego ilustraremos el empleo del método de los coeficientes de comparación, que hemos agrupado en la tabla 6.1. Este último procedimiento para calcular una suma de cuadrados solamente funciona cuando la suma de cuadrados involucra un solo grado de libertad.

Suma de Cuadrados para la Implantación

$$SCI = \frac{(212 + 228)^2 + (236 + 252)^2}{8} - \frac{(928)^2}{16} = 144$$

Al utilizar los coeficientes de comparación, aplicamos la siguiente regla para calcular una suma de cuadrados:

$$SC = \frac{\left[\Sigma\left(c_i T_i\right)\right]^2}{r\left(\Sigma c_i^2\right)}$$, donde c_i = coeficientes de comparación de la tabla 6.1, T_i = totales del tratamiento, y r = número de repeticiones.

$$SCI = \frac{[1(212) - 1(236) + 1(228) - 1(252)]^2}{[(1)^2 + (-1)^2 + (+1)^2 + (-1)^2]\,4} = \frac{(-48)^2}{4(4)} = 144$$

Suma de Cuadrados para el Sexo

$$SCS = \frac{(212 + 236)^2 + (228 + 252)^2}{8} - \frac{(928)^2}{16} = 64,$$ o, por el método del coeficiente,

$$SCS = \frac{[1(212) + 1(236) - 1(228) - 1(252)]^2}{4(4)} = \frac{(-32)^2}{16} = 64$$

Tabla 6.1. Coeficientes de comparación.

Comparación	Tratamientos y totales del tratamiento			
	FS_0 212	FS_3 236	MS_0 228	MS_3 252
I ← Implantación	+1	−1	+1	−1
S ← Sexo	+1	+1	−1	−1
I × S	+1	−1	−1	+1

En lo que a la implantación se refiere, estamos comparando a todas las ovejas de ambos sexos tratadas con estilbestrol, con todas aquellas ovejas no tratadas, a fin de comparar el efecto promedio de la implantación. Esta es una comparación válida, puesto que grupos iguales de ovejas machos y hembras recibieron cada nivel de estilbestrol.

El cuanto al aumento de peso para cada sexo, estamos comparando la tasa promedio de aumento de todas las ovejas hembras con el de todas las ovejas machos para ambos niveles de estilbestrol. Esta también es una comparación válida, puesto que fueron tratados grupos iguales de ovejas de cada sexo.

Si la implantación diera lugar a una tasa de aumento significativamente mayor en un sexo que en el otro, diríamos que existe una considerable interacción entre el sexo de las ovejas y la implantación. Los coeficientes para esta comparación (I × S) se determinan mediante la multiplicación de los coeficientes para cada tratamiento en las dos primeras líneas de la tabla 6.1.

Las comparaciones son independientes, y por lo mismo ortogonales, cuando: a) la suma de los coeficientes de la comparación es igual a cero; y b) la suma de los productos de los coeficientes correspondientes de dos comparaciones cualesquiera es igual a cero.

Los estudiantes que analizan por primera vez este método, frecuentemente se confunden en cuanto al origen de los coeficientes. Las siguientes reglas son de utilidad en ese sentido:

1. Si se van a comparar dos grupos de igual tamaño, simplemente asígnense coeficientes + 1 a los miembros de un grupo y −1 a los integrantes del otro. No importa a qué grupo se asignen los coeficientes positivos.
2. En la comparación de grupos que contienen distintos números de tratamientos, asígnense al primer grupo tantos coeficientes como número de tratamientos tenga el segundo grupo; y a este último, tantos coeficientes del signo opuesto como número de tratamientos tenga el primer grupo. Entonces, si entre cinco tratamientos se quiere comparar los dos primeros con los tres últimos, los coeficientes serían +3, +3, −2, −2, −2.
3. Redúzcanse los coeficientes a los enteros más pequeños posibles; por ejemplo, en la comparación de un grupo de dos tratamientos con un grupo de cuatro tendremos, en virtud de la regla 2, coeficientes +4, +4, −2, −2, −2, −2, pero éstos pueden reducirse a: +2, +2, −1, −1, −1, −1.
4. Los coeficientes de interacción siempre pueden determinarse mediante la multiplicación de los coeficientes correspondientes de los efectos principales.

d) Indicamos la significación estadística mediante líneas o letras:

MS_3	FS_3	MS_0	FS_0	o	MS_3	FS_3	MS_0	FS_0
63	59	57	53		63a	59ab	57bc	53c

Las medias conectadas por la misma línea o seguidas por una letra común no son significativamente diferentes en el nivel de 5%. Si se utilizan letras, pueden verse más claramente las diferencias significativas, incluso cuando las medias no están ordenadas.

En nuestro experimento, nótese que las comparaciones de medias a través de la prueba de rango múltiple de Duncan o de DSM conducen a las mismas conclusiones ($MS_3 > MS_0$ y $FS_3 > FS_0$) aunque ambas pruebas llevan a la conclusión de que no existen diferencias significativas de aumento de peso entre machos y hembras ($MS_3 \not> FS_3$ y $MS_0 \not> FS_0$).

ANÁLISIS FUNCIONAL DE VARIANZA —PRUEBAS F PLANEADAS

En la planificación de un experimento, frecuentemente podemos preparar pruebas F para responder a preguntas pertinentes. Esto involucra la partición de los grados de libertad y de la suma de cuadrados para los **tratamientos** en comparaciones componentes. Los componentes pueden consistir en comparaciones de clase o de tendencia de respuesta. Estas pueden probarse mediante la partición de los grados de libertad y de la suma de cuadrados, para efectos del tratamiento, en grados de libertad apropiados y en sumas de cuadrados asociadas. Los tratamientos hábilmente seleccionados pueden responder a tantas preguntas como grados de libertad existan. Cuando son independientes, las comparaciones reciben el nombre de ortogonales; ésta es una característica deseable, pues entonces las comparaciones conducen a afirmaciones de probabilidad bien definidas.

La eficacia y sencillez de este método de separación de medias no son tan apreciadas entre los investigadores como deberían de serlo. Quizás el término **coeficientes ortogonales** cree la impresión de que el método es complicado y difícil; esto dista de ser cierto. En realidad, el método tiene tres importantes ventajas: a) nos permite responder importantes preguntas específicas sobre los efectos del tratamiento, b) los cálculos son sencillos, y c) suministra un útil control sobre la suma de cuadrados del tratamiento.

La construcción de una tabla de coeficientes de comparación resulta útil para comprobar la ortogonalidad, así como en el cálculo de las sumas de cuadrados componentes.

Comparaciones de clase

Para ilustrar lo anterior, nuevamente utilizaremos el experimento de implantación de hormonas en ovejas. Nótese que en la selección de los tratamientos se plantearon tres preguntas específicas: a) considerando a todas las ovejas, ¿afecta la implantación a la capacidad de aumento de peso?, b) ¿existen diferencias en la capacidad de aumento de peso entre las ovejas machos y hembras?, y c) ¿es el efecto de la implantación de hormonas, el mismo para ambos sexos? La respuesta a cada una de estas preguntas involucra un solo grado de libertad. Los coeficientes para las tres comparaciones están dados en la tabla 6.1.

PRUEBA DE RANGO MÚLTIPLE DE DUNCAN

Esta prueba es la más ampliamente utilizada entre diversas pruebas de rango múltiple disponibles. Evita la comisión de errores inherentes al empleo indiscriminado de la prueba de DSM. La prueba es idéntica a la de DSM para medias adyacentes de un arreglo ordenado, pero requiere valores progresivamente mayores para la significación entre medias, en la medida en que éstas se encuentran más ampliamente separadas en el arreglo. Esta prueba se utiliza más apropiadamente cuando diversos tratamientos no relacionados se incluyen en un experimento (por ejemplo, para efectuar todas las comparaciones posibles entre las capacidades de producción de diversas variedades). Para ilustrar su aplicación, emplearemos el experimento de implantación de hormonas en ovejas.

La prueba incluye el cálculo de las diferencias significativas mínimas (DSMn) para todas las posiciones relativas posibles entre las medias del tratamiento cuando éstas se encuentran dispuestas en orden de magnitud. Las DSMn se utilizan entonces en un procedimiento ordenado para determinar diferencias estadísticas entre las medias. La fórmula DSMn = R(DSM), donde R es un valor extraído de una tabla de **factores studentizados significativos** (tabla A. 4), que se elige de acuerdo con el nivel de significación deseado, con los grados de libertad para el error y con la disposición relativa de las medias en el arreglo.

Utilizando como ejemplo nuestro experimento con ovejas, el procedimiento es como sigue:

a) Calculamos la diferencia significativa mínima:

$$DSM_{.05} = t\sqrt{\frac{2s^2}{r}} = 2.262\sqrt{\frac{2(7.78)}{4}} = 4.46$$

b) Calculamos la DSMn para la posición relativa en el arreglo ordenado de medias. Puesto que existen cuatro medias, éstas pueden encontrarse separadas por 2, 3 o 4 valores. (Nota: las medias adyacentes se denominan separadas por 2).

Posición relativa en la ordenación (p de la tabla A. 4)	2	3	4
Valores de R, nivel de 5%, tabla A. 4	1.00	1.04	1.07
DSMn = R(DSM)	4.5	4.6	4.8

c) Disponemos las medias por orden de magnitud y probamos para diferencias significativas:

Tratamiento	FS_0	MS_0	FS_3	MS_3
Media	53	57	59	63

Empezamos por comparar las medias mayor y menor, aplicando la DSMn para sus posiciones relativas a cada una de las demás en el arreglo ordenado (en este caso, p =4, por tanto: DSMn = 4.8). Si la diferencia entre dichas medias es igual o mayor que la DSM, las medias serán significativamente diferentes. (63 – 53 = 10, DSMn = 4.8, por tanto, 63 es significativamente diferente de 53). A continuación comparamos la media mayor con la penúltima de las menores 57 = 6, DSMn = 4.6; 63 es significativamente mayor que 57). Luego, la mayor con la penúltima menor (63 — 59 = 4, DSMn = 4.5; 63 no es significativamente diferente de 59). **Cuando se encuentre una diferencia no significativa, podemos trazar una línea conectando (e interviniendo) dichas medias.** Luego repetimos el proceso; empezamos comparando la segunda media mayor con la menor, etc. Este procedimiento evita la realización de pruebas entre medias que están ya conectadas por una línea. Existe una **regla de excepción** utilizada con la prueba de rango múltiple de Duncan. Esta establece que una diferencia entre dos medias no puede declararse significativa si ambas medias están contenidas en un subconjunto de medias con un rango no significativo. Entonces, si entre cinco medias en un arreglo ordenado, A se ha encontrado no significativamente diferente de D (es decir, A B C D E) y B es significativamente diferente de E, no resulta necesario probar B contra D y C, puesto que éstas se encuentran en un subconjunto con un rango no significativo. El próximo paso consistirá en probar C contra E y si la diferencia no es significativa, C y E se conectan, A B C D E, resultando innecesarias pruebas ulteriores.

DIFERENCIA SIGNIFICATIVA MÍNIMA

Esta prueba no debe utilizarse, a menos que la prueba F sea significativa. Estrictamente hablando, la DSM sólo debe emplearse para comparar medias adyacentes en un arreglo ordenado (medias dispuestas por orden de magnitud). Cuando éste se usa indiscriminadamente para probar todas las posibles diferencias entre diversas medias, ciertas diferencias serán significativas, pero no en el nivel de significación que hemos escogido. En vez de efectuarse el nivel de 5%, las comparaciones entre medias con una separación mayor de dos en un arreglo ordenado, se realizarán en un nivel de significación más bajo. La DSM puede utilizarse para comparar medias adyacentes, y la mayoría de los especialistas coinciden en que ésta es adecuada para comparar un **tratamiento estándar** con otros tratamientos; tal es el caso de la comparación de variedades con una variedad testigo o estándar.

No obstante la gran cantidad de críticas, la DSM aún se utiliza ampliamente en las revistas profesionales. Si se emplea con cuidado, no deberá conducir a demasiados errores. La gran ventaja de la DSM es que resulta fácil de calcular y emplea un estimador único para efectuar comparaciones.

Como se indicó con anterioridad, la DSM es una forma de la prueba t. Su fórmula se deriva de la fórmula de t en la prueba de la significación estadística de la diferencia entre dos medias,

$t = \dfrac{\bar{d} - \mu_{\bar{d}}}{s_{\bar{d}}}$ Sea la diferencia entre dos medias ($\bar{X}_1 - \bar{X}_2 = \bar{d}$) el límite inferior de los valores que cabría esperar en un 5% o más de las veces, sólo por casualidad, al obtener muestras de diferencias de medias pertenecientes a una población de diferencias de medias, donde la media es igual a cero ($\mu_{\bar{d}} = 0$). Sustituimos $\bar{d}$ por la DSM y $\mu_{\bar{d}}$ por cero, y la fórmula se transforma en: $t = \dfrac{DSM}{s_{\bar{d}}}$, resolviendo para DSM:

$DSM = t\, s_{\bar{d}}$, donde $s_{\bar{d}} = \dfrac{s_1^2}{r_1} + \dfrac{s_2^2}{r_2}$, s_1^2 y s_2^2 son las varianzas estimadas de parcelas que reciben los tratamientos 1 y 2 respectivamente, y r_1 y r_2 son los números de unidades experimentales que reciben los tratamientos 1 y 2 respectivamente. En un análisis de varianza se supone que s_1^2 estima la misma varianza que s_2^2 y que r_1 es, por regla general, igual a r_2, por tanto, $DSM = t\sqrt{\dfrac{2s^2}{r}}$, donde s^2 es el cuadrado medio para el error, r es el número de repeticiones, y t es el valor tabular de t para los grados de libertad del error.

A fin de ilustrar la utilización de la DSM, la aplicaremos para separar las medias de nuestro experimento con ovejas y estilbestrol, tabla 5.2 (capítulo 5). Los efectos medios son: $FS_0 = 53$; $MS_0 = 57$; $FS_3 = 59$; $MS_3 = 63$ libras aumentadas por oveja por cada 100 días.

$$DSM_{.05} = t_{.05}\sqrt{\frac{2s^2}{r}} = 2.262\sqrt{\frac{2(7.78)}{4}} = 2.262\,(1.972) = 4.46 \text{ libras por animal por cada 100 días.}$$

Si utilizamos la DSM sólo para comparar medias adyacentes, concluimos que no hay diferencias; pero el valor F revela que éstas sí existen. Si empleamos la DSM para comparar todas las medias, llegamos a la conclusión de que el estilbestrol mejoró la capacidad de aumento de peso tanto en las ovejas hembras (59 – 53 = 6) como en los machos (63 – 57 = 6). Las diferencias en la capacidad de aumento de peso asociadas con el sexo no son significativas.

Capítulo 6

Separación de medias

Como hemos visto, un experimento se lleva a cabo para responder ciertas preguntas que el investigador plantea por adelantado. Estas preguntas resultan importantes para la determinación de los tratamientos que se van a incluir en el experimento y para la elección del método apropiado, mediante la comparación de los tratamientos. Una cuidadosa consideración de los siguientes principios para decidir sobre los tratamientos, puede ayudarnos a seleccionar un plan para la separación de medias.

El arreglo factorial. Cuando se han de estudiar dos o más factores simultáneamente, el arreglo factorial hace posible el estudio de las interacciones entre factores. Cuando las interacciones no están presentes, los resultados son más ampliamente aplicables, puesto que los tratamientos principales se prueban para un rango de condiciones más amplio.

La selección de niveles de tratamiento para obtener información sobre tendencias o respuestas óptimas. Cuando los tratamientos consisten en dosis, como los niveles de nutrientes de una planta, de una hormona de crecimiento, de un fungicida o de un herbicida, la pregunta fundamental no es si alguna dosis resulta **significativamente** mejor que otras, sino si existe o no una respuesta al factor estudiado y cómo se caracteriza mejor esa respuesta. Una atención cuidadosa en la selección de los niveles de dosificación, es importante para caracterizar una función dosis-respuesta.

La utilización de un tratamiento estándar. Puede resultar deseable comparar cada uno de los diversos tratamientos con un tratamiento estándar a fin de decidir si cualesquiera de los nuevos tratamientos son mejores que el que se está utilizando actualmente.

Cuando estos principios no pueden emplearse para establecer tratamientos y no existen relaciones lógicas entre los mismos, se requieren pruebas de comparación múltiple para la separación de medias. Cuando pueden aplicarse uno o más de los principios, suele haber métodos para la separación de medias más precisos que aquellos suministrados por las pruebas de comparación múltiple.

Un valor F significativo plantea la siguiente pregunta: ¿Cuáles de los valores medios son significativamente diferentes? A continuación se estudiarán tres métodos ampliamente utilizados en la separación de medias.

Nótese que el total de grados de libertad $(rn - 1)$ debe reducirse mediante los grados de libertad implicados en la remoción de los efectos de bloque $(r - 1)$ y de tratamiento $(n - 1)$

$$\text{Así, } CME = \frac{(58-58)^2 + (59-58)^2 + \ldots + (56-58)^2}{[4(4)-1] - (4-1) - (4-1)} = \frac{70}{9} = 7.78$$

VALORES F

Las razones F se utilizan para evaluar las probabilidades de obtención de medias del tratamiento y del bloque que varíen tanto como aquellas de nuestro experimento, si no existen tratamientos reales o diferencias de bloque. Hemos estimado σ^2, la varianza de la población por unidad experimental, en tres formas: a) basada en la variación entre medias del tratamiento (CMT), b) basada en la variación entre medias del bloque (CMB), y c) basada en la variabilidad entre las unidades experimentales, una vez eliminados los efectos de bloques y de tratamientos, (CME). Si no existen diferencias debidas a los bloques y a los tratamientos, la totalidad de los tres cuadrados medios deben ser aproximadamente iguales.

$$F \text{ (bloques)} = CMB/CME = 192.0/7.78 = 24.69$$

$$F \text{ (tratamientos)} = CMT/CME = 69.3/7.78 = 8.91$$

Los valores F requeridos para la significación estadística, para grados de libertad 3 (numerador) y 9 (denominador), se encuentran en la tabla A.3 y se han anotado en la tabla de análisis de varianza (tabla 5.3). Puesto que nuestro valor de F observado para los bloques, así como para los tratamientos, excede al requerido para la significación en el nivel del 1%, podemos decir que si la hipótesis nula es verdadera, las probabilidades son menores que 1 entre 100, de que nuestra muestra particular de bloques o tratamientos pueda haber ocurrido sólo por casualidad. Estamos dispuestos a afirmar que estas probabilidades no se registran, a rechazar la hipótesis nula y a concluir que existen diferencias reales entre bloques y tratamientos. El próximo paso consiste en determinar cuáles de los tratamientos son significativamente diferentes. Esto se estudiará en el capítulo 6.

Antes de concluir el diseño de bloques completos al azar, debemos comentar el mejoramiento en la eficiencia con respecto al diseño completamente aleatorio. Debido a la existencia de considerables diferencias de bloque y a la eliminación de los efectos de bloque, la precisión de nuestro experimento se vio incrementada, permitiéndonos detectar diferencias entre tratamientos las cuales no podrían advertirse mediante el diseño completamente aleatorio.

RESUMEN

En el diseño del bloques completos al azar, los bloques son conjuntos de unidades experimentales dispuestas o seleccionadas con anterioridad a la asignación de tratamientos, de modo que la variabilidad existente es minimizada dentro de los bloques y maximizada entre los mismos. Los tratamientos se asignan aleatoriamente el mismo número de veces (usualmente una vez) a las unidades experimentales dentro de un bloque. Una selección aleatoria independiente se realiza para cada bloque. Comparados con el diseño completamente aleatorio, los grados de libertad para el error experimental se ven reducidos por el número de grados de libertad para los bloques. La variabilidad del bloque se elimina a partir del error experimental. Así, cuanto mayor sea la variabilidad entre bloques, más eficiente será el proyecto en lo que se refiere a su capacidad para detectar posibles diferencias del tratamiento.

Tabla 5.4. Elementos de la tabla 5.2, con los efectos de bloque eliminados.

Tratamiento		Bloque				Tratamiento	
		I	II	III	IV	Total	Media
A	FS_0	53	54	52	53	212	53
B	MS_0	56	56	57	59	228	57
C	FS_3	63	55	59	59	236	59
D	MS_3	60	67	64	61	252	63
Total de bloque		232	232	232	232	928	
Media del bloque		58	58	58	58		

57 – 58 = –1; C = 59 – 58 = 1 y D = 63 – 58 = 5. Ahora podemos eliminar los efectos de tratamiento de los elementos de la tabla 5.4 para completar la tabla 5.5: una tabla de elementos individuales corregidos para los efectos de bloque y de tratamiento. La primera celda de la tabla 5.5 es 53 – (–5) = 58 y la última es 61 – 5 = 56. Las medias del tratamiento ahora también son iguales a la media general. La variabilidad subsistente entre los elementos de la tabla 5.5 es una variabilidad no asociada con los efectos de bloque o de tratamiento. La suma de cuadrados $[\Sigma(X - \bar{X})^2]$ entre dichos elementos dividida entre los grados de libertad apropiados, se denomina error experimental (CME o s^2).

Tabla 5.5. Elementos de la tabla 5.2, con los efectos de bloque y del tratamiento eliminados.

Tratamiento		Bloque				Tratamiento	
		I	II	III	IV	Total	Media
A	FS	58	59	57	58	232	58
B	MS	57	57	58	60	232	58
C	FS	62	54	58	58	232	58
D	M	55	62	59	56	232	58
Total del bloque		232	232	232	232	928	
Media del bloque		58	58	58	58		

A partir de la tabla 5.5, calculamos

$$s^2 = CME = \frac{\Sigma(X - \bar{X})^2}{(rn - 1) - (r - 1) - (n - 1)}$$

$s^2 = n\ s^2_{\bar{X}}$. Puesto que $s^2_{\bar{X}_b} = \dfrac{\Sigma\left(\bar{X}_b - \bar{X}\right)^2}{r - 1}$, la fórmula para el CMB pasa a ser: $CMB = n\left[\dfrac{\Sigma\left(\bar{X}_b - \bar{X}\right)^2}{r - 1}\right]$

donde $\bar{X}_b$ representa cada media del bloque, $\bar{X}$ es la media general, y r es el número de medias del bloque. Calculando el CMB, obtenemos:

$$CMB = \frac{4\left[(52-58)^2 + (56-58)^2 + \ldots + (56-58)^2\right]}{4-1} = \frac{4(144)}{3} = 192.0$$

Cuadrado medio para los tratamientos

Recurriendo nuevamente a la hipótesis nula y suponiendo que no hay diferencias reales entre las medias del tratamiento, $s^2 = CMT = r\ s^2_{\bar{X}_t}$ donde r = número de repeticiones y $s^2_{\bar{X}_t}$ es la varianza de las medias de tratamiento. Esta es otra estimación de la varianza por unidad experimental, basada en la variabilidad entre medias del **tratamiento**.

Nuevamente se utiliza la relación entre una varianza de medias estimada ($s^2_{\bar{X}}$) y la varianza estimada de los elementos individuales de la población original (s^2). Desarrollando la fórmula, obtenemos:

$CMT = \dfrac{r\,\Sigma\left(\bar{X}_t - \bar{X}\right)^2}{n - 1}$, donde $\bar{X}_t$ = cada una de las medias del tratamiento, y n es el número de tratamientos. SCT es el numerador y el denominador está constituido por los gl para el tratamiento. Su cálculo da como resultado:

$$CMT = \frac{4\left[(53-58)^2 + (57-58)^2 + \ldots + (63-58)^2\right]}{4-1} = \frac{4(52)}{3} = 69.3$$

Cuadrado medio para el error

El CME representa la variabilidad entre las unidades experimentales que subsisten después de que las otras fuentes de variación han sido eliminadas. Resulta ilustrativo observar que está implicado en la **eliminación** de los efectos del bloqueo y del tratamiento.

Remoción de los efectos de bloque

Cada efecto de bloque se define como la media del bloque menos la media general (es decir $\bar{X}_b - \bar{X}$). Para el bloque I, el efecto de bloque es 52 – 58 = – 6. Si sustraemos —6 de cada elemento del bloque I en la tabla 5.2 obtendremos: 47 – (–6) = 53; 50 – (–6) = 56; 57 – (–6) = 63; 54 – (–6) = 60. Los efectos de bloque para los demás bloques son: II = 56 – 58 = –2; III = 68 – 58 = 10; y IV = 56 – 58 = –2. La **eliminación** de los efectos de bloque de los elementos de la tabla 5.2 da lugar a la tabla 5.4. Nótese que ahora todas las medias del bloque son iguales a la media general, pero que las medias del tratamiento permanecen inalteradas.

Remoción de los efectos de tratamiento

Un efecto de tratamiento se elimina mediante la sustracción de la desviación de la media del tratamiento de la media general de cada elemento de dicho tratamiento. Los efectos del tratamiento son: A = 53 – 58 = –5; B =

Sumas de cuadrados y cuadrados medios

Bloques. $SCB = \dfrac{\Sigma\left(T_b\right)^2}{n} - C$

$SCB = \dfrac{(208)^2 + \ldots + (224)^2}{4} - 53\,824 = 54\,400 - 53\,824 = 576.$ Nótese que en el término $\dfrac{\Sigma\left(T_b\right)^2}{n}$ el divisor n es el número de elementos que integran cada total en el numerador; en este caso, **el número de tratamientos.**

$CMB = SCB/gl\,(B) = 576/3 = 192.0$

Tratamientos. $SCT = \dfrac{\Sigma\left(T_t\right)^2}{r} - C$

$SCT = \dfrac{(212)^2 + \ldots + (252)^2}{4} - 53\,824 = 54\,032 - 53\,824 = 208$

$CMT = SCT/gl\,(T) = 208/3 = 69.3$

Total. $SC = \Sigma(X)^2 - C$

$SC = (47)^2 + (52)^2 + \ldots + (59)^2 - C = 54\,678 - 53\,824 = 854.$

Error $SCE = SC - SCT - SCB$

$SCE = 854 - 208 - 576 = 70.$

Si las diversas sumas de cuadrados se efectúan en el orden anterior, la SCE se obtiene rápidamente por sustracción, tan pronto como la suma total de cuadrados se obtenga.

$CME = SCE/gl\,(E) = 70/9 = 7.78$

EL QUÉ Y EL PORQUÉ DEL ANÁLISIS

Antes de continuar con otros aspectos del análisis de varianza, será de utilidad echar una ojeada a lo que se hizo e indagar por qué se hizo en el cálculo de cada cuadrado medio.

Cuadrado medio para los bloques

Suponiendo la ausencia de diferencias reales entre las medias del bloque (nuevamente la hipótesis nula), una estimación de la variabilidad por unidad experimental se calcula a partir de la varianza de las medias del **bloque.** Así, $s^2 = CMB = n\, s^2_{\bar{x}_b}$ donde n = número de tratamientos y $s^2_{\bar{x}_b}$ es la varianza de las medias del bloque. Nótese que esto aplica la relación entre una varianza de medias y la varianza por unidad experimental,

Tabla 5.2. Aumentos de peso de ovejas agrupadas por tratamiento y por bloque (libras por ovejas por 100 días).

Tratamiento		Bloque I	Bloque II	Bloque III	Bloque IV	Tratamiento total (T_t)	Tratamiento media ($\bar{X}_t$)
A	FS_0	47	52	62	51	212	53
B	MS_0	50	54	67	57	228	57
C	FS_3	57	53	69	57	236	59
D	MS_3	54	65	74	59	252	63
Total del bloque (T_b)		208	224	272	224	928 = ΣX	
Media del bloque ($\bar{X}_b$)		52	56	68	56	Media principal →	58 ($\bar{X}$)

Tabla 5.3. Analisis de varianza.

Fuente de variación	gl	SC	CM	F observado	F requerido 5%	F requerido 1%
Total	15	854				
Bloques	3	576	192.0	24.69	3.86	6.99
Tratamientos	3	208	69.3	8.91		
Error	9	70	7.78			

Fuentes de variación y grados de libertad

Tenemos ahora una fuente adicional de variación... aquella debida a los bloques. Puesto que cada tratamiento ocurre el mismo número de veces en cada bloque, las diferencias entre bloques no se deben a los tratamientos, sino a otras diferencias asociadas con los bloques. Este componente de la suma total de cuadrados puede eliminarse y el error inexplicado (error experimental) puede reducirse como corresponde.

Los grados de libertad son uno menos que el números de observaciones asociadas con cada fuente de variación. Existen 16 unidades experimentales (grupos de ovejas): por tanto, los grados de libertad serán 15. Existen cuatro bloques y cuatro tratamientos y, por ende, tres gl para cada fuente de variación. Los grados de libertad del error pueden encontrarse por sustracción, 15 – 3 – 3 = 9 o multiplicando los gl para los bloques por los gl para los tratamientos, 3 × 3 = 9. En este diseño, cuando cada tratamiento se repite una sola vez en cada bloque, los gl para el error son siempre: gl de los bloques × gl de los tratamientos.

Término de corrección

$$C = \frac{(\Sigma X)^2}{rn}$$ donde r = número de repeticiones y n = número de tratamientos.

$$C = \frac{(928)^2}{4(4)} = 53\,824$$

I	II	III	IV
D	A	C	C
A	D	D	B
B	C	B	D
C	B	A	A

Baja fertilidad ⟶ Alta fertilidad

Figura 5.1. Cuatro tratamientos repetidos igual número de veces en un diseño de bloques completos al azar.

bloque; por ejemplo, los cuatro tratamientos de la figura 5.1 fueron aleatoriamente asignados de la siguiente manera: empezando arbitrariamente con la hilera 15 de la tabla A.1, continuamos a lo largo de la misma hasta encontrar los dígitos del 1 al 4, que representan los tratamientos A al D: 4, 1, 2, 3 ... . Este es el orden en que asignaremos los tratamientos en el bloque I. Luego continuamos a través de la hilera 15 y regresamos (de derecha a izquierda) sobre la hilera 16 para encontrar los dígitos 1, 4, 3, 2, orden en que asignaremos los tratamientos en el bloque II. Análogamente, la distribución aleatoria se completa para los bloques III y IV.

ANÁLISIS DE VARIANZA

Los datos que analizaremos serán los mismos que utilizamos en el capítulo 4. El experimento buscaba determinar el efecto de la implantación de una hormona, el estilbestrol, sobre la capacidad para aumentar el peso de ovejas machos y hembras. Entonces, los tratamientos fueron el conjunto factorial de la tabla 5.1, siendo los dos factores el sexo y el estilbestrol, y teniendo cada factor dos niveles.

Tabla 5.1. Tratamientos para determinar el efecto que produce el estilbestrol implantado en la oreja sobre la capacidad para aumentar de peso, de ovejas machos y hembras.

Sexo	Estilbestrol	
	0	3 mg/animal
hembra	FS_0	FS_3
macho	CM_0	CM_3

En este caso, los **bloques** fueron cuatro ranchos diferentes. Por tanto, las **repeticiones** de la tabla 4.1 pasan a ser bloques y los **tratamientos** pasan a ser el conjunto factorial de la tabla 5.1. Los datos se reorganizan en la tabla 5.2.

Capítulo 5

Diseño en bloques completos al azar

En este diseño, los tratamientos se asignan aleatoriamente, a un grupo de unidades experimentales denominado bloque o repetición. Bloque es el término más adecuado, puesto que evita confusión con las repeticiones del diseño completamente aleatorio.

El objetivo consiste en mantener la variabilidad entre unidades experimentales dentro de un bloque tan pequeño como sea posible, y maximizar las diferencias entre bloques. Si no hay diferencia entre los bloques, este diseño no contribuirá a la precisión para detectar las diferencias de tratamientos.

Cada tratamiento es asignado el mismo número de veces a unidades experimentales dentro de un bloque, usualmente una vez; pero todos o ciertos tratamientos pueden repetirse dos o más ocasiones dentro de un bloque. Por regla general, es más eficiente tener una sola repetición de cada tratamiento por bloque. A fin de minimizar el error experimental, deben tomarse todas las precauciones para tratar las unidades experimentales dentro de un bloque lo más uniformemente posible.

Los bloques pueden estar constituidos por áreas compactas de un campo, grupos de animales que pueden manipularse de un modo uniforme, o diferentes tiempos de aplicación de tratamientos a unidades experimentales.

Por lo que respecta a sembradíos, las parcelas de campo adyacentes suelen producir en forma más parecida que aquellos separados por alguna distancia. Los bloques se pueden mantener compactos, disponiendo las parcelas, usualmente de forma larga y estrecha, cercanas unas a las otras. El número de tratamientos debe ser el menor posible; no obstante, debe ser suficiente para lograr los objetivos del experimento. Cuando el tamaño del bloque aumenta, se incrementará la variabilidad dentro de éste. No es necesario que cada bloque sea de la misma forma; pero en los experimentos de campo con sembradíos, esto es normalmente deseable, puesto que las diferencias en las formas de los bloques generalmente incrementan la variabilidad dentro del bloque.

Cuando se prevé un gradiente de productividad dentro del área experimental, los bloques deben extenderse a lo largo del gradiente y las parcelas dentro de un bloque deben disponerse paralelamente respecto al gradiente, como se ve en la figura 5.1

MUESTREO ALEATORIO

Después de que las unidades experimentales han sido agrupadas en los bloques deseados, los tratamientos se asignan aleatoriamente a las unidades dentro de cada bloque, con una distribución aleatoria hecha para cada

El denominador 6(30), se determina sumando los cuadrados de los coeficientes de los términos del numerador, y multiplicando el resultado por el número de elementos que integran cada término del numerador; así

$$[(5^2) + (-1)^2 + (-1)^2 + (-1)^2 + (-1)^2 + (-1)^2]\ 6 = 30(6).$$

$$\text{SC (N orgánico y N inorgánico)} = \frac{(188.9)^2}{6} + \frac{(186.1 + 182.1 + 183.8 + 182.2)^2}{24} - \frac{(188.9 + 186.1 + 182.1 + 183.8 + 182.2)^2}{30} = \frac{(188.9)^2}{6} + \frac{(734.2)^2}{24} - \frac{(923.1)^2}{30}$$

Nótese que el tercer término es un nuevo término de corrección:

$$= 5\,947.202 + 22\,460.402 - 28\,403.787 = 3.816.$$

El cálculo más corto es:

$$\frac{[4(188.9) - 186.1 - 182.1 - 183.8 - 182.2]^2}{6(20)} = 3.816$$

$$\text{SC}(NH_4 - \text{N vs } NO_3 - \text{N}) = \frac{(186.1 + 182.1)^2 + (183.8 + 182.2)^2}{12} - \frac{(186.1 + 182.1 + 183.8 + 182.2)^2}{24}$$

$$= \frac{(368.2)^2 + (366.0)^2}{12} - \frac{(734.2)^2}{24} = 22\,460.603 - 22\,460.402 = 0.201$$

o:

$$\frac{(186.1 + 182.1 - 183.8 - 182.2)^2}{6(4)} = \frac{(2.2)^2}{24} = 0.202$$

$$\text{SC}[(NH_4)_2SO_4 \text{ vs } NH_4NO_3] = \frac{(186.1)^2 + (182.1)^2}{6} - \frac{(186.1 + 182.1)^2}{12} = 11\,298.937 - 11\,297.603 = 1.334$$

o:

$$\frac{(186.1 - 182.1)^2}{2(6)} = \frac{(4.0)^2}{12} = 1.333$$

$$\text{SC } Ca(NO_3)_2 \text{ vs } NaNO_3 = \frac{(183.8)^2 + (182.2)^2}{6} - \frac{(183.8 + 182.2)^2}{12} = 11\,163.213 - 11\,163.000 = 0.213$$

o:

$$\frac{(183.8 - 182.2)^2}{2(6)} = \frac{(1.6)^2}{12} = 0.213$$

Los cuadrados medios se obtienen mediante la división de las sumas de cuadrados entre sus grados de libertad asociados; puesto que, en este caso, cada comparación involucra un solo grado de libertad, SC = CM.

Los valores F se calculan mediante la división de cada CM entre el CM para el error. Los valores F requeridos son valores tabulares de la tabla A.3 para 1 y 20 gl. Tenemos ahora una prueba F para responder cada una de las preguntas que se plantearon cuando se planeó el experimento. El único valor F significativo se emplea para la comparación de no N y N; todos los demás son bastante bajos, y llevan a la conclusión de que hubo una respuesta al nitrógeno, pero que las remolachas respondieron análogamente a todas las fuentes de N.

RESUMEN

En un cuadro latino:

Las unidades experimentales se clasifican en dos categorías, además de los tratamientos. Tales categorías se conocen generalmente como **hileras** y **columnas**, con respecto a la organización de los datos en una tabla de doble entrada.

Cada tratamiento se asigna el mismo número de veces (usualmente una vez) dentro de cada categoría, de modo que las diferencias entre categorías no se deban a efectos del tratamiento.

Se requieren al menos tantas repeticiones como tratamientos existan. Por regla general, los cuadros latinos dejan de ser prácticos con más de ocho tratamientos.

Sólo cuando ambas categorías (hileras y columnas) varían apreciablemente, el diseño de cuadro latino superará al de bloques completos al azar, en cuanto a la detección de diferencias de tratamientos.

Capítulo **8**

Diseño de parcelas divididas

Los diseños de parcelas divididas y una variante de éstos, el bloque dividido, se emplean frecuentemente en experimentos factoriales, en los que la naturaleza del material experimental o las operaciones contempladas dificultan el manejo de todas las combinaciones de factores en una misma forma. El diseño básico de parcelas divididas involucra la asignación de tratamientos de un factor a parcelas principales dispuestas en un diseño completamente aleatorio, de bloques completos al azar o de cuadro latino. Los tratamientos del segundo factor se asignan a subparcelas dentro de cada parcela principal. El proyecto suele sacrificar la precisión en la estimación de los efectos promedio de los tratamientos asignados a las parcelas principales, aunque frecuentemente aumenta la precisión para comparar los efectos promedio de tratamientos asignados a subparcelas; y cuando existen interacciones, para comparar los efectos de tratamientos de subparcelas en un tratamiento de una parcela principal dada. Esto proviene del hecho de que el error experimental para las parcelas principales suele ser mayor que el error experimental utilizado para comparar tratamientos de subparcelas. A menudo, el término de error para tratamientos de subparcelas es inferior al que se obtendría si todas las combinaciones de tratamientos fuesen dispuestas en un diseño de bloques completos al azar.

Cada variación del diseño de parcelas divididas impone ciertas restricciones en cuanto al término de error que puede utilizarse para probar los efectos del tratamiento. Es importante, entonces, asignar los factores en tal forma que obtengamos la mayor precisión al comparar las interacciones y los efectos promedio de los tratamientos en los que estamos más interesados.

MUESTREO ALEATORIO

La distribución aleatoria de los tratamientos asignados a las parcelas principales se lleva a cabo en la forma prescrita para el diseño seleccionado para los mismos. Los tratamientos de subparcelas serán distribuidos aleatoriamente dentro de cada parcela principal, efectuando una distribución aleatoria independiente para cada parcela principal.

ANÁLISIS DE VARIANZA

Un experimento diseñado para probar el efecto de tres cultivos de abono vegetal sobre la producción subsecuente de remolacha azucarera, con dos niveles de fertilización de nitrógeno, fue planificado con un diseño de parcela dividida, como el de la figura 8.1. Al principio se supuso que la remolacha azucarera respondería en diversas formas a los abonos vegetales, dependiendo del nivel de fertilidad del nitrógeno; por

tanto, el objetivo consistió en comparar tan precisamente como fuese posible el efecto de los abonos vegetales en cada nivel de fertilidad. En consecuencia, las parcelas principales hubieron de ser dos niveles de fertilización de nitrógeno, aplicados a la remolacha de azúcar en poco tiempo y repetidos tres veces en un proyecto de bloque aleatorio completo. Las subparcelas fueron los abonos vegetales que crecieron durante el otoño y el invierno anteriores a la siembra de la remolacha azucarera. Los tratamientos de abono vegetal fueron cebada (C), vicia (V), cebada y vicia creciendo juntas (CV) y barbecho (B). No se permitió que creciera nada en las parcelas en barbecho, antes de sembrar la remolacha azucarera. Las producciones de las parcelas de remolacha azucarera subsecuentes a los abonos vegetales se muestran en la figura 8.1 y se han ordenado para su análisis en la tabla 8.1.

El primer paso consiste en determinar las fuentes de variación y los grados de libertad asociados; véanse las dos primeras columnas de la tabla 8.2. Este y los pasos subsecuentes en el análisis se presentan en dicha tabla.

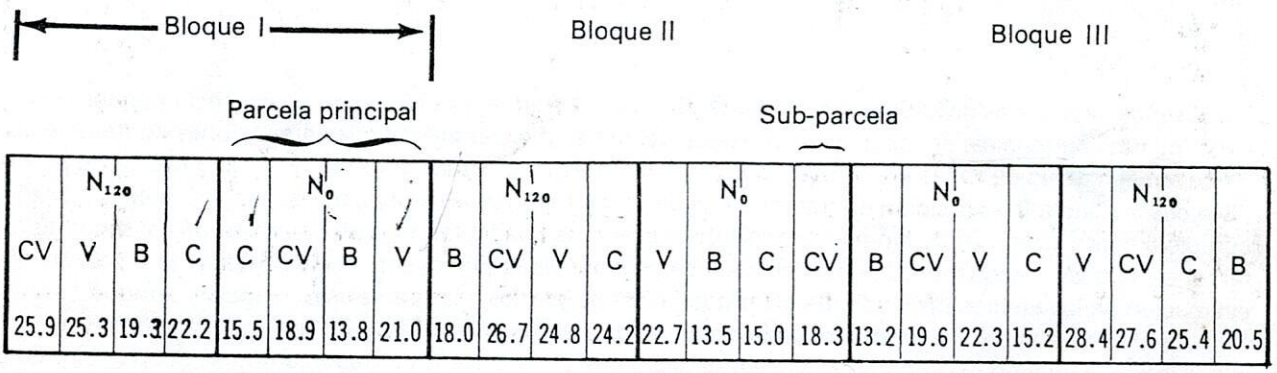


Figura 8.1. Proyecto de parcela dividida. Las parcelas principales, (N_{120}, N_0), son los niveles de fertilidad del nitrógeno. Las subparcelas CV, V, B, C son tratamientos de abono vegetal. Todas las parcelas están planificadas en franjas a través del terreno. Las producciones de las parcelas de cultivo de remolacha de azúcar a continuación de los tratamientos de abono vegetal están dadas en toneladas de raíces por acre.

Fuentes de variación y grados de libertad

Los grados de libertad totales del experimento son (niveles de N × tratamientos de AV × repeticiones) − 1 = (2 × 4 × 3) − 1. Esta variación total se denomina "parcelas de N × AV" o subparcelas. Los grados de libertad que involucran las parcelas principales, (N × repeticiones) − 1 = (2 × 3) − 1 se separan de acuerdo con el diseño en el cual éstos se encuentran dispuestos; en este caso en el de bloques completos al azar: gl (bloques) = 3 − 1; gl (N) = 2 − 1; y gl (Ea) = gl (parcelas de N) − gl (B) − gl (N) = 5 − 2 - 1.

Los grados de libertad para el AV = 4 - 1; para N y AV = glN × gl (AV) = 1 × 3; y para el error (b) = gl (parcelas de N y parcelas de AV) − gl (parcelas de N) − gl(AV) − gl (N y AV) = 23 − 5 − 3 − 3.

Término de corrección

$C = \dfrac{(\Sigma X)^2}{rn(AV)}$, donde r = número de repeticiones (3), n = número de niveles de N (2), y (AV) = número de tratamientos de abono vegetal (4).

$C = (497.3)^2/[3(2)\ (4)] = 10\ 304.47$

Tabla 8.1. Producciones de raíces de remolacha de azúcar (tons/acre) organizadas por tratamiento, parcela principal y bloques.

Tratamientos		Bloques			Tratamiento	
Libras de N/acre	Abono vegetal	I	II	III	Totales (T_t)	Medias ($\bar{X}_t$)
0	Barbecho	13.8	13.5	13.2	40.5	13.5
	Cebada	15.5	15.0	15.2	45.7	15.2
	Vicia	21.0	22.7	22.3	66.0	22.0
	Cebada-vicia	18.9	18.3	19.6	56.8	18.9
Totales de las parcelas principales (Tpp)		69.2	69.5	70.3	209.0 = Tn_1	17.4 = $\bar{X}n_1$
120	Barbecho	19.3	18.0	20.5	57.8	19.3
	Cebada	22.2	24.2	25.4	71.8	23.9
	Vicia	25.3	24.8	28.4	78.5	26.2
	Cebada-vicia	25.9	26.7	27.6	80.2	26.7
Totales de las parcelas principales (Tpp)		92.7	93.7	101.9	288.3 = Tn_2	24.0 = $\bar{X}n_2$
Totales del bloque (Tb)		161.9	163.2	172.2	497.3 = ΣX	20.7 = $\bar{X}$

Abonos vegetales

	B	C	V	CV
Totales (TAV)	98.3	117.5	144.5	137.0
Medias ($\bar{X}$AV)	16.4	19.6	24.1	22.8

Tabla 8.2. Análisis de varianza. Experimento de la rernolacha de azúcar, nitrógeno × abono vegetal.

Fuente de variación	gl	SC	CM	F Observado	F requerido 5%	F requerido 1%
Parcelas: N x parcelas de AV (subparcelas)	23	516.12				
Parcelas de nitrógeno: (parcelas principales)	5	274.92				
Bloques	2	7.87	3.935	1.56	19.00	99.00
Nitrógeno	1	262.02	262.020	104.18	18.51	98.49
Error (a)	2	5.03	2.515			
Abonos vegetales	3	215.26	71.753	118.99	3.49	5.95
N × AV	3	18.70	6.233	10.34		
Error (b)	12	7.24	0.603			

Sumas de Cuadrados y Cuadrados Medios

Bloques

$$SCB = \frac{\Sigma(T_b)^2}{n(AV)} - C = \frac{(161.9)^2 + \ldots + (172.2)^2}{2(4)} - FC = 7.87$$

Nótese que el denominador (8) es el número de elementos que integra cada uno de los términos del numerador.

$$CMB = SCB/gl(B) = 7.87/2 = 3.935$$

Nitrógeno

$$SCN = \frac{\Sigma(T_n)^2}{r(AV)} - C = \frac{(209.0)^2 + (288.3)^2}{3(4)} - FC = 262.02$$

$$CMN = SCN/gl(N) = 262.02/1 = 262.02$$

Parcelas de N (parcelas principales)

$$SC\ (\text{parcelas de N}) = \frac{\Sigma(T_{pm})^2}{(AV)} - C = \frac{(69.2)^2 + \ldots + (101.9)^2}{4} - FC = 274.92$$

Error (a)

$$SC(Ea) = SC\ (\text{parcelas de N}) - SCB - SCN = 274.92 - 7.87 - 262.02 = 5.03$$

$$CM(Ea) = SC(Ea)/gl(Ea) = 5.03/2 = 2.515$$

Tratamientos de abono vegetal

$$SC(AV) = \frac{\Sigma(T_{av})^2}{rn} - C = \frac{(98.3)^2 + \ldots + (137.0)^2}{3(2)} - FC = 215.26$$

$$CM(AV) = SC(AV)/gl(AV) = 215.26/3 = 71.753$$

N × AV

$$SC(N \times AV) = \frac{\Sigma(T_t)^2}{r} - C\text{-}SCN\text{-}SC(AV) = \frac{(40.5)^2 + \ldots + (80.2)^2}{3} - FC - 262.02 - 215.26 = 18.70$$

$$CM(N \times AV) = SC(N \times AV)/gl\,(N \times AV) = 18.70/3 = 6.233$$

Parcelas de (N × AV) (subparcelas)

$$SC\ \text{parcelas de}\ (N \times AV) = \Sigma(X)^2 - C = (13.8)^2 + (15.5)^2 + \ldots + (27.6)^2 - C = 516.12$$

Error (b)

$$SC(Eb) = SC\ \text{parcelas de}\ (N \times AV) - SC\ (\text{parcelas de N}) - SC(AV) - SC(N \times AV)$$
$$= 516.12 - 274.92 - 215.26 - 18.70 = 7.24$$

$$CM(Eb) = SC(Eb)/gl(Eb) = 7.24/12 = 0.603$$

VALORES F

Los efectos del nitrógeno y de bloque se prueban utilizando el error (a); los del abono vegetal y de la interacción del nitrógeno y los abonos vegetales se prueban mediante el error (b).

F para N = 262.02/2.515 = 104.18 F para N × AV = 6.233/0.603 = 10.34. El valor F altamente significativo para N × AV indica una diferencia en la respuesta comparativa del cultivo de remolacha de azúcar a los abonos vegetales en los diferentes niveles de fertilidad.

SEPARACIÓN DE MEDIAS

Pruebas F pertinentes

Mediante la partición de la suma de cuadrados para la interacción de N x AV, obtendremos un conocimiento mayor acerca de la naturaleza de la interacción. Diversas son las formas en que esto puede realizarse, pero la partición para responder a las siguientes preguntas parece lógica: ¿respondió la remolacha de azúcar en forma diferente en los dos niveles de nitrógeno a la vicia versus no vicia?, ¿al barbecho versus la cebada? ¿a la vicia versus la cebada y la vicia? La tabla 8.3 contiene los totales del tratamiento y un conjunto de coeficientes ortogonales a ser utilizados en el cálculo de la interacción, así como en otras comparaciones con un solo grado de libertad. Los resultados de la partición se muestran en la tabla 8.4 y los cálculos se indican a continuación de dicha tabla.

Tabla 8.3. Coeficientes ortogonales para las comparaciones indicadas.

	Tratamiento y totales del tratamiento							
	N_0				N_{120}			
Comparación	B 40.5	C, 45.7	V 66.0	CV 56.8	B 57.8	C 71.8	V 78.5	CV 80.2
N_0 vs N_{120}	−1	−1	−1	−1	1	1	1	1
V y No V	−1	−1	1	1	−1	−1	1	1
B y C	−1	1	0	0	−1	1	0	0
V y CV	0	0	−1	1	0	0	−1	1
N × (V y No V)	1	1	−1	−1	−1	−1	1	1
N × (B y C)	1	−1	0	0	−1	1	0	0
N × (V y CV)	0	0	1	−1	0	0	−1	1

$$SC[N \times (V \text{ y NO } V)] = \frac{(40.5 + 45.7 - 66.0 - 56.8 - 57.8 - 71.8 + 78.5 + 80.2)^2}{3(8)} = \frac{(7.5)^2}{24} = 2.344$$

Nuevamente, nótese la aplicación de la fórmula para calcular una suma de cuadrados con un solo grado de libertad:

$SC = \frac{[\Sigma(c_i T_i)]^2}{r(\Sigma c_i^2)}$. En estos casos, todos los coeficientes (c_i's) son ± 1 y no resulta necesario escribirlos en los numeradores.

Tabla 8.4. Componentes de varianza para la interacción.

Fuente de variación	gl	SC	CM	F obser-vado	F requerido 5%	F requerido 1%
N × AV	3	18.70	6.233	10.34	3.49	5.95
N × (V y No V)	1	2.34	2.34	3.88	4.75	9.33
N × (B y C)	1	6.453	6.453	10.70		
N × (V y CV)	1	9.901	9.901	16.42		
Error (b)	12	7.24	0.603			

$$SC[N \times (B \text{ y } C)] = \frac{(40.5 - 45.7 - 57.8 + 71.8)^2}{3(4)} = 6.453$$

$$SC[N \times (V \text{ y } CV)] = \frac{(66.0 - 56.8 - 78.5 + 80.2)^2}{3(4)} = 9.901$$

Ahora podemos contestar las tres interrogantes planteadas anteriormente:

1) La diferencia en respuesta de la vicia en N_0 en comparación con N_{120} no es significativamente diferente, es decir, en referencia a las medias de la tabla 8.6, 22.0 + 18.9 — 13.5 — 15.2 = 12.2 no es significativamente mayor que 26.2 + 26.7 — 19.3 — 23.9 = 9.7.

2) La diferencia entre el barbecho y la cebada en N_0 es significativamente menor que en N_{120}, es decir, en la tabla 8.6, 15.2 — 13.5 = 1.7 es significativamente menor que 23.9 — 19.3 = 4.6.

3) La diferencia entre la vicia y la cebada-vicia es significativamente diferente en N_0 que en N_{120} es decir, en la tabla 8.6, 22.0 — 18.9 = 3.1 es significativamente diferente que 26.2 — 26.7 = —0.5.

Errores estándar

Tanto en la prueba de DSM como en la de rango múltiple de Duncan, los errores estándar se calculan en base a la variabilidad dentro de las unidades experimentales a las cuales los tratamientos se aplican. En el diseño de parcelas divididas, el cálculo de los errores estándar para ciertos tipos de comparaciones de tratamientos se vuelve más complicado, como se puede observar en la tabla 8.5, puesto que tenemos dos fuentes de error experimental... que incluyen parcelas principales y subparcelas.

Diferencias Significativas Mínimas

En la tabla 8.6 se muestran las DSM para todas las comparaciones posibles de medias en el experimento de abono vegetal de la remolacha de azúcar por fertilidad del nitrógeno. Los procedimientos para el cálculo se indican más abajo. En realidad, puesto que la interacción de N x AV fue bastante fuerte, hay poca justificación para presentar los efectos promedio del AV y del N, pero así lo haremos aquí para ilustrar el procedimiento.

Tabla 8.5. Errores estándar para un diseño de parcela dividida. [a]

Ea = error (a), Eb = error (b), a = número de tratamientos de parcela principal, b = número de tratamientos de subparcelas, r = número de réplicas

A = tratamientos aplicados a las parcelas principales B = tratamientos aplicados a las subparcelas.

Medias comparadas	Error estándar o una media $(s_{\bar{x}})$
Tratamientos de parcela principal: $A_1 - A_2$	$\sqrt{\frac{Ea}{rb}}$
Tratamientos de subparcelas: $B_1 - B_2$	$\sqrt{\frac{Eb}{ra}}$
Tratamientos de subparcelas para el mismo tratamiento de parcela principal: $B_1A_1 - B_2A_1$	$\sqrt{\frac{Eb}{r}}$
Tratamientos de subparcelas para diferentes tratamientos de parcela principal: $B_1A_1 - B_1A_2$ o $B_1A_1 - B_2A_2$	$\sqrt{\frac{(b-1)\,Eb + Ea}{rb}}$

[a] Nótese la aplicación de $s_{\bar{x}}$ en la determinación de la DSM o la SCD: $DSM = t\sqrt{2}\ s_{\bar{x}}$; SCD = R (DSM).

Tabla 8.6. Aplicación de las diversas DSM's para comparar medias del tratamiento en el experimento de la remolacha de azúcar, N x AV en el proyecto de parcela dividida.

Libras de nitrógeno/acre	Abonos vegetales				Medias del nitrógeno
	Barbecho	Cebada	Vicia	Cebada y Vicia	
			Raíces (tons/acre)		
0	13.5	15.2	22.0	18.9	17.4
120	19.3	23.9	26.2	26.7	24.0
Medias del abono vegetal	16.4	19.6	24.1	22.8	

DSM 5%: medias del N, 2.8; medias del abono vegetal, 1.0; entre abonos vegetales en el mismo nivel de N, 1.4; entre abonos vegetales en diferentes niveles de N, 2.9.

DSM Para diferencias entre tratamientos de parcelas principales (entre medias del nitrógeno).

$$DSM_{.05} = t_a \sqrt{\frac{2(Ea)}{rb}} \text{ donde } t_a \text{ es el valor tabular de t para los gl para el Ea.}$$

$$= 4.303 \sqrt{\frac{2(2.515)}{3(4)}} = 4.303\,(0.647) = 2.8 \text{ tons/acre.}$$

DSM para diferencias entre tratamientos de subparcelas (entre medias del abono vegetal).

$$DSM_{.05} = t_b \sqrt{\frac{2\,Eb}{ra}} \quad \text{donde } t_b = \text{valor tabular de t para los gl para Eb.}$$

$$= 2.179 \sqrt{\frac{2(0.603)}{3(2)}} = 2.179\,(0.448) = 1.0 \text{ tons/acre.}$$

DSM para diferencias entre tratamientos de subparcela para el mismo tratamiento de parcela principal (entre medias del abono vegetal para el mismo nivel de nitrógeno).

$$DSM_{.05} = t_b \sqrt{\frac{2\,Eb}{r}} = 2.179 \sqrt{\frac{2(0.603)}{3}} = 2.179\,(0.634) = 1.4 \text{ tons/acre.}$$

DSM para diferencias entre tratamientos de subparcelas en distintos tratamientos de parcela principal (para comparar diferentes medias del abono vegetal en distintos niveles de nitrógeno, o para comparar medias para el mismo tratamiento de abono vegetal en diferentes niveles de nitrógeno).

$$DSM_{.05} = t_{ab} \sqrt{\frac{2[(b-1)\,Eb + Ea]}{rb}} \quad \text{donde } t_{ab} \text{ es un valor ponderado de t en algún punto entre los}$$

valores tabulares para t_a y t_b que se calcula como sigue:

$$t_{ab} = \frac{(b-1)\,(Eb)\,(t_b) + Ea\,(t_a)}{(b-1)\,Eb + Ea} = \frac{(4-1)\,(0.603)\,(2.179) + 2.515\,(4.303)}{(4-1)\,(0.603) + 2.515}$$

$$= \frac{14.764}{4.324} = 3.414$$

$$DSM_{.05} = 3.414 \sqrt{\frac{2[(4-1)\,(0.603) + 2.515]}{3(4)}} = 3.414\,(0.849) = 2.9 \text{ tons/acre}$$

RESUMEN

El diseño de parcelas divididas algunas veces resulta útil para un conjunto de tratamiento factorial de dos factores. Respecto al diseño de bloques completos al azar, se pierde precisión al hacer comparaciones entre tratamientos de parcelas principales, pero frecuentemente ésta se mejora para comparaciones entre tratamientos de subparcela y para la interacción de tratamientos de subparcela dentro de tratamientos de parcela principal.

Capítulo 9

Diseño de parcelas subdivididas

La adición de un tercer factor mediante la división de las subparcelas de un diseño de parcelas divididas da lugar a un diseño de parcelas sub-divididas. Este diseño suele ser bastante útil en un experimento de tres factores, a fin de facilitar las operaciones de campo o cuando resulta deseable mantener agrupadas combinaciones de tratamientos; sin embargo, la restricción adicional a la distribución aleatoria hace necesario el cálculo de un tercer término de error, que se utiliza para probar los efectos principales del factor aplicado a la segunda división, así como todas las interacciones que incluyen dicho factor. El proyecto puede tener ciertas ventajas en cuanto a las operaciones físicas con las unidades experimentales, pero la necesidad de un tercer término de error puede hacer bastante complicada la separación de medias. Se deberá consultar a un biometrista antes de emplear este diseño.

El procedimiento para la distribución aleatoria es el mismo que para el diseño de parcelas divididas, estando las subparcelas divididas en sub-subparcelas, cuyo número coincide con los niveles de los tres factores, a los cuales el tercer factor es aleatoriamente asignado... con una nueva distribución aleatoria para cada conjunto de subparcelas. La figura 9.1 ilustra la distribución parcial de un diseño de parcelas subdivididas, para evaluar los efectos de las fechas de siembra, del control del pulgón y de la fecha de recolección, sobre el control del virus portador del pulgón en la remolacha de azúcar. El procedimiento para la manipulación gradual de los datos de tal experimento se ilustrará con el efecto que ejercen estos tratamientos en la producción de raíces de remolacha.

ORGANIZACIÓN DE DATOS

Los datos que se ven en la tabla 9.1 están organizados por tratamientos y por bloques. La tabla 9.2 está diseñada para suministrar los totales de las interacciones en dos sentidos y de los efectos principales.

ANÁLISIS DE VARIANZA

El análisis completo de varianza se muestra en la tabla 9.3. El procedimiento gradual para completar la tabla es el siguiente:

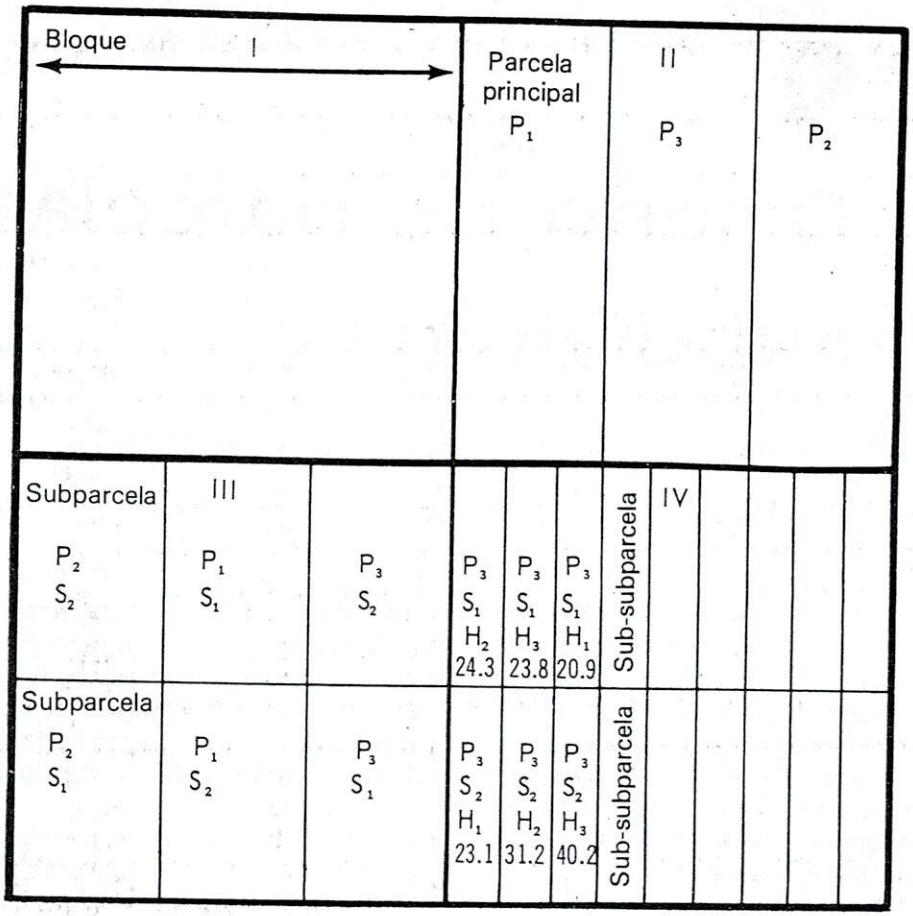


Figura 9.1. Fisonomía de un proyecto de parcela sub-subdividida, para un experimento de control viral de la remolacha de azúcar. Las parcelas principales son fechas de siembra (P_1, P_2, P_3) dispuestas en bloques aleatorios completos (I, II, III, IV). Las subparcelas son rociadas (S_1) y no rociadas (S_2) para el control del pulgón. Las sub-subparcelas con fechas de recolección a intervalos de cuatro semanas (H_1, H_2, H_3). Las producciones de raíz de remolacha de azúcar se muestran para las sub-subparcelas de la parcela principal P_3 del bloque IV. Los datos completos de este experimento aparecen en la tabla 9.1.

Fuentes de variación y grados de libertad

La variación entre las 72 unidades experimentales o sea las sub-subparcelas, se desglosa como se indica en la tabla 9.3. Los grados de libertad para los tres términos de error pueden encontrarse por sustracción, atendiendo a la distribución de las fuentes de variación. Error (a) = 11 – 3 – 2 = 6 error (b) = 23 – 11 – 1 – 2 = 9 y error (c) = 71 – 23 – 2 – 4 – 2 – 4 = 36.

Término de corrección

$$C = \frac{\Sigma X^2}{pshb},$$ donde p = número de fechas de siembra: 3; s = número de tratamientos de rociado: 2; h = número de fechas de cosecha: 3; y b = número de bloques: 4.

$$C = \frac{(2\,227.4)^2}{3\,(2)\,(3)\,(4)} = 68\,907.0939.$$

Tabla 9.1. Producciones de raíz de remolacha de azúcar (tons/acre), diseño de parcela subdividida, organizadas por tratamiento y por bloque

Tratamientos			Bloques					
P	S	H	I	II	III	IV	Totales	Medias
1	1	1	25.7	25.4	23.8	22.0	96.9 [bc]	24.2
		2	31.8	29.5	28.7	26.4	116.4	29.1
		3	34.6	37.2	29.1	23.7	124.6	31.2
		Tsp	92.1	92.1	81.6	72.1	337.9 [a]	28.2
	2	1	27.7	30.3	30.2	33.2	121.4 [b]	30.4
		2	38.0	40.6	34.6	31.0	144.2	36.0
		3	42.1	43.6	44.6	42.7	173.0	43.2
		Tsp	107.8	114.5	109.4	106.9	438.6	36.6
		Tmp	199.9	206.6	191.0	179.0	776.5	
2	1	1	28.9	24.7	27.8	23.4	104.8 [c]	26.2
		2	37.5	31.5	31.0	27.8	127.8	32.0
		3	38.4	32.5	31.2	29.8	131.9	33.0
		Tsp	104.8	88.7	90.0	81.0	364.5	30.4
	2	1	38.0	31.0	29.5	30.7	129.2	32.3
		2	36.9	31.9	31.5	35.9	136.2	34.0
		3	44.2	41.6	38.9	37.6	162.3	40.6
		Tsp	119.1	104.5	99.9	104.2	427.7	35.6
		Tmp	223.9	193.2	189.9	185.2	792.2	
3	1	1	23.4	24.2	21.2	20.9	89.7 [c]	22.4
		2	25.3	27.7	23.7	24.3	101.0	25.2
		3	29.8	29.9	24.3	23.8	107.8	27.0
		Tsp	78.5	81.8	69.2	69.0	298.5	24.9
	2	1	20.8	23.0	25.2	23.1	92.1	23.0
		2	29.0	32.0	26.5	31.2	118.7	29.7
		3	36.6	37.8	34.8	40.2	149.4	37.4
		Tsp	86.4	92.8	86.5	94.5	360.2	30.0
		Tmp	164.9	174.6	155.7	163.5	658.7	
		Tb	588.7	574.4	536.6	527.7	2 227.4 = ΣX	

$C = (2\,227.4)^2/72 = 68\,907.0939 \quad \Sigma(X)^2 = 71\,747.70$

[a, b, c] **Véanse las notas al pie de la tabla 9.2.**

Tabla 9.2.

Totales para las interacciones en dos sentidos									
P×S			P×H				S×H		
	S_1	S_2		H_1	H_2	H_3		S_1	S_2
P_1	337.9[a]	438.6	P_1	218.3[b]	260.6	297.6	H_1	291.4[c]	342.7
P_2	364.5	427.7	P_2	234.0	264.0	294.2	H_2	345.2	399.1
P_3	298.5	360.2	P_3	181.8	219.7	257.2	H_3	364.3	484.7

Totales para los efectos principales							
Fecha de plantación			Tratamiento de rociado		Fecha de recolección		
P_1	P_2	P_3	S_1	S_2	H_1	H_2	H_3
776.5	792.2	658.7	1 000.9	1 226.5	634.1	744.3	849.0

[a] Tomado de la tabla 9.1: Total para P_1S_1 incluyendo todas las cosechas y bloques.

[b] Total para P_1H_1 incluyendo todos los rociados y bloques = 96.9 + 121.4 = 218.3.

[c] Total para S_1H_1 incluyendo todas las fechas de plantación y bloques = 96.9 + 104.8 + 89.7 = 291.4.

Tabla 9.3. Análisis de varianza y diseño de parcela subdividida.

Fuente de variación	gl	SC	CM	F observado	F requerido 5%	F requerido 1%
P×S×H parcelas (sub-subparcelas)	71	2 840.6061				
P×S parcelas (subparcelas)	23	1 542.8128				
Parcelas de P (parcelas principales)	11	698.9028				
Bloques	3	143.4561	47.8187			
Fechas de siembra	2	443.6886	221.8443	11.91	5.14	10.92
Error (a)	6	111.7581	18.6264			
Tratamiento de rociado	1	706.8800	706.8800	81.21	5.12	10.56
P×S	2	40.6875	20.3438	2.34	4.26	8.02
Error (b)	9	78.3425	8.7047			
Fechas de recolección	2	962.3353	481.1676	102.80	3.26	5.25
P×H	4	13.1097	3.2774	0.70	2.63	3.89
S×H	2	127.8308	63.9154	13.66	3.26	5.25
P×S×H	4	44.0192	11.0048	2.35	2.63	3.89
Error (c)	36	168.4983	4.6805			

Sumas de Cuadrados

$$SCB = \frac{\Sigma T_b^2}{psh} - C = \frac{(588.7)^2 + \ldots + (527.7)^2}{3(2)(3)} - C = 143.4561$$

b = bloque o repetición

$$SCP = \frac{\Sigma T_p^2}{shb} - C = \frac{(776.5)^2 + \ldots + (658.7)^2}{2(3)(4)} - C = 443.6886$$

parcelas principales

$$SC(MP) = \frac{\Sigma T_{mp}^2}{sh} - C = \frac{(199.9)^2 + \ldots + (163.5)^2}{2(3)} - C = 698.9028$$

parcel. princ.

$$SC(Ea) = SC(MP) - SSB - SCP = 111.7581$$

$$SCS = \frac{\Sigma T_s^2}{phb} - C = \frac{(1000.9)^2 + (1226.5)^2}{3(3)(4)} - C = 706.8800$$

$$SC(P \times S) = \frac{\Sigma T_{p\times s}^2}{hb} - C - SCP - SCS = \frac{(337.9)^2 + \ldots + (360.2)^2}{3(4)} - C - SCP - SCS = 40.6875$$

$$SC\ (Subparcelas) = \frac{\Sigma T_{sp}^2}{h} - C = \frac{(92.1)^2 + \ldots + (94.5)^2}{3} - C = 1524.8128$$

$$SC(Eb) = SC\ (subparcelas)\ SC - (MP) - SCS - SC(P \times S) = 78.3425$$

$$SCH = \frac{\Sigma T_h^2}{psb} - C = \frac{(634.1)^2 + \ldots + (849.0)^2}{3(2)(4)} - C = 962.3353$$

$$SC(P \times H) = \frac{\Sigma T_{p\times h}^2}{sb} - C - SCP - SCH = \frac{(218.3)^2 + \ldots + (257.2)^2}{2(4)} - C - SCP - SCH = 13.1097$$

$$SC(S \times H) = \frac{\Sigma T_{s\times h}^2}{pb} - C - SCS - SCH = \frac{(291.4)^2 + \ldots + (484.7)^2}{3(4)} - C - SCS - SCH = 127.8308$$

$$SC(P \times S \times H) = \frac{\Sigma T_{p\times s\times h}^2}{b} - C - SCP - SCS - SCH - SC(P \times S) - SC(P \times H) - SC(S \times H)$$

$$= \frac{(96.9)^2 + \ldots + (149.4)^2}{4} - C - SCP - SCS - SCH - SC(P \times S) - SC(P \times H) - SC(S \times H)$$

$$= 44.0192$$

$$SC\ (sub\text{-}subparcelas) = \Sigma X^2 - C = (25.7)^2 + \ldots + (40.2)^2 - C = 2\,840.6061$$

$$SC\ (Ec) = SC(sub\text{-}\ subparcelas) - SC\ (subparcelas) - SCH - SC(P \times H) - SC(S \times H) - SC\ (P \times S \times H)$$
$$= 168.4983$$

Cuadrados Medios

Los cuadrados medios se obtienen, como es usual, mediante la división de los SC entre los grados de libertad apropiados, es decir, CM (Ec) = 168.4983/36 = 4.6805.

VALORES F

El error (a), o sea la interacción de B × P, se utiliza para probar los efectos de las fechas de siembra y de los bloques; el error (b), que corresponde a las interacciones combinadas de B × S más B × P × S, se emplea para probar los efectos del tratamiento de rociado y la interacción P × S; y el error (c), compuesto por las interacciones combinadas de B × H + B × P × H + B × S × H + B × P × S × H, se utiliza para probar las restantes fuentes de variación... aquellas asociadas con los tratamientos de sub-subparcelas.

SEPARACIÓN DE MEDIAS

El procedimiento empleado para la separación de medias dependerá de la naturaleza de los tratamientos, de las preguntas que el investigador se haya propuesto contestar y de los resultados del análisis inicial. Para nuestro ejemplo, el análisis revela que los efectos de los tratamientos de rociado y de las fechas de cosecha fueron los mismos para todas las fechas de siembra (no se registraron valores F significativos para P × S, P × H y P × S × H), pero que las plantas que fueron rociadas para el control del pulgón se condujeron en forma bastante diferente con respecto a la fecha de recolección que las plantas que no fueron rociadas (valores F altamente significativos para S × H).

Una práctica generalmente aconsejable en el resumen de los resultados de un experimento consiste en presentar las medias de la combinación de factores de mayor orden conjuntamente con las medias de la combinación de factores que parecen particularmente relevantes para las conclusiones que se van a obtener. En nuestro ejemplo, la tabla 9.4 contiene las medias de la fecha de siembra × el tratamiento de rociado × la fecha de recolección, las medias de la fecha de siembra y las medias para la integración de más alta significación, S × H.

Un gráfico de la respuesta de las plantas rociadas en contraposición a las plantas no rociadas a la fecha de recolección, muestra la naturaleza de la interacción (figura 9.2). Al parecer, la disminución de la infección viral a través del control del vector habilitaron a las plantas para crecer en una proporción más rápida durante el periodo de tres cosechas. Este efecto puede examinarse cuantitativamente, y la naturaleza de la función de respuesta de los tratamientos de rociado a la fecha de recolección se determina mediante la separación de dos grados de libertad para la misma, tanto para plantas rociadas (S_2) como para plantas no rociadas (S_1) en componentes que probarán hasta qué punto las respuestas observadas corresponden a una función rectilínea. Para separar los dos grados de libertad y las sumas de cuadrados de las fechas de recolección para cada tratamiento de rociado, en componentes de regresión lineal y residual (tabla 9.5), utilizaremos los métodos abreviados de análisis de regresión para observaciones igualmente espaciadas, que se estudiarán en el capítulo 15. Nótese que las fechas de recolección tienen lugar a intervalos de cuatro semanas (figura 9.2).

$$SC(H:S_1) = \frac{(291.4)^2 + (345.2)^2 + (364.3)^2}{3(4)} - \frac{(1\,000.9)^2}{3(12)} = 238.15$$

$$SC(H_L:S_1) = \frac{[-1(291.4) + 0(345.2) + 1(364.3)]^2}{12(2)} = 221.43$$

Tabla 9.4.

Fecha de siembra	Tratamiento de rociado	Fecha de recolección 8/27	9/24	10/22	Medias de la fecha de siembra [a]
		Raíces, tons/acre [b]			
3/2	No	24.2	29.1	31.2	32.3
	Sí	30.4	36.0	43.2	
4/2	No	26.2	32.0	33.0	33.0
	Sí	32.3	34.0	40.6	
5/2	No	22.4	25.2	27.0	27.4
	Sí	23.0	29.7	37.4	
		Tratamiento de rociado × medias de la fecha de recolección [c]			
Sin rociado		24.3	28.8	30.4	
Rociado		28.6	33.3	40.4	

a Significativo en el nivel del 1%: DSM, 5% — 3.0.
b DSM, 3% entre tratamientos de rociado para la misma planta y fecha de recolección — 3.7; entre fechas de siembra para el mismo tratamiento de rociado y fecha de recolección — 4.4. La interacción P × S × H no es significativa en el nivel del 5%.
c Significativo en el nivel del 0.1%: DSM, 5% entre fechas de H para el mismo tratamiento de rociado — 1.8; entre las fechas de H para diferentes tratamientos de rociado o entre tratamiento de rociado para la misma fecha de recolección — 2.1.

Tabla 9.5. Sumas de cuadrados de fechas de recolección para tratamientos de rociado, separados con el fin de determinar la naturaleza de la respuesta.

Fuente de variación	gl	SC	CM	F observado	F requerido 5%	F requerido 1%
$H:S_1$	2	238.15				
Regresion lineal	1	221.43	221.43	47.31	4.11	7.39
Residual	1	16.72	16.72	3.57		
$H:S_2$	2	852.01				
Regresión lineal	1	840.17	840.17	179.52		
Residual	1	11.84	11.84	2.53		
Error (c)	36		4.68			

Nótese que aquí se aplica la fórmula para calcular una suma de cuadrados de un solo grado de libertad:

$SC = \frac{[\Sigma(c_i T_i)]^2}{r(\Sigma c_i^2)}$. Las c_i son polinomios ortogonales de la tabla A.11, $n = 3$ (−1, 0, +1) y r es el número de elementos que integran cada total del numerador (3 fechas de siembra × 4 bloques = 12).

$$SC(H_R:S_1) = \frac{[1(291.4) - 2(345.2) + 1(364.3)]^2}{12(6)} = 16.72$$

. A modo de comprobación del cálculo, nótese que

$$SC(H_L:S_1) + SC(H_R:S_1) = SC(H:S_1)$$

Análogamente, las sumas de cuadrados de la fecha de recolección para S_2 se determinan y separan:

$$SC(H:S_2) = \frac{(342.7)^2 + (399.1)^2 + (484.7)^2}{3(4)} - \frac{(1226.5)^2}{3(12)} = 852.01$$

$$SC(H_L:S_2) = \frac{[-1(342.7) + 0(399.1) + 1(484.7)]^2}{12(2)} = 840.17$$

$$SC(H_R:S_2) = \frac{[1(342.7) - 2(399.1) + 1(484.7)]^2}{12(6)} = 11.84$$

Los valores F de la tabla 9.5 indican efectos altamente significativos de regresión lineal de la producción de raíz en fechas de recolección para ambos tratamientos de rociado. Los valores F no significativos del residuo indican que las desviaciones de las medias con respecto a una línea recta pueden ocurrir frecuentemente por azar.

Para hacer que estos datos coincidan con líneas rectas, utilizaremos la fórmula para observaciones igualmente espaciadas, presentada en el capítulo 15: $\hat{Y}_L = \overline{Y} + (K_2 P_1)X'$ donde: $\hat{Y}_L$ = producción de raíz estimada para la función lineal. Una línea recta que conecte a dos $\hat{Y}_L$ cualesquiera de la línea recta que deseamos calcular. $\overline{Y}$ = la media de las producciones para las cuales se ajusta la recta. K_2 se obtiene a partir de la tabla A.11. P_1 = la suma de cada media × un coeficiente de polinomio extraído de la tabla A.11. X' es el valor codificado de cada fecha de recolección en la escala de los polinomios de regresión lineal para $n = 3$, tabla A. 11.

Para hacer corresponder una línea recta que represente toneladas por acre en cada fecha de recolección para plantas que no fueron rociadas (S_1):

$$\overline{Y} = (24.3 + 28.8 + 30.4)/3 = 27.8 \text{ tons/acre}$$

$K_2 = ½$, véase la tabla A. 11 debajo de $n = 3$

$$P_1 = -1(24.3) + 0(28.8 + 1(30.4) = 6.1$$

Luego: $\hat{Y}_L = 27.8 + [½(6.1)]X' = 27.8 + 3.05\,X'$. Nótese que esta ecuación se da en términos de los valores codificados para las fechas de recolección, y que los valores de X' para fechas de recolección son: $H_1 = -1$, $H_2 = 0$ y $H_3 = +1$. El cálculo de un valor $\hat{Y}_L$ para dos valores de X' nos permite trazar la recta S_1 de la figura 9.2:

$$\hat{Y}_L = 27.8 + 3.05(-1) = 24.75 \quad \text{y} \quad \hat{Y}_L = 27.8 + 3.05(1) = 30.85$$

Tabla 9.6. Errores estándar para comparaciones que involucran al factor C, tratamientos aplicados a las sub-subparcelas. Los factores A y B son tratamientos aplicados a las parcelas principales y a las subparcelas respectivamente. Ea = error (a), Eb = error (b), Ec = error (c), donde a, b y c son respectivamente, los números de la parcela principal, de la subparcela y de los tratamientos de las subparcelas. Nótese que DSM = $t\sqrt{2}\ s_{\bar{x}}$ y SCD = R (DSM).

Comparación	Error estándar ($s_{\bar{x}}$)	t calculado [a]
$C_1 - C_2$	$\sqrt{\frac{Ec}{rab}}$	
$A_1C_1 - A_1C_2$	$\sqrt{\frac{Ec}{rb}}$	
$B_1C_1 - B_1C_2$	$\sqrt{\frac{Ec}{rc}}$	
$B_1C_1 - B_2C_1$ $B_1C_1 - B_2C_2$	$\sqrt{\frac{(c-1)\ Ec + Eb}{rac}}$	$t = \frac{(c-1)\ Ec(tc)^c + Eb(tb)^b}{(c-1)\ Ec + Eb}$
$A_1C_1 - A_2C_1$ $A_1C_1 - A_2C_2$	$\sqrt{\frac{(c-1)\ Ec + Ea}{rbc}}$	$t = \frac{(c-1)\ Ec(tc) + Ea(ta)^b}{(c-1)\ Ec + Ea}$
$A_1B_1C_1 - A_1B_1C_2$	$\sqrt{\frac{Ec}{r}}$	
$A_1B_1C_1 - A_1B_2C_1$	$\sqrt{\frac{(c-1)\ Ec + Eb}{rc}}$	$t = \frac{(c-1)\ Ec(tc) + Eb(tb)}{(c-1)\ Ec + Eb}$
$A_1B_1C_1 - A_2B_1C_1$	$\sqrt{\frac{b(c-1)\ Ec + (b-1)\ Eb + Ea}{rbc}}$	$t = \frac{b(c-1)\ Ec(tc) + (b-1)\ Eb(tb) + Ea(ta)}{b(c-1)\ Ec + (b-1)\ Eb + Ea}$

a En el cálculo de la DSM, t = valor tabular de los gl para el término de error apropiado. Cuando una comparación incluye dos o más terminos de error debe calcularse un valor t.

b Los símbolos ta, tb y tc son valores t tabulares, provenientes de la tabla A. 2 de los grados de libertad para Ea, Eb y Ec, respectivamente.

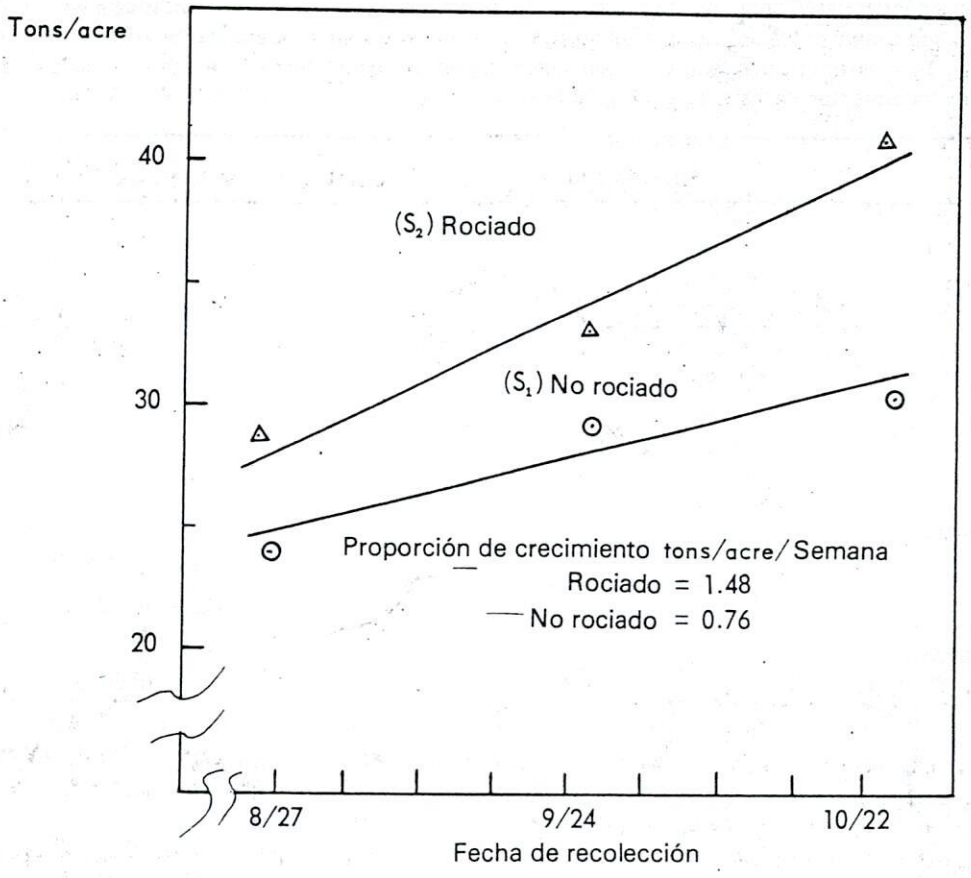


Figura 9.2. Efecto que ejerce el control del portador sobre la disminución de la proporción de crecimiento de la remolacha de azúcar. La diferencia entre las pendientes de las dos rectas (proporción de crecimiento) muestra la naturaleza de la interacción S × H.

Análogamente, se hace corresponder una recta con S_2 como sigue:

$$\bar{Y} = (28.6 + 33.3 + 40.4)/3 = 34.1;\ K_2 = \tfrac{1}{2};\ P_1 = -1(28.6) + 0(33.3) + 1(40.4) = 11.8$$

Por tanto: $\hat{Y}_L = 34.1 + [\tfrac{1}{2}(11.8)]X'$, $\hat{Y}_L = 34.1 + 5.9X'$

$\hat{Y}_L$ para $X' = -1$ es 28.2 y $\hat{Y}_L$ para $X' = 1$ es 40.0.

Para algunos experimentos que incluyen el diseño de parcelas subdivididas puede resultar deseable separar ciertas medias a través de la DSM o la prueba de rango múltiple de Duncan. Los cálculos de errores estándar y de ciertos valores t se indican en la tabla 9.6.

RESUMEN

El diseño de parcelas sub-subdivididas es una extensión del principio de parcelas divididas con subparcelas, que a su vez dividen en sub-subparcelas, al cual se le asigna un tercer factor de tratamiento. El análisis de varianza es más complicado que el de parcelas divididas, puesto que existen tres términos de error para probar efectos de tratamientos. Usualmente, el factor asignado a las sub-subparcelas y las interacciones que incluyen dicho factor, se evalúan con mayor precisión que los demás tratamientos.

Capítulo **10**

Diseño de bloques divididos

DISEÑO DE BLOQUES DIVIDIDOS

En esta variante del diseño de parcelas divididas, los niveles del factor B se aplican en franjas a lo largo de un bloque completo de parcelas de factor A. Si las parcelas de A están dispuestas en cuadro latino, las parcelas B pueden estar en franjas a lo largo de una hilera o columna completa.

Este diseño facilita frecuentemente las operaciones físicas, aunque, comparado con el de subparcelas, sacrifica la precisión para medir los efectos principales del factor B. La comparación entre las medias de B para un nivel de A dado se determina con bastante precisión; cuando éste es el efecto principal que el experimentador desea determinar, el diseño resulta bastante útil. La figura 10.1 presenta la disposición de un

Columnas →	I					II					III					IV					T
Hileras	H_4	H_5	H_1	H_3	H_2	H_4	H_2	H_3	H_5	H_1	H_1	H_5	H_2	H_3	H_4	H_4	H_3	H_5	H_1	H_2	Hileras
			N80					N160					N-0					N320			
IV	26.4	29.3	10.1	23.1	18.2	34.2	18.5	22.4	30.3	10.8	8.4	29.2	15.6	20.7	24.8	30.2	24.0	30.8	10.4	22.4	
			(107.1)					(116.2)					(98.7)					(117.8)			439.8
			(N total parcela)																		
			N320					N-0					N80					N160			
III	31.2	34.2	10.3	25.9	19.2	21.3	12.5	16.7	19.1	5.2	10.8	31.0	16.9	21.2	26.0	29.2	24.3	35.2	11.2	20 9	
			(120.8)					(74.8)					(105.9)					(120.8)			422.3
			N160					N80					N320					N-0			
II	28.0	31.2	10.2	22.3	16.9	29.5	16.9	20.4	26.6	9.5	9.8	30.9	18.1	23.9	28.8	13.6	13.9	16.4	6.1	10.5	
			(109.6)					(102.9)					(111.5)					(60.5)			384.5

Figura 10.1

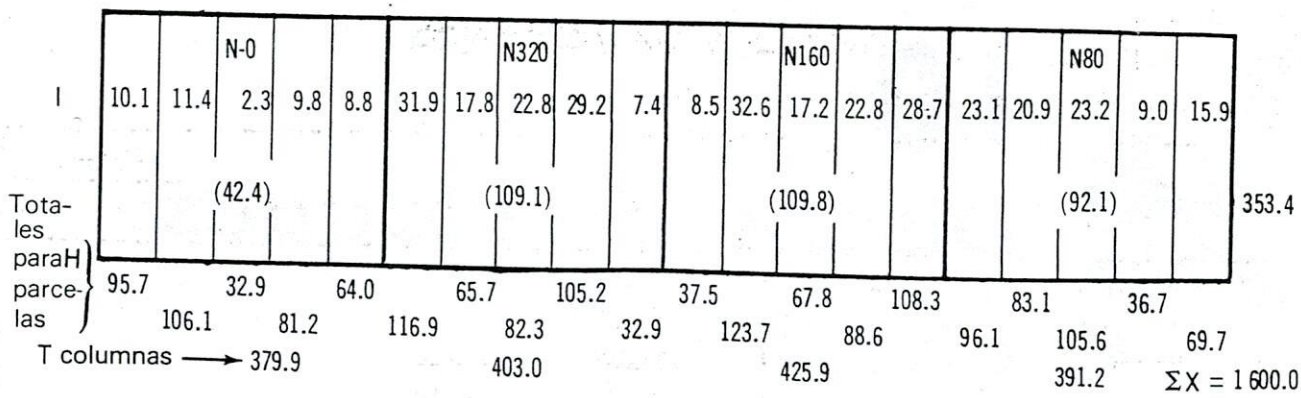


Figura 10.1. Disposición del experimento con la remolacha de azúcar, de las producciones de parcelas (toneladas de raíces/acre) y de totales. Proyecto: bloque dividido. Las parcelas A son proporciones de N en un cuadro latino de 4 x 4. Las parcelas B son cosechas a intervalos de tres semanas. Nótese que las fechas de recolección, $H_1 \ldots H_5$ son las mismas a través de cada columna de parcelas de N.

experimento de dosis de nitrógeno × fecha de recolección de remolacha azucarera y de producciones de raíces por subparcelas.

Las parcelas de A son las parcelas de dosis de nitrógeno dispuestas en un cuadro latino 4 x 4. Las parcelas de B son fechas de recolección dispuestas en línea a través de las columnas, pero distribuidas al azar en forma independiente para cada columna del cuadro latino. Las operaciones de recolección resultan fáciles de realizar cuando las parcelas que van a ser cosechadas en una fecha dada forman una columna continua; sin embargo, este arreglo necesita el cálculo de un término separado de error, para probar el efecto principal de la fecha de recolección, y reduce los grados de libertad para probar la interacción N × H. Si el diseño con respecto a las parcelas principales de nitrógeno fuese de bloques completos al azar, las columnas recibirían el nombre de

Tabla 10.1. Totales del tratamiento y medias; experimento con la remolacha azucarera de la figura 10.1.

Dosis de N	Fecha de recolección 1	2	3	4	5	T_N
			Totales			
0	22.0	47.4	61.1	69.8	76.1	276.4
80	39.4	67.9	85.6	105.0	110.1	408.0
160	40.7	74.4	91.9	120.1	129.3	456.4
320	37.9	77.5	96.6	122.1	125.1	459.2
T_H	140.0	267.2	335.2	417.0	440.6	1600.0
			Medias			
0	5.5	11.8	15.3	17.4	19.0	
80	9.8	17.0	21.4	26.2	27.5	
160	10.2	18.6	23.0	30.5	32.3	
320	9.5	19.4	24.2	30.5	31.3	

Tabla 10.2. Análisis de varianza. Diseño de bloque dividido.

Fuente de variación	gl	SC	CM	F calculado	F requerido 5%	F requerido 1%
Parcelas de N x parcelas de H (subparcelas)	79	5 542.680				
Parcelas de N	15	1 503.720				
Hileras	3	224.657	74.886			
Columnas	3	58.063	19.354			
Niveles de N	3	1 101.328	367.109	18.41	4.76	9.78
Error(a)	6	119.672	19.945			
Fechas de (H)	4	3710.765	927.691	111.92	3.26	5.41
Error(b) (C × H)	12	99.467	8.289			
N × H	12	157.147	13.096	6.59	2.03	2.72
Error(c) (C × N × H)	36	71.581	1.988			

bloques y las hileras representarían efectos que no podrían eliminarse; no obstante, en otros aspectos el análisis estadístico sería similar al de la tabla 10.2.

Los totales necesarios para completar el análisis de varianza se muestran en la figura 10.1 y en la tabla 10.1.

Término de corrección

$C = \frac{(\Sigma X)^2}{rnh}$, donde r = número de repeticiones, n = número de niveles de N, y h = número de fechas de recolección.

$$C = \frac{(1600)^2}{4(4)(5)} = 32\,000.00$$

Sumas de cuadrados

$$SCR = \frac{\Sigma T^2_r}{nh} - C$$

$$= \frac{(439.8)^2 + \ldots + (353.4)^2}{4(5)} - C = 32\,224.657 - C = 224.657$$

$$SCC = \frac{\Sigma T^2_c}{nh} - C$$

$$= \frac{(379.9)^2 + \ldots + (391.2)^2}{4(5)} - C = 32058.063 - C = 58.063$$

$$SCN = \frac{\Sigma T^2_n}{rh} - C$$

$$= \frac{(276.4)^2 + \ldots + (459.2)^2}{4(5)} - C = 33101.328 - C = 1101.328$$

$$SC(\text{parcelas de N}) = \frac{\Sigma T^2_{np}}{h} - C, \text{ donde } T_{np} = \text{totales para parcelas de N}$$

$$= \frac{(107.1)^2 + \ldots + (92.1)^2}{5} - C = 33503.720 - C = 1503.720$$

$$SC(Ea) = SC(\text{parcelas de N}). - SCR - SCC - SCN$$

$$= 1503.720 - 224.657 - 58.063 - 1101.328 = 119.672$$

$$SCH = \frac{\Sigma T^2_h}{rn} - C$$

$$= \frac{(140.0)^2 + \ldots + (440.6)^2}{4(4)} - C = 35710.765 - C = 3710.765$$

$$SC(Eb) = \frac{\Sigma T^2_{hp}}{r} - C - SCC - SCH, \text{ donde } T_{hp} = \text{totales para las parcelas de fecha de recolección.}$$

$$= \frac{(95.7)^2 + \ldots + (69.7)^2}{4} - C - SCC - SCH$$

$$= 35868.295 - C - SCC - SCH = 99.467$$

$$SC(N \times H) = \frac{\Sigma T^2_{n \times h}}{r} - C - SCN - SCH, \text{ donde } T_{n \times h} = \text{totales para los tratamientos de N x H.}$$

$$= \frac{(22.0)^2 + \ldots + (125.1)^2}{4} - C - SCN - SCH$$

$$= 36969.240 - C - SCN - SCH = 157.147$$

$$SC(N \times H \text{ parcelas}) = \Sigma X^2 - C$$

$$= (26.4)^2 + (29.3)^2 + \ldots + (15.9)^2 - C = 37\,542.68 - C$$

$$= 55\,42.680$$

$$SC(Ec) = SC(\text{parcelas de N x parcelas de H}) - SC(\text{parcelas de N}) - SCH - SC(Eb) - SC(N \times H)$$

$$= 5542.680 - 1503.720 - 3710.765 - 99.467 - 157.147$$

$$= 71.581$$

Cuadrados medios

Los cuadrados medios se obtienen mediante la división de las sumas de cuadrados entre los grados de libertad asociados a cada uno de los mismos; por ejemplo:

$$CM(Ec) = \frac{SC(Ec)}{gl\ (Ec)} = \frac{71.581}{36} = 1.988$$

Valores de F y comparación de medias

Los valores F se determinan mediante la división de los cuadrados medios entre los términos de error apropiados. Ea se utiliza para evaluar la hilera, la columna y los efectos del nivel de N; Eb, para las fechas de H; y Ec, para la interacción de los niveles de N con las fechas de recolección.

El valor F altamente significativo para la interacción N × H, indica una respuesta diferente a la fecha de recolección, dependiendo del nivel de N. Las medias de N × H en la tabla 10.1 muestran diferencias muy pequeñas en la producción de raíces en cualquier fecha de recolección entre plantas fertilizadas con 160 y 320 libras de N por acre, pero revelan diferencias crecientes en la producción de raíz entre estos dos niveles y los demás niveles de N, a medida que la época de recolección avanzó. La presentación gráfica de las medias (figura 10.2) muestra la diferencia de las razones de crecimiento. Las curvas de crecimiento de la figura 10.2, polinomios de segundo grado, parecen ajustar adecuadamente los datos. Estos son los tipos de funciones de crecimiento que cabe esperar. Se ajusta una curva única a las medias de producción de raíces en los niveles N 160 y N 320.

La justificación para el ajuste de curvas cuadráticas de regresión se obtiene mediante la separación de las sumas de cuadrados para el efecto de la fecha de recolección para cada nivel de N, determinándose las varianzas para las pruebas F a fin de indicar qué tan bien se ajustan los datos mediante los polinomios de grados sucesivos.
Usando los métodos de regresión del capítulo 15, se ha efectuado, en la tabla 10.3, la partición de las sumas de cuadrados y se calcularon las ecuaciones de la figura 10.2. El procedimiento se muestra sólo para el caso de N = 0.

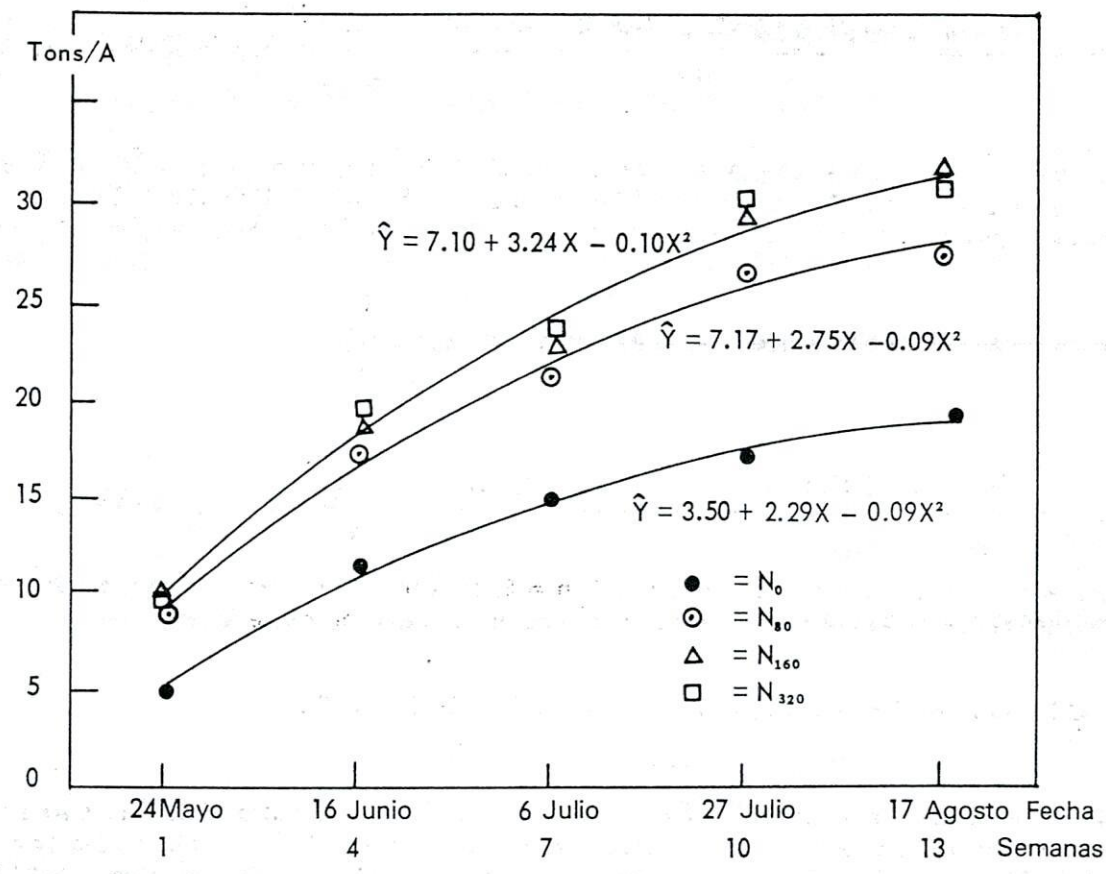


Figura 10.2. Producciones de raíces de remolacha de azúcar, tal como son influidas por la fertilización con nitrógeno y el tiempo de recolección. La X en las ecuaciones de regresión es la semana de recolección.

Tabla 10.3 Sumas de cuadrados de la fecha de recolección para cada nivel de N, separadas para determinar la función de respuesta apropiada.

Fuente de variación	gl	Nivel de N							
		0		80		160		320	
		SC	CM	SC	CM	SC	CM	SC	CM
Fecha de H	4	461.957		836.085		1 279.942		1 289.928	
Lineal	1	426.409	426.409	796.556	796.556	1 242.110	1 242.110	1 199.025	1 199.025
Cuadrática	1	33.326	33.326	35.045	35.045	26.194	26.194	79.683	79.683
Residual	2	2.222	1.111	4.631	2.316	11.638	5.819	11.220	5.610

Sumas de cuadrados. N = 0 (veánse los totales de la Tabla 10.1).

$$SCH = \frac{(22.0)^2 + \ldots + (76.1)^2}{4} - \frac{(276.4)^2}{4(5)} = 4\,281.805 - 3819.484 = 461.957$$

$$SCH_L = \frac{[-2(22.0) - 1(47.4) + 0(61.1) + 1(69.8) + 2(76.1)]^2}{10(4)} = \frac{(130.6)^2}{40} = 426.409,$$ donde los coeficientes del numerador (−2, −1, 0 + 1, + 2) provienen de la tabla A.11 para c_1 bajo n = 5; el 10 en el denominador es la suma de cuadrados de los coeficientes del numerador $[(-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2]$ y 4 es el número de observaciones que integran cada total del numerador. Nótese la aplicación de la ecuación para determinar una suma de cuadrados con un solo grado de libertad: $SC = \frac{[\Sigma(c_i T_i)]^2}{r(\Sigma c_i^2)}$

$$SCH_Q = \frac{[2(22.0) - 1(47.4) - 2(61.1) - 1(69.8) + 2(76.1)]^2}{14(4)} = \frac{(-43.2)^2}{56} = 33.326.$$ donde los coeficientes del numerador provienen de la tabla A.11 para c_2 bajo n = 5; el 14 en el denominador es la suma de cuadrados de dichos coeficientes; y 4 es el número de observaciones en cada total del numerador:

$$SCH_R = SCH - SCH_L - SCH_Q = 461.957 - 426.409 - 33.326 = 2.222$$

Los 16 grados de libertad (4 por nivel de N) y las sumas de cuadrados asociadas, desglosadas en la tabla 10.3, son aquellos para las fechas H, más la interacción de N × H. Nótese que la suma total de cuadrados de la tabla 10.3 (461.957 + ... + 1289.928 = 3867.912) es igual a la SC(fechas H) + SC(N × H) de la tabla 10.2. Por tanto, un cuadrado medio para las pruebas F implicará ambos errores (b) y (c). Se obtiene un cuadrado medio del error apropiado para las pruebas F, referentes a las fechas de recolección dentro de cada nivel de N, mediante la combinación de las sumas de cuadrados y los grados de libertad para el error (b) + error (c). Este es el mismo término de error que se obtendría si el error (b) estuviera en un diseño de parcela dividida.

Errores (b + c) = (99.467 + 71.581)/(12 + 36) = 3.564. En consecuencia, $F(H_Q : N_O) = 33.326/3.564 = 9.35$.

F tabulado está basado en 1 y 48 gl y es igual a 7.19 para el nivel de significación del 1%. Análogamente, se puede verificar que los componentes cuadráticos de las fechas de recolección para cada nivel de N son altamente significativos y que las sumas residuales de cuadrados son todas no significativas.

Aplicando el método abreviado contenido en el capítulo 15, la ecuación cuadrática de regresión para N = 0 de la figura 10.2, se calcula como sigue:

$\hat{Y}_Q = (\bar{Y} - K_1P_2) + (K_2P_1)X' + (K_4P_2)X'^2$ donde $\bar{Y} = (5.5 + 11.8 + 15.3 + 17.4 + 19.0)/5 = 13.8$; de la tabla A.11, $K_1 = 1/7$, $K_2 = 1/10$, y $K_4 = 1/14$; $P_1 = -2(5.5) - 1(11.8) + 0(15.3) + 1(17.4) + 2(19.0) = 32.6$; $P_2 = 2(5.5) - 1(11.8) - 2(15.3) - 1.(17.4) + 2(19.9) = -10.8$

Por tanto, $\hat{Y}_Q = [13.8 - (1/7)(-10.8)] + [(1/10)(32.6)]X' + [(1/14)(-10.8)]X'^2 = 15.34 + 3.26X' - 0.77X'^2$

Esta ecuación puede reescribirse en términos de X, que es la semana de recolección, haciendo $X' = (X - 7)/3$. (Véase el capítulo 15.)

$$\hat{Y}_Q = 15.34 + 3.26[(X - 7)/3] - 0.77[(X - 7)/3]^2 = 3.50 + 2.29X - 0.086X^2$$

Capítulo 11

Subparcelas como observaciones repetidas

El principio de parcela dividida puede aplicarse a ciertos experimentos donde se hacen observaciones sucesivas de la misma unidad experimental durante cierto periodo. Por ejemplo, con un cultivo de forraje perenne, los datos de producción suelen obtenerse varias veces al año por un periodo de dos o más años. Steel y Torrie (A.13) se refieren a tal experimento como **parcelas divididas en el tiempo**. El procedimiento escalonado para la manipulación de tales datos se ilustrará en el siguiente ejemplo. Los datos se refieren a producción de materia seca de forraje en un experimento con variedades de alfalfa, incluyendo cuatro variedades semilatentes,* dispuestas en un diseño de bloques completos al azar. Para simplificar la exposición, sólo consideraremos los datos obtenidos en dos años de experimentación.

ANÁLISIS PARA CADA CONJUNTO DE OBSERVACIONES

Se realiza un análisis de varianza para cada año por corte.

a) Se tabulan los datos por variedad, corte y bloques, como en la tabla 11.1. (Para ilustrar el procedimiento, mostramos únicamente los datos de corte para el primer año. Los datos para el segundo año se manejan en forma análoga.)

b) Se realiza un análisis de varianza para cada corte, de acuerdo con el diseño básico (tabla 11.2).

El procedimiento para calcular el término de corrección y las sumas de cuadrados se muestra más abajo, para el corte 1.

$$C_1 = (61.50)^2/20 = 189.1125$$

$$SC(\text{bloques}) = [(11.91)^2 + \ldots + (12.49)^2]/4 - C = 0.1651$$

$$SC(\text{elementos}) = [(14.46)^2 + \ldots + (15.68)^2]/5 - C = 0.4729$$

$$SC(\text{total}) = (2.69)^2 + \ldots + (3.05)^2 - C = 1.1801$$

$$SC(\text{error}) = SC(\text{total}) - SC(\text{bloques} - SC(\text{elementos}) = 0.5421$$

*En México se llaman así a aquellas con capacidad de rebrote. (N. del R. T.)

Tabla 11.1. Producción de material seco por parcela (tons/acre) en el primer año de un experimento con una variedad de alfalfa.

Variedad	Bloques I	II	III	IV	V	$T_{v \times c}$	$\bar{X}_{v \times c}$	Bloques I	II	III	IV	V	$T_{v \times c}$	$\bar{X}_{v \times c}$
	Corte 1							Corte 2						
1	2.69	2.40	3.23	2.87	3.27	14.46	2.89	2.74	1.91	3.47	2.87	3.43	14.42	2.88
2	2.87	3.05	3.09	2.90	2.98	14.89	2.98	2.50	2.90	3.23	2.98	3.05	14.66	2.93
3	3.12	3.27	3.41	3.48	3.19	16.47	3.29	2.92	2.63	3.67	2.90	3.25	15.37	3.07
4	3.23	3.23	3.16	3.01	3.05	15.68	3.14	3.50	2.89	3.39	2.90	3.16	15.84	3.17
$T_{b \times c}$	11.91	11.95	12.89	12.26	12.49	61.50		11.66	10.33	13.76	11.65	12.89	60.29	
$C_1 = (61.50)^2/20 = 189.1125$			$\Sigma(X)^2 = 190.2926$					$C_2 = (60.29)^2/20 = 181.7442$			$\Sigma(X)^2 = 184.8487$			
	Corte 3							Corte 4						
1	1.67	1.22	2.29	2.18	2.30	9.66	1.93	1.92	1.45	1.63	1.60	1.96	8.56	1.71
2	1.47	1.85	2.03	1.82	1.51	8.68	1.74	2.00	2.03	1.71	1.60	1.96	9.30	1.86
3	1.67	1.42	2.81	1.51	1.76	9.17	1.83	2.03	1.96	1.85	1.82	2.40	10.06	2.01
4	2.60	1.92	2.36	1.92	2.14	10.94	2.19	2.07	1.89	1.92	1.82	1.78	9.48	1.90
$T_{b \times c}$	7.41	6.41	9.49	7.43	7.71	38.45		8.02	7.33	7.11	6.84	8.10	37.40	
$C_3 = (38.45)^2/20 = 73.9201$			$\Sigma(X)^2 = 77.2217$					$C_4 = (37.40)^2/20 = 69.9380$			$\Sigma(X)^2 = 70.7756$			

Tabla 11.2. Análisis de varianza para la producción de cada corte en el primer año.

Fuente de variación	gl	Suma de cuadrados Corte: #1	#2	#3	#4	Cuadrados medios Corte: #1	#2	#3	#4
Total	19	1.1801	3.1045	3.3016	0.8376				
Bloques	4	0.1651	1.7249	1.2562	0.3112	0.0413	0.4312	0.3140	0.0778
Variedades	3	0.4729	0.2547	0.5660	0.2295	0.1576	0.0849	0.1887	0.0765
Error (B x V)	12	0.5421	1.1249	1.4794	0.2969	0.0452	0.0937	0.1233	0.0247

Los cuadrados medios se obtienen mediante la división de las SC entre los gl apropiados, es decir,

$$(CMV) = \frac{SCV}{gl(V)} = \frac{0.4729}{3} = 0.1576.$$

ANÁLISIS ANUAL

Construimos una tabla de producciones totales por parcela, por año (tabla 11.3). Para un año dado, las producciones totales son análogas a los totales de parcela principal del diseño de parcelas divididas.

Tabla 11.3. Producción total (tons/acre) por parcela por año.

	Bloques						
Variedad	I	II	III	IV	V	$T_{v\times y}$	$\bar{X}_{v\times y}$
				Año 1			
1	9.02[a]	6.98	10.62	9.52	10.96	47.10	9.42
2	8.84	9.83	10.06	9.30	9.50	47.53	9.51
3	9.74	9.28	11.74	9.71	10.60	51.07	10.21
4	11.40	9.93	10.83	9.65	10.13	51.94	10.39
$T_{b\times y}$	39.00	36.02	43.25	38.18	41.19	197.64	
$C_1 = (197.64)^2/20 = 1\,953.0785$				$\Sigma(X)^2 = 1\,973.3862$			
				Año 2			
1	11.88	11.33	11.81	12.22	10.65	57.89	11.58
2	12.15	10.98	12.20	11.30	12.54	59.15	11.83
3	12.92	11.95	12.05	11.88	13.19	61.99	12.40
4	11.74	11.62	11.54	12.00	11.74	58.64	11.73
$T_{b\times y}$	48.69	45.86	47.60	47.40	48.12	237.67	
$C_2 = (237.67)^2/20 = 2\,824.3514$				$\Sigma(X)^2 = 2\,831.1491$			

[a] Tomado de la tabla 11.1: 2.69 + 2.74 + 1.67 + 1.92 = 9.02

El análisis de varianza de parcelas divididas se calcula para cada año, como en la tabla 11.4.

Tabla 11.4. Análisis de varianza. Primer año, ensayo de variedades de alfalfa

Fuente de variación	gl	SC	CM	F calculado	F requerido 5%	F requerido 1%
Total (variedad × parcelas de corte)	79	34.8690				
Parcelas principales (parcelas de variedad)	19	5.0770				
Bloques	4	1.9453	0.6484			
Variedades	3	0.9014	0.3005	1.62	3.49	5.95
Error (a) (B × V)	12	2.2303	0.1859			
Cortes	3	26.4452	8.8150	69.96	3.49	5.95
Cortes × bloques	12	1.5122	0.1260	3.74	2.03	2.72
Cortes × variedades	9	0.6217	0.0690	2.05	2.15	2.94
Error (b) (B × V × C)	36	1.2129	0.0337			

El término de corrección y las SC para el primer año se obtienen como sigue (los valores utilizados provienen de las tablas 11.1 y 11.3):

$C = (197.64)^2/[4(20)] = 488.2696$. Nótese que el factor 4 en el denominador representa el número de cortes por observación

$$SCB = [(39.00)^2 + \ldots + (41.19)^2]/[4(4)] - C = 490.2092 - C = 1.9453$$

$$SCV = [(47.10)^2 + \ldots + (51.94)^2]/[4(5)] - C = 489.1710 - C = 0.9014$$

$$SC(MP) = [(9.02)^2 + \ldots + (10.13)^2]/4 - C = 493.3466 - C = 5.0770$$

$$SC(Ea) = SC(MP) - SCB - SCV = 2.2303$$

$$SC(cortes) = [(61.50)^2 + \ldots + (37.40)^2]/20 - C = 514.7148 - C = 26.4452$$

Nótese que $SCC = (C_1 + C_2 + C_3 + C_4) - C$, o sea, es la suma de los términos de corrección para cada análisis de fecha de corte, menos el término de corrección total:

$$SC(C \times B) = [(11.91)^2 + \ldots + (8.10)^2]/4 - C - SCC - SCB$$

$$= 518.1723 - C - SCC - SCB = 1.5122$$

La eliminación de SC(C x B) del error experimental es una variación del análisis de parcelas divididas usual, donde dicho término suele dejarse en el error (b). Esto se elimina para parcelas divididas en el tiempo, puesto que esa interacción es, en general, significativa. Se puede ver el valor de F altamente significativo para C × B en la tabla 11.4. Dado que es significativamente distinto del error (b), CM(C × B) es el término de error apropiado para probar el efecto principal de corte; la F calculada para cortes será = 8.8150/.1260 = 69.96.

$$SC(CXV) = [(14.46)^2 + \ldots + (9.48)^2]/5 - C - SCC - SCV$$

$$= 516.2379 - C - SCC - SCV = 0.6217$$

$$SC(Total) = (2.69)^2 + \ldots + (1.78)^2 - C = 523.1386 - C$$

$$= [\Sigma(X^2)\, c_1 + \ldots + \Sigma(X^2)\, c_4] - C$$

$$= [\,190.2926 + \ldots + 70.7756] - C = 34.8690$$

además, $SC(Total) = SCc_1 + \ldots + SCc_4 + SCC$

$$= 1.1801 + \ldots + 0.8376 + 26.4452 = 34.8690$$

$$SC(Eb) = SC(Total) - SC(MP) - SCC - SC(C \times B) - SC(C \times V)$$

$$= 34.8690 - 5.0770 - 26.4452 - 1.5122 - 0.6217 = 1.2129$$

Separación de medias

El error (a) es el CM apropiado para emplearlo en la comprobación de las diferencias entre medias de las variedades para un solo año durante todos los cortes. El valor de F para variedades no es significativo. Normalmente, el análisis de los efectos de las variedades anuales terminaría aquí; sin embargo, las variedades 1 y 2 están estrechamente relacionadas, dado que la 2 es una selección a partir de la 1, para una alta producción de semillas. Por tanto, los tres gl y las SCV pueden desglosarse como en la tabla 11.5.

Tabla 11.5. Partición de la suma de cuadrados para las variedades de la tabla 11.4.

Comparación	gl	CM	F observado	F requerido 5%	F requerido 1%
1 y 2 vs 3 y 4	1	0.8778	4.72	4.75	9.38
1 vs 2	1	0.0046			
3 vs 4	1	0.0189			
Error(a)	12	0.1859			

Cálculos de la suma de cuadrados y de los cuadrados medios de la tabla 11.5. Estos totales utilizados provienen de la tabla 11.3:

$$(1\ y\ \ 2\ vs\ 3\ \ y\ \ 4) = (47.10 + 47.53 - 51.07 - 51.94)^2/[4(20)] = 0.8778$$

$$SC(1\ vs\ 2) = (47.10 - 47.53)^2/[2(20)] = 0.0046$$

$$SC(3\ vs\ 4) = (51.07 - 51.94)^2/[2(20)] = 0.0189$$

El valor de F para comparar las dos variedades similares con las demás es casi significativo en el nivel de 5%; por tanto, resultaría inconveniente concluir, sin mayores pruebas, que no existen diferencias entre las variedades. Además, el valor de F para variedades × corte (tabla 11.4) es casi significativo en el nivel de 5%, y sugiere la consideración de posibles interacciones. La naturaleza de este tipo de experimento suele producir diversos CM del error que difieren considerablemente entre fechas de corte (tabla 11.2); por ende, el error apropiado para examinar medias de variedad para una fecha de corte dada es el CM del error de dicha fecha de corte. En la tabla 11.6, las SC de variedades para cada fecha de corte han sido separadas y sometidas a prueba de significación mediante la aplicación del término de error para fecha de corte. Esto conduce a suponer la existencia de una interacción variedad × corte, ya que las dos variedades estrechamente relacionadas (1 y 2), produjeron menos en la época temprana y en la tardía, aunque se condujeron en forma comparable a las otras dos variedades a mitad de estación. Aunque las cuatro variedades se clasifican como semilatentes, la 1 y la 2 parecen más latentes que la 3 y la 4.

El procedimiento para desglosar las SC para el primer corte se presenta a continuación (tabla 11.6). Los totales utilizados provienen de la tabla 11.1.

Tabla 11.6. Cuadrados medios por cortes primer año, ensayo de variedades de alfalfa.

		SC[a] y CM			
Fuente de variación	gl	#1	#2	#3	#4
Variedades	3	(0.4729)	(0.2547)	(0.5660)	(0.2295)
V_1 y V_2 vs V_3 y V_4	1	0.3920*	0.2268	0.1566	0.1411*
V_1 vs V_2	1	0.0185	0.0058	0.0960	0.0548
V_3 vs V_4	1	0.0624	0.0221	0.3133	0.0336
Error	12	0.0452	0.0937	0.1233	0.0247

[a] Sólo los valores entre paréntesis son SC

*Se excede el valor F requerido para la significación en el nivel del 5%. $F_{.05}$ requerido, gl $^1/_{12}$ = 4.75

Sumas de cuadrados para los cortes número 1:

$$SC(V_1 \text{ y } V_2 \text{ vs } V_3 \text{ y } V_4) = (14.46 + 14.89 - 16.47 - 15.68)^2/[4(5)]$$

$$SC(V_1 \text{ vs } V_2) = (14.46 - 14.89)^2/[2(5)]$$

$$SC(V_3 \text{ vs } V_4) = (16.47 - 15.68)^2/[2(5)]$$

Errores estándar. Los errores estándar para la separación de medias a través de la DSM o del método de la prueba de rango múltiple de Duncan para la comparación, son:

1. Dos medias de variedad, incluyendo todos los cortes.

   a) Para medias con base en un corte:

   $s_{\bar{x}} = \sqrt{\frac{Ea}{rc}}$ donde r = número de bloques, y c = número de cortes. En el cálculo de los valores de la DSM o de la SCD de Duncan para esta y las otras separaciones de medias más abajo, los valores tabulados de t, o factores studentizados, están basados en los gl para el CM del error correspondiente. Nótese que DSM= $t\sqrt{2}\, s_{\bar{x}}$ y SCD = R (DSM).

   b) Para medias con base en el total por estación, es decir, tons por acre por año: $s_{\bar{x}} = \sqrt{\frac{c\,Ea}{r}}$

2. Para dos medias de corte: $s_{\bar{x}} = \sqrt{\frac{MC(C \times B)}{rv}}$

3. Para dos medias de variedad en el mismo corte: $s_{\bar{x}} = \sqrt{\frac{E}{r}}$ donde E es el cuadrado medio del error para el corte particular (tabla 11.2).

4. Para dos medias de corte en la misma variedad o en otra diferente $s_{\bar{x}} = \sqrt{\frac{E_1 + E_2}{2r}}$ donde E_1 y E_2 son los CM del error para los dos cortes. Nótese que E_1 y E_2 se promedian en el cálculo de este error estándar de una media.

COMBINACIÓN DE DOS O MÁS AÑOS

Además de analizar el rendimiento de las variedades en cada año, el investigador suele estar interesado en el rendimiento de las variedades durante una serie de años, así como en la posible interacción de las variedades en el transcurso de éstos. Los resultados de diversos años, incluyendo varios cortes por año, pueden combinarse como un análisis de parcelas sub-subdivididas, las variedades como parcelas principales, los años como subparcelas, y los cortes como sub-subparcelas; pero, la interacción de variedades × años × corte no es, por regla general, de primera importancia. Los análisis anuales más un análisis individual de los totales de todas las parcelas durante una serie de años es lo que se requiere, generalmente, para tomar una decisión en cuanto a la variedad más apropiada.

Para ilustrar el procedimiento de combinación de los rendimientos anuales de cada variedad durante varios años, utilizaremos solamente los datos obtenidos en dos años.

a) Se calcula un análisis de varianza para **parcelas principales** de cada año, como en la tabla 11.7 (véase la tabla 11.3 para los totales anuales).

Tabla 11.7. Análisis de varianza de la producción total por parcela para cada año.

Fuente de variación	gl	SC		CM	
		Año 1	Año 2	Año 1	Año 2
Total	19	20.3077	6.7978		
Bloques	4	7.7544	1.1261	1.9386	0.2815
Variedades	3	3.6054	1.9254	1.2018	0.6418
Error (B × V)	12	8.9479	3.7462	0.7457	0.3120

Los cálculos para el término de corrección y las SC son idénticos a los efectuados a continuación de la tabla 11.4, excepto en cuanto a la omisión del factor "4", el número de cortes, del denominador. Esto mantiene a las SC sobre la base de producción total por parcela por año, en vez de hacerlo sobre la base de corte de la tabla 11.4. Los cálculos para el primer año se presentan a continuación (los totales utilizados provienen de la tabla 11.3):

$$C_1 = (197.64)^2/20 = 1\,953.0785$$

$$SCB = [(39.00)^2 + \ldots + (41.19)^2]/4 - C = 7.7544$$

$$SCV = [(47.10)^2 + \ldots + (51.94^2]/5 - C = 3.6054$$

$$SC(Total) = [(9.02)^2 + \ldots + (10.13)^2] - C = 20.3077$$

$$SC(Error) = SC(Total) - SCB - SCV = 8.9479$$

b. Se construye una tabla de producción total por variedad durante todos los años que se van a combinar (tabla 11.8):

Tabla 11.8. Producción total por parcela por variedad durante dos años.

Variedad	Bloques						Xv, tons/ acre/año
	I	II	III	IV	V	Tv	
1	20.90[a]	18.31	22.43	21.74	21.61	104.99	10.50
2	20.99	20.79	22.26	20.60	22.04	106.68	10.67
3	22.66	21.23	23.79	21.59	23.79	113.06	11.31
4	23.14	21.55	22.37	21.65	21.87	110.58	11.06
Tb	87.69	81.88	90.85	85.58	89.31	435.31	

[a] Tomada de la tabla 11.3: 9.02 + 11.88 = 20.90

c) Se calcula el análisis combinado de varianza de la tabla 11.9.

Tabla 11.9. Análisis de varianza de producciones anuales durante un periodo de dos años.

Fuente de variación	gl	SC	CM	F observado	F requerido 5%	F requerido 1%
Total(variedad × parcelas por año)	39	67.1654				
Parcelas principales (parcelas de variedad)	19	14.0138				
Bloques	4	6.1058	1.5264			
Variedades	3	4.0323	1.3441	4.16	3.49	5.95
B × V (Error (a))	12	3.8757	0.3230			
Años	1	40.0600	40.0600	55.29	4.49	8.53
V × Y	3	1.4985	0.4995	0.69	3.24	5.29
B × Y	4	2.7747	0.6937			
B × V × Y	12	8.8184	0.7349			
Error (b)	16	11.5931	0.7246			

Los totales utilizados en los cálculos se encuentran en las tablas 11.3 y 11.8. El término de corrección y las SC se calculan como sigue: $C = (435.31)^2/yrv$ donde y = número de años, r = número de bloques, y v = número de variedades. Nótese que el número de cortes por año no forma parte del denominador. Las observaciones se encuentran sobre la base de producción total por parcela por año.

$$C = (435.31)^2/[2(5)(4)] = 4737.3699$$

$$SCB = [(87.69)^2 + \ldots + (89.31)^2]/[4(2)] - C = 4743.4757 - C = 6.1058$$

$$SCV = [(104.99)^2 + \ldots + (110.58)^2]/[2(5)] - C = 4741.4022 - C = 4.0323$$

$$SC(MP) = [(20.90)^2 + \ldots + (21.87)^2]/2 - C = 4751.3837 - C = 14.0138$$

$$SC(Ea) = SS(MP) - SCB - SCV = 3.8757$$

$$SCY = [(197.64)^2 + (237.67)^2]/[4(5)] - C = 4777.4299 - C = 40.0600$$

$$SC(V \times Y) = [(47.10)^2 + \ldots + (58.64)^2]/5 - C - SCV - SCY$$
$$= 4782.9607 - C - SCV - SCY = 1.4985$$

$$SC(B \times Y) = [(39.00)^2 + \ldots + (48.12)^2]/4 - C - SCB - SCY$$
$$= 4786.3104 - C - SCB - SCV = 2.7747$$

$$SC(total) = [(9.02)^2 + \ldots + (11.74)^2] - C = 4804.5353 - C = 67.1654$$

$$SC(B \times V \times Y) = SC(total) - SC(MP) - SCy - SC(V \times Y) - SC(B \times Y) = 8.8184$$

La razón F, CM(B × Y)/CM(B × V × Y), se acerca a la unidad, indicando la ausencia de la interacción B × Y. Esto justifica la combinación de las SC y los gl para estas dos fuentes de variación, a fin de formar el error (b), que puede utilizarse para probar tanto la interacción V × Y como el efecto de los años.

$$F(V \times Y) = 0.4995/0.7246 = 0.69$$

$$F(Y) = 40.06/0.7246 = 55.29$$

No hay indicios de una interacción V × Y, aunque existe un efecto de año altamente significativo.

RESUMEN

(Parcelas divididas como observaciones repetidas)

El muestreo periódico de la producción de parcelas, como los cortes repetidos en parcelas de variedades perennes, la repetida recolección de frutos de los mismos árboles o el muestreo cíclico del contenido de nutrientes en el suelo de parcelas en cierto periodo se pueden analizar más apropiadamente en un proyecto de parcela dividida. Frecuentemente la interacción de bloque × tiempo de recolección (o muestreo) es significativa y se utiliza como un término de error para probar el efecto principal de la observación repetida.

Capítulo **12**

Transformaciones

(Que hacer cuando los datos
violan las reglas)

Un investigador que se conforma con aprender las "recetas" para llevar a cabo un análisis de varianza, sin buscar el dominio y la comprensión de los principios inherentes al mismo, puede encontrarse con serios problemas. Sea que los comprenda o no, el investigador hará ciertas suposiciones acerca de sus datos cuando realice un análisis de varianza. Si los datos no concuerdan con estas suposiciones, dicho análisis puede dar lugar a que el investigador llegue a conclusiones que no tienen justificación. Asimismo, el investigador puede descuidar conclusiones importantes que se alcanzarían si los datos fuesen analizados adecuadamente.

SUPUESTOS DEL ANÁLISIS DE VARIANZA

Los supuestos sobre los que se basa un análisis de varianza son, en resumen:

1. Los términos de error son aleatoria, independiente y normalmente distribuidos.
2. Las varianzas de las diferentes muestras son homogéneas.
3. Las varianzas y las medias de las distintas muestras no están correlacionadas.
4. Los efectos principales son aditivos.

A continuación, estudiaremos detalladamente estos supuestos.

Normalidad

Afortunadamente, las desviaciones del supuesto de normalidad no afectan muy seriamente la validez del análisis de varianza. Existen pruebas de normalidad, pero resulta bastante fútil aplicarlas, a menos que el número de muestras con las que estamos trabajando sea definitivamente grande. La independencia implica que no hay relación entre el tamaño de los términos de error y la agrupación experimental a la cual pertenecen. Puesto que las parcelas adyacentes de un campo tienden a estar más estrechamente relacionadas entre sí que las parcelas aleatoriamente distribuidas, resulta importante evitar que todas las parcelas que reciben el

mismo tratamiento ocupen posiciones adyacentes en el campo. Esta es una de las principales razones para insistir en no dividir en subparcelas una parcela que recibe cierto tratamiento y referirnos a las mismas como repeticiones. La mayor seguridad contra violaciones evidentes del primer supuesto del análisis de varianza, consiste en llevar a cabo la distribución aleatoria adecuada al diseño experimental en particular, que estamos utilizando.

Homogeneidad de varianzas

La primera referencia en este libro al análisis de varianza (capítulo 3) versó sobre un ejemplo simple con dos tratamientos, repetido cada uno cinco veces. Notará el lector que supusimos que las varianzas dentro de cada tratamiento estimaron una varianza común. Por tanto, nos sentimos justificados para utilizar el promedio de estas dos varianzas como la mejor estimación de σ^2 en vez de hacerlo con cualquiera de las otras dos por separado. Análogamente, en el capítulo 4 empleamos un "cuadrado medio de error combinado", o un promedio de cuatro varianzas, para obtener la mejor estimación de la varianza común.

Si las varianzas dentro de tratamientos diferentes fuesen de hecho distintas, no tendríamos justificación para combinarlas. Supongamos, por ejemplo, que las repeticiones de dos de los tratamientos fueron en realidad muestras de poblaciones con grandes varianzas, mientras que aquellas de los otros dos tratamientos se obtuvieron de poblaciones con varianzas mucho menores. Resulta obvio que la diferencia requerida para la significación sería mayor para los dos tratamientos con más alta variabilidad que para los dos de menor variabilidad. Promediar las varianzas mayores y menores podría arrojar resultados engañosos. La diferencia entre dos tratamientos con varianzas grandes puede ser considerada significativa, cuando en realidad ésta pudo haber ocurrido fácilmente, por casualidad. Por otro lado, la diferencia entre dos tratamientos con varianzas pequeñas puede ser declarada no significativa cuando, de hecho, lo era. Los siguientes datos de un experimento hipotético con cuatro tratamientos, cada uno repetido cinco veces, ilustrarán esta situación.

Tratamiento	Repetición					Total	Media	s^2
	1	2	3	4	5			
A	3	1	5	4	2	15	3	2.5
B	6	8	7	4	5	30	6	2.5
C	12	6	9	3	15	45	9	22.5
D	20	14	11	17	8	70	14	22.5

Realizando el análisis de varianza en la forma habitual, obtenemos:

Fuente de variación	gl	SC	CM	F
Tratamientos	3	330	110	8.8** [a]
Error	16	200	12.5	

[a] Un procedimiento común es usar * ** *** para denotar la significación estadística en el nivel de 5, 1, y 0.1%, respectivamente.

Nótese que el cuadrado medio del error es el promedio de las cuatro varianzas individuales dentro de los tratamientos. El valor F es altamente significativo. Calculemos ahora una DSM:

$$DSM_{.05} = t\sqrt{2ESM/r} = 2.12\sqrt{5} = 4.74$$

Puesto que la diferencia de medias entre los tratamientos A y B es de solamente 3, podríamos concluir que ésta no fue significativa. La diferencia de medias entre C y D es igual a 5; por tanto, ésta podría denominarse significativa en el nivel de 5%; sin embargo, notamos que las varianzas de C y D son nueve veces mayores que las de A y B. El supuesto de que las varianzas son homogéneas es

puesto en seria duda. Por tanto, resultaría más razonable analizar A y B por separado de C y D.

El análisis para A y B es:

Fuente de variación	gl	SC	CM	F
Tratamientos	1	22.5	22.5	9*
Error	8	20.0	2.5	

Para C y D:

Fuente de variación	gl	SC	CM	F
Tratamientos	1	62.5	62.5	2.78ns
Error	8	180	22.5	

Esto nos conduce precisamente a las conclusiones opuestas respecto de las diferencias entre A y B y entre C y D. Posteriormente veremos cómo probar los datos para la homogeneidad de varianzas. En cuanto a lo que podemos hacer al encontrar datos en los que las varianzas no son homogéneas, hay diversos cursos que podemos seguir. Primero, podemos separar los datos en grupos, de modo que las varianzas dentro de cada grupo sean homogéneas. Luego cada grupo puede analizarse por separado, como lo hicimos en el ejemplo anterior. Segundo, podemos utilizar un método descrito en textos más avanzados de estadística, el cual contempla un procedimiento bastante complicado para ponderar medias de acuerdo a sus varianzas. Tercero, podemos transformar los datos en forma tal que éstos sean homogéneos. Este método se analizará más ampliamente en el presente capítulo.

Independencia de medias y varianzas

En algunos datos existe una relación definida entre las medias de las muestras y sus varianzas. Este es un caso especial y la causa más común de heterogeneidad de varianza. Una correlación positiva entre medias y varianzas suele encontrarse cuando existe un amplio rango de medias de la muestra.

Supongamos, por ejemplo, que estuvimos probando los efectos de diversos insecticidas sobre el pulgón, y que estuvimos midiendo su eficacia mediante el conteo del número de pulgones por hoja, después de la aplicación. Si las medias de dos tratamientos bastante ineficaces fueran 305 y 315, naturalmente vacilaríamos en otorgar demasiada importancia a esta diferencia. Por otro lado, si las medias de otros dos tratamientos fuesen 5 y 15, tenderíamos a pensar que esta diferencia fue apreciable, impresionados por el hecho de que una de ellas fue tres veces mayor que la otra. Bajo el supuesto de que las varianzas son homogéneas y no relacionadas con las medias, tendríamos que otorgar tanta importancia a la diferencia entre 305 y 315 como a la diferencia entre 5 y 15, puesto que las diferencias reales son las mismas en ambos casos. Probablemente enfrentaríamos un sentimiento de intranquilidad de que algo estuvo mal. El examen de las diversas muestras revelaría casi seguramente que en general las muestras con medias grandes presentarían también grandes varianzas y que aquellas con medias pequeñas presentarían varianzas reducidas. Entonces el supuesto de que las medias y las varianzas no están correlacionadas resultaría falso, y un análisis de varianza ordinario de los datos en bruto no tendría validez.

Tomemos un ejemplo más extremo. Un investigador desea probar el efecto de una nueva vitamina sobre el peso de algunos animales. Es su deseo incluir un amplio rango de animales en sus pruebas, de modo que elige ratones, gallinas y ovejas. El sentido común diría que una diferencia de media libra en los pesos medios de dos lotes de ovejas se consideraría despreciable y se atribuiría fácilmente a la casualidad. Una diferencia de media libra en los pesos medios de dos lotes de gallinas se consideraría muy grande, pero no más allá del reino de las posibilidades. Una diferencia de media libra en los pesos medios de dos lotes de ratones, se consideraría algo absolutamente fantástico. Este es un ejemplo reconocidamente extremo y casi absurdo, pero sirve para enfatizar el punto de que el supuesto de independencia de varianzas y medias no debe aceptarse ciegamente. Debemos examinar los datos y, si es necesario, probar la validez del supuesto, antes de proceder con el análisis de varianza.

Otros tipos de datos, que frecuentemente muestran una relación entre varianzas y medias, son aquellos basados en conteos y los que consisten en proporciones o porcentajes. Ahora, supóngase que encontramos

que existe una relación entre varianzas y medias; ¿significa esto que estamos forzados a abandonar el análisis de varianza como método para analizar los datos? Afortunadamente, esto no sucede con frecuencia. A menudo podemos transformar los datos en forma tal que el supuesto de independencia entre varianzas y medias sea válido. Luego podemos proceder con el análisis de varianza de los datos trasformados.

Aditividad

Para cada diseño experimental existe un modelo matemático denominado **modelo lineal aditivo.** Para un diseño completamente aleatorio, este modelo es $X_i = \bar{X} + t_i + e_i$, que expresa que el valor de cualquier unidad experimental está compuesto por la media general más el efecto de tratamiento, más un término de error. El modelo correspondiente para un diseño de bloque completos al azar es: $X_{ij} = \bar{X} + t_i + b_j + e_{ij}$; que expresa que cualquier unidad experimental está compuesta por la media general más un efecto de tratamiento, más un efecto de bloque, más un término de error. El aspecto importante, que debe notarse en estos modelos, es que los términos se **suman;** de ahí el término aditividad.

El modelo para un diseño en bloques completos al azar, por ejemplo, implica que un efecto de tratamiento es el mismo para todos los bloques y que el efecto de bloque es el mismo para todos los tratamientos. En otras palabras, si se encuentra que un tratamiento incrementa la producción en cierta cantidad promedio por encima de la media general, suponemos que éste tiene el mismo efecto en los bloques de alta producción que en los bloques de baja producción.

Podemos concebir diversas situaciones en las que este supuesto **no sería** correcto; por ejemplo, en un experimento para probar el efecto de N sobre la producción, algunos bloques pueden producir menos que otros, debido a un bajo nivel natural de N en el suelo. Cabe esperar que las parcelas de dichos bloques se beneficien más con la adición de nitrógeno que las parcelas de bloques, en que la reserva natural de nitrógeno era ya adecuada. Por otro lado, supóngase que la baja producción se debió a una reserva inadecuada de humedad. Entonces, cabe esperar que el suministro de nitrógeno tenga resultados poco halagüeños en estos bloques de baja producción, pero que origine un notable incremento de la producción en los bloques donde hubo suficiente agua. Otra situación puede ser aquella en la cual el efecto de un tratamiento es incrementar la producción en cierto porcentaje o proporción. Esto recibe el nombre de **efecto multiplicativo de tratamiento.**

En cualquiera de los casos anteriores, el supuesto de aditividad sería incorrecto; este hecho debe reconocerse en el análisis de los datos. En el caso de efectos multiplicativos de tratamiento, se registran nuevamente transformaciones que cambiarán los datos para ajustarlos al modelo aditivo.

PRUEBAS PARA LAS VIOLACIONES DE LOS SUPUESTOS

Ahora estamos en condiciones de brindar algunos ejemplos específicos de datos que no cumplen uno o más de los supuestos del análisis de varianza. Mostraremos cómo probar dichos supuestos y las formas en que pueden transferirse los datos, de modo que resulten adecuados. A continuación presentamos algunos datos hipotéticos que pueden obtenerse en un experimento como el estudiado en párrafos anteriores, sobre los efectos de una nueva vitamina en ratones, gallinas y ovejas:

Tabla 12.1. Pesos en libras, de animales tratados con vitaminas y de control, en un experimento de bloque completo aleatorio.

Especies—tratamiento	Bloque I	II	III	IV	Total	Media
Ratones—control	0.18	0.30	0.28	0.44	1.2	0.3
Ratones—vitamina	0.32	0.40	0.42	0.46	1.6	0.4
Subtotales	0.50	0.70	0.70	0.90	2.8	0.35
Gallinas—control	2.0	3.0	1.8	2.8	9.6	2.40
Gallinas—vitamina	2.5	3.3	2.5	3.3	11.6	2.90
Subtotales	4.5	6.3	4.3	6.1	21.2	2.65
Ovejas—control	108.0	140.0	135.0	165.0	548.0	137.0
Ovejas—vitamina	127.0	153.0	148.0	176.0	604.0	151.0
Subtotales	235.0	293.0	283.0	341.0	1152.0	144.0
Totales principales	240.0	300.0	288.0	348.0	1176.0	49.0

Examinando los datos mediante los métodos utilizados en los capítulos 5 y 6, dan como resultado el siguiente análisis de varianza:

Fuente de variación	gl	SC	CM	F
Bloques	3	984.00	328.00	2.63
Especies	2	108 321.16	54 160.58	434.51**
Vitaminas	1	142.11	142.11	1.14
Especies x vitaminas	2	250.41	125.20	1.00
Error	15	1 869.72	124.65	

La diferencia altamente significativa entre especies no debe sorprendernos en lo más mínimo. Puede parecer muy extraño que no hayamos encontrado una diferencia significativa debida a las vitaminas, especialmente porque cada animal que recibió la vitamina mostró en cada repetición un mayor peso que el animal de control correspondiente. Parece también extraño que no encontremos evidencias de interacción entre los efectos de la vitamina y las especies, puesto que la respuesta aparente a las vitaminas es sumamente diferente en las distintas especies. Si aceptamos este análisis en su valor nominal, tendríamos que concluir que el experimento fue virtualmente un fracaso total. Al parecer, todo lo que aprendimos fue que los ratones, las gallinas y las ovejas difieren de peso. Incluso si separáramos aquí los efectos de especies en dos comparaciones, una comparando a las ovejas con las gallinas y los ratones, y la otra comparando a las gallinas con los ratones, encontraríamos que no podríamos siquiera mostrar una diferencia significativa entre gallinas y ratones.

Fijémonos en los datos con los supuestos del análisis de varianza en mente, y veamos qué puede hacerse si algunos de los mismos resultan falsos.

Primero, podemos fijarnos en los términos de error para comprobar si los mismos están aleatoria, independiente y normalmente distribuidos. Para hacerlo, removeremos la media general, los efectos de tratamiento y los efectos de bloque de cada celda de la tabla, como hicimos en el capítulo 5. Esto da la siguiente tabla de términos de error.

Tabla 12.2. Componentes del error en un experimento con vitaminas.

Especies – tratamiento	Bloque I	II	III	IV	Total
Ratones – control	8.88	−1.00	.98	−8.86	0
Ratones – vitamina	8.92	−1.00	1.02	−8.94	0
Gallinas – control	8.60	− .40	.40	−8.60	0
Gallinas – vitamina	8.60	− .60	.60	−8.60	0
Ovejas – control	−20.00	2.00	−1.00	19.00	0
Ovejas – vitamina	−15.00	1.00	−2.00	16.00	0
Totales	0	0	0	0	

Estos términos de error no parecen estar aleatoriamente distribuidos. En apariencia, éstos no son independientes, puesto que en cada bloque los términos de error para los dos miembros de cada especie están estrechamente relacionados. Por último, parece que su distribución se desviara considerablemente de la normal, puesto que existen dos clases de modelo, uno entre 8.5 y 9.0, y el otro entre −8.5 y −9.0. El primer supuesto del análisis de varianza no logró sostenerse muy bien bajo una inspección minuciosa.

A continuación, examinaremos el supuesto de homogeneidad de varianzas. Para hacerlo, necesitamos aprender una prueba conocida como **prueba de Bartlett para la homogeneidad de varianzas.**

Primero, necesitamos calcular la varianza entre las cuatro repeticiones de cada combinación de tratamiento. Para los controles del ratón, ésta será:

$$[.18^2 + .30^2 + .28^2 + .44^2 - (1.2^2\ 4)] \quad \text{número de repeticiones} - 1) = .0115.$$

Después de calcular cada una de dichas varianzas, los valores obtenidos se agrupan en una tabla como la que se presenta a continuación:

Tabla 12.3. Varianza y sus logaritmos para grupos en un experimento con vitaminas.

Tratamiento	gl	s_i^2	codificado s_i^2	logaritmos de s_i^2 codificado
Ratón – control	3	0.0115	11.5	1.06
Ratón – vitamina	3	0.0035	3.5	0.54
Gallinas – control	3	0.3467	346.7	2.54
Gallinas-vitamina	3	0.2133	213.3	2.33
Oveja-control	3	546.0	546 000.	5.74
Oveja-vitamina	3	425.3	425 300.	5.63
Totales	18		971 875.	17.84
Media			161 979.	
Logaritmo de la media			5.209	

El propósito de codificar las varianzas es evitar los logaritmos negativos. Podemos multiplicar las varianzas por una constante cualquiera, sin alterar la prueba. Resulta deseable hacer todos los valores codificados iguales o mayores que 1, de modo que realizamos nuestra codificación multiplicando cada s_i^2 por 1 000. Resulta más sencillo utilizar logaritmos comunes; dos dígitos en la mantisa suelen ser suficientes. La media de las varianzas codificadas se encuentra al dividir sus totales entre el número de muestras, y el log de esta media se incluye en la tabla. Ahora estamos listos para calcular lo que se denomina ji cuadrada no ajustada, a partir de la fórmula:

$$\chi^2 = 2.3026\ [(\Sigma\ \text{gl} \times \text{log de la media}) - (\text{gl por muestra} \times \Sigma\ \text{log s})]$$

$$= 2.3026\ [(18 \times 5.209) - (3 \times 17.84)]$$

$$= 92.66.$$

El factor 2.3026 es el factor para convertir logaritmos comunes a logaritmos naturales.

El valor de ji cuadrada que hemos calculado requiere un ajuste, y para hacerlo necesitamos:

$$C = 1 + \frac{1}{3(\text{número de muestras} - 1)} \left[\frac{\text{número de muestras}}{\text{gl por muestra}} - \frac{1}{\Sigma\ \text{gl}}\right]$$

$$= 1 + \frac{1}{3\ (6-1)} \left[\frac{6}{3} - \frac{1}{18}\right] = 1.13$$

entonces χ^2 ajustada $= \chi^2$ no ajustada $/C = 92.66/1.13 = 82.00$.

Consultamos ahora la tabla A.6 de ji cuadrada con 5 grados de libertad (uno menos que el número de muestras), y encontramos que 82 excede ampliamente el valor tabular en el nivel de significación de 1%. La evidencia de que las varianzas son heterogéneas resultan, por tanto, muy convincentes.

(Nota: hemos presentado aquí una forma simplificada de la prueba de Bartlett, basada en tamaños iguales de muestra, puesto que este es el caso más comúnmente encontrado. La prueba puede realizarse con muestras de tamaño desigual, pero los cálculos son más laboriosos. Pueden hallarse mayores detalles sobre esta prueba en los textos de Steel y Torrie o de Snedecor y Cochran, recomendados en el apéndice A.13.)

El próximo supuesto que examinaremos es el relativo a la independencia entre medias y varianzas. Una rápida ojeada a los datos es suficiente para convencernos de que dicho supuesto resulta ciertamente incorrecto, puesto que las medias más elevadas tienen varianzas muy grandes y las medias reducidas presentan varianzas muy pequeñas.

Una pregunta importante que debe contestarse con el fin de decidir la transformación que ha de utilizarse, es la de cuáles están más cercanamente proporcionales a las medias: si las varianzas o las desviaciones estándar. Hemos construido una tabla de proporciones:

Tabla 12.4. Proporciones de varianza y desviaciones estándar para medias en un experimento con vitaminas

Tratamiento	$\bar{X}$	s_i^2	s_i	$s^2_i/\bar{X}$	$s_i/\bar{X}$
R-C	0.3	0.01147	0.107	0.04	0.36
R-V	0.4	0.00347	0.059	0.01	0.15
G-C	2.4	0.3467	0.589	0.14	0.24
G-V	2.9	0.2133	0.462	0.07	0.16
O-C	137.0	546.0	23.367	3.98	0.17
O-V	151.0	425.3	20.624	2.82	0.14

Puede advertirse que la proporción de varianzas para las medias se incrementa marcadamente al hacerlo las medias, mientras que la proporción de desviaciones estándar permanece absolutamente constante. A propósito, si las varianzas y las medias no estuviesen relacionadas, cabría esperar que ambas razones disminuyeran cuando las medias aumentaran.

El supuesto que falta examinar es el de aditividad. Uno de los aspectos que notamos en los datos originales es que los efectos de bloque difieren ampliamente de especie a especie. Bajo el supuesto de aditividad, sustrajimos el efecto de bloque promedio de todas las parcelas, para calcular los términos de error. Esta fue la razón principal para que en el análisis de varianza se registrara un término de error grande poco usual.

La prueba formal para la aditividad recibe el nombre de **prueba de Tukey**. Esta puede llevarse a cabo para probar la no aditividad de dos factores principales cualesquiera. Esto se ilustrará mediante la prueba de los efectos principales de especies y vitaminas. Primero, agrupamos los totales de cada combinación especies-vitaminas en una tabla, y luego calculamos los principales efectos de especies y los efectos de vitamina en los márgenes.

Debemos tener en cuenta que cada casilla de la tabla es la suma de cuatro repeticiones; esto debe considerarse en el cálculo de las medias.

Tabla 12.5. Totales de tratamiento - combinaciones de especies.

Especies	Control	Vitamina	Total	$\bar{X}_{sp}$ = Total 8	$\bar{X}_{sp} - \bar{X} = Sp_i$
Ratones	1.2	1.6	2.8	.35	–48.65
Gallinas	9.6	11.6	21.2	2.65	–46.35
Ovejas	548.0	604.0	1 152.0	144.00	95.00
Total	558.8	617.2	1 176.0		0
$\bar{X}_v$ = Total/12	46.567	51.433		$\bar{X}$ = 49.00	
$\bar{X}_v - \bar{X} = V_j$	–2.433	2.433	0		

Si el cálculo se ha llevado a cabo correctamente, las sumas, tanto de los efectos de especies como de los efectos de vitamina deben ser iguales a cero. La media general se obtiene al dividir el total principal entre 24, el número total de parcelas en el experimento. Debemos calcular ahora $Q = \Sigma X_{ij}\ Sp_i\ V_j$, según la cual multiplicamos cada casilla de la tabla por los efectos de especies y de vitaminas correspondientes:

$$Q = [1.2 \times (-48.65) \times (-2.433)] + \ldots + [604 \times 95.00 \times 2.433] = 12672.4.$$

La suma de cuadrados para la no aditividad se encuentra entonces como sigue:

SC no aditividad = (Q^2 × total de unidades experimentales)/SCEp x SCV), donde SCEp es la suma de cuadrados para las especies, y SCV es la suma de cuadrados para las vitaminas en el análisis de varianza.

Aplicando esta ecuación obtenemos:

SC no aditividad = $[(12\,672.4)^2 \times 24] / (108\,321.16 \times 142.11) = 250.375$

Esta es una porción de la suma de cuadrados de C x V, de modo que puede probarse como sigue:

Fuente de variación	gl	SC	CM	F
C × V	2	250.41		
No aditividad	1	250.375	250.375	7153.6
residual	1	0.035	0.035	

El valor F incluso excede el valor F requerido de 4 052 en el nivel de 1% para 1 y 1 grado de libertad, de modo que existe considerable evidencia de que el supuesto de aditividad es incorrecto.

Tenemos ahora comprobados todos los supuestos del análisis de varianza y encontramos que nuestros datos no satisfacen a ninguno de los mismos. No debe extrañar que el análisis de varianza haya arrojado resultados desilusionantes.

Quizá la forma más sensible de analizar estos datos consista en manejar por separado cada especie. Los análisis son:

Especies	Fuentes de variación	gl	SC	CM	F
Ratones	Bloques	3	0.0400	0.0133	8.31
	Vitaminas	1	0.0200	0.0200	12.50*
	Error	3	0.0048	0.0016	
Gallinas	Bloques	3	1.64	0.547	41.00**
	Vitaminas	1	0.50	0.500	37.5**
	Error	3	0.04	0.013	
Ovejas	Bloques	3	2834.0	944.7	157.4**
	Vitaminas	1	392.0	392.0	66.3**
	Error	3	18.0	6.0	

Estos resultados son, ciertamente, mucho más satisfactorios que el análisis de varianza general original. Estos análisis son válidos, ya que dentro de alguna especie los datos se ajustan bastante bien a los supuestos básicos. El único defecto de dichos análisis es que revelan muy poco acerca de si las diferentes especies reaccionan análogamente a las vitaminas. Quizá este no sea un aspecto demasiado importante, y en la práctica el investigador estaría satisfecho, sin lugar a dudas, de detenerse en este punto; sin embargo, seguiremos el otro procedimiento de transformación de los datos, a fin de mostrar los resultados notables que pueden obtenerse.

TRANSFORMACIÓN LOGARÍTMICA

Debemos afrontar ahora el problema de cómo transformar los datos. Siempre que tengamos datos en los que las desviaciones estándar (no las varianzas) de las muestras sean aproximadamente proporcionales a las medias, la transformación más efectiva será la de tipo logarítmico. Otro criterio para la elección de esta transformación es la evidencia de efectos principales multiplicativos, en vez de aditivos. Ambos criterios se encuentran en los datos con los que estamos trabajando, de modo que intentaremos transformarlos en logaritmos y observar qué sucede.

Antes de empezar, hagamos algunas consideraciones acerca de la aplicación de esta transformación. Los datos con valores negativos no pueden transformarse en esta forma. Si existen ceros entre los datos, afrontaremos el problema de que el logaritmo de cero es menos infinito. Para evitar esta situación, se recomienda sumar 1 a cada dato antes de la transformación. Los datos que contienen un gran número de ceros probablemente se manejarían mejor mediante algún otro método. Se pueden utilizar logaritmos de cualquier base, pero los logaritmos comunes (de base 10) son generalmente los más sencillos. Antes de la transformación, es legítimo multiplicar todos los datos por una constante, puesto que la misma no ejerce ningún efecto sobre el análisis subsecuente. Es una buena idea que ninguno de los datos sea menor que 1, pues en esta forma se pueden evitar los logaritmos negativos.

En los datos con los que estamos trabajando no existen ceros, pero el menor valor es 0.18, de modo que multiplicaremos todos nuestros valores por 10, antes de obtener los logaritmos. Esto da la siguiente tabla de valores transformados:

Tabla 12.6. Datos del experimento con vitaminas, transformados a log 10X.

	Bloque					
Especies—tratamiento	I	II	III	IV	Total	Media
Ratones—control	0.26	0.48	0.45	0.64	1.83	0.4575
Ratones—vitamina	0.51	0.60	0.62	0.66	2.39	0.5975
Subtotales	0.77	1.08	1.07	1.30	4.22	0.5275
Gallinas—control	1.30	1.48	1.26	1.45	5.49	1.3725
Gallinas—vitamina	1.40	1.52	1.40	1.52	5.84	1.4600
Subtotales	2.70	3.00	2.66	2.97	11.33	1.41625
Ovejas—control	3.03	3.15	3.13	3.22	12.53	3.1325
Ovejas—vitamina	3.10	3.18	3.17	3.25	12.70	3.1750
Subtotales	6.13	6.33	6.30	6.47	25.23	3.15375
Totales	9.60	10.41	10.03	10.74	40.74	
Medias	1.60	1.735	0.672	0.790		1.69917

El análisis de varianza es como se presenta a continuación.

Fuente de variación	gl	SC	CM	F
Bloques	3	0.12075	0.04025	13.77**
Vitaminas	1	0.04860	0.04860	16.62**
Especies	2	28.54926	14.27463	4883.00**
E × V	2	0.009525	0.00476	1.63
Error	15	0.04385	0.00292	

Este es ciertamente un resultado más satisfactorio que el análisis de los datos originales hasta donde los resultados positivos están implicados. No hemos obtenido aún una interacción significativa entre especies y vitaminas, pero ahora estamos planteando la pregunta en forma diferente. Antes preguntábamos: "¿Varía de especie a especie la cantidad de cambio de peso debido a la adición de vitaminas?" Ahora preguntamos: "¿Varía de especie a especie la **proporción** o el **porcentaje** de cambio de peso debido a las vitaminas?"

¿Obtuvimos un resultado más evidente esta vez, simplemente porque estuvimos "disponiendo las cosas" hasta alcanzar un resultado deseado?, ¿o fue justificada la transformación que utilizamos y es válido el nuevo análisis? Para estar seguros, comprobaremos los supuestos del análisis de varianza con los nuevos datos.

Como antes, construiremos una tabla de términos de error, sustrayendo la media, los efectos de tratamiento y los efectos de bloque de cada casilla de la tabla.

Tabla 12.7 Componentes del error de los datos transformados.

	Bloque			
Especies – vitaminas	I	II	III	IV
R – C	– 0.10	– 0.01	0.2	0.9
R – V	0.01	– 0.03	0.5	– 0.0.3
G – C	0.03	0.07	– 0.8	– 0.01
G – V	0.04	0.02	– 0.03	– 0.03
O – C	0.00	– 0.02	– 0.02	0.00
O – V	0.02	– 0.03	0.02	– 0.02

Estos términos de error parecen estar más aleatoria y normalmente distribuidos que aquellos de los datos originales.

Para probar la homogeneidad de la varianza, nuevamente llevamos a cabo la prueba de Bartlett:

$\chi^2 = 2.3026 \quad [(18 \times .9614) - (3 \times 5.11)] = 4.548$

C = 1.13 como antes.

χ^2 ajustada = 4.548 1.13 = 4.03

la cual, de acuerdo con χ^2 en la tabla A.6, se excedería por casualidad en más de 50% de las oportunidades.

Tabla 12.8. Prueba de Bartlett aplicada a los datos transformados del experimento con vitaminas.

Tratamiento	Media	s_i^2	codificado s_i^2	Logaritmo de s_i^2 codificado
R – C	0.4575	0.0243	24.3	1.39
R – V	0.5975	0.0040	4.0	0.60
G – C	1.3725	0.0118	11.8	1.07
G – V	1.4600	0.0048	4.8	0.68
O – C	3.1325	0.0062	6.2	0.79
O – V	3.1750	0.0038	3.8	0.58
Totales			54.9	5.11
Media			9.15	
Log. de la media			0.9614	

Una ojeada a la próxima tabla revela que no existen indicaciones de ninguna relación entre las medias y las varianzas.

Realizando la prueba para la aditividad obtenemos, como antes, los siguientes resultados:

Fuente de variación	gl	SC	CM	F
S × T	2	0.009525		
sin aditividad	1	0.009035	0.009035	18.44
residual	1	0.000490	0.000490	

El valor F ni siquiera se aproxima al nivel de significación de 10% para 1 y 1 gl .C $F_{.10}$ requerido = 39.86).

Ahora nos sentimos confiados en que el nuevo análisis es válido, puesto que los datos transformados satisficieron **todos** los supuestos del análisis de varianza. Con los datos originales, ninguno de los supuestos fue verdadero.

TRANSFORMACIÓN DE LA RAÍZ CUADRADA

Siempre que estamos tratando con cómputos de acontecimientos poco comunes, los datos tienden a seguir una distribución especial, denominada **distribución de Poisson**. Entendemos por acontecimiento poco común aquel que tiene muy baja probabilidad de ocurrir en cualquier individuo; por ejemplo, supongamos que en una partida de semillas de lechuga, 0.1% de las semillas llevaban el virus de la enfermedad del mosaico. La probabilidad de que cualquier semilla en particular contenga el mosaico es entonces de sólo 1/1 000, de modo que en lo que se refiere a una sola semilla, éste es un acontecimiento poco común. Si tomamos 100 muestras de 1 000 semillas de dicho lote, obtendremos aproximadamente estos resultados:

37 muestras contendrán 0 semillas infectadas
37 muestras contendrán 1 semilla infectada
18 muestras contendrán 2 semillas infectadas
6 muestras contendrán 3 semillas infectadas
2 muestras contendrán 4 semillas infectadas

Resulta obvio que esto se parece muy poco a un distribución normal. Esta distribución de Poisson tiene características muy interesantes: la varianza es igual a la media. En la práctica, la varianza es generalmente algo mayor que la media, debido a otros factores, además de la variación de muestreo, que afectan la ocurrencia de los acontecimientos objeto de cómputo. En cualquier proporción, la varianza tiende a ser proporcional a la media.

Cuando analizamos datos de este tipo, estamos violando diversos supuestos hechos en un análisis de varianza. Los errores no están normalmente distribuidos y las varianzas están relacionadas con las medias (siendo, por tanto, homogéneas).

Otro ejemplo de datos de este tipo se encuentra en el conteo de insectos, como el realizado a partir de números estándar de barridas con una malla. Aquí resulta bastante difícil definir qué se entiende por observación individual. Podemos considerarla como un sitio en particular sobre el cual podría hallarse un insecto. Al barrer con una malla, estamos haciendo un muestreo de miles de sitios semejantes y encontrando solamente algunos insectos. Entonces, la probabilidad de hallar un insecto en un punto particular, aleatoriamente escogido en un instante dado es, en realidad, un acontecimiento poco común.

Los datos de este tipo pueden hacerse más normales y al mismo tiempo las varianzas pueden hacerse relativamente independientes de las medias a través de su transformación en raíces cuadradas. En realidad, es mejor utilizar $\sqrt{X + \frac{1}{2}}$, especialmente si existen conteos por debajo de 10.

Los datos presentados a continuación muestran el número de insectos lygus obtenidos en 50 barridas en cada

parcela de un experimento para probar 10 insecticidas y un tratamiento de control, repetido cuatro veces en un diseño de bloques completos al azar.

Tabla 12.9. Número de lygus por 50 barridas.

	Bloque						
tratamiento	I	II	III	IV	Total	Media	s_i^2
A	7	5	4	1	17	4.25	6.25
B	6	1	2	1	10	2.50	5.67
C	6	2	1	0	9	2.25	6.92
D	0	1	2	0	3	0.75	0.92
E	1	0	1	2	4	1.00	0.67
F	5	14	9	15	43	10.75	21.58
G	8	6	3	6	23	5.75	4.25
H	3	0	5	9	17	4.25	14.25
I	4	10	13	5	32	3.00	18.00
J	6	11	5	2	24	6.00	14.00
K	8	11	2	6	27	6.75	14.25

El análisis de varianza es:

Fuente de variación	gl	SC	CM	F
Bloques	3	12.25	4.08	0.40
Tratamientos	10	380.00	38.00	3.70**
Error	30	308.00	10.27	

Transformando los datos mediante la aplicación de $\sqrt{X + \frac{1}{2}}$ obtenemos:

Tabla 12.10. Datos transformados del lygus.

Tratamiento	Bloques				Total	Media	s_i^2
	I	II	III	IV			
A	2.74	2.35	2.12	1.22	8.43	2.11	0.41
B	2.55	1.22	1.58	1.22	6.57	1.65	0.39
C	2.55	1.58	1.22	0.71	6.06	1.52	0.60
D	0.71	1.22	1.58	0.71	4.22	1.06	0.18
E	1.22	0.71	1.22	1.58	4.73	1.18	0.13
F	2.35	3.81	3.08	3.94	13.18	3.29	0.54
G	2.92	2.55	1.87	2.55	9.89	2.45	0.19
H	1.87	0.71	2.35	3.08	8.01	2.00	0.99
I	2.12	3.24	3.67	2.35	11.38	2.84	0.53
J	2.55	3.39	2.35	1.58	9.87	2.47	0.55
K	2.92	3.39	1.58	2.55	10.44	2.61	0.59

y este análisis de varianza:

Fuente de variación	gl	SC	CM	F
Bloques	3	0.532	0.177	0.36
Tratamientos	10	19.993	1.999	4.04**
Error	30	14.841	/492	

Los dos análisis no son demasiado diferentes, puesto que ambos muestran un efecto de tratamiento altamente significativo. El valor F es aproximadamente 10% mayor después de la transformación. Se registrarán algunas diferencias importantes en la separación de medias:

Tabla 12.11. Prueba de rango múltiple de Duncan, aplicada a datos en bruto y transformados (nivel del 5%).

Separación de medias	Tratamientos y medias										
de:	D	E	C	B	A	H	G	J	K	I	F
	0.75	1.00	2.25	2.50	4.25	4.25	5.75	6.00	6.75	8.00	10.0
Datos en bruto											
Datos transformados											

Notaremos que en los datos transformados, G y D, G y E, J y D, fueron declarados significativamente diferentes, puesto que no se encontraban entre los datos en bruto. El efecto general de la transformación es incrementar la precisión con la cual podemos medir las diferencias entre medias pequeñas. Esto es altamente deseable en el trabajo de control de insectos, ya que por lo general no estamos tan interesados en las diferencias entre dos tratamientos relativamente ineficaces como en comparar tratamientos que permitan un buen control.

Una ojeada a las varianzas en las dos tablas mostrará que antes de la transformación existió una estrecha relación positiva entre medias y varianzas. El coeficiente de correlación lineal entre las mismas es de 0.89, significativo en el nivel de 0.1%. Después de la transformación, la correlación fue de solamente 0.37, ni siquiera significativa en el nivel de 10%. Por tanto, uno de los supuestos del análisis de varianza fue violado en los datos originales, lo cual se subsanó mediante la transformación.

En general, podemos decir que los datos que requieren la transformación de raíz cuadrada no violan los supuestos del análisis de varianza casi tan drásticamente como los datos que requieren una transformación logarítmica. Consecuentemente, los cambios en el análisis provocados por la transformación no son tan espectaculares.

TRANSFORMACIÓN ANGULAR O ARCOSENO

Otro tipo de datos que pueden requerir transformación es el basado en conteos expresados como porcentajes o proporciones de la muestra total. Por regla general, tales datos tienen una distribución binomial, en vez de una distribución normal. Una de las características de esta distribución es que las varianzas se hallan relacionadas con las medias, pero en forma bastante diferente a la de los tipos de datos que hemos estado considerando. Hasta el momento, los casos que hemos estudiado son aquellos en que medias grandes tienden a tener varianzas grandes, y viceversa. En los datos binomiales, las varianzas tienden a ser pequeñas en los dos extremos de los rangos de valores (cercanos a cero y a 100%), pero mayores en el medio (alrededor del 50%). En realidad, esta es una idea bastante natural, incluso para quienes no son matemáticos. Tendemos a otorgarle más importancia a una diferencia entre cero y 6%, o entre 94% y 100%, que a una diferencia entre 47% y 53%, aunque todas ellas sean de la misma magnitud.

La transformación apropiada para este tipo de datos recibe el nombre de angular o arcoseno. Esta se obtiene mediante la determinación del ángulo cuyo seno es la raíz cuadrada de la proporción (porcentaje/100). Expresada en notación matemática, ésta es arcoseno $\sqrt{X}$ o seno $^{-1}\sqrt{X}$. La tabla A.8 puede utilizarse para encontrar las transformaciones directamente de los porcentajes.

Los datos deben transformarse si el rango de porcentajes es mayor que 40. Por otro lado, esto apenas es necesario. Los datos siguientes corresponden a un experimento con diseño completamente aleatorio, con semillas de lechuga en el que se incluyen 24 tratamientos, cada uno de ellos repetido tres veces. Los tratamientos se encuentran dispuestos por el orden de la magnitud de sus medias.

Tabla 12.12. Número de semillas de lechuga que germinan, en muestras de 50.

Tratamiento	Repeticiones 1	2	3	Media	s_i^2	Log $(10 \times s_i^2)$
1	0	0	1	0.33	0.33	0.519
2	0	1	0	0.33	0.33	0.519
3	0	0	1	0.33	0.33	0.519
4	0	2	0	0.67	1.33	1.124
5	2	0	0	0.67	1.33	1.124
6	0	2	3	1.67	2.33	1.367
7	7	10	7	8.00	3.00	1.477
8	11	12	15	12.67	4.33	1.637
9	13	18	18	16.33	8.33	1.921
10	22	16	13	17.00	21.00	2.322
11	24	13	18	18.33	30.33	2.482
12	23	21	29	24.33	17.33	2.239
13	24	29	29	27.33	8.33	1.921
14	37	28	27	30.67	30.33	2.482
15	42	41	40	41.00	1.00	1.000
16	39	41	45	41.67	9.33	1.970
17	41	45	40	42.00	7.00	1.845
18	47	41	43	43.67	9.33	1.970
19	45	42	48	45.00	9.00	1.954
20	46	42	48	45.33	9.33	1.970
21	49	46	48	47.67	2.33	1.367
22	48	49	48	48.33	0.33	0.519
23	50	49	48	49.00	1.00	1.000
24	49	49	50	49.33	0.33	0.519
				Totales	178.00	35.767
				10 × media	74.167	
				Log (10 × media)	1.8702	

Nótese que hay una marcada tendencia, en las varianzas de los extremos, a ser más pequeñas que aquellas en la mitad de la distribución. Esto es clásico de los datos binomiales. Los logaritmos de las varianzas (codificados mediante la multiplicación por 10) se han incluido de manera que la prueba de Bartlett pueda llevarse a cabo.

χ^2 no ajustada = 2.3026 (log de la media $\times \Sigma$ gl − gl por muestra $\times \Sigma$ log codificado s_i^2)

$$= 2.3026\,(1\,8702 \times 48 - 2 \times 35.767)$$

$$= 41.99$$

$$C = 1 + \frac{1}{3\,(\text{muestras} - 1)} - \left(\frac{\text{número de tratamientos}}{\text{gl por tratamiento}} - \frac{1}{\Sigma\ \text{gl}}\right)$$

$$= 1 + \frac{1}{3 \times 23}\left(\frac{24}{2} - \frac{1}{48}\right) = 1.1736$$

χ^2 ajustada = $\chi^2 / C = 35.78$

Este sólo es significativo en el nivel de 5% (valor requerido 35.172), de modo que contamos con muy buena evidencia de que las varianzas no son homogéneas.

Analizando los datos en bruto, obtenemos los siguientes resultados:

Fuente de variación	gl	SC	CM	F
Tratamientos	23	25266.0	1098.52	148.12**
Error	48	356.0	7.42	

Transformando los datos, ejercicio que dejaremos al estudiante, aparentemente eliminamos toda relación entre las varianzas y las medias. Una prueba de Bartlett sobre estas cifras transformadas da un valor de ji cuadrada, ajustado igual a 9.00, el cual puede ser excedido por casualidad en más de un 99% de las veces.

Un análisis de varianza de los datos transformados no parece conducirnos a una conclusión distinta de la del análisis de los datos en bruto:

Fuente de variación	gl	SC	CM	F
Tratamientos	23	60 725.7	2 640.25	102.37**
Error	48	1 237.9	25.79	

La diferencia importante no se encuentra en el análisis total, sino en la separación de medias. Una prueba de rango múltiple de Duncan muestra que:

1. Cinco diferencias fueron declaradas significativas antes de la transformación, y no después: 7 − 8, 8 − 11, 10 − 12, 11 − 12 y 12 − 14.

2. Cinco diferencias fueron declaradas significativas después, y no antes, de la transformación: 18 − 22, 19 − 23, 19 − 24, 20 − 23 y 20 − 24.

¿Qué conjunto de conclusiones debe ser aceptado por nosotros? La respuesta es sencilla: debemos aceptar las conclusiones basadas en el análisis de mayor validez (en este caso, el análisis de los datos transformados).

Recuérdese que no transformamos los datos para obtener resultados que resulten más agradables, sino que los transformamos de modo que el análisis sea **válido** y las conclusiones **correctas**.

ESCALAS PRETRANSFORMADAS

Frecuentemente sucede que nos gustaría expresar los datos en porcentajes, pero encontramos este procedimiento muy difícil y engorroso para hacer mediciones precisas. Consideremos, por ejemplo, el problema de evaluar la cantidad de escara en tubérculos de papa. Una medida conveniente sería el porcentaje de área del tubérculo cubierta por la escara; no obstante, resulta muy difícil realizar esta medición con exactitud. Otro ejemplo sería el porcentaje de área de la hoja cubierta por lesiones provocadas por la enfermedad. Un último ejemplo sería el porcentaje de control de la maleza obtenido mediante la aplicación de diversos herbicidas. En todos estos casos podríamos, realizando un gran esfuerzo, medir dichos porcentajes con bastante precisión; pero el trabajo contemplado en esta tarea consumiría tal cantidad de tiempo que el número de parcelas se vería drásticamente limitado. Es una práctica común para realizar un mayor número de mediciones en un tiempo dado, hacer estimaciones visuales aproximativas de los porcentajes, en vez de efectuar mediciones precisas.

Por regla general, se construye una escala, como la escala de cero a 10 comúnmente utilizada en el trabajo de control de la maleza, donde el cero representa ausencia de control y el 10 indica un 100% de control. Si los escalones de dicha escala representan incrementos iguales de porcentajes, los datos deben ser transformados mediante la transformación angular, específicamente como debe ocurrir para medidas de porcentaje precisas.

¿Por qué no pretransformar nuestra escala? En otras palabras, podríamos escoger escalones de porcentajes tales que, al ser transformados por la transformación angular, resultaran en escalones igualmente incrementados que podrían reducirse a enteros.

La tabla siguiente presenta los porcentajes que producirán tales escalas.

Tabla 12.13. Escalas de clasificación pretransformadas. Escala de cero a:

Clasificación	4	5	6	8	10	18	20	24
0	0	0	0	0	0	0	0	0
1	15	10	7	4	2.5	0.75	0.7	0.5
2	50	35	25	15	10	3	2.5	2
3	85	65	50	30	21	6.7	5.5	4
4	100	90	75	50	35	12	10	7
5		100	93	70	50	18	15	10
6			100	85	65	25	20	15
7				96	79	33	27	20
8				100	90	42	35	25
9					97.5	50	42	31
10					100	58	50	37
11						67	58	43
12						75	65	50
13						82	73	57
14						88	80	63
15						93.3	85	69
16						97	90	75
17						99.25	94.5	80
18						100	97.5	85
19							99.3	90
20							100	93
21								96
22								98
23								99.5
24								100

Dichas escalas aprovechan el hecho de que generalmente es más fácil advertir pequeñas diferencias en las inmediaciones de cero y 100% que alrededor del 50%. En realidad, algunas escalas se han utilizado en el pasado, donde fueron deliberada o subconscientemente designadas para ajustarse a estos tipos de porcentajes. Con las papas se ha empleado una escala de 0 a 10, basada en patrones fotográficos que representan aproximadamente los porcentajes mostrados en la tabla de la página anterior. Con manzanas, se ha usado una clasificación de almidón que corresponde estrechamente a la escala de 0 a 8 mostrada en la tabla 12.13. En el trabajo con maleza, donde se utiliza una escala de 0 a 10, hay la tendencia a usar la clasificación 1, en vez de 10%, para un pequeño indicio de control, y la clasificación 9 para un control casi total.

Al analizar datos basados en tales clasificaciones de escala, éstos **no** deben ser transformados.

Debemos hacer hincapié en las clasificaciones de las parcelas de **control.** Hay una diferencia entre si éstas se incluyen en el experimento como un nivel cero de algún factor y se encuentran sujetas a la misma variación que la totalidad de los demás niveles de tratamiento, o si se incluyen como parcelas de referencia con las cuales comparar las otras parcelas. En el último caso, éstas suelen clasificarse como cero, y las demás parcelas de un bloque se comparan con las mismas. Siendo éste el caso, los datos de las parcelas de control **no** deben incluirse en un análisis de varianza. Las parcelas de control, con valores de cero arbitrariamente asignados, no tienen varianza. Por tanto, su varianza difiere de la de otros tratamientos, de modo que el supuesto de homogeneidad de la varianza es automáticamente violado.

RESUMEN

1. Los principales supuestos básicos para un análisis de varianza son: distribución aleatoria y normal de los términos de error, homogeneidad de las varianzas, independencia de varianzas y medias, y aditividad de los efectos principales.

2. Cuando estos supuestos son seriamente violados, el análisis de varianza no es válido.

3. Frecuentemente se pueden hacer transformaciones que corregirán las deficiencias de los datos, a fin de satisfacer los supuestos.

4. Cuando las **desviaciones** estándar están linealmente relacionadas con las medias, y los efectos principales parecen ser multiplicativos, una trasformación logarítmica usualmente corregirá ambas situaciones.

5. Los datos basados en conteos de acontecimientos raros, donde las **varianzas** están relacionadas con las medias, deben someterse a la transformación de raíz cuadrada.

6. Los datos basados en proporciones o porcentajes deben tener en cuenta la transformación angular o arcoseno.

7. Las escalas de clasificación pueden ser pretransformadas, basándolas en una escala variable de porcentajes.

Capítulo

13 Correlación lineal y regresión

EL CONCEPTO

Los términos **correlación** y **regresión** pueden parecer complicados, pero las ideas básicas implicadas en los mismos son tan sencillas que todo mundo las utiliza en sus conversaciones diarias. Consideremos, por ejemplo, las siguientes expresiones familiares:

"Cuanto mayor sea la altura, más fuerte será la caída."

"Cuanto más haya, mejor."

"Así como viene, se va."

"Cuanto mejor sea el día, mejor será la obra."

"Según se doble la rama, así estará inclinado el árbol."

Todos estos asertos contienen diversas ideas en común. Cada uno implica dos cantidades variables: la magnitud de una, dependiendo de la magnitud de otra. Los estadísticos se refieren a las mismas como a las **variables independiente** y **dependiente;** además, en estas oraciones en particular, hay la idea de que cuando una variable se incrementa, así lo hará la otra. En estadística esto recibe el nombre de **correlación directa** o **positiva.**

Tomemos otro grupo de frases:

"A mucha prisa, poca velocidad."

"Al buen entendedor, pocas palabras."

"Los buenos regalos vienen en estuches pequeños."

Tenemos aquí la misma idea general de dos variables, una dependiente de la otra, pero hay un leve giro en las relaciones entre las mismas. El **incremento** de una variable se acompaña por la disminución de la otra. Esto se denomina correlación **inversa** o **negativa.**

La idea de correlación no se limita solamente a estas sentencias. Recordemos los problemas que encontramos a cada minuto en el trabajo agrícola que tratan con relaciones entre dos variables. ¿Cuál es la cantidad de fertilizante aplicado, relacionada con la producción del cultivo?, ¿cuál es la relación entre la temperatura del horno y la suavidad de un asado?, ¿qué relación existe entre la cantidad de alimento consumido y el aumento

de peso del ganado?, ¿cuál es el precio de una mercancía afectada por la oferta?, ¿cuál es la dosificación de insecticida relacionada con el porcentaje de control o con la cantidad de residuo?, ¿cuál es la correlación entre el tamaño de la granja y la renta?, ¿está el realistamiento de miembros estrechamente relacionado con el realistamiento de líderes en los clubes 4-H?, ¿la cantidad de actividad de extensión agrícola se ve reflejada por la cantidad y calidad de la producción agrícola?

La lista de tales preguntas puede extenderse indefinidamente, pero debe resultar evidente, por el momento, que cualquiera de ellas contempla el tema de la correlación, llamémosle o no a ésta por su nombre.

Otro ejemplo de correlación que encontramos casi a diario es el gráfico común. Prácticamente, todo gráfico es, en esencia, una representación de la correlación entre dos variables. La escala sobre la línea horizontal, o abscisa, suele ser el rango de valores de la variable independiente. Los valores sobre la escala vertical, u ordenada, corresponden a la variable dependiente. La representación gráfica de los datos es frecuentemente un punto de partida muy útil para la realización de un análisis de correlación.

Ahora que hemos visto algunos ejemplos comunes de correlación, debemos ser capaces de formular una definición abstracta del término: la tendencia de dos variables a estar relacionadas en una forma definida. En realidad, la idea puede extenderse a más de dos variables, como en la ley de la oferta y la demanda, donde están contenidas tres variables: el precio, la oferta y la demanda. Para mantener el estudio tan simple como sea posible, nos limitaremos por el momento a la correlación entre dos variables.

Es costumbre considerar a una de las variables como dependiente de la otra. La elección sobre la variable que llamaremos dependiente o independiente, según el caso, suele ser obvia; por ejemplo, al estudiar la relación entre la producción y el fertilizante, sería lógico considerar a la primera como dependiente de lo segundo. Tratándose del precio y la oferta, generalmente pensamos en el precio como dependiente de la oferta. Por otro lado, existen situaciones en que la oferta es dependiente del precio. Frecuentemente existe un espacio de tiempo entre las mediciones de una variable y las mediciones correspondientes de la otra. En tales casos, la variable objeto de la primera medición recibe el nombre de independiente. Algunas veces resulta útil estudiar la correlación entre pares de mediciones sobre la misma variable; por ejemplo, un estudio de la correlación entre los precios de una mercancía en años sucesivos con los precios correspondientes en el año anterior, puede revelar una tendencia cíclica en el patrón de precios.

Existen situaciones en las que realmente no nos preocupa cuál de las variables se designe como dependiente. Simplemente podemos desear describir la distribución conjunta de dos variables, donde cada una de las mismas se encuentra normalmente distribuida. Dicha distribución recibe el nombre de **distribución normal bivariada**. Para describir esta distribución, necesitamos una estimación de ρ (rho), que no es más que uno de los parámetros de la población. El coeficiente de correlación r es la mejor estimación de ρ. Estudiar la correlación entre la longitud del antebrazo y su peso sería un ejemplo de situación en la que no importaría qué variable fue denominada dependiente.

CÓMO MEDIR LA CORRELACIÓN

Hasta el momento, nos hemos referido a la correlación como la idea general de dos variables relacionadas en alguna forma definida. No ha habido en la misma mucho de matemáticas o estadística. La simple observación de que dos variables parecen estar relacionadas no revela gran cosa. Necesitamos respuestas a dos importantes preguntas: ¿qué tan estrechamente relacionadas se encuentran las variables?; y ¿es real la relación, o podría haber ocurrido por un accidente debido a la casualidad? Para responder a la primera pregunta, necesitamos una medida definida de la estrechez de la relación entre dos variables. Esta medida recibe el nombre de **coeficiente de correlación**, representado por la letra r. Después de definir algunos otros términos, estaremos listos para mostrar cómo se calcula este valor y cómo se interpreta. Podemos obtener la respuesta a la segunda pregunta si consultamos las tablas de probabilidad apropiadas.

REGRESIÓN

El término **regresión** no se ha utilizado en este estudio desde la oración inicial. ¿Qué significa regresión? El diccionario no resulta de gran utilidad, pues el vocablo regresión es uno de aquellos desafortunados términos (como el término "error") que ha sufrido una evolución, de modo que su significado actual guarda poca semejanza con su significado original. En pocas palabras, regresión es la **cantidad de cambio** de una variable asociada a un **cambio único** de otra variable. Esta definición es susceptible de crítica en aquellas áreas en que resulta insuficientemente precisa o general desde el punto de vista matemático; sin embargo, para nuestros propósitos debe servir con el fin de puntualizar la principal distinción entre correlación y regresión. Nótese que la correlación se refiere al hecho de que dos variables se encuentran relacionadas y a la **estrechez** de dicha relación. La regresión, a su vez, se refiere a la **naturaleza** de la relación.

Volvamos a considerar algunos adagios familiares y veamos cómo el concepto de regresión aflora en nuestro pensamiento diario:

"Un centavo ahorrado es un centavo ganado."

"Más vale pájaro en mano que ciento volando."

"Un espacio de tiempo ahorra nueve."

"Un cuadro vale más que mil palabras."

Nótese que todos estos refranes implican la correlación de dos variables, pero van más allá y nos dicen en términos numéricos **cómo** están relacionadas ambas variables. Tomando estas oraciones literalmente, podemos construir una tabla:

Tabla 13.1. Asertos en términos matemáticos.

Variable independiente (X)	Variable dependiente (Y)	Ecuación de regresión	Coeficiente de regresión
Centavos ahorrados	Centavos ganados	$Y = X$	1
Pájaros en mano	Pájaro volando	$Y = 2X$	2
Un espacio de tiempo	Espacios ahorrados	$Y = 9X$	9
Cuadros	Palabras	$Y = 1000\ X$	1000

Hemos seguido la costumbre de denominar X a la variable independiente y Y a la variable dependiente. En los capítulos precedentes de este libro hemos considerado principalmente a una sola variable. sin hacer referencias a la correlación. A ésta la hemos denominado variable X. Luego, si estamos realizando un simple análisis de varianza con producciones de un grupo de parcelas, denominaremos a aquellas producciones de parcelas X; sin embargo, si deseamos estudiar la correlación entre producciones y cantidades de fertilizante aplicado, denominaremos Y a las producciones y X a las cantidades de fertilizante. No debemos permitir que este cambio aparentemente súbito de notación nos confunda. Este nos parecerá bastante natural después de que obtengamos alguna experiencia en el trabajo con correlación.

La tercera columna de la tabla lleva como título **ecuación de regresión.** Todas éstas corresponden a ecuaciones de rectas. La ecuación general de la recta es $Y = a + b\ X$. El símbolo *a* recibe el nombre de **intercepto,** puesto que cuando X es igual a cero, $Y = a$; de donde la recta corta al eje de las X en *a* unidades a partir del origen. Cuando *a* es igual a cero, la recta pasa a través del origen, pues cuando X es igual a cero, Y es también igual a cero. El símbolo *b* se denomina **pendiente,** puesto que determina la inclinación de la recta.

Resulta fácil apreciar que *b* es la cantidad de cambio de Y, asociada a un cambio unitario de X; y ésta es exactamente la forma en que hemos definido la regresión. Por tanto, resulta lógico denominar a *b* **coeficiente de regresión.**

CÓMO CALCULAR LA CORRELACIÓN LINEAL

Un ejemplo común de correlación es la relación entre la oferta y el precio. La tabla que presentaremos a continuación muestra las ofertas y precios del cerdo desde 1950 hasta 1959.

¿Existe una relación real entre oferta y precio durante este periodo? Una de las primeras observaciones que hacemos es que el precio más alto estuvo acompañado por las más baja producción, y viceversa. Esta es una alentadora evidencia de la correlación negativa que podemos esperar. A continuación, obtengamos una mejor idea acerca de los datos mediante el "trazado de un dibujo". Esto lo hacemos fácilmente situando puntos sobre un papel cuadriculado, dejando que la altura por encima del eje de las X represente el precio, y la distancia a la derecha del eje de las Y represente el número de cerdos en el año correspondiente (figura 13.1).

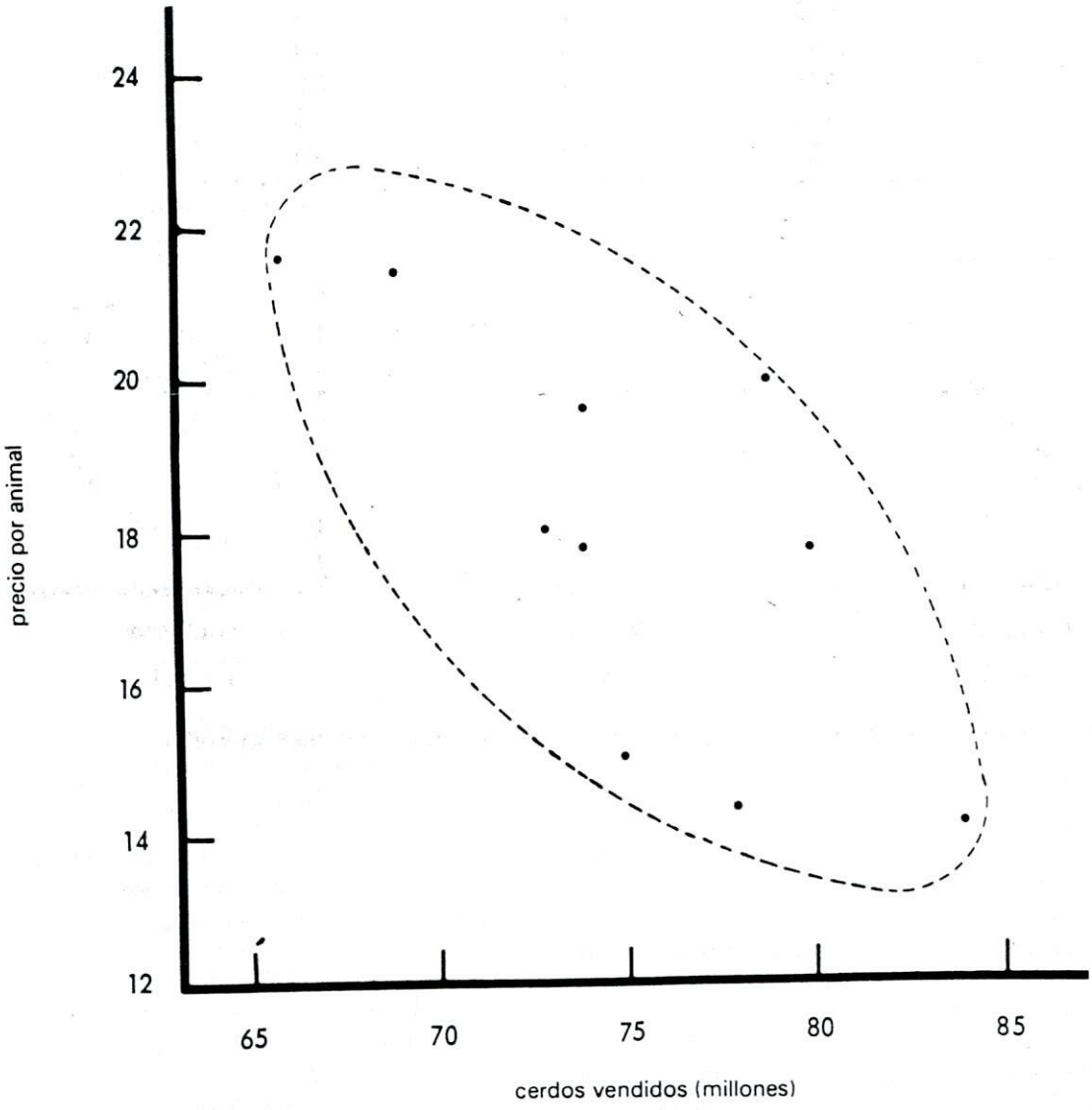


Figura 13.1. Diagrama de dispersión que muestra la relación entre el precio de los cerdos y el número de cerdos vendidos anualmente.

Un gráfico de este tipo recibe el nombre de **diagrama de dispersión**. Si tuvimos la idea de que la correlación entre la oferta y el precio fue muy estrecha, la dispersión bastante fortuita de estos puntos puede resultar desilusionante; sin embargo, parece existir una tendencia general entre los puntos a la izquierda, a ser mayores que aquellos a la derecha. Los puntos parecen caer dentro de una elipse bastante alargada (figura 13.1), típica de los diagramas que representan una correlación alta media. Otros tipos de diagramas de dispersión (figura 13.2) son guías para interpretar tales gráficos. La dirección del eje de la elipse en nuestro ejemplo indica una correlación negativa.

Tabla 13.2. Ofertas y precios del cerdo.

Año	Cerdos vendidos (millones) (X)	Precio por animal (dólares) (Y)
1950	73	18.0
1951	79	20.0
1952	80	17.8
1953	69	21.4
1954	66	21.6
1955	75	15.0
1956	78	14.4
1957	74	17.8
1958	74	19.6
1959	84	14.1

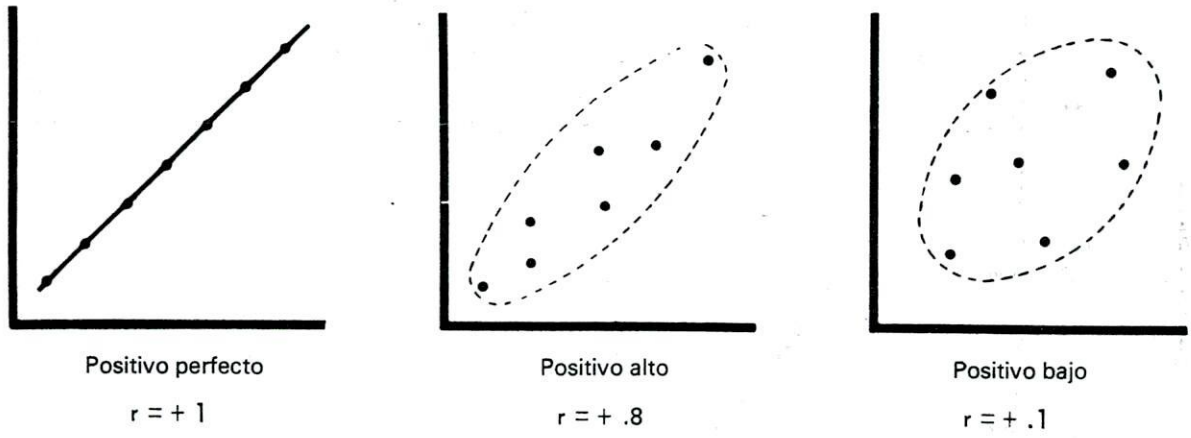


Figura 13.2. Diversos tipos de diagramas de dispersión con sus coeficientes de correlación asociados.

Tabla 13.3. Datos de los cerdos, por rangos.

Rango de ofertas	Rango de precios	Diferencia de rangos (d)	d^2
8	5	3	9
3	3	0	0
2	6.5	-4.5	20.25
9	2	7	49
10	1	9	81
5	8	-3	9
4	9	-5	25
6.5	6.5	0	0
6.5	4	2.5	6.25
1	10	-9	81
	Totales	0.0	280.5

Primero, ordenamos las observaciones en cada columna, de mayor a menor. En el caso de ligaduras, otórguese a cada miembro de la ligadura el rango promedio; por ejemplo, en ambas columnas anteriores, los rangos 6 y 7 presentan ligadura, de modo que ambos se denominan 6.5

Segundo, sustraemos el segundo número del primero en cada hilera y anotamos la diferencia en la columna encabezada por la letra d.

Tercero, elevamos al cuadrado las cifras de la columna d y anotamos el resultado en la columna d^2.En realidad, el segundo paso puede omitirse, puesto que resulta sencillo elevar los números al cuadrado mentalmente y anotarlos directamente en la columna d^2.*

Cuarto, obtenemos el total de la columna d^2. Este total se plantea en la forma Σd^2.

Quinto, calculamos el coeficiente de correlación, r, empleando la fórmula:

$$r = 1 - [6\Sigma d^2/n(n - 1)(n + 1)]$$

donde n es el número de pares de observaciones.

En nuestro ejemplo,

$$r = 1 - [(6 \times 280.5) / (10 \times 9 \times 11)]$$

$$= 1 - 1.70$$

$$= -.70$$

* La mayoría de las *d* deberán ser enteros pequeños. Si terminan en .5, se pueden elevar mentalmente al cuadrado usando la siguiente fórmula $(X + .5)^2 = X(X + 1) + .25$. Así, $4.5^2 = 4 \times 5 + .25 = 20.25$, $7.5^2 = 7 \times 8 + .25 = 56.25$, etc.

La respuesta se encontrará siempre entre + 1 y − 1. El uno, positivo o negativo, representa una correlación perfecta, mientras que el cero indica ausencia total de correlación. Por tanto, en nuestro ejemplo parece existir una correlación negativa bastante alta, de modo que calcularemos el coeficiente con mayor precisión utilizando el método estándar.

METODO ESTÁNDAR

Este se conoce más apropiadamente como **método de producto-momento para el coeficiente de correlación lineal.**

En el capítulo 2 indicamos que la desviación de una X individual de la media de las X's $(X - \bar{X})$ puede representarse por una x minúscula cursiva. Análogamente, podemos utilizar el símbolo y para representar $(Y - \bar{Y})$.

Adoptando estos símbolos más cortos, simplificamos grandemente muchas de las expresiones que encontraremos; dichos símbolos se emplearán a menudo en este capítulo y en los sucesivos.

La fórmula para el coeficiente de correlación puede plantearse de diversas formas. Es conveniente escribirla primero en términos de r^2 y luego encontrar r, obteniendo la raíz cuadrada de la respuesta final:

$$r^2 = [\Sigma (X - \bar{X})(Y - \bar{Y})]^2 / [\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2]. \quad (1)$$

Puesto que $x = X - \bar{X}$ y $y = Y - \bar{Y}$, podemos escribir (1) en forma abreviada:

$$r^2 = (\Sigma xy)^2 / \Sigma x^2 \Sigma y^2 \quad (2)$$

A pesar de que estas formas son sencillas, en general no es fácil calcularlas directamente, puesto que incluyen la elevación al cuadrado de decimales difíciles de manejar. Para evitar esto, aprovechamos la relación $\Sigma x^2 = \Sigma (X - \bar{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$. Sustituyendo Y por X donde sea necesario, podemos replantear (2) en la forma siguiente:

$$r^2 = \left[\Sigma XY - \frac{\Sigma X \Sigma Y}{n}\right]^2 \Big/ \left[\left(\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right)\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}\right)\right] \quad (3)$$

Esta recibe el nombre de "forma operacional".

Aplicando la fórmula (3), podemos calcular ahora el coeficiente de correlación para los datos de nuestro ejemplo, usando el método estándar. Necesitaremos ΣX, ΣY, ΣX^2, ΣY^2 y ΣXY. Encontramos, a partir de los datos, que $\Sigma X = 752$, $\Sigma Y = 179.7$, $\Sigma X^2 = 56\ 804.0$, $\Sigma Y^2 = 3\ 297.53$ y $\Sigma XY = 13\ 420.40$. Por tanto:

$$r^2 = \left[13\ 420.40 - \frac{752 \times 179.7}{10}\right]^2 \Big/ \left[\left(56\ 804.0 - \frac{752^2}{10}\right)\left(3\ 297.53 - \frac{179.7^2}{10}\right)\right]$$

$$= [13\ 420.40 - 13\ 513.44]^2 / [(56\ 804.0 - 56\ 550.4)(3\ 297.53 - 3\ 229.21)]$$

$$= (-93.04)^2 / (253.6 \times 68.32)$$

$$= 0.4996$$

$$r = \sqrt{r^2} = \sqrt{0.4996} = -0.707$$

Nótese que el signo de r debe ser el mismo que el signo de Σxy; en este caso, negativo. La respuesta al método abreviado fue −0.70, muy cercana a la respuesta por el método estándar, −0.707. No debemos entusiasmarnos demasiado con esta coincidencia. Las respuestas por los dos métodos usualmente no son tan cercanas. En el capítulo 14 ilustraremos un caso en que el método abreviado arroja una correlación perfecta, siendo ésta extremadamente engañosa. Se podrían encontrar otros casos en que el método abreviado arrojara una respuesta que sería demasiado baja.

Podríamos utilizar el método abreviado para una rápida comprobación sin el empleo de una máquina calculadora o sólo cuando una respuesta aproximada se estimara suficiente. Para una estimación más eficiente del coeficiente de correlación y para una prueba de significación, debemos utilizar el método estándar.

SIGNIFICACIÓN ESTADÍSTICA

En el último párrafo mencionamos la **significación**. La idea general es la misma que encontramos en el análisis de varianza. Formulamos la hipótesis de que no hay correlación entre las dos variables y que la aparente relación se debe simplemente a la casualidad. Esta suele recibir el nombre de **hipótesis nula.** Entonces planteemos la siguiente pregunta: "Si esta hipótesis nula fuese verdadera, ¿cuál es la probabilidad de que un valor de r obtenido fuera tan grande o mayor que el observado?" Si esta probabilidad es de un 5%, decimos entonces que la correlación es **significativa.** Si afirmamos que la correlación es real, corremos un 5% de riesgo de estar equivocados. Si la probabilidad es de un 1% o menor, decimos que la correlación es **altamente significativa** y rechazamos la hipotesis nula, con sólo un 1% de riesgo de estar equivocados.

Afortunadamente, ya se han hecho y resumido los difíciles cálculos requeridos para la determinación de las probabilidades específicas (tabla A.7). Fijándonos en la tabla sobre la línea opuesta a 8 grados de libertad, encontramos que el coeficiente de correlación de 0.7 podría ocurrir por casualidad en algún lugar entre el 1 y el 5% de las oportunidades; por tanto, podemos decir que la correlación es significativa. Debemos ser muy cuidadosos al interpretar datos de este tipo. Incluso si la correlación es significativa, necesitamos ser cautelosos al afirmar que una fluctuación de la oferta **causa** una fluctuación en el precio. El precio y la oferta pueden estar relacionados con el tiempo, una tercera variable que no se ha considerado en los cálculos. Al final de este capítulo analizaremos algunas de las trampas encontradas en el trabajo con correlación, y daremos un ejemplo para mostrar cuán arriesgado es interpretar la correlación entre dos variables que se hallan relacionadas con el tiempo.

¿Por qué 8 grados de libertad? Hemos estado acostumbrados a utilizar como grados de libertad a **uno** menos el número de observaciones, pero ahora, con 10 pares de observaciones, empleamos como el número de grados de libertad a **dos** menos el número de ítemes u 8. Por primera vez resulta obvio por qué hubo el cuidado de afirmar que los grados de libertad fueron **usualmente** iguales a uno menos el número de observaciones. Esta es la primera excepción que hemos encontrado. La razón comúnmente presentada para la sustracción de dos en vez de uno, es que se pierde un grado en el cálculo de la media y el otro se pierde con la regresión.

Para facilitar aún mas el tema, estudiémoslo en otra forma. Supóngase que tenemos dos pares de observaciones, dos pares cualesquiera, siempre y cuando no sean idénticos. Estos pueden representarse en un gráfico como dos puntos, y puede trazarse una línea a través de los mismos. A dicha línea la denominamos recta de regresión, y los dos puntos se ajustan perfectamente a la misma. Dado que esto podría ser verdadero para cualesquiera dos pares de observaciones, sin importar cuán poca fuera su relación, resultaría absolutamente ridículo otorgarle cualquier significado a un coeficiente de correlación basado únicamente en dos pares de observaciones. Así como una observación resulta sin valor para revelar algo acerca de la variabilidad, dos pares de observaciones nada indican acerca de la correlación.

Para utilizar un ejemplo sencillo sobre este aspecto, los periódicos de la mañana informan que los "Dodgers" hicieron 8 carreras en la noche de ayer y que cierta acción cerró en 51. El día anterior, los "Dodgers" hicieron 4 carreras y la misma acción cerró en 49. A partir de estos datos podemos concluir que tanto las carreras de

los "Dodgers" como el precio de dicha acción están sujetos a variación. Incluso podemos estimar la cantidad de variación en ambos casos, pero la estimación será muy inexacta, puesto que en cada caso estará basada sólo en un grado de libertad, (n — 1). Probablemente podríamos incluso estar seguros en concluir que las carreras de los "Dodgers" son más variables que el precio de la acción; sin embargo, ¿qué podemos decir acerca de la relación entre las dos variables? Resulta sencillo verificar que:

$$r^2 = \frac{(\Sigma xy)^2}{\Sigma x^2 \, \Sigma y^2} = \frac{4^2}{(8)\ (2)} = 1$$

De aquí $r = 1$.

¿No sería absurdo sostener que existió una perfecta correlación entre el número de carreras de los "Dodgers" y el precio de cierta acción en el mismo día?; no obstante, esto es lo que el coeficiente de correlación revela aparentemente. Evitaremos tal absurdo si decimos que dicha correlación estuvo basada en (n — 2) o ceros grados de libertad y que, por tanto, no tiene sentido.

¿Con qué frecuencia el lector ha escuchado a la gente hacer extensas conclusiones respecto a correlaciones basadas en un reducido número de observaciones? Imagínese a una persona que vuela por primera vez entre San Francisco y Denver y que generaliza: "Cuanto más al este nos desplazamos, más frío sentimos." (O si desea que sus palabras tengan más profundidad, podría afirmar: "He observado una correlación positiva entre la temperatura y la longitud.") Este ejemplo no es tan descabellado si consideramos que no pocas veces encontramos a personas que hacen amplias generalizaciones a partir de escasas observaciones. Este es un defecto que debemos tratar de evitar, y la ciencia estadística puede ayudarnos a evadir dichas trampas.

LA RECTA DE REGRESIÓN

Hasta el momento, en nuestro ejemplo sobre la oferta y el precio sólo hemos determinado la estrechez de la relación y la probabilidad de que se debiera a la casualidad. No hemos aprendido nada acerca de **cómo** las dos variables están relacionadas.

Si suponemos que la relación es lineal, es decir, se describe mejor mediante una línea recta, el problema se reduce a encontrar la línea recta en particular que se ajusta más estrechamente a los datos. ¿Qué entendemos por **más estrechamente ajustada**? Si observamos el gráfico de los datos, resulta obvio que no se puede construir ninguna recta que pase a través de todos los puntos. No importa qué recta construyamos, diversos puntos se desviarán de dicha recta. Medimos la variación entre un solo conjunto de observaciones y obtenemos la suma de cuadrados de las desviaciones de la media. Luego, parece lógico medir la variación de una recta, obteniendo la suma de cuadrados de las desviaciones de la recta. Utilizando esta medida como el criterio para la exactitud del ajuste, trataremos de encontrar la línea recta que hará la suma de cuadrados de las desviaciones tan pequeña como sea posible. Tal procedimiento recibe el nombre de **método de mínimos cuadrados.** Quienes estén familiarizados con el cálculo, inmediatamente reconocerán este problema como un caso típico que contempla la determinación del valor mínimo de una función.

La solución al problema resulta muy sencilla. En términos de desviaciones de las medias de X y Y, la ecuación de la recta más apropiada es:

$$\hat{y} = \left(\frac{\Sigma xy}{\Sigma x^2}\right)x$$

($\hat{y}$ equivale a: "el valor estimado de y").

La expresión $\frac{\Sigma xy}{\Sigma x^2}$ es el **coeficiente de regresión,** puesto que indica el cambio estimado de y, para un cambio unitario de x. Esto se ajusta a nuestra definición de regresión, y ya hemos denominado b al coeficiente de regresión, de modo que ahora podemos decir: $b = \Sigma xy / \Sigma x^2$. Más precisamente, debemos denominarlo "el coeficiente de regresión de Y sobre X", y utilizar el símbolo b_{yx}. En general si b, se utiliza sin subíndice, éste es el coeficiente subentendido.

La ecuación presentada en el párrafo anterior, puede replantearse en términos de las observaciones en sí mismas, en vez de en términos de desviaciones de las medias.

Podemos escribir: $(\hat{Y} - \bar{Y}) = b\,(X - \bar{X})$

Que puede replantearse como: $\hat{Y} = (\bar{Y} - b\bar{X}) + bX$

Si hacemos $\bar{Y} - b\bar{X} = a$, la ecuación puede escribirse como $\hat{Y} = a + bX$, que es la forma intercepto-pendiente de la ecuación de la recta mencionada al comienzo de nuestro estudio sobre regresión.

Ahora veamos cómo se aplica esta ecuación a nuestros datos. Tenemos ya todas las sumas que necesitamos para el cálculo de r, es decir, el coeficiente de correlación. Ahí encontramos que:

$\Sigma X = 752$; entonces: $\bar{X} = 752/10 = 75.2$

$\Sigma Y = 179.7$; entonces: $\bar{Y} = 17.97$

$\Sigma xy = -93.04$

$\Sigma x^2 = 253.6$; entonces: $b = -93.04/253.6 = -0.367$

Por tanto, sustituyendo en la ecuación:

$\hat{Y} = (\bar{Y} - b\bar{X}) + bX$, obtenemos

$\hat{Y} = [17.97 - (-0.367)\ \ 75.2] \quad +(-0.367)X$

$\hat{Y} = 45.57 - 0.367X$

Esta ecuación puede expresarse de la siguiente manera: "Iniciando con un precio base de 45.57 dólares por quintal, cada incremento unitario (millón) en las ventas anuales de cerdo está asociado con una reducción promedio de 0.367 dólares, del precio por quintal."

Comparemos los valores de Y observados con los valores estimados, ($\hat{Y}$'s), basados en la ecuación de regresión.

Tabla 13.4. Precios observados y estimados del cerdo.

X	Y	$\hat{Y} = 45.57 - .367X$		$d = Y - \hat{Y}$	d^2
73	18.0	18.8		- 0.8	0.64
79	20.0	16.6		3.4	11.56
80	17.8	16.2		1.6	2.56
69	21.4	20.2		1.2	1.44
66	21.6	21.4		0.2	0.04
75	15.0	18.1		-3.1	9.61
78	14.4	16.9		-2.5	6.25
74	17.8	18.4		- 0.6	0.36
74	19.6	18.4		1.2	1.44
84	14.1	14.7		- 0.6	0.36
			Totales	0.0	34.26

El hecho de que la suma de las desviaciones sea igual a cero sirve como una comprobación de los cálculos. Esto será siempre verdadero (excepto para errores de aproximación). La suma de cuadrados de las desviaciones puede calcularse en una forma muy simple, a partir de la siguiente fórmula:

$$\Sigma d^2 = (1 - r^2) \Sigma y^2$$

En nuestro ejemplo:

$$\Sigma d^2 = (1 - 0.4996)\, 68.32 = 34.19,$$

una respuesta muy cercana a 34.26, mostrada en la tabla 13.4. La pequeña diferencia se debe a la aproximación.

Esta suma de cuadrados, Σd^2, recibe el nombre de **suma de cuadrados debida a la desviación de la regresión,** y la raíz cuadrada de la cantidad $\Sigma d^2 \ (n-2)$ se denomina **error estándar de la estimación.** Este es precisamente otro tipo de error estándar similar a aquellos que hemos encontrado anteriormente. Es una medida de la cantidad de variación de la recta de regresión.

En general, no es necesario afrontar todos los problemas para construir una tabla como la 13.4, a fin de verificar la corrección de la recta de regresión. La construcción de la recta sobre el diagrama de dispersión revelará, por regla general, cualquier error craso. La construcción de la recta es muy simple, puesto que para la determinación de una recta cualquiera sólo se requieren dos puntos. Un punto puede estar sobre el eje de las Y, a a unidades (en este caso, 45.57) del origen. El otro puede ser el punto que representa a $\overline{X}$ (la media de X) y a Y (la media de Y). La recta que pasa a través de estos dos puntos recibirá el nombre de recta de regresión. La figura 13.3 muestra la recta de nuestro ejemplo trazada entre los puntos observados. Las líneas punteadas entre los puntos observados y la recta de regresión representan las desviaciones.

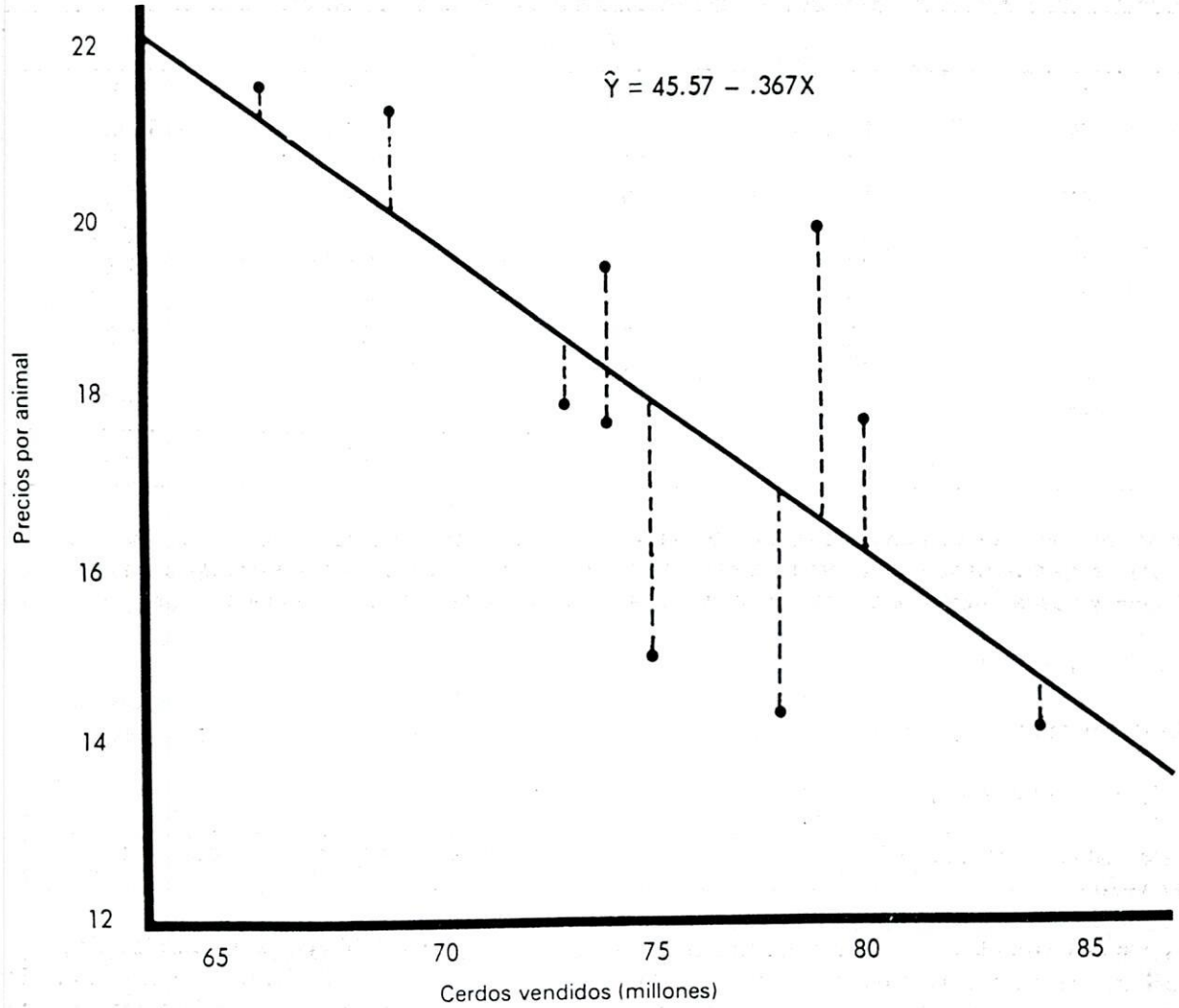


Figura 13.3. Recta de regresión para los datos del cerdo, que muestra desviaciones a partir de la misma.

Notará el lector que las desviaciones están representadas como rectas **verticales.** Esta es la suma de cuadrados de tales desviaciones que hemos minimizado para obtener la **recta más estrechamente ajustada.** Supóngase que decidimos construir una recta de modo que la suma de cuadrados de las desviaciones horizontales de los puntos a la recta sea mínima. ¿Obtendremos así la misma recta? La respuesta es no, a menos que exista una correlación perfecta. Esta nueva recta tendrá la ecuación:

$$\hat{x} = \left(\frac{\Sigma xy}{\Sigma y^2}\right) Y.$$

La expresión $\Sigma xy / \Sigma y^2$ recibe el nombre de **coeficiente de regresión de X sobre Y**, y se denota b_{XY}. Ahora debe ser evidente por qué fuimos cuidadosos al puntualizar que el símbolo b se sobreentiende por b_{YX} la regresión de Y sobre X, a menos que se haga alguna otra especificación.

Existe una razón para mencionar que hay dos **rectas más apropiadas,** de acuerdo al sentido en que se toman las desviaciones. Nótese que:

$$b_{YX} \cdot b_{XY} = \frac{\Sigma xy}{\Sigma x^2} \cdot \frac{\Sigma xy}{\Sigma y^2} = r^2.$$

Esto hace resaltar la relación entre los coeficientes de regresión y el coeficiente de correlación.

Podemos ahora responder a las preguntas planteadas acerca de los datos de nuestro ejemplo.

1. ¿Cuán estrecha fue la relación entre la oferta y el precio?

   Respuesta: totalmente estrecha. El coeficiente de correlación fue de -0.7, y ± 1 sería el perfecto.

2. ¿Cuál es la probabilidad de que tal correlación pudiera deberse a la casualidad?

   Respuesta: una correlación de este tamaño de 10 pares de observaciones sólo podría ocurrir por casualidad entre el 5 y el 1% de las veces.

3. ¿Qué ecuación describiría mejor la relación entre el precio (Y) y la oferta (X) para dichos datos?

   Respuesta: $\hat{Y} = 45.57 - .367X$

4. ¿Hasta qué punto se ajusta esta recta a los datos?

   Respuesta: la suma de cuadrados de las desviaciones de los puntos observados de la recta fue igual a 34.19 o aproximadamente la mitad de la variación total del precio. Luego, sólo la mitad de la variación del precio estuvo asociada de algún modo con la variación de la oferta. Una simple tabla de análisis de varianza revela esto (tabla 13.5).

El hecho de que el valor F de 7.99 es sólo ligeramente mayor que el F requerido en el punto de 2.5% para 1 y 8 grados de libertad (7.57), comprueba nuestro hallazgo previo en la respuesta de la pregunta número 2.

Tabla 13.5. Análisis de regresión dispuesto en forma de análisis de varianza.

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F
Total	9	$\Sigma y^2 = 68.32$		
Regresión	1	$r^2 \Sigma y^2 = 34.13$	34.13	7.99*
Desviación de la regresión	8	$(1 - r^2) \Sigma y^2 = 34.19$	4.27	

*Significativo en el nivel del 5%

TRAMPAS

Probablemente ninguna otra parte de la estadística esté sujeta a tantos abusos y malas interpretaciones como la correlación y la regresión. El aserto de que "se puede probar cualquier cosa con la estadística" es cierto sólo si ignoramos algunos de los principios básicos contemplados. Los principios más frecuentemente ignorados en correlación son:

1. El nombre completo del coeficiente de correlación es coeficiente de correlación **lineal,** y

2. Nada en la definición de correlación indica o implica que la relación entre dos variables sea de tipo causal. A continuación damos algunos ejemplos de lo fácil que resulta meterse en problemas.

Una correlación baja no siempre significa ausencia de relación. Fijémonos en los siguientes pares de cifras:

X	Y
0	0
1	144
2	256
3	336
4	384
5	400
6	384
7	336
8	256
9	144
10	0

Si calculamos el coeficiente de correlación entre X y Y, encontraremos que éste es igual a cero; sin embargo, si concluimos que no existe relación entre X y Y, podríamos estar completamente equivocados. X es el tiempo transcurrido en segundos después del disparo vertical de una flecha a 160 pies/seg. Y es la elevación de la flecha en pies. Desde luego, resulta absolutamente ridículo afirmar que no hay relación entre la altura de una flecha y su tiempo de vuelo. ¿Cuál es el quid de esta paradoja? La importante palabra **lineal**, implícita cuando hablamos del coeficiente de correlación, fue ignorada. Es cierto que ninguna línea recta resultará adecuada a estos datos, pero la ecuación $Y = 160X - 16X^2$ dará un ajuste perfecto. Esta es la ecuación de la parábola.

La moraleja de este ejemplo es que debemos estar a la expectativa de correlaciones curvilíneas que puedan ajustarse a los datos mejor que una simple correlación lineal. Las formas de manipulación de datos de este tipo se presentarán más adelante.

Una alta correlación no necesariamente significa una relación de causa y efecto. Considérese la tabla 13.6 a partir de la cual podemos calcular el coeficiente de correlación y la ecuación de regresión.

Tabla 13.6. Quince pares de datos altamente correlacionados.

X	Y	
295	73	$\Sigma X = 5669$
339	78	$\bar{X} = 377.9$
343	85	$\Sigma X^2 = 2\ 163\ 935$
344	91	$(\Sigma X)^2/15 = 2\ 142\ 504$
357	100	$\Sigma x^2 = 21\ 431$
359	109	$\Sigma Y = 1768$
368	119	$\bar{Y} = 117.9$
395	125	$\Sigma Y^2 = 218\ 482$
414	129	$(\Sigma Y)^2/15 = 208\ 388$
406	135	$\Sigma y^2 = 10\ 094$
385	142	$\Sigma XY = 681\ 962$
394	139	$(\Sigma X \Sigma Y)/15 = 668\ 186$
404	140	$\Sigma xy = 13\ 776$
420	147	
446	156	

$r^2 = (\Sigma xy)^2/\Sigma x^2 \Sigma y^2$

$= (13\ 776)^2/(21\ 431 \times 10\ 094)$

$= 189\ 778\ 176/216\ 324\ 514 = .8773$

$r = \sqrt{.8773} = .937$ (coeficiente de correlación)

$b = \Sigma xy/\Sigma x^2 = 13\ 776/21\ 431 = .643$ (coeficiente de regresión)

$a = \bar{Y} - b\bar{X} = 117.9 - 243.0 = 125.1$ (intercepto)

$\hat{Y} = -125.1 + .643X$ (ecuación de regresión)

Este alto valor del coeficiente de correlación, 0.937, indica una estrecha relación entre X y Y. Podemos tener la sensación de afirmar que cada cambio unitario de X **causa** un cambio de 0.643 en Y. Ahora veamos qué representan X y Y. Las X son el número de cigarros consumidos anualmente en los Estados Unidos (en millares de millón) entre 1944 y 1958. Las Y son las cifras del índice de producción por hora-hombre en cultivos de heno y forraje durante el mismo periodo. Se requeriría un gran alcance de imaginación para pensar en cualquier causa directa y relación de efecto entre el consumo de cigarros y la eficiencia del negocio de heno. Precisamente sucedió que ambas variables mostraron un incremento estable con el tiempo durante el periodo considerado.

La moraleja de este ejemplo es que el coeficiente de correlación medirá la estrechez de la **relación** entre dos variables, pero nada revela acerca de si dicha relación es de tipo **causal**. Esta decisión queda en manos del investigador, y debe estar basada en una gran cantidad de conocimientos de las variables objeto de estudio.

Vigilar las correlaciones entre partes del todo. Hace algunos años, en una reunión meteorológica se presentó un artículo referente a los estudios sobre la longitud de las estaciones en crecimiento entre las heladas dañinas. Se informó que existió poca o ninguna correlación entre la última helada de la primavera y la primera helada del otoño durante un largo periodo. La siguiente conclusión reportada fue que existió una correlación bastante alta entre las fechas de la última helada de la primavera y la longitud de las estaciones.

Si examinamos esta segunda conclusión, notaremos que la longitud de la estación está completamente determinada por dos partes: el comienzo (última helada de primavera) y el final (primera helada del otoño). Se puede demostrar fácilmente que si una variable está constituida por dos o más partes independientes, automáticamente existe una correlación entre cualesquiera de las partes y el todo. La relación es sencilla: r = (desviación estándar de la parte)/(desviación estándar del todo). En el caso de la fecha de la helada, si las fechas de las heladas de primavera y las fechas de la heladas de otoño son casi igualmente variables, entonces esperamos que la correlación entre las fechas de las heladas de primavera y la longitud de la estación sea de aproximadamente $\sqrt{.5}$ o 0.707.

La conclusión acerca de la correlación entre la helada de primavera y la duración de la estación, a pesar de ser correcta, fue trivial.

La extrapolación es tentadora pero peligrosa. Frecuentemente, una serie de observaciones caen dentro de un rango bastante restringido de valores para las dos variables objeto de estudio. Si éstas muestran un alto coeficiente de correlación, hay una gran tendencia a extender la recta de regresión más allá del rango de observaciones y de tratar de predecir qué les sucedería a los valores de Y si X fuese tomado en valores por encima o por debajo de aquellos en realidad observados. Esto recibe el nombre de **extrapolación.**

Esta es una práctica peligrosa, puesto que muchas variables que se encuentran relacionadas en forma curvilínea darán una alta correlación lineal sólo si una pequeña sección de la curva se toma como muestra.

La tabla 13.7 resume las mediciones de diez bulbos de cebolla con diámetros entre 50 y 70 milímetros con sus correspondientes pesos en gramos.

Tabla 13.7. Mediciones de diez bulbos de cebolla.

Diámetro (X)	Peso (Y)
51.0	63.4
66.2	115.3
69.2	146.6
69.5	132.6
56.9	80.7
67.1	125.6
58.1	80.0
53.9	78.7
63.0	112.8
60.0	96.2

El cálculo de r, el coeficiente de correlación, y de la ecuación de regresión es como sigue:

$\Sigma X = 614.9$ $\quad\quad$ $\Sigma Y = 1031.9$

$\bar{X} = 61.49$ $\quad\quad$ $\bar{Y} = 103.19$

$\Sigma X^2 = 38\ 192.17$ $\quad\quad$ $\Sigma Y^2 = 113\ 247.79$ $\quad\quad$ $\Sigma XY = 65\ 014.60$

$(\Sigma X)^2/n = 37\ 810.20$ $\quad\quad$ $(\Sigma Y)^2/n = 106\ 481.76$ $\quad\quad$ $\Sigma X \Sigma Y/10 = 63\ 451.53$

$\Sigma x^2 = 381.97$ $\quad\quad$ $\Sigma y^2 = 6\ 766.03$ $\quad\quad$ $\Sigma xy = 1\ 563.07$

$r^2 = (1\ 563.07)^2/(381.97 \times 6\ 766.03) = .9454$

$r = \sqrt{.9454} = .97$ (coeficiente de correlación)

$b = 1\ 563.07/381.97 = 4.092$ (coeficiente de regresión)

$a = 103.19 - (4.092)\ (61.49) = -\ 148.43$ (intercepto)

$\hat{Y} = 4.092X - 148.43$ (ecuación de regresión)

La correlación de 0.97 entre el diámetro y el peso es muy alta. (Esto no es sorprendente.) Dentro del rango de 50 a 70 milímetros, una ecuación de la recta describe muy bien la relación entre las dos variables.

Realicemos ahora la extrapolación y veamos qué sucede. Un bulbo que midió 92.4 milímetros arrojó un peso de 300.2 gramos; pero nuestra estimación del peso a partir de la ecuación de regresión equivale a 229.7 gramos. La extrapolación hizo que nos equivocáramos en 70.5 gramos en nuestra estimación. Por otro lado, un bulbo que midió 37.8 milímetros pesó 27.8 gramos, pero la extrapolación dio una estimación de 6.2 gramos. La extrapolación para valores aún menores de X pronto dará estimaciones completamente absurdas de Y; por ejemplo, un bulbo de 36.27 milímetros sería estimado sin peso y todos los bulbos menores que éste obtendrían estimaciones negativas. La figura 13.4 muestra la recta ajustada a los datos y los efectos de la extrapolación.

Es fácil observar por qué la extrapolación nos conduce por caminos tan equivocados en este caso. La ecuación de regresión lineal implica que una cantidad dada, añadida al diámetro de un bulbo, acrecentará cierta cantidad fija de peso; sin embargo, debe resultar obvio que esto no puede ser así.

Un centímetro acrecentado a un bulbo de 9 centímetros, ciertamente resultará en un mayor aumento de peso que un centímetro añadido a un bulbo de 2 centímetros.

Si deseamos determinar en qué forma están relacionadas dos variables fuera del rango de nuestras observaciones, el procedimiento más seguro consiste en realizar mayores observaciones en el área que nos interesa.

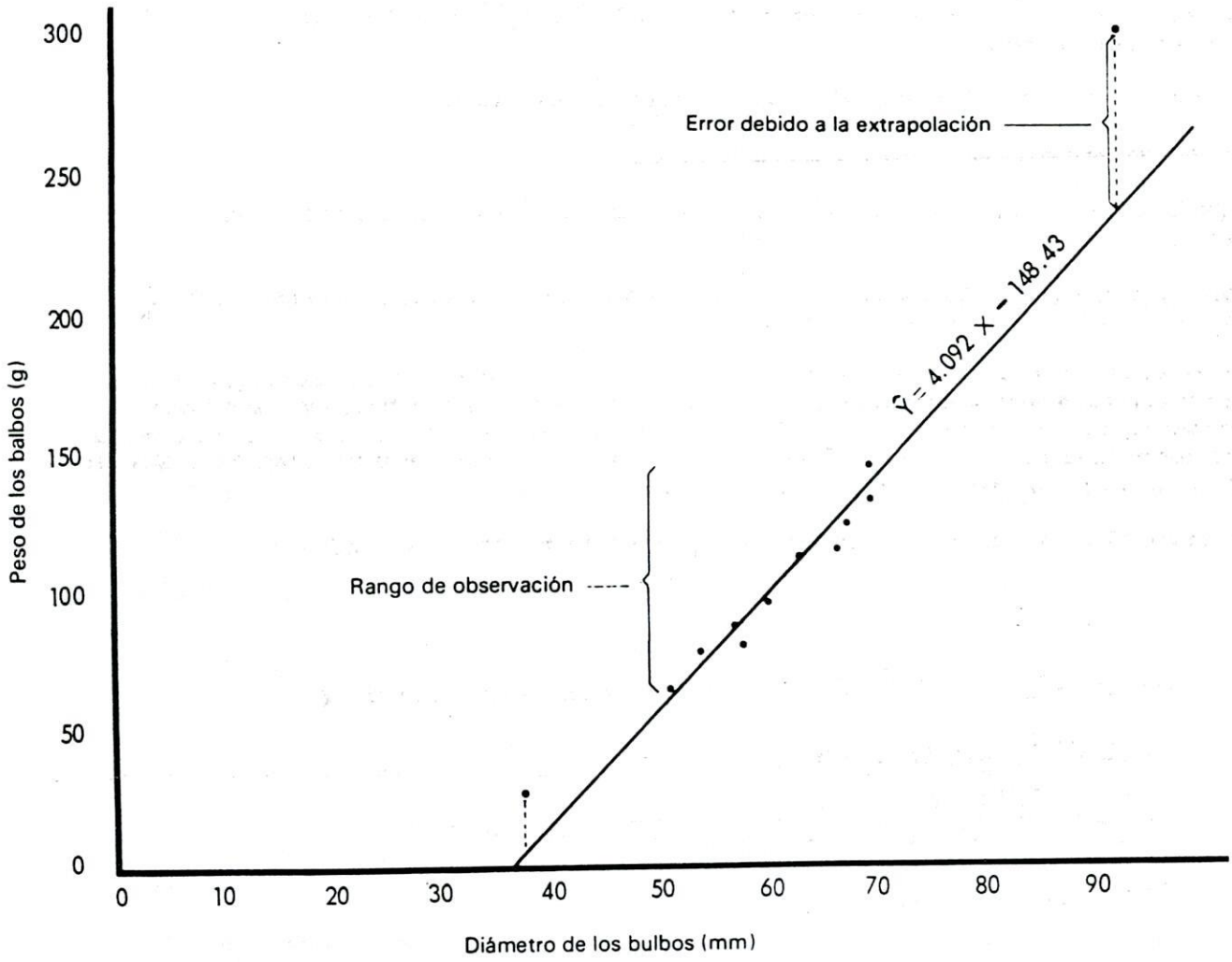


Figura 13.4. Recta de regresión trazada a través de los datos de la cebolla para un rango limitado, mostrando el peligro de la extrapolación a partir de observaciones limitadas.

RESUMEN

Correlación es la tendencia de dos variables a estar relacionadas en una forma definida.

Las dos variables reciben el nombre de **independiente** y **dependiente,** de acuerdo con la manera en que se conciba a una dependiendo de la otra.

La variable independiente se denomina X y la variable dependiente se denota por Y.

El **coeficiente de correlación** mide la **estrechez** de la relación.

Regresión es la cantidad de cambio de la variable dependiente asociado con un cambio unitario de la variable independiente.

La ecuación de regresión lineal se escribe $\hat{Y} = a + bX$, donde $\hat{Y}$ es el **valor estimado de Y,** *a* es el **intercepto** o punto donde la recta corta al eje de las *x,* y *b* es la **pendiente** o **coeficiente de regresión.**

La representación gráfica de un conjunto de datos constituidos por pares de elementos da lugar a un **diagrama de dispersión.** Por regla general, éste constituye un conveniente primer paso en el análisis de regresión. Un **método abreviado rápido** conocido como **método de diferencia de rango** facilita el cálculo de una aproximación al coeficiente de correlación. La fórmula es: $r = 1 - [6\Sigma d^2/n(n-1)(n+1)]$ donde r es el coeficiente de correlación, d es la diferencia de rango en cada par de observación, y n es el número de pares.

El **método estándar** o **método de producto-momento** puede expresarse con diversas fórmulas:

$$r^2 = [\Sigma(X - \bar{X})(Y - \bar{Y})]^2 \Big/ [\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2] \quad \text{(forma de observación directa)}$$

$$r^2 = (\Sigma xy)^2 / \Sigma x^2 \Sigma y^2 \quad \text{(forma de desviación de la media)}$$

$$r^2 = \left[\Sigma XY - \frac{\Sigma X \Sigma Y}{n}\right]^2 \Big/ \left[\left(\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right)\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}\right)\right] \quad \text{(forma operacional)}$$

$$r^2 = b_{YX} b_{XY} \quad \text{(forma de regresión)}$$

$r = \pm \sqrt{r^2}$ (El signo corresponderá al signo del número dentro de los corchetes de los numeradores de las ecuaciones 1 a 3. Este corresponderá al signo de b_{XY} en la ecuación 4.)

La significación del coeficiente de correlación puede determinarse mediante una tabla especial de r, utilizando n—2 grados de libertad, donde n es el número de pares de observaciones.

Las correlaciones basadas en sólo dos pares de observaciones serán siempre más o menos uno, pero éstas no tienen sentido.

El **coeficiente de regresión** es: $b = \Sigma xy / \Sigma x^2$

El intercepto es: $a = \bar{Y} - b\bar{X}$.

Cuando *a* y *b* están determinados, podemos plantear la ecuación de regresión como $\hat{Y} = a + bX$.

La falta de coincidencia entre los valores de Y observados y estimados se mide por la **suma de cuadrados debida a la desviación de la regresión,** obtenida a partir de la siguiente relación:

$$\Sigma d^2 = (1 - r^2)\ \Sigma y^2.$$

La suma de cuadrados debida a las desviaciones dividida entre sus grados de libertad (n — 2) da el **cuadrado medio debido a las desviaciones de la regresión.**

La raíz cuadrada del cuadrado medio, debida a las desviaciones de la regresión, recibe el nombre de **error estándar de la estimación.**

La suma de cuadrados debida a la regresión puede obtenerse directamente:

$SCR = r^2 \Sigma y^2$ o, por sustracción: $SCR = \Sigma y^2 - \Sigma d^2$.

El **cuadrado medio debido a la regresión** es igual a la suma de cuadrados, puesto que éste sólo tiene un grado de libertad.

Una prueba de significación basada en $F = \dfrac{\text{(cuadrado medio para la regresión)}}{\text{(cuadrado medio para las desviaciones)}}$ puede verificarse en una tabla de F bajo 1 y (n — 2) grados de libertad. Esto suministra la misma prueba que si consultáramos una tabla de r.

Siempre debe recordarse que el coeficiente de correlación ordinario supone una relación **lineal** entre las dos variables; además, ésta no puede ayudarnos a decidir si la relación es de tipo **causal** y **efecto.**

Un bajo coeficiente de correlación no siempre significa ausencia de relación. Puede existir una **relación curvilínea** muy estrecha.

Un alto coeficiente de correlación no implica una relación directa de causa y efecto. Las dos variables pueden estar simplemente relacionadas con una tercera variable, como el tiempo.

Evítese la correlación de una variable con una de sus componentes. Las conclusiones alcanzadas son igualmente triviales.

Evítese la extrapolación de una recta de regresión más allá del rango de observaciones.

Capítulo 14

Relaciones curvilíneas

En el capítulo anterior insistimos repetidamente en la necesidad de tener en mente que los coeficientes de correlación y regresión usuales están basados en una relación **lineal** entre dos variables. La relación lineal es el tipo más sencillo de relación encontrada entre dos variables. Incluso si existen desviaciones pronunciadas de la linealidad para valores extremos de X y Y, frecuentemente sucede que, dentro del rango útil o práctico de valores de las variables, una recta es suficiente para caracterizar la relación; por ejemplo, en pruebas con fertilizantes, frecuentemente notamos que se registra un incremento sostenido en la producción con aplicaciones cada vez mayores de algún nutriente hasta un punto dado. Más allá de dicho punto, el incremento de la producción puede ser menos pronunciado, y finalmente la producción en realidad disminuye cuando utilizamos excesivas cantidades de abono. Si estamos interesados exclusivamente en aplicaciones bajas a medianas de fertilizantes, una recta puede resultar satisfactoria para describir la relación entre la producción y el fertilizante. Si deseamos describir esta relación a través de todo el rango de aplicaciones desde cero hasta extremadamente altas, probablemente tendremos que utilizar una curva que alcance un punto máximo y luego disminuya.

CÓMO DECIDIR QUÉ CURVA UTILIZAR

Puesto que existen múltiples tipos de curvas que podemos utilizar para expresar la relación entre dos variables, primero debemos decidir qué tipo de curva vamos a tratar de ajustar a los datos. Resultaría deseable encontrar una curva que expresara alguna relación **natural** estre las dos variables; pero esto no siempre es posible. Algunas veces, el conocimiento cabal y la experiencia con las variables que estamos estudiando nos capacita para elegir un tipo de curva que es más logica que las demás. Citaremos más ejemplos de esto a medida que avancemos en nuestro estudio. Otras veces lo inverso es verdadero. Encontrar una curva que se ajuste estrechamente a los datos puede darnos una importante pista, como la relación natural que existe entre dos variables. Muchas de nuestras leyes naturales fueron descubiertas en esta forma; por ejemplo, la ley de Boyle, la de Charles y la ley de la caída de los cuerpos.

Tratándose de datos biológicos, la relación entre dos variables puede ser tan compleja que ninguna ecuación simple satisfaga la descripción de la misma. A menudo debemos conformarnos con hallar una ecuación que se ajuste razonablemente a los datos, sin exigir que la ecuación exprese cualquier relación natural. Siempre es posible encontrar una curva que se ajuste de manera perfecta a los datos, pero tal curva puede ser estrictamente artificial y completamente desprovista de significado físico o biológico.

Entre una multitud de tipos de curvas, hemos seleccionado cuatro para su consideración, por dos razones: primera, porque éstas son las más comúnmente encontradas tratándose de datos biológicos y económicos; y, segunda, porque para su estudio sólo se requieren ideas matemáticas elementales.

La línea recta de tipo logarítmica

Por línea recta de tipo logarítmica entendemos cualquier curva que pueda transformarse en una línea recta mediante la toma de los logaritmos de X y Y como variables, en vez de los valores originales. La forma general de la ecuación de una curva de este tipo es:

$$Y = aX^b$$

Si aplicamos logaritmos a ambos miembros de esta ecuación, obtendremos:

$$\log Y = \log a + b \log X^*$$

Si hacemos que los logaritmos de X y de Y sean las variables, denominándolas X' y Y', y la constante log a se denomina a', podemos replantear la ecuación de la siguiente forma:

$$Y' = a' + bX'$$

Esta es fácilmente reconocible como la ecuación general de la recta, estudiada en el capítulo anterior; por tanto, todo lo que tenemos que hacer para analizar los datos de este tipo es transformar las observaciones en logaritmos y luego proceder exactamente como lo hicimos con la correlación lineal y la regresión.

El valor de b puede ser positivo o negativo y contener números enteros o fracciones. La figura 14.1 muestra algunos ejemplos de la gran variedad de formas de curvas que resultan de diferentes valores de b. Después de la transformación de X y Y a logaritmos, todas estas curvas se vuelven líneas rectas con pendiente b, como se muestra a la derecha de la figura.

El efecto de a en las curvas originales es comprimir o expandir la escala en uno de los ejes, mientras que su efecto sobre la recta logarítmica transformada es simplemente desplazarla hacia arriba o hacia abajo sin cambiar su pendiente.

Puesto que sólo los números positivos tienen logaritmos, la forma logarítmica de las ecuaciones no tiene significado para valores negativos de X. Por tanto, debemos aplicar la transformación logarítmica sólo a aquellos datos donde todas las observaciones de X y Y sean positivas. Esta no es realmente una restricción muy seria, puesto que diversas mediciones físicas, como el peso, la longitud, el área, etc., toman sólo valores positivos.

¿Cómo sabemos que es plausible emplear la transformación logarítmica? Nuevamente aquí, la utilización de un gráfico brinda un buen comienzo. La representación gráfica puede hacerse en dos formas. Los valores observados de X y Y pueden convertirse en logaritmos y marcarse en un papel cuadriculado ordinario. Un método aún más simple es trazar los valores originales en un papel cuadriculado denominado papel logarítmico. Con cualquiera de los dos métodos, el resultado obtenido será un diagrama de dispersión. Si este diagrama de dispersión tiene la apariencia de una larga y estrecha elipse, típica de datos correlacionados en forma lineal, podemos proceder al análisis de los logaritmos de X y Y.

Desde un punto de vista lógico, cabría esperar datos basados en mediciones que incluyeran dos números diferentes en dimensiones para ajustar curvas de la forma $Y = aX^b$; por ejemplo, la altura es unidimensional, mientras que el peso, estando relacionado con el volumen, es tridimensional; por tanto, al correlacionar la altura con el peso, sería lógico probar la transformación logarítmica. Lo mismo ocurriría con mediciones de amplitud y área, longitud y volumen, superficie y diámetro, etc.

En el capítulo anterior, al examinar los peligros de la extrapolación, presentamos algunos datos de los diámetros y pesos de los bulbos de cebolla. Puntualizamos ahí que una línea recta describe bastante bien la

*Para que quienes no recuerdan las reglas de logarítmos y exponentes, un repaso resultará beneficioso para este estudio. En ese sentido, puede consultarse cualquier texto de álgebra elemental.

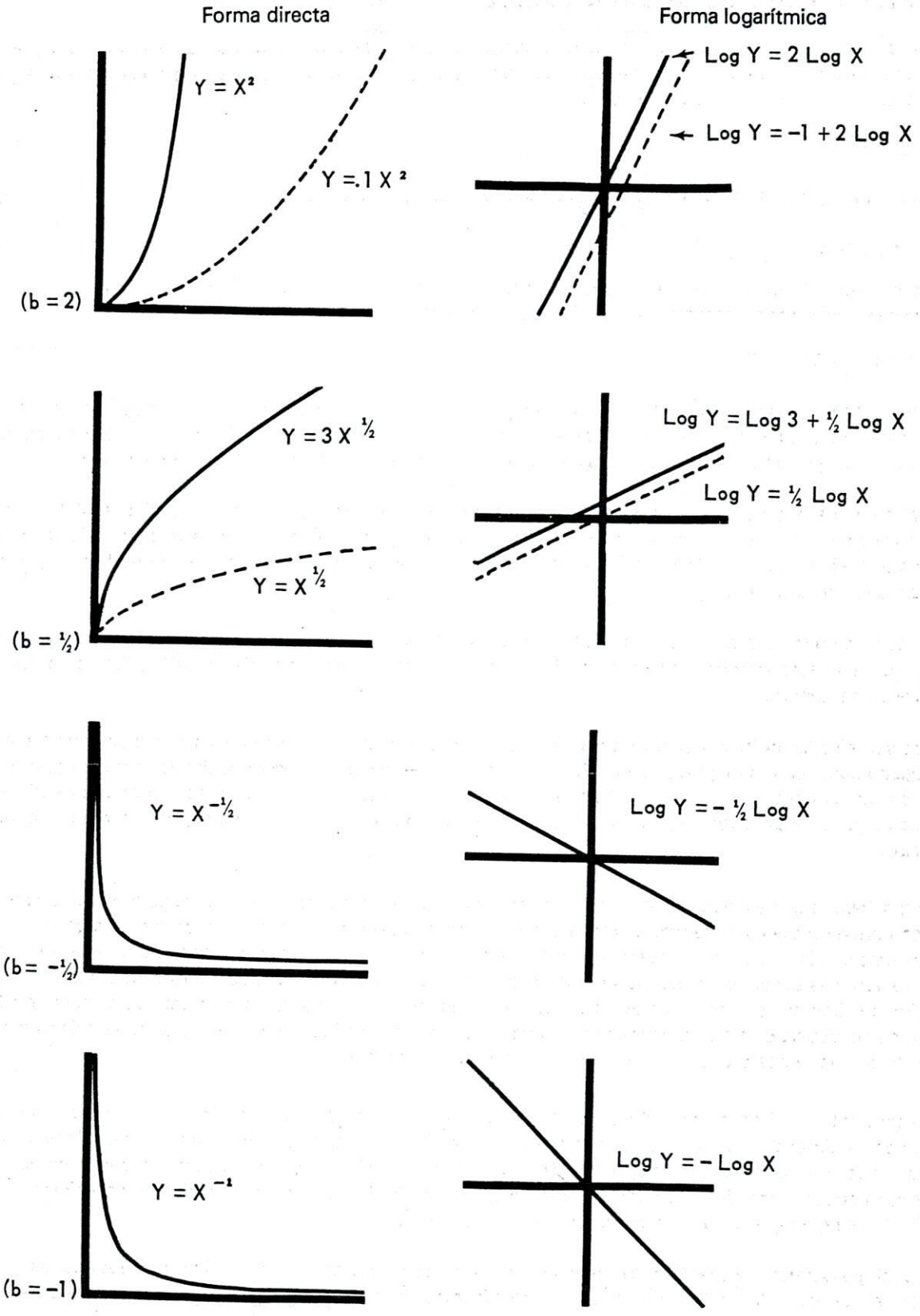


Figura 14.1. Diversas curvas formadas con sus transformaciones logarítmicas, mostrando cómo los logaritmos pueden convertir curvas en líneas rectas.

relación si sólo consideramos un rango pequeño de diámetros. Si esta recta fuese extendida hacia cualquier dirección más allá del rango de observaciones, fracasaría en brindar una buena representación de la relación entre el diámetro y el peso. Si nos detenemos y pensamos al respecto, esto es en realidad lo que cabría esperar. Se esperaría que un centímetro añadido a un gran bulbo significara un mayor aumento de peso que un centímetro añadido al diámetro de un bulbo pequeño; además, si los bulbos fueran esféricos, la relación entre el diámetro y el volumen sería:

$$V = \pi d^3 / 6.$$

Si la gravedad específica de los bulbos permaneciera completamente constante a través de todos los tamaños de bulbo, el peso sería una función lineal directa del volumen; por tanto, esperaríamos que el peso (Y) fuese una función del cubo del diámetro (X).

Tabla 14.1 Diámetros y pesos de bulbos de cebolla.

Diámetro (X)	Peso (Y)	Log X(X′)	Log Y(Y′)
35.1	24.3	1.54531	1.38561
35.3	24.1	1.54777	1.38202
35.5	24.4	1.55023	1.38739
37.8	27.8	1.57749	1.44404
37.8	28.7	1.57749	1.45788
41.4	42.0	1.61700	1.62325
41.7	34.5	1.62014	1.53782
44.8	56.1	1.65128	1.74896
44.9	49.0	1.65225	1.69020
47.9	58.4	1.68034	1.76641
51.0	63.4	1.70757	1.80209
53.9	78.7	1.73159	1.89597
56.9	80.7	1.75511	1.90687
58.1	80.0	1.76418	1.90309
60.0	96.2	1.77815	1.98318
63.0	112.8	1.79934	2.05231
66.2	115.3	1.82086	2.06183
67.1	125.6	1.82672	2.09899
69.2	146.6	1.84011	2.16613
69.5	132.6	1.84198	2.12254
70.7	142.8	1.84942	2.15473
73.1	137.1	1.86392	2.13704
73.1	163.2	1.86392	2.21272
77.4	180.0	1.88874	2.25527
81.7	198.0	1.91222	2.29667
81.7	207.8	1.91222	2.31765
82.3	190.8	1.91540	2.28058
83.1	225.5	1.91960	2.35315
84.6	237.0	1.92737	2.37475
92.4	300.2	1.96567	2.47741
Totales 1817.2	3383.6	52.90339	58.27655
Sumas de cuadrados 118 958.58	542 675.26	93.80268806	116.4541216
Sumas de productos en cruz 241.772.67		104.0495715	

La verdadera situación con las cebollas no es tan simple, puesto que éstas raramente tienen forma esférica, pero son esferoides con una sección longitudinal elíptica; además, cuando los bulbos crecen, cambian continuamente de forma, siendo esferoides prolato cuando son pequeños, casi esféricos en algún tamaño mediano, y esferoides achatados cuando son grandes. Este constante cambio en la forma resulta del hecho de que los bulbos crecen más rápidamente en diámetro que en longitud. No obstante estas complejidades, parece que el tipo de datos con los que estamos tratando puede simplificarse considerablemente mediante una transformación logarítmica.

La tabla 14.1 muestra los diámetros y pesos observados con 30 bulbos, dispuestos por orden de sus diámetros.
Primero calcularemos los coeficientes de correlación y la ecuación de regresión para los datos originales.

$$\Sigma x^2 = 118\,958.58 - \frac{(1817.2)^2}{30} = 8\,884.72$$

$$\Sigma y^2 = 542\,675.26 - \frac{(3383.6)^2}{30} = 161\,050.29$$

$$\Sigma xy = 241\,772.67 - \frac{1817.2(3383.6)}{30} = 36\,816.74$$

$$r^2 = \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2} = \frac{(36\,816.74)^2}{8\,884.72(161\,050.29)} = 0.9473$$

$$r = \sqrt{0.9473} = .973$$

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{36\,816.74}{8\,884.72} = 4.144$$

$$a = \overline{Y} - b\overline{X} = \frac{3383.6}{30} - \left[4.144\left(\frac{1817.2}{30}\right)\right] = -138.20$$

$$\hat{Y} = 4.144\,X - 138.20$$

A primera vista parece como si la línea recta nos diera un excelente ajuste a los datos. El coeficiente de correlación, 0.973, es muy alto; sin embargo, si nos fijamos en el gráfico de los datos con la recta de regresión superpuesta (figura 14.2), notamos algo perturbador. Todas las desviaciones de la recta en los extremos del rango son positivas, mientras que aquellas en el medio del rango son negativas. Si las desviaciones fueran más o menos aleatorias, estaríamos satisfechos; pero esta agrupación sistemática de las desviaciones nos conduce a pensar que una curva describiría aún mejor las observaciones. Existe otra razón más apremiante para tratar de ajustar una curva. La recta que hemos ajustado a los datos, sencillamente no tiene sentido para diámetros inferiores a casi 34 mm, pues indica que los bulbos menores que lo señalado tendrían pesos negativos.

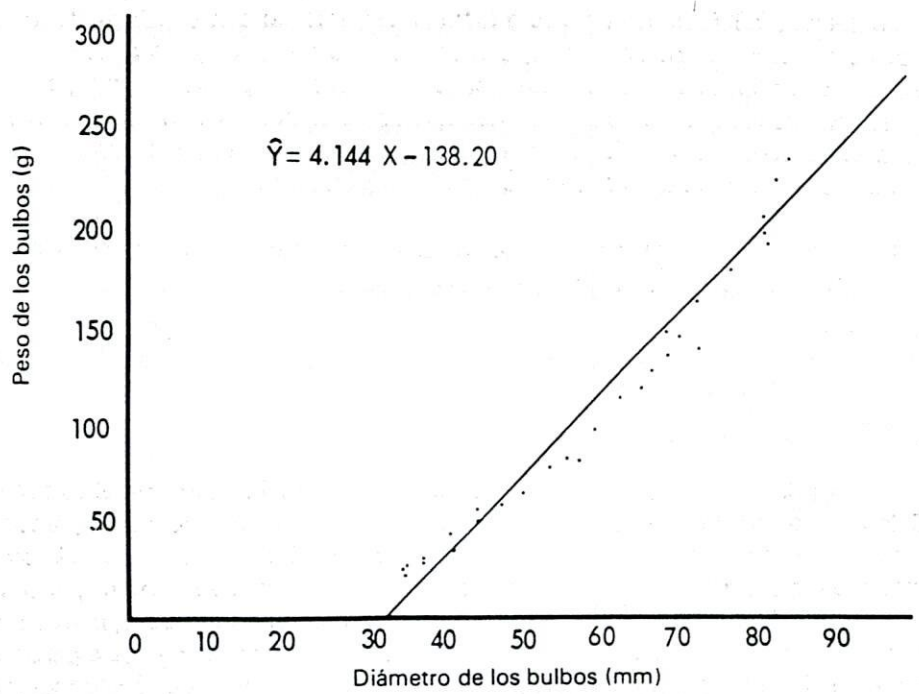


Figura 14.2. Datos de la cebolla a partir de un rango más amplio de observaciones que el de la figura 13.4, mostrando desviaciones no aleatorias de la recta de regresión.

Ahora ajustaremos una línea recta a los logaritmos de X y Y, y veremos si estas dificultades son superables. Los cálculos son exactamente los mismos, excepto que reemplazamos X por $X' = \log X$, y Y por $Y' = \log Y$.

$$\Sigma x'^2 = 93.80268806 - \frac{(52.90339)^2}{30} = .51039894$$

$$\Sigma y'^2 = 116.4541216 - \frac{(58.27655)^2}{30} = 3.2489123$$

$$\Sigma x'y' = 104.0495715 - \frac{52.90339(58.27655)}{30} = 1.2820031$$

$$r^2 = \frac{1.2820031^2}{.51039894(3.2489123)} = .991129$$

$$r = \sqrt{.9911} = .996$$

$$b = \frac{1.2820031}{.51039894} = 2.5118$$

$$a' = \frac{58.27655}{30} - 2.5118\left(\frac{52.90339}{30}\right) = -2.4869$$

$$\hat{Y}' = 2.5118X' - 2.4869$$

El coeficiente de correlación, 0.996, indica un ajuste extremadamente estrecho, incluso mayor que el obtenido a partir de los datos no transformados; sin embargo, el perfeccionamiento de la correlación no es la razón principal para preferir el uso de los datos transformados en este caso. Se puede observar, a partir de la figura 14.3, que las desviaciones de los puntos de la recta de regresión son más o menos aleatoriamente distribuidos en cuanto a su dirección; además, la relación entre X y Y, expresada en una nueva ecuación, implica que cuando el diámetro se aproxima a cero, el peso también se aproxima a cero.

La ecuación de regresión en forma logarítmica puede volverse a transformar para las mediciones originales, tomando el antilogarítmo de *a'* para encontrar *a* y sustituyendo:

ecuación: $Y = aX^b$

forma logarítmica: $Y' = 2.5118X' - 2.4869$

forma original: $Y = .00326(X^{2.5118})$

El exponente de alrededor de 2.5 es interesante, por cuanto revela aproximadamente el patrón de crecimiento de las cebollas. Si los bulbos crecieran en la misma proporción en todas las dimensiones, la forma permanecería constante y el peso debería ser una función del cubo del diámetro de X^3. Si la profundidad permaneciera constante y el crecimiento contemplado sólo se incrementara en diámetro, el peso debería ser una función del cuadrado del diámetro de X^2. Si los bulbos aumentan en profundidad, pero en una proporción más lenta que su incremento en diámetro, la forma debe cambiar de prolato a esférica a achatada, y el peso debe ser una función de alguna potencia del diámetro entre 2 y 3. La última situación concuerda exactamente con las observaciones. La ecuación que hemos desarrollado no sólo se ajusta estrechamente a los datos, sino que también expresa una relación natural entre el diámetro y el peso, que concuerda con otros hechos referentes a la geometría del crecimiento.

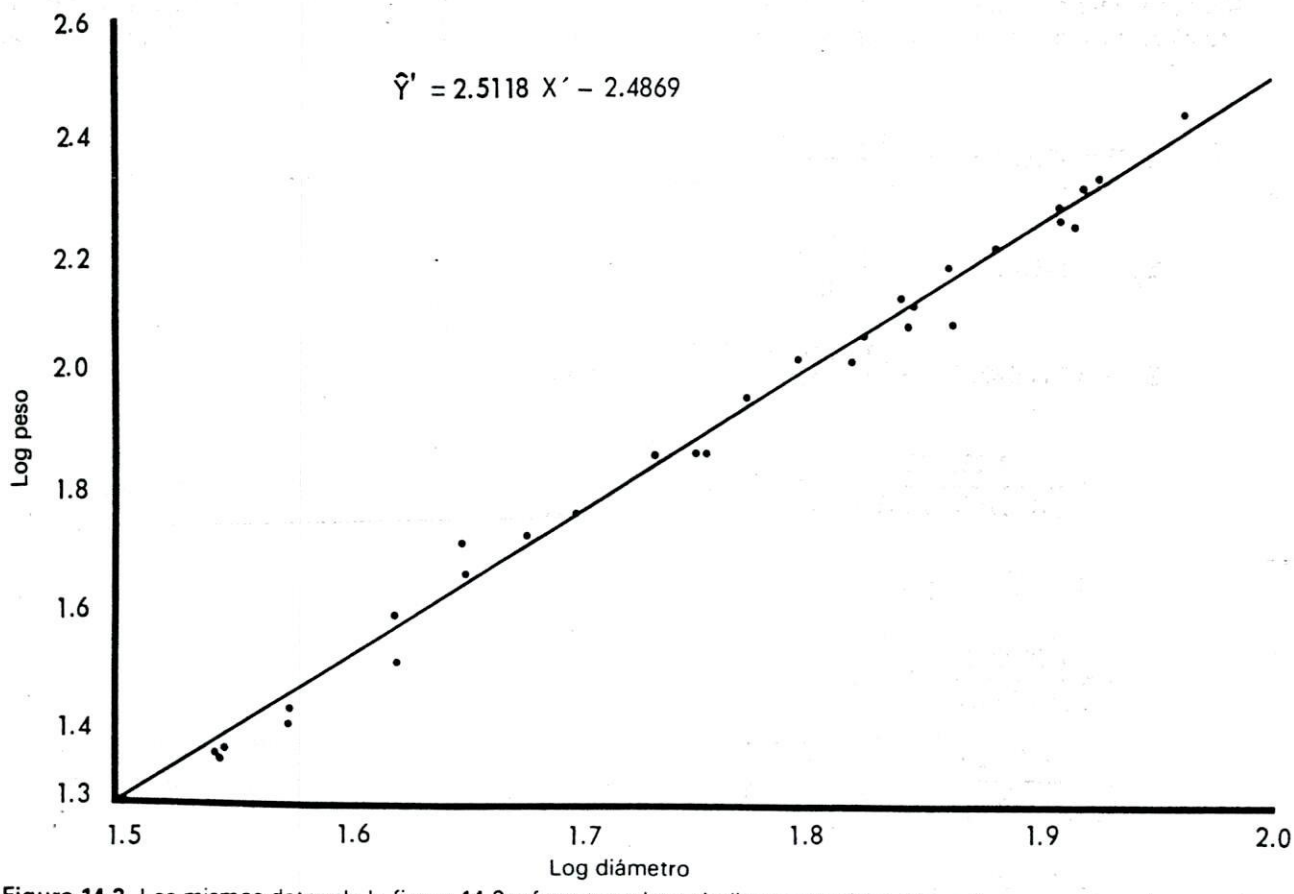


Figura 14.3. Los mismos datos de la figura 14.2 referentes a las cebollas convertidos a logaritmos, mostrando un mejor ajuste a una línea recta.

La línea recta de tipo semilogarítmica

La ecuación general para este tipo de curva es:

$$Y = ab^X$$

Si aplicamos logaritmos a ambos miembros de la ecuación, obtenemos:

$$\log Y = \log a + (\log b)X$$

Haciendo $Y = Y'$, $\log a = a'$, y $\log b = b'$, entonces:

$$Y' = a' + b'X$$

Nuevamente, la transformación ha producido una línea recta; pero en este tipo de curva son el logaritmo de X y los valores originales de X los que están siendo utilizados, en vez de los logaritmos de ambas variables. Por ese motivo, ésta recibe el nombre de tipo **semilogarítmica.** Existe papel cuadriculado semilogarítmico con una escala logarítmica sobre el eje de las Y y una escala ordinaria sobre el eje de las X. Los datos pueden anotarse en un papel semilogarítmico, o los valores Y pueden llevarse a logaritmos y trazarse sobre un papel cuadriculado ordinario. En cualquiera de los casos, si el diagrama de dispersión resultante se parece al de datos lineales, vale la pena calcular los coeficientes de correlación lineal y la regresión del logaritmo de Y sobre X.

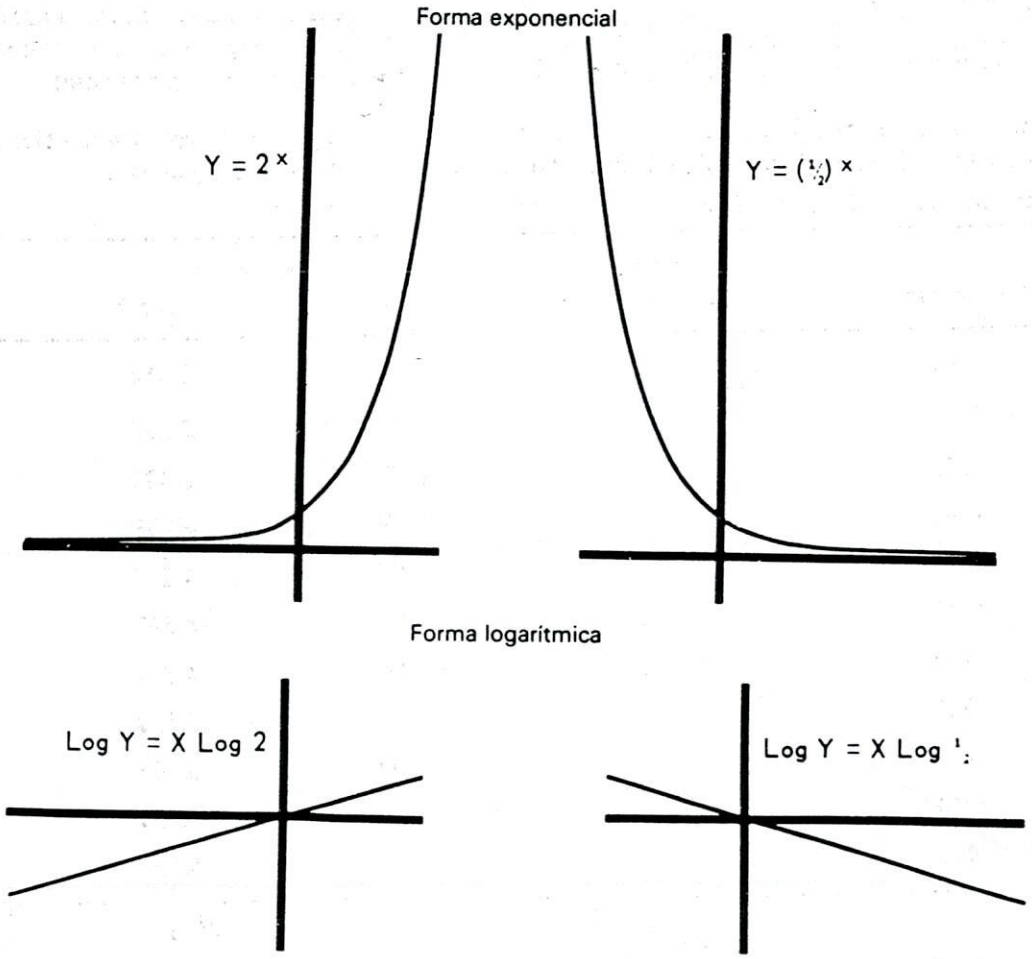


Figura 14.4 Curvas exponenciales típicas, con su transformación logarítmica.

Una curva de este tipo suele denominarse **curva exponencial,** puesto que la variable X aparece como un exponente en la ecuación. Los valores de X pueden ser positivos o negativos, fracciones o números enteros; pero b sólo puede ser un número positivo. La figura 14.4 muestra dos curvas exponenciales típicas, una con $b = 2$, y la otra con $b = \frac{1}{2}$. La figura revela también la recta resultante de la transformación de Y a log Y.

Los datos que más probablemente se ajustan a este tipo de curva son aquellos relacionados con **proporciones de interés.** La fórmula para el cambio en el capital con el tiempo, invertido a una proporción constante de interés compuesto anualmente, es:

$$A = P(1 + r)^t$$

donde A es la cantidad al final del tiempo t, P es el capital original, r es la proporción anual de intereses, y t es el tiempo en años.

¿Dónde encontramos algo semejante en la agricultura? Muchos organismos tienen un crecimiento bastante constante al menos durante las primeras etapas de crecimiento, siguiendo, por tanto, la ley de composición del interés. Si estamos estudiando la relación entre tiempo y tamaño de un organismo o de una población, a menudo resulta útil ver si los datos se ajustan a este tipo de curva.

Otra situación en la que este tipo de curva será útil es en el tratamiento con leyes físicas que tienen la característica de ser **exponenciales.** Consideremos, por ejemplo, la ley de Van Hoff, que establece que la proporción de reacción aproximadamente se duplica para cada elevación de 10°C. de temperatura. Se sabe que muchas respuestas de plantas cumplen bastante bien esta ley, al menos a través de un rango limitado de temperaturas. Luego, la temperatura y la proporción de pudrición de frutas y vegetales, frecuentemente puede estudiarse con facilidad, suponiendo que ésta se encuentra relacionada **en forma exponencial.**

Como un ejemplo de datos que pueden analizarse transformando Y en log Y, tomaremos la relación de población (Y) y tiempo (X) para la ciudad de San Diego, California, a través de once censos.

Tabla 14.2. Población de San Diego, California, desde 1860 hasta 1960.

Año del censo	Décadas Desde 1860 (X)	Población (Y)	Log Y
1860	0	731	2.864
1870	1	2 300	3.362
1880	2	2 636	3.421
1890	3	16 159	4.208
1900	4	17 700	4.248
1910	5	39 578	4,597
1920	6	74 361	4.871
1930	7	147 995	5.170
1940	8	203 341	5.308
1950	9	334 387	5.524
1960	10	573 224	5.758
Totales	55		49.331
Sumas de cuadrados	385		230.393503
Suma de X log Y			277.981

Un gráfico de las poblaciones contra el tiempo (figura 14.5) muestra en seguida que es inútil calcular una ecuación de regresión lineal para estos datos. Este es un ejemplo clásico de un caso en el que el método abreviado daría resultados extremadamente engañosos. Puesto que el rango de las poblaciones es exactamente igual al rango de los años, el método abreviado daría un coeficiente de correlación de +1. Este no sería capaz de revelar que los datos son decididamente curvilíneos; sin embargo, cuando se grafica el logaritmo de la población contra el tiempo, observamos que una línea recta parece razonable para representar la relación.

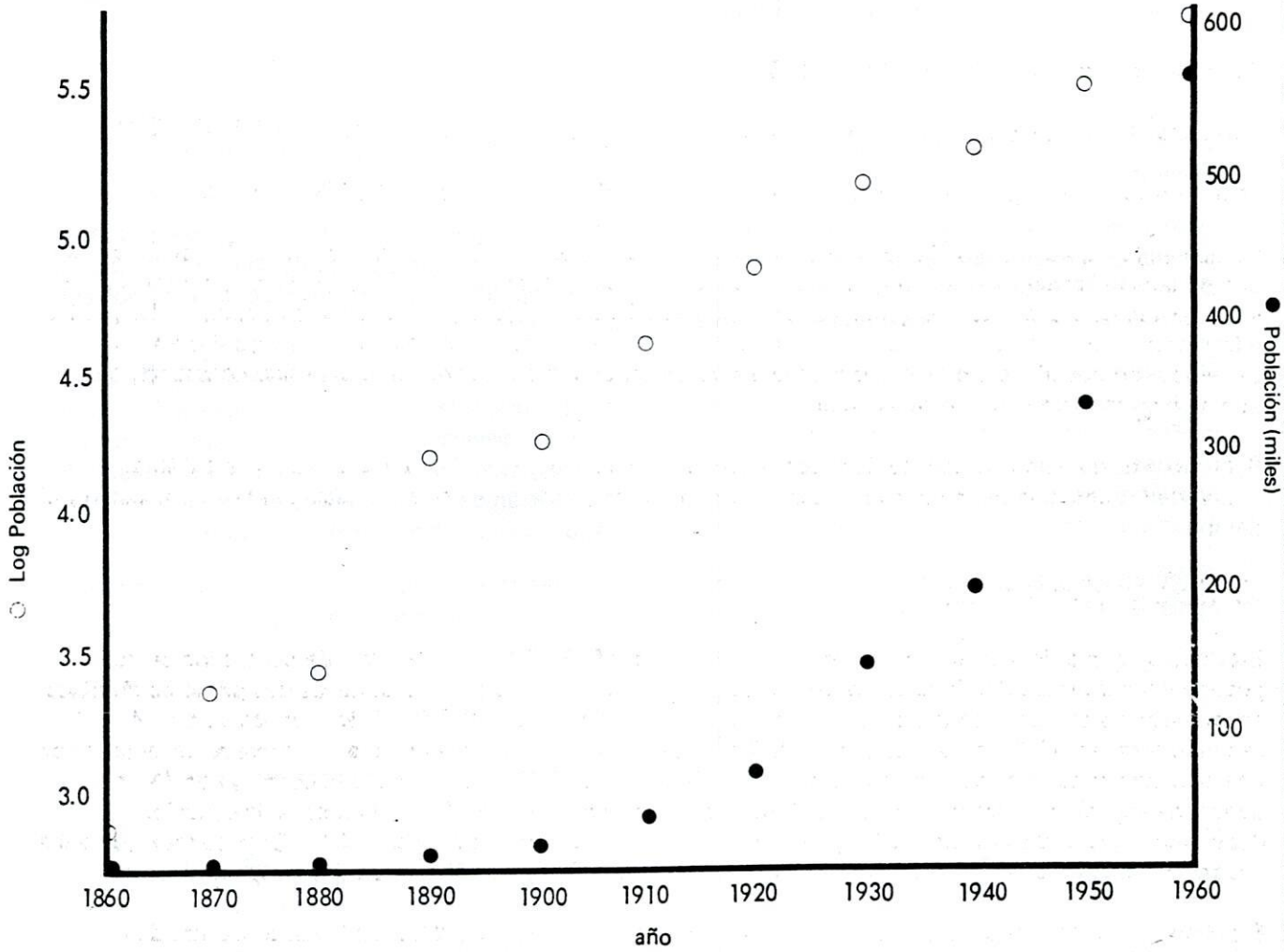


Figura 14.5. Poblaciones de San Diego para once décadas trazadas directamente sobre una escala logarítmica.

Los cálculos son directos si sustituimos Y por $Y' = \log Y$ como una de las variables:

$$\Sigma x^2 = 385 - \frac{(55)^2}{11} = 110$$

$$\Sigma y'^2 = 230.393503 - \frac{(49.331)^2}{11} = 9.161907$$

$$\Sigma xy' = 277.981 - \frac{55(49.331)}{11} = 31.326$$

$$r^2 = \frac{(\Sigma xy')^2}{\Sigma x^2 \Sigma y'^2} = (31.326)^2/110(9.161907) = .9737$$

$$r = \sqrt{.9737} = .987$$

$$b = \frac{\Sigma xy'}{\Sigma x^2} = \frac{31.326}{110} = .2848$$

$$a = \bar{Y}' - b\bar{X} = \frac{49.331}{11} - .2848\left(\frac{55}{11}\right) = 3.0606$$

Ecuación de regresión: $\hat{Y}' = 3.0606 + .2848X$

Tomando el antilogaritmo de ambos miembros obtenemos la ecuación exponencial: $\hat{Y} = 1\ 150(1.927)^{X}$

Esta ecuación revela que, en promedio, la población registró un incremento de 92.7% cada 10 años.

No hay lugar a dudas de que la curva exponencial se ajusta mucho mejor a los datos que cualquier línea recta que pudiera utilizarse; sin embargo, incluso cuando empleamos los logaritmos de la población contra el tiempo y los ajustamos a una recta, el ajuste no es ideal y se registra una ligera, pero definida tendencia de los puntos a formar una curva. Las desviaciones en el medio de la recta son positivas, mientras que aquellas en los extremos son negativas. Por el gráfico, parece que la proporción de crecimiento no ha sido constante, pero que ha mostrado una tendencia a decaer.

Si la curva fuese extrapolada a 1970, la población estimada sería de 1 561 000 habitantes. Más adelante mostraremos cómo se puede construir una ecuación aún mejor para expresar la relación entre la población y el tiempo.

El tipo polinómico

Este tipo de curva tiene la ecuación general $Y = a + bX + cX^2 + dX^3 + \ldots$. La hilera de puntos indica que podemos tener tantos términos como deseemos. Si la ecuación presenta solamente los dos primeros términos en su miembro derecho, podemos reconocerla como la ecuación de una recta. Si ésta concluye con el segundo término (cX^2), se trata de una ecuación de **segundo grado** o **cuadrática**. La curva representada por una ecuación cuadrática tiene en especial el nombre de **parábola**. Una ecuación que concluye en dX^3 se denomina ecuación de **tercer grado** o **cúbica**. La potencia más alta de X que aparece en la ecuación determina el grado de la misma, y los grados más comunes reciben nombres especiales. Correspondiendo a los cuatro primeros grados, encontramos los términos lineal, cuadrática, cúbica y cuártica, respectivamente.

El polinomio es, con mucho, la expresión más ampliamente utilizada para describir la relación entre dos variables. Algunas veces puede no ser una expresión particularmente "natural", es decir, aquella que expresa una relación de causa y efecto entre las variables; sin embargo, es tan flexible y tan fácilmente manejable en forma matemática, que resulta de gran utilidad.

La figura 14.6 muestra algunas de las muchas formas de curvas que pueden representarse mediante una ecuación polinómica. Una propiedad de este tipo de ecuación es que, sin importar cuántos pares de observaciones tengamos, resulta posible calcular una curva polinómica que se ajustará exactamente a cada punto, siempre que sólo exista un valor de Y para cada valor de X. El grado del polinomio, requerido para que esto se realice, es, a lo sumo, uno menos que el número de pares de observaciones. En la práctica, rara vez calculamos ecuaciones que sean mayores de tercer o cuarto grado. Los cálculos más allá de esto son formidables y los resultados suelen ser curvas serpenteantes, sin sentido.

Notamos que una línea recta fue simplemente un caso especial de la ecuación general polinómica; un polinomio de primer grado o lineal. A fin de encontrar una expresión para la relación curvilínea de dos variables, tratamos de hacer lo mismo que cuando la ajustábamos a una línea recta; es decir, buscábamos la curva de cierto grado que hiciera mínima la suma de cuadrados de las desviaciones. A propósito, las

transformaciones logarítmicas que se utilizaron para simplificar los dos primeros tipos de curva, no se emplean en el ajuste al polinomio general. Deben usarse otros métodos.

El problema consiste en encontrar los coeficientes a, b, c, d, etc., que darán un polinomio que cumplirá el requisito de que la suma de cuadrados de las desviaciones sea mínima. Para esto, hacemos uso de lo que conocemos como **ecuaciones normales.** Necesitamos tantas ecuaciones como coeficientes haya, o una más que el grado de la ecuación que deseamos ajustar.

Las ecuaciones normales son:

$$an + b\Sigma X + c\Sigma X^2 + d\Sigma X^3 + \ldots = \Sigma Y$$

$$a\Sigma X + b\Sigma X^2 + c\Sigma X^3 + d\Sigma X^4 + \ldots = \Sigma XY$$

$$a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 + d\Sigma X^5 + \ldots = \Sigma X^2Y$$

$$a\Sigma X^3 + b\Sigma X^4 + c\Sigma X^5 + d\Sigma X^6 + \ldots = \Sigma X^3Y$$

$$\ldots$$

Los puntos significan que continuamos con el mismo patrón hasta que tengamos tantos términos a la izquierda del signo de igualdad y tantas ecuaciones, como coeficientes haya. Entonces, para una línea recta sólo necesitamos los primeros dos términos de las primeras dos ecuaciones. Para una curva cuadrática o de segundo grado, necesitamos los tres primeros términos de las tres primeras ecuaciones, y así sucesivamente.

A partir de los datos, necesitamos calcular las sumas de potencias de X y las sumas de productos requeridos en la ecuación. Para una ecuación de enésima potencia, necesitamos calcular las sumas de todas las potencias de X hasta X^{2n}, y las sumas de productos hasta X^nY. La matemática es sencilla, pero la aritmética resulta abrumadora si tratamos de ajustar polinomios de grados elevados.

A modo de ejemplo, utilizaremos algunos datos de la producción de frijol de media luna verde en diferentes edades del campo, en tiempo de recolección. La fecha de la primera recolección se emplea como la fecha base, y se le otorga un valor de X igual a cero. Los valores de X para las recolecciones subsecuentes son el número de días transcurridos desde la fecha base. La producción en libras es la variable dependiente, designada por Y. Se espera que los datos sean curvilíneos, puesto que se debió registrar un incremento de la producción con la edad del campo; pero cuando los frijoles aumentan en madurez, pasan de verde a pálido y a blanco. Por tanto, la producción de frijoles verdes disminuirá después de alcanzar un máximo.

Tabla 14.3. Producción en libras de frijol de media luna verde (Y) en seis fechas (X)

X	Y	X^2	X^3	X^4	X^5	X^6	XY	X^2Y	X^3Y
0	27.4	0	0	0	0	0	0	0	0
4	39.3	16	64	256	1 024	4 096	157.2	628.8	2 515.2
7	46.2	49	343	2 401	16 807	117 649	323.4	2 263.8	15 846.6
10	47.8	100	1 000	10 000	100 000	1 000 000	478.0	4 780.0	47 800.0
13	44.5	169	2 197	28 561	371 293	4 826 809	578.5	7 520.5	97 766.5
18	24.5	324	5 832	104 976	1 889 568	34 012 224	441.0	7 938.0	142 884.0
Totales 52	229.7	658	9 436	146 194	2 378 692	39 960 778	1 978.1	23 131.1	306 812.3

Ahora tenemos todas las sumas que necesitamos para las ecuaciones normales, hasta el tercer grado. En primer lugar, ajustaremos una línea recta a los datos, utilizando las ecuaciones normales:

$$an + b\Sigma X = \Sigma Y$$

$$a\Sigma X + b\Sigma X^2 = \Sigma XY$$

Sustituyendo en los valores conocidos de estas ecuaciones, tenemos:

$$6a + 52b = 229.7 \quad (1)$$

$$52a + 658b = 1\,978.1 \quad (2)$$

Multiplicando la ecuación (1) por 52 y la ecuación (2) por 6, obtenemos:

$$312a + 2\,704b = 11\,944.4 \quad (3)$$

$$312a + 3\,948b = 11\,868.6 \quad (4)$$

y sustrayendo (3) de (4),

$$1\,244b = -75.8$$

$$b = -75.8/1\,244 = -.06093$$

Sustituyendo este valor de b en la ecuación (1), obtenemos:

$$6a = 229.7 + 52(.0609) = 232.868$$

$$a = 38.8114$$

La ecuación de regresión es, por tanto,

$$\hat{Y} = 38.81 - .0609X$$

Podríamos haber llegado a la misma ecuación, utilizando las fórmulas estándar:

$$b = \frac{\Sigma xy}{\Sigma x^2} \quad \text{y} \quad a = \overline{Y} - b\overline{X}$$

La idea de seguir el procedimiento de la ecuación normal tuvo como objetivo obtener cierta destreza en el proceso que seguiremos para curvas de grados más elevados.

Podemos ver, por el gráfico de esta recta (figura 14.7), que la misma se ajusta deficientemente. Para el coeficiente de correlación, necesitamos ΣY^2, que es 9 295.03. Entonces,

$$r^2 = \frac{(\Sigma xy)^2}{\Sigma x^2 \Sigma y^2} = \left[1978.1 - \left(\frac{52(229.7)}{6}\right)\right]^2 \Big/ \left(658 - \frac{(52)^2}{6}\right)\left(9295.03 - \frac{(229.7)^2}{6}\right)$$

$$= (-12.6)^2/(207.33)(501.35) = .00153$$

$$r = \sqrt{.00153} = -.039$$

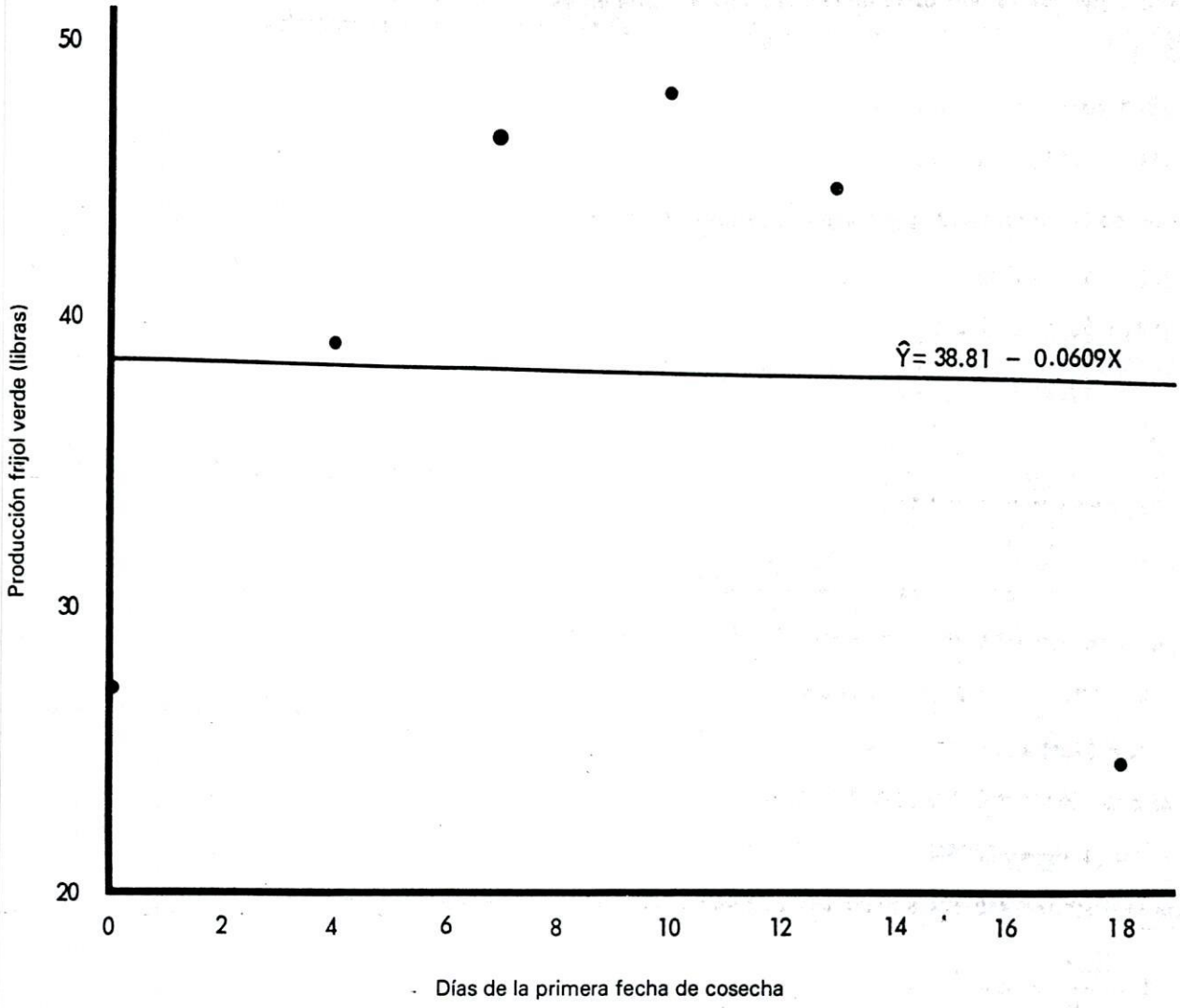


Figura 14.7. Gráfico de los datos del frijol de media luna, mostrando todas las deficiencias de la regresión lineal para expresar la relación entre la producción y la edad del cultivo.

El coeficiente se aproxima a cero y, obviamente, no es significativo. Tenemos un buen ejemplo de una de las trampas descritas en el capítulo 13: "un bajo coeficiente de correlación no necesariamente significa ausencia de relación". Aunque el coeficiente del presente ejemplo es casi igual a cero, sería ridículo concluir que no hubo relación entre la producción de frijol de media luna verde y la edad del cultivo en la recolección.

Ahora ajustaremos los datos a una curva de segundo grado o cuadrática. Necesitamos tres ecuaciones normales:

$$an + b\Sigma X + c\Sigma X^2 = \Sigma Y$$

$$a\Sigma X + b\Sigma X^2 + c\Sigma X^3 = \Sigma XY$$

$$a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 = \Sigma X^2Y$$

Sustituyendo en los valores de la tabla observados, obtenemos:

$$6a + 52b + 658c = 229.7 \quad (1)$$

$$52a + 658b + 9436c = 1\,978.1 \quad (2)$$

$$658a + 9436b + 146\,194c = 23\,131.1 \quad (3)$$

Multiplicando (1) por 52 y (2) por 6 y sustrayendo:

$$312a + 2704b + 34\,216c = 11\,944.4$$

$$\underline{312a + 3948b + 56\,616c = 11\,868.6}$$

$$1244b + 22\,400c = -75.8 \quad (4)$$

Ahora, multiplicando (1) por 658 y (3) por 6, y sustrayendo:

$$3948a + 34\,216b + 432\,964c = 151\,142.6$$

$$\underline{3948a + 56\,616b + 877\,164c = 138\,786.6}$$

$$22\,400b + 444\,200c = -12\,356.0 \quad (5)$$

Los dos pasos anteriores eliminan a a y dan dos ecuaciones de dos incógnitas. Ahora, multiplicando (4) por 22 400; (5) por 1 244, y sustrayendo:

$$27\,865\,600b + 501\,760\,000c = -1\,697\,920$$

$$\underline{27\,865\,600b + 552\,584\,800c = -15\,370\,864}$$

$$50\,824\,800c = -13\,672\,944$$

$$c = -0.2690$$

Sustituyendo c en (4): $1244b - 6025.6 = -75.8$

$$1244b = 5949.8$$

$$b = 4.7828$$

Sustituyendo b y c en (1): $6a + 248.7056 - 177.0020 = 229.7$

$$6a = 157.9964$$

$$a = 26.3327$$

Podemos escribir ahora la ecuación de segundo grado:

$$\hat{Y} = 26.3327 + 4.7828X - 0.2690X^2$$

Veamos hasta qué punto esto significa un mejoramiento sobre la ecuación lineal. Denominamos $\hat{Y}_L$ a la estimación lineal y $\hat{Y}_c$ a la estimación cuadrática. La siguiente tabla muestra estas dos estimaciones, comparadas con los valores originales.

Tabla 14.4. Producciones observadas y calculadas del frijol de media luna.

X	Y	$\hat{Y}_L$	$d_L = Y - \hat{Y}_L$	d_L^2	$\hat{Y}_Q$	$d_Q = Y - \hat{Y}_Q$	d_Q^2
0	27.4	38.81	−11.41	130.19	26.33	1.07	1.14
4	39.3	38.57	0.73	0.53	41.16	−1.86	3.46
7	46.2	38.38	7.82	61.15	46.63	− 0.43	0.18
10	47.8	38.20	9.60	92.16	47.26	0.54	0.29
13	44.5	38.02	6.48	41.99	43.05	1.45	2.10
18	24.5	37.71	−13.21	174.50	25.27	− 0.77	0.59
Totales			0.01	500.52		0.00	7.76

Los resultados pueden resumirse en un cuadro de análisis de varianza como sigue:

Fuente de variación	SC	gl
Total	501.35	5
Lineal	0.83	1
Desviaciones de la lineal	500.52	4
Componente cuadrático	492.76	1
Desviaciones de la cuadrática	7.76	3

Luego, observamos que el ajuste a una línea recta sólo explicó un 0.2% de la variabilidad de Y (0.83/501.35), mientras que la curva cuadrática explicó (492.76 + 0.83)/501.35, o 98.5%. No habría mucho sentido en ajustar los datos a una curva de grado más elevado, puesto que sólo el 1.5% de la variabilidad de Y permanece inexplicable.

Sin embargo, para ilustrar el método, ajustaremos a una curva de tercer grado. Las ecuaciones normales son:

$$an + b\Sigma X + c\Sigma X^2 + d\Sigma X^3 = \Sigma Y$$
$$a\Sigma X + b\Sigma X^2 + c\Sigma X^3 + d\Sigma X^4 = \Sigma XY$$
$$a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 + d\Sigma X^5 = \Sigma X^2Y$$
$$a\Sigma X^3 + b\Sigma X^4 + c\Sigma X^5 + d\Sigma X^6 = \Sigma X^3Y$$

Sustituyendo los valores observados, obtenemos las siguientes ecuaciones que deseamos resolver para a, b, c y d:

$$6a + 52b + 658c + 9\,436d = 229.7 \quad (1)$$
$$52a + 658b + 9\,436c + 146\,194d = 1\,978.1 \quad (2)$$
$$658a + 9\,436b + 146\,194c + 2\,378\,692d = 23\,131.1 \quad (3)$$
$$9\,436a + 146\,194b + 2\,378\,692c + 39\,960\,778d = 306\,812.3 \quad (4)$$

Eliminemos primero a a, como sigue:

La ecuación (2) multiplicada por 6, menos la ecuación (1) multiplicada por 52, da:

$$1\,244b + 22\,400c + 386\,492d = -75.8 \quad (5)$$

La ecuación (3) multiplicada por 6, menos la ecuación (1) multiplicada por 658, da:

$$22\,400b + 444\,200c + 8\,063\,264d = -12\,356.0 \quad (6)$$

La ecuación (4) multiplicada por 6, menos la ecuación (1) multiplicada por 9 436, da:

$$386\,492b + 8\,063\,264c + 150\,726\,572d = -326\,575.4 \quad (7)$$

Ahora eliminamos b mediante los siguientes pasos:

La ecuación (6) multiplicada por 1 244, menos la ecuación (5) multiplicada por 22 400, da:

$$50\,824\,800c + 1\,373\,279\,616d = -13\,672\,944 \quad (8)$$

La ecuación (7) multiplicada por 1 244, menos la ecuación (5) multiplicada por 386 492, da:

$$1\,373\,279\,616c + 38\,127\,789\,500d = -276\,963\,704 \quad (9)$$

Para eliminar c, tomamos la ecuación (8) multiplicada por 1 373 279 616, menos la ecuación (9) multiplicada por 50 824 800 y dividimos ambos de sus miembros entre 10 000 000 y redondeamos para reducir los grandes números de 10 dígitos.

Esto da:

$$5\,194\,037\,206d = -38\,232\,948$$

$$d = -.00736$$

Sustituyendo d en la ecuación (8) y resolviendo para c, obtenemos:

$$c = -.07015$$

Sustituyendo d y c en la ecuación (5), obtenemos:

$$b = 3.48886$$

Finalmente, sustituyendo d, c y b en la ecuación (1), obtenemos:

$$a = 27.31449$$

Y la ecuación de tercer grado, o cúbica, es:

$$\hat{Y}_c = 27.31449 + 3.48886X - .07015X^2 - .00736X^3$$

Calculando los valores estimados $\hat{Y}_c$, encontramos un mejoramiento sustancial sobre el ajuste de la curva cuadrática.

X	Y	$\hat{Y}_c$	$d = Y - \hat{Y}_c$	d^2
0	27 4	27 31	0.09	0.01
4	39 3	39 68	− 0.38	0.14
7	46 2	45 78	0.42	0.18
10	47.8	47.83	− 0.03	0.00
13	44.5	44.64	− 0.14	0.02
18	24.5	24.46	0.04	0.00
Totales			0.00	0.35

La suma de cuadrados para la desviación de la curva cuadrática puede ser ahora separada como sigue:

Fuente de variación	SC	gl	CM	F	F requerido 5%	F requerido 1%
Desviación de la cuadrática	7.76	3				
Componente cúbico	7.41	1	7.41	42.3*	18.51	98.49
Desviación de la cúbica	0.35	2	0.175			

*Significativo en el nivel del 5%.

El ajuste mejorado, obtenido mediante el cálculo de una ecuación cúbica, a pesar de ser apreciable, fue significativo sólo en el punto del 5%. Con tan pocos grados de libertad, esto no es sorprendente, puesto que se requiere un valor F de 98.49 para la significación en el nivel del 1%.

La figura 14.8 muestra las curvas cuadráticas y cúbicas trazadas sobre un rango más amplio que las observaciones, para destacar su diferencia en cuanto a la forma. Durante todo el rango de observaciones, las dos curvas no difieren notablemente, pero el ajuste superior de la curva cúbica es evidente.

El lector probablemente notó cómo los cálculos se volvieron cada vez más difíciles cuando pasamos de las curvas lineales a las cuadráticas y cúbicas. Se han diseñado diversos métodos para sistematizar dichos cálculos, siendo los más comunes el de Doolittle y los métodos abreviados de Doolittle. Un estudio de estos métodos se encuentra más allá del alcance de este libro, pero puede hallarse en casi todos los textos recientes de estadística. También se encuentran disponibles los programas destinados a calcular los coeficientes para casi todos los grados deseados en una computadora electrónica.

En los casos donde los valores de X son igualmente espaciados, existen métodos abreviados muy sencillos, mismos que se estudiarán en el siguiente capítulo.

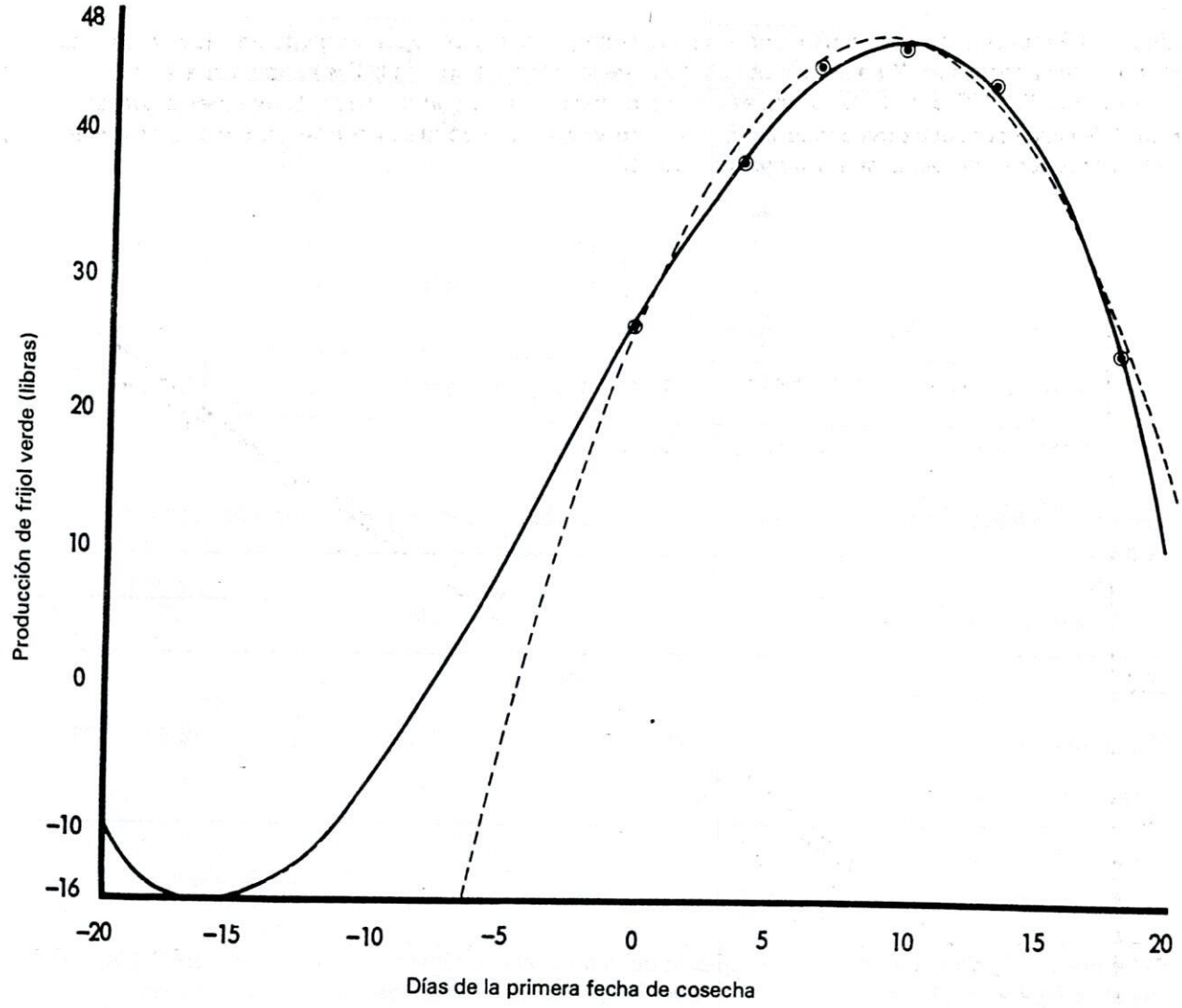


Figura 14.8. Los mismos datos de la figura 14.7 sobre el frijol de media luna, mostrando el ajuste adecuado de una curva cuadrática (línea de puntos) y el ajuste aún más estrecho de la curva cúbica (línea continua).

Hemos examinado tres tipos generales de curvas y mostrado cómo ajustarlas a los datos observados. Algunas veces resulta ventajoso utilizar una combinación de dos tipos de curvas; por ejemplo, en los datos sobre la población de San Diego, encontramos que trazar los **logaritmos** de la población contra años, dio una aproximación mucho más cercana a una línea recta que cuando graficamos simplemente población contra años.

Sin embargo, una ojeada a la figura 14.5 muestra que incluso los datos transformados no forman una línea completamente recta, sino que tiene más bien una tendencia definida a la curva. La proporción de incremento parece estar decayendo con el tiempo.

Nuevamente podemos ajustar con facilidad una curva de segundo grado a los datos, utilizando $Y' = \log Y$ como la variable dependiente, en vez de Y. Los cálculos se dejan a los lectores interesados, como un buen ejercicio del ajuste a una curva de segundo grado. La ecuación obtenida es:

$$Y' = 2.87906 + .40590X - .01211X^2$$

La figura 14.9 muestra la comparación entre la línea recta y la curva de segundo grado en relación con los logaritmos de la población. Ya hemos puntualizado que la extrapolación de la línea recta arrojaría una predicción de 1 561 000 para 1970. La extrapolación de la curva de segundo grado da una predicción de 756 800*. En vista de la estrecha concordancia de la curva de segundo grado con las tendencias observadas, la predicción menor es probablemente la más razonable.

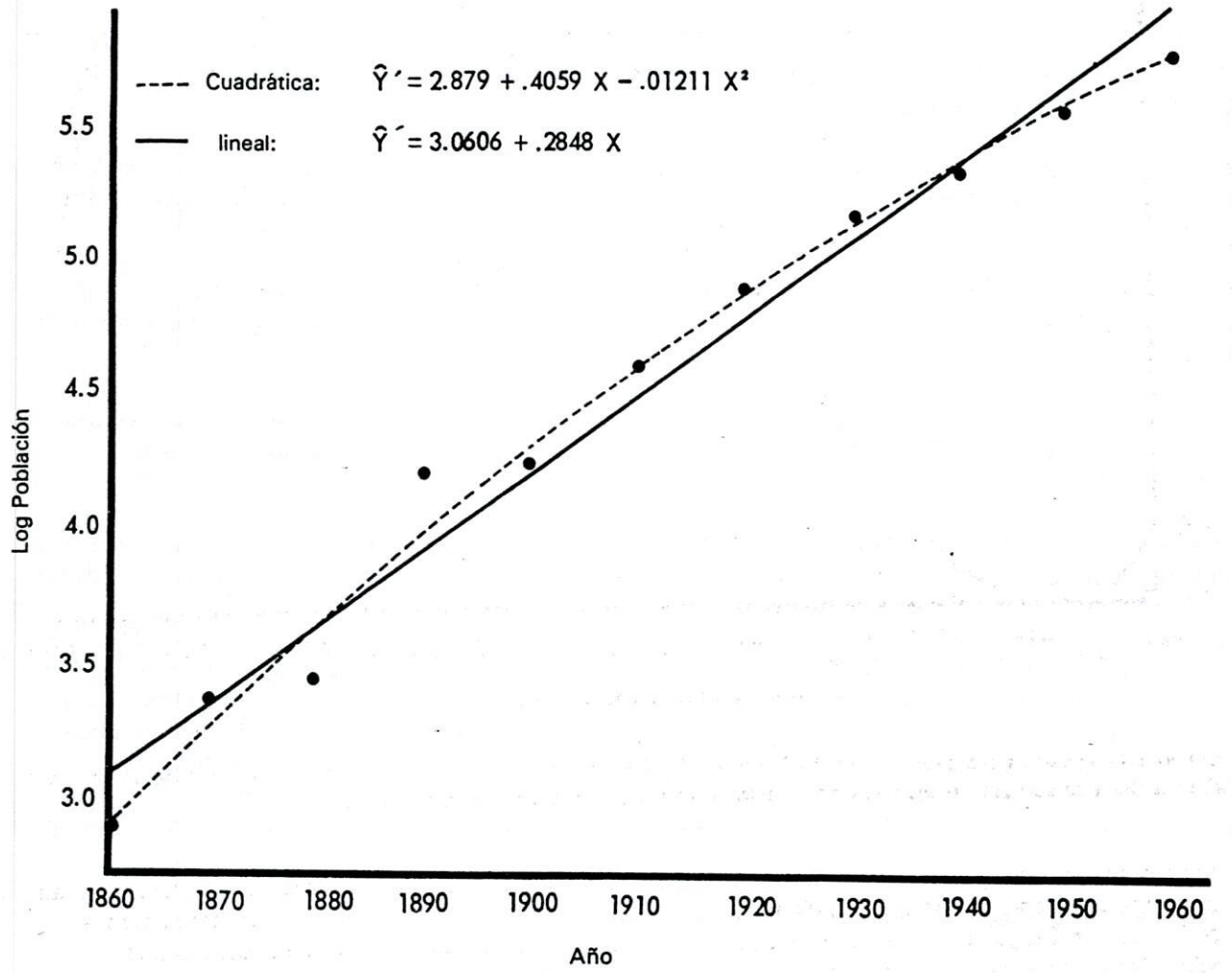


Figura 14.9. Datos de la población de San Diego con una línea recta ajustada a los logaritmos de la población (línea continua), y el mejoramiento obtenido con el ajuste de una ecuación cuadrática (línea de puntos).

* Las cifras del censo de 1970 se encuentran ahora disponibles y revelan que la pobalción de San Diego consta de 697 000 habitantes, la cual está a 8% por debajo de la cifra predicha.

El tipo periódico

Esta es una curva que relaciona alguna variable con el tiempo y que se repite a intervalos de tiempo fijos. Se conoce en los textos matemáticos como **curva de Fourier**, y resulta útil para cualquier tipo de datos que tienden a fluctuar hacia arriba y hacia abajo a intervalos regulares. Muy pocos textos estadísticos estudian el ajuste de datos de este tipo; no obstante, lo hemos creído tan ventajoso para diversas clases de datos agrícolas, que presentaremos un breve esbozo del método general. En el próximo capítulo examinaremos un método abreviado para la manipulación de casos especiales.

La ecuación general para una curva periódica es:

$Y = a_0 + a_1 \cos CX + b_1 \operatorname{sen} CX + a_2 \cos 2CX + b_2 \operatorname{sen} 2CX + a_3 \cos 3CX + b_3 \operatorname{sen} 3CX \ldots$, donde X es tiempo expresado como unidades de algún tiempo inicial arbitrario, y C es una constante igual a 360° dividida entre el número de unidades en un ciclo.

Supóngase, por ejemplo, que estamos estudiando las fluctuaciones por hora de alguna variable en ciclos de 24 horas y que tomamos la medianoche como punto de partida. Una observación hecha a las 9 A.M: tendría un valor X de 9 y C sería igual a 360°/24 o 15°; por tanto, el valor de CX sería 9 x 15° o 135°.

La línea de puntos a la derecha de la ecuación general significa que podemos seguir añadiendo pares de términos hasta que el número total de los mismos no exceda el número de periodos para los cuales tenemos observaciones.

Esta curva tiene muchas características similares a la curva polinómica. Presenta la misma notable propiedad de que si existe un solo valor de Y para cada valor de X, puede encontrarse una ecuación que pasará exactamente a través de cada punto.

Recordará el lector que un polinomio de primer grado es una línea recta con la ecuación general $Y = a + bX$. Esta recta es descrita completamente por dos valores: el intercepto a y la pendiente b. Una curva de Fourier de primer grado es una curva ondulatoria simple, con la ecuación general $Y = a_0 + a_1 \cos CX + b_1 \operatorname{sen} CX$. Para describir esta curva, necesitamos **tres** valores. El término a_0 da el valor central alrededor del cual la onda fluctúa. Esto puede verse como una media ponderada. Un segundo valor $A = \sqrt{a_1^2 + b_1^2}$, recibe el nombre de **semiamplitud** y revela hasta dónde la curva fluctúa hacia arriba y hacia abajo del punto central. El rango total desde el punto más alto hasta el punto más bajo de la onda es 2A y se denomina **amplitud**. El tercer valor que se necesita para describir la onda es el **ángulo de fase**. Este indica el punto en el ciclo donde la onda alcanza su valor máximo. Para encontrarlo, primero determinamos θ' (theta) = arc tan (b_1/a_1), que significa "el ángulo cuya tangente es b_1/a_1.". Luego encontramos el ángulo de fase mediante la aplicación de las siguientes reglas:

Si b_1 es positivo y a_1 es positiva	$\theta = \theta'$
Si b_1 es positivo y a_1 es negativa	$\theta = 180^\circ - \theta'$
Si b_1 es negativo y a_1 es negativa	$\theta = 180^\circ + \theta'$
Si b_1 es negativo y a_1 es positiva.	$\theta = 360^\circ - \theta'$

En el tipo polinómico, obtuvimos curvas más complicadas añadiendo términos con potencias sucesivas de X, como cX^2, dX^3, etc. Con la curva de Fourier, obtenemos formas de ondas más complicadas añadiendo pares de términos, como $a_2 \cos 2CX + b_2 \operatorname{sen} 2X$, $a_3 \cos 3CX + b_3 \operatorname{sen} 3CX$, etc. El efecto del par de segundo grado es sobreponer en la primera onda una segunda onda con dos oscilaciones completas por ciclo. El par de tercer grado sobrepone otra curva con tres oscilaciones completas por ciclo, y así sucesivamente.

El método para ajustar una curva de Fourier es similar al método para ajustar un polinomio. Utilizamos un

conjunto de ecuaciones normales en las que sustituimos sumas calculadas a partir de los datos observados, y las resolvemos para los coeficientes requeridos

Tabla 14.5. Temperaturas medias mensuales para nueve meses en Stockton, California

			(Ciclo = 12 meses C = 360°/12 = 30°)						
Y (Tiempo)	Mes	X	CX	U_i = cos (CX)	V_i = sen (CX)	U_i^2	V_i^2	U_iV_i	YU_i	YV_i
44.7	Ene.	0	0	1.000	0.000	1.00	0.00	0.000	44.700	0.000
49.0	Feb.	1	30°	0.866	0.500	0.75	0.25	0.433	42.434	24.500
53.7	Mar.	2	60°	0.500	0.866	0.25	0.75	0.433	26.850	46.504
59.7	Abr.	3	90°	0.000	1.000	0.00	1.00	0.000	0.000	59.700
76.2	Ags.	7	210°	−0.866	−0.500	0.75	0.25	0.433	−65.989	−38.100
72.7	Sep.	8	240°	−0.500	−0.866	0.25	0.75	0.433	−36.350	−62.958
64.0	Oct.	9	270°	0.000	−1.000	0.00	1.00	0.000	0.000	−64.000
53.0	Nov.	10	300°	0.500	−0.866	0.25	0.75	− 0.433	26.500	−45.898
45.9	Dic.	11	330°	0.866	−0.500	0.75	0.25	− 0.433	39.749	−22.950
Totales										
518.9				2.366	−1.366	4.00	5.00	0.866	77.894	−103.202

Para simplificar las ecuaciones normales, resulta conveniente adoptar dos símbolos, U y V:

$U_i = \cos i\,(CX)$

$V_i = \text{sen } i\,(CX)$ Entonces ΣU_2V_1 significa $\Sigma \cos 2(CX) \text{ sen } (CX)$.

Las ecuaciones normales son:

$$a_0 n + a_1\Sigma U_1 + b_1\Sigma V_1 + a_2\Sigma U_2 + b_2\Sigma V_2 + \ldots = \Sigma Y$$

$$a_0\Sigma U_1 + a_1\Sigma U_1^2 + b_1\Sigma U_1V_1 + a_2\Sigma U_1U_2 + b_2\Sigma U_1V_2 + \ldots = \Sigma U_1Y$$

$$a_0\Sigma V_1 + a_1\Sigma U_1V_1 + b_1\Sigma V_1^2 + a_2\Sigma U_2V_1 + b_2\Sigma V_1V_2 + \ldots = \Sigma V_1Y$$

$$a_0\Sigma U_2 + a_1\Sigma U_1U_2 + b_1\Sigma U_2V_1 + a_2\Sigma U_2^2 + b_2\Sigma U_2V_2 + \ldots = \Sigma U_2Y$$

$$a_0\Sigma V_2 + a_1\Sigma U_1V_2 + b_1\Sigma V_1V_2 + a_2\Sigma U_2V_2 + b_2\Sigma V_2^2 + \ldots = \Sigma V_2Y$$

Como en el caso del polinomio, necesitamos tantos términos en el miembro izquierdo de estas ecuaciones y tantas ecuaciones como coeficientes tengamos que calcular. Para un polinomio de grado enésimo necesitamos n + 1 ecuaciones, cada una con n + 1 términos en el miembro de la izquierda. Para las curvas de Fourier necesitamos 2n + 1 ecuaciones, cada una con 2n + 1 términos.

Para ilustrar el procedimiento, ajustaremos una curva de Fourier de primer grado a las temperaturas medias observadas durante nueve meses en Stockton, California. La tabla 14.5 muestra los datos observados y las columnas necesarias para sustituir en los términos de las ecuaciones normales.

Ahora podemos escribir las tres ecuaciones normales requeridas para encontrar a_0, a_1, y b_1.

$$9\ a_0 + 2.366\ a_1 - 1.366\ b_1 = 518.9 \quad (1)$$

$$2.366\ a_0 + 4\ a_1 + .866\ b_1 = 77.894 \quad (2)$$

$$-1.366\ a_0 + .866\ a_1 + 5\ b_1 = -103.202 \quad (3)$$

Multiplicando la ecuación (1) por 0.866 y la ecuación (2) por 1.366 y sumándolas, obtenemos:

$$11.026\ a_0 + 7.513\ a_1 = 555.771 \quad (4)$$

Multiplicando la ecuación (1) por 5 más la ecuación (3) por 1.366, obtenemos:

$$43.134\ a_0 + 13.013\ a_1 = 2453.526 \quad (5)$$

Multiplicando la ecuación (4) por 13.013 menos la ecuación (5) por 7.513, obtenemos:

$$-180.584\ a_0 = -11\ 201.093 \text{ and } a_0 = 62.027$$

Sustituyendo este valor de a_0 en la ecuación (4), obtenemos:

$$(11.026 \times 62.027) + 7.513\ a_1 = 555.771$$

$$7\ 513\ a_1 = -128.139$$

$$a_1 = -17.056$$

Sustituyendo a_0 y a_1 en la ecuación (3), obtenemos:

$$(-1.366 \times 62.027) + (.866 \times -17.057) + 5\ b_1 = -103.202$$

$$-84.729 - 14.770 + 5\ b_1 = -103.202$$

$$5\ b_1 = 84.729 + 14.770 - 103.202 = -3.703$$

$$b_1 = -.741$$

Podemos ahora plantear nuestra ecuación como sigue:

$$Y = 62.027 - 17.056 \cos (CX) - .741 \text{ sen } (CX)$$

Sustituyendo los valores de cos (CX) y sen (CX) para cada mes, obtenemos los valores predichos, que podemos comparar con los valores observados.

Las cifras entre paréntesis de la tabla 14.6 representan los datos para los meses que supusimos que no estuvieron disponibles cuando calculamos la curva y, por tanto, no fueron integrados a los cálculos. Se notará que la curva que calculamos a partir de los datos disponibles sobreestimó las medias reales para los meses perdidos.

El ajuste de la curva a los datos observados es muy estrecho. La suma total de cuadrados de las temperaturas observadas es 1 032.942, y podemos separarla en un análisis de varianza como el siguiente:

Fuente de variación	gl	SC	CM	F
Total	8	1032.942		
Debido a la regresión	2	1016.187	508.094	181.85***
Desviación de la regresión	6	16.755	2.794	

Tabla 14.6. Temperaturas observadas y predichas para un periodo de nueve meses en Stockton, California

Mes	Y (Observada)	$\hat{Y}$ (predicha)	$(Y - \hat{Y})$	$(Y - \hat{Y})^2$
Enero	44.7	44.97	−0.27	0.0729
Febrero	49.0	46.89	2.11	4.4521
Marzo	53.7	52.86	0.84	0.7056
Abril	59.7	61.29	−1.59	2.5281
(Mayo)	(66.2)	(69.91)	(−3.71)	
(Junio)	(72.8)	(76.43)	(−3.63)	
(Julio)	(78.2)	(79.08)	(−0.88)	
Agosto	76.2	77.17	−0.97	0.9409
Septiembre	72.7	71.20	1.50	2.2500
Octubre	64.0	62.77	1.23	1.5129
Noviembre	53.0	54.14	−1.14	1.2996
Diciembre	45.9	47.63	−1.73	2.9929
Totales	518.9	518.92	−0.02	16.7550

La regresión tiene dos grados de libertad, puesto que calculamos dos parámetros, a_1 y b_1 además de la media. La suma de cuadrados para la regresión se obtiene al sustraer la suma de cuadrados de las desviaciones del total. La proporción de la suma de cuadrados total, asociada a la regresión, es 1 016.187/1 032.942 = 0.9838.

El valor de 62.027 para a_0 es interesante. Nos referimos anteriormente a la misma como una media ponderada. Esta es una estimación de lo que sería la media, si tuvieramos datos para todo el año. En realidad, es una estimación muy cercana a la verdadera media anual de 61.34, basada en registros completos. Obviamente, la media de los datos observados, 518.9/9 = 57.656 sería una estimación muy inexacta de la media anual, puesto que los datos perdidos fueron todos de meses calientes; sin embargo, el valor de a_0 obtenido por el ajuste de la curva de Fourier, nos permite alcanzar una estrecha estimación a pesar de los datos perdidos.

Los valores de a_1 y b_1 pueden utilizarse para encontrar la semiamplitud y el ángulo de fase:

La semiamplitud $= A = \sqrt{a_1^2 + b_1^2} = \sqrt{(-17.056)^2 + (-.741)^2} = 17.1$

$\theta' = \tan^{-1} b_1/a_1$ = ángulo cuya tangente es $-.741/-17.056 = 2.5°$

por las reglas de los signos $\theta = 180° + \theta' = 182.5°$.

Puesto que un mes = 30°, 182.5° es equivalente a 6.1 meses. Esto expresa que el punto máximo de la curva se registra a aproximadamente 6.1 meses después de la fecha inicial. Utilizamos la **media** de enero como fecha inicial, de modo que denominamos a la misma enero 15. Por tanto, nuestra máxima calculada es 6.1 meses después de enero 15, es decir, aproximadamente julio 18.

Hemos seguido los pasos para el ajuste de datos a una curva de Fourier simple de un grado. Si resulta necesario ajustar los datos en esta forma a una curva de dos o más grados, los cálculos se vuelven bastante complicados, puesto que deben añadirse dos ecuaciones para cada grado. Tales problemas pueden manipularse muy fácilmente en una computadora. La figura 14.10 muestra una curva que relaciona la fecha de siembra con la longitud de tiempo para la cosecha del apio que fue calculada y trazada en una computadora. Diez años de datos se utilizaron para calcular esta curva.

Afortunadamente, si tenemos datos obtenidos a intervalos iguales a través de un ciclo completo, los cálculos llegan a ser grandemente simplificados. En el próximo capítulo describiremos los métodos abreviados para manipular datos de este tipo.

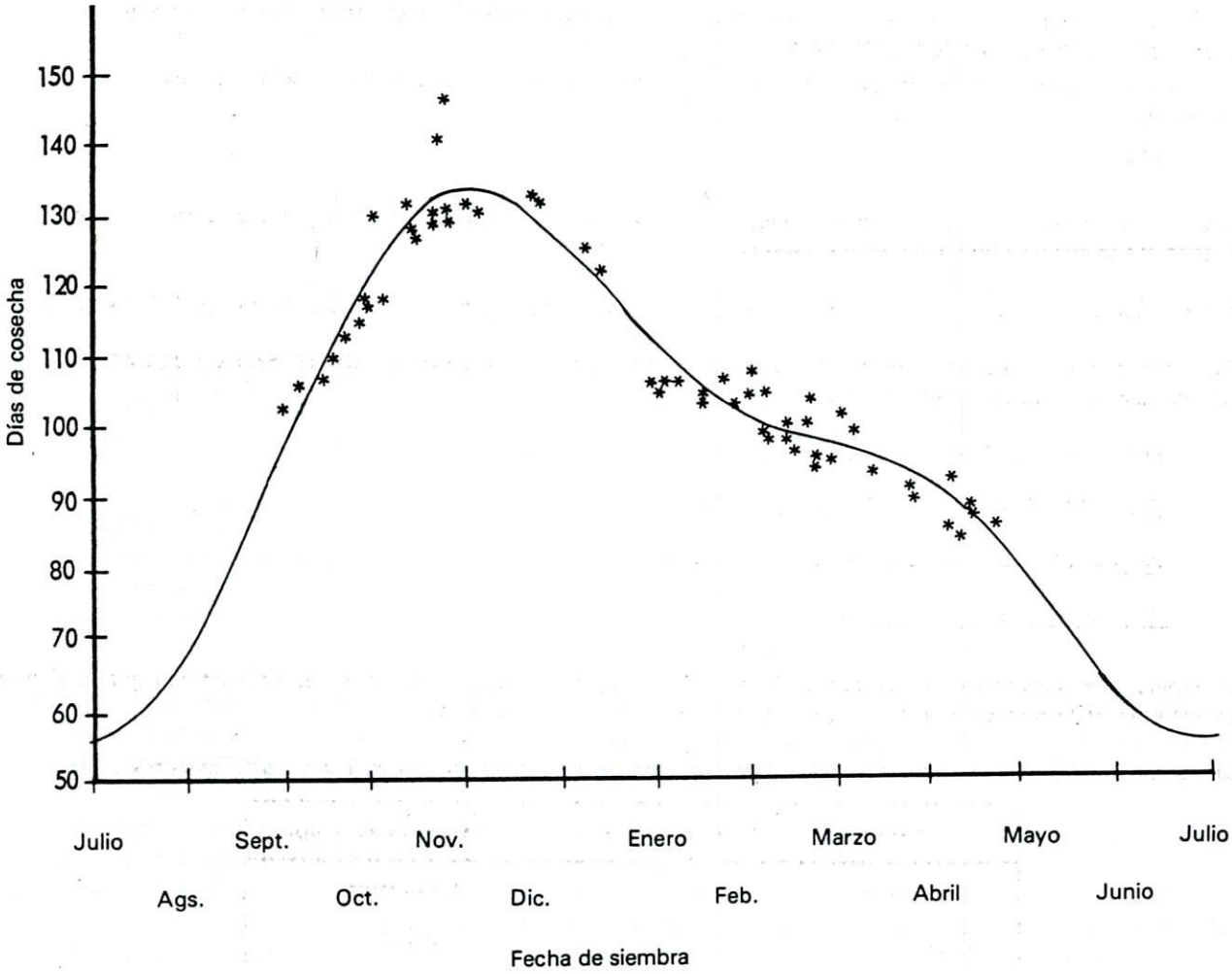


Figura 14.10. Fecha de siembra y días para cosechar el apio de Ventura, California.

RESUMEN

Si el diagrama de dispersión de dos variables muestra una tendencia de los puntos a estar dispersos alrededor de una curva, en vez de alrededor de una línea recta, resulta aconsejable analizar la relación curvilínea entre las variables. Las equivocaciones al hacerlo pueden ser muy engañosas.

Si los logaritmos de dos variables dan lugar a un diagrama de dispersión que parece ajustarse a una línea recta, la curva que describe la relación es de la forma:

$$Y = aX^b.$$

Las variables que incluyen diferentes números de dimensiones, parecen ajustarse mejor a este tipo de curva.

Para analizar tales datos, transformamos las variables originales X y Y en nuevas variables $X' = \log X$ y $Y' = \log Y$. Entonces procedemos a calcular la correlación lineal y la regresión, determinando la ecuación de regresión para la recta:

$$Y' = a' + bX'$$

Si el logaritmo de Y trazado contra X forma un diagrama de dispersión de línea recta, la curva apropiada es de la forma:

$$Y = ab^X$$

Cabe esperar que los datos en que la variable Y tiende a tener una proporción bastante constante de incremento o disminución, se ajusten a este tipo de curva.

Para analizar, solamente transformamos Y en $Y' = \log Y$ y procedemos como en la regresión lineal, ajustando a la ecuación:

$$Y' = a' + b'X$$

Los datos curvilíneos que no se aproximan a los datos lineales bajo una transformación logarítmica o semilogarítmica, pueden ser ajustados a un polinomio de la forma:

$Y = a + bX + cX^2 + dX^3 + \ldots$, utilizando tantos términos como sea necesario para obtener un ajuste satisfactorio.

Para encontrar los coeficientes desconocidos a, b, c, d, etc., resolvemos el conjunto de ecuaciones simultáneas, conocidas como ecuaciones **normales**:

$$an + b\Sigma X + c\Sigma X^2 + d\Sigma X^3 + \ldots = \Sigma Y$$

$$a\Sigma X + b\Sigma X^2 + c\Sigma X^3 + d\Sigma X^4 + \ldots = \Sigma XY$$

$$a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 + d\Sigma X^5 + \ldots = \Sigma X^2Y$$

$$a\Sigma X^3 + b\Sigma X^4 + c\Sigma X^5 + d\Sigma X^6 + \ldots = \Sigma X^3Y$$

El número de ecuaciones y el de términos a la izquierda del signo de igualdad deben ser respectivamente iguales al número de coeficientes requeridos, o uno más que el grado de la ecuación de regresión.

Las ecuaciones de los primeros cuatro grados reciben nombres especiales, al igual que algunas de las curvas:

Grado	Nombre de la ecuación	Nombre de la curva
Primero	Lineal	Línea recta
Segundo	Cuadrático	Parábola
Tercero	Cúbica	Parábola cúbica
Cuarto	Cuártica	Parábola cuártica
Quinto	Quíntica	Parábola quíntica

Si las desviaciones de las observaciones de una curva calculada parecen ser más o menos aleatorias, usualmente no vale la pena el ajuste a una curva de grado más elevado. Si las desviaciones son sistemáticas o en grupos definidos en cuanto al signo, generalmente resulta ventajoso calcular la ecuación del grado próximo más alto.

Los cálculos de coeficientes para ecuaciones mayores que la cúbica sólo deben llevarse a cabo por medio del dominio de métodos especiales (como el de Doolittle) o mediante una computadora electrónica.

Cuando los valores de X son igualmente espaciados, se ahorrará mucho tiempo utilizando los metodos abreviados descritos en el capítulo 15.

La combinación de los métodos logarítmico y polinómico, algunas veces dará como resultado un ajuste mucho mejor a los datos que cualquiera de los dos métodos por separado.

Los datos que fluctúan hacia arriba y hacia abajo con el tiempo en un patrón bastante regular, pueden ajustarse a una curva periódica (de Fourier) de la forma:

$$Y = a_0 + a_1 \cos CX + b_1 \operatorname{sen} CX + a_2 \cos 2CX + b_2 \operatorname{sen} 2CX + \ldots$$

Las ecuaciones **normales** para determinar los coeficientes desconocidos son:

$$a_0 n + a_1 \Sigma U_1 + b_1 \Sigma V_1 + a_2 \Sigma U_2 + b_2 \Sigma V_2 + \ldots = \Sigma Y$$

$$a_0 \Sigma U_1 + a_1 \Sigma U_1^2 + b_1 \Sigma U_1 V_1 + a_2 \Sigma U_1 U_2 + b_2 \Sigma U_1 V_2 + \ldots = \Sigma U_1 Y$$

$$a_0 \Sigma V_1 + a_1 \Sigma U_1 V_1 + b_1 \Sigma V_1^2 + a_2 \Sigma U_2 V_1 + b_2 \Sigma V_1 V_2 + \ldots = \Sigma V_1 Y$$

$$a_0 \Sigma U_2 + a_1 \Sigma U_1 U_2 + b_1 \Sigma U_2 V_1 + a_2 \Sigma U_2^2 + b_2 \Sigma U_2 V_2 + \ldots = \Sigma U_2 Y$$

$$a_0 \Sigma V_2 + a_1 \Sigma U_1 V_2 + b_1 \Sigma V_1 V_2 + a_2 \Sigma U_2 V_2 + b_2 \Sigma V_2^2 + \ldots = \Sigma V_2 Y$$

donde $U_i = \cos i(CX)$ y $V_i = \text{sen } i(CX)$

En el capítulo 15 se describirán métodos abreviados para cuando los datos se obtienen a partir de intervalos de tiempo igualmente espaciados a través de un ciclo completo.

Capítulo 15

Métodos abreviados de regresión para observaciones o tratamientos a intervalos iguales

Frecuentemente sucede que hacemos observaciones de una variable dependiente, Y, asociada con valores igualmente espaciados de una variable independiente, X; por ejemplo, si la variable independiente es el tiempo y hacemos lecturas de Y a intervalos diarios, semanales, mensuales o anuales, las X o tiempos son igualmente espaciados. Otro caso donde a menudo tenemos intervalos igualmente espaciados de X, es el de experimentos que contemplan proporciones de fungicidas, insecticidas, fertilizantes, etc. Un experimento en el que las proporciones de tratamiento son igualmente espaciadas, presenta ventajas reales desde el punto de vista de la facilidad del análisis.

El método abreviado que vamos a describir se estudió bajo el título de **comparaciones de tendencias** en el capítulo 6. Dicho método es tan útil, que nos parece válido ampliar aquel estudio y relacionarlo con el capítulo anterior de dicha sección acerca de la regresión curvilínea.

Usualmente, los estadísticos se refieren a éste como el **método de polinomios ortogonales.** Si aquellos que padecen un bloqueo mental al verse enfrentados a un título tan imponente, pueden pensar en el mismo como en el "método abreviado para medir tendencias" y lo encontrarán fácil de utilizar y efectivo para el ahorro de tiempo.

AJUSTE DE LA CURVA POLINÓMICA

La base del método para el ajuste de polinomios es la tabla A.11*, cuya utilización elimina muchos de los laboriosos cálculos comúnmente requeridos en la regresión curvilínea. La tabla puede utilizarse para: a) encontrar las ecuaciones de regresión lineal, cuadrática, cúbica y cuártica para cualquier número de observaciones igualmente espaciadas hasta 25, y b) separar la suma de cuadrados del tratamiento en un análisis de varianza en componentes lineales, cuadráticos, cúbicos, cuárticos y residuales, hasta para 25 tratamientos u observaciones igualmente espaciados.

En la parte superior de la tabla se encuentran los valores de n, es decir, el número de observaciones o tratamientos. Para cualquier problema dado, necesitamos utilizar solamente la porción de la tabla por debajo del valor apropiado de n. La primera columna de coeficientes, encabezada por c_1 además de ser utilizada para diversos cálculos, consiste en **valores codificados de X.** La codificación se hace en forma tal que su

*Esta tabla, calculada por los autores, se utiliza en vez de una de las diversas tablas similares que aparecen en otras publicaciones. Para una mejor comprensión, los valores de K no aparecen en cualesquier otras tablas publicadas.

resultado son números enteros lo más pequeños posibles. A pesar de los valores de X igualmente espaciados, si n es impar, podemos tomar: $X' = (X - \bar{X})/L$, donde L es el intervalo entre valores sucesivos de X. Si n es par, tomamos: $X' = (X - \bar{X})2/L$. Estas transformaciones darán los valores de la columna c_1.

Los pasos para determinar las ecuaciones de regresión lineal, cuadrática, cúbica y cuártica son los siguientes:

1. Disponer los valores de Y en una columna, de acuerdo con los valores ascendentes de las X asociadas, empezando con la Y correspondiente al menor valor de X.
2. Multiplicar los valores de Y por los coeficientes para c_1, c_2, c_3 y c_4 mostrados en la tabla, obteniéndose cuatro columnas.
3. Encontrar la suma de cada columna, observando los signos más y menos. Estas sumas se denotan por ΣY, P_1, P_2, P_3 y P_4.
4. Aplicando los valores obtenidos de P y los valores de K provenientes de la tabla, las ecuaciones lineales, cuadráticas, cúbicas y cuárticas pueden ser planteadas a partir de las siguientes relaciones:

   Ecuacion lineal: $\hat{Y}_L = \bar{Y} + (K_2P_1)X'$

   Cuadrática: $\hat{Y}_Q = (\bar{Y} - K_1P_2) + (K_2P_1)X' + (K_4P_2)X'^2$

   Cúbica: $\hat{Y}_C = (\bar{Y} - K_1P_2) + (K_2P_1 - K_3P_3)X' + (K_4P_2)X'^2 + (K_5P_3)X'^3$

   Cuártica: $\hat{Y}_4 = (\bar{Y} - K_1P_2 + K_8P_4) + (K_2P_1 - K_3P_3)X' + (K_4P_2 - K_7P_4)X'^2 + (K_5P_3)X'^3 + (K_6P_4)X'^4$

   Nótese que dichas ecuaciones están expresadas en términos de **valores codificados de X.**
5. Si los valores de Y en el paso 1 fuesen totales de diversas observaciones o repeticiones en cada nivel de X, y si deseamos que las ecuaciones estén dadas en términos de las medias, debemos dividir cada término de las ecuaciones entre el número de repeticiones. (Esto debe ser igual para todos los niveles de X.)

La tabla 15.1 muestra la producción total de leche diaria de 37 vacas, en libras, registrada una vez al mes para los diez meses desde el parto hasta el final de la lactancia. Aplicaremos los cinco pasos anteriores a estos datos.

Tabla 15.1. Registros de producción de leche de 37 vacas durante 10 meses.

Producción de leche (Y)	Mes (X)	X' (c_1)	c_1Y	c_2	c_2Y	c_3	c_3Y	c_4	c_4Y
2 442.3	1	−9	−21 980.7	6	14 653.8	−42	−102 576.6	18	43 961.4
2 517.6	2	−7	−17 623.2	2	5 035.2	14	35 246.4	−22	−55 387.2
2 334.4	3	−5	−11 672.0	−1	−2 334.4	35	81 704.0	−17	−39 684.8
2 166.1	4	−3	−6 498.3	−3	−6 498.3	31	67 149.1	3	6 498.3
2 030.0	5	−1	−2 030.0	−4	−8 120.0	12	24 360.0	18	36 540.0
1 903.9	6	1	1 903.9	−4	−7 615.6	−12	−22 846.8	18	34 270.2
1 779.5	7	3	5 338.5	−3	−5 338.5	−31	−55 164.5	3	5 338.5
1 630.6	8	5	8 153.0	−1	−1 630.6	−35	−57 071.0	−17	−27 720.2
1 485.7	9	7	10 399.9	2	2 971.4	−14	−20 799.8	−22	−32 685.4
1 304.7	10	9	11 742.3	6	7 828.2	42	54 797.4	18	23 484.6
Totales 19 594.8			P_1 = −22 266.6		P_2 = −1 048.8		P_3 = 4 798.2		P_4 = −5 384.6

Los coeficientes c_1, c_2, c_3 y c_4 fueron copiados de la tabla A.11 y multiplicados por los valores correspondientes de Y (producción de leche). Los totales de dichas columnas arrojaron los valores para ΣY, P_1, P_2, P_3 y P_4. Estamos ahora preparados para aplicar el paso 4 y plantear las ecuaciones.

$$\hat{Y}_L = 1\,959.48 + (^1/_{330})\ (-22\,266.6)X' = 1\,959.48 - 67.475\,X'$$

$$\hat{Y}_Q = [1\,959.48 - (^1/_{32})\ (-1\,048.8)] - 67.475X' + (^1/_{1,056})(-1\,048.8)X'^2$$

$$= 1\,992.26 - 67.475X' - .9932X'^2$$

$$\hat{Y}_C = 1\,992.26 + [-67.475 - (293\ 205\ 920)\ (4\,798.2)]\,X' - .9932X'^2 + (1\ 46\ 184)\ (4\,798.2)X'^3$$

$$= 1\,992.26 - 74.302X' - .9932X'^2 + .11651X'^3$$

$$\hat{Y}_4 = [1\,992.26 + (9\ 1\ 280)\ (-5\,384.6)] - 74.302X' + [-.9932 - (41\ 54912)\ (-5384.6)]\,X'^2$$

$$+ .11651X'^3 + (1\ 109\ 824)\ (-5\,384.6)X'^4$$

$$= 1\,954.40 - 74.302X' + 3.0272X'^2 + .11651X'^3 - .049029X'^4$$

Estas ecuaciones están basadas en la producción total de leche de 37 vacas. Si las deseamos con base en una por vaca, simplemente dividimos cada término entre 37 y obtenemos:

$$\hat{Y}_L = 52.959 - 1.8236X'$$

$$\hat{Y}_Q = 53.845 - 1.8236X' - .02684X'^2$$

$$\hat{Y}_C = 53.845 - 2.0082X' - .02684X'^2 + .003149X'^3$$

$$\hat{Y}_4 = 52.822 - 2.0082X' + .08182X'^2 + .003149X'^3 - .0013251X'^4$$

En la práctica, no es necesario construir una tabla como la 15.1, puesto que los valores requeridos de P pueden encontrarse por acumulación de los productos en una máquina calculadora, sin anotar cada producto por separado. Debe darse especial atención a los signos de los coeficientes. Cuando un coeficiente es negativo, su producto con el valor correspondiente de Y debe ser sustraído de la suma acumulada.

Es muy importante tener en cuenta que las ecuaciones que hemos calculado se encuentran en términos de X', es decir, los **valores codificados** de X. Estos son idénticos para los coeficientes c_1. Supóngase que en nuestro ejemplo deseamos calcular la producción de leche predicha por vaca a partir de la ecuación cuadrática para el tercer mes. Remitiéndonos a la tabla 15.1, vemos que la X' para el tercer mes es −5, de modo que sustituimos −5 por X' en la ecuación cuadrática:

$$\hat{Y}_Q = 53.845 - 1.8236(-5) - .02684(-5)^2$$

$$= 53.845 + 9.118 - .671 \quad 62.292$$

Generalmente, es más fácil trabajar con las ecuaciones en esta forma; pero si por alguna razón deseamos expresar las ecuaciones en términos de los valores originales de X, es necesario sustituir $(X - \bar{X})$ L o $(X - \bar{X})2$ L por X' en las ecuaciones, dependiendo de que n sea impar o par. Para ilustrar el procedimiento, plantearemos nuestra ecuación cuadrática $\hat{Y}_Q = 53.845 - 1.8236X' - .02684X'^2$ en términos de X.

En este caso, $n = 10$ fue par, de modo que sustituimos $(X - \bar{X})2\ L$ por X'. El intervalo entre los valores sucesivos de X fue 1, de manera que $L = 1$. El valor de $\bar{X}$ fue 5.5, de modo que tenemos $X' = (X - 5.5)2\ 1$ o $2X - 11$. Sustituyendo estos resultados en nuestra ecuación, obtenemos:

$$\hat{Y}_Q = 53.845 - 1.8236(2X - 11) - .02684(2X - 11)^2$$

$$= 53.845 - 1.8236(2X - 11) - .02684(4X^2 - 44X + 121)$$

$$= 53.845 - 3.6472X + 20.0596 - .10736X^2 + 1.18096X - 3.23764$$

Agrupando los términos, obtenemos:

$$\hat{Y}_Q = 70.65696 - 2.46624X - .10736X^2$$

Utilicemos esta ecuación con el fin de calcular nuevamente Y_Q para el tercer mes. Sustituyendo 3 por X en esta nueva ecuación, obtenemos:

$$\hat{Y}_Q = 70.65696 - 2.46624(3) - .10736(3)^2 = 62.292 \quad \text{igual que antes.}$$

Veamos cuánto trabajo hemos ahorrado. Aplicando los métodos del capítulo 14 (que debemos usar si las X no están igualmente espaciadas) para encontrar las cuatro ecuaciones de regresión, necesitaríamos hallar $\Sigma X, \Sigma X^2, \Sigma X^3, \Sigma X^4, \Sigma X^5, \Sigma X^6, \Sigma X^7, \Sigma X^8, \Sigma Y, \Sigma XY, \Sigma X^2Y, \Sigma X^3Y$ y ΣX^4Y. Estos valores deberían ser sustituidos en las ecuaciones normales y tendríamos que resolver conjuntos de ecuaciones simultáneas, desde 2 para los coeficientes lineales hasta 5 para los cuárticos. Si el lector trabajó en los ejemplos del capítulo 14, podrá apreciar cuán laboriosa sería esta tarea. Compárense todos estos cálculos con el método abreviado. Aplicándolo, sólo necesitamos ΣY, P_1, P_2, P_3, y P_4. Sustituyendo estos valores en las ecuaciones estándar del paso 4, obtenemos directamente las cuatro ecuaciones de regresión requeridas. Sólo tenemos que calcular cinco sumas, en lugar de 13, y no existen ecuaciones simultáneas que resolver.

Ahora que tenemos las cuatro ecuaciones, podemos notar cómo los valores calculados a partir de las mismas se comparan con la producción de leche observada para cada mes. Es mejor trabajar con los totales, en vez de hacerlo con las medias, puesto que se introducen muy pocos errores de aproximación. La tabla 15.2 muestra los valores calculados a partir de cada ecuación, así como sus desviaciones de los valores observados.

Debemos considerar diversos aspectos respecto de esta tabla. La suma de las desviaciones de todas las curvas debe ser igual a cero, excepto por pequeños errores de aproximación. Esto suministra una comprobación de los cálculos. La suma de cuadrados de las desviaciones de una curva suministra una medida de la estrechez del ajuste; cuanto menor sea la suma de cuadrados, más estrecho será el ajuste de la curva a los datos. Cada grado añadido resulta en una reducción de esta suma de cuadrados. Esto siempre debe ocurrir; en caso contrario, búsquese un error en los cálculos. (El problema consiste en si el mejoramiento del ajuste es significativo; demostraremos cómo comprobarlo en forma abreviada.) Por ahora, nótese simplemente que existe una reducción moderada en la suma de cuadrados cuando pasamos de la curva lineal a la cuadrática; una reducción muy pequeña cuando pasamos de la cuadrática a la cúbica; y una gran reducción cuando pasamos de la cúbica a la cuártica. Finalmente, nótese que en los primeros tres grados, los signos de las desviaciones parecen distribuirse en patrones bastante definidos, mientras que aquellos para las del cuarto son más o menos aleatorios.

Tabla 15.2. Producción mensual de leche de 37 vacas, observada y calculada

Y observada	$\hat{Y}_L$	$Y - \hat{Y}_L$	$\hat{Y}_Q$	$Y - \hat{Y}_Q$	$\hat{Y}_C$	$Y - \hat{Y}_C$	$\hat{Y}_4$	$Y - \hat{Y}_4$
2 442.3	2 566.8	−124.5	2 519.1	−76.8	2 495.6	−53.3	2 461.7	−19.4
2 517.6	2 431.8	85.8	2 415.9	101.7	2 423.7	93.9	2 465.2	52.4
2 334.4	2 296.9	37.5	2 304.8	29.6	2 324.4	10.0	2 356.4	−22.0
2 166.1	2 161.9	4.2	2 185.7	−19.6	2 203.1	−37.0	2 197.4	−31.3
2 030.0	2 027.0	3.0	2 058.7	−28.7	2 065.5	−35.5	2 031.6	−1.6
1 903.9	1 892.0	11.9	1 923.8	−19.9	1 917.1	−13.2	1 883.2	20.7
1 779.5	1 757.1	22.4	1 780.9	−1.4	1 763.6	15.9	1 757.9	21.6
1 630.6	1 622.1	8.5	1 630.1	0.5	1 610.5	20.1	1 642.5	−11.9
1 485.7	1 487.2	−1.5	1 471.3	14.4	1 463.4	22.3	1 504.9	−19.2
1 304.7	1 352.2	−47.5	1 304.5	0.2	1 328.0	−23.3	1 294.1	10.6
Σ desv		−0.2		0.0		−0.1		−0.1
Σ (desv)^2		27 268.90		18 930.76		16 258.59		6 105.23
Σ (desv)^2/37		737.00		511.64		439.42		165.01

Cómo separar la suma de cuadrados

Determinar todos los valores calculados y sus desviaciones de los valores observados, y luego encontrar las sumas de cuadrados de dichas desviaciones, es un laborioso procedimiento. La segunda característica del método abreviado para analizar datos igualmente espaciados es la facilidad con que pueden calcularse dichas sumas de cuadrados. Si observamos la tabla A.11 por debajo de cualquier valor de n, podemos reconocer que los valores de c son en realidad conjuntos ortogonales de coeficientes. La suma de cada columna de coeficientes es igual a cero, y los productos de los coeficientes correspondientes de dos columnas cualesquiera son también iguales a cero. En el capítulo 6 aprendimos que la suma de cuadrados asociada con un solo grado de libertad puede encontrarse a partir de un conjunto de coeficientes, mediante la fórmula general.

$$SC = \left(\Sigma c_i T_i\right)^2 / \left(r \Sigma c_i^2\right)$$

Tal como fue previamente calculado, P_1 es igual a $\Sigma c_i T_i$ cuando las c son los coeficientes lineales.

Análogamente, $P_2 = \Sigma c_i T_i$ cuando usamos los coeficientes cuadráticos, y así sucesivamente. Los divisores mostrados en la tabla A.11 son las sumas de cuadrados de los coeficientes; por tanto, la suma de cuadrados debida a la regresión lineal es simplemente P_1^2 (divisor por número de repeticiones). Análogamente, la suma de cuadrados para la regresión cuadrática es: P_2^2 (divisor por número de réplicas), y así sucesivamente hasta el componente cuártico. Después de calcular las sumas de cuadrados para cada componente, podemos encontrar la suma de cuadrados residual al sustraer las sumas de cuadrados componentes de la suma de cuadrados total. Esta suma de cuadrados residual es igual a la suma de cuadrados de las desviaciones de la curva de los datos observados.

Apliquemos este método de separación a los datos de la producción de leche. El valor de P_1 que encontramos fue −22 266.6, de modo que la SC lineal es:

$$(-22\,266.6)^2 / (330 \times 37) = 40\,606.18$$

La suma total de cuadrados de Y fue 41 343.01, de manera que la suma de cuadrados residual es 41 343.01 − 40606.18 = 736.83. Esta es igual (excepto por una pequeña diferencia debida a la aproximación) a la suma de cuadrados de las observaciones del componente lineal encontrado por un método mucho más difícil en la tabla 15.2.

Puesto que P_2 fue igual a −1 048.8, la suma de cuadrados para el componente cuadrático es

$$(-1\,048.8)^2 / (132 \times 37) = 225.22$$

Sustrayendo este resultado de 736.83, obtenemos un residuo de 511.62. El valor calculado en la tabla 15.2 fue 511.64.

P_3 fue igual a 4 798.2, de modo que la suma de cuadrados para el componente cúbico es:

$(4\,798.2)^2 / (8\,580 \times 37) = 72.52$ arrojando un residuo de 439.09 (en comparación con 439.42 de la tabla 15.2).

Finalmente, P_4 fue igual a −5 384.6, de manera que la suma de cuadrados para el componente cuártico es:

$(-5\,384.6)^2 / (2\,860 \times 37) = 273.99$, arrojando un residuo de 165.10.

Todos estos resultados pueden resumirse en una tabla de análisis de varianza, en la que tanto las sumas de cuadrados para las vacas como el error, se obtuvieron de los registros de vacas individuales.

Tabla 15.3. Análisis de varianza de los registros de producción de leche.

Fuente de variación	gl	SC	MC	F
Total	369	76 167.74		
Vacas	36	23 464.56	651.79	18.59**
Meses	9	41 343.01	4 593.67	131.02**
Componente lineal	1	40 606.18	40 606.18	1 158.19**
Desviación del componente lineal	8	736.83	92.10	2.63*
Componente cuadrático	1	225.22	225.22	6.42*
Desviación del componente cuadrático	7	511.61	73.09	2.08*
Componente cúbico	1	72.52	72.52	2.07ns
Desviación del componente cúbico	6	439.09	73.18	2.09ns
Componente cuártico	1	273.99	273.99	7.81**
Desviación del componente cuártico	5	165.10	33.02	0.94ns
Error	324	11 360.17	35.06	

Se registró una diferencia altamente significativa entre vacas y entre meses. Ninguno de estos resultados es sorprendente, pero deseamos saber más acerca del patrón de cambio de la producción de leche mes a mes. El valor F muy alto para los componentes lineales revela la existencia de una tendencia descendente altamente significativa. La desviación significativa del componente lineal indica que una línea recta no explica completamente la variación mensual. El componente cuadrático significativo muestra que una curva simple representa un mejoramiento sobre una línea recta, pero aún persiste una cantidad significativa de variación residual. El ajuste de una curva cúbica no resultó en un mejoramiento significativo, y el residuo dejado no es significativo. En este punto, diversos investigadores optan por detener su trabajo. Frecuentemente, como en este caso, esto es una equivocación. El componente cuártico explicó tal proporción elevada de la suma de cuadrados restante, en el sentido de que fue altamente significativa. La desviación del componente cuártico no es significativa. La probabilidad de encontrar otro componente significativo es muy pequeña, pues incluso si un solo componente explica el 80% de la variabilidad restante, ésta no sería significativa. Por tanto, hay justificación para concluir el análisis en este punto.

AJUSTE DE LA CURVA PERIÓDICA

La tabla A.12 presenta conjuntos de coeficientes ortogonales para ajustar datos periódicos cuando las observaciones son igualmente espaciadas a través de un ciclo completo. La tabla se ha construido para valores escogidos de n más comúnmente encontrados en el tratamiento con ciclos diarios, semanales o anuales.

Al contrario de los conjuntos de coeficientes con los que hemos estado tratando, éstos no pueden ser reducidos a números enteros pequeños. Por esta razón, los cálculos de los valores de P son de alguna forma más difíciles; pero en otros aspectos, el cálculo de las ecuaciones y la partición de las sumas de cuadrados es aún más fácil que tratándose de polinomiales, puesto que no se necesitan divisores o valores de K especiales.

La razón por la cual tratar con intervalos igualmente espaciados es mucho más simple que tratar con datos irregulares, es que se elimina la mayoría de los términos de las ecuaciones normales presentadas en el capítulo

14. Luego, $\Sigma U_i = \Sigma V_i = 0$ donde i es cualquier subíndice; además, $\Sigma U_i^2 = \Sigma V_i^2 = n/2$. Por tanto, la primera ecuación normal, que es:

$$n\,a_0 + a_1\,\Sigma U_1 + b_1\,\Sigma V_1 + a_2\,\Sigma U_2 + b_2\,\Sigma V_2 + \ldots = \Sigma Y$$

se reduce a: $n\,a_0 = \Sigma Y$, o $a_0 = \Sigma Y/n = \bar{Y}$

Análogamente, las otras ecuaciones normales se reducen a:

$$a_1\,(n/2) = \Sigma U_1 Y \text{ or } a_1 = (2\Sigma U_1 Y)/n$$

$$b_1 = (2\Sigma V_1 Y)/n$$

$$a_2 = (2\Sigma U_2 Y)/n$$

$$b_2 = (2\Sigma V_2 Y)/n$$

y así sucesivamente, siguiendo el mismo patrón, excepto cuando n es par, en cuyo caso el último coeficiente que puede calcularse es:

$$a_{(n/2)} = \left[\Sigma U_{(n/2)} Y\right]/n$$

(Rara vez conduciríamos un análisis hasta este punto, puesto que entonces no existiría suma de cuadrados residual. En otras palabras, una ecuación conducida hasta aquí se ajustaría exactamente a todos los puntos de los datos, lo cual equivale a ajustar una línea recta a dos puntos.)

Adoptaremos un símbolo similar a uno de los utilizados en el ajuste de polinomios, denotando $\Sigma U_i Y$ como PU_i y $\Sigma V_i Y$ como PV_i. Nótese que en el caso del polinomio tuvimos un solo valor P para cada grado, pero al ajustar una curva periódica necesitamos dos valores P, denominados PU y PV, para cada grado* de ajuste.

Los términos generales de la ecuación son:

$$a_0 = \bar{Y},\ a_i = 2PU_i/n,\ b_i = 2PV_i/n$$

Apliquemos este método para ajustar una curva periódica a los datos completos de temperaturas medias mensuales en Stockton, California, mostradas en la tabla 15.4.

Lo que hemos calculado es una ecuación de tipo general en la cual podemos sustituir cualquier valor de X y buscar los senos y cosenos apropiados en una tabla trigonométrica; sin embargo, si sólo estamos interesados en calcular los valores correspondientes a los puntos de datos observados, podemos simplemente sustituir U_1 por cos CX, V_1 por sen CX, U_2 por cos 2CX y V_2 por sen 2CX en la ecuación; por ejemplo, para encontrar Y para marzo (mes número 2, puesto que enero fue denominado mes 0), calculamos:

$$\hat{Y}_2 = 61.34 - 16.0751\,(.5) - 1.5788\,(.866) - .275\,(-.5) + 1.9774\,(.866) = 53.79$$

Si deseamos el valor calculado sólo para la curva de primer grado, sencillamente utilizamos los tres primeros términos de la ecuación anterior:

$$\hat{Y}_1 = 61.34 - 16.0751\,(.5) - 1.5788\,(.866) = 51.94$$

*Hemos designado a cada par de términos añadidos a la ecuación general de regresión periódica como **grado**, para mantener la analogía con el polinomio general. Técnicamente hablando, éstas reciben el nombre de **armónicas.**

Tabla 15.4. Temperaturas medias mensuales en Stockton, California, con los cálculos para el ajuste de una curva periódica de segundo grado. ($C = {}^{1}/_{12} \times 360° = 30°$)

Mes (X)	Temp (Y)	cos CX (U_1)	U_1Y	sen CX (V_1)	V_1Y	cos 2CX (U_2)	U_2Y	sen 2CX (V_2)	V_2Y
0	44.7	1.0	44.7000	0.0	0.0000	1.0	44.7000	0.0	0.0000
1	49.0	0.866	42.4340	0.5	24.5000	0.5	24.5000	0.866	42.4340
2	53.7	0.5	26.8500	0.866	46.5042	−0.5	−26.8500	0.866	46.5042
3	59.7	0.0	0.0000	1.0	59.7000	−1.0	−59.7000	0.0	0.0000
4	66.2	−0.5	−33.1000	0.866	57.3292	−0.5	−33.1000	−0.866	−57.3292
5	72.8	−0.866	−63.0448	0.5	36.4000	0.5	36.4000	−0.866	−63.0448
6	78.2	−1.000	−78.2000	0.0	0.0000	1.0	78.2000	0.0	0.0000
7	76.2	−0.866	−65.9892	−0.5	−38.1000	0.5	38.1000	0.866	65.9892
8	72.7	−0.5	−36.3500	−0.866	−62.9582	−0.5	−36.3500	0.866	62.9582
9	64.0	0.0	0.0000	−1.0	−64.0000	−1.0	−64.0000	0.0	0.0000
10	53.0	0.5	26.5000	−0.866	−45.8980	−0.5	−26.5000	−0.866	−45.8980
11	45.9	0.866	39.7494	−0.5	−22.9500	0.5	22.9500	−0.866	−39.7494
Totales	736.1		$PU_1 = -96.4506$		$PV_1 = -9.4728$		$PU_2 = -1.6500$		$PV_2 = 11.8642$
$a_0 =$	61.34		$a_1 = -16.0751$		$b_1 = -1.5788$		$a_2 = -0.2750$		$b_2 = 1.9774$

$$\hat{Y} = 61.34 - 16.0751 \cos CX - 1.5788 \operatorname{sen} CX - .275 \cos 2CX + 1.9774 \operatorname{sen} 2CX$$

Los valores calculados para las ecuaciones de primer y segundo grados se muestran en la tabla 15.5, conjuntamente con la desviación de los valores observados de estas dos curvas.

Tabla 15.5. Temperaturas medias mensuales observadas y calculadas en Stockton, California.

Mes	Y observada	$\hat{Y}_1$ primer grado	$(Y - \hat{Y}_1)$	$\hat{Y}_2$ segundo grado	$(Y - \hat{Y}_2)$
Enero	44.7	45.26	− 0.56	44.99	− 0.29
Febrero	49.0	46.63	2.37	48.20	0.80
Marzo	53.7	51.94	1.76	53.79	− 0.09
Abril	59.7	59.76	0.06	60.04	− 0.34
Mayo	66.2	68.01	−1.81	66.44	− 0.24
Junio	72.8	74.47	−1.67	72.62	− 0.18
Julio	78.2	77.42	0.78	77.14	1.06
Agosto	76.2	76.05	0.15	77.63	−1.43
Septiembre	72.7	70.74	1.96	72.59	− 0.11
Octubre	64.0	62.92	1.08	63.19	0.81
Noviembre	53.0	54.67	−1.67	53.09	− 0.09
Diciembre	45.9	48.21	−2.31	46.36	− 0.46
Totales			0.02		0.02
Σd^2			28.86		4.99

Cómo separar la suma de cuadrados

Al igual que con el polinomio, existe una forma muy fácil para separar la suma de cuadrados total, sin construir una tabla como la 15.5.

La suma de cuadrados para la regresión de primer grado es $2(PU_1^2 + PV_1^2)/n$ para la de segundo grado es $2(PU_2^2 + PV_2^2)/n$ y así sucesivamente. Al contrario del polinomio, no necesitamos un divisor diferente para cada grado. Las sumas de cuadrados para las desviaciones de los datos observados pueden obtenerse por sustracción. A partir de la tabla 15.4, encontramos que PU_1 fue −96.4506 y PV_1 fue −9.4728; por tanto, la suma de cuadrados de primer grado es:

$$2[(-96.4506)^2 + (-9.4728)^2]/12 = 1565.41$$

La suma de cuadrados total para Y fue 1 594.33, de modo que la suma de cuadrados para la desviación es:

1 594.33 − 1 565.41 = 28.92 un resultado que difiere del valor 28.86 encontrado en la tabla 15.5, debido a la aproximación.

Análogamente, la suma de cuadrados debida a la regresión de segundo grado es:

$$2[(-1.65)^2 + (11.8642)^2]/12 = 23.91$$

El residuo o la desviación de la suma de cuadrados de segundo grado es: 28.92 — 23.91 = 5.01 (en comparación con 4.99 en la tabla 15.5). Estos resultados se resumen a continuación:

Tabla 15.6. Análisis de varianza de los datos de temperatura.

Fuente de variación	gl	SC	CM	Valor de F
Meses	11	1594.33		
Primer grado	2	1565.41	782.705	243.61**
Desviación	9	28.92	3.213	
Segundo grado	2	23.91	11.955	16.70**
Desviación	7	5.01	0.716	

Nótese que cada término de la ecuación tiene **dos** grados de libertad. Esto se debe a que tienen que calcularse dos coeficientes, *a* y *b,* para cada grado. El cuadrado medio para cada grado se comprueba contra su componente residual, para constituir una prueba F. En este caso, tanto el primero como el segundo grados fueron altamente significativos.

Hemos ajustado una curva a las temperaturas medias mensuales y separado la suma de cuadrados para los meses en diversos componentes. Si deseamos tener en consideración los registros anuales individuales a partir de los cuales se calcularon dichas medias, el análisis de varianza es considerablemente más complicado. Para un estudio más detallado del tema, el lector debe consultar el Boletín 615 de la Estación Experimental de Agricultura de Connecticut, 1958, que lleva por título *Regresión periódica en biología y climatología,* de C.I. Bliss.

La curva de segundo grado que hemos calculado está en realidad formada por dos curvas sinusoidales simples, una sobre la otra. La primera tiene una semiamplitud

$$A = \sqrt{a_1^2 + b_1^2}\text{, de modo que: } A = \sqrt{(-16.0751)^2 + (-1.5788)^2} = 16.13$$

El ángulo de fase es $\tan^{-1}(b_1/a_1) + 180^\circ =$ ángulo cuya tangente es .0982 + 180° = 185° 36' que, convertido a unidades de tiempo, equivale a aproximadamente 6 meses y 5 días después de la iniciación del ciclo. Puesto que nuestro ciclo comienza con la media de enero, podemos denominarla enero 15, de manera que la máxima de nuestra curva caerá en julio 20 y la mínima seis meses antes de enero 20.

Si consultamos la figura 15.1, y nos fijamos en la curva continua en la mitad inferior de la misma, veremos que las temperaturas observadas tienden a encontrarse por encima de la curva en su primera y tercera partes, y por debajo de la misma en la segunda y cuarta partes. La curva de segundo grado se ajusta ampliamente a estas discrepancias. Tiene una semiamplitud

$$A = \sqrt{a_2^2 + b_2^2} = \sqrt{(-.275)^2 + (1.9774)^2} = 2.00$$ y un ángulo de fase de

$180^\circ - \tan^{-1}(b_2/a_2) = 180^\circ - \tan^{-1} 7.1905 = 180^\circ - 82^\circ\ 5' = 97^\circ 5'$. Este ángulo debe dividirse entre 2, puesto que ahora estamos tratando con una curva de 2 ciclos, de modo que tenemos una máxima en 48°32.5' o aproximadamente un mes y 18 días después de enero 15. Existe otra máxima 6 meses después y una mínima en los 3 meses siguientes a cada máxima. Esto se traza como la curva de puntos de la parte inferior de la figura 15.1.

Sumando estas dos curvas con la media 61.34, obtenemos la curva resultante en la mitad superior de la figura 15.1.

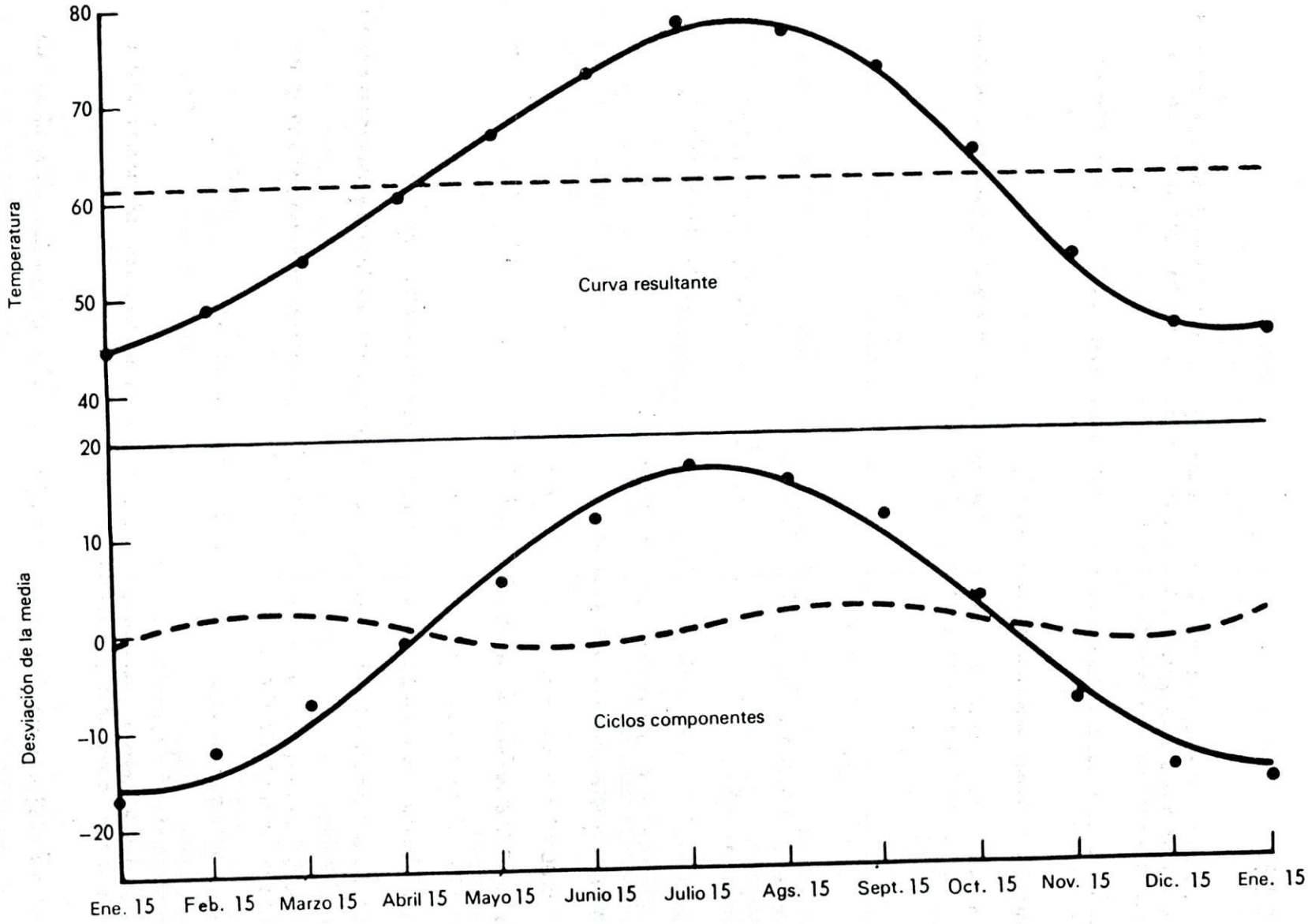


Figura 15.1. Temperaturas medias mensuales en Stockton, California. Curva de Fourier de segundo grado y sus componentes.

RESUMEN

Para observaciones o tratamientos igualmente espaciados, se construye una tabla (A.11) que simplifica grandemente los cálculos para derivar ecuaciones de regresiones lineales, cuadráticas, cúbicas y cuárticas o para separar sumas de cuadrados de tratamientos en componentes de tendencia.

La tabla consta de tres partes por debajo de cada número de observaciones desde 3 hasta 25: los coeficientes c, los divisores y los valores de K.

Los valores de P se obtienen a partir de la ecuación $P_i = \Sigma c_i T_i$.

Después de que los valores P se obtienen a partir de las observaciones, las ecuaciones de regresión lineal, cuadrática, cúbica y cuártica pueden obtenerse de las siguientes ecuaciones:

$$\hat{Y}_L = \overline{Y} + (K_2 P_1)X'$$

$$\hat{Y}_Q = (\overline{Y} - K_1 P_2) + (K_2 P_1)X' + (K_4 P_2)X'^2$$

$$\hat{Y}_C = (\overline{Y} - K_1 P_2) + (K_2 P_1 - K_3 P_3)X' + (K_4 P_2)X'^2 + (K_5 P_3)X'^3$$

$$\hat{Y}_4 = (\overline{Y} - K_1 P_2 + K_8 P_4) + (K_2 P_1 - K_3 P_3)X' + (K_4 P_2 - K_7 P_4)X'^2 + (K_5 P_3)X'^3 + (K_6 P_4)X'^4$$

Los valores de X' en las ecuaciones de regresión son **valores codificados de X,** iguales a los coeficientes c_i

Las sumas de cuadrados para los tratamientos pueden separarse en:

SC lineal'SC = P_1^2/(divisor veces número de repeticiones)

SC cuadrática SC = P_2^2/(divisor veces número de repeticiones)

SC cúbica SC = P_3^2/(divisor veces número de repeticiones)

SC cuártica SC = P_4^2/(divisor veces número de repeticiones)

SC residual = SC del tratamiento — SC lineal — SC cuadrática — SC cúbica — SC cuártica.

La tabla A.12 presenta conjuntos de coeficientes para calcular curvas periódicas para datos igualmente espaciados a través de un ciclo de tiempo.

La tabla contiene dos conjuntos de coeficientes, denominados U y V, para cada uno de los primeros cuatro grados (armónicos) para valores seleccionados de n.

Se calculan los valores de P para cada grado de ajuste, a partir de las ecuaciones

$$PU_i = \Sigma U_i Y \text{ y } PV_i = \Sigma V_i Y.$$

Después de que son determinados los valores de P, puede plantearse directamente una ecuación de cualquier grado deseado hasta el cuarto, a partir de la siguiente ecuación:

$$Y = \overline{Y} + (2PU_1/n)\cos CX + (2PV_1/n)\,\text{sen}\,CX + \ldots + (2PU_i/n)\cos i\,CX + (2PV_i/n)\,\text{sen}\,i\,CX$$

donde X es el número de unidades de tiempo a partir del comienzo de un ciclo, y C es la longitud de cada unidad en grados.

La suma de cuadrados para cualquier grado tiene 2 grados de libertad y se encuentra a partir de la relación

SC para el i-ésimo grado = $2(PU_i^2 + PV_i^2)/n$

y la suma de cuadrados para las desviaciones de la curva puede obtenerse por sustracción de aquellos componentes de la regresión de la SC total.

Los métodos de este capítulo sólo son aplicables cuando los valores de X se encuentran **igualmente espaciados.**

Capítulo **16**

Correlación y regresión para más de dos variables

Hasta el momento hemos estudiado solamente las relaciones entre dos variables. Frecuentemente, estamos interesados en la relación entre una variable dependiente y más de una variable independiente. La ley de oferta y demanda, por ejemplo, implica una relación entre precio (la variable dependiente) y dos variables: oferta y demanda. En ganadería, podemos estar interesados en el aumento de peso en relación con los diversos componentes de la alimentación. Tratándose de cultivos, podemos desear estudiar el efecto sobre la producción cuando los niveles N, P y K varían.

COEFICIENTES DE CORRELACIÓN

La correlación entre dos variables, pasando por alto cualesquiera otras variables que puedan variar simultáneamente, recibe el nombre de **correlación simple o total**. La correlación entre dos variables, cuando una o más variables permanecen fijas a un nivel constante, se denomina **correlación parcial.** La relación combinada entre una variable y dos o más variables que varían simultáneamente recibe el nombre de **correlación múltiple.**

Supóngase que tenemos una variable dependiente, Y, y para cada valor de Y existen valores correspondientes de otras dos variables X_1 y X_2.

La correlación simple o total entre Y y X_1 es el coeficiente de correlación lineal que estudiamos en el capítulo 13. Recuérdese que su fórmula es: *

$$r^2 = (\Sigma xy)^2/\Sigma x^2 \Sigma y^2$$

Para mostrar claramente que ésta es la correlación simple de Y con X_1, es usual incluir subíndices explicativos, de modo que escribimos la fórmula de la siguiente manera:

$r^2_{Y X_1} = (\Sigma x_1 y)^2/\Sigma x_1^2 \Sigma y^2$ Análogamente, la correlación simple entre Y y X_2 se escribe:

$r^2_{Y X_2} = (\Sigma x_2 y)^2/\Sigma x_2^2 \Sigma y^2$ Finalmente, para calcular la correlación **parcial** y **múltiple**, necesitamos una tercera correlación **simple**; aquella entre X_1 y X_2: $r^2_{X_1 X_2} = (\Sigma x_1 x_2)^2/\Sigma x_1^2 \Sigma x_2^2$

*Como antes, las fórmulas se expresan en términos de r^2 en vez de r. Debe recordarse que r, el coeficiente de correlación, es la raíz cuadrada de r^2.

La correlación parcial entre Y y X_1 con un X_2 fijo se denota $r_{YX_1 \cdot X_2}$, y se calcula a partir de las correlaciones simples de la siguiente manera:

$$r^2_{YX_1 \cdot X^2} = \frac{\left(r_{YX_1} - r_{YX_2} r_{X_1X_2}\right)^2}{\left(1 - r^2_{YX_2}\right)\left(1 - r^2_{X_1X_2}\right)}$$

. Análogamente, $r^2_{YX_2 \cdot X_1} = \frac{\left(r_{YX_2} - r_{YX_1} r_{X_1X_2}\right)^2}{\left(1 - r^2_{YX_1}\right)\left(1 - r^2_{X_1X_2}\right)}$

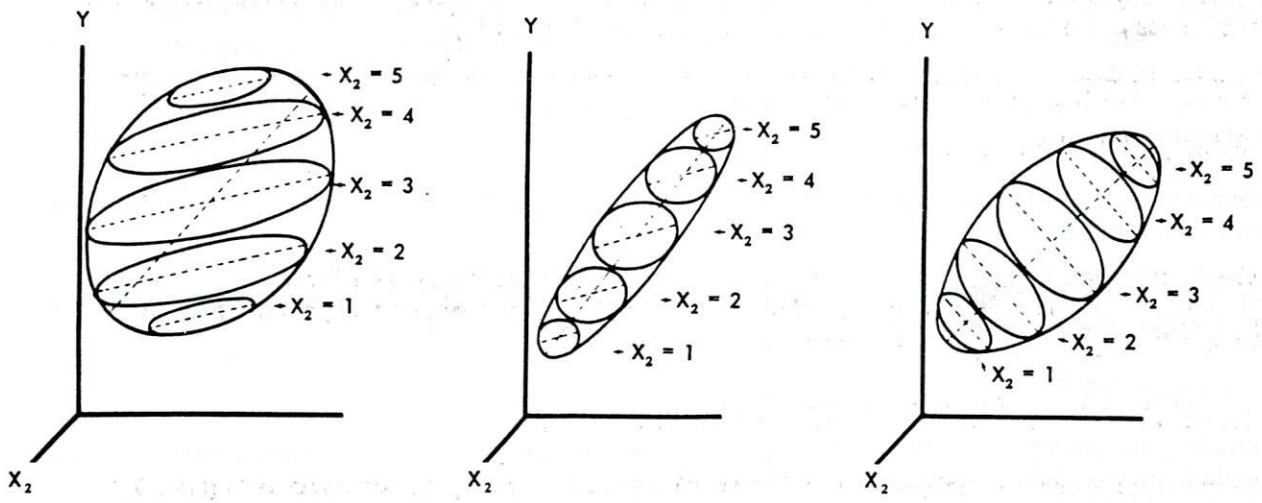
El coeficiente de correlación múltiple, denotado $R_{Y \cdot X_1X_2}$, mide la correlación combinada de X_1 y X_2 con Y. Esto se determina al obtener la raíz cuadrada de:

$$R^2_{Y \cdot X_1X_2} = \left(r^2_{YX_1} + r^2_{YX_2} - 2r_{YX_1} r_{YX_2} r_{X_1X_2}\right) \Big/ \left(1 - r^2_{X_1X_2}\right)$$

Nótese como la adición de sólo una variable más ha incrementado la complejidad de la correlación. Con dos variables, X y Y, sólo teníamos un coeficiente de correlación. Con tres variables, X_1, X_2 y Y, tenemos tres coeficientes simples, tres coeficientes parciales y el coeficiente múltiple.

El problema de visualizar una relación de tres variables es también mucho más difícil que cuando se tienen dos variables. En el caso de dos variables podemos representar las observaciones sobre un gráfico bidimensional. La relación se describe por una recta de regresión y, con muchas observaciones, el diagrama de dispersión de los puntos aparece como una elipse. Cuanto más estrecha sea la elipse, mayor será la correlación. Con tres variables, la relación debe describirse como un plano en el espacio tridimensional. La dispersión de los puntos alrededor de este plano tendrá la forma de una elipsoide. La proyección del elipsoide sobre el plano X_1Y muestra la correlación simple de X_1 y Y. Una sección a través de la elipsoide paralela al plano X_1Y proyectada sobre el mismo, mostrará la correlación parcial de X_1 con X_2 fija, denotada $r_{YX_1 \cdot X_2}$

En la figura 16.1 se muestran gráficamente diversas situaciones. Nótese que la correlación simple puede ser baja, pero la correlación parcial puede ser alta, o viceversa. Pueden incluso ser diferentes en signo.



Parcial positiva alta
Total positiva baja

Parcial positiva baja
Total positiva alta

Parcial negativa media
Total positiva media

Figura 16.1. Diagrama de diversas combinaciones de correlaciones parcial y total, incluyendo tres variables.

El coeficiente de correlación múltiple, R, muestra cuán estrechamente los puntos de la elipsoide están agrupados alrededor del plano de regresión. El valor de R es siempre positivo, fluctuando entre cero y uno; además, éste es siempre cuando menos tan grande como el mayor de los coeficientes simples y parciales. Este hecho sirve como una buena comprobación de los cálculos.

COEFICIENTES DE REGRESIÓN

Hasta el momento, sólo hemos hablado acerca de correlaciones, la estrechez de las relaciones entre las variables. Deseamos también conocer la naturaleza de las relaciones. ¿Qué cambio de Y está asociado con cambios unitarios de las variables independientes? Para contestar esta pregunta, necesitamos una ecuación de la forma

$$\hat{Y} = a + b_1 X_1 + b_2 X_2.$$

Los términos b_1 y b_2 reciben el nombre de **coeficientes de regresión parcial.** La ecuación mejor ajustada de esta forma será aquella que haga mínima la suma de cuadrados de las desviaciones de las Y observadas y de las Y estimadas. Para encontrar los valores de a, b_1 y b_2 que cumplirán este requerimiento, resolvemos **ecuaciones normales** muy similares a aquellas que resolvimos para la regresión curvilínea:

$$an + b_1 \Sigma X_1 + b_2 \Sigma X_2 + \ldots = \Sigma Y$$

$$a\Sigma X_1 + b_1 \Sigma X_1^2 + b_2 \Sigma X_1 X_2 + \ldots = \Sigma X_1 Y$$

$$a\Sigma X_2 + b_1 \Sigma X_2 X_1 + b_2 \Sigma X_2^2 + \ldots = \Sigma X_2 Y$$

$$\ldots$$

Los puntos indican cómo pueden ampliarse estas ecuaciones para incluir a más de tres variables.

Los cálculos pueden reducirse replanteando la ecuación en términos de las desviaciones de las medias, en vez de hacerlo en términos de los valores originales. Puesto que la suma de las desviaciones de cualquier variable de su media es igual a cero, $\Sigma x_1 = \Sigma x_2 = \Sigma y = 0$. Por tanto, la primera ecuación normal se elimina, como lo hacen todos los primeros términos de las ecuaciones restantes. Hemos dejado:

$$b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 + \ldots = \Sigma x_1 y$$

$$b_1 \Sigma x_2 x_1 + b_2 \Sigma x_2^2 + \ldots = \Sigma x_2 y$$

$$\ldots$$

Resolviendo estas ecuaciones para las b obtenemos una ecuación de regresión de la forma $\hat{y} = b_1 x_1 + b_2 x_2 + \ldots$. Si deseamos una ecuación en términos de las observaciones originales, podemos calcular: $a = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 \ldots$. Entonces $\hat{Y} = a + b_1 X_1 + b_2 X_2 + \ldots$

UN EJEMPLO CON TRES VARIABLES

Para ilustrar la correlación parcial y múltiple y la regresión, analizaremos algunos datos sobre la gravedad específica de papas (Y), el contenido de nitrógeno (X_1) y el contenido de fósforo (X_2). Se codificarán las observaciones, a fin de simplificar los cálculos. (Véase la tabla 16.1.)

Primero calcularemos los diversos coeficientes de correlación. Las correlaciones simples o totales son:

$$r^2_{Y X_1} = (\Sigma y x_1)^2 / \Sigma y^2 \Sigma x_1^2 = (-29\ 218.35)^2 / (51\ 172.95)(21\ 240.55) = 0.7854$$

$$r_{Y X_1} = \sqrt{r^2_{Y X_1}} = -.8862 \text{ (Nótese que ésta es negativa porque } \Sigma y x_1 \text{ fue negativa)}$$

$$r^2_{Y X_2} = (\Sigma y x_2)^2 / \Sigma y^2 \Sigma x_2^2 = (-\ 6\ 611.8)^2 / (51\ 172.95)(1\ 663.2) = 0.5136$$

$$r_{Y X_2} = \sqrt{r^2_{Y X_2}} = -.7167$$

$$r^2_{X_1 X_2} = (\Sigma x_1 x_2)^2 / \Sigma x_1^2 \Sigma x_2^2 = (2\ 584.4)^2/(21\ 240.55)(1\ 663.2) = .1891$$

$$r_{X_1 X_2} = \sqrt{r^2_{X_1 X_2}} = .4348$$

Tabla 16.1. Gravedad específica, contenidos de nitrógeno y fósforo de 20 muestras de papas.

Y (Gr. Es. − 1.07)10⁴	X_1 (Nitrógeno − 1)100	X_2 (Fósforo)100
2	96	40
14	82	36
15	121	30
15	88	42
16	100	28
27	114	26
48	71	33
54	94	26
58	74	15
68	36	35
82	36	25
83	73	15
91	58	26
97	31	25
98	38	24
101	56	11
128	24	22
140	37	11
163	10	24
179	14	10
Totales 1 479	1 253	504
$\Sigma Y^2 = 160\ 545$	$\Sigma X_1^2 = 99\ 741$	$\Sigma X_2^2 = 14\ 364$
$(\Sigma Y)^2/20 = 109\ 372.05$	$(\Sigma X_1)^2/20 = 78\ 500.45$	$(\Sigma X_2)^2/20 = 12\ 700.8$
$\Sigma y^2 = 51\ 172.95$	$\Sigma x_1^2 = 21\ 240.55$	$\Sigma x_2^2 = 1\ 663.2$
$\Sigma Y X_1 = 63\ 441$	$\Sigma Y X_2 = 30\ 659$	$\Sigma X_1 X_2 = 34\ 160$
$\Sigma Y \Sigma X_1/20 = 92\ 659.35$	$\Sigma Y \Sigma X_2/20 = 37\ 270.8$	$\Sigma X_1 \Sigma X_2\ 20 = 31\ 575.6$
$\Sigma y x_1 = -29\ 218.35$	$\Sigma y x_2 = -6\ 611.8$	$\Sigma x_1 x_2 = 2\ 584.4$

Los coeficientes de correlación parcial son:

$$r^2_{Yx_1 \cdot x_2} = (r_{Yx_1} - r_{Yx_2} r_{x_1x_2})^2 \Big/ (1 - r^2_{Yx_2})(1 - r^2_{x_1x_2}) = \frac{[-0.8862 - (-0.7167)(0.4338)]^2}{(1 - 0.5136)(1 - 0.1891)}$$

$$= \frac{(-0.5746)^2}{(0.4864)(0.8109)} = 0.8371$$

$$r_{Yx_1 \cdot x_2} = \sqrt{r^2_{Yx_1 \cdot x_2}} = -0.9149$$

$$r^2_{Yx_2 \cdot x_1} = (r_{Yx_2} - r_{Yx_1} r_{x_1x_2})^2 \Big/ (1 - r^2_{Yx_1})(1 - r^2_{x_1x_2}) = \frac{[-0.7167 - (-0.8862)(0.4348)]^2}{(1 - 0.7854)(1 - 0.1891)}$$

$$= \frac{(-0.3314)^2}{(0.2146)(0.8109)} = 0.6310$$

$$r_{Yx_2 \cdot x_1} = \sqrt{r^2_{Yx_2 \cdot x_1}} = -0.7944$$

Finalmente, calculamos R, el coeficiente múltiple:

$$R^2_{Y \cdot x_1x_2} = (r^2_{Yx_2} + r^2_{Yx_1} - 2r_{Yx_1} r_{Yx_2} r_{x_1x_2}) \Big/ (1 - r^2_{x_1x_2})$$

$$= [0.5136 + 0.7854 - 2(-0.8862)(-0.7167)(0.434)]/(1 - 0.1891)$$

$$= 0.7467/0.8109 = 0.9208$$

$$R_{Y \cdot x_1x_2} = \sqrt{0.9208} = 0.9596$$

Las correlaciones simples de cada contenido de nitrógeno o fósforo por separado con la gravedad específica no son muy grandes; pero cuando las dos variables se consideran simultáneamente, la relación con la gravedad específica es muy estrecha.

Expresados en cifras de porcentaje, el nitrógeno sólo explica el 78.54% de la variabilidad de la gravedad específica, $\left(100 \times r^2_{Yx_1}\right)$

El fósforo explica el 51.36%. El nitrógeno y el fósforo conjuntamente explican el 92.08%.

Ahora necesitamos describir la relación mediante el cálculo de la ecuación de regresión. Utilizando las ecuaciones normales basadas en las desviaciones de las medias, obtenemos:

$$b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 = \Sigma x_1 y$$

$$b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2 = \Sigma x_2 y$$

Sustituyendo los valores observados de los datos:

$$21\ 240.55 b_1 + 2\ 584.4 b_2 = -29\ 218.35$$

$$2\ 584.40 b_1 + 1\ 663.2 b_2 = -6\ 661.8$$

Multiplicando la primera ecuación por 2 584.4, la segunda ecuación por 21 240.55 y sustrayendo, obtenemos:

$$28\ 648\ 159.4b_2 = -64\ 926\ 364.75$$

$$b_2 = -2.266$$

Sustituyendo este valor de b_2 en cualquiera de las ecuaciones originales, y resolviendo para b_1, encontramos:

$$b_1 = -1.100$$

Para expresar una ecuación de regresión en términos de los valores originales, necesitamos determinar el *a:*

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

$$= \frac{1\,479}{20} - \left(-1.100\ \frac{1\,253}{20}\right) - \left(-2.266\ \frac{504}{20}\right) = 199.968$$

Podemos ahora plantear la ecuación de regresión como sigue: $\hat{Y} = 199.968 - 1.100X_1 - 2.266X_2$

A partir de esta ecuación podemos calcular los valores de $\hat{Y}$ y compararlos con los valores observados.

Tabla 16.2. Gravedad específica observada y calculada de 20 muestras de papas.

Y	$\hat{Y}$	$d = Y - \hat{Y}$
2	3.7	−1.7
14	28.2	−14.2
15	−1.1	16.1
15	8.0	7.0
16	26.5	−10.5
27	15.7	11.3
48	47.1	0.9
54	37.7	16.3
58	84.6	−26.6
68	81.1	−13.1
82	103.7	−21.7
83	85.7	−2.7
91	77.2	13.8
97	109.2	−12.2
98	103.8	−5.8
101	113.4	−12.4
128	123.7	4.3
140	134.3	5.7
163	134.6	28.4
179	161.9	17.1

La suma de las desviaciones es igual a cero, como debería ser. Esto suministra una buena combinación de los cálculos. La suma de cuadrados de las desviaciones es 4 051.16. Esto representa la variación de la gravedad específica (Y), no asociada con la variación del contenido de nitrógeno (X_1) o del contenido de fósforo (X_2). Esto puede calcularse sin determinar cada Y, tomando $(1 - R^2)\Sigma y^2$, que es:

$$(1 - 0.9208)\ 51\ 172.95 = 4\ 052.90$$

Las dos respuestas coinciden estrechamente, debiéndose la pequeña diferencia, a la aproximación.

Los resultados que hemos obtenido pueden resumirse en una tabla de análisis de varianza como la siguiente:

Fuente de variación	Método para calcular la SC	SC	gl	CM	F
Total	Σy^2	51 172.95	19		
Regresión debida a X_1	$r^2_{Y X_1}(\Sigma y^2)$	40 191.23	1	40 191.23	65.9**
Desviación de la regresión simple	$\left(1 - r^2_{Y X_1}\right)\Sigma y^2$	10 981.72	18	610.10	
Regresión adicional debida a X_2	$r^2_{Y X_2 . X_1}\left(1 - r^2_{Y X_1}\right)\Sigma y^2$	6 929.47	1	6 929.47	29.07**
Desviación de la regresión múltiple	$\left(1 - R^2_{Y . X_1 X_2}\right)\Sigma y^2$	4 052.90	17	238.41	

La última suma de cuadrados puede obtenerse por sustracción: 10 981.72 – 6 929.47 = 4 052.25. La discrepancia entre este valor y el que aparece en la tabla se debe a la aproximación y no ejercerá un efecto importante sobre el valor de F. La raíz cuadrada de 238.41 o 15.44 recibe el nombre de **error estándar de la estimación**, y se denota por el símbolo $s_{Y . X_1 X_2}$.

Nueva tabla, que muestra valores de F completamente diferentes:

Fuente de variación	Método para calcular la SC	SC	gl	CM	F
Total	Σy^2	51 172.95	19		
Regresión debida a X_2	$r^2_{Y X_2}(\Sigma y^2)$	26 282.43	1	26 282.43	19.01**
Desviación de la regresión simple	$\left(1 - r^2_{Y X_2}\right)\Sigma y^2$	24 890.52	18	1 382.81	
Regresión adicional debida a X_1	$r^2_{Y X_1 . X_2}\left(1 - r^2_{Y X_2}\right)\Sigma y^2$	20 835.85	1	20 835.85	87.40**
Desviación de la regresión múltiple	$\left(1 - R^2_{Y . X_1 X_2}\right)\Sigma y^2$	4 052.90	17	238.41	

En la primera de estas dos tablas, consideramos el efecto total del nitrógeno y luego el efecto adicional del fósforo. En la segunda tabla, consideramos el efecto total del fósforo y luego el efecto adicional del nitrógeno. El hecho de que el orden en que se considera a las variables da lugar a una marcada diferencia en los resultados del análisis, puede resultar confuso para cualquiera durante esta primera exposición de la regresión múltiple.

Un ejemplo sencillo puede ayudar a aclarar parte de la confusión. Es perfectamente conocido que la producccción de muchos cultivos está influida tanto por la temperatura como por la duración del día. Supóngase que tenemos numerosos registros de producción de un cultivo, que crece en diferentes estaciones del año. Para cada registro de producción, tenemos un registro de la duración media del día y de la temperatura media durante la estación de crecimiento. Esperamos que la duración del día y la temperatura

estén estrechamente correlacionadas entre sí. Dado que esto es cierto, no deberíamos sorprendernos si encontráramos que la producción estuvo estrechamente correlacionada con la temperatura, pero que la consideración adicional de la duración del día explicaría poco la variación de la producción aún no explicada. Al mismo tiempo, la duración del día por separado puede estar estrechamente correlacionada con la producción, mientras que la temperatura puede tener un pequeño efecto adicional. La conclusión sería que los días largos y calientes están asociados con la producción más elevada que los días cortos y fríos. Podríamos decir muy poco acerca de qué factor fue el más importante: si la temperatura o la duración del día. Para responder a esta pregunta, necesitaríamos un experimento en que la duración del día y/o la temperatura fuesen controladas de modo que éstas podrían estar menos estrechamente correlacionadas de lo que se encuentran por naturaleza.

En el capítulo 13 presentamos un ejemplo de una correlación espuria entre el consumo de cigarrillos y la producción de heno. Esta alta correlación se debió, en apariencia, al hecho de que ambas variables estuvieron estrechamente relacionadas con una tercera variable: el tiempo. Un análisis de regresión múltiple mostraría una notable diferencia entre los dos análisis, dependiendo de qué variable independiente se considerara primero.

Tabla 16.3. Análisis de regresión múltiple de la producción de heno (Y), consumo de cigarrillos (X_1) y tiempo (X_2).

Fuente de variación	gl.	SC	CM	F
X_1 considerada primero				
Total	14	10 094.00		
Regresión debida a X_1	1	8 855.31	8 855.31	92.94**
Desviación de la regresión simple	13	1 238.69	95.28	
Regresión adicional debida a X_2	1	918.01	918.01	34.35**
Desviación de la regresión múltiple	12	320.67	26.72	
X_2 considerada primero				
Total	14	10 094.00		
Regresión debida a X_2	1	9 723.21	9 723.21	340.90**
Desviación de la regresión simple	13	370.79	28.52	
Regresión adicional debida a X_1	1	50.11	50.11	1.88ns
Desviación de la regresión múltiple	12	320.67	26.72	

En el segundo análisis, donde primero eliminamos la regresión con el tiempo, vemos que no existe una regresión adicional significativa relacionada con el consumo de cigarrillos.

MÁS DE TRES VARIABLES

Por el bien de la simplicidad, la mayor parte de nuestro estudio y los ejemplos ilustrativos han estado basados en tres variables: una dependiente y dos independientes. En realidad, los coeficientes de correlación múltiple y parcial y las ecuaciones de regresión pueden calcularse para cualquier número de variables. Un reciente estudio en la Universidad de California incluyó 35 variables. Solamente podemos indicar aquí, en forma

general, que los métodos descritos pueden extenderse a más de tres variables, y puntualizar algunas de las dificultades contempladas.

Hemos visto cómo las ecuaciones normales para calcular los coeficientes de regresión b_1, b_2, etc., pueden ampliarse para incluir tantas variables como se deseen. Cada nueva variable requiere solamente la adición de un nuevo término en el miembro izquierdo de cada ecuación, y la adición de una nueva ecuación siguiendo el mismo patrón de las anteriores. Para m variables, la última ecuación normal sería:

$$b_1\Sigma x_1 x_m + b_2\Sigma x_2 x_m + b_3\Sigma x_3 x_m + \ldots + b_m\Sigma x^2_m = \Sigma x_m y$$

El álgebra no cambia, pero la aritmética implicada en la resolución de las ecuaciones se vuelve más difícil cuando añadimos nuevas variables. Por esta razón, se sugiere que se utilice uno de los procedimientos sistemáticos mencionados en el capítulo anterior o, si es posible que se emplee una computadora electrónica.

Hemos observado cómo, con solamente dos variables existió un solo coeficiente, pero con tres variables existieron siete, incluyendo un coeficiente múltiple, tres simples y tres parciales. Con cuatro variables, el total se incrementa a 25, y con 5 a 81. Una de las razones de este gran incremento es que tenemos la adición de **coeficientes parciales de orden elevado.** El **orden** de un coeficiente de correlación parcial es el número de variables que permanecen fijas. Con tres variables, sólo tuvimos coeficientes parciales de primer orden, como $r_{Yx_1 \cdot x_2}$. Con cuatro variables, tenemos coeficientes simples y parciales de primer orden y coeficientes parciales de segundo orden, como $r_{Yx_1 \cdot x_2 x_3}$, es decir, "la correlación de Y y X_1 para valores fijos de X_2 y X_3."

Existe una ecuación general que permite calcular un coeficiente de correlación parcial de cualquier orden, si conocemos tres coeficientes parciales de un orden inferior:

$$r^2_{Yx_1 \cdot x_2 x_3 \ldots x_m} = \frac{\left(r_{Yx_1 \cdot x_3 \ldots x_m} - r_{Yx_2 \cdot x_3 \ldots x_m} r_{x_1 x_2 \cdot x_3 \ldots x_m}\right)^2}{\left(1 - r^2_{Yx_2 \cdot x_3 \ldots x_m}\right)\left(1 - r^2_{x_1 x_2 \cdot x_3 \ldots x_m}\right)}$$

Las ecuaciones dadas para encontrar los coeficientes parciales de primer orden que incluyen tres variables de las tres correlaciones simples, fueron casos especiales de esta ecuación general.

Una ecuación general para encontrar el coeficiente de correlación múltiple que incluye m variables independientes es:

$$1 - R^2_{Y \cdot x_1 \ldots x_m} = \left(1 - r^2_{Yx_1}\right)\left(1 - r^2_{Yx_2 \cdot x_1}\right)\left(1 - r^2_{Yx_3 \cdot x_1 x_2}\right) \cdots \left(1 - r^2_{Yx_m \cdot x_1 \ldots x_{m-1}}\right)$$

En el caso de dos variables independientes, se reduce a la forma bastante simple ya presentada para $R^2_{Y \cdot x_1 x_2}$.

Hemos visto que la aritmética se vuelve cada vez más difícil cuando consideramos un mayor número de variables; pero quizá la mayor dificultad encontrada cuando consideramos más de tres variables consiste en visualizar las relaciones. La relación entre dos variables puede trazarse en un gráfico bidimensional. Las relaciones entre tres variables pueden representarse en un diagrama tridimensional; no obstante, ¿cómo trazamos la representación de las relaciones entre cuatro o más variables? La respuesta es que ni siquiera lo intentamos. Tenemos que aprender a no incomodarnos por ser incapaces de visualizar relaciones que incluyen cuatro o más dimensiones. En su lugar, necesitamos pensar en términos de las ecuaciones, antes que en los diagramas. Después de todo, no tenemos problemas en captar la idea de que la producción de un cultivo está relacionada con los niveles de N, de P y de K del suelo, la cantidad de agua aplicada, la competencia de la maleza, la cantidad de enfermedad, el número de insectos perjudiciales, la temperatura y la duración del día. Con datos suficientes, incluso podemos plantear una ecuación relativamente sencilla que describa dichas relaciones. ¿Deberíamos estar preocupados si no pudiésemos trazar un esquema descriptivo de esta compleja interacción de factores? Una ecuación puede ser mejor que mil dibujos.

Algo más debemos decir acerca de la correlación y de la regresión, que incluyen un gran número de variables.

Mostramos que con tres variables, pueden realizarse dos análisis diferentes, dependiendo de cuál de las variables independientes consideremos primero. Con tres variables independientes, el número de análisis posibles se eleva a seis, y con m variables independientes existen m! formas posibles de ordenar las variables. (El símbolo "m!" representa **m factorial** y significa el producto de todos los números desde 1 hasta m. Luego, $10! = 1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8 \times 9 \times 10 = 3\,628\,800$.) ¿Cuál es el **mejor** orden para considerar a las variables? Una pregunta relacionada es: "Entre un número de variables independientes, ¿cómo podemos encontrar el mejor conjunto de un tamaño dado?" Hallar un método directo simple para obtener el mejor conjunto es uno de los mayores problemas no resueltos de la estadística. En las computadoras electrónicas se encuentran disponibles programas para alcanzar esta solución, pero el tiempo es el factor limitante.

RESUMEN

Cuando consideramos más de dos variables, existen tres tipos de coeficientes de correlación.

La **correlación simple o total** es la correlación lineal entre cualquier par de variables, sin importar los valores de las variables restantes.

La correlación **parcial** es la relación entre dos variables cuando una o más de las variables restantes se mantienen constantes.

La correlación **múltiple** es la relación conjunta entre la variable dependiente y todas las variables independientes.

La ecuación para el coeficiente de correlación simple al cuadrado es:

$$r^2_{Yx_i} = \left(\Sigma x_i y\right)^2 \Big/ \Sigma x^2_i \, \Sigma y^2$$

La ecuación general para un coeficiente de correlación parcial de primer orden al cuadrado es:

$$r^2_{Yx_i \cdot x_j} = \frac{\left(r_{Yx_i} - r_{Yx_j} r_{x_i x_j}\right)^2}{\left(1 - r^2_{Yx_j}\right)\left(1 - r^2_{x_i x_j}\right)}$$

El **orden** de un coeficiente de correlación parcial es el número de variables mantenidas constantes, expresadas simbólicamente por el número de subíndices siguientes al punto. Con tres variables, sólo podemos obtener coeficientes parciales de primer orden.

El **coeficiente de correlación múltiple** entre tres variables se encuentra a partir de:

$$R^2_{Y \cdot x_1 x_2} = \left(r^2_{Yx_1} + r^2_{Yx_2} - 2r_{Yx_1} r_{Yx_2} r_{x_1 x_2}\right) \Big/ \left(1 - r^2_{x_1 x_2}\right)$$

El coeficiente de correlación múltiple es siempre positivo y al menos tan grande como los mayores coeficientes simple y parcial.

Una **ecuación de regresión** describe la relación entre la variable dependiente y todas las variables independientes. Esta es de la forma:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \ldots\ldots$$

Los símbolos b_1, b_2, etc., reciben el nombre de **coeficientes de regresión parcial.** Para encontrar la ecuación de regresión que se ajuste mejor a los datos observados, resolvemos la siguientes **ecuaciones normales** para los coeficientes de regresión parcial:

$$b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 + \ldots + b_m \Sigma x_1 x_m = \Sigma x_1 y$$
$$b_1 \Sigma x_2 x_1 + b_2 \Sigma x_2^2 + \ldots + b_m \Sigma x_2 x_m = \Sigma x_2 y$$
$$\ldots$$
$$b_1 \Sigma x_m x_1 + b_2 \Sigma x_m x_2 + \ldots + b_m \Sigma x_m^2 = \Sigma x_m y$$

donde m es el número de variables independientes. Para su resolución, necesitamos m ecuaciones con m términos en el miembro izquierdo de cada una de ellas.

Capítulo 17

Análisis de conteos

La mayor parte de lo presentado en este libro ha tratado acerca del análisis de mediciones, como el peso, la producción, la altura, etc.; sin embargo, no siempre medimos alguna característica de un individuo. Algunas veces, simplemente clasificamos individuos en dos o más grupos, como muertos o vivos; saludables o enfermos; macho o hembra; rojo, rosado o blanco; novato, estudiante; joven o adulto. Incluso con las características que pueden medirse, algunas veces resulta más conveniente clasificar a los individuos en amplios grupos; por ejemplo, podemos desear conducir un estudio que incluya una medida de los ingresos de las personas. Muchas personas de nuestra muestra pueden molestarse al interrogarles sobre la cantidad exacta de sus ingresos; pero no vacilarían si se les preguntara a cuál de las tres o cuatro categorías de ingresos pertenecen, y dicha clasificación puede ser suficiente para los propósitos de nuestro estudio.

Los datos basados en cómputos de individuos que pertenecen a cada una de diversas clases, generalmente requiere un tipo diferente de análisis estadístico que el comúnmente utilizado para las mediciones. Tomemos, por ejemplo, un estudio para determinar algo acerca de las características de los huevos puestos por un hato de gallinas. Podemos pesar cada huevo de una muestra y determinar que la media o el peso promedio por huevo fue, digamos, 21 gramos. Podríamos también clasificar cada huevo como cuarteado o sano y encontrar que el 5% de los huevos estaban cuarteados. No tendría sentido afirmar que el promedio de huevos estuvo en un 5% de cuarteados. Nuestro promedio se aplica a la proporción de individuos en la muestra que poseen esta característica.

En el capítulo sobre transformaciones, mostramos cómo los datos basados en cómputos pueden ser algunas veces transformados y analizados válidamente como si se tratara de datos de medición. En este capítulo describiremos un método denominado **ji cuadrada** (representado por el símbolo χ^2) para analizar los datos de enumeración.

Antes de estudiar este método, debemos primero considerar qué nos gustaría aprender al clasificar y computar a los individuos. Los propósitos de recolectar tales datos generalmente incluyen uno o más de los siguientes objetivos: a) probar una o más hipótesis no sugeridas por los datos, b) determinar si diferentes características están interrelacionadas, y c) probar si las muestras se obtuvieron a partir de distintas poblaciones.

JI CUADRADA

La fórmula general de ji-cuadrada utilizada en la resolución de todos estos problemas es: $\chi^2 = \Sigma \frac{(Ob - Ex)^2}{Ex}$

donde Ob es el valor observado para cada una de dos o más clases, y Ex es el valor esperado correspondiente.

Para evaluar esta expresión, primero debemos determinar el valor esperado para cada clase de individuos, de acuerdo con nuestra hipótesis. El valor esperado se sustrae entonces del valor observado y la diferencia resultante se eleva al cuadrado y se divide entre el valor esperado. Estos cocientes se suman para todas las clases. Entonces, la suma se compara con los valores de una tabla de χ^2 con los grados de libertad apropiados. Esto revela la probabilidad aproximada de obtener desviaciones de las expectativas, tan grandes o mayores que aquellas observadas por casualidad.

La aritmética es bastante simple, y para ciertos casos especiales se encuentran disponibles métodos de cálculo abreviados: sin embargo, necesitamos considerar diversos aspectos para utilizar apropiadamente las pruebas de ji cuadrada:

1. Debemos tener cuidado al seleccionar la hipótesis que vamos a probar. Esta hipótesis debería ser razonable y estar basada en hechos o principios previamente conocidos.

2. Necesitamos estar conscientes de que la ji cuadrada es una distribución continua y se encuentra, de hecho, relacionada con la **distribución normal**. Por otro lado, la distribución de las muestras basadas en cómputos es una distribución discreta o discontinua. Si se clasifica a los individuos en una de dos clases, estamos tratando con lo que se denomina una **distribución binomial**. Las distribuciones normal y binomial son similares, pero no idénticas; de ahí que haya sido anteriormente establecido que la referencia a la tabla de ji cuadrada da una probabilidad **aproximada**. Necesitamos saber qué situaciones resultan en aproximaciones deficientes, de modo que podamos evadir estas situaciones o quizá hacer ajustes para obtener aproximaciones más apegadas a la probabilidad real.

3. Dada una hipótesis, necesitamos saber cómo calcular correctamente los valores esperados para cada clase.

4. El número de grados de libertad para completar la tabla de ji cuadrada no es siempre obvio. Necesitamos aprender ciertas reglas para su determinación.

5. Al interpretar los resultados de una prueba de ji cuadrada debemos tener precaución y buen juicio. Incluso cuando nuestras observaciones no difieran significativamente de nuestra hipótesis, podemos no estar justificados para aceptar la hipótesis si los datos se ajustan también a otra hipótesis igualmente lógica.

Ilustremos estos diversos aspectos con un ejemplo. Supóngase que estamos trabajando con alguna planta que presenta influorescencias rojas y blancas. Hemos cruzado plantas de línea de raza de las dos formas y la generación de F_1 fue toda roja. Desarrollamos una generación de F_2 de ocho plantas y encontramos que cuatro registraron influorescencias rojas y cuatro blancas. Con base en lo que hemos aprendido hasta el momento, nos sentimos bastante seguros de que el rojo es dominante sobre el blanco, y además suponemos que esto está determinado por un gene en particular. Nuestro conocimiento de genética nos conduce a formular la hipótesis de que los F_2 se segregarán a una proporción de 3:1 de rojas a blancas.

En base a esta hipótesis esperamos obtener, entre ocho plantas, seis rojas y dos blancas, de modo que nuestros números observados se desviarán en dos de lo esperado. Nos preguntamos: "¿Cuál es la probabilidad de que hubiéramos obtenido una desviación de lo esperado tan grande o mayor de la que observamos sólo por casualidad?" Si esta probabilidad es muy pequeña, rechazaremos nuestra hipótesis.

Reconociendo que la ji cuadrada dará sólo una aproximación de la probabilidad deseada, calcularemos la probabilidad exacta basada en la distribución binomial. Para hacerlo, necesitamos encontrar la probabilidad de cada resultado posible y combinar todos los casos que sean iguales o excedan las desviaciones observadas de la esperada.

Primero debemos definir algunos símbolos. Denominamos a la proporción hipotética $r_1:r_2$. La probabilidad de un individuo perteneciente a la primera clase recibe el nombre de p y es igual a $r_1/(r_1 + r_2)$. La probabilidad de pertenecer a la segunda clase se denomina q y es igual a $r_2/(r_1 + r_2)$ o $1 - p$. Los números observados en cada clase reciben el nombre de n_1 y n_2, y $n_1 + n_2 = n$, el número total de la muestra. El símbolo n! se denomina **n factorial** y se calcula al obtener el producto de todos los números enteros desde 1 hasta n. Cero factorial es igual a uno, por definición.

En una distribución binomial, la probabilidad de obtener una muestra con n_1 en la primera clase y n_2 en la segunda es:

$$p^{n_1}q^{n_2}n!/n_1!n_2!$$

En nuestro ejemplo $r_1 = 3$, $r_2 = 1$, $p = r_1/(r_1 + r_2) = \frac{3}{4}$, $q = r_2/(r_1 + r_2) = \frac{1}{4}$.

La probabilidad de obtener una muestra, en la cual $n_1 = 4$ y $n_2 = 4$ es:

$$p^{n_1}q^{n_2}n!/n_1!n_2! = (\tfrac{3}{4})^4 \cdot (\tfrac{1}{4})^4 \cdot 1.2.3.4.5.6.7.8/1.2.3.4.1.2.3.4 = 81/256 \times 1/256 \times 70 = .0865.$$

Análogamente, podemos calcular la probabilidad para cada uno de los demás resultados y construir las tres primeras columnas de la tabla 17.1.

Tabla 17.1

Resultado	Probabilidad	Desviación de n_1, de $(n_1 - 6)$ esperado	Intervalo de clase	Probabilidad basada en una curva normal
8:0	0.1001	2	> 1.5	0.1104
7:1	0.2670	1	0.5 to 1.5	0.2312
6:2	0.3115	0	− 0.5 to 0.5	0.3168
5:3	0.2076	−1	− 0.5 to − 1.5	0.2312
4:4	0.0865	−2	− 1.5 to − 2.5	0.0897
3:5	0.0231	−3	− 2.5 to − 3.5	0.0186
2:6	0.0038	−4	− 3.5 to − 4.5	0.0020
1:7	0.0004	−5	− 4.5 to − 5.5	0.0001
0:8	0.0000	−6	< − 5.5	0.0000
	Total 1.0000			

La última probabilidad no es en realidad cero, pero es menor que 0.00005.

Nótese que la suma de todas las probabilidades es igual a 1, lo que brinda una comprobación de los cálculos.

El valor esperado de n_1 es $np = 8 \times \frac{3}{4} = 6$, de modo que hacemos una tercera columna en la tabla, mostrando las diferencias entre los valores observados de n_1 y sus valores esperados.

Ahora podemos responder a nuestra pregunta original. La probabilidad de obtener una desviación de dos o más de lo esperado es igual a la suma de las probabilidades del primero y de los últimos cinco de los nueve casos de la tabla. Esto es .1001 + .0865 + . . . + .0000 = .2139.

Veamos cómo este resultado se compara con la prueba de ji cuadrada. Nuestra fórmula es:

$$X^2 = \Sigma \frac{(Ob - Ex)^2}{Ex} = \frac{(4 - 6)^2}{6} + \frac{(4 - 2)^2}{2} = \frac{4}{6} + \frac{4}{2} = 0.67 + 2 = 2.67$$

Buscando este valor en la tabla A.6 de ji cuadrada, para un grado de libertad, vemos que nuestra ji cuadrada observada es muy cercana al valor 2.706, encontrado en el punto de 10%, indicando que la probabilidad de obtener una desviación al menos tan grande como la observada por casualidad es de 0.10. (Un valor más preciso obtenido a partir de las tablas más extensas es 0.1025). Este es considerablemente menor que la probabilidad exacta de 0.2139 que encontramos.

Corrección de Yates para la continuidad

Existe una corrección que recibe el nombre de **corrección de Yates para la continuidad,** que reducirá considerablemente la discrepancia entre los dos métodos. Supóngase que estamos utilizando la distribución normal para obtener una estimación de la probabilidad de cada resultado. Para hacerlo, primero debemos encontrar la varianza y la desviación estándar de la distribución. Esto puede hallarse al elevar al cuadrado la desviación de la media (valor esperado) para cada resultado y al multiplicar este resultado por la probabilidad correspondiente. Estos productos se suman para todos los resultados.

Varianza = $2^2 \times .1001 + 1^2 \times .2670 + \ldots + (5)^2 \times .0004 = 1.4997$. Puesto que estamos tratando con una distribución binomial, existe una fórmula mucho más simple para obtener la varianza: $\sigma^2 = npq$. Luego, en este ejemplo:

$$\sigma^2 = 8 \times \tfrac{3}{4} \times \tfrac{1}{4} = 1.5$$

Desviación estándar: $\sigma = \sqrt{\sigma^2} = \sqrt{1.5} = 1.225$

Los intervalos de clase pueden expresarse ahora en términos de valores z, dividiendo los límites de cada intervalo entre la desviación estándar. El área bajo la curva normal para cada intervalo puede entonces determinarse mediante una tabla de funciones de probabilidad encontrada en la mayoría de los libros de tablas matemáticas.

Dichas estimaciones se muestran en la tabla 17.1, para puntualizar cómo difieren las distribuciones normal y binomial. Puesto que la normal es una distribución continua, tenemos que agrupar conjuntamente todas las porciones de la curva normal, desde $n_1 - Ex = -1.5$ hasta $n_1 - Ex = -2.5$ y determinar el área de esta porción con el fin de encontrar una probabilidad de — 2 para $n_1 - Ex$. Análogamente, la probabilidad de 2 para $n_1 - Ex$ es el área bajo la curva normal, desde $n_1 - Ex = 1.5$ hasta el infinito. Entonces, nuestra pregunta acerca de la probabilidad de obtener una desviación de 2 o mayor del valor esperado cuando utilizamos una curva normal, debe replantearse para inquirir cuál es la probabilidad de que la desviación del valor esperado exceda de 1.5. La corrección de Yates tiene en cuenta esto y consiste simplemente en sustraer 0.5 del valor absoluto (sin importar el signo) de las diferencias entre los valores observados y los esperados.

Aplicando esta corrección, calculamos una ji cuadrada ajustada como sigue:

$$X^2 = \Sigma \frac{(|Ob - Ex| - .5)^2}{Ex} = \frac{(2 - .5)^2}{6} + \frac{(2 - .5)^2}{2} = \frac{(1.5)^2}{6} + \frac{(1.5)^2}{2} = .375 + 1.125 = 1.50$$

(**Nota**: el símbolo |x| significa valor absoluto de x)

Buscando este valor en una tabla de ji cuadrada (A.6), encontramos que la probabilidad está entre 0.10 y 0.50, pero es mucho mayor de lo que fue anteriormente. Tablas más extensas arrojan un valor de P de 0.2207, muy cercano a la probabilidad calculada exacta de 0.2139. La probabilidad basada en una distribución normal puede obtenerse también en la misma forma como lo hicimos con la distribución binomial, es decir, añadiendo las probabilidades de la primera línea y las últimas cinco líneas de la tabla 17.1. Esto da 0.2208, que es, como debería ser, igual (considerando errores de aproximación) al resultado obtenido mediante la prueba de ji cuadrada.

GUÍAS PARA LA UTILIZACIÓN DE JI CUADRADA

Hemos visto que incluso con una muestra tan pequeña como 8, no es muy grande la diferencia entre la distribución binomial exacta y la distribución normal sobre la cual la ji cuadrada está basada. Las siguientes reglas ayudarán a saber si la ji cuadrada dará una aproximación lo suficientemente cercana a la respuesta correcta:

1. Cuanto mayor sea el tamaño de la muestra, más estrecha será la coincidencia entre las dos distribuciones.

2. Cuanto mayor sea la proporción entre r_1 y r_2 en nuestra hipótesis, mayor será la discrepancia entre las dos distribuciones para un tamaño de muestra dado. Luego, si hipotetizamos una proporción de 1:1, la coincidencia será estrecha incluso para pequeñas muestras; pero si hipotetizamos una proporción de 15:1, es necesario un tamaño de muestra mucho mayor.

3. Un método práctico consiste en evitar la utilización de ji cuadrada si la menor clase esperada es inferior a 5. Si tenemos más de dos clases, podemos combinar aquellas cuyos valores esperados son menores que 5. También pueden emplearse tamaños de muestra cada vez mayores, a fin de incrementar el tamaño del valor más pequeño esperado.

4. Utilícese siempre la corrección de Yates para determinar la ji cuadrada con un solo grado de libertad. **Nunca** debe emplearse para problemas en los que se incluya más de un grado de libertad.

Los grados de libertad pueden definirse, en general, como el número de clases al que se le puede asignar un valor arbitrario. Luego, si tenemos dos clases, como en el ejemplo que hemos estado utilizando, podemos asignar cualquier valor a n_1, pero n_2 es entonces fijo, ya que debe incluir a los miembros restantes de la muestra, de donde $n_2 = n - n_1$. Por tanto, la ji cuadrada tiene un grado de libertad. **En la comprobación de cualquier hipótesis exterior a los datos, los grados de libertad son siempre uno menos que el número de clases.** Se estudiarán posteriormente otras situaciones.

CÓMO INTERPRETAR LOS RESULTADOS

La interpretación es el último y el más importante paso en nuestro análisis de los datos. Hemos visto que la discrepancia entre lo que fue observado y lo que fue esperado, fácilmente podría haberse debido sólo a la casualidad: por tanto, no tenemos pruebas para rechazar nuestra hipótesis. ¿Esto quiere decir que tenemos pruebas contundentes para **apoyar** nuestra hipótesis? No necesariamente, y este es un punto que suele entenderse mal. Considérese de esta forma. Existen muchas otras hipótesis que podríamos plantear, a partir de la cual dicha muestra no representaría una desviación significativa. Si tenemos una prueba contundente de que el rojo y el blanco están determinados por un solo par de genes, entonces una proporción de 3:1 es la hipótesis más razonable y nuestra muestra podría considerarse como suministrando una buena evidencia de

apoyo. Por otro lado, la evidencia que tenemos para postular un solo par de genes puede ser muy débil. Entonces, debemos considerar tales posibilidades cuando dos pares de genes dan lugar a una proporción de 9:7 o de 13:3. Nuestra muestra observada de 4 rojos: 4 blancos daría una "buena aproximación" a cada una de estas proporciones. Pruebas más amplias o muestras mucho mayores de F_2 habrían de emplearse para distinguir entre las diversas hipótesis plausibles.

La tabla 17.2 contiene los tamaños de muestra que se requieren para distinguir entre diversas razones comunes; por ejemplo, la tabla revela que una muestra de 105 debe estar necesariamente asegurada de que o una proporción de 3:1 o una de 9:7 se rechazará en el nivel del 5%. El valor de rechazo en la tabla de ji cuadrada es de 3.84. Si observamos una proporción de 70:35, el valor de ji cuadrada al probar la hipótesis 3:1 sería de 3:46, no lo suficientemente grande para rechazarla en el nivel del 5%. Nuevamente probada la hipótesis de 9:7, obtenemos un valor de ji cuadrada de 4.22, lo suficientemente grande para rechazar a la misma en el nivel del 5%. Por otro lado, una proporción observada de 69:36 daría valores de ji cuadrada de 4.34 y 3.45 para las hipótesis de 3:1 y 9:7, respectivamente. Por tanto, rechazaríamos la hipótesis de 3:1. La verificación de estos valores de ji cuadrada es un ejercicio que dejaremos al estudiante. En bien de la continuidad, asegúrese de utilizar la corrección.

Tabla 17.2. Tamaño de la muestra para estar seguros de que al menos una de las dos hipótesis alternativas será rechazada. (Números superiores en el nivel del 5%, y los inferiores en el nivel del 1%)

	15:1	7:1	13:3	3:1	11:5	5:3	9:7
1:1	16	24	38	62	112	254	1 008
	24	38	61	101	186	428	1 718
9:7	20	33	56	105	243	977	
	31	53	92	174	407	1 664	
5:3	27	49	94	223	915		
	42	79	155	374	1558		
11:5	39	80	195	823			
	61	130	326	1 398			
3:1	60	159	699				
	97	264	1 184				
13:3	114	543					
	186	915					
7:1	354						
	589						

CÓMO PROBAR LA INDEPENDENCIA

Uno de los aspectos que frecuentemente deseamos conocer cuando realizamos conteos, es si dos variables están relacionadas; por ejemplo, una variable utilizada para clasificar individuos puede ser el nivel de educación y la otra el nivel de ingreso. Podríamos realizar una prueba para verificar si la educación y el ingreso están relacionados.

Podemos imponer deliberadamente dos niveles de una variable, cómo la inoculación, sobre dos grupos, tratando a uno de ellos y dejando al otro sin tratamiento. Podríamos entonces clasificar a cada uno de los grupos en saludables y enfermos después de cierto tiempo y someter a prueba cualquier relación entre el tratamiento y la incidencia de la enfermedad. En la investigación genética, con frecuencia se desea determinar si dos características se heredan independientemente o si muestran evidencias de eslabonamiento. Todos estos problemas son similares en cuanto al análisis de correlación.

En el análisis de una relación entre dos variables, resulta mas conveniente plantear la hipótesis nula de que las mismas son independientes. Si la desviación de la independencia es mucho mayor de lo que esperaríamos por casualidad, rechazamos la hipótesis de que las dos variables son independientes y aceptamos la hipótesis alternativa de que éstas se encuentran relacionadas.,

A fin de hallar los valores esperados para aplicar la fórmula de ji cuadrada, utilizamos un principio de la teoría de la probabilidad que establece: **si dos acontecimientos son independientes,** la probabilidad de que ambos ocurran simultáneamente es el producto de sus probabilidades individuales de ocurrir. Ilustraremos este principio con un ejemplo y mostraremos cómo se realiza la prueba de ji cuadrada.

Cien animales fueron tratados con un antibiótico y, después de cierto tiempo, examinados en busca de síntomas de enfermedad. Hubo 88 animales que estaban saludables, y 12 mostraron síntomas de enfermedad. A otro grupo de 200 animales no se les suministró antibiótico y al ser examinados posteriormente, se encontró que 143 estaban saludables y 57 enfermos. Estos resultados pueden resumirse en lo que se denomina una tabla de contingencia de 2 x 2 (tabla 17.3).

Tabla 17.3. Incidencia de enfermedad en reses tratadas y no tratadas.

	Categorías de enfermedad		
Tratamiento	Saludable	Enfermo	Total
Tratado	88	12	100
Esperado	(77)	(23)	
No tratado	143	57	200
Esperado	(154)	(46)	
Totales	231	69	300

Someteremos a prueba la hipótesis de que no existe relación entre el tratamiento con antibiótico y la incidencia de enfermedad. Si estas dos variables son independientes, la proporción esperada de animales tratados saludables será igual a la proporción de animales saludables por la proporción de animales tratados. Esto es igual a $231/300 \times 100/300 = 77/300$. Puesto que existen 300 animales todos juntos, $77/300 \times 300 = 77$ es el número de animales que esperamos que estén tratados y saludables. Los cálculos pueden reducirse considerablemente, notando que el total principal aparece como el denominador de ambas fracciones que son multiplicadas para dar la probabilidad conjunta. La **proporción** resultante se multiplicó entonces por el total principal, para obtener el **número** esperado. Podemos cancelar uno de los totales

principales de nuestro cálculo y encontrar el número esperado de(100 × 231)/300 = 77. Esto puede expresarse en palabras: el número esperado de animales saludables tratados es igual al número total de animales tratados por el número total de animales saludables, dividido entre el total principal. Cada una de las demás clases esperadas puede calcularse en forma análoga. En realidad, en una tabla de 2 x 2, sólo necesita calcularse un valor esperado. Puesto que esperamos que 77 de los animales tratados fueran saludables, esperamos que los 23 restantes estuvieran enfermos. Análogamente, si esperamos que 77 de los animales saludables estuvieran en la clase de tratados, debemos esperar que el restante de los 231 animales saludables, 154, estuvieran en la clase de no tratados. Luego, de los 200 animales no tratados, puesto que esperamos que 154 sean saludables, esperamos que los 46 restantes estén enfermos. Nótese que se asigna un solo número a una de las clases, permaneciendo fijas las tres clases restantes. Luego, sólo tenemos un grado de libertad en una tabla de 2 x 2. La regla general para una tabla de contingencia r x c (r hileras y c columnas) es que los grados de libertad son iguales a $(r - 1) \times (c - 1)$.

Una característica especial de una tabla de 2 x 2 es que la diferencia entre lo observado y lo esperado es igual para cada casilla de la tabla, excepto que dos de las diferencias son positivas y las otras dos negativas. Esta diferencia común en nuestro ejemplo es igual a 11 (o sea, 88 − 77 = 11, etc.) y, puesto que estamos tratando con un grado de libertad, deberíamos aplicar la corrección de Yates y considerar las diferencias como 10.5.

Aplicando nuestra fórmula de ji cuadrada, obtenemos:

$$\chi^2 = \sum \frac{(|Ob - Ex| - .5)^2}{Ex} = \frac{(10.5)^2}{77} + \frac{(10.5)^2}{23} + \frac{(10.5)^2}{154} + \frac{(10.5)^2}{46} = 9.34$$

Si consultamos la tabla A.6 de ji cuadrada, por debajo de un grado de libertad, observaremos que podemos esperar que un valor de ji cuadrada de 6.635 ocurra por casualidad en sólo el 1% de las veces, y uno de 10.827 en sólo el 0.1% de las veces. Por tanto, podemos decir que la probabilidad de obtener un valor de ji cuadrada tan grande como 9.34, es sólo ligeramente mayor a uno entre mil, de modo que rechazamos la hipótesis de independencia y decimos que existe una relación entre el antibiótico y la incidencia de la enfermedad.

A fin de mostrar cómo se emplea la ji cuadrada para probar la independencia entre dos pares de genes, analizaremos algunos datos de una larga progenie de caléndulas, segregadas por dos factores: precocidad y virescencia (una ligera deficiencia de clorofila). Se sabe que la precocidad es recesiva al desarrollo posterior y está determinada en su material genético por un solo par de genes. La virescencia es recesiva en la planta normal y también se controla por un solo par de genes. Deben contestarse las siguientes preguntas: ¿se ajusta la razón de tarde: temprano a la razón 3:1?, ¿se ajusta la razón de la planta normal: virescente a la razón de 3:1?, ¿son los dos pares de características heredados independientemente o hay evidencia de eslabonamiento?

Los datos dispuestos en una tabla de contingencias (tabla 17.4) fueron:

Tabla 17.4. Segregación de dos características en una progenie de caléndulas.

	Normal	Virescente	Total	Ex 3:1
Tardío	3470	910	4380	4275
Esperado	(3457.9)	(922.1)		
Temprano	1030	290	1320	1425
Esperado	(1042.1)	(277.9)		
Totales	4500	1200	5700	
Ex (3:1)	4275	1425		

Para responder a la primera pregunta respecto a la razón de tarde: temprano, calculamos la ji cuadrada:

$$\chi^2 = (4\,380 - 4\,275 - .5)^2/4\,275 + (1\,320 - 1425 - .5)^2\ 1\,425$$

$$= (104.5)^2/4\,275 + (104.5)^2/1\,425$$

$$= 10.22$$

Este valor es casi igual al valor requerido de ji cuadrada de 10.827 en el nivel de 0.1%. Esto significa que si 3:1 fuese la razón verdadera, la probabilidad de encontrar una desviación tan grande como la que observamos fue sólo de aproximadamente uno entre mil; por tanto, rechazamos la hipótesís de que 3:1 es la razón verdadera. En realidad, no se rechazó la hipótesis de que la inflorescencia tardía fue un simple factor dominante sobre la inflorescencia temprana, puesto que se observó que (como con diversas características recesivas), las primeras plantas son en cierto modo más débiles que las posteriores. La pequeña, pero significativa desviación de una razón de 3:1 fue, por ende, atribuida a proporciones diferenciales de sobrevivencia. Vale la pena notar que ésta fue una progenie desacostumbradamente larga. Si ésta hubiera tenido un décimo de su tamaño (570 plantas) y la razón de tarde: temprano hubiera sido la misma, el valor de ji cuadrada habría sido de sólo 0.94, no aproximándose a la significación.

La pregunta acerca de la razón de normal: virescente se contesta en la misma forma, y el valor de ji cuadrada pasa a ser 47.16, de nuevo muy altamente significativo. Las plantas virescentes, que registran una pérdida parcial de clorofila, muestran incluso una pérdida mayor del vigor en comparación con las normales, de la que registran las plantas tempranas en comparación con las tardías.

Al someter a prueba la independencia, aceptamos las razones observadas, en vez de aceptar una razón de 3:1, y calculamos los valores esperados sobre el supuesto de independencia. Luego, el número esperado de plantas tardías normales es:

$$(\text{total normal} \times \text{total tardío})/\text{total principal} = (4\,500 \times 4\,380)/5\,700 = 3457.9.$$

Los valores esperados para las tres casillas restantes de la tabla 17.4 pueden calcularse en una forma similar, u obtenerse por sustracción a partir de los totales marginales. La utilización de ambos métodos suministra una comprobación de la exactitud de los cálculos. Nótese que (Ob — Ex) es igual a 12.1 en las casillas superior izquierda e inferior derecha de la tabla e igual a —12.1 en las otras dos casillas. Los numeradores de los términos para determinar la ji cuadrada serán iguales para cada clase. Aplicando la corrección de Yates en cada casilla de la tabla, obtenemos$(12.1 - .5)^2 = (11.6)^2$.Por tanto, ji cuadrada es:

$(11.6)^2/3\,456.9 + (11.6)^2/922.1 + (11.6)^2/1\,042.1 + (11.6)^2/277.9 = .80$. La probabilidad de obtener un valor de esta magnitud sólo por casualidad es del 10 al 50%, de modo que no tenemos ninguna evidencia para justificar el rechazo de la hipótesis de independencia.

Otro ejemplo mostrará cómo calcular la ji cuadrada cuando se incluye más de un grado de libertad, y cómo una tabla de contingencia puede "disolverse".Tres grupos de 39 reses fueron alimentados con raciones diferentes. Se midió condición de salud de cada animal mediante el registro del número de veces que el mismo tuvo que ser tratado por enfermedad. Se obtuvieron los siguientes resultados:

Tabla 17.5. Condición de salud de reses alimentadas con tres raciones. Los valores esperados se indican entre paréntesis.

Número de veces tratado	Ración 1	Ración 2	Ración 3	Total
0	19(17.3)	16(17.3)	17(17.3)	52
1	1 (.3)	0 (.3)	0 (.3)	1
2	0 (1.3)	3 (1.3)	1 (1.3)	4
3	7 (5.7)	9 (5.7)	1 (5.7)	7
4	3 (4.7)	5 (4.7)	6 (4.7)	14
5	4 (3.3)	1 (3.3)	5 (3.3)	10
6	2 (2.0)	1 (2.0)	3 (2.0)	6
7	0 (1.3)	2 (1.3)	2 (1.3)	4
8	1 (2.3)	2 (2.3)	4 (2.3)	7
10	2 (.7)	0 (.7)	0 (.7)	2
Totales	39	39	39	117

En este caso es fácil calcular los valores esperados, puesto que exactamente un tercio de todas las reses estaban en cada clase de ración. Esto significa que deberíamos esperar que un tercio de los animales en cada clase de frecuencia del tratamiento se ubicara en cada clase de ración, si ésta y la frecuencia del tratamiento fueran independientes. Notamos que muchos de los valores esperados son menores que cinco, de modo que, realmente, no hay justificación para aplicar la fórmula de ji cuadrada a los datos, tal como se encuentran; sin embargo, avanzaremos a través de los cálculos y veremos cómo los resultados se comparan con aquellos obtenidos a partir de una tabla disuelta:

$$\chi^2 = \Sigma \frac{(Ob - Ex)^2}{Ex} = \frac{(19 - 17.3)^2}{17.3} + \frac{(16 - 17.3)^2}{17.3} + \ldots + \frac{(0 - .7)^2}{.7} \quad 24.5$$

Grados de libertad = $(r - 1)(c - 1) = (10 - 1)(3 - 1) = 18$.

Buscando nuestro valor calculado de ji cuadrada de 24.5 en la tabla A.6 en el lado opuesto a 18 grados de libertad, encontramos que la probabilidad de obtener los resultados observados solo por casualidad, se halla ligeramente sobre el 10%; por tanto, nuestra evidencia es suficiente para rechazar la hipótesis de que la salud del animal no estuvo relacionada con la ración.

Para satisfacer la regla de que ninguna clase esperada debería ser menor que 5, podemos disolver la tabla, combinando las clases de frecuencia 1,2 y 3;4 y 5; y 6,7,8 y 10. Esto da lugar a la nueva tabla que presentamos a continuación:

Tabla 17.6. Versión disuelta de la tabla 17.5.

Número de veces tratado	Ración 1	Ración 2	Ración 3	Total
0	19 (17.3)	16 (17.3)	17 (17.3)	52
1 – 3	8 (7.3)	12 (7.3)	2 (7.3)	22
4 – 5	7 (8.0)	6 (8.0)	11 (8.0)	24
6 – 10	5 (6.3)	5 (6.3)	9 (6.3)	19
Totales	39	39	39	117

El cálculo de ji cuadrada da un valor de 10.61 que buscamos en el lado opuesto a 6 grados de libertad. Encontramos que éste es casi exactamente igual al valor tabular en el 10% de probabilidad. Por ende, nuestras conclusiones serán iguales a aquellas alcanzadas por la tabla original, aunque este no será siempre el caso. Es siempre más seguro disolver una tabla para evitar las clases esperadas demasiado pequeñas; además, esto reduce el número de cálculos que se necesitan para obtener la ji cuadrada. Nótese que la corrección para la continuidad no se utilizó en este ejemplo, puesto que estamos tratando con más de un solo grado de libertad.

HETEROGENEIDAD

La tercera y última utilización que consideraremos en relación con la ji cuadrada es probar si un grupo de muestras se podria haber obtenido de la misma población

Consideremos ocho progenies de caléndulas, cada una segregada entre normal y virescente, como sigue:

Tabla 17.7. Caléndulas normales y virescentes en ocho progenies.

Progenie	Normal	Virescente	χ^2 (3:1)	χ^2 (3 106:854)
1	315	85	3.00	0.023
2	602	170	3.65	0.094
3	868	252	3.73	0.578
4	174	42	3.56	0.575
5	192	48	3.20	0.348
6	165	39	3.76	0.723
7	161	43	1.67	0.028
8	629	175	4.48	0.019
Totales			27.05	2.388
Combinado	3 106	854	24.91	0.000
Heterogeneidad			2.14	2.388

Llevaremos a cabo dos tipos de análisis. Primero someteremos a prueba a cada progenie y los datos combinados de todas las progenies, en busca de la desviación de una razón hipotética de 3:1.

La ji cuadrada calculada para cada progenie se muestra en la columna cuatro. Estas se calcularon sin la corrección para la continuidad, puesto que deseamos sumarlas, y sólo las ji cuadradas no ajustadas son

aditivas. Nótese que sólo una de ellas excede el valor requerido de 3.84 para la significación en el nivel del 5%. Por tanto, tenemos muy poca evidencia de las progenies individuales para el rechazo de nuestra hipótesis. Aún no tenemos justificación para concluir que, puesto que siete de las ocho progenies dieron un "buen ajuste" (es decir, no se desviaron significativamente de 3:1), existe una prueba contundente para apoyar nuestra hipótesis. Debemos llevar el análisis más adelante. Sumando las ocho ji cuadradas individuales, cada una con un grado de libertad, obtenemos una ji cuadrada total de 27.05 con ocho grados de libertad. Esto excede el valor tabular de ji cuadrada de 26.125 en el nivel de 0.001. En otras palabras, la probabilidad es menor de una entre mil de que tal valor pudiera ser simplemente el resultado de la casualidad. Otra prueba puede aplicarse al total de 3 106 normales y de 854 virescentes. Los números esperados son

$3\,960 \times \frac{3}{4} = 2\,970$ y $3\,960 \times \frac{1}{4} = 990$, de modo que $\chi^2 = \frac{(3\,106 - 2\,970)^2}{2\,970} + \frac{(854 - 990)^2}{990} = 24.91$. Esto excede ampliamente el valor tabular de ji cuadrada para un grado de libertad en el nivel de 0.001, de manera que ahora rechazamos definitivamente la hipótesis de que todas las progenies son muestras de una población con una razón de 3:1. Aún nos gustaría saber si todas estas progenies pueden representar muestras de una sola población. Para someter esta hipótesis a prueba, calculamos lo que se denomina la **ji cuadrada de heterogeneidad.**

Ji cuadrada de heterogeneidad = ji cuadrada total — ji cuadrada combinada.

Puesto que la ji cuadrada total fue de 27.05 y la ji cuadrada es igual a 24.91, la ji cuadrada de heterogeneidad es de 2.14 con siete grados de libertad. Si consultamos la tabla, observaremos que éste es incluso menor que el de 2.167 requerido en el nivel de 0.95% la probabilidad de que un valor de ji cuadrada de este tamaño o menor pudiera provenir de un conjunto homogéneo de muestras sólo por casualidad, es de aproximadamente un 95%. Todas estas pruebas pueden resumirse en una tabla similar a una tabla de análisis de varianza.

Tabla 17.8. Resumen de los datos de ocho progenies de caléndulas, basados en una proporción de 3:1.

Fuente	gl	ji cuadrada
Total	8	27.05***
Combinado	1	24.91***
Heterogeneidad	7	2.14 ns

En vez de probar cada progenie contra una razón hipotética, podemos probarla contra la razón observada de los totales. Esto se efectúa en la última columna de la tabla 17.7. La ji cuadrada combinada tiene, desde luego, un valor de cero, puesto que la razón observada es aquella contra la cual estamos probando. Una tabla análoga a la anterior podría expresar lo siguiente:

Tabla 17.9. Resumen de los datos de las caléndulas, basado en totales observados.

Fuente	gl	ji cuadrada
Total	8	2.388
Combinado	1	0.000
Heterogeneidad	7	2.388

No tenemos aún evidencia de la heterogeneidad, concluyendo que estamos tratando con un conjunto homogéneo de progenies y que nuestra mejor estimación de la verdadera razón es 3 106:854.

Nótese que en la última prueba, los cálculos fueron exactamente iguales a los realizados para probar la independencia. En otras palabras, al probar cada muestra contra la razón total observada, **ji cuadrada de heterogeneidad = ji cuadrada de independencia.**

Sólo necesitamos separar la ji cuadrada total en dos componentes, cuando las muestras y los totales han sido probados contra una razón hipotética.

Veamos qué aspecto habría tenido el análisis si las primeras cuatro progenies hubieran mostrado iguales desviaciones de la razón 3:1, pero en el sentido opuesto:

Tabla 17.10. Conjunto hipotético de datos de las caléndulas, que muestran heterogeneidad.

Progenie	Normal	Virescente	χ^2 (3:1)	χ^2 (2 950:1 010)
1	285	115	3.00	2.05
2	556	216	3.65	2.49
3	812	308	3.73	2.35
4	150	66	3.56	2.90
5	192	48	3.20	3.82
6	165	39	3.76	4.38
7	161	43	1.67	2.10
8	629	175	4.48	5.92
Totales			27.05	26.01
Combinado	2950	1 010	0.54	0.00
Heterogeneidad			26.51	26.01

Nótese que los datos combinados están ahora muy cercanos a ajustarse a una razón de 3:1, pero que la ji cuadrada de heterogeneidad es altamente significativa. Nuevamente rechazamos la hipótesis de que todas las progenies son muestras de una población en la cual la razón es de tres normales para uno virescente. El rechazo se debe, en este caso, a la existencia de una fuerte evidencia de que las muestras no constituyen un conjunto homogéneo, de modo que la combinación de los datos no está justificada.

A través de todo este estudio hemos utilizado una sola fórmula:

$\chi^2 = \Sigma \frac{(Ob - Ex)^2}{Ex}$ con sólo una ligera modificación para los casos en que se requiere la corrección para la continuidad. Existen muchas modificaciones de esta fórmula que proporcionan métodos abreviados de cálculo para casos especiales. Una persona que debe calcular un gran número de jies cuadradas deberá recurrir a un texto más avanzado para las fórmulas abreviadas apropiadas. Para el lector que sólo ocasionalmente encuentra problemas que requieren el análisis de ji cuadrada, creemos que resulta preferible aprender únicamente esta fórmula básica.

RESUMEN

La fórmula general para calcular la ji cuadrada es $\chi^2 = \Sigma \dfrac{(Ob - Ex)^2}{Ex}$.

Los individuos clasificados en dos o más clases pueden ser comparados con una razón hipotética. Los grados de libertad son uno menos que el número de clases.

Si comparamos la ji cuadrada calculada con una tabla, podremos encontrar la probabilidad de ocurrencia de una desviación al menos tan grande como aquella observada sólo por casualidad.

Los individuos clasificados en dos formas en clases r y c, pueden ser sometidos a una prueba de independencia entre los dos criterios de clasificación. Los grados de libertad son $(r - 1) \times (c - 1)$.

Si se prueban dos o más muestras contra una razón hipotética común, la suma de las jies cuadradas resultantes puede separarse en dos componentes como sigue:

Fuente	gl
Total	$r(c - 1)$
Combinado	$(c - 1)$
Heterogeneidad	$(r - 1)(c - 1)$

El número de clases en que se clasifica cada muestra se denota por c, y r es el número de muestras.

Apéndice de tablas

Tabla A.1. Números aleatorios.

Para distribuir aleatoriamente cualquier conjunto de diez observaciones o menos, comiéncese por un punto aleatorio de la tabla y sígase cada una de las hileras, columnas o diagonales en cada dirección. Anótense los números en el orden en que éstos aparecen, sin importar aquellos que sean mayores que el número que está siendo distribuido aleatoriamente y aquellos que han aparecido antes en la serie. Si se desea distribuir aleatoriamente más de diez números, se pueden combinar pares de columnas o hileras para formar números de dos dígitos, siguiéndose el mismo proceso que el descrito anteriormente.

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

Tabla A.2. Distribución de t.

Grados de libertad	Probabilidad de obtener un valor tan grande o mayor.			
	0.100	0.050	0.010	0.001
1	6.314	12.706	63.657	
2	2.920	4.303	9.925	31.598
3	2.353	3.182	5.841	12.941
4	2.132	2.776	4.604	8.610
5	2.015	2.571	4.032	6.859
6	1.943	2.447	3.707	5.959
7	1.895	2.365	3.499	5.405
8	1.860	2.306	3.355	5.041
9	1.833	2.262	3.250	4.781
10	1.812	2.228	3.169	4.587
11	1.796	2.201	3.106	4.437
12	1.782	2.179	3.055	4.318
13	1.771	2.160	3.012	4.221
14	1.761	2.145	2.977	4.140
15	1.753	2.131	2.947	4.073
16	1.746	2.120	2.921	4.015
17	1.740	2.110	2.898	3.965
18	1.734	2.101	2.878	3.922
19	1.729	2.093	2.861	3.883
20	1.725	2.086	2.845	3.850
21	1.721	2.080	2.831	3.819
22	1.717	2.074	2.819	3.792
23	1.714	2.069	2.807	3.767
24	1.711	2.064	2.797	3.745
25	1.708	2.060	2.787	3.725
26	1.706	2.056	2.779	3.707
27	1.703	2.052	2.771	3.690
28	1.701	2.048	2.763	3.674
29	1.699	2.045	2.756	3.659
30	1.697	2.042	2.750	3.646
35	1.690	2.030	2.724	3.591
40	1.684	2.021	2.704	3.551
45	1.680	2.014	2.690	3.520
50	1.676	2.008	2.678	3.496
55	1.673	2.004	2.669	3.476
60	1.671	2.000	2.660	3.460
70	1.667	1.994	2.648	3.435
80	1.665	1.989	2.638	3.416
90	1.662	1.986	2.631	3.402
100	1.661	1.982	2.625	3.390
120	1.658	1.980	2.617	3.373
∞	1.6448	1.9600	2.5758	3.2905

Partes de esta tabla fueron tomadas de la obra de Fisher y Yates: *Statistical Tables for Biological, Agricultural, and Medical Research,* publicada por Oliver and Boyd Limited, Edimburgo (1948), y parte de la obra de Bernard Ostle, *Statistics in Research,* Iowa State University Press (1954), con autorización de los autores y editores.

Tabla A.3. Puntos de 10%, 5% y 1% para la distribución F

Grados de libertad para el denominador	P	Grados de libertad para el numerador (mayor cuadrado medio)															
		1	2	*3	4	5	6	7	8	9	10	11	12	14	15	16	20
1	0.10	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.89	60.20		60.70		61.22		61.74
	0.05	161	200	216	225	230	234	237	239	241	242	243	244	245		246	248
	0.01	4,052	4,999	5,403	5,625	5,764	5,859	5,928	5,981	6,022	6,056	6,082	6,106	6,142		6,169	6,208
2	0.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39		9.41		9.42		9.44
	0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42		19.43	19.44
	0.01	98.49	99.00	99.17	99.25	99.30	99.33	99.34	99.36	99.38	99.40	99.41	99.42	99.43		99.44	99.45
3	0.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23		5.22		5.20		5.18
	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71		8.69	8.66
	0.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05	26.92		26.83	26.69
4	0.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92		3.90		3.87		3.84
	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87		5.84	5.80
	0.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.54	14.45	14.37	14.24		14.15	14.02
5	0.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30		3.27		3.24		3.21
	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64		4.60	4.56
	0.01	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.27	10.15	10.05	9.96	9.89	9.77		9.68	9.55
6	0.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94		2.90		2.87		2.84
	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96		3.92	3.87
	0.01	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.60		7.52	7.39
7	0.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70		2.67		2.63		2.59
	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52		3.49	3.44
	0.01	12.25	9.55	8.45	7.85	7.46	7.19	7.00	6.84	6.71	6.62	6.54	6.47	6.35		6.27	6.15
8	0.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54		2.50		2.46		2.42
	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23		3.20	3.15
	0.01	11.26	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.91	5.82	5.74	5.67	5.56		5.48	5.36
* 9	0.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42		2.38		2.34		2.30
	*0.05	5.12	4.26	*3.86*	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02		2.98	2.93
	*0.01	10.56	8.02	*6.99*	6.42	6.06	5.80	5.62	5.47	5.35	5.26	5.18	5.11	5.00		4.92	4.80

Tabla A.3. Puntos de 10%, 5% y 1% para la distribución F (continuación).

Grados de libertad para el denominador	P	Grados de libertad para el numerador (mayor cuadrado medio)															
		1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	20
10	0.10	3.28	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32		2.28		2.24		2.20
	0.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86		2.82	2.77
	0.01	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.06	4.95	4.85	4.78	4.71	4.60		4.52	4.41
11	0.10	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25		2.21		2.17		2.12
	0.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74		2.70	2.65
	0.01	9.65	7.20	6.22	5.67	5.32	5.07	4.88	4.74	4.63	4.54	4.46	4.40	4.29		4.21	4.10
12	0.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19		2.15		2.10		2.06
	0.05	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64		2.60	2.54
	0.01	9.33	6.93	5.95	5.41	5.06	4.82	4.65	4.50	4.39	4.30	4.22	4.16	4.05		3.98	3.86
13	0.10	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14		2.10		2.05		2.01
	0.05	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55		2.51	2.46
	0.01	9.07	6.70	5.74	5.20	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.85		3.78	3.67
14	0.10	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10		2.05		2.01		1.96
	0.05	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48		2.44	2.39
	0.01	8.86	6.51	5.56	5.03	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.70		3.62	3.51
15	0.10	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06		2.02		1.97		1.92
	0.05	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43		2.39	2.33
	0.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.56		3.48	3.36
16	0.10	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03		1.99		1.94		1.89
	0.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37		2.33	2.28
	0.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.61	3.55	3.45		3.37	3.25
17	0.10	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00		1.96		1.91		1.86
	0.05	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33		2.29	2.23
	0.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.45	3.35		3.27	3.16
18	0.10	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98		1.93		1.89		1.84
	0.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29		2.25	2.19
	0.01	8.28	6.01	5.09	4.58	4.25	4.01	3.85	3.71	3.60	3.51	3.44	3.37	3.27		3.19	3.07

19	0.10	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96		1.91		1.86		1.81
	0.05	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26		2.21	2.15
	0.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.19		3.12	3.00
20	0.10	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94		1.89		1.84		1.79
	0.05	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23		2.18	2.12
	0.01	8.10	5.85	4.94	4.43	4.10	3.87	3.71	3.56	3.45	3.37	3.30	3.23	3.13		3.05	2.94
21	0.10	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92		1.88		1.83		1.78
	0.05	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20		2.15	2.09
	0.01	8.02	5.78	4.87	4.37	4.04	3.81	3.65	3.51	3.40	3.31	3.24	3.17	3.07		2.99	2.88
22	0.10	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90		1.86		1.81		1.76
	0.05	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23	2.18		2.13	2.07
	0.01	7.94	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.02		2.94	2.83
23	0.10	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89		1.84		1.80		1.74
	0.05	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20	2.14		2.10	2.04
	0.01	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	2.97		2.89	2.78
24	0.10	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88		1.83		1.78		1.73
	0.05	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13		2.09	2.02
	0.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.25	3.17	3.09	3.03	2.93		2.85	2.74
25	0.10	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87		1.82		1.77		1.72
	0.05	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11		2.06	2.00
	0.01	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.21	3.13	3.05	2.99	2.89		2.81	2.70
26	0.10	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86		1.81		1.76		1.71
	0.05	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10		2.05	1.99
	0.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.17	3.09	3.02	2.96	2.86		2.77	2.66
27	0.10	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85		1.80		1.75		1.70
	0.05	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08		2.03	1.97
	0.01	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.14	3.06	2.98	2.93	2.83		2.74	2.63
28	0.10	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84		1.79		1.74		1.69
	0.05	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06		2.02	1.96
	0.01	7.64	5.45	4.57	4.07	3.76	3.53	3.36	3.23	3.11	3.03	2.95	2.90	2.80		2.71	2.60
29	0.10	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83		1.78		1.73		1.68
	0.05	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05		2.00	1.94
	0.01	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.08	3.00	2.92	2.87	2.77		2.68	2.57

Tabla A.3. Puntos de 10%, 5% y 1% para la distribución F (continuación)

Grados de libertad para el denominador	P	Grados de libertad para el numerador (mayor cuadrado medio)															
		1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	20
30	0.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82		1.77		1.72		1.67
	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04		1.99	1.93
	0.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.06	2.98	2.90	2.84	2.74		2.66	2.55
32	0.05	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02		1.97	1.91
	0.01	7.50	5.34	4.46	3.97	3.66	3.42	3.25	3.12	3.01	2.94	2.86	2.80	2.70		2.62	2.51
34	0.05	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00		1.95	1.89
	0.01	7.44	5.29	4.42	3.93	3.61	3.38	3.21	3.08	2.97	2.89	2.82	2.76	2.66		2.58	2.47
36	0.05	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03	1.98		1.93	1.87
	0.01	7.39	5.25	4.38	3.89	3.58	3.35	3.18	3.04	2.94	2.86	2.78	2.72	2.62		2.54	2.43
38	0.05	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96		1.92	1.85
	0.01	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.91	2.82	2.75	2.69	2.59		2.51	2.40
40	0.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76		1.71		1.66		1.61
	0.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95		1.90	1.84
	0.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.88	2.80	2.73	2.66	2.56		2.49	2.37
42	0.05	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94		1.89	1.82
	0.01	7.27	5.15	4.29	3.80	3.49	3.26	3.10	2.96	2.86	2.77	2.70	2.64	2.54		2.46	2.35
44	0.05	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.92		1.88	1.81
	0.01	7.24	5.12	4.26	3.78	3.46	3.24	3.07	2.94	2.84	2.75	2.68	2.62	2.52		2.44	2.32
46	0.05	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91		1.87	1.80
	0.01	7.21	5.10	4.24	3.76	3.44	3.22	3.05	2.92	2.82	2.73	2.66	2.60	2.50		2.42	2.30
48	0.05	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96	1.90		1.86	1.79
	0.01	7.19	5.08	4.22	3.74	3.42	3.20	3.04	2.90	2.80	2.71	2.64	2.58	2.48		2.40	2.28
50	0.05	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.90		1.85	1.78
	0.01	7.17	5.06	4.20	3.72	3.41	3.18	3.02	2.88	2.78	2.70	2.62	2.56	2.46		2.39	2.26
55	0.05	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93	1.88		1.83	1.76
	0.01	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66	2.59	2.53	2.43		2.35	2.23

60	0.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71		1.66		1.60		1.54
	0.05	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.86		1.81	1.75
	0.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.40		2.32	2.20
65	0.05	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90	1.85		1.80	1.73
	0.01	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.79	2.70	2.61	2.54	2.47	2.37		2.30	2.18
70	0.05	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84		1.79	1.72
	0.01	7.01	4.92	4.08	3.60	3.29	3.07	2.91	2.77	2.67	2.59	2.51	2.45	2.35		2.28	2.15
80	0.05	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.82		1.77	1.70
	0.01	6.96	4.88	4.04	3.56	3.25	3.04	2.87	2.74	2.64	2.55	2.48	2.41	2.32		2.24	2.11
100	0.05	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79		1.75	1.68
	0.01	6.90	4.82	3.98	3.51	3.20	2.99	2.82	2.69	2.59	2.51	2.43	2.36	2.26		2.19	2.06
120	0.10	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65		1.60		1.54		1.48
125	0.05	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86	1.83	1.77		1.72	1.65
	0.01	6.84	4.78	3.94	3.47	3.17	2.95	2.79	2.65	2.56	2.47	2.40	2.33	2.23		2.15	2.03
150	0.05	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.76		1.71	1.64
	0.01	6.81	4.75	3.91	3.44	3.14	2.92	2.76	2.62	2.53	2.44	2.37	2.30	2.20		2.12	2.00
200	0.05	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80	1.74		1.69	1.62
	0.01	6.76	4.71	3.88	3.41	3.11	2.90	2.73	2.60	2.50	2.41	2.34	2.28	2.17		2.09	1.97
400	0.05	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78	1.72		1.67	1.60
	0.01	6.70	4.66	3.83	3.36	3.06	2.85	2.69	2.55	2.46	2.37	2.29	2.23	2.12		2.04	1.92
1000	0.05	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76	1.70		1.65	1.58
	0.01	6.66	4.62	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.26	2.20	2.09		2.01	1.89
∞	0.10	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60		1.55		1.49		1.42
	0.05	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69		1.64	1.57
	0.01	6.64	4.60	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.24	2.18	2.07		1.99	1.87

Los valores para las probabilidades de 5% y 1% se han copiado de la obra de Snedecor, *Statistical Methods,* y los valores para una probabilidad de 10% se han tomado de la obra *Tables of Percentage Points of the Inverted Beta (F) Distribution,* de Maxine Merrington y Catherine M. Thompson, Biometrika, 33:73, 1943, con autorización de los autores y editores.

PRUEBA DE DUNCAN.

Tabla A.4. Valores studentizados significativos (R) para multiplicar por DSM, para las medias en varios rangos (p), nivel del 5%; n = grados de libertad para el "error".

Grados Libertad Error

	# de tratamientos menos el 1ero.															
	p:															
n:	2	3	4	5	6	7	8	9	10	12	14	16	18	20	50	100
4	1.00	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.02
5	1.00	1.03	1.04	1.05	1.05	1.05	1.05	1.05	1.05	1.05	1.05	1.05	1.05	1.05	1.05	1.05
6	1.00	1.03	1.05	1.06	1.06	1.06	1.06	1.06	1.06	1.06	1.06	1.06	1.06	1.06	1.06	1.06
7	1.00	1.04	1.06	1.07	1.07	1.08	1.08	1.08	1.08	1.08	1.08	1.08	1.08	1.08	1.08	1.08
8	1.00	1.04	1.06	1.08	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09
9	1.00	1.04	1.07	1.08	1.09	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10
10	1.00	1.05	1.07	1.09	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10	1.10
11	1.00	1.05	1.08	1.09	1.10	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.12	1.12	1.12	1.12
12	1.00	1.05	1.08	1.09	1.10	1.11	1.12	1.12	1.12	1.12	1.12	1.12	1.13	1.13	1.13	1.13
13	1.00	1.05	1.08	1.09	1.10	1.11	1.12	1.12	1.13	1.13	1.13	1.13	1.13	1.13	1.13	1.13
14	1.00	1.05	1.08	1.10	1.11	1.12	1.13	1.13	1.14	1.14	1.14	1.14	1.15	1.15	1.15	1.15
15	1.00	1.05	1.08	1.10	1.12	1.12	1.13	1.14	1.14	1.14	1.15	1.15	1.15	1.15	1.15	1.15
16	1.00	1.05	1.08	1.10	1.12	1.12	1.13	1.14	1.14	1.15	1.15	1.16	1.16	1.16	1.16	1.16
17	1.00	1.05	1.08	1.10	1.12	1.13	1.13	1.14	1.15	1.15	1.16	1.16	1.16	1.16	1.16	1.16
18	1.00	1.05	1.08	1.10	1.12	1.13	1.13	1.14	1.15	1.15	1.16	1.16	1.17	1.17	1.17	1.17
19	1.00	1.05	1.08	1.10	1.12	1.13	1.14	1.15	1.15	1.16	1.16	1.17	1.17	1.17	1.17	1.17
20	1.00	1.05	1.08	1.10	1.12	1.13	1.14	1.15	1.15	1.16	1.17	1.17	1.17	1.18	1.18	1.18
22	1.00	1.05	1.08	1.10	1.12	1.13	1.14	1.15	1.16	1.17	1.18	1.18	1.18	1.18	1.18	1.18
24	1.00	1.05	1.08	1.10	1.12	1.13	1.14	1.15	1.16	1.17	1.18	1.18	1.18	1.19	1.19	1.19
26	1.00	1.05	1.08	1.10	1.12	1.13	1.15	1.15	1.16	1.17	1.18	1.19	1.19	1.19	1.19	1.19
28	1.00	1.05	1.08	1.10	1.12	1.14	1.15	1.16	1.16	1.17	1.18	1.19	1.19	1.20	1.20	1.20
30	1.00	1.05	1.08	1.11	1.12	1.14	1.15	1.16	1.17	1.18	1.19	1.19	1.20	1.20	1.20	1.20
40	1.00	1.05	1.08	1.11	1.13	1.14	1.15	1.16	1.17	1.19	1.20	1.20	1.21	1.21	1.21	1.21
60	1.00	1.05	1.09	1.11	1.13	1.14	1.16	1.17	1.18	1.19	1.20	1.21	1.22	1.23	1.23	1.23
100	1.00	1.05	1.09	1.11	1.14	1.15	1.16	1.18	1.19	1.20	1.21	1.22	1.23	1.24	1.26	1.26
∞	1.00	1.05	1.09	1.12	1.14	1.15	1.17	1.18	1.19	1.21	1.22	1.23	1.24	1.25	1.30	1.32

Tabla A.5. Valores studentizados significativos (R) para multiplicar por DSM, para las medias en varios rangos (p), nivel del 1%; n = grados de libertad para el "error".

n	p:															
	2	3	4	5	6	7	8	9	10	12	14	16	18	20	50	100
3	1.00	1.03	1.04	1.05	1.07	1.08	1.08	1.09	1.09	1.09	1.10	1.11	1.13	1.13	1.13	1.13
4	1.00	1.04	1.06	1.08	1.09	1.09	1.11	1.11	1.12	1.12	1.14	1.14	1.15	1.15	1.15	1.15
5	1.00	1.05	1.07	1.08	1.10	1.11	1.12	1.13	1.14	1.16	1.16	1.18	1.18	1.19	1.19	1.19
6	1.00	1.05	1.08	1.09	1.11	1.12	1.14	1.15	1.15	1.16	1.18	1.18	1.20	1.20	1.20	1.20
7	1.00	1.05	1.08	1.10	1.12	1.13	1.15	1.16	1.17	1.17	1.19	1.19	1.21	1.21	1.21	1.21
8	1.00	1.05	1.08	1.10	1.12	1.14	1.15	1.16	1.17	1.18	1.20	1.20	1.22	1.22	1.22	1.23
9	1.00	1.06	1.08	1.10	1.12	1.14	1.16	1.17	1.17	1.20	1.20	1.22	1.24	1.24	1.24	1.24
10	1.00	1.06	1.09	1.11	1.13	1.15	1.16	1.17	1.18	1.20	1.21	1.22	1.24	1.24	1.24	1.24
11	1.00	1.05	1.09	1.11	1.13	1.14	1.15	1.17	1.17	1.19	1.20	1.22	1.23	1.23	1.23	1.23
12	1.00	1.05	1.08	1.10	1.12	1.14	1.15	1.16	1.17	1.19	1.20	1.21	1.21	1.22	1.22	1.22
13	1.00	1.05	1.08	1.10	1.12	1.14	1.15	1.16	1.17	1.18	1.19	1.20	1.21	1.21	1.21	1.21
14	1.00	1.05	1.08	1.10	1.12	1.14	1.15	1.16	1.17	1.18	1.19	1.20	1.20	1.20	1.20	1.20
15	1.00	1.05	1.08	1.10	1.11	1.13	1.14	1.15	1.16	1.18	1.18	1.19	1.20	1.20	1.20	1.20
16	1.00	1.05	1.08	1.10	1.11	1.13	1.14	1.15	1.16	1.17	1.18	1.19	1.19	1.20	1.20	1.20
17	1.00	1.05	1.08	1.10	1.11	1.13	1.14	1.15	1.16	1.17	1.18	1.19	1.19	1.19	1.19	1.19
18	1.00	1.05	1.08	1.10	1.11	1.13	1.14	1.15	1.16	1.17	1.18	1.18	1.19	1.19	1.19	1.19
19	1.00	1.05	1.07	1.09	1.11	1.13	1.14	1.15	1.16	1.17	1.18	1.18	1.19	1.19	1.19	1.19
20	1.00	1.05	1.07	1.09	1.11	1.13	1.14	1.15	1.15	1.17	1.18	1.18	1.19	1.19	1.19	1.19
22	1.00	1.05	1.07	1.09	1.11	1.12	1.14	1.15	1.15	1.17	1.17	1.18	1.19	1.19	1.19	1.19
24	1.00	1.05	1.07	1.09	1.11	1.12	1.13	1.15	1.15	1.17	1.17	1.18	1.19	1.19	1.20	1.20
26	1.00	1.05	1.07	1.09	1.11	1.12	1.13	1.15	1.15	1.17	1.18	1.18	1.19	1.19	1.20	1.20
28	1.00	1.04	1.07	1.09	1.11	1.12	1.13	1.14	1.15	1.17	1.18	1.18	1.19	1.19	1.21	1.21
30	1.00	1.04	1.07	1.09	1.11	1.12	1.13	1.14	1.15	1.17	1.18	1.19	1.19	1.20	1.21	1.21
40	1.00	1.04	1.07	1.09	1.11	1.12	1.14	1.14	1.15	1.17	1.18	1.19	1.20	1.20	1.23	1.23
60	1.00	1.04	1.07	1.09	1.11	1.12	1.14	1.15	1.15	1.17	1.18	1.19	1.20	1.20	1.24	1.24
100	1.00	1.04	1.07	1.09	1.11	1.12	1.14	1.15	1 15	1.17	1.18	1.19	1.20	1.21	1.25	1.25
∞	1.00	1.04	1.07	1.09	1.11	1.12	1.14	1.15	1.15	1.17	1.18	1.19	1.20	1.21	1.26	1.29

Tabla A.6. Distribución de χ^2 (ji cuadrada).

Grados de libertad	Probabilidad de obtener un valor tan grande o mayor.							
	0.99	0.95	0.90	0.50	0.10	0.05	0.01	0.001
1	0.0002	0.00393	0.0158	0.455	2.706	3.841	6.635	10.827
2	0.0201	0.103	0.211	1.386	4.605	5.991	9.210	13.815
3	0.115	0.352	0.584	2.366	6.251	7.815	11.345	16.268
4	0.297	0.711	1.064	3.357	7.779	9.488	13.277	18.465
5	0.554	1.145	1.610	4.351	9.236	11.070	15.086	20.517
6	0.872	1.635	2.204	5.348	10.645	12.592	16.812	22.457
7	1.239	2.167	2.833	6.346	12.017	14.067	18.475	24.322
8	1.646	2.733	3.490	7.344	13.362	15.507	20.090	26.125
9	2.088	3.325	4.168	8.343	14.684	16.919	21.666	27.877
10	2.558	3.940	4.865	9.342	15.987	18.307	23.209	29.588
11	3.053	4.575	5.578	10.341	17.275	19.675	24.725	31.264
12	3.571	5.226	6.304	11.340	18.549	21.026	26.217	32.909
13	4.107	5.892	7.042	12.340	19.812	22.362	27.688	34.528
14	4.660	6.571	7.790	13.339	21.064	23.685	29.141	36.123
15	5.229	7.261	8.547	14.339	22.307	24.996	30.578	37.697
16	5.812	7.962	9.312	15.338	23.542	26.296	32.000	39.252
17	6.408	8.672	10.085	16.338	24.769	27.587	33.409	40.790
18	7.015	9.390	10.865	17.338	25.989	28.869	34.805	42.312
19	7.633	10.117	11.651	18.338	27.204	30.144	36.191	43.820
20	8.260	10.851	12.443	19.337	28.412	31.410	37.566	45.315
21	8.897	11.591	13.240	20.337	29.615	32.671	38.932	46.797
22	9.542	12.338	14.041	21.337	30.813	33.924	40.289	48.268
23	10.196	13.091	14.848	22.337	32.007	35.172	41.638	49.728
24	10.856	13.848	15.659	23.337	33.196	36.415	42.980	51.179
25	11.524	14.611	16.473	24.337	34.382	37.652	44.314	52.620
26	12.198	15.379	17.292	25.336	35.563	38.885	45.642	54.052
27	12.879	16.151	18.114	26.336	36.741	40.113	46.963	55.476
28	13.565	16.928	18.939	27.336	37.916	41.337	48.278	56.893
29	14.256	17.708	19.768	28.336	39.087	42.557	49.588	58.302
30	14.953	18.493	20.599	29.336	40.256	43.773	50.892	59.703

Esta tabla se ha resumido de la tabla 4 de Fisher y Yates: *Statistical Tables for Biological, Agricultural and Medical Research,* publicada por Oliver and Boyd Limited, Edimburgo, con autorización de los autores y editores.

Tabla A.7. Valores del coeficiente de correlación r, para ciertos niveles de significación.

Grados de libertad	Probabilidad de obtener un valor tan grande o mayor			
	0.1	0.05	0.01	0.001
1	0.9879	0.9969	0.9999	1.0000
2	0.9000	0.9500	0.9900	0.9990
3	0.8054	0.8783	0.9587	0.9912
4	0.7293	0.8114	0.9172	0.9741
5	0.6694	0.7545	0.8745	0.9507
6	0.6215	0.7067	0.8343	0.9249
7	0.5822	0.6664	0.7977	0.8982
8	0.5494	0.6319	0.7646	0.8721
9	0.5214	0.6021	0.7348	0.8471
10	0.4973	0.576C	0.7079	0.8233
11	0.4762	0.552S	0.6835	0.8010
12	0.4575	0.5324	0.6614	0.7800
13	0.4409	0.5139	0.6411	0.7603
14	0.4259	0.4973	0.6226	0.7420
15	0.4124	0.4821	0.6055	0.7246
16	0.4000	0.4683	0.5897	0.7084
17	0.3887	0.4555	0.5751	0.6932
18	0.3783	0.4438	0.5614	0.6787
19	0.3687	0.4329	0.5487	0.6652
20	0.3598	0.4227	0.5368	0.6524
25	0.3233	0.3809	0.4869	0.5974
30	0.2960	0.3494	0.4487	0.5541
35	0.2746	0.3246	0.4182	0.5189
40	0.2573	0.3044	0.3932	0.4896
45	0.2428	0.2875	0.3721	0.4684
50	0.2306	0.2732	0.3541	0.4433
60	0.2108	0.2500	0.3248	0.4078
70	0.1954	0.2319	0.3017	0.3799
80	0.1829	0.2172	0.2830	0.3568
90	0.1726	0.2050	0.2673	0.3375
100	0.1638	0.1946	0.2540	0.3211

Esta tabla se ha resumido de la tabla 4 de Fisher y Yates: ***Statistical Tables for Biological Agricultural, and Medical Research,*** publicada por Oliver and Boyd Limited, Edimburgo, con autorización de los autores y editores.

Tabla A.8. Transformación angular de porcentajes a grados.

%	0	1	2	3	4	5	6	7	8	9
0	0	5.7	8.1	10.0	11.5	12.9	14.2	15.3	16.4	17.5
10	18.4	19.4	20.3	21.1	22.0	22.8	23.6	24.4	25.1	25.8
20	26.6	27.3	28.0	28.7	29.3	30.0	30.7	31.3	31.9	32.6
30	33.2	33.8	34.4	35.1	35.7	36.3	36.9	37.5	38.1	38.6
40	39.2	39.8	40.4	41.0	41.6	42.1	42.7	43.3	43.9	44.4
50	45.0	45.6	46.1	46.7	47.3	47.9	48.4	49.0	49.6	50.2
60	50.8	51.4	51.9	52.5	53.1	53.7	54.3	54.9	55.6	56.2
70	56.8	57.4	58.1	58.7	59.3	60.0	60.7	61.3	62.0	62.7
80	63.4	64.2	64.9	65.6	66.4	67.2	68.0	68.9	69.7	70.6
90	71.6	72.5	73.6	74.7	75.8	77.1	78.5	80.0	81.9	84.3
100	90.0	—	—	—	—	—	—	—	—	—

Esta tabla se ha resumido de la tabla 12 de Fisher y Yates: *Statistical Tables for Biological, Agricultural, and Medical Research,* publicada por Oliver and Boyd Limited, Edimburgo, con autorización de los autores y editores.

Tabla A.9. Logaritmos.

Números naturales	0	1	2	3	4	5	6	7	8	9	Partes proporcionales								
											1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	16	19	21
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	8152	1	2	3	3	4	5	6	7	8
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1	2	2	3	4	5	6	7	7
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7

Tabla A.9. Logaritmos (continuación).

Números naturales	0	1	2	3	4	5	6	7	8	9	Partes proporcionales								
											1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	1	2	3	4	4	5	6	7
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	1	2	3	4	4	5	6	7
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	5	5	6
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	3	3	4	4	5	6
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1	1	2	2	3	4	4	5	6
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1	1	2	2	3	4	4	5	6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	5
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	5
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1	2	2	3	4	4	5	5
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	5
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	5
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	0	1	1	2	2	3	3	4	4
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0	1	1	2	2	3	3	4	4
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0	1	1	2	2	3	3	4	4
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0	1	1	2	2	3	3	3	4

Tabla A.10. Cuadrados y raíces cuadradas.

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	N	N^2	$\sqrt{N}$	$\sqrt{10N}$
1.00	1.0000	1.00000	3.16228	1.45	2.1025	1.20416	3.80789
1.01	1.0201	1.00499	3.17805	1.46	2.1316	1.20830	3.82099
1.02	1.0404	1.00995	3.19374	1.47	2.1609	1.21244	3.83406
1.03	1.0609	1.01489	3.20936	1.48	2.1904	1.21655	3.84708
1.04	1.0816	1.01980	3.22490	1.49	2.2201	1.22066	3.86005
1.05	1.1025	1.02470	3.24037	1.50	2.2500	1.22474	3.87298
1.06	1.1236	1.02956	3.25576	1.51	2.2801	1.22882	3.88587
1.07	1.1449	1.03441	3.27109	1.52	2.3104	1.23288	3.89872
1.08	1.1664	1.03923	3.28634	1.53	2.3409	1.23693	3.91152
1.09	1.1881	1.04403	3.30151	1.54	2.3716	1.24097	3.92428
1.10	1.2100	1.04881	3.31662	1.55	2.4025	1.24499	3.93700
1.11	1.2321	1.05357	3.33167	1.56	2.4336	1.24900	3.94968
1.12	1.2544	1.05830	3.34664	1.57	2.4649	1.25300	3.96232
1.13	1.2769	1.06301	3.36155	1.58	2.4964	1.25698	3.97492
1.14	1.2996	1.06771	3.37639	1.59	2.5281	1.26095	3.98748
1.15	1.3225	1.07238	3.39116	1.60	2.5600	1.26491	4.00000
1.16	1.3456	1.07703	3.40588	1.61	2.5921	1.26886	4.01248
1.17	1.3689	1.08167	3.42053	1.62	2.6244	1.27279	4.02492
1.18	1.3924	1.08628	3.43511	1.63	2.6569	1.27671	4.03733
1.19	1.4161	1.09087	3.44964	1.64	2.6896	1.28062	4.04969
1.20	1.4400	1.09545	3.46410	1.65	2.7225	1.28452	4.06202
1.21	1.4641	1.10000	3.47851	1.66	2.7556	1.28841	4.07431
1.22	1.4884	1.10454	3.49285	1.67	2.7889	1.29228	4.08656
1.23	1.5129	1.10905	3.50714	1.68	2.8224	1.29615	4.09878
1.24	1.5376	1.11355	3.52136	1.69	2.8561	1.30000	4.11096
1.25	1.5625	1.11803	3.53553	1.70	2.8900	1.30384	4.12311
1.26	1.5876	1.12250	3.54965	1.71	2.9241	1.30767	4.13521
1.27	1.6129	1.12694	3.56371	1.72	2.9584	1.31149	4.14729
1.28	1.6384	1.13137	3.57771	1.73	2.9929	1.31529	4.15933
1.29	1.6641	1.13578	3.59166	1.74	3.0276	1.31909	4.17133
1.30	1.6900	1.14018	3.60555	1.75	3.0625	1.32288	4.18330
1.31	1.7161	1.14455	3.61939	1.76	3.0976	1.32665	4.19524
1.32	1.7424	1.14891	3.63318	1.77	3.1329	1.33041	4.20714
1.33	1.7689	1.15326	3.64692	1.78	3.1684	1.33417	4.21900
1.34	1.7956	1.15758	3.66060	1.79	3.2041	1.33791	4.23084
1.35	1.8225	1.16190	3.67423	1.80	3.2400	1.34164	4.24264
1.36	1.8496	1.16619	3.68782	1.81	3.2761	1.34536	4.25441
1.37	1.8769	1.17047	3.70135	1.82	3.3124	1.34907	4.26615
1.38	1.9044	1.17473	3.71484	1.83	3.3489	1.35277	4.27785
1.39	1.9321	1.17898	3.72827	1.84	3.3856	1.35647	4.28952
1.40	1.9600	1.18322	3.74166	1.85	3.4225	1.36015	4.30116
1.41	1.9881	1.18743	3.75500	1.86	3.4596	1.36382	4.31277
1.42	2.0164	1.19164	3.76829	1.87	3.4969	1.36748	4.32435
1.43	2.0449	1.19583	3.78153	1.88	3.5344	1.37113	4.33590
1.44	2.0736	1.20000	3.79473	1.89	3.5721	1.37477	4.34741

Tabla A.10 Cuadrados y raíces cuadradas (continuación).

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	N	N^2	$\sqrt{N}$	$\sqrt{10N}$
1.90	3.6100	1.37840	4.35890	2.35	5.5225	1.53297	4.84768
1.91	3.6481	1.38203	4,37035	2.36	5.5696	1.53623	4.85798
1.92	3.6864	1.38564	4.38178	2.37	5.6169	1.53948	4.86826
1.93	3.7249	1.38924	4.39318	2.38	5.6644	1.54272	4.87852
1.94	3.7636	1.39284	4.40454	2.39	5.7121	1.54596	4.88876
1.95	3.8025	1.39642	4.41588	2.40	5.7600	1.54919	4.89898
1.96	3.8416	1.40000	4.42719	2.41	5.8081	1.55242	4.90918
1.97	3.8809	1.40357	4.43847	2.42	5.8564	1.55563	4.91935
1.98	3.9204	1.40712	4.44972	2.43	5.9049	1.55885	4.92950
1.99	3.9601	1.41067	4.46094	2.44	5.9536	1.56205	4.93964
2.00	4.0000	1.41421	4.47214	2.45	6.0025	1.56525	4.94975
2.01	4.0401	1.41774	4.48330	2.46	6.0516	1.56844	4.95984
2.02	4.0804	1.42127	4.49444	2.47	6.1009	1.57162	4.96991
2.03	4.1209	1.42478	4.50555	2.48	6.1504	1.57480	4.97996
2.04	4.1616	1.42829	4.51664	2.49	6.2001	1.57797	4.98999
2.05	4.2025	1.43178	4.52769	2.50	6.2500	1.58114	5.00000
2.06	4.2436	1.43527	4.53872	2.51	6.3001	1.58430	5.00999
2.07	4.2849	1.43875	4.54973	2.52	6.3504	1.58745	5.01996
2.08	4.3264	1.44222	4.56070	2.53	6.4009	1.59060	5.02991
2.09	4.3681	1.44568	4.57165	2.54	6.4516	1.59374	5.03984
2.10	4.4100	1.44914	4.58258	2.55	6.5025	1.59687	5.04975
2.11	4.4521	1.45258	4.59347	2.56	6.5536	1.60000	5.05964
2.12	4.4944	1.45602	4.60435	2.57	6.6049	1.60312	5.06952
2.13	4.5369	1.45945	4.61519	2.58	6.6564	1.60624	5.07937
2.14	4.5796	1.46287	4.62601	2.59	6.7081	1.60935	5.08920
2.15	4.6225	1.46629	4.63681	2.60	6.7600	1.61245	5.09902
2.16	4.6656	1.46969	4.64758	2.61	6.8121	1.61555	5.10882
2.17	4.7089	1.47309	4.65833	2.62	6.8644	1.61864	5.11859
2.18	4.7524	1.47648	4.66905	2.63	6.9169	1.62173	5.12835
2.19	4.7961	1.47986	4.67974	2.64	6.9696	1.62481	5.13809
2.20	4.8400	1.48324	4.69042	2.65	7.0225	1.62788	5.14782
2.21	4.8841	1.48661	4.70106	2.66	7.0756	1.63095	5.15752
2.22	4.9284	1.48997	4.71169	2.67	7.1289	1.63401	5.16720
2.23	4.9729	1.49332	4.72229	2.68	7.1824	1.63707	5.17687
2.24	5.0176	1.49666	4.73286	2.69	7.2361	1.64012	5.18652
2.25	5.0625	1.50000	4.74342	2.70	7.2900	1.64317	5.19615
2.26	5.1076	1.50333	4.75395	2.71	7.3441	1.64621	5.20577
2.27	5.1529	1.50665	4.76445	2.72	7.3984	1.64924	5.21536
2.28	5.1984	1.50997	4.77493	2.73	7.4529	1.65227	5.22494
2.29	5.2441	1.51327	4.78539	2.74	7.5076	1.65529	5.23450
2.30	5.2900	1.51658	4.79583	2.75	7.5625	1.65831	5.24404
2.31	5.3361	1.51987	4.80625	2.76	7.6176	1.66132	5.25357
2.32	5.3824	1.52315	4.81664	2.77	7.6729	1.66433	5.26308
2.33	5.4289	1.52643	4.82701	2.78	7.7284	1.66733	5.27257
2.34	5.4756	1.52971	4.83735	2.79	7.7841	1.67033	5.28205

Tabla A.10. Cuadrados y raíces cuadradas (continuación).

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	N	N^2	$\sqrt{N}$	$\sqrt{10N}$
2.80	7.8400	1.67332	5.29150	3.25	10.5625	1.80278	5.70088
2.81	7.8961	1.67631	5.30094	3.26	10.6276	1.80555	5.70964
2.82	7.9524	1.67929	5.31037	3.27	10.6929	1.80831	5.71839
2.83	8.0089	1.68226	5.31977	3.28	10.7584	1.81108	5.72713
2.84	8.0656	1.68523	5.32917	3.29	10.8241	1.81384	5.73585
2.85	8.1225	1.68819	5.33854	3.30	10.8900	1.81659	5.74456
2.86	8.1796	1.69115	5.34790	3.31	10.9561	1.81934	5.75326
2.87	8.2369	1.69411	5.35724	3.32	11.0224	1.82209	5.76194
2.88	8.2944	1.69706	5.36656	3.33	11.0889	1.82483	5.77062
2.89	8.3521	1.70000	5.37587	3.34	11.1556	1.82757	5.77927
2.90	8.4100	1.70294	5.38516	3.35	11.2225	1.83030	5.78792
2.91	8.4681	1.70587	5.39444	3.36	11.2896	1.83303	5.79655
2.92	8.5264	1.70880	5.40370	3.37	11.3569	1.83576	5.80517
2.93	8.5849	1.71172	5.41295	3.38	11.4244	1.83848	5.81378
2.94	8.6436	1.71464	5.42218	3.39	11.4921	1.84120	5.82237
2.95	8.7025	1.71756	5.43139	3.40	11.5600	1.84391	5.83095
2.96	8.7616	1.72047	5.44059	3.41	11.6281	1.84662	5.83952
2.97	8.8209	1.72337	5.44977	3.42	11.6964	1.84932	5.84808
2.98	8.8804	1.72627	5.45894	3.43	11.7649	1.85203	5.85662
2.99	8.9401	1.72916	5.46809	3.44	11.8336	1.85472	5.86515
3.00	9.0000	1.73205	5.47723	3.45	11.9025	1.85742	5.87367
3.01	9.0601	1.73494	5.48635	3.46	11.9716	1.86011	5.88218
3.02	9.1204	1.73781	5.49545	3.47	12.0409	1.86279	5.89067
3.03	9.1809	1.74069	5.50454	3.48	12.1104	1.86548	5.89915
3.04	9.2416	1.74356	5.51362	3.49	12.1801	1.86815	5.90762
3.05	9.3025	1.74642	5.52268	3.50	12.2500	1.87083	5.91608
3.06	9.3636	1.74929	5.53173	3.51	12.3201	1.87350	5.92453
3.07	9.4249	1.75214	5.54076	3.52	12.3904	1.87617	5.93296
3.08	9.4864	1.75499	5.54977	3.53	12.4609	1.87883	5.94138
3.09	9.5481	1.75784	5.55878	3.54	12.5316	1.88149	5.94979
3.10	9.6100	1.76068	5.56776	3.55	12.6025	1.88414	5.95819
3.11	9.6721	1.76352	5.57674	3.56	12.6736	1.88680	5.96657
3.12	9.7344	1.76635	5.58570	3.57	12.7449	1.88944	5.97495
3.13	9.7969	1.76918	5.59464	3.58	12.8164	1.89209	5.98331
3.14	9.8596	1.77200	5.60357	3.59	12.8881	1.89473	5.99166
3.15	9.9225	1.77482	5.61249	3.60	12.9600	1.89737	6.00000
3.16	9.9856	1.77764	5.62139	3.61	13.0321	1.90000	6.00833
3.17	10.0489	1.78045	5.63028	3.62	13.1044	1.90263	6.01664
3.18	10.1124	1.78326	5.63915	3.63	13.1769	1.90526	6.02495
3.19	10.1761	1.78606	5.64801	3.64	13.2496	1.90788	6.03324
3.20	10.2400	1.78885	5.65685	3.65	13.3225	1.91050	6.04152
3.21	10.3041	1.79165	5.66569	3.66	13.3956	1.91311	6.04979
3.22	10.3684	1.79444	5.67450	3.67	13.4689	1.91572	6.05805
3.23	10.4329	1.79722	5.68331	3.68	13.5424	1.91833	6.06630
3.24	10.4976	1.80000	5.69210	3.69	13.6161	1.92094	6.07454

Tabla A.10. Cuadrados y raíces cuadradas (continuación).

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	N	N^2	$\sqrt{N}$	$\sqrt{10N}$
3.70	13.6900	1.92354	6.08276	4.15	17.2225	2.03715	6.44205
3.71	13.7641	1.92614	6.09098	4.16	17.3056	2.03961	6.44981
3.72	13.8384	1.92873	6.09918	4.17	17.3889	2.04206	6.45755
3.73	13.9129	1.93132	6.10737	4.18	17.4724	2.04450	6.46529
3.74	13.9876	1.93391	6.11555	4.19	17.5561	2.04695	6.47302
3.75	14.0625	1.93649	6.12372	4.20	17.6400	2.04939	6.48074
3.76	14.1376	1.93907	6.13188	4.21	17.7241	2.05183	6.48845
3.77	14.2129	1.94165	6.14003	4.22	17.8084	2.05426	6.49615
3.78	14.2884	1.94422	6.14817	4.23	17.8929	2.05670	6.50384
3.79	14.3641	1.94679	6.15630	4.24	17.9776	2.05913	6.51153
3.80	14.4400	1.94936	6.16441	4.25	18.0625	2.06155	6.51920
3.81	14.5161	1.95192	6.17252	4.26	18.1476	2.06398	6.52687
3.82	14.5924	1.95448	6.18061	4.27	18.2329	2.06640	6.53452
3.83	14.6689	1.95704	6.18870	4.28	18.3184	2.06882	6.54217
3.84	14.7456	1.95959	6.19677	4.29	18.4041	2.07123	6.54981
3.85	14.8225	1.96214	6.20484	4.30	18.4900	2.07364	6.55744
3.86	14.8996	1.96469	6.21289	4.31	18.5761	2.07605	6.56506
3.87	14.9769	1.96723	6.22093	4.32	18.6624	2.07846	6.57267
3.88	15.0544	1.96977	6.22896	4.33	18.7489	2.08087	6.58027
3.89	15.1321	1.97231	6.23699	4.34	18.8356	2.08327	6.58787
3.90	15.2100	1.97484	6.24500	4.35	18.9225	2.08567	6.59545
3.91	15.2881	1.97737	6.25300	4.36	19.0096	2.08806	6.60303
3.92	15.3664	1.97990	6.26099	4.37	19.0969	2.09045	6.61060
3.93	15.4449	1.98242	6.26897	4.38	19.1844	2.09284	6.61816
3.94	15.5236	1.98494	6.27694	4.39	19.2721	2.09523	6.62571
3.95	15.6025	1.98746	6.28490	4.40	19.3600	2.09762	6.63325
3.96	15.6816	1.98997	6.29285	4.41	19.4481	2.10000	6.64078
3.97	15.7609	1.99249	6.30079	4.42	19.5364	2.10238	6.64831
3.98	15.8404	1.99499	6.30872	4.43	19.6249	2.10476	6.65582
3.99	15.9201	1.99750	6.31664	4.44	19.7136	2.10713	6.66333
4.00	16.0000	2.00000	6.32456	4.45	19.8025	2.10950	6.67083
4.01	16.0801	2.00250	6.33246	4.46	19.8916	2.11187	6.67832
4.02	16.1604	2.00499	6.34035	4.47	19.9809	2.11424	6.68581
4.03	16.2409	2.00749	6.34823	4.48	20.0704	2.11660	6.69328
4.04	16.3216	2.00998	6.35610	4.49	20.1601	2.11896	6.70075
4.05	16.4025	2.01246	6.36396	4.50	20.2500	2.12132	6.70820
4.06	16.4836	2.01494	6.37181	4.51	20.3401	2.12368	6.71565
4.07	16.5649	2.01742	6.37966	4.52	20.4304	2.12603	6.72309
4.08	16.6464	2.01990	6.38749	4.53	20.5209	2.12838	6.73053
4.09	16.7281	2.02237	6.39531	4.54	20.6116	2.13073	6.73795
4.10	16.8100	2.02485	6.40312	4.55	20.7025	2.13307	6.74537
4.11	16.8921	2.02731	6.41093	4.56	20.7936	2.13542	6.75278
4.12	16.9744	2.02978	6.41872	4.57	20.8849	2.13776	6.76018
4.13	17.0569	2.03224	6.42651	4.58	20.9764	2.14009	6.76757
4.14	17.1396	2.03470	6.43428	4.59	21.0681	2.14243	6.77495

Tabla A.10. Cuadrados y raíces cuadradas (continuación).

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	N	N^2	$\sqrt{N}$	$\sqrt{10N}$
4.60	21.1600	2.14476	6.78233	5.05	25.5025	2.24722	7.10634
4.61	21.2521	2.14709	6.78970	5.06	25.6036	2.24944	7.11337
4.62	21.3444	2.14942	6.79706	5.07	25.7049	2.25167	7.12039
4.63	21.4369	2.15174	6.80441	5.08	25.8064	2.25389	7.12741
4.64	21.5296	2.15407	6.81175	5.09	25.9081	2.25610	7.13442
4.65	21.6225	2.15639	6.81909	5.10	26.0100	2.25832	7.14143
4.66	21.7156	2.15870	6.82642	5.11	26.1121	2.26053	7.14843
4.67	21.8089	2.16102	6.83374	5.12	26.2144	2.26274	7.15542
4.68	21.9024	2.16333	6.84105	5.13	26.3169	2.26495	7.16240
4.69	21.9961	2.16564	6.84836	5.14	26.4196	2.26716	7.16938
4.70	22.0900	2.16795	6.85565	5.15	26.5225	2.26936	7.17635
4.71	22.1841	2.17025	6.86294	5.16	26.6256	2.27156	7.18331
4.72	22.2784	2.17256	6.87023	5.17	26.7289	2.27376	7.19027
4.73	22.3729	2.17486	6.87750	5.18	26.8324	2.27596	7.19722
4.74	22.4676	2.17715	6.88477	5.19	26.9361	2.27816	7.20417
4.75	22.5625	2.17945	6.89202	5.20	27.0400	2.28035	7.21110
4.76	22.6576	2.18174	6.89928	5.21	27.1441	2.28254	7.21803
4.77	22.7529	2.18403	6.90652	5.22	27.2484	2.28473	7.22496
4.78	22.8484	2.18632	6.91375	5.23	27.3529	2.28692	7.23187
4.79	22.9441	2.18861	6.92098	5.24	27.4576	2.28910	7.23878
4.80	23.0400	2.19089	6.92820	5.25	27.5625	2.29129	7.24569
4.81	23.1361	2.19317	6.93542	5.26	27.6676	2.29347	7.25259
4.82	23.2324	2.19545	6.94262	5.27	27.7729	2.29565	7.25948
4.83	23.3289	2.19773	6.94982	5.28	27.8784	2.29783	7.26636
4.84	23.4256	2.20000	6.95701	5.29	27.9841	2.30000	7.27324
4.85	23.5225	2.20227	6.96419	5.30	28.0900	2.30217	7.28011
4.86	23.6196	2.20454	6.97137	5.31	28.1961	2.30434	7.28697
4.87	23.7169	2.20681	6.97854	5.32	28.3024	2.30651	7.29383
4.88	23.8144	2.20907	6.98570	5.33	28.4089	2.30868	7.30068
4.89	23.9121	2.21133	6.99285	5.34	28.5156	2.31084	7.30753
4.90	24.0100	2.21359	7.00000	5.35	28.6225	2.31301	7.31437
4.91	24.1081	2.21585	7.00714	5.36	28.7296	2.31517	7.32120
4.92	24.2064	2.21811	7.01427	5.37	28.8369	2.31733	7.32803
4.93	24.3049	2.22036	7.02140	5.38	28.9444	2.31948	7.33485
4.94	24.4036	2.22261	7.02851	5.39	29.0521	2.32164	7.34166
4.95	24.5025	2.22486	7.03562	5.40	29.1600	2.32379	7.34847
4.96	24.6016	2.22711	7.04273	5.41	29.2681	2.32594	7.35527
4.97	24.7009	2.22935	7.04982	5.42	29.3764	2.32809	7.36206
4.98	24.8004	2.23159	7.05691	5.43	29.4849	2.33024	7.36885
4.99	24.9001	2.23383	7.06399	5.44	29.5936	2.33238	7.37564
5.00	25.0000	2.23607	7.07107	5.45	29.7025	2.33452	7.38241
5.01	25.1001	2.23830	7.07814	5.46	29.8116	2.33666	7.38918
5.02	25.2004	2.24054	7.08520	5.47	29.9209	2.33880	7.39594
5.03	25.3009	2.24277	7.09225	5.48	30.0304	2.34094	7.40270
5.04	25.4016	2.24499	7.09930	5.49	30.1401	2.34307	7.40945

Tabla A.10. Cuadrados y raíces cuadradas (continuación).

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	N	N^2	$\sqrt{N}$	$\sqrt{10N}$
5.50	30.2500	2.34521	7.41620	5.95	35.4025	2.43926	7.71362
5.51	30.3601	2.34734	7.42294	5.96	35.5216	2.44131	7.72010
5.52	30.4704	2.34947	7.42967	5.97	35.6409	2.44336	7.72658
5.53	30.5809	2.35160	7.43640	5.98	35.7604	2.44540	7.73305
5.54	30.6916	2.35372	7.44312	5.99	35.8801	2.44745	7.73951
5.55	30.8025	2.35584	7.44983	6.00	36.0000	2.44949	7.74597
5.56	30.9136	2.35797	7.45654	6.01	36.1201	2.45153	7.75242
5.57	31.0249	2.36008	7.46324	6.02	36.2404	2.45357	7.75887
5.58	31.1364	2.36220	7.46994	6.03	36.3609	2.45561	7.76531
5.59	31.2481	2.36432	7.47663	6.04	36.4816	2.45764	7.77174
5.60	31.3600	2.36643	7.48331	6.05	36.6025	2.45967	7.77817
5.61	31.4721	2.36854	7.48999	6.06	36.7236	2.46171	7.78460
5.62	31.5844	2.37065	7.49667	6.07	36.8449	2.46374	7.79102
5.63	31.6969	2.37276	7.50333	6.08	36.9664	2.46577	7.79744
5.64	31.8096	2.37487	7.50999	6.09	37.0881	2.46779	7.80385
5.65	31.9225	2.37697	7.51665	6.10	37.2100	2.46982	7.81025
5.66	32.0356	2.37908	7.52330	6.11	37.3321	2.47184	7.81665
5.67	32.1489	2.38118	7.52994	6.12	37.4544	2.47386	7.82304
5.68	32.2624	2.38328	7.53658	6.13	37.5769	2.47588	7.82943
5.69	32.3761	2.38537	7.54321	6.14	37.6996	2.47790	7.83582
5.70	32.4900	2.38747	7.54983	6.15	37.8225	2.47992	7.84219
5.71	32.6041	2.38956	7.55645	6.16	37.9456	2.48193	7.84857
5.72	32.7184	2.39165	7.56307	6.17	38.0689	2.48395	7.85493
5.73	32.8329	2.39374	7.56968	6.18	38.1924	2.48596	7.86130
5.74	32.9476	2.39583	7.57628	6.19	38.3161	2.48797	7.86766
5.75	33.0625	2.39792	7.58288	6.20	38.4400	2.48998	7.87401
5.76	33.1776	2.40000	7.58947	6.21	38.5641	2.49199	7.88036
5.77	33.2929	2.40208	7.59605	6.22	38.6884	2.49399	7.88670
5.78	33.4084	2.40416	7.60263	6.23	38.8129	2.49600	7.89303
5.79	33.5241	2.40624	7.60920	6.24	38.9376	2.49800	7.89937
5.80	33.6400	2.40832	7.61577	6.25	39.0625	2.50000	7.90569
5.81	33.7561	2.41039	7.62234	6.26	39.1876	2.50200	7.91202
5.82	33.8724	2.41247	7.62889	6.27	39.3129	2.50400	7.91833
5.83	33.9889	2.41454	7.63544	6.28	39.4384	2.50599	7.92465
5.84	34.1056	2.41661	7.64199	6.29	39.5641	2.50799	7.93095
5.85	34.2225	2.41868	7.64853	6.30	39.6900	2.50998	7.93725
5.86	34.3396	2.42074	7.65506	6.31	39.8161	2.51197	7.94355
5.87	34.4569	2.42281	7.66159	6.32	39.9424	2.51396	7.94984
5.88	34.5744	2.42487	7.66812	6.33	40.0689	2.51595	7.95613
5.89	34.6921	2.42693	7.67463	6.34	40.1956	2.51794	7.96241
5.90	34.8100	2.42899	7.68115	6.35	40.3225	2.51992	7.96869
5.91	34.9281	2.43105	7.68765	6.36	40.4496	2.52190	7.97496
5.92	35.0464	2.43311	7.69415	6.37	40.5769	2.52389	7.98123
5.93	35.1649	2.43516	7.70065	6.38	40.7044	2.52587	7.98749
5.94	35.2836	2.43721	7.70714	6.39	40.8321	2.52784	7.99375

Tabla A.10. Cuadrados y raíces cuadradas (continuación).

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	N	N^2	$\sqrt{N}$	$\sqrt{10N}$
6.40	40.9600	2.52982	8.00000	6.85	46.9225	2.61725	8.27647
6.41	41.0881	2.53180	8.00625	6.86	47.0596	2.61916	8.28251
6.42	41.2164	2.53377	8.01249	6.87	47.1969	2.62107	8.28855
6.43	41.3449	2.53574	8.01873	6.88	47.3344	2.62298	8.29458
6.44	41.4736	2.53772	8.02496	6.89	47.4721	2.62488	8.30060
6.45	41.6025	2.53969	8.03119	6.90	47.6100	2.62679	8.30662
6.46	41.7316	2.54165	8.03741	6.91	47.7481	2.62869	8.31264
6.47	41.8609	2.54362	8.04363	6.92	47.8864	2.63059	8.31865
6.48	41.9904	2.54558	8.04984	6.93	48.0249	2.63249	8.32466
6.49	42.1201	2.54755	8.05605	6.94	48.1636	2.63439	8.33067
6.50	42.2500	2.54951	8.06226	6.95	48.3025	2.63629	8.33667
6.51	42.3801	2.55147	8.06846	6.96	48.4416	2.63818	8.34266
6.52	42.5104	2.55343	8.07465	6.97	48.5809	2.64008	8.34865
6.53	42.6409	2.55539	8.08084	6.98	48.7204	2.64197	8.35464
6.54	42.7716	2.55734	8.08703	6.99	48.8601	2.64386	8.36062
6.55	42.9025	2.55930	8.09321	7.00	49.0000	2.64575	8.36660
6.56	43.0336	2.56125	8.09938	7.01	49.1401	2.64764	8.37257
6.57	43.1649	2.56320	8.10555	7.02	49.2804	2.64953	8.37854
6.58	43.2964	2.56515	8.11172	7.03	49.4209	2.65141	8.38451
6.59	43.4281	2.56710	8.11788	7.04	49.5616	2.65330	8.39047
6.60	43.5600	2.56905	8.12404	7.05	49.7025	2.65518	8.39643
6.61	43.6921	2.57099	8.13019	7.06	49.8436	2.65707	8.40238
6.62	43.8244	2.57294	8.13634	7.07	49.9849	2.65895	8.40833
6.63	43.9569	2.57488	8.14248	7.08	50.1264	2.66083	8.41427
6.64	44.0896	2.57682	8.14862	7.09	50.2681	2.66271	8.42021
6.65	44.2225	2.57876	8.15475	7.10	50.4100	2.66458	8.42615
6.66	44.3556	2.58070	8.16088	7.11	50.5521	2.66646	8.43208
6.67	44.4889	2.58263	8.16701	7.12	50.6944	2.66833	8.43801
6.68	44.6224	2.58457	8.17313	7.13	50.8369	2.67021	8.44393
6.69	44.7561	2.58650	8.17924	7.14	50.9796	2.67208	8.44985
6.70	44.8900	2.58844	8.18535	7.15	51.1225	2.67395	8.45577
6.71	45.0241	2.59037	8.19146	7.16	51.2656	2.67582	8.46168
6.72	45.1584	2.59230	8.19756	7.17	51.4089	2.67769	8.46759
6.73	45.2929	2.59422	8.20366	7.18	51.5524	2.67955	8.47349
6.74	45.4276	2.59615	8.20975	7.19	51.6961	2.68142	8.47939
6.75	45.5625	2.59808	8.21584	7.20	51.8400	2.68328	8.48528
6.76	45.6976	2.60000	8.22192	7.21	51.9841	2.68514	8.49117
6.77	45.8329	2.60192	8.22800	7.22	52.1284	2.68701	8.49706
6.78	45.9684	2.60384	8.23408	7.23	52.2729	2.68887	8.50294
6.79	46.1041	2.60576	8.24015	7.24	52.4176	2.69072	8.50882
6.80	46.2400	2.60768	8.24621	7.25	52.5625	2.69258	8.51469
6.81	46.3761	2.60960	8.25227	7.26	52.7076	2.69444	8.52056
6.82	46.5124	2.61151	8.25833	7.27	52.8529	2.69629	8.52643
6.83	46.6489	2.61343	8.26438	7.28	52.9984	2.69815	8.53229
6.84	46.7856	2.61534	8.27043	7.29	53.1441	2.70000	8.53815

Tabla A.10. Cuadrados y raíces cuadradas (continuación)

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	N	N^2	$\sqrt{N}$	$\sqrt{10N}$
7.30	53.2900	2.70185	8.54400	7.75	60.0625	2.78388	8.80341
7.31	53.4361	2.70370	8.54985	7.76	60.2176	2.78568	8.80909
7.32	53.5824	2.70555	8.55570	7.77	60.3729	2.78747	8.81476
7.33	53.7289	2.70740	8.56154	7.78	60.5284	2.78927	8.82043
7.34	53.8756	2.70924	8.56738	7.79	60.6841	2.79106	8.82610
7.35	54.0225	2.71109	8.57321	7.80	60.8400	2.79285	8.83176
7.36	54.1696	2.71293	8.57904	7.81	60.9961	2.79464	8.83742
7.37	54.3169	2.71477	8.58487	7.82	61.1524	2.79643	8.84308
7.38	54.4644	2.71662	8.59069	7.83	61.3089	2.79821	8.84873
7.39	54.6121	2.71846	8.59651	7.84	61.4656	2.80000	8.85438
7.40	54.7600	2.72029	8.60233	7.85	61.6225	2.80179	8.86002
7.41	54.9081	2.72213	8.60814	7.86	61.7796	2.80357	8.86566
7.42	55.0564	2.72397	8.61394	7.87	61.9369	2.80535	8.87130
7.43	55.2049	2.72580	8.61974	7.88	62.0944	2.80713	8.87694
7.44	55.3536	2.72764	8.62554	7.89	62.2521	2.80891	8.88257
7.45	55.5025	2.72947	8.63134	7.90	62.4100	2.81069	8.88819
7.46	55.6516	2.73130	8.63713	7.91	62.5681	2.81247	8.89382
7.47	55.8009	2.73313	8.64292	7.92	62.7264	2.81425	8.89944
7.48	55.9504	2.73496	8.64870	7.93	62.8849	2.81603	8.90505
7.49	56.1001	2.73679	8.65448	7.94	63.0436	2.81780	8.91067
7.50	56.2500	2.73861	8.66025	7.95	63.2025	2.81957	8.91628
7.51	56.4001	2.74044	8.66603	7.96	63.3616	2.82135	8.92188
7.52	56.5504	2.74226	8.67179	7.97	63.5209	2.82312	8.92749
7.53	56.7009	2.74408	8.67756	7.98	63.6804	2.82489	8.93308
7.54	56.8516	2.74591	8.68332	7.99	63.8401	2.82666	8.93868
7.55	57.0025	2.74773	8.68907	8.00	64.0000	2.82843	8.94427
7.56	57.1536	2.74955	8.69483	8.01	64.1601	2.83019	8.94986
7.57	57.3049	2.75136	8.70057	8.02	64.3204	2.83196	8.95545
7.58	57.4564	2.75318	8.70632	8.03	64.4809	2.83373	8.96103
7.59	57.6081	2.75500	8.71206	8.04	64.6416	2.83549	8.96660
7.60	57.7600	2.75681	8.71780	8.05	64.8025	2.83725	8.97218
7.61	57.9121	2.75862	8.72353	8.06	64.9636	2.83901	8.97775
7.62	58.0644	2.76043	8.72926	8.07	65.1249	2.84077	8.98332
7.63	58.2169	2.76225	8.73499	8.08	65.2864	2.84253	8.98888
7.64	58.3696	2.76405	8.74071	8.09	65.4481	2.84429	8.99444
7.65	58.5225	2.76586	8.74643	8.10	65.6100	2.84605	9.00000
7.66	58.6756	2.76767	8.75214	8.11	65.7721	2.84781	9.00555
7.67	58.8289	2.76948	8.75785	8.12	65.9344	2.84956	9.01110
7.68	58.9824	2.77128	8.76356	8.13	66.0969	2.85132	9.01665
7.69	59.1361	2.77308	8.76926	8.14	66.2596	2.85307	9.02219
7.70	59.2900	2.77489	8.77496	8.15	66.4225	2.85482	9.02774
7.71	59.4441	2.77669	8.78066	8.16	66.5856	2.85657	9.03327
7.72	59.5984	2.77849	8.78635	8.17	66.7489	2.85832	9.03881
7.73	59.7529	2.78029	8.79204	8.18	66.9124	2.86007	9.04434
7.74	59.9076	2.78209	8.79773	8.19	67.0761	2.86182	9.04986

Tabla A.10. Cuadrados y raíces cuadradas (continuación).

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	N	N^2	$\sqrt{N}$	$\sqrt{10N}$
8.20	67.2400	2.86356	9.05539	8.65	74.8225	2.94109	9.30054
8.21	67.4041	2.86531	9.06091	8.66	74.9956	2.94279	9.30591
8.22	67.5684	2.86705	9.06642	8.67	75.1689	2.94449	9.31128
8.23	67.7329	2.86880	9.07193	8.68	75.3424	2.94618	9.31665
8.24	67.8976	2.87054	9.07744	8.69	75.5161	2.94788	9.32202
8.25	68.0625	2.87228	9.08295	8.70	75.6900	2.94958	9.32738
8.26	68.2276	2.87402	9.08845	8.71	75.8641	2.95127	9.33274
8.27	68.3929	2.87576	9.09395	8.72	76.0384	2.95296	9.33809
8.28	68.5584	2.87750	9.09945	8.73	76.2129	2.95466	9.34345
8.29	68.7241	2.87924	9.10494	8.74	76.3876	2.95635	9.34880
8.30	68.8900	2.88097	9.11043	8.75	76.5625	2.95804	9.35414
8.31	69.0561	2.88271	9.11592	8.76	76.7376	2.95973	9.35949
8.32	69.2224	2.88444	9.12140	8.77	76.9129	2.96142	9.36483
8.33	69.3889	2.88617	9.12688	8.78	77.0884	2.96311	9.37017
8.34	69.5556	2.88791	9.13236	8.79	77.2641	2.96479	9.37550
8.35	69.7225	2.88964	9.13783	8.80	77.4400	2.96648	9.38083
8.36	69.8896	2.89137	9.14330	8.81	77.6161	2.96816	9.38616
8.37	70.0569	2.89310	9.14877	8.82	77.7924	2.96985	9.39149
8.38	70.2244	2.89482	9.15423	8.83	77.9689	2.97153	9.39681
8.39	70.3921	2.89655	9.15969	8.84	78.1456	2.97321	9.40213
8.40	70.5600	2.89828	9.16515	8.85	78.3225	2.97489	9.40744
8.41	70.7281	2.90000	9.17061	8.86	78.4996	2.97658	9.41276
8.42	70.8964	2.90172	9.17606	8.87	78.6769	2.97825	9.41807
8.43	71.0649	2.90345	9.18150	8.88	78.8544	2.97993	9.42338
8.44	71.2336	2.90517	9.18695	8.89	79.0321	2.98161	9.42868
8.45	71.4025	2.90689	9.19239	8.90	79.2100	2.98329	9.43398
8.46	71.5716	2.90861	9.19783	8.91	79.3881	2.98496	9.43928
8.47	71.7409	2.91033	9.20326	8.92	79.5664	2.98664	9.44458
8.48	71.9104	2.91204	9.20869	8.93	79.7449	2.98831	9.44987
8.49	72.0801	2.91376	9.21412	8.94	79.9236	2.98998	9.45516
8.50	72.2500	2.91548	9.21954	8.95	80.1025	2.99166	9.46044
8.51	72.4201	2.91719	9.22497	8.96	80.2816	2.99333	9.46573
8.52	72.5904	2.91890	9.23038	8.97	80.4609	2.99500	9.47101
8.53	72.7609	2.92062	9.23580	8.98	80.6404	2.99666	9.47629
8.54	72.9316	2.92233	9.24121	8.99	80.8201	2.99833	9.48156
8.55	73.1025	2.92404	9.24662	9.00	81.0000	3.00000	9.48683
8.56	73.2736	2.92575	9.25203	9.01	81.1801	3.00167	9.49210
8.57	73.4449	2.92746	9.25743	9.02	81.3604	3.00333	9.49737
8.58	73.6164	2.92916	9.26283	9.03	81.5409	3.00500	9.50263
8.59	73.7881	2.93087	9.26823	9.04	81.7216	3.00666	9.50789
8.60	73.9600	2.93258	9.27362	9.05	81.9025	3.00832	9.51315
8.61	74.1321	2.93428	9.27901	9.06	82.0836	3.00998	9.51840
8.62	74.3044	2.93598	9.28440	9.07	82.2649	3.01164	9.52365
8.63	74.4769	2.93769	9.28978	9.08	82.4464	3.01330	9.52890
8.64	74.6496	2.93939	9.29516	9.09	82.6281	3.01496	9.53415

Tabla A.10. Cuadrados y raíces cuadradas (continuación).

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	N	N^2	$\sqrt{N}$	$\sqrt{10N}$
9.10	82.8100	3.01662	9.53939	9.55	91.2025	3.09031	9.77241
9.11	82.9921	3.01828	9.54463	9.56	91.3936	3.09192	9.77753
9.12	83.1744	3.01993	9.54987	9.57	91.5849	3.09354	9.78264
9.13	83.3569	3.02159	9.55510	9.58	91.7764	3.09516	9.78775
9 14	83.5396	3.02324	9.56033	9.59	91.9681	3.09677	9.79285
9.15	83.7225	3.02490	9.56556	9.60	92.1600	3.09839	9.79796
9.16	83.9056	3.02655	9.57079	9.61	92.3521	3.10000	9.80306
9.17	84.0889	3.02820	9.57601	9.62	92.5444	3.10161	9.80816
9.18	84.2724	3.02985	9.58123	9.63	92.7369	3.10322	9.81326
9.19	84.4561	3.03150	9.58645	9.64	92.9296	3.10483	9.81835
9.20	84.6400	3.03315	9.59166	9.65	93.1225	3.10644	9.82344
9.21	84.8241	3.03480	9.59687	9.66	93.3156	3.10805	9.82853
9.22	85.0084	3.03645	9.60208	9.67	93.5089	3.10966	9.83362
9.23	85.1929	3.03809	9.60729	9.68	93.7024	3.11127	9.83870
9.24	85.3776	3.03974	9.61249	9.69	93.8961	3.11288	9.84378
9.25	85.5625	3.04138	9.61769	9.70	94.0900	3.11448	9.84886
9.26	85.7476	3.04302	9.62289	9.71	94.2841	3.11609	9.85393
9.27	85.9329	3.04467	9.62808	9.72	94.4784	3.11769	9.85901
9.28	86.1184	3.04631	9.63328	9.73	94.6729	3.11929	9.86408
9.29	86.3041	3.04795	9.63846	9.74	94.8676	3.12090	9.86914
9.30	86.4900	3.04959	9.64365	9.75	95.0625	3.12250	9.87421
9.31	86.6761	3.05123	9.64883	9.76	95.2576	3.12410	9.87927
9.32	86.8624	3.05287	9.65401	9.77	95.4529	3.12570	9.88433
9.33	87.0489	3.05450	9.65919	9.78	95.6484	3.12730	9.88939
9.34	87.2356	3.05614	9.66437	9.79	95.8441	3.12890	9.89444
9.35	87.4225	3.05778	9.66954	9.80	96.0400	3.13050	9.89949
9.36	87.6096	3.05941	9.67471	9.81	96.2361	3.13209	9.90454
9.37	87.7969	3.06105	9.67988	9.82	96.4324	3.13369	9.90959
9.38	87.9844	3.06268	9.68504	9.83	96.6289	3.13528	9.91464
9.39	88.1721	3.06431	9.69020	9.84	96.8256	3.13688	9.91968
9.40	88.3600	3.06594	9.69536	9.85	97.0225	3.13847	9.92472
9.41	88.5481	3.06757	9.70052	9.86	97.2196	3.14006	9.92975
9.42	88.7364	3.06920	9.70567	9.87	97.4169	3.14166	9.93479
9.43	88.9249	3.07083	9.71082	9.88	97.6144	3.14325	9.93982
9.44	89.1136	3.07246	9.71597	9.89	97.8121	3.14484	9.94485
9.45	89.3025	3.07409	9.72111	9.90	98.0100	3.14643	9.94987
9.46	89.4916	3.07571	9.72625	9.91	98.2081	3.14802	9.95490
9.47	89.6809	3.07734	9.73139	9.92	98.4064	3.14960	9.95992
9.48	89.8704	3.07896	9.73653	9.93	98.6049	3.15119	9.96494
9.49	90.0601	3.08058	9.74166	9.94	98.8036	3.15278	9.96995
9.50	90.2500	3.08221	9.74679	9.95	99.0025	3.15436	9.97497
9.51	90.4401	3.08383	9.75192	9.96	99.2016	3.15595	9.97998
9.52	90.6304	3.08545	9.75705	9.97	99.4009	3.15753	9.98499
9.53	90.8209	3.08707	9.76217	9.98	99.6004	3.15911	9.98999
9.54	91.0116	3.08869	9.76729	9.99	99.8001	3.16070	9.99500

Reproducida de la obra *Introduction to Probability and Statistics* de Henry L. Alder y Edward B. Roessler, 1960, W.H. Freeman & Company, con autorización de los autores.

Tabla A.11. Coeficientes, divisores y valores de K para ajustar las curvas cuárticas a datos de intervalos iguales, y para separar la suma de cuadrados.

n:	3		4			5				6			
	c_1	c_2	c_1	c_2	c_3	c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
	−1	1	−3	1	−1	−2	2	−1	1	−5	5	−5	1
	0	−2	−1	−1	3	−1	−1	2	−4	−3	−1	7	−3
	1	1	1	−1	−3	0	−2	0	6	−1	−4	4	2
			3	1	1	1	−1	−2	−4	1	−4	−4	2
						2	2	1	1	3	−1	−7	−3
										5	5	5	1
Divisores													
	2	6	20	4	20	10	14	10	70	70	84	180	28
K_1		1/3		5/16				1/7				5/96	
K_2		1/2		1/20				1/10				1/70	
K_3				41/240				17/60				101/4320	
K_4		1/2		1/16				1/14				1/224	
K_5				1/48				1/12				1/864	
K_6								1/24				1/768	
K_7								31/168				95/2688	
K_8								3/35				27/256	

n:	7				8				9			
	c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
	−3	5	−1	3	−7	7	−7	7	−4	28	−14	14
	−2	0	1	−7	−5	1	5	−13	−3	7	7	−21
	−1	−3	1	1	−3	−3	7	−3	−2	−8	13	−11
	0	−4	0	6	−1	−5	3	9	−1	−17	9	9
	1	−3	−1	1	1	−5	−3	9	0	−20	0	18
	2	0	−1	−7	3	−3	−7	−3	1	−17	−9	9
	3	5	1	3	5	1	−5	−13	2	−8	−13	−11
					7	7	7	7	3	7	−7	−21
									4	28	14	14
Divisores												
	28	84	6	154	168	168	264	616	60	2772	990	2002
K_1				1/21				1/32				5/693
K_2				1/28				1/168				1/60
K_3				7/36				37/3168				59/5940
K_4				1/84				1/672				1/924
K_5				1/36				1/3168				1/1188
K_6				1/264				1/16896				1/3432
K_7				67/1848				179/59136				115/24024
K_8				3/77				9/512				9/1001

Tabla A.11. Continuación.

n:	10			11			
c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
−9	6	−42	18	−5	15	−30	6
−7	2	14	−22	−4	6	6	−6
−5	−1	35	−17	−3	−1	22	−6
−3	−3	31	3	−2	−6	23	−1
−1	−4	12	18	−1	−9	14	4
1	−4	−12	18	0	−10	0	6
3	−3	−31	3	1	−9	−14	4
5	−1	−35	−17	2	−6	−23	−1
7	2	−14	−22	3	−1	−22	−6
9	6	42	18	4	6	−6	−6
				5	15	30	6
Divisores							
330	132	8580	2860	110	858	4290	286
K_1			1/32			5/429	
K_2			1/330			1/110	
K_3			293/205920			89/25740	
K_4			1/1056			1/858	
K_5			1/41184			1/5148	
K_6			1/109824			1/3432	
K_7			41/54912			25/3432	
K_8			9/1280			3/143	

n:	12			13			
c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
−11	55	−33	33	−6	22	−11	99
−9	25	3	−27	−5	11	0	−66
−7	1	21	−33	−4	2	6	−96
−5	−17	25	−13	−3	−5	8	−54
−3	−29	19	12	−2	−10	7	11
−1	−35	7	28	−1	−13	4	64
1	−35	−7	28	0	−14	0	84
3	−29	−19	12	1	−13	−4	64
5	−17	−25	−13	2	−10	−7	11
7	1	−21	−33	3	−5	−8	−54
9	25	−3	−27	4	2	−6	−96
11	55	33	33	5	11	0	−66
				6	22	11	99
Divisores							
572	12012	5148	8008	182	2002	572	68068
K_1			1/336			1/143	
K_2			1/572			1/182	
K_3			85/61776			25/3432	
K_4			1/16016			1/2002	
K_5			1/61776			1/3432	
K_6			1/439296			1/116688	
K_7			419/1537536			19/62832	
K_8			27/7168			3/2431	

Tabla A.11. Continuación.

n:	14				15			
	c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
	−13	13	−143	143	−7	91	−91	1001
	−11	7	−11	−77	−6	52	−13	−429
	−9	2	66	−132	−5	19	35	−869
	−7	−2	98	−92	−4	−8	58	−704
	−5	−5	95	−13	−3	−29	61	−249
	−3	−7	63	63	−2	−44	49	251
	−1	−8	24	108	−1	−53	27	621
	1	−8	−24	108	0	−56	0	756
	3	−7	−67	63	1	−53	−27	621
	5	−5	−95	−13	2	−44	−49	251
	7	−2	−98	−92	3	−29	−61	−249
	9	2	−66	−132	4	−8	−58	−704
	11	7	11	−77	5	19	−35	−869
	13	13	143	143	6	52	13	−429
					7	91	91	1001
Divisores								
	910	728	97240	136136	280	37128	39780	6466460
K_1				5/448				1/663
K_2				1/910				1/280
K_3				581/2333760				167/238680
K_4				1/5824				1/12376
K_5				1/466752				1/47736
K_6				1/3734016				1/2217072
K_7				575/13069056				331/15519504
K_8				3/3584				27/230945

Tabla A.11. Continuación.

n:	16			17			
c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
−15	35	−455	273	−8	40	−28	52
−13	21	−91	−91	−7	25	−7	−13
−11	9	143	−221	−6	12	7	−39
−9	−1	267	−201	−5	1	15	−39
−7	−9	301	−101	−4	−8	18	−24
−5	−15	265	23	−3	−15	17	−3
−3	−19	179	129	−2	−20	13	17
−1	−21	63	189	−1	−23	7	31
1	−21	−63	189	0	−24	0	36
3	−19	−179	129	1	−23	−7	31
5	−15	−265	23	2	−20	−13	17
7	−9	−301	−101	3	−15	−17	−3
9	−1	−267	−201	4	−8	−18	−24
11	9	−143	−221	5	1	−15	−39
13	21	91	−91	6	12	−7	−39
15	35	455	273	7	25	7	−13
				8	40	28	52
Divisores							
1360	5712	1007760	470288	408	7752	3876	16796
K_1			5/1344				1/323
K_2			1/1360				1/408
K_3			761/12093120				43/23256
K_4			1/22848				1/7752
K_5			1/2418624				1/23256
K_6			1/12899328				1/201552
K_7			755/45147648				61/201552
K_8			3/7168				9/4199

Tabla A.11. Continuación.

n:	18			19			
c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
−17	68	−68	68	−9	51	−204	612
−15	44	−20	−12	−8	34	−68	−68
−13	23	13	−47	−7	19	28	−388
−11	5	33	−51	−6	6	89	−453
−9	−10	42	−36	−5	−5	120	−354
−7	−22	42	−12	−4	−14	126	−168
−5	−31	35	13	−3	−21	112	42
−3	−37	23	33	−2	−26	83	227
−1	−40	8	44	−1	−29	44	352
1	−40	−8	44	0	−30	0	396
3	−37	−23	33	1	−29	−44	352
5	−31	−35	13	2	−26	−83	227
7	−22	−42	−12	3	−21	−112	42
9	−10	−42	−36	4	−14	−126	−168
11	5	−33	−51	5	−5	−120	−354
13	23	−13	−47	6	6	−89	−453
15	44	20	−12	7	19	−28	−388
17	68	68	68	8	34	68	−68
				9	51	204	612
Divisores							
1938	23256	23256	28424	570	13466	213180	2288132
K_1			1/576				5/2261
K_2			1/1938				1/570
K_3			193/558144				269/1279080
K_4			1/62016				1/13566
K_5			1/558144				1/255816
K_6			1/5457408				1/3922512
K_7			137/2728704				535/27457584
K_8			9/5632				9/52003

Tabla A.11. Continuación.

n:	20			21			
c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
−19	57	−969	1938	−10	190	−285	969
−17	39	−357	−102	−9	133	−114	0
−15	23	85	−1122	−8	82	12	−510
−13	9	377	−1402	−7	37	98	−680
−11	−3	539	−1187	−6	−2	149	−615
−9	−13	591	−687	−5	−35	170	−406
−7	−21	553	−77	−4	−62	166	−130
−5	−27	445	503	−3	−83	142	150
−3	−31	287	948	−2	−98	103	385
−1	−33	99	1188	−1	−107	54	540
1	−33	−99	1188	0	−110	0	594
3	−31	−287	948	1	−107	−54	540
5	−27	−445	503	2	−98	−103	385
7	−21	−553	−77	3	−83	−142	150
9	−13	−591	−687	4	−62	−166	−130
11	−3	−539	−1187	5	−35	−170	−406
13	9	−377	−1402	6	−2	−149	−615
15	23	−85	−1122	7	37	−98	−680
17	39	357	−102	8	82	−12	−510
19	57	969	1938	9	133	114	0
				10	190	285	969
Divisores							
2660	17556	4903140	22881320	770	201894	432630	5720330
K_1			1/528				5/9177
K_2			1/2660				1/770
K_3			1193/58837680				329/2595780
K_4			1/70224				1/67298
K_5			1/11767536				1/519156
K_6			1/251040768				1/9806280
K_7			1187/878642688				131/13728792
K_8			3/56320				27/260015

Tabla A.11. Continuación.

n:	22			23			
c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
−21	35	−133	1197	−11	77	−77	1463
−19	25	−57	57	−10	56	−35	133
−17	16	0	−570	−9	37	−3	−627
−15	8	40	−810	−8	20	20	−950
−13	1	65	−775	−7	5	35	−955
−11	−5	77	−563	−6	−8	43	−747
−9	−10	78	−258	−5	−19	45	−417
−7	−14	70	70	−4	−28	42	−42
−5	−17	55	365	−3	−35	35	315
−3	−19	35	585	−2	−40	25	605
−1	−20	12	702	−1	−43	13	793
1	−20	−12	702	0	−44	0	858
3	−19	−35	585	1	−43	−13	793
5	−17	−55	365	2	−40	−25	605
7	−14	−70	70	3	−35	−35	315
9	−10	−78	−258	4	−28	−42	−42
11	−5	−77	−563	5	−19	−45	−417
13	1	−65	−775	6	−8	−43	−747
15	8	−40	−810	7	5	−35	−955
17	16	0	−570	8	20	−20	−950
19	25	57	57	9	37	3	−627
21	35	133	1197	10	56	35	133
				11	77	77	1463
Divisores							
3542	7084	96140	8748740	1012	35420	32890	131231100
K_1			1/352				1/805
K_2			1/3542				1/1012
K_3			289/2307360				79/197340
K_4			1/56672				1/35520
K_5			1/2307360				1/197340
K_6			1/239965440				1/22496760
K_7			1439/839879040				787/157477320
K_8			3/36608				1/15295

Tabla A.11. Continuación.

n:	24			25			
c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
−23	253	−1771	253	−12	92	−506	1518
−21	187	−847	33	−11	69	−253	253
−19	127	−133	−97	−10	48	−55	−517
−17	73	391	−157	−9	29	93	−897
−15	25	745	−165	−8	12	196	−982
−13	−17	949	−137	−7	−3	259	−857
−11	−53	1023	−87	−6	−16	287	−597
−9	−83	987	−27	−5	−27	285	−267
−7	−107	861	33	−4	−36	258	78
−5	−125	665	85	−3	−43	211	393
−3	−137	419	123	−2	−48	149	643
−1	−143	143	143	−1	−51	77	803
1	−143	−143	143	0	−52	0	858
3	−137	−419	123	1	−51	−77	803
5	−125	−665	85	2	−48	−149	643
7	−107	−861	33	3	−43	−211	393
9	−83	−987	−27	4	−36	−258	78
11	−53	−1023	−87	5	−27	−285	−267
13	−17	−949	−137	6	−16	−287	−597
15	25	−745	−165	7	−3	−259	−857
17	73	−391	−157	8	12	−196	−982
19	127	133	−97	9	29	−93	−897
21	187	847	33	10	48	55	−517
23	253	1771	253	11	69	253	253
				12	92	506	1518
Divisores							
4600	394680	17760600	394680	1300	53820	1480050	14307150
K_1			5/13728				1/1035
K_2			1/4600				1/1300
K_3			1721/213127200				467/8880300
K_4			1/526240				1/53820
K_5			1/42625440				1/1776060
K_6			1/75778560				1/34337160
K_7			49/7577856				133/34337160
K_8			27/73216				1/16675

Tabla A.12. Coeficientes para ajustar curvas periódicas, y para separar sumas de cuadrados, para datos tomados a intervalos de tiempo iguales durante un ciclo completo.

Valores de X para n †												
4	6	8	12	24	U_1	V_1	U_2	V_2	U_3	V_3	U_4	V_4
0	0	0	0	0	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
				1	0.966	0.259	0.866	0.500	0.707	0.707	0.500	0.866
			1	2	0.866	0.500	0.500	0.866	0.000	1.000	− 0.500	0.866
		1		3	0.707	0.707	0.000	1.000	− 0.707	0.707	− 1.000	0.000
	1		2	4	0.500	0.866	− 0.500	0.866	− 1.000	0.000	− 0.500	− 0.866
				5	0.259	0.966	− 0.866	0.500	− 0.707	− 0.707	0.500	− 0.866
1		2	3	6	0.000	1.000	− 1.000	0.000	0.000	− 1.000	1.000	0.000
				7	− 0.259	0.966	− 0.866	− 0.500	0.707	− 0.707	0.500	0.866
	2		4	8	− 0.500	0.866	− 0.500	− 0.866	1.000	0.000	− 0.500	0.866
		3		9	− 0.707	0.707	0.000	− 1.000	0.707	0.707	− 1.00	0.000
			5	10	− 0.866	0.500	0.500	− 0.866	0.000	1.000	− 0.500	− 0.866
				11	− 0.966	0.259	0.866	− 0.500	− 0.707	0.707	0.500	− 0.866
2	3	4	6	12	− 1.000	0.000	1.000	0.000	− 1.000	0.000	1.000	1.000
				13	− 0.966	− 0.259	0.866	0.500	− 0.707	− 0.707	0.500	0.866
			7	14	− 0.866	− 0.500	0.500	0.866	0.000	− 1.000	− 0.500	0.866
		5		15	− 0.707	− 0.707	0.000	1.000	0.707	− 0.707	− 0.000	0.000
	4		8	16	− 0.500	− 0.866	− 0.500	0.866	1.000	0.000	− 0.500	− 0.866
				17	− 0.259	− 0.966	− 0.866	0.500	0.707	0.707	0.500	− 0.866
3		6	9	18	0.000	− 1.000	− 1.000	0.000	0.000	1.000	1.000	1.000
				19	0.259	− 0.966	− 0.866	− 0.500	− 0.707	0.707	0.500	0.866
	5		10	20	0.500	− 0.866	− 0.500	− 0.866	− 1.000	1.000	− 0.500	0.866
		7		21	0.707	− 0.707	1.000	− 1.000	− 0.707	− 0.707	− 1.000	0.000
			11	22	0.866	− 0.500	1.500	− 0.866	1.000	− 1.000	− 0.500	− 0.866
				23	0.966	− 0.259	0.866	− 0.500	0.707	0.707	0.500	− 0.866

†Para un valor dado de n, úsense solamente las líneas de la tabla para las cuales se han dado valores de X. Cuando n = 4, empléense sólo las columnas para U_2. Cuando n = 6, utilícense sólo las columnas para U_3. Cuando n = 8, úsense sólo las columnas para U_4.

Valores de X para n = 7

	U_1	V_1	U_2	V_2	U_3	V_3
0	− 1.000	0.000	1.000	0.000	1.000	0.000
1	0.623	0.782	− 0.223	0.975	0.901	0.434
2	− 0.223	0.975	0.901	− 0.434	0.623	− 0.782
3	− 0.901	0.434	0.623	− 0.782	− 0.223	0.975
4	− 0.901	− 0.434	0.623	0.782	− 0.223	− 0.975
5	− 0.223	− 0.975	− 0.901	0.434	0.623	0.782
6	0.623	− 0.782	− 0.223	− 0.975	− 0.901	− 0.434

Tabla A.12. Continuación.

Valor de X para n = 52.

	U_1	V_1	U_2	V_2	U_3	V_3	U_4	V_4
0.	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
1	0.993	0.121	0.971	0.239	0.935	0.355	0.885	0.465
2	0.971	0.239	0.885	0.465	0.749	0.663	0.568	0.823
3	0.935	0.355	0.749	0.663	0.465	0.885	0.121	0.993
4	0.885	0.465	0.568	0.823	0.121	0.993	− 0.355	0.935
5	0.823	0.568	0.355	0.935	− 0.239	0.971	− 0.749	0.663
6	0.749	0.663	0.121	0.993	− 0.568	0.823	− 0.971	0.239
7	0.663	0.749	− 0.121	0.993	0.823	0.568	− 0.971	− 0.239
8	0.568	0.823	− 0.355	0.935	− 0.971	0.239	− 0.749	− 0.663
9	0.465	0.885	− 0.568	0.823	− 0.993	− 0.121	− 0.355	− 0.935
10	0.335	0.935	− 0.749	0.885	− 0.885	− 0.465	0.121	− 0.993
11	0.239	0.971	− 0.885	0.465	− 0.663	− 0.749	0.568	− 0.823
12	0.121	0.993	− 0.971	0.239	− 0.355	− 0.935	0.885	− 0.465
13	0.000	1.000	− 1.000	0.000	0.000	− 1.000	1.000	0.000
14	− 0.121	0.993	− 0.971	− 0.239	0.355	− 0.935	0.885	0.465
15	− 0.239	0.971	− 0.885	− 0.465	0.663	− 0.749	0.568	0.823
16	− 0.355	0.935	− 0.749	− 0.663	0.885	− 0.465	0.121	0.993
17	− 0.465	0.885	− 0.568	− 0.823	0.993	− 0.121	− 0.355	0.935
18	− 0.568	0.823	− 0.355	− 0.935	0.971	0.239	− 0.749	0.663
19	− 0.663	0.749	− 0.121	− 0.993	0.823	0.568	− 0.971	0.239
20	− 0.749	0.663	0.121	− 0.993	0.568	0.823	− 0.971	− 0.239
21	− 0.823	0.568	0.355	− 0.935	0.239	0.971	− 0.749	− 0.663
22	− 0.885	0.465	0.568	− 0.823	− 0.121	0.993	− 0.355	− 0.935
23	− 0.935	0.355	0.749	− 0.663	− 0.465	0.885	0.121	− 0.993
24	− 0.971	0.239	0.885	− 0.465	− 0.749	0.663	0.568	− 0.823
25	− 0.993	0.121	0.971	− 0.239	− 0.935	0.355	0.885	− 0.465
26	− 1.000	0.000	1.000	0.000	− 1.000	0.000	1.000	0.000
27	− 0.993	− 0.121	0.971	0.239	− 0.935	− 0.355	0.885	0.465
28	− 0.971	− 0.239	0.885	0.465	− 0.749	− 0.663	0.568	0.823
29	− 0.935	− 0.355	0.749	0.663	− 0.465	− 0.885	0.121	0.993
30	− 0.885	− 0.465	0.568	0.823	− 0.121	− 0.121	− 0.355	0.935
31	− 0.823	− 0.568	0.355	0.935	0.239	− 0.971	− 0.749	0.663
32	− 0.749	− 0.663	0.121	0.993	0.568	− 0.823	− 0.971	0.239
33	− 0.663	− 0.749	− 0.121	0.993	0.823	− 0.568	− 0.971	− 0.239
34	− 0.568	− 0.823	− 0.355	0.935	0.971	− 0.239	− 0.749	− 0.663
35	− 0.465	− 0.885	− 0.568	0.823	0.993	0.121	− 0.355	− 0.935
36	− 0.355	− 0.935	− 0.749	0.663	0.885	0.465	0.121	− 0.993
37	− 0.239	− 0.971	− 0.885	0.465	0.663	0.749	0.568	− 0.823
38	− 0.121	− 0.993	− 0.971	0.239	0.355	0.935	0.885	− 0.465
39	0.000	− 1.000	− 1.000	0.000	0.000	1.000	1.000	0.000
40	0.121	− 0.993	− 0.971	− 0.239	− 0.355	0.935	0.885	0.465

Tabla A.12. Continuación.

Valor de X para n = 52.

	U_1	V_1	U_2	V_2	U_3	V_3	U_4	V_4
41	0.239	− 0.971	− 0.885	− 0.465	− 0.663	0.749	0.568	0.823
42	0.355	− 0.935	− 0.749	− 0.663	− 0.885	0.465	0.121	0.993
43	0.465	− 0.885	− 0.568	− 0.823	− 0.993	0.121	0.355	0.935
44	0.568	− 0.823	− 0.355	− 0.935	− 0.971	− 0.239	0.749	0.663
45	0.663	− 0.749	− 0.121	− 0.993	− 0.823	− 0.568	− 0.971	0.239
46	0.749	− 0.663	0.121	− 0.993	− 0.568	− 0.823	− 0.971	− 0.239
47	0.823	− 0.568	0.355	− 0.935	− 0.239	− 0.971	− 0.749	− 0.663
48	0.885	− 0.465	0.568	− 0.823	0.121	− 0.993	− 0.355	− 0.935
49	0.935	− 0.355	0.749	− 0.663	− 0.465	− 0.885	0.121	− 0.993
50	0.971	− 0.239	0.885	− 0.465	0.749	− 0.663	0.568	− 0.823
51	0.993	− 0.121	0.971	− 0.239	0.935	− 0.355	0.885	− 0.465

Tabla A.13. Lecturas seleccionadas

Alder, H. L. y E. V. Roessler, 1968, 4ª ed., *Introduction to Probability and Statistics* W.H. Freeman, & Co., San Francisco, pág. 333.

Cochran, W.G. y G. M. Cox, 1964, 2ª ed., *Experimental Design.* John Wiley & Sons, Inc., Nueva York, pág. 617.

Finney, D. J., 1962, 2ª ed., *An Introduction to Statistical Science in Agriculture.* Munksgaard, Copenhague, Dinamarca, y Oliver & Boyd Ltda., Edimburgo, Escocia, pág. 216.

LeClerg, E. L., W. H. Leonard y A. G. Clark, 1962, 2ª ed., *Field Plot Technique*. Burgess Publishing Co., Minneapolis, Minnesota, pág. 373.

Snedecor, G.W. y W.G. Cochran, 1967, 6ª ed., *Statistical Methods.* The Iowa State University Press, Ames, Iowa, pág. 593.

Sokal, R. R. y F. J. Rohlf, 1969, *Biometry, The principles and Practice of Statistics in Biological Research.*. W. H. Freeman & Co., San Francisco, pág. 776.

Steel, R. G. D. y J. H. Torrie, 1960, *Principles and Procedures of Statistics With Special Reference to the Biological Sciences.* McGraw Hill Book Co., Inc., Nueva York, pág. 481.