Interpretación fácil de la bioestadística

La conexión entre la evidencia y las decisiones médicas

Gail F. Dawson, MD, MS, FAAEP

ELSEVIER SAUNDERS

Interpretación fácil de la bioestadística

Interpretación fácil de la bioestadística

La conexión entre la evidencia y las decisiones médicas

Gail F. Dawson, MD, MS, FAAEP
University of Michigan, Ann Arbor MI
Wayne State University, Detroit MI

ELSEVIER

Ámsterdam Barcelona Beijing Boston Filadelfia Londres Madrid
México Milán Múnich Orlando París Roma Sídney Tokio Toronto

Dedico este libro a todos aquellos a quienes les gusta enseñar,
por el puro placer que proporciona compartir el conocimiento.

Un agradecimiento especial a la gente de mi entorno
que me ha hecho llegar su sabiduría, dedicación,
alegría y compasión, y sobre todo a...

Alan, Gary, Diane,
Dennis y Joyce
Drew, Alex, Carl
y Jim

por su cariño y apoyo.

Nuestra vida se desperdicia en los detalles...
hay que simplificar. Simplificar.

—Henry David Thoreau

PREFACIO

Somos afortunados por vivir en una época en la cual los increíbles avances científicos y tecnológicos permiten alargar la vida humana más que nunca. El uso apropiado de estos avances ha tenido un impacto asombroso en la condición humana. Debido a la adecuada asistencia médica, la esperanza de vida casi se ha doblado en el siglo pasado. Además de vivir más años, la gente también es más productiva debido a su mejor estado de salud general.

Los profesionales sanitarios proporcionan los medios para que estos avances puedan aplicarse a la población. A través de la investigación, de la atención individualizada del paciente y de las políticas sanitarias, intentamos proporcionar la mejor asistencia al mayor número de personas. Hay dos motivos por los cuales los profesionales sanitarios hacemos lo que hacemos: ayudar a la gente a vivir más tiempo y ayudarla a vivir mejor. Cualquier intervención que recomendamos o realizamos, desde la simple aspirina diaria hasta el trasplante de células madre, debería estar justificada por la contribución realizada en al menos uno de estos dos propósitos. Éste es el objetivo común de los profesionales médicos.

Cuando recomendamos una intervención, lo hacemos con la intención de obtener una buena respuesta en un individuo concreto. Es decir, queremos que el individuo viva más tiempo, se sienta mejor, o ambas cosas. Sin embargo, una cierta proporción de individuos tendrá complicaciones debidas a la intervención, aun cuando el cribado se realice excluyendo a aquellos que probablemente obtendrán un resultado adverso. ¿Significa esto que se debe abandonar la intervención? No si podemos demostrar que la mayoría de los pacientes sometidos a la intervención tendrán una mejoría en su estado que compensa el riesgo de acontecimientos adversos. De esta manera, vivir más tiempo y vivir mejor requieren una dedicación completa porque el beneficio de la mayor parte de la población tiene prioridad sobre los acontecimientos adversos de unos pocos.

Mi hijo Alex y yo reflexionábamos recientemente sobre los grandes descubrimientos de los últimos siglos. Él pensaba que el automóvil era la mayor contribución, ya que proporcionaba una vida mejor a muchas personas. A causa del empleo masivo en fábricas, hubo un cambio en la población de las ciudades donde los recursos comunes, como el agua, el alcantarillado y la energía podían concentrarse. Los trabajadores disfrutaron de salarios sustanciales estables y podían permitirse más servicios. La limitación de la jornada laboral proporcionó más tiempo para el entretenimiento. En su tiempo de ocio, podían relacionarse más aprovechando el transporte disponible. La industria de la automoción también era indirectamente responsable de la asistencia médica de alta calidad, a través del poder negociador de sus sindicatos. Tuve que estar de acuerdo con Alex. Es una verdad indudable que el automóvil tenía (y todavía tiene) un impacto enorme en la calidad de vida de muchas personas.

Hay muchos otros inventos que han contribuido en beneficio de la especie humana. Estoy seguro de que el lector puede nombrar algunos. Personalmente, mi favorito es el mando a distancia. (Hay un placer regocijante en hacer que algo ocurra sin levantarse. ¡Y no es necesario pedirlo amablemente!) En el contexto global, sin embargo, mantengo un profundo respeto por el desarrollo de la ciencia bioestadística y su aplicación a la medicina.

La bioestadística nos permite cuantificar el hecho de vivir más tiempo y vivir mejor. Éste es el sistema a través del cual verificamos que las intervenciones que recomendamos, que consumen nuestro valioso tiempo, nuestra energía y nuestros

recursos, son realmente la mejor opción para la mayor parte de las personas. Midiendo respuestas con valores matemáticos podemos ponderar el beneficio de una intervención sobre otra, o comparar una intervención con no hacer nada. El resultado final es nuestra verificación de por qué hacemos lo que hacemos. La aplicación de esta ciencia se refleja en el incremento de vida útil y la gran calidad de vida de que disfrutamos actualmente.

Así pues, las ventajas de esta ciencia aplicada no se deben subestimar. La bioestadística proporciona el sistema de trabajo a todos los tipos de investigación médica. Y nos permite convertir este conocimiento en prácticas que mejoren nuestras condiciones. Esto ha mejorado nuestra capacidad de vivir más y sentirnos mejor. En un nivel filosófico, sin embargo, la bioestadística aplicada tiene una justificación aún más noble.

Actualmente, existen diversas teorías sobre cómo medir la inteligencia. Hay algún desacuerdo sobre la definición exacta de la palabra inteligencia, especialmente según los distintos lugares del planeta. Las culturas orientales tienden a enfatizar el autoconocimiento y las relaciones interpersonales, mientras que en Occidente la identifican con el aprendizaje de habilidades[1]. A pesar de las diferencias en creencias filosóficas, no obstante, hay un consenso general acerca de que la inteligencia implica algo más que memorizar y repetir la información, como un animal que realiza trucos circenses. Está comúnmente aceptado que la inteligencia incluye la capacidad de razonar y hacer valoraciones. Los seres más inteligentes no sólo asimilan la información, sino que también la procesan y la emplean en su provecho. Algunos psicólogos miden la inteligencia por la capacidad de una criatura para comunicarse, como podrían hacer los lobos entre sí mientras cazan una presa; otros mantienen que es un indicador de cómo los animales usan sus medios para interactuar con el entorno en su propio beneficio, como una nutria marina que usa una roca para abrir una almeja y poder comerla.

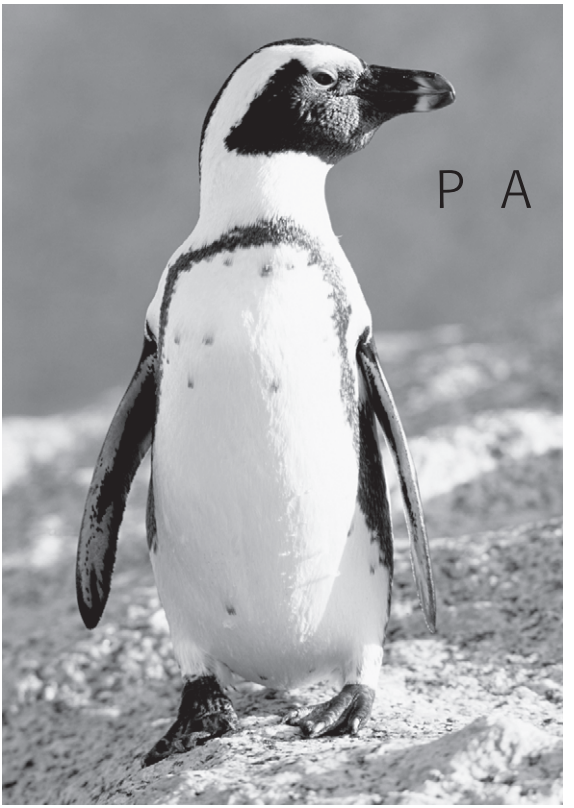
Otra manifestación propuesta de inteligencia es la capacidad para procesar la información, del pasado y del presente, y elegir la opción que con menos probabilidad vaya a causarnos la muerte o algún perjuicio (¿recuerda lo de vivir más tiempo y vivir mejor?). Esta teoría declara que las especies más inteligentes son aquellas que reconocen y utilizan el hecho que pueda afectar a su propio devenir reaccionando en una situación dada. Cuando encuentran una nueva situación, analizan las circunstancias e integran la información que aprendieron en el pasado para decidir la mejor actuación. La bioestadística nos permite recoger y utilizar lógicamente los datos para incrementar nuestro conocimiento sobre una situación concreta y, por tanto, podremos usarla para alcanzar el máximo beneficio. Mi argumento es que la aplicación de la bioestadística es la quintaesencia de la inteligencia; es decir, proporciona el paradigma dentro del cual la raza humana puede lograr su máximo potencial intelectual.

La bioestadística es una ciencia absolutamente fundamentada en la lógica y el razonamiento. Justifica las decisiones que tomamos con nuestros pacientes y nos conduce a esforzarnos por alcanzar la imparcialidad en el desarrollo de políticas sanitarias. Como hemos visto, también conlleva ideas filosóficas apasionantes. Espero haber transmitido un poco del entusiasmo que siento por esta disciplina, una materia muy atractiva. ¡Vamos a por ella!

Gail F. Dawson

1. Sternberg R. J. and Kaufman J. C. 1998. Human Abilities. *Annu. Rev. Psychol*. 49:479–502

ÍNDICE DE CAPÍTULOS



PARTE I

Aprender lo esencial

Sorprende saber lo que necesitas saber
para saber lo poco que sabes.
—Anónimo

INTRODUCCIÓN A LA PARTE I

Como profesionales sanitarios, es importante entender por qué hacemos lo que hacemos. Los pacientes acuden a nosotros para que les aconsejemos el escenario con la mejor evolución posible. La ciencia de la bioestadística se usa para verificar las recomendaciones que les damos, basadas en estudios de grandes muestras de individuos similares.

El ejemplo típico de la bioestadística es el ensayo clínico. Los sujetos se dividen en dos o más grupos, y cada grupo se somete a un tratamiento diferente. Nosotros cuantificamos la respuesta al tratamiento mediante fórmulas matemáticas que comparan la diferencia entre los grupos. Si podemos demostrar una diferencia significativa, sabremos qué opción es más verosímil que cause un resultado favorable, y podremos hacer entonces una recomendación basada en dichos resultados.

Los profesionales sanitarios suelen basarse en estudios científicos para responder a la pregunta «¿Qué debería decir a este paciente que haga?». Con el fin de aconsejar correctamente a un paciente y recomendar un tratamiento, es esencial saber interpretar los resultados de los estudios publicados en revistas. Esta fuente de información se conoce como «evidencia»; en general, es más fiable que la confianza en uno mismo o que la experiencia personal de sus colegas.

La bioestadística también se usa en la determinación de políticas de salud. Los administradores confían en los estudios como una guía en el desarrollo de programas diseñados para proporcionar el cuidado más eficaz con el coste más bajo. Por ejemplo, los programas de cribado deberían ser capaces de demostrar un beneficio significativo antes de ponerlos en práctica. Los gestores sanitarios emplean esta información para asegurarse de que distribuyen los recursos para beneficiar al máximo de usuarios.

Esta ciencia también ha encontrado su sitio en las salas de los tribunales. A los jueces se les está presionando para que admitan únicamente las pruebas más sólidas en procesos judiciales sobre aspectos médicos, en lugar de confiar en aquel que pueda contar la historia más convincente o aquel que proporcione la interpretación más brillante. Una reciente sentencia del Tribunal Supremo[1] determinó que ahora los magistrados tienen la responsabilidad de escrutar los testimonios en busca de una evidencia basada en la validez y la fiabilidad científicas. Así, para evaluar la admisibilidad del experto, es necesario disponer de los mínimos conocimientos de bioestadística necesarios para interpretar datos y validar decisiones.

Los periodistas médicos también deben entender estos conceptos, ya que los lectores confían en la prensa para recibir una información fiable. Un articulista responsable debería ser capaz de presentar al público una interpretación apropiada de un estudio científico. Las conclusiones inválidas y el sensacionalismo sólo confunden y enajenan a un auditorio ya receloso. De hecho, a todos los que leemos la prensa nos corresponde ser capaces de criticar el razonamiento usado en un artículo, y decidir de manera independiente si los datos apoyan la conclusión de los investigadores.

La literatura médica proporciona una valiosa información que nos ayuda en la toma de decisiones. Sin embargo, muchos de nosotros no hemos tenido la formación adecuada para interpretar dicha literatura. Nos sentimos intimidados por las complicadas fórmulas estadísticas empleadas. Estamos incómodos con términos como «intervalo de confianza» y «potencia», ya que están basados en conceptos que parecen verdaderamente complejos. Debido a que somos pocos los que hemos teni-

do una adecuada formación estadística, puede ser complicado interpretar la literatura médica.

Imagínese en una ciudad desconocida, intentando ir de un lugar a otro. En vez de partir sin rumbo, usted encuentra que es mucho más eficaz obtener un mapa de la ciudad para orientarse. Mientras usted sepa interpretar el mapa, podrá utilizarlo en su beneficio. Usted no creó el mapa. Usted dejó esta tarea al cartógrafo, y esperó que él o ella recogiera la información adecuada. Mientras usted tenga las habilidades necesarias para interpretar el mapa, podrá alcanzar su destino.

La comprensión de los conceptos de bioestadística es como el conocimiento que le permite leer un mapa correctamente. Las habilidades para interpretar la literatura le permitirán alcanzar la respuesta que usted busca, guiándole en sus decisiones y recomendaciones. No es necesario dominar las matemáticas que se esconden tras una prueba estadística. Deje esto a los expertos estadísticos. Confíe en que ellos le proporcionarán la prueba más apropiada para un tipo de datos y con unos resultados fidedignos. No todos los profesionales médicos pueden ser estadísticos, ni necesitan serlo. Si asimila los fundamentos, usted debería navegar cómodamente por la literatura médica.

El objetivo de este libro es presentar los conceptos de bioestadística de una manera simple, organizada y fácil de entender. Cada capítulo se concentra en un solo concepto, y la información se asimila paso a paso reforzándola a medida que se avanza. Los conceptos principales están enfatizados, pero las fórmulas matemáticas se han minimizado. Esto no es un libro de texto de estadística (¡no sería tan divertido!). El objetivo no es convertirle en un bioestadístico, sino proveerle de las habilidades necesarias para interpretar con seguridad la literatura, y así practicar la medicina basada en la evidencia. Usted aprenderá el razonamiento que hay detrás del proceso de inferencia y desarrollará un vocabulario laboral que le permitirá participar en discusiones referentes a la literatura médica. Por el camino, disfrutará de unas cuantas anécdotas históricas.

Este libro está pensado para usted. Soy consciente de las restricciones de tiempo, por eso los capítulos son concisos y se concentran en los conceptos principales. También sé que muchos profesionales sanitarios pueden disponer de unos minutos de tiempo potencialmente reflexivos entre las actividades programadas, ya sean visitas, conferencias, trámites o reuniones. Este libro está deliberadamente diseñado así para que estos breves intervalos de tiempo puedan ser aprovechados de manera eficaz examinando los puntos clave.

Los capítulos se centran en conceptos. Siempre que ha sido posible, se ilustran las ideas con gráficos. Se recomienda entender a fondo un capítulo antes de pasar al siguiente. Al final de cada capítulo hay preguntas de revisión de los puntos clave. Comprenda el enunciado y busque la mejor respuesta para reforzar el aprendizaje. Si debe releer el texto, no dude en hacerlo. Entenderá mejor el material si intenta responder antes de mirar la solución. Algunas preguntas pueden tener más de una respuesta correcta, pero estas respuestas deberían ser intercambiables. Puede intentar responder algunas ahora mismo.

PREGUNTAS DE REVISIÓN

1. Usamos la bioestadística para justificar _____.

por qué hacemos lo que hacemos
lo que recomendamos a los pacientes

2. Todo tipo de _____ usan la bioestadística para tomar decisiones acerca de tratamientos o políticas de salud.

personas, profesionales, sanitarios

3. La bioestadística usa fórmulas matemáticas para _____ en diferentes grupos de tratamiento de la misma población.

comparar resultados o respuestas
medir la diferencia

4. En general, las decisiones sobre el tratamiento o las políticas de salud basadas en la _____ son más fidedignas que la experiencia personal.

evidencia

5. No es necesario saber las _____ para entender los conceptos básicos de bioestadística.

fórmulas matemáticas

6. Me gusta entender los conceptos de bioestadística porque _____.

es divertido
es intrigante
me ayuda a aconsejar a mis pacientes
me ayuda a tomar decisiones médicas
me ayuda a valorar la credibilidad de lo que leo en la prensa
me permite juzgar la fuerza de la evidencia

BIBLIOGRAFÍA

1. Daubert, V. Merrell Dow Pharmaceuticals 509 US 579 [1993].

CAPÍTULO 1

Medidas de enfermedad

La premisa subyacente de la epidemiología es que la enfermedad y la insalubridad no están distribuidas aleatoriamente en una población.

—Leon Gordis[1]

Los primeros médicos quedaron fascinados al observar que ciertos tipos de enfermedades tendían a agruparse. Buscaban explicaciones para las tendencias que veían. Tenían la sensación de que si conseguían entender el proceso de propagación de una enfermedad, quizá entonces la enfermedad podría prevenirse interrumpiendo este proceso.

Sin embargo, para obtener un programa de tratamiento o de prevención acertado, era necesario medir la frecuencia de una enfermedad. Sólo entonces sería posible verificar si una intervención evitaba la transmisión de una enfermedad. La ciencia de la epidemiología evolucionó a través del estudio de los patrones de las enfermedades y del deseo de controlar las tasas de las enfermedades. La epidemiología usa métodos bioestadísticos para estudiar tendencias en grandes poblaciones y evaluar la eficacia de programas de prevención y tratamiento. Por eso hay que conocer cómo medir y cuantificar la enfermedad en una población y entender cómo se aplica la bioestadística en esta ciencia.

La epidemiología, para estudiar la distribución de la enfermedad en poblaciones, parte del principio según el cual las enfermedades a menudo se presentan en un tipo concreto de personas. Esto es especialmente notable en las enfermedades infecciosas, transmitidas de persona a persona, como la tuberculosis. Sin embargo, la tendencia de las enfermedades a agruparse no se basa necesariamente en la localización. La agrupación también puede depender de otros factores ambientales, denominados exposiciones. Por ejemplo, los fumadores, a diferencia de quienes no lo son, tienden a contraer ciertas enfermedades pulmonares, independientemente de donde vivan. La epidemiología intenta identificar qué exposiciones están asociadas con la enfermedad para controlar estos factores y reducir así las tasas de enfermedad. A través de métodos de supervivencia también podemos identificar epidemias potenciales, sobre todo en enfermedades de obligada declaración.

INCIDENCIA Y PREVALENCIA

Las enfermedades pueden cuantificarse por sus tasas de incidencia. La palabra «tasa» implica que algo está ocurriendo a lo largo del tiempo. La incidencia se define como el número de casos nuevos de una enfermedad que ocurren durante un cierto período de tiempo —por lo general, 1 año—, dividido entre el número de personas en riesgo. Por ejemplo, si 15 personas desarrollan la gripe de 100 que están expuestas durante 1 año, la incidencia de nuevos casos de gripe es de 0,15/año. Como la incidencia mide la ocurrencia de un acontecimiento (enfermedad) a lo largo del tiempo, es una medida del riesgo de contraer la enfermedad si se pertenece a la población.

- *La incidencia es el número de casos nuevos de una enfermedad dividido entre el número de personas en riesgo de padecerla, durante un período de tiempo dado.*
- *La incidencia es una medida de riesgo.*

La incidencia se usa para estimar las tasas habituales de una enfermedad y medir la eficacia de programas de prevención, tales como las vacunaciones. *Haemophilus influenzae* puede causar meningitis y epiglotis en niños pequeños. Ambas enfermedades tienen un riesgo vital. Cuando apareció la vacuna de *H. influenzae*, la incidencia de estas enfermedades disminuyó de manera ostensible (fig. 1-1). Además de una reducción general en la incidencia, se observó un cambio en la edad media de presentación de estas enfermedades —desde niños pequeños que habían sido protegidos por la vacuna, hasta niños más mayores que no habían sido inmunizados (fig. 1-2)—. Si se ha hecho un seguimiento de la incidencia de una enfermedad durante varios años, conocemos su tasa habitual. No esperamos un gran cambio de un año a otro. Si esto ocurre, se investiga su causa para intervenir y prevenir una epidemia aún peor. A mediados de los años ochenta, un aumento alarmante en la incidencia de la tuberculosis (fig. 1-3) alertó a los expertos de salud pública de una epidemia inminente. Previamente, parecía que la tuberculosis estaba en camino de erradicarse en Estados Unidos gracias a los esfuerzos de la sanidad pública que estaban, en efecto, controlándola. La creciente incidencia fue parcialmente atribuida a la aparición del virus de la inmunodeficiencia humana (VIH), infección que aumenta la susceptibilidad a con-

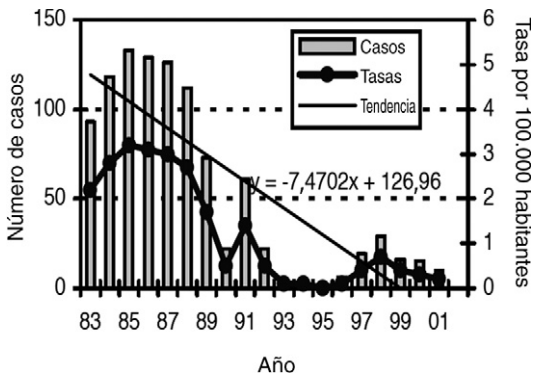


FIGURA 1-1 Casos de *H. influenzae* causantes de gripe y tasas de incidencia, Luisiana, 1983-2001. (De www.oph.dhh.louisiana.gov.)

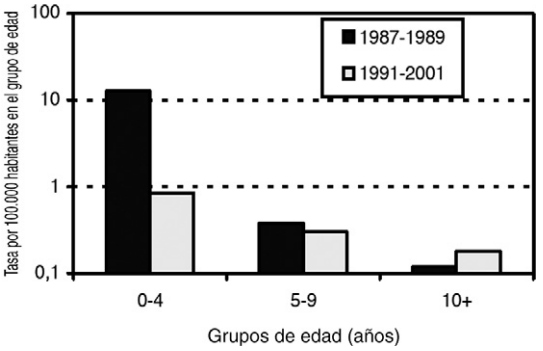


FIGURA 1-2 Enfermedades causadas por *H. influenzae*, incidencias anuales medias por edad y período de tiempo, Luisiana, 1987-1989 y 1991-2000. (De www.oph.dhh.louisiana.gov.)

traer la tuberculosis. Esta observación originó programas para identificar a los individuos afectados y asegurarse que fueran tratados.

La *prevalencia*, por otra parte, es el número de personas afectadas por una enfermedad, respecto al número de personas en la población que podrían padecerla, en un momento dado. No tiene en cuenta la duración de la enfermedad. Es como mirar una foto de la población en un instante de tiempo para ver quién tiene la enfermedad y quién no. Como esta medida no identifica los nuevos casos, no es una medida de riesgo, pero refleja el peso real de la enfermedad en la comunidad.

- *La prevalencia es el número de personas afectadas divididas entre los individuos totales en un instante de tiempo.*

La prevalencia de la diabetes mostrada en la figura 1-4 incluye a todos los adultos que han sido diagnosticados con esta enfermedad en comparación con el número de personas en la población, expresada como un porcentaje. En 2000, el número de diabéticos, incluyendo a los no diagnosticados, en Estados Unidos era asombroso:

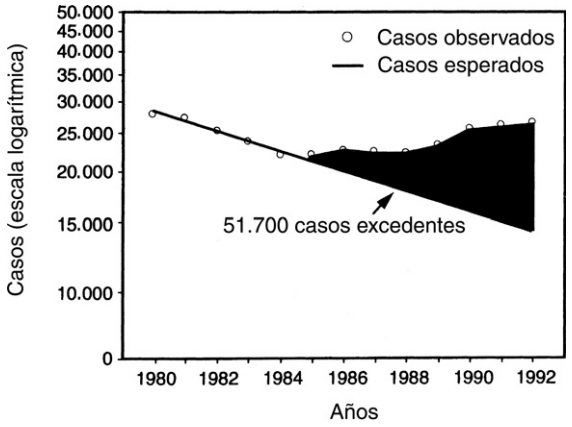


FIGURA 1-3 Número esperado y observado de casos de tuberculosis, Estados Unidos, 1980-1992. (De Centers for Disease Control and Prevention. 1993. MMWR, 42:696.)

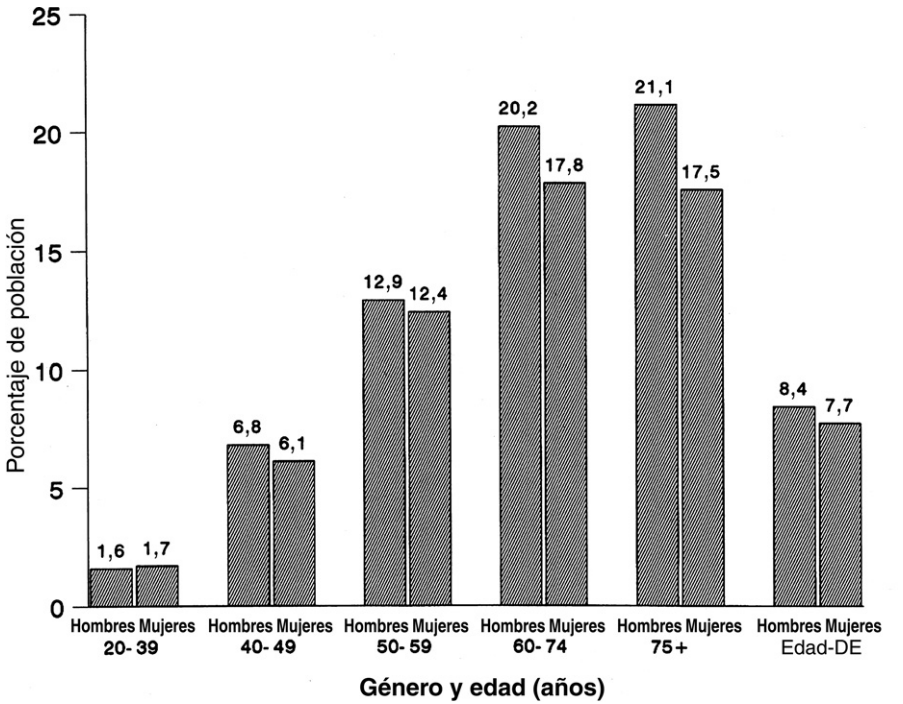


FIGURA 1-4 Prevalencia de la diabetes en hombres y mujeres (edad: >20 años) en la población estadounidense. La diabetes incluye la enfermedad previamente diagnosticada y no diagnosticada definida por glucosa en sangre en ayunas >126 mg/dl; la edad está estandarizada. (De Colwell, J. A. 2003. *Diabetes hot topics*. Philadelphia: Hanley & Belfus, p. 2.)

17 millones (el 6,2% de la población). También se puede examinar la prevalencia de la diabetes en otros grupos, como aquellos que están en diálisis por insuficiencia renal. En este caso, el denominador no sería la población entera, sino todos aquellos que están en la fase terminal de la enfermedad renal (ERFT). Conocer la prevalencia de una enfermedad nos debe permitir proporcionar servicios de asistencia médica adecuados, como centros de diálisis. También es útil a la hora de tomar decisiones sobre la distribución de fondos financieros para la asistencia médica.

La incidencia y la prevalencia están relacionadas. Una manera de aumentar la prevalencia de una enfermedad es añadir nuevos casos, por una mayor incidencia. Este aumento puede deberse a que más personas cumplen los criterios diagnósticos, ya sea porque han variado, ya sea porque programas intensivos de cribado han aumentado el número de individuos detectados. Este cambio en las tasas de incidencia podría provocar, a su vez, un aumento en la prevalencia.

Un *desgaste*, o una disminución en la prevalencia, sucede cuando la población afectada o se cura o muere. Un estado estable ocurre cuando la incidencia es igual al desgaste. El «recipiente» de la prevalencia de la figura 1-5 ilustra la relación entre incidencia, prevalencia, curación y mortalidad. Una disminución en la prevalencia podría resultar de un programa de prevención acertado que causase una disminución en la incidencia de nuevos casos. Tenga presente, sin embargo, que una dismi-

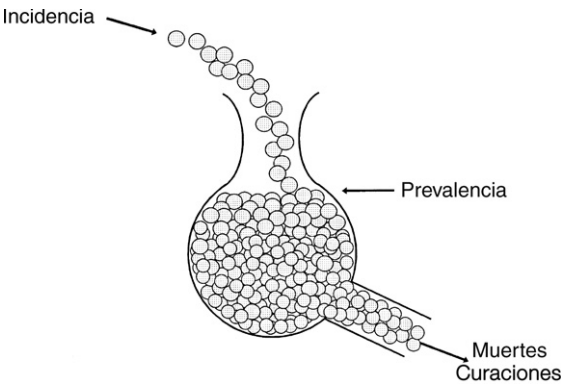


FIGURA 1-5 Relación entre incidencia y prevalencia: IV. (De Gordis, L. 2004. *Epidemiology*, 3rd ed. actualizada Philadelphia: WB Saunders, p. 37.)

nución en la prevalencia también podría deberse a una mortalidad excesiva de las personas afectadas.

Por otra parte, cuando la prevalencia sube, puede deberse a que la incidencia ha aumentado, pero también a otros factores. Existen muchas enfermedades muy prevalentes que no son curables, pero que pueden ser tratadas durante muchos años. En este caso, las personas con dichas enfermedades se van acumulando en el tiempo. El paro cardíaco congestivo, la ERFT, la enfermedad vascular y la diabetes son algunos ejemplos. Incluso si los programas preventivos controlan con éxito la incidencia de nuevos casos, la prevalencia podría aumentar debido al mejor tratamiento y al alargamiento de la vida útil del afectado. Independientemente de la causa, un aumento de la prevalencia refleja una mayor carga social, que se traduce en gastos de asistencia médica más elevados.

Tipos de estadística

En general, hay dos grandes usos de la estadística. Un conjunto de datos que describen las características de una muestra en estudio se denomina *estadística descriptiva*. Este tipo de análisis no extrae conclusiones a través de pruebas estadísticas, sino que únicamente tiene un propósito informativo.

- *La estadística descriptiva es un conjunto de observaciones que describen las características de una muestra.*

En el campo médico, puede usarse para obtener perfiles de los individuos (o instituciones) con un cierto tipo de la patología, como el número o el tipo de cánceres de pulmón diagnosticados en un período de tiempo. Los datos de la tabla 1-1 fueron recogidos por el National Cancer Institute (Instituto Nacional del Cáncer). Aportan la incidencia anual media de cánceres de pulmón diagnosticados por cada 100.000 personas en hombres y mujeres en estratos de edades diferentes. Los datos muestran el riesgo de que una persona media desarrolle el cáncer de pulmón en un año dado.

Observamos que los picos de incidencia en los hombres se dan en edades más avanzadas (75-79 años) que en las mujeres (70-74 años), y la incidencia total del

TABLA 1-1 Tasa de incidencia anual por edad y sexo (por 100.000) de cáncer de pulmón y bronquios

Edad	*Hombres*	*Mujeres*
0 a 4	0,00	0,00
5 a 9	0,00	0,00
10 a 14	0,00	0,80
15 a 19	0,15	0,12
20 a 24	0,16	0,23
25 a 29	0,46	0,56
30 a 34	1,76	1,26
35 a 39	5,40	4,15
40 a 44	21,59	14,46
45 a 49	51,31	31,84
50 a 54	104,60	56,94
55 a 59	198,15	89,97
60 a 64	283,17	123,10
65 a 69	380,18	142,84
70 a 74	459,64	153,08
75 a 79	469,67	135,20
80 a 84	441,69	108,03
85 y más	314,12	75,17

De Sondik, E. ed. (1988). *Annual Cancer Statistics Review.* Washington, D. C.: Division of Cancer Prevention and Control, National Cancer Institute. 1989.

cáncer de pulmón es más alta en los hombres que en las mujeres en los años estudiados. En los datos biológicos es habitual distribuirse a lo largo de un rango de valores, pero con picos en ciertos valores, como vemos aquí. Observe que este tipo de tabla no nos dice si hay una diferencia significativa entre sexos en la incidencia del cáncer de pulmón, o si estos datos son *significativamente* diferentes en cuanto al número de cánceres en la población general respecto a otros años. Son únicamente datos descriptivos. Los números generalmente cambiarán de año a año, pero ver si una intervención (como una campaña antitabaco) tiene un efecto beneficioso significativo requeriría una recogida adicional de datos durante muchos más años, dado el amplio lapso de tiempo entre la exposición al tabaco y el desarrollo del cáncer de pulmón. Para hacer formalmente esta comparación, usted debería realizar una prueba estadística para comparar los datos.

La *inferencia estadística*, por otra parte, hace justamente esto: aplica fórmulas matemáticas a los datos para hacer comparaciones formales entre dos o más grupos.

- *La inferencia estadística usa datos de una muestra para hacer comparaciones y sacar conclusiones sobre un grupo más grande del que representa la muestra.*

Una aplicación fundamental de la investigación clínica es identificar un camino o una intervención que beneficie a un grupo de personas. El grupo inicial está dividido en dos subgrupos, como se muestra en la figura 1-6, y cada subgrupo está expuesto a

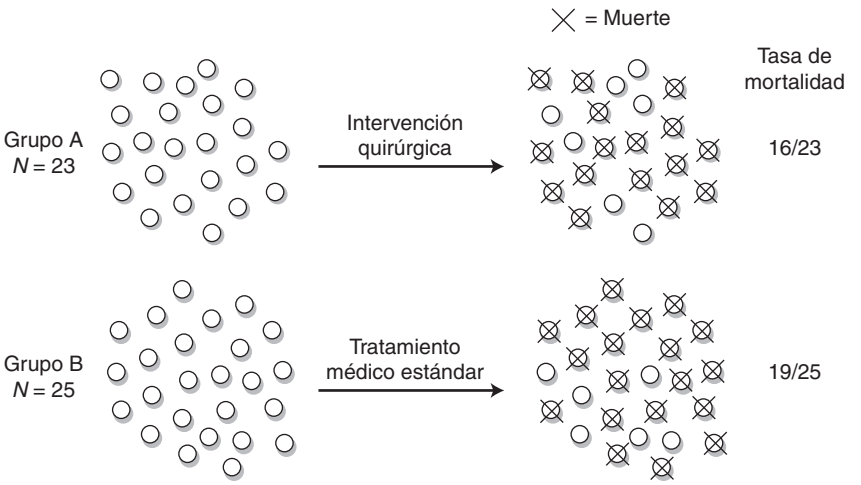


FIGURA 1-6 Un protocolo estándar de investigación donde cada grupo recibe un tratamiento diferente, y el resultado final se compara después de un cierto tiempo.

un camino o tratamiento diferente. Al final del ensayo, el efecto de cada tratamiento se mide en los subgrupos. Por ejemplo, para comparar una nueva intervención quirúrgica con el tratamiento médico estándar contra el cáncer de pulmón, usted podría asignar aleatoriamente a los pacientes de cáncer de pulmón a recibir uno u otro tratamiento, y comparar las tasas de mortalidad después de un período de tiempo.

A través de la comparación de los resultados, la inferencia estadística responde si alguna opción aportó un beneficio significativo. Se usan distintas fórmulas para contestar tipos diferentes de preguntas pero con el mismo razonamiento detrás de cada prueba estadística. Es necesario saber cuándo y cómo usar estas fórmulas para analizar usted mismo los datos. Sin embargo, la mayor parte de investigadores confían en los estadísticos para aplicar las fórmulas apropiadas a sus datos. Si usted entiende la lógica común que hay detrás de la inferencia estadística, será capaz de interpretar cualquier tipo de estudio.

PUNTOS CLAVE

- Las enfermedades tienden a agruparse en grupos de personas.
- La epidemiología es el estudio de la expansión de una enfermedad en una población.
- Las medidas de enfermedad en una población son necesarias para seguir tendencias, identificar epidemias y valorar los programas de tratamiento y prevención.
- La incidencia es el número de nuevos casos de una enfermedad dividida entre el número de personas en riesgo de padecerla, durante un período de tiempo dado. Como mide un acontecimiento a lo largo del tiempo, se expresa como una tasa.
- La incidencia es una medida del riesgo de contraer la enfermedad en un período de tiempo para cada individuo de la población.

- La incidencia se usa para conocer las tasas de enfermedad basales y para medir la eficacia de programas de prevención.
- La prevalencia es el número de personas afectadas divididas entre los individuos totales en un instante de tiempo.
- La prevalencia refleja el peso de la enfermedad en la comunidad.
- La prevalencia aumenta con incrementos de la incidencia o con programas de prevención acertados que alargan las vidas útiles.
- La prevalencia disminuye con una mortalidad excesiva, con curaciones o con programas preventivos acertados.
- La estadística descriptiva describe las características de un conjunto de datos tomados de una muestra.
- La inferencia estadística realiza comparaciones y saca conclusiones sobre un grupo más grande en función de los datos de una muestra.

BIBLIOGRAFÍA

1. Gordis, L. 2004. *Epidemiology,* Philadelphia PA: WB Saunders.

PREGUNTAS DE REVISIÓN

1. La tasa y el riesgo hacen referencia al número de _____ que ocurren en el tiempo.

2. La incidencia es el número de _____ casos de una enfermedad que aparecen en el tiempo.

3. La incidencia es una _____ y una medida de _____.

4. La prevalencia es el número de casos de una enfermedad en un _____ de tiempo.

5. El gráfico de la figura 1-4 muestra la prevalencia de diabetes en estratos de edad diferentes. Esto es un ejemplo de estadística _____.

6. Para hallar una diferencia significativa en la prevalencia de diabetes entre estratos de edad diferentes, requeriría una _____ estadística.

7. Si la incidencia aumenta y todas las demás cosas no varían, la prevalencia _____.

8. Si la mortalidad aumenta y todas las demás cosas no varían, la prevalencia _____.

9. Cuando el tratamiento alarga la vida útil en enfermedades crónicas, la prevalencia _____.

RESPUESTAS

1. acontecimientos
2. nuevos
3. tasa, riesgo
4. instante
5. descriptiva
6. inferencia
7. aumenta
8. disminuye
9. aumenta

CAPÍTULO 2

Principios matemáticos

Nuestras vidas están llenas de números, pero a veces olvidamos que los números son sólo herramientas.

—Peter L. Bernstein[1]

Una aplicación habitual de la inferencia bioestadística consiste en comparar varias opciones para ver si alguna es mejor que otra. Se selecciona una muestra y se divide en grupos. Los grupos son expuestos a intervenciones diferentes, y se observan y comparan los resultados finales. Los resultados, como la tasa de mortalidad, se expresan con un valor numérico. De este modo, las matemáticas forman parte del asunto. Sin embargo, las matemáticas que usted tiene que saber son básicas e implican comparaciones en forma de ratios o fórmulas algebraicas simples. En este capítulo se presenta una revisión de estos conceptos y una discusión de varios tipos de gráficos usuales.

RATIOS

Cuando se desea comparar las proporciones de una determinada característica en dos o más grupos, se suele recurrir a las ratios (o razones), como podría ser la razón de mortalidad de las personas tratadas médicamente frente a las tratadas quirúrgicamente. Las ratios permiten comparar fracciones. Cada fracción tiene un numerador, *N*, y un denominador, *D*.

$$\frac{N}{D}$$

D representa, por ejemplo, el número de sujetos que reciben el tratamiento médico estándar. *N* es el número de sujetos con la característica medida en el grupo (p. ej., la mortalidad). La figura 2-1 ilustra los resultados de un estudio donde los sujetos con cáncer de pulmón recibieron o bien tratamiento médico o bien cirugía, y se compararon las tasas de mortalidad.

Es más fácil comparar fracciones cuando el denominador de cada fracción es el mismo. Esto se realiza convirtiendo las fracciones de manera que los denominadores se expresan como valores iguales. La multiplicación del numerador y del denomina-

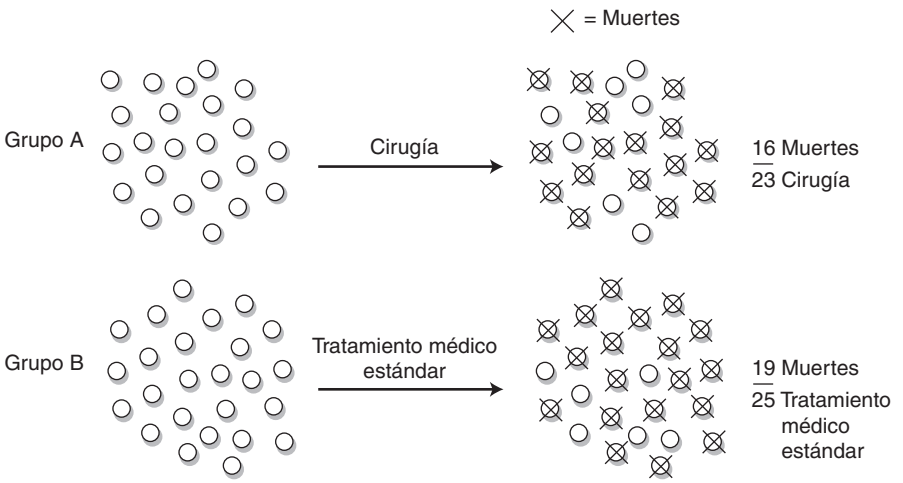


FIGURA 2-1 Las tasas de mortalidad de cada grupo pueden compararse usando fracciones.

dor por el mismo número no cambia el valor de la fracción (ya que se está multiplicando la fracción por 1). Entonces, las proporciones pueden compararse a través del valor de los numeradores.

Por ejemplo, el grupo A contaba con 16 muertes en 23 individuos en 5 años, y el grupo B contaba con 19 muertes en 25 individuos en el mismo período de tiempo. ¿Qué grupo corrió peor suerte?

$$\frac{16}{23} \text{ comparado con } \frac{19}{25}$$

Este problema es más fácil si los denominadores son iguales. Multiplicaremos la primera fracción por 25/25 y la segunda fracción por 23/23 para conseguir denominadores iguales, sin cambiar el valor de cada fracción.

$$\text{Grupo A } \frac{16}{23} \times \frac{25}{25} = \frac{400}{575}$$

$$\text{Grupo B } \frac{19}{25} \times \frac{23}{23} = \frac{437}{575}$$

El grupo B resultó peor: tuvo una tasa de mortalidad más elevada.

Otra manera de comparar ratios es reducir cada fracción de modo que el denominador sea igual a 1. Cualquier numerador sobre un denominador de 1 es igual a sí mismo (se deduce que cualquier número está realmente dividido entre 1). Realizamos esta operación dividiendo el numerador entre el denominador. Cuando todas las fracciones en un grupo son simplificadas de modo que sus denominadores sean iguales (tal como 1 o 100), es legítimo comparar sus numeradores.

$$\text{Grupo A } \frac{16}{23} = 0{,}70$$

$$\text{Grupo B } \frac{19}{25} = 0{,}76$$

Vemos de nuevo que el grupo B tuvo una tasa de mortalidad más elevada.

Las fracciones también se pueden expresar como porcentajes. Esto implica que todos los denominadores se han transformado en 100. Es fácil comparar valores relativos usando la escala porcentual, ya que nuestro dinero se basa en dicha escala y todos nosotros sabemos comparar precios.

$$\text{Grupo A } \frac{16}{23} = 0{,}70$$

$$\text{Grupo B } \frac{19}{25} = 0{,}76$$

$$\text{Grupo A } \frac{16}{23} \times 100 = 70\%$$

$$\text{Grupo B } \frac{19}{25} \times 100 = 76\%$$

El grupo A tuvo una tasa de mortalidad del 70%, mientras que el grupo B tuvo una tasa de mortalidad del 76%.

Podemos combinar las tasas de mortalidad de los dos grupos en un único número denominado ratio (o razón). La tasa de mortalidad de un grupo se convierte en el numerador, *N*, y la tasa de mortalidad del otro grupo pasa a ser el denominador, *D*. Tanto el numerador como el denominador son fracciones en sí mismos. Para nuestro propósito, cuando una fracción se divide entre otra, se representará por una doble línea. De la división de las dos tasas de mortalidad se obtiene un único número, la ratio de mortalidad.

$$\frac{\frac{16}{23}}{\frac{19}{25}} = 0{,}92$$

No siempre está claro qué grupo fue asignado al numerador y cuál fue asignado al denominador. Si el resultado es menor que 1, sabemos que la tasa de mortalidad del grupo representado en el numerador fue menor que la del otro grupo, pero a menudo tenemos que confiar en una explicación en el texto para identificar exactamente qué grupo fue mejor.

TABLAS 2 × 2

Una forma habitual de mostrar los datos que miden la prevalencia de una característica en cada uno de los dos grupos es una tabla conocida como «tabla 2 × 2». Estas tablas pueden ser confusas a primera vista porque es intuitivo dar a las cuatro celdas el mismo énfasis, pero realmente deberían interpretarse secuencialmente. Para el ejemplo, las celdas contienen el número de sujetos que murieron y aquellos que estaban vivos después del período de estudio.

	Muertos	Vivos
Grupo A	16	7
Grupo B	19	6

En primer lugar, identifique los grupos comparados imaginando una doble línea. Tenemos que añadir a las filas el número total en cada grupo. Éstos serán los denominadores.

	Muertos	Vivos	
Grupo A	16	7	**23**
Grupo B	19	6	**25**

A continuación identifique el numerador, que es la característica que se está midiendo. Este ejemplo compara las tasas de mortalidad, así que el número de fallecidos en cada grupo serán los numeradores.

Grupo A	(16)	7	(23)	$\frac{16}{23}$ = 0,92
Grupo B	(19)	6	(25)	$\frac{19}{25}$

Las tablas 2 × 2 son muy comunes; al igual que el clásico vestido negro, son ideales para muchas situaciones diferentes (son fáciles de construir y elegantes mostrando los resultados). Existen muchas aplicaciones que usan estas tablas, así que usted la encontrará en numerosas ocasiones. Recuerde que debe interpretarla identificando en primer lugar los denominadores.

LÓGICA Y DIAGRAMAS DE VENN

Ciertos grupos de individuos tienden a presentar características comunes. Por ejemplo, las personas con diabetes tienden a necesitar diálisis en la fase terminal de la enfermedad renal (ERFT); pero no toda la gente que recibe diálisis tiene diabetes, ni todos los diabéticos tienen fallo renal. Estas relaciones pueden expresarse gráficamente con diagramas de Venn, donde un círculo representa la condición 1 (p. ej., diabetes) y otro círculo representa la condición 2 (p. ej., diálisis). Un ejemplo de diagrama de Venn es el mostrado en la figura 2-2. Éstos son diagramas aproximados que no tienen un significado numérico, pero en ocasiones los tamaños relativos de los círculos pueden representar la prevalencia de cada condición. La zona común representa aquellos sujetos con ambas condiciones.

Los diagramas de Venn no tienen que estar limitados a dos condiciones. Por ejemplo, la figura 2-3 ilustra la relación entre asma, bronquitis crónica y enfisema. Algu-

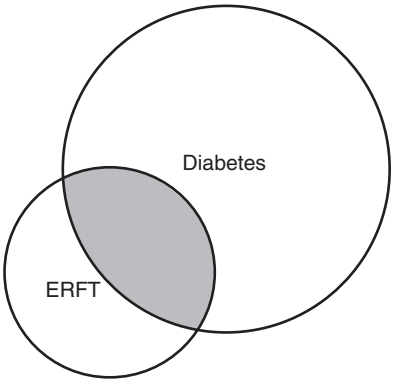


FIGURA 2-2 Relación ordinaria entre diabetes y ERFT. Vemos que la prevalencia de diabetes es más alta que la de ERFT, y que aproximadamente la mitad de los sujetos con ERFT tienen diabetes. (Arkey, R. A., ed. consultor. 2001. *MKSAP 12: Nephrology and hypertension.* Philadelphia: American College of Physicians-American Society of Internal Medicine.)

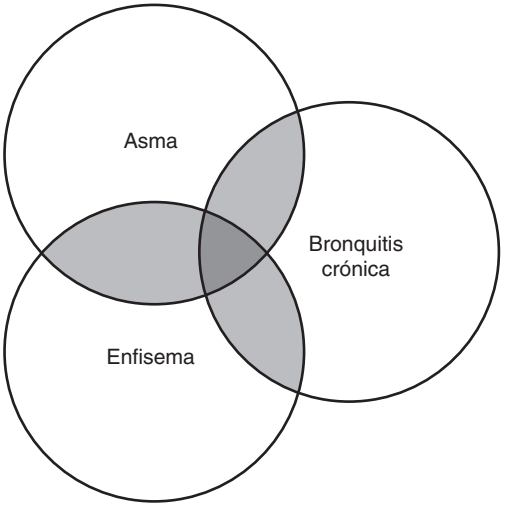


FIGURA 2-3 Asma, bronquitis crónica y enfisema pueden darse individualmente, o dos de ellos en el mismo individuo, o los tres al mismo tiempo.

nas personas tienen sólo una de estas condiciones, otras tienen dos, y las representadas en el centro del diagrama han sido diagnosticadas con las tres. Este diagrama de Venn también ilustra que ninguna de estas condiciones es exclusiva. La presencia de una condición no excluye la posibilidad de tener otra.

Los diagramas de Venn se usan para ilustrar un primer conocimiento sobre la relación entre las condiciones. Cuando un círculo invade a otro o se incrementa el solapamiento, quizá pidamos datos adicionales para cuantificar su relación. De hecho, volveremos a los diagramas de Venn mientras profundizamos en las enfermedades y su epidemiología.

Cuando un círculo está completamente dentro de otro círculo más grande, como ocurre en la figura 2-4, todos los que tengan la condición representada por el círculo más pequeño también tendrán la otra condición. Éste es un ejemplo de inclusión: si se tiene la condición B, seguro que se tiene la condición A. Son bastante inusuales, ya que en general puede haber alguna excepción a la regla. Por otra parte, las condiciones mutuamente excluyentes tienen dos círculos separados que no se tocan (círculos C y D). Tener una condición implica no tener la otra.

El diagrama de Venn de la figura 2-5 describe la relación entre la arteriosclerosis coronaria (AC) y la presencia de uno o varios factores de riesgo: hipertensión, diabetes, hiperlipidemia o antecedentes familiares de AC.

Los diagramas de Venn pueden ser muy eficaces para describir relaciones influidas por algún sesgo. Por ejemplo, las primeras investigaciones psiquiátricas que se realizaron en homosexuales concluían que todos ellos tenían un trastorno mental. De hecho, hasta 1973, el *Manual Diagnóstico y Estadístico de los Trastornos Mentales*,

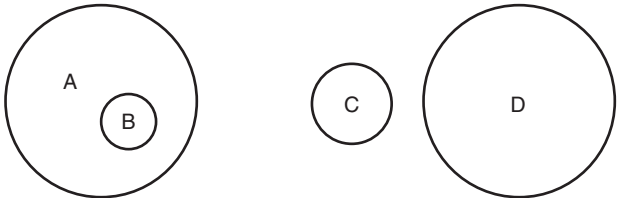
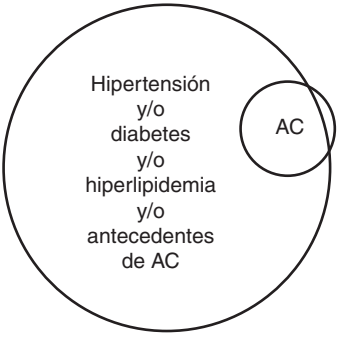


FIGURA 2-4 En el primer diagrama, el círculo B está dentro de A. Tener la característica A es un requisito necesario para tener también la B. El segundo diagrama ilustra condiciones mutuamente excluyentes, donde tener una condición (representada por C) excluye tener la otra (representada por D).

FIGURA 2-5 Este diagrama de Venn muestra que la AC está sólidamente ligada con los factores de riesgo ya que el círculo de la AC está casi completamente cercado por el círculo de factores de riesgo. Casi todos los que tienen AC tienen al menos un factor de riesgo, aunque haya muchas personas alrededor con uno o varios factores de riesgo y que no tienen AC. Hay muy pocos casos de personas con AC que no tienen ninguno de estos factores de riesgo. Obsérvese que este diagrama no indica que los factores de riesgo *causen* AC, sólo que hay una asociación observada.



publicado por la American Psychiatric Association, incluía la homosexualidad como una enfermedad, y el tratamiento se centraba en la «conversión» de estos individuos hacia una orientación heterosexual[2]. El diagrama de Venn que representa esta percepción se parecería al de la figura 2-6.

Este diagrama representa la creencia de que todos los homosexuales son enfermos mentales. Los psiquiatras no se percataban de que no estaban contemplando todo el denominador de homosexuales, sino sólo los que habían sido remitidos por una enfermedad psiquiátrica, pero muchos homosexuales no tenían ningún trastorno mental y, por lo tanto, no existía ninguna razón para consultar al psiquiatra. Se necesitaron años de esfuerzos para obtener el diagrama de Venn correcto (parecido al de la fig. 2-7) y eliminar la clasificación de la homosexualidad como una enfermedad.

¡Éste es un dibujo muy diferente! Los psiquiatras que al principio clasificaron la homosexualidad como un trastorno mental no consideraban toda la realidad, pero también podían haber estado influidos por el movimiento social y religioso contra los homosexuales. Los diagramas de Venn pueden usarse para representar observaciones objetivas, pero también podrían reflejar asociaciones originadas por un sesgo. Es un ejercicio fácil generalizar sobre una profesión, una religión o una raza basándose en unas pocas observaciones. Tenga siempre en mente la posibilidad de un círculo más grande cuando realice juicios y forme opiniones. Puede existir una llamada

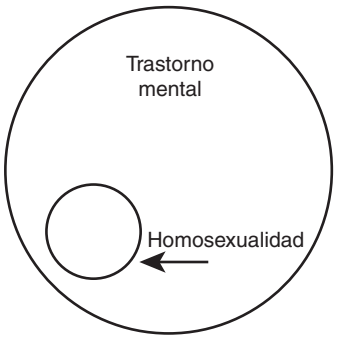


FIGURA 2-6 Este diagrama de Venn describe una relación debida a un sesgo.

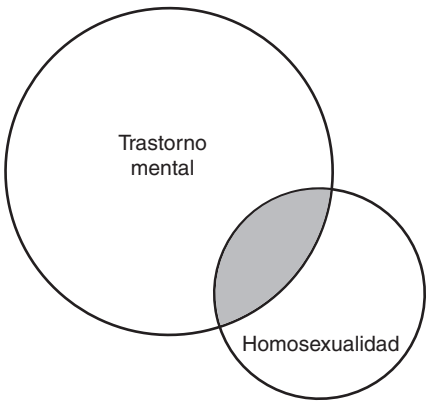


FIGURA 2-7 Un diagrama de Venn con clasificaciones cambiadas para revertir la vieja creencia de que todos los homosexuales son enfermos mentales.

CUADRO 2-1

Recientemente tuve una discusión con un residente que pasaba la mañana evaluando a pacientes admitidos en el servicio de urgencias con dolor en el pecho la noche anterior. Estaba frustrado por el número de pacientes admitidos con un dolor que aparentemente no tenía un origen cardíaco. Comentó que los médicos de urgencias admitían a todo aquel que presentaba dolor en el pecho. Le dibujé el diagrama de Venn de su percepción del grupo P (aquellos con dolor en el pecho que acuden a urgencias) y del grupo A (aquellos con dolor en el pecho que eran admitidos). Asumí que todas las admisiones por dolor en el pecho se realizaban a través de urgencias.

Todos aquellos que se presentan (grupo P) son admitidos (grupo A). Por lo tanto, el grupo P está contenido completamente dentro del grupo A. Por otra parte, ¿todos los que son admitidos con dolor en el pecho (grupo A) acuden a urgencias (grupo P)? Si suponemos que todas las admisiones proceden de urgencias, entonces el grupo A estará contenido dentro del grupo P. Los dos círculos estarán superpuestos.

Luego dibujé el diagrama de Venn para la percepción del médico que realmente trabajaba en el turno de noche de urgencias para presentados (P) y admitidos (A). Sospecho que el grupo P será más grande que el grupo A. Puede que existan varios pacientes que fueron evaluados con dolor en el pecho y enviados a casa, sin saberlo el residente.

«mayoría silenciosa» que no se había tenido en cuenta. Nunca subestime el verdadero denominador. Ésta es una práctica muy sana, que conviene aplicar en la vida profesional y en nuestras relaciones sociales. Considere la situación descrita en el cuadro 2-1.

RELACIONES ENTRE CARACTERÍSTICAS

A menudo, el valor de una característica está asociado con el de otra de forma constante y expresable numéricamente. Por ejemplo, el nivel de la educación está relacionado con los ingresos anuales: la gente con más educación generalmente gana más dinero.

Esta relación numérica puede expresarse de tres modos:

1. *Palabras*. Podemos hacer una declaración sobre la fuerza de la relación. Por ejemplo, podríamos decir que cada año adicional de formación tras los estudios secundarios aumenta los ingresos una cierta cantidad media, por ejemplo 10.000 dólares, que se añaden a un sueldo base que podría ser de 30.000 dólares (todos ellos son datos ficticios).
2. *Fórmulas*. Son más satisfactorias intelectualmente, pero requieren que el lector entienda los símbolos usados. Por ejemplo, la fórmula familiar $y = bx + a$ representa una asociación lineal entre y (variable *dependiente*) y x (variable *independiente*): y depende del valor de x. La fuerza del efecto que x tiene en y se observa en el valor de b; a se denomina constante o valor basal y es la intersección en y cuando $x = 0$. Usando el ejemplo de los ingresos, la variable independiente x es el número de años de formación adicional a la enseñanza secundaria, y b es el efecto en los ingresos por cada año de formación adicional; a representa los ingresos basales de 30.000 dólares para aquéllos sin más formación;

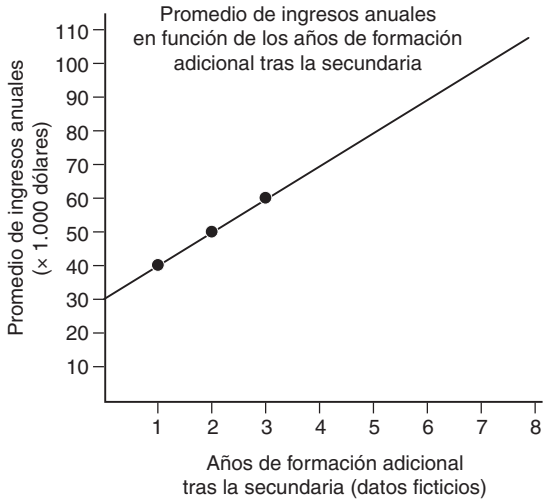


FIGURA 2-8 Gráfico de la ecuación $y = 10.000x + 30.000$. Por cada incremento unitario en x, y aumenta en 10.000. La constante de y para $x = 0$ es 30.000.

y representa los ingresos anuales medios que dependen del valor de *x*, añadido a los ingresos basales.

Cuando queremos saber múltiples valores de *y*, puede ser más fácil consultar una tabla que contenga la solución de la ecuación para diferentes valores de *x*. Un ejemplo habitual es una tabla que convierte la temperatura de grados Fahrenheit a grados Celsius (o centígrados). También veremos ejemplos de tablas estadísticas que toman distintos valores y los transforman en probabilidades.

3. *Gráficos*. Éste es el modo más informativo de mostrar una relación entre variables. Transmiten fácilmente el tipo y la fuerza de la relación existente entre valores. Es más informativo ver una ilustración que tratar de interpretar descripciones verbales o numéricas.

- *El eje x es la abscisa.*
- *El eje y es la ordenada.*

El gráfico de la figura 2-8 muestra que la cantidad de incremento de *y* depende del valor de *b*, que es la pendiente de la recta.

Cómo leer un gráfico: aproximación paso a paso

Aunque parezca rudimentario, es una buena costumbre hacer secuencialmente cada uno de estos simples pasos siempre que usted se encuentre con un gráfico. Conviértalo en una rutina.

1. *Mire el título*. Sorprendentemente, muchos de nosotros hacemos esto en último lugar porque nos atrae la parte gráfica de los datos. Es mucho más eficiente concentrarse primero en qué contiene antes de mirar el gráfico. A veces tendrá que buscar el título. Puede estar en la cabecera, como en la figura 2-9. En ocasiones sólo se puede encontrar en el texto.

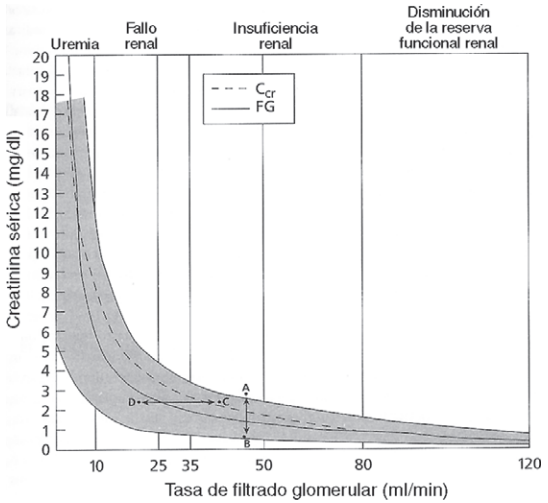


FIGURA 2-9 Relaciones entre la tasa de filtrado glomerular (FG) por ^{125}I-iotalamato y la concentración de creatinina sérica (C_{cr}) y la etapa terminal de la enfermedad renal crónica.

2. *Mire la etiqueta de la abscisa (eje x)*. Representa la variable independiente. Esta variable puede estar trazada de manera continua, o bien puede representar grupos diferentes (p. ej., en un gráfico de barras).
3. *Mire la etiqueta de la ordenada (eje y).* Representa la variable dependiente. Preste atención al rango de valores, ya que puede no ser el que usted pensaría intuitivamente. Algunos valores pueden haber sido excluidos del rango, especialmente si contenían únicamente unos pocos datos. El gráfico de ingresos anuales medios (v. fig. 2-8) podría mostrarse sin los niveles inferiores de ingresos ya que son redundantes en nuestro ejemplo.
4. *Mire la recta o la barra que define la relación*. Aquí está la sustancia del gráfico. Es más fácil interpretarla después de haber pasado por los pasos previos. Algunos gráficos contienen mucha información y pueden mostrar diversas relaciones en un cuadro, usando múltiples líneas o barras coloreadas para representar grupos diferentes. Observe que cuando x se traza de manera continua, la pendiente de la recta representa el efecto que un incremento de x tiene en y. Si es una línea recta, x tiene un efecto constante en y. Si la línea está curvada, la magnitud del efecto en y cambia con el valor de x, como sucede en la ecuación $y = x^2$.
5. *Saque sus propias conclusiones*. ¡Esto es como el postre! Es la suma de los pasos previos. Compare sus conclusiones con las del autor. La figura 2-9 puede parecer un poco confusa a primera vista. Mirémosla usando esta aproximación paso a paso.

Primero encuentre el título (en este caso, está en la cabecera). Ayuda a saber que la tasa de filtrado glomerular (FG) es un reflejo directo de la función renal, y que la manera más precisa de medir el FG es inyectando una sustancia conocida como ^{125}I-iotalamato y luego medir el volumen de sangre que los riñones son capaces de filtrar

de esta sustancia en 1 min. Debido a que es un método muy engorroso, no se usa en la práctica médica. Un método mucho más simple consiste en hacer un análisis de sangre midiendo la creatinina sérica. Este gráfico representa la relación entre el FG medido por la sustancia altamente precisa ^{125}I-iotalamato y la creatinina sérica. Es esencialmente una valoración de la fiabilidad de la creatinina para estimar la función renal.

A continuación, observe que la abscisa está trazada sobre valores continuos, de 0 ml/min (función renal nula) a 120 ml/min. ¿Es ésta la mejor función renal posible? Probablemente no, pero valores más altos no aportarían más información. El FG ha sido subdividido en niveles de la función renal, pero los grupos no son iguales. «Uremia» y «Fallo» tienen un rango más pequeño de valores. Las divisiones se van haciendo más pequeñas para valores inferiores de FG, justo cuando la creatinina comienza a aumentar rápidamente. Esto puede ayudar a identificar a las personas con enfermedad grave antes de que llegue el fallo renal.

La ordenada contiene los valores para la creatinina sérica. Los valores se cortan en 20 mg/dl, ya que valores superiores no aportarían información.

¿Cuál es la relación? Fíjese que hay dos líneas que comprenden un área de gris, que implica que existen puntos de datos entre estas líneas. Esto significa que hay un rango de valores de creatinina para un FG dado en diferentes individuos, y que el rango se hace realmente ancho cuando el FG decrece. Advierta también que las líneas son encorvadas. El efecto de la función renal en el nivel de creatinina sérica no es lineal. Para descensos del FG, la creatinina en un principio varía muy poco, pero en los niveles más bajos de función renal puede subir drásticamente.

¿Cuál es la conclusión? Los puntos A y B de la figura 2-9 representan el rango de creatinina que podría observarse en un grupo de individuos con la misma función renal. Si el FG es normal, este rango es muy estrecho. En cambio, para niveles inferiores de función renal, es más amplio: hay otros factores que afectan más a la creatinina sérica que los descensos en la función renal.

Los puntos C y D de la figura 2-9 muestran cómo individuos con diferentes niveles de capacidad de filtración renal pueden tener el mismo nivel de creatinina. Los autores atribuyen este hecho a diferencias en la masa muscular, que explica la mayor parte de la creatinina en sangre. Cuando la función renal falla, la creatinina aumenta. De hecho, cuando se estima el FG, las fórmulas usadas no solamente tienen en cuenta la creatinina sérica sino también otros factores, como la masa muscular, la altura, el peso, el sexo y, en ocasiones, la raza. Este punto podría resaltarse bajando el punto D a 1,5 mg/dl y colocando el punto C en una posición paralela a la categoría de Insuficiencia renal.

En general, el uso de la creatinina sérica para estimar la función renal es muy eficiente. Un valor de creatinina mayor que 2 mg/dl debería levantar la sospecha como mínimo de una enfermedad renal moderada. Estos puntos de datos representan un gran número de individuos en un instante de tiempo, pero no nos cuentan lo que sucede con los valores de creatinina a medida que el fallo renal crece. ¿Podría alguien medir la creatinina sérica en el tiempo en un individuo dado para supervisar la función renal, o el rango de valores variaría demasiado como para ser útil? (De hecho, es de gran utilidad supervisar la creatinina sérica en un individuo durante un período de tiempo, aunque este gráfico no dé esta información.)

¿Está de acuerdo con las conclusiones?

ECUACIONES MATEMÁTICAS Y DATOS

Una de las aplicaciones de las matemáticas es la resolución de ecuaciones. Cuando nos dan los valores de una variable, podemos hallar el valor de otra, como en $y = bx + a$. Cuando tratamos con datos, sin embargo, comenzamos con los valores de x e y para cada sujeto y tratamos de encontrar la ecuación que mejor describa la relación. Esto es particularmente desafiante porque los valores que observamos en la naturaleza no se ajustan a una relación matemática estricta. Para cualquier valor de x, habrá más de un valor de y debido a la variabilidad natural de los datos biológicos. Por ejemplo, ¿existe asociación entre peso y altura? Si medimos altura y peso en cada sujeto, veremos que no todos los individuos con la misma altura pesarán lo mismo. Para estudiar si esta relación realmente existe, trataremos de encontrar el modelo que mejor se ajuste a las observaciones que tenemos.

- *La estadística usa observaciones para definir relaciones matemáticas.*

PUNTOS CLAVE

- Las fracciones pueden compararse cuando sus denominadores expresan los mismos valores. Si una fracción representa el resultado final de una intervención en cada grupo, las fracciones pueden compararse y aportarse como un único valor o ratio.
- Las tablas 2 × 2 organizan los datos por grupos, y muestran la frecuencia de una característica dentro de cada grupo.
- Los diagramas de Venn se usan para mostrar aproximadamente las asociaciones entre características.
- Siempre considere todo el denominador cuando examine los datos o al sacar conclusiones.
- Cuando las características están relacionadas numéricamente, la mejor manera de describir la relación es con un gráfico.
- El eje x es la abscisa.
- El eje y es la ordenada.
- Aprenda a leer un gráfico paso a paso:
  1. Lea el título para orientarse.
  2. Identifique las unidades en la abscisa.
  3. Identifique las unidades en la ordenada.
  4. Fíjese en la relación entre los puntos trazados.
  5. Saque sus propias conclusiones.
- La estadística usa observaciones para tratar de encontrar el mejor modelo matemático que describe una relación.

BIBLIOGRAFÍA

1. Bernstein, P. L. 1998. *Against the gods*. New York: John Wiley & Sons, Inc., p. 7.
2. Kaplan, B. J. y V. A. Sadock. 2000. *Comprehensive textbook of psychiatry*. Philadelphia: Lippincott, Williams & Wilkins.

PREGUNTAS DE REVISIÓN

1. Una _____ es una manera de comparar dos tasas de mortalidad mediante un solo número.

2. Se compararon dos regímenes médicos diferentes para el cáncer de mama. La razón de las tasas de mortalidad era 1,23. El grupo que obtuvo una mejor respuesta estaba en el _____ del cálculo de la ratio.

3. Un estudio que revela una ratio de mortalidad de 1 muestra que los grupos tenían tasas de mortalidad _____.

4. Los diagramas de Venn son un modo de ilustrar _____ relaciones entre características.

5. Los resultados que corresponden a múltiples valores de una variable independiente de una ecuación pueden mostrarse en una _____.

6. Cuando usted lea un gráfico, antes de interpretar la relación entre las variables, mire las _____ y _____ de la abscisa y la ordenada, porque entonces tendrá los datos en el contexto apropiado.

7. Las observaciones en la naturaleza no se ajustan a una ecuación _____ estricta, pero tratamos de encontrar el mejor modelo que la describa.

RESPUESTAS

1. razón

2. denominador

3. iguales

4. aproximadamente

5. tabla

6. unidades, la escala

7. matemática

CAPÍTULO 3

Poblaciones

Población: 1. Número total de personas que habitan un país, ciudad o distrito. 2. Masa de habitantes de un lugar. 3. Conjunto de organismos en consideración. 4. Conjunto de casos estadísticos.

—The American College Dictionary[1]

Los estadísticos piensan en términos de grandes números. Se centran en la multitud, más que en lo individual. Cuando usted realice esta transición, estará en camino de comprender la teoría que hay detrás de la bioestadística y del proceso de inferencia.

CONCEPTOS TEÓRICOS SOBRE POBLACIONES

La mayoría de nosotros tratamos casi exclusivamente con situaciones individuales a lo largo del día. Las recomendaciones que damos a los pacientes dependen de sus condiciones particulares. A través del proceso educativo convencional, nos hemos entrenado para responder de cierto modo a cada situación. Además, nuestras acciones están muy influidas por una experiencia personal y que guardamos como un valioso tesoro.

La teoría que hay detrás de la bioestadística amplía este concepto. Imagínese que es capaz de observar la respuesta a un tratamiento dado (de una manera totalmente imparcial) *no sólo en un* paciente, sino también en un número *infinito* de pacientes. No todos los pacientes tendrían la misma respuesta. Sin embargo, los resultados tenderían a agruparse, y si el tratamiento fuese beneficioso, el resultado en general sería mejor que cualquier otra opción inferior. Provisto de este holgado conocimiento, usted podría predecir un resultado individual basándose en las respuestas observadas en el gran grupo.

En general, pensamos en las poblaciones como personas que ocupan una misma parcela de terreno. Pero, en términos estadísticos, la definición es mucho más extensa. En la literatura médica, las poblaciones estudiadas suelen ser grupos de personas con características comunes. (Los libros de estadística definen las poblaciones como conjuntos de datos u observaciones; es decir, como el conjunto de números obtenidos de las unidades. En nuestro caso usaremos la definición más universal de poblaciones considerándolas conjuntos de unidades en sí mismas.)

- *Una población es un conjunto de entes que tienen alguna característica cuantificable en común.*

Identificamos a las poblaciones según lo que nos gustaría estudiar. Una población podría estar constituida casi de cualquier cosa. Podrían ser las uvas de un cierto viñedo que nos gustaría probar. Podrían ser los muelles recién fabricados cuya resistencia a la tensión nos gustaría testar. Podrían ser huevos de insecto o gotas de lluvia. Podrían ser niños asmáticos o mujeres con osteoporosis.

Una población real es un concepto teórico. Muchas poblaciones no pueden enumerarse completamente por su elevado tamaño; por ejemplo, las gotas de lluvia. Si aparte de considerar todas las gotas de lluvia que caen en el presente, consideramos también las que han caído en el pasado y aquellas que aún están por caer, el número tiende a infinito. Muchas poblaciones, como los pacientes con paro cardíaco congestivo, son tan enormes que es imposible tener en cuenta a cada uno de sus miembros. Un diagrama de Venn de una gran población debería tener los bordes borrosos, difuminados o flechas alejándose del centro, como en la figura 3-1.

Está consensuado el uso de un círculo distinto para representar una población. Cuando vea una población representada por un círculo tradicional, piense en estos conceptos.

Otra razón por la cual las poblaciones no tienen bordes marcados es que, en muchos casos, cambian constantemente. Piense en la población de niños asmáticos. Los miembros de la población comparten una característica medible, que es o bien un conjunto de síntomas recurrentes, o bien los resultados de una prueba respiratoria. La definición de asma es fija. Sin embargo, resulta complicado contar los miembros de la población en un momento dado. Establecemos criterios para distinguir a los miembros que la componen, pero cada día nuevos individuos cumplen estos criterios mientras que otros dejan de hacerlo. Por ejemplo, continuamente se añaden a la población nuevos casos de asma. Por otra parte, debido a que los niños crecen, habrá un flujo continuo de población que alcanzará el momento (¿un minuto o quizá un segundo?) en que «un niño» se convierta en un «no niño».

PARÁMETROS POBLACIONALES

Todas las poblaciones tienen atributos que intentamos medir. Aunque las características de una población y los atributos pueden tener aspectos en común, no son lo mismo. Los atributos hacen referencia a *parámetros*. Mientras las poblaciones «fluyen», sus atributos son completamente estables. Por ejemplo, considere el parámetro «edad media» en la población de ciudadanos estadounidenses. Si usted fuera capaz de medir la edad media de la población de Estados Unidos, ésta no variaría de un día para otro, aunque se incorporarán muchas personas a esta población por nacimiento o por inmigración, y muchas otras saldrán por fallecimiento o por cambio de nacionalidad. El flujo de gente entrante y saliente de la población tiene un efecto insignificante

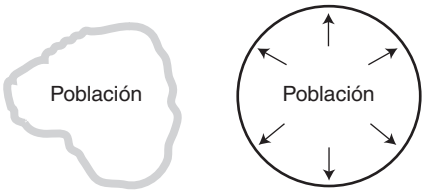


FIGURA 3-1 Diagramas de Venn de grandes poblaciones con bordes borrosos, difuminados y flechas que se alejan del centro.

en la edad media. Aunque las poblaciones puedan ser entes inmensurables, son muy reales y sus parámetros son inflexibles; existe una edad media de la población estadounidense que no variará ni en unos días ni en varias semanas, aunque sí podrá cambiar gradualmente a lo largo de los años.

Un parámetro como la edad media es fácil de entender. Pero también hay parámetros algo más abstractos, como el riesgo de sufrir un accidente en una población de conductores borrachos o la tasa de supervivencia en una población de personas con un cierto tipo de cáncer. Un parámetro poblacional es numéricamente muy sólido, porque si no somos capaces de saberlo exactamente, sí que podemos estimarlo con cierto grado de certidumbre usando técnicas estadísticas.

El hecho de que los atributos o parámetros poblacionales sean constantes es un concepto esencial en la bioestadística. Considere el parámetro de riesgo de accidente en una población de conductores borrachos. Definiremos la población por la característica común de nivel de alcohol en sangre por encima de un cierto umbral. Además, deberán cumplir dicho criterio estando al volante de un vehículo en movimiento. Podrá darse cuenta de que es una población que cambia constantemente. Los individuos entran y salen de esta población continuamente. Es imposible tenerlos en cuenta a todos en el mismo momento.

El parámetro *riesgo* es un número con un valor particular. Es la probabilidad de que un conductor borracho sufra un accidente en una unidad de tiempo mientras está conduciendo. El riesgo existe y es un valor concreto, aunque la población fluya (¡en diversas direcciones!). No podemos medir este riesgo con exactitud, pero la idea clave es que cualquier parámetro de una población es estable, y aunque no tengamos el número exacto, poseemos los métodos para estimarlo a través de observaciones.

- *Un parámetro poblacional es un valor numérico que resume los datos poblacionales.*

No estamos diciendo que no podamos cambiar este riesgo. De hecho, disminuir el riesgo en una situación concreta es una de las aplicaciones quintaesenciales de la bioestadística. Como los conductores ebrios tienen un riesgo más alto que los sobrios, hemos incrementado la conciencia social sobre el problema y hemos promulgado leyes para desalentar a conducir en esta situación. Además, se han diseñado coches más seguros para reducir el riesgo de morbilidad y mortalidad en caso de accidente.

CONCEPTOS PRÁCTICOS SOBRE POBLACIONES

Una población se define como un conjunto de entes que tienen características comunes. De hecho, los miembros de una población pueden compartir (y ocurre a menudo) más de una característica. Éstas deberían ser medibles (es decir, debería haber una forma razonable de distinguir si un individuo posee o no una característica).

Por ejemplo, podemos querer estudiar el efecto de un nuevo agente anticanceroso en un grupo de personas con un cierto tipo de cáncer. Definiríamos nuestra población como las personas con este tipo de cáncer. Debería demostrarse que los miembros de la población tienen esta característica medible (un tipo de cáncer) a través del diagnóstico de una biopsia realizada por un patólogo. Podemos querer excluir a aquellos con cáncer en fase más avanzada si creemos que podrían no responder positivamente al tratamiento. Si observamos una respuesta favorable en nuestro experimento, posteriormente podremos estudiar a aquéllos con metástasis. De momento, podríamos excluirlos de la población.

En los ensayos clínicos, las poblaciones se definen por los criterios de selección o de elegibilidad. Por una parte, estos criterios deberían ser muy restrictivos ya que los resultados podrían no aportar nada si existe demasiada variación entre las características de los individuos. Por ejemplo, si se incluyen en la población pacientes con diferentes tipos de leucemia, podemos pasar por alto el hecho de que el nuevo agente sea eficaz contra un tipo concreto de leucemia, ya que el efecto quedaría diluido entre los otros sujetos. También ha de considerarse la financiación y los recursos. Si el ensayo propuesto es demasiado general, los costes pueden ser prohibitivos.

Por otra parte, si los criterios de exclusión son demasiado rigurosos, podemos no ser capaces de reclutar suficientes sujetos que cumplan todos los criterios, o puede llevarnos demasiado tiempo hacerlo. También existe la posibilidad de no observar un efecto potencial en un subconjunto de pacientes que habrían participado si no hubieran sido excluidos desde un principio.

Fíjese que es un reto identificar la población de estudio. En ocasiones, las poblaciones se componen de aquellos que tienen más probabilidad de responder al tratamiento. Observaciones previas o estudios piloto proporcionan dicha información. Que la condición estudiada sea frecuente facilita la inclusión de individuos. Mientras dispongamos de suficiente financiación, la población podrá ampliarse relajando los criterios de selección. Tener más individuos nos permite analizar el efecto del tratamiento en subgrupos de pacientes después de obtener los resultados. Remítase al caso de estudio del cuadro 3-1.

CUADRO 3-1

En el estudio Heart Outcomes Prevention Evaluation (HOPE)*, los investigadores estudiaron el efecto del agente antihipertensivo ramiprilato en la población de pacientes con enfermedad cardiovascular o diabetes[2]. Se demostró, en estudios previos, que tenía efectos beneficiosos en las personas con insuficiencia cardíaca y que prevenía infartos de miocardio. Los investigadores cambiaron esta población para estudiar a aquellos pacientes con una función cardíaca normal.*

Criterios de selección

- *Hombres y mujeres de al menos 55 años.*
- *Antecedentes de enfermedad coronaria, ictus, vasculopatía periférica o diabetes, además de al menos uno de estos síntomas: hipertensión, colesterol total elevado, lipoproteínas de alta densidad (HDL), fumador o microalbuminuria.*
- *Ningún paro cardíaco.*
- *No tomar inhibidores de la angiotensina II.*
- *No tomar vitamina E.*
- *No tener hipertensión inestable.*
- *No tener nefropatía manifiesta.*
- *Ningún infarto de miocardio previo o ictus en las cuatro últimas semanas.*

Como las patologías cardiovasculares y la diabetes son frecuentes, no sorprendió la inclusión de 9.297 pacientes que también cumplían los otros criterios. A causa del gran número de pacientes incluidos, el efecto del fármaco también pudo estudiarse en subgrupos de pacientes. En total se estudiaron diez subgrupos, como aquellos pacientes con/sin hipertensión, con/sin diabetes y con/sin ataque cardíaco previo. Los resultados mostraron un efecto global beneficioso en los individuos y en nueve de los diez subgrupos. Actualmente, esta información se usa para tratar a aquellos pacientes que cumplen los criterios de la población estudiada.

CARACTERÍSTICAS NO IDENTIFICADAS

Sin duda existen características entre los miembros de la población que no estarán identificadas. Los criterios de selección no tienen en cuenta *todas* las características posibles. Por ejemplo, en un ensayo sobre el tratamiento del cáncer puede no preguntarse sobre el consumo diario de vitaminas, que podría variar considerablemente de unos individuos a otros. ¿Podría esto influir en el resultado observado? Teóricamente, las vitaminas *podrían* cambiar la respuesta al tratamiento.

Si el ensayo comparara un tratamiento farmacológico con un remedio naturalista, y los sujetos fueran libres de elegir la opción que prefirieran, los que toman vitaminas podrían preferir la sustancia natural. Si este tratamiento resultara mejor que el farmacológico, podría ser consecuencia de una de las vitaminas en vez del remedio naturalista. En este caso se adjudicaría al tratamiento un beneficio que podría deberse parcialmente a un factor que no se tuvo en consideración.

Los resultados podrían haber sido distintos; el efecto de una vitamina podría haber disminuido el efecto de la sustancia natural o provocar una toxicidad nociva. En este caso se hubiese afirmado erróneamente que la sustancia farmacológica tuvo un beneficio añadido sobre la sustancia natural, cuando, en realidad, fue la vitamina (que no se tuvo en consideración) la que influyó en los resultados.

Las características no identificadas que podrían afectar potencialmente al resultado se denominan *confusoras*. Sin embargo, si los factores no considerados se reparten equitativamente entre los grupos, se anula su efecto. Cuando los individuos se asignan aleatoriamente a los grupos y no son ellos los que deciden, los factores no considerados se distribuyen equitativamente y su efecto se anulará en los resultados.

EL ARTE DE LA MEDICINA

Es sencillo tratar a los propios pacientes cuando cumplen los criterios de un estudio relevante. No es complicado aplicar los resultados de un estudio bien diseñado a los pacientes apropiados. No obstante, habrá muchas ocasiones en que tenga que tomar una decisión sobre un tratamiento para un individuo que cumple algunos (pero no todos) de los criterios que definieron la población. Otro hecho habitual consiste en que la compañía de seguros no cubra el coste del medicamento del estudio, pero sí el de un sustituto más barato.

Generalmente no es una buena práctica extrapolar los resultados de un estudio a un paciente que no forma parte de la población. Además, existe un argumento válido en contra de prescribir un sustituto y esperar los mismos resultados. Aquí es donde el *arte* de la práctica médica entra en juego. En estas situaciones se necesita sentido común y pensamiento racional para aconsejar a los pacientes. Le será de ayuda mantenerse al corriente de la bibliografía en su ámbito de trabajo. También debe hacer partícipes a los pacientes en la discusión de las opciones. Se les debería animar a tomar parte en sus propias decisiones de tratamiento y a expresar una opinión después de proveerles de la información disponible. Avíseles, sin embargo, de que los resultados pueden no ser completamente transferibles a su caso particular.

En resumen, la bioestadística no trata con el individuo, sino con las masas. Distánciese de lo individual; aprenda a pensar en *grande*. En vez de pensar en un paciente concreto, piense en cientos de pacientes como ése. Entonces podrá aprovechar esta clarividente experiencia y usarla en beneficio de su paciente. Aprenda a confiar en la

CUADRO 3-2

Un ejemplo que ilustra las avanzadas revisiones en el pensamiento médico está extraído de un artículo recientemente publicado sobre la disección aórtica aguda, un trastorno de riesgo vital[3]. Un retraso en el diagnóstico puede tener un efecto deletéreo en la probabilidad de supervivencia. Esta entidad se conoce de antaño, pero el diagnóstico puede ser esquivo. Los investigadores estudiaron los síntomas presentados en 464 pacientes con una disección aórtica aguda que se incluyeron en un registro de doce centros de referencia en seis países.

Descubrieron que muchos de los síntomas clásicos descritos en los libros de texto no eran tan frecuentes como antes se pensaba. Clásicamente, el dolor característico lo habían definido como lacerante y migratorio en la cavidad torácica. Sin embargo, la mayor parte de sujetos definieron su dolor como agudo y no migratorio. Además, a menudo otros síntomas típicos —disminución del pulso, soplo cardíaco y radiografía de tórax anormal— estaban ausentes. Curiosamente, muchos pacientes presentaron pérdida del conocimiento, un síntoma previamente subestimado.

Esta información sin duda será útil para los clínicos en un futuro cuando evalúen a un paciente con dolor de pecho agudo o síncope. Se reescribirán los libros de texto cuando se disponga de información más fiable. Los síntomas clásicos considerados previamente serán reemplazados por una descripción diferente basada en la evidencia reciente.

bibliografía para guiar sus decisiones. Esto no implica disminuir la calidad humana de la práctica médica, sino distanciarse emocionalmente de las recomendaciones que da.

La bioestadística requiere, sobre todo, que mantengamos una mentalidad abierta. No es en absoluto incorrecto cambiar las recomendaciones siempre y cuando sea en beneficio del paciente. A medida que se aporta nueva evidencia, ésta debería incorporarse a la práctica médica. Los mejores profesionales de la asistencia médica trabajan en un entorno muy fluido y continuamente actualizan su base de conocimiento. Ellos tienen presente que la evidencia más fiable proviene, por lo general, de grandes poblaciones de pacientes con las mismas características que el individuo que ellos tratan. Considere el ejemplo del cuadro 3-2.

PUNTOS CLAVE

- Una población es un conjunto de entes que tienen alguna característica cuantificable en común. Pueden ser múltiples características.
- Las poblaciones se definen por sus características comunes, denominadas criterios de selección.
- Las poblaciones también se definen por las características comunes de las que carecen; se denominan criterios de exclusión.
- Las características no medidas en una población son un aspecto muy importante en el diseño experimental.
- La ciencia de la inferencia estadística se centra en el estudio de poblaciones; los resultados pueden aplicarse a todos los individuos que pertenezcan a la población.
- Muchas poblaciones son tan grandes que no podemos enumerar a todos sus miembros.

- Las poblaciones tienen atributos denominados parámetros, que son valores numéricos que resumen los datos. Éstos se diferencian claramente de sus características cuantificables.
- Aunque las poblaciones fluyan, sus parámetros son estables.
- Tenga cuidado cuando extrapole los resultados de estudios a una población específica de pacientes que no cumplan los criterios.
- Continuamente se usa la evidencia reciente para actualizar las fuentes de información médica.

BIBLIOGRAFÍA

1. *The American College Dictionary*. 1969. New York: Random House, Inc., p. 943.
2. The HOPE Study Investigators. 2000. Effect of an angiotensin-converting enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. *New Engl J Med*, 342:3, 145.
3. Hagan, P. G. et al. 2000. The International Registry of Acute Aortic Dissection. *JAMA*, 283:7, 897.

PREGUNTAS DE REVISIÓN

1. Las _____ son conjuntos de personas o cosas con características cuantificables comunes.

2. Los criterios de_____ definen una población.

3. Las características _____ podrían influir potencialmente en el resultado de un estudio si no están distribuidas equitativamente entre los grupos.

4. Las poblaciones son un concepto teórico, pero sus _____ son muy reales y medibles.

5. Si un paciente que es miembro de una población estudiada prefiere un medicamento más barato que aquél para el cual se demostró el beneficio, debería decirle que _____.

RESPUESTAS

1. poblaciones

2. selección

3. no identificadas

4. parámetros

5. no espere los mismos resultados que en el estudio. Sin embargo, en muchos casos se puede realizar, dependiendo de la información disponible.

CAPÍTULO 4

Muestras

Una característica básica de la ciencia experimental es la necesidad de obtener conclusiones a partir de información incompleta.

—M. Anthony Schork y Richard D. Remington[1]

Las poblaciones se basan en conceptos teóricos, pero los individuos que las componen son muy reales. Necesitamos un método concreto para estudiar una población identificada. Como hemos visto, las poblaciones pueden ser muy grandes; algunas tienen tantos miembros que no podemos tenerlos a todos en cuenta. También son muy inestables, con un flujo continuo de individuos entrante y saliente. A menudo es imposible estudiar a cada individuo, pero esto no significa que sea imposible estudiar las poblaciones; para ello confiaremos en una *muestra*.

Una muestra es un grupo de individuos que representan a la población. Si seleccionamos la muestra correctamente, los resultados que obtendremos serán muy parecidos a los que habríamos obtenido si hubiésemos observado la población entera. El atributo medido en la muestra se acercará al valor del atributo en la población.

Por ejemplo, si quisiéramos saber la distribución de las diferentes religiones en Estados Unidos, podríamos preguntar a cada persona para averiguar cuál es su creencia religiosa. Por supuesto, esto es imposible. Sin embargo, si escogiésemos una muestra de gente correctamente, las proporciones de las diferentes religiones observadas en la muestra reflejarían de forma precisa las proporciones en la población.

SELECCIÓN DE LA MUESTRA

Lo ideal es seleccionar una muestra que represente equitativamente a todos los individuos de la población; es decir, que cada miembro tenga la misma probabilidad de ser escogido. Esto se conoce como *muestreo aleatorio simple*. La *condición de independencia* establece que la elección de un individuo no influye en las opciones de cualquier otro de ser elegido. Cuando el muestreo se realiza siguiendo estas reglas, gracias a las leyes de la probabilidad, sabemos el grado de certidumbre de nuestras observaciones respecto al verdadero resultado de la población. No obstante, los números no serían idénticos. Por ejemplo, la proporción de baptistas puede ser del 10% en nuestra muestra, mientras que el valor real en la población puede ser del 12% (datos ficticios).

- *Una muestra aleatoria simple se escoge de modo que cada posible combinación de* N *individuos tiene la misma probabilidad de ser seleccionada.*

Si cada individuo de una población tiene, teóricamente, la misma probabilidad de ser escogido, ¿cuántas combinaciones de muestras de tamaño *N* existen? Si la población es grande (que es lo habitual), entonces existe un número increíblemente grande de combinaciones posibles que podrían formar la muestra. Por ejemplo, en una población pequeña de 20 individuos, si quisiéramos estudiar una muestra de tamaño 5, ¡habría 15.504 combinaciones posibles! Existe una fórmula para calcularlas*, pero la lección importante no es cuántas combinaciones diferentes de individuos podrían escogerse como muestra, sino el hecho de que, en un muestreo aleatorio simple, cada combinación tiene la misma probabilidad de estar en la muestra.

Puede parecer que al utilizar muestras estamos comprometiendo nuestro rigor científico para no complicarnos la vida. Es cierto que en las muestras no disponemos de la información completa que tendríamos si estudiáramos toda la población, y por lo tanto no obtendremos una respuesta exacta. Sin embargo, necesitamos llegar a conclusiones a partir de esta información incompleta. Así, es totalmente aceptable estudiar una población a partir de una muestra siempre y cuando aceptemos un poco de incertidumbre en la interpretación de los resultados. Hemos de ser conscientes que con las muestras estamos estimando el resultado que obtendríamos si estudiáramos a toda la población.

Éste es un punto crucial en la bioestadística. Es importante saber distinguir entre las poblaciones y las muestras que las representan. Los atributos que caracterizan una población son los *parámetros*. En cambio, a los atributos observados en la muestra los denominamos *estadísticos muestrales* o *estadísticos*. En la práctica, usamos los estadísticos para estimar los parámetros. En el ejemplo de las religiones, el estadístico del 10% de baptistas en la muestra es una estimación del parámetro poblacional del 12% de baptistas (datos ficticios).

- *El valor de un atributo en una muestra se denomina estadístico. Un estadístico es una estimación del parámetro, o valor real, de aquel atributo en la población.*

ESTIMACIONES E INCERTIDUMBRE

Ya que el estadístico nos proporciona una estimación del parámetro poblacional, podría desviarse de éste en una u otra dirección; es decir, el estadístico podría ser una sobrestimación, una subestimación o el mismo parámetro. No sabemos con certeza qué escenario es el correcto, pero estamos dispuestos a aceptar un margen de error. Los estadísticos a menudo se notifican con su valor al que se le suma o resta una cierta cantidad. Este rango de valores es el *intervalo de confianza*. Se trata de una zona de seguridad que intenta contener el verdadero parámetro poblacional. Un intervalo de confianza estrecho significa que nuestro estadístico estará, muy probablemente, bastante próximo al parámetro poblacional.

*La fórmula es: $M!/N!\ (M\text{-}N)!$, donde M = tamaño de la población; N = tamaño de la muestra, y *!* representa la operación de multiplicar el número dado consecutivamente por los enteros menores del valor hasta alcanzar el 1.

- *El intervalo de confianza es el estadístico de la muestra al que se le suma o resta el margen de error.*

El intervalo de confianza es un rango de valores deducidos de la muestra que tiene una probabilidad dada de contener el valor real que buscamos; refleja el margen de error que los estadísticos poseen intrínsecamente para estimar el parámetro poblacional. Más adelante veremos cómo se calcula, pero de momento el concepto clave consiste en que existe una probabilidad conocida de que los límites del intervalo de confianza contengan el verdadero parámetro poblacional.

Se suele usar un grado de incertidumbre del 5% en el estadístico. Si el experimento se repitiese varias veces con muestras aleatorias simples del mismo tamaño, obtendríamos resultados diversos. Tendríamos diferentes estimaciones del verdadero parámetro, cada una con su propio intervalo de confianza.

La figura 4-1 representa los posibles resultados de un experimento que se repitiese múltiples veces. La línea negra gruesa representa el verdadero parámetro, que realmente desconocemos. Cada intervalo de confianza está entre paréntesis, con el esta-

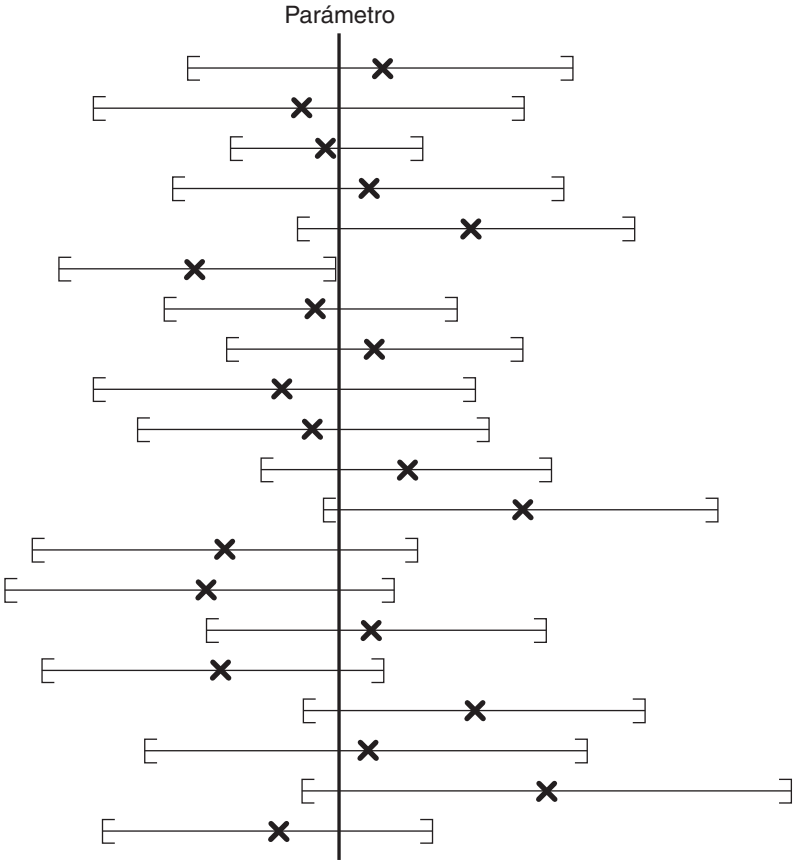


FIGURA 4-1 Veinte muestras del mismo tamaño de una misma población. Cada muestra nos da un estadístico representado por **X**. Los límites de los intervalos de confianza se marcan con corchetes, [ ].

dístico en el centro, representado por una **X**. En general, el 95% (o 19 de cada 20) de los intervalos de confianza contendrán (a la larga) el verdadero parámetro. Debido a que sólo realizamos el experimento una vez, podría producirse cualquiera de las situaciones de la figura.

En el ejemplo sobre las creencias religiosas, no somos capaces de saber el verdadero 12% de baptistas, pero nos permitimos decir que nuestro estadístico es el 10% ± 3%. Así, el intervalo de confianza va del 7 al 13%. No estamos al 100% seguros que el intervalo de confianza incluya el parámetro real, pero sí tenemos una certeza del 95% si hemos seguido las reglas del muestreo aleatorio y hemos aplicado la prueba estadística correcta. En este caso, el intervalo de confianza incluye el parámetro verdadero (12% de baptistas). No obstante, tenga presente que no podemos saber realmente dónde está el parámetro dentro del intervalo. Sin embargo, podemos decir, con una confianza del 95%, que el intervalo con los límites del 7 y el 13% contiene el valor real.

Es intuitivo que una muestra aleatoria más grande proporcione un resultado más preciso que una más pequeña. Realmente es así. Una muestra más grande estimará el parámetro, en promedio, de manera más precisa y tendrá un intervalo de confianza más estrecho debido a que las muestras más grandes tienen una variabilidad global menor de la característica estudiada. Sin embargo, llegará un punto en el que un aumento del tamaño muestral aporte un beneficio menor. A partir de entonces, el mayor grado de certeza que pueda obtenerse ampliando el tamaño muestral no compensará la inversión de tiempo, dinero y molestias a las personas. También se han de tener en cuenta los aspectos éticos, ya que los individuos que participan en experimentos aceptan el riesgo potencial de exponerse a una intervención aún no probada. Si el tamaño muestral es más grande de lo necesario para obtener la respuesta que buscamos, ¿por qué exponer a este riesgo desconocido a más individuos de los necesarios?

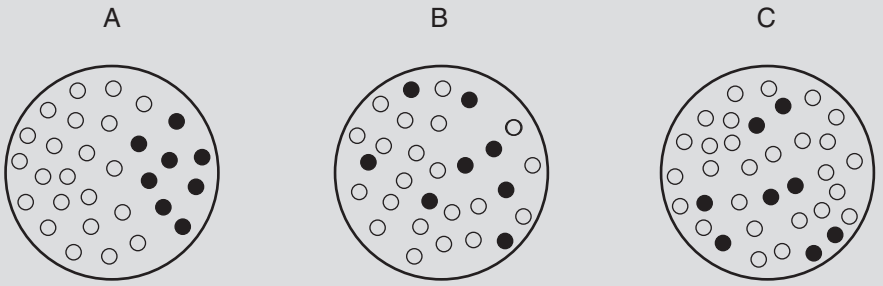
Por otra parte, si el estudio tiene pocos individuos, puede no ser capaz de descubrir una diferencia en el efecto del tratamiento que realmente existe. Una vez diseñado el experimento, los profesionales estadísticos pueden estimar el número de participantes que deben incluirse. Existe una fórmula (v. cap. 14) para calcular el número mínimo de participantes para obtener una respuesta *razonablemente* precisa sobre si una opción concreta es mejor que otra. Fíjese en la palabra «razonablemente» de la oración precedente. Hemos mencionado que los investigadores se sienten cómodos con un pequeño grado de incertidumbre en la estimación del comportamiento de una población. Precisamente, uno de los elementos que intervienen en el cálculo del tamaño muestral es el grado aceptable de incertidumbre, que por lo general es del 5%.

MUESTREO ALEATORIO Y PRECISIÓN

Considere el ejemplo del cuadro 4-1 sobre un control de calidad en una cadena de producción. Con el objetivo de probar la resistencia a la tensión de una población de muelles producidos en una fábrica, usamos una muestra aleatoria para ver cuántos cumplen las especificaciones. Si cada muelle tiene la misma probabilidad de representar bien la variación de la resistencia a la tensión, entonces escoger un grupo de muelles producidos consecutivamente sería una buena representación de la población. No obstante, hay muchos motivos por los cuales el valor de la resistencia podría agruparse

CUADRO 4-1

Cada uno de estos círculos representa la misma población. Los círculos pequeños representan a individuos, y de éstos, los sombreados son individuos invitados a participar en un estudio. ¿Cuál de estos ejemplos representa de forma más adecuada el muestreo aleatorio?



Es realmente una pregunta delicada. Pensemos sobre ello. La respuesta más cauta podría ser B, ya que visualmente garantiza que la muestra representa la diversidad de la población. Sin embargo, hay que tener en cuenta otra cosa. Si los miembros de una población están distribuidos aleatoriamente con respecto a las características estudiadas, entonces coger un pedazo de la población (como en A) también representará una muestra aleatoria. El problema radica en que en la bioestadística las características tienden a agruparse en relación con su valor. Los valores, por lo general, no están distribuidos aleatoriamente. Los baptistas, al igual que los miembros de otras religiones, tienden a agruparse en ciertas regiones. Las parejas mostradas en el diagrama C tendrían vínculos entre ellos, por lo que no cumplirían la condición de independencia e introducirían sesgo en los resultados.

durante la producción. Puede haber un técnico que esté especialmente atento a las especificaciones de la producción y los muelles sean mejores cuando esta persona está en el lugar de trabajo, o podría haber una alteración en la maquinaria a lo largo del día debida al cambio de temperatura del equipo o a la cantidad de lubricante en los engranajes. Hasta podrían haber diferencias entre los distintos turnos, ya que los trabajadores nocturnos podrían dormir menos y, por lo tanto, no estar tan atentos.

En este caso, una muestra de muelles producidos en todos los turnos y todos los días de la semana sería más representativa de toda la población. Por otra parte, este ejemplo ilustra de qué manera un estudio diseñado cuidadosamente podría proporcionar una estimación errónea del parámetro poblacional debido a un error en el plan de muestreo. Los estudios clínicos publicados deben describir el método de muestreo usado. Cuando lea este apartado en un estudio, sea consciente de que la muestra realmente podría no representar a la población.

VALIDACIÓN EXTERNA E INTERNA

Es habitual incluir en los estudios a los pacientes de centros de referencia, que son los que usualmente realizan dichos estudios. A menudo se incluyen de forma consecu-

tiva. Si los individuos cumplen los criterios de estudio, se les invita a participar uno tras otro. Si los pacientes no tuvieran ninguna otra conexión, las leyes del muestreo aleatorio no se habrían violado, sobre todo cuando el estudio se realiza simultáneamente en distintos emplazamientos (ensayo multicéntrico). Pero ¿representan estos pacientes a la población de pacientes de la comunidad promedio? Los pacientes atendidos podrían ser diferentes de los pacientes generales por muchos motivos, aun teniendo la misma enfermedad. Para empezar, fueron al médico. Este hecho podría indicar solamente que estos pacientes son más cuidadosos con su atención médica o que podrían tener un acceso más fácil al sistema sanitario. Podrían ser más obedientes y prestar más atención a la dieta y a otros factores ambientales que no se tienen en cuenta en el estudio. Podrían tener un seguro con mayor cobertura y permitirse una mejor asistencia y todos los medicamentos recomendados. También podrían estar menos enfermos que aquellos que no fueron al médico, o estar más preocupados por su salud. Éstos podrían ser algunos de los factores no considerados que podrían influir potencialmente en los resultados de un estudio.

De alguna manera, el hecho de que los pacientes pasen por la puerta de la consulta los hace diferentes de sus equivalentes en la comunidad. Nos estamos dando cuenta de que los resultados de los estudios procedentes de centros de referencia no siempre son reproducibles en la comunidad. A menudo es necesario observar los resultados

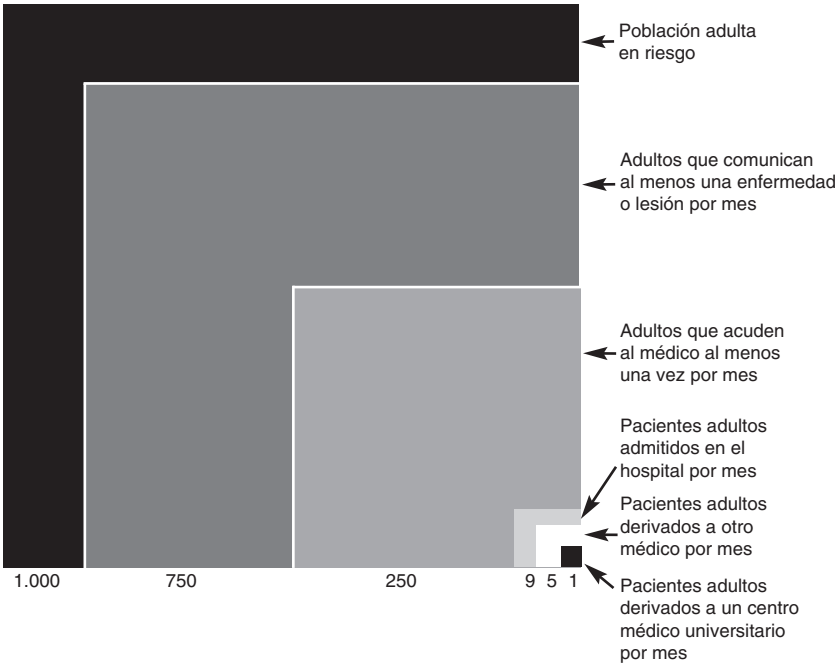


FIGURA 4-2 Modelo de referencia de selección en la asistencia sanitaria. Los pacientes enviados a un centro médico universitario son sólo una parte muy pequeña de todos los que tienen la misma enfermedad en la comunidad. Pueden no representar una muestra completamente aleatoria de la población. (Redibujado de la figura 1 de White, K., T. Williams, y B. Greenberg. 1961. The ecology of medical care. *N Engl J Med*, 265:885-892, con autorización.)

de una cierta intervención en una comunidad durante varios años para conseguir una estimación precisa del efecto en la práctica. Este proceso se conoce como *validez externa*, a diferencia de la *validez interna*, que se refiere a la correcta metodología estadística seguida durante el estudio. Véase la figura 4-2.

PUNTOS CLAVE

- Normalmente es imposible estudiar a cada individuo de una población.
- Las muestras deben representar a todos los individuos de la población.
- La ley del muestreo aleatorio afirma que cada individuo de una población tiene la misma probabilidad de ser incluido en la muestra.
- La condición de independencia requiere que la elección de un individuo de la muestra no influya en la posterior elección de otro.
- Si se cumplen estas normas, cada muestra de tamaño *N* tiene la misma probabilidad de ser seleccionada.
- Las muestras le permiten *estimar* el resultado que usted observaría si hubiese estudiado toda la población.
- Los atributos observados en una muestra son los estadísticos, mientras que los auténticos valores de la población son los parámetros.
- Como no estudiamos toda la población, no sabemos el valor del parámetro.
- Los estadísticos son una estimación del parámetro.
- Como el estadístico es una estimación, se representa con un rango de valores a ambos lados conocido como intervalo de confianza. Si el experimento se repitiera en múltiples ocasiones, el intervalo de confianza dejaría de incluir el parámetro poblacional en el 5% de las repeticiones.
- Las muestras más grandes producen estimaciones más precisas del parámetro poblacional; tienen intervalos de confianza más estrechos.
- Las muestras pequeñas pueden no tener suficientes individuos para hallar una diferencia en el efecto del tratamiento aunque realmente exista. Existe una fórmula para estimar el número de individuos necesarios para minimizar este tipo de error.
- Normalmente, los resultados de estudios realizados en grandes centros de referencia no son exactamente reproducibles en la comunidad porque la muestra puede no ser representativa de la población.

BIBLIOGRAFÍA

1. Schork, M. A. y R. D. Remington 2000. *Statistics with applications to the biological and health sciences*, 3rd ed. Upper Saddle River, NJ: Prentice Hall.

PREGUNTAS DE REVISIÓN

1. Como no podemos estudiar a cada individuo de una población, elegimos una _____ que represente a la población.

2. Los datos de una muestra producen un _____, que es una _____ del parámetro poblacional.

3. Si una muestra se escoge siguiendo las leyes del _____ _____, el intervalo de confianza del estadístico contendrá, muy a menudo, el _____ poblacional si el experimento se repitiese en múltiples ocasiones.

4. En general, las muestras más grandes tienen intervalos de confianza _____.
5. Las muestras más pequeñas tienen limitada su capacidad de hallar una _____ _____, si realmente existe.
6. Los pacientes que participan en estudios de centros de referencia grandes pueden no representar la población porque el proceso de selección puede ser _____.

RESPUESTAS

1. muestra
2. estadístico, estimación
3. muestreo aleatorio, parámetro
4. más estrechos
5. diferencia en el tratamiento
6. sesgado o imperfecto o no aleatorio

CAPÍTULO 5

Invariablemente variables

La estadística es una herramienta que permite a los datos generar conocimiento en lugar de confusión.

—David S. Moore y George P. McCabe[1]

Si usted cogiera un libro de estadística, el primer capítulo sería invariablemente sobre variables. Es el momento de empezar a pensar en las poblaciones en términos de sus variables e incorporar esta idea a su conocimiento. Es aquí donde radica el fundamento de la verdadera comparación entre grupos. El concepto de variable es realmente simple.

- *Una variable es cualquier característica medible hallada en todos los individuos de una población.*

Fíjese que esta definición es muy amplia. Podrían identificarse una cantidad ilimitada de variables (o de características medibles) en cualquier población, pero no todas las variables son importantes en el tema tratado. Los investigadores eligen aquellas variables que más probablemente les ayudarán a determinar el resultado final.

Las variables se usan para estimar los parámetros poblacionales. Una variable específica puede medirse en cada individuo de la muestra. Un ejemplo es la variable «edad». Cada miembro tendrá un valor asociado a la edad. A partir de estos valores podrá obtenerse un estadístico que estime el parámetro poblacional *edad media*, o también la *desviación típica* (o *estándar*) *de la edad*.

TIPOS DE VARIABLES

Las variables en estudio se miden en todos los individuos de la muestra que se ha seleccionado para representar a la población. Los valores numéricos se asignan como representantes del valor de una variable en un individuo. Los números son necesarios para poder hacer los cálculos, aun si los números no representan directamente el valor de la característica. Considere el género, por ejemplo. Cada individuo es hombre o mujer; no obstante, para fines estadísticos, a menudo se registran los hombres con un 0 y las mujeres con un 1 (o viceversa).

Otro ejemplo es la facultad en una muestra de estudiantes universitarios. Esta variable no puede medirse numéricamente, pero para propósitos estadísticos, puede adjudicarse un 1 a Psicología, un 2 a Historia, un 3 a Biología, etc. Observe que en estos ejemplos el número adjudicado a la variable no representa un valor numérico real. Este tipo de variables se denominan *categóricas*. Los números se asignan a los subconjuntos de la categoría, de modo que un ordenador pueda realizar los cálculos para analizar los datos. Este tipo de variables tienen un código de categorías para descifrar lo que significan los valores numéricos.

Otras variables pueden representarse fácilmente por números que directamente reflejan el valor del individuo. Estas variables se denominan *cuantitativas*. Los ingresos anuales son un buen ejemplo. El número de dólares por año ganados por un individuo es una representación directa del valor real de la característica. Otro ejemplo es el número de kilómetros conducidos en un día de trabajo. El número tiene un sentido real: más kilómetros realizados se corresponden con un valor numérico más elevado.

Usted puede encontrarse con otras definiciones de tipos de variables. Así, el género puede clasificarse como una variable *binaria* o *dicotómica*, ya que sólo hay dos valores posibles. Las respuestas a las preguntas de «Sí» o «No» («¿Posee vehículo propio?») también pueden considerarse variables dicotómicas. No es importante recordar las diversas formas de clasificar las variables, pero sí debería saber qué clase de pruebas estadísticas son idóneas para analizar cada tipo de variable.

Inherente a la definición de variable es el hecho de que, efectivamente, *varían*. Individuos diferentes tendrán valores diferentes para una variable concreta. En algunos casos, como en el género, el valor será una de las dos opciones posibles. Otras variables tendrán un número finito de opciones, como la carrera universitaria. Las variables con un número finito de valores posibles se denominan *discretas*. Sin embargo, las variables como los ingresos anuales o el nivel de colesterol sérico pueden tener un amplia variedad de respuestas dentro de un rango numérico. Éstas se denominan variables *continuas* ya que existe un amplio rango de valores posibles dentro de una escala continua.

ESCALAS DE MEDIDA

Los diferentes tipos de variables nos han llevado a otra clasificación de las variables basada en la escala de medida que usan para determinar su valor. No es necesario memorizarla, pero es útil saber que existe. Hay cuatro escalas clásicas de medida:

Escala nominal. Realmente no es ni muchísimo menos una escala; más bien es un sistema de etiquetaje. Las variables categóricas, como el género y la carrera universitaria, se miden de esta forma. Aunque se asignen números a las categorías, éstos no tienen un valor numérico real. Son como los dorsales de las camisetas de fútbol que distinguen a los diferentes integrantes de un equipo: carecen de un valor cuantitativo.

Escala ordinal. En esta escala se da un valor a la variable basado en la posición dentro de una serie. La posición relativa de la variable tiene algún sentido numérico, por ejemplo, si queremos medir el puesto en que los corredores de una carrera cruzan la línea de meta. El primer y segundo corredor pueden alcanzar la meta más juntos que el segundo y tercer corredor, pero este tipo de

escala sólo presta atención a la posición. Usando esta escala, la diferencia entre el primer y segundo puesto es la misma que entre el segundo y tercer puesto. Los aspectos relacionados con la calidad de vida a menudo se miden con escalas ordinales. Considere una escala que mida la satisfacción con respecto a la salud personal, siendo 0 el mínimo y 100 el máximo. En una población de pacientes que se han sometido a una amputación, una persona que obtiene una calidad de vida de 95 tiene una mayor satisfacción que alguien con un 85, que a su vez está más satisfecho que alguien que obtiene un 75. No obstante, en la escala que mide la satisfacción no podemos decir que la diferencia de 10 unidades entre los tres individuos sea equivalente.

Escala de intervalo. Esta escala se usa para medir variables continuas que tienen valores matemáticos legítimos. La diferencia entre dos valores consecutivos es constante en cualquier punto de la escala. Muchas variables pueden medirse con esta escala. Por ejemplo, alguien que gana 60.000 dólares por año tiene el doble poder adquisitivo que alguien que gana 30.000 dólares y la mitad de aquellos que ganan 120.000 dólares anuales.

Escala de razón. Esta escala tiene un valor cero real (que indica «ausencia de»). Es útil para realizar comparaciones entre conjuntos de variables que usan escalas diferentes. La distancia relativa entre el valor y el cero para cualquier caso puede compararse con otra mediante cocientes. Esta escala no se usa muy a menudo en bioestadística.

El *valor* de una variable es obviamente importante. Estos valores se incorporan a las fórmulas matemáticas que finalmente producirán la respuesta buscada. Sin embargo, hay muchos modos diferentes de mirar los datos y realizar comparaciones. Cada tipo particular de variable usa una prueba estadística concreta para sacar conclusiones. Estas pruebas emplean fórmulas matemáticas diferentes según el tipo de espacio entre valores consecutivos. Por lo tanto, *es el tipo de variable y la escala de medida lo que determina la prueba estadística idónea que finalmente se usará para sacar una conclusión sobre los datos*. De hecho, la capacidad de afirmar que una diferencia es *estadísticamente significativa* puede depender de la escala de medida y del tipo de análisis. Corresponde al investigador escoger entre los métodos alternativos de tratar los datos cuál es el que proporcionará la conclusión más válida.

Nuevamente, no es necesario saber matemáticas para poder aplicar los conceptos. Sin embargo, sí es útil saber que los datos categóricos (como el género) se tratan de forma distinta a los datos ordinales (como los corredores de una carrera), que a su vez se manejan diferente que los datos continuos (como los ingresos). También tenga presente que para facilitar el análisis, algunos investigadores pueden tratar datos continuos como ordinales agrupándolos. En ocasiones, los ingresos se representan así. Estamos familiarizados con encuestas que preguntan por un grupo de ingresos anuales en vez de por los ingresos reales (p. ej., 0-20.000 dólares, 20.000-50.000 dólares, etcétera). La agrupación de los datos concentra la información y proporciona conclusiones válidas, pero en el análisis se usará una prueba estadística diferente.

La tabla 5-1 se publicó recientemente en el *Seventh Report of the Joint National Committee on the Prevention, Detection, Evaluation and Treatment of High Blood Pressure* (Séptimo Informe del Comité Nacional Conjunto para la Prevención, la Detección, la Evaluación y el Tratamiento de la Hipertensión). Este grupo revisó la bibliografía sobre la hipertensión y recomendó pautas para tratar a los pacientes. Aunque la presión arterial es una variable continua, se marcaron algunos intervalos de valores para crear cate-

TABLA 5-1 Clasificación y tratamiento de la hipertensión para mayores de 18 años

					*Tratamiento**	
					Tratamiento farmacológico inicial	
Clasificación de la PA	***PA sistólica, mmHg****		***PA diastólica, mmHg****	***Modificar el estilo de vida***	***Sin obligaciones indicadas***	***Con obligaciones indicadas***
Normal	<120	y	<80	Se alienta		
Prehipertensión	120-139	o	80-89	Sí	Sin medicamentos antihipertensivos indicados	Medicamento(s) para las indicaciones obligadas[†]
Hipertensión fase I	140-159	o	90-99	Sí	Para la mayoría, diuréticos tipo tiazida; pueden considerarse inhibidores ECA, BRA, β- bloqueadores, BCC, o combinaciones	Medicamento(s) para las obligaciones indicadas Otros medicamentos antihipertensivos (diuréticos, inhibidores ECA, BRA, β-bloqueadores o BCC) que se necesiten
Hipertensión fase II	>160	o	>100	Sí	Combinación de dos o más medicamentos (normalmente un diurético tipo tiazida y un inhibidor ECA o BRA o β-bloqueadores o BCC)[§]	Medicamento(s) para las obligaciones indicadas Otros medicamentos antihipertensivos (diuréticos, inhibidores ECA, BRA, β-bloqueadores o BCC) que se necesiten

*Tratamiento determinado para la categoría de PA más alta.
[†] Pacientes tratados con enfermedad renal crónica o diabetes con PA objetivo menor que 130/80 mmHg.
[§] Terapia combinada que debería usarse con cautela en aquellos pacientes con riesgo de hipotensión ortostática.
BCC: bloqueador de los canales del calcio; BRA: bloqueador del receptor de la angiotensina; ECA: enzima de conversión de la angiotensina; PA: presión arterial.
(De Chobanian AV, Bakris GL, et al., 2003. *JAMA*, 289:29, 2561, con autorización).

gorías. Ésta es una práctica habitual que recurre al instinto humano para agrupar y clasificar. Otro ejemplo que hemos visto son las agrupaciones de filtrado glomerular para los niveles de enfermedad renal de la figura 2-9 del capítulo 2.

En este punto (si se siente algo aventurero), vaya al apéndice A. Este flujograma muestra los diferentes tipos de pruebas estadísticas que se usan para analizar datos según el tipo de variables, la escala de medida y otros factores de los cuales aún no hemos hablado. Puede parecer intimidador, pero ilustra los pasos que siguen los estadísticos para decidir las fórmulas apropiadas que aplican en cada caso. Aunque las fórmulas difieran, todas producen un valor *p* que se interpreta de la misma manera. No es necesario memorizar la prueba apropiada para cada situación. Usted debería ser capaz de interpretar el resultado sea cual sea la prueba usada.

Una versión mucho más simple del diagrama se muestra en el apéndice B. Fíjese que la primera bifurcación del árbol depende del tipo de variable y de la escala de la medida. Aunque esta presentación visual no sea tan completa como la otra, deja claro que cada tipo de datos tiene sus propias pruebas estadísticas.

CONJUNTOS DE DATOS

A primera vista, los conjuntos de datos parecen bastante...

repetitivos	***agotadores***	***aburridos***
monótonos	tediosos	cansinos
redundantes	rutinarios	sosos...

…¡y lo son!, por eso tenemos ordenadores que hacen todo el trabajo. Sin embargo, está bien saber cómo se organizan los conjuntos de datos, pues realmente no son tan confusos como aparentan. Un conjunto de datos es una cuadrícula de casillas que contienen los valores de las variables. En la cabecera están los nombres de las variables, que constituyen las columnas. La primera columna se reserva para los individuos, que a menudo se designan por un número. Obviamente, estos números no se incluyen en el análisis final. Las filas, alineadas horizontalmente, muestran todos los valores de las variables para un individuo particular. El conjunto de datos más simple tiene cuatro casillas (dos individuos con un valor de una variable cada uno). Los más complejos pueden almacenar millones de casillas.

La siguiente muestra de datos se tomó de Moore y McCabe[1]. Los investigadores se preguntaban por qué una proporción particularmente grande de estudiantes de primer año de ciencias informáticas de una gran universidad no llegaron a graduarse. Los datos de la tabla 5-2 se recogieron de estudiantes de primer año de ciencias informáticas después de tres semestres.

Este extracto incluye los cinco primeros individuos de un conjunto de datos típico. En la cabecera de las columnas están las etiquetas de las variables. La primera columna es Observaciones o Individuos (OBS, *Observation or Subjects*). Los números del 001 al 005 se asignan rutinariamente a los estudiantes y no tienen valor numérico. GPA es el promedio de las notas *(grade point average)* después de tres semestres. Podría considerarse un intervalo de datos con un rango de 0 a 4,0. Las variables son continuas ya que podrían tomar cualquier valor dentro de este rango.

TABLA 5-2 Datos de los estudiantes de primer año de ciencias informáticas después de tres semestres

OBS	*GPA*	*HSM*	*HSS*	*HSE*	*SATM*	*SATV*	*GÉNERO*
001	3,32	10	10	10	670	600	1
002	2,26	6	8	5	700	640	1
003	2,35	8	6	8	640	530	1
004	2,08	9	10	7	670	600	1
005	3,38	8	9	8	540	580	1

Los resultados completos del estudio están en Campbell, P. F. y G. P. McCabe, 1984. Predicting the success of freshmen in a computer science major. *Communications of the ACM*, 27, 1108-1113, con autorización.

HSM, HSS y HSE representan las notas en matemáticas, ciencias e inglés, respectivamente, en los estudios secundarios *(high school scores)*. La escala va de 0 a 10, siendo 10 una A, 9 una A–, 8 una B+, etc. Como las variables tienen un conjunto de valores limitado, podrían considerarse discretas (no hay ningún valor de 9,3, p. ej.). El tipo de escala puede considerarse de intervalo si la diferencia entre una A y una B se considera igual que la diferencia entre una B y una C. SATM y SATV representan las puntuaciones en las pruebas de matemáticas y capacidad verbal. Éstas son variables continuas medidas en una escala de intervalo. La última columna es el género, que se registró con un 1 para los hombres y con un 2 para las mujeres. Éste es un ejemplo de una variable discreta y dicotómica en escala nominal.

Este conjunto limitado de datos ilustra los diferentes tipos de variables que pueden estudiarse. También muestra que hasta los datos más simples pueden tratarse de modos diferentes —posiblemente todos correctos— según sean la cuestión planteada y las premisas (o asunciones) razonables. Es bastante habitual que dos estadísticos expertos discrepen en los puntos más sutiles del análisis. Una manera de evitar estos conflictos en un estudio es decidir el tipo de prueba estadística *antes* de recoger los datos.

De nuevo, no se necesita saber cómo se desarrolló la parte matemática; sólo saber que existen diferentes vías de tratar los datos que dependen de los tipos de variables y de las premisas aceptadas por los investigadores. Cuando usted evalúe un estudio, será capaz de identificar estas premisas observando el tipo de prueba estadística usada.

VARIABLES EXPLICATIVAS Y VARIABLE RESPUESTA

Cuando se diseña un estudio para medir el efecto de una o varias variables en otra, se asume que ciertas variables pueden influir en el resultado (variable dependiente). Los estadísticos denominan a estas variables influyentes como *explicativas,* ya que ayudan a explicar la respuesta y afectan al resultado final. La variable dependiente se denomina variable *respuesta*. Existe algo más que una simple asociación entre las variables explicativas y la variable respuesta: la presencia y el valor de las variables explicativas *determinan* el valor de la respuesta.

Por ejemplo, un investigador podría querer estudiar el efecto del consumo de alcohol en el tiempo de reacción del conductor. Él o ella podría clasificar el nivel de alcohol en sangre como la variable explicativa. El tiempo transcurrido entre ver una luz roja y pisar el pedal de freno podría ser la variable respuesta. Parece probable que un nivel elevado de alcohol en sangre esté asociado a un tiempo de reacción más prolongado. Pero hay algo más que una asociación entre estas dos variables; una tiene un efecto sobre la otra.

El ejemplo citado podría sugerir que la presencia de una variable *causa* el efecto en la otra. Sin embargo, demostrar la causalidad es un poco más complicado que una mera observación de que una variable dada tiene una relación con la respuesta. Han de realizarse varios estudios aleatorizados con resultados concluyentes antes de que la comunidad científica acepte una relación de causa-efecto entre dos variables. En el estudio de los acontecimientos adversos, uno de los argumentos más sólidos para respaldar una relación de causa-efecto es que el suceso desaparezca cuando se elimina la variable explicativa.

PUNTOS CLAVE

- Una variable es cualquier característica medible que se encuentra en todos los individuos de una población.
- Las variables categóricas son asignaciones de números que representan categorías. Estas cifras no tienen ningún sentido numérico.
- Las variables cuantitativas tienen un número que representa un valor real.
- Existen cuatro escalas clásicas de medida para clasificar las variables: Nominal, Ordinal, Intervalo y Razón. (NOIR. ¿Alguien sabe francés?)
- El tipo de variable y la escala de medida determinan la prueba estadística apropiada para contestar a la cuestión investigada.
- Es posible agrupar una variable continua de modo que use una escala de medida diferente, según las premisas hechas por los investigadores que deciden la manera más válida de tratar los datos.
- Los conjuntos de datos se organizan en casillas que contienen los valores de las variables del estudio.
- La variable explicativa tiene una relación con la variable respuesta, pero no demuestra necesariamente causalidad.

BIBLIOGRAFÍA

1. Moore, D. S. y G. P. McCabe. 1999. *Introduction to the practice of statistics,* 3rd ed. New York: W. H. Freeman and Co.

PREGUNTAS DE REVISIÓN

1. Una variable es una característica medible con _____ valores posibles.

2. A las variables _____ se les asigna una cifra aunque no tengan ningún sentido numérico.

3. La _____ de medida y el _____ de variable determinan la prueba estadística.

4. En el gráfico de ingresos según la educación (fig. 2-8, cap. 2), los _____ es la variable respuesta ya que los _____ ayudan a explicar el nivel de ingresos.

5. En el gráfico del fallo renal (fig. 2-9, cap. 2), la _____ es la variable respuesta ya que _____.
6. Los resultados de un estudio pueden variar si una variable se mide en escalas diferentes. Corresponde a los _____ y a los _____ determinar la escala de medida y la prueba estadística más válida.

RESPUESTAS

1. dos o más (según el tipo de variable)
2. categóricas
3. escala, tipo
4. ingresos anuales, años de educación secundaria
5. creatinina sérica, la función renal produce un efecto en los niveles de creatinina
6. investigadores, estadísticos

CAPÍTULO 6

Variables respuesta

Salud es «un estado de completo bienestar físico, mental y social, y no solamente la ausencia de afecciones o enfermedades».

—Organización Mundial de la Salud

Hemos afirmado que el objetivo del profesional sanitario es ayudar a la gente a vivir más tiempo y a sentirse mejor identificando la forma de conseguirlo. La investigación halla esta forma observando la evolución de una muestra dividida en varios grupos expuestos a diferentes intervenciones. Si un grupo tiene un beneficio cuantificable sobre otro en «vivir más» o en «sentirse mejor», recomendaremos esta intervención a un individuo que pertenezca a la población en estudio.

Este capítulo trata de un tipo concreto de variable que tiene el honor de representar a nuestro objetivo. Pretende capturar la condición de vivir más o sentirse mejor en las personas. Esta variable única se denomina *variable respuesta*. Las variables respuesta miden este objetivo.

AÑOS DE VIDA AJUSTADOS POR LA CALIDAD

El «vivir más» es fácilmente medible a través de los años de vida. En conjunto, un grupo será mejor si hay más personas vivas al final del estudio. La medida del tiempo de vida no es complicada; esencialmente, es la tasa de mortalidad. El beneficio de vivir más años es evidente y está públicamente aceptado en la mayoría de casos.

De todas maneras, es importante considerar la calidad de los años de vida que se otorgan a los individuos. No siempre es beneficioso alargar la vida si cada día es una carga. Disfrutar es una parte importante en la satisfacción global de un individuo. Ha aparecido una reciente oleada de interés por los métodos que cuantifican la «calidad de vida» (CV). Obviamente, esta variable respuesta es más difícil de medir que los años de vida, ya que tiende a ser más subjetiva y variable entre individuos en las mismas condiciones. A pesar de las limitaciones inherentes a este tipo de investigación, se están realizando grandes avances en este ámbito a través de técnicas sofisticadas que intentan cuantificar los llamados factores de bienestar.

La mayor parte de las medidas de CV se realizan a través de cuestionarios, que pueden rellenar los afectados bien directamente, bien con ayuda de profesionales convenientemente adiestrados. Estos instrumentos con frecuencia usan escalas tipo

Likert para medir las respuestas (p. ej., malo, normal, bueno, muy bueno, excelente). Antes de aprobarse su uso, son sometidos a rigurosas pruebas de validez y fiabilidad. La *validez* contesta a la pregunta de si un instrumento mide lo que desea medir. La mayoría de métodos para comprobar la validez emplean otra determinación del mismo objetivo y estudian su correlación. La *fiabilidad* es el espejo de la reproducibilidad. Un instrumento fiable mostrará cierta variabilidad, pero será más consistente que uno no fiable.

Algunos cuestionarios tratan aspectos genéricos de la CV, como el Cuestionario Breve del Estado de Salud (*SF-36^R: Short Form Health Status Questionnaire*). Este cuestionario plantea 36 preguntas que abarcan múltiples categorías relacionadas con la salud (también llamadas dominios); intentan medir los niveles físicos, emocionales, sociales y cognoscitivos del funcionalismo del paciente. Por lo tanto, pueden usarse en varias situaciones para valorar el estado funcional general y el bienestar global. Hay varias versiones de este tipo de cuestionario. La figura 6-1 desglosa los tipos de dominios del cuestionario.

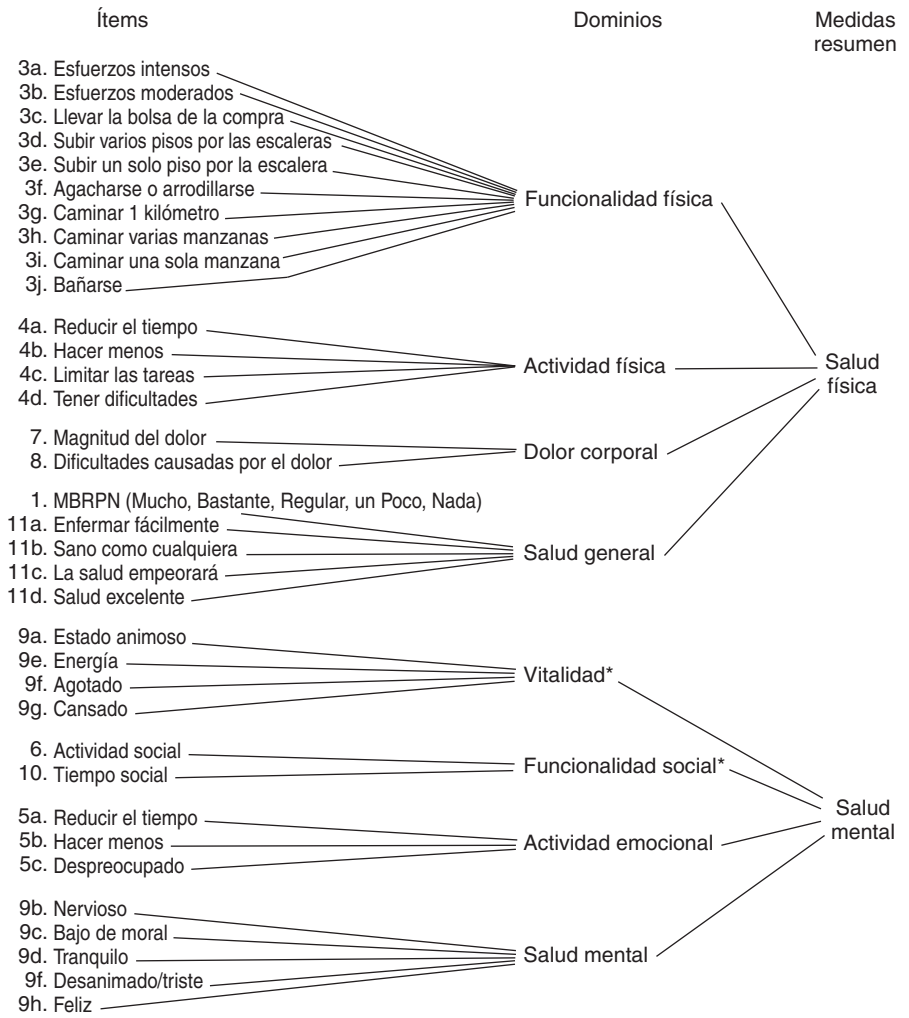
La figura 6-2 muestra los resultados de un estudio transversal de pacientes con diferentes enfermedades de los cuales se evaluó su CV con una medida genérica conocida como el *Sickness Impact Profile* (perfil del impacto de la enfermedad). Son datos descriptivos que comparan diferentes enfermedades en función del bienestar total percibido por el paciente. Los resultados no se ajustaron ni por la edad ni por otros factores, y los estados de las enfermedades mostradas no incluyen todas sus manifestaciones. Las barras más largas representan una percepción de la CV más pobre.

La mayor parte de los resultados son previsibles. Los pacientes con enfermedades crónicas que causan un dolor incontrolable aportaron, en general, una CV menor. Los que manifestaban una peor CV en este estudio padecían un proceso degenerativo neurológico llamado esclerosis lateral amiotrófica (ELA, o enfermedad de Lou Gehrig), que produce falta de fuerza y disminución de la capacidad psicomotriz, pero (quizá lamentablemente) no afecta a la función intelectual. Puede darse el desafortunado caso de que estos individuos soliciten ayuda para poner fin a sus vidas.

Por otra parte, un resultado bastante sorprendente es que los que han sufrido un paro cardíaco aportan una sorprendente ¡buena CV! Estas personas a menudo no tienen ninguna lesión permanente una vez que se les ha controlado el episodio, y pueden reflexionar sobre su fortuna. Quizá los que han visto la luz al final de túnel valoran más el hecho de estar vivos, y sus otras enfermedades llegan a ser banales en comparación.

Además de las medidas de CV generales, otras herramientas se centran en aspectos específicos de la enfermedad relacionados con la CV. Son más aplicables a enfermedades crónicas y frecuentes, y también sirven para comparar diferentes momentos de tiempo. Por ejemplo, se han desarrollado cuestionarios para medir la CV experimentada por pacientes con paro cardíaco congestivo (PCC). La conocida clasificación NYHA (New York Heart Association; Asociación del Corazón de Nueva York)[1] es una herramienta simple que evalúa el nivel de los síntomas experimentados por el paciente desde la Clase I (asintomático con las actividades del día a día) hasta la Clase IV (síntomas en reposo). Los médicos usan esta clasificación para verificar el éxito de un tratamiento en un paciente concreto. También se aplica a grupos de pacientes con PCC para evaluar la eficacia de programas e intervenciones a gran escala.

Las enfermedades físicas y psíquicas concurrentes que afectan a la autoperspectiva de un individuo pueden afectar a la validez de la medida de CV. Por ejemplo, una



* Correlación significativa con otra medida resumen.

FIGURA 6-1 El SF-36^{R} es un cuestionario para valorar el bienestar global (calidad de vida).

persona con dolor crónico por una patología ortopédica puede también sufrir una depresión, que disminuiría su CV. Puede ser complicado evitar estos factores interdependientes. Si los investigadores quieren excluir de la muestra a los pacientes con depresión clínica, deberían cribar por esta condición.

A diferencia de los años de vida, las medidas de CV también intentan tener en cuenta los inoportunos efectos secundarios de un tratamiento, desde una desagradable erupción cutánea hasta un rechazo de médula ósea con riesgo vital. Si un efecto secundario provoca el ingreso en un hospital, las medidas de CV deben reflejar este acontecimiento adverso, aunque la vida del paciente no corra peligro. La mayoría de

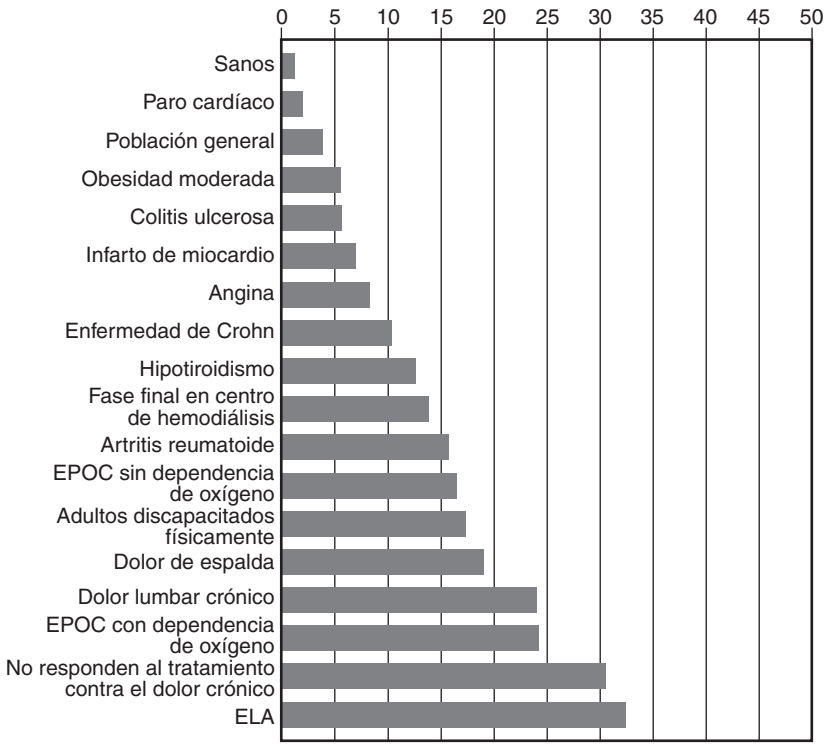


FIGURA 6-2 Puntuaciones del perfil del impacto de la enfermedad (SIP: *Sickness Impact Profile*) para distintas enfermedades o grupos poblacionales. Puntuaciones más altas reflejan una calidad de vida (CV) más pobre. (Patrick D. L. y R. A. Deyo, 1989. Generic and disease: specific measures in assessing health status and quality of life. *Medical Care*, 27[suppl]:3, 5220.)

los estudios mencionarán la gravedad y la frecuencia de los acontecimientos adversos aun cuando no se mida la CV. Se supone que estas complicaciones tienen un impacto negativo en la CV global.

Los años de vida y la CV se combinan en la variable «años de vida ajustados por la calidad» (AVAC), que no sólo mide la longevidad sino también el nivel de bienestar experimentado en cada año. En teoría, la variable AVAC concentra la respuesta global a un tratamiento y una unidad representa vivir un año más en plena capacidad.

- *Los años de vida ajustados por la calidad (AVAC) es la medida esencial.*

En la práctica, sin embargo, la CV es difícil de medir. En muchos casos no se dispone de los datos para realizar el análisis estadístico. En estos casos sólo se usan las tasas de mortalidad para obtener los resultados, y lo único que se considera es que es mejor estar vivo que muerto (¡creo que la mayoría de nosotros estará de acuerdo!). Esta simplificación se realiza cuando las condiciones tienen un impacto insignificante en la CV total, o si la ausencia de comodidad en una persona se limita a un período de tiempo relativamente corto. Sin embargo, si usted ve un estudio que no está enfocado a la CV, pero cree que uno de los tratamientos en comparación puede

tener un impacto considerablemente negativo en el bienestar de un paciente, incorporará su evaluación a las recomendaciones finales de los autores.

EL PODER DE LA PREVENCIÓN

La importancia de la variable AVAC puede apreciarse cuando un individuo se somete a un procedimiento de vida o muerte. Por ejemplo, una operación quirúrgica exitosa de rotura aórtica salva a la víctima de una muerte segura en una intervención dramática. El AVAC otorgado durante este procedimiento puede apreciarse hasta por aquellos que son remotamente conscientes de la gravedad de esta situación.

Sin embargo, no subestime el poder de la *prevención* en el incremento de AVAC. Por ejemplo, la prevención de una enfermedad vascular a través de unos hábitos de vida saludables es una forma muy rentable de proporcionar AVAC. Al paciente fumador con una hipertensión leve, le sería mucho más beneficioso renunciar a la nicotina que empezar un tratamiento para minimizar el aumento de la presión arterial. A los profesionales sanitarios nos conviene recalcar a nuestros pacientes la importancia de los hábitos saludables.

El efecto beneficioso de la prevención es aún más evidente en ciertas enfermedades infantiles. La vacunación ha proporcionado un inmenso número de AVAC adicionales —más de los que cualquier intervención invasiva podrá alcanzar jamás—. Una única dosis o una aspiración nasal, cuando se administra en masa, tiene un efecto potencial beneficioso no sólo en el individuo sino también en la población. No todo el mundo habría contraído la enfermedad, pero vacunando impedimos que la enfermedad la sufran más personas. Debido a que la mayoría de estas enfermedades se contraen en la infancia, el número de AVAC concedidos a aquellos que se habrían perdido su vida adulta (como los supervivientes de la polio, aunque deban vivir con una deficiencia) es enorme.

La diferencia entre una intervención espectacular que salva a un individuo y la prevención de una enfermedad en la comunidad, con el consecuente ahorro de múltiples vidas, es que normalmente el beneficio de la prevención no es tan perceptible. Tendemos a fijarnos en un individuo que se debate entre la vida y la muerte y se salva milagrosamente, pero no somos tan conscientes del impacto de la prevención en evitar malos resultados. La prevención es un instrumento muy potente en el suministro de AVAC adicionales a grandes grupos de personas.

RESPUESTAS INTERMEDIAS O SUBROGADAS

Sin duda, encontrará muchos estudios clínicos que no usan la variable respuesta ideal, el AVAC. Si todos los estudios esperaran pacientemente a medir las tasas de mortalidad antes de publicarse, probablemente nos moriríamos antes de disponer de los resultados. Si un investigador sabe que ciertas variables respuesta están asociadas con baja CV o con elevadas tasas de mortalidad, podría sustituir los AVAC por estas variables representantes. Es una práctica aceptada usar respuestas intermedias que reflejen indirectamente el AVAC, sobre todo cuando no puede obtenerse una medida directa del AVAC de una forma razonable y la respuesta subrogada es relativamente fácil de medir. La respuesta intermedia debería tener una correlación directa con el

resultado que representa. Por supuesto, cuantificar el impacto de una intervención en el AVAC mediante una respuesta subrogada es más delicado.

Un ejemplo de esta aplicación se ve en estudios que usan el grosor de la pared carótida como un indicador de arteriosclerosis. El grosor de las partes internas y medias de las paredes del vaso sanguíneo puede medirse fácilmente con técnicas de ultrasonido no invasivas. Debido a que esto está asociado con el avance de la enfermedad de la arteria coronaria, se ha usado como un indicador para seguir la evolución subclínica de la enfermedad cardiovascular. También se ha usado como una medida intermedia alternativa de la evolución de la enfermedad cardiovascular[2].

SUPERVIVENCIA LIBRE DE ACONTECIMIENTOS

Los investigadores que supervisan estudios clínicos están interesados en minimizar la probabilidad de una mala respuesta en una población. La mejor opción tendrá más supervivientes sanos al final del estudio. Normalmente, otra respuesta que aparece en la bibliografía es la *supervivencia libre de acontecimientos*, que es parecida al AVAC. Esta respuesta sigue la pista a los principales acontecimientos adversos que disminuyen el AVAC, pero no los pondera según su efecto individual en el AVAC. Alguien que evita un suceso adverso importante, como un ataque cardíaco, un ictus o la muerte, se denomina un superviviente libre de acontecimientos cuando las comparaciones entre intervenciones se realizan al final del estudio. Este tipo de respuesta explica los principales factores que condicionan la variable AVAC, pero no tiene en consideración disminuciones más pequeñas, como las relacionadas con efectos secundarios de la medicación.

RESPUESTAS COMBINADAS

También es habitual ver estudios que persiguen lo opuesto al AVAC: acontecimientos adversos y muertes que ocurren en los grupos. Son los llamados riesgos *(hazards)*. No siempre es necesario concentrarse en un único factor responsable de un mal resultado, sobre todo si cualquiera de las enfermedades relacionadas podría haber sido la culpable.

Por ejemplo, las enfermedades vasculares afectan a los vasos sanguíneos de muchos órganos. A medida que la enfermedad vascular evoluciona, la persona tiene un riesgo creciente de ataque cardíaco o de ictus. Ambos pueden causar la muerte o la invalidez permanente. No importa dónde ocurre anatómicamente el acontecimiento deletéreo; lo importante es que hay una evolución negativa. Cuando las respuestas potenciales están relacionadas, la variable respuesta puede combinarse incluyendo varias respuestas. Entonces se observa a los individuos para ver si se produce cualquiera de las respuestas. Esto se denomina *respuesta combinada*.

Cada una de las respuestas se identifica como un acontecimiento deletéreo, y la diferencia de riesgo entre los dos grupos se muestra antes que en el caso que la variable principal fuera un suceso solitario. Se emplea la regla aditiva de la probabilidad. Por ejemplo (puramente especulativo), si en una rama del estudio se observara que el riesgo de ictus es del 2% anual y el riesgo de ataque cardíaco, del 3% anual, el riesgo de ataque cardíaco *o* ictus sería del 5% anual, o 5 de cada 100 (suponiendo que ningún individuo contrae ambos).

Los acontecimientos combinados no son concurrentes en un individuo. Si seguimos la pista al riesgo de ataque cardíaco *e* ictus, multiplicaríamos los riesgos individuales. En el ejemplo citado, el riesgo de tener ambas patologías sería aproximadamente del 0,06% anual, o 6 de cada 10.000 (asumiendo que estas enfermedades son independientes y no están relacionadas, aunque lo más probable es que lo estén; esto modificaría un poco el cálculo).

Si buscáramos obtener respuestas por separado, tendríamos que esperar mucho tiempo antes de poder apreciar algún beneficio en un grupo sobre el otro, ya que la frecuencia de la respuesta concurrente sería muy baja. La combinación de respuestas de forma aditiva permite que una diferencia en el riesgo aparezca bastante más pronto. Muchos estudios publicados usan la conjunción *y* cuando definen sus acontecimientos combinados, cuando en realidad la conjunción más apropiada es *o*. La ventaja principal de usar acontecimientos combinados es acortar el tiempo necesario de observación. La respuesta a la pregunta de qué opción proporciona un mayor beneficio se observará antes si se siguen simultáneamente varios efectos adversos relacionados.

El ensayo HOPE de 2000[3] utilizó un acontecimiento combinado de infarto de miocardio, ictus *o* muerte por enfermedad cardiovascular, estudiando el efecto del ramipril en una población de personas de riesgo seguida, en promedio, unos 5 años. Tanto el infarto de miocardio como el ictus pueden tener un fuerte impacto en la CV. Se combinaron con la mortalidad por enfermedad cardiovascular para formar la *variable respuesta principal*. Cada respuesta puede analizarse por separado. En este estudio, el ramipril tuvo un efecto beneficioso en reducir la tasa del acontecimiento combinado y también los acontecimientos individuales.

PUNTOS CLAVE

- La variable respuesta es la medida de vivir más y/o sentirse mejor al final del estudio.
- «Vivir más tiempo» se mide a través de la mortalidad.
- «Sentirse mejor» se mide a través de la calidad de vida.
- La combinación de estas respuestas proporciona los años de vida ajustados por la calidad (AVAC), que es la medida respuesta esencial.
- Las respuestas intermedias pueden sustituir a las variables respuesta si guardan alguna correlación con «vivir más tiempo» o «sentirse mejor», pero el efecto puede ser más difícil de cuantificar que en el caso de usar una verdadera medida de AVAC.
- Los acontecimientos combinados albergan múltiples acontecimientos adversos en una sola medida.
- La utilización de acontecimientos combinados puede disminuir el tiempo en observar una diferencia de tratamiento entre los grupos.
- La prevención puede ser un modo muy eficaz de otorgar AVAC adicionales, aunque su impacto no sea evidente.

BIBLIOGRAFÍA

1. www.hcoa.org/hcoacme/chf.cme/chf 00070.htm, 2005.
2. O'Leary, D. H. et al. 2002. Intima-media thickness: A tool for atherosclerosis imaging and event prediction. *Am J Cardiol*, 90(suppl), 18L–21L.
3. The HOPE Study Investigators. 2000. Effect of an angiotensin-converting enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. *New Engl J Med*, 342:3, 145.

PREGUNTAS DE REVISIÓN

1. El objetivo de los profesionales sanitarios es mejorar la condición humana ayudando a la gente a _____ _____ y a _____ _____.

2. «Vivir más tiempo» se mide a través de la _____.

3. «Sentirse mejor» se mide a través de la _____ ___ _____.

4. Los _____ es un intento de combinar las dos medidas anteriores.

5. La _____ es un modo eficaz de alcanzar más AVAC.

6. A la larga, una medida de tratamiento o de prevención eficaz disminuye el _____ de una mala respuesta en la población.

7. Las respuestas _____ pueden acortar la duración de un estudio.

RESPUESTAS

1. vivir más, sentirse mejor

2. mortalidad

3. calidad de vida

4. años de vida ajustados por la calidad o AVAC

5. prevención

6. riesgo

7. combinadas

CAPÍTULO 7

Probabilidad

La idea revolucionaria que marca el límite entre los tiempos modernos y el pasado es el control del riesgo: el concepto de que el futuro es algo más que un capricho de los dioses y que los hombres y las mujeres no son entes pasivos ante la naturaleza.

—Peter L. Bernstein[1]

La ciencia de la inferencia estadística no se desarrolló en una noche. Se necesitaron cientos de años de principios matemáticos básicos dentro de una disciplina que usa las observaciones para predecir el futuro. Las teorías probabilísticas sobre las que se sustenta la inferencia estadística fueron surgiendo paulatinamente a través de un proceso diligente que refleja varios siglos de búsqueda y curiosidad intelectual.

LA HISTORIA DE LA TEORÍA DE PROBABILIDADES

La estadística es un producto relativamente reciente de unos pocos siglos de antigüedad. Los primeros sistemas de cálculo eran capaces de contar un gran número de ítems, pero no reconocían ideas abstractas como «ninguno» (significado de 0) o los números negativos. La aritmética moderna de hoy en día surgió hace aproximadamente 1.500 años, cuando los hindúes desarrollaron un sistema de numeración que incorporaba el concepto del cero. Esta noción permitió el desarrollo de un sistema contable más estándar que permitía incluir un número limitado de dígitos (del 0 al 9) uno al lado del otro, y su valor relativo dependía de la posición que ocupaban (unidades, decenas, centenas, etc.). Desde ese momento no había ningún límite en el valor que podía expresar un número.

Los árabes adoptaron el nuevo sistema cuando viajaron a través de la India, y aún lo perfeccionaron más. En el siglo IX, un matemático árabe llamado Al-Khowarizmi escribió una de las primeras obras matemáticas sobre ecuaciones algebraicas simples. Su nombre evolucionó hacia la palabra *algoritmo*, que significa «reglas de cálculo». Otra de sus obras, *Hisab al-jabr w' almuqabalah (Ciencia de la transposición y la cancelación)*, origina el término *álgebra* (del árabe *al-jabr*). Los nuevos números encontraron alguna resistencia en Europa occidental, pero finalmente fueron aceptados como una forma lógica de expresar valores y solucionar ecuaciones. Este nuevo sistema desencadenó la teoría de probabilidades, basada en fórmulas capaces de cuantificar el riesgo.

A los humanos nos gusta la comodidad que proporcionan los resultados previsibles, pero vivimos en un mundo imperfecto. Nos esforzamos en construir dispositivos infalibles, pero cualquier equipamiento que fabricamos tendrá una tasa de fallos. No podemos asegurar que no nos ocurrirá nada malo cuando viajamos. Tampoco podemos garantizar que un tratamiento tendrá efecto. No nos queda otra opción que aceptar un poco de incertidumbre sobre el futuro. Sin embargo, gracias a las leyes de la probabilidad, podemos tomar decisiones informadas que nos permiten maximizar la probabilidad de un buen resultado.

El concepto revolucionario de que los humanos podemos predecir el futuro es relativamente reciente. Hasta hace pocos siglos, la gente creía en un Dios omnipotente que controlaba los acontecimientos futuros. Ellos eran incapaces de moldear su destino. Fuera del marco de la ciencia lógica, confiaban en las fascinantes predicciones de místicos y adivinos. Al descubrirse las leyes de la física en la Edad Media, la gente empezó a darse cuenta de que existía algún orden en los acontecimientos celestiales y en el mundo físico. Fue entonces cuando comenzó a disolverse el aura de misticismo que rodeaba los acontecimientos mundanos.

Las personas siempre nos hemos sentido intrigadas por los juegos de azar. Se han hallado pruebas de la existencia de estos juegos en pinturas egipcias que se remontan al 3500 a. C. y hay constancia de ellos en todas las sociedades y clases sociales a lo largo de la historia. Al igual que empezamos a percibir que el mundo se rige por leyes físicas, también descubrimos que ciertos juegos podrían ser predecibles usando fórmulas matemáticas. El deseo de predecir los resultados conllevó un interés hacia la «ciencia de decidir».

Se escribieron grandes obras durante el Renacimiento sobre temática probabilística. Una de las más relevantes fue un tratado de un médico italiano del siglo XVI llamado Girolamo Cardano. Fue un ávido jugador que desarrolló las fórmulas algebraicas para calcular los resultados esperados al lanzar dados. Actualmente, los dueños de los casinos usan estas fórmulas para asegurarse de que tendrán un beneficio a largo plazo y que no perderán dinero.

Los resultados en los juegos de azar (como los dados) pueden calcularse matemáticamente porque se conoce la probabilidad de cada acontecimiento. En una partida de dados, sabemos la probabilidad de conseguir una suma de 7 cuando tiramos dos dados. Existen seis combinaciones que podrían producir un 7 de un total de 36 combinaciones posibles. Las fórmulas probabilísticas nos permiten calcular la frecuencia con la que esperamos sacar un 7 cuando lanzamos los dados más de una vez. Una situación análoga es el lanzamiento de una moneda. Sabemos que la probabilidad de conseguir cara es del 50%, pero usando las fórmulas de la binomial, también podemos calcular la probabilidad de conseguir, por ejemplo, una cuarta parte de caras en múltiples tiradas.

Un siglo antes de que se presentaran las fórmulas de Cardano, un monje italiano llamado Luca Pacioli había planteado la pregunta de cómo se debía dividir el dinero apostado en un juego cuando uno de los jugadores se retiraba. Este enigma permaneció sin resolver durante dos siglos, hasta 1654, cuando los célebres matemáticos franceses Pascal y Fermat fueron capaces de solucionar el rompecabezas prediciendo lo que era más probable que ocurriese si el juego proseguía.

La predicción de sucesos aleatorios en situaciones como los lanzamientos de moneda o de dados conllevó más preguntas hipotéticas. ¿Hay algún modo de predecir acontecimientos de los cuales se desconoce la probabilidad de un resultado concreto? Jacob Bernoulli fue un matemático interesado en la idea de aplicar la teoría de

probabilidades a estas situaciones abstractas. En 1713 se publicó su ensayo *Ars Conjectandi (Arte de las conjeturas)*; sugería que los datos de una muestra podían servir para calcular probabilidades desconocidas. Varios años más tarde, un abad llamado Thomas Bayes publicó un documento que le distinguió como un estadístico prominente. En *An Essay Towards Solving a Problem in the Doctrine of Chances (Ensayo para la solución de un problema en la doctrina de las probabilidades)*, presentó un sistema que permitía modificar las predicciones incorporando nuevas observaciones. Éstas han sido algunas de las contribuciones más notables a la teoría de probabilidades, que nos permiten cuantificar el riesgo.

EL CONCEPTO MODERNO DE RIESGO

Estos descubrimientos tuvieron grandes consecuencias que han condicionado el desarrollo de la sociedad moderna. Las personas estamos dispuestas a aceptar un riesgo si sabemos la probabilidad del resultado deseado. Nadie toma un riesgo en una situación completamente desconocida. Pero calcular la probabilidad del resultado deseado puede justificar una decisión, incluso sin unas garantías del 100%. La práctica de la medicina moderna no tiene ninguna garantía. Hay un riesgo inherente en cada tratamiento, pero ser capaz de predecir este riesgo justifica una elección. La inferencia estadística, con sus raíces en la teoría de probabilidades, permite calcular este riesgo. El innato interés humano por los juegos de azar ha conllevado descubrimientos extraordinarios sobre la ciencia de la predicción. Aunque algunos conceptos de la teoría probabilística parezcan relativamente simples, se necesitaron varios siglos para descubrirlos. Cuando aplicamos modernos cálculos de probabilidad para interpretar los resultados de estudios que guiarán nuestras decisiones médicas diarias, nos apoyamos en los hombros de gigantes matemáticos como Pascal y Fermat.

La aplicación moderna de la inferencia estadística ha seguido desarrollándose durante el siglo pasado. Contribuciones recientes han provocado el refinamiento de un proceso que actualmente define el procedimiento estadístico estándar. R. A. Fisher fue un genetista inglés que vivió durante la primera mitad del siglo xx y realizó experimentos agrícolas. Se le atribuye la introducción de la aleatorización en el diseño experimental y el desarrollo de métodos estadísticos para las pruebas de significación. También presentó el concepto de prueba de significación. Los célebres estadísticos Jerzy Neyman y Egon Pearson profundizaron en esta idea en la década de los años treinta. Promovieron una teoría de pruebas de hipótesis basada en un modelo de toma de decisiones que incorporaba las denominadas hipótesis nula y alternativa. Todo ello nos ha llevado a la era de la teoría estadística moderna.

Éstas son sólo algunas de las contribuciones recientes con un impacto significativo en la evolución del análisis estadístico. Esta ciencia continúa evolucionando actualmente, con el desarrollo de nuevas fórmulas matemáticas para interpretar diferentes tipos de datos. Y se formulan argumentos filosóficos en favor de alternativas metodológicas o de interpretaciones de resultados.

CUANTIFICACIÓN DE LOS RESULTADOS ALEATORIOS

La inferencia estadística se basa en la probabilidad de que un cierto resultado ocurra por azar. En la teoría de probabilidades, la palabra *suceso* se refiere al resultado obser-

vado. No necesariamente reflejará una variable respuesta como «años de vida ajustados por la calidad» (AVAC) que vimos en los ensayos clínicos. Es simplemente el resultado de un acontecimiento. El rango de probabilidades de un suceso oscila entre 0 (probabilidad nula de que ocurra) y 1 (pasará siempre). Es raro encontrar sucesos en la naturaleza con probabilidades iguales a 0 o 1. En tal caso, no sería necesario aplicar la teoría de probabilidades. Todos los acontecimientos estudiados en medicina tienen una probabilidad entre 0 y 1, como 0,35. Convertir estos valores en porcentajes nos da una interpretación más coloquial de la probabilidad de un acontecimiento. Se dice que un resultado con una probabilidad de 0,35 tiene una probabilidad del 35% de ocurrir o que, en promedio, pasará 35 de cada 100 veces. Se deduce que un suceso con una probabilidad del 100% equivale a que no existe ninguna posibilidad de que el resultado no ocurra (¡pero esto nunca ocurre!).

Un valor *p* en realidad es la probabilidad de que un suceso dado ocurra por casualidad. Habitualmente se expresa con decimales, como 0,07. Cuando multiplicamos un valor *p* por 100, es un porcentaje. En el ejemplo anterior, el valor *p* de 0,07 significa que hay una probabilidad del 7% de que el resultado observado ocurra por azar. (Estamos asumiendo que se cumplen ciertas premisas que veremos más adelante.) Otra manera de expresarlo es: si el estudio se repitiera cientos de veces en las mismas circunstancias, usando individuos de la misma población, en promedio sólo en 7 de cada 100 estudios obtendríamos el resultado observado por azar. La razón por la cual cada estudio no da el mismo resultado es que se usan muestras diferentes que, a su vez, dan estimaciones diferentes del parámetro. Hablaremos de este concepto posteriormente, pero de momento recordaremos que el valor *p* representa una probabilidad que puede expresarse como un porcentaje.

SUCESOS

Cuando yo era estudiante de bioestadística y estudiaba las leyes probabilísticas, aprendí un concepto muy aleccionador que merece la pena enfatizar:

- *Si algo puede pasar, pasará.*

Si es posible que un acontecimiento ocurra, no importa cuán mínima sea la probabilidad, finalmente ocurrirá. Podemos aún no haberlo visto e incluso no llegar a verlo en nuestra vida, pero es una certeza matemática de que si se cumplen las condiciones para producirse y con suficientes oportunidades, éste ocurrirá aunque tengan que pasar cientos de años.

El reto, por supuesto, es saber de antemano que algo puede pasar. Si nunca se ha observado, sólo podemos conjeturar que es posible si la lógica no nos dice lo contrario. En los casos de acontecimientos poco frecuentes, dependemos de la experiencia de otros que nos informan, pero debemos considerar la fiabilidad de la fuente. Con todo, otros acontecimientos pueden no haberse identificado por lo que realmente eran, y pueden haberse perdido o haberse pasado por alto[2]. No siempre es evidente que un acontecimiento concreto pueda ocurrir y, lamentablemente, en algunas circunstancias puede conllevar consecuencias desastrosas.

Por ejemplo, es posible que un empleado distraído de una fábrica que trabaja en una prensa meta su mano en ella y sufra una amputación. Lo sabemos porque se observó en cierta ocasión. Como podía pasar, pasó. Esto ha aumentado la precaución y la formación obligatoria, pero sólo podrá prevenirse completamente si es *imposible*

que ocurra. Éste es el motivo por el cual se inventaron los dispositivos mecánicos de seguridad. Si es posible que una enfermera confunda una solución de cloruro potásico con otro medicamento, y que lo inyecte a un paciente, causándole un paro cardíaco, esto pasará. Puede que no pase en todos los hospitales, pero mientras existan múltiples lugares donde se repiten acontecimientos en los cuales podría pasar, el número de afectados se incrementará. Por este motivo, el cloruro potásico en forma inyectable no debería tener un acceso fácil.

Cualquiera de los que trabajamos en el campo de la asistencia médica deberíamos guardar un sano respeto por la ley de probabilidades. Las conferencias sobre morbilidad y mortalidad son útiles desde una óptica retrospectiva, pero imposibilitar un suceso antes de que ocurra es el único modo de asegurarse de que no pasará (si algo no puede pasar, no pasará). Intente predecir los posibles acontecimientos adversos que podría producir una acción y estime la probabilidad de que ocurran. Si un acontecimiento terrible puede prevenirse, encuentre la forma más fácil de hacerlo. Hay muchas situaciones donde todos los acontecimientos adversos simplemente no pueden ser eliminados; hay un riesgo inherente en cualquier intervención. Tampoco existe ningún sustituto del sentido común ni de la atenta dedicación a los detalles. Pero el error humano ocurre y conviene evitar el guión que lo propicia en vez de culpar al individuo. Siempre que haya una manera de reducir o eliminar el riesgo sin un esfuerzo excesivo, debería intentarse.

PUNTOS CLAVE

- La ciencia de la bioestadística se desarrolló a lo largo de muchos siglos.
- El desarrollo de la teoría de probabilidades surgió del interés por los juegos de azar.
- La cuantificación del riesgo nos permite tomar decisiones en circunstancias inciertas.
- Los métodos estadísticos modernos son el resultado de las múltiples contribuciones de científicos insignes que aplicaron la teoría de probabilidades a las pruebas de hipótesis.
- El valor *p* es la probabilidad de que un acontecimiento se deba al azar.
- Si algo puede pasar, pasará.

BIBLIOGRAFÍA

1. Bernstein, P. L. 1988. *Against the gods: The remarkable story of risk*. New York: John Wiley & Sons, Inc. Todas las referencias históricas parten de esta obra.
2. Dr. Bob Wolfe, comunicación personal, agosto, 2005.

PREGUNTAS DE REVISIÓN

1. La probabilidad de ciertos acontecimientos puede predecirse con _____ matemáticas.

2. Si conocemos la probabilidad de que un suceso ocurra (como acertar cada número de la lotería), podemos calcular la _____ de acertar todos los números.

3. Aunque tengamos la intuición de que jugar a ciertos números nos reportará más ganancias, probablemente ganaremos lo mismo que si jugamos a otro conjunto de _____.

4. En el caso de no conocer la probabilidad de un acontecimiento, basaremos nuestras predicciones en _____ previas.
5. Podemos afinar el proceso de cuantificar el riesgo incorporando _____ que afectan al suceso.
6. Si quisiéramos predecir la probabilidad de tener un accidente de coche, podríamos calcular nuestro riesgo global basándonos en datos previos. Si también consideráramos la hora, la edad del conductor y las condiciones meteorológicas, podríamos precisar _____ utilizando métodos bayesianos considerando observaciones que incluyan estos factores.
7. El valor *p* de <0,01 significa que la probabilidad de que este acontecimiento ocurra por azar es menor del _____.

RESPUESTAS

1. fórmulas
2. probabilidad
3. números
4. observaciones
5. factores, variables
6. el riesgo, la predicción
7. 1%

CAPÍTULO 8

Distribuciones

Una imagen vale más que mil palabras.

—Anónimo

Los números en un conjunto de datos son los valores de ciertos atributos. Cada individuo de la población tendrá un valor diferente asociado a aquel atributo. Sin embargo, es habitual que los valores tiendan a agruparse. Por ejemplo, en una población de 30 personas, los valores de la presión arterial sistólica variarán pero tenderán a situarse alrededor de los 120 mmHg debido a la naturaleza de los datos biológicos.

Cuando viajamos durante las vacaciones a un lugar lejano, queremos explicar la experiencia a nuestros amigos. Recordamos la topografía del área, la temperatura y los habitantes. Para captar los detalles, tomamos fotografías de las plantas y las flores, del paisaje y de la mirada de los lugareños. Para nuestros amigos, es mucho más fácil hacerse una idea del lugar mirando las fotografías que escuchando una descripción verbal.

Al igual que una colección de fotografías contiene mucha información sobre un lugar, un conjunto de datos contiene mucha información sobre una población. Sin embargo, es difícil ver cómo se distribuye una determinada característica mirando una simple lista de números. Es más sencillo hacerse una idea de la distribución de las características transformando los datos en fotos a través de gráficos.

En la metáfora de las vacaciones, cada tipo de foto representará de una forma más adecuada cada característica concreta del lugar. Por ejemplo, los primeros planos de caras son apropiados para mostrar las facciones de los individuos, mientras que los planos de ángulo abierto son ideales para plasmar la geografía del área. Del mismo modo, distintos gráficos pueden representar la distribución de los valores. El gráfico ilustra el patrón de variabilidad, que se conoce como la *distribución* de la variable.

- *Una distribución es el patrón observado en un conjunto de valores de una variable.*

Hay diversas formas de mostrar el patrón que sigue una variable, según sea el tipo de variable, su escala de medida y el rango de valores. Los siguientes ejemplos incluyen los tipos de distribuciones más habituales e ilustran cómo apreciar la distribución a través de un gráfico. Como veremos, en el proceso de inferencia, las distribuciones son una conexión importante entre los datos y las conclusiones.

TIPOS DE DISTRIBUCIONES

Cuando hablamos sobre los tipos de variables en el capítulo 5, afirmamos que puede asignarse un número a las variables categóricas pero que este número representa una categoría y no un valor real. La tabla 8-1 es un conjunto de datos que contiene información sobre el estado civil (variable categórica) de los norteamericanos de 18 o más años. En este ejemplo se ha elaborado una lista de las categorías del estado civil en vez de adjudicarles un número. Los mismos datos se muestran en un formato distinto en la figura 8-1.

Otra manera de ilustrar la distribución de estas variables es con un diagrama de barras, como el de la figura 8-2. La altura de la barra permite comparar fácilmente las distintas categorías. Fíjese que la ordenada (eje *y*) representa la frecuencia de la categoría. Una barra más alta indica que hay más individuos en los datos que tienen aquel valor. Este gráfico se conoce como *histograma*.

TABLA 8-1 Número y porcentaje del estado civil de los norteamericanos de 18 o más años

Estado civil	*Número (millones)*	*Porcentaje*
Soltero(a)	43,9	22,9
Casado(a)	116,7	60,9
Viudo(a)	13,4	7,0
Divorciado(a)	17,6	9,2

Datos tomados de Moore, D. S. y G. P. McCabe. 1999. *Introduction to the practice of statistics*, 3rd ed. New York: W. H. Freeman and Co., p. 6.

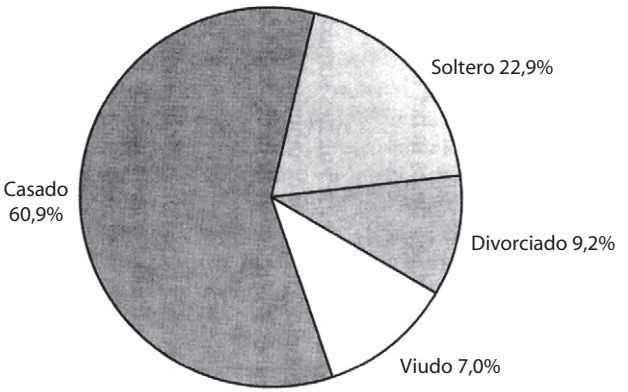


FIGURA 8-1 Diagrama de sectores con los mismos datos que la tabla 8-1. El ojo humano inmediatamente compara el área contenida dentro de las diferentes cuñas. Las categorías tienen que ser mutuamente excluyentes a no ser que haya una cuña dedicada a una combinación de categorías. (De Moore, D. S. y G. P. McCabe. 1999. *Introduction to the practice of statistics*, 3rd ed. New York: W. H. Freeman and Co., p. 7, con autorización.)

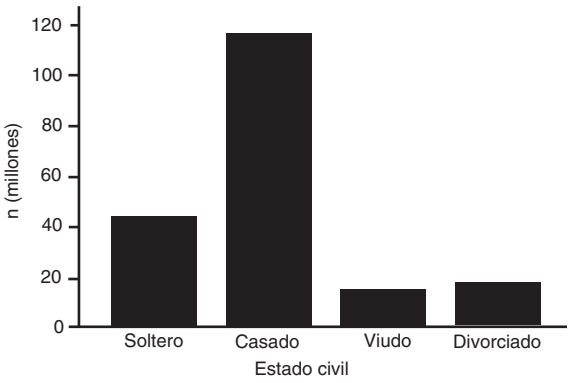


FIGURA 8-2 Diagrama de barras del estado civil de los adultos estadounidenses. (De Moore, D. S. y G. P. McCabe. 1999. *Introduction to the practice of statistics*, 3rd ed. New York: W. H. Freeman and Co., p. 6, con autorización.)

- *Un histograma es un tipo de gráfico de barras en el cual la altura de la barra representa la frecuencia relativa del valor en el conjunto de datos.*

Las variables ordinales también pueden representarse con un diagrama de barras, como el de la figura 8-3. Recuerde que estas variables tienen valores numéricos que indican su posición relativa dentro del grupo y que las diferencias matemáticas entre los valores numéricos no son constantes. El valor observado de la variable se ha sustituido por una posición, de menor a mayor. Es fácil ver dónde tienden a agruparse los valores.

Recuerde que las variables de intervalo tienen valores que se representan en una escala continua. En un gráfico, estas variables pueden visualizarse de manera eficiente poniendo la escala de intervalo en la abscisa (eje *x*), con lo que aumentan los valores de forma continua a lo largo del eje. Como en el diagrama de barras, la altura

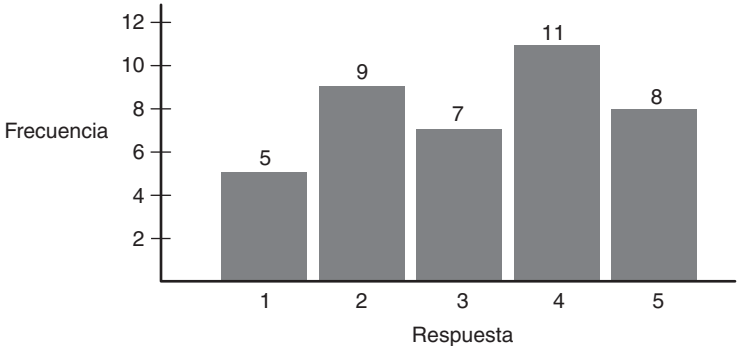


FIGURA 8-3 Resultados de una encuesta a 40 personas en un centro comercial que preguntó: «¿Está de acuerdo con la postura del alcalde respecto al desarrollo del centro de la ciudad?» 1, totalmente en desacuerdo; 2, en desacuerdo; 3, ni de acuerdo ni en desacuerdo; 4, de acuerdo; 5, totalmente de acuerdo.

de la curva representa la frecuencia con la que observamos aquel valor. Este gráfico se denomina *distribución empírica* o *polígono de frecuencias*. Este tipo particular de representación gráfica es un componente crucial de la inferencia estadística. La usaremos para representar probabilidades, que son la base de la inferencia estadística.

- *Una distribución empírica es un gráfico que representa la frecuencia en los datos del valor de una variable.*

Un histograma es un tipo de distribución empírica. La altura de las barras representa la frecuencia relativa de cada valor. Al juntar las barras, surge un patrón. En la figura 8-4, a medida que dividimos los valores de la abscisa en intervalos más pequeños, la altura de las barras se va convirtiendo en una línea continua.

Un *diagrama de tallo* (stem) *y hojas* (plot) (también llamado *stemplot*) es una forma creativa de mostrar no sólo la frecuencia relativa, sino también los valores individuales. Permite comparar una variable entre dos grupos poniendo las distribuciones una al lado de otra. Este tipo de gráfico toma los datos tal cual y los organiza según su valor relativo (de menor a mayor). El tallo contiene los dígitos principales —hasta las centenas o las decenas—. Las unidades se dibujan individualmente al lado del tallo formando las hojas. En la tabla 8-2 se comparó a aquellos estudiantes que asistían a clase regularmente con aquellos que no lo hacían. Se representaron gráficamente las puntuaciones totales de todos los alumnos de un curso de psicología con un diagrama de tallo y hojas. A la derecha están los asistentes regulares y a la izquierda están los que no lo son.

El tallo es el centro del gráfico. El inicio del tallo, 18, representa realmente 180. Las hojas son las unidades. El primer valor del lado izquierdo es un 8, que, junto al 18, representa un único dato de 188 —la puntuación más baja obtenida—. Esta persona resultó ser del grupo de los «novilleros». Hay dos individuos que sacaron 195 puntos que se colocan uno al lado del otro. Fíjese que ambos están, otra vez, en el grupo de los «novilleros». La puntuación más baja en el grupo de los asistentes fue de 241 puntos y la más alta, de 328, que a su vez fue la puntuación más alta global.

Al observar este gráfico, inmediatamente se hacen patentes diversos puntos. Los «novilleros» consiguieron en general una menor puntuación que los estudiantes más aplicados. Mirando las agrupaciones, las puntuaciones de los «novilleros», en promedio, son menores. Al mismo tiempo, la dispersión de las puntuaciones de este grupo es mayor que la de los asistentes. Las puntuaciones de éstos van desde la peor pun-

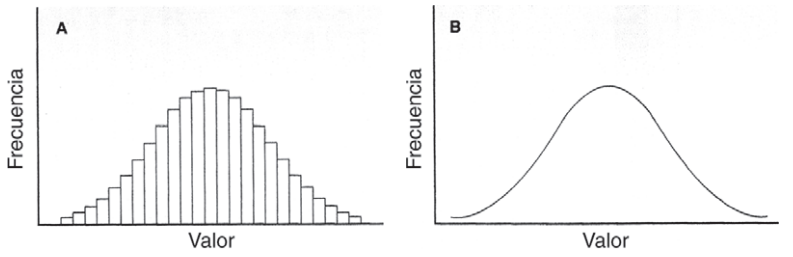


FIGURA 8-4 Ejemplo de una distribución empírica. La frecuencia relativa de cada valor tiene un patrón simétrico, con el valor más frecuente en el centro del gráfico. A menudo se observa esta distribución en la naturaleza. (De Jekel, J. et al. 2001. *Epidemiology, biostatistics, and preventive medicine*, 2nd ed. Philadelphia: Saunders, p. 142.)

TABLA 8-2 Puntuación en la clase de psicología

No asisten a clase	*Tallo*	*Asistencia regular*
8	18	
5 5	19	
	20	
	21	
8 5	22	
9 7 3 2	23	
0	24	1 3 6 9
6 6 6 0	25	0 2 4 4 5 6
8 4 4 1	26	1 2 3 4 4 4 5 7 7
7 4 4 0 0	27	0 1 2 3 6 6 7 8 8
	28	0 1 2 4 8 8
	29	0 1 1 2 3 4 6 6 7 8
8	30	
	31	0
	32	0 1 8

Código |25| 6 = 256

La puntuación total en una clase de psicología agrupada según si la asistencia a las clases fue regular o esporádica.
Datos tomados de Howell, D. C. 1999. *Fundamental statistics for the behavioral sciences*, 4th ed. Pacific Grove, CA: Duxbury Press, p. 37.

tuación hasta la quinta más alta, lo cual indica que algunos eran capaces de sacar buena nota; quizá tenían un método de aprendizaje alternativo. Sin embargo, sólo uno fue capaz de hacerlo muy bien. Estas apreciaciones, obviamente, ayudan a ir a clase, a menos que uno sea de los privilegiados. Obsérvese que el diagrama de tallo y hojas es realmente un histograma, aunque retenga los valores individuales y facilite las comparaciones visuales entre grupos.

No todas las distribuciones son simétricas; de hecho, muchas no lo son. Cuando miramos una distribución empírica, podemos ver un vacío en uno de los lados, como en la figura 8-5. Este rasgo se denomina *asimetría*. Una distribución empírica con asimetría positiva tiene una cola alargada hacia la derecha, y cuando la asimetría es negativa, hacia la izquierda.

El diagrama de caja (en inglés, *boxplot*) es un tipo de gráfico útil para resumir la forma de una distribución asimétrica, ya que usa los percentiles. Los datos se ordenan de mayor a menor y se hacen cuatro particiones, de modo que cada una contenga una cuarta parte de los datos totales. En una distribución asimétrica habrá el mismo número de puntos a ambos lados del percentil 50, pero la anchura será diferente a ambos lados. El percentil 25 y el percentil 75 no estarán a la misma distancia del centro.

- *Los percentiles usan marcas que dividen el número total de observaciones en partes iguales.*

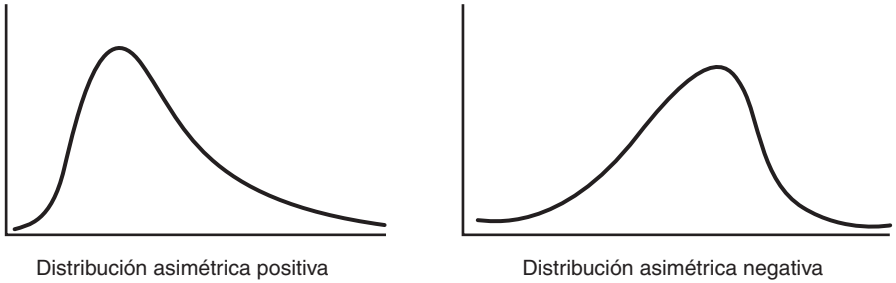


FIGURA 8-5 Ejemplos de distribuciones asimétricas.

En un diagrama de caja, ésta (del percentil 25 al 75) contiene la mitad de las observaciones. La línea horizontal representa la mediana o percentil 50. Las líneas que se proyectan desde la caja (los «bigotes») representan el resto de valores —aquellos que no están próximos al centro—. Una limitación potencial de este tipo de gráfico es la incapacidad de resaltar más de un agrupamiento. Un diagrama de tallo y hojas sería más informativo en estas situaciones. Los valores extremos (*outliers* o datos «raros») se representan con puntos más allá de los bigotes. Los diagramas de caja son ideales para un rápido resumen visual de la distribución global. Una ventaja es su capacidad para realizar comparaciones cuando se colocan uno al lado de otro, como en la figura 8-6.

Los *outliers* son observaciones inesperadamente alejadas del resto. Se apodan «banderas rojas» ya que deberían levantar sospechas. Pueden reflejar un problema en la recogida o en la entrada de los datos. Los *outliers* no deberían desecharse automáticamente, pero sí estudiarse a fondo ya que podrían influir en la validez de los resultados.

Estos gráficos ilustran algunas formas visuales de representar distribuciones. Nos muestran el patrón de dispersión y de agrupamiento, y son una buena herramienta para detectar *outliers*. Las distribuciones ayudan a los estadísticos a decidir qué prueba estadística deberían usar en el análisis final. Además del tipo de variable y la escala de medida, se ha de tener en cuenta el tipo de distribución de la variable.

ESTADÍSTICOS DESCRIPTIVOS NUMÉRICOS DE UNA DISTRIBUCIÓN

En bioestadística usamos fórmulas para comparar grupos. Los valores reales de la variable son necesarios en el análisis, ya que los datos se resumen a través de *estadísticos descriptivos* que describen la forma de la distribución. Necesitamos describir matemáticamente la distribución de una variable. La mayor parte de las variables tienen dos atributos que explican la forma global de una distribución:

1. La tendencia a agruparse. Aparece como un pico en un histograma o como una gran cuña en un diagrama de sectores.
2. La variabilidad. El rango de valores puede ser ancho o estrecho.

Pueden existir unos pocos datos fuera de los límites, que se destaquen visualmente en los extremos de un gráfico. Cada uno de estos atributos pueden describirse mediante números, que se incluirán en las fórmulas.

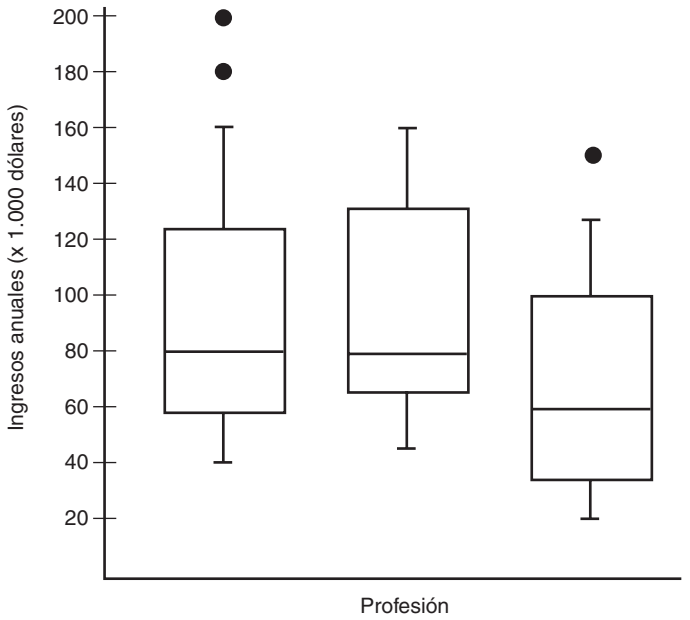


FIGURA 8-6 Diagrama de caja *(bloxplot)* de los ingresos anuales de tres profesiones. Estos gráficos representan los ingresos anuales ×1.000 dólares para universitarios con tres titulaciones diferentes. Las medianas son las líneas de dentro de las cajas. El rango intercuartílico es el mismo para los tres diagramas, pero con menos ingresos para el tercer tipo de formación. El primer y tercer gráfico muestran *outliers* (valores extremos) que representan aquellos que ganan mucho más que los demás.

MEDIDAS DE TENDENCIA CENTRAL

Hay tres medidas de tendencia central para describir la posición de una variable.

La *moda* describe el punto más alto en una distribución empírica. Una distribución con un único pico se denomina unimodal. No obstante, puede haber más de un pico: una distribución trimodal, por ejemplo, tiene tres picos. En una distribución unimodal, la moda es el valor que ocurre con más frecuencia. Esta medida es útil en la descripción de datos nominales.

- *MOda = Más Ocurrente (más frecuente).*

La figura 8-7 muestra la distribución de edades de los casos de tuberculosis en Estados Unidos por año de estudio: 1985, 1988 y 1991. La incidencia de la tuberculosis tiene una distribución de edades trimodal. Afecta a los niños pequeños, pero no tanto como a los adultos jóvenes, que es otro de los picos. Hay una última agrupación en los ancianos. También vemos que en 1991 hubo un aumento en la incidencia de la tuberculosis en la categoría de edad más afectada, las personas de 30 a 40 años.

Muchas variables tienen distribuciones que tienden a agruparse alrededor del valor medio. La mediana es el verdadero punto medio. El número de valores que son más pequeños que este punto es el mismo que el número de valores más altos. La mediana también se llama percentil 50. Este estadístico descriptivo se usa en datos ordinales (¿recuerda la escala ordinal? Las diferencias entre valores adyacentes no son

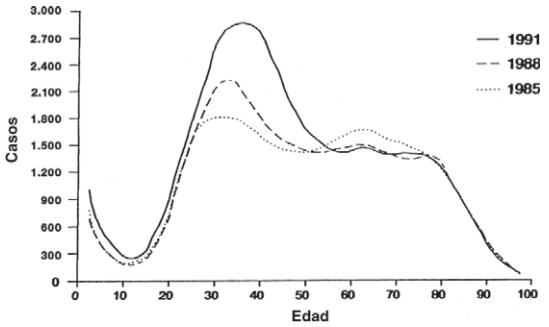


FIGURA 8-7 Distribución por edades de casos de tuberculosis en Estados Unidos en los años de estudio: 1985, 1988 y 1991. (De Kent, J. H. 1993. The epidemiology of multidrug-resistant tuberculosis in the United States. *Medical Clinics of North America*, 77:6, 1393, con autorización.)

constantes). También es más representativa del valor medio en una distribución asimétrica, como la de la figura 8-5. Cuando hay un número impar de observaciones, la mediana es la observación que queda en el medio después de ordenar los datos de forma creciente. Si el número de observaciones es par, la mediana es el promedio de las dos observaciones centrales. La figura 8-3 es un histograma que ilustra una distribución asimétrica donde la mediana no es igual a la moda.

- *MEDIana = MEDIo.*

La *media* de una distribución es el valor promedio observado, representado simbólicamente por $\bar{X}$. Para obtenerla, se suman (Σ) los valores y se divide entre el número de observaciones. La fórmula es:

$$\text{Media muestral} = \frac{\text{suma de todos los valores de la muestra}}{\text{número de observaciones en la muestra}}$$

o, en términos matemáticos,

$$\bar{X} = \frac{\sum X}{n}$$

donde X representa cada observación y *n* es el número total de observaciones.

- *MEDIA = valor proMEDIAdo de las observaciones.*

La mediana y la media serán similares en distribuciones simétricas. Cuando la distribución es asimétrica, se abren más posibilidades. La cola de la curva tira de la media, como en la figura 8-8. La media es una medida de tendencia central fiable en una distribución simétrica. Con asimetría, sin embargo, la mediana es una medida mejor ya que está menos afectada por los valores extremos u *outliers*.

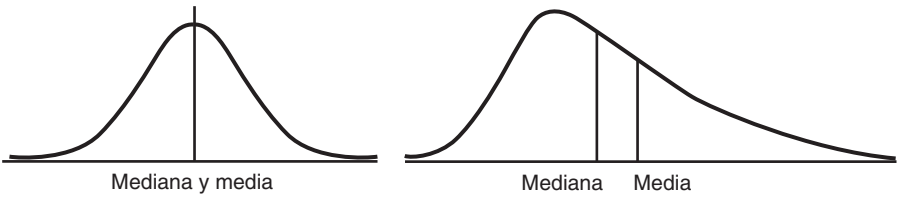


FIGURA 8-8 En una distribución asimétrica, la cola de la curva tira de la media.

MEDIDAS DE DISPERSIÓN

Al igual que hay estadísticos descriptivos para la tendencia central de una distribución, también puede describirse la dispersión o la variabilidad. Es evidente que, para cualquier distribución, habrá muchos valores distintos a la moda, la mediana o la media. Podemos comenzar describiendo el rango de valores, del más bajo al más alto. Algunos valores extremos *(outliers)* pueden tirar del límite del rango hacia un extremo, sin reflejar exactamente la proximidad global de la mayoría de valores.

Un estadístico descriptivo más fiable sería la distancia media de todos los valores al centro. Esto se denomina *desviación estándar*. Se calcula como el promedio de las diferencias entre cada valor y la media.

Debido a que la media generalmente estará próxima a la mediana, al calcular las diferencias entre cada valor y la media, aproximadamente la mitad darán un resultado negativo. ¡Cuando se sumen los resultados de las diferencias, se anularán los unos con los otros y la suma total será 0! Este desaguisado se evita elevando al cuadrado todas las diferencias antes de sumarlas, lo que asegura que sólo sumaremos valores positivos. Después de sumarlos, se divide la suma entre el número total de observaciones menos 1 *(n – 1)*. El motivo por el cual se divide por *n – 1* y no por *n* puede demostrarse con argumentos matemáticos complejos, pero una explicación intuitiva es que es la manera de compensar la menor variabilidad en la muestra respecto a la población total. Esto se denomina *varianza*, s^2, de una distribución. Si calculamos la raíz cuadrada de la varianza, obtenemos la desviación estándar, *s*.

Para aquellos que estén interesados, a continuación se muestran las fórmulas:

$$\text{Desviación estándar muestral} = \sqrt{\frac{\text{suma de (valores de la muestra} - \text{media muestral})^2}{\text{número de observaciones} - 1}}$$

o, en términos matemáticos,

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

La desviación estándar es útil cuando la distribución es simétrica, pero es menos fiable en una distribución asimétrica. En este caso, la variabilidad no es consistente a

ambos lados de la mediana y no hay realmente ninguna desviación «estándar». En este caso es más apropiado usar cuantiles como descriptores de la dispersión, como en el ejemplo del diagrama de caja. Unos pocos *outliers* también pueden tener un gran impacto en el valor de *s*.

La distribución de una variable es uno de los factores determinantes a la hora de decidir el tipo de prueba estadística más apropiada para el análisis de los datos. Es habitual representar gráficamente los datos antes de realizar el análisis para comprobar la simetría y detectar *outliers*.

Por ejemplo, cuando queremos estudiar el efecto de una intervención, tomamos una muestra que sea representativa de la población en estudio. Los estadísticos distinguen entre la media y la desviación estándar de una variable de la muestra (fácil de medir) y la de la población (que normalmente sólo puede estimarse). Recuerde que las medidas muestrales son los *estadísticos* que estiman los *parámetros* poblacionales. Los estadísticos se representan con letras latinas. La media muestral se representa por $\bar{X}$ y la desviación estándar muestral, por *s*. Son estimaciones de los parámetros poblacionales, que se representan con letras griegas. La media poblacional es μ, mientras que la desviación estándar poblacional usa el símbolo sigma (σ).

	Parámetro poblacional	se estima a través del	*Estadístico muestral*
Media	μ (pronunciada [mu])		$\bar{X}$
Desviación estándar	σ (sigma)		*s*

PRUEBAS PARAMÉTRICAS Y NO PARAMÉTRICAS

Ciertos tipos de datos no mostrarán simetría al representarse con una distribución empírica. Además, algunas variables (p. ej., las categóricas) no se miden con escalas de intervalo. Estos tipos de datos a menudo se analizan con pruebas que no asumen premisas sobre la simetría de la distribución de la variable. Estas pruebas se llaman *no paramétricas* o técnicas de *libre distribución*. El gráfico del apéndice A ilustra cómo la distribución de estas variables puede afectar al tipo de prueba estadística. En general, los métodos no paramétricos son más fáciles de aplicar, pero se consideran más rudimentarios que las pruebas paramétricas y son menos sensibles a detectar efectos significativos. Sin embargo, se usan frecuentemente porque muchos tipos de datos no satisfacen la simetría requerida en las pruebas paramétricas.

TIPOS DE BIOESTADÍSTICA

En el capítulo 1 presentamos los conceptos de inferencia y estadística descriptiva. Ahora que hemos ampliado nuestro vocabulario estadístico, podemos mirar en qué se diferencian de un modo más formal. Las distribuciones muestran todos los valores de una sola variable. No realizan ninguna comparación entre variables o grupos distintos dentro de la población. Los gráficos son informativos, pero no podemos sacar ninguna conclusión sobre diferencias o semejanzas entre grupos.

- *Los estadísticos descriptivos describen las variables de un conjunto de datos.*

La estadística descriptiva permite ojear la distribución de una variable. De aquí se determinará la prueba estadística para realizar comparaciones, pero la estadística descriptiva no puede decirnos si una opción es mejor que otra o si existe alguna relación entre variables. Estas preguntas las contesta la inferencia estadística.

- *La inferencia estadística compara los valores de las variables de un conjunto de datos para sacar conclusiones.*

Un reciente artículo de periódico examinaba el efecto en la seguridad de diferentes estrategias para dar el permiso de conducir[1]. Inicialmente otorgaban un permiso de conducir provisional, que, entre otras restricciones, requería el consentimiento paterno. Al periodista le pareció que estos programas funcionaban —en tres de los estados que habían implantado los programas se había reducido la tasa de accidentes en los que estaban implicados conductores noveles—. Sin embargo, 38 estados habían adoptado algún tipo de estrategia. ¿El mejor resultado de los tres estados era debido a su sistema, o podría explicarse únicamente por el azar? Para hallar la respuesta, tenemos que ser capaces de estimar la probabilidad de obtener los mismos resultados en el caso de que los programas sean ineficaces. En los siguientes capítulos averiguaremos cómo hacerlo construyendo un tipo de distribución empírica de las posibles respuestas.

PUNTOS CLAVE

- Una distribución es el patrón observado en un conjunto de valores de una variable.
- Un histograma es un diagrama de barras que representa gráficamente valores observados en la abscisa y frecuencias en la ordenada.
- Una distribución empírica es un tipo de histograma que representa gráficamente los valores de una variable continua en la abscisa y su frecuencia relativa en los datos en la ordenada.
- Las distribuciones empíricas se usan para representar la probabilidad de un acontecimiento aleatorio representando gráficamente el valor de una variable respuesta continua en la abscisa y la frecuencia con la que observaríamos esta respuesta por azar en la ordenada.
- El conjunto de valores de una variable puede representarse como una distribución visualmente satisfactoria, pero el ordenador necesita estadísticos descriptivos para poder hacer los cálculos necesarios.
- Los estadísticos descriptivos de tendencia central son la media, la mediana y la moda.
- El indicador de dispersión que hemos visto es la desviación estándar.
- Los *outliers* son datos extremos que deberían estudiarse a fondo, ya que pueden afectar al análisis final. Tienen un mayor impacto en la media y en la desviación estándar que en la mediana y en la moda.
- Los cuantiles ordenan los valores de una variable de mayor a menor, y posteriormente los divide en grupos con igual número de observaciones.
- Los cuantiles son sobre todo útiles para variables ordinales o si hay *outliers* válidos (que no se pueden eliminar) en la distribución.
- La media muestral $\bar{X}$ es una estimación del parámetro poblacional μ.
- La desviación estándar muestral s es una estimación del parámetro poblacional σ.

- Las pruebas paramétricas se usan en variables con distribuciones que cumplen ciertos criterios, como simetría y medidas de escala de intervalo; por otra parte, las pruebas no paramétricas pueden usarse, a pesar de que, en general, son menos sensibles a la hora de detectar efectos significativos.

BIBLIOGRAFÍA

1. Durbin, D. 2003. Graduated licensing is working. Flint, MI: *The Flint Journal*, February 18.

PREGUNTAS DE REVISIÓN

1. La _____ de una variable es una representación visual de todos los _____ de aquella variable.

2. Muchos tipos de variables tienen valores que tienden a agruparse. Las medidas de tendencia central son la _____, la _____ y la _____.

3. Una medida de dispersión es la _____ _____.

4. Un *outlier* tiene más efecto sobre la _____ que sobre la mediana.

5. La media y la desviación estándar de la muestra son _____ de los parámetros poblacionales.

Verdadero o falso:

6. En cualquier distribución siempre habrá un valor igual a la moda.

7. En cualquier distribución siempre habrá un valor igual a la media.

8. En cualquier distribución siempre habrá un valor igual a la mediana.

RESPUESTAS

1. distribución, valores

2. media, mediana, moda

3. desviación estándar

4. media

5. estimadores

Verdadero o falso:

6. verdadero

7. falso

8. verdadero si el número total de observaciones es impar; puede ser falso si el número total de observaciones es par.

CAPÍTULO 9

Distribución normal

El proceso comienza con la curva de campana, cuyo objetivo principal no es dónde apunta sino con qué error.

—Peter L. Bernstein[1]

Hay un tipo de distribución empírica que se presenta en una gran variedad de datos biológicos, sobre todo en variables que siguen una escala continua o de intervalo. También es una de las distribuciones más habituales en la estadística porque retrata la frecuencia relativa con la que podría ocurrir un resultado particular en muchos tipos de situaciones. Por esta razón, se usa como prototipo para demostrar cómo funciona el proceso de inferencia estadística.

La distribución normal la presentó inicialmente el matemático francés Abraham de Moivre (1667-1754). Se percató de que, para usar muestras que representasen a poblaciones mucho más grandes, debía emerger un patrón de los datos, como Jacob Bernoulli había sugerido. Demostró que los valores de una variable en una muestra se distribuían alrededor de un valor promedio. Cuando representaba los valores de una distribución empírica, el resultado era una distribución unimodal y simétrica. Los puntos se repartían equitativamente a ambos lados de la media. Los extremos disminuían aproximándose a la abscisa porque los valores más distanciados de la media eran menos probables. A esta distribución se le llamó distribución *normal*, ya que de Moivre atribuyó este patrón al orden del mundo natural. Como el célebre matemático Carl Friedrich Gauss disertó sobre ella, razón de que se conozca por su nombre y suela tener una forma elegante de campana, se denomina *campana de Gauss* (fig. 9-1).

ATRIBUTOS DE LA DISTRIBUCIÓN NORMAL

Hay una ecuación matemática para la distribución normal que describe la relación entre los puntos de la abscisa y la ordenada. Un conjunto de valores con una distribución normal tendrá una media μ y una desviación estándar σ. Estos parámetros determinan el pico y la dispersión de la curva. Hay muchas campanas de Gauss diferentes para diferentes valores de μ y σ. No obstante, para una curva concreta, la altura de la curva y para cualquier punto dado depende del valor de x. El cuadro 9-1 muestra la fórmula de las distribuciones normales, aunque no es necesario retenerla.

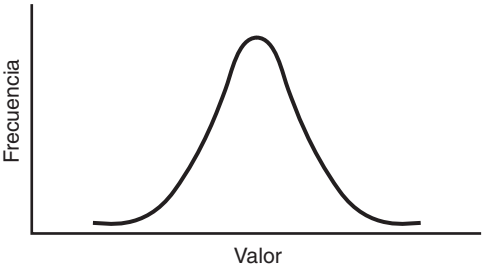


FIGURA 9-1 La distribución normal tiene una forma de campana característica.

CUADRO 9-1

La fórmula de la distribución normal es:

$$y = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$$

Todos los valores del lado derecho de la ecuación son conocidos excepto x*, ya que se deducen de los datos (*π *es una constante). De este modo, para cualquier valor de* x *a lo largo de la abscisa podemos calcular la frecuencia de que ocurra.*

En una distribución normal, la moda es el pico de la montaña. Es el valor que ocurre más frecuentemente. Coincide con la media (μ) o promedio de todos los valores de la distribución. Además, debido a que los puntos se distribuyen regularmente a ambos lados, el mismo punto representa la mediana, que es el percentil 50. Si cambia μ pero σ permanece constante, la curva se moverá a lo largo de la abscisa pero la forma no variará, como en la figura 9-2.

La figura 9-3 muestra lo que le ocurre a la forma de la curva para distintos valores de σ. Cuando la desviación estándar es más pequeña, la distancia promedio a la media es menor y habrá más puntos con un valor cercano a la media. Esto produce una curva más estrecha con un pico más alto. La media permanece intacta, pero la curva se estira hacia arriba.

En realidad, la distribución normal es una colección de curvas con forma de campana, que dependen de μ y σ. Nosotros usaremos la curva como una distribución

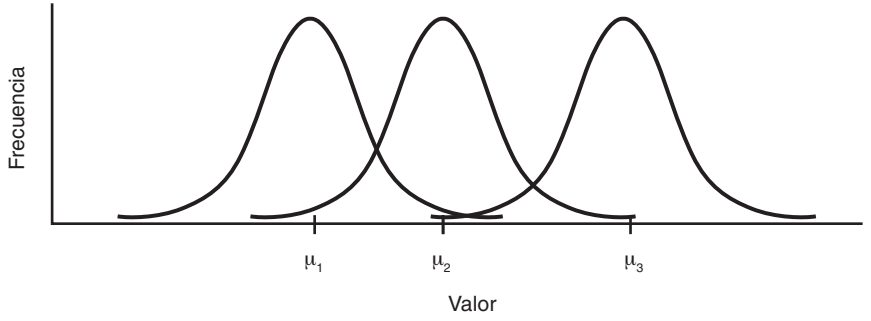


FIGURA 9-2 Efecto de diferentes valores de μ con la misma σ.

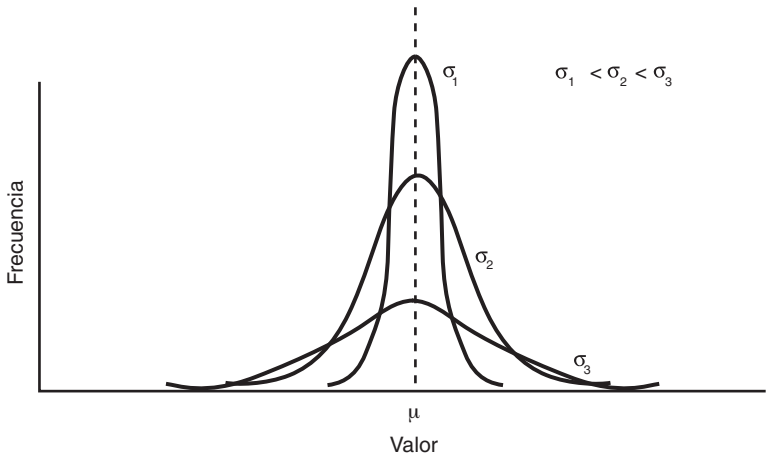


FIGURA 9-3 En una distribución normal, moda = media = mediana = percentil 50 = μ. Si la desviación estándar σ es pequeña, la curva es alta y estrecha. Cuando σ se hace más grande, la curva se aplana y se ensancha.

empírica donde pondremos los valores en la abscisa para ver la probabilidad de su ocurrencia en la ordenada. Sería interesante que existiera alguna manera de estandarizar todas estas curvas para poder fijarnos en sólo una. Lo lograremos realizando algunas simples operaciones matemáticas en cada uno de los datos. Esto es posible gracias a que la familia de distribuciones normales se caracterizan por la manera en que los datos se distribuyen bajo ciertas porciones de la curva.

Una de las peculiaridades de la distribución normal tiene que ver con su simetría. La mitad de los datos se encuentran a la izquierda de μ y la mitad restante está a la derecha, sin que importe su dispersión. Cuando representamos todas las observaciones, vemos que el 68% de éstas caen dentro de una desviación estándar (σ) desde la media. Cuando nos vamos a dos desviaciones estándar a ambos lados de la media (concretamente, a 1,96 desviaciones estándar, aunque a menudo se redondea a 2), aproximadamente el 95% de todas las observaciones estarán contenidas en esta área. La misma distancia (o desviación) desde la media contendrá aproximadamente el mismo número de datos a ambos lados. De hecho, un método para comprobar la normalidad es calcular los percentiles del conjunto de valores de una variable y ver si aproximadamente el 68 y el 95% de los valores caen dentro de una y dos desviaciones estándar, respectivamente.

La figura 9-4 ilustra este principio. La media μ es también la mediana: marca el percentil 50. Hay la mitad de los datos a cada lado de μ. Cuando desplazamos una desviación estándar (σ) a cada lado, nos encontramos los percentiles 16 y 84. Esto significa que el 68% de los datos caen entre estos dos puntos (84 – 16 = 68). Del mismo modo, los puntos para el 95% de los datos son 97,5 y 2,5. Cuando se excluyen las observaciones fuera de estos puntos en las dos colas del gráfico, se conservan el 95% de las observaciones originales.

Esto hay que tenerlo presente, sobre todo si calculamos probabilidades con estos gráficos. Aunque sea correcto decir que el 95% de los datos permanecen por *debajo* del percentil 95, a menudo los cálculos probabilísticos usan las observaciones entre dos puntos situados a la misma distancia en desviaciones estándar de la media. La

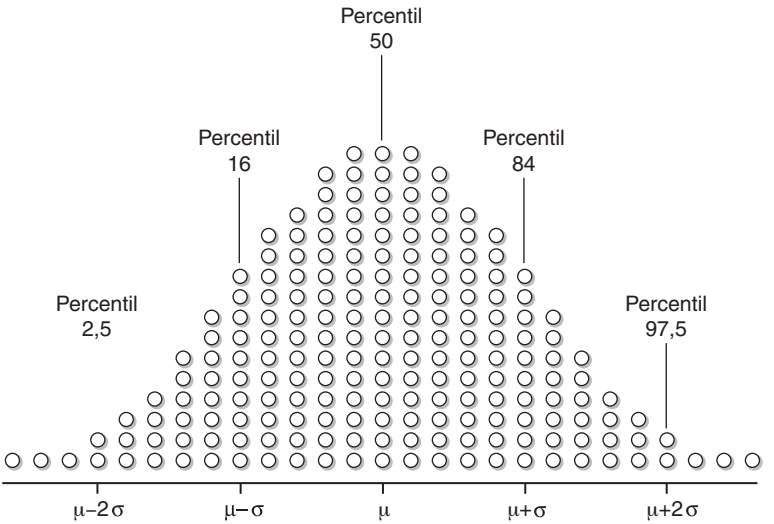


FIGURA 9-4 Cuando la distribución de una variable es normal, el porcentaje de observaciones que caen dentro de una o dos desviaciones estándar siempre es el mismo, sea cual sea el valor de σ o μ. (De Glantz, S. A. 1992. *Primer of biostatistics*, 3rd ed. New York: McGraw-Hill, p. 20, con autorización.)

figura 9-5 muestra que el área bajo cualquier distribución normal que comprenda el 95% de los datos es ±1,96 desviaciones estándar. Esto equivale a los percentiles 2,5 y 97,5. Como veremos, éste es el principio básico que se esconde tras las pruebas unilaterales y bilaterales.

DISTRIBUCIÓN NORMAL ESTANDARIZADA

Todas las distribuciones normales tienen la misma forma, aunque pueda variar la altura del pico y la distancia entre percentiles pueda diferir de una distribución

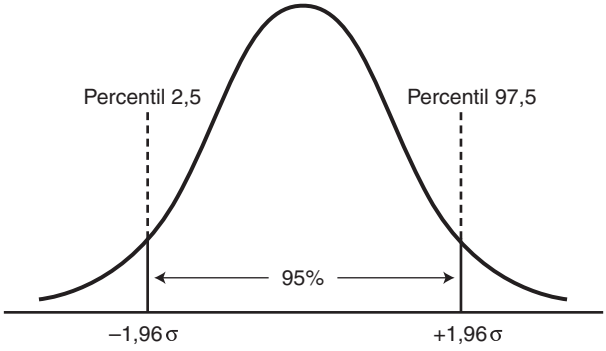


FIGURA 9-5 El área entre los percentiles 2,5 y 97,5 contiene el 95% de los puntos en cualquier distribución normal. Estos puntos pueden encontrarse entre ± 1,96 σ.

normal a otra. Sin embargo, los percentiles 68 y 95 (que contienen aproximadamente el 68 y el 95% del total de las observaciones, respectivamente) estarán situados sobre los valores correspondientes a la primera y segunda desviación estándar. Esta propiedad permite la *estandarización* de cualquier distribución normal. Podemos definir la distancia a lo largo de la abscisa en términos de desviaciones estándar a la media en vez de con el valor real de la observación. Esto condensa todas las distribuciones normales en una, gracias a una manipulación matemática de los datos.

Cada observación se convierte en un valor estandarizado simbolizado por *Z*:

$$Z = \frac{\chi - \mu}{\sigma}$$

El nuevo valor, *Z*, equivale al número de desviaciones estándar que estaba distanciado del valor original de la media. Cuando se representan los nuevos valores, obtenemos la distribución normal estandarizada, como la de la figura 9-6. Cualquier distribución normal puede estandarizarse.

- *En cualquier distribución normal, la escala a lo largo del eje x puede transformarse en desviaciones estándar lejos de la media, sin cambiar la forma relativa de la curva. Esto se denomina estandarización o tipificación.*

La distribución normal estandarizada tiene una media de 0 y una desviación estándar de 1. Esta distribución normal estandarizada pone las observaciones en

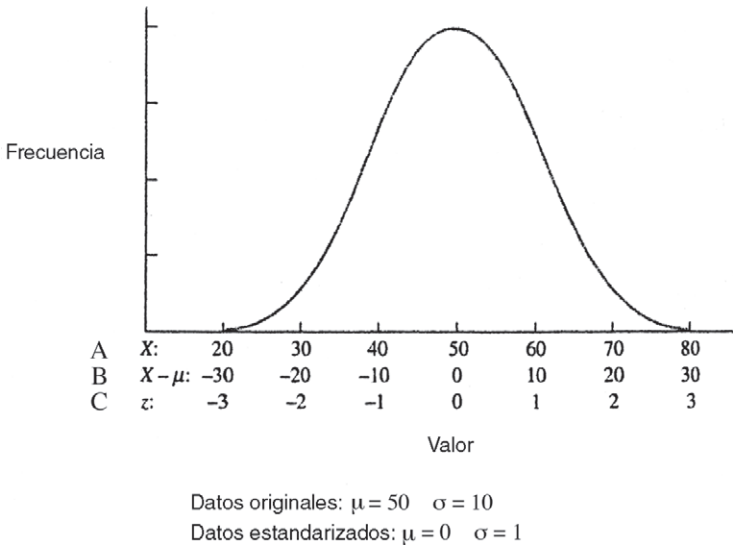


FIGURA 9-6 En la distribución normal estandarizada, μ = 0. Cada valor de *x* se estandariza a la media (μ) restando μ de *x*, como en la línea B. En la distribución normal estandarizada, σ = 1. Entonces, cada valor se acaba de estandarizar dividiendo entre σ, como en la línea C. Ahora cada observación está totalmente estandarizada y tiene un nuevo valor *Z*. (Adaptado de Howell, D. C. 1999. *Fundamental statistics for the behavioral sciences*, 4th ed. Pacific Grove, CA: Brooks/Cole Publishing Co., p. 93, con autorización.)

términos de distancia a la media o desviaciones estándar. Una vez sabemos la media y la desviación estándar de una distribución normal, podemos calcular el valor *Z* para cualquier valor *x*.

- *El valor Z es el número de desviaciones estándar por encima o por debajo de la media en una distribución normal estandarizada o tipificada.*

Z = 1 significa que se está a una desviación estándar a la derecha de la media. Si *Z* = –2,5, se está a 2,5 desviaciones estándar a la izquierda de la media. La media, por definición, no tiene desviación, y por lo tanto su valor *Z* es 0. La manera en que los datos se distribuyen es la misma para cualquier distribución normal, y ahora los percentiles pueden expresarse en términos de *Z*.

CÓMO USAR EL VALOR *Z*

Ya estamos preparados para hacer un simple problema con datos descriptivos que ilustran una aplicación de la distribución normal. Gracias a la recogida de datos durante muchos años, sabemos que el cociente intelectual (CI) se mide en una escala continua de media 100 y desviación estándar 15, y que se distribuye normalmente. Si usted sabe el CI de un individuo, puede convertirlo en un valor *Z* y encontrar el correspondiente percentil para aquel individuo consultando una tabla de la normal estandarizada. Para un CI de 105, la conversión sería:

$$Z = \frac{x - \mu}{\sigma} = \frac{105 - 100}{15} = \frac{5}{15} = 0{,}33$$

Esta puntuación individual del test está 0,33 desviaciones estándar por encima de la media. (Un valor *Z* negativo estaría por debajo de la media.) Para convertir a percentiles, tenemos que consultar una tabla que nos dé las áreas relativas de la distribución normal estandarizada. El apéndice C contiene esta tabla abreviada. Sólo muestra los valores *Z* positivos —los situados a la derecha de la media—. La primera columna representa los valores del primer decimal de *Z* (en este caso, 0,3). Siga esta fila hasta que encuentre el área para *Z* = 0,33. El área es 0,1293. Para encontrar el área total, tenemos que añadir el 0,5 del área que queda a la izquierda de la media. Se obtiene un total de 0,6293, que podemos redondear a 0,63. El percentil de una persona que sacó 105 en la prueba de CI es el 63%. Él o ella obtuvo una puntuación mejor que el 62,9% de todos los que hicieron el test.

Se necesita una pequeña operación para hallar los percentiles de la parte izquierda de la curva, que son los valores menores de 100. Como la curva es simétrica, sólo se tienen que tabular la mitad de los datos; el resto puede deducirse a través de técnicas llamadas de imagen especular. Para los detalles de su realización, consulte la bibliografía.

LA DISTRIBUCIÓN NORMAL COMO PREDICTORA DE ACONTECIMIENTOS

La estandarización se usa para convertir las distribuciones normales que contienen datos en una distribución más manejable para calcular probabilidades. Recuerde que

la distribución normal es una distribución empírica, de modo que la altura de la curva en cualquier punto representa la frecuencia relativa de aquel valor. Como cada dato representa una observación, si escogiéramos uno al azar, podríamos estimar la probabilidad que tuviera un valor dado. Sabemos, por ejemplo, que cualquier valor escogido al azar tiene una probabilidad del 95% de caer dentro de 1,96 desviaciones estándar de la media.

Si está familiarizado con el análisis matemático, sabrá que es posible calcular el área bajo una curva. La distribución normal nos lo facilita. Una vez que estandarizamos, podemos convertir cualquier observación en un valor *Z*. Esta propiedad, una vez que hemos estandarizado, permite encontrar estas áreas consultando una sola tabla. Podríamos usar el análisis matemático para encontrar las áreas con la fórmula de la distribución normal, pero es mucho más fácil estandarizar y usar una tabla publicada.

La tabla de la normal estandarizada es útil para calcular percentiles de individuos cuando se conoce la media y la desviación estándar de la población. Sin embargo, muchos procedimientos estadísticos se centran más en probar diferencias entre grupos que en concentrarse en un único individuo. Por ejemplo, estamos más interesados en una intervención médica que proporciona un beneficio en un individuo *promedio* del grupo antes que en un individuo concreto. Este tipo de razonamiento usa parámetros que se basan en medias muestrales y no en un valor individual. En este proceso veremos cómo la media muestral de una variable usa la distribución normal. También veremos que las distribuciones normales pueden aplicarse a distribuciones asimétricas si tomamos muchas muestras y representamos gráficamente sus medias.

Resulta que la distribución normal tiene muchas aplicaciones directas en la estadística. Hemos visto cómo la distribución normal estandarizada permite que calculemos probabilidades. Muchas pruebas estadísticas usan esta distribución de una manera análoga en las pruebas de significación. Aunque esto pueda parecer endeble, la distribución normal está considerada la columna vertebral de la inferencia estadística. No sólo representa el patrón de muchas variables en la naturaleza, sino que también es el modelo matemático de numerosos cálculos de probabilidad usados en la predicción de respuestas.

La distribución normal se usará en este libro como prototipo para ilustrar los conceptos que hay detrás de la inferencia estadística. Esta elegante curva representa la probabilidad de los acontecimientos, hecho que nos lleva hacia la inferencia.

PUNTOS CLAVE

- La distribución normal es una distribución empírica habitual en la naturaleza.
- Es unimodal y simétrica. Media = mediana = moda = percentil 50. Sus extremos decrecen a medida que se alejan hacia el infinito a ambos lados de la media.
- En todas las distribuciones normales, el 68% de los valores caen dentro de 1 desviación estándar, y el 95% caen dentro de 2 (realmente, 1,96) desviaciones estándar.
- El área bajo la curva normal está directamente relacionada con la probabilidad.
- Si un valor se escogiera al azar, habría una probabilidad del 95% de que cayera entre los percentiles 2,5 y 97,5 (±1,96 desviaciones estándar de la media).

- Si obtuviéramos un valor aleatorio y nos preguntáramos si pertenece a una cierta distribución, y cuando se colocase en la abscisa estuviese por debajo del percentil 2,5, concluiríamos que es muy improbable que el valor proviniese de aquella distribución (aunque esto pasará un 2,5% de las veces).
- La distribución normal estandarizada convierte todos los valores en valores *Z*, que son unidades de desviaciones estándar. La media tendrá un valor *Z* de 0 ya que no se desvía de ella misma. La desviación estándar de la distribución original toma un valor de 1.
- La distribución normal estándar tiene un área bajo su curva igual a 1. Así, cualquier porción del área de la curva tiene un valor entre 0 y 1.
- En vez de hacer operaciones para calcular el área de una distribución normal, estandarizamos de manera que usamos una sola tabla (basada en el valor *Z*) para encontrar el área de una parte de la curva que se traducirá en una probabilidad.

BIBLIOGRAFÍA

1 Bernstein, P. L. 1988. *Against the gods: The remarkable story of risk*. New York: John Wiley & Sons, Inc., p. 141.

PREGUNTAS DE REVISIÓN

1. Muchos tipos de variables en la naturaleza tienen valores que, cuando se representan en una _____ _____, toman la forma de una distribución normal.

2. En una distribución normal, el _____% de los datos permanecen entre los percentiles 2,5 y 97,5. En una distribución normal, el_____% de los datos permanecen entre −1,96 y +1,96 desviaciones estándar.

3. Cuando estandarizamos una distribución normal, ya sabemos que el _____% de los datos permanece entre $Z = +1{,}96$ y −1,96.

4. El área bajo una parte cualquiera de la distribución normal puede hallarse con métodos de _____.

5. Es mucho más simple convertir un valor de una distribución normal en un valor *Z* y consultar una tabla de la distribución normal _____ que contenga el área a ambos lados del mismo.

6. Para un CI equivalente a $Z = 0$, el percentil es _____.

7. El área bajo una distribución normal guarda correlación con la _____ que ocurra un acontecimiento.

8. La distribución normal estandarizada tiene un área igual a _____ bajo su curva.

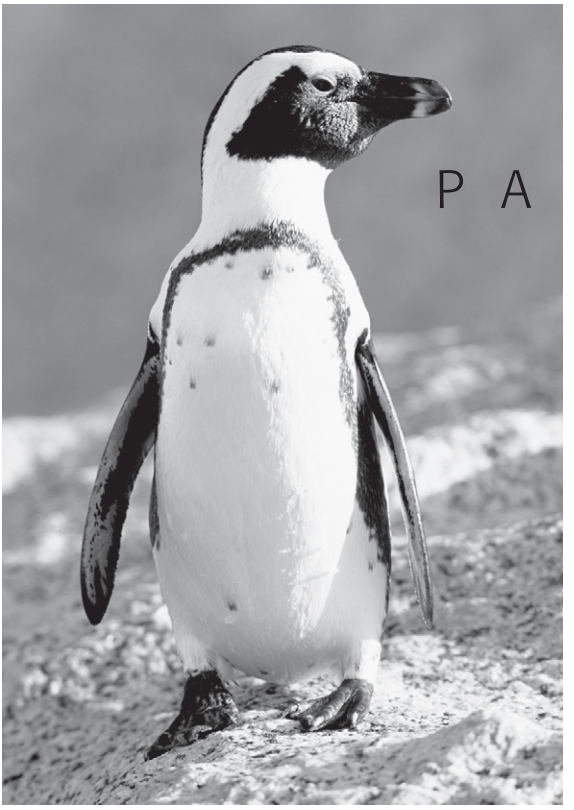
9. Si $Z > 0$, la observación se sitúa a la derecha de la _____ de la distribución normal estandarizada.

10. Si escogiera al azar un valor estandarizado de la distribución normal, el valor esperado sería $Z =$ _____.

11. Para un solo valor escogido al azar de una distribución normal estandarizada, obtener un valor $Z < -3$ sería altamente _____.

RESPUESTAS

1. distribución empírica
2. 95, 95
3. 95
4. análisis matemático
5. estandarizada
6. 50
7. probabilidad
8. 1
9. media, moda o mediana
10. 0
11. improbable



PARTE II

Entender la inferencia

La capacidad de determinar lo que pasará en el futuro y elegir entre diferentes alternativas subyace en el corazón de la sociedad contemporánea.

—Peter L. Bernstein[1]

INTRODUCCIÓN A LA PARTE II

La inferencia estadística puede decirnos, con cierto grado de confianza, si existe una diferencia real entre dos alternativas, o si la diferencia puede ser explicada por el azar. También sirve para determinar la probabilidad de una relación real entre dos o más variables. Nosotros usamos una muestra para representar la población, y sólo una. Sabemos que cada muestra será diferente, pero aceptaremos un grado de incertidumbre a cambio de no tener que observar a cada individuo de la población.

Para ilustrar el grado de confianza requerido para aceptar un resultado como válido, considere el escenario tan elocuentemente presentado por Stanton Glantz en la figura II-1.

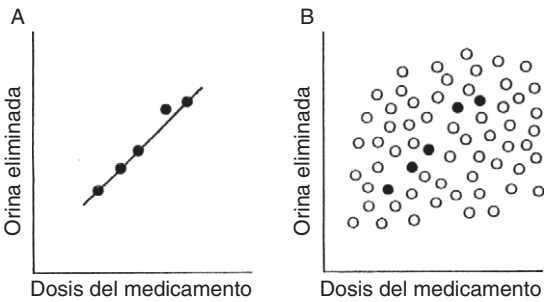


FIGURA II-1 (A) Resultados de un experimento en el cual los investigadores administraron cinco dosis diferentes de un medicamento a cinco personas diferentes y midieron su producción de orina diaria. Ésta aumentaba a medida que se incrementaba la dosis de medicamento, sugiriendo que el medicamento era un diurético eficaz en todas las personas similares a los muestreados. (B) Si los investigadores hubieran podido administrar el medicamento a toda la población y medir su producción de orina diaria, se habría deducido claramente que no existe ninguna relación entre la cantidad de orina y el medicamento. Los cinco individuos seleccionados del estudio del gráfico A se muestran con puntos sombreados. Es posible, pero no probable, obtener una muestra no representativa que lleve a creer que existe una relación entre las dos variables cuando no la hay. Un conjunto de procedimientos estadísticos llamados pruebas de hipótesis permiten estimar la probabilidad de conseguir una muestra no representativa. (De Glantz, S. A. 1992. *Primer of Biostatistics*, 3rd ed. New York, NY: McGraw-Hill, pp. 5-6.)

La moraleja que hay detrás de esta historia es que realmente no hay ningún efecto real entre la dosis del medicamento y la cantidad de orina eliminada. Esto sería evidente si hubiésemos estudiado la población entera. La mala suerte nos llevó a elegir una muestra que nos hizo creer en un efecto diurético real del medicamento. No obstante, el punto a enfatizar es que sería muy improbable que elegiríamos tal muestra aleatoriamente. Podría ocurrir por casualidad, pero es abrumadoramente más probable que nuestra muestra represente con más precisión la ausencia de efecto del medicamento que se observaría en la población.

Las muestras no actúan irracionalmente. Se rigen por modelos de comportamiento previsibles. Pueden salir mal de vez en cuando, pero normalmente seguirán las reglas si se han elegido según las leyes del muestreo aleatorio. Este hecho reconforta. Usted podrá depender de los datos de una muestra para que le proporcionen una conclusión fiable.

BIBLIOGRAFÍA

1. Bernstein P. L. 1998. *Against the gods*. New York: John Wiley & Sons, Inc., p. 7.

CAPÍTULO 10

Pruebas de hipótesis

Anime a los seres humanos, afectados por algo desagradable, a encontrar su causa de manera que puedan evitarla en el futuro.

—Robert Hooke[1]

Los valores de muchas variables en la naturaleza tienen una distribución gaussiana o distribución de campana. La altura de la curva es menor en los extremos, dado que estos valores no ocurren frecuentemente. Si escogiéramos un valor al azar de esta distribución, sería improbable que el valor estuviese en los extremos de la distribución; se hallará con mayor probabilidad en algún sitio próximo al medio.

Asimismo, si ocurre un acontecimiento con múltiples respuestas posibles y la variable respuesta tiene una distribución empírica normal, ciertos resultados ocurrirán con mayor probabilidad que otros. La probabilidad de que un resultado concreto ocurra por azar depende de dónde esté localizado en la distribución normal. Los resultados más probables están en el medio, donde la ordenada tiene valores más altos. Aquellos que ocurrirán con menor probabilidad estarán en los extremos de la distribución. Probablemente, un valor escogido al azar estará en la parte central del gráfico.

Veamos cómo podemos aplicar la lógica probabilística con la distribución normal para decidir cuán verosímil es que una intervención médica tenga un beneficio real. La inferencia estadística nos proporciona los medios para contestar preguntas como: ¿es una opción mejor que otra, en término medio? Más concretamente, podemos contestar estas preguntas:

1. ¿Hay una diferencia significativa en la proporción de una variable entre los grupos?
2. ¿Es el tratamiento o la intervención más beneficioso que otra opción como el cuidado estándar actual?
3. ¿Están dos variables relacionadas?
4. ¿Tiene una variable un efecto sobre otra (variable respuesta)? ¿Cuál es la magnitud del efecto?

Éstas son cuestiones muy representativas de los estudios de investigación. Las preguntas se enfocan a diferentes problemas, pero la forma global de sumergirse para obtener una respuesta es la misma. El tipo de prueba estadística usada es diferente, como se muestra en los flujogramas de los apéndices A y B. La prueba específica

depende del tipo de variable y de la cuestión planteada. Sin embargo, la lógica que se esconde tras las pruebas utiliza el mismo razonamiento estándar. Consideraremos más detalladamente cada una de estas situaciones concentrándonos en el tipo de prueba estadística. Sin embargo, antes de ahondar en las situaciones individuales, primero debemos entender el proceso usado al realizar pruebas estadísticas.

Cuando un investigador diseña un ensayo clínico, sigue un protocolo básico con pasos que son comunes a todos los estudios, aunque no sean del mismo tipo. Incluyen pasos de procedimiento así como pasos logísticos del proceso estadístico que se concentra en las pruebas de hipótesis (cursiva en la lista). Ya hemos hablado de muchos de estos conceptos, y ahora revisaremos la metodología estadística.

- Plantee la hipótesis de investigación. Debe ser una pregunta pertinente que pueda ser contestada.
- Defina la población y las variables del estudio.
- Identifique la variable respuesta y, si es aplicable, la prueba estadística que debe ser usada.
- *Establezca la hipótesis nula.*
- *Declare el nivel de significación.*
- Estime el número de individuos necesarios para lograr una respuesta fiable.
- Obtenga la financiación y la aprobación de un comité ético.
- Obtenga una muestra.
- Recoja los datos.
- *Obtenga el estadístico de la prueba y el método estadístico basándose en el tipo de variables y en el tipo de cuestión planteada.*
- *Aplique la prueba estadística a una distribución muestral.*
- *Obtenga el valor* p.
- *Tome la decisión de aceptar o rechazar la hipótesis nula.*
- Extraiga una conclusión que conteste la hipótesis de investigación.

Hay un número casi infinito de temas que podrían ser potencialmente estudiados. No obstante, no todas las preguntas en una investigación son contestables: muchas son demasiado amplias como para considerarse una única hipótesis de investigación. La mayoría de los principales investigadores elegirán un tema de su ámbito donde sean capaces de plantear preguntas particularmente adecuadas. Tendrán en cuenta las investigaciones previas que informan sobre la evidencia actual de un tema concreto. Los investigadores son capaces de identificar el siguiente paso lógico en el proceso que conduce a un entendimiento útil del conocimiento y son conscientes de las limitaciones de tiempo y coste que determinan la viabilidad de un proyecto de investigación y deben considerar las cuestiones éticas. Muchos estudios no pueden justificarse debido a que podrían interpretarse como un aprovechamiento sobre un cierto grupo de personas, sobre todo los pertenecientes a clases sociales bajas o los que no pueden dar el consentimiento, como un feto.

Entonces, el primer paso es elegir una pregunta contestable (considerando las restricciones). Después se identifica la población del estudio y las variables pertinentes para la hipótesis de investigación. El siguiente paso es enunciar la hipótesis nula, para comprobar si los datos recogidos la respaldan. No podemos llegar a una conclusión a menos que hayamos seguido esta ruta.

HIPÓTESIS NULA E HIPÓTESIS ALTERNATIVA

Todas las pruebas estadísticas comienzan con la premisa de que los datos son consecuencia de fluctuaciones aleatorias (*hipótesis nula* o H_0); es diferente de la *hipótesis de investigación*, para cuya contestación fue diseñado el experimento. La hipótesis nula es el primer paso en el proceso específico del análisis estadístico.

En las dos primeras cuestiones que se plantearon preguntamos si hay alguna diferencia entre dos o más grupos con respecto a la variable del interés. Otro modo de enunciarlo es: «¿Es verosímil que los grupos sean diferentes con respecto a la intervención, o es más probable que la diferencia se deba al azar?». En este último caso, la hipótesis nula afirmaría que el tratamiento no tenía ningún efecto.

Siempre habrá una diferencia cuando comparemos las variables respuesta en los grupos de una muestra, porque es la naturaleza de las variables: varían. Lo que realmente debemos preguntarnos es: «La diferencia de la variable respuesta en la muestra, ¿cae fuera del rango esperado si suponemos que no existe realmente ninguna diferencia en la población?».

Además de pruebas para hallar diferencias entre grupos expuestos o no, podemos usar la bioestadística para demostrar relaciones entre variables. Las preguntas que se plantean en estas circunstancias serían similares a las preguntas 3 y 4. El modo de contestar este tipo de pregunta es cuestionándose lo siguiente: «¿Qué probabilidad hay de que la relación observada se deba únicamente a la variación aleatoria del valor de las variables?». En este caso, la hipótesis nula (H_0) afirmaría que no existe relación alguna entre las variables en la población, y la prueba estadística usada hallaría la probabilidad de que la aparente relación se deba a la aleatoriedad en la muestra.

- *La hipótesis nula, simbolizada como H_0, enuncia que no hay ninguna diferencia o relación entre los grupos.*

En teoría, podemos obtener *cualquier* resultado como consecuencia de la variación aleatoria. Asumiendo que el resultado observado se debe realmente al azar, nos gustaría saber la probabilidad de que ocurra. Si es muy improbable obtener un resultado cuando en realidad no hay ninguna diferencia (o relación) entre los grupos en la población, rechazaremos la hipótesis nula por ser demasiado inverosímil. Los métodos estadísticos para testar la hipótesis nula son las *pruebas de significación*, de las cuales hablaremos más en el siguiente capítulo.

¿Por qué usamos la hipótesis nula? Puede parecer una lógica contradictoria suponer que los resultados no dependen de una intervención y luego tratar de demostrar lo contrario. ¿Por qué no suponer que existe una diferencia e intentar demostrar que es cierto? El razonamiento filosófico que hay detrás de este enfoque lo propuso Fisher a principios del siglo pasado al desarrollar el concepto de las pruebas de significación. Simplemente, su argumento es que resulta más fácil rechazar una afirmación encontrando datos que no la respalden. Sin embargo, para reconocer que algo siempre se cumple, hay que considerar todos los casos en los cuales la afirmación podría ser cierta. Tendríamos que probar las múltiples hipótesis —revisando todas las posibles relaciones numéricas que pudieran existir—, lo cual es una tarea imposible. Por lo tanto, es mucho más productivo enunciar una hipótesis estadística de no diferencia o no relación y luego ver si los datos respaldan o contradicen esta hipótesis. Sin profundizar en los detalles del argumento a favor de la hipótesis nula, es importante saber que es un punto de partida para prácticamente cualquier prueba estadística. La

lógica de la bioestadística aplicada se basa en la hipótesis nula, y está ampliamente aceptada como la base de las pruebas estadísticas.

Es habitual enunciar una hipótesis alternativa, que se acepte si se demuestra que la hipótesis nula es muy improbable. En este caso, rechazaríamos la hipótesis nula de no diferencia (o relación) y aceptaríamos la hipótesis alternativa de que existe realmente una diferencia (o relación entre variables) entre los tratamientos. El símbolo típico para la hipótesis alternativa es H_a. En ocasiones se usa el símbolo H_1. La hipótesis alternativa que especificaremos depende de la cuestión planteada.

- *La hipótesis alternativa, simbolizada con H_a o H_1, se acepta cuando se rechaza la hipótesis nula. Es la hipótesis de existencia de diferencia o relación, y a menudo replica la hipótesis de investigación.*

Si los datos no inducen a rechazar la hipótesis nula, esto será todo lo que podamos decir. No podemos afirmar que la hipótesis nula es cierta, sólo que no se ha rechazado. Similar a un proceso judicial que intenta demostrar la culpabilidad más allá de una duda razonable, los datos se usan para establecer una diferencia de tratamiento real más allá de la duda razonable. La ausencia de culpabilidad no demuestra la inocencia, y la ausencia de evidencia para rechazar la hipótesis nula no demuestra que sea cierta.

ANÁLISIS DE DATOS

El siguiente paso en el proceso de inferencia involucra el tipo específico de prueba estadística usada. Esta ecuación depende del tipo de variable y su distribución. Se entran los datos de la muestra en esta ecuación obteniendo un solo número, que se denomina el *estadístico de la prueba* y tiene una conexión directa con el valor *p*. En el capítulo 11 veremos exactamente cómo se usan los estadísticos de las pruebas para obtener los valores *p*. De momento, es suficiente pensar en el estadístico de la prueba como un solo número que es fruto del análisis estadístico.

EL VALOR *p*

Presentamos el valor *p* como representación de una probabilidad, con un rango que va de 0 (probabilidad nula que ocurra el acontecimiento) a 100 (el acontecimiento ocurrirá siempre). La mayoría de probabilidades caen entre estos valores extremos.

El valor *p* representa la probabilidad de que ocurra un acontecimiento. En las pruebas estadísticas, el valor *p* es la probabilidad de observar el resultado obtenido si no hay ninguna diferencia (o relación) real entre los grupos (es decir, si la hipótesis nula es correcta). Si el valor *p* es muy bajo, significa que sería muy inverosímil obtener el resultado observado si en realidad no existiera ninguna diferencia entre los grupos. En este caso rechazaremos la hipótesis nula de no diferencia, y aceptaremos la hipótesis alternativa de diferencia entre tratamientos.

- *El valor* p *es la probabilidad del resultado observado, si H_0 es cierta.*

El valor *p* puede representarse por el área bajo una curva. Por lo tanto, normalmente no toma un valor exacto, sino que es más correcto denotar una probabilidad mayor o menor que un valor dado. Es más correcto expresar «$p < 0{,}05$» que «$p = 0{,}05$».

NIVEL DE SIGNIFICACIÓN

Una pregunta muy lógica en este punto es: «Si el valor *p* puede tomar valores entre 0 y 1, ¿a partir de qué punto pondremos la frontera y concluiremos que los resultados son demasiado improbables para ocurrir por azar?». O bien: «*¿Cuál es la probabilidad con la que estamos dispuestos a rechazar la hipótesis nula, aun siendo cierta?*». Esta probabilidad se denomina nivel de significación y se simboliza con α.

- *El nivel de significación* α *es el valor* p *a partir del cual estaremos dispuestos a rechazar* H_0 *aun siendo cierta.*

Si realmente decidimos rechazar H_0, corremos el riesgo, mínimo, de equivocarnos. Sin embargo, cuando se usan muestras para estudiar poblaciones, *se espera* que tengan un poco de variación respecto a la población (en el caso de que pudiéramos medir a cada individuo). Se espera también que las muestras difieran entre sí, ya que es la naturaleza del muestreo aleatorio. Debemos decidir cuándo consideramos que los resultados observados son demasiado improbables para respaldar la hipótesis nula. El nivel de significación comúnmente aceptado es $\alpha = 0{,}05$ o el 5%. Esto significa que la probabilidad de observar los resultados obtenidos sin existir realmente ningún efecto del tratamiento (si la hipótesis nula fuera cierta) es menor del 5%. En otras palabras, es posible que nos equivoquemos al rechazar la hipótesis nula, pero no es muy probable ya que esto se producirá, en promedio, en sólo 5 de cada 100 (o en 1 de cada 20) repeticiones del mismo experimento usando muestras diferentes del mismo tamaño.

No hay justificación científica para el nivel de significación del arbitrario 0,05. Es un valor muy cómodo (como una temperatura ambiental de 22 °C es, por consenso, un ambiente confortable) y 0,05 se acepta históricamente como el nivel de significación estándar para rechazar la hipótesis nula. Se acepta que el riesgo de equivocarse 1 de cada 20 veces es suficientemente riguroso como para prevenir grandes errores en el proceso médico. Cuando estudios posteriores apoyan la conclusión del primero, la evidencia a favor de la conclusión original es aplastante. También hay referencias legales sobre márgenes de error aceptables. *Castaneda versus Partida* fue un caso de 1976 de presunta segregación étnica en contra de los mexicanos americanos en la selección de jurados. En la revisión del caso, el Tribunal Supremo aprobó «dos o tres desviaciones estándar» como criterio para la significación estadística[2].

El consenso general es que cuando $p < 0{,}05$, existe una diferencia significativa entre dos alternativas. Entonces, la opción con el mejor resultado se considera superior. Un valor $p < 0{,}01$ afianza aún más una diferencia real debida a una intervención. En general, un valor *p* más pequeño implica una menor probabilidad de que el resultado se deba al azar y una menor probabilidad de error al rechazar H_0.

La distribución normal puede considerarse una distribución de probabilidades que ilustra estos conceptos. Cuando usamos la distribución normal estandarizada, la *Z* se aplica a una tabla que nos da un área directamente proporcional a la probabilidad. Ahora podemos usar una situación análoga para contestar las preguntas que planteamos al principio de este capítulo.

El valor *p* se determina a través de una prueba estadística, que es específica para cada tipo de variable y pregunta. Un valor *p* que cae en los extremos del gráfico es altamente improbable que se deba al azar si la hipótesis nula es cierta. En la figura 10-1 se representa con zonas sombreadas. Que un valor cayera dentro de una de estas áreas podría pasar por azar en menos del 5% de las veces.

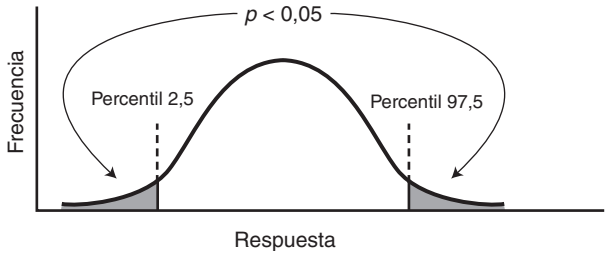


FIGURA 10-1 Los posibles valores de una prueba estadística se representan en una distribución empírica. Si el resultado cae en una de las áreas sombreadas, la probabilidad de obtener este resultado únicamente por azar es baja y se refleja en un valor *p* bajo. La probabilidad conjunta de observar un resultado en los extremos es <5%.

Ésta es una situación donde la probabilidad de un acontecimiento sigue una distribución normal. Si con nuestra prueba estadística obtuviéramos una $p < 0,15$, veríamos que cae en los extremos de la curva, pero no lo suficientemente como para sentirnos cómodos rechazando la hipótesis nula. Concluiríamos que no existe ninguna diferencia significativa en la respuesta entre los dos grupos. No podríamos rechazar la hipótesis nula (que es como aceptar la hipótesis nula de no diferencia de tratamientos) y diríamos que ningún tratamiento mostró un beneficio sobre el otro con un nivel de significación del 0,05.

TIPOS DE ERROR

Si usted es un individuo a quien le gusta tener controlado el objetivo, puede sentirse un poco reacio a dejar al azar el estar equivocado y usar el valor *p* para tomar una decisión. Tenga presente, sin embargo, que la inferencia estadística es lo que es y usa la estadística para inferir una conclusión. La inferencia bioestadística no es cien por cien perfecta, pero la metodología es muy fiable siempre y cuando aceptemos este posible pequeño grado de error. Potencialmente, hay dos maneras de equivocarse en nuestra conclusión de rechazar o aceptar la H_0. Podríamos afirmar que H_0 es falsa cuando obtenemos un valor *p* pequeño, pero podríamos estar equivocados ya que un valor *p* pequeño podría producirse aun cuando H_0 sea cierta. En este caso, rechazaríamos H_0 inadecuadamente. Cuando afirmamos una diferencia que no existe, hacemos un error de tipo I.

Un error de tipo II, por otra parte, es aceptar H_0 cuando realmente es falsa, y H_a cierta. Nos estaremos perdiendo una diferencia real entre los tratamientos. Como veremos, un escenario donde esto podría producirse es cuando el tamaño muestral es demasiado pequeño para descubrir una diferencia significativa, aun cuando exista. Los *cálculos de potencia* se diseñan para minimizar la probabilidad de que esto ocurra, de modo que cuando no se detecte ninguna diferencia significativa y no rechacemos H_0, sea muy improbable que cometamos un error. La tabla 10-1 clasifica los diferentes tipos de errores en la toma de decisión estadística.

PRUEBAS UNILATERALES Y BILATERALES

Muchos estudios observan diferencias entre distintas opciones o intervenciones. Por ejemplo, pueden estudiar el efecto de un nuevo medicamento en pacientes con

TABLA 10-1 Pruebas estadísticas

		Situación real en la población	
Decisión tomada en función de los datos		**H_0 cierta**	**H_a cierta**
	Rechazar H_0	Error tipo I	Sin error
	Aceptar H_0	Sin error	Error tipo II

La interpretación de las pruebas estadísticas implica la aceptación de una pequeña probabilidad de error. Este «error aceptable» se llama error alfa cuando H_0 es cierta y error beta cuando lo es H_a.

cáncer, comparándolo con la anterior medicación, que es el tratamiento estándar. Los investigadores pueden tener una idea de cómo serán los resultados o qué opción podría tener un mayor beneficio. Si sólo preguntan si una terapia específica es mejor que otra, la hipótesis de investigación podría preguntar: «¿Tiene el nuevo medicamento un beneficio añadido respecto al tratamiento estándar?».

En la mayoría de circunstancias, no obstante, es completamente factible que el tratamiento estándar pueda resultar mejor. Si esto no se considera, entonces los investigadores no contemplarán esta posibilidad, e incluso si el medicamento habitual es mejor, la prueba mostrará que no hay beneficio alguno en ningún sentido. Ellos (y, consecuentemente, nosotros) dejarían pasar este conocimiento potencialmente importante. Esto se denomina una *prueba unilateral* o *prueba de una sola cola*. Se rechaza H_0 en el caso de que el resultado caiga en sólo una de las colas de la figura 10-2.

- *Una prueba unilateral, o de una sola cola, considera una diferencia de tratamiento en sólo un sentido, incluso si lo contrario es cierto.*

Aunque los investigadores puedan sospechar que el nuevo medicamento es mejor, la probabilidad del resultado inverso a favor del medicamento estándar, que es el actual número 1, debería considerarse. De hecho, ¿no es éste el motivo principal por el cual se está realizando el estudio? Un enfoque imparcial que considere una u otra posibilidad se denomina *equipoise*. La prueba estadística que refleja *equipoise* se llama *prueba bilateral* o *prueba de dos colas*. H_0 se rechazará si el resultado cae en cualquiera de las dos colas, como en la figura 10-3.

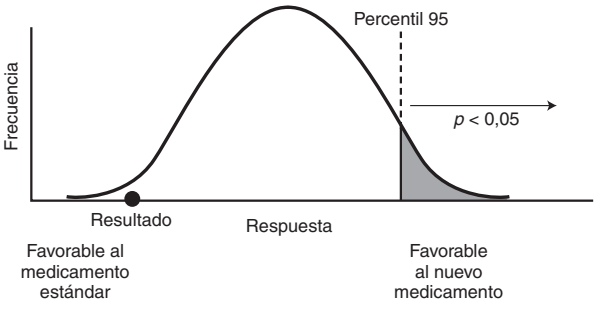


FIGURA 10-2 En una prueba para hallar diferencias entre grupos, si una opción es mejor, el resultado caerá en una de las colas del gráfico. En este caso, el medicamento estándar es mejor, pero este hecho pasó desapercibido porque la prueba se construyó para recoger una diferencia sólo en el caso de que el nuevo medicamento fuese mejor. Se usó un nivel de significación del 5%.

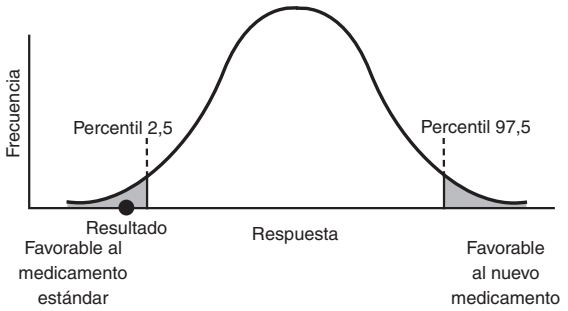


FIGURA 10-3 En una prueba bilateral, H_0 será rechazada si el resultado cae en alguno de los extremos. El error permitido, o error alfa, debe repartirse entre las dos colas.

- *Una prueba bilateral, o prueba de dos colas, considera una diferencia de tratamiento en uno u otro sentido.*

Rechazamos H_0 si parece demasiado improbable. Si nos acogemos al nivel de significación del 5% en una prueba unilateral, podemos aplicar todo el error del 5% a uno de los extremos del gráfico. Esto implica que las condiciones para rechazar H_0 y encontrar una diferencia entre tratamientos son menos exigentes que en una prueba bilateral. La figura 10-4 ilustra cómo una prueba unilateral podría llegar a una conclusión completamente opuesta a una prueba bilateral con los mismos datos.

Muchos expertos estadísticos estarían de acuerdo en que las reglas de *equipoise* deberían seguirse y, siempre que fuera posible, debería realizarse una prueba bilateral antes que una prueba unilateral. Sin embargo, no sería extraño encontrar argumentos a favor de las pruebas unilaterales en ciertas circunstancias, sobre todo cuando la investigación previa ha apoyado resultados en un sentido. No hay ningún enfoque correcto o incorrecto, pero las pruebas unilaterales son más clementes en su capacidad de interpretar una diferencia entre tratamientos porque aplican toda «la incertidumbre aceptable» a una circunstancia que favorece una respuesta concreta sin contemplar la posibilidad contraria. Tenga en cuenta estos puntos de vista cuando haga una valoración crítica de la bibliografía.

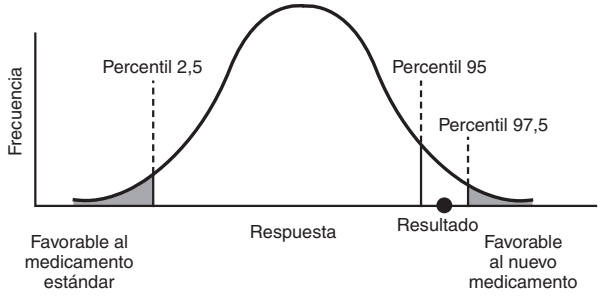


FIGURA 10-4 Los datos produjeron un punto que cayó entre las probabilidades del 5 y el 2,5%. Una prueba bilateral sería favorable a la H_0 y afirmaría que no existe ningún beneficio del nuevo medicamento. Una prueba unilateral se interpretaría de manera completamente diferente; se rechazaría H_0 y el nuevo medicamento se consideraría significativamente mejor.

SIGNIFICACIÓN ESTADÍSTICA FRENTE A CLÍNICA

Es importante observar que la significación *estadística* no siempre puede traducirse en significación *clínica*. Gracias a la búsqueda de datos es posible encontrar significación estadística a muchos niveles, sobre todo en grandes bases de datos con numerosas variables. Los grandes tamaños muestrales tenderán a recoger diferencias estadísticamente significativas entre variables, aunque la diferencia sea diminuta. Que se demuestre que una opción es estadísticamente beneficiosa no implica que sea la mejor decisión. Siempre considere lo que está siendo comparado, así como el coste de tratamiento, los efectos secundarios potenciales y el beneficio global en la población objeto de estudio. Si el beneficio clínico es realmente pequeño, puede no estar justificado el gasto y el esfuerzo de poner en práctica el cambio.

REGLA NEMOTÉCNICA

Muchos de nosotros que no participamos con regularidad en la investigación podemos no acordarnos de la definición exacta del valor *p* cuando la necesitamos. Es comprensible, ya que el concepto es algo complejo. Aquí presentamos un modo fácil de recordar exactamente lo que representa el valor *p*:

- El valor *p* es una probabilidad. Esto es fácil, ya que es lo que significa la *p*.
- Pero ¿probabilidad de qué?
- Ahora que usted ha llegado tan lejos en bioestadística, ha de PREVALecer en su mente la definición de valor *p*: PRobabilidad de un EVento ALeatorio.

Con esto lo tiene casi todo, pero no es suficiente. Ya que le he «dado» esta frase nemotécnica, le daré más detalles:

- Un valor *p* es la PRobabilidad de EVento ALeatorio, «dado» que la hipótesis nula es cierta.

Ahora PREVALece en su mente la definición «dada».

PUNTOS CLAVE

- La inferencia estadística puede contestar a preguntas sobre relaciones entre variables o diferencias entre grupos.
- Los estudios científicos comienzan planteando una hipótesis de investigación que debería ser una pregunta apropiada y contestable.
- Una de las partes habituales en el proceso de inferencia son las pruebas de hipótesis.
- La prueba de hipótesis empieza con el planteamiento de la hipótesis nula, H_0, de no relación o no diferencia.
- La prueba estadística realizada depende del tipo de cuestión planteada y de los tipos de variables estudiadas.
- El resultado de la prueba estadística determina el valor *p*.
- El valor *p* es la probabilidad de un acontecimiento aleatorio, dado que la hipótesis nula es cierta.
- Un valor *p* pequeño supone una probabilidad pequeña de acontecimiento aleatorio, y rechazamos la hipótesis nula de no relación o no diferencia por ser dema-

siado inverosímil, y afirmamos que existe una diferencia real entre los grupos (como cuando funciona una intervención) o una relación entre las variables.
- La hipótesis alternativa, H_a, se acepta si H_0 se rechaza. En general, repite la hipótesis de investigación.
- El nivel de significación es el valor p a partir del cual rechazaremos H_0 aun siendo cierta. Se acostumbra a usar un nivel de significación del 5%, que equivale a $p < 0{,}05$.
- Podemos equivocarnos al rechazar H_0. Esto es un error de tipo I.
- Si no rechazamos H_0 cuando realmente existe una diferencia o relación, cometemos un error de tipo II. Los cálculos de potencia intentan asegurar que este tipo de error no sea consecuencia de un tamaño muestral inadecuado.
- Una prueba unilateral es una prueba de una sola cola que busca el beneficio de un tratamiento en un solo sentido.
- *Equipoise* es una actitud que no realiza ninguna asunción sobre los resultados antes de obtenerlos.

BIBLIOGRAFÍA

1. Hooke, R. 1983. *How to tell the liars from the statisticians*. New York: Marcel Dekker, Inc., p. 137.
2. Moore D. S. y G. P. McCabe. 1999. *Introduction to the practice of statistics*, 3rd ed. New York: W. H. Freeman and Co., p. 619.

PREGUNTAS DE REVISIÓN

1. La _____ _____ supone que no hay ningún beneficio del tratamiento o relación entre las variables.

2. Si el resultado de la prueba es demasiado incompatible con la hipótesis nula, la _____ y asumiremos que la hipótesis _____ es correcta.

3. La prueba estadística desemboca en un valor *p*, que es la probabilidad de un _____ _____, suponiendo correcta la hipótesis nula.

4. Una prueba _____ considerará un beneficio en cualquier sentido.

RESPUESTAS

1. hipótesis nula

2. rechazaremos, alternativa

3. acontecimiento aleatorio

4. bilateral / de dos colas

CAPÍTULO 11

Conexión con la probabilidad

Las estadísticas son la historia, y las reunimos para poder aprender cómo nuestras probabilidades de éxito dependen de nuestras acciones.

—Robert Hooke[1]

Ahora sabemos que *p* representa la probabilidad de un acontecimiento aleatorio. También sabemos que la interpretación de los datos, en todas las pruebas estadísticas, se basa en el valor *p* y en la probabilidad de obtener el resultado observado si la hipótesis nula de no efecto es cierta. Sabemos que la probabilidad puede representarse con una distribución empírica. Ahora estamos preparados para ver cómo se obtienen los valores *p*. El razonamiento es el mismo aunque se usen diferentes pruebas estadísticas.

Se dispone de varios métodos estadísticos con distintas fórmulas para obtener un valor *p*. Debido a que éste no es un libro que incluya toda la estadística, no es lo suficientemente completo para abordar toda la metodología. Sin embargo, al final de este capítulo estará familiarizado con algunos de los métodos estadísticos más habituales.

DISTRIBUCIONES MUESTRALES

Una de las ventajas de trabajar con muestras es que el investigador no tiene que obtener la respuesta para cada individuo de la población. Una muestra, cuando es aleatoria, representa la población. La muestra puede estudiarse y permite extraer conclusiones sobre la población.

Centrémonos en el grupo que constituye la muestra. No estamos tan interesados en la respuesta individual como en la del grupo. Por supuesto, los valores individuales se contabilizan en el grupo, pero para comparar los resultados se mira la respuesta global. Las observaciones de los grupos se usan para estimar un parámetro. Como ya vimos, los parámetros son los promedios de un conjunto de valores, como la edad media o su desviación estándar (que realmente es la «desviación promedio»).

Para ver cómo se hace, veamos primero una situación hipotética. Considere lo que pasaría si estudiáramos una variable poblacional con distribución normal. Si tomá-

ramos múltiples muestras de esta población, cada una de ellas tendría una media y una desviación estándar ligeramente diferente. Si pudiéramos representar todas las medias muestrales ($\overline{X}$s), tenderían a agruparse alrededor de la media poblacional, μ. Muchas incluso coincidirían con ella.

Por supuesto, el problema consiste en que no sabemos *con certeza* cómo nos aproximamos observando sólo una muestra. La media de cualquier muestra dada ($\overline{X}$) podría estar a cualquier lado de μ y a una distancia diferente de μ.

La figura 11-1 ilustra lo que podría pasar cuando tomamos una muestra. Ésta es una forma muy brillante de ilustrar el concepto de distribuciones muestrales que aparece en el *Primer of Biostatistics* de Stanton Glantz[2]. En este caso, la población son marcianos. La figura 11-1 A muestra la distribución empírica de las alturas de *cada miembro* de la población, suponiendo que se pudieran medir. Pero como realmente no podemos medirlas, fíjese que se tomaron varias *muestras* representadas en los gráficos B, C y D. Los puntos pequeños son las medias de cada muestra, que están próximos pero no llegan a coincidir. Las barras a ambos lados de las medias representan las desviaciones estándar de las muestras.

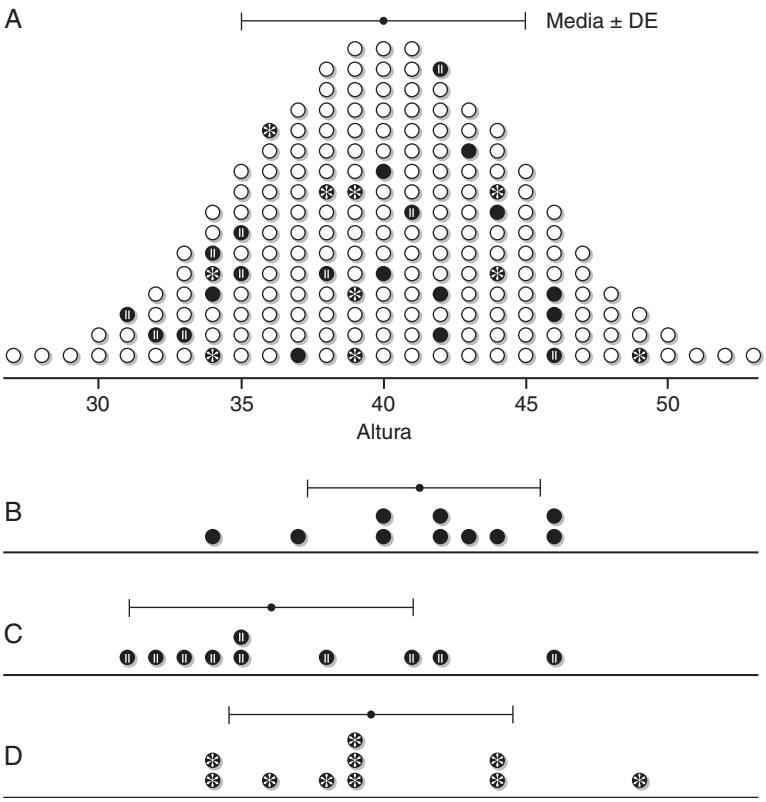


FIGURA 11-1 Si se representan tres muestras diferentes de 10 individuos de una única población, se obtendrán tres estimaciones diferentes de la media y de la desviación estándar poblacionales. (De Glantz, S. A. 1992. *Primer of Biostatistics*, 3rd ed. New York: McGraw-Hill, p. 23, con autorización.)

Ahora observe el gráfico A. Las leyes del muestreo aleatorio nos dicen que cada una de estas muestras tiene la misma probabilidad de ser escogida. Nos proporcionan estimadores bastante buenos del parámetro de altura media. Es muy improbable escoger una muestra constituida únicamente por miembros del final de la curva. En este caso, la estimación del parámetro no sería muy buena.

La inferencia estadística no se concentra en «¿cuál es el parámetro verdadero?», sino en «¿con qué probabilidad estaremos a una cierta distancia del verdadero parámetro?». Lo que realmente tenemos que saber es el grado de variabilidad entre las muestras debida a la aleatoriedad y la probabilidad de obtener una muestra extrema. El método que usemos depende de la *distribución muestral del estadístico* de la prueba. Es la base de toda la teoría sobre inferencia.

- *La distribución muestral de un estadístico es la distribución de todos sus valores para cada muestra de un tamaño concreto de la misma población.*

En este caso, el estadístico tiene un significado especial —es el resultado del análisis usado para estimar el parámetro—. La distribución muestral es un concepto abstracto. La distribución del estadístico es el resultado de lo que pasaría si se estudiasen todas las muestras de un tamaño concreto. Aunque estudiemos una sola muestra aleatoria, es sólo una de un número increíblemente grande de muestras posibles, cada una de ellas con su propio valor del estadístico.

El resto del capítulo trata sobre cómo se generan las distribuciones muestrales para diferentes tipos de estadísticos. Éste no es un paso real en el proceso de inferencia. No creamos esta distribución porque sólo tenemos una muestra para trabajar. Los estadísticos observan el tamaño de la muestra y el tipo y la variabilidad de los datos para ver qué distribución usar.

En el ejemplo citado hablamos de la utilización de una media muestral, designada por $\overline{X}$, para estimar el parámetro μ. Muchas pruebas estadísticas usarán medias muestrales en el análisis, aunque existen alternativas que desembocan en otro tipo de pruebas. Exploraremos algunas de ellas cuando veamos los diferentes tipos de datos, pero de momento nos concentraremos sólo en las medias muestrales como la forma de estimar medias poblacionales.

Para cualquier muestra de un tamaño concreto, podemos calcular su media, $\overline{X}$. Si representáramos el valor de la media en una distribución empírica, para todos los valores de $\overline{X}$ de muestras del mismo tamaño, surgiría un modelo. La distribución de las medias muestrales tiene la misma media que la población, pero con una dispersión mucho más pequeña que la muestra original. Esto tiene sentido porque las medias de cada muestra no tendrán el mismo grado de dispersión que los valores individuales. Las medias en los extremos de la distribución serán menos frecuentes, ya que es muy improbable que se desvíen significativamente de la media poblacional. Las muestras se comportan de manera previsible. Tendrán alguna variabilidad, pero si vienen de la misma población, el valor del estadístico caerá en un rango de valores esperado. La *distribución muestral* es la plasmación gráfica de esta frecuencia y variabilidad esperada.

La figura 11-2 es un diagrama de puntos de las alturas medias de los marcianos en 25 muestras. Al representar gráficamente estas medias, surge una distribución normal y forma un modelo previsible. Observe que la media de todas estas muestras, designadas con $\overline{X}_{\overline{X}}$, es la misma que la media poblacional, μ, pero la desviación estándar de las medias es mucho más pequeña que la de los datos originales. La desviación de las medias muestrales (en este caso, de 25 medias) se conoce como el *error están-*

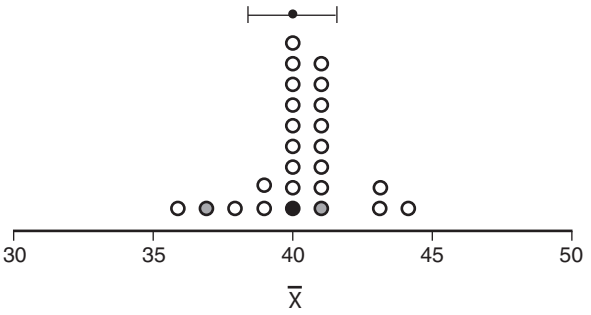


FIGURA 11-2 Distribución de las medias muestrales de las alturas de los marcianos. (De Glantz, S. A. 1992. *Primer of Biostatistics*, 3rd ed. New York: McGraw-Hill, p. 25, con autorización.)

dar (o típico) de la media, y hay que distinguirlo de la desviación estándar (o típica) de los valores de los casos en una sola muestra. El error estándar de la media se indica por $SE_{\bar{x}}$. Como Glantz afirma acertadamente, «a diferencia de la desviación estándar, que cuantifica la variabilidad en la población, el error estándar de la media cuantifica la incertidumbre en la estimación de la media»[3].

En teoría, si tomáramos las medias de una variable concreta para cada muestra posible del mismo tamaño, podríamos representarlas en una distribución empírica. Actualmente, la simulación por ordenador puede mostrar este proceso. Lo que surge es un patrón que encaja con la distribución normal, *incluso si la distribución original de los valores no fuera normal*. ¿Sorprendido? Yo al principio lo estaba. Pero tiene sentido que las *medias* muestrales tiendan a acercarse al parámetro, y que lo hagan con igual probabilidad de infraestimar o sobrestimar la media poblacional.

La figura 11-3 es una distribución de medias muestrales generadas por ordenador. La población original (gráfico A) tenía una media de 50 y una desviación estándar de 29. Era una distribución uniforme. Todos los valores entre 0 y 100 tenían la misma frecuencia.

En los gráficos B y C, cada punto representa una media muestral. Éstas tienen una distribución más dispersa para un tamaño muestral más pequeño de 5 (gráfico B), con una distribución aproximadamente normal. Cuando el tamaño de la muestra se incrementa a 30 (gráfico C), la distribución de las medias es menos dispersa.

Resulta que las muestras actúan de una forma previsible. Esta previsibilidad no sólo se da con medias muestrales, sino también con otros parámetros. Recuerde que algunos parámetros pueden ser completamente abstractos, como el «riesgo de un accidente». Para todas las posibles muestras del mismo tamaño de una población, el riesgo calculado formará un conjunto de valores previsibles. El riesgo actual en la población es un valor fijo y la muestra proporciona una estimación de este riesgo.

Piense en las distribuciones muestrales como conjuntos de números previsibles que siguen un patrón. A partir de aquí, el resto del proceso va rodado: si es demasiado improbable que nuestro estadístico provenga del conjunto previsible que esperaríamos si el tratamiento no tuviese ningún efecto, rechazaremos la hipótesis nula y afirmaremos que un tratamiento es significativamente mejor. El razonamiento oculto tras todas las pruebas estadísticas se basa en este sistema.

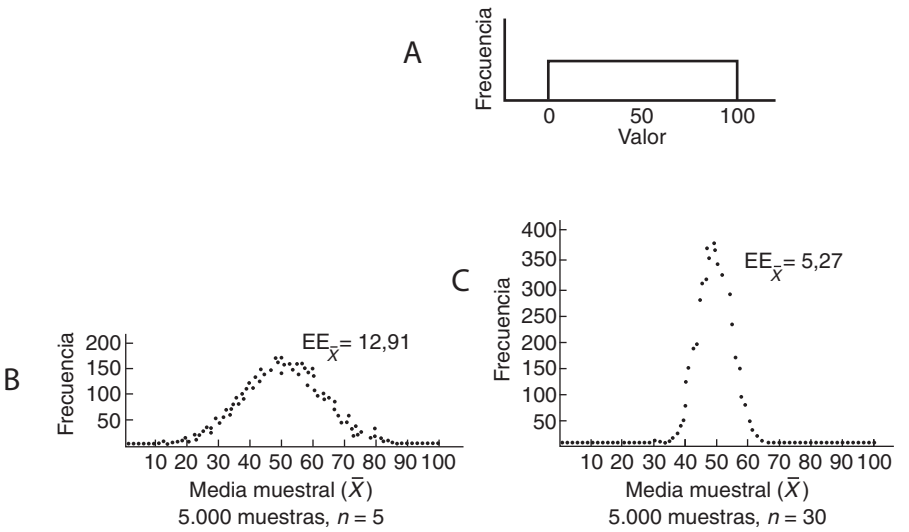


FIGURA 11-3 Distribuciones de medias muestrales generadas por ordenador. (De Howell, D. C. 1999. *Fundamental statistics for the behavioral sciences*, 4th ed. Pacific Grove, CA: Brooks/Cole Publishing Co., p. 226, con autorización).

ESTADÍSTICO DE PRUEBA

Los datos de una sola muestra pueden analizarse a través de medias o con otras formas alternativas, que dependerán del tipo de variable(s) y de la cuestión planteada. Estos análisis aplican diferentes tipos de fórmulas que generan un único número a partir de los datos. Este número se conoce como el *estadístico de prueba*.

El estadístico, por sí solo, no tiene sentido. Se interpreta por su lugar relativo dentro de una distribución de todos los posibles estadísticos de prueba que resultarían de todas las muestras del mismo tamaño, si la hipótesis nula fuera cierta. Su importancia se determina por la distancia a la que está de la mayoría de estadísticos de prueba que esperaríamos obtener si no hubiese ninguna diferencia de tratamiento o ninguna relación entre variables.

- *El estadístico de prueba es un solo número que evalúa la compatibilidad de los datos con la hipótesis nula.*

El tipo de prueba depende del tipo de variable(s) y de la cuestión planteada. Nosotros aplicamos una fórmula específica a los datos. En general, el análisis se realiza con ordenadores, con programas que, además de almacenar los datos, realizan los cálculos solicitados. El resultado es el estadístico (de prueba). Por ejemplo, una prueba chi-cuadrado de diferencia entre grupos produce un estadístico chi-cuadrado. Una prueba de correlación produce un coeficiente de correlación, otro tipo de estadístico. Algunas pruebas analizan diferencias en la media muestral, que produce un estadístico *t*. A este estadístico le aplicamos la distribución correspondiente, que es una distribución muestral específica que depende del tamaño muestral y de la variabilidad de datos.

CUADRO 11-1

Teorema central del límite

La relación entre la media de una población y su distribución muestral se conoce como el «teorema central del límite». Esencialmente, proporciona un método para construir la distribución muestral de la media poblacional. Además, proporciona un método para contrastar la hipótesis nula.

El teorema central del límite establece que:

1. *La distribución de las medias muestrales de una población tenderá a la distribución normal, aunque los datos originales no se distribuyan normalmente.*
2. *El valor de la media de todas las posibles medias muestrales ($\bar{X}_{\bar{x}}$) coincidirá con la media poblacional, μ.*
3. *El error estándar de la media ($ES_{\bar{x}}$) depende de la desviación estándar original y del tamaño muestral. El error estándar será mucho más pequeño que la desviación estándar de la muestra. Un tamaño muestral mayor comportará un error estándar aún más pequeño y una curva más estrecha.*

El error estándar de la media ($ES_{\bar{x}}$) puede calcularse a partir de la desviación estándar y del tamaño muestral. No es necesario saberla, pero para aquellos que estén interesados, la fórmula del error estándar de la media es:

$$ES_{\bar{x}} = \frac{S}{\sqrt{n}}$$

El hecho de que múltiples muestras den diferentes estimaciones de μ debido a su variabilidad innata se denomina error de muestreo. No es realmente un error; es una función de las diferencias que veríamos si estudiáramos múltiples muestras del mismo tamaño, como en la figura 11-3. La distribución de todas las posibles medias muestrales tiene una aplicación especial. Si tomamos sólo una muestra y encontramos la media y la desviación estándar, podemos usar los datos muestrales como pruebas para respaldar o rechazar la hipótesis nula basándonos en los resultados que esperaríamos de la distribución muestral.

Debido a que el estadístico estima los parámetros poblacionales, hay un margen de error a ambos lados de esta estimación. Esto es el intervalo de confianza, que ahora nos resulta un concepto familiar:

- *El intervalo de confianza es la estimación del parámetro poblacional ± un margen de error.*

EN BUSCA DE LA CORRECTA DISTRIBUCIÓN MUESTRAL

La distribución muestral tiene gran variedad de formas y tamaños. Escoger la correcta dependerá del tipo de prueba. Usted puede depender de su asesor estadístico para elegir la prueba correcta y el tipo de distribución muestral. He incluido aquí una breve exposición sobre este tema para aquellos que estén interesados en conocer las más habituales. Lo que realmente tiene que entender es que todas sirven para medir

el estadístico y determinar la probabilidad de obtener nuestro resultado fruto del azar.

Muchas distribuciones muestrales son normales, como vimos en la prueba *Z* del capítulo 9. Usamos el estadístico *Z* cuando se conoce la desviación estándar poblacional, μ. El estadístico *t* se usa para comparar medias entre grupos cuando no se conoce σ, que es el caso más habitual. Este estadístico es útil cuando comparamos dos tratamientos y la variable respuesta se mide por una escala de intervalo. Usa distribuciones de la *t*-Student, que son acampanadas y análogas a la distribución *Z*, pero con la diferencia de que hay una distribución específica para cada tamaño muestral.

Cuando queremos comparar proporciones en dos o más grupos, generamos un estadístico chi-cuadrado. Lo interpretamos dentro de una distribución chi-cuadrado, que es asimétrica. Es útil, sobre todo, para datos categóricos. Por ejemplo, la prueba chi-cuadrado puede comparar las proporciones de diferentes religiones en varios estados estadounidenses, como se discutió en el capítulo 4.

Otras dos distribuciones con importancia estadística son las distribuciones de Poisson y binomial. Se usan en situaciones específicas que implican variables discretas. Se mencionan aquí brevemente para que se familiarice con las situaciones donde se aplican.

La distribución binomial sirve cuando la respuesta tiene dos valores posibles, como «sí» o «no», «sano» o «enfermo», o «éxito» o «fracaso». Las distribuciones ilustran la probabilidad de observar un cierto número de éxitos en un número dado de intentos, como echar una moneda a cara o cruz. Si usted tirara la moneda 10 veces, esperaría aproximadamente 5 caras, que se expresa matemáticamente como un 50% de éxitos. Las fórmulas de la binomial pueden darnos la probabilidad de ver 2, 1 o hasta ninguna cara (sí, esto podría pasar: la probabilidad es 0,001, o 1 entre 1.000). Del mismo modo, usted podría determinar el resultado esperado de un suceso con una probabilidad del 20% de éxitos sobre muchos intentos, y ver si su observación se desvió significativamente de lo que esperaba basándose únicamente en la probabilidad.

La distribución de Poisson sirve para evaluar un pequeño número de entidades discretas repartidas regularmente sobre una media, relativamente grande, como partículas dispersas en el aire o en el agua. También se usa para observar acontecimientos particulares en el tiempo, como mutaciones genéticas aleatorias. En cualquier caso, deberá conocerse el parámetro poblacional, ya sea el número esperado de entidades en un espacio concreto, o los acontecimientos esperados en un intervalo de tiempo. Esta distribución evalúa si los datos se desvían de los valores esperados.

Cada tipo de distribución (p. ej., normal, binomial, Poisson o chi-cuadrado) tiene su propia forma. Las características de la curva (altura, dispersión y pendiente) dependerán del tamaño muestral, *n*. Estos detalles tienen un mayor efecto en la forma de la distribución cuando *n* es relativamente pequeño. Para grandes *n*, los efectos son más sutiles. Estas diferencias en la forma no cambian el área total bajo la curva, que siempre valdrá 1.

¿Cómo tratamos con tantas curvas posibles al interpretar el estadístico? De hecho, hay tablas publicadas para cada tipo de distribución. Por ejemplo, si tenemos un valor chi-cuadrado, consultamos una tabla chi-cuadrado, que puede encontrarse en el apéndice de la mayoría de libros de estadística. La tabla contiene varias distribuciones chi-cuadrado para diferentes tamaños muestrales. El valor de chi-cuadrado tendrá un área correspondiente bajo la curva que la define. *Z* es un tipo de estadístico

que se usa para comparar valores de una población con una desviación estándar conocida. La tabla lista el área a ambos lados de un *Z* estandarizado. En este caso no se necesita adaptar a cada tamaño muestral.

VALOR CRÍTICO

Cuando un estadístico analiza los datos y coloca los valores en la fórmula apropiada, obtiene el valor del estadístico. Entonces consulta la tabla apropiada para hacerse una idea de cuál es la probabilidad de obtener ese resultado por azar. (Muchos programas informáticos tienen estas tablas almacenadas y aplicarán el resultado a la distribución adecuada comunicando el valor *p* correspondiente.) Sabemos que cualquier valor del estadístico es posible, pero unos son más probables que otros por fluctuaciones aleatorias de las muestras. Nosotros decidimos de antemano qué error de tipo I estamos dispuestos a asumir y a partir de qué punto afirmaremos que una diferencia de tratamiento es verdadera o significativa. Esto se llama *nivel de significación*, y a menudo se fija en $\alpha = 0{,}05$.

El valor del *estadístico* que define este punto, a partir del cual diremos que el resultado es muy improbable que se haya producido por azar, se llama *valor crítico*. Podemos cometer un error de tipo I rechazando la hipótesis nula cuando es cierta, pero asumimos este pequeño riesgo. Si nuestro estadístico va más allá de este valor crítico, entonces la probabilidad de que este resultado se deba al azar es muy pequeña.

- *El valor crítico es el valor del estadístico que fija el nivel de significación especificado, que a menudo es del 5%.*

PROBABILIDAD E INFERENCIA

Recuerde que la probabilidad de un suceso o de una respuesta se expresa como una fracción entre 0 y 1. En el caso de las distribuciones muestrales, se han considerado todas las respuestas de todas las posibles muestras de un tamaño uniforme. El fundamento lógico afirma que cuando se tienen en cuenta todas las respuestas posibles, la probabilidad de que ocurra al menos una de ellas es 1. El área total bajo cualquier curva de una distribución es siempre igual a 1. Cualquier porción de esta área es una fracción entre 0 y 1. Esta área puede traducirse directamente en una probabilidad. Para un segmento de curva que contenga el 50% del área, la probabilidad de que ocurra uno de los sucesos contenidos en dicho segmento es 0,5. Para un segmento con menos del 5%, hay una probabilidad menor a 0,05 de que ocurra uno de estos acontecimientos. Cuando $p < 0{,}05$, decimos que hay una probabilidad menor del 5% de obtener únicamente por azar un estadístico al menos tan extremo como éste.

- *El área de un segmento de la distribución muestral es directamente proporcional a la probabilidad.*

El área bajo una porción de curva puede calcularse con métodos numéricos. Este laborioso proceso puede evitarse ya que, como hemos visto, hay tablas publicadas que proporcionan el área de los segmentos de estas distribuciones. Un punto importante que hay que destacar es que todo este proceso sólo funciona bajo la premisa de

que la muestra es aleatoria simple. Esto asegura que el estadístico se comporta como era de esperar. Si no se ha cumplido la selección aleatoria, entonces la lógica se estropea y el proceso es inválido. En realidad, hay métodos para corregir matemáticamente muestras «aleatorias más sofisticadas». Está más allá del alcance de este libro hablar de cuándo o cómo se hace esta corrección matemática, pero sea consciente de que los resultados serán válidos siempre y cuando se reconozca y se tenga en cuenta esta transgresión.

Si estamos investigando una asociación entre variables, usamos una prueba diferente pero la interpretación es análoga. Si no hubiera ninguna conexión real entre el valor de una variable y la otra, los pares de números no serían similares, no estarían relacionados. Habría un modelo *aleatorio predecible* en las parejas de valores. Si en cambio obtenemos un patrón sólido de relación (muy improbable si no existiera relación real), rechazaremos la hipótesis nula de no asociación y afirmaremos que la hipótesis alternativa es cierta.

PUNTOS CLAVE

- Los datos de la muestra se usan para calcular una respuesta global.
- Si estudiáramos múltiples muestras de la misma población, obtendríamos diferentes resultados.
- Si se estudiaran todas las combinaciones de posibles muestras del mismo tamaño, cada resultado tendría una frecuencia de aparición.
- Si representáramos la frecuencia de un cierto resultado que obtuviéramos estudiando todas las muestras posibles, observaríamos un patrón llamado distribución muestral.
- La distribución muestral puede seguir una distribución normal o puede tomar otra forma, según sea el tipo de datos y la cuestión planteada.
- La inferencia estadística nos permite contestar «¿Cuál es la probabilidad de observar este resultado si la hipótesis nula es cierta?».
- Las muestras actúan de forma previsible. La distribución muestral es un conjunto de valores esperados.
- Introducimos los datos de la muestra en una fórmula para obtener un estadístico. La fórmula particular depende del tipo de variable(s) y de la cuestión planteada.
- El estadístico es un único número que evalúa la compatibilidad de los datos con la hipótesis nula. Representamos el estadístico en su distribución correspondiente para ver la verosimilitud, o probabilidad *p*, de obtener este resultado si la hipótesis nula fuera cierta.
- Si el resultado es demasiado improbable, rechazaremos la hipótesis nula de no diferencia entre tratamientos, o de no relación entre variables, y afirmaremos que la hipótesis alternativa es cierta.

BIBLIOGRAFÍA

1. Hooke, R. 1983. *How to tell the liars from the statisticians*. New York: Marcel Dekker, Inc.
2. Glantz, S. A. 1992. *Primer of Biostatistics*, 3rd ed. New York: McGraw-Hill, p. 23.
3. Glantz, S. A. 1992. *Primer of Biostatistics*, 3rd ed. New York: McGraw-Hill, p. 28.

PREGUNTAS DE REVISIÓN

1. La _____ _____ es una distribución empírica de todos los estadísticos que podrían obtenerse de muestras del mismo tamaño.

2. La probabilidad de obtener un estadístico de una muestra concreta está relacionada con el _____ bajo la curva en aquel punto.

3. Los _____ que pertenecen a los extremos de la distribución son menos probables que ocurran por azar.

4. El valor del estadístico que define el punto a partir del cual rechazaremos la hipótesis nula es el _____ _____.

5. Habitualmente, el valor crítico define una probabilidad de _____ de que el resultado ocurra por azar si la hipótesis nula es cierta.

RESPUESTAS

1. distribución muestral

2. área

3. estadísticos

4. valor crítico

5. 0,05 o del 5%

CAPÍTULO 12

Tipos de pruebas estadísticas

La estadística es la esencia del método científico: se usa en todas las principales disciplinas.

—D. B. Owen y Nancy R. Mann[1]

Las pruebas estadísticas se diseñan para contestar a la pregunta de si se ha de rechazar la hipótesis nula. El matiz de cada hipótesis nula cambia ligeramente según la aplicación, pero en general afirma que la distancia observada en los datos entre el estadístico y su valor esperado se debe a fluctuaciones aleatorias. Recuerde que la variación aleatoria sigue el patrón que recoge la distribución muestral. Si es muy inverosímil que los datos se produzcan por azar, rechazaremos la hipótesis nula y aceptaremos la hipótesis alternativa. Esta lógica es consistente en todas las pruebas estadísticas.

Se pueden realizar muchos tipos de pruebas estadísticas, según el tipo de variables y la cuestión planteada. Este capítulo tratará sobre algunas de las pruebas más habituales. Aquí no se han incluido las fórmulas porque no son fundamentales para entender el proceso de las pruebas de hipótesis. Como usted sabe, el valor *p* se interpreta del mismo modo sea cual sea el tipo de prueba realizada.

CHI-CUADRADO

Esta prueba, usada frecuentemente, puede decirnos si existe una diferencia en la proporción de una variable categórica que se desvíe por encima de lo previsible respecto al caso de una distribución regular entre los diferentes grupos, como enunciaría la hipótesis nula. Esta prueba se basa en la «bondad del ajuste» porque evalúa si los datos observados encajan con un modelo esperado cuando la hipótesis nula es cierta. También se usa para estudiar tendencias.

La fórmula del estadístico chi-cuadrado χ^2 es una suma de las diferencias para cada casilla entre lo que se observa y lo que se esperaría si la hipótesis nula fuera correcta. Su cálculo será sumamente complicado si el número de casillas es muy elevado. El estadístico puede alcanzar valores altos, hasta de dos o tres cifras. Cuanto mayor sea el estadístico, menor será la probabilidad de que la hipótesis nula sea cierta y los datos sean consecuencia del azar. La figura 12-1 es un ejemplo de un conjunto simple de datos que sirven para contestar a la pregunta: «¿Hay una diferencia significativa entre hombres y mujeres con respecto a su probabilidad de ser pobres?».

En esta muestra, hay 300 mujeres y 200 hombres. ¿Existe una diferencia significativa en la probabilidad de ser pobre entre hombres y mujeres?

	Mujeres	Hombres	Total
En la pobreza	150 (cell a)	50 (cell b)	200
Fuera de la pobreza	150 (cell c)	150 (cell d)	300
Total	300	200	500

FIGURA 12-1 Si no hubiera ninguna diferencia entre hombres y mujeres con respecto a su probabilidad de ser pobres, la proporción de hombres y mujeres pobres no debería ser significativamente diferente.
Véase www.brynmawr.edu/Acads/GSSW/Vartanian/Handouts/ChiSqu.Examples (2007) para una completa explicación.

Con el estadístico chi-cuadrado, si hay más de dos grupos y obtenemos un resultado estadísticamente significativo, sólo podemos decir que al menos uno de los grupos es significativamente diferente del resto. Hacer comparaciones entre todos los grupos requiere métodos estadísticos adicionales.

PRUEBA *T*-STUDENT

Esta prueba sirve, en variables continuas, para comprobar si hay una diferencia entre las medias muestrales de dos grupos. Se suele usar para demostrar una diferencia significativa debida al tratamiento, o para ver si hubo un cambio en un grupo antes y después del tratamiento.

La muestra se divide en dos grupos: uno que recibe una intervención, por ejemplo, y otro que no. Si la intervención no tiene ningún efecto, entonces las medias de la variable respuesta en los dos grupos no se diferenciarían sensiblemente. Cuando una media muestral se resta de otra, la diferencia está cerca del cero. La hipótesis nula afirma que las medias de las variables respuesta en los dos grupos serán la misma y que no existe diferencia entre los tratamientos.

El estadístico *t* es análogo al estadístico *Z,* y a medida que la muestra se hace más grande, la distribución *t*-Student se parece cada vez más a la distribución *Z*. Podemos realizar pruebas bilaterales con estas distribuciones o demostrar diferencias en uno u otro sentido.

CORRELACIÓN

En el mundo real observamos que ciertas variables parecen estar relacionadas. Por ejemplo, el color del cabello y el de los ojos tienden a estar relacionados. Las variables que están mutuamente relacionadas se dice que están *correlacionadas*. Una no necesariamente causa la otra, pero tienden a producirse al unísono. Cuando se representan gráficamente los valores de cada una para un individuo dado, se presenta una relación ya que los puntos tienden a agruparse a lo largo de una línea recta. La fuerza de la correlación se expresa con un número llamado *coeficiente de correlación de Pearson* (simbolizado con *r*).

El coeficiente de correlación de Pearson puede variar entre –1 y 1. Una relación más sólida y con menos parejas discordantes tendrá un coeficiente de correlación de Pearson más próximo a –1 o a 1. Cuando $r = 1$, hay una correlación perfecta (¡inexistente en la naturaleza!). Si representásemos los datos, obtendríamos una línea absolutamente recta con una pendiente positiva. Cuando $r = -1$, se obtiene una línea absolutamente recta pero con una pendiente negativa. En este caso, las variables tienen una correlación negativa, es decir, un valor más alto de una de las variables está asociado a un valor inferior de la otra. Cuando $r = 0$, no hay correlación lineal y los puntos se dispersan aleatoriamente. La hipótesis nula declara que $r = 0$.

Cuando las variables tienen una relación lineal directa, el aumento de una afecta a la otra. La pendiente de la recta explica el grado de impacto. La proximidad de los puntos alrededor de la recta reflejan un valor de r más alto (positivo o negativo). Algunas variables pueden relacionarse de forma no lineal. Una línea curva puede ser el mejor ajuste en variables pareadas, lo que significa que el valor de una se relaciona con una función mayor de la otra. Esto no afecta ni a la interpretación del coeficiente de correlación de Pearson ni a su valor p correspondiente. La figura 12-2 muestra ejemplos de asociaciones entre dos variables.

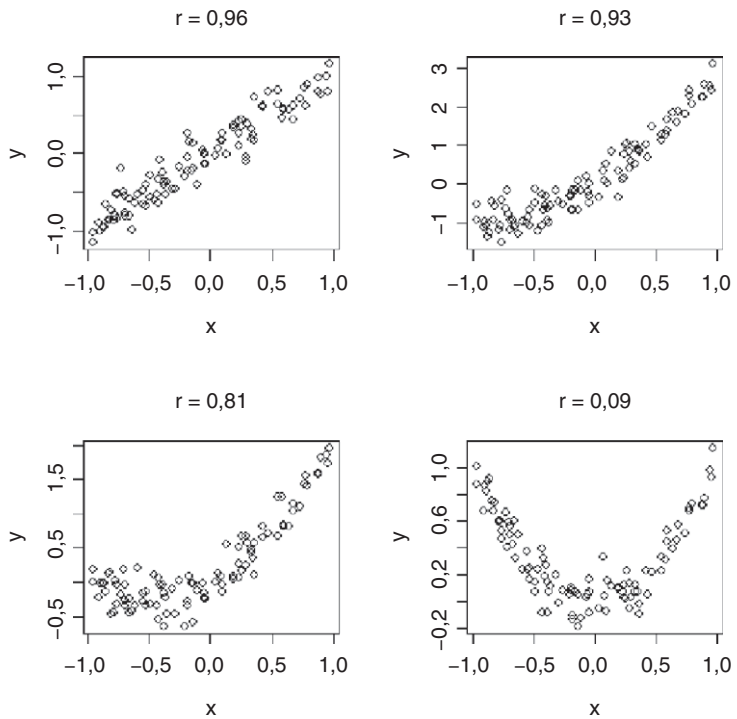


FIGURA 12-2 Ejemplos de gráficos de correlación y coeficiente de correlación de Pearson para varias asociaciones teóricas entre las variables *x* e *y*. De: http://math.uprm.edu/~wrolke/esma3103/graphs/correlationfig4.png (agosto de 2007).

REGRESIÓN

La correlación mide la fuerza y la dirección de dos variables relacionadas. Por otra parte, la regresión mide la fuerza de una asociación cuando una variable ayuda a explicar o a predecir otra. Las ecuaciones de regresión se modelan en busca de la conocida ecuación algebraica de una recta, $y = bx + a$, de la cual se habló en el capítulo 2, «Principios matemáticos». Si resolviéramos la ecuación hallando y (variable respuesta) para un valor dado de x (variable explicativa), tendríamos una respuesta exacta. Pero la bioestadística empieza con valores de x e y para cada individuo e intenta hallar la ecuación o modelo que mejor explique la relación. El resultado es la recta que mejor se ajusta a los datos, que será aquella con menor desviación de los puntos.

La hipótesis nula afirma que no hay ninguna relación significativa entre las variables y que la pendiente es $b = 0$. Un resultado significativo haría verosímil la existencia de una relación, entre los valores de x e y, no atribuible al azar. La regresión es un instrumento potente porque puede demostrar la asociación entre muchas variables explicativas y la variable respuesta, y puede ponderar el efecto independiente de cada una de ellas. También permite mostrar relaciones no lineales entre las variables explicativas y la respuesta. La regresión múltiple se usa si la variable respuesta es continua, mientras que la regresión logística sirve para respuestas dicotómicas.

ANÁLISIS DE LA VARIANZA (ANOVA)

Esta prueba estadística, al igual que la *t*-Student, trata con diferencias de medias pero tiene la capacidad de comparar varios grupos. Lo hace comparando la variabilidad dentro de cada grupo con la variabilidad dentro de toda la muestra. Hay algunas premisas subyacentes en el ANOVA: los valores se distribuyen normalmente y las varianzas de los grupos son similares. La hipótesis nula afirma que todas las *medias* de los grupos son iguales. Si esto es verdad, las *varianzas* de las medias de los grupos no deberían ser significativamente diferentes de la varianza de la muestra entera. La prueba *F* es la prueba de significación que compara las varianzas.

El ANOVA es un instrumento muy útil, ya que puede evaluar simultáneamente los efectos de múltiples variables independientes. También puede detectar una interacción, es decir, si el efecto de una variable concreta cambia por la presencia de una tercera variable.

PRUEBA DE MANN-WHITNEY

Las variables que no se distribuyen normalmente contrastan la significación a través de una medida de tendencia central que no es la media poblacional. Estas pruebas se denominan *no paramétricas*.

La prueba de Mann-Whitney ordena todos los resultados individuales en la muestra y luego compara la suma de los rangos (posiciones) entre los grupos. La hipótesis nula establece que las sumas de rangos en cada grupo no difieren, ya que si los grupos son iguales no debería haber ninguna tendencia de los rangos a ser más altos o más bajos en ninguno de ellos. Si los grupos son diferentes, entonces los rangos divergirán. La prueba de Wilcoxon es similar, pero para datos pareados. La prueba de Kruskal-Wallis

es parecida al ANOVA, pero con datos ordenados: compara varios grupos para comprobar si sus sumas de rangos son similares, como afirma la hipótesis nula.

PRUEBA DE CORRELACIÓN DE SPEARMAN

Una vez ordenados los datos también puede estudiarse su correlación. En este caso puede calcularse el coeficiente de correlación de Spearman y contrastarse la hipótesis nula de no correlación entre los datos ordenados. Esta hipótesis asume que un aumento en la posición o rango de una variable en un individuo no tiene relación con el aumento o la disminución en la otra variable. Se interpreta igual que el coeficiente de correlación de Pearson. Se usa cuando los datos no son normales o cuando se ha recogido únicamente el orden.

PUNTOS CLAVE

- La prueba chi-cuadrado evalúa diferencias entre grupos en la proporción de una variable.
- Cuando dos variables están relacionadas para todos los individuos, se dice que están correlacionadas.
- La regresión evalúa el efecto independiente de varias variables explicativas en la variable respuesta.
- El ANOVA es una prueba que evalúa diferencias entre varios grupos examinando las varianzas de cada uno de ellos y comparándolas con la varianza de la muestra.
- Los rangos se evalúan con pruebas no paramétricas que no necesitan asumir que la variable siga una distribución normal.

BIBLIOGRAFÍA

1. Owen, D. B. y N. R. Mann (Series Editors) en: Hooke, R. 1983. *How to tell the liars from the statisticians*. New York: Marcel Dekker, Inc.

LECTURAS RECOMENDADAS

Howell, D. C. 1999. *Fundamental statistics for the behavioral sciences*, 4th ed. Pacific Grove, CA: Brooks/Cole Publishing Co.

Moore D. S. y G. P. McCabe. 1999. *Introduction to the practice of statistics*, 3rd ed. New York: W. H. Freeman and Co.

www.statsoft.com/textbook 2005

www.udel.edu/~mcdonald 2005

PREGUNTAS DE REVISIÓN

1. La _____ implica una asociación entre dos variables, pero no causa y efecto.

2. Los modelos de regresión miden la _____ del efecto de una variable en la respuesta.

3. Los métodos de regresión también pueden comprobar el efecto independiente de varias _____ en la respuesta.

4. Para datos no normales, se usan pruebas _____ que consideran otras medidas como el rango o la posición en vez de las medias muestrales.
5. La prueba chi-cuadrado evalúa las diferencias en la _____ de una variable entre diferentes grupos.

RESPUESTAS

1. correlación
2. fuerza
3. variables
4. no paramétricas
5. proporción

CAPÍTULO 13

Propiedades de los intervalos de confianza

La inferencia estadística … proporciona, en lenguaje probabilístico, cuánta confianza podemos tener en nuestras conclusiones.

—David S. Moore y George P. McCabe[1]

Ya sabemos manejar los estadísticos para estimar el parámetro poblacional desconocido, como la media poblacional, μ. Este parámetro es muy real e invariable, pero no podemos medirlo, sólo podemos estimarlo. ¿Cuánta precisión tiene nuestra estimación? Podemos contestar a esta pregunta mediante los intervalos de confianza, un concepto introducido por el distinguido estadístico polaco Jerzy Neyman en 1937. Los intervalos de confianza son como paraguas que intentan «atrapar» el parámetro poblacional que estimamos.

Cuando estamos bajo un gran paraguas, nos sentimos protegidos y salvaguardados. Es el mismo sentimiento que tenemos con un gran intervalo de confianza que contenga un amplio abanico de valores, ya que probablemente contendrá el parámetro poblacional que tratamos de estimar. Por supuesto, un amplio rango de valores ha limitado la utilidad del intervalo para estimar el parámetro. No sabemos con certeza dónde está el parámetro dentro del intervalo, pero confiamos en que el intervalo lo contenga. Es el precio que pagamos por sentirnos seguros en la estimación. Si queremos una confianza del 99%, el intervalo será mayor que si nos conformamos con una confianza del 90%. Como hemos visto, es habitual aceptar un intervalo de confianza del 95%. Esto significa que si el estudio se repitiese con el mismo tamaño muestral, el 95% de las veces nuestro intervalo de confianza incluiría el verdadero parámetro poblacional.

Por ejemplo, para medir los pesos de los neonatos de madres fumadoras, podríamos tomar una muestra aleatoria de 25 niños, pesarlos al nacer y promediar los pesos. Pretendemos estimar el verdadero peso medio de la población de todos los neonatos de madres fumadoras. Asumimos que los pesos se distribuirán normalmente. Suponga que obtuvimos una media muestral de 6 libras, con una desviación estándar de 1 libra (datos ficticios). Con el *teorema central del límite* podemos calcular el error estándar de la distribución muestral, que será de 0,2 libras. Para calcular el intervalo de confianza, la fórmula es:

Estimación ± Margen de error

La estimación es la del estadístico muestral que se genera a partir de los datos. En este caso, es la media muestral. El margen de error depende de tres aspectos:

1. La desviación estándar de la muestra.
2. El tamaño muestral.
3. La confianza que deseemos, que normalmente será del 95%.

La fórmula exacta variará según el tipo de datos que tengamos, pero estos tres valores son los que contribuyen a la amplitud del rango que se obtiene independientemente de la fórmula específica. A quienes deseen conocer estas fórmulas les remito a cualquier libro de texto estándar sobre estadística.

El intervalo de confianza calculado para los datos citados resultó ser:

$6 \pm 0{,}40$ libras, o de 5,60 a 6,40 libras con un nivel de confianza del 95%

Este intervalo tiene una probabilidad del 95% de incluir μ. *Confiamos* al 95% que nuestro intervalo contendrá μ. (No es lo mismo que afirmar que μ tiene una probabilidad del 95% de estar dentro de estos límites, ya que μ es constante, y está o no está dentro de los límites. Es un matiz sutil pero válido.) Si repitiéramos el experimento múltiples veces con muestras aleatorias del mismo tamaño, los intervalos de confianza se situarían alrededor del valor μ invisible. El 95% de los intervalos de confianza de las diferentes muestras contendrían μ. La figura 4-1 del capítulo 4 representa el estadístico muestral y los intervalos de confianza del 95% de varias muestras del mismo tamaño. Fíjese en que los intervalos de confianza tienen longitudes ligeramente diferentes. La causa es que cada muestra tiene una desviación estándar diferente, que es uno de los factores que afectan a la amplitud del intervalo de confianza.

Una manera de estrechar el intervalo es reducir el nivel de significación para el mismo tamaño de muestra. Si nos sentimos cómodos perdiendo el verdadero parámetro el 10% de las veces en vez del 5%, podemos reducir el nivel de confianza al 90%. Esto resulta en un intervalo de confianza más pequeño de:

$6 \pm 0{,}34$ libras, o de 5,66 a 6,34 libras

Nuestro intervalo se ha estrechado, pero a expensas de aumentar la probabilidad de perder μ.

Para aumentar nuestras posibilidades de capturar a μ, podemos incrementar el nivel de confianza al 99%. El resultado es:

$6 \pm 0{,}56$ libras, o de 5,44 a 6,56 libras

La figura 13-1 ilustra cómo el cambio en las exigencias del intervalo de confianza cambia el margen de error y, en consecuencia, la probabilidad de capturar a μ.

Mayores tamaños muestrales generalmente dan resultados más precisos. Cuantos más datos poseemos, más capaces somos de estimar el verdadero parámetro poblacional. Un aumento del tamaño muestral, por tanto, también estrechará el intervalo de confianza. La figura 13-2 muestra tres tamaños muestrales diferentes tomados de una población de neonatos de madres fumadoras. El mayor tamaño muestral tiene un intervalo de confianza más pequeño. No obstante, tamaños muestrales muy grandes tendrán menos efecto en la amplitud del intervalo de confianza.

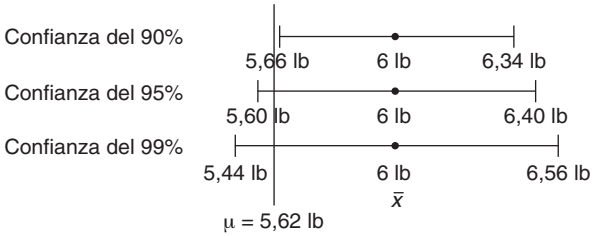


FIGURA 13-1 Intervalos de confianza para los pesos de neonatos de madres fumadoras (datos ficticios). Si μ es 5,62 libras, nuestro intervalo de confianza del 95% con estimación puntual de 6 libras contenía μ, pero se perdió en el intervalo de confianza del 90%. Aunque los intervalos con bajos niveles de confianza sean más estrechos, corren un mayor riesgo de no contener el verdadero parámetro poblacional.

El intervalo también será más pequeño si hay menos desviación o dispersión de los valores en la muestra. Esto no es algo manipulable, como el tamaño muestral; es una característica de la población en estudio. Las fórmulas exactas para calcular los intervalos de confianza tienen en cuenta la desviación estándar, el tamaño de la muestra y el nivel de confianza. En el resumen hay tres motivos por los cuales un intervalo de confianza será más estrecho:

- Tamaño de muestra grande.
- Desviación estándar pequeña en los datos de la muestra.
- Nivel de confianza relativamente bajo.

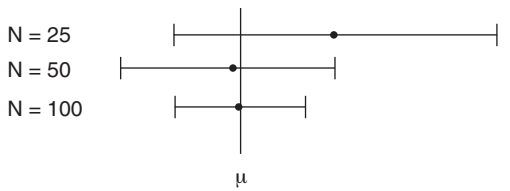


FIGURA 13-2 Medias muestrales e intervalos de confianza que intentan capturar el verdadero parámetro poblacional μ con una confianza del 95% con tamaños muestrales diferentes. La muestra más grande tiene el intervalo más estrecho.

El intervalo de confianza es una forma de expresar la amplitud o estrechez relativa de la distribución muestral. Los intervalos más anchos son menos útiles para precisar μ que los intervalos más compactos. En general, los intervalos estrechos nos dan más tranquilidad y nos dicen qué datos serán más fiables en la estimación del parámetro.

PUNTOS CLAVE

- El intervalo de confianza intenta capturar el parámetro poblacional.
- La amplitud del intervalo depende de tres cosas: la desviación estándar de los datos, el tamaño muestral y el nivel de confianza exigido (normalmente del 95%).

- Podemos estrechar el intervalo de confianza disminuyendo el nivel de confianza.
- Un modo más eficaz de estrechar el intervalo es aumentar el tamaño muestral.
- Los tamaños de muestra muy grandes tendrán un rendimiento decreciente en el efecto de la anchura del intervalo de confianza.

BIBLIOGRAFÍA

1. Moore, D. S. y G. P. McCabe. 1999. *Introduction to the Practice of Statistics*, 3rd ed. New York: W. H. Freeman and Co.

PREGUNTAS DE REVISIÓN

1. El intervalo de confianza es la estimación del parámetro más o menos el _____ __ _____.

2. Los _____ __ _____ más estrechos son más útiles.

3. El rango del intervalo de confianza depende de la _____ _____ de los datos, del _____ muestral y del _____ __ _____ deseado.

4. Una manera de disminuir el margen de error en nuestra estimación del parámetro es _____.

RESPUESTAS

1. margen de error

2. intervalos de confianza

3. desviación estándar, tamaño, nivel de confianza

4. aumentar el tamaño de muestra

CAPÍTULO 14

Potencia

Los que apuestan en las carreras de caballos no ignorarían el hecho de que hasta el mejor caballo a veces pierde, pero a menudo ignoramos el hecho de que hasta los mejores experimentos a veces no producen ninguna diferencia significativa.

—David C. Howell[1]

En todo momento, durante nuestro estudio de la bioestadística, nos hemos dado cuenta de que en cualquier experimento existe la posibilidad de llegar a una conclusión errónea. Nos sentimos cómodos sabiendo que nuestros datos pueden conducirnos a una inferencia incorrecta. Hasta tenemos nombres para los tipos de errores que podríamos cometer. Sin embargo, mientras la probabilidad de error sea escasa (menos del 5%), estamos dispuestos a aceptar el riesgo.

Hemos considerado detenidamente la posibilidad de cometer un error de tipo I. Ésta es la situación en la que los datos no respaldan la hipótesis nula, y a pesar de no existir ningún efecto real del tratamiento o relación entre las variables, hemos «decidido» que sí que existe. Hemos rechazado la hipótesis nula de no efecto y hemos afirmado erróneamente que es cierta la hipótesis alternativa. No lo hemos hecho intencionadamente; actuamos con la posibilidad de que los datos puedan, con poca frecuencia, llevarnos por el mal camino.

En la figura 14-1 se ilustra de manera gráfica cómo podríamos cometer este tipo de error mediante la distribución normal de un estadístico muestral. Si el tratamiento realmente no produce ningún efecto, entonces la hipótesis nula es cierta pero nuestra prueba estadística (aun siendo improbable) nos hizo pensar lo contrario. Se debió a la mala fortuna de escoger una muestra que nos incitó a rechazar la hipótesis nula. Si esto realmente pasara, probablemente se nos corregiría en el futuro. Los resultados observados en los subsecuentes experimentos no apoyarían estas conclusiones originales, o los resultados del tratamiento aplicados a la comunidad no serían positivos.

Si estuviéramos muy preocupados por cometer un error de tipo I, podríamos poner nuestro nivel α muy bajo (p. ej., a 0,01) y minimizar esta probabilidad. Sin embargo, cuando hacemos esto, será menos probable que descubramos un beneficio del tratamiento si realmente existe. Una prueba con un nivel α muy pequeño no es muy útil porque casi nunca rechaza H_0. Cuando disminuimos la probabilidad de un error de tipo I, estamos aumentando la probabilidad de un error de tipo II, que afir-

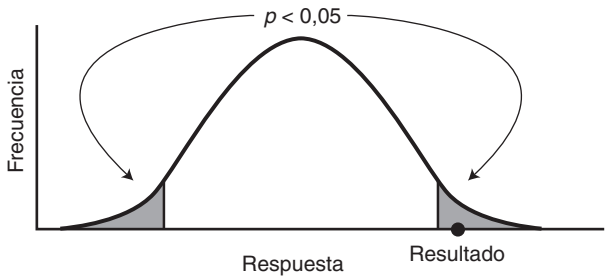


FIGURA 14-1 Distribución de la media muestral si H_0 es cierta. Este resultado improbable provocó que rechazáramos la hipótesis nula de no efecto del tratamiento, aunque era cierta. $\alpha = 0{,}05$, usando una prueba bilateral. Éste es un error de tipo I. Esto pasará, en promedio, una vez de cada 20.

maría incorrectamente que *no* hay ningún efecto del tratamiento. Nos gustaría una prueba que minimizara el margen de error evitando cometer un error de tipo I, pero también queremos una prueba que sea capaz de detectar un beneficio real del tratamiento.

Si realmente el tratamiento produjese un beneficio, entonces habría dos distribuciones muestrales diferentes: una para el grupo control y otra para el grupo tratado. Éste es un concepto importante que explica la esencia de las pruebas estadísticas. La razón por la cual obtenemos una diferencia en los resultados (aun teniendo en cuenta la variabilidad entre los grupos) es que las distribuciones de la variable respuesta son diferentes para cada grupo, ya que uno de ellos ha cambiado debido a la intervención. Ésta es una forma matemática compleja de argumentarlo; podría simplemente decirse que hay una diferencia real en la respuesta debido a la intervención.

Cuando comparamos dos grupos que han sido expuestos a una intervención, miramos la diferencia de medias en la variable respuesta entre los dos grupos, que se representa con δ. La hipótesis nula enunciaría que los grupos son iguales, $\delta = 0$. Si el tratamiento tiene un efecto en la respuesta, las medias de estos grupos estarán más separadas, como en la figura 14-2. El estadístico muestral (y la prueba estadística resultante) que obtengamos probablemente no será compatible con los resultados que obtendríamos si la hipótesis nula fuera cierta.

A veces el tratamiento funciona pero es más sutil. La diferencia entre las respuestas entre los grupos no será tan perceptible y las curvas tendrán más solapamiento, como en la figura 14-3. Como la diferencia es más pequeña, es más fácil dejar pasar una diferencia real aunque exista, porque nuestros resultados podrían todavía estar dentro de la zona donde no esperaríamos ningún beneficio del tratamiento.

Si nuestro estadístico muestral cayera dentro del rango de valores consistentes con H_0, supondríamos incorrectamente que no hay ningún efecto del tratamiento. No rechazaríamos H_0 y afirmaríamos que los datos respaldan que no existe efecto del tratamiento, pero nos equivocaríamos al hacerlo. Esto es un error de tipo II. El hecho de que la diferencia en el resultado observado sea pequeña conlleva problemas.

Los errores de tipo II se comunican como si no hubiera ningún efecto del tratamiento, si es que llegan a ser publicados. Estos resultados pueden terminar en una estantería polvorienta sin disfrutar de la notoriedad de sus estudios hermanos que son capaces de mostrar resultados significativos. Éste es el mayor incentivo del inves-

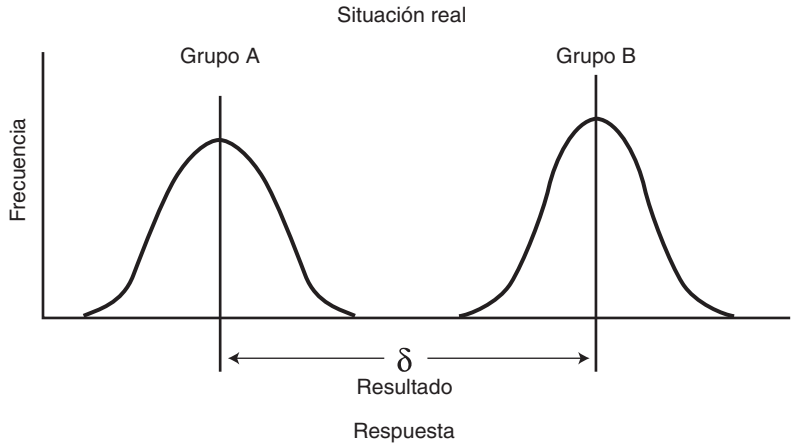


FIGURA 14-2 En un efecto real del tratamiento, la diferencia observada en la respuesta se debe, en gran parte, al efecto de la intervención más que a la variabilidad aleatoria de los datos. El resultado es más compatible con la hipótesis alternativa que con la nula. Hemos rechazado correctamente la hipótesis nula y hemos hallado un efecto real del tratamiento.

tigador para diseñar un estudio que minimice la probabilidad de un error de tipo II. Esto mejoraría la potencia de un estudio, que es la probabilidad de rechazar acertadamente una H_0 falsa. Se suele aceptar una probabilidad de que esto ocurra del 80%.

- *La potencia es la probabilidad de rechazar correctamente H_0 cuando es verdadero el valor alternativo del parámetro.*

Si quisiéramos incrementar la potencia de un estudio, ¿cómo lo haríamos? Si cambiamos el nivel de significación de modo que $\alpha = 0{,}1$, con mayor probabilidad hallaremos una diferencia significativa y rechazaremos H_0 porque ahora nuestro

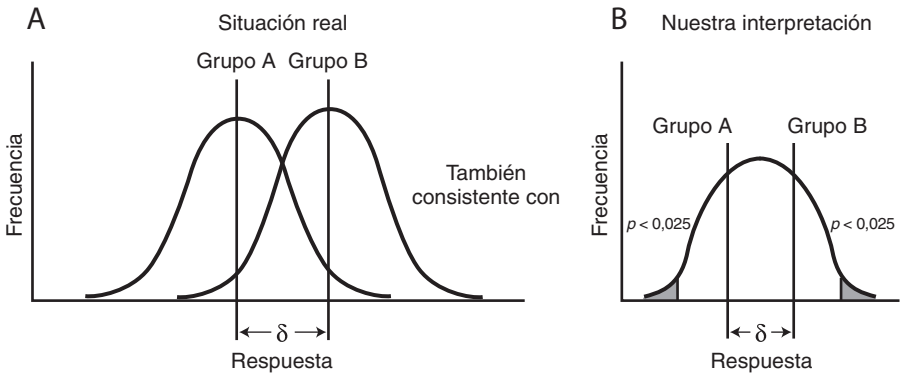


FIGURA 14-3 Existe un efecto real pero sutil del tratamiento. Las medias de las variables respuesta son diferentes, pero están próximas. No podemos rechazar la hipótesis nula con un nivel de significación del 5%. Éste es un error de tipo II.

estadístico puede caer más allá del valor crítico. Sin embargo, esto podría causar sin querer un error de tipo I. Rechazaríamos más probablemente H_0, pero hemos aumentado la probabilidad de estar equivocados de 1 de cada 20 veces a 1 de cada 10. Deberíamos seguir con la significación convencional del 95%.

Para aumentar la potencia, efectivamente queremos separar las curvas que representan las respuestas. Sabemos que una distribución será más estrecha si tiene una desviación o error estándar más pequeño, pero esto no lo podemos controlar.

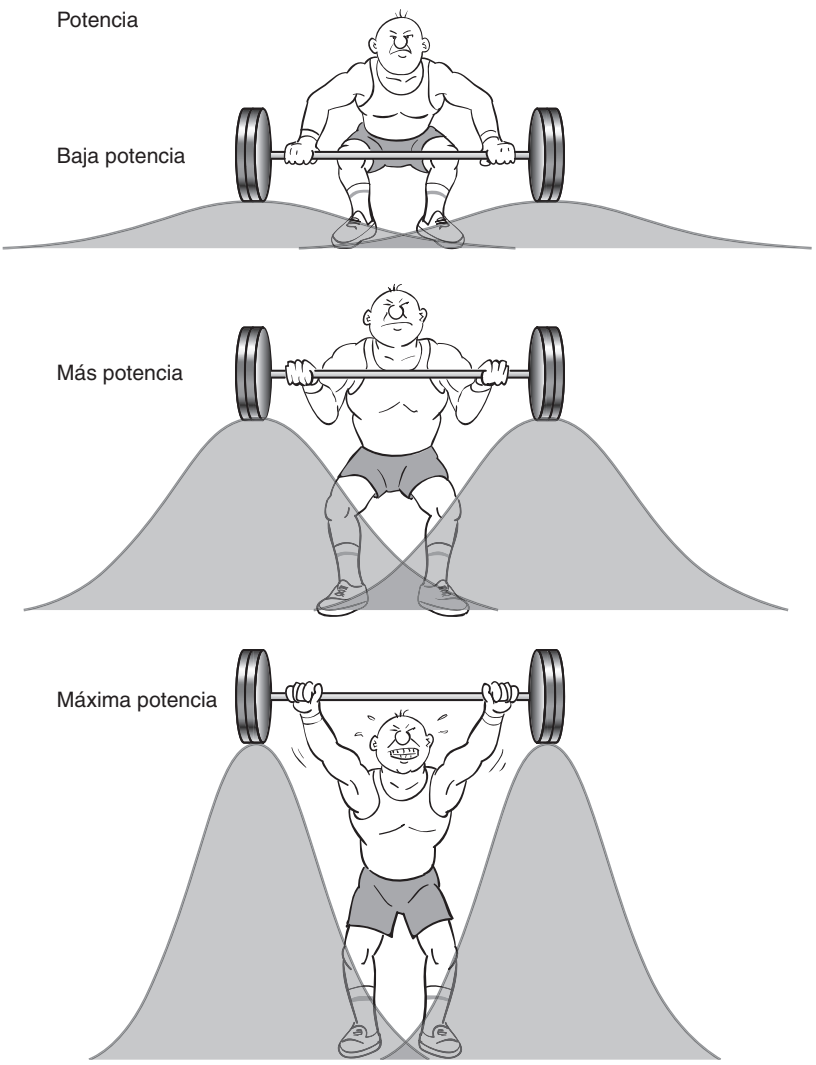


FIGURA 14-4 Cuando hay un beneficio real del tratamiento, H_0 no es cierta, y nuestra capacidad de detectarlo aumenta cuando hay menos solapamiento entre las distribuciones muestrales. Al aumentar la potencia, el levantador de pesas es capaz de «alcanzar más altura» para aislar las curvas minimizando la probabilidad de un error de tipo II.

¿Y el tamaño muestral? Es verdad que un mayor tamaño producirá una distribución muestral más estrecha, y éste es un factor *controlable*. De hecho, hay fórmulas para calcular el tamaño necesario para hallar un efecto significativo (si existe) a un nivel de potencia dado. El cálculo también tiene en cuenta el nivel de significación (normalmente del 5%), el valor de δ y la desviación o error estándar. Estos últimos pueden estimarse usando datos de estudios publicados similares o haciendo un estudio piloto más pequeño.

La figura 14-4 ilustra lo que le pasa a las distribuciones muestrales cuando se aumenta la potencia de un estudio para un valor concreto de δ. Aunque el parámetro δ permanezca inalterable, se hallará con mayor probabilidad una diferencia significativa al haber menos solapamiento entre las curvas. Esto puede llevarse a cabo aumentando el tamaño muestral. Así, el estadístico muestral se convierte en un indicador más fiable de la existencia real de una diferencia significativa en las respuestas.

Podemos calcular el número de individuos que tendremos que incluir considerando la potencia de la prueba de significación, el valor de α, el valor de δ y la desviación estándar. Esto es como estimar la cantidad de comida necesaria para ir de picnic. Podemos estimar la cantidad que necesitamos basándonos en la comida consumida en otros picnics y en cuántas personas esperamos que vengan. Siempre es mejor tener comida de sobra, ya que al final siempre se acaba añadiendo alguien con quien no contábamos. Es decir, tenemos que contar con el hecho de que nuestra estimación del error estándar puede ser ligeramente inexacta y que algunos individuos abandonarán el estudio antes del final.

CUADRO 14-1

Los cálculos del tamaño muestral varían según el tipo de análisis. La fórmula general es:

$$N = \left[\frac{(z' + z'')\sigma}{\mu_2 - \mu_1}\right]2$$

Donde (z' + z") es una función del nivel de significación y de la potencia deseada que puede obtenerse de una tabla. Vemos que el tamaño muestral, a fin de reducir un error de tipo II, puede determinarse con la potencia, el nivel de significación, la desviación estándar estimada y el tamaño del efecto detectable reflejado en el denominador. Hay varios sitios web que hacen el cálculo introduciendo estos números.

Por ejemplo, si tuviéramos que realizar un experimento con 25 individuos, para una diferencia en el efecto de 5 en un conjunto de datos con distribución normal, media de 50 y desviación estándar de 15, con un nivel de significación del 0,05, la potencia de la prueba sería de 0,38. Esto significa que cuando H_0 es realmente falsa y hay un efecto real del tratamiento, sólo en el 38% de las veces se obtendrá una diferencia significativa. Usando estos cálculos, la potencia puede incrementarse a 0,80 aumentando el tamaño muestral a 71 individuos.

PUNTOS CLAVE

- Nos sentimos cómodos con el hecho de que los datos pueden llevarnos a rechazar H_0 aun siendo cierta. Este error de tipo I pasa el 5% de las veces si aceptamos el nivel de significación estándar del 0,05.
- Cuando hay un efecto real del tratamiento y no rechazamos H_0, cometemos un error de tipo II.
- La potencia es la probabilidad de rechazar correctamente H_0 cuando es falsa.
- Un manera de incrementar la potencia es aumentar el tamaño de la muestra.
- El cálculo de tamaño muestral halla *N*, que es el número de individuos necesarios para minimizar la probabilidad de un error de tipo II.

BIBLIOGRAFÍA

1. Howell, D. C. 1999. *Fundamental statistics for the behavioral sciences*, 4th Ed. Pacific Grove, CA: Duxbury Press.

PREGUNTAS DE REVISIÓN

1. La potencia de un estudio es su capacidad de _____ H_0 correctamente y descubrir que existe un efecto real del tratamiento.

2. Un tamaño muestral _____ provoca un incremento de potencia, si los demás factores permanecen invariables.

3. Cuando _____ incorrectamente H_0, cometemos un error de tipo II.

RESPUESTAS

1. rechazar

2. mayor

3. rechazamos

CAPÍTULO 15

Sesgo

La ciencia de la estadística ha progresado mucho en este siglo, pero el avance ha ido acompañado por un incremento correspondiente del mal uso de la estadística.

—Robert Hooke[1]

El prejuicio es un rasgo común en los humanos. Es la expectativa de ciertas respuestas en determinados escenarios, basándonos en experiencias previas. Moldea nuestro comportamiento alentándonos a elegir situaciones cómodas y previsibles. Esto influye, inconscientemente, en nuestras decisiones sobre la forma de actuar.

Por ejemplo, si tiene que asistir a un banquete, es muy probable que decida sentarse a una mesa con la gente que conoce. Usted se siente cómodo con estas personas y, debido a la experiencia pasada, subconscientemente prevé pasar un rato agradable. De esta manera, su decisión sobre dónde sentarse está influida por su experiencia previa.

Ahora tome el ejemplo de una cirujana que está a punto de rehacer una herniorrafia. Ésta es una intervención quirúrgica compleja sobre una operación de hernia previa. Suponga que hay dos posibles procedimientos para esta intervención y la cirujana es experta en ambos. Ella pide opinión a dos de sus colegas sobre el mejor procedimiento. Uno es un cirujano adjunto con gran experiencia y el otro es un cirujano residente con menos experiencia.

Basándose en sus respectivos niveles de experiencia, el adjunto recomienda el primer procedimiento, en contraste con el residente, que recomienda el segundo. La cirujana está más sesgada (aunque no en detrimento de su paciente) hacia la opinión del cirujano adjunto. Sin embargo, hay un estudio disponible que demuestra que la preferencia del cirujano más experto tiene una tasa más alta de complicaciones si no intervienen otros factores. La indecisa cirujana debe tomar una decisión por el mejor interés de su paciente, pero si decide realizar la operación preferida por el adjunto, estará introduciendo un sesgo en su decisión y poniendo a su paciente en una desventaja potencial.

El sesgo a menudo es subconsciente e involuntario. Sin embargo, debido a que la gente que busca asistencia médica debe confiar en las decisiones de los profesionales, el sesgo puede causar respuestas menos favorables. Aunque la interpretación vulgar de sesgo tiene una connotación negativa próxima a prejuicio, el sesgo no es una decisión consciente de tratar mal a un paciente. Más bien es un acto cómodo basado

en la experiencia previa con pacientes similares. A fin de minimizar el sesgo en la práctica médica, debemos reconocer su existencia y estar dispuestos a aceptar pruebas que puedan modificar nuestra forma de pensar.

Los experimentos se diseñan para recrear la realidad. Tomamos una muestra de individuos y los estudiamos bajo condiciones controladas pero procuramos no alterar la situación real, de modo que cualquier conclusión obtenida de nuestras observaciones sea válida y pueda aplicarse a toda la población. Sin embargo, los experimentos no son situaciones reales; introducen elementos que no reflejan el contexto real con una exactitud del 100%. El sesgo en la investigación es cualquier influencia que convierta un resultado observado en un mal representante del efecto real en estudio. Una vez más, el sesgo en la investigación no es intencionado, pero un estudio sesgado puede desembocar en una conclusión errónea.

- *El sesgo en la estadística es una influencia, posiblemente no reconocida, que introduce un error sistemático en el proceso.*

Ciertos diseños experimentales son más susceptibles de sesgarse que otros, dependiendo del sistema de recogida de los datos. El ensayo aleatorio, ciego y controlado se diseñó para minimizar el sesgo y, por lo tanto, tiene más crédito que otros tipos de diseños. Además de en el diseño, el sesgo puede afectar a un estudio en otras etapas. Puede producirse en cualquier fase: en la idea inicial, en cualquier punto del proceso de selección y recogida de los datos, en el análisis y hasta en la publicación. El sesgo en la investigación es más perjudicial que el sesgo individual porque los resultados publicados pueden llegar a influir en un gran número de profesionales.

A continuación presentamos algunas áreas donde podría identificarse una fuente de sesgo. Cuando criticamos la bibliografía, deberíamos hallar las fuentes potenciales de sesgo y entender que podrían tener un efecto en las conclusiones en las cuales basamos nuestras recomendaciones.

Sesgo de selección, en el cual difieren las poblaciones de los grupos tratado y control. Es más frecuente en los estudios de casos y controles debido a que los casos y los controles son seleccionados y no aleatorizados. Los estudios de cohortes prospectivos, no aleatorizados, también son propensos, ya que los individuos autoseleccionan sus intervenciones y sus decisiones pueden reflejar alguna característica intrínseca de los grupos. En los ensayos donde los individuos son asignados aleatoriamente se minimiza este tipo de sesgo. El sesgo de selección también puede influir en los resultados de un ensayo que no seleccionó la muestra aleatoriamente, ya que el proceso de inferencia que usa distribuciones muestrales se basa en el muestreo aleatorio.

Sesgo de información, en el cual los datos (sobre todo en estudios retrospectivos) se obtienen de fuentes que nos son 100% fiables. Por ejemplo, los certificados de defunción no siempre precisan la causa exacta de la muerte. El *sesgo de memoria* es un tipo de sesgo de información que se da en situaciones donde se pide a los individuos que recuerden la exposición que tenían a ciertas sustancias en el pasado. Los sujetos pueden no recordar exactamente estas exposiciones.

Sesgo de seguimiento, en el cual hay diferencias en el seguimiento de los grupos estudiados o en los métodos de medición de las variables. Todos los tipos de estudios son vulnerables a este sesgo, pero el enmascaramiento (donde pacientes e investigadores desconocen el tratamiento administrado) minimiza este efecto.

Sesgo estadístico, en el cual los resultados se comunican con el fin de promover las creencias de los investigadores. Por extraño que parezca, los datos pueden manipularse para dar resultados diferentes. Una manera de evitarlo es decidir el tipo del análisis antes de que empiece el estudio.

Sesgo de publicación, en el cual las revistas podrían no publicar estudios válidos con resultados negativos, o podrían decidir no publicar estudios que puedan parecer poco importantes en ese momento. En contraste, un estudio de un investigador reconocido podría tener más probabilidades de ser aceptado que el trabajo de un investigador con menos renombre, aunque ambos pudieran tener contribuciones igualmente válidas.

Sesgo de financiación, en el cual un investigador reconocido es más factible que obtenga una subvención que un colega igualmente calificado pero relativamente desconocido. Esto permite dirigir los proyectos a los investigadores más influyentes. Lo cual no es necesariamente perjudicial para el conocimiento médico, ya que suelen ser hábiles diseñando ensayos, y a menudo tendrán un historial productivo. Sin embargo, los investigadores menos experimentados no podrán disfrutar de las mismas oportunidades para lanzar sus proyectos. También es una práctica habitual de las compañías farmacéuticas financiar estudios de investigación. Ésta es una conocida fuente de sesgo potencial: las farmacéuticas quieren que su nuevo medicamento supere al actual. Este entusiasmo podría influir en el diseño del estudio y en la comunicación de los resultados. Igualmente, un investigador podría tener un interés financiero en el resultado de un ensayo. Actualmente es una práctica habitual que un investigador incluya un descargo de responsabilidad sobre sus conflictos de intereses.

Sesgo del lector, en el cual los lectores dan más crédito a los estudios realizados en grandes centros de investigación. Como otros tipos de sesgo, esto no es necesariamente incorrecto, pero debería admitirse. Similar a la cirujana en el ejemplo anterior, si un lector no está seguro de la validez de un estudio, el historial pasado de una fuente de información puede ayudar a guiar sus decisiones. Además, los estudios publicados en idiomas extranjeros (aun disponiendo de la traducción) podrían ser infraestimados respecto a los escritos en la lengua materna del lector.

EXACTITUD Y PRECISIÓN

Usamos las muestras para estimar un parámetro poblacional. Si un estudio ha sido sesgado, podemos no estar apuntando al parámetro y no darnos cuenta. La estimación del parámetro no será correcta. Incluso si el estudio se repite por un defecto en el diseño o en la realización, el dardo se desviará sistemáticamente del objetivo. Este defecto se manifiesta en la exactitud del estudio.

La variabilidad de los datos también afecta a los resultados. Si una variable concreta tiene una gran dispersión en sus valores, será complicado que la estimación apunte al parámetro aunque el diseño del estudio no sea sesgado. El resultado carecerá de precisión y se reflejará en un amplio intervalo de confianza. La exactitud de la estimación no estará afectada, pero no podremos confiar en los resultados de un solo estudio. Cada repetición del estudio dará un resultado diferente, pero múltiples estudios finalmente atinarán con el objetivo. Esto puede corregirse aumentando el tamaño muestral.

Las figuras 15-1 a 15-4 ilustran de qué manera la exactitud y la precisión son importantes en la estimación de parámetros poblacionales. Si un estudio tiene exactitud y precisión, la estimación estará próxima al parámetro y los estudios posteriores mostrarán poca variabilidad.

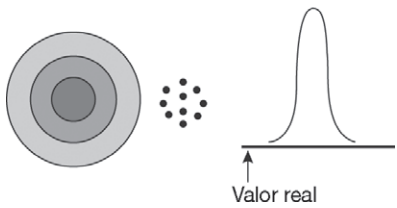


FIGURA 15-1 Alta precisión pero poca exactitud. (De Jekel, J. et al. 2001. *Epidemiology, biostatistics, and preventive medicine*, 2nd ed. Philadelphia: W. B. Saunders Co.)

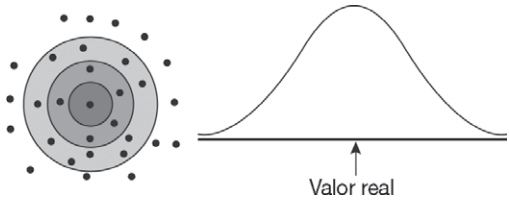


FIGURA 15-2 Alta exactitud pero baja precisión. (De Jekel, J. et al. 2001. *Epidemiology, biostatistics, and preventive medicine*, 2nd ed. Philadelphia: W. B. Saunders Co.)

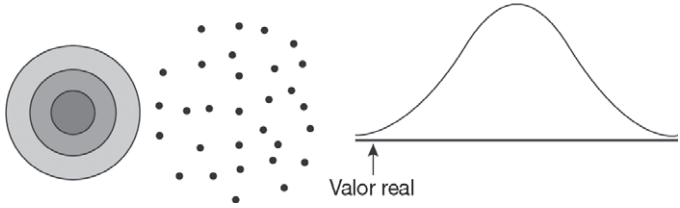


FIGURA 15-3 Ni exactitud ni precisión. (De Jekel, J. et al. 2001. *Epidemiology, biostatistics, and preventive medicine*, 2nd ed. Philadelphia: W. B. Saunders Co.)

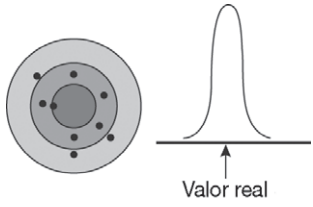


FIGURA 15-4 Alta exactitud y precisión. (De Jekel, J. et al. 2001. *Epidemiology, biostatistics, and preventive medicine*, 2nd ed. Philadelphia: W. B. Saunders Co.)

PUNTOS CLAVE

- El sesgo en la práctica médica podría poner al paciente en desventaja.
- El sesgo en la literatura médica puede ser muy influyente.
- El sesgo en la estadística es una influencia, posiblemente no reconocida, que introduce error en el proceso.
- El sesgo puede estar presente en cualquier fase de un proyecto de investigación, desde su inicio hasta su publicación y en sus lectores.
- Sólo admitiendo que el sesgo existe, y tratando de identificar sus fuentes, podremos minimizarlo.
- El sesgo puede invalidar un estudio por proporcionar resultados no del todo exactos, e incluso si se repite, dará el mismo resultado.
- Un estudio impreciso intenta aproximarse al parámetro, pero cada vez que se repite, apunta en una dirección. Esto es fruto de la alta variabilidad dentro de las muestras.
- La exactitud y la precisión se mejoran minimizando el sesgo y usando métodos estadísticos apropiados, como el muestreo aleatorio y tamaños de muestra apropiados.

BIBLIOGRAFÍA

1. Hooke, R. 1983. *How to tell the liars from the statisticians*. New York: Marcel Dekker, Inc.

PREGUNTAS DE REVISIÓN

1. Los experimentos de investigación intentan estudiar una situación real usando _____ y observando los resultados.

2. Hay muchos motivos por los cuales un estudio puede no recrear exactamente la situación real. Se dice, entonces, que los resultados son _____.

3. El _____ no es necesariamente intencionado.

4. El sesgo afecta a la _____ de un estudio.

5. Cuando hay menos variabilidad en los datos, un estudio será más preciso, y se obtendrá un _____ __ _____ más pequeño.

RESPUESTAS

1. muestras

2. sesgados

3. sesgo

4. validez, exactitud

5. intervalo de confianza

CAPÍTULO 16

Ética en la investigación médica

Los mentirosos son retratados en la gran pantalla y en la ficción como pícaros encantadores, mientras los estadísticos son siempre personas de una sosería infinita. En la vida real, ésta no es la forma como usted describiría a unos y a otros.

—Robert Hooke[1]

Los médicos, en su profesión, siempre se han guiado por un código ético. Los profesionales médicos están en una posición privilegiada en la que toman decisiones que afectan a la salud y el bienestar de un paciente confiado. Damos recomendaciones a la gente que espera que actuemos en pro de sus intereses. Ya desde el siglo IV a. C., el médico griego Hipócrates se percató de la importante responsabilidad moral del médico frente al paciente. Produjo una letanía de documentos dirigidos al principio básico de la conducta ética, «ayudar y no dañar». El Juramento Hipocrático es un producto de esta época que enfatiza la importancia de respetar los derechos éticos del paciente. Todavía hoy, titulados de las facultades de medicina de todo el mundo recitan partes de este juramento en su graduación.

Existe la posibilidad de un abuso de confianza en la relación médico-paciente cuando los profesionales sanitarios sitúan al paciente en un riesgo excesivo de sufrir daños. Uno de los ejemplos más flagrantes de incumplimiento del principio ético básico en el tratamiento de personas fue la brutal tortura que los nazis infligieron a las víctimas de los campos de concentración durante la Segunda Guerra Mundial. Estas acciones atroces incluían «experimentos» hipotérmicos en los cuales se sumergía a las personas en tanques de agua helada y se les dejaba morir. A otros se les quitaba el oxígeno y eran lentamente asfixiados. Algunos presos fueron deliberadamente infectados por el cólera o sufrieron absurdas intervenciones de trasplante de órganos o de esterilización. Estas espantosas intervenciones se realizaron en víctimas cautivas contra su voluntad y bajo el pretexto de la «experimentación médica».

Al hacerse públicas estas acciones inhumanas, hubo una enérgica protesta internacional para que se instauraran directrices que guiasen la ética en el tratamiento de las personas en la experimentación médica. El Tribunal Militar de Nuremberg, que investigó estos crímenes de guerra, finalmente condenó a ejecución a los médicos responsables y confeccionó una lista de condiciones que trazan las líneas maestras de

una ética aceptable en estas circunstancias. Se dio a conocer como Código de Nuremberg (cuadro 16-1). Cuando se hizo público en 1947, el mundo se introdujo en la era moderna de la ética médica. Proporcionó las expectativas básicas que debían cumplirse al realizar experimentos con humanos, haciéndose eco de la filosofía hipocrática de «no dañar» y acentuando la importancia del consentimiento informado de las personas.

CUADRO 16-1

Código de Nuremberg

Experimentos médicos permitidos

Son abrumadoras las pruebas que demuestran que algunos tipos de experimentos médicos en seres humanos, cuando se mantienen dentro de límites bien definidos, satisfacen —generalmente— la ética de la profesión médica. Los protagonistas de la práctica de experimentos en humanos justifican sus puntos de vista basándose en que dan resultados provechosos para la sociedad, que no pueden ser procurados mediante otros métodos de estudio. Sin embargo, se está de acuerdo en que deben conservarse ciertos principios básicos para poder satisfacer conceptos morales, éticos y legales:

1. *El consentimiento voluntario del sujeto humano es absolutamente esencial. Esto quiere decir que la persona implicada debe tener capacidad legal para dar su consentimiento; que debe estar en una situación tal que pueda ejercer su libertad de escoger, sin la intervención de cualquier elemento de fuerza, fraude, engaño, coacción o algún otro factor coercitivo o coactivo, y que debe tener el suficiente conocimiento y comprensión del asunto en sus distintos aspectos para que pueda tomar una decisión consciente. Esto último requiere que antes de aceptar una decisión afirmativa del sujeto que va a ser sometido al experimento, hay que explicarle la naturaleza, duración y propósito del mismo, el método y las formas mediante las cuales se llevará a cabo, todos los inconvenientes y riesgos que pueden presentarse, y los efectos sobre su salud o persona que puedan derivarse de su participación en el experimento. El deber y la responsabilidad de determinar la calidad del consentimiento recaen en la persona que inicia, dirige o implica a otro en el experimento. Es un deber personal y una responsabilidad que no puede ser delegada con impunidad a otra persona.*
2. *El experimento debe realizarse con la finalidad de obtener resultados fructíferos para el bien de la sociedad que no sean asequibles mediante otros métodos o medios de estudio, y no debe ser de naturaleza aleatoria o innecesaria.*
3. *El experimento debe diseñarse y basarse en los resultados obtenidos mediante la experimentación previa con animales y el pleno conocimiento de la historia natural de la enfermedad o del problema en estudio, de manera que los resultados anticipados justifiquen la realización del experimento.*
4. *El experimento debe guiarse de tal forma que evite todo sufrimiento o daño innecesario físico o mental.*
5. *No debe realizarse experimento alguno si hay una razón* a priori *para suponer que puede producirse alguna muerte o lesión irreparable; excepto, quizá, en los experimentos en los que los médicos investigadores son también sujetos de experimentación.*

Continúa

CUADRO 16-1 *(continuación)*

6. *El riesgo asumido no debe nunca exceder el determinado por la importancia humanitaria del problema que ha de resolver el experimento.*
7. *Se deben tomar las precauciones adecuadas y disponer de las instalaciones óptimas para proteger al individuo implicado de las posibilidades de lesión, incapacidad o muerte, aunque sean remotas.*
8. *El experimento debe ser dirigido únicamente por personas científicamente cualificadas. En todas las fases del experimento se requiere la máxima precaución y capacidad técnica de los que lo dirigen o forman parte del mismo.*
9. *Durante el transcurso del experimento, el sujeto debe tener la libertad de finalizarlo si llega a un estado físico o mental en el que le parece imposible continuar el experimento.*
10. *En cualquier momento durante el transcurso del experimento, el científico que lo realiza debe estar preparado para interrumpirlo si tiene razones para creer —en el ejercicio de su buena fe, habilidad técnica y juicio cuidadoso— que la continuación del experimento puede provocar lesión, incapacidad o muerte al sujeto en experimentación.*

(De Trials of War Criminals before the Nuremberg Military Tribunals under Control Council Law No. 10. Nuremberg, October 1946-April 1949. Washington, D. C.: U. S. Government Printing Office, 1949-4953.)

La comunidad médica internacional también reaccionó ante estas atrocidades desarrollando un conjunto de directrices que asegurasen la realización de la experimentación bajo unos estándares éticos aceptables. Este documento se publicó inicialmente en Helsinki (Finlandia) en 1964 durante una convención internacional de la World Medical Association. Actualmente se conoce como la Declaración de Helsinki, aunque desde entonces se haya sometido a varias revisiones. Contiene muchos de los principios establecidos en el Código de Nuremberg, pero además distingue entre la investigación terapéutica y no terapéutica.

A pesar del reconocimiento formal y de la aprobación de los principios éticos expuestos en el Código de Nuremberg y la Declaración de Helsinki, se han producido múltiples incidentes en Estados Unidos en los que los experimentos médicos han usado estándares éticos cuestionables en su diseño. Uno de los más famosos fue el ensayo patrocinado por el gobierno federal sobre hombres afroamericanos que habían contraído la sífilis. El estudio comenzó en Tuskegee (Alabama) en 1932. Era un ensayo observacional que estudió a más de 600 hombres durante varias décadas para hallar los efectos fisiológicos de la sífilis. La mayoría eran pobres y analfabetos, y no se les informó de su enfermedad. Además, no se les ofreció el tratamiento una vez que estuvo disponible, un par de décadas más tarde del inicio del estudio.

El infame estudio de Tuskegee fue uno de los catalizadores para el desarrollo de pautas en la experimentación humana promovido por el gobierno federal estadounidense. En 1974 se aprovó la National Research Act (Acta Nacional de Investigación) estableciendo la National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (Comisión Nacional para la Protección de los Humanos en la Investigación Biomédica y del Comportamiento). Durante los siguientes años, la National Commission publicó recomendaciones que perfilaron los principios éticos del tratamiento en humanos. Uno de sus informes se ha convertido en la piedra angular de la ética

experimental. Se conoce como el *Informe de Belmont*, que nació después de la Belmont Conference Center en la Smithsonian Institution. Este informe promulga tres principios filosóficos básicos en los cuales deberían basarse los estándares éticos:

1. Mantener el *Respeto por las personas*. Este principio protege la dignidad de las personas y también incluye consideraciones sobre su autonomía y su derecho a tomar decisiones informadas en las intervenciones que condicionan su destino. Este principio enfatiza la necesidad del consentimiento informado. Los individuos deberían entender los riesgos y beneficios potenciales de participar en un proyecto de investigación y deberían estar dispuestos a participar.
2. Considerar la *Beneficencia*. Garantiza que la salud del individuo tenga prioridad sobre la hipótesis de investigación. Deberían tenerse en cuenta todos los posibles tipos de daño y evaluarse sistemáticamente. El riesgo de inclusión debería sopesarse con el beneficio potencial del individuo por participar y con el beneficio social del conocimiento adquirido durante el estudio.
3. Mantener la *Justicia*. Ningún grupo debería seleccionarse injustamente por motivos raciales, por su sexo o por su nivel socioeconómico. Por otra parte, cuando se percibe que la participación ofrece un beneficio significativo al individuo, todos los grupos elegibles deberían ser invitados a participar.

La National Commission requiere que estos principios fundamentales se apliquen en todos los tipos de diseños de investigación y en la realización de los experimentos. El órgano gubernamental que garantiza estos principios es el Institutional Review Board (IRB). Éste es un comité local específico para cada institución; está compuesto por un grupo diverso de miembros médicos y no médicos, que a menudo incluye a un clérigo o a otras personas con formación en ética médica. El IRB evalúa la investigación sistemáticamente, desde su inicio hasta su finalización. Se considera la ética de los proyectos en relación con los principios expuestos en el *Informe de Belmont*, y tiene la potestad de interrumpir la investigación si los miembros creen que hay una transgresión de cualquiera de estos principios.

Cuando evalúe la bibliografía, penetrará en algunas áreas oscuras donde la respuesta a si se ha seguido el procedimiento ético apropiado no es evidente. A menudo, las personas cautivas, como los presos o los pacientes hospitalarios, se escogen como población objetivo. Incluso si son capaces de dar el consentimiento informado, ¿no se está quebrantando el principio de justicia? ¿No es posible que piensen que conseguirán un trato mejor si participan, y por tanto, que se sientan presionados? Estos ejemplos ilustran por qué actualmente hay debates sobre principios éticos y por qué es crucial la sabiduría colectiva del IRB. Existen múltiples opiniones sobre estas cuestiones, y nos conviene entretenernos en ellas. La ética de la experimentación médica se basa en unos sanos principios filosóficos que concentran la sabiduría e imparcialidad del adagio «Trate a todo el mundo como usted querría ser tratado». Cuando se enfrente a una cuestión ética sin una respuesta fácil, pregúntese cómo se sentiría si usted fuese el sujeto. Su respuesta debería guiarle en su decisión.

PUNTOS CLAVE

- La práctica y la experimentación médica se guían por un código ético.
- Pioneros en el estudio de la ética médica, como Hipócrates, adoptaron la máxima de «ayudar y no dañar», que aún hoy subraya la filosofía de la ética médica moderna.

- Las acciones atroces contra personas realizadas bajo el pretexto de la «experimentación médica» durante la Segunda Guerra Mundial incitaron al Tribunal Militar Internacional en Nuremberg a redactar un código ético (Código de Nuremberg). Se resalta, entre otros, el concepto del consentimiento informado.
- La World Medical Association ideó un código ético para la experimentación médica: la Declaración de Helsinki. Respalda muchos conceptos del Código de Nuremberg y distingue entre investigación terapéutica y no terapéutica.
- El gobierno estadounidense también desarrolló pautas para la experimentación humana con la creación de la National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. En su *Informe de Belmont*, identifican tres principios éticos en la experimentación: Respeto por las personas, Beneficencia y Justicia.
- El Institutional Review Board (IRB) supervisa la investigación institucional para garantizar que no se quebrante la ética médica.

BIBLIOGRAFÍA

1. Hooke, R. 1983. *How to tell the liars from the statisticians*. New York: Marcel Dekker, Inc.

PREGUNTAS DE REVISIÓN

1. El _____ _____ _____ es un comité que supervisa la investigación médica en las instituciones.

2. El _____ __ _____ es un código ético ideado por el Tribunal Militar Internacional en Nuremberg después de la Segunda Guerra Mundial en respuesta a las atrocidades de los campos de concentración.

3. El estudio de la sífilis en Tuskegee no obtuvo el _____ _____ de sus participantes, con lo que se infringió el principio de Respeto a las Personas mencionado en el *Informe de Belmont*.

4. El principio de _____ que aparece en el *Informe de Belmont* garantiza que ningún grupo será injustamente seleccionado en la investigación médica.

5. El principio de Beneficencia cuida del _____ de los individuos.

RESPUESTAS

1. Institutional Review Board

2. Código de Nuremberg

3. consentimiento informado

4. Justicia

5. bienestar, la salud

CAPÍTULO 17

Tipos de estudios de investigación

Los estudios científicos persiguen la verdad. En la evaluación de un estudio, la consideración inicial consiste en si el estudio alcanzó esta verdad con éxito.

—D. J. Friedland[1]

¡Felicidades! Usted ya tiene el mapa y las habilidades necesarias para navegar por la bibliografía. Cuando lea detenidamente las revistas y otras fuentes de información, verá que hay varios tipos de estudios, cada uno con sus ventajas e inconvenientes. Por ejemplo, el ensayo clínico controlado aleatorizado y enmascarado (EC) se considera el patrón de oro en la investigación clínica. Está expresamente diseñado para minimizar las fuentes potenciales de sesgo. Con la potencia suficiente y correctamente ejecutado, tiene tanto exactitud como precisión.

Lamentablemente, por varios motivos, no todos los ensayos pueden ser EC. Han de cumplirse ciertos criterios, como poder enmascarar a los participantes sobre la intervención asignada. Si desconocen el tratamiento al cual están expuestos, no habrá ninguna expectativa subconsciente que pueda influir potencialmente en los resultados. Sin embargo, algunos tipos de experimentos no son enmascarables. Por ejemplo, en los estudios que comparan intervenciones quirúrgicas con tratamientos médicos, ni los pacientes ni los investigadores pueden estar enmascarados.

Los estándares éticos también deben cumplirse en los ensayos clínicos. Conviene que todos los investigadores inicien cualquier estudio con *equipoise* (v. cap. 10), de modo que se considere en el análisis tanto la posibilidad de que el tratamiento sea beneficioso como potencialmente dañino. Después de todo, el estudio se realiza porque a menudo no sabemos si una intervención proporciona un beneficio o alberga posibles efectos deletéreos. Los participantes deben conocer la posibilidad de sufrir algún perjuicio, aunque sea mínimo.

Aunque un EC tenga la propiedad de producir resultados más fiables, hay muchos otros tipos de experimentos que ocupan un lugar importante en la literatura clínica. No todos los ensayos pueden ser un EC, ni tampoco tienen que serlo. Los estudios observacionales, en los cuales los individuos eligen su propio curso y son seguidos en el tiempo, pueden contribuir enormemente al fundamento del conocimiento médico. La mayoría de estudios epidemiológicos se realizan a través de esta vía, como los

famosos estudios de Framingham iniciados en 1948[2]. Esta base de datos registró, entre otras cuestiones, los hábitos de estilo de vida de miles de individuos y luego examinaron su subsecuente tasa de enfermedad cardiovascular. Esta inestimable información nos ha proporcionado el conocimiento que tenemos actualmente sobre factores de riesgo cardiovascular.

INFORMES DE CASOS CLÍNICOS

¿Qué incita a un individuo a perseguir una particular pregunta de investigación? El primer indicio sobre una asociación entre unas variables concretas puede comunicarse en un simple, pero estimulante, informe de casos. Cuando un clínico ve algo fuera de lo normal, comunicarlo en una revista alerta a otros profesionales sobre una posible conexión. Por ejemplo, en una conferencia sobre dermatología en San Francisco, a principios de los años ochenta, se comunicó la existencia de varios casos de hombres homosexuales con un extraño tipo de cáncer de piel llamado «sarcoma de Kaposi»[3]. ¿Había alguna conexión entre el cáncer de piel y la homosexualidad? Esta observación conllevó una investigación que resultó en la identificación de un virus que ha provocado una gran epidemia con inmensas ramificaciones, conocida como el síndrome de inmunodeficiencia adquirida (sida).

ESTUDIOS OBSERVACIONALES

Una vez que se ha identificado una posible conexión entre variables gracias a los informes de casos, se realizan estudios observacionales para intentar dilucidar la presencia y la fuerza de esta asociación. Los diagramas de Venn son una basta forma de ilustrar conexiones entre variables, pero tienen poca validez científica. La correlación es una manera de expresar la fuerza y la dirección de la asociación. El coeficiente de correlación nos dice no sólo si hay una asociación significativa, sino también el grado de influencia de una variable sobre otra y la dirección (positiva o negativa). La correlación *no* nos dice si una variable *causó* el efecto en la otra, sólo que hay una relación. Cuando la asociación se mide en un instante de tiempo (en vez de a lo largo del tiempo), hablamos de un estudio transversal.

Tenga presente que dos variables pueden estar conectadas a través de una tercera variable «escondida». Es lo que pasaba con los homosexuales y el sarcoma de Kaposi. Finalmente se identificó la conexión como el virus de inmunodeficiencia humana (VIH). Si nos hubiéramos conformado con asumir que no había ninguna otra conexión entre la homosexualidad y el cáncer de piel, y no se hubiera buscado más, habríamos perdido esta información. Estas variables escondidas se llaman *confusoras*; explican una parte principal de la asociación entre otras variables por su conexión independiente con cada una de ellas. ¿Rinden los niños afroamericanos en la escuela menos que los caucásicos debido a su color de piel? La mayoría de expertos estarían de acuerdo en que, a pesar de que se puede demostrar una asociación entre el color de la piel y el rendimiento escolar, la conexión tiene que ver con variables confusoras como la clase social.

ESTUDIOS DE CASOS Y CONTROLES

Los estudios observacionales también pueden hacerse retrospectivamente. Son los estudios de casos y controles, en los que las respuestas ya se conocen. Tomamos a las personas enfermas e intentamos emparejarlas según sus características con equivalentes sanos. Entonces retrocedemos para ver si hay una correlación entre una variable particular y una respuesta adversa. De ser así, deducimos que si quitamos aquella variable podremos mejorar la respuesta. Por ejemplo, un reciente estudio retrospectivo observó una muestra de una población de individuos con ataques cardíacos. Manteniendo los otros factores inalterables, ¿influyó negativamente la anemia en la mortalidad? El análisis demostró que aquellos con anemia grave evolucionaron peor que aquellos con concentraciones sanguíneas normales. Esto respalda el uso de la transfusión de sangre en aquellos pacientes con ataques cardíacos y anemia severa[4].

Me gusta pensar en los estudios de casos y controles como personas que se bajan de un tren en su destino. Usted se los encuentra en la estación Enfermedad y mira en sus maletas para ver qué artículos han llevado con ellos. Después, va a la estación Salud y mira lo que habían empaquetado los individuos de esta estación. La diferencia entre sus equipajes podría explicar su destino. Por ejemplo, la gente que llega a la estación Enfermedad de Arteria Coronaria tendrá una mayor tendencia a llevar cigarrillos, huevos con bacón e insulina para su diabetes. También es posible que lleven los certificados de defunción de sus padres y hermanos que murieron por enfermedad cardíaca. Aquellos que se bajan en la estación Salud tendrán más ropa de *footing* y sándwiches de pavo. Habrá alguna coincidencia de artículos, pero en general verá diferencias en los equipajes de los dos destinos.

Los estudios retrospectivos son sobre todo aplicables a respuestas relativamente raras o que tardan mucho en manifestarse, ya que este tipo de estudio puede realizarse sin esperar a que los individuos desarrollen la enfermedad. Una de las limitaciones de este tipo de estudio radica en encontrar un emparejamiento apropiado en el grupo control. Tienen que ser similares al grupo enfermo en todos los aspectos (excepto en la enfermedad y en la variable en estudio). Si hay alguna otra diferencia no considerada, una confusora no identificada podría ser la responsable del resultado. Esto podría llevarnos a una conclusión incorrecta respecto a la otra variable en estudio.

Los estudios retrospectivos también son propensos al sesgo de memoria cuando confían en el recuerdo que los individuos tienen sobre sus exposiciones a sustancias. Los enfermos podrían tener una memoria más precisa de sus exposiciones, pero, por otra parte, podrían sobrestimar una exposición que la crean causante de la enfermedad. Además, existen algunas limitaciones matemáticas. El cálculo del riesgo implica el desarrollo de la enfermedad a lo largo del tiempo, con un punto de partida común tanto para el grupo de enfermos como para el de no enfermos. Debido a que retrocedemos en el tiempo, no tenemos ningún punto de partida real. Por eso es mejor calcular *odds ratios* (medida de asociación) que un riesgo (que refleja una tasa).

ESTUDIOS DE COHORTES

Cuando tenemos que esperar un tiempo a que se desarrolle la respuesta, puede realizarse un estudio prospectivo que vaya hacia delante en el tiempo. Los estudios de cohortes comienzan con un conjunto de individuos y miden la incidencia relativa

de la enfermedad en los expuestos a una variable en comparación con los no expuestos. Los grupos en un estudio prospectivo no necesariamente están aleatorizados o enmascarados, ya que pueden elegir sus exposiciones. El estudio de Framingham es un buen ejemplo de diseño de cohortes prospectivo. Como la exposición precedió al resultado, hay un argumento más fuerte en favor de que la exposición causó la enfermedad, aunque esto sea insuficiente para demostrar causalidad. Hay un punto de partida en el tiempo; por tanto, el riesgo puede calcularse. El sesgo de memoria se minimiza en este diseño. Sin embargo, el gasto y el tiempo prolongado que requieren estos estudios para llevarlos a cabo (a la espera de la respuesta observada) son un inconveniente.

ENSAYOS CONTROLADOS ALEATORIZADOS

Hemos llegado al prototipo de la medicina basada en la evidencia, el ensayo controlado aleatorizado enmascarado (EC). El sesgo tiene muchas oportunidades para aparecer sigilosamente en los estudios y causar una conclusión final errónea. El artificioso EC, aunque no garantiza ser completamente insesgado, elimina muchos tipos de sesgos que se entremeten en otros diseños. Este estudio es parecido a un estudio de cohortes prospectivo, pero con algunas distinciones importantes.

Cuando las personas dan su consentimiento informado, son asignadas aleatoriamente a una de las intervenciones. Si se permitiese a los individuos elegir su intervención, como en un estudio observacional, se introducirían otros factores que podrían explicar la respuesta. Considere un ensayo sobre el cáncer que permita a los pacientes elegir entre el medicamento estándar o uno nuevo. Dentro de esta muestra de pacientes con cáncer, los más sanos podrían decidir, hasta subconscientemente, adherirse al tratamiento estándar, mientras que los más desesperados podrían arriesgarse con el nuevo tratamiento. El punto que se debe enfatizar aquí es que, en este caso, el tratamiento estándar está destinado a tener un mejor resultado, aun si no es mejor, porque *los grupos no son iguales*. La aleatorización elimina la posibilidad de que un tratamiento pueda mostrar un beneficio erróneo por este tipo de sesgo de selección. Si las confusoras potenciales están presentes (aun sin ser identificadas), deberían estar distribuidas equitativamente entre los grupos y, de este modo, no propiciar una diferencia en las respuestas.

Es propio de la naturaleza humana querer complacer a los demás. Cuando las personas participan en un ensayo, puede haber un deseo subconsciente por parte de los investigadores de justificar los esfuerzos y proporcionar esperanza a las futuras víctimas de la misma enfermedad. Esta admirable cualidad puede ser deletérea al proceso de la investigación científica; sin embargo, produce conclusiones erróneas sobre el beneficio real de una intervención concreta. Si los participantes o los investigadores conocen la intervención asignada, puede haber una tendencia a minimizar los síntomas comunicados. A través de mecanismos de *feedback* biológico, es posible cambiar psicológicamente procesos medibles como la presión arterial. Esto puede provocar un beneficio falsamente magnificado en favor de un tratamiento particular. En un ensayo doblemente enmascarado (doble ciego), se elimina este tipo de sesgo porque tanto los investigadores como los participantes desconocen el tratamiento que están administrando y recibiendo, respectivamente. Lo normal es que todos los grupos reciban el mismo número y tipo de intervenciones. Por ejemplo, si se compara un fármaco oral con uno intravenoso, ambos grupos recibirán una píldora y

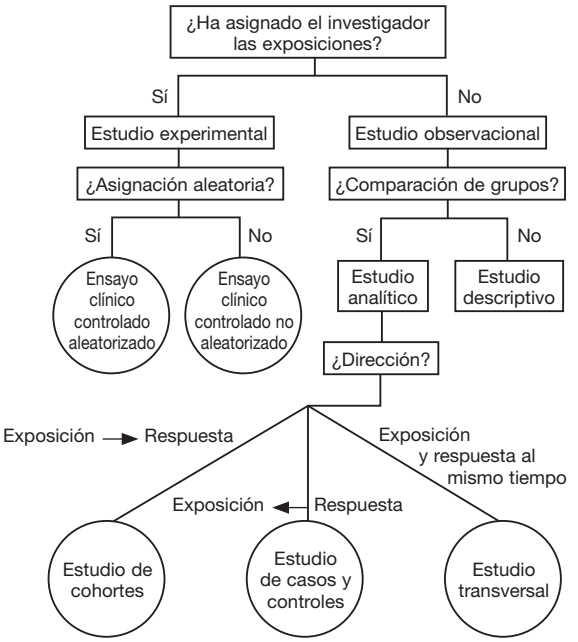


FIGURA 17-1 Algoritmo para la clasificación de tipos de investigación clínica. (De Grimes, D. A. et al. 2002. An overview of clinical research: The lay of the land. *Lancet*, 359:9300, 58, con autorización.)

una inyección. Nadie sabe cuál es la sustancia inocua. Algunos placebos pasan por un proceso de diseño riguroso que los empareja a sus equivalentes activos en tamaño, color y consistencia.

Uno de los inconvenientes de un EC, aunque esté bien ejecutado y sea científicamente sólido (esto se conoce por validez interna), es el proceso de selección. Todos los participantes son voluntarios que quieren y son capaces de contribuir al proceso del conocimiento científico. Pueden tener más formación y ser más sanos en general que los que no pueden participar o deciden no hacerlo. Los sujetos pueden no ser representativos de todos los miembros de la comunidad que cumplen los mismos criterios de selección. Por tanto, los resultados de un EC no siempre pueden tener validez externa o ser extrapolables a toda la comunidad.

El esquema de la figura 17-1 es una forma de clasificar el tipo de investigación hallada en la bibliografía. El diseño del estudio es una llave del tipo de conclusiones que pueden dibujarse, así como de la fuerza de la evidencia. Estudios de revisión, como los metaanálisis, no se mencionan expresamente ya que algunos incluso los consideran un tipo de estudio observacional retrospectivo o transversal.

FACTORES EN EL ANÁLISIS FINAL

Los estadísticos tienen que explicar los factores que pueden complicar el resultado final en un ensayo clínico. En ensayos no enmascarados, algunos pacientes pasan al otro tratamiento por la recomendación de sus médicos o por su propio deseo. Hay formas de tener en cuenta estos cambios en el análisis final de los datos, siempre que se reconozcan. El método más fiable es el análisis por «intención de tratar», que su-

pone que los participantes siguieron el tratamiento inicialmente asignado, con independencia de lo que hicieran después.

Por varios motivos, también habrá abandonos inevitables a lo largo del seguimiento. Es importante considerar este hecho en la estimación del tamaño muestral. Tampoco debe subestimarse la razón por la cual los sujetos abandonan. La intervención podría tener un efecto secundario adverso tan extremo que fuese insoportable. Sería importante que los investigadores supieran esto. Salvo la retirada de un estudio, un individuo puede seguir participando en el ensayo incluso sin cumplir con el protocolo del tratamiento. Esto puede pasar por varios motivos, incluyendo efectos secundarios molestos. Los estudios más audaces incorporarán un sistema de control sobre dicho incumplimiento.

Hay muchos casos en que el efecto de una intervención será mayor en una parte de la muestra. Por ejemplo, un descenso de colesterol sérico puede tener un mayor efecto en la respuesta en un paciente hipertenso que en un normotenso. Es posible estudiar un efecto en muchos subconjuntos de la muestra realizando *análisis por subgrupos*. Es un modo de extraer la máxima información del diseño de un estudio. Es importante tener una muestra bastante grande para que todos los subgrupos tengan un número suficiente de sujetos para realizar un análisis por subgrupos significativo. Si un subgrupo contiene sólo unos pocos individuos, la potencia puede ser insuficiente para hallar un beneficio del tratamiento en el subgrupo, aunque exista. Esto puede tenerse en cuenta en la fase de planificación de los estudios en el cálculo del tamaño muestral. O bien los diseñadores pueden ponderar el reclutamiento de sujetos de modo que cada subgrupo tenga la representación adecuada.

ARTÍCULOS DE REVISIÓN

Hay varias situaciones en las cuales podríamos tener que leer detenidamente una colección de artículos de un tema particular. Por ejemplo, las enfermedades frecuentes que tienen un gran impacto en la asistencia médica general, como el asma o el paro cardíaco, han sido extensamente estudiadas. Podemos encontrar varios grandes ensayos sobre el mismo problema. Normalmente, los administradores se enfrentan con problemas en la política sanitaria que afectan a grandes poblaciones y que consumen una parte importante de los recursos asistenciales. Sus decisiones sobre la organización y los gastos asistenciales deberían basarse en la información colectiva disponible más reciente.

Los artículos de revisión ya han hecho este trabajo por nosotros. Los autores de un artículo de revisión deberían informar al lector del método de búsqueda usado para localizar los artículos y el proceso de toma de decisiones empleado para considerar qué artículos se incluían en la revisión. Las poblaciones no tienen que encajar exactamente, ya que los artículos no combinan los datos; los autores sólo reúnen y comunican los resultados de varios estudios con intereses similares. Tenga presente que la búsqueda puede haberse realizado de un modo un poco diferente a como usted la hubiera planteado. Delegar todo ello a otro podría haber resultado en artículos ligeramente diferentes.

Varios organismos, como la Cochrane Collaboration (www.cochrane.org/reviews) y la United Health Foundation (www.clinicalevidence.org), realizan un servicio de revisión formal para nuestra conveniencia. Como emplean un método científico estricto en la revisión, las llaman revisiones sistemáticas. Todos los artículos se some-

ten a un proceso de cribado riguroso que busca defectos potenciales en el diseño y en las conclusiones, incluyendo sólo los artículos más válidos. Este enfoque realza la fiabilidad del artículo de revisión minimizando el sesgo que podría encontrarse si un individuo realizara la revisión.

METAANÁLISIS

La palabra *metaanálisis* significa «después del análisis». Este tipo de estudio integra los resultados de varios estudios previos similares y analiza la información combinada. Es el más aplicable en el ámbito de la administración de la asistencia médica, ya que evalúa la utilidad de las intervenciones en la configuración de pautas y fijación de políticas sanitarias. A través de integrar varios estudios epidemiológicos, también puede señalar con exactitud el riesgo real de exposiciones ambientales y del desarrollo de enfermedades. El objetivo de un metaanálisis es conseguir la respuesta de una vez para siempre, sobre todo en situaciones donde los datos parecen aportar conclusiones contradictorias.

Como en cualquier otro tipo de estudio, un metaanálisis correctamente ejecutado comienza con una hipótesis de investigación. La población se identifica en función de unos criterios de selección. No obstante, en vez de obtener una muestra de individuos de la población, los autores hacen una búsqueda de la evidencia existente para hallar estudios con poblaciones similares. Generalmente se prefieren los datos procedentes de ensayos clínicos controlados aleatorizados (ECA) que de estudios observacionales, aunque los ECA disponibles puedan tener limitaciones que impedirían su uso exclusivo en un metaanálisis. Los datos se combinan con técnicas estadísticas sofisticadas que observan las respuestas de estos ensayos individuales e integran los resultados. El resultado total se conoce como *medida resumen*. Debido a que estos ensayos serán de diferentes tamaños, cada resultado se ponderará según su contribución a la medida resumen.

Este método puede ser muy potente debido al gran número de sujetos obtenidos al combinar los estudios. Hemos visto que la potencia de un estudio (la capacidad de descubrir un beneficio del tratamiento si realmente existe) está condicionada por el número de participantes. Un aumento del tamaño muestral separará metafóricamente las curvas acampanadas y moverá nuestro resultado de una zona gris a una conclusión en blanco y negro.

Los tamaños muestrales más grandes también proporcionan una estimación más exacta del parámetro poblacional porque tienen menos variabilidad que muestras más pequeñas de la misma población. En el capítulo 13 vimos, matemáticamente, que el intervalo de confianza se estrechaba para mayores tamaños de muestra. Los metaanálisis también poseen la ventaja de tener menos sesgo de selección. Los estudios combinados incluirán probablemente una muestra representativa de la población y, por tanto, se aumentará la validez externa de los resultados.

La figura 17-2 muestra los resultados de un metaanálisis que combinó los datos de estudios sobre la exposición al humo ambiental del tabaco y el cáncer de pulmón. La medida del efecto —el resultado global de los nueve estudios— es la *odds ratio*. Se contestó a la pregunta: «En una persona que desarrolla el cáncer de pulmón, ¿cuál es la probabilidad de que estuviera expuesto al humo ambiental del tabaco?». Este gráfico muestra que la medida resumen concuerda con los estudios individuales y que el intervalo de confianza se ha estrechado notablemente al combinar los datos.

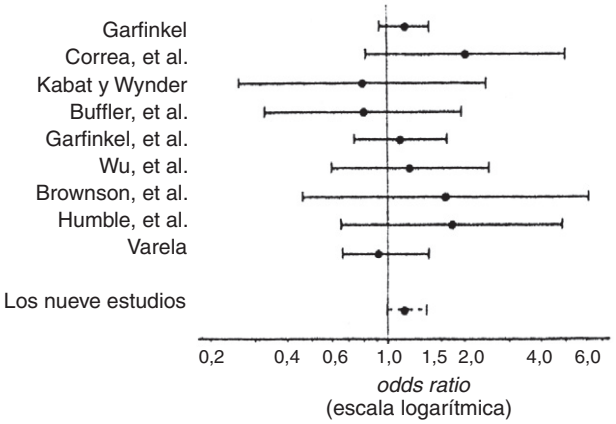


FIGURA 17-2 Metaanálisis: *odds ratios* y sus intervalos de confianza del 95% para nueve estudios epidemiológicos estadounidenses sobre la supuesta asociación entre la exposición al humo ambiental del tabaco y el cáncer de pulmón. (De Fleiss, J. L. y A. J. Gross. 1991. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: A critique. *J Clin Epidemiol,* 44:2, 127-139, con autorización.)

Tenemos una confianza del 95% de que este intervalo contiene la *odds ratio* poblacional verdadera.

Este tipo de análisis es atractivo debido a su fuerza científica. Además, es relativamente barato ya que el trabajo lo realizan básicamente los ordenadores y hay más bien poco trabajo de campo. Aunque pueda parecerlo, combinar los datos no es trivial. Existen dificultades inherentes a la situación concreta de hacer un estudio basándonos en pruebas preexistentes. Cada fase debe ejecutarse cuidadosamente, empezando por la búsqueda de bibliografía. Puede ser un desafío únicamente identificar y cribar apropiadamente los artículos pertinentes. Discrepancias en el proceso de selección pueden convertir a algunos estudios en inelegibles debido a que no representan la población en estudio del metaanálisis. También deben tenerse en cuenta los diferentes tamaños muestrales y metodologías al combinar estudios. Las medidas de la respuesta pueden ser diferentes en cada estudio. En resumen, el metaanálisis puede ser una herramienta potente, pero la metodología es complicada y requiere técnicas estadísticas sofisticadas. Hoy en día aún se sigue estudiando el enfoque adecuado para este complejo proceso[5-7].

DIAGRAMAS DE EMBUDO

Es posible que algunos artículos relevantes para la hipótesis de investigación de un metaanálisis no se hayan publicado debido a sus resultados negativos. Se puede intentar identificar el sesgo de publicación existente en la bibliografía usando un diagrama de embudo. La idea es que los estudios más pequeños tendrán, lógicamente, mayor variabilidad en sus resultados que estudios más grandes (principio de potencia).

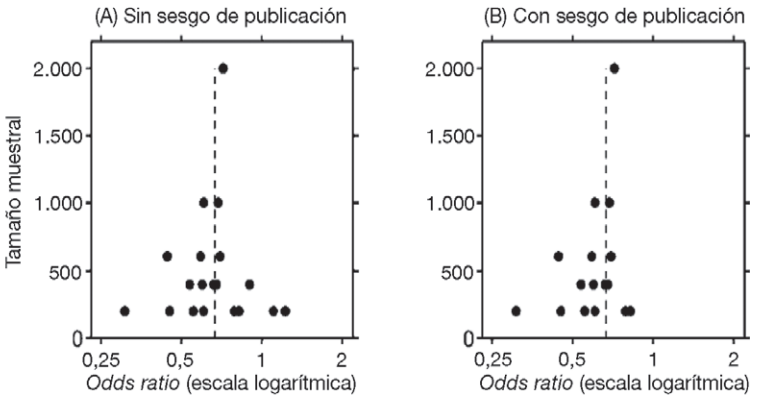


FIGURA 17-3 Se representan los resultados de los estudios en función de su tamaño muestral. El gráfico A debería tener una forma de embudo si no existe sesgo de publicación; en el gráfico B, sólo los estudios más pequeños muestran resultados favorables, lo que sugiere un sesgo de publicación. Los estudios más grandes también se aproximan al parámetro poblacional real de forma más precisa que los estudios más pequeños, que, como era de esperar, tienen una mayor variabilidad en los resultados. (De Dwamena, B. 2004. *Metagraphiti by Stata*. 3rd North American Stata Users Group Meeting, August, University of Michigan, con autorización.)

Si varios estudios más pequeños tienden a comunicar resultados favorables de manera aplastante, esto apoya la premisa de que algunos estudios pequeños con resultados menos convincentes no se habrán examinado detalladamente, dando lugar a un sesgo de publicación. Se debería ampliar la búsqueda de artículos a revistas menos conocidas y a cualquier resultado no publicado que pudiera ajustarse a la hipótesis de investigación. En la figura 17-3 se muestra un ejemplo de diagrama de embudo.

REALIZACIÓN DE LA BÚSQUEDA BIBLIOGRÁFICA

Cuando nos encontramos con una situación que requiere un análisis pormenorizado de la bibliografía, tenemos varias opciones. Podemos hacer una búsqueda nosotros mismos o pedir ayuda a un bibliotecario. Se dispone de varios mecanismos de búsqueda para seleccionar e imprimir los artículos apropiados sobre un tema particular. Se puede acceder a varias bases de datos por Internet, incluso a sitios web como la Cochrane Library (www.cochrane.org) o la base de datos Ovid (www.ovid.com). Otro magnífico sistema de búsqueda es la National Library of Medicine's PubMed (www.ncbi.nlm.nih.gov/pubmed). La mayoría de mecanismos de búsqueda buscan frases o palabras que pueden aplicarse a una guía de términos médicos. Los resultados obtenidos dependerán de las palabras clave introducidas. La búsqueda puede realizarse cruzando términos con operadores boleanos tales como «AND», «OR» o «NOT». Puede que tenga que ampliar su búsqueda si los resultados originales fueran escasos, o reducirla si fueran excesivos. El sitio web PubMed incluye un tutorial básico que le guiará en este proceso.

CUADRO 17-1

Cómo se estudian los medicamentos

Antes de que un medicamento nuevo salga al mercado, normalmente ha sido probado en animales para obtener una respuesta deseable y conocer la dosis y los efectos secundarios. A partir de aquí debe someterse a tediosas pruebas en humanos que suelen implicar varias fases durante bastantes años[8-10].

Fase I. Farmacología clínica y toxicidad

Determinar la dosis mejor tolerada, destacando los efectos secundarios y farmacocinéticos de dosis incrementales. A menudo se realizan en voluntarios sanos, pero también pueden participar sujetos con la enfermedad objetivo. Habitualmente son estudios observacionales y la información sirve para diseñar ensayos en la fase II.

Fase II. Evaluación inicial de los efectos terapéuticos

Evaluar la eficacia en una situación clínica dada y continuar evaluando las toxicidades a corto plazo junto con la mejor relación dosis-respuesta. En estos estudios observacionales se incluye un número limitado de sujetos humanos (normalmente, menos de 100).

Fase III. Evaluación del tratamiento a gran escala

Comparar la eficacia y la toxicidad del nuevo medicamento con el medicamento estándar. Por lo general, es un ensayo controlado aleatorizado con cientos de participantes. El medicamento se administrará del mismo modo que cuando se ponga a la venta. Después de esta fase, el patrocinador puede solicitar la aprobación de la U. S. Food and Drug Administration (FDA).

Fase IV. Observación sistemática posterior a la puesta en el mercado

Comprobar los efectos secundarios tardíos y comparar los resultados en la comunidad con los de los ensayos (validación interna y externa).

Un método es empezar buscando artículos de revisión sobre un tema en particular. Esto probablemente será productivo si la enfermedad estudiada es frecuente y es un problema de salud fundamental. Los recursos mencionados proporcionan este servicio, al igual que el sitio web United Health Foundation's Clinical Evidence (www.clinicalevidence.com). Otra posibilidad es buscar únicamente metaanálisis o EC, ya que son generalmente más fiables. Muchos mecanismos de búsqueda tienen un comando que limita la búsqueda a estos tipos de estudios. Sin embargo, debido a que muchos temas tienen una información limitada en estos formatos, la búsqueda también tendrá que incluir estudios observacionales. Si usted ha notado un acontecimiento extraño en un paciente concreto, puede únicamente querer buscar informes de casos similares. Una vez completada la búsqueda, los artículos pueden evaluarse por su validez y aplicabilidad.

PUNTOS CLAVE

- Todos los tipos de estudios de investigación proporcionan una información importante.
- El tipo de estudio que puede realizarse depende de factores como el marco temporal, la capacidad de enmascarar a los participantes y aspectos éticos como el consentimiento informado.

- Los informes de casos pueden identificar una asociación entre variables no conocida previamente.
- Un estudio de casos y controles es un tipo de estudio observacional retrospectivo que compara la diferencia en la exposición a una sustancia empezando a trabajar con individuos enfermos y comparando el grado de exposición a la misma sustancia con sus equivalentes sanos. Puede calcularse la *odds ratio,* que es la probabilidad de que un enfermo haya estado expuesto respecto a un individuo no enfermo.
- Un estudio de cohortes es un estudio observacional prospectivo. Los casos se seleccionan según la exposición y se siguen durante un período de tiempo. Debido a que tenemos un punto de partida, podemos medir el riesgo de enfermedad (la probabilidad de contraer la enfermedad con el tiempo) en individuos expuestos comparado con el riesgo de enfermedad en individuos sin la exposición.
- El ensayo clínico controlado aleatorizado está diseñado para minimizar el sesgo. Las confusoras potenciales deberían distribuirse equitativamente entre los grupos y no deberían influir en los resultados.
- Un metaanálisis combina los datos de múltiples estudios después de ser realizados. Su fuerza radica en la potencia que hay tras el gran número de individuos que debería comportar una estimación precisa del parámetro.
- Los diagramas de embudo sirven para identificar el sesgo de publicación.
- La revisión de artículos es una recopilación a partir de una búsqueda bibliográfica. Una revisión sistemática usa un diseño científico estricto para garantizar la inclusión de toda la información apropiada y protegerse del sesgo.
- Antes de poner los medicamentos en el mercado, deben someterse a rigurosas pruebas en varias fases consecutivas para evaluar la eficacia y la toxicidad.

BIBLIOGRAFÍA

1. Friedland, D. J. 1998. *Evidence-based medicine: A framework for clinical practice*. Stanford, CT: Appleton and Lange.
2. Dawber, T. W. 1980. *The Framingham study*. Cambridge, MA: Harvard University Press.
3. http://en.wikipedia.org/wiki/Kaposi's_sarcoma. 2005.
4. Wu, W. C. et al. 2001. Blood transfusion in elderly patients with acute myocardial infarction. *New Engl J Med*, 345:17, 1230.
5. Kjaergard, L. L. et al. 2001. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Ann Intern Med*, 135:11, 982–989.
6. Moher, D. et al. 1999. Improving the quality of reports of meta-analyses of randomized controlled trials: The QUORUM statement. Quality of Reporting of meta-analyses. *Lancet*, 354:9193, 1896–1899.
7. Stroup, D. F. et al. 2000. Meta-analysis of observational studies in epidemiology: A proposal for reporting. *JAMA*, 283:15, 2008–2012.
8. http://en.wikipedia.org/wiki/Clinical_trial. 2005.
9. www.clinicaltrials.gov/ct/info/phase. 2005.
10. Dr. S. Murray, lecture notes on Clinical Trials, University of Michigan School of Public Health, 2000–2001.

PREGUNTAS DE REVISIÓN

1. Los estudios de casos y controles empiezan con el diagnóstico y miran hacia atrás en busca de variables asociadas, por eso son _____.

2. Los estudios de casos y controles no tienen un punto de partida uniforme; por tanto, no tienen una medida de _____ _____, pero miden _____ _____.

3. Los ensayos aleatorizados no dejan a los participantes elegir el tratamiento, lo que implica que las variables _____ deberían distribuirse equitativamente y no afectar al resultado.

4. En un ensayo _____, los participantes no saben a qué intervención les asignaron; por tanto, no sesgarán su respuesta.

5. A veces, los resultados de un ensayo bien ejecutado no son reproducibles en la _____. Esto puede ser consecuencia de un sesgo de selección.

6. Un _____ no selecciona a sus propios participantes, sino que combina datos ya recogidos y analizados.

7. Si un diagrama de embudo es asimétrico, puede haber un sesgo de _____.

8. Una vez que un medicamento se ha estudiado en humanos, la _____ III es un ensayo controlado aleatorizado que compara el nuevo medicamento con el estándar.

RESPUESTAS

1. retrospectivos

2. riesgo relativo, *odds ratios*

3. confusoras

4. enmascarado

5. comunidad

6. metaanálisis

7. publicación

8. fase

CAPÍTULO 18

Evaluación de la evidencia

La medicina basada en la evidencia es la práctica de tomar decisiones médicas con la identificación, la evaluación y la aplicación razonable de la información más relevante.

—D. J. Friedland, et al.[1]

Frecuentemente, los profesionales médicos se encontrarán en la situación de tener que formarse una opinión o hacer una recomendación a un paciente basándose en la bibliografía. Un juez puede tener que decidir la validez y admisibilidad de un testimonio experto. Un periodista puede tener que escribir un artículo sobre un gran avance médico para satisfacer la curiosidad del público. En cualquier caso, el profesional recurre a la bibliografía esperando hallar una respuesta. Existe una increíble cantidad de investigación ahí fuera. ¿Cómo empieza uno a criticar estos estudios?

Cualquier estudio bien diseñado resulta en una conclusión sólida. Sin embargo, el tipo del diseño limita la fuerza de la conclusión. Un estudio es lo que es. Por ejemplo, la metodología más precisa en un estudio de casos y controles no puede eliminar todo el sesgo de memoria, o un estudio de cohortes meticuloso podría estropearse por no considerar ciertas variables confusoras. Incluso los ensayos controlados y aleatorizados tienen sus limitaciones en el análisis y en la selección. Cada tipo de estudio hace una importante contribución a la base del conocimiento. No obstante, la fuerza de la conclusión depende de la metodología y de la cantidad de sesgo que conlleve.

El estudio más preciso es como el oro de 24 quilates. El sesgo representa las impurezas encontradas en el producto final. Requiere un esfuerzo concentrado quitar las impurezas durante el diseño y la ejecución de un ensayo. Aunque algunos estudios no puedan refinarse más debido a las limitaciones inherentes del diseño, el producto todavía puede tener un valor significativo. El sesgo no puede eliminarse totalmente pero la investigación será más sólida si se reduce lo máximo posible, como en un ensayo clínico controlado aleatorizado y enmascarado (EC).

Varios autores y comités han publicado un sistema para evaluar la fuerza de la evidencia. Por ejemplo, el U. S. Preventive Services Task Force (Grupo Especial de Servicios Preventivos de Estados Unidos) ha publicado una simple jerarquía de la calidad de la evidencia según el diseño del estudio (resumida en la tabla 18-1). En este esquema, un EC bien diseñado descrito en una revisión literaria tendría más credibilidad que un estudio observacional. Las recomendaciones se clasifican en una escala alfabética según la evidencia en su favor. La fuerza de la recomendación es

TABLA 18-1 Sistema de evaluación de la evidencia clínica del U. S. Preventive Services Task Force

Calidad de la evidencia	
I	Evidencia obtenida de al menos un ensayo controlado aleatorizado y correctamente diseñado
II-1	Evidencia obtenida a partir de ensayos controlados no aleatorizados y bien diseñados
II-2	Evidencia obtenida a partir de estudios de cohortes o de casos y controles bien diseñados, realizados preferentemente en más de un centro o por más de un grupo de investigación
II-3	Evidencia obtenida a partir de múltiples series comparadas en el tiempo con o sin intervención. Resultados importantes en experimentos no controlados (como la introducción del tratamiento con penicilina en los años cuarenta) también podrían considerarse en este tipo de pruebas
III	Opiniones de autoridades respetadas basadas en la experiencia clínica, estudios descriptivos o informes de comités expertos
Fuerza de las recomendaciones	
A	Adecuada evidencia que respalda la intervención
B	Cierta evidencia que respalda la intervención
C	Evidencia insuficiente para recomendar o desaconsejar la intervención, pero podría recomendarse en otro ámbito
D	Cierta evidencia en contra de la intervención
E	Adecuada evidencia en contra de la intervención

(De Friedland, D. J. et al. 1998. *Evidence-based medicine: a framework for clinical practice.* Stamford CT: Appleton & Lange, p. 229, con autorización.)

directamente proporcional a la calidad de la evidencia. La mayoría de escalas publicadas están de acuerdo con este sistema de clasificación. En algunas de éstas, el metaanálisis se coloca en lo alto de la lista como la evidencia más fiable.

Otro enfoque es evaluar las indicaciones para realizar un cierto procedimiento. Estas guías aseguran que los recursos se distribuyen adecuadamente a los individuos que ganarán más con la intervención o la prueba. También intentan identificar a los pacientes con un riesgo de perjuicio más alto durante el procedimiento. El American College of Cardiology y la American Hearth Association (ACC/AHA) han puesto en práctica un sistema de clasificación (resumido en la tabla 18-2) a fin de evaluar las indicaciones para realizar un procedimiento o tratamiento particular.

TABLA 18-2 Clasificación de recomendaciones

Clasificaciones	
Clase I	Hay evidencia y/o consenso de que un procedimiento o tratamiento es útil y eficaz
Clase II	Hay un conflicto de evidencia y/o una divergencia de opinión sobre la utilidad/eficacia de un procedimiento o tratamiento
Clase IIa	La mayor parte de evidencia/opinión es favorable a la utilidad/eficacia
Clase IIb	La mayor parte de evidencia/opinión es contraria a la utilidad/eficacia
Clase III	Hay evidencia y/o consenso de que el procedimiento/tratamiento no es útil/eficaz y en algunos casos puede ser perjudicial
Nivel de evidencia	
Nivel de pruebas A	Datos procedentes de múltiples ensayos clínicos aleatorizados o metaanálisis
Nivel de pruebas B	Datos procedentes de un solo ensayo clínico aleatorizado o de estudios no aleatorizados
Nivel de pruebas C	Sólo opinión consensuada de expertos, estudios de casos y controles o tratamiento estándar

(Adaptado de www.acc.org/qualityandscience/clinical/statements, 2006.)

CÓMO CRITICAR LA BIBLIOGRAFÍA

En una búsqueda literaria nos encontraremos con varios artículos apropiados. Sabemos clasificar la evidencia, pero ¿cómo vamos a evaluar un artículo concreto? Para empezar, ayuda tener un enfoque organizado al criticar la bibliografía. En este momento usted tiene las habilidades necesarias para hacerlo con confianza. Se han propuesto varios métodos para evaluar un estudio; algunos expertos recomiendan un acercamiento específico según sea el diseño del estudio. Esto conlleva una profunda crítica del estudio y es útil en revisiones de la bibliografía, así como en revistas especializadas. Sin embargo, estos acercamientos requieren un compromiso de tiempo significativo y una consulta continuada a las guías. Puede que no sea necesario criticar todos los artículos exhaustivamente. Me gustaría proponer un método simple aplicable a cualquier tipo de estudio. Esto no requiere ninguna lista de comprobación impresa ni ninguna plantilla.

Hay cinco preguntas que debería hacerse a sí mismo siempre que evalúe un estudio:

¿*POR QUÉ* se realizó el estudio? A menudo puede encontrarse en el título del artículo. La respuesta es la hipótesis de investigación. La introducción forma al lector sobre la información existente y justifica los esfuerzos de los investigadores y de los participantes. Normalmente, la hipótesis de investigación es muy específica y sirve para colocar una pieza en un rompecabezas más grande.

¿*QUIÉN* fue estudiado? Aparece en los criterios de selección. Es la definición de la población. Identifica el tipo de persona que puede beneficiarse del conocimiento.

¿*CÓMO* se hizo el estudio? Es el engranaje del artículo. Los datos concretos pueden hallarse en la sección de Métodos. ¿Cómo se reclutó a los participantes? ¿Cómo se diseñó el estudio para contestar la hipótesis de investigación? Pregúntese si es un estudio observacional o un ensayo controlado, teniendo presente las limitaciones de cada tipo de diseño. Si fuera un ensayo prospectivo, ¿cuánto tiempo se siguió a los sujetos?

¿*CUÁLES* fueron los resultados? Lo encontrará en la sección de Resultados. Debería mencionarse la medida (como el riesgo relativo o la *odds ratio*) con un valor *p*, así como los resultados de los análisis por subgrupos. Un intervalo de confianza más amplio de la medida reflejará un mayor margen de error en la estimación del parámetro.

¿*DÓNDE* estaban las fuentes potenciales de sesgo? Éstas son las limitaciones del estudio. La mayoría de los autores tratarán estas cuestiones en la Discusión, pero usted podría identificar sus propias inquietudes. Recuerde, el sesgo no puede eliminarse totalmente de ningún estudio, pero ciertos diseños pueden minimizarlo. Al lector también le conviene conocer quién financió el estudio. Si la financiación la proporcionó un individuo o una institución que podría beneficiarse del resultado, existe una fuente de sesgo potencial.

USTED Y LA MEDICINA BASADA EN LA EVIDENCIA

Ya hemos llegado al punto donde hemos adquirido las habilidades necesarias para navegar por la literatura médica. Sabemos cómo y por qué se usan las muestras para estudiar a las poblaciones. Conocemos la teoría que hay detrás de las pruebas estadísticas que contrastan la hipótesis nula y cómo usar las distribuciones muestrales para calcular la probabilidad del resultado observado. Nos sentimos cómodos con la definición de valor *p* (recuerde, PREVALece en su mente) y con el concepto de paraguas del intervalo de confianza. También hemos explorado algunas cuestiones filosóficas y hemos revisado la historia del desarrollo de la investigación médica. Naturalmente, el motivo por el cual pasamos el trago de este aprendizaje es para poder tomar decisiones que permitan a las personas vivir más y mejor.

Dave Sackett es un médico del Reino Unido que ha sido durante mucho tiempo un abanderado de la aplicación de la medicina basada en la evidencia en nuestra práctica profesional. El hecho de ser mencionado al final de esta sección no subestima de ninguna manera las significativas contribuciones que ha hecho en este ámbito. Ha dedicado su carrera a promover el proceso de toma de decisiones médicas basadas en la mejor evidencia. Define cinco pasos[2] que deberíamos incorporar al realizar nuestro trabajo:

1. Convierta las necesidades informativas en preguntas contestables.
2. Encuentre la mejor evidencia para contestarlas.
3. Valore críticamente la evidencia por su validez y utilidad.
4. Aplique los resultados a nuestra práctica.
5. Evalúe nuestra actuación.

Estos enunciados resumen más o menos los objetivos de este libro, uno de los cuales es proveerle de la habilidad necesaria para interpretar la bibliografía. Esto le dará la confianza que necesita para practicar la medicina basada en la evidencia. Ahora que entiende el propósito, observará que se produce un cambio de paradigma

al enfrentarse con problemas médicos. Cuando tenga que tomar una decisión, la bibliografía será su mejor amigo. Este recurso fiable estará inmediatamente disponible y espero que no le resulte tan intimidador como lo era hasta ahora. Disfrute de su nuevo conocimiento.

PUNTOS CLAVE

- Una escala comúnmente aceptada en la evaluación de la calidad de la bibliografía se basa en el diseño del estudio.
- La calidad de la evidencia disponible se usa para evaluar la fuerza de una recomendación. Estas recomendaciones tienen en cuenta quién será el mayor beneficiado de una intervención dada, quién se beneficiará menos y quién probablemente saldrá perjudicado.
- Un acercamiento fácil a la evaluación de un artículo individual es:
  - ¿Por qué se hizo el estudio?
  - ¿Cuál es la población?
  - ¿Cómo se hizo el estudio?
  - ¿Cuáles fueron los resultados?
  - ¿Dónde están las fuentes potenciales de sesgo?
- La medicina basada en la evidencia usa las habilidades que usted acaba de adquirir para integrar el conocimiento validado en la práctica médica; por tanto, la gente puede vivir más y mejor y los recursos pueden distribuirse justamente.

BIBLIOGRAFÍA

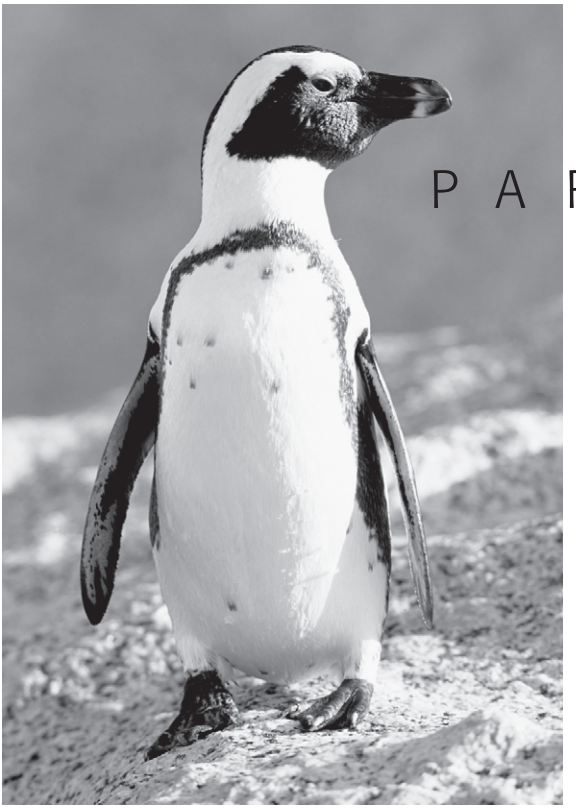
1. Friedland, D. J. et al. 1998. *Evidence-based medicine: a framework for clinical practice*. Stamford CT: Appleton & Lange.
2. Sackett D. L. et al. 1996. *Evidence-based medicine: how to practice and teach EPBM*. New York, NY: Churchill Livingstone.

PREGUNTA DE REVISIÓN

1. Ahora que puedo evaluar la bibliografía, me siento cómodo usándola para tomar decisiones que ayuden a la gente a _____ _____ y _____.

RESPUESTA

1. vivir más y mejor



PARTE III

Aplicaciones habituales

La estadística es arte y ciencia.
—M. Anthony Schork y Richard D. Remington[1]

INTRODUCCIÓN A LA PARTE III

El primer paso en un estudio de investigación es hacerse una pregunta pertinente. La medida escogida dependerá del tipo de estudio que se esté realizando. Hay muchas aplicaciones que pueden causar una gran variedad de medidas para estimar los parámetros poblacionales en el análisis final. Puede haber más de un enfoque correcto.

El siguiente capítulo describe las medidas más habituales y cómo se utilizan para hacer comparaciones, sobre todo si se trata de estimar un parámetro como el riesgo. También se presenta el concepto fundamental de análisis económico. Ésta es una técnica estadística emergente que cuantifica el coste-beneficio. El último capítulo se centra en los diferentes modos de evaluar la precisión de las pruebas diagnósticas que tienen un gran impacto en las decisiones sobre tratamientos.

Estos capítulos deben ser una guía. Cuando se encuentre con un tipo de estudio o prueba diagnóstica concreto, lea la sinopsis del tema para entender cómo interpretar los resultados y familiarizarse con sus aplicaciones. Si quiere profundizar más en un tema, tendrá que consultar un libro de texto de epidemiología o de estadística.

BIBLIOGRAFÍA

1. Schork, M. A. y Remington, R. D. 2000. *Statistics with applications to the biological and health sciences*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, p. 5.

CAPÍTULO 19

Medidas de efecto

Tomar decisiones es la parte más importante de la mayoría de trabajos.

—Robert Hooke[1]

Cuando estimamos un parámetro, a menudo intentamos estimar la fuerza de una relación o la magnitud del efecto de una variable sobre otra. Incluso si decidimos que los datos respaldan un efecto significativo, es útil estimar la magnitud de dicho efecto. Por ejemplo, en un estudio que compara el medicamento estándar con uno nuevo, ambas intervenciones pueden tener un resultado positivo, pero nos gustaría saber el grado de beneficio de una opción sobre otra. Las medidas usadas dependen del diseño del estudio.

ODDS RATIO

Es necesario usar este tipo de medida en estudios de casos y controles. En la metáfora de la estación de tren, los estudios de casos y controles empiezan en las destinaciones de la Enfermedad, como Cáncer y No Cáncer. Entonces retrocedemos para revisar qué llevaban los viajeros en sus maletas (o a lo que estuvieron expuestos). Comparamos el número de individuos en cada grupo que llevaban un cierto artículo, como cigarrillos.

La *odds* de exposición es la probabilidad de haber estado expuesto dividida entre la probabilidad de no haber estado expuesto. Por ejemplo, en los pacientes con un cierto tipo de cáncer de pulmón, las probabilidades de haber estado expuestos al tabaco son:

$$Odds = \frac{\text{Probabilidad de exposición al tabaco}}{\text{Probabilidad de no exposición al tabaco}}$$

Puede hacerse el mismo cálculo para los individuos sanos. Ambos grupos tienen algún grado de exposición. Cuando comparamos las fracciones, como en la figura 19-1, estamos comparando la *odds* de los individuos con cáncer expuestos al tabaco contra la *odds* de los sanos expuestos. Este cociente se conoce como la *odds ratio*.

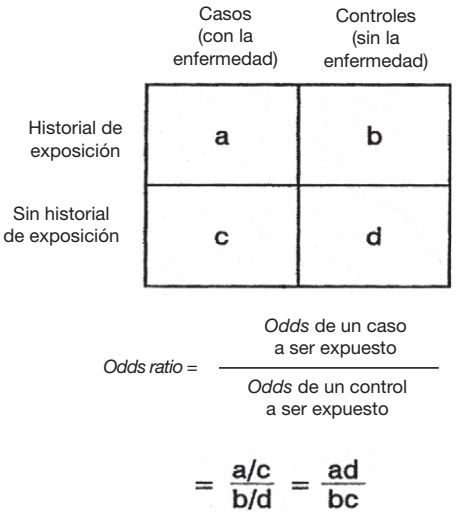


FIGURA 19-1 En una *odds ratio*, primero dividimos los grupos en enfermos y sanos, como indica la doble línea. Entonces revisamos el número de expuestos y de no expuestos en cada grupo. El resultado final nos da la *odds* de los casos que fueron expuestos comparada con la *odds* de los que no fueron expuestos. (Adaptado de Gordis, L. 2001. *Epidemiology*, 3rd ed., Philadelphia: Elsevier.)

Si no hay ninguna diferencia en la exposición, el resultado estará próximo a 1, que apoya la conclusión de que la exposición no tiene ninguna asociación con la enfermedad. Cuanto más lejos estemos del 1, más fuerte será la asociación. Usando la fórmula de la figura 19-1, un número mayor que 1 respalda una asociación entre la exposición al tabaco y el cáncer de pulmón. En algunos casos la exposición puede ser protectora, como el uso del cinturón de seguridad en las personas que han sufrido accidentes de tráfico mortales en relación con los que sobrevivieron. En este caso, la *odds ratio* sería menor que 1 e indicaría que la *odds* de haber usado cinturón de seguridad en víctimas mortales de accidente de tráfico es menor que la *odds* del uso del cinturón en aquellos que sobrevivieron al accidente.

La *odds ratio* no es una medida de riesgo. Recuerde que el riesgo es la tasa de acontecimientos a lo largo del tiempo. Los estudios de casos y controles no pueden proporcionar esta información ya que son estudios retrospectivos; no hay ningún punto de partida real en el tiempo. Sin embargo, la *odds ratio* en un estudio retrospectivo se aproximará al riesgo si la incidencia de la enfermedad en la población es bastante baja.

También es posible calcular una *odds ratio* en un estudio prospectivo. La figura 19-2 muestra este escenario. Las *odds* relativas, tanto en los casos y controles como en los estudios de cohortes, dan el mismo número. Ésta es una manera eficaz de medir si una exposición específica está asociada con una enfermedad.

Normalmente, la *odds ratio* se proporciona con un solo número seguido de un intervalo de confianza entre paréntesis. En un gráfico, el intervalo de confianza puede representarse por una barra. Si el intervalo de confianza incluye el número 1, entonces el resultado no es estadísticamente significativo. La figura 19-3 contiene las *odds ratio* de algunos estudios que observaron la mortalidad a corto plazo en pacientes con infarto agudo de miocardio (IAM) y el beneficio de empezar a tomar inhibidores de la enzima de conversión de la angiotensina (ECA) en estos individuos.

	Desarrolla la enfermedad	No desarrolla la enfermedad
Expuesto	a	b
No expuesto	c	d

$$\textit{Odds ratio} = \frac{\textit{Odds} \text{ de una persona expuesta a desarrollar la enfermedad}}{\textit{Odds} \text{ de una persona no expuesta a desarrollar la enfermedad}}$$

$$= \frac{a/b}{c/d} = \frac{ad}{bc}$$

FIGURA 19-2 La *odds ratio* puede provenir de un estudio de cohortes. La fórmula es la misma que en un estudio de casos y controles, pero comparamos la *odds* de desarrollar la enfermedad en el grupo expuesto contra la *odds* de los no expuestos. (Adaptado de Gordis, L. 2001. *Epidemiology*, 3rd ed., Philadelphia: Elsevier.)

RIESGO RELATIVO

Esta medida de riesgo representa la probabilidad de que un acontecimiento o respuesta concreta sucedan a lo largo del tiempo. Así, el riesgo sólo puede determinarse en un estudio que considere a las personas de manera prospectiva. Los cocientes de riesgos comparan el riesgo para dos o más grupos.

$$\text{Riesgo relativo} = \frac{\text{Incidencia de la enfermedad en expuestos}}{\text{Incidencia de la enfermedad en no expuestos}}$$

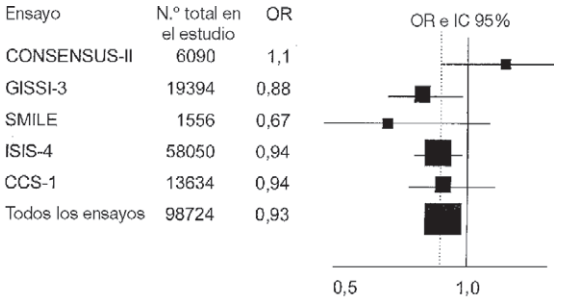


FIGURA 19-3 Las *odds ratio* de estar expuesto a un inhibidor de la ECA en la supervivencia a corto plazo después de un IAM frente a los que murieron. El estudio CONSENSUS-II usó un preparado intravenoso, mientras que en otros estudios se usó un preparado oral. El tamaño de la caja es proporcional al número de sujetos. Si un intervalo de confianza cruza el 1, los resultados no son significativos. Sabemos que si un estudio tiene poca potencia podemos no recoger un efecto significativo aun si existe (error de tipo II). Cuando se combinan los datos, la *odds ratio* real puede estimarse con mayor precisión. Estos resultados sugieren un beneficio real en la supervivencia en un IAM por la administración temprana de un inhibidor de la ECA. (Datos de Zipes, D. P. et al. 2004. *Braunwald's Heart Disease*, 7th ed. Philadelphia: Elsevier/Saunders, p. 1192.)

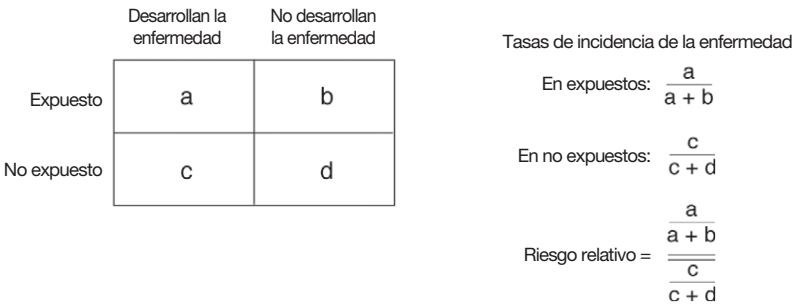


FIGURE 19-4 El riesgo relativo compara la incidencia de la enfermedad entre los individuos expuestos y los no expuestos. (Adaptado de Gordis, L. 2001. *Epidemiology*, 3rd ed., Philadelphia: Elsevier.)

Cuando calculamos un riesgo relativo, comparamos la incidencia de una enfermedad en el grupo expuesto con la del grupo no expuesto. La tabla 2 × 2 de la figura 19-4 muestra cómo se calcula.

La interpretación del riesgo relativo es similar a la de la *odds ratio*. Si el riesgo relativo está próximo a 1, no hay ninguna asociación entre exposición y desarrollo de la enfermedad. Un riesgo relativo mayor que 1 indica una asociación positiva entre la exposición y la enfermedad. Un riesgo relativo menor que 1 significa que la exposición está asociada con menos enfermedad, y que puede ser protectora. La figura 19-5 muestra el efecto de los betabloqueantes en el tratamiento del IAM. El riesgo relativo de mortalidad se compara entre los que tomaron y no tomaron este medicamento.

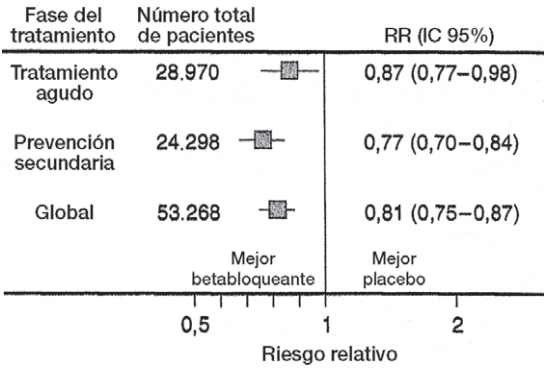


FIGURA 19-5 Efecto de los betabloqueantes en la tasa de mortalidad en pacientes con infarto de miocardio. El riesgo relativo de mortalidad se reduce con betabloqueantes tanto durante la fase aguda del tratamiento como cuando prescribió la prevención secundaria después del IAM. (Datos de Zipes, D. P. et al. 2004. *Braunwald's Heart Disease*, 7th ed. Philadelphia: Elsevier/ Saunders, p. 1190.)

RIESGO ABSOLUTO Y RIESGO ATRIBUIBLE

El riesgo relativo considera el hecho de que algunas enfermedades ocurren en individuos no expuestos, como la incidencia del cáncer de pulmón en no fumadores. Refleja el aumento en la cantidad de enfermedad observada en los individuos expuestos (o la disminución en la cantidad de enfermedad si la exposición es protectora). El riesgo relativo es una medida de la fuerza de la asociación.

El riesgo histórico en individuos no expuestos se conoce como el *riesgo absoluto*. Otra medida útil es la cantidad adicional de enfermedad observada en la población debida a la exposición de ciertos individuos, llamada *riesgo atribuible*. Esta medida tiene en cuenta la carga de la enfermedad en la población y cuánta enfermedad puede prevenirse si se elimina la exposición. La incidencia de la enfermedad atribuible a una exposición se calcula como:

$$\text{Riesgo atribuible} = \frac{(\text{Incidencia en el grupo expuesto})}{(\text{Incidencia en el grupo no expuesto})}$$

Si hay una fuerte asociación entre la exposición y el desarrollo de la enfermedad, el riesgo atribuible será alto. Quitar la exposición tendrá un impacto significativo en la incidencia total de la enfermedad y en la carga del sistema sanitario. El riesgo atribuible también puede entenderse como la proporción de riesgo adicional en individuos expuestos, que puede expresarse como un porcentaje. La fórmula es:

$$\frac{(\text{Incidencia en el grupo expuesto}) - (\text{Incidencia en el grupo no expuesto})}{\text{Incidencia en el grupo expuesto}}$$

REDUCCIÓN DE RIESGO

Algunas exposiciones son protectoras y causan una mejoría en la respuesta, como vimos con los inhibidores de la ECA y los betabloqueantes en los pacientes con IAM. La reducción de riesgo absoluto considera la disminución en la incidencia global de enfermedad o muerte para aquellos que recibieron la terapia. La fórmula es:

$$\text{Reducción de riesgo absoluto} = (\text{Incidencia en el grupo placebo}) - (\text{Incidencia en el grupo tratado})$$

La reducción del riesgo relativo por una terapia puede parecer bastante grande, pero cuando se aplica a la población y se calcula la reducción absoluto del riesgo, el beneficio puede ser bastante más modesto. Esto ocurre si la terapia es eficaz pero la incidencia de la enfermedad en el grupo placebo es pequeña. Considere el ejemplo de un nuevo medicamento del cual se está probando su capacidad de prevenir la toxicidad renal en pacientes que están tomando un contraste (como el yodo) durante una exploración con tomografía computarizada (TC). El estudio observa a 200 personas. La mitad toman el nuevo medicamento y la otra mitad el placebo, y se observa la incidencia de toxicidad renal.

$N = 200$

	Toxicidad	No toxicidad
Nuevo medicamento	2	98
Placebo	4	96

Riesgo relativo Reducción: Medicación $\frac{2}{100}$ / Placebo $\frac{4}{100}$ = $\frac{0,02}{0,04}$ = 0,5 o 50%

Riesgo absoluto Reducción: $\frac{4}{100} - \frac{2}{100} = \frac{2}{100}$ = 0,02 o 2%

La incidencia de toxicidad renal es de 4/100 en el grupo placebo y de 2/100 en el grupo tratado. La medicación redujo el número de acontecimientos a la mitad; así, la reducción en el riesgo relativo es del 50%, que parece bastante. La reducción de riesgo absoluto considera la tasa de acontecimientos en la población. En este contexto, la medicación redujo los acontecimientos un 2%, que es una medida más razonable del rendimiento en una población.

NÚMERO NECESARIO QUE HAY QUE TRATAR

La reducción en el riesgo absoluto tiene en cuenta el número de acontecimientos que ocurren en una población con y sin tratamiento. Esto significa que los acontecimientos pasarán tanto si intervenimos como si no, pero podemos reducir esta tasa con un tratamiento eficaz. Esto también significa que algunas personas tomarán el tratamiento a pesar de que nunca hubieran tenido el acontecimiento. Una forma más intuitiva de expresar este concepto es con el *número necesario que hay que tratar* (NNT), que refleja el número de personas que tomarán el tratamiento sin necesitarlo por cada persona en la cual sí se previene el acontecimiento. El NNT es el recíproco de la reducción de riesgo absoluto. La fórmula es:

$$\text{NNT} = \frac{1}{(\text{Incidencia en el grupo placebo}) - (\text{Incidencia en el grupo de tratamiento})}$$

El NNT nos da una idea de cuántos individuos tendrán que ser expuestos al tratamiento para poder recoger algún beneficio en un solo individuo. En el ejemplo sobre la prevención de la toxicidad renal en el contraste de exploraciones con TC, el NNT = 1/0,02 o 50. Tendríamos que tratar a 50 personas para prevenir un acontecimiento.

Incluso aunque el NNT no tenga en cuenta el coste o los efectos secundarios del tratamiento, nos podemos hacer una idea de éstos con esta medida. Si el tratamiento es seguro y barato, probablemente lo recomendaremos aun si el NNT es relativamente grande. Tenga presente que podemos reducir el NNT restringiendo la población tratada. Por ejemplo, sabemos que los diabéticos y la gente con enfermedad renal previa tienen una incidencia más alta de toxicidad renal tras estos contrastes. Si eliminásemos su uso en estos individuos, seguramente veríamos una mayor reducción del riesgo absoluto y, por tanto, un NNT más pequeño.

CURVAS DE SUPERVIVENCIA DE KAPLAN-MEIER

Si seguimos a dos grupos de individuos durante un período de tiempo dado, podemos analizar el beneficio en la supervivencia comparando las proporciones de supervivientes en cada grupo al final del período de tiempo. Sin embargo, muchos estudios que estudian la mortalidad no incluyen a todos los pacientes en el mismo momento, y además se siguen durante tiempos diferentes. Es habitual usar el método de análisis de supervivencia de Kaplan-Meier para analizar estos tipos de estudios longitudinales. A diferencia de los métodos que calculan las tasas de supervivencia sobre intervalos de tiempo fijos, como tablas actuariales, esta medida calcula una nueva tasa de supervivencia cada vez que se produce una muerte. Debido a que las muertes no ocurren en períodos fijos, hay intervalos de tiempo desiguales y el gráfico es un escalonado irregular. El gráfico representa la proporción de supervivientes cada vez que se produce una muerte, y es esencialmente la probabilidad de supervivencia en cualquier intervalo de tiempo dado empezando desde el cero (cuando el individuo entra en el estudio).

A menudo estos estudios duran muchos años y es bastante habitual perder la pista de algunos participantes. Sin embargo, la información importante es el tiempo que se observaron. El análisis de Kaplan-Meier tiene en cuenta las pérdidas (como datos *CENSURADOS*) asignando la probabilidad ponderada de muerte según las tasas de mortalidad de los que permanecieron bajo observación. Estos datos se incorporan al volver a calcular la probabilidad de supervivencia cada vez que se produce una muerte.

Cuando se estudian dos o más grupos, las curvas se representan juntas. Se pueden comparar las pendientes de las curvas con un análisis estadístico, como la prueba del *log rank*. Si las curvas están próximas, o se cruzan, será poco probable que tengan tasas de supervivencia significativamente diferentes y las dos opciones estudiadas tendrán el mismo beneficio. La figura 19-6 es un ejemplo de curva de Kaplan-Meier en mujeres con cáncer de mama que fueron aleatorizadas a participar o bien en el grupo de terapia de soporte o bien en el grupo control. Aunque las curvas son di-

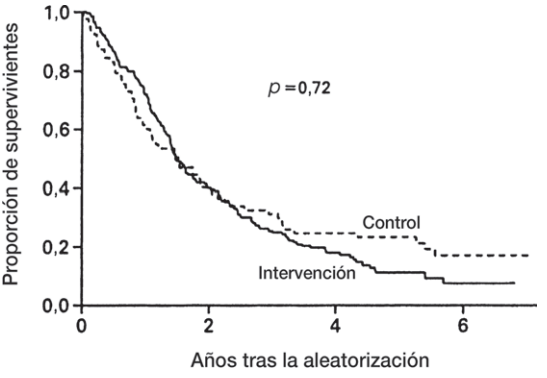


FIGURA 19-6 Curvas de supervivencia de Kaplan-Meier de mujeres asignadas al grupo intervenido y al grupo control. No había ninguna diferencia significativa en la supervivencia entre los dos grupos. (Adaptado de Gordis, L. 2001. *Epidemiology*, 3rd ed., Philadelphia: Elsevier.)

ferentes, no hay ninguna diferencia significativa en la supervivencia en los dos grupos ($p = 0{,}72$).

A veces, en la ordenada tendrá la supervivencia libre de acontecimientos. Si alguien está vivo, pero tiene un acontecimiento controlado, como un ictus, se traducirá en un descenso en la curva. Además de para la supervivencia, este método también sirve en el estudio de otros resultados dicotómicos que suceden a lo largo del tiempo, como el desarrollo o no de la diabetes.

ANÁLISIS DE RIESGOS PROPORCIONALES DE COX

Esta metodología es una prueba habitual para demostrar diferencias en la supervivencia, pero al mismo tiempo considera otras variables que podrían afectar a la respuesta y que no pudieron controlarse al incluir a los participantes. Por ejemplo, la edad de las personas en el momento de la inclusión en un estudio podría influir en la diferencia entre tasas de supervivencia si un grupo incluyera individuos significativamente más viejos que el otro. El método de Cox permite que identifiquemos qué variables están asociadas con una mayor supervivencia. También puede servir para cuantificar el efecto de diversas variables en otros resultados dicotómicos, como el peso y los hábitos alimentarios en el desarrollo de la diabetes.

El *hazard ratio* (o cociente de riesgos instantáneos) es una comparación del efecto de diversas variables en la supervivencia o en otras respuestas que se desarrollen en el tiempo. Se asemeja al riesgo relativo en que un *hazard ratio* más alto conlleva una mayor probabilidad de que el individuo expuesto desarrolle la respuesta.

BIBLIOGRAFÍA

1. Hooke, R. 1983. *How to tell the liars from the statisticians*. New York: Marcel Dekker, Inc.

CAPÍTULO 20

Análisis económico

Los recursos para proporcionar asistencia médica son enormes, pero no ilimitados.

—Mike F. Drummond et al.[1]

Las medidas del efecto consideradas hasta ahora son puramente científicas. Todas ellas nos conducen al noble principio de proveer un tiempo de vida de calidad adicional que mejore la existencia de las personas. Como todavía vivimos en un mundo de recursos limitados, la práctica médica requiere de un cierto grado de consideración económica. Como la población crece y envejece, el coste de los tratamientos seguirá aumentando, sobre todo en enfermedades prevalentes y con tratamientos relativamente caros. El análisis económico es un enfoque que tiene en cuenta no sólo vivir más y mejor, sino también el coste de implantar las intervenciones.

El análisis económico del *coste-eficiencia* cuantifica el beneficio a través de una de las medidas de respuesta estándares que recogen la mortalidad o la calidad de vida (CV). La medida puede reflejar la CV reducida resultante de los efectos secundarios de la medicación, de un empeoramiento de la enfermedad o de una complicación del proceso que afecte al bienestar global. Además, se ha de tener en cuenta el coste económico del tratamiento. Esto incluye no sólo los gastos directos del proceso y de la medicación, sino también los gastos ocultos desde una perspectiva social, como los ingresos perdidos debidos a la baja laboral o el coste de un asistente social mientras un paciente se recupera. Los estudios de *coste-utilidad* tienen en cuenta expresamente el número de años vividos con salud plena (AVAC) y se consideran un tipo de análisis de coste-eficiencia, pero normalmente se usan de modo intercambiable[2].

Por ejemplo, consideremos la situación teórica de una cara intervención quirúrgica que mejore la visión en una enfermedad frecuente que cause deficiencia visual. Los pacientes también pueden usar colirio y conseguir una mejoría, pero ésta será mayor si se someten a la operación. Hasta ahora parecería que debería recomendarse la operación. Sin embargo, si durante varios meses se necesitaran varios procedimientos para producir el resultado deseado, y cada operación conllevara un pequeño pero posible riesgo de ceguera permanente en aquel ojo, ¿cuál sería el coste relativo y el beneficio de cada tratamiento? Ahora la decisión no es tan sencilla. Aproximémonos a este problema a través del análisis económico.

En primer lugar, es importante saber que muchos de estos estudios consideran el coste y el beneficio durante varios años, a menudo hasta que todos los sujetos han

muerto por una complicación de su enfermedad o por alguna otra causa. Como es habitual que los datos exactos no estén disponibles, los modelos incorporan los datos disponibles basados en las observaciones de los sujetos (durante varios años) que han seguido varias intervenciones. Entonces se realizan simulaciones con programas informáticos que proyectan estos datos en el futuro, para el resto de vida de los individuos.

Los resultados comparados incluyen el coste y el beneficio para cada tratamiento. La fórmula de la comparación es:

$$\frac{\text{Coste del tratamiento 2} - \text{Coste del tratamiento 1}}{\text{Resultado del tratamiento 2} - \text{Resultado del tratamiento 1}}$$

Los datos se extraen de ensayos publicados o de otras bases de datos (o de opiniones expertas cuando no hay más opción). Si quisiéramos considerar la mortalidad en el ejemplo citado, iríamos a la bibliografía para encontrar la tasa de mortalidad en la cirugía ocular. Por otra parte, si la enfermedad no tiene ningún efecto en la mortalidad, supondremos que la persona vivirá lo esperado para su sexo y enfermedad. A menudo pueden hallarse estos datos en tablas actuariales que proporcionan las compañías de seguros de vida.

Nuestra medida de la respuesta también debería considerar la CV. Ya que la CV está fuertemente relacionada con el grado de visión, podemos usar una medida objetiva, como una prueba visual, para cuantificarla. Además, podemos considerar el impacto en la CV de una ceguera en el ojo sometido a cirugía. Si la cirugía se realiza más de una vez, el impacto potencial de esta complicación debería considerarse cada vez.

Para cuantificar la CV, se usa un sistema de punto relativo. De nuevo, estos datos se suelen tomar de revisiones publicadas en la bibliografía. En el ejemplo, la CV durante un año después de perder la visión en un ojo puede ser 0,7 en una escala relativa de 0 a 1, con 0 = muerte y 1 = salud plena. Para alguien que evite la cirugía y use colirio, el daño continuado en su visión podría dar una CV = 0,5. Si su visión se deteriorara hasta el punto de ser incapaz de conducir o trabajar, la CV sería aún más baja.

Como la CV es subjetiva, el valor adjudicado a una enfermedad dada (como la ceguera) puede variar de una persona a otra. Por ejemplo, es posible que alguien que pierda la vista pueda tener una CV peor que alguien que nació ciego. Las medidas más precisas de CV se basan en numerosas observaciones provenientes de grandes poblaciones.

Por varias razones, es un desafío calcular los costes. El coste de proporcionar un servicio o producto no es necesariamente lo que se le cobra al consumidor. Además, en Estados Unidos no siempre se cobra lo que se factura. La estimación de los gastos médicos en un hospital es bastante trivial; se tabula lo que se reembolsa el hospital y los médicos a través de las compañías de seguros. Sin embargo, se requiere un entendimiento minucioso del complejo proceso de reembolso y facturación. También se deben considerar los gastos por consultas externas, equipamiento médico, productos farmacéuticos y transporte para las visitas. Debería tenerse en cuenta la posibilidad de necesitar el cuidado domiciliario de una enfermera. También ha de evaluarse la pérdida de salario por culpa de la enfermedad. Si se hacen proyecciones de futuro, las cifras tienen que ajustarse por la inflación. Probablemente se estudiarán más las enfermedades frecuentes o las intervenciones costosas porque su coste social es elevado comparado con otras enfermedades con menor impacto económico.

¿Cómo incorporamos todas estas cifras en una fórmula? Empezaremos con un árbol de decisión que compare las diversas vías. Éste es un gráfico que contempla todos los escenarios posibles. Se usa un círculo (nodo) para representar una intervención que podría tener más de un resultado. Al realizar una intervención, los individuos se dividen en el nodo, y el número de sujetos que seguirán un camino concreto dependerá de la probabilidad de que el acontecimiento se produzca.

La figura 20-1 es un árbol de decisión simplificado del ejemplo de la discapacidad visual. Todos los sujetos son de la misma población y cumplen los criterios de selección.

Teóricamente, si cada camino lo siguen 100 pacientes, podríamos tabular los gastos y cuantificar el beneficio para cada rama. Para cada intervención tenemos que conocer la probabilidad de mejorar la visión, de permanecer igual y de empeorarla (asumimos que la mortalidad debida a la cirugía sea insignificante). También tenemos que saber la probabilidad de ceguera como consecuencia de una complicación quirúrgica. Quieres se sometieron a cirugía y no manifestaron ningún cambio en su visión o la empeoraron se someterán a una segunda operación, y volverán hacia atrás en el árbol. Podemos decidir parar el proceso en cualquier punto temporal. Si queremos seguir a los sujetos durante 10 años, podemos proyectar los datos hacia ese punto. También es posible simular el proceso hasta que todos hayan muerto.

El beneficio incluye las medidas de CV que antes consideramos. El beneficio será mayor en la rama quirúrgica porque más personas han mejorado su visión. Sin embargo, el beneficio de los que usaron colirio es bastante significativo, ya que el 25% mejorará únicamente con las gotas.

Hemos dicho que la cirugía es cara y algunos sujetos la sufrirán varias veces. Así, el coste de la rama quirúrgica será elevado. También es posible que el coste del tratamiento con colirio sea alto si las gotas son caras y tienen que ponerse varias veces al día, o si un número considerable de personas necesitan ayuda continuamente debido a una visión deficiente. Es difícil predecir qué tratamiento tendrá una proporción coste-beneficio más favorable, pero ésta es exactamente la clase de pregunta que intenta contestar el análisis económico.

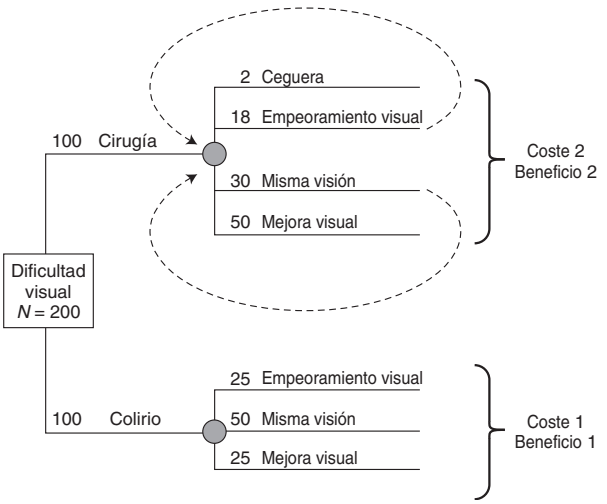


FIGURA 20-1 Árbol de decisión para la cirugía ocular o el tratamiento médico.

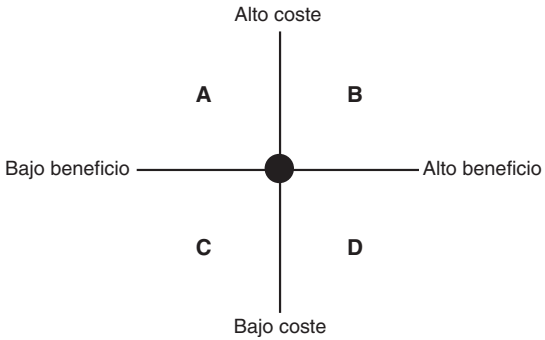


FIGURA 20-2 Beneficio de un tratamiento alternativo en relación con el actual tratamiento estándar si consideramos el coste-beneficio. El número en el cuadrante A refleja un bajo beneficio y un coste elevado, y por tanto debería abandonarse esta alternativa. Un resultado en el cuadrante D no tiene una interpretación difícil —la opción alternativa es mejor y más barata y debería recomendarse—. Las recomendaciones para los resultados de los cuadrantes B y C están menos claras y requieren una discusión sobre si el elevado coste del tratamiento alternativo justifica el beneficio (cuadrante B) o si el bajo coste compensa la pérdida de beneficio (cuadrante C). (De apuntes de Análisis Económico, Prof. Michael Chernow, *University of Michigan School of Public Health*, 2000.)

Se suman los gastos y los beneficios de cada camino con una fórmula compleja que pondera el resultado por la probabilidad de que el acontecimiento ocurra. El resultado final es la cantidad de coste de un camino por unidad de beneficio, como un año de vida ajustado por calidad (AVAC). Es intuitivo pensar que, normalmente, una vía con mejores resultados será más cara. En esta situación, la sociedad debe ser capaz de decidir si el coste justifica el beneficio. La figura 20-2 es un gráfico que representa el coste incremental por unidad de beneficio de una alternativa en comparación con el tratamiento estándar, que tiene adjudicado un coste-beneficio basal situado en el medio del gráfico.

La fuerza del análisis económico está limitada por la disponibilidad de los datos. El proceso requiere integrar los resultados de varios ensayos. Los datos usados en las comparaciones de varias intervenciones podrían provenir de ensayos en los cuales las poblaciones no eran exactamente las mismas o en los que se consideraron diferentes vías de tratamientos. Las estimaciones de los costes pueden diferir según la fuente de información y la perspectiva (p. ej., usar cifras de coste basadas en lo que se factura en vez de lo que se cobra). El hecho de asumir varias premisas es inherente a este tipo de estudio. Una forma de acercarse a la incertidumbre en la proyección del coste y del beneficio es hacer un análisis de sensibilidad empleando un amplio rango de valores para ciertas variables claves. Por ejemplo, si la tasa de complicaciones de un procedimiento puede mantenerse muy baja (mejorando la CV y disminuyendo el coste), la intervención puede cruzar el umbral a partir del cual se considera rentable. El análisis de sensibilidad considera varios valores de las variables que podrían tener un impacto en los resultados y en las recomendaciones finales.

BIBLIOGRAFÍA

1. Drummond, M. F. et al. 1997. Users' Guide to the Medical Literature, *JAMA*, 277:10.
2. wikipedia.org/wiki/cost-utility_analysis 2007.

CAPÍTULO 21

Evaluación de pruebas

Las pruebas diagnósticas son más útiles si cambian una probabilidad a través de un umbral de toma de decisiones.

Lee Goldman et al.[1]

Usamos medidas de precisión para determinar la fiabilidad del resultado de una prueba. Esta información proviene del estudio de una muestra de individuos de una población definida por los criterios de selección. Por ejemplo, la gente a menudo acude al hospital con síntomas de dolor torácico. Si excluyéramos a los menores de 18 años y estudiáramos una muestra de estas personas con dos pruebas distintas —la prueba de referencia y la prueba en estudio—, sabríamos la precisión de la prueba en comparación con los resultados de la prueba de referencia.

El individuo está enfermo (E) o sano (S). Sólo el estándar de referencia puede determinar esto exactamente. Ésta es la prueba más fiable en la discriminación de E y S, pero normalmente es cara e invasiva. Por este motivo no es la primera prueba que se realiza. Cuando nos encontramos con los pacientes, no sabemos si son E o S. A veces, únicamente podemos confiar en los síntomas y en los signos para llegar a un diagnóstico; por ejemplo, un paciente con síntomas y signos de gastritis leve puede tratarse sin exponerlo a pruebas adicionales. Sin embargo, en otros casos queremos asegurar más el diagnóstico, sobre todo si el tratamiento es caro, de larga duración o tiene importantes efectos secundarios.

Las pruebas que usamos darán o bien un resultado positivo (+), o bien negativo (–). No todos los individuos que den + en la prueba estarán E, y no todos los que den – estarán S. Habrá falsos positivos y falsos negativos, pero una prueba precisa tendrá el mínimo número de resultados falsos. La siguiente tabla 2 × 2 muestra cómo los sujetos E y S podrían distribuirse entre los que han realizado la prueba.

		Estado de la enfermedad: Enfermedad	Estado de la enfermedad: Salud
Resultado de la prueba	+	Verdadero positivo	Falso positivo
	–	Falso negativo	Verdadero negativo

SENSIBILIDAD Y ESPECIFICIDAD

La sensibilidad es la probabilidad de que un E dé positivo. Es el número de verdaderos positivos dividido entre todos los E, que incluye no sólo a los E que dan positivo (verdaderos positivos) sino también a los E que dan negativo (falsos negativos). Fíjese que la sensibilidad sólo se refiere a aquellos que están en la columna E de la tabla.

La fórmula de la sensibilidad es:

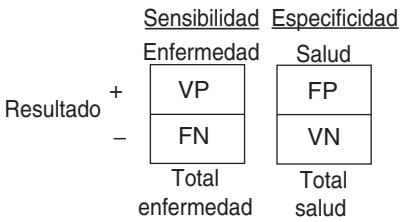
$$\text{Sensibilidad} = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{Falsos negativos}}$$

La especificidad, por otra parte, es la probabilidad que un S dé negativo. Es el número de verdaderos negativos divididos entre todos los S, que incluye tanto a los verdaderos negativos como a los falsos positivos. La especificidad se refiere sólo a la columna S.

La fórmula de la especificidad es:

$$\text{Especificidad} = \frac{\text{Verdaderos negativos}}{\text{Verdaderos negativos} + \text{Falsos positivos}}$$

Un modo de entender la sensibilidad y la especificidad es dividir las columnas, como en la siguiente tabla, porque los individuos enfermos no se solapan con los sanos.



Los conceptos de sensibilidad y especificidad son contraintuitivos porque los pacientes no presentan un signo que los identifica como E o como S, pero pueden someterse a pruebas que los sitúan en la fila de positivos o negativos. Veremos que los valores predictivos positivos y negativos pueden ser más útiles para decidir si un paciente tiene una enfermedad basándose en los resultados de la prueba.

Conviene que estas pruebas orienten el proceso diagnóstico de la enfermedad. Una prueba muy sensible se aproximará a una sensibilidad del 100%. Esto significa que el valor del numerador será parecido al del denominador, o dicho de otra manera, que el número de falsos negativos será muy bajo. Entonces, dentro de todos los resultados negativos, una prueba negativa será creíble. Un resultado negativo indicará de forma fidedigna que alguien está S. Para una prueba con una alta sensibilidad, un resultado negativo excluye la enfermedad.

SANE: con *S*ensibilidad *A*lta, un resultado *N*egativo *E*xcluye la Enfermedad[2]

Por otra parte, una prueba muy específica no tendrá muchos falsos positivos, y aquellos que realmente den positivo probablemente serán verdaderos positivos, es decir, serán E. Para una prueba muy específica con un resultado positivo, el individuo probablemente tendrá la enfermedad.

PAPA: con *P*recisión *A*lta, un resultado *P*ositivo *A*barca la Enfermedad[2]

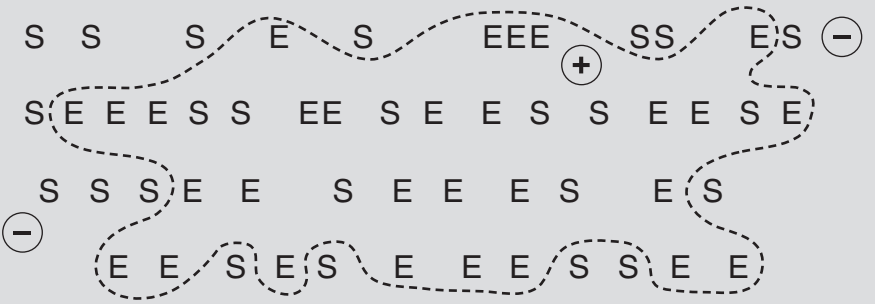
Si usted quiere que SANE el PAPA, usted será capaz de aplicar estas pruebas apropiadamente. Al considerar la sensibilidad y la precisión de una prueba, pregúntese siempre: «¿Cuál es el estándar de referencia con el que se comparó la prueba?» (piense que algunas pruebas de referencia pueden no ser cien por cien exactas). También fíjese en la población estudiada al determinar la sensibilidad y la especificidad de la

CUADRO 21-1

La historia del pastor

La sensibilidad y la especificidad son más fáciles de entender si usamos una metáfora. Piense en dos tipos de pastores, cada uno preocupado por un rebaño de animales domésticos (E), como ovejas. Cada uno quiere construir una valla alrededor de su rebaño para garantizar su alimentación. El problema consiste en que hay algunos animales silvestres (S), como ratones, conejos, zorros, mapaches, etc., que se mezclan con las ovejas (E). En esta metáfora, piense que algo que esté dentro de la valla es un resultado + y algo que esté fuera es un resultado –. Los verdaderos positivos serán las ovejas (E) dentro de la valla y los verdaderos negativos serán los animales silvestres (S) que estén fuera. Un falso positivo será un S dentro de la valla y un falso negativo será una E fuera de la valla.

Un pastor MUY SENSIBLE es generoso. Quiere estar seguro de que todo su rebaño E está bien alimentado. Sabe que puede construir un recinto que incluya todas las E, pero incluirá también algunos S ya que éstos se mezclan con las E. También sabe seguro que algo fuera de la valla será un S. Construirá una gran valla alrededor de la multitud, atrapando todas las E y excluyendo algunos S.



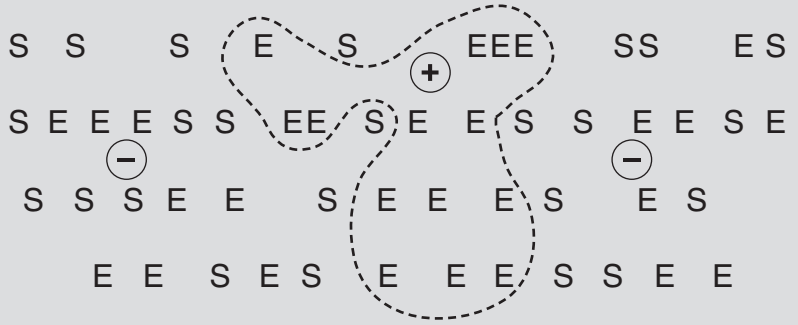
Estar fuera de la valla es análogo a un resultado negativo. Un falso negativo sería una oveja (E) que está fuera de la valla. Fíjese, sin embargo, que fuera de la valla todos son S (no hay ninguna E), así que no habrá falsos negativos.

- *En una prueba muy sensible: aquellos que den negativo probablemente serán negativos (sanos).*

Continúa

CUADRO 21-1 *(continuación)*

Ahora cambie el enfoque. Un pastor MUY ESPECÍFICO es tacaño. No quiere alimentar a otros animales. Construye una cerca mucho más pequeña, de modo que sólo contenga algunas E. A cambio, tiene la compensación de que en el interior todas serán E. Construye la valla para incluir exclusivamente E.



El interior de la valla es análogo a un resultado positivo. Fíjese que dentro de la valla sólo hay E, así que no hay ningún falso positivo, aunque pueden existir muchos falsos negativos que son las E en el exterior.

- *En una prueba muy específica: aquellos que den positivo probablemente serán positivos (enfermedad).*

prueba. Los individuos eran S o E según el estándar de referencia, pero todos ellos tuvieron que cumplir los mismos criterios de selección para considerarse comparables. Ciertas poblaciones con factores de riesgo de la enfermedad tendrán una mayor proporción de E, que puede provocar una capacidad exagerada de la prueba para detectar estos casos.

PRUEBAS DE CRIBADO

Las pruebas pueden usarse con propósitos diagnósticos en individuos sintomáticos, pero también sirven para proteger a una población sana. Las pruebas de cribado deberían ser relativamente seguras, baratas y sencillas. Recuerde la metáfora del pastor: si usted quisiera protegerse de una enfermedad, ¿qué preferiría: una prueba con sensibilidad alta o con alta especificidad? (Pista: las pruebas de cribado deberían ser totalmente inclusivas y no perder a nadie que pueda estar E.)

Leemos la prueba o como positiva (dentro de la cerca) o negativa (fuera de la cerca). Si protegemos a los E, no queremos perder a nadie que esté E, entonces queremos que todos los E den positivo, aunque puedan incluirse algunos falsos positivos. Queremos una prueba de cribado inicial muy sensible ya que identificará a los verdaderos positivos entre todos los resultados positivos. Para confirmar la presencia verdadera de la enfermedad, podemos usar una prueba más específica sobre los individuos cuya prueba inicial dio positivo.

CURVAS ROC

Es muy común en una prueba tener un rango de valores para el resultado. Considerar la prueba positiva o negativa depende de dónde tracemos la línea de separación (o cómo se limite nuestra cerca). Los niveles de tiroides, los niveles de colesterol y las medidas de presión arterial son ejemplos de estas pruebas. La figura 21-1 muestra las distribuciones de los resultados que podríamos obtener para personas E o S.

Queremos un punto de corte que minimice tanto los falsos negativos como los falsos positivos. Idealmente, nos gustarían una sensibilidad y una especificidad del 100%, pero a menudo tenemos que conformarnos con menos. Un aumento de la sensibilidad a menudo comporta una disminución en la especificad. Una curva característica de la operatividad de un receptor (ROC, *Receiver Operating characteristic Curve*) puede ayudarnos a elegir el mejor punto de corte mostrando gráficamente la compensación entre sensibilidad y especificidad.

Las curvas ROC representan las tasas de verdaderos positivos contra las tasas de falsos positivos y sirven para determinar el punto de corte más preciso. El nombre proviene de un sistema usado por los británicos durante la Segunda Guerra Mundial para distinguir las verdaderas señales de radio del ruido de fondo. Usaron estos gráficos para evaluar su capacidad de leer precisamente las señales que los alertaban de una invasión aérea alemana. Si acertaban, era un verdadero positivo; de ser incorrecto, era un falso positivo. El objetivo de su sistema de defensa era minimizar las lecturas falsas. Las curvas ROC también se han utilizado para evaluar la precisión de los motores de búsqueda en Internet[3,4].

La figura 21-2 muestra algunas curvas ROC. En la evaluación de pruebas diagnósticas, la {sensibilidad} se coloca en la ordenada y {1 – especificidad} en la abscisa. La sensibilidad más alta es del 100%, representada por un 1, y la especificidad más alta de 1 resulta en una {1 – especificidad} = 0. Las pruebas más precisas tendrán un punto que se acerque a la parte superior izquierda del gráfico donde la {sensibilidad} = 1 y

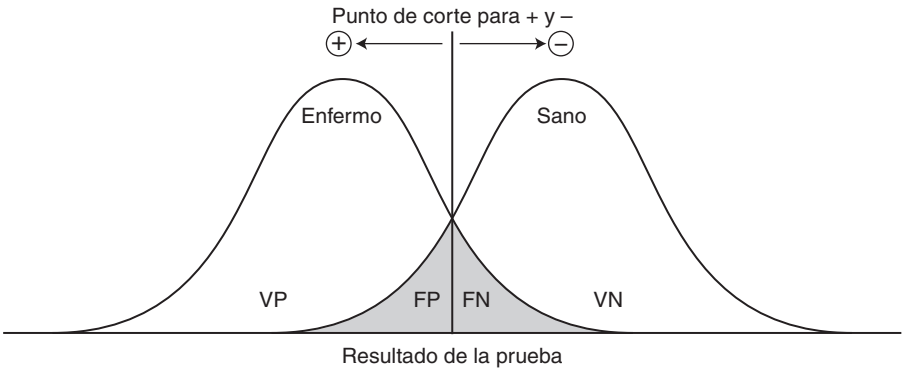


FIGURA 21-1 Distribución de los resultados de una prueba en una escala continua para los E y los S. Los valores se solapan; la distinción entre un resultado positivo y uno negativo depende de dónde coloquemos el punto de corte. Cuando lo movemos en una u otra dirección, cambiará el número de falsos negativos y falsos positivos y afectará a la sensibilidad y la especificidad. VP, verdadero positivo; FP, falso positivo; FN, falso negativo; VN, verdadero negativo. (De www.wikipedia.org/image:Roc-general 2008.)

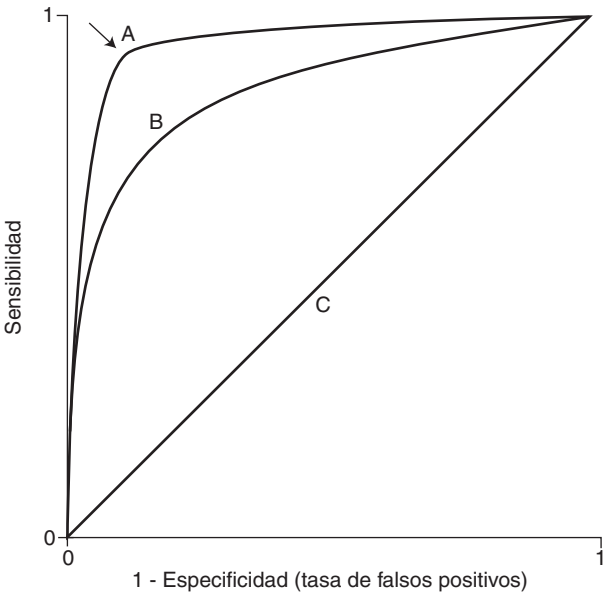


FIGURA 21-2 La curva ROC A corresponde a una prueba con sensibilidad y especificidad próximas al 100% para un punto de corte dado. Ésta es una situación ideal poco habitual y que se acerca a la exactitud del estándar de referencia. Prueba B, con sensibilidad y especificidad más altas en el punto donde la curva está más próxima a la esquina superior izquierda. Esto es el punto que minimiza los falsos positivos y los falsos negativos. La prueba C no tiene mejores resultados que el propio azar. Una pendiente positiva o negativa de esta línea no discrimina en absoluto entre E y S. (De uptodate.com/ROC curve 2008.)

{1 – especificidad} = 0. Las pruebas menos precisas no funcionarán mejor que los resultados al azar, representados por una línea recta con una pendiente de 45 grados.

El gráfico se crea representando {1 – especificidad} contra la {sensibilidad} para diferentes puntos de corte. Cuanto más cerca esté la curva de la parte superior izquierda del gráfico, más precisa será la prueba para aquel punto de corte. En la figura 21-2, la curva ROC A ilustra una prueba con una sensibilidad y especificidad casi perfectas, junto con otras curvas ROC más frecuentes en la vida real.

Un modo de medir la precisión de diferentes pruebas, considerando tanto la sensibilidad como la especificidad, es comparar las áreas bajo las curvas. El área total del gráfico es = 1, entonces una prueba perfecta tendrá un área = 1. Cuanto más cerca esté el área de una prueba concreta a 1, mejor será la prueba. Para una prueba con resultados que no son mejores que el azar (como la curva C de la fig. 21-2), el área será 0,5[5].

VALORES PREDICTIVOS Y RAZÓN DE VEROSIMILITUD

Cuando los pacientes acuden al médico, tienen un conjunto de síntomas que sugieren una enfermedad concreta. Las pruebas diagnósticas nos ayudan a decidir si

tienen la enfermedad. Las pruebas pueden compararse con el estándar de referencia para determinar su precisión. Un valor predictivo positivo es la probabilidad de que un paciente que dé positivo tenga realmente la enfermedad. Un valor predictivo negativo es la probabilidad de que alguien que dé negativo esté realmente sano. Son conceptos análogos a la sensibilidad y la especificidad, pero son más intuitivamente útiles en el modo en que los pacientes se presentan.

La fórmula para el valor predictivo positivo es:

$$\text{VPP} = \frac{\text{Verdadero positivo}}{\text{Todos los positivos (verdaderos y falsos)}}$$

La fórmula para el valor predictivo negativo es:

$$\text{VPN} = \frac{\text{Verdaderos negativos}}{\text{Todos los negativos (verdaderos y falsos)}}$$

La misma tabla 2 × 2 puede usarse para clasificar a los enfermos y los sanos y aquellos que dan + y –. Fíjese que los denominadores están en las filas en vez de las columnas.

		Enfermedad	Salud	
Resultado de la prueba	+	VP	FP	Total +
	–	FN	VN	Total –

Las pruebas con alta sensibilidad tendrán pocos falsos negativos y, por tanto, un mejor valor predictivo negativo, mientras que aquellas con alta especificidad tendrán un mejor valor predictivo positivo. Un rasgo interesante de los valores predictivos es que no sólo dependen de la precisión de la prueba, sino también de la prevalencia de la enfermedad en la población sometida a prueba. Esto tiene un impacto en si un resultado positivo es o no creíble.

Incluso en una prueba con una alta especificidad, si la enfermedad no es muy prevalente, un resultado positivo todavía tendrá una probabilidad relativamente alta de ser un falso positivo. La siguiente tabla muestra cómo el valor predictivo positivo está amortiguado por la baja prevalencia del 1% de la enfermedad, hasta para una prueba con sensibilidad elevada y especificidad del 99%[6].

		Enfermedad	
		E (100 puntos)	S (10.000 puntos)
Resultado de la prueba	+	99	100
	–	1	9.900
		100	10.000

$$\text{Valor predictivo positivo} = \frac{99}{100 + 99} = 50\%$$

La *razón de verosimilitud* (LR, *Likelihood Ratio*) es una evaluación más robusta de la precisión de una prueba diagnóstica. Combina las características de una prueba con sensibilidad y especificidad relativamente estables con la utilidad clínica de los valores predictivos. Esto nos permite explicar el *incremento predictivo* que se ganaría usando una prueba en una población particular[7]. Una LR se define como la probabilidad de que el resultado de una prueba (positivo o negativo) se produzca en un paciente enfermo (E), comparado con la probabilidad de que el mismo resultado ocurra en alguien sano (S).

$$\text{Valor predictivo positivo} = \frac{\text{Probabilidad del resultado de la prueba en pacientes E}}{\text{Probabibidad de los mismos hallazgos en pacientes S}}$$

El rango de la LR va de 0 hasta el infinito. Cuanto más grande sea la LR positiva, más fuerte será el argumento de que estamos ante un E. Cuando la LR negativa se aproxime a 0, más sólido será el argumento de que estamos ante un S. Cuando LR = 1, la prueba no tiene ningún valor diagnóstico[8].

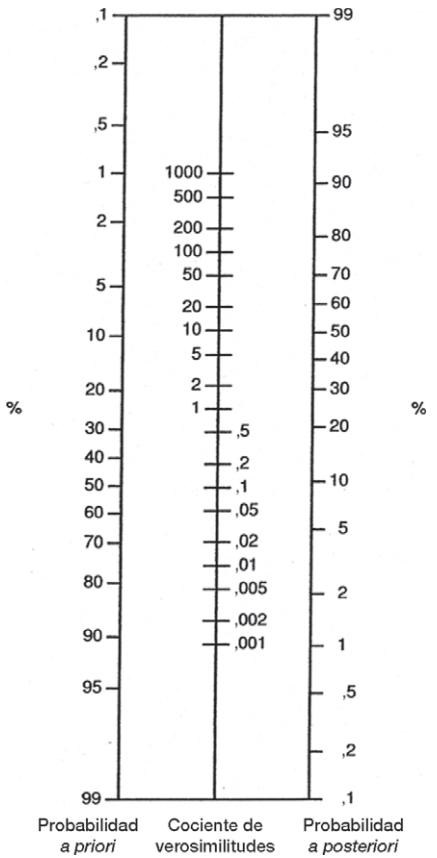


FIGURA 21-3 Nomograma para convertir la probabilidad *a priori* en probabilidad *a posteriori*, después de una prueba diagnóstica con una razón de verosimilitud dada. Las *odds* se han transformado en probabilidades dentro del nomograma. (De Fagan, T. J. 1975. Nomogram for Bayes theorem. Letter: *N Engl J Med,* 293:5, 257, con autorización).

LÓGICA BAYESIANA

La prevalencia de una enfermedad también se conoce como la probabilidad *a priori* de una enfermedad. Esto es como preguntarse: «En una cierta población de individuos con los mismos síntomas, si escogiéramos uno al azar, ¿cuál sería la probabilidad de que estuviera E?». El análisis bayesiano es un instrumento que podemos usar para mejorar nuestras habilidades diagnósticas. El teorema de Bayes incorpora el conocimiento obtenido de los resultados de una prueba diagnóstica y nos permite calcular la probabilidad *a posteriori* de la enfermedad usando la LR.

Una versión simplificada del teorema Bayes es:

Probabilidad *a priori* × LR = Probabilidad *a posteriori*

El nomograma de la figura 21-3 se ha desarrollado para representar la utilidad total de una prueba diagnóstica en escenarios con diferentes probabilidades *a priori* (o prevalencias) de la enfermedad. La probabilidad *a posteriori* refleja el conocimiento adicional ganado haciendo la prueba. Fíjese que para cualquier línea que cruce una LR de 1, las probabilidades *a priori* y *a posteriori* no cambian. LR más extremas tienen un impacto más grande en la diferencia entre las probabilidades *a priori* y *a posteriori*.

En general, las pruebas diagnósticas proporcionan más información adicional al realizarse en individuos que tienen una probabilidad intermedia de la enfermedad, aproximadamente entre el 20 y el 80%. La figura 21-4 ilustra la ganancia de informa-

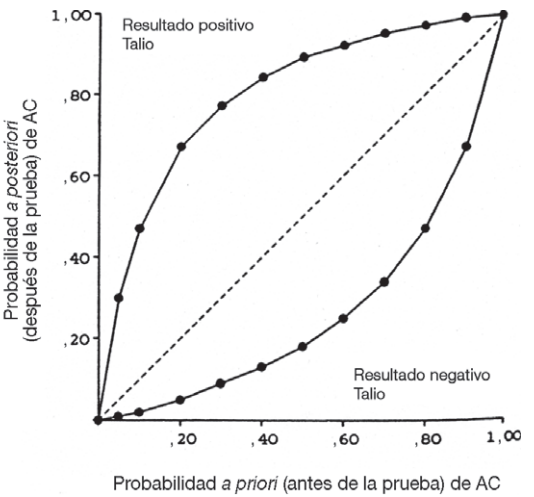


FIGURA 21-4 La probabilidad *a posteriori* de enfermedad de la arteria coronaria (AC) en función de la probabilidad *a priori* y los resultados de una gammagrafía. La mayor parte del conocimiento se gana cuando la probabilidad *a priori,* estimada a partir de los síntomas del paciente, está en un rango moderado. (De Goldman, L. 1986. Noninvasive tests for diagnosing the presence and extent of coronary artery disease: Exercise, electrocardiography, thallium, scintigraphy, and radionuclide ventriculography. *J Gen Intern Med,* 1(4): 258-65.)

ción adicional al hacer una prueba de representación del talio (sensibilidad del 75% y especificidad del 90% comparado con el estándar de referencia del angiograma cardíaco) para diagnosticar enfermedad de la arteria coronaria (AC). Cambios incrementales más grandes de probabilidades *a priori* y *a posteriori* ocurren cuando la prevalencia (probabilidad *a priori*) de la enfermedad es moderada. Esto es cierto tanto para resultados + como –.

Puede ser tentador realizar una prueba que confirme o rechace un diagnóstico, pero una prueba no debería realizarse si su resultado no va a cambiar el tratamiento recomendado. Si sabemos que la probabilidad *a priori* de E o S es muy alta o muy baja, no es probable que una prueba diagnóstica añada un conocimiento considerable sobre el estado del paciente. Hay que ir con cuidado con este principio general si una enfermedad podría resultar mortal si no se diagnostica. En estos casos puede bajarse el listón para la realización de pruebas diagnósticas, aunque sólo puedan añadir unos pocos grados de certeza. En estos casos también puede seguirse directamente el estándar de referencia.

BIBLIOGRAFÍA

1. Goldman, L. et al. 1998. *Primary Cardiology*. Philadelphia, PA: WB Saunders Co.
2. Sackett, D. L. et al. 2000. *Evidence-based medicine: How to practice and teach EBM*. New York: Churchill Livingstone.
3. www.blackwellpublishing.com/specialarticles/jcn_11_136, 2006.
4. www.wikipedia.org "receiver operating characteristic", 2006.
5. www.uptodate.com "evaluating diagnostic tests", 2006.
6. Fleming, M. D. et al. 2004. Pulmonary embolism: Diagnostic evaluation. *Amer J Clin Med*, 1:1, 6.
7. Gallagher, E. J. 1998. Clinical utility of likelihood ratios. *Ann Emerg Med*, 31:3, 391.
8. McGee, S. 2002. Simplifying likelihood ratios. *J Gen Intern Med*, 17:8, 646–649.

Epílogo

No heredamos la tierra de nuestros padres, la tomamos prestada de nuestros hijos.

—Viejo refrán Amish

Tenemos la oportunidad de oro de otorgar muchos años de vida a través de la aplicación del conocimiento adquirido al usar la bioestadística en su plenitud. Muchos de nosotros viviremos nuestras vidas hasta los últimos días, con la mayoría de los años llenos de experiencias productivas y provechosas. Esta ciencia puede servir no sólo para obtener un beneficio a través de los tratamientos médicos, sino también a través de medidas preventivas.

Hemos visto avances asombrosos en el siglo pasado. Somos muchos los que recordamos la llegada de la vacuna de la polio, los antibióticos y la instauración de leyes de salud pública que garantizaban una alimentación y agua salubres. Los tratamientos contra el cáncer y la fase terminal de un paro cardíaco pueden alargar la vida más que nunca hasta la fecha. Sin embargo, no todos los habitantes de este planeta tienen el mismo acceso a los beneficios en el campo de la salud de los cuales sólo algunos disfrutamos. El desafío del próximo siglo será aplicar el conocimiento para otorgar AVAC a todas las personas de cualquier parte del mundo.

Si somos capaces de llevar a cabo este ambicioso proyecto, se habrán cumplido los mejores intereses globales. Sin embargo, el poder de la intervención para mejorar puede ampliarse más allá de los que vivimos en la actualidad. Podemos usar este conocimiento para moldear el futuro de modo que nuestros descendientes sigan apreciando los beneficios que actualmente experimentamos.

La población de la Tierra ha aumentado exponencialmente en el siglo pasado a más de 6.000 millones de habitantes, y sigue creciendo. El consumo de nuestros recursos naturales se está llevando a cabo a un ritmo cada vez más rápido[1]. La consecuencia ha sido un efecto negativo en nuestro entorno natural que ha trastornado el enrevesado equilibrio entre el medio ambiente, las plantas y los animales que han evolucionado durante cientos de miles de años.

El siglo pasado también ha visto acelerar la carrera armamentística hasta el punto de haber creado armas que, si se usan con el propósito para el que fueron diseñadas, son extremadamente destructivas y hasta podrían ser capaces de extinguir la raza humana.

El daño causado por los conflictos políticos puede minar el sistema sanitario y la felicidad de las personas. Todo el AVAC potencial en la vida de un individuo, y sobre todo en la infancia, se devalúa de repente en el escenario de una guerra.

Tenemos la capacidad y los instrumentos para aplicar nuestro conocimiento en la mejora de la especie humana, no sólo en poblaciones limitadas, sino también para todos los presentes y futuros habitantes de la Tierra. Esto es un enorme desafío, pero que debemos afrontar. La extraordinaria capacidad de las personas para resolver problemas, aun si se necesitan varios siglos, es un indicador de que cualquier cosa es posible. Si tenemos la capacidad de alcanzar nuestro potencial intelectual máximo, lo alcanzaremos cuando seamos capaces de reconocer situaciones que tienen un impacto global perjudicial e intervenir con eficacia, protegiendo el derecho del individuo de perseguir una experiencia vital de calidad.

Gail F. Dawson

---

1. Wilson, E. O. 2002. *The Future of Life*. New York: Random House.

APÉNDICE **A**

Flujograma de tipos de pruebas estadísticas

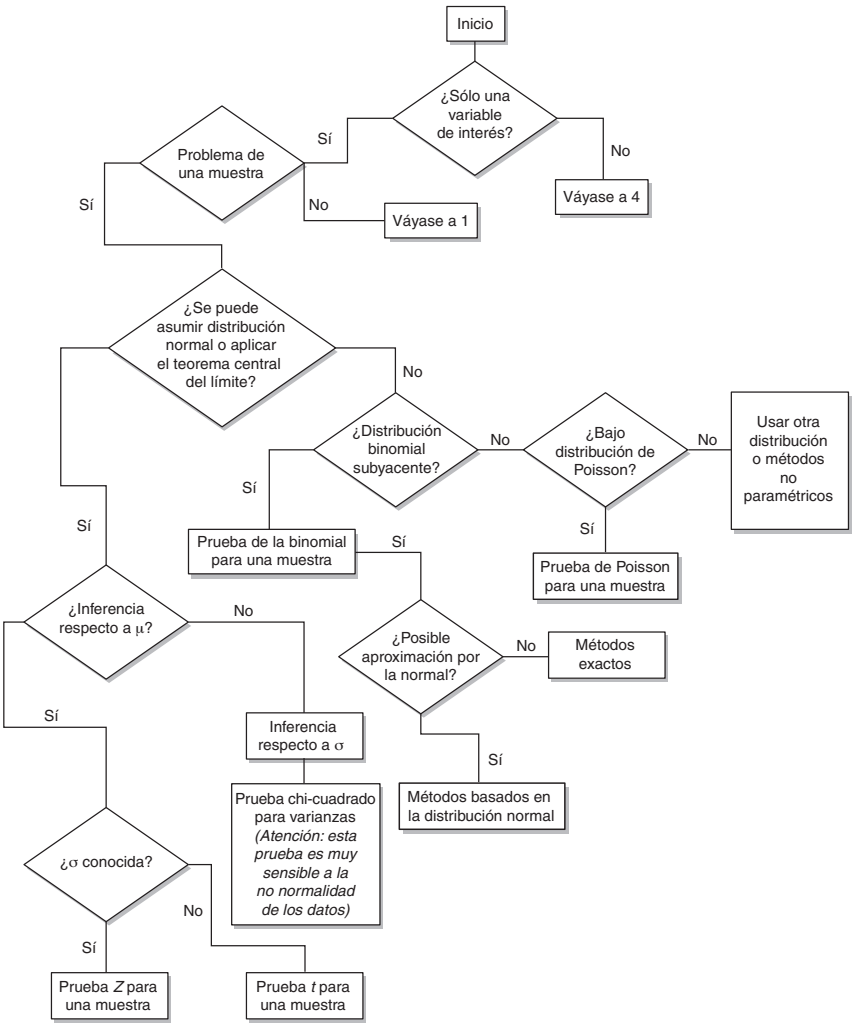


FIGURA AP A-1 Flujograma: métodos de inferencia estadística. (De Rosner B: *Fundamentals of Biostatistics,* Duxbury Press, 1999, con autorización.)

(Continúa)

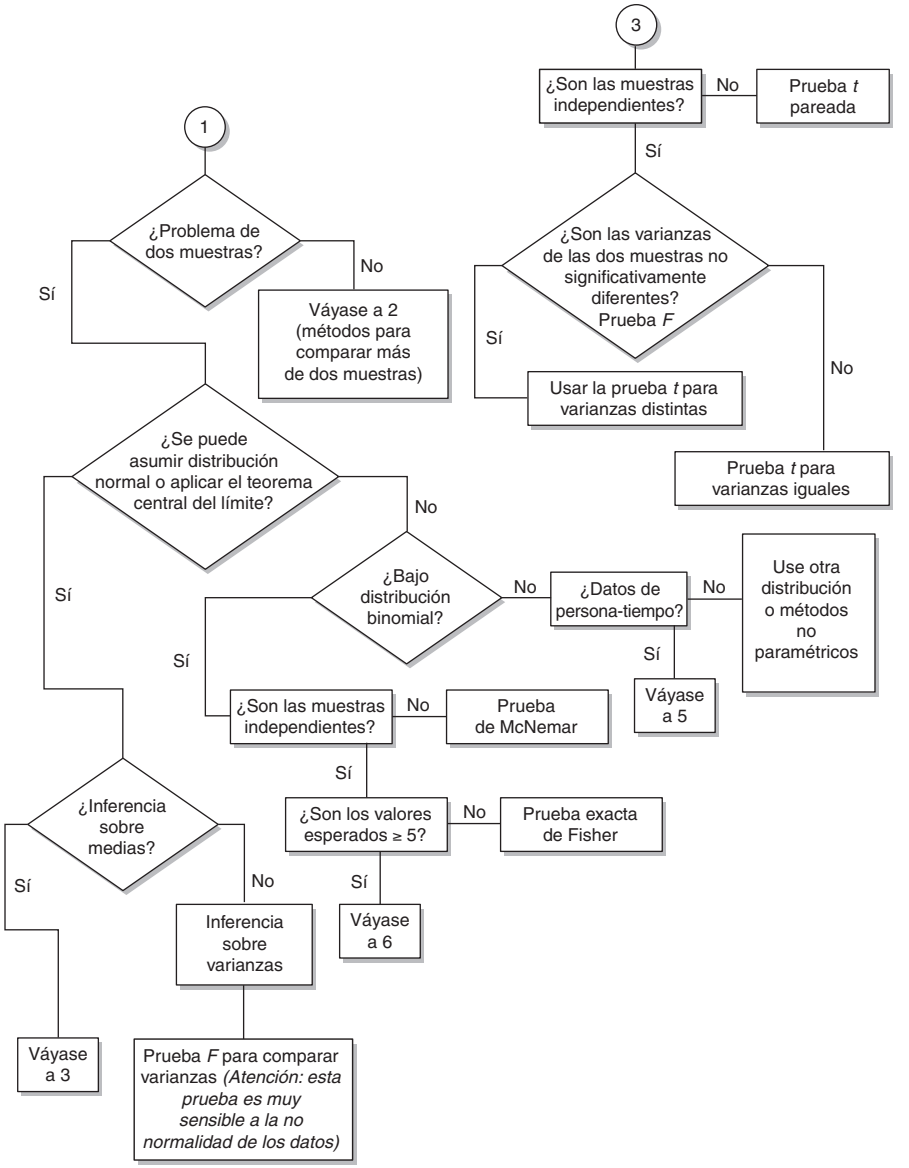


FIGURA AP A-1 ***(continuación)***

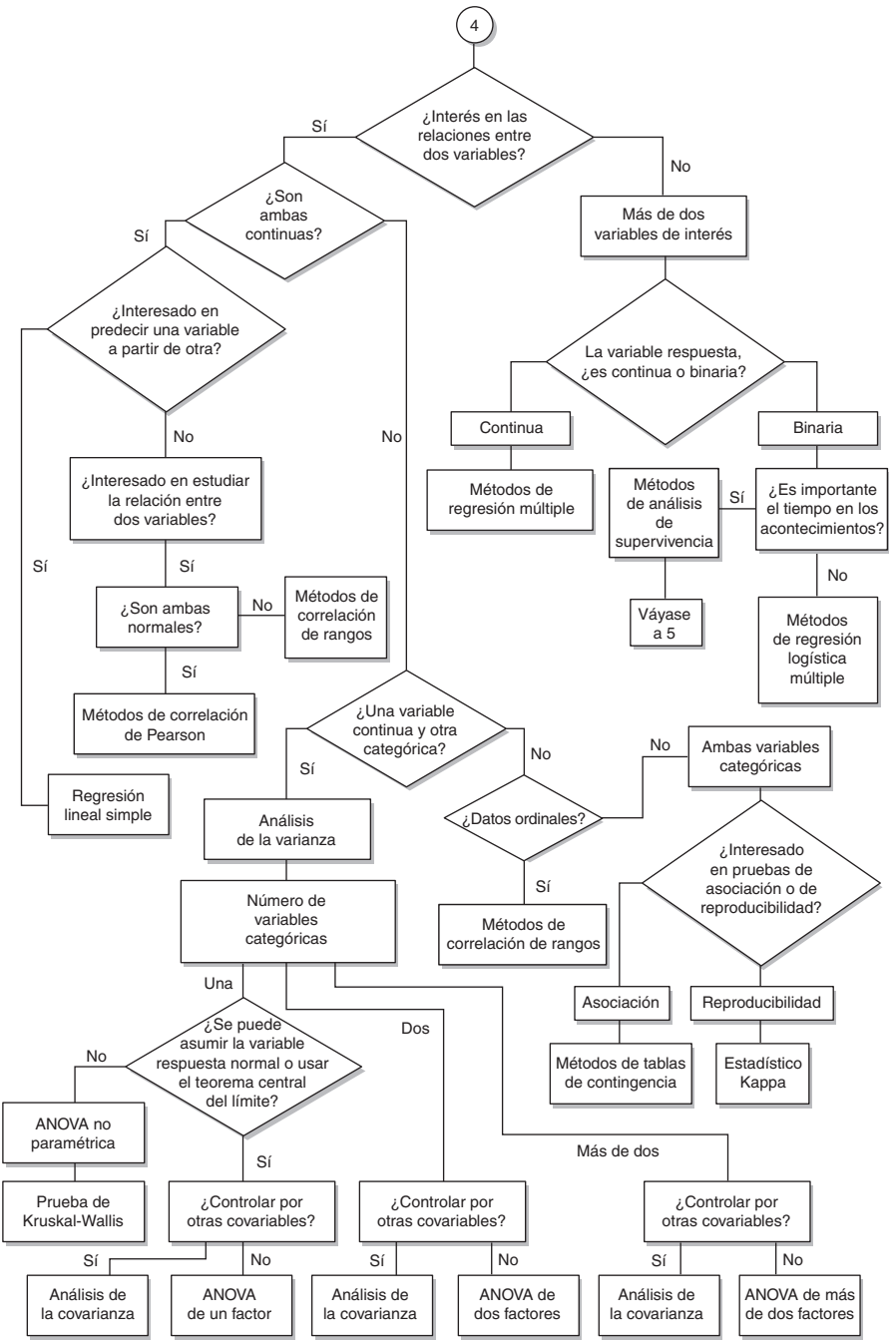


FIGURA AP A-1 ***(continuación)***

(Continúa)

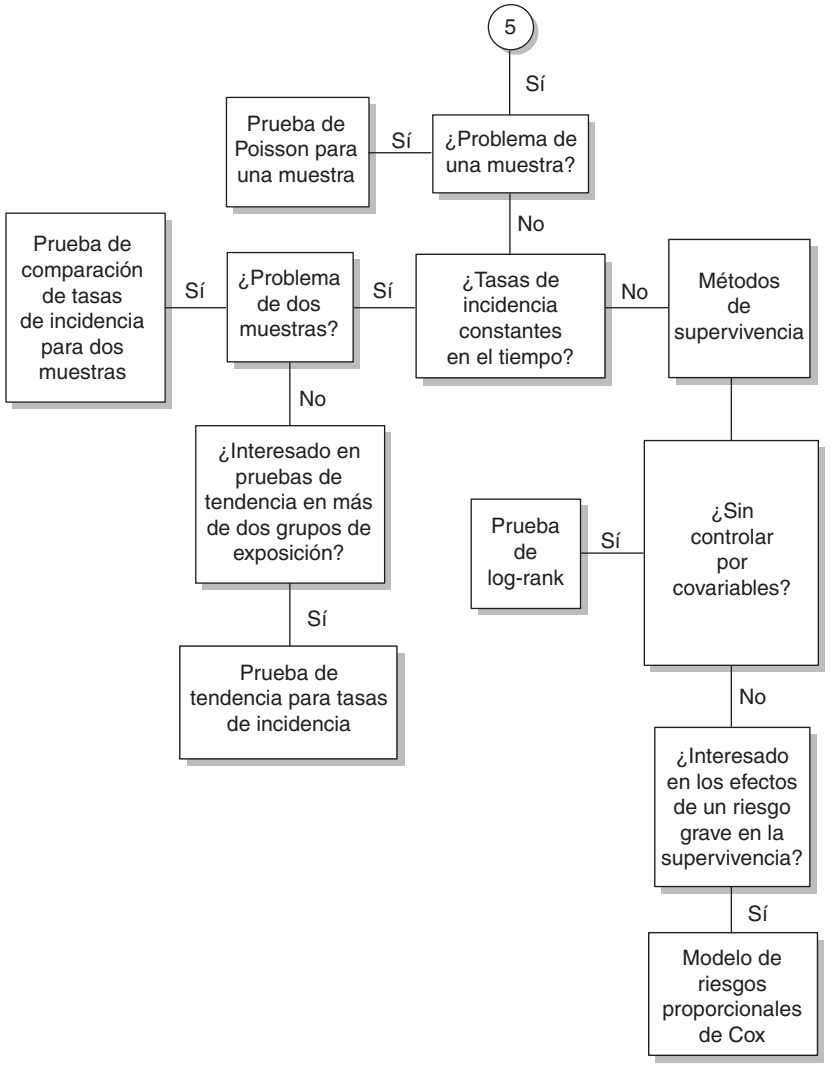


FIGURA AP A-1 ***(continuación)***

APÉNDICE B

Flujograma simplificado de tipos de pruebas estadísticas

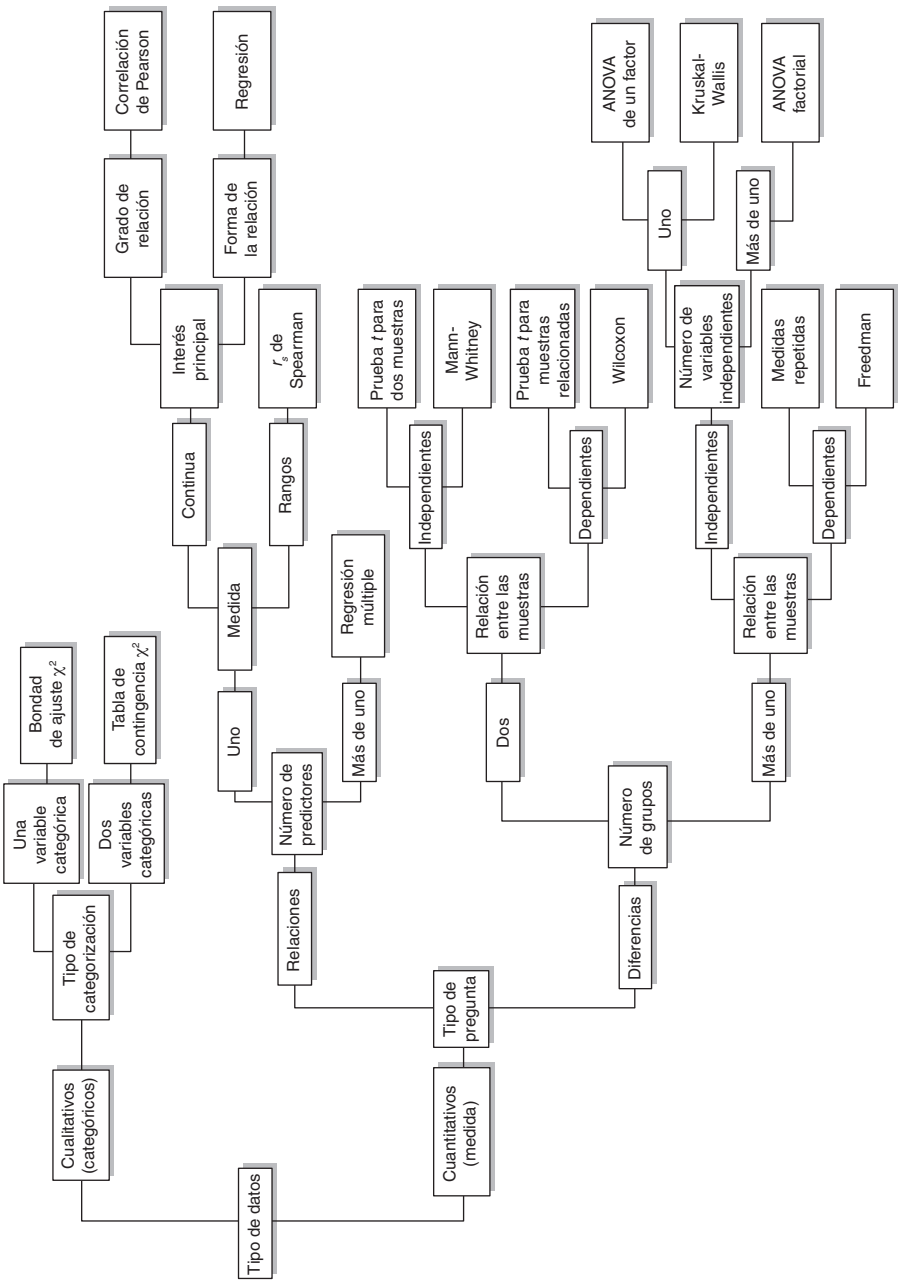
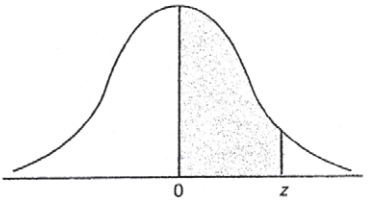


FIGURA AP B-1 Enfoque más simplificado del flujograma del Apéndice A. (De Howell D: *Fundamental Statistics for the Behavioral Sciences*, 4th ed, Brooks/Cole Publishing Co., Pacific Grove, CA, 1999, con autorización.)

APÉNDICE C

Valores de la distribución normal

Área de la distribución normal



z	,00	,01	,02	,03	,04	,05	,06	,07	,08	,09
0,0	,0000	,0040	,0080	,0120	,0160	,0199	,0239	,0279	,0319	,0359
0,1	,0398	,0438	,0478	,0517	,0557	,0596	,0636	,0675	,0714	,0753
0,2	,0793	,0832	,0871	,0910	,0948	,0987	,1026	,1064	,1103	,1141
0,3	,1179	,1217	,1255	,1293	,1331	,1368	,1406	,1443	,1480	,1517
0,4	,1554	,1591	,1628	,1664	,1700	,1736	,1772	,1808	,1844	,1879
0,5	,1915	,1950	,1985	,2019	,2054	,2088	,2123	,2157	,2190	,2224
0,6	,2257	,2291	,2324	,2357	,2389	,2422	,2454	,2486	,2517	,2549
0,7	,2580	,2611	,2642	,2673	,2704	,2734	,2764	,2794	,2823	,2852
0,8	,2881	,2910	,2939	,2967	,2995	,3023	,3051	,3078	,3106	,3133
0,9	,3159	,3186	,3212	,3238	,3264	,3289	,3315	,3340	,3365	,3389
1,0	,3413	,3438	,3461	,3485	,3508	,3531	,3554	,3577	,3599	,3621
1,1	,3643	,3665	,3686	,3708	,3729	,3749	,3770	,3790	,3810	,3830
1,2	,3849	,3869	,3888	,3907	,3925	,3944	,3962	,3980	,3997	,4015
1,3	,4032	,4049	,4066	,4082	,4099	,4115	,4131	,4147	,4162	,4177
1,4	,4192	,4207	,4222	,4236	,4251	,4265	,4279	,4292	,4306	,4319
1,5	,4332	,4345	,4357	,4370	,4382	,4394	,4406	,4418	,4429	,4441
1,6	,4452	,4463	,4474	,4484	,4495	,4505	,4515	,4525	,4535	,4545
1,7	,4554	,4564	,4573	,4582	,4591	,4599	,4608	,4616	,4625	,4633
1,8	,4641	,4649	,4656	,4664	,4671	,4678	,4686	,4693	,4699	,4706
1,9	,4713	,4719	,4726	,4732	,4738	,4744	,4750	,4756	,4761	,4767
2,0	,4772	,4778	,4783	,4788	,4793	,4798	,4803	,4808	,4812	,4817
2,1	,4821	,4826	,4830	,4834	,4838	,4842	,4846	,4850	,4854	,4857
2,2	,4861	,4864	,4868	,4871	,4875	,4878	,4881	,4884	,4887	,4890
2,3	,4893	,4896	,4898	,4901	,4904	,4906	,4909	,4911	,4913	,4916
2,4	,4918	,4920	,4922	,4925	,4927	,4929	,4931	,4932	,4934	,4936
2,5	,4938	,4940	,4941	,4943	,4945	,4946	,4948	,4949	,4951	,4952
2,6	,4953	,4955	,4956	,4957	,4959	,4960	,4961	,4962	,4963	,4964
2,7	,4965	,4966	,4967	,4968	,4969	,4970	,4971	,4972	,4973	,4974
2,8	,4974	,4975	,4976	,4977	,4977	,4978	,4979	,4979	,4980	,4981
2,9	,4981	,4982	,4982	,4983	,4984	,4984	,4985	,4985	,4986	,4986
3,0	,4987	,4987	,4987	,4988	,4988	,4989	,4989	,4989	,4990	,4990

FIGURA AP C-1 Valores de la distribución normal. (De Sternstein M: *Barron's EZ 101 Statistics*, Barron's Educational Series Inc., Hauppauge, NY, 1994, con autorización.)

Índice alfabético

D

E

F

G

H

R

S

T

U

V