Extraído de: "Los Matemáticos en la vida cotidiana" 3ª Ed. Ediciones de la UAM' 1999

Capítulo

5

La producción de los datos

MATEMATICAS

Las noticias aparecen a menudo en forma de números. Los titulares anuncian que el desempleo ha bajado hasta el 7,2%. Las encuestas de la empresa Gallup afirman que el 45% de los ciudadanos americanos tienen miedo de salir de noche a causa de la delincuencia. ¿De dónde provienen estos números? La mayor parte de la gente no es encuestada personalmente para determinar su estatus laboral. Y la empresa Gallup solamente pregunta a unos cuantos estadounidenses si el miedo a ser atracados les mantiene en casa por la noche.

El porcentaje de desempleados y otros datos de la población activa son estimados en Estados Unidos por el Departamento de Estadísticas Laborales con datos obtenidos a partir del Estudio de la Población Actual. Este estudio lo realiza el Departamento del Censo, que entrevista a 60.000 hogares cada mes. Considerando que Estados Unidos tiene cerca de 95 millones de hogares, resulta extraordinario que datos acerca de un grupo tan pequeño —menos de uno de cada 1.000 hogares— puedan representar el porcentaje de desempleados de toda la nación. ¿Cómo se pueden justificar conclusiones para un grupo numeroso usando datos extraídos de un número pequeño de personas?

Los números afectan nuestras vidas aun cuando no aparezcan en los titulares. Cuando una medicina para el reúma tiene que ser retirada de las farmacias porque hay serias dudas acerca de su seguridad, o cuando el Departamento de Alimentos y Medicinas aprueba un nuevo fármaco para el tratamiento del SIDA, sus decisiones están basadas en experimentos clínicos realizados en un número relativamente pequeño de pacientes. La confianza pública acerca de la seguridad y efectividad de nuevas medicinas está basada en conclusiones estadísticas, o *inferencias*, extraídas de datos clínicos. Una vez más, debemos extraer conclusiones fiables a partir de una cantidad relativamente pequeña de datos.

Cualquier uso de los números como evidencia depende de la correcta producción de los datos. El método directo para calcular el porcentaje de desempleados de una nación consistiría en entrevistar a cada una de las personas. El coste de esta operación a gran escala sería enorme y el esfuerzo requerido lo hace poco práctico. En su lugar, debemos fiarnos de información obtenida a partir de una parte de la población, en lenguaje estadístico, una *muestra*.

A menudo extraemos conclusiones acerca del todo basándonos en una muestra. Todos hemos probado una cucharada de sopa y juzgado el contenido de toda la taza basándonos en ella. Pero una taza de sopa es homogénea; el sabor de una cucharada repre-

senta a toda ella. Otro tipo de observaciones, tales como el miedo hacia la delincuencia o la reacción producida por un medicamento, no se pueden medir tan fácilmente. Varían de persona en persona o de región en región.

La parte central de la estadística consiste en manejar este tipo de variabilidad. El primer paso es producir datos que representen a un grupo numeroso de individuos. Para hacer esto, debemos establecer cuidadosamente de qué grupo queremos tener información. Los estadísticos llaman a este grupo la población.

Por ejemplo, si queremos conocer el porcentaje de desempleados, debemos determinar la población que tratamos de describir. ¿Qué grupos de edad estarán incluidos? ¿Incluiremos extranjeros o personas encarceladas? ¿Y los estudiantes? El Departamento de Estadísticas Laborales debe responder a estas preguntas antes de recoger la información mensual sobre el desempleo (véase Primer Plano 5.1). De hecho, el actual censo de los Estados Unidos de América define la población como el conjunto de todos los residentes del país (ya sean ciudadanos o no) que tengan 16 o más años. El porcentaje de desempleo que se publica en los periódicos y en otros reportajes se refiere siempre a esta población específica.

Los experimentos clínicos de un nuevo fármaco para el tratamiento de úlceras también tienen en mente una población específica. Esta población podría consistir en todos los individuos que tienen úlcera, o puede restringirse a los pacientes con un tipo especial de ella, o a aquéllos en un cierto grupo de edad. El plan detallado del experimento clínico debe especificar exactamente cuál es la población elegida para realizar el estudio.

Muestreo

No es necesario que una población consista de personas; puede consistir de animales o de cosas. Supongamos que queremos probar una población de fusibles que se acaban de fabricar para confirmar, como debe suceder, que se funden cuando pasa por ellos una corriente elevada. En este caso, si recogiéramos información acerca de cada miembro de la población se destruiría la población entera. En otros casos, como, por ejemplo, para hacer el inventario de un gran almacén, el aburrimiento y la fatiga nos impedirían obtener un resultado preciso. Y el coste de producir los datos requiere a menudo que no examinemos a toda la población —ni siquiera el Gobierno de Estados Unidos puede permitirse entrevistar a todos los individuos cada mes para medir el desempleo. Por todas estas razones, normalmente recogemos información acerca de algunos elementos seleccionados de la población. La parte de la población que se usa para extraer conclusiones acerca de toda ella se llama **muestra**.

¿Cómo podemos elegir una muestra que represente verdaderamente a la población? El método más sencillo —pero no el mejor— para seleccionar una muestra es elegir los individuos más cercanos. Si estamos interesados en encontrar, por ejemplo, cuántas personas tienen trabajo, podemos ir a un centro comercial y preguntar a sus visitantes si tienen trabajo. Una muestra en la que se elige a los miembros de la población que estén más cercanos se llama *muestra de conveniencia*.

Muestras de conveniencia producen a menudo datos que no representan a la población. Las personas en un centro comercial suelen ser mas prósperas que el ciudadano medio; también es más probable que sean jóvenes o jubilados. Y, cuando decidimos a qué personas entrevistar, tendemos a elegir personas bien vestidas, con un salario medio, mientras que evitamos las mal vestidas, hostiles, o huraños. Resumiendo, nuestras entrevistas en un centro comercial no ofrecen una muestra que sea representativa de toda la población, y, por tanto, no podrán representar de manera precisa el porcentaje de desempleados.

Cuando nos permitimos tomar una muestra de conveniencia nos encontramos —ya sea consciente o inconscientemente— con una muestra basada en amistad, apariencia o nivel económico, dejando, por tanto, partes de la población sin representar. Nuestra muestra del centro comercial estará probablemente sesgada hacia la clase media y los jubilados, y no servirá para los trabajadores y los pobres. Este problema surgirá cada vez que tomemos este tipo de muestras. Es decir, es un error sistemático producido por un método de muestreo erróneo, y no consecuencia de mala fortuna en la muestra. Esta diferencia sistemática entre los resultados obtenidos con la muestra y los datos correctos de toda la población se llama **sesgo**. Para producir datos precisos debemos tomar precauciones para eliminar el sesgo. En particular, los estadísticos gastan mucho tiempo para eliminar cualquier preferencia personal en la selección de la muestra con la que se desea trabajar.

PRIMER PLANO 5.1 El Estadístico del Gobierno

El comisario de las estadísticas laborales es uno de los estadísticos más influyentes de la nación. En su cargo de director del Departamento de Estadísticas Laborales, el comisario supervisa la recogida e interpretación de los datos del desempleo, del salario, y de muchas otras tendencias económicas y sociales. Estos datos tienen grandes repercusiones en la economía de Estados Unidos. Por ejemplo, los precios de venta al por menor recogidos por este Departamento se utilizan para ajustar muchos pagos estatales y privados, incluidos los de la Seguridad Social y los salarios mínimos, a los efectos de la inflación.

Los datos recogidos por el Departamento de Estadísticas Laborales tienen a menudo significado político, como cuando un informe sobre el crecimiento del desempleo se hace público justo antes de unas elecciones. Por esta razón, el Departamento debe mantenerse independiente de influencias políticas. Para salvaguardar su independencia, el comisario es nombrado por el Presidente y confirmado por el Senado por un período de cuatro años. El comisario debe tener experiencia en estadística, habilidad administrativa y facilidad para trabajar tanto con el Congreso como con el Presidente.

La doctora Janet Norwood ha servido durante tres mandatos como comisaria, desde 1979 hasta 1991, bajo tres Presidentes. Cuando abandonó este puesto, el periódico *New York Times* dijo que ella dejó «una casi legendaria reputación de no alineamiento y alabanzas entre las que se incluyen el que un senador dijera que ella era un 'tesoro nacional'» (31 de diciembre de 1993). La doctora Norwood dijo: «Ha habido momentos en el pasado en los que los comisarios han discrepado con el Ministro de Trabajo o, en algunos casos, con el Presidente. Guardamos nuestro profesionalismo con mucho cariño».

La doctora Janet Norwood, comisario del Departamento de Estadísticas de los Estados Unidos de América, 1979-1991.

EJEMPLO: Llamadas telefónicas voluntarias

La televisión utiliza a menudo las encuestas voluntarias, en las que los espectadores son invitados a declarar sus opiniones por teléfono. Algunas cadenas de televisión realizan encuestas diariamente haciendo una pregunta en las noticias de la tarde y comentando los resultados de las respuestas en las noticias de la noche. A los espectadores se les pide que llamen a números especiales indicando sus respuestas. Marcar uno de estos números significa una respuesta afirmativa y marcar otro negativa. El sistema funciona de manera que no es necesario hablar. La respuesta se graba una vez terminada la llamada telefónica. ▲

Las encuestas mediante llamadas telefónicas voluntarias tienen varias fuentes de sesgo. Familias sin teléfono quedan automáticamente excluidas. (A pesar de que el 94% de las familias en Estados Unidos tiene teléfono, alrededor de un cuarto en Alaska y un quinto en Mississipi no lo tiene). Puesto que estas llamadas tienen un pequeño coste, las personas con bajo salario pueden ser reacias a pagar una cuota por llamar a la cadena de televisión.

La fuente más seria de sesgo en las encuestas mediante llamadas telefónicas es la **respuesta voluntaria**; esto es, los encuestados se autoseleccionan. Sólo aquéllos que tienen tiempo y dinero para pagar la llamada son tenidos en cuenta. Cualquier muestra elegida mediante llamada voluntaria atrae a personas con fuertes opiniones, la mayoría de las veces de carácter negativo. Así pues, cuando el presentador pregunta a la audiencia si tienen miedo de salir por la noche a causa de la delincuencia, las personas más enojadas son más propensas a llamar que las menos enfadadas. La respuesta voluntaria es una fuente común y preocupante de sesgo en las encuestas.

También es posible manipular las encuestas mediante llamadas telefónicas voluntarias. Si el crimen es un asunto local con repercusiones políticas, un partido político puede pedir a todos sus miembros que pasen la tarde haciendo llamadas telefónicas. Asimismo, puesto que no es necesario hablar para registrar una respuesta, un computador puede ser programado para marcar el número en repetidas ocasiones.

La elección personal es una fuente común de sesgo en un muestreo. Ya sea expresada en la forma en que un encuestador o grupo de trabajo selecciona sus sujetos o en respuestas voluntarias, las elecciones personales pueden distorsionar los resultados. Para reducir sesgos, es vital reducir la influencia personal. La forma en que los estadísticos eliminan la elección personal es seleccionando una muestra *aleatoria*.

MUESTREO ALEATORIO

Imaginemos que tenemos una urna de cristal que contiene miles de bolas. Las bolas son todas idénticas, excepto que la mayoría son blancas y algunas son negras. Las bolas constituyen una población. Quizá las bolas representan la población laboral de Estados Unidos y las bolas negras representan aquellos que no tienen trabajo (véase la figura 5.1).

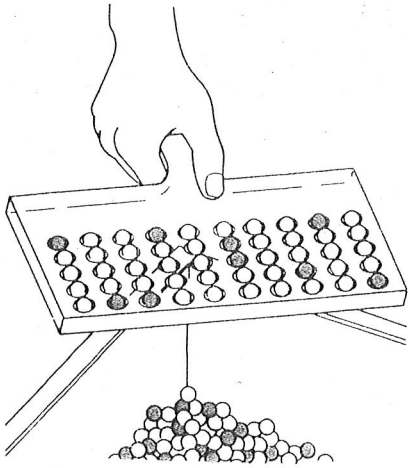

Figura 5.1 Extracción de una muestra aleatoria.

Nuestra tarea es estimar qué porcentaje de la población de bolas es negra sin examinar cada una de las bolas. Supongamos que hemos mezclado las bolas en la urna y a continuación seleccionado una muestra, usando una tabla con 50 huecos. Cada bola tiene la misma posibilidad de ser escogida, y cada grupo de 50 bolas también tiene la misma posibilidad de ser seleccionado que cualquier otro grupo de 50. Esta es una muestra aleatoria simple de tamaño 50.

> Una **muestra aleatoria simple** de tamaño n es una muestra extraída de tal manera que cada posible grupo de n elementos de la población tiene la misma posibilidad de estar en la muestra finalmente elegida.

Puesto que cada una de las bolas tiene la misma posibilidad de ser elegida, la muestra aleatoria simple ha eliminado el sesgo en la selección de la muestra. En un estudio del desempleo, una muestra aleatoria simple da a cada uno —rico o pobre, hembra o varón, con trabajo o desempleado— la misma posibilidad de ser seleccionado.

Supongamos que extraemos una muestra de nuestra urna de bolas y obtenemos 12 bolas negras. Puesto que 12 entre 50 es la fracción 12/50, o 0,24, el 24% de esta muestra son bolas negras. Los resultados de una muestra aleatoria simple están libres de sesgo, de manera que podemos utilizar el porcentaje de la muestra para estimar resultados acerca de toda la población. Si el 24% de las bolas en la muestra son negras, estimamos que el 24% de las bolas en la urna son negras. Esta es una técnica importante en estadística: *para estimar una característica de una población, se elige una muestra aleatoria simple y se usa la correspondiente característica de la muestra como estimación para toda la población.*

En teoría, ahora sabemos cómo estimar el porcentaje de desempleados de un país. Se ponen los nombres de todos los trabajadores en un sombrero, se mezclan bien, y se extraen 60.000 nombres. Puesto que ésta es una muestra aleatoria simple, el porcentaje de desempleados entre las 60.000 personas es una muestra no sesgada del porcentaje de desempleados de toda la fuerza laboral.

En la práctica, cuando una población es demasiado grande para ser puesta en un sombrero, necesitamos una forma más sencilla de seleccionar una muestra aleatoria simple. Por ejemplo, cada miembro de la población puede ser identificado con un número, y se puede elegir una muestra del conjunto de números de tal manera que cada número tenga la misma posibilidad de ser seleccionado. Si la población es relativamente pequeña, identificar y extraer muestras puede hacerse de manera sencilla con una tabla de números aleatorios como la de la Tabla 5.1. Para grandes poblaciones, un computador puede ser programado para que genere números aleatorios e imprima la muestra elegida.

> Una *tabla de dígitos aleatorios* es una sucesión de números, cada uno elegido independientemente de los otros de entre los dígitos 0, 1, 2, 3, 4, 5, 6, 7, 8, y 9, y donde cada dígito tiene la misma probabilidad de ser elegido.

Se puede pensar en escribir cada uno de los 10 dígitos en una tarjeta, poner las tarjetas en un sombrero, mezclarlas, y extraer una. Este es el primer número aleatorio en la tabla. Devolver la tarjeta extraída, mezclar de nuevo, y extraer otro número. Este es el segundo número aleatorio. Las extracciones y mezclas se continúan durante varias horas y tendremos así una tabla de dígitos aleatorios como la de la tabla 5.1. Los dígitos de la tabla 5.1 aparecen en grupos de 5 para que la tabla sea más fácil de leer y las filas están numeradas para que nos podamos referir a ellas, pero los grupos y columnas se han hecho porque resultan convenientes. Toda la tabla es una larga sucesión de dígitos elegidos aleatoriamente.

Son necesarias dos etapas para elegir una muestra aleatoria simple usando la tabla de dígitos aleatorios.

ETAPA 1. Dar a cada miembro de la población un valor numérico de la *misma longitud*. Dos dígitos son suficientes para numerar hasta 100 elementos, tres son suficientes para numerar hasta 1.000, etc.

ETAPA 2. Para elegir una muestra aleatoria simple, mirar en la tabla 5.1 grupos sucesivos de dígitos de la

TABLA 5.1 Dígitos aleatorios

101	03918	86495	47372	21870	28522	99445	38783	83307
102	10041	35095	66357	64569	08993	20429	28569	63809
103	43537	58268	80237	17407	89680	04655	24678	61932
104	64301	47201	31905	60410	80101	33382	95255	10353
105	43857	42186	77011	93839	28380	49296	63311	49713
106	91823	39794	47046	78563	89328	39478	04123	19287
107	34017	87878	35674	39212	98246	29735	09924	27893
108	49105	00755	39242	50472	39581	44036	54518	46865
109	72479	02741	75732	99808	02382	77201	44932	88978
110	84281	45650	28016	77753	39495	41847	19634	82681
111	61589	35486	59500	20060	89769	54870	75586	07853
112	25318	01995	87789	41212	74907	90734	31946	24921
113	40113	37395	51406	98099	43023	70195	07013	72306
114	58420	43526	15539	24845	15582	16780	95286	69021
115	18075	45894	09875	42869	20618	07699	80671	54287
116	52754	73124	93276	71521	59618	44966	37502	15570
117	05255	53579	08239	99174	75548	95776	42314	13093
118	76032	35569	28738	38092	74669	00749	17832	64855
119	97050	31553	32350	51491	53659	89336	36912	05292
120	29030	43074	84602	95131	22769	44680	68492	33987
121	28124	29686	63745	12313	15745	11570	50953	17149
122	97469	41277	90524	36459	22178	63785	20466	67130
123	91754	40784	38916	12949	76104	20556	34001	59133
124	84599	29798	57707	57392	91757	76994	43827	69089
125	06490	42228	94940	10668	62072	58983	10263	08832
126	30666	02218	89355	76117	75167	69005	42479	79865
127	87228	15736	08506	29759	74257	85594	75154	48664
128	45133	49229	32502	99698	68202	44704	39191	73740
129	55713	98670	57794	64795	27102	83420	26630	95009
130	20390	38266	30138	61250	07527	02014	43972	49370
131	13400	68249	32459	41627	56194	93075	50520	96784
132	08900	87788	73717	19287	69954	45917	80026	55598
133	86757	47905	16890	99047	78249	73739	97076	00525
134	19862	54700	18777	22218	25414	13151	54954	80615
135	96282	11576	59837	27429	60015	40338	39435	94021
136	17463	26715	71680	04853	55725	87792	99907	67156
137	44880	55285	95472	57551	24602	98311	63293	58110
138	61911	78152	96341	31473	58398	61602	38143	93833
139	07769	22819	58373	88466	71341	32772	93643	92855
140	73063	63623	29388	89507	78553	62792	89343	27401
141	24187	60720	74055	36902	22047	09091	79368	35408
142	06875	53335	91274	87824	04137	77579	54266	38762
143	23393	37710	46457	03553	58275	11138	18521	59667
144	00980	73632	88008	10060	48563	31874	90785	78923
145	46611	39359	98036	25351	88031	72020	13837	03121
146	56644	79453	49072	30594	73185	81691	29225	70495
147	98350	36891	04873	71321	29929	37145	95906	41005
148	17444	61728	86112	76261	92519	61569	65672	95772
149	45785	21301	89563	23018	60423	50801	70564	45398
150	54369	08513	36838	19805	67827	74938	66946	01206

00 01 02 03 04 05 06 07 08 09
10 11 12 13 14 15 16 17 18 19
20 21 22 23 24 25 26 27 28 29
30 31 32 33 34 35 36 37 38 39
40 41 42 43 44 45 46 47 48 49

Figura 5.2 La primera etapa en un muestreo aleatorio: asignar números a los 50 automóviles.

longitud utilizada en la numeración. Esto produce el resultado deseado porque, por ejemplo, cualquier grupo de dos dígitos tiene la misma posibilidad de ser uno de los 100 posibles grupos 00, 01,...,99. Cualquier grupo de dígitos que no sea utilizado como identificación o que duplique un número que ya está en la muestra deber ser ignorado.

Damos dos ejemplos que ilustran esta técnica.

Ejemplo: Muestreo de automóviles

Un fabricante de automóviles desea seleccionar 5 de los últimos 50 coches fabricados en una cadena de producción para realizar una inspección detallada de su calidad. Para evitar sesgos, se utilizará un muestreo aleatorio simple.

ETAPA 1. Asigne a cada coche un número. Puesto que son necesarios dos dígitos para numerar 50 coches, todos los números que se usen deberán tener dos dígitos. Comenzamos con 00. El resto de la numeración se muestra en la figura 5.2. También es correcto usar los números desde 01 hasta 50 en lugar de 00 hasta 49. Es importante anotar cómo se han numerado los elementos de la población.

ETAPA 2. Observe los grupos sucesivos de dos cifras en la tabla 5.1. Comenzando con la línea 140 (cualquier línea es válida), leemos

73063 63623 29388

Los coches numerados 06, 36, 23, 29 y 38 son los elegidos. El número inicial 73 se ignora porque no se ha usado en la numeración, y el segundo 36 se ignora porque está duplicado. ▲

Ejemplo: Muestreo de hogares

Una ciudad tiene 4.756 hogares. Como parte de un muestreo nacional, debemos elegir una muestra aleatoria simple de 3 de estos hogares para ser entrevistados. Son necesarios cuatro dígitos para numerar 4.756 hogares. Se asignan los números desde el 0000 hasta el 4755 (también es valido desde el 0001 hasta el 4756) a los hogares, y a continuación se leen grupos de cuatro dígitos en la Tabla 5.1. Si comenzamos con la línea 130, los hogares elegidos son los que corresponden a los números 2039, 0382 y 1386. ▲

Claramente, el Departamento de Estadísticas Laborales no pone bolas en una urna o nombres en

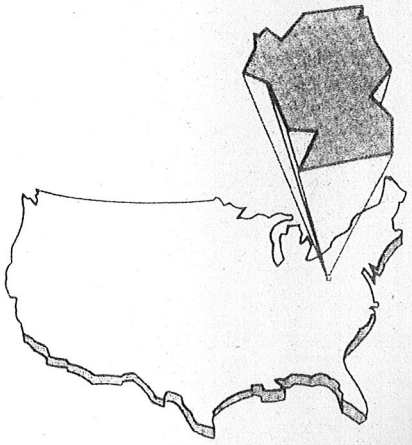

Figura 5.3 Una unidad de muestreo primaria para el Estudio de la Población Actual.

PRIMER PLANO 5.2 Muestreos y encuestas: midiendo el pulso nacional

La Doctora Janet L. Norwood, comisaria de Estadísticas Laborales desde 1979 hasta 1991, nos cuenta algunas observaciones acerca de su Agencia y de la tarea de hacer un muestreo de una nación a través del Estudio de la Población Actual.

Cómo procedemos para hacer muestreos

La encuesta de hogares es un estudio básico de la fuerza de trabajo, que produce una enorme cantidad de datos. De manera más precisa, produce información sobre las características demográficas de la población y acerca de la situación de su fuerza de trabajo: su situación laboral, el tiempo que están empleados, si están buscando trabajo, si han perdido el trabajo, si trabajan la jornada completa o a tiempo parcial, etc. Esta encuesta, que nosotros llamamos Estudio de la Población Actual, es realizada bajo contrato por el Departamento del Censo. Los entrevistadores visitan 60.000 hogares diseminados por todo Estados Unidos y pertenecientes a más de 400 áreas. La razón por la que esta muestra es tan grande —es realmente grande para un estudio de hogares— es porque se pueden extraer de ella una gran cantidad de datos. No solamente coleccionamos datos de si una persona está trabajando o desempleada. Debemos también tener información sobre hombres y mujeres, sobre la población negra, blanca, hispana, y sobre las tendencias de la población juvenil. Necesitamos tener información zona por zona y estado por estado. Y cuando analizamos grupos pequeños es necesario tener una muestra de tamaño suficiente para conseguir datos adecuados.

un sombrero para determinar quiénes serán entrevistados para el Estudio de la Población Actual. De hecho, ni siquiera utiliza muestras aleatorias simples. Las encuestas a nivel nacional y el Departamento de Estadísticas Laborales usan versiones más sofisticadas de muestreos aleatorios. La aleatoriedad está aún determinada por la muestra, pero el proceso de selección para toda la nación se hace en varias etapas, en lugar de en una sola. El Departamento del Censo, contratado por el Departamento de Estadísticas Laborales, realiza este trabajo. En primer lugar, toda la nación se divide en cerca de 1.900 unidades de muestreo primarias o UMP (véase la figura 5.3). Cada UMP es un grupo de municipios limítrofes. El Departamento del Censo selecciona una muestra aleatoria de estas unidades de muestreo. En cada una de ellas, se seleccionan aleatoriamente zonas más pequeñas de cerca de 500 individuos, llamadas distritos numerados de censos. Finalmente el Departamento del Censo elige, también aleatoriamente, casas individuales dentro de cada uno de los distritos.

Tal *muestra aleatoria multietapas* ofrece varias ventajas prácticas sobre una muestra aleatoria simple. Por mencionar una de ellas, en la elección de la muestra no se requiere enumerar cada unidad familiar de la nación. Una lista de unidades de muestreo primarias (UMP) se usa en la primera etapa y una lista de distritos numerados en la segunda. Solamente necesitamos una lista de los hogares en un número relativamente pequeño de distritos elegidos; en caso necesario, podemos hacer esta lista caminando por estos distritos. Además, los hogares que han de ser encuestados están agrupados en unos cuantos lugares, de manera que los costes de viajes de los encuestadores son reducidos. Sin embargo, el precio pagado por tener estas ventajas prácticas es complejidad al elegir la muestra y al interpretar los resultados. Puesto que una muestra aleatoria simple es el principio fundamental que hay detrás de toda muestra aleatoria y puesto que es el eslabón principal para hacer diseños más complejos, centraremos nuestro estudio en los muestreos aleatorios simples.

El impacto de nuestro Departamento

Nuestros datos sobre los precios son, probablemente, uno de nuestros mejores indicadores. Esto es así porque el índice de precios al consumo —la mejor medida que tenemos de la inflación en este país— se utiliza en varios programas del gobierno como un indicador para mantener el salario a la par con la inflación. De hecho, si se tienen en cuenta todas las personas que dependen de una unidad familiar, hemos estimado que más de la mitad de la población de este país tiene un salario que, de alguna manera, se ve afectado por el índice de precios al consumo. Y en 1985, los intervalos de impuestos se habrán comenzado a ajustar al índice de precios al consumo para calcular los impuestos.

Tratando con la incertidumbre

¡Admitámoslo! Todo el mundo quiere vivir en un mundo en el que reine la certeza. Incluso los estadísticos quieren vivir en un mundo certero. Nuestro trabajo es explicar a todo el mundo que no existe la certeza absoluta. Todas las palabras que usamos tenemos que examinarlas para estar seguros que las entendemos y que otras personas no van a entender más de lo que se pretendía expresar.

Lo que podemos hacer es proveer un puñado de estadísticos que contengan algún error, algo que a menudo llamamos error de muestreo o varianza, que nos indicará cuáles son las tolerancias básicas. Si, por ejemplo, el porcentaje de desempleados varía una décima parte, en el Departamento de Estadísticas decimos que es más o menos el mismo. Si sube dos décimas partes, decimos que ha habido una variación, porque el error de muestreo asociado con el porcentaje es tal que una variación de dos décimos está fuera del límite.

Si alguien quiere saber algo sobre los datos recogidos por el Departamento de Estadísticas Laborales, coge el teléfono y me llama o llama a alguien del Departamento que sea experto en el área. Es nuestro trabajo proveer información a la mayor brevedad posible y con el mayor número de datos que podamos. Creo que una agencia estadística sólo puede ser buena si provee una información sin trabas.

VARIABILIDAD DE UNA MUESTRA

Para exponer la idea de una muestra aleatoria, elegimos 50 bolas de una gran urna de bolas (la población). Nuestra primera muestra contenía 12 bolas negras. Si extraemos otra muestra, probablemente no obtendremos 12 bolas negras. Por ejemplo, cuando volvemos a mezclar las bolas y tomamos una segunda muestra podemos encontrar solamente que 8 de las 50, o el 16%, son negras, un número bastante diferente del 24% que obtuvimos anteriormente. Los resultados obtenidos al repetir muestras aleatorias varían.

Supongamos que la empresa Gallup hace su encuesta semanal sobre la opinión pública dos veces, seleccionando dos muestras simples, llamando por teléfono a los elegidos, y haciendo la misma pregunta a las dos muestras. Dos muestras aleatorias, cada una conteniendo 1.500 de entre más de 190 millones de personas que residen en Estados Unidos que tienen 18 años o más, probablemente contendrán personas diferentes. Y dos grupos diferentes de personas tendrán opiniones diferentes.

Cuando la empresa Gallup informa que el 45% de los americanos temen salir por las noches porque tienen miedo de la delincuencia se refiere a los 1.500 individuos en una muestra específica. Una segunda muestra producirá, sin duda, resultados diferentes. Si el Estudio de la Población Actual fuera realizado dos veces por mes, las dos muestras serían diferentes, produciendo dos porcentajes de desempleados diferentes. Las muestras aleatorias eliminan el sesgo, pero no eliminan la *variabilidad*. Las variaciones entre muestras en una serie de muestras aleatorias es una consecuencia inevitable de las variaciones en la población.

¿Cómo es posible confiar en los resultados de una muestra aleatoria, conociendo que una segunda muestra producirá resultados diferentes? ¿Cómo podemos basar decisiones económicas y políticas en el porcentaje de desempleados, sabiendo que el porcentaje sería diferente si el Departamento de Estadísticas Laborales tomara una segunda muestra? De hecho, *podemos* tener confianza en las muestras aleatorias. Para cerciorarnos de esto, debemos estudiar meticulosamente el problema de la variabilidad de la muestra.

Hay varios tipos de variabilidad. Las respuestas obtenidas enviando un encuestador a un centro comercial varían de una forma caprichosa y difícil de analizar. Sin embargo, las muestras aleatorias simples varían de forma regular, debido a que un mecanismo específico se utiliza en la selección aleatoria. Los resultados a largo plazo no son caprichosos. Vemos este tipo de variación regular en los resultados obtenidos al lanzar una moneda o en los sucesivos giros de una ruleta. Lanzar una moneda equilibrada 1.500 veces es como elegir una muestra aleatoria simple de 1.500 elementos de una gran población, siempre que imaginemos que la opinión en esta población está repartida equilibradamente de manera que las caras representan «sí» y las cruces representan «no».

Miremos más detenidamente los resultados de lanzar una moneda. Es improbable que una persona que lance 1.500 veces una moneda obtenga 750 caras (un 50% de respuestas afirmativas). La primera muestra puede producir 732 caras (49% a favor), una segunda 781 caras (52% a favor), y así sucesivamente. Pero el lanzamiento de monedas está sujeto a las leyes de *probabilidad*, que determinan con qué frecuencia cada resultado aparecerá después de muchos lanzamientos. Presentaremos la probabilidad en el capítulo 7, y mostraremos en el capítulo 8 cómo el lenguaje de la probabilidad describe la fidelidad de las conclusiones estadísticas. En particular, la teoría de la probabilidad nos dice que 1.500 lanzamientos de una moneda equilibrada casi nunca (sólo una de cada 1.000 veces) producen menos del 46% o más del 54% de caras.

Por tanto, podemos confiar en que 1.500 lanzamientos de una moneda equilibrada producirán un resultado cercano al 50% de caras. De manera similar, podemos confiar en que una encuesta de la empresa Gallup con 1.500 ciudadanos de Estados Unidos producirá un resultado cercano al resultado obtenido si se hubiera encuestado a los 190 millones de adultos de Estados Unidos. La probabilidad trata con fenómenos que son variables, pero que sin embargo muestran un comportamiento regular a largo plazo. Lanzar una moneda y elegir una muestra aleatoria son ejemplos de tales fenómenos.

TABLA 5.2 Resultados de un muestreo con bolas

Porcentaje de bolas negras	Número de muestras	Porcentaje de bolas negras	Número de muestras
8	1	22	14
10	1	24	14
12	5	26	8
14	8	28	3
16	11	30	3
18	19	32	0
20	12	34	1

En el caso de tener que elegir una muestra aleatoria simple de 50 bolas de una urna, podemos fácilmente tomar repetidas muestras y observar que existe un comportamiento regular.

EJEMPLO: Un experimento de muestreo con bolas

Tomamos 100 muestras aleatorias simples de tamaño 50 de una urna con bolas. Nuestra primera muestra tenía 12 bolas negras de las 50, o el 24%. La segunda muestra tenía el 16%, la tercera el 26% y así sucesivamente. La tabla 5.2 muestra los resultados de nuestros 100 muestreos. Podemos exhibir los resultados de las 100 muestras en un **histograma** (figura 5.4). La altura de cada barra en el histograma indica cuántas veces ha aparecido el resultado que está en la base de la barra. Por ejemplo, la altura de la barra marcada con un 16% es 11 porque 11 de nuestras muestras tenían el 16% de bolas negras. En el próximo capítulo daremos más detalles acerca de los histogramas. ▲

El histograma de nuestras 100 muestras en la figura 5.4 está muy diseminado, pero la mayor parte de las muestras están alrededor de la marca 20%. De hecho, el 20% es la proporción exacta de bolas negras en la urna de la que fueron extraídas estas muestras. Solamente 12 de las muestras que elegimos contenían exactamente el 20%, pero 70 de las 100 muestras tenían entre el 16% y el 24% de bolas negras.

Los resultados de las muestras aleatorias producen normalmente resultados muy cercanos a la verdad acerca de la población. Descubriremos en el capítulo 7 que las leyes de la probabilidad permiten hacer afirmaciones correctas acerca de la precisión de una muestra aleatoria simple. Veremos, por ejemplo, que el resultado de una muestra de tamaño 1.500 es más probable que nos dé resultados más cercanos a la verdad que una muestra de tamaño 50. Esta es la razón por la que las encuestas de opinión se hacen con 1.500 personas y no con 50.

El uso deliberado del azar para producir datos es una de las ideas fundamentales de la Estadística. Es una forma muy importante de eliminar sesgos. Más importante aún, el uso adecuado de las muestras aleatorias nos permite extraer conclusiones seguras a partir de los datos, porque las matemáticas de la probabilidad sustentan estas conclusiones.

Como hemos visto anteriormente, las muestras aleatorias simples son los elementos básicos de todos los diseños de muestras. Por tanto, aunque el Estudio de la Población Actual y la empresa de encuestas Gallup utilizan muestreos más complejos, los resultados de su muestreo aleatorio multietapas son descritos también por las leyes de la probabilidad. Así pues, estas muestras normalmente producen resultados cercanos a la verdad acerca de la población. Una

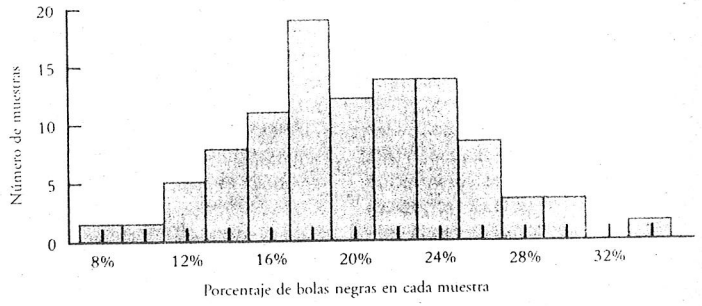

Figura 5.4 El histograma muestra la variabilidad de los resultados de las muestras en un muestreo con bolas.

afirmación cuidadosa de los resultados de un muestreo expresa «normalmente» en términos de probabilidad y «cercanos a la verdad» en términos de *margen de error*.

Por ejemplo, la empresa de encuestas Gallup toma una muestra de cerca de 1.500 personas cada semana. Para describir la precisión de los resultados de la encuesta, nos preguntamos: «¿Qué sucedería si la empresa Gallup tomara muchas muestras de 1.500 personas, haciendo la misma pregunta cada vez?» Esto es similar en principio a tomar muchas muestras de la urna de bolas. Los resultados son también similares: hay una regularidad en los resultados que nos permite conocer cuántas veces el resultado del muestreo se acercará a la verdad con respecto a la población. De manera más precisa, la probabilidad de que un porcentaje de los muestreos de la empresa Gallup esté dentro del ±3% del valor verdadero de la población es 0,95.

La probabilidad describe el comportamiento regular que aparece cuando un muestreo aleatorio se repite muchas veces. La probabilidad 0,95 significa que después de muchos muestreos el 95% de los resultados de las muestras de la empresa Gallup tienen un margen de error que está dentro de ±3 puntos acerca de la verdad de la población global. Si la encuesta revela que el 45% de la muestra tiene miedo de salir por la noche, podemos estar bastante seguros que entre el 42% y el 48% de la población comparte este miedo. La variabilidad del muestreo no ha desaparecido, pero la probabilidad nos permite describir la variabilidad anunciando un margen de error que será satisfecho por la mayoría de las muestras. Gallup y otras organizaciones dedicadas a hacer encuestas indican estos hechos en todas sus noticias de prensa, pero los editores eliminan muchas veces esta parte de la historia.

¿Qué sucede con la variabilidad en el porcentaje de desempleados? Aquí están los hechos en este caso. El Departamento del Censo prefiere anunciar un margen de error que será válido para el 90% de las muestras del Estudio de la Población Actual: es alrededor de ±0,2%, o dos décimos de 1 por ciento. Por tanto, cuando el porcentaje de desempleados baja de 7,4% a 7,3%, la variabilidad de la muestra puede ser el motivo para este cambio. Pero una variación de 7,4% a 7,0% probablemente refleja un descenso real en el porcentaje de la población activa de Estados Unidos que no tiene trabajo.

El pequeño margen de error en el porcentaje de desempleados se debe al gran tamaño de la muestra. Un margen de error de ±3% es aceptable en encuestas de opinión pública, pero no para medir el desempleo. Puesto que los datos de trabajadores y desempleados son importantes para planear la economía, el gobierno prefiere soportar el gasto que supone hacer mensualmente un gran muestreo. El margen de error alcanzado con una probabilidad deseada (tal como 0,90 o 0,95) depende del tamaño de la muestra y del diseño del muestreo. Las matemáticas de la probabilidad describen de manera precisa estas relaciones.

Experimentos

Las encuestas de muestreo recogen información de parte de la población para extraer conclusiones acerca del todo. Cuando el objetivo es describir una población, tal como el Estudio de la Población Actual tiene como objetivo describir ocupación y desempleo en Estados Unidos, la herramienta correcta que se debe usar es el muestreo estadístico.

Supongamos, sin embargo, que queremos estudiar la respuesta a un estímulo, para ver cómo una variable afecta a otra cuando cambiamos las condiciones existentes. ¿Un nuevo plan de estudios de matemáticas mejorará los resultados de los alumnos en un examen estándar de conocimientos matemáticos? ¿Tomar pequeñas dosis de aspirina reduce el riesgo de sufrir un ataque al corazón? ¿El fumar incrementa el riesgo de tener cáncer de pulmón? Estudios de observación, tales como las encuestas, son herramientas inútiles para responder a estas cuestiones. En su lugar es necesario realizar experimentos.

Un **experimento** difiere de una observación en que el experimentador interviene activamente imponiendo un *tratamiento* a los sujetos. Un tratamiento puede ser cualquier condición en la que el experimentador está interesado, como por ejemplo un nuevo plan de estudios de matemáticas o una aspirina cada día. En los muestreos, por el contrario, observamos o medimos el estado de los sujetos sin intentar cambiarlo mediante un tratamiento.

Los experimentos son el método preferido para examinar el efecto de una variable sobre otra. Imponiendo un tratamiento específico, y controlando otro tipo de influencias, podemos descubrir la causa y el efecto. Una muestra simple, por el contrario, puede mostrar que dos variables están relacionadas, pero no puede demostrar que una es consecuencia de la otra. La estadística tiene algo que decir acerca de cómo organizar experimentos, de la misma manera que sugiere métodos de muestreo.

PRIMER PLANO 5.3 Sir Ronald A. Fisher, 1890-1962

Mientras trabajaba en la estación agrícola de experimentación de Rothamsted en los años 1920, el estadístico y genético británico R.A. Fisher revolucionó la estrategia de experimentación. Los experimentadores estaban comparando los efectos de varios tratamientos, tales como diferentes fertilizantes, en las cosechas. Puesto que la fertilidad y otras variables pueden cambiar cuando nos movemos en una dirección a través de la plantación, los experimentadores usaban elaborados métodos de plantación en forma de tablero de ajedrez para evitar el sesgo. Fisher se dio cuenta de que unas asignaciones aleatorias de tratamientos a cuadrados de plantas era simple y mejor. Introdujo la aleatoriedad, describió asignaciones más complicadas, como los bloques y los cuadrados latinos, y elaboró las matemáticas del *análisis de varianza* para analizar los datos de experimentos aleatorios que se podían comparar.

Fisher contribuyó con muchas otras ideas, tanto matemáticas como prácticas, a la nueva ciencia de la estadística. Sus influyentes libros sirvieron para organizar este área. Fisher tenía fuertes opiniones y era muy combativo. Desde los años 1930 hasta su muerte, se enzarzó en virulentos debates sobre el uso apropiado del razonamiento estadístico en la inferencia científica.

Sir Ronald. A. Fisher.

Ejemplo: Un experimento incontrolado

El distrito escolar de Bigfoot Mountain, preocupado por la defectuosa preparación matemática de los

niños y niñas de Estados Unidos, adoptó un nuevo y ambicioso plan de estudios de matemáticas. Después de tres años del nuevo plan de estudios, los estudiantes que están terminando el sexto año de escuela primaria tienen una media de conocimientos que es el 10% superior a la que tenían antes del tratamiento. El distrito de Bigfoot Mountain declara que el nuevo plan de estudios es un éxito y otros sistemas lo adoptan.

Este experimento tiene un diseño muy simple. Un grupo de sujetos (estudiantes) son expuestos a un tratamiento (el nuevo plan de estudios), y se observa el resultado (medida del conocimiento de los exámenes). Este es el diseño:

Nuevo plan de estudios → observar resultados de exámenes

o, en forma general,

Tratamiento → observar respuesta ▲

Muchos de los experimentos realizados en los laboratorios usan un diseño como el del ejemplo: aplicar un tratamiento y medir la respuesta. En la atmósfera controlada del laboratorio, diseños simples son la mayoría de las veces adecuados. Pero experimentos de campo y experimentos con sujetos humanos están expuestos a condiciones más variables y tratan con sujetos muy diferentes. Requieren un control de factores externos que puedan influir en el resultado. Con mayor posibilidad de variaciones está asociada una mayor necesidad de diseño estadístico.

En Bigfoot Mountain, un ambiente de preocupación por la educación produjo un buen número de cambios simultáneos que pudo ser la causa del aumento de conocimientos de los estudiantes. A los maestros de la escuela primaria se les dio entrenamiento adicional en matemáticas. Un grupo de padres comenzó a organizar tutorías en las clases para dar a cada niño y niña ayuda individual con las matemáticas. La preocupación pública hizo que los padres se preocuparan más del progreso de sus hijos y que los profesores asignaran más trabajos para hacer en casa.

Bajo estas circunstancias, una mejora matemática se habría alcanzado sin necesidad de un nuevo plan de estudios. De hecho, el nuevo plan de estudios puede que sea *menos* efectivo que el viejo. El experimento de Bigfoot Mountain no puede distinguir los efectos en los cambios en los padres y profesores de los efectos del nuevo plan de estudios. Diremos que el nuevo plan de estudios ha sido confundido con los otros cambios que tuvieron lugar al mismo tiempo.

Las variables, ya sean parte de un estudio o no, se dicen *confundidas* cuando sus efectos en el resultado no pueden distinguirse entre sí.

El remedio para evitar tener variables confundidas es hacer *experimentos comparativos* en los que algunos estudiantes son enseñados con el nuevo plan de estudios y otros con el viejo. Los cambios en la actitud de los padres y en su participación, reenseñar a los profesores, y otras variables operan ahora de la misma manera en ambos grupos de estudiantes, de manera que es posible una comparación directa de los dos posibles planes. Casi todos los experimentos bien diseñados comparan dos o más tratamientos.

Una vez que se decide hacer un experimento comparativo necesitamos encontrar una forma de dividir a los estudiantes en dos grupos. Si los grupos son muy diferentes cuando el experimento comienza aparecerá algún sesgo. Por ejemplo, si permitimos a los estudiantes apuntarse voluntariamente al nuevo plan de estudios solamente los más aventureros que a la vez estén interesados en las matemáticas son los que se apuntarán a nuestro tratamiento experimental, y es muy probable que éstos saquen más provecho del experimento. Una elección personal sesgará los resultados de la misma manera que las llamadas voluntarias sesgan los resultados en las encuestas de opinión. La solución al problema del sesgo es la misma para los experimentos que para los muestreos: realizar la selección aleatoriamente.

Supongamos que el distrito escolar de Bigfoot Mountain decide comparar el progreso de 100 estudiantes que aprenderán con el nuevo plan de estudios de matemáticas con otros 100 estudiantes que

seguirán las enseñanzas con el plan de estudios antiguo. Seleccionaremos los estudiantes que trabajarán con el nuevo plan de estudios tomando una muestra aleatoria simple de tamaño 100 de entre los 200 sujetos posibles. Los restantes 100 estudiantes forman el **grupo de control**; continuarán con el viejo plan de estudios.

El procedimiento de selección es exactamente el mismo que el utilizado en los muestreos: los 200 miembros de la población son numerados, comenzando con 000 y terminando con 199. A continuación, consultamos una tabla de dígitos aleatorios, inspeccionando sucesivamente los grupos de tres dígitos. Las primeras 100 etiquetas que aparezcan seleccionan el grupo que será formado con el nuevo plan de estudios. Las etiquetas repetidas y los grupos de dígitos no usados son ignorados. Por ejemplo, si comenzamos en la línea 125 de la tabla 5.1 los primeros estudiantes elegidos serán aquellos etiquetados 064, 106, 102, 022, 188. Los 100 estudiantes restantes forman el grupo de control, que serán enseñados con el plan de estudios antiguo.

Lo que tenemos es un **experimento aleatorio comparativo** con dos grupos. La figura 5.5 esboza el diseño de manera gráfica. El experimento es comparativo porque dos tratamientos (dos planes de estudios de matemáticas) son comparados; es aleatorio porque los sujetos son asignados a los tratamientos aleatoriamente.

Los experimentos aleatorios comparativos son usados cuando las variables de entorno, tales como cambios en el comportamiento de los padres y profesores de Bigfoot Mountain, amenazan con confundir los resultados. Los resultados de experimentos comparativos aleatorios son tan fiables como los resultados de los muestreos aleatorios, y por las mismas razones: la selección aleatoria está gobernada por las leyes de la probabilidad. A continuación presentamos un ejemplo con tres tratamientos.

Ejemplo: Criando pavos

A los pavos criados para carne se les da a menudo el antibiótico salinomicina para prevenir que las

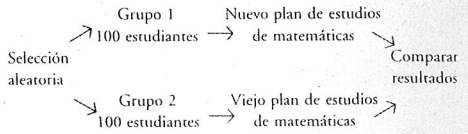

Figura 5.5 Esbozo del diseño de un experimento comparativo aleatorio para evaluar un nuevo plan de estudios de matemáticas.

infecciones se propaguen entre ellos. La salinomicina puede dañar los órganos internos de los pavos, especialmente el páncreas. Un investigador cree que el peligro se puede prevenir añadiendo vitamina E a la dieta. Quiere explorar el efecto de añadir a la dieta de los pavos tres diferentes niveles de tratamiento con vitamina E, junto con la dosis normal de salinomicina. Hay treinta pavos con los que se puede realizar el estudio. Al final del estudio, los pavos serán sacrificados y sus páncreas examinados bajo el microscopio.

El investigador decide hacer un diseño comparativo aleatorio que asigne 10 pavos elegidos al azar a cada uno de los tres tratamientos de vitamina E. En la figura 5.6 se muestra un esbozo del diseño. Los pavos son numerados del 00 al 29 (también es aceptable del 01 al 30, pero hay que asegurarse que cada número tiene dos dígitos). Se leen grupos de dos dígitos comenzando con la línea 115 de la tabla 5.1 hasta que se han elegido 10 pavos para formar el primer grupo. Los elegidos tienen los números 18, 07, 09, 28, 20, 15, 24, 27, 21 y 05. Después se continúa en la tabla para elegir 10 pavos más para el segundo grupo. Los 10 restantes forman el tercer grupo. ▲

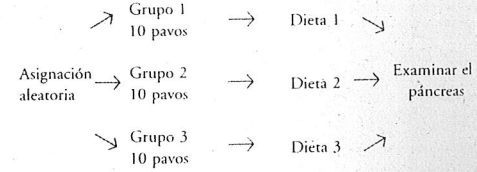

Figura 5.6. Diseño de un experimento comparativo aleatorio para estudiar tres dietas para pavos.

Los experimentos comparativos aleatorios son herramientas comunes en las investigaciones académicas e industriales. También son ampliamente usados en investigaciones médicas. Por ejemplo, las leyes requieren que la seguridad y efectividad de los nuevos medicamentos sea demostrada con experimentos comparativos aleatorios. Examinemos un típico experimento médico.

Ejemplo: El estudio sobre la salud de los médicos

Existe alguna evidencia de que tomando pequeñas dosis de aspirina de manera regular se reduce el riesgo de ataques al corazón. Algunos médicos también sospechan que dosis regulares de beta caroteno (que el cuerpo transforma en vitamina A) ayudan a prevenir algunos tipos de cáncer. El estudio sobre la salud de los médicos fue un gran experimento diseñado para analizar estas afirmaciones (véase Primer Plano 5.4). Los sujetos de este estudio fueron 22.000 hombres médicos con más de 40 años de edad. Cada médico tomó una pastilla un día sí y otro no durante un período de varios años. Hubo cuatro tratamientos: sólo aspirina, sólo beta caroteno, ambos, y ninguno. Los sujetos fueron asignados aleatoriamente a uno de estos tratamientos al comienzo del experimento. ▲

El estudio sobre la salud de los médicos introduce varias ideas nuevas importantes para el diseño adecuado de los experimentos. La primera es la importancia de contrarrestar el **efecto placebo**, para que no tengamos variables confundidas. Un placebo es un tratamiento fingido, una pastilla que no contiene ingredientes activos pero con apariencia y sabor similar a la pastilla real. El efecto placebo es la tendencia de los sujetos a responder favorablemente a cualquier tratamiento, incluso a un placebo. Si, por ejemplo, los sujetos a los que se les ha dado aspirina son comparados con aquéllos que no recibieron ningún tratamiento, el primer grupo tiene el beneficio tanto de la aspirina como del efecto placebo. Cualquier efecto beneficioso que la aspirina pudiera tener se confunde con el efecto placebo. Para prevenir la confusión, es por tanto muy importante que *todos* los sujetos reciban algún tratamiento en un experimento médico.

En este estudio médico, todos los sujetos tomaron pastillas que eran idénticas en apariencia, pero algunas contenían aspirina o beta caroteno y otras eran un placebo. El estudio fue diseñado como un **experimento ciego:** los sujetos no sabían qué tratamiento iban a recibir, porque este conocimiento podía influir en su reacción. Además, el conocimiento del tratamiento de cada sujeto puede influir a los experimentadores que les entrevistan y les examinan. Por tanto, las personas que trabajaron en el experimento también desconocían el tratamiento recibido por cada sujeto. Solamente el estadístico del experimento conocía el tratamiento recibido por cada sujeto.

Este estudio médico es un experimento más detallado que nuestros ejemplos anteriores. No sólo son comparados cuatro tratamientos, sino que dos variables diferentes están presentes: aspirina o no, y beta caroteno o no. Un experimento con dos variables, llamado normalmente **experimento con dos factores**, nos permite estudiar la interacción, o efecto conjunto, de los dos medicamentos, así como el efecto de cada uno de ellos por separado. Por ejemplo, el beta caroteno puede reforzar (o contrarrestar) el efecto de la aspirina en futuros ataques al corazón. Comparando estos cuatro grupos podemos estudiar todas las posibles interacciones. El esbozo del diseño (figura 5.7) es similar a nuestros ejemplos anteriores porque permanecen las ideas básicas de aleatoriedad y comparación de varios tratamientos.

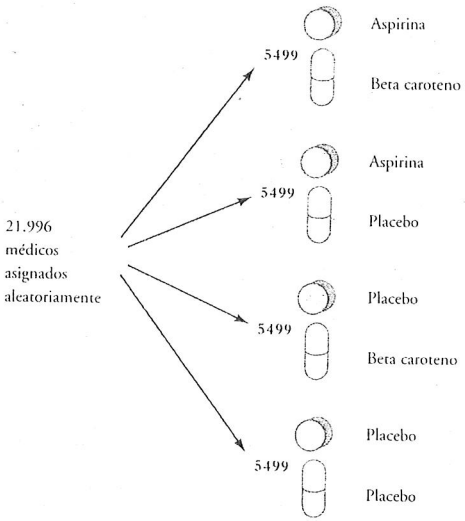

Figura 5.7. El diseño del estudio sobre la salud de los médicos, un experimento con dos factores.

PRIMER PLANO 5.4 Las estadísticas se usan para estudiar enfermedades

La doctora Julie Buring, directora asociada del estudio sobre la salud de los médicos, explica cómo obtienen los epidemiólogos datos acerca de enfermedades.

Como epidemióloga, siempre busco factores que estén involucrados en la distribución y en la frecuencia de las enfermedades de las poblaciones humanas. ¿Qué es lo que hay en nuestros hábitos, en lo que comemos, en nuestro entorno, en nuestras ocupaciones, en nuestra historia —médica o familiar— que hace que un grupo de personas sea más o menos propenso que otro a desarrollar una enfermedad? Son estos factores los que intentamos identificar.

Hacemos esto desde un par de ángulos diferentes. Uno se llama *epidemiología descriptiva*, y busca la tendencia de las enfermedades con el tiempo, la tendencia de las enfermedades en una población relativa a otra.

La otra se llama *epidemiología observacional*, y trata de observar lo que hacen las personas. Elegimos un grupo de personas que tienen una enfermedad y un grupo de personas que no la tienen. Estudiamos sus hábitos de comida o bebida, historia médica, y a qué han podido estar expuestos. La forma de proceder es tomar un grupo de personas que han sido expuestas a algo como fumar y un grupo de personas que no y hacer un seguimiento en el tiempo para ver si desarrollan una enfermedad o no. Hagamos uno u otro diseño, ambos son observacionales. Esto significa que nosotros no interferimos en el proceso. Solamente observamos.

Otro tipo de epidemiología, de la que es un ejemplo el estudio sobre la salud de los médicos, es la *epidemiología experimental*, algunas veces llamada estudio con intervención. Elegimos un grupo de personas que tiene un tratamiento y otro que no. La diferencia entre los enfoques experimental y observacional es que en la epidemiología experimental los investigadores determinan quiénes estarán en los grupos que se forman, quién recibirá el tratamiento y el placebo, quién será sometido al tratamiento y quién no.

A partir de estos dos enfoques diferentes —descriptivo y observacional— podemos juzgar si un cierto factor causa o previene la enfermedad que estamos estudiando.

Evidencia Estadística

Para un estadístico, un experimento diseñado adecuadamente es aquel que usa los principios de *comparación* y *aleatoriedad*: la comparación de varios tratamientos y la aleatoriedad en la asignación de sujetos a los tratamientos. Como vimos en el ejemplo del plan de estudios de matemáticas, la comparación elimina la confusión con variables ambientales. El efecto inmediato de la aleatoriedad es la eliminación

PRIMER PLANO 5.5 Experimentos y ética

Doctor Charles Hennekens.

El Doctor Charles Hennekens, director del estudio sobre la salud de los médicos, tuvo que ocuparse de los objetivos, el diseño y la implementación a gran escala del estudio. Pero en el curso del experimento aparecieron otras preguntas. El Doctor Hennekens fue interrogado acerca de la ética de realizar experimentos sobre la salud humana:

Mucho se ha dicho de la ética que rodea a los experimentos aleatorios. Hay ocasiones en las que no sería ético hacer un experimento aleatorio. Cuando la penicilina fue lanzada para el tratamiento de la neumonía, que producía muertes prácticamente en el cien por cien de los casos, la mortandad bajó enormemente. Por supuesto que hubiera sido poco ético hacer un experimento aleatorio, e impedir el tratamiento efectivo a personas que lo necesitaban.

Hay un delicado equilibrio entre cuándo hacer y cuándo no hacer un experimento aleatorio. Por un lado, debe haber suficiente confianza en la efectividad del agente para justificar que la mitad de los sujetos sean expuestos a su tratamiento. Por el otro, debe haber suficientes dudas acerca de su eficacia para justificar el que no se le administre a la otra mitad de los sujetos, a los que se les asignarían los placebos, las pastillas con ingredientes inertes. Eran estas circunstancias las que pensamos que existían en las hipótesis de la aspirina y del beta caroteno.

Un experimento aleatorio realizado con una nueva terapia o medicamento se hace mejor cuando el procedimiento se introduce por primera vez. Se hace muy difícil, pensando en su viabilidad y ética, hacer tales experimentos después de haber pasado un largo período de tiempo. El tratamiento comienza a ser tan aceptado por la población que es difícil encontrar gente que quiera estar sin él. Otros pueden pensar que no sería ético eliminar el tratamiento.

Un ejemplo de nuestra época es el tratamiento del cáncer de pecho. William Halsted de la Universidad Johns Hopkins, el padre de la cirugía americana, inventó la mastectomía radical al comienzo de este siglo como tratamiento para el cáncer de pecho. Se ha utilizado ampliamente durante más de medio siglo. Sin embargo, después de todo este tiempo, a algunos investigadores les parecía evidente que procedimientos menos radicales podían alcanzar los mismos resultados, esto es, mantener a un bajo nivel la mortandad en mujeres afectadas de cáncer de pecho en una edad específica. Ha sido sólo recientemente cuando se han realizado experimentos aleatorios sobre el cáncer de pecho con tratamientos menos radicales. Es importante hacerlo cuanto antes para que podamos tener una respuesta clara al comienzo del desarrollo de nuevos procedimientos y tratamientos.

La Doctora Julie Buring, directora asociada del estudio sobre la salud de los médicos, añade:

En algunas ocasiones, el hecho de que algunos procedimientos no sean probados inmediatamente con un experimento se debe a que los procedimientos parece que intuitivamente tienen sentido. Por ejemplo, tiene sentido que si se elimina mucho tejido del pecho se reduce el riesgo de que se reproduzca la enfermedad. Sin embargo la mayor parte de los médicos sólo ve a un pequeño número de pacientes. Es muy difícil observar una pauta para ver que realmente no hay diferencia entre aquellas personas a las que se les practica una mastectomía radical y las que reciben un tratamiento menos radical. Solamente cuando somos capaces de hacer estos experimentos en miles de pacientes es cuando podemos ver que no hay realmente diferencia, o que solamente hay una pequeña diferencia. Debemos tener grandes números para ser capaces de observar tales pautas.

de sesgo al formar los grupos de sujetos que recibirán diferentes tratamientos.

La salud futura de los sujetos en el estudio sobre la salud de los médicos puede depender de la edad, de su historia clínica, de su estado emocional, de sus hábitos de fumadores, y de muchas otras variables conocidas o desconocidas. En promedio, la aleatoriedad equilibrará los grupos simultáneamente en todas estas variables. Puesto que los grupos están expuestos a exactamente las mismas variables ambientales, excepto por el contenido de las pastillas, las diferencias entre los grupos puede ser atribuida a los efectos del medicamento. Esta es la lógica de los experimentos aleatorios comparativos.

Seamos un poco más precisos: cualquier diferencia entre los grupos se debe *o bien* al medicamento *o bien* a un fallo del azar en la asignación aleatoria de los sujetos. Puede suceder, por ejemplo, que personas que están a punto de sufrir un ataque al corazón sean asignadas en su mayoría, por mala suerte, a uno de los grupos. De nuevo, las estadísticas echan mano de la probabilidad. Puesto que la aleatoriedad fue deliberadamente usada en las asignaciones, las leyes de la probabilidad nos dicen lo grandes que deberían ser las diferencias entre los cuatro grupos si sólo actuara el azar. Si observamos diferencias tan grandes que casi nunca ocurrirían aleatoriamente, estamos bastante seguros de que estamos presenciando los efectos de los tratamientos. Las diferencias entre los distintos grupos de tratamientos que sean tan grandes que casi nunca ocurrirían por azar se dice que son **estadísticamente significativas.** Al igual que en los muestreos, un número grande de sujetos incrementa nuestra confianza en los resultados. El estudio sobre la salud de los médicos utilizó 22.000 sujetos para poder estar seguros de que cualquier diferencia clínica importante entre los grupos sería detectada y que estas diferencias podían ser atribuidas a la aspirina o a la vitamina A. De hecho, hubo significativamente menos ataques al corazón entre los hombres que

tomaron la aspirina que entre los que recibieron el placebo. Como consecuencia del estudio sobre la salud de los médicos los doctores recomiendan a menudo que los hombres de más de 50 años tomen pequeñas cantidades de aspirina regularmente.

La lógica de los experimentos, su diseño estadístico, y las matemáticas de la probabilidad se combinan para establecer una evidencia prácticamente irrefutable entre causa y efecto. Sólo los experimentos pueden producir una completa y convincente evidencia de las causas.

EJEMPLO: Tabaco y salud

Por otro lado, consideremos la evidencia estadística que une el fumar cigarrillos y el cáncer de pulmón. Esta evidencia está basada en observaciones en lugar de en experimentos. El estudio más cuidadoso ha seleccionado muestras de fumadores y no fumadores, los ha seguido durante varios años, anotando la causa de su muerte cuando ésta ocurría. Estos son llamados *estudios prospectivos* porque siguen a los sujetos en el tiempo. Los estudios prospectivos son comparativos, pero no son experimentos porque fueron los sujetos los que libremente eligieron fumar o no fumar. Recuérdese que un experimento debe imponer un tratamiento en los sujetos. Un gran estudio prospectivo llevado a cabo por doctores británicos permitió descubrir que la mortandad por cáncer de pulmón entre los fumadores era 20 veces superior a la de los no fumadores; otro estudio entre hombres americanos entre 40 y 79 años descubrió que la mortandad por cáncer de pulmón era 11 veces superior entre fumadores que entre no fumadores. Por tanto, la conexión observada entre el fumar y el cáncer de pulmón es fuerte. ▲

Esta conexión es estadísticamente significativa; esto es, es mucho más fuerte de lo que ocurriría por pura casualidad. Podemos estar seguros que algo más que el azar une el fumar con el cáncer. Pero observaciones de muestras no pueden decirnos *qué* otros factores intervienen además de la suerte. Quizá haya algo en la estructura genética de algunas personas que les hace estar más predispuestas tanto a hacerse adictas a la nicotina como a tener cáncer de pulmón. En este caso, se observaría una fuerte conexión incluso si el fumar no tuviera efecto en los pulmones.

La evidencia estadística que apunta al fumar como causa del cáncer de pulmón es casi tan fuerte como la que se puede alcanzar con la evidencia no experimental. En primer lugar, la conexión se ha observado en múltiples estudios en mucho países. Esto elimina factores peculiares de un grupo de personas, o de un diseño específico. En segundo lugar, se han identificado formas precisas en las que el fumar causa cáncer. El humo de los cigarrillos contiene alquitrán, y se ha mostrado experimentalmente que causa tumores en animales. En tercer lugar, no hay ninguna otra explicación plausible. Por ejemplo, la hipótesis genética no puede explicar el aumento del porcentaje de cáncer de pulmón entre mujeres, que sucedió cuando más y más mujeres comenzaron a fumar. El cáncer de pulmón, que ha sido durante mucho tiempo la primera causa de muertes por cáncer entre los hombres, está compitiendo ahora con el cáncer de pecho como el cáncer más fatal entre mujeres. Además, la genética no puede explicar por qué la mortandad por cáncer de pulmón aumenta entre no fumadores que están expuestos al humo del cigarro de otras personas.

Esta evidencia es convincente para muchos, y casi todos los médicos la aceptan. Pero no es tan fuerte como la conclusión estadística que obtenemos realizando experimentos comparativos aleatorios.

OPCIONAL EXPERIMENTOS MÁS ELABORADOS

Muchos experimentos son más complejos que un simple diseño aleatorio comparativo, de la misma manera que muchos muestreos no son muestreos aleatorios simples. Si pensamos que los pacientes masculinos y femeninos van a responder de manera diferente a la aspirina y al beta caroteno, podemos primero dividir el grupo de pacientes en dos **bloques** (hombres y mujeres) y después asignar aleatoriamente pacientes a tratamientos separados dentro de cada bloque. La división en bloques controla una variable, el sexo del paciente, incluyéndola en el diseño. Si la aspirina, por ejemplo, tiene efectos diferentes en hombres y en mujeres, estos efectos están separados en dos bloques. Otras influencias están equilibradas por la aleatoriedad. (El estudio sobre la salud de los médicos fue restringido a sujetos masculinos porque no había suficientes mujeres doctoras con más de 40 años para formar un segundo bloque. Un estudio sobre la salud de las enfermeras está explorando la respuesta de las mujeres a la aspirina y a otros tratamientos.) El uso de los bloques es común. Supóngase, para dar otro ejemplo, que un agrónomo quiere comparar distintas variedades de soja. Puesto que el tipo de terreno y el clima local pueden afectar la cosecha, el experimento se lleva a cabo en tres lugares diferentes. Cada lugar es un bloque, y las variedades se asignan aleatoriamente a pequeños trozos de terreno por separado en cada lugar.

Incluso más común que hacer bloques es el estudio simultáneo de diversos *factores*, o variables experimentales. Vimos antes que el estudio sobre la salud de los médicos fue un experimento con dos factores. La industria nos da otro ejemplo.

EJEMPLO: Un experimento industrial

Un ingeniero químico trata de determinar los valores del conjunto temperatura-presión más eficientes para un proceso de producción. Sería erróneo realizar experimentos teniendo en cuenta cada variable por separado, puesto que la temperatura más productiva variará de acuerdo con la presión y viceversa. Un experimento adecuado deberá variar a la vez la temperatura y la presión de un tratamiento a otro. Hay por tanto dos factores. El ingeniero elige tres valores para la temperatura y dos valores para la presión, de manera que cubre la gama en la que está interesado. Las combinaciones de estas temperaturas y presiones forman seis tratamientos, que se exhiben en la figura 5.8. ▲

Factor 2: Presión	Factor 1: Temperatura (°C) 95°	110°	120°
100 psi	1	2	3
150 psi	4	5	6

Figura 5.8 Los tratamientos en un experimento con dos factores.

La combinación de distintas variables experimentales usando bloques puede hacer que los experimentos sean laboriosos y caros. Sin embargo, *la combinatoria*, que es la parte de la matemática que estudia la manera en que varios elementos pueden combinarse, puede prever a veces soluciones inteligentes que permiten reducir el tamaño y el coste de los experimentos. Ilustraremos este hecho comparando varios aceites para motores.

EJEMPLO: Comparación de aceites para motores

Algunos fabricantes de aceite para motores aseguran que usando su producto se disminuye el gasto de gasolina en los coches. Para verificar esta afirmación queremos comparar el efecto de cuatro aceites diferentes en el gasto de gasolina. Sin embargo, puesto que el modelo de coche y los hábitos del conductor influyen en el gasto, el efecto de un aceite puede variar de coche a coche y de conductor a conductor. Para obtener resultados de interés general, debemos comparar los aceites en varios coches diferentes y con varios conductores.

Si elegimos 4 modelos de coches y 4 conductores, tendríamos 16 combinaciones de coches-conductores. El tipo de coche y los hábitos del conductor influyen tanto que cada una de estas 16 combinaciones es considerada como un bloque. Si hacemos los experimentos con 4 aceites (en orden aleatorio) en cada uno de los bloques, necesitaríamos 4×16, es decir, 64, experimentos.

Sin embargo, la combinatoria nos proporciona una manera de probar los efectos de coches, conductores y aceites que requiere solamente 16 experimentos. Llamemos A, B, C y D a los aceites y asignemos *un* aceite a cada una de las 16 combinaciones coche-conductor como en la tabla de la figura 5.9. Estudiemos esta tabla: cada aceite aparece 4 veces, exactamente una vez en cada fila (para cada coche) y también exactamente una vez en cada columna (para cada conductor). Esta distribución puede, de hecho, mostrar cómo se comporta cada aceite con cada uno de los coches y con cada uno de los conductores. ▲

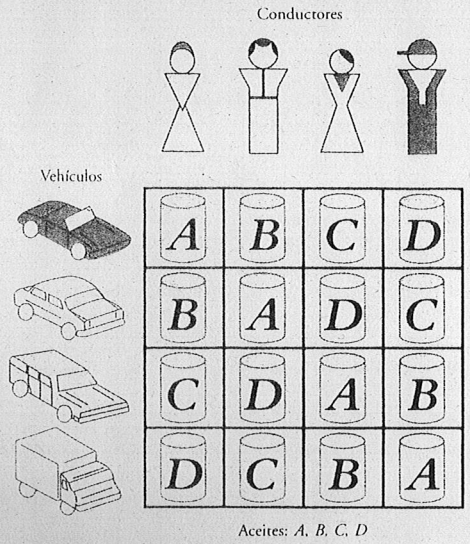

Figura 5.9 La distribución de un cuadrado latino diseñado para comparar 4 aceites para motores en cuatro coches con cuatro conductores diferentes.

Una distribución como la del ejemplo de los aceites para motores se llama un **cuadrado latino.** La figura 5.9 es un cuadrado latino de 4×4 puesto que distribuye 4 tipos de objetos (los aceites) en una matriz cuadrada de 4×4 de manera que cada objeto aparece exactamente una sola vez en cada fila y en cada columna. Hay varias formas de asignar las letras A, B, C, D de manera que cada una aparezca una sola vez en cada fila y una sola vez en cada columna; esto es, hay varios cuadrados latinos diferentes de 4×4. El ejercicio 43 pide encontrar uno que no sea el de la figura 5.9. Los matemáticos estudiaron los cuadrados latinos y han catalogado muchos de ellos antes de que el diseño estadístico de experimentos fuera inventado. El uso de los cuadrados latinos para diseñar experimentos ilustra cómo el estudio de la matemática pura, que en un principio se realizan simplemente por curiosidad, muchas veces tiene importancia práctica.

El cuadrado latino es una distribución inteligente de comparaciones. Es casi lo opuesto de una asignación aleatoria. Sin embargo, el principio de aleatoriedad de diseños estadísticos todavía gobierna la aplicación de los cuadrados latinos. La distribución de letras en filas y columnas que aparece en la figura 5.9 es el cuadrado latino. Cuando se aplica este cuadrado latino a un experimento es esencial asignar los cuatro conductores aleatoriamente a cada una de las columnas, los cuatro coches aleatoriamente a cada una de las filas y las cuatro letras A, B, C, D, aleatoriamente a cada uno de los aceites para motores. En esta situación las leyes de la probabilidad nos dirán si el gasto de gasolina con los diferentes aceites es mayor del que podría esperarse por azar. Si esto sucede, las diferencias son estadísticamente significativas. ◀

La Estadística en la Práctica

El uso correcto de la estadística requiere muchas más cosas que los diseños con cuadrados latinos. Un estadístico debe saber cuándo se puede aplicar cada técnica y si es apropiada. En el experimento de los aceites para motores sabemos que los coches y los conductores influirán en el gasto de gasolina. Un cuadrado latino nos permitirá comparar diferentes aceites de motor sin ser confundidos por los diferentes coches y los diferentes conductores; tenemos tres variables conocidas. Sin embargo, en un estudio de problemas del corazón o de cáncer, ni siquiera conocemos todas las variables que pueden influir. En estos casos, el cuadrado latino no se puede usar.

Además, una buena acumulación de datos requiere mucho más que un buen experimento o una muestra bien diseñada. También tenemos que *medir* las variables que nos interesan. Esto es fácil en el caso del gasto de combustible, pero medir «inteligencia» o «actitud hacia el aborto» es mucho mas difícil. Debemos decidir exactamente qué queremos medir y diseñar un procedimiento para medirlo.

Incluso con un buen diseño y con cuidadosos procedimientos de medida, otras fuentes de error pueden aparecer. Algunos de los sujetos en una muestra aleatoria de personas puede no estar en su casa. Otros pueden interpretar equivocadamente las preguntas, negarse a cooperar, o no decir la verdad. En experimentos con sujetos humanos es difícil aplicar buenos tratamientos. Un psicólogo que estudia la tensión nerviosa exponiendo a estudiantes voluntarios a un situación artificial debe preguntarse si unas cuantas horas de tensión en un laboratorio pueden simular meses o años de una vida difícil y obtener las mismas respuestas.

Cuando estamos planeando un estudio estadístico, también pueden aparecer algunas cuestiones de ética. ¿Justifica el conocimiento obtenido del experimento o estudio el posible riesgo a que se exponen los sujetos? En el estudio sobre la salud de los médicos, los doctores dieron su consentimiento por escrito para tomar aspirina, beta caroteno, o placebos, en cualquiera de sus posibles combinaciones, tal como describía el diseño del estudio. Cuando se vio claro que los hombres tomando aspirina tenían menos ataques al corazón, el experimento se interrumpió para que todos los sujetos pudieran obtener ventaja de este nuevo conocimiento. En Primer Plano 5.5 los directores del estudio sobre la salud de los médicos explican por qué los experimentos comparativos aleatorios son importantes en medicina, agricultura, y muchos otros tipos de investigación, y en qué ocasiones tales estudios clínicos están justificados. Los problemas éticos y prácticos nunca están lejos de la superficie cuando la estadística se aplica a problemas reales.

Vocabulario

Bloque Grupo de sujetos de un experimento que es homogéneo en una o mas variables, por ejemplo, del mismo sexo. La división en bloques controla los efectos de las variables. Los experimentos complejos asignan a menudo tratamientos diferentes de manera aleatoria a sujetos en cada uno de los bloques.

Confundidas Dos variables están confundidas cuando sus efectos en el resultado de un estudio no pueden ser distinguidos entre sí.

Cuadrado latino Tabla de k tipos de objetos distribuidos en una matriz cuadrada de k filas y k columnas, de manera que cada tipo aparece exactamente una vez en cada fila y exactamente una vez en cada columna.

Efecto placebo Efecto de un tratamiento ficticio (como una pastilla inerte en un experimento médico) en la respuesta de los sujetos.

Estadísticamente significativo Un efecto observado es estadísticamente significativo si es tan grande que es poco probable que ocurra como un resultado aleatorio de la producción de los datos en ausencia de un efecto real en la población de la que fueron extraídos.

Estudio mediante respuesta voluntaria Estudio en el que una muestra elige responder a una invitación general para escribir o llamar dando sus opiniones. Tales estudios están normalmente muy sesgados.

Experimento Estudio en el que varios tratamien-