Publicación Científica No. 531

Richard K. Riegelman

Robert P. Hirsch

Cómo estudiar un estudio y probar una prueba: lectura crítica de la literatura médica

Organización Panamericana de la Salud

Cómo estudiar un estudio y probar una prueba:

lectura crítica de la literatura médica

Richard K. Riegelman y Robert P. Hirsch

Publicación Científica 531

ORGANIZACION PANAMERICANA DE LA SALUD
Oficina Sanitaria Panamericana, Oficina Regional de la
ORGANIZACION MUNDIAL DE LA SALUD
525 Twenty-third Street, NW
Washington, DC, 20037, EUA

1992

Edición original en inglés:
Studying a Study and Testing a Test.
How to Read the Medical Literature, 2nd edition.

Publicada por Little, Brown and Company
34 Beacon St., Boston, MA 02108, EUA

Catalogación por la Biblioteca de la OPS

Riegelman, Richard K.
Cómo estudiar un estudio y probar una prueba : lectura crítica de la literatura médica / Richard K. Riegelman ; Robert P. Hirsh. — 2a. ed.
Washington, D.C. : OPS, 1992. — 260p.
(Publicación Científica ; 531)

ISBN 92 75 31531 0

I. Hirsch, Robert P. II. Organización Panamericana de la Salud
III. Título IV. (Serie)
1. LITERATURA DE REVISION 2. LECTURA—métodos
NLM WZ345

Primera reimpresión, 1995

Traducción de Josep María Borrás, revisada por el Servicio Editorial de la Organización Panamericana de la Salud. Esta versión en español se publica con permiso de Little, Brown and Company.

ISBN 92 75 31531 0

Contenido

PRÓLOGO A LA EDICIÓN EN ESPAÑOL

La administración del conocimiento constituye una de las siete orientaciones estratégicas de la Organización Panamericana de la Salud para el cuadrienio 1991–1994. Esta orientación desempeña un papel fundamental en el desarrollo científico y técnico del sector salud y, al mismo tiempo, es uno de los pilares que deben regir el desarrollo integral de los países de América Latina.

Somos conscientes de que existen escollos que obstruyen, y a veces impiden, la difusión de los resultados de las investigaciones científicas. Uno de ellos es el uso inapropiado de los métodos para redactar trabajos científicos. Otro, probablemente con repercusiones más graves, es el desconocimiento de los métodos estadísticos y del diseño experimental, necesarios para llevar a cabo un estudio riguroso desde el punto de vista científico. Ello explica por qué los resultados de muchas investigaciones nunca alcanzan las páginas impresas de una revista biomédica o, si lo hacen, pierden su validez al sucumbir frente a una evaluación metodológica estricta. A ello se añade el desconocimiento de muchos profesionales sobre la forma de leer y revisar críticamente un trabajo científico.

Consciente de estos problemas, en los dos últimos años, la Organización Panamericana de la Salud ha intensificado notablemente sus esfuerzos por consolidar la producción editorial destinada a ofrecer libros que permitan mejorar la comunicación entre los profesionales de las ciencias de la salud. Con la edición en español del libro *Cómo escribir y publicar trabajos científicos*, de Robert A. Day, la Organización puso al alcance de los autores de habla española una de las obras de consulta básicas sobre redacción científica. La edición en español de *Cómo estudiar un estudio y probar una prueba: lectura crítica de la literatura médica*, de Richard K. Riegelman y Robert P. Hirsch, es el fruto de la decisión de la Organización de ofrecer a los profesionales de la salud de habla española una obra que aporta los principios básicos sobre los que debe descansar la crítica metodológica precisa de un trabajo de investigación médica.

Estamos convencidos de que un libro que presenta los principios y la aplicación de los métodos estadísticos y epidemiológicos de forma sencilla, clara y, sobre todo, destinado a un público no matemático, es una herramienta de inigualable valor. Nuestro deseo es que la obra contribuya a mejorar el proceso de administración del conocimiento en el campo de la salud pública y, por ende, nos aproxime a la meta de salud para todos en el año 2000.

Carlyle Guerra de Macedo
Director

Acerca de los autores

RICHARD K. RIEGELMAN es Profesor de Ciencias de Atención de la Salud en la Facultad de Medicina y Ciencias de la Salud de la Universidad George Washington y Director del Programa de Maestría en Salud Pública de la misma universidad. Además, es médico del Servicio de Medicina del Hospital de la Universidad George Washington de Washington, DC, EUA.

ROBERT P. HIRSCH es Profesor Asociado y Jefe Asociado del Servicio de Ciencias de Atención de la Salud de la Facultad de Medicina y Ciencias de la Salud de la Universidad George Washington de Washington, DC, EUA.

Prefacio

Enfrentado a una avalancha de investigaciones médicas, ¿qué debe hacer el clínico? ¿Cómo podemos evaluar con exactitud y de forma eficiente esta información e incorporarla en nuestra práctica clínica? El fin que persigue *Cómo estudiar un estudio y probar una prueba: lectura crítica de la literatura médica*, Segunda edición, es enseñar un método práctico y progresivo de lectura cuidadosa, crítica y, en última instancia, más eficaz de la literatura médica.

Esta segunda edición está basada en los enfoques utilizados en la primera edición. En el nuevo material introducido sobre los ensayos clínicos controlados se examina lo que debe hacerse e ilustra los errores que se pueden cometer en cada paso del proceso. A lo largo de todo el libro se encuentran ejercicios para detectar errores, que se han preparado expresamente para esta segunda edición. Un esquema reorganizado y ampliado sobre pruebas estadísticas resume los nuevos capítulos de la Sección 4, *La selección de una prueba estadística*. Esta segunda edición contiene numerosas notas a pie de página diseñadas para el lector interesado en la estadística o para utilizarlas en las aulas. No obstante, el texto básico sigue estando orientado hacia los clínicos que desean leer la literatura médica por sí solos o en un club de revistas.

Si bien se presentan conceptos estadísticos y se definen sus implicaciones y suposiciones, no se hace hincapié en los cálculos matemáticos. Por el contrario, el énfasis recae en la pregunta que se intenta responder con el estudio, en determinar si se utiliza una prueba estadística correcta y en entender el significado de los resultados.

A pesar de que los clínicos con experiencia pueden resumir sistemáticamente los puntos pertinentes de un documento de dos volúmenes, a menudo tienen dificultades para presentar un artículo de investigación de cuatro páginas. Para satisfacer esta necesidad, al lector de *Cómo estudiar un estudio y probar una prueba* se le ofrece un marco uniforme sobre el que puede basar su crítica de cualquier artículo de investigación. Una lista de preguntas que el lector debe formularse cuando evalúe un estudio también le ayuda a desarrollar un enfoque sistemático que acelerará su análisis de los artículos.

Cualquier método de lectura de la literatura médica debe basarse en, y ser compatible con, la capacitación clínica. En el diagnóstico diferencial de las enfermedades, un marco organizado y estructurado ayuda a los clínicos a reflexionar sobre los problemas con rapidez. Del mismo modo que al aprender a realizar una exploración física, la atención del lector se dirige inicialmente a los componentes individuales del proceso. Los resúmenes en cápsulas de los artículos de revistas biomédicas, cada uno de ellos dedicados a un tipo de error específico, ilustran y cristalizan los conceptos esenciales. El tiempo dedicado a aprender los principios básicos se traduce en la capacidad de comprender más rápidamente el significado y las limitaciones de cualquier estudio que uno pueda encontrar.

El método del estudio de casos utilizado en el libro es paralelo a la formación en las salas de urgencias o en las conferencias clinicopatológicas, pues ofrece la capacitación activa necesaria para internalizar los conceptos y ganar dominio en su uso. A final de las primeras tres secciones figuran ejercicios para detectar errores —artículos simulados repletos de distintos errores que constituyen un campo de prácticas para la aplicación del marco uniforme. Un ejemplo de crítica de cada uno de esos ejercicios permite a los lectores evaluar su progreso en el análisis del material. Al final

de la Sección 4, se presenta un esquema que resume las diversas pruebas estadísticas y guía al lector a través de las diferentes etapas del razonamiento estadístico. Como en la capacitación clínica, la meta consiste en el análisis organizado de los datos, el cual proporcionará una base para la toma de decisiones.

El objetivo global es aprender los tipos de preguntas que los estudios y la estadística pueden responder y las que nosotros mismos debemos contestar. Es importante reconocer que los estudios que no alcanzan el ideal deseable no invalidan necesariamente la investigación o eximen a los clínicos de la responsabilidad de extraer conclusiones clínicas. Los clínicos que pueden leer críticamente la literatura médica y entender y aceptar la incertidumbre están en mejores condiciones de llegar a conclusiones llenas de sentido y de integrar los resultados de la investigación médica en la práctica clínica. La lectura de la literatura médica puede ser más que una responsabilidad. Esperamos haber aliviado un poco el dolor del proceso.

Sección 1

El estudio de un estudio

CAPÍTULO 1

INTRODUCCIÓN Y EJERCICIO DE PRUEBA

El curso tradicional de lectura de la literatura médica consiste en: "¡Aquí tiene la *New England Journal of Medicine*"; léala! Este método es análogo al de aprender a nadar por el método de la inmersión total. Desde luego, ciertas personas pueden aprender a nadar de esa forma, pero algunas se ahogan y muchas le toman miedo al agua. Leer solo los resúmenes de los artículos médicos es como tener miedo al agua.

En contraposición al método de la inmersión total, aquí presentaremos un método gradual y de participación activa para analizar la literatura médica. Con estas técnicas analíticas, el clínico debe ser capaz de leer un artículo de una revista crítica y eficientemente. Si bien se subrayarán los errores que pueden aparecer en los diversos tipos de estudios, recuerde que no todos los errores son fatales.

No obstante, antes de desarrollar e ilustrar los elementos de un análisis crítico, comencemos con un ejercicio para detectar errores constituido por un artículo de revista simulado y veamos qué tal lo hace. Lea el siguiente estudio y trate de responder a las preguntas que figuran a continuación.

UN ESTUDIO DEL TAMIZAJE MÉDICO EN UNA POBLACIÓN MILITAR

Durante el primer año de su servicio militar se ofreció a 10 000 soldados de 18 años de edad la oportunidad de someterse a un examen médico anual que constaba de una historia clínica, una exploración física y diversas pruebas de laboratorio. El primer año participaron 5 000 reclutas y los 5 000 restantes no lo hicieron. Los 5 000 participantes fueron seleccionados como grupo de estudio y los 5 000 que no participaron, como grupo control. A los que participaron durante el primer año se les ofreció la oportunidad de someterse anualmente a exámenes médicos de mantenimiento de la salud durante el resto de su servicio militar.

Al finalizar el servicio, se preparó la historia clínica completa de los 5 000 integrantes del grupo de estudio y de los 5 000 del grupo control y se les practicó una exploración física y una evaluación de laboratorio para determinar si las visitas anuales habían producido alguna diferencia en su salud y en su estilo de vida.

Los investigadores obtuvieron la siguiente información:

1. A partir del consumo de alcohol declarado, la tasa de alcoholismo de los participantes fue la mitad de la de los no participantes.
2. El número de diagnósticos establecidos entre los participantes fue el doble del de los realizados en los no participantes.
3. Los participantes habían tenido un promedio de ascensos dos veces más alto que los no participantes.
4. No se observaron diferencias estadísticamente significativas entre las tasas de infarto de miocardio (IM) de ambos grupos.
5. Tampoco se encontraron diferencias entre los grupos en cuanto a las tasas de aparición de cáncer de testículo o de enfermedad de Hodgkin, los dos tipos de cáncer más frecuentes en los hombres jóvenes.

En consecuencia, los autores llegaron a las siguientes conclusiones:

1. El tamizaje anual puede reducir a la mitad la tasa de alcoholismo en la población de militares.
2. Dado que el número de diagnósticos realizados en los participantes fue el doble del de los no participantes, sus enfermedades se diagnosticaron en una fase temprana, cuando el tratamiento es más beneficioso.
3. Como a los participantes se les habían concedido dos veces más ascensos que a los no participantes, el programa de tamizaje debe haber mejorado la calidad de su trabajo.
4. El tamizaje y la intervención sobre los factores de riesgo coronarios no se deben incluir en un futuro programa de tamizaje para el mantenimiento de la salud, ya que no se observaron diferencias entre las tasas de IM de los dos grupos.
5. Dado que la frecuencia de la enfermedad de Hodgkin y del cáncer de testículo fue igual en ambos grupos, los futuros exámenes para el mantenimiento de la salud no deben incluir pruebas para el diagnóstico de estas enfermedades.

Ahora veamos si usted puede responder a las siguientes preguntas, que forman parte del marco uniforme de revisión de los estudios médicos.

1. ¿Estaba el estudio diseñado apropiadamente para responder a las preguntas planteadas?
2. ¿Fue apropiado el método de asignación de los pacientes al grupo de estudio y al de control?
3. ¿Fue correcta la valoración de los resultados de los grupos de estudio y de control?
4. ¿Se comparó apropiadamente en el análisis el desenlace (*outcome*) del grupo de estudio con el del grupo control?
5. ¿Se obtuvo una interpretación válida basada en las comparaciones realizadas entre el grupo de estudio y el de control?
6. ¿Se efectuaron correctamente las extrapolaciones a los individuos no incluidos en el estudio?

¿Cómo le fue? Si usted cree que ya puede responder a estas preguntas, pase a la crítica que aparece en el capítulo 10 y compare sus respuestas. Cuando esté listo, ¡prosigamos!

CAPÍTULO 2

LOS MARCOS UNIFORMES

EL MARCO UNIFORME

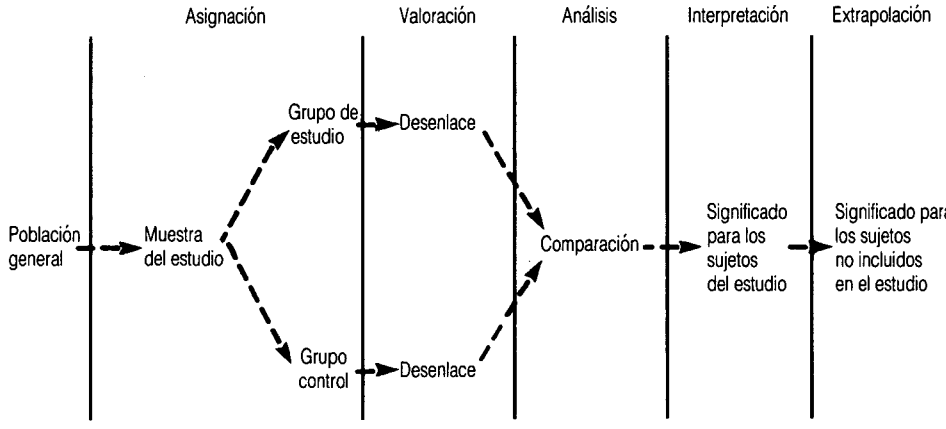
En la literatura médica se encuentran con frecuencia tres tipos básicos de estudios de investigación clínica: estudios retrospectivos o de casos y controles, estudios de cohortes o prospectivos, y ensayos clínicos aleatorios o ensayos clínicos controlados. Para evaluar los tres tipos de estudios se puede utilizar un marco uniforme. Este marco constituirá el fundamento de todo el proceso de *estudiar un estudio*. En la figura 2-1 se esboza la aplicación del marco uniforme a una investigación.

El marco uniforme contiene los siguientes elementos básicos:

ASIGNACIÓN. Selección de los individuos del grupo de estudio y del grupo control.
VALORACIÓN. Determinación de los resultados de la investigación en el grupo de estudio y en el de control.
ANÁLISIS. Comparación de los resultados del grupo de estudio y del grupo control.
INTERPRETACIÓN. Extracción de conclusiones sobre las diferencias encontradas entre el grupo de estudio y de control, y sobre su significado para los sujetos estudiados.
EXTRAPOLACIÓN. Extracción de conclusiones sobre el significado del estudio para los individuos o situaciones no incluidos en el mismo.

Para ilustrar la aplicación del marco uniforme a los estudios de casos y controles (retrospectivos), de cohortes (prospectivos) y a los ensayos clínicos aleatorios (ensayos clínicos controlados), primero esbozaremos las características distintivas de cada tipo de estudio y luego veremos cómo podríamos aplicar cada tipo de estudio al problema concreto de la relación entre el uso de estrógenos aislados (estrógenos sin progesterona) y el cáncer de endometrio.

FIGURA 2-1. Marco uniforme para la revisión de un estudio



Estudio de casos y controles (estudio retrospectivo)

La característica específica de los *estudios de casos y controles* o *retrospectivos* es que se inician después de que los individuos hayan desarrollado (o hayan dejado de hacerlo) la enfermedad investigada. Estos estudios se dirigen hacia atrás en el tiempo para determinar las características que esos individuos presentaban antes del inicio de la enfermedad. En los estudios de casos y controles, los "casos" son los individuos que ya han desarrollado la enfermedad y los controles, los que no la han desarrollado. Para utilizar este tipo de estudio con objeto de examinar la relación entre la toma de estrógenos y el cáncer de endometrio, un investigador procedería de la siguiente forma:

ASIGNACIÓN. Seleccionar un grupo de estudio formado por mujeres que actualmente tienen un cáncer de endometrio (casos) y un grupo de mujeres sin cáncer de endometrio, pero similares a las primeras respecto a las demás características (controles). Como la enfermedad ha evolucionado sin la participación del investigador, el proceso se denomina *asignación observada*.

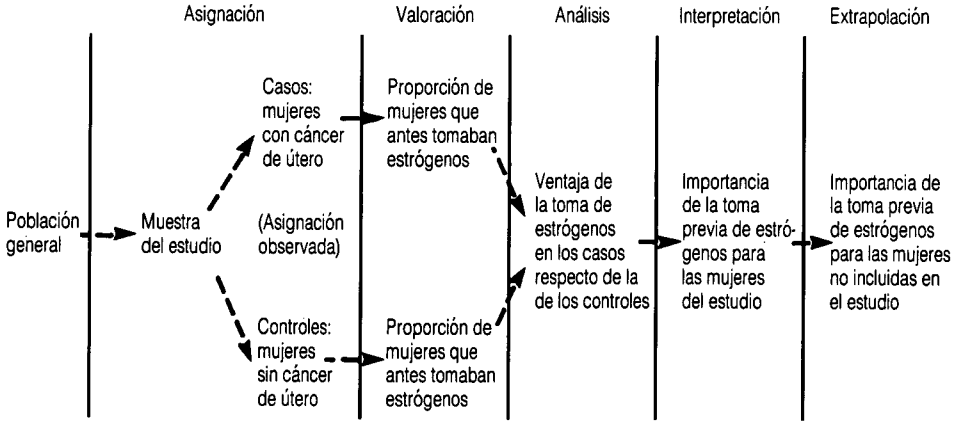
VALORACIÓN. Determinar si cada mujer del grupo de estudio y del grupo control había tomado antes estrógenos sin progesterona y, si así fuese, calcular la cantidad.

ANÁLISIS. Calcular la *ventaja* (*odds*)[1] de que las mujeres del grupo con cáncer endometrial hayan tomado estrógenos sin progesterona respecto de la ventaja de que los hayan tomado las mujeres del grupo sin cáncer endometrial.

INTERPRETACIÓN. Extraer conclusiones sobre el significado de la toma de estrógenos en las mujeres estudiadas.

EXTRAPOLACIÓN. Extraer conclusiones sobre el significado de la toma de estrógenos para las categorías de mujeres no incluidas en el estudio, tales como las tratadas con una dosis igual o distinta, o las que toman estrógenos combinados con progesterona. La figura 2-2 muestra la aplicación del marco uniforme a este estudio.

FIGURA 2-2. Aplicación del marco uniforme a un estudio retrospectivo o de casos y controles



[1] N del E. *Ventaja* (*odds*) = razón entre la probabilidad de que se produzca un hecho (*p*) y la probabilidad de que no se produzca (1-*p*). Por lo tanto, ventaja = *p*/(1-*p*).

Estudio de cohortes (prospectivo)

Los *estudios de cohortes* o *prospectivos* se diferencian de los de casos y controles en que se inician antes de que los individuos hayan desarrollado la enfermedad investigada, a los cuales se sigue durante un período para determinar quiénes desarrollarán la enfermedad. Una *cohorte* es un grupo de individuos que comparten una experiencia. En estos estudios se sigue a una cohorte que posee las características estudiadas y a una cohorte que no las posee. Para utilizar un estudio de cohortes con objeto de examinar la relación entre la toma de estrógenos y el cáncer endometrial, un investigador procedería de la siguiente forma:

ASIGNACIÓN. Seleccionar a un grupo de estudio formado por mujeres que toman estrógenos sin progesterona y a un grupo de control integrado por mujeres similares, pero que no han tomado estrógenos. Como las primeras toman estrógenos sin la intervención del investigador, el proceso se denomina también asignación observada.

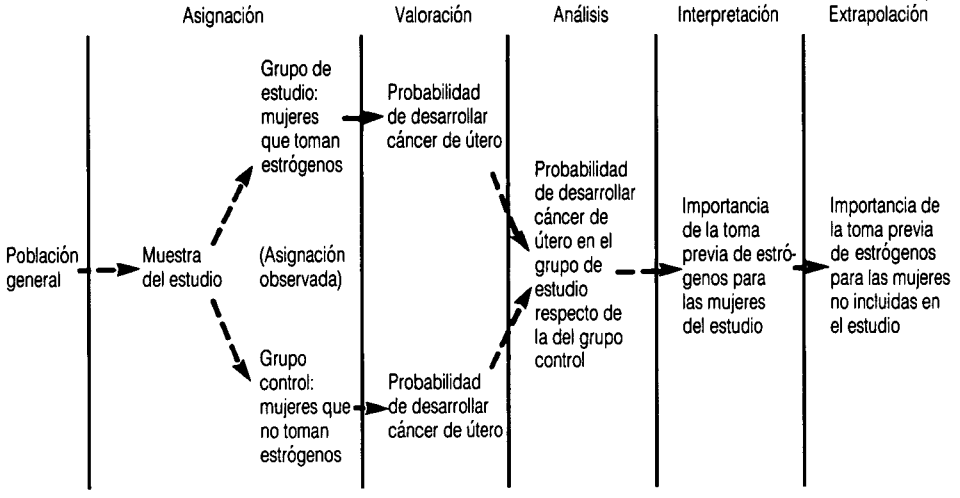
VALORACIÓN. Seguir a las mujeres del grupo de estudio y del grupo control para determinar cuáles desarrollarán cáncer de endometrio.

ANÁLISIS. Calcular la probabilidad de desarrollar cáncer de endometrio en el grupo de mujeres que toman estrógenos en relación con la de las que no los toman.

INTERPRETACIÓN. Extraer conclusiones sobre el significado de la toma de estrógenos en las mujeres estudiadas.

EXTRAPOLACIÓN. Extraer conclusiones sobre la toma de estrógenos para las mujeres no incluidas en el estudio, como las que toman una dosis igual o distinta, o las que toman estrógenos combinados con progesterona. La figura 2-3 muestra la aplicación del marco uniforme a un estudio de cohortes o prospectivo.

FIGURA 2-3. Aplicación del marco uniforme a un estudio prospectivo o de cohortes



Ensayo clínico aleatorio (ensayo clínico controlado)

Los *ensayos clínicos aleatorios* también se denominan *ensayos clínicos controlados*. Como en los estudios de cohortes, los individuos se siguen durante un período para determinar si desarrollan la enfermedad concreta o trastorno investigado. La

característica distintiva de los estudios experimentales, sin embargo, es el método de asignación de los individuos a los grupos de estudio y de control. En condiciones ideales, los individuos se asignan al azar y a ciegas tanto al grupo de estudio como al de control. La *asignación al azar* significa que cualquier individuo tiene una probabilidad conocida de ser asignado al grupo de estudio o al de control. La *asignación a doble ciego* —que es el método ideal de enmascaramiento (*matching o blinding*)— indica que ni los participantes ni los investigadores saben si un participante concreto ha sido asignado al grupo de estudio o al de control. Para utilizar un ensayo clínico aleatorio con el fin de investigar la relación entre la toma de estrógenos y el cáncer de endometrio, un investigador procedería de la siguiente forma:

ASIGNACIÓN. Mediante la asignación al azar, las mujeres se asignan al grupo de estudio, en el que tomarán estrógenos, o al grupo control, en el que no los tomarán.

VALORACIÓN. Seguimiento de esas mujeres en el tiempo para determinar cuál desarrollará cáncer.

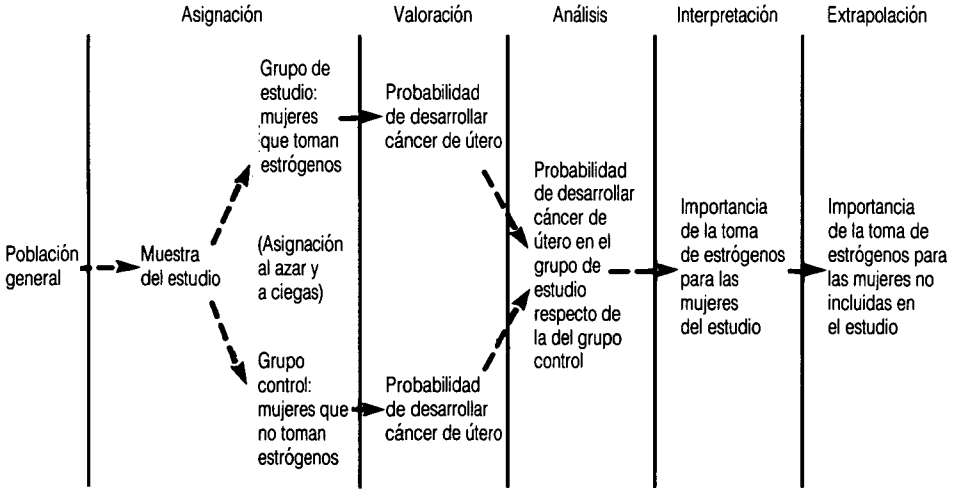
ANÁLISIS. Cálculo de la probabilidad de que las mujeres que toman estrógenos desarrollen un cáncer de endometrio respecto a la de las que no los toman.

INTERPRETACIÓN. Extracción de conclusiones sobre el significado de la toma de estrógenos sin progesterona en las mujeres estudiadas.

EXTRAPOLACIÓN. Extracción de conclusiones sobre el significado de la toma de estrógenos sin progesterona para las mujeres no incluidas en el estudio, como las que toman dosis distintas o estrógenos combinados con progesterona. La figura 2-4 ilustra la aplicación del marco uniforme a los ensayos clínicos aleatorios o ensayos clínicos controlados.

Esta breve presentación de los tres tipos básicos de estudios utilizados en la investigación clínica tiene por objeto mostrar cómo se pueden analizar estos estudios utilizando el marco uniforme. En el capítulo 8 se comentarán las ventajas y desventajas de cada tipo de estudio. Antes de proseguir con el diseño general de los estudios, revisaremos los requisitos necesarios para aplicar correctamente cada componente del marco uniforme y ejemplificaremos los errores que se cometen con más frecuencia.

FIGURA 2-4. Aplicación del marco uniforme a un estudio experimental



CAPÍTULO 3

ASIGNACIÓN

ESTUDIOS OBSERVACIONALES

En los próximos seis capítulos aplicaremos el marco uniforme a los estudios de casos y controles y a los de cohortes. Estos estudios se conocen conjuntamente bajo el nombre de *estudios observacionales*. En un estudio observacional no se intenta intervenir ni alterar el curso de la enfermedad. Los investigadores observan el curso de la enfermedad en los grupos con y sin las características estudiadas.

Las investigaciones se realizan generalmente utilizando una muestra o subgrupo de individuos seleccionados a partir de una población mayor. Los sujetos elegidos pueden o no haber sido seleccionados de la población mediante un proceso aleatorio. Por eso, los grupos de estudio y de control no son necesariamente representativos de todos los de la población. No obstante, el investigador define las características de los individuos elegibles para el grupo de estudio y de control con objeto de formar grupos de estudio y de control tan idénticos como sea posible excepto por la característica estudiada. Una posible razón de que no se cumpla con este objetivo es la presencia de un sesgo de selección.

En medicina hay pocos términos que se entiendan menos claramente o que se usen con menos precisión que la palabra sesgo. El *Webster's New World Dictionary* define sesgo (*bias*) como "un prejuicio, juicio u opinión formado antes de que se conozcan los hechos".[1] Según la mejor tradición de la investigación científica, un estudio debe estar libre de prejuicio. Incluso con la mejor intención científica, los investigadores pueden introducir en la investigación de forma no intencionada factores capaces de predeterminar el resultado del estudio. La presencia de estos factores puede crear un sesgo de selección. En el siguiente estudio hipotético se muestran los elementos del sesgo de selección.[2]

En un estudio de casos y controles para investigar las causas del cáncer de mama en la premenopausia se comparó la toma anterior de píldoras anticonceptivas en 500 mujeres con cáncer de mama, que fueron apareadas con 500 mujeres hospitalizadas por diabetes o hipertensión. Los investigadores observaron que 40% de las mujeres con cáncer de mama habían tomado píldoras anticonceptivas durante los cinco años anteriores, mientras que solo 5% de las mujeres con diabetes o hipertensión del grupo control las habían tomado. Los autores concluyeron que existía una fuerte asociación entre la toma de píldoras anticonceptivas y la aparición de cáncer de mama en la premenopausia.

Para averiguar si existió un sesgo de selección en la asignación de las pacientes al grupo control, debemos preguntarnos, en primer lugar, si las pacientes de ese grupo eran similares a la población de mujeres sin cáncer de mama. La respuesta es negativa, ya que esas mujeres eran singulares por el hecho de haber sido hospitalizadas por diabetes o hipertensión. Hemos de preguntarnos si era probable que esta sin-

[1] *Webster's New World Dictionary of the American Language*. (College Edition). Cleveland: World Publishing Company; 1966, p. 1150.

[2] En la revisión de este caso hipotético, así como en los restantes del libro, el lector debe suponer que las partes del estudio omitidas se realizaron correctamente.

gularidad influyese en el uso de la característica estudiada, esto es, de las píldoras anticonceptivas. La respuesta es positiva. Dado que es ampliamente conocido que las píldoras anticonceptivas elevan la presión arterial y la glucemia, no es probable que los médicos receten píldoras anticonceptivas a las mujeres con hipertensión o diabetes. La singularidad de la salud de esas mujeres contribuyó a que tomaran menos píldoras de lo esperado. Por consiguiente, este estudio creó el potencial para un sesgo de selección en la asignación de las pacientes.

Así, el sesgo de selección puede aparecer cuando el grupo de estudio y el de control difieren entre sí en algún factor que puede influir en la medición del desenlace estudiado. En otras palabras, el sesgo de selección aparece cuando la forma en que difieren los grupos origina una diferencia en el desenlace.

El siguiente ejemplo ilustra el sesgo de selección que puede aparecer en un estudio de cohortes:

Para estudiar el efecto del consumo de cigarrillos sobre el desarrollo del infarto de miocardio, se seleccionaron 10 000 fumadores de cigarrillos y 10 000 fumadores de pipa, pero no de cigarrillos. Los investigadores observaron que la tasa de infartos de miocardio en los fumadores de cigarrillos fue 4 por 100 en 10 años, mientras que la de los fumadores de pipa fue 7 por 100 en 10 años. Los resultados fueron estadísticamente significativos. Los investigadores concluyeron que los fumadores de cigarrillos estaban en menor riesgo de padecer un infarto de miocardio que los fumadores de pipa.

A pesar de que la diferencia fue estadísticamente significativa, la conclusión se contradice con los resultados de muchos otros estudios. Veamos si un sesgo de selección pudo influir en los resultados obtenidos.

Al analizar este estudio se deben reconocer dos hechos generalmente aceptados: la gran mayoría de los fumadores de pipa son hombres y en ellos la tasa de infartos de miocardio es más alta que en las mujeres.

Teniendo en cuenta este hecho, la primera cuestión que surge es si los grupos de control y de estudio difieren. La respuesta es afirmativa, dado que los hombres constituyen la gran mayoría de fumadores de pipa mientras que las mujeres fuman más cigarrillos que en pipa. Para investigar la posibilidad de la presencia de un sesgo de selección, tenemos que preguntarnos si esta diferencia influye en el desenlace que se pretende medir. Nuevamente, la respuesta es positiva. El riesgo de infarto de miocardio en los hombres es mayor. Por lo tanto, los dos elementos de un sesgo de selección están presentes. Los grupos difieren de una forma que influye en el desenlace.

Aun cuando el sesgo de selección sea improbable, el azar por sí solo puede contribuir a que los grupos de estudio y de control difieran en los factores de riesgo del desarrollo de la enfermedad o en los factores pronósticos que influyen en el desenlace de la enfermedad. Cuando estas diferencias en los factores de riesgo influyen en el desenlace, se denominan *variables de confusión* (*confounding variables*). De esta forma, el sesgo de selección es un tipo especial de variable de confusión, que resulta de la presencia de un sesgo basado en la forma en que se han seleccionado los sujetos del grupo de control o de estudio. Recuerde que, incluso en ausencia de un sesgo de selección, las diferencias en las variables de confusión pueden ser producidas solo por azar. Es importante comparar los sujetos del grupo de estudio y del de control para determinar si la forma en que difieren puede influir en el desenlace. En los capítulos 5 y 29 se presentarán los métodos que se pueden emplear para tratar las variables de confusión, pero, antes, daremos un vistazo a los tipos de problemas que pueden surgir en la valoración del desenlace.

CAPÍTULO 4

VALORACIÓN DEL DESENLACE

Para valorar el desenlace de una investigación, los investigadores deben definir el *desenlace* o *resultado* (*outcome*) que pretenden medir. El término *desenlace* resulta un tanto confuso, porque tiene diferentes significados en los diversos tipos de estudios. Analicemos qué quiere decir *desenlace* en un estudio de casos y controles y en uno de cohortes, y definamos, luego, los criterios para efectuar una medición válida del desenlace. Los estudios de cohortes se inician con un grupo de estudio que posee la característica estudiada y un grupo de control que no la tiene. Los individuos del grupo de estudio y del de control se siguen durante un período para determinar quiénes desarrollan una enfermedad concreta. La aparición de la enfermedad que se estudia se conoce como desenlace o resultado.

El investigador debe emplear una medida válida de la aparición de la enfermedad. Por ejemplo, en el caso de los ejemplos referentes a los estrógenos y el cáncer endometrial, el desarrollo del cáncer es el desenlace estudiado por los investigadores.

Los estudios de casos y controles se inician con personas que ya han desarrollado una determinada enfermedad o trastorno (casos) y con personas que no la han desarrollado (controles). Los investigadores examinan la historia previa de los grupos de casos y de controles para determinar si los individuos poseían o habían estado expuestos anteriormente a una característica. En un estudio de casos y controles, esta característica previa es el desenlace del estudio. Los investigadores deben utilizar una medida válida del desenlace o de la característica previa. En el caso de los estrógenos y el cáncer endometrial, la toma de estrógenos sin progesterona es la característica previa que se debe valorar.

¿Qué es una medida adecuada de un desenlace? La que cumple todos los criterios siguientes:

1. El investigador debe usar una medida apropiada para responder a la cuestión planteada en el estudio.
2. La medida del desenlace debe ser exacta. (Ha de aproximarse a la medición verdadera del fenómeno.)
3. La medida del desenlace debe ser completa.
4. La medida del desenlace utilizada en el estudio no debe estar influida por el proceso de observación.

MEDIDA ADECUADA DEL DESENLACE

Para comprender la importancia que tiene el disponer de una medida apropiada del desenlace, primero analizaremos un ejemplo que ilustra cómo el empleo de una medida inadecuada del desenlace puede invalidar las conclusiones de un estudio.

Un investigador realizó un estudio para averiguar si el uso de espermicidas de la marca A estaba asociado con una probabilidad menor de desarrollar infecciones tubáricas por *Chlamydia* que el uso de espermicidas de la marca B. Para ello, se seleccionaron 100 mujeres que usaban una u otra marca del espermicida, se tomaron

frotis cervicales para cultivo y se siguió su evolución durante 5 años. El investigador observó que las mujeres que usaban la marca A del espermicida tenían la mitad de cultivos positivos para *Chlamydia* y concluyó que el espermicida de la marca B se asociaba con una menor tasa de infecciones tubáricas.

Los cultivos de cervix para *Chlamydia* no son adecuados para detectar la presencia de infección tubárica. El estudio puede contribuir a establecer una frecuencia más alta de infección por *Chlamydia*. Sin embargo, el investigador no escogió una medida apropiada del desenlace, si su intención era la de estudiar la frecuencia relativa de esta infección.

MEDIDA EXACTA DEL DESENLACE

Seguidamente, veremos cómo una medición inexacta puede influir en la valoración de un desenlace. La información para medir un desenlace puede proceder de tres fuentes distintas:

1. Lectura de los instrumentos de medida
2. Mediciones del investigador
3. Informes o registros obtenidos de individuos

La información obtenida puede ser inexacta, porque los datos producidos estén sistemáticamente fuera del objetivo y en la misma dirección debido a un sesgo en la forma como se recogieron los datos. Otra posibilidad es que los datos sean inexactos debido a una variación al azar en cualquier dirección.

La información de los individuos estudiados está sujeta a los sesgos de recuerdo y de declaración. El *sesgo de recuerdo* presupone defectos de memoria, en particular, cuando es más probable que los individuos de un grupo recuerden ciertos sucesos que los de otros grupos. El *sesgo de declaración* se produce en los estudios de casos y controles cuando los sujetos de un grupo de estudio relatan con más exactitud sus recuerdos que los del otro grupo. Considere el siguiente ejemplo sobre la forma como puede aparecer un sesgo de recuerdo.

En un estudio de casos y controles sobre la causa de la espina bífida se estudiaron 100 madres cuyos hijos nacieron con la enfermedad y 100 madres cuyos hijos nacieron sin la enfermedad. De las madres de hijos con espina bífida, 50% declararon haber padecido dolor de garganta durante el embarazo, mientras que entre las madres de hijos sin la enfermedad solo lo declararon 5%. Los investigadores llegaron a la conclusión de que habían demostrado una asociación entre el dolor de garganta y la espina bífida.

Antes de aceptar las conclusiones del estudio, uno debe preguntarse si los resultados podrían explicarse por la presencia de un sesgo de recuerdo. Se puede afirmar que es más probable que las madres que experimentaron el traumatismo de tener un hijo con espina bífida recavaran en su memoria y recordaran sucesos que habitualmente no se recuerdan con más intensidad que el resto de mujeres. Por lo tanto, es más probable que el sesgo de recuerdo aparezca cuando los sucesos son traumáticos, ya que estas experiencias motivan a recordar subjetivamente sucesos que ocurren frecuentemente y que en circunstancias normales se olvidarían. Por consiguiente, el resultado de este estudio de casos y controles se puede atribuir, al menos parcialmente, al sesgo de recuerdo. La posibilidad de que exista un sesgo de recuerdo arroja dudas sobre la supuesta asociación entre el dolor de garganta y la espina bífida.

El sesgo de declaración, como el sesgo de recuerdo, puede disminuir la exactitud de la valoración del desenlace, como se muestra en el siguiente ejemplo.

En un estudio sobre la relación entre la gonorrea y el número de compañeros sexuales, se comparó a 100 mujeres recién diagnosticadas de gonorrea con 100 mujeres a las que no se diagnosticó la enfermedad y que fueron atendidas en el mismo consultorio. A las mujeres diagnosticadas de gonorrea se les informó que solo se podían prevenir las graves consecuencias de la enfermedad si se localizaba y se trataba a sus compañeros sexuales. A las mujeres de ambos grupos se les preguntó el número de compañeros sexuales que habían tenido en los dos meses precedentes. Las mujeres con gonorrea declararon haber tenido en promedio el doble de compañeros sexuales que las mujeres sin gonorrea. Los investigadores concluyeron que las mujeres con gonorrea tenían el doble de compañeros sexuales que las mujeres sin gonorrea.

Puede suponerse que, en este estudio, las mujeres con gonorrea se sintieron más obligadas y, por tanto, menos reacias a ofrecer información acerca de sus compañeros sexuales que las mujeres sin gonorrea. Es más probable que el sesgo de declaración se cometa cuando la información que se busca es personal o delicada. Además, un grupo ha sido presionado más que el otro para que informe exactamente sobre los sucesos anteriores. De este modo, es posible, simplemente, que las mujeres con gonorrea hayan sido más cuidadosas al declarar el número de compañeros sexuales sin que en realidad hayan tenido más relaciones. El error de declaración juntamente con el de recuerdo pueden alterar la exactitud de la valoración de los estudios de casos y controles.

Error del instrumento

El error de medición también se puede deber a la falta de exactitud de los instrumentos empleados, como se demuestra en el siguiente ejemplo.

Para valorar los efectos secundarios gastrointestinales de dos medicamentos antiinflamatorios no esteroideos en el tratamiento de la artritis, se tomaron radiografías esófago-gastro-duodenales de varios pacientes. El investigador no encontró pruebas que apoyaran la existencia de una asociación entre esos fármacos y la gastritis.

No obstante, no tuvo en cuenta que una radiografía esófago-gastro-duodenal es una prueba inadecuada para diagnosticar la gastritis. Aunque un fármaco causara gastritis, esta prueba no sería suficiente para identificar su presencia. De este modo, es probable que cualquier conclusión basada en esa medición sea inexacta. Cuando se comete un craso error del instrumento, como en este caso, la medida del desenlace también se puede considerar inadecuada.

Sesgo del investigador

La posibilidad de un sesgo del investigador existe en todos aquellos casos en que la medida del desenlace depende de que el investigador interprete subjetivamente los datos. Sin embargo, es posible reconocer y corregir un principio fundamental de la psicología humana: el de que las personas, investigadores incluidos, ven aquello que quieren o esperan ver. Esto se consigue evitando que el investigador que valora el desenlace conozca el grupo al que se ha asignado un individuo. La valoración a ciegas puede emplearse en los estudios de casos y controles y en los de cohortes, así como en los ensayos clínicos aleatorios. El no utilizarla puede conducir al siguiente tipo de sesgo.

En un estudio sobre el uso de antiinflamatorios no esteroideos, los investigadores —que eran los médicos que atendían a los pacientes— interrogaron a todos sus pacientes para determinar si la administración de alguno de esos medicamentos estaba asociada con una frecuencia más elevada de síntomas compatibles con

gastritis. Después de interrogar a los pacientes sobre sus síntomas, determinaron que no existía ninguna diferencia en el número de casos con gastritis. Los investigadores concluyeron que los dos fármacos producían el mismo número de casos de gastritis sintomáticos.

En este estudio, los investigadores que realizaron la valoración del desenlace sabían cuáles eran los pacientes que tomaban cada fármaco y, por tanto, no la estaban efectuando "a ciegas". Además, valoraron síntomas subjetivos como la náusea, el dolor de estómago o la indigestión, para determinar la presencia de gastritis. Esta es la situación en la cual el enmascaramiento desempeña el papel más importante. Aunque los pacientes no supieran qué fármaco estaban tomando, la valoración de los investigadores podía estar sesgada. Si su valoración fuera compatible con su propia hipótesis, sus resultados serían especialmente cuestionables. Esto no quiere decir que sean fraudulentos, sino que solamente muestran la tendencia natural de los seres humanos a ver lo que se quiere o se espera ver. Las conclusiones de los investigadores podrían ser ciertas, pero sus técnicas imperfectas harían difícil o imposible aceptarlas. De esta forma, el enmascaramiento en el proceso de valoración es una medida importante para eliminar el sesgo.

INTEGRIDAD DE LA VALORACIÓN

Cuando el seguimiento de los pacientes es incompleto, existe la posibilidad de que la frecuencia del desenlace en los que no fueron incluidos en la valoración final sea distinta de la frecuencia en los que fueron incluidos. El siguiente estudio ilustra un error debido a la valoración incompleta.

En un estudio de una cohorte de pacientes positivos al virus de la inmunodeficiencia humana (VIH), se comparó la historia natural de la enfermedad en pacientes asintomáticos con un recuento de células T4 entre 100 y 200 con la de un grupo de pacientes asintomáticos positivos al VIH y con un recuento entre 200 y 400 células. Los investigadores siguieron a 50% de los que tenían recuentos bajos de células T4 y a 60% del otro grupo. No encontraron diferencias entre ambos grupos y llegaron a la conclusión de que el recuento de células T4 no era un factor de riesgo para desarrollar el síndrome de la inmunodeficiencia adquirida (SIDA).

En este estudio es lícito argumentar que no se pudo seguir la evolución de varios pacientes porque habían fallecido. Si este fuera el motivo, el seguimiento completo podría haber modificado espectacularmente los resultados del estudio. El seguimiento incompleto puede distorsionar las conclusiones de una investigación.

El seguimiento incompleto no significa necesariamente que no se pueda seguir a los pacientes, como ocurrió en el ejemplo anterior; el seguimiento puede ser distinto para cada paciente, como se muestra en el siguiente ejemplo.

Se realizó un estudio de cohortes sobre los efectos secundarios de los anticonceptivos orales comparando 1 000 mujeres jóvenes que los tomaban con 1 000 que utilizaban otros métodos de planificación familiar. Los datos se obtuvieron de los registros de sus médicos privados durante un pèriodo de un año. Se citó a tres visitas de seguimiento durante el año a las mujeres que tomaban anticonceptivos orales, mientras que a las restantes se les pidió que volvieran si tenían problemas. Entre las usuarias de anticonceptivos orales, 75 mujeres declararon haber padecido cefaleas, 90, fatiga y 60, depresión. Entre las no usuarias, 25 declararon haber padecido cefaleas, 30, fatiga y 20, depresión. Las usuarias de anticonceptivos orales realizaron en promedio tres visitas al médico durante el año por cada visita de las no usuarias. El investigador conluyó que el uso de los anticonceptivos orales estaba asociado con un aumento de la frecuencia de cefaleas, fatiga y depresión.

El problema de una observación desigual de los dos grupos puede invalidar los resultados. El hecho de que las usuarias de anticonceptivos orales hicieran 3 veces más visitas que las no usuarias puede explicar que las cefaleas, la fatiga y la depresión se registraran con mayor frecuencia. Mientras más frecuentes son las observaciones, mayor es la posibilidad de que se declaren los síntomas más comunes.

EFECTO DE LA OBSERVACIÓN

Aunque el resultado de un estudio cumpla los difíciles criterios de una valoración apropiada, exacta y completa, todavía existe un área de preocupación. Los investigadores intentan medir los sucesos como hubieran ocurrido si nadie los hubiese observado. Por desgracia, el proceso real de llevar a cabo un estudio puede incluir la introducción de un observador en los sucesos que se miden. El revisor de un artículo debe determinar si el proceso de observación modificó el resultado. A continuación figura un ejemplo en el que esto pudo haber ocurrido.

En un estudio de cohortes para investigar la relación entre la obesidad y la regularidad de las menstruaciones, se compararon 100 mujeres obesas con irregularidades menstruales que se habían inscrito en un grupo para bajar de peso con 100 mujeres obesas con el mismo patrón de irregularidades menstruales, pero que no se habían inscrito en dicho grupo. Los grupos se compararon para valorar los efectos de la pérdida de peso sobre las irregularidades menstruales. La frecuencia de retorno a los ciclos menstruales regulares de las mujeres del grupo de reducción de peso fue la misma que la de las controles.

Es posible que las mujeres del grupo control perdieran el mismo peso que las mujeres del grupo de reducción de peso, ya que estaban siendo observadas como integrantes del estudio. Los efectos de la observación pueden influir en una investigación cuando es posible que los sujetos del estudio cambien de grupo o modifiquen su comportamiento. La posibilidad de que esto ocurra es mayor cuando los pacientes del grupo control son conscientes de las consecuencias adversas de su comportamiento actual y se encuentran en una situación de presión directa o indirecta para cambiar debido al proceso de observación.

CAPÍTULO 5

ANÁLISIS

En este capítulo introduciremos tres funciones fundamentales del análisis:

1. Eliminar los efectos de las variables de confusión.
2. Contrastar las hipótesis que permiten al investigador extraer conclusiones relacionadas con diferencias entre poblaciones a partir de muestras de esas poblaciones.
3. Medir la magnitud de las diferencias entre grupos o la fuerza de las relaciones entre las variables observadas en el estudio.

Como hemos comentado en el capítulo 3, las variables de confusión pueden ser el resultado tanto del azar como de un sesgo. El azar es un problema inevitable siempre que se obtienen muestras de poblaciones y se desea extraer conclusiones sobre esas poblaciones. En contraposición al sesgo, el efecto del azar es impredecible.[1] Además, puede favorecer o contradecir la hipótesis del estudio en una forma que no podemos conocer de antemano.

El sesgo, por otro lado, traduce la existencia de un efecto sistemático sobre los datos en una dirección determinada que, de forma predecible, favorece o contradice la hipótesis del estudio. El sesgo es el resultado del modo en que se asigna o se evalúa a los pacientes.

El sesgo y el azar pueden producir diferencias entre las variables de confusión que ocasionen que los grupos de estudio y de control difieran de una forma tal que pueda afectar al desenlace del estudio. Comencemos nuestros comentarios sobre los análisis examinando las técnicas disponibles para tratar las variables de confusión. Las técnicas básicas para eliminar los efectos de los sesgos son el apareamiento (*matching*) de las muestras del grupo de estudio y del grupo control al inicio del estudio y el ajuste de los datos como parte del análisis.

APAREAMIENTO PREVIO

Un método para solventar el problema de las variables de confusión consiste en aparear individuos similares respecto de las variables de confusión potenciales. Por ejemplo, si la edad se relaciona con la probabilidad de pertenecer a un grupo dado y con el desenlace, el investigador puede aparear a los sujetos del estudio según la edad. Por cada persona de 65 años de edad en el grupo control, el investigador puede escoger una de 65 años para el grupo de estudio, y proceder del mismo modo con las de 30 años, 40 años, etc. Cuando el apareamiento se realiza correctamente, su resultado garantiza que la distribución de la edad en cada grupo será similar.

El apareamiento no se limita a formar grupos uniformes según la edad. También puede emplearse para factores de riesgo o de pronóstico; es decir, para

[1] Algunos efectos del azar son predecibles. Por ejemplo, en un estudio de casos y controles, la determinación imprecisa de la presencia o ausencia de enfermedad subestimará la ventaja (*odds*).

cualquier factor relacionado con la probabilidad de experimentar el desenlace estudiado. El apareamiento está especialmente indicado para reducir la probabilidad del sesgo de selección. Por ejemplo, si se estudia la relación entre los anticonceptivos orales y el accidente vascular cerebral (AVC), la presión arterial podría considerarse como un factor de riesgo o de pronóstico importante. Dado que la presión arterial elevada es una contraindicación relativa del uso de anticonceptivos orales, la presión arterial elevada debe reducir la probabilidad de que una mujer los tome. Además, el hecho de que la presión arterial elevada aumenta la probabilidad de padecer un AVC, podría influir en la probabilidad de que se produzca el desenlace. Por lo tanto, la presión arterial elevada es una variable de confusión según la cual se deben aparear los grupos.

Una desventaja del apareamiento por grupos es que los investigadores no pueden estudiar los efectos del factor de apareamiento respecto del desenlace.[2] Por ejemplo, si se realiza el apareamiento según la edad y la presión arterial, se pierde la capacidad de estudiar el efecto de estas dos variables sobre la aparición de AVC. También se pierde la capacidad de estudiar los factores que están estrechamente asociados con el factor por el que se aparea. El peligro de intentar estudiar el factor de apareamiento o los factores estrechamente relacionados con él se demuestra en el siguiente ejemplo.

Cien sujetos con diabetes del adulto se compararon con cien sin diabetes para estudiar los factores asociados con esa enfermedad. Los grupos fueron apareados con objeto de garantizar que la distribución del peso fuese similar en cada grupo. Los autores observaron que la distribución de las calorías totales consumidas era muy similar en ambos grupos y concluyeron que el número de calorías consumidas no estaba relacionado con la posibilidad de desarrollar diabetes del adulto.

Los autores del estudio, que aparearon a los pacientes según su peso, intentaron estudiar, a continuación, las diferencias en el consumo de calorías. No sorprende que no encontraran diferencias en el consumo calórico entre los dos grupos apareados según el peso, dado que existe una alta correlación entre el peso y las calorías consumidas. No es posible investigar la posibilidad de que los factores de apareamiento o los factores estrechamente asociados con ellos estén asociados con la frecuencia del desenlace.

El tipo de apareamiento discutido en nuestro ejemplo de la diabetes se denomina *apareamiento por grupos*. Un segundo tipo de apareamiento es conocido simplemente como apareamiento (esto es, cuando en la investigación se incluye un grupo de estudio y un grupo control). Este tipo de apareamiento exige identificar a un individuo del grupo de estudio que pueda ser comparado con uno o más individuos del grupo control. El apareamiento es un método eficiente para eliminar sesgos.

A pesar de sus ventajas, este apareamiento presenta una desventaja peculiar. Muchas veces, identificar a un paciente del grupo control que posea los mismos factores de riesgo conocidos que el sujeto del grupo de estudio con el que se aparea plantea un problema. A veces, este problema puede solventarse utilizando un paciente como su propio control. Ello se puede realizar mediante el denominado *estudio cruzado* (*cross-over study*), en el cual se compara a los individuos con ellos mismos cuando toman y cuando no toman la medicación. Cuando estos tipos de estudios se realizan correctamente, permiten utilizar los mismos sujetos en el grupo de estudio y en el de control y aparear sus resultados, manteniendo de este modo muchos factores constantes. Como el individuo es su propio control, el apareamiento permite el empleo de prue-

[2] La variable que se aparea se puede estudiar en el contexto de su interacción con otras variables.

bas de significación estadística potentes que aumentan la probabilidad de detectar diferencias estadísticamente significativas para un determinado tamaño del grupo de estudio. Estas pruebas habitualmente se denominan *pruebas apareadas*.

Los estudios cruzados se deben usar con sumo cuidado, ya que pueden producir resultados erróneos, como se muestra en el siguiente estudio hipotético.

En una investigación sobre el efecto beneficioso de una medicación no narcótica para aliviar el dolor posoperatorio, se administró a 100 pacientes la medicación el primer día posoperatorio y el placebo, el segundo día. El grado de dolor de cada paciente se midió en el primero y el segundo día con una escala bien establecida de medición del dolor. Los investigadores no encontraron diferencias entre los niveles de dolor con y sin la medicación.

Cuando se evalúa un diseño cruzado es preciso tener presente la posibilidad de un efecto del tiempo y de un efecto tardío (*carry-over effect*) del tratamiento. Es de esperar que el dolor disminuya con el paso del tiempo después de la cirugía y, por lo tanto, no es correcto comparar el nivel de dolor del primer día con el del segundo. Además, se debe tener cuidado al valorar si puede existir un efecto tardío por el cual la medicación del primer día continúe siendo activa el segundo. De este modo, la ausencia de beneficio en este estudio cruzado no debe dar a entender que la medicación contra el dolor del primer día no es más eficaz que el placebo.

PRUEBAS DE SIGNIFICACIÓN ESTADÍSTICA

La mayor parte de las investigaciones se realizan en una muestra o subgrupo de un grupo mayor de individuos que podrían haber sido incluidos en el estudio. Por lo tanto, los investigadores se enfrentan a menudo con la pregunta de si hubieran obtenido resultados similares con toda la población o si el azar pudo haber producido unos resultados inusitados en su muestra. Lamentablemente, no existe un método directo para responder a esta pregunta. En su lugar, los investigadores están obligados a poner a prueba sus hipótesis de estudio empleando un método indirecto de prueba por eliminación. Este método se conoce como *prueba de significación estadística*.

En su forma habitual, las pruebas de significación estadística cuantifican, a partir de los datos del estudio, la probabilidad de obtener los datos observados o un resultado más extremo si realmente no existiera una asociación entre los factores estudiados en la población. Estas pruebas se basan en el supuesto de que los individuos que toman parte en la investigación son representativos o seleccionados al azar de una población mayor. Esta acepción del término *azar* es confusa, porque las pruebas de significación estadística se emplean en estudios en los cuales los individuos asignados a un grupo de estudio o de control no son seleccionados al azar. Esta aparente contradicción se puede reconciliar si se observa que la población está compuesta por individuos que tienen las mismas características que las requeridas para participar en el estudio. De este modo, las pruebas de significación estadística realmente se aplican a cuestiones sobre poblaciones compuestas por individuos como los que participan en la investigación.

Procedimientos de las pruebas de significación estadística

Las pruebas de significación estadística o pruebas de hipótesis se basan en la premisa de que el mundo está formado por dos tipos de relaciones: o dos factores están asociados o no lo están. Estos factores, también denominados variables o características, están asociados si aparecen a la vez con más frecuencia de lo esperado

exclusivamente por azar. El papel de las pruebas de significación estadística o de hipótesis es determinar si los resultados son tan inusuales que, si no existe realmente una asociación, estamos dispuestos a suponer que existe una asociación. No obstante, en las pruebas de significación estadística se supone al principio que esa asociación no existe. Observe que el problema consiste en averiguar si la asociación existe o no. La prueba de significación por sí misma no dice nada sobre la fuerza o la importancia de la posible asociación.

Estas pruebas empiezan con la formulación de una hipótesis de estudio que afirma que existe una asociación entre factores en la población. Al realizar una prueba de significación, inicialmente se supone que la hipótesis de estudio es falsa, y se formula una hipótesis nula según la cual no existe dicha asociación o diferencia en la población. Los métodos estadísticos se emplean entonces para calcular la probabilidad de obtener los resultados observados en la muestra estudiada o resultados más extremos si realmente no existe la asociación en la población.

Cuando solo existe una pequeña probabilidad de obtener los resultados observados si la hipótesis nula fuese verdadera, los investigadores pueden entonces rechazar la aseveración de que la hipótesis nula es cierta y con ello la hipótesis nula. Al hacerlo, aceptan por eliminación la existencia de su única alternativa, la de una asociación o diferencia entre la población. Las etapas específicas de las pruebas de significación estadística son las siguientes:

1. Formulación de una hipótesis. Antes de recoger los datos, los investigadores plantean una hipótesis de estudio que postula la existencia de una diferencia entre el grupo de estudio y el de control.
2. Formulación de la hipótesis nula. Los investigadores suponen que no existe una verdadera diferencia entre el grupo de estudio y el de control. Esto se conoce como hipótesis nula.
3. Decisión sobre el nivel de significación estadística. Los investigadores deciden cuál es el nivel de probabilidad que se considerará suficientemente pequeño para rechazar la hipótesis nula. En la mayoría de los estudios de investigación médica se considera que una probabilidad de 5% o inferior es suficientemente baja para permitir el rechazo de la hipótesis nula. El 5% es el valor aceptado generalmente. Sin embargo, cualquier nivel nos deja siempre alguna posibilidad de que el azar por sí solo haya producido un conjunto de datos inusuales. Por eso, una hipótesis nula, que de hecho sea verdadera, puede ser rechazada en favor de la hipótesis del estudio hasta 5% de las veces.
4. Recogida de los datos. Se pueden recoger datos utilizando los diseños de los estudios de casos y controles, de cohortes o de los ensayos clínicos aleatorios.
5. Aplicación de la prueba de significación estadística. Si existen diferencias entre los grupos de la muestra, los investigadores calculan la probabilidad de observar esas diferencias si no existieran diferencias reales en la población de la que fueron seleccionados los individuos de los grupos de estudio y de control. Esta probabilidad se denomina valor *P*. En otras palabras, calculan la probabilidad de obtener los valores observados u otros más extremos si la hipótesis nula de no diferencia fuese cierta. Para ello, los investigadores deben escoger entre las diversas pruebas estadísticas aquella que sea apropiada para su tipo concreto de datos. Por tanto, deben asegurarse cuidadosamente de que seleccionan la prueba apropiada, como se comentará en los capítulos 27 a 30.

Para entender la forma como una prueba de significación estadística calcula probabilidades o valores *P*, veamos un ejemplo en el que se utilizan cifras

suficientemente pequeñas para facilitar los cálculos. Suponga que un investigador quiere responder a la siguiente pregunta: "*¿Es el número de hombres nacidos en los Estados Unidos de América el mismo que el de mujeres?*". En primer lugar, el investigador plantea la hipótesis de que hay más hombres que mujeres nacidos en los Estados Unidos, y luego formula la hipótesis nula de que el número de hombres nacidos en los Estados Unidos es el mismo que el de mujeres. Seguidamente, decide el nivel de significación estadística, que se sitúa habitualmente en 5% o $P = 0,05$. Después extrae una muestra de 4 certificados de nacimiento y encuentra que en su muestra hay 4 hombres y ninguna mujer. Vamos a calcular cuál es la probabilidad de obtener 4 hombres y ninguna mujer, si la hipótesis nula de igualdad en el número de hombres y mujeres fuese cierta:

Probabilidad de un hombre	0,50 ó 50%
Probabilidad de dos hombres consecutivos	0,25 ó 25%
Probabilidad de tres hombres consecutivos	0,125 ó 12,5%
Probabilidad de cuatro hombres consecutivos	0,0625 ó 6,25%

Si el número de hombres nacidos en los Estados Unidos fuese el mismo que el de mujeres, la probabilidad de obtener 4 hombres consecutivos sería 6,25%. Por eso, el "valor *P*" es igual a 0,0625.[3,4] Con una prueba de significación estadística tan simple como esta se calcula la probabilidad de obtener los datos observados suponiendo que la hipótesis nula es cierta. Con la mayor parte de las pruebas de significación estadística se obtiene el mismo tipo de resultados. Todas miden la probabilidad de obtener los datos observados o más extremos si en la población no existieran diferencias reales entre los grupos.

6. Rechazar o no la hipótesis nula. Una vez calculada la probabilidad de que los resultados pudieran haber ocurrido por azar si no existieran verdaderas diferencias en la población, los investigadores proceden a rechazar o no la hipótesis nula. Si la probabilidad de obtener los resultados por azar es menor o igual que 0,05, los investigadores pueden rechazar la hipótesis nula. Es decir, es muy poco probable que la hipótesis nula sea cierta y que los resultados obtenidos sean solo producto del azar. Por eliminación aceptan que existe una diferencia verdadera en el desenlace entre la población de la que proceden los individuos del grupo de estudio y la población de la que provienen los individuos del grupo control de las cuales fueron seleccionados los individuos investigados.

¿Qué pasa cuando la probabilidad de que la diferencia observada ocurra por azar es mayor que 0,05, como en el ejemplo precedente? Los investigadores no pueden rechazar la hipótesis nula. Esto no quiere decir que la hipótesis nula de que no hay verdaderas diferencias en la población sea cierta o incluso probable. Simplemente indica que la probabilidad de obtener los resultados observados, si la hipótesis nula fuese cierta, es demasiado grande para rechazarla en favor de la hipótesis del estudio. El peso de la prueba, por tanto, recae sobre los investigadores, que deben demostrar que la hipótesis nula es bastante improbable, antes de rechazarla en favor de la hipótesis del estudio. El siguiente ejemplo ilustra cómo funciona en la práctica el procedimiento de las pruebas de significación.

Un investigador quería poner a prueba la hipótesis de que el cáncer de la cavidad oral está asociado con fumar en pipa. Para ello, formuló una hipótesis

[3] Hemos realizado una prueba de significación estadística unilateral.
[4] N. del E. Las pruebas de significación estadística unilaterales también se denominan "*pruebas de una cola*".

nula según la cual el fumar en pipa no estaba asociado con el cáncer de la cavidad oral en la población general. Luego decidió que si observaba unos datos que solo se obtuvieran 5% o menos de las veces si la hipótesis nula fuese cierta, rechazaría la hipótesis nula. Seguidamente, recogió los datos en una muestra de la población general de fumadores de pipa y de no fumadores. Mediante la prueba de significación estadística apropiada observó que, si no existiera una asociación entre el fumar en pipa y el cáncer de la cavidad oral en la población general, datos tan extremos o más que los obtenidos se observarían por azar solo el 3% de las veces. Por último, rechazó la hipótesis nula, dado que era bastante improbable obtener esos datos si no existiera una asociación entre fumar en pipa y cáncer de boca. De este modo, el investigador aceptó por eliminación la hipótesis del estudio según la cual existe una asociación entre el fumar en pipa y el cáncer de boca en la población general.

Recuerde que hemos definido *pequeña* como una probabilidad de 5% o menos de que los resultados observados se hubiesen obtenido si no existiera una verdadera diferencia en la población. La cifra 5% puede ser demasiado grande o demasiado pequeña si de los resultados dependen decisiones importantes. El valor 5% se basa en la conveniencia de algunas propiedades estadísticas; de todos modos, no es un valor mágico. Es posible definir como *pequeña* a una probabilidad de 1%, 0,1% o a cualquier otra probabilidad que se escoja. Recuerde, sin embargo, que, independientemente del nivel escogido, siempre habrá alguna probabilidad de rechazar la hipótesis nula cuando no exista una verdadera diferencia. Las pruebas de significación estadística pueden medir esta probabilidad, pero no eliminarla.

En el cuadro 5-1 se repasan y resumen las etapas para llevar a cabo una prueba de significación estadística.

CUADRO 5-1. Cómo funciona una prueba de significación estadística

1. *Formular una hipótesis*
   Desarrollar la pregunta del estudio: existe una asociación entre factores[a] o una diferencia entre grupos de la población general.
2. *Formular la hipótesis nula*
   Invertir la hipótesis: no existe una asociación entre factores o una diferencia entre los grupos de la población general.
3. *Decidir el nivel de significación*
   5%, si no se indica y justifica lo contrario.
4. *Recoger los datos*
   Determinar si existe una asociación entre los factores o una diferencia entre los grupos a partir de los datos recogidos de las muestras de la población.
5. *Aplicar la prueba de significación estadística*
   Calcular la probabilidad de obtener los datos observados o más extremos si la hipótesis nula fuese verdadera (esto es, escoger y aplicar la prueba de significación estadística adecuada).
6. *Rechazar o no rechazar la hipótesis nula*
   Rechazar la hipótesis nula y aceptar por eliminación la hipótesis del estudio, si se alcanza el nivel de significación estadística. No rechazar la hipótesis nula si la probabilidad de observar los datos por azar es mayor que 5% cuando no existe una asociación entre factores o diferencia entre grupos en la población.

[a] En realidad se trata de una asociación entre variables, como se verá más adelante.

ERRORES EN LAS PRUEBAS DE SIGNIFICACIÓN ESTADÍSTICA

Cuando se utilizan las pruebas de significación estadística, con frecuencia se pueden cometer diversos errores:

1. No formular la hipótesis antes de realizar el estudio.
2. No interpretar correctamente los resultados de una prueba de significación estadística al no considerar el error de tipo I.
3. No interpretar correctamente los resultados de las pruebas de significación estadística al no considerar el error de tipo II.

Empezaremos observando las consecuencias que acarrea el no formular la hipótesis antes de realizar el estudio.

Un investigador seleccionó al azar 100 individuos con hipertensión arterial esencial conocida y 100 sin hipertensión. Para averiguar en qué diferían ambos grupos, los comparó en función de una lista de 100 variables. Utilizando las pruebas estadísticas habituales, de las 100 variables estudiadas, solo dos fueron estadísticamente significativas a un nivel de 0,05: (1) los hipertensos tenían en general más letras en su apellido que los normotensos, y (2) los hipertensos nacieron en uno de los primeros tres días y medio de la semana, mientras que los normotensos lo hicieron en la segunda mitad. El autor concluyó que, a pesar de que estas diferencias no habían sido previstas, los apellidos largos y haber nacido en la primera mitad de la semana se asociaban con la hipertensión esencial.

Este ejemplo muestra la importancia de establecer la hipótesis de antemano. Cuando se contrasta un elevado número de variables, es probable que algunas de ellas sean estadísticamente significativas solo por azar. Si la hipótesis no se ha formulado de antemano, no existe una hipótesis nula que se pueda rechazar. Además, puede ser confuso aplicar los niveles habituales de significación estadística si no se establece la hipótesis antes de recoger y analizar los datos. Si se buscan las asociaciones una vez recogidos los datos, es posible que se tengan que aplicar criterios más estrictos que la probabilidad habitual de 0,05.

Cuando se formula una sola hipótesis, una regla práctica que se puede sugerir al lector de la literatura médica es la de dividir el valor *P* observado por el número de variables estudiadas. El valor *P* resultante se puede utilizar para rechazar o no la hipótesis nula. Por ejemplo, imagine un estudio en el que se analizan cinco variables en cada uno de los grupos, sin formular la hipótesis de estudio. Para obtener un valor *P* global de 0,05, cualquier variable determinada debe tener un valor *P* igual a[5]

$$\frac{0{,}05}{\text{número de variables}} = \frac{0{,}05}{5} = 0{,}01$$

Este valor *P* de 0,01 debe interpretarse de la misma forma que se interpretaría el valor *P* de 0,05 si la hipótesis del estudio se hubiera formulado antes de iniciarlo.[6] Este enfoque reduce la potencia estadística de un estudio para demostrar la significación estadística de la diferencia entre los valores de cualquier variable. Por eso,

[5] Si hubiera más de dos grupos, la ecuación correspondiente sería igual a la probabilidad deseada de cometer un error de tipo I dividida por el número de comparaciones que se van a realizar.

[6] Este método es una aproximación útil cuando el número de variables es bajo. Cuando el número de variables aumenta bastante por encima de 5, tiende a requerir un valor *P* demasiado pequeño antes de que se alcance una significación estadística.

muchos bioestadísticos consideran que es mejor emplear un método multivariante, que se describirá en la *Parte 4, La selección de una prueba estadística.*

Recuerde que las pruebas de significación estadística o de hipótesis constituyen un método para realizar inferencias en un mundo en el cual debemos decidir entre la hipótesis de estudio y la hipótesis nula basándonos *exclusivamente en los datos del estudio.*[7] Sin embargo, es posible contemplar la inferencia como un proceso que incorpora una probabilidad de que la hipótesis sea cierta. En este proceso, el investigador, antes de iniciar el estudio, debe estimar la probabilidad de que la hipótesis sea cierta. Ello puede realizarse a partir de los resultados de estudios anteriores o de conocimientos médicos previos. Cuando se calcula la probabilidad previa, existen métodos estadísticos para estimar la probabilidad de que la hipótesis sea cierta después de obtener los datos del estudio. Este proceso *bayesiano* es paralelo al uso de pruebas diagnósticas que se discutirán en la *Segunda parte, La prueba de una prueba.* Una ventaja del método bayesiano es que no exige ajustar los valores *P* según el número de variables.

Error de tipo I

Algunos errores son inherentes al método empleado en las pruebas de significación estadística. La posibilidad de que la hipótesis nula pueda ser falsamente rechazada y de que la hipótesis del estudio sea falsamente aceptada es un concepto fundamental subyacente en las pruebas de significación estadística. Esta posibilidad se conoce como el *error de tipo I.* En las pruebas de significación tradicionales, las posibilidades de aceptar incorrectamente una hipótesis de estudio, pueden alcanzar hasta 5% aun cuando no exista una verdadera diferencia en la población de la que se ha extraído la muestra del estudio. Este nivel de error de tipo I se denomina nivel alfa. Las pruebas de significación estadística no eliminan la incertidumbre. Los lectores cuidadosos de los artículos científicos pueden por tanto apreciar el grado de duda existente y decidir por sí mismos si están dispuestos a tolerar o actuar con tal grado de incertidumbre.

En determinadas circunstancias, un nivel alfa de 0,05 puede sobrepasar lo que uno está dispuesto a tolerar, mientras que en otras puede tolerar incluso más de 5%. Por ejemplo, antes de introducir un nuevo método de potabilización del agua en una comunidad con una baja frecuencia de infecciones transmitidas por el agua, puede ser inaceptable una probabilidad de 5% de que el nuevo método no mate a los organismos patógenos. Por otro lado, en una comunidad donde el agua es la principal fuente de transmisión de enfermedades se puede tolerar una probabilidad más alta de que el nuevo método no consiga eliminar las enfermedades transmitidas por el agua, especialmente si no existe otro método disponible. Veamos cómo el hecho de soslayar la posibilidad de un error de tipo I puede conducir a una interpretación errónea de los resultados del estudio.

El autor de un artículo médico de revisión evaluó 20 estudios bien realizados en los que se examinaba la relación entre la lactancia materna y el cáncer de mama. En 15 estudios no se encontró una asociación entre ambas variables. En un estudio se observó una asociación significativa a un nivel de 0,05 entre la lactancia ma-

[7] Es posible incorporar indirectamente información externa en el método estadístico mediante la elección del valor *P* que se utilizará para declarar la existencia de significación estadística. Por ejemplo, la utilización de una prueba de significación estadística unilateral presupone que los datos anteriores ya implican que la hipótesis es verdadera y que el estudio se realiza para determinar la fuerza de la relación. No obstante, esta aproximación incorpora mucha menos información externa que la que es posible aportar con el empleo del método bayesiano.

terna y el cáncer de mama. El autor del artículo de revisión concluyó que se debía desaconsejar la lactancia materna, dada la existencia de un estudio en el que se sugería que esta estaba asociada con un aumento del riesgo de cáncer de mama.

Cuando se llevan a cabo correctamente 20 estudios para probar la existencia de una asociación que en realidad no existe, hay una posibilidad sustancial de que uno de los estudios muestre una asociación a un nivel de 5% simplemente por azar. Recuerde el significado de la significación estadística a un nivel de 0,05: ello implica que los resultados tienen una probabilidad de 5%, o de 1 entre 20, de observarse solo por azar cuando no existe una asociación en la población. Por eso, el que un estudio entre 20 muestre una asociación no debe interpretarse como prueba de que esta exista. Es importante tener en cuenta la posibilidad de que no exista una asociación aunque así lo indiquen los resultados de las pruebas de significación estadística. Si se hubiera aceptado sin mayor cuestionamiento el único estudio que muestra dicha asociación, la lactancia materna se habría desaconsejado por no producir ningún beneficio en la prevención del cáncer.

Error de tipo II

Según el error de *tipo II*, la ausencia de pruebas suficientes para rechazar la hipótesis nula no significa necesariamente que no exista una verdadera diferencia. Recuerde que las pruebas de significación estadística solo hacen referencia a la hipótesis nula. El proceso de la significación estadística nos permite únicamente rechazar o no rechazar la hipótesis nula; no nos permite confirmarla. No rechazar la hipótesis nula quiere decir simplemente que los datos no son lo suficientemente convincentes como para rechazar el supuesto de que no hay diferencias entre los grupos o asociación entre los factores en la población.

Dos factores pueden impedir al investigador demostrar la existencia de una diferencia estadísticamente significativa aunque esta exista. El azar por sí solo puede producir un grupo de datos tan inusual que no indique la existencia de una diferencia sustancial aunque esta realmente exista en la población. Este tipo de error es paralelo al error de tipo I e indica que el azar ha desempeñado algún papel. Un estudio determinado puede aportar un resultado inusual que solo podría ocurrir en un bajo porcentaje de ocasiones. Esto no quiere decir que se hayan cometido errores en el diseño o en la interpretación del estudio, simplemente indica que, a pesar de nuestros mejores esfuerzos e intenciones, los métodos estadísticos pueden conducirnos a extraer conclusiones incorrectas por lotería. Este factor es intrínseco a los conceptos estadísticos: el resultado de realizar pruebas de significación estadística siempre conlleva esta probabilidad de error.

Los investigadores pueden ir contra sus propios intereses al realizar estudios de pocos individuos. Cuanto menor sea el número de individuos que tomen parte en un estudio, mayor será el impacto de la aparición por azar de algunos individuos con valores infrecuentes. La inclusión de observaciones inusuales dificulta el rechazo de la hipótesis nula. Cuanto menor sea el número de individuos incluidos en un estudio, mayor deberá ser la verdadera diferencia en promedio antes de poder demostrar resultados estadísticamente significativos.

Por el contrario, cuanto mayor sea el número de individuos participantes en un estudio, menor será la magnitud de la verdadera diferencia que pueden demostrarse como estadísticamente significativa, utilizando los datos. En su forma extrema, según este concepto, cualquiera que sea la verdadera diferencia, por pequeña que sea, puede ser estadísticamente significativa si el número de individuos que participan en el estudio es lo suficientemente elevado.

Se dispone de pruebas estadísticas para calcular la probabilidad de que un estudio detecte una diferencia estadísticamente significativa si realmente existe en la población una verdadera diferencia de un tamaño especificado. Estas pruebas miden la "potencia" estadística de un estudio. La probabilidad de que muchos estudios no detecten una diferencia estadísticamente significativa, cuando realmente existe una verdadera diferencia, es bastante alta. Ningún número arbitrario puede indicar cuál es la magnitud del error de tipo II que uno debe tolerar. Sin afirmarlo realmente, los investigadores que utilizan muestras relativamente pequeñas aceptan un riesgo de 20%, 30% o incluso más alto de no demostrar una diferencia estadísticamente significativa que realmente existe en la población. En el cuadro 5-2 se resumen y comparan los errores de tipo I y II.

El siguiente ejemplo muestra el efecto del tamaño muestral sobre la capacidad de una prueba para detectar diferencias estadísticamente significativas entre grupos.

En un estudio de los efectos perjudiciales de los cigarrillos sobre la salud se siguió a 100 fumadores de cigarrillos y a 100 no fumadores durante 20 años. En ese período, 5 fumadores desarrollaron cáncer de pulmón y los no fumadores, ninguno. En el mismo período, 10 fumadores y 9 no fumadores padecieron infarto de miocardio. Los resultados de los sujetos con cáncer de pulmón fueron estadísticamente significativos, pero los del infarto de miocardio no. Los autores concluyeron que existía una asociación entre los cigarrillos y el cáncer de pulmón, y rebatieron la existencia de una asociación entre los cigarrillos y el infarto de miocardio.

Cuando las diferencias entre los grupos son sustanciales, como sucede entre los fumadores y los no fumadores en relación con el cáncer de pulmón, solo se necesita una muestra pequeña para demostrar una significación estadística. Cuando las verdaderas diferencias son pequeñas, se precisan más sujetos para demostrarla. No se puede decir que este estudio haya refutado una asociación entre los cigarrillos y el infarto de miocardio. Es muy probable que el tamaño muestral utilizado fuera demasiado pequeño para que el estudio tuviera suficiente potencia estadística y pudiera demostrar una asociación entre los cigarrillos y el infarto de miocardio, aunque otros estudios hayan sugerido que dicha asociación existe en la población general. Un estudio con potencia limitada para demostrar la existencia de una diferencia tiene también una potencia limitada para rebatirla.

CUADRO 5-2. Errores inherentes a las pruebas de significación estadística

	Error de tipo I	Error de tipo II
Definición:	Rechazar la hipótesis nula cuando no existe una verdadera diferencia en la población general	No rechazar la hipótesis nula cuando existe una verdadera diferencia en la población general
Causa:	Azar	Azar o tamaño muestral demasiado pequeño
Probabilidad de presentación:	La determinación del nivel de significación indicará la magnitud del error tolerado	Las pruebas estadísticas permiten estimar la probabilidad de cometerlo a partir del tamaño de los grupos (la probabilidad de cometer este error puede ser bastante grande si el número de individuos estudiados es bajo)

AJUSTE

En el capítulo 3 se señaló que el investigador está obligado a comparar las características de los individuos del grupo de estudio con las de los del grupo control, para determinar si difieren en alguna de ellas. Si los grupos difieren, aunque no sea de forma estadísticamente significativa, el investigador debe considerar si estas diferencias pudieron haber influido en los resultados. Las características que indican diferencias entre los grupos y que pueden influir en los resultados del estudio son variables de confusión potenciales. Estas variables de confusión potenciales pueden ser el resultado de un sesgo de selección en los estudios de casos y controles o de cohortes o de diferencias aleatorias en los tres tipos de estudios básicos. Si el investigador detecta una variable de confusión potencial, está obligado a tomarla en consideración en el análisis mediante un proceso denominado *ajuste de los datos*.[7]

Al realizar un ajuste, el investigador separa en grupos a los sujetos que poseen niveles diferentes de la variable de confusión. A continuación, compara los grupos con el mismo nivel de la variable de confusión, para investigar si existe una asociación entre la exposición y la enfermedad. Por ejemplo, si la edad es una variable de confusión potencial, el investigador puede subdividir los grupos según la edad en diversas categorías; entonces puede comparar los grupos de estudio y de control en cada categoría de la variable edad para determinar si existen diferencias cuando se comparan los grupos de edad similar. Como se verá en el capítulo 29, se dispone de técnicas estadísticas conocidas como *métodos multivariantes* para ajustar los datos según una o más variables al mismo tiempo. Si la variable de confusión no se identifica y los datos no se ajustan según dicha variable, se pueden cometer errores graves, como ilustra el siguiente ejemplo.

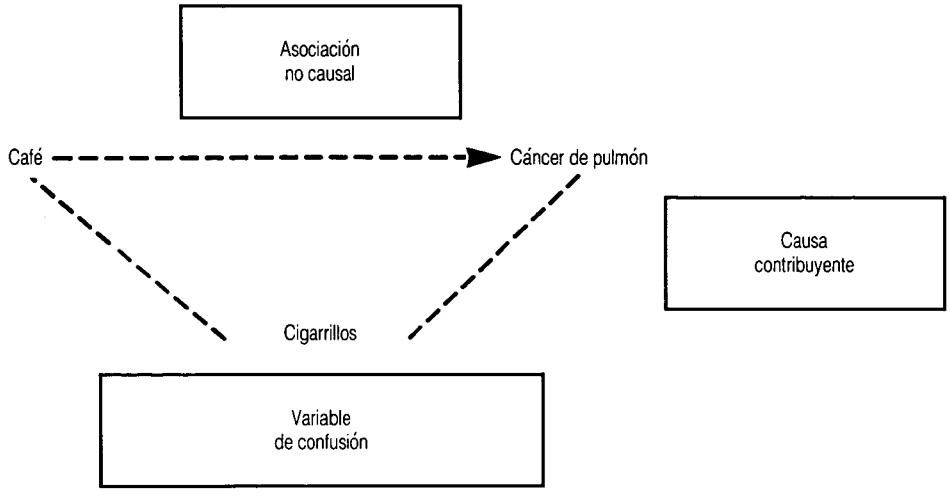
Un investigador estudió la relación entre el consumo de café y el cáncer de pulmón mediante el seguimiento durante 10 años de 500 bebedores de café y 500 no bebedores. En este estudio de cohortes, el riesgo de padecer cáncer de pulmón de los bebedores asiduos fue 10 veces el de los no bebedores. El autor concluyó que el café, juntamente con los cigarrillos, era un factor de riesgo del desarrollo de cáncer de pulmón.

En este estudio, el consumo de cigarrillos puede ser considerado como variable de confusión si se supone que fumar cigarrillos y beber café están asociados. En otras palabras, es más probable que los bebedores de café fumen que los no bebedores de café. Además, fumar cigarrillos está asociado con el cáncer de pulmón. Por eso, los cigarrillos son un factor de confusión potencial, dado que están relacionados tanto con el desenlace —cáncer de pulmón— como con el consumo de café. En la figura 5-1 se representa la relación entre el consumo de café, el de cigarrillos y el cáncer de pulmón. Si fumar cigarrillos es una variable de confusión, el ajuste según este factor debe formar parte del análisis.

Para efectuar el ajuste según el consumo de cigarrillos, el investigador dividiría a los bebedores de café en fumadores y no fumadores, y haría lo mismo con los no bebedores. Entonces compararía a los bebedores de café que no fuman con los que no beben café y no fuman, para determinar si continúa existiendo la relación entre consumo de café y cáncer de pulmón. Solo después de demostrar que el ajuste según el consumo de cigarrillos no elimina la relación entre el consumo de café y el cán-

[7] Muchos bioestadísticos son partidarios del uso de las técnicas de ajuste, incluso cuando las diferencias no son aparentes, a causa de la posibilidad de que existan interacciones entre las variables.

FIGURA 5-1. Relación entre causa contribuyente, variable de confusión y asociación no causal



cer de pulmón, el autor puede concluir que el consumo de café está asociado con el desarrollo del cáncer de pulmón.

FUERZA DE LA RELACIÓN

Una vez examinado el uso de los métodos estadísticos para tener en cuenta las variables de confusión y realizar pruebas de significación estadística, centremos nuestra atención en la manera como los métodos estadísticos nos ayudan a medir la fuerza de una asociación observada. Primero, veremos la medida fundamental de la fuerza de una asociación que se emplea la mayor parte de las veces en los estudios de cohortes. Luego revisaremos la medida básica usada en los estudios de casos y controles. Recuerde que por *asociación* entendemos que un factor, con frecuencia llamado factor de *riesgo*, se observa juntamente con una enfermedad con mayor frecuencia que la esperada solo por azar. Observe que una *asociación* no implica necesariamente una relación de causa-efecto, como examinaremos con más detalle en el capítulo 6.

Supongamos que estamos estudiando la asociación entre los anticonceptivos orales y la tromboflebitis y que queremos medir la fuerza de la asociación para determinar cómo afecta el uso de anticonceptivos orales al desarrollo de tromboflebitis. Por lo tanto, primero debemos clarificar el concepto de *riesgo*.

El riesgo mide la probabilidad de desarrollar una enfermedad durante un determinado período de tiempo. El riesgo es igual al número de individuos que desarrollan la enfermedad dividido por el número de individuos que podían desarrollar la enfermedad al inicio del período. Para estimar el riesgo de tromboflebitis en 10 años, dividiríamos el número de mujeres que tomaban anticonceptivos orales y que desarrollaron tromboflebitis durante el período de 10 años por el total de mujeres del grupo de estudio que tomaban anticonceptivos al inicio del período.

Para medir el grado relativo de asociación entre la tromboflebitis en las mujeres que tomaban anticonceptivos orales y en las que no los tomaban, se debe efectuar un cálculo adicional. Una de tales medidas se denomina *riesgo relativo*. El riesgo

relativo mide el riesgo de desarrollar tromboflebitis si se toman anticonceptivos respecto del riesgo si no se toman, y se define de la siguiente manera:

$$\text{Riesgo relativo} = \frac{\text{Riesgo de desarrollar tromboflebitis si se toman anticonceptivos orales}}{\text{Riesgo de desarrollar tromboflebitis si no se toman anticonceptivos orales}}$$

En general,

$$\text{Riesgo relativo} = \frac{\text{Riesgo del desenlace en presencia del factor de riesgo}}{\text{Riesgo del desenlace en ausencia del factor de riesgo}}$$

Veamos ahora, mediante un ejemplo hipotético, cómo se calculan el riesgo y el riesgo relativo.

Durante 10 años un investigador siguió a 1 000 mujeres jóvenes seleccionadas al azar que tomaban píldoras anticonceptivas y a 1 000 que no las tomaban. Observó que 30 de las mujeres que tomaban anticonceptivos desarrollaron tromboflebitis durante dicho período, mientras que solo lo hicieron 3 de las que no los tomaban. A continuación, presentó sus datos empleando la siguiente tabla de 2 × 2:

	Con tromboflebitis	Sin tromboflebitis	
Tomaban píldoras anticonceptivas	a = 30	b = 970	a + b = 1 000
No tomaban píldoras anticonceptivas	c = 3	d = 997	c + d = 1 000

El riesgo de desarrollar tromboflebitis a los 10 años de las mujeres que tomaban píldoras anticonceptivas es igual al número de mujeres que desarrollaron la enfermedad y tomaban las píldoras dividido por el número total de mujeres que tomaban la píldora. De este modo,

Riesgo de desarrollar tromboflebitis de las mujeres que tomaban píldoras anticonceptivas

$$= \frac{a}{a + b} = \frac{30}{1\,000} = 0{,}03$$

De forma similar, el riesgo de desarrollar tromboflebitis a los 10 años de las mujeres que no tomaban las píldoras anticonceptivas es igual al número de mujeres que no las tomaban y que desarrollaron la enfermedad dividido por el número total de mujeres que no las tomaban. Por consiguiente,

Riesgo de desarrollar tromboflebitis de las mujeres que no tomaban píldoras

$$= \frac{c}{c + d} = \frac{3}{1\,000} = 0{,}003$$

El riesgo relativo es igual a la razón entre esos dos riesgos; entonces,

$$\text{Riesgo relativo} = \frac{a/a + b}{c/c + d} = \frac{0{,}03}{0{,}003} = 10$$

Un riesgo relativo de 1 significa que la toma de anticonceptivos orales no aumenta el riesgo de padecer tromboflebitis. Un riesgo relativo de 10 significa que, como término medio, las mujeres que toman la píldora tienen un riesgo de tromboflebitis 10 veces más elevado que las que no la toman.

Ahora veamos cómo se mide la fuerza de una asociación en un estudio retrospectivo o de casos y controles observando la asociación entre los anticonceptivos orales y la tromboflebitis.

Un investigador seleccionó 100 mujeres jóvenes con tromboflebitis y 100 sin tromboflebitis. Seguidamente, elaboró con cuidado la historia del uso previo de píldoras anticonceptivas. Encontró que 90 de las 100 mujeres con la enfermedad tomaban la píldora, en comparación con las 45 que las tomaban entre las que no tenían la enfermedad. A continuación, representó los datos mediante una tabla de 2 × 2:

	Con tromboflebitis	Sin tromboflebitis
Tomaban píldoras anticonceptivas	a = 90	b = 45
No tomaban píldoras anticonceptivas	c = 10	d = 55
	a + c = 100	b + d = 100

Observe que en los estudios de casos y controles el investigador puede escoger el número total de pacientes de cada grupo (con y sin tromboflebitis). En este caso, pudo haber escogido a 200 pacientes con tromboflebitis y a 100 sin la enfermedad u otras combinaciones. Por ello, el número final de cada columna se puede modificar a voluntad del investigador. En otras palabras, en un estudio de casos y controles el número de individuos que padecen y no padecen la enfermedad no refleja necesariamente la frecuencia natural de la enfermedad. Por tanto, es incorrecto sumar las casillas horizontalmente en un estudio de casos y controles (como hicimos en el estudio de cohortes precedente). Esto permitiría al investigador manipular la dimensión del riesgo relativo resultante.

Lamentablemente, sin números en el marginal derecho de la tabla 2 × 2 no es posible calcular el riesgo, como lo hicimos en el estudio de cohortes. Sin embargo, en los estudios de casos y controles existe una buena aproximación al riesgo relativo que resulta muy útil para realizar análisis estadísticos. Esta aproximación al riesgo relativo se denomina *razón de productos cruzados o de ventajas (odds ratio)*.

En primer lugar, ¿qué queremos decir con *ventaja (odds)* y en qué se diferencia de la probabilidad o del riesgo? El riesgo es una medida de probabilidad cuyo numerador contiene el número de veces que un suceso como la tromboflebitis ocurre en un determinado período de tiempo. El denominador del riesgo es el número de ve-

ces que el suceso pudo haber ocurrido. La ventaja, como la probabilidad, tiene por numerador el número de veces que el suceso ha ocurrido. Sin embargo, el denominador es el *número de veces que el suceso no ha ocurrido*. La diferencia entre ventaja y probabilidad se puede apreciar pensando en la probabilidad de sacar un as de una baraja de 52 cartas. La probabilidad de sacar un as es el número de veces que saldrá un as dividido por el total de cartas; es decir, 4 entre 52 ó 1 entre 13. La ventaja, por su parte, es el número de veces que saldrá un as dividido por el número de veces que no saldrá, o sea, 4 entre 48 ó 1 entre 12. Por eso, la ventaja es ligeramente distinta de la probabilidad, pero cuando el suceso o la enfermedad estudiada es poco frecuente, la ventaja es una buena aproximación al riesgo o a la probabilidad.

La razón de productos cruzados o de ventajas mide la ventaja de tener el factor de riesgo si la enfermedad está presente dividida por la ventaja de tener el factor de riesgo si la enfermedad no está presente. La ventaja de haber tomado la píldora en presencia de tromboflebitis es igual a:

$$\frac{a}{c} = \frac{90}{10} = 9$$

De forma similar, la ventaja de tomar la píldora para las mujeres que no desarrollan la enfermedad se calcula dividiendo el número de mujeres que no tienen tromboflebitis y están tomando la píldora por el número de mujeres que no tienen tromboflebitis y no están tomando la píldora. De este modo, la ventaja de estar tomando la píldora en ausencia de tromboflebitis es igual a:

$$\frac{b}{d} = \frac{45}{55} = 0{,}82$$

Paralelamente al cáculo del riesgo relativo, se puede desarrollar una medida de la ventaja relativa de estar tomando la píldora en presencia de tromboflebitis respecto a la de tomar la píldora en ausencia de tromboflebitis. Esta medida de la fuerza de la asociación se conoce como razón de productos cruzados o de ventajas (*odds ratio*). De este modo,

$$\text{Razón de productos cruzados o de ventajas} = \frac{\text{Ventaja de estar tomando la píldora en presencia de tromboflebitis}}{\text{Ventaja de estar tomando la píldora en ausencia de tromboflebitis}}$$

$$= \frac{a/c}{b/d} = \frac{ad}{cb} = \frac{9}{0{,}82} = 11$$

De forma semejante a nuestra interpretación del riesgo relativo, una razón de productos cruzados de 1 indica que la ventaja de tomar la píldora si la tromboflebitis está presente es la misma que la de tomarla si la tromboflebitis está ausente. Nuestra razón de productos cruzados o de ventajas de 11 significa que la ventaja de tomar píldoras anticonceptivas aumenta 11 veces en las mujeres con tromboflebitis.

La razón de productos cruzados sirve como medida básica del grado de asociación en los estudios de casos y controles. Por sí misma, es una medida útil y válida de la fuerza de la asociación. Además, mientras la enfermedad (tromboflebitis)

sea rara, la razón de productos cruzados se aproxima al riesgo relativo.

Es posible contemplar la razón de productos cruzados de forma inversa a la que se haría en un estudio de cohortes y obtener el mismo resultado. Por ejemplo,

$$\text{Razón de productos cruzados o de ventajas} = \frac{\text{Ventaja de desarrollar la tromboflebitis si se toma la píldora}}{\text{Ventaja de desarrollar la tromboflebitis si no se toma la píldora}}$$

La razón de productos cruzados entonces es igual a

$$\frac{a/b}{c/d} = \frac{ad}{cb} = 11$$

Observe que esta es la misma fórmula de la razón de productos cruzados que la mostrada previamente. Esta útil propiedad permite calcular la razón de productos cruzados en un estudio de cohortes o en un ensayo controlado aleatorio en lugar del riesgo relativo, y compararla directamente con la razón de productos cruzados calculada en un estudio de casos y controles.

El riesgo relativo y la razón de productos cruzados son, por consiguiente, medidas fundamentales que empleamos para cuantificar la fuerza de una asociación entre un factor de riesgo y una enfermedad.

INTERVALOS DE CONFIANZA

Cuando hablamos de las pruebas de significación estadística, señalábamos que estas pruebas no proporcionan información acerca de la fuerza de una asociación. Por ello, es interesante utilizar un método que proporcione una medida de síntesis —con frecuencia denominada estimación puntual— de la fuerza de una asociación y que, al mismo tiempo, nos permita realizar una prueba de significación estadística.

Los *intervalos de confianza* son un método que combina información obtenida en muestras sobre la fuerza de la asociación con información sobre los efectos del azar en la probabilidad de obtener los resultados observados. Si bien es posible calcular el intervalo de confianza para cualquier porcentaje de confianza entre 0 y 100, el de 95% es el utilizado con más frecuencia.

El intervalo de confianza de 95% nos permite tener una "confianza" de 95% de que el valor de la población (parámetro) se halla dentro del intervalo.

Frecuentemente se calculan intervalos de confianza para la razón de productos cruzados y para el riesgo relativo. El cálculo de estos intervalos es complejo. El lector puede encontrarse con una expresión como 10(8,12), que expresa la razón de productos cruzados (límite inferior del intervalo, límite superior).

El término *límite de confianza* se emplea para indicar los límites superior e inferior del intervalo de confianza. Esta expresión indica habitualmente la razón de productos cruzados observada y su intervalo de confianza de 95%. Cuando se empleen otros intervalos de confianza, se deben indicar específicamente.

Imagine un estudio en el que la razón de productos cruzados para las píldoras anticonceptivas y la tromboflebitis fue 10(8,12). ¿Cómo interpretaría este intervalo de confianza?

El intervalo de confianza de esta razón de productos cruzados nos permite afirmar con una confianza de 95% que la razón de productos cruzados poblacional se encuentra entre 8 y 12. Esto nos permite estar bastante seguros de que se ha observado una razón de productos cruzados importante, no solo en nuestra muestra, sino también en la población de la cual se extrajo dicha muestra.

Estas expresiones de los límites de confianza tienen otra ventaja para el lector de la literatura clínica: le permiten realizar pruebas de hipótesis y sacar rápidamente conclusiones sobre la significación estadística de los datos observados. Cuando se emplea un intervalo de confianza de 95% podemos inmediatamente concluir si los datos observados son estadísticamente significativos con un valor *P* igual a 0,05 o menor.

Este cálculo es particularmente sencillo para las razones de productos cruzados y para los riesgos relativos. Para la razón de productos cruzados y para los riesgos relativos, 1 representa el punto en el cual la ventaja o el riesgo de la enfermedad son iguales tanto si está como si no está presente el factor de riesgo. De este modo, una razón de productos cruzados de 1 es, en realidad, la expresión de la hipótesis nula según la cual la ventaja de la enfermedad es la misma, tanto si el factor de riesgo está presente como si está ausente.

Por consiguiente, como la razón de productos cruzados es estadísticamente significativa si su intervalo de confianza de 95% se aleja de 1 o no lo incluye, sería correcto concluir que la razón de productos cruzados es estadísticamente significativa con un valor *P* igual a 0,05 o menor. Los mismos principios son ciertos para el riesgo relativo. Veamos las siguientes razones de productos cruzados y sus intervalos de confianza:

A. 4(0,9–7,1)
B. 4(2–6)
C. 8(1–15)
D. 8(6–10)
E. 0,8(0,5–1,1)
F. 0,8(0,7–0,9)

Dado que la razón de productos cruzados es estadísticamente significativa si el intervalo de confianza de 95% no incluye a 1, los ejemplos B, C, D y F son estadísticamente significativos con un valor *P* igual a 0,05 o menor. Los ejemplos A y E no son estadísticamente significativos, porque su valor *P* correspondiente es mayor de 0,05.

Como lector de la literatura clínica, usted encontrará cada vez más valores observados e intervalos de confianza en la sección de resultados de los artículos científicos. Esto es una gran ayuda, porque permite hacerse una idea o formarse una "gestalt" sobre los datos.[8] Le permite sacar, además, sus propias conclusiones sobre la importancia clínica de la dimensión o de la fuerza de la estimación puntual observada. Por último, para aquellos que desean convertirlo al formato tradicional de las pruebas de significación estadística para contraste de hipótesis, muchas veces se puede realizar un cálculo aproximado para determinar si los resultados son estadísticamente significativos con un valor *P* de 0,05 o menor.

[8] Nota del T. La formación de una "gestalt" se refiere a la tendencia a organizar sensaciones en una totalidad o en un patrón significativo.

CAPÍTULO 6

INTERPRETACIÓN

SIGNIFICACIÓN ESTADÍSTICA E IMPORTANCIA CLÍNICA

Las pruebas de significación estadística están diseñadas para ayudarnos a valorar el papel del azar cuando en una investigación observamos una diferencia o una asociación. Como ya hemos comentado, estas pruebas nos dicen muy poco de la magnitud de una diferencia o de la fuerza de una asociación. Por ello, es importante preguntarse no solo si una diferencia o una asociación es estadísticamente significativa, sino también si es lo suficientemente grande para ser útil en el medio clínico. El mundo está lleno de miríadas de diferencias entre individuos y grupos. Sin embargo, muchas de ellas no son lo suficientemente grandes como para permitirnos separar a los individuos en grupos con fines diagnósticos y terapeúticos.[1]

Como hemos visto, la probabilidad de cometer un error de tipo II puede ser muy grande cuando el tamaño de la muestra es pequeño. Recuerde que el error de tipo II es la probabilidad de no demostrar la existencia de una significación estadística cuando existe una verdadera diferencia o asociación. Por el contrario, como se ilustra en el siguiente ejemplo, cuando la muestra es bastante grande, es posible obtener una diferencia o asociación estadísticamente significativa, aun cuando esta es demasiado pequeña o débil para ser clínicamente útil.

Unos investigadores siguieron a 100 000 hombres de edad media durante 10 años para determinar los factores asociados con la enfermedad coronaria. Establecieron de antemano la hipótesis de que la concentración del ácido úrico en la sangre podría ser un factor predictivo de la enfermedad. Observaron que, en los hombres que desarrollaron enfermedad coronaria, la concentración de ácido úrico era 7,8 mg/dl, mientras que entre los que no la desarrollaron era de 7,7mg/dl. La diferencia fue estadísticamente significativa a un nivel de 0,05. Los autores llegaron a la conclusión de que los resultados eran clínicamente útiles, ya que habían encontrado una diferencia estadísticamente significativa.

Las diferencias observadas en este estudio fueron estadísticamente significativas, pero tan pequeñas que probablemente no tenían importancia clínica. El elevado número de hombres incluidos en el estudio permitió que los investigadores detectaran una diferencia muy pequeña entre los grupos. Sin embargo, la pequeña magnitud de la diferencia hace improbable que la medición de la concentración de ácido úrico pueda ser clínicamente útil para predecir quién padecerá una enfermedad coronaria. Esa pequeña diferencia no ayuda al clínico a separar a aquellos que desarrollarán la enfermedad de los que no la desarrollarán. De hecho, cuando la prueba se realiza en el laboratorio, la diferencia observada es probablemente menor que la magnitud del error de laboratorio que se comete al medir la concentración de ácido úrico.

[1] En ocasiones, es necesario distinguir entre lo que es estadísticamente significativo y sustancial y lo que es clínicamente importante. A veces hay diferencias entre grupos que son grandes y estadísticamente significativas, pero no útiles para tomar decisiones. Desde el punto de vista médico o social, podemos decidirnos a tratar igualmente a los individuos sin tener en cuenta grandes diferencias en factores como la inteligencia, la estatura o la edad.

CAUSA CONTRIBUYENTE

¿Pueden causar cáncer los cigarrillos? ¿Puede el colesterol ser la causa de la enfermedad coronaria? ¿Pueden los productos químicos causar trastornos congénitos? El clínico tiene que hacer frente constantemente a controversias relacionadas con causas y efectos. Por esta razón, el lector de la literatura médica debe comprender el concepto de causalidad manejado por los investigadores.

Un concepto clínico práctico de causalidad es el denominado *causa contribuyente*. Se trata de una definición empírica que requiere el cumplimiento de los siguientes criterios: 1) la característica referida como la "causa" está asociada con la enfermedad (efecto); esto es, la causa y la enfermedad afectan al mismo individuo con más frecuencia que la esperada solo por azar; 2) demostración de que la causa precede al efecto; es decir, la causa actúa antes de que se desarrolle la enfermedad, y 3) demostración de que la modificación exclusiva de la causa altera la probabilidad del efecto (enfermedad). El proceso de análisis, que incluye pruebas de significación estadística y ajustes, ayuda a determinar si existe una asociación y si esta es producida por algún sesgo conocido. Sin embargo, para cumplir el segundo y el tercer criterio necesitamos basarnos en algo más que en análisis estadísticos. Demostrar que una causa precede a una enfermedad puede parecer sencillo, pero veamos dos estudios hipotéticos en los cuales los autores pudieron haberse dejado engañar al creer que habían demostrado que la causa precedió al efecto.

Dos investigadores realizaron un estudio de casos y controles sobre los fármacos que tomaron varios pacientes con infarto de miocardio (IM) durante la semana anterior al infarto. Mediante el estudio intentaban buscar las causas que desencadenaron la enfermedad. Se hizo una comparación de estos pacientes con otros que habían sido hospitalizados para cirugía electiva. Los autores observaron que la probabilidad de tomar aspirina o antiácidos de los pacientes con IM fue 10 veces más alta que la de los controles durante la semana precedente al ingreso y concluyeron que la toma de aspirinas y de antiácidos se asociaba con IM posterior.

Los autores creyeron que habían demostrado el cumplimiento no solo del primer criterio de causalidad (esto es, una asociación), sino también del segundo criterio, según el cual la causa precede al efecto. Pero, ¿lo hicieron? Los individuos que padecen angina antes de un IM, pueden malinterpretar el dolor y tratar de aliviarlo automedicándose con aspirinas o antiácidos. Por lo tanto, toman la medicación para tratar la enfermedad y no, verdaderamente, antes de tener la enfermedad. Con este estudio no se consiguió demostrar que la "causa" precede al "efecto", porque no se aclaró si la enfermedad condujo a los pacientes a tomar la medicación o si la medicación desencadenó la enfermedad. Este ejemplo muestra las dificultades potenciales halladas al tratar de separar la causa y el efecto en los estudios de casos y controles. No obstante, estos estudios pueden aportar pruebas convincentes de que la "causa" precede al "efecto". Este es el caso cuando se dispone de buena documentación sobre características previas que no son influidas por el conocimiento de la aparición de la enfermedad.

Los estudios de cohortes o prospectivos frecuentemente ofrecen ventajas para demostrar que la posible causa antecede al efecto. No obstante, el siguiente ejemplo sirve para indicar que, incluso en los estudios de cohortes, podemos encontrarnos con dificultades a la hora de determinar si la causa precede al efecto.

En un estudio se comparó a un grupo de 1 000 pacientes que habían dejado de fumar en el último año con 1 000 fumadores de cigarrillos, apareados según el total de paquetes de cigarrillos-año fumados. Los dos grupos se siguieron durante 6 meses, para determinar la frecuencia con que desarrollaron cáncer de pulmón.

En el estudio se observó que 5% de los que habían dejado de fumar y 0,1% de los controles desarrollaron cáncer de pulmón. Los autores concluyeron que dejar de fumar era una característica previa asociada al desarrollo de cáncer de pulmón, y, consecuentemente, aconsejaron a los fumadores actuales que continuaran fumando.

En este ejemplo, los casos parecen haber dejado de fumar antes de desarrollar cáncer de pulmón, pero ¿qué ocurre si los fumadores dejan de fumar a causa de los síntomas producidos por el cáncer de pulmón? Si esto es cierto, entonces el cáncer de pulmón es el que obliga a la gente a dejar de fumar, y no al revés. Por eso, se debe tener cuidado al aceptar que la causa hipotética precede al efecto. La capacidad de los estudios de cohortes para establecer que la causa precede al efecto aumenta cuando el tiempo transcurrido entre la causa y el efecto es superior al del ejemplo. Los intervalos cortos de tiempo dejan todavía abierta la posibilidad de que la causa haya sido influenciada por el efecto en lugar de lo contrario.

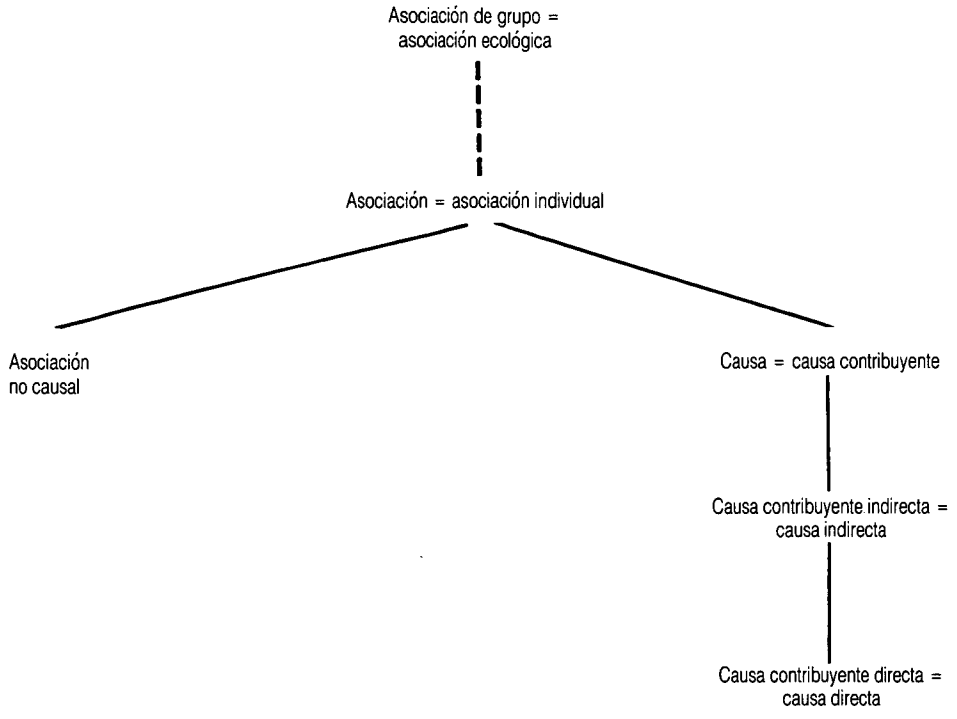
Aunque se haya establecido firmemente que la causa precede al efecto, es necesario demostrar que al modificar la causa se altera la probabilidad de que ocurra el efecto. Este criterio puede cumplirse realizando un estudio de intervención en el cual se modifique la causa y se determine si ello contribuye posteriormente a modificar la probabilidad de que ocurra el efecto. Idealmente, este criterio se cumple llevando a cabo un ensayo clínico controlado, como comentaremos en el capítulo 11. Es importante reconocer que la definición de causa contribuyente es empírica y que no requiere que comprendamos el mecanismo intermedio por el que esta produce el efecto. Existen numerosos ejemplos en los cuales las acciones basadas en una demostración de la causa contribuyente redujeron la frecuencia de la enfermedad, a pesar de no comprender científicamente el mecanismo por el que se produjo el resultado. La fiebre puerperal se controló mediante el lavado de las manos, antes de que se identificaran los agentes bacterianos. La malaria se controló desecando los pantanos antes de identificar su transmisión por el mosquito. Los frutos cítricos permitieron prevenir el escorbuto antes de que los ingleses oyeran hablar de la vitamina C. En la figura 6-1 se presentan las relaciones entre asociación, causa contribuyente, causa directa y causa indirecta.

Es posible que estos estudios promuevan investigaciones subsiguientes que permitan determinar los mecanismos directos mediante los cuales la causa contribuyente produce el efecto. Puede estar justificado actuar sobre la base de esta definición de causa y efecto antes de comprender los mecanismos inmediatos involucrados. Empero, los investigadores deben asegurarse de que cualquier cambio que hayan observado no esté asociado con otros cambios no medidos, que constituyan la causa contribuyente "real". En el siguiente ejemplo se ilustra esta trampa.

En una comunidad rural pobre donde la dieta era muy baja en proteínas la diarrea estaba muy difundida. Se estudió el efecto del aumento de proteínas en la dieta en zonas seleccionadas al azar mediante la introducción de cultivos de alto contenido en proteínas con métodos agrícolas modernos. El seguimiento posterior reveló una reducción de 70% en la incidencia de diarrea en las zonas de estudio y un escaso cambio en las otras zonas. Los autores concluyeron que la dieta con alto contenido proteínico previene la diarrea.

Es probable que la introducción de la agricultura moderna se asociara con cambios en el suministro de agua y saneamiento, que también pudieron haber contribuido a la reducción de la incidencia de diarrea. Por este motivo, es preciso cerciorarse cuidadosamente de que la característica seleccionada como causa es verdaderamente el factor que produjo el efecto. En otras palabras, el investigador y el lector deben ser cautelosos al considerar que la causa ha precedido verdaderamente al efecto

FIGURA 6-1. Relación entre diferentes tipos de asociaciones y distintas definiciones de causalidad



y que la presunta modificación del efecto no se ha producido por otros cambios que en realidad son causales. Aun cuando la causa contribuyente no se pueda establecer definitivamente, puede que sea necesario ejercitar nuestra mejor capacidad para decidir si existe una relación de causa y efecto. Se han establecido una serie de criterios auxiliares, accesorios y de apoyo para definir la causa contribuyente en esas situaciones. Estos criterios son los siguientes:

1. *Fuerza de la asociación*. La fuerza de una asociación entre el factor de riesgo y la enfermedad medida, por ejemplo, por la magnitud del riesgo relativo.
2. *Consistencia de la asociación*. Existe una asociación consistente cuando las investigaciones realizadas en diferentes lugares sobre diferentes tipos de pacientes producen resultados similares.
3. *Plausibilidad biológica*. La plausibilidad biológica de la relación se evalúa a partir de los principios de las ciencias básicas o clínicas.
4. *Relación dosis-respuesta*. La presencia de una relación de dosis-respuesta significa que los cambios en los niveles de exposición al factor de riesgo están asociados con cambios de dirección consistente en la frecuencia de la enfermedad.

Los datos que apoyan cada uno de estos criterios ayudan a reforzar el argumento de que un factor es realmente una causa contribuyente. Estos criterios reducen la probabilidad de que la asociación observada se deba al azar o al efecto de un sesgo. Sin embargo, los criterios no prueban la existencia de una causa contribuyente. Además, ninguno de esos cuatro criterios para establecer una causa contribuyente es esencial. Un factor de riesgo con una asociación moderada pero real puede ser, de he-

cho, una causa que forme parte de un conjunto de causas contribuyentes de una enfermedad. La consistencia no es esencial, ya que es posible que un factor de riesgo actúe en una comunidad pero no en otra. Esto puede ocurrir porque en una comunidad existan otras condiciones necesarias. La plausibilidad biológica supone que comprendemos el proceso biológico que nos ocupa. Por último, una relación dosis-respuesta, aunque sea frecuente en medicina, no es necesaria para establecer la existencia de una relación de causa-efecto, como se muestra en el siguiente estudio.

Un investigador realizó un estudio de cohortes sobre la asociación entre la radiación y el cáncer de tiroides. Encontró que el riesgo relativo de padecer cáncer de tiroides de las personas irradiadas con dosis bajas era 5. No obstante, observó también que, con niveles moderados de radiación, el riesgo relativo era 10 y que a niveles elevados era 1. El investigador concluyó que la radiación no podía causar cáncer de tiroides, dado que no existía una relación dosis-respuesta que demostrara más casos de cáncer entre las personas sometidas a radiaciones más altas.

El riesgo relativo de 10 implica una fuerte asociación entre cáncer de tiroides y radiación. Esta asociación no debe ser descartada simplemente porque el riesgo relativo disminuya a dosis más altas. Es posible que la radiación a dosis bajas y moderadas sea una causa contribuyente del cáncer de tiroides, mientras que las dosis altas maten realmente a las células y, por ello, no contribuyan a su desarrollo.

De este modo, los criterios auxiliares, accesorios y de apoyo para juzgar las causas contribuyentes son solo eso: por sí mismos no resuelven el problema. Sin embargo, pueden ayudar a apoyar el argumento a favor o en contra de la causa contribuyente. La consideración de estos criterios nos ayuda a comprender las controversias y las limitaciones de los datos.

OTROS CONCEPTOS DE CAUSALIDAD

El concepto de causa contribuyente ha sido muy útil para estudiar la causalidad de las enfermedades. No obstante, este no es el único concepto de causa que se ha utilizado en medicina clínica. En el siglo XIX, Robert Koch formuló una serie de condiciones que era necesario satisfacer antes de que un microorganismo pudiera ser considerado la causa de una enfermedad.[2] Entre las condiciones conocidas como *postulados de Koch* hay una que exige que el organismo siempre se encuentre asociado con la enfermedad. Esta condición se denomina con frecuencia *causa necesaria*.

La causa necesaria va más allá de los requisitos para establecer una causa contribuyente. Con el paso del tiempo, este requisito ha demostrado ser muy útil para estudiar las enfermedades infecciosas en circunstancias en que un solo agente es responsable de una sola enfermedad. Sin embargo, si el concepto de causa necesaria se aplica al estudio de las enfermedades crónicas, es casi imposible probar la existencia de una relación causal. Por ejemplo, aunque se haya reconocido que los cigarrillos son una causa contribuyente del cáncer de pulmón, fumar cigarrillos no es una condición necesaria para el desarrollo de dicho cáncer; no todos los que tienen cáncer de pulmón han fumado cigarrillos.

Según las reglas de la lógica estricta, la causalidad también exige el cumplimiento de una segunda condición conocida como *causa suficiente*. Esta condición afirma que si está presente la causa, también lo debe estar la enfermedad. En nuestro ejemplo de los cigarrillos y el cáncer de pulmón, esto implicaría que, si se fuman cigarrillos, siempre se desarrollará cáncer de pulmón. Por citar otro ejemplo, la mono-

[2] Last JM. *A Dictionary of Epidemiology*. Nueva York: Oxford University Press; 1988.

nucleosis infecciosa es una enfermedad clínicamente bien establecida en la que se ha aceptado que el virus de Epstein-Barr es una causa contribuyente. No obstante, también se ha demostrado que otros virus, como el citomegálico, producen el síndrome de la mononucleosis. Por añadidura, existen pruebas que demuestran que el virus de Epstein-Barr ha estado presente sin provocar la aparición del síndrome de la mononucleosis o que puede manifestarse como causa contribuyente de otras enfermedades como el linfoma de Burkitt. Por este motivo, y a pesar de que se ha demostrado que el virus de Epstein-Barr es una causa contribuyente del síndrome de la mononucleosis, no es una causa ni necesaria ni suficiente del mismo. El siguiente ejemplo ilustra las consecuencias de aplicar el concepto de la causa necesaria de la lógica formal a los estudios médicos.

En un estudio sobre los factores de riesgo de la enfermedad coronaria, los investigadores identificaron a 100 individuos de una población de 10 000 pacientes que padecieron un infarto de miocardio, a pesar de tener la tensión arterial y la concentración plasmática de colesterol normales, practicar regularmente ejercicio, no fumar, tener una personalidad de tipo B y no presentar una historia familiar de enfermedad coronaria. Los autores concluyeron que habían demostrado que la hipertensión, la concentración plasmática de colesterol elevada, la falta de ejercicio, el consumo de cigarrillos, la personalidad de tipo A y la historia familiar de enfermedad coronaria no eran las causas de la enfermedad coronaria, porque no todos los pacientes con infarto de miocardio poseían uno de esos factores de riesgo.

Los autores del estudio estaban usando el concepto de causa necesaria como concepto de causalidad. En lugar de la causa necesaria, supongamos que se ha demostrado que todos esos factores cumplían los criterios de causa contribuyente de la enfermedad coronaria. La causa contribuyente, a diferencia de la causa necesaria, no requiere que todos los que no tienen la causa tampoco tengan la enfermedad. La incapacidad de las causas contribuyentes para predecir todos los casos de la enfermedad subraya las limitaciones de nuestro conocimiento actual sobre todas las causas contribuyentes de la enfermedad coronaria. Asimismo, demuestra nuestro estado de ignorancia actual, porque si se conocieran todas la causas contribuyentes, en todos los que padecen la enfermedad se observaría al menos uno de esos factores. Por eso, ni siquiera el establecimiento de una causa contribuyente implica que esta estará presente necesariamente en todos y cada uno de los casos.

En resumen, la definición clínicamente útil de causalidad se conoce como causa contribuyente. Esta definición exige demostrar que la presunta causa precede al efecto y que la modificación de la causa modifica el efecto en algunos individuos. No requiere, sin embargo, que todos los sujetos que no presentan la causa contribuyente no presenten el efecto, ni que todos los que la posean desarrollen el efecto. En otras palabras, una causa clínicamente útil puede no ser ni necesaria ni suficiente, pero tiene que ser contribuyente. Su presencia debe aumentar la probabilidad de que se manifieste la enfermedad y su ausencia debe reducir dicha probabilidad.

CAPÍTULO 7

EXTRAPOLACIÓN

En los capítulos precedentes hemos ilustrado los errores que se pueden cometer al asignar pacientes a los grupos de estudio y de control, al valorar el desenlace de un estudio y al analizar e interpretar sus resultados. Una vez completado este proceso, el investigador se pregunta qué significado tiene todo esto para los individuos no incluidos en el estudio y para las situaciones no abordadas directamente en el estudio. Para averiguar el significado que tiene la investigación en otros grupos o situaciones, el investigador debe extrapolar los resultados del estudio a situaciones nuevas y potencialmente diferentes.[1] Empezaremos viendo cómo podemos usar los resultados de un estudio para extrapolarlos a individuos, grupos de riesgo o comunidades similares compuestas por individuos con y sin los factores que se han estudiado. Luego examinaremos la extrapolación a nuevas situaciones y a nuevos tipos de individuos.

EXTRAPOLACIÓN A INDIVIDUOS

El primer paso para extrapolar los resultados de un estudio consiste en valorar el significado global de los resultados para un individuo concreto, similar a los individuos incluidos en la investigación. Al hacerlo, suponemos que los hallazgos del estudio son tan válidos para otros individuos con el factor de riesgo estudiado como lo fueron para los individuos que realmente participaron en la investigación.

En muchos estudios de casos y controles y de cohortes se estima la razón de productos cruzados o el riesgo relativo asociado con el desarrollo de la enfermedad cuando el factor de riesgo está presente comparado con cuando no está presente. La razón de productos cruzados y el riesgo relativo nos informan sobre la fuerza de la relación entre el factor de riesgo y la enfermedad. Si existe una relación de causa-efecto y el efecto del factor de riesgo es completamente reversible, los riesgos relativos nos ofrecen una información importante para el paciente individual. Un riesgo relativo de 10 indica al paciente individual que, en promedio, se multiplicará por 10 el riesgo de desarrollar la enfermedad en un determinado período de tiempo si tiene el factor de riesgo, comparado con su riesgo si no tiene el factor.[2]

Sin embargo, el riesgo relativo no nos dice nada acerca de la magnitud absoluta de la probabilidad o riesgo de desarrollar la enfermedad si el factor de riesgo está presente comparado con cuando no está presente. Un riesgo relativo de 10 puede indicar un aumento de riesgo de 1 por 1 000 000 para aquellos sin el factor de riesgo a 1 por 100 000 para aquellos con el factor. Por otro lado, puede indicar un aumento de 1 por 100 para los que no tienen el factor de riesgo a 1 por 10 para los que lo

[1] Se puede argüir que el uso de datos para sacar conclusiones sobre individuos no incluidos en un estudio es siempre una extrapolación. En la presente situación, estamos extrapolando de un período a otro.

[2] La corrección con que una estimación del riesgo relativo es aplicable a un individuo está determinada, de hecho, por el grado de semejanza entre los individuos incluidos en el estudio y aquel al que deseamos aplicar los resultados. La aplicación de los resultados a un individuo supone que la muestra estudiada se compone totalmente de personas exactamente iguales a aquel individuo. No es suficiente con que se hayan incluido en el estudio personas como el individuo en cuestión.

tienen. Por ello, y a pesar de tener el mismo riesgo relativo, el riesgo absoluto puede ser muy diferente para los individuos.

La incapacidad de comprender el concepto de riesgo absoluto puede conducir al siguiente tipo de error de extrapolación.

Un paciente leyó que el riesgo relativo de leucemia aumenta cuatro veces con el uso de una nueva quimioterapia para el tratamiento del cáncer de mama, mientras que el riesgo relativo de curar dicho cáncer con la quimioterapia es de 3. Por consiguiente, pensó que la quimioterapia no merecía la pena.

Sin embargo, el riesgo de morir por cáncer de mama para el paciente es bastante mayor que el riesgo de padecer leucemia. La aparición infrecuente y tardía de la leucemia quiere decir que, incluso en presencia de un factor que aumenta cuatro veces el riesgo de padecerla, el riesgo absoluto todavía será muy pequeño comparado con el alto riesgo de morir por cáncer de mama. El paciente no ha comprendido la importante diferencia entre el riesgo relativo y el absoluto. Por eso, cuando se extrapolan los resultados de un estudio a un individuo concreto, es deseable tener información tanto sobre el riesgo relativo como sobre el riesgo absoluto.

EXTRAPOLACIÓN A LOS GRUPOS DE RIESGO

El riesgo relativo y el riesgo absoluto se emplean con frecuencia para efectuar estimaciones acerca de pacientes individuales. A veces, sin embargo, estamos más interesados en el efecto que puede tener un factor de riesgo sobre grupos de individuos con el factor de riesgo o sobre una comunidad de individuos con y sin el factor de riesgo.

Cuando se valora el efecto de un factor de riesgo sobre un grupo de individuos, empleamos el concepto conocido como *riesgo atribuible porcentual*.[3] El cálculo del riesgo atribuible porcentual no requiere que exista una relación de causa-efecto. No obstante, cuando existe una causa contribuyente, el riesgo atribuible porcentual nos informa del porcentaje de una enfermedad que puede eliminarse entre los que tienen el factor de riesgo si se pueden suprimir completamente los efectos del factor de riesgo.[4]

El riesgo atribuible porcentual se define como:

$$\frac{\begin{array}{c}\text{Riesgo de la enfermedad}\\ \text{si el factor de riesgo}\\ \text{está presente}\end{array} - \begin{array}{c}\text{Riesgo de la enfermedad}\\ \text{si el factor de riesgo}\\ \text{está ausente}\end{array}}{\text{Riesgo de la enfermedad si el factor de riesgo está presente}} \times 100\%$$

El riesgo atribuible porcentual se calcula más fácilmente a partir del riesgo relativo mediante la siguiente fórmula:

$$\text{Riesgo atribuible porcentual} = \frac{\text{Riesgo relativo} - 1}{\text{Riesgo relativo}} \times 100\%$$

Cuando el riesgo relativo es menor que 1

$$\text{Riesgo atribuible porcentual} = 1 - \text{riesgo relativo} \times 100\%$$

[3] El *riesgo atribuible porcentual* también se denomina *fracción atribuible* (en los expuestos), *fracción etiológica* (en los expuestos), *proporción atribuible* (en los expuestos), *porcentaje de reducción del riesgo* y *tasa de eficacia protectora*.

[4] Esta interpretación del riesgo atribuible exige que los efectos del factor de riesgo se puedan eliminar inmediata y completamente.

Esto es, si

Riesgo relativo	Riesgo atribuible porcentual
1	0
2	50%
4	75%
10	90%
20	95%

Observe que incluso un riesgo relativo de 2 puede estar asociado con una reducción de 50% en la enfermedad entre los sujetos con el factor de riesgo.

La incapacidad para entender este concepto puede conducir al siguiente error de extrapolación:

Se realizó un estudio de cohortes bien diseñado comparando hombres que hacían ejercicio regularmente con hombres que no lo hacían, los cuales fueron apareados según los factores de riesgo de la enfermedad coronaria. En el estudio se encontró que el riesgo relativo de enfermedad coronaria en los que no hacían ejercicio físico era de 1,5. Los investigadores concluyeron que, aunque fuera verdad, este riesgo relativo era demasiado pequeño para tener importancia práctica.

A pesar de que el riesgo relativo era solamente 1,5, observe que se convierte en un riesgo atribuible porcentual sustancial:

$$\text{Riesgo atribuible porcentual} = \frac{1{,}5 - 1}{1{,}5} = \frac{0{,}5}{1{,}5} = 33\%$$

Este resultado indica que, entre los que no hacían ejercicio físico regularmente, se podía eliminar como máximo un tercio de su riesgo, si se pudiera suprimir la falta de ejercicio. Ello puede representar un elevado número de individuos, dado que la enfermedad coronaria es frecuente y la falta de ejercicio físico es un factor de riesgo común.

Con frecuencia, es difícil transmitir la información contenida en el riesgo absoluto, el riesgo relativo y el riesgo atribuible porcentual. Otro modo de expresar esta información —que es aplicable a los estudios de cohortes y a los ensayos clínicos controlados— se conoce como el *número de pacientes que es preciso tratar*.[5] Esta cifra establece una información clínicamente importante: ¿cuántos pacientes similares a los del estudio es necesario tratar, considerando el paciente promedio del estudio, para obtener un desenlace malo menos o uno bueno más? El número se calcula suponiendo que el grupo A tiene un mejor desenlace que el grupo B:

$$\text{Número de pacientes que es preciso tratar} = \frac{1}{\text{Probabilidad del desenlace en el grupo A} - \text{Probabilidad del desenlace en el grupo B}}$$

De este modo, si una investigación demostró una reducción de la enfermedad coronaria en un período de 5 años de 20 por 1 000 a 10 por 1 000, el número de pacientes que es preciso tratar durante 5 años para reducir un caso de la enfermedad se calcularía del siguiente modo:

[5] Véase Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med*. 1988;318:1728–1733.

$$\text{Número de pacientes que es preciso tratar} = \frac{1}{20/1\,000 - 10/1\,000} = \frac{1}{10/1\,000} = 100$$

El número de pacientes que es preciso tratar ofrece a menudo una información clínicamente más útil para interpretar los datos de la investigación clínica que otros estadísticos de síntesis tales como el riesgo relativo de 2, un riesgo atribuible porcentual de 50% o incluso un riesgo atribuible de 20 por 1 000 frente a otro de 10 por 1 000.[6]

EXTRAPOLACIÓN A UNA COMUNIDAD

Cuando se extrapolan los resultados de un estudio a una comunidad de individuos con y sin el factor de riesgo, necesitamos utilizar otra medida de riesgo conocida como *riesgo atribuible poblacional porcentual (RAP)*.[7] El riesgo atribuible poblacional porcentual indica el porcentaje del riesgo de enfermedad en una comunidad que está asociado con la exposición a un factor de riesgo.[8] El cálculo del riesgo atribuible poblacional porcentual requiere que conozcamos otros datos además del riesgo relativo. Exige conocer o que seamos capaces de estimar el porcentaje de individuos de la comunidad que poseen el factor de riesgo. Si conocemos el riesgo relativo y el porcentaje de individuos con el factor de riesgo (b), podemos calcular el riesgo atribuible poblacional porcentual empleando la siguiente fórmula:

$$\text{Riesgo atribuible poblacional porcentual} = \frac{b\,(\text{riesgo relativo}) - 1}{b\,(\text{riesgo relativo} - 1) + 1} \times 100\%$$

Esta fórmula nos permite relacionar el riesgo relativo, b, y el riesgo atribuible poblacional del siguiente modo:

Riesgo relativo	b	Riesgo atribuible poblacional (aproximado)
2	1%	1%
4	1%	3%
10	1%	8%
20	1%	16%
2	10%	9%
4	10%	23%
10	10%	46%
20	10%	65%
2	50%	33%
4	50%	60%
10	50%	82%
20	50%	90%
2	100%	50%
4	100%	70%
10	100%	90%
20	100%	95%

[6] El número de pacientes que es preciso tratar también se puede calcular para los efectos adversos. Esto permite realizar una comparación directa entre el número necesario para prevenir un efecto adverso y el número de pacientes que es preciso tratar para producir un efecto secundario.

[7] El *riesgo atribuible poblacional porcentual* también se denomina *fracción atribuible* (poblacional), *proporción atribuible* (en la población) y *fracción etiológica* (en la población).

[8] Esta interpretación del RAP exige la existencia de una relación de causa-efecto y que las consecuencias de la "causa" sean inmediata y completamente reversibles.

Observe que, si el factor de riesgo es poco frecuente (1%, por ejemplo), el riesgo relativo tiene que ser considerable para que el riesgo atribuible poblacional porcentual alcance una magnitud importante. Por otro lado, si el factor de riesgo es común, por ejemplo, 50%, incluso un riesgo relativo pequeño indica que el impacto potencial en la comunidad puede ser sustancial. Además, fíjese que, cuando la prevalencia del factor de riesgo es de 100% (esto es, cuando todo el mundo posee el factor de riesgo), el riesgo atribuible poblacional porcentual iguala al riesgo atribuible porcentual.

La incapacidad de comprender el concepto de riesgo atribuible poblacional puede conducir al siguiente error de extrapolación:

Unos investigadores informaron que una forma hereditaria de hipercolesterolemia conocida como hiperlipidemia de tipo III aparece en 1 de cada 100 000 estadounidenses. También notificaron que el riesgo relativo de desarrollar la enfermedad coronaria de los que padecen esta enfermedad es 20. Los autores concluyeron que la curación de la hiperlipidemia de tipo III tendría un impacto sustancial sobre el problema nacional de la enfermedad coronaria.

Con estos datos y con nuestra fórmula para calcular el riesgo atribuible poblacional porcentual, podemos ver que la eliminación de la enfermedad coronaria secundaria a la hiperlipidemia de tipo III está asociada con un riesgo atribuible poblacional de un cincuentavo de 1%. Por tanto, por el hecho de que la hiperlipidemia de tipo III sea un factor de riesgo tan raro, no se puede esperar que la eliminación de su impacto tenga consecuencias sustanciales sobre la frecuencia global de la enfermedad coronaria.

EXTRAPOLACIÓN A SITUACIONES NUEVAS

La extrapolación a situaciones nuevas o a diferentes tipos de individuos es aun más difícil, y muchas veces es la etapa más complicada de una investigación para el lector. Resulta difícil ya que, por lo general, el investigador y los revisores no pueden tratar adecuadamente los aspectos que interesan a un lector concreto. Esto le corresponde a usted, lector. El investigador no conoce la comunidad ni los pacientes de los investigadores. A pesar de la dificultad que entraña extrapolar los datos de una investigación, es imposible practicar la medicina sin extrapolar los resultados de investigaciones clínicas. Con frecuencia, debemos ir más allá de los datos basados en supuestos razonables. Si uno está poco dispuesto a realizar extrapolaciones, se limitará a aplicar solamente los resultados de investigaciones a los pacientes que son prácticamente idénticos a los de un estudio.

A pesar de la importancia de la extrapolación, es preciso conocer los tipos de errores que se pueden cometer, si esta no se realiza cuidadosamente. Cuando se extrapola a grupos o a situaciones diferentes, se pueden cometer dos tipos de errores: uno, porque se extrapole más allá de los datos y, dos, porque existan diferencias entre el de estudio y el *grupo objetivo*, siendo este el grupo sobre el que deseamos sacar conclusiones.

MÁS ALLÁ DEL INTERVALO DE LOS DATOS

En los estudios clínicos, los individuos suelen estar expuestos durante un período de tiempo determinado y con un intervalo limitado de exposición a los factores que se consideran asociados con el desenlace. Los investigadores pueden estudiar un factor como la hipertensión, que produce un accidente vascular cerebral, o un agente terapéutico como un antibiótico, que está asociado con la curación de una infección. En cada caso, la interpretación debe limitarse al intervalo y a la duración de

la hipertensión padecida por los sujetos o a la dosis y duración del antibiótico empleado en el estudio. Cuando los investigadores extraen conclusiones que exceden los límites del intervalo y de la duración experimentados por los sujetos, frecuentemente están haciendo suposiciones injustificadas. Pueden suponer que una exposición más prolongada continuará produciendo el mismo efecto experimentado por los sujetos del estudio. El siguiente ejemplo muestra un error potencial que resulta de extrapolar más allá del intervalo de los datos.

En 100 hipertensos resistentes a la medicación, se probó un nuevo fármaco antihipertensor. Se observó que este medicamento reducía la tensión arterial diastólica en los 100 hipertensos de 120 a 110 mmHg cuando se administraba a una dosis de 1 mg/kg, y de 110 a 100 mmHg, a dosis de 2 mg/kg. Los autores concluyeron que este agente, administrado a dosis de 3 mg/kg, sería capaz de disminuir la tensión diastólica de 100 a 90 mmHg.

Es posible que mediante la evidencia clínica se pueda documentar la eficacia del nuevo medicamento cuando se administra a dosis de 3 mg/kg. Sin embargo, esa documentación depende de los resultados de la prueba empírica. Muchos fármacos antihipertensores alcanzan su máxima eficacia a cierta dosis, y esta no aumenta a dosis más elevadas. Concluir sin pruebas experimentales que dosis más altas de ese hipotensor producen efectos más intensos es lo mismo que efectuar una extrapolación lineal que sobrepase el intervalo de los datos observados.

Otro tipo de error asociado con la extrapolación más allá de los datos se relaciona con los efectos indeseables potenciales experimentados a una exposición más prolongada, como puede apreciarse en el siguiente ejemplo.

En un estudio de un año de duración sobre los efectos de la administración de estrógenos aislados a 100 mujeres menopáusicas, se observó que estos fármacos aliviaban las sofocaciones y reducían la tasa de osteoporosis, en contraposición a la ausencia de mejoramiento sintomático de las mujeres, apareadas según la edad, a las que se administró placebo. Los autores no detectaron efectos indeseables atribuibles a los estrógenos y concluyeron que estos agentes eran seguros y eficaces. Por lo tanto, recomendaron su administración a largo plazo, iniciando el tratamiento al comienzo de la menopausia.

Los autores extrapolaron los resultados de un período de seguimiento de un año a la administración de los estrógenos a largo plazo. No existen pruebas que demuestren que si su administración durante un año es segura, también lo será su administración a largo plazo. Es improbable que todos los efectos adversos a largo plazo aparezcan en un año de estudio. Por consiguiente, los autores, al sobrepasar el intervalo de los datos observados, realizaron una extrapolación potencialmente peligrosa.

A veces, la extrapolación lineal puede ser necesaria en la práctica de la medicina, pero los clínicos deben reconocer que este es el tipo de extrapolación realizada y estar a la expectativa de nuevos datos que pueden socavar esos supuestos y cuestionar la conclusión obtenida mediante extrapolación lineal.

DIFERENCIAS EN EL GRUPO OBJETIVO

Cuando se extrapola a un grupo objetivo, es importante considerar la forma como ese grupo se diferencia del grupo de la investigación. El siguiente caso ilustra de qué forma las diferencias entre países pueden complicar las extrapolaciones de un país a otro.

En un estudio sobre la sociedad japonesa y la americana, se estimó que las prevalencias de hipertensión y de tabaquismo entre los japoneses eran de 20 y 80%, respectivamente. Ambas son causas contribuyentes conocidas de enfermedad coronaria en los estadounidenses. En estos últimos, la prevalencia de hipertensión era de 10% y la de tabaquismo, 40%. Varios estudios de casos y controles realizados en el Japón no demostraron una asociación entre hipertensión o consumo de tabaco y enfermedad coronaria, mientras que estudios similares efectuados en los Estados Unidos de América demostraron una asociación estadísticamente significativa. Los autores concluyeron que la hipertensión y el consumo de cigarrillos deben proteger a los japoneses del infarto de miocardio.

Los autores extrapolaron los resultados de una cultura a otra muy diferente, sin tener en cuenta que hay otras formas de explicar los datos observados. Si los estadounidenses poseen con frecuencia otro factor de riesgo —como la concentración plasmática de colesterol elevada— que es raro en el Japón, este factor podría invalidar el papel que desempeñan el tabaquismo y la hipertensión y ser en parte responsable de la tasa elevada de infartos de miocardio de la población estadounidense.

Realizar una extrapolación dentro de cada país también puede ser difícil cuando existen diferencias entre el grupo investigado y el grupo objetivo al que se quieren aplicar los hallazgos. Este principio se ejemplifica a continuación.

En un estudio llevado a cabo durante un año de esquimales de Alaska se investigó el efecto del tratamiento con isoniazida de las personas con resultados limítrofes (6–10 mm) a las pruebas de intradermorreacción a la tuberculina. La prevalencia en dicha población de las pruebas intradérmicas limítrofes fue 2 por 1 000. Para realizarlo, se administró isoniazida a 200 esquimales con pruebas intradérmicas limítrofes y placebo a 200 esquimales con el mismo resultado en esas pruebas. Entre los pacientes tratados con placebo aparecieron 20 casos de tuberculosis activa y solo uno entre los tratados con isoniazida. Los resultados fueron estadísticamente significativos a un nivel de significación de 0,05. Un funcionario de salud de Georgia, donde la frecuencia de las pruebas intradérmicas con resultados limítrofes es de 300 por 1 000, quedó muy impresionado por esos resultados. Por ello, decidió tratar a todos los pacientes de ese estado que tenían pruebas intradérmicas limítrofes con isoniazida durante 1 año.

Al extrapolar los resultados de aquel estudio a la población de Georgia, el funcionario de salud supuso que las pruebas intradérmicas con resultados limítrofes tenían el mismo significado en la población de Georgia que en los esquimales de Alaska. Sin embargo, otros datos sugieren que muchos resultados limítrofes en Georgia no son debidos a la exposición a la tuberculosis. En lugar de indicar la presencia de tuberculosis, son causados frecuentemente por una micobacteria atípica cuya infección conlleva un pronóstico mucho mejor y que no responde con seguridad a la isoniazida. El funcionario de salud ignoraba el hecho de que las pruebas intradérmicas limítrofes tienen un significado muy distinto en los esquimales que en los residentes en Georgia. Por desconocer este nuevo factor entre los habitantes de Georgia, el funcionario de salud corrió el riesgo de someter a un número elevado de individuos a un tratamiento inútil y potencialmente peligroso.

La extrapolación de los resultados de un estudio siempre es un paso difícil, aunque extremadamente importante, de la lectura de la literatura médica. La extrapolación exige, en primer lugar, preguntarse qué significan los resultados para los individuos semejantes y promedio incluidos en la investigación. En segundo término, uno puede preguntarse lo que significan los resultados para los grupos en riesgo similares y, finalmente, para las comunidades compuestas por individuos con y sin las características estudiadas. A menudo, el lector deseará avanzar y extender la extrapola-

ción a los individuos y situaciones que son diferentes de las estudiadas. Al extrapolar más allá de los datos observados, se deben tener en cuenta las diferencias entre los tipos de individuos incluidos en la investigación y el grupo objetivo. El reconocimiento de los supuestos que realizamos al extrapolar nos obliga a mantener los ojos bien abiertos ante la eventual aparición de nueva información que cuestione estos supuestos e invalide potencialmente nuestras conclusiones.

CAPÍTULO 8

DISEÑO DEL ESTUDIO

Una vez revisados los requisitos de la aplicación correcta de los componentes del marco uniforme, volvamos al principio para formular algunas preguntas básicas.

1. ¿Estaban definidos adecuadamente los objetivos del estudio?
2. ¿Cuál es el tipo de estudio? ¿Es apropiado para responder a las preguntas planteadas?
3. ¿Cuál es el tamaño de la muestra? ¿Es suficiente para responder a las preguntas del estudio?

Las respuestas a estos interrogantes le dirán al lector si los investigadores escogieron un diseño de estudio apropiado; esto es, aquel que define y puede responder a las preguntas planteadas.

OBJETIVO DEL ESTUDIO

Supongamos que unos investigadores desean estudiar los efectos orgánicos de la hipertensión arterial. La imposibilidad de realizar biopsias renales y angiografías cerebrales puede obligarlos a explorar detenidamente el fondo del ojo. Supongamos que otros desean investigar los efectos a largo plazo de un nuevo fármaco para prevenir la osteoporosis y que el tiempo, el dinero y el deseo de publicar limitan su investigación a sus efectos a corto plazo sobre el metabolismo y la densidad ósea. Estos ejemplos ilustran la importancia de que los investigadores y el lector distingan entre lo que idealmente desearían estudiar aquellos y lo que de hecho estudian.

Al definir los objetivos del estudio es esencial formular una hipótesis específica. Cuando se estudia el daño orgánico producido por la hipertensión, los investigadores pueden formular la hipótesis de que el grado de daño orgánico está asociado con el grado de hipertensión. Sin embargo, esta hipótesis no es suficientemente concreta para ser contrastada. Por el contrario, para ello, es preciso formular una hipótesis específica como: el aumento del estrechamiento de las arterias de la retina, medido mediante fotografías sucesivas tomadas durante tres años de observación, está asociado con un aumento de la tensión arterial diastólica, utilizando como medida la media de tres mediciones realizadas al inicio del estudio. Esta última afirmación constituye una hipótesis de estudio específica que se puede abordar por medio de una investigación.

La incapacidad para definir claramente las hipótesis que se desean contrastar dificulta al investigador y al lector la selección y la valoración del diseño del estudio, respectivamente. También hace más difícil determinar si se alcanzaron los objetivos del estudio. En última instancia, como se señaló al presentar las pruebas de sig-

nificación estadística, las pruebas de significación habituales no se pueden aplicar si no se define un resultado o desenlace específico que se pueda valorar.

EVALUACIÓN DEL TIPO DE ESTUDIO

Una vez definidas las hipótesis específicas del estudio, el lector está preparado para identificar el tipo de estudio realizado y evaluar su idoneidad. Rara es la ocasión en la que solo hay un tipo de diseño apropiado para responder a la pregunta del estudio. A veces, las desventajas de un tipo de diseño pueden obstaculizar notablemente el cumplimiento de los objetivos del estudio. Para ayudar al lector a juzgar la idoneidad del diseño escogido, esbozaremos las ventajas y desventajas de los tipos básicos de estudio.

Los estudios de casos y controles o retrospectivos presentan la ventaja distintiva de que permiten estudiar enfermedades muy poco frecuentes. Si la enfermedad es rara, con los estudios de casos y controles se pueden detectar diferencias entre los grupos empleando muchos menos individuos de los que se necesitarían con otros diseños. El tiempo necesario para realizar un estudio de casos y controles es mucho menor, porque la enfermedad ya se ha manifestado. Este tipo de estudio permite a los investigadores examinar simultáneamente asociaciones entre varios factores y una enfermedad. Por ejemplo, es posible examinar diversas variables que puedan estar asociadas con el cáncer de colon. En el mismo estudio, se podrían investigar la dieta anterior, la cirugía, la colitis ulcerosa, los pólipos, el alcohol, los cigarrillos, los antecedentes familiares y muchas otras variables.

La mayor objeción a estos estudios es su tendencia a presentar una serie de errores metodológicos y sesgos, que ya se indicaron en los estudios hipotéticos de capítulos anteriores. Muchos sesgos, como el de declaración y el de recuerdo, comprometen la exactitud de los datos referentes a las características previas. Sin embargo, el estudio de casos y controles puede ser el método adecuado para revelar la existencia de una asociación previa, especialmente cuando no hay razones para creer que el conocimiento del investigador o de los sujetos estudiados sobre la presencia de la enfermedad influye en la valoración de los datos del pasado.

La ventaja principal de los estudios de cohortes es que ofrecen más garantías de que la característica estudiada precede al desenlace estudiado. Esta es una distinción fundamental cuando se valora una relación de causa-efecto. Los *estudios de cohortes concurrentes*, en los que se sigue la evolución de los pacientes durante largos períodos, son caros y requieren mucho tiempo. No obstante, es posible realizar un estudio de cohortes sin un período de seguimiento tan largo. Cuando existen datos fiables de épocas anteriores sobre la presencia o ausencia de la característica estudiada, estos se pueden utilizar para realizar un *estudio de cohortes no concurrentes*. En un estudio de cohortes no concurrentes la asignación de los individuos a los grupos se lleva a cabo a partir de los datos del pasado. Después de la asignación, el investigador puede investigar si la enfermedad se desarrolló posteriormente.

Por ejemplo, si conociéramos las concentraciones de colesterol de un grupo de adultos jóvenes medidas 15 años antes del inicio del estudio actual, podríamos seguir prospectivamente a los pacientes que todavía no han desarrollado la consecuencia clínica de la hipercolesterolemia para valorar el desarrollo ulterior de enfermedad coronaria, accidentes vasculares cerebrales u otras consecuencias que podrían aparecer poco tiempo después de iniciar el estudio. El elemento fundamental que caracteriza a todos los estudios de cohortes es la identificación de los individuos del gru-

po de estudio y del grupo control sin conocer si se ha desarrollado la enfermedad estudiada.[1]

Los estudios de cohortes permiten delimitar diversas consecuencias que pueden estar asociadas con un único factor de riesgo. Los investigadores pueden estudiar simultáneamente la relación entre la hipertensión y el accidente vascular cerebral, el infarto de miocardio, la insuficiencia cardíaca o la enfermedad renal. Los estudios de cohortes pueden ayudar a comprender con más detalle el efecto de un factor etiológico sobre varios desenlaces. No obstante, la posibilidad de que con estos estudios se descubran nuevos factores etiológicos es menor que con los de casos y controles.

Ambos tipos de estudios son observacionales; esto es, en ellos las características y los desenlaces de los individuos no se imponen, sino que se observan. Los ensayos clínicos aleatorios se distinguen de los estudios observacionales en que el investigador interviene asignando al azar a los individuos al grupo de control y al de estudio. La capacidad de asignar a los individuos contribuye a asegurar que la característica estudiada, y no alguna predisposición subyacente, produce los resultados del estudio. Cuando se realizan adecuadamente, los ensayos clínicos aleatorios pueden cumplir los tres criterios de causa contribuyente.

En los capítulos 11 y 12 examinaremos en profundidad las ventajas e inconvenientes de los ensayos clínicos controlados.

Puede ser útil examinar una posible secuencia de estudios realizados para comprobar la existencia de una causa contribuyente. Muchas veces, los investigadores inician una investigación con un estudio de casos y controles con objeto de indagar la existencia de posibles causas.[2] Estos estudios ofrecen la ventaja de la rapidez, el bajo costo y la capacidad de investigar numerosas causas a la vez. Además, tienen por objeto demostrar la existencia de asociaciones o relaciones entre factores. A veces, pueden ser fiables para garantizar que la causa precede al efecto, si bien pueden dejar algunas dudas sobre cuál precede a cuál.

Una vez que se ha comprobado la existencia de una asociación en uno o más estudios de casos y controles, los investigadores llevan a cabo frecuentemente un estudio de cohortes concurrentes. A pesar de la necesidad de interpretar los datos cuidadosamente —como se demostró en el ejemplo del abandono del tabaco—, con los estudios de cohortes concurrentes a menudo es posible comprobar que la causa precede al efecto.

Después de demostrar que una posible causa precede al efecto, los investigadores pueden utilizar un estudio de intervención, por ejemplo, un ensayo clínico aleatorio, para comprobar que la modificación de la causa altera el efecto. En este estudio, los individuos se asignan al azar y a ciegas al grupo de estudio y al de control. Solo el grupo de estudio es expuesto a la posible causa o al tratamiento propuesto. El ensayo clínico aleatorio cumple idealmente con los tres criterios de causa contribuyente y, por ello, es un instrumento potente para demostrar que una determinada causa es contribuyente.

En teoría, esta secuencia de estudios funcionaría de la siguiente manera: para comprobar que los estrógenos sin progesterona son una causa contribu-

---

[1] Los estudios de cohortes se realizan cada vez con más frecuencia utilizando bases de datos que se han completado antes de iniciar el estudio. Esta situación representa el caso extremo de los estudios no concurrentes, a veces denominados *estudios de cohortes retrospectivos*. El elemento clave que transforma a estos estudios en estudios de cohortes es el hecho de que la identificación de los sujetos para su inclusión en el estudio se realiza sin saber si han desarrollado la enfermedad.

[2] Con la creciente disponibilidad de grandes bases de datos, los investigadores pueden empezar realizando un estudio de cohortes no concurrentes, que también puede llevarse a cabo rápidamente y a bajo costo.

yente del cáncer de útero, un investigador podría utilizar, en primer lugar, un estudio de casos y controles con el que se examinarían diversas variables, incluyendo la asociación hipotética entre los estrógenos y el cáncer de útero. Si se encontrara una asociación, se podría realizar un estudio de cohortes concurrentes para establecer con más firmeza que la toma de estrógenos sin progesterona precede al desarrollo de un cáncer de útero. Los investigadores desearían estar seguros de que los estrógenos no se están administrando para tratar una pérdida de sangre que pudiera ser un signo de cáncer de útero. En un estudio de cohortes concurrente se seleccionarían grupos similares de mujeres que han tomado estrógenos y de mujeres que no los han tomado; se seguiría a ambos grupos durante un período de tiempo y se investigaría si las mujeres que toman estrógenos desarrollan cáncer de útero con más frecuencia que las que no los toman. Este estudio de cohortes concurrente puede demostrar más firmemente que la toma de estrógenos precede al desarrollo de un cáncer de útero.

En teoría, la investigación proseguiría con un ensayo clínico controlado en el cual las mujeres se asignarían al azar al grupo de las que toman estrógenos sin progesterona o al de las que toman placebo. Sin embargo, después de obtener pruebas de que los estrógenos son peligrosos, no sería ético o podría ser imposible realizar un ensayo clínico controlado sobre la relación entre los estrógenos y el cáncer de útero. En este caso, los investigadores podrían realizar un *experimento natural*, para respaldar la idea de que los estrógenos son una causa contribuyente del cáncer de útero. Este experimento natural se podría efectuar en el caso de que un grupo de mujeres dejase de tomar estrógenos como resultado de la publicidad generada por los estudios. Si la tasa de cáncer de útero del grupo de mujeres que dejan de tomar estrógenos disminuyera y no lo hiciera la de las mujeres que continúan tomándolos, este experimento aportaría la prueba más convincente disponible de que la modificación de la causa altera el efecto.

Los tipos básicos de estudios presentados en este libro no son los únicos que se encuentran en la literatura médica. Muchas veces se llevan a cabo estudios transversales. En estas investigaciones, la característica estudiada y el desenlace se miden en el mismo momento; en otras palabras, la asignación y la valoración se realizan en el mismo momento. Los estudios transversales son relativamente baratos y rápidos. Son útiles cuando se espera que es improbable que la exposición cambie con el tiempo o que el tiempo entre la exposición y el desarrollo de la enfermedad sea muy corto. Cuando se quiere estudiar la relación entre la tromboflebitis y la toma de píldoras anticonceptivas se puede usar un estudio transversal. Uno podría desear estudiar si es más probable que las mujeres con tromboflebitis estén tomando píldoras anticonceptivas en el momento en que aparece la tromboflebitis.

TAMAÑO DE LA MUESTRA

Una vez que se han valorado los objetivos y el tipo de estudio, el lector debe concentrarse en el tamaño de la muestra de individuos seleccionada. Además, ha de preguntarse si el número de pacientes incluidos en el estudio es suficiente para demostrar con una probabilidad razonable la existencia de una diferencia estadísticamente significativa entre las muestras del estudio y si dicha diferencia existe realmente en la población de la cual se han extraído las muestras.

Cuando nos preguntamos por la idoneidad del tamaño de la muestra, es preciso distinguir entre los estudios de casos y controles, por un lado, y los ensayos clínicos aleatorios y los estudios de cohortes, por el otro. Recuerde que en los estudios de casos y controles el desenlace es una característica del paciente, mientras que en los estudios de cohortes y en los ensayos clínicos aleatorios el desenlace es una

enfermedad. Por consiguiente, en el estudio de casos y controles sobre la tromboflebitis, estaríamos interesados en conocer la magnitud de la verdadera diferencia en la toma de píldoras anticonceptivas que, dado el tamaño de la muestra utilizado en el estudio, es probable demostrar como estadísticamente significativa. Al realizar un estudio de cohortes o un ensayo clínico aleatorio, uno está interesado en saber cuán pequeña debe ser una verdadera diferencia en la probabilidad de desarrollar una enfermedad como la tromboflebitis para que sea probable demostrar que es estadísticamente significativa, dado el tamaño de la muestra empleado en el estudio.

Las respuestas a estas cuestiones dependen de la magnitud de los errores de tipo I y de tipo II que el lector y el investigador estén dispuestos a tolerar. Recuerde que el error de tipo II es la probabilidad de no demostrar una diferencia estadísticamente significativa cuando realmente existe una diferencia en la población de la que se ha extraído la muestra del grupo de estudio y del de control.

El error de tipo I aceptado habitualmente es 5%. El error de tipo II aceptado está abierto a discusión. La mayoría de los investigadores desearían que la probabilidad de no demostrar una diferencia estadísticamente significativa cuando realmente existe una verdadera diferencia fuese 10% o menor. Si aceptamos un error de tipo I de 5% y uno de tipo II de 10% y utilizamos las tablas estadísticas estándar, se pueden extraer las siguientes conclusiones sobre el tamaño de la muestra.[3]

1. Si el grupo de estudio y el de control están formados por 100 individuos cada uno, el estudio tiene potencia estadística para detectar una diferencia estadísticamente significativa, si la frecuencia real de un desenlace, como la muerte en una población, es de 20% o más alta en una población y de 5% o menor en la otra.
2. Si tanto el grupo de estudio como el de control están formados por 250 individuos cada uno, la investigación tiene una potencia estadística para detectar una diferencia estadísticamente significativa, si la frecuencia real de un desenlace en una población es de 20% o más alta y en la otra población, de 10% o menor.
3. Si el grupo de estudio y el de control están formados por 500 individuos cada uno, la investigación tiene una potencia estadística para detectar una diferencia estadísticamente significativa, si la verdadera frecuencia de un desenlace en una población es de 10% o más alta y en la otra, de 5% o menor.
4. Cuando la frecuencia de ambos desenlaces es baja y la diferencia entre los porcentajes de los desenlaces es pequeña, se necesitan muestras grandes para detectar una diferencia significativa. Por ejemplo, para detectar que una verdadera diferencia entre dos poblaciones es estadísticamente significativa, cuando la frecuencia del desenlace en un grupo es de 2% y en el otro, de 1%, se necesitarían más de 3 500 individuos en cada grupo. Cuando utilice estas orientaciones, recuerde que, incluso con una potencia estadística alta, una muestra concreta extraída de poblaciones en las que existen verdaderas diferencias todavía puede ser insuficiente para detectar una diferencia estadísticamente significativa.

Estas estimaciones son útiles para el lector de la literatura médica, porque le permiten estimar si el estudio tiene una posibilidad real de demostrar una significación estadística a partir del tamaño de su muestra.

[3] Fleiss JL. *Statistical methods for rates and proportions*. 2a. ed. Nueva York: Wiley; 1981, pp. 260–280.

Ahora aplicaremos estos principios para demostrar por qué los estudios de casos y controles son útiles para estudiar enfermedades raras que afectan a un número relativamente bajo de individuos. No olvide que el término *desenlace* se refiere a una característica del paciente en los estudios de casos y controles y a la enfermedad misma en los de cohortes y en los ensayos clínicos aleatorios. El siguiente ejemplo hipotético muestra la dificultad de demostrar una diferencia estadísticamente significativa cuando se emplea un estudio de cohortes con el fin de investigar una enfermedad muy poco frecuente.

Los investigadores deseaban estudiar si la toma de píldoras anticonceptivas está asociada con la infrecuente aparición de accidentes vasculares cerebrales (AVC) en la mujer joven. Para ello, siguieron durante 10 años a 20 000 mujeres que tomaban píldoras anticonceptivas y a 20 000 que utilizaban otros métodos de planificación familiar. Después de gastar varios millones de dólares en el seguimiento, encontraron 2 casos de accidente vascular cerebral entre las usuarias de las píldoras y uno entre las no usuarias. La diferencia no fue estadísticamente significativa.

Cuando una enfermedad es muy rara, como los AVC en las mujeres jóvenes, muchas veces es preciso estudiar a un número muy alto de individuos para detectar una diferencia estadísticamente significativa, si se utiliza un estudio de cohortes. Suponga, por ejemplo, que la proporción de accidentes vasculares cerebrales en las mujeres jóvenes que no toman la píldora es 1 por 100 000 ó 0,001%. Supongamos también que la píldora aumenta 10 veces el riesgo de padecer la enfermedad, es decir, hasta 1 por 10 000 ó 0,01%. La diferencia en el desenlace es de 0,01% a 0,001% ó 0,009%. El uso de un estudio de cohortes para demostrar una diferencia estadísticamente significativa, existiendo una diferencia verdadera tan pequeña, puede requerir más de 100 000 mujeres en cada grupo.

Por otro lado, si se realiza un estudio de casos y controles en mujeres jóvenes con un AVC como grupo de estudio y mujeres jóvenes sin AVC como grupo de control, el desenlace que se medirá será la toma de píldoras anticonceptivas, en lugar de los AVC. La inclusión de 100 mujeres en cada grupo sería suficiente para detectar una diferencia estadísticamente significativa si existiera una diferencia real en la toma de píldoras anticonceptivas de 20% entre las que padecen AVC y de 5% entre el grupo sin AVC. En este ejemplo, es factible realizar un estudio de casos y controles sobre la relación entre las píldoras anticonceptivas y los AVC, utilizando solo una pequeña proporción de los individuos requeridos para estudiar la misma cuestión con un estudio de cohortes. Por lo tanto, aquel estudio de cohortes estaba condenado al fracaso desde el principio; un estudio de casos y controles habría sido mucho más apropiado. Siempre que en una investigación no se logre detectar una diferencia estadísticamente significativa, el lector se debe preguntar si el tamaño de la muestra del estudio era suficiente para detectarlas.

En el capítulo 11 exploraremos en mayor profundidad las implicaciones del tamaño de la muestra. La evaluación del diseño de un estudio exige que el lector valore sus objetivos, la idoneidad del tipo de estudio utilizado y la suficiencia del tamaño de la muestra. El lector capaz de comprender estos problemas básicos puede evaluar los resultados de un estudio de forma más inteligente.

CAPÍTULO 9

RESUMEN: EL ESTUDIO DE UN ESTUDIO

DISEÑO DEL ESTUDIO

Al analizar si un estudio se diseñó adecuadamente para responder a las preguntas planteadas, el revisor debe determinar, en primer lugar, si los objetivos del estudio se definieron con suficiente precisión y si la hipótesis se formuló de forma clara. A continuación, debe preguntarse si el tamaño de la muestra fue suficiente para reponder a la pregunta planteada en el estudio.

El lector de la literatura también debe decidir si el diseño empleado fue el apropiado para contestar a la cuestión planteada, teniendo en cuenta las ventajas y desventajas de cada tipo de estudio.

ASIGNACIÓN

Los investigadores intentan formar grupos de estudio y de control que sean semejantes en todas las características excepto en la estudiada. Los estudios de casos y controles y los de cohortes pueden contener un *sesgo de selección*. Este sesgo se produce cuando el grupo de estudio y el de control se escogen de tal forma que las frecuencias de un factor de riesgo o pronóstico que influya en el resultado de la investigación son distintas en ambos grupos. El sesgo de selección es un tipo especial de variable de confusión producida por diferencias aleatorias entre el grupo de estudio y el de control que están relacionadas con el desenlace estudiado. Cuando aparecen variables de confusión potenciales es importante identificarlas para poder incluirlas en el análisis.

VALORACIÓN

Para valorar el desenlace de un estudio, el lector debe considerar si se han cumplido los criterios de una valoración válida. Los investigadores deben demostrar que escogieron una medida adecuada del desenlace, aquella que mide lo que se propone medir. Deben haber realizado una valoración exacta; esto es, aquella medición que se aproxima a la medida verdadera del fenómeno. La medición de un desenlace debe ser completa. Por último, deben haber considerado si el proceso de la observación influyó en el desenlace valorado.

ANÁLISIS

El análisis implica el uso de métodos estadísticos para investigar el efecto al azar y el de los sesgos, así como para realizar estimaciones puntuales sobre los datos de la muestra. Es posible que el sesgo o el azar produzcan variables de confusión que pueden prevenirse al inicio del estudio apareando a los grupos de estudio y de control o bien apareando a los individuos de cada grupo. Las pruebas de significación estadística son métodos de contraste de hipótesis que sirven para valorar los efectos del azar en los resultados de una investigación. Estas pruebas suponen una hipótesis y conllevan errores de tipo I y de tipo II. Son un método de prueba por eliminación. En los estudios clínicos, la razón de productos cruzados y el riesgo relativo son las medidas

básicas de la fuerza de una asociación. Los intervalos de confianza de 95% están sustituyendo paulatinamente a las pruebas de significación o se dan como información adicional. Estos intervalos proporcionan el valor numérico observado o estimación puntual del valor de la población, así como el intervalo de valores que contiene el verdadero valor poblacional (parámetro) con un nivel de confianza de 95%. Las pruebas de significación estadística y los intervalos de confianza se calculan con los mismos métodos estadísticos. A veces, el lector puede usar rápidamente los intervalos de confianza para realizar una prueba de significación estadística.

INTERPRETACIÓN

Los autores de un estudio deben preguntarse qué significan los resultados para las personas incluidas en la investigación. Deben cuestionar también si la magnitud de las diferencias o la fuerza de la asociación es tal que los resultados son clínicamente útiles o importantes. Asimismo, han de plantearse si se han cumplido los criterios de una relación de causa-efecto.

Es preciso, además, que los autores y el lector apliquen el concepto clínico de causa contribuyente. La causa contribuyente requiere que la supuesta causa esté asociada con el efecto y lo preceda y, por añadidura, que la modificación de la causa altere el efecto. No se exige que la causa sea necesaria o suficiente para producir el efecto. Cuando no se consigue cumplir los criterios definidos, los criterios auxiliares, accesorios o de apoyo pueden ayudar a respaldar la observación de la existencia de una relación de causa-efecto. Estos criterios son la fuerza de la asociación, la consistencia, la plausibilidad biológica y la relación dosis-respuesta.

EXTRAPOLACIÓN

Finalmente, el lector debe preguntarse qué significan los resultados del estudio para los individuos no incluidos en el mismo. Al extrapolar los resultados a un individuo, es preciso que el lector distinga entre el riesgo relativo y el absoluto. Cuando se extrapola a nuevos grupos de sujetos con el factor de riesgo, el número de pacientes que es preciso tratar ofrece una medida de síntesis útil sobre el número de individuos que es necesario tratar para obtener un desenlace negativo menos o uno positivo más. También es preciso considerar el riesgo atribuible porcentual. Cuando se extrapola a poblaciones compuestas por individuos con y sin el factor de riesgo, se debe considerar el riesgo atribuible poblacional porcentual. Es importante reconocer el peligro que supone la extrapolación lineal más allá del intervalo de los datos observados. También hay que tener en cuenta cómo las distintas características de una nueva población objetivo pueden influir en la capacidad de extrapolar los resultados.

Pocas investigaciones pueden zafarse de estos errores; no obstante, su presencia no invalida automáticamente una investigación. Es responsabilidad del lector atento identificar estos errores y tenerlos en cuenta cuando se aplican los resultados del estudio.

PREGUNTAS ACERCA DEL ESTUDIO DE UN ESTUDIO

Ahora reuniremos el material precedente y veremos si usted puede aplicar lo que ha aprendido a varios artículos de investigación simulados. El método crí-

tico para evaluar un estudio de investigación se perfila en la siguiente lista de preguntas que uno debe formularse cuando está estudiando un estudio.

1. Diseño del estudio: ¿estaba diseñado adecuadamente?
   a. Los objetivos del estudio, ¿estaban definidos correctamente? Las hipótesis del estudio, ¿estaban formuladas con claridad?
   b. ¿Cuál era el tipo de estudio? ¿Era el adecuado para responder a las preguntas planteadas?
   c. ¿Cuál fue el tamaño de los grupos de estudio? ¿Era suficiente para contestar las preguntas del estudio?

2. Asignación: ¿se asignaron adecuadamente los pacientes al grupo de estudio y al de control?
   a. Si el estudio fue de casos y controles o de cohortes, ¿pudo existir un sesgo de selección?
   b. Si el estudio fue un ensayo clínico aleatorio, ¿se mantuvo la asignación al azar y a ciegas?
   c. Sin tener en cuenta el tipo de estudio, ¿los grupos de estudio y de control fueron comparables respecto a características distintas del factor estudiado o pudo haber influido en los resultados una variable de confusión?

3. Valoración: ¿se valoró el desenlace adecuadamente en los grupos de estudio y de control?
   a. La medida del desenlace, ¿era apropiada para los objetivos del estudio?
   b. La medida del desenlace, ¿fue exacta, reflejando entonces el verdadero valor del fenómeno?
   c. La medida del desenlace, ¿fue completa?
   d. ¿Afectó el proceso de observación al desenlace?

4. Análisis: ¿comparó correctamente el desenlace en los grupos de estudio y de control en el análisis?
   a. ¿Se ajustaron los resultados para tener en cuenta el efecto de posibles variables de confusión?
   b. La prueba de significación estadística, ¿fue realizada correctamente para valorar la probabilidad de que la diferencia o la asociación observada fuese debida al azar si la hipótesis nula fuera verdadera?
   c. ¿Se proporcionó la estimación puntual del valor de la población (parámetro) y su intervalo de confianza de 95%?
   d. ¿Se consideró el número de hipótesis formuladas? Usando el enfoque bayesiano, ¿se asignó a cada hipótesis la probabilidad previa antes de empezar el estudio para poder calcular la probabilidad de la hipótesis después de obtener los datos?
   e. ¿Podría el error de tipo I o el de tipo II explicar los resultados?

5. Interpretación: ¿se llegó a conclusiones válidas sobre el significado de la investigación para los sujetos incluidos en el estudio?
   a. ¿Es la magnitud de la diferencia o de la fuerza de la asociación lo suficientemente grande como para ser clínicamente importante o útil?
   b. ¿Se cumplieron los tres criterios de causa contribuyente?
   c. ¿Los investigadores distinguieron entre causa contribuyente y causa necesaria y suficiente?

d. Si no se cumplieron los tres criterios de causa contribuyente, ¿se cumplieron los criterios auxiliares?

6. Extrapolación: ¿se realizaron correctamente las extrapolaciones a los individuos y situaciones no incluidos en el estudio?
   a. ¿Consideraron los investigadores tanto el riesgo relativo como el absoluto al extrapolar los resultados a los individuos?
   b. Cuando se extrapoló a nuevos grupos con el factor de riesgo, ¿los investigadores tomaron en consideración el riesgo atribuible porcentual?
   c. Cuando se extrapoló a nuevos grupos formados por individuos con y sin el factor de riesgo, ¿los autores tuvieron en cuenta el riesgo atribuible poblacional porcentual?
   d. ¿Los autores realizaron una extrapolación más allá del intervalo de los datos?
   e. ¿Los autores consideraron las diferencias entre el grupo de estudio y la población objetivo?

CAPÍTULO 10

EJERCICIOS PARA DETECTAR ERRORES: ESTUDIOS OBSERVACIONALES

Los siguientes estudios hipotéticos incluyen errores del tipo ejemplificado en cada uno de los componentes del marco básico. Estos ejercicios para detectar errores se han diseñado con el fin de comprobar su capacidad para aplicar el marco básico al examen crítico de un estudio. Se presentan ejemplos de estudios de casos y controles y de cohortes. Por favor, lea los ejercicios y escriba una crítica de cada estudio. Al final de cada ejercicio encontrará una crítica en la que se señalan los errores más importantes.

Observe que el último ejercicio es el mismo que leyó en el primer capítulo. Compare su crítica actual de este ejercicio con la que escribió previamente para ver el progreso realizado.

EJERCICIO No. 1: ESTUDIO DE CASOS Y CONTROLES

Se llevó a cabo un estudio de casos y controles para estudiar los factores asociados con el desarrollo en el feto de enfermedades cardíacas congénitas (ECC). El grupo de estudio estaba formado por 200 mujeres que habían tenido abortos espontáneos durante el primer trimestre, en los que se detectaron malformaciones cardíacas congénitas. El grupo control estaba compuesto por 200 mujeres con abortos inducidos en el primer trimestre y en los que no se hallaron esas malformaciones.

Se intentó entrevistar a todas las mujeres durante el primer mes posterior al aborto, para determinar qué factores del embarazo podrían estar asociados con una ECC. Se estudiaron 100 variables. Los encuestadores consiguieron que participaran 120 de las 200 mujeres del grupo de estudio y 80 de las 200 del grupo de control. El resto de mujeres rehusaron participar.

Los investigadores encontraron las siguientes diferencias entre las mujeres cuyos fetos tenían ECC y aquellas cuyos fetos no la tenían.

1. La *ventaja* de tomar medicamentos contra la náusea de las mujeres con fetos que presentaban ECC fue tres veces más elevada que la de las mujeres con fetos sin ECC. Esta diferencia fue estadísticamente significativa.

2. No se observaron diferencias en el uso de tranquilizantes entre el grupo de estudio y el de control.

3. La media de la edad de las mujeres cuyos fetos presentaron ECC fue de 23 años y la de las mujeres del grupo control, 18. Los resultados fueron estadísticamente significativos.

4. Las mujeres del grupo de estudio bebían una media de 3,7 tazas de café diarias, mientras que las mujeres con fetos sin ECC bebían una media de 3,5 tazas. Esa diferencia también fue estadísticamente significativa.

5. Entre las 96 variables restantes, los autores observaron que la *ventaja* de tener el pelo rubio y de medir más de 167 centímetros era el doble en las mujeres que dieron a

luz fetos con ECC. Ambas diferencias fueron estadísticamente significativas empleando los métodos estadísticos habituales.

Los autores llegaron a las siguientes conclusiones.

1. La medicación contra la náusea causa ECC, porque las mujeres que dan a luz fetos con ECC la toman con frecuencia.
2. Los tranquilizantes se pueden usar con seguridad en las mujeres embarazadas, ya que no están asociadas con un aumento del riesgo de ECC.
3. Dado que es más probable que las mujeres de 20 años de edad tengan fetos con ECC, se debe animar a las mujeres a que tengan sus hijos antes de los 20 años.
4. Como el café aumenta el riesgo de ECC, su consumo se debe eliminar completamente durante el embarazo, lo cual eliminaría en gran parte el riesgo de ECC.
5. A pesar de que no se haya formulado la hipótesis de que el pelo rubio y la talla pueden ser factores de riesgo de ECC, se ha demostrado que pueden ser factores predictivos importantes de la enfermedad.

CRÍTICA: EJERCICIO No. 1

Diseño del estudio

Los investigadores no formularon claramente los objetivos de su estudio. ¿Estaban interesados en un tipo específico de ECC? Las enfermedades congénitas del corazón son una serie de enfermedades que afectan a las válvulas, el septo y los vasos sanguíneos. Al reunir todas las enfermedades bajo el encabezamiento de ECC, estaban suponiendo la existencia de una etiología común para todas ellas. Además, no quedó clara la hipótesis concreta que se pretendía contrastar en el estudio. Los grupos escogidos consistían en uno de estudio, cuyas integrantes tuvieron un aborto espontáneo, y uno de control, a cuyas participantes se les indujo un aborto con su consentimiento. Es de esperar que estos grupos difirieran en varios aspectos. Hubiera sido preferible escoger grupos más comparables de mujeres, por ejemplo, aquellas que habían tenido un aborto inducido con y sin ECC o aquellas que habían tenido un aborto espontáneo con y sin ECC.

Con este diseño de estudio debe recordarse que solo se podían estudiar las ECC que eran suficientemente graves como para producir un aborto espontáneo. Aunque ello puede proporcionar información importante, los factores que causan ECC suficientemente grave como para producir un aborto pueden ser distintos de los que causan ECC en los recién nacidos a término.

Asignación

Para determinar si existió un sesgo de selección, primero debemos preguntarnos si el grupo de estudio y el de control difirieron en algunos aspectos. Segundo, si estas diferencias influyeron en los resultados. Es probable que las experiencias de las mujeres que padecieron un aborto espontáneo difirieran en múltiples aspectos de las que lo tuvieron por inducción. Es probable también que las actitudes de las mujeres acerca de sus embarazos fueran distintas y que estas pudieran influir en la toma de medicamentos durante el embarazo. Tales diferencias entre el grupo de estudio y el de control pudieron influir en el resultado. Por consiguiente, en este estudio se pudo haber introducido un sesgo de selección.

Valoración

La elevada tasa de pérdidas en el seguimiento de las participantes sugiere la posibilidad de que las mujeres a las que no se pudo seguir tuvieran características diferentes. Una tasa elevada de pérdidas en el seguimiento debilita las conclusiones que es posible extraer a partir de cualquier diferencia observada. Una posibilidad es el sesgo de recuerdo por parte de las participantes, especialmente cuando estas mujeres experimentaron una experiencia traumática, como el de tener un aborto con ECC, y se les solicitó varias veces que recordaran sucesos subjetivamente como el uso de medicamentos o el consumo de café. La notificación retrospectiva de la toma de fármacos, por ejemplo, puede estar influida por las emociones que provoca la pérdida del feto en las mujeres que experimentan un aborto espontáneo inesperado. El resultado puede ser un escrutinio más detallado de la memoria que conduzca a un recuerdo más preciso del uso de medicamentos.

Análisis, interpretación y extrapolación

1. Aunque se supusiera que la inferencia sobre la relación entre la medicación contra la náusea y la ECC es correcta, con ello no se demostraría la existencia de una relación de causa-efecto. Con los estudios de casos y controles no se pueden afirmar definitivamente qué factor es la "causa" y cuál es el "efecto". Es posible que las mujeres que dieron a luz fetos con ECC tuvieran más náuseas y, por lo tanto, que tomaran más antieméticos. Antes de establecer una relación causal en el sentido clínico de causalidad (causa contribuyente), los investigadores deben demostrar que la causa postulada precede al efecto y que su modificación lo modifica. Los autores de este estudio hicieron una interpretación que no está necesariamente justificada por los datos. El ajuste de los resultados de los grupos de estudio y de control según la diferencia en la frecuencia de náusea como parte del análisis sería un método para evaluar más a fondo la posible relación entre los antieméticos y la ECC, aunque todavía se usara un estudio de casos y controles.

2. La ausencia de una diferencia entre los grupos en términos del uso de tranquilizantes no garantiza necesariamente la seguridad de estos medicamentos. Las muestras pueden ser demasiado pequeñas para examinar completamente el riesgo de que los tranquilizantes causen ECC. Un pequeño aumento del riesgo requiere muchos más sujetos de estudio antes de que la investigación tenga la potencia estadística suficiente y con un alto grado de certeza para demostrar una diferencia entre los grupos. Por esta razón, se pudo haber cometido un error de tipo II. Incluso aunque no existiera el riesgo de que los tranquilizantes causen ECC, no tendríamos garantías de que estos medicamentos produzcan otros efectos adversos sobre el feto que los conviertan en inseguros durante el embarazo. Por consiguiente, los investigadores extrapolaron los resultados bastante más allá de sus datos.

3. La diferencia de edad entre los dos grupos de mujeres puede estar relacionada tanto con el tipo de aborto como con la presencia o ausencia de ECC. Por ello, la edad puede constituir un sesgo de selección, si es más probable que las mujeres tengan un aborto inducido en la adolescencia que en edades más avanzadas; esta relación puede explicar, por sí sola, las diferencias de edad observadas entre ambos grupos. Aunque el estudio haya mostrado que el riesgo de ECC era menor para las embarazadas adolescentes, los riesgos médicos y sociales pueden superar este beneficio. La mera presencia de una diferencia estadísticamente significativa no significa que se haya alcanzado una conclusión clínicamente importante.

4. La diferencia en el consumo de café fue estadísticamente significativa, aunque no muy grande. Un resultado estadísticamente significativo es aquel cuya probabilidad de observarse por azar es baja, si no existen verdaderas diferencias en la población de la que se han extraído los datos de la muestra. Sin embargo, es clínicamente improbable que una reducción tan pequeña tenga un efecto notable sobre el riesgo de ECC. La significación estadística debe distinguirse de la importancia clínica y de la causa contribuyente. Beber café puede tener un efecto, pero con diferencias tan pequeñas como las observadas, uno debe tener cuidado en extraer demasiadas conclusiones.

5. No es sorprendente que, al analizar 100 variables, se encuentren algunas asociaciones estadísticamente significativas. Cuando se utilizan muchas variables, uno no puede aplicar el nivel de significación estadística habitual para rechazar la hipótesis nula de no asociación. El nivel de 5% habitual supone que se ha formulado una hipótesis antes de iniciar el estudio. Por ello, los autores no pueden concluir con seguridad que la estatura y el color del pelo sean factores de riesgo de la ECC.

Con un enfoque bayesiano se puede decir que la probabilidad de que el color del cabello y la estatura estén asociados con la ECC es menor. Por lo tanto, el observar una asociación puede representar un error de tipo I, dado que la probabilidad de detectar una asociación después de obtener los resultados del estudio todavía es relativamente baja.

EJERCICIO No. 2: ESTUDIO DE COHORTES

Con el fin de estudiar los efectos de una unidad coronaria (UC) que funcione adecuadamente sobre los infartos de miocardio (IM), varios investigadores realizaron un estudio de cohortes concurrentes de los efectos de una nueva UC.

Durante el primer año de funcionamiento de la UC, se ingresaron 100 pacientes del grupo de estudio remitidos por sus médicos con el diagnóstico de sospecha de IM. En salas hospitalarias fuera de la UCC, se ingresaron 100 pacientes del grupo de control en los que se había descartado el diagnóstico de IM.

A los pacientes de la UC se les administró lidocaína si sus enzimas cardíacas eran positivas para IM a las 24 horas de su ingreso. Además, se les administraron tratamientos invasivos para valorar y tratar las oclusiones de sus arterias coronarias. Las complicaciones de los pacientes de sala se monitorearon y se trataron cuando se presentaron.

Al comparar a los pacientes de sala con los de la UC, los investigadores observaron que la media de la edad de los de la UC era de 58 años y la de los controles, 68. Una cuarta parte de los pacientes de la UC y un veinteavo de los de sala desarrollaron hipotensión. El 80% de los pacientes de la UC y 20% de los de sala presentaron arritmias ventriculares. Los investigadores siguieron la evolución de los pacientes durante su hospitalización y el año posterior, y recogieron los siguientes datos sobre los desenlaces.

1. En 36% de los pacientes de la UC y en 30 de los de sala se encontraron finalmente pruebas enzimáticas o electrocardiográficas definitivas de IM.
2. Ocho pacientes de la UC y cuatro de los de sala fallecieron en el hospital. Estas diferencias no fueron estadísticamente significativas.
3. Los pacientes de la UC permanecieron hospitalizados durante una media de 12 días y los de la sala, durante 15 días. Las diferencias fueron estadísticamente significativas.

4. Ninguno de los pacientes que recibieron lidocaína en la UC falleció.
5. Un año después del alta médica, los antiguos pacientes de la UC eran capaces de hacer, en promedio, 20% más ejercicio que los de sala.

Los autores llegaron a las siguientes conclusiones:

1. La atención de la UC aumenta la tasa de desarrollo de IM entre los pacientes ingresados en el hospital con dolor de pecho.
2. Dado que las diferencias entre las tasas de mortalidad no fueron estadísticamente significativas, las tasas de mortalidad fueron idénticas en ambos grupos.
3. Como las diferencias en la duración de la hospitalización fueron estadísticamente significativas, los investigadores concluyeron que, mediante la creación de la UC, habían demostrado un importante ahorro en los costos.
4. Habida cuenta de que la lidocaína previno todas las muertes, si se empleaba después del diagnóstico definitivo de IM, el uso de este medicamento en el momento del ingreso eliminaría toda la mortalidad debida al IM.
5. Ya que los pacientes de la UC toleraron mejor el ejercicio durante el año posterior al alta, la UC causa una mejora de la supervivencia a largo plazo.

CRÍTICA: EJERCICIO No. 2

Diseño del estudio

Los investigadores intentaban estudiar el efecto de una UC que funcionara bien y, para ello, decidieron realizar su estudio en una UC nueva. Sin embargo, las nuevas instalaciones no pueden operar a pleno rendimiento en su primer año. De este modo, los investigadores no seleccionaron las mejores condiciones para estudiar los efectos de una UC que funcionara bien. Además, no formularon sus hipótesis específicas antes de iniciar el estudio.

Asignación

El sesgo de selección puede estar presente en este estudio, si los individuos con mal pronóstico fueron ingresados en la UC por sus médicos. Este factor puede ser importante, si los médicos ingresaron selectivamente en la UC a los pacientes más enfermos. En este caso, sería de esperar que el sesgo de selección influyese en los resultados.

Valoración

Los investigadores encontraron una tasa más elevada de arritmias entre los pacientes de la UC, lo que podría ser el resultado del método empleado para valorarlas en dicha unidad. Si los pacientes de la UC fueron monitoreados continuamente —al contrario que los de la sala—, es posible que dada la intensidad con que fueron observados, se descubriera un porcentaje más alto de las arritmias desarrolladas.

Análisis

Los investigadores también observaron que la media de la edad de los pacientes de la UC era menor que la de los de sala. Este factor pudo haber aparecido por azar o ser resultado del deseo del médico de ofrecer una asistencia más intensiva a los pacientes jóvenes. Es probable que la edad de los pacientes se asocie a desenlaces

tales como la tolerancia al ejercicio después del IM, porque los hombres jóvenes toleran mejor el ejercicio. Este hecho podría explicar la diferencia observada, sin tener en cuenta si la diferencia ocurrió a causa de un sesgo o por azar. Las diferencias en la edad son una variable de confusión potencial que debe ser motivo de ajuste en el análisis.

Interpretación y extrapolación

1. Los investigadores encontraron que un porcentaje menor de pacientes de la sala presentaron finalmente pruebas de IM y concluyeron que la tasa más elevada de IM en la UC había sido causada por la atención médica prestada en ella. El primer requisito para establecer una relación de causa-efecto es que la causa preceda al efecto. En esta situación, es probable que los pacientes ya hubiesen padecido o estuviesen padeciendo el IM cuando ingresaron en el hospital. Por eso, en muchos casos, el efecto (IM) puede haber precedido a la causa (ingreso en la UC). Hay pocas pruebas que apoyen la interpretación de que la UC esté asociada con una tasa de IM más elevada.

2. Los autores concluyeron que las tasas de mortalidad se debían considerar similares, porque no se observó una diferencia estadísticamente significativa entre dichas tasas. No demostrar una diferencia estadísticamente significativa no implica que esta diferencia no exista. Cuando el número de individuos incluidos en una muestra es muy bajo, se necesita una diferencia muy grande para demostrar que es estadísticamente significativa. Los autores no pensaron en la posibilidad de cometer un error de tipo II. Es posible que los pacientes más graves fueran ingresados en la UC y que, por este motivo, se esperara una tasa de mortalidad más alta entre ellos. Cuando el número de individuos estudiados es tan bajo, como en este estudio, es preferible presentar los resultados sin aplicar ninguna prueba de significación estadística. En este estudio existía una diferencia; aunque no fue estadísticamente significativa, el número de muertes observado no puede considerarse idéntico.

3. La estancia hospitalaria de los pacientes de la UC fue menos prolongada que la de los de la sala. Los resultados fueron estadísticamente significativos, lo que indica que no es probable que las diferencias fueran debidas al azar. El que estas diferencias sean importantes desde el punto de vista de los costos es otra cuestión, porque los costos adicionales de la atención en la UC pueden superar las pequeñas diferencias en la duración de la estancia hospitalaria. Esta consideración puede ilustrar la distinción que debe hacerse entre una diferencia estadísticamente significativa y una clínicamente importante.

4. La lidocaína se administró a los pacientes de la UC solamente después de establecer un diagnóstico enzimático definitivo, y el IM estaba presente desde hacía 24 horas. En ese momento, el riesgo de morir de IM había descendido en gran parte, especialmente el riesgo de arritmia. Es probable, por consiguiente, que la administración de lidocaína tuviera poco que ver con el hecho de que no se produjeran muertes entre aquellos a los que se administró. Los autores fueron más allá de los datos al extrapolar sus resultados a todos los pacientes de la UC y al no observar que el grupo de pacientes que recibieron lidocaína era diferente del de los pacientes recién ingresados en la UC.

5. Los autores concluyeron que la UC había mejorado la perspectiva de recuperación, dado que los pacientes ingresados en ella toleraban mejor el ejercicio un año más tarde. Existen pocas pruebas para establecer una relación de causa-efecto. Como los pacientes de la UC eran más jóvenes que los de la sala, era previsible que toleraran mejor el ejercicio. Además, en la UC murieron más pacientes. La capacidad de sobrevivir a un IM pudo haber seleccionado a un grupo de pacientes con una mayor tolerancia al ejer-

cicio. Finalmente, no se aportaron pruebas para respaldar que la tolerancia al ejercicio un año después del IM estuviese realmente asociada con una supervivencia a largo plazo. Los autores debieron haber tenido más cuidado al relacionar un mejor pronóstico con la atención prestada en la UC; al hacerlo, extrapolaron bastante más allá de los datos que observaron.

EJERCICIO DE REPASO: ESTUDIO DEL TAMIZAJE MÉDICO EN UNA POBLACIÓN MILITAR

En el primer año del servicio militar se ofreció la posibilidad a 10 000 soldados de 18 años de edad de participar en un examen médico de salud anual constituido por una historia clínica, un examen físico y diversas pruebas de laboratorio. El primer año participaron 5 000 y los 5 000 restantes no lo hicieron. Los 5 000 que participaron fueron escogidos como grupo de estudio y los 5 000 que no participaron formaron el grupo de control. A los que participaron el primer año se les practicaron exámenes médicos anuales durante su servicio militar.

Al finalizar el servicio, tanto a los 5 000 del grupo de estudio como a los 5 000 del grupo control se les hizo una amplia evaluación de su historia clínica y se les practicó un examen físico y una evaluación de laboratorio para determinar si las visitas anuales habían producido alguna diferencia en su salud y en sus estilos de vida.

Los investigadores obtuvieron la siguiente información:

1. Según el consumo de alcohol declarado, la tasa de alcoholismo de los no participantes fue dos veces más alta que la de los participantes.
2. Se establecieron el doble de diagnósticos en los participantes que en los no participantes.
3. Los participantes tuvieron un promedio de ascensos dos veces más alto que los no participantes.
4. No se observaron diferencias estadísticamente significativas entre las tasas de infarto de miocardio (IM) de ambos grupos.
5. No se encontraron diferencias entre los grupos respecto a la tasa de aparición de cáncer de testículo o de enfermedad de Hodgkin, que son los dos tipos de cáncer más frecuentes en la gente joven.

Después, los autores extrajeron las siguientes conclusiones:

1. El tamizaje anual puede reducir a la mitad la tasa de alcoholismo en la población en el servicio militar.
2. Dado que el número de enfermedades diagnosticadas en los participantes fue el doble que en los no participantes, sus enfermedades se estaban diagnosticando en un estadio temprano del proceso patológico, momento en el cual el tratamiento es más beneficioso.
3. Como a los participantes se les concedieron el doble de ascensos que a los no participantes, el programa de tamizaje tuvo que haber contribuido a la calidad de su trabajo.
4. Habida cuenta de que no se observaron diferencias entre las tasas de IM de ambos grupos, el tamizaje y la intervención sobre los factores de riesgo de la enfermedad coronaria no se deben incluir en un futuro programa de tamizaje sanitario.
5. Dado que la frecuencia de la enfermedad de Hodgkin y del cáncer de testículo fue igual en ambos grupos, los futuros exámenes de salud no deben incluir esfuerzos para diagnosticar estas enfermedades.

CRÍTICA: EJERCICIO DE REPASO

Diseño del estudio

Los investigadores solo establecieron como objetivo general el estudio del valor anual de los exámenes de salud. No definieron la población a la que deseaban aplicar sus resultados, no formularon una hipótesis específica, ni identificaron claramente las preguntas específicas de su estudio.

Si el objetivo de los investigadores era estudiar los efectos de los exámenes anuales de salud, no cumplieron su objetivo, ya que no disponían de pruebas de que los participantes del primer año tomaran realmente parte en los exámenes siguientes.

Por añadidura, la elección de los participantes pudo no ser la más apropiada para responder a la pregunta formulada en la investigación. En el estudio se seleccionaron personas jóvenes que ya habían sido sometidas a un tamizaje de enfermedades crónicas en virtud de haber superado las pruebas físicas para entrar en el servicio militar. Siendo un grupo joven y sano, es posible que el grupo estudiado no fuera la población adecuada para probar la utilidad de un examen de salud en poblaciones de más edad o en alto riesgo, como las de militares de más edad, en las que es de esperar que la frecuencia de enfermedades sea más alta.

Asignación

Los individuos de este estudio se seleccionaron a sí mismos; ellos decidieron si participaban o no. Por lo tanto, los participantes pueden considerarse voluntarios. Los investigadores no presentaron ninguna prueba indicativa de que los que decidieron participar se diferenciasen en algún aspecto de los que no participaron. Es probable que aquellos tuvieran hábitos de salud distintos de los no participantes. Estas diferencias pudieron haber contribuido a las del desenlace. Dado que no se dispone de una evaluación inicial del grupo de control, no se sabe si sus integrantes eran diferentes y en qué forma de los del grupo de estudio. Por este motivo desconocemos si el grupo de control y el de estudio eran comparables.

Los individuos de ambos grupos se asignaron a sí mismos a partir de su participación en los exámenes de salud del primer año. Como esos exámenes se llevaron a cabo anualmente, los que habían participado al principio se pudieron haber retirado del estudio. Por ello, el estado de los individuos del grupo de estudio y del de control puede no reflejar de forma válida su participación real en el tamizaje.

Valoración

La valoración del desenlace se realizó solo en aquellos que fueron licenciados del servicio militar. No se incluyó a los que permanecieron en el ejército. Los individuos que murieron durante el servicio militar no se hubieran incluido en la valoración de los que dejaron el servicio. Estos pudieron haber sido los sujetos más adecuados para valorar los beneficios potenciales conseguidos con el tamizaje.

Los que participaron en los múltiples exámenes de salud estaban sometidos a una observación más intensa que los que no lo hicieron. Esta diferente intensidad podría explicar que en ellos se establecieran más diagnóticos durante su servicio militar. Si bien los no participantes podían haber tenido el mismo número de enfermedades, no todas resultaron diagnosticadas.

Análisis, interpretación y extrapolación

1. La tasa de alcoholismo de los participantes fue más baja que la de los no participantes, quizá a causa de las diferencias entre los grupos antes de su entrada en el estudio. Si fuese menos probable que los grandes bebedores participaran en el tamizaje, este hecho solo hubiera modificado la frecuencia de alcoholismo. En el análisis no se incluyeron datos comparativos de los participantes a su entrada en el estudio, ni se ajustaron según las diferencias. Además, la validez del método utilizado para valorar el consumo de alcohol es cuestionable. Como no existe un criterio uniforme de diagnóstico, es posible que existieran diferencias en el recuerdo y en la declaración. Aunque no se hubiera cometido ninguno de estos errores potenciales, no hay pruebas en el estudio de que el tamizaje por sí mismo fuera el factor causante de una tasa de alcoholismo más baja. La extrapolación a los militares en general excedió el intervalo de los datos.

2. Si el nivel de motivación más elevado estuviera asociado con la participación en el estudio y con los ascensos en el ejército, la motivación sería una variable de confusión, al estar relacionada con la participación y con el desenlace. Sin ajustar los datos según esta variable de confusión potencial, no se puede llegar a ninguna conclusión sobre la relación entre la participación y los ascensos.

3. Muchos de los que tuvieron IM podrían haber muerto y, de ese modo, estar excluidos de la valoración. Además, uno esperaría que la tasa de IM fuera más baja en una población joven. Incluso con el elevado número de participantes estudiados, el tamaño de la muestra pudo haber sido insuficiente para detectar diferencias estadísticamente significativas para diferencias reales pero pequeñas entre los grupos. Si se supone que las modificaciones de los factores de riesgo del IM también alteran el pronóstico, en este estudio no existe una indicación de que los que participaron tuvieran más factores de riesgo identificados o bien más factores de riesgo alterados. Es posible que los efectos de las posibles modificaciones de los factores de riesgo no se hagan aparentes hasta años después de que los participantes hayan abandonado el servicio militar. Por lo tanto, con este estudio no se puede determinar si el tamizaje de los factores de riesgo coronarios modifica el pronóstico de la enfermedad.

4. La ausencia de diferencias entre las tasas de aparición de cáncer de testículo y de la enfermedad de Hodgkin no se puede valorar a partir de los que dejaron el servicio con vida. Aunque esas tasas fueran idénticas, dicen poco sobre el fracaso o el éxito del programa de tamizaje. Un programa de tamizaje del cáncer pretende detectar la enfermedad en un estadio temprano, pero no intenta prevenirla. Por eso, la tasa de aparición de cáncer no se puede utilizar para evaluar el éxito de un programa de tamizaje. Uno esperaría encontrar tasas idénticas de desarrollo de ambos tipos de cáncer. El estadio de la enfermedad en el momento del diagnóstico y el pronóstico de los que desarrollaron una de las dos enfermedades serían medidas más apropiadas para valorar el éxito de dicho programa. Como no se presentan estos datos, no es posible hacer interpretaciones.

Una vez criticados estos ejercicios de detección de errores, el lector puede sentirse desanimado, pero sepa que la mayor parte de los estudios de investigación contienen bastantes menos errores que los ejercicios que acabamos de presentar. Sin embargo, puede ser de ayuda para el lector recordar que algunos errores son inevitables y que su detección no es sinónimo de invalidez de la investigación.

La práctica de la medicina clínica exige que los clínicos actúen sobre la base de probabilidades y la lectura crítica de la literatura médica les ayuda a de-

finir con más exactitud esas probabilidades. El arte de la lectura de la literatura médica consiste en la capacidad de extraer conclusiones útiles a partir de datos inciertos. Aprender a detectar errores no solo ayuda al clínico a identificar las limitaciones de un estudio concreto, sino también a moderar la tendencia natural a poner en práctica automáticamente los resultados más recientes de la investigación.

CAPÍTULO 11

ESTUDIOS DE INTERVENCIÓN: ENSAYOS CLÍNICOS CONTROLADOS

Los ensayos clínicos controlados se han convertido paulatinamente en el criterio de referencia (*gold standard*) mediante el cual juzgamos los beneficios de un tratamiento. La Administración de Alimentos y Medicamentos de los Estados Unidos de América (Food and Drug Administration, FDA) exige su realización para aprobar la comercialización de los fármacos, los Institutos Nacionales de Salud (National Institutes of Health, NIH) los premian con becas, las revistas los promueven mediante su publicación y, cada vez más, los médicos los leen en busca de certeza. Los ensayos clínicos controlados se han transformado en una fase estándar de la investigación clínica cuando son viables y éticos. Por eso, es de fundamental importancia reconocer lo que estos estudios nos dicen, los errores que se pueden cometer al realizarlos y las cuestiones que no se pueden resolver con ellos. Para cumplir con estos objetivos, emplearemos el marco uniforme de los estudios clínicos y comentaremos los elementos del diseño del estudio, de la asignación, de la valoración, del análisis, de la interpretación y de la extrapolación en relación con los ensayos clínicos controlados.

MARCO UNIFORME EN LOS ENSAYOS CLÍNICOS

Diseño del estudio

Los ensayos clínicos controlados son capaces de demostrar los tres criterios de causa contribuyente. Cuando se aplican a un tratamiento, se emplea el término *eficacia* en lugar del de *causa contribuyente*.[1] Por *eficacia* se quiere indicar que el tratamiento reduce en el grupo de estudio la probabilidad o el riesgo de experimentar un desenlace adverso. No obstante, es preciso distinguir *eficacia* de *efectividad*. La *efectividad* implica que el tratamiento funciona en las condiciones normales de la práctica clínica, en contraposición a las condiciones de una investigación. Habitualmente, nuestro objetivo es utilizar los ensayos clínicos controlados para determinar si un tratamiento funciona de acuerdo con una dosis dada, a través de una vía de administración y para un tipo de paciente concreto.[2]

Los ensayos clínicos controlados no están indicados en las investigaciones iniciales de un nuevo tratamiento. Cuando se utilizan como parte del proceso de aprobación de un nuevo fármaco, se conocen como *ensayos de fase III*. De acuerdo con la definición de la FDA, los ensayos de *fase I* hacen referencia a los esfuerzos iniciales para administrar el tratamiento a seres humanos. Su finalidad es establecer la dosifi-

[1] Una técnica que elimina la causa contribuyente es eficaz por definición. No obstante, la eliminación de una causa contribuyente indirecta también puede ser eficaz, incluso después de que el estado de los conocimientos nos haya permitido definir una causa contribuyente más directa.

[2] Es posible realizar un ensayo clínico controlado para valorar la efectividad de un tratamiento mediante el empleo de una muestra representativa de los tipos de pacientes que se han de tratar con él y los métodos habituales que se usarán en la práctica clínica.

cación y evaluar sus posibles efectos tóxicos. Estos estudios solo proporcionan una visión preliminar de la eficacia del fármaco. Los *ensayos de fase II* están destinados a establecer las indicaciones y el régimen de administración del nuevo tratamiento, y a determinar si está justificado realizar más estudios. Estos estudios son generalmente ensayos de pequeña escala, controlados o no, que permiten juzgar si se debe realizar un estudio controlado a gran escala.

Idealmente, un ensayo clínico controlado o de fase III debe realizarse después de haber establecido las indicaciones y el régimen de administración, pero antes de que el tratamiento haya pasado a formar parte de la práctica clínica. Este proceso es automático para los nuevos fármacos que todavía no están comercializados. Sin embargo, para muchos procedimientos terapeúticos y fármacos que ya están comercializados, el tratamiento puede haberse empleado extensamente antes de que se realicen ensayos clínicos controlados. Esto constituye un problema, porque al haberse empleado, los médicos y, frecuentemente, los propios pacientes, ya tienen ideas firmes sobre su valor. Cuando esto sucede, los médicos o los pacientes pueden pensar que no es ético participar en un ensayo experimental o continuar su participación, si descubren que el paciente ha sido asignado al grupo de control.

Una vez decidido que ha llegado el momento de realizar un ensayo clínico controlado, la siguiente pregunta referente al diseño es: ¿es viable el estudio? Para entender lo que es viable, se debe definir la cuestión que se quiere estudiar con un ensayo clínico controlado.

La mayor parte de los estudios clínicos controlados tienen como objetivo determinar si el nuevo tratamiento produce un resultado mejor que el placebo o el tratamiento estándar. Para decidir si el ensayo es viable, es preciso estimar el tamaño de la muestra necesaria. En otras palabras, los investigadores deben averiguar cuántos pacientes es necesario estudiar para tener una probabilidad razonable de demostrar una diferencia estadísticamente significativa entre el nuevo tratamiento y el placebo o el tratamiento estándar.

El tamaño necesario de la muestra depende de los siguientes factores:[3]

1. La magnitud del error de tipo I tolerado por los investigadores. Esta es la probabilidad de demostrar una diferencia estadísticamente significativa en las muestras cuando no existe una verdadera diferencia entre los tratamientos en las poblaciones. El nivel alfa correspondiente al error de tipo I se sitúa habitualmente en el 5%.

2. La magnitud del error de tipo II tolerado por los investigadores. Esta es la probabilidad de no detectar una diferencia estadísticamente significativa en las muestras cuando realmente existe una verdadera diferencia de una determinada magnitud entre los tratamientos. Muchos investigadores tienen como objetivo un error de tipo II no superior a 20%. Un error de 20% también se denomina potencia estadística de 80%. La potencia de 80% implica que existe una probabilidad de 80% de demostrar una diferencia estadísticamente significativa, cuando existe realmente.

3. El porcentaje de individuos en el grupo de control que se espera que experimentarán el desenlace adverso estudiado (muerte o desarrollo de la enfermedad). Con frecuencia, esta cifra puede estimarse a partir de estudios anteriores.

[3] Esta es toda la información necesaria para una variable con dos posibles resultados. Cuando se calcula el tamaño muestral para variables con múltiples resultados posibles, se ha de estimar también la desviación estándar de la variable.

4. La mejora en el desenlace entre los miembros del grupo de estudio que se pretende demostrar como estadísticamente significativa. A pesar del deseo de demostrar una diferencia estadísticamente significativa, incluso para cambios reales pequeños, es necesario que los investigadores decidan la magnitud de la diferencia que sería considerada como clínicamente importante. Cuanto menor sea la diferencia que se pretende observar entre el grupo de control y el de estudio, mayor será el tamaño de la muestra requerido.

Echemos una ojeada a la forma en que estos factores influyen en el tamaño necesario de la muestra. El cuadro 11-1 ofrece unas orientaciones generales sobre el tamaño de la muestra necesario para diferentes niveles de estos factores. Este mismo cuadro presupone que el grupo de estudio y el de control tienen el mismo tamaño. También presupone que estamos interesados en los resultados del estudio, tanto si se producen en la dirección del tratamiento en estudio como en la opuesta. Los estadísticos denominan las pruebas de significación estadística que consideran los datos que se desvían de la hipótesis nula en ambas direcciones pruebas bilaterales.[4] En el cuadro 11-1 también se supone un error de tipo I de 5%.

Veamos el significado de esas cifras en los diferentes tipos de estudios. Imagine que deseamos realizar un ensayo clínico controlado con un tratamiento destinado a reducir la mortalidad en un año por adenocarcinoma del ovario. Supongamos que la mortalidad anual utilizando el tratamiento estándar es 40%. En este estudio esperamos ser capaces de reducir la mortalidad en un año hasta 20% mediante un nuevo tratamiento. No obstante, creemos que el tratamiento puede aumentar la tasa de mortalidad. Si estamos dispuestos a tolerar una probabilidad de 20% de no obtener resultados estadísticamente significativos aunque exista una diferencia verdadera de esa magnitud en la población, ¿cuántos pacientes es necesario incluir en los grupos de estudio y de control?

Para responder a esta pregunta podemos utilizar el cuadro 11-1 del siguiente modo. Localice, en el eje horizontal, la probabilidad de 20% de un efecto adverso en el grupo de estudio. Seguidamente, localice en el eje vertical la probabilidad de 40% de un efecto adverso en el grupo de control. Estas probabilidades se intersecan en las casillas que contienen la cifras 117, 90 y 49. La respuesta correcta es la que se alinea con el error de tipo II de 20%. La respuesta es 90. Por lo tanto, se necesitan 90 mujeres con adenocarcinoma de ovario avanzado en el grupo de estudio y 90 en el de control para tener una probabilidad de 20% de no demostrar una diferencia estadísticamente significativa si la verdadera tasa de mortalidad en un año es realmente 40% con el tratamiento estándar y 20% con el nuevo tratamiento.

En los ensayos clínicos controlados generalmente se utilizan muestras de 100. Esta es una estimación aproximada del número de individuos necesarios en cada grupo cuando la probabilidad de un efecto adverso es sustancial y los investigadores esperan reducirla a la mitad con el nuevo tratamiento, mientras mantienen la magnitud del error de tipo II por debajo de 20%.

Ahora contrastaremos esta situación con aquella en la que la probabilidad de un efecto adverso es mucho menor, incluso sin intervención.

Un investigador desea estudiar el efecto de un nuevo tratamiento sobre el riesgo de sepsis neonatal secundaria a un retraso en la visita al ginecólogo tras la rotura de aguas. Supondremos que el riesgo de sepsis neonatal empleando el trata-

[4] N del E. Estas pruebas también se denominan "pruebas de dos colas".

CUADRO 11-1. Tamaño de la muestra necesario en los ensayos clínicos controlados

		Error de tipo II	Probabilidad de un desenlace adverso en el grupo de estudio			
			1%	5%	10%	20%
Probabilidad de un desenlace adverso en el grupo de control	2%	10%	3,696	851	207	72
		20%	2,511	652	161	56
		50%	1,327	351	90	38
	10%	10%	154	619		285
		20%	120	473		218
		50%	69	251		117
	20%	10%	62	112	285	
		20%	49	87	218	
		50%	29	49	117	
	40%	10%	25	33	48	117
		20%	20	26	37	90
		50%	12	16	22	49
	60%	10%	13	16	20	34
		20%	11	13	16	27
		50%	7	8	10	16

miento estándar es de 10% y que el estudio pretende reducirlo a 5%, aunque es posible que el nuevo tratamiento aumente la tasa de mortalidad. El investigador está dispuesto a aceptar una probabilidad de 10% de no demostrar una diferencia estadísticamente significativa.

Usando los datos del mismo cuadro, como lo hicimos antes, encontramos 619, 473 y 251. Por tanto, necesitamos 619 individuos en el grupo de estudio y 619 en el de control para garantizar una probabilidad de 10% de cometer un error de tipo II. Si estamos dipuestos a tolerar una probabilidad de 20% de no demostrar una diferencia estadísticamente significativa, en el caso de que exista realmente en la población, necesitaríamos 473 individuos en cada grupo. Quinientos individuos en cada grupo es una cifra grande para un ensayo clínico controlado. Esta es la cifra aproximada que necesitamos si queremos ser capaces de demostrar una diferencia estadísticamente significativa cuando la verdadera diferencia en la población es solo de 10% frente a 5%. El ejemplo de la sepsis neonatal es un problema típico que estudiamos en la práctica clínica. Demuestra por qué en muchos estudios clínicos controlados son necesarias grandes muestras antes de que sea posible demostrar una diferencia estadísticamente significativa. Por eso, generalmente no es viable someter a la prueba de un ensayo clínico controlado las mejoras terapéuticas de poca magnitud.

Avancemos un paso más y veamos qué le sucede al tamaño de la muestra requerido cuando un ensayo clínico controlado se realiza sobre una intervención preventiva en la cual el efecto adverso es infrecuente incluso sin la prevención.

Imaginemos un nuevo fármaco que previene los efectos adversos del embarazo de las mujeres que eran hipertensas antes de quedar embarazadas y con el que se pretende reducir los riesgos de los resultados adversos del embarazo de 2% a 1%, aunque sea posible que el nuevo tratamiento aumente la tasa de mortalidad. Los investigadores están dispuestos a tolerar una probabilidad de 20% de no demostrar una diferencia estadísticamente significativa.

En el cuadro 11-1 podemos ver que se necesitan 2 511 individuos en cada grupo. Estas cifras tan altas señalan la dificultad de realizar ensayos clínicos

controlados cuando uno desea aplicar tratamientos preventivos, especialmente si el riesgo de desenlaces adversos ya es bastante bajo.

Aun cuando un ensayo clínico controlado sea viable, es posible que no sea ético realizarlo. Los ensayos clínicos controlados no se consideran éticos si exigen someter a los individuos a riesgos importantes sin una previsión realista de beneficios sustanciales. Por ejemplo, un ensayo con estrógenos sin progesterona a altas dosis no sería permitido hoy día por un comité institucional de revisión cuya aprobación es necesaria para utilizar voluntarios en un estudio. Por eso, a pesar de las ventajas de los ensayos clínicos controlados para definir la eficacia de un tratamiento, estos estudios no son siempre viables o éticos.

Asignación

Los individuos incluidos en un ensayo controlado aleatorio habitualmente no se seleccionan al azar de la población. Generalmente son voluntarios que cumplen una serie de criterios de inclusión y de exclusión establecidos por los investigadores.

Los voluntarios de una investigación deben dar su consentimiento informado, cuyo formulario debe contener una explicación de los riesgos conocidos y de las opciones disponibles. Los voluntarios pueden abandonar el estudio en cualquier momento y por cualquier razón; sin embargo, no tienen derecho a saber a qué grupo han sido asignados mientras estén en el estudio y no pueden recibir indemnizaciones por los efectos secundarios causados por el tratamiento.

La asignación al azar de los pacientes a los grupos de estudio y de control es la característica distintiva de los ensayos clínicos aleatorios. La asignación al azar implica que todo individuo tiene una probabilidad predeterminada de ser asignado a un grupo, sea de estudio o de control. Esto puede significar probabilidades idénticas o diferentes de ser asignado a uno de los dos grupos.

La asignación al azar es un instrumento poderoso para eliminar el sesgo de selección en la asignación de los individuos a los grupos de control o de estudio. En los grandes estudios permite reducir la posibilidad de que los efectos del tratamiento sean debidos a los tipos de individuos que reciben el tratamiento de estudio o el de control. Es importante distinguir entre la asignación al azar, que es una parte esencial de un ensayo clínico controlado, y la selección al azar, que no forma parte habitualmente de un ensayo clínico controlado. La selección al azar, al contrario de la asignación al azar, supone que el individuo seleccionado para un estudio es escogido al azar de un grupo o población más grande. Así, la selección al azar es un método dirigido a obtener una muestra representativa (esto es, aquella que refleja las características del grupo más grande).

La asignación al azar, por otro lado, no dice nada acerca de las características de la población de la que se extraen los individuos de la investigación. Se refiere al mecanismo mediante el cual los individuos son asignados a los grupos de estudio y de control, una vez que son elegibles para el estudio y aceptan participar voluntariamente en él. El siguiente estudio hipotético muestra la diferencia entre la selección al azar y la asignación al azar.

Un investigador desea valorar los beneficios de un nuevo fármaco denominado "*Surf-ez*", elaborado con el propósito de mejorar la capacidad de hacer *surfing*. Para valorar los efectos del *Surf-ez*, el investigador realiza un ensayo clínico controlado con un grupo de voluntarios que son campeones de *surfing* en Hawai. Una vez asignados aleatoriamente, la mitad al grupo que toma *Surf-ez* y la otra mitad al grupo que toma

un placebo, se mide la capacidad de realizar *surfing* en todos los individuos mediante un sistema de puntuación estándar. Los calificadores desconocen quiénes toman *Surf-ez* y quiénes reciben el placebo. Los que toman *Surf-ez* muestran una mejora estadísticamente significativa y considerable en comparación con los que toman el placebo. A partir de estos resultados, los autores recomiendan a todos los que practican *surfing* que tomen *Surf-ez* como medio para mejorar su capacidad en este deporte.

Este ensayo clínico controlado ha demostrado la eficacia del *Surf-ez* entre campeones de *surfing* mediante el uso de la asignación al azar. Sin embargo, dado que en este estudio el grupo de estudio y el de control difícilmente constituían una muestra representativa de aficionados, hemos de ser muy cuidadosos al sacar conclusiones sobre los efectos del *Surf-ez* como ayuda al aprendizaje de todos los que practican ese deporte.[5]

La asignación al azar no elimina la posibilidad de que los grupos de estudio y de control difieran en cuanto a factores que influyen en el pronóstico (variables de confusión). Los factores pronósticos conocidos también deben ser medidos, y muchas veces se encontrarán diferencias entre los grupos de estudio y de control debidas solo al azar, especialmente en estudios pequeños. Si existen diferencias sustanciales entre los grupos, es preciso tomarlas en cuenta en el análisis mediante un proceso de ajuste.[6] Sin embargo, muchas de las características que influyen en el pronóstico no se conocen. En estudios de grupos grandes, la asignación al azar tiende a equilibrar la multitud de características que podrían estar relacionadas con el desenlace, incluidas las que desconoce el investigador. Sin la asignación al azar, el investigador necesitaría tener en cuenta todas las diferencias conocidas y potenciales entre los grupos. Dado que es difícil, si no imposible, tenerlo todo en cuenta, la asignación al azar ayuda a equilibrar los grupos, especialmente en los estudios grandes.

Valoración

En el diseño de los ensayos clínicos controlados, el *enmascaramiento* (*blinding* o *masking*) de los sujetos de estudio y de los investigadores se suele considerar una característica importante para prevenir errores en la valoración de los desenlaces. El *enmascaramiento simple ciego* significa que el paciente no sabe qué tratamiento recibe y *doble ciego*, que ni el paciente ni el investigador saben a qué grupo ha sido asignado.

Se pueden cometer errores en la valoración del desenlace o resultado de un ensayo clínico controlado cuando el paciente o la persona que efectúa la valoración sabe cuál es el tratamiento administrado. Es muy probable que esto ocurra cuando el desenlace o resultado medido es subjetivo o está influido por el conocimiento del grupo de tratamiento, como se muestra en el siguiente estudio hipotético.

En un ensayo clínico controlado de un nuevo tratamiento quirúrgico del cáncer de mama, se comparó el edema y la fuerza en el brazo con el nuevo procedimiento respecto del tradicional. Las pacientes sabían cuál procedimiento se les había practicado, y el edema y la fuerza del brazo eran los resultados valorados por ellas y por los cirujanos. El estudio mostró que las pacientes a las que se había practicado el

---

[5] Se debe tener cuidado incluso al extrapolar los resultados a los campeones de *surfing*, puesto que no se ha llevado a cabo una selección al azar entre ellos. Esta limitación se produce en muchos ensayos clínicos controlados en los que se selecciona a los pacientes de un hospital o de una clínica en particular.

[6] Muchos bioestadísticos recomiendan usar técnicas de análisis multivariante, como el análisis de regresión, incluso cuando no existen diferencias sustanciales entre los grupos. El uso de análisis multivariantes permite ajustar según las interacciones. La interacción se produce cuando, por ejemplo, ambos grupos contienen idénticas distribuciones de edad y sexo, pero uno contiene mayoritariamente hombres jóvenes y el otro mujeres jóvenes. El análisis multivariante permite separar los efectos de la interacción de la edad y el sexo.

nuevo procedimiento tenían menos edema y más fuerza en el brazo que aquellas a las que se había practicado la mastectomía tradicional.

En este estudio, el hecho de que tanto las pacientes como los cirujanos que realizaron la operación y que valoraron el desenlace sabían qué procedimiento se había llevado a cabo pudo haber influido en el grado de objetividad con que se midieron y notificaron el edema y la fuerza del brazo. Este efecto se podría haber minimizado, pero no suprimido totalmente, si el edema y la fuerza del brazo hubieran sido valorados mediante un sistema de puntuación estandarizado por individuos que no sabían qué terapia habían recibido las pacientes. Este sistema de enmascaramiento simple y de puntuación objetiva minimizaría el impacto del hecho de que las pacientes y los cirujanos sabían qué técnica quirúrgica se le practicó a cada una. No obstante, también es posible que las pacientes sometidas a la nueva técnica pusieran más de su parte para aumentar la fuerza del brazo y reducir el edema. Esto podría suceder si el cirujano que realiza el nuevo procedimiento pone un enérgico énfasis en los ejercicios posoperatorios de las pacientes.

En la práctica, el enmascaramiento muchas veces no tiene sentido o es infructuoso. Los procedimientos quirúrgicos no se enmascaran fácilmente. El sabor o los efectos secundarios de los medicamentos constituyen un indicio para el paciente o el médico, o para ambos. La necesidad de titular una dosis para conseguir el efecto deseado hace más difícil enmascarar al médico y, en algunos casos, al paciente. El acatamiento estricto del enmascaramiento contribuye a garantizar la objetividad del proceso de valoración. Por añadidura, ayuda a eliminar la posibilidad de que las diferencias en el cumplimiento, seguimiento y valoración del desenlace estén influidas por el conocimiento del tratamiento que se está recibiendo.

Aunque se pueda garantizar una valoración objetiva, un cumplimiento excelente y un seguimiento completo, el enmascaramiento es deseable dado que contribuye a controlar el efecto placebo. El efecto placebo es un potente proceso biológico que produce una serie de efectos biológicos objetivos y subjetivos, y que trasciende el control del dolor. Un porcentaje sustancial de los pacientes que creen estar recibiendo un tratamiento efectivo obtienen beneficios terapéuticos objetivos. Cuando el enmascaramiento efectivo no forma parte de un ensayo clínico controlado, queda abierta la posibilidad de que el beneficio observado en los sujetos estudiados sea realmente resultado del placebo.

De modo que, cuando es imposible enmascarar, queda una duda acerca de la exactitud de las mediciones del desenlace. Esta incertidumbre puede reducirse pero no suprimirse con el uso de medidas objetivas de los resultados, con el monitoreo cuidadoso del cumplimiento y con un seguimiento completo de los pacientes.

La valoración válida del desenlace requiere de medidas apropiadas, precisas y completas que no estén influidas por el proceso de observación. Estos requisitos son tan importantes en un ensayo clínico controlado como en un estudio de cohortes o de casos y controles, como hemos comentado en el capítulo 4.

En un ensayo clínico controlado ideal, todos los individuos serían tratados y seguidos de acuerdo con el protocolo del estudio. Sus desenlaces se valorarían desde el momento de su entrada en el estudio hasta su finalización. En realidad, la valoración difícilmente es tan completa o perfecta. Los pacientes muchas veces reciben tratamientos que se desvían de los predefinidos en el protocolo. Los investigadores suelen denominarlos *desviaciones del protocolo*. Además, en ocasiones no es posible seguir a todos los pacientes antes de finalizar el estudio.

En los ensayos clínicos controlados pueden surgir sesgos a partir de estas desviaciones del protocolo y de los pacientes que no se han podido incluir en

el seguimiento. Veamos un ejemplo en el siguiente estudio hipotético.

En un ensayo clínico controlado sobre la diálisis renal, se asignaron al azar 100 pacientes a una sesión diaria de diálisis y otros 100 a una sesión semanal de diálisis intensiva. Durante el estudio, dos pacientes del primer grupo se desviaron del protocolo y recibieron un trasplante de riñón, mientras que 20 pacientes del segundo grupo se desviaron del protocolo y también recibieron trasplantes renales. Los investigadores eliminaron del estudio a los que habían recibido trasplantes, creyendo que su inclusión influiría negativamente sobre los resultados del estudio.

Es posible que muchos de los que recibieron trasplantes estuvieran respondiendo mal al tratamiento con diálisis. Si este fuera el caso, la exclusión de los que se desviaron del protocolo sesgaría los resultados del estudio a favor del grupo sometido a diálisis semanal. Esto ocurriría si los que continuaron en el grupo de diálisis semanal eran principalmente los que estaban respondiendo bien al tratamiento.

A causa del sesgo potencial, generalmente se recomienda que las personas desviadas del protocolo continúen participando en la investigación y se analicen como si hubiesen continuado en el grupo al que fueron asignadas al azar. Esto se conoce como *análisis de acuerdo con la intención de tratar* (*analysis according to the intention to treat*). Sin embargo, al retener a los que se desvían del protocolo, la cuestión planteada en el estudio cambia ligeramente. Ahora, lo que se plantea es si la política de administrar en lo posible el nuevo tratamiento es mejor que la de administrar el tratamiento estándar tanto como sea posible. Esta modificación ayuda realmente a mejorar la aplicabilidad de la investigación a las cuestiones clínicas reales o, en otras palabras, a la efectividad del tratamiento tal como se utiliza en la práctica clínica.

Las desviaciones del protocolo son relativamente frecuentes en los ensayos clínicos controlados, ya que no se considera ético evitarlas cuando el médico de un participante opina que el acatamiento prolongado está contraindicado por el estado del paciente o cuando el propio paciente no desea seguir por más tiempo el protocolo. Por eso, al evaluar un ensayo clínico controlado, el lector debe entender el grado de adherencia al protocolo y determinar cómo manejaron los investigadores los datos de aquellos que se desviaron del protocolo.

Es posible que surja un problema similar cuando el seguimiento de algunos individuos se ha visto interrumpido antes de terminar el estudio. Incluso pérdidas moderadas en el seguimiento pueden ser desastrosas, si los que se pierden han emigrado a lugares salubres, como Arizona, porque ha empeorado su salud, abandonan el tratamiento por la toxicidad de los fármacos o no regresan al estudio porque les es difícil cumplir con alguno de los protocolos de tratamiento.

En los estudios bien realizados se toman precauciones extremas para minimizar las pérdidas en el seguimiento. En algunos casos, el seguimiento puede completarse mediante una entrevista telefónica o un cuestionario enviado por correo. En otros casos, puede ser necesario realizar una búsqueda de certificados de defunción de los que no han podido ser localizados. Cuando, a pesar de todas estas precauciones, algunos pacientes quedan excluidos del seguimiento, es importante determinar, en lo posible, las características iniciales de esas personas. Esto se hace para intentar averiguar si es probable que los perdidos sean diferentes de los que continúan en el estudio. Si los perdidos en el seguimiento tienen un pronóstico especialmente desfavorable, poco se puede ganar analizando los datos de los que siguen en el estudio, como sugiere el siguiente estudio hipotético.

En un estudio sobre los efectos de un nuevo programa de tratamiento contra el alcoholismo, se asignó al azar a 100 pacientes a ese nuevo tratamiento y 100 al convencional. Los investigadores visitaron los domicilios de todos los pacientes

un sábado a las 9 de la noche y extrajeron una muestra de sangre de todos los que encontraron para medir la alcoholemia. De los pacientes del grupo con el nuevo tratamiento, 30 estaban en su domicilio y, de ellos, un tercio tenían alcohol en la sangre. Entre los pacientes tratados convencionalmente, 40 estaban en su domicilio y la mitad tenían alcohol en la sangre.

Siempre que ocurra una pérdida importante o desproporcionada en el seguimiento de un grupo, conviene preguntarse qué ha ocurrido con los perdidos. En este estudio, si los perdidos en el seguimiento estaban fuera de su domicilio bebiendo, los resultados que solo tuviesen en cuenta a los que se hallaban en la casa serían especialmente desorientadores.

Un método para tratar las pérdidas en el seguimiento consiste en suponer lo peor de aquellos que se han perdido. Por ejemplo, se podría suponer que todos los pacientes que no se encontraban en su domicilio estaban bebiendo. Se puede entonces repetir el análisis y comparar el desenlace en los grupos de estudio y de control. Cuando las pérdidas en el seguimiento son grandes, este procedimiento generalmente nos deja sin una diferencia sustancial o estadísticamente significativa entre el grupo de estudio y el de control. Sin embargo, cuando las pérdidas son pequeñas, puede continuar existiendo una diferencia estadísticamente significativa entre los grupos de estudio y de control. Cuando se mantienen las diferencias estadísticamente significativas entre ambos grupos después de suponer el peor resultado con respuesto a los perdidos, el lector puede tener total confianza en que las pérdidas en el seguimiento no explican las diferencias observadas.

Análisis

El investigador debe responder a dos cuestiones básicas al realizar un ensayo clínico controlado: cuándo y cómo analizar los datos.

Cuándo analizar los datos

Esta cuestión aparentemente simple ha provocado una considerable controversia metodológica y ética. Cuantas más veces analice uno los datos, más probable es que llege un momento en que el valor *P* alcance el valor 0,05 de significación estadística.

Cuándo analizar es un problema ético, dado que se desea establecer la existencia de una verdadera diferencia lo antes posible para evitar someter a los pacientes a un tratamiento menos efectivo. Además, es de desear que otros pacientes reciban un tratamiento efectivo cuanto antes.

En un intento de hacer frente a estos problemas, se ha desarrollado una serie de métodos "secuenciales". Estos métodos han tenido mucho éxito cuando se han aplicado en estudios de enfermedades agudas en las cuales el desenlace se conoce en un período muy breve. Sin embargo, la mayor parte de los estudios se basan en la técnica de realizar análisis en momentos predeterminados. Por eso, es importante entender cuándo y con cuánta frecuencia es preciso analizar los datos. En una situación ideal, los momentos se han de determinar antes de iniciar el estudio y de acuerdo con los períodos en que se esperaría un efecto terapéutico. Por ejemplo, en el tratamiento antibiótico de una enfermedad aguda, el resultado puede valorarse diariamente. En el estudio de la mortalidad por cáncer, el resultado solo puede medirse anualmente. Cuando

se prevé realizar varias comparaciones en el análisis, existen técnicas estadísticas para tomarlas en consideración al calcular el valor *P*.[7]

Cómo analizar los datos

Las tablas de vida (*life tables*) son el método de análisis diseñado para los ensayos clínicos controlados más empleado. Se usan para mostrar cuándo y con qué frecuencia se producen los desenlaces adversos.

En este caso, cuando hablemos del efecto adverso estudiado nos referiremos a la muerte. Sin embargo, las tablas de vida pueden utilizarse para presentar otros efectos, como la pérdida permanente de la visión o la aparición de consecuencias deseables como el embarazo después de un tratamiento contra la infertilidad.

Empezaremos comentando por qué muchas veces son necesarias las tablas de vida en los ensayos clínicos controlados. A continuación, examinaremos los supuestos en que se basa su utilización y mostraremos cómo se deben interpretar.

En la mayor parte de los ensayos clínicos controlados, los individuos que ingresan en el estudio son seleccionados al azar durante cierto tiempo, a medida que acuden a recibir asistencia. Además, a causa de ingresos tardíos, muerte o pérdidas en el seguimiento, el tiempo de seguimiento de cada individuo puede variar. Por lo tanto, muchos pacientes no son seguidos durante todo el estudio.

Si se sigue a todos los individuos durante el mismo espacio de tiempo, el cálculo de la probabilidad de morir es igual al número de personas que han fallecido al término del estudio dividido por el total de participantes iniciales. Sin embargo, todos los individuos no son seguidos durante el mismo período y las tablas de vida proporcionan un método para utilizar los datos de aquellos individuos que solo han participado en una parte de la duración total del estudio.[8] Por lo tanto, las tablas de vida permiten al investigador utilizar todos los datos que ha recogido tan laboriosamente.

El método de las tablas de vida se basa en el supuesto fundamental de que quienes participaron en la investigación durante períodos más cortos tuvieron la misma experiencia ulterior que los que fueron seguidos durante períodos más prolongados. En otras palabras, los de "períodos cortos" tendrían los mismos resultados que los de "períodos largos", si fueran seguidos durante más tiempo.

Este supuesto fundamental puede no ser cierto si los individuos seguidos durante períodos cortos tienen un pronóstico mejor o peor que los de períodos largos. Esto puede ocurrir si la rigurosidad de los criterios de inclusión disminuye durante el curso del estudio. El siguiente estudio hipotético ilustra esta posibilidad.

Mediante un ensayo clínico controlado, se comparó con un tratamiento estándar un nuevo tratamiento hormonal destinado a tratar la infertilidad secundaria a la endometriosis grave. Tras la dificultad inicial de reclutar pacientes y los fracasos iniciales para conseguir mujeres embarazadas entre las integrantes del grupo de estudio, una mujer de este grupo quedó embarazada. La noticia de que dio a luz se publicó en la primera plana de los periódicos. Si bien las siguientes pacientes reclutadas para el estudio tenían endometriosis menos graves, los investigadores las aceptaron y combinaron los resultados con los del grupo original de pacientes.

---

[7] Nota del E. Esta situación se conoce como "problema de las comparaciones múltiples".

[8] Existen varios tipos de tablas de vida para cohortes, dos de las cuales se denominan tablas de vida de Kaplan-Meier y de Cutler-Ederer. Las tablas de vida de cohortes deben distinguirse de las tablas de vida transversales, que se emplean para estimar la esperanza de vida.

Como demuestra este estudio, la rigurosidad de los criterios de inclusión puede disminuir si al inicio de la investigación solo se incluyen pacientes gravemente enfermos. A medida que el tratamiento se da a conocer en la comunidad, en una institución concreta o en la literatura, los médicos y también los pacientes pueden tener la tendencia a remitir enfermos menos graves para ser tratados. En este caso, es más probable que las mujeres seguidas por períodos cortos tuvieran una enfermedad menos grave y, por lo tanto, mejores desenlaces que las seguidas durante períodos largos. Este problema puede minimizarse si los investigadores definen claramente en el protocolo el tipo de pacientes que son elegibles para el estudio a partir de las características relacionadas con el pronóstico y se ciñen estrictamente a esa definición. Otra opción consiste en reconocer el problema y ajustar los datos por medio de técnicas estadísticas, para tener en cuenta la gravedad de la enfermedad de los pacientes en el momento de su entrada en el estudio.

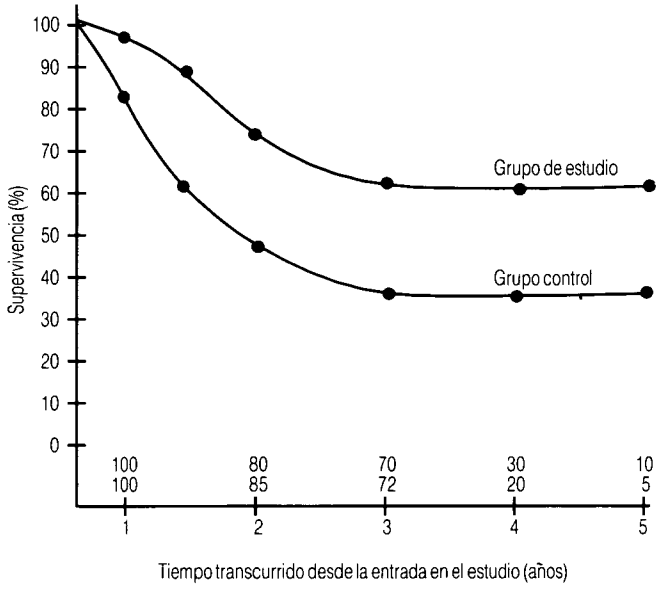
Las pérdidas en el seguimiento también pueden producir diferencias entre los individuos seguidos durante períodos cortos y los que se han seguido durante más tiempo. Es probable que esto se produzca cuando las pérdidas de seguimiento se producen preferentemente entre aquellos con peor evolución o entre los que presentan reacciones adversas al tratamiento. Ya hemos hablado de la importancia de las pérdidas de seguimiento y subrayado la necesidad de valorar si los pacientes perdidos son o no similares a los que permanecen en el estudio.

Generalmente, los datos de las tablas de vida se presentan como una curva de supervivencia. Se trata de un gráfico en cuyo eje vertical se representa el porcentaje de supervivencia, que va de 0% en la base a 100%. Así, al inicio del estudio, tanto el grupo de estudio como el de control parten del 100% señalado en la parte superior del eje vertical.[9] En el eje horizontal se representa el tiempo de seguimiento. El tiempo se cuenta para cada individuo a partir de su entrada en el estudio. De esta forma, el tiempo cero no es el momento cuando empieza la investigación. Las curvas de supervivencia también deben incluir el número de individuos que se han seguido en cada intervalo de tiempo. En condiciones ideales, esto debe presentarse separadamente para el grupo de estudio y el de control. De este modo, una curva típica de tabla de vida comparando datos de 5 años del grupo de estudio y del de control podría parecerse a la de la figura 11-1. Cuando los datos de la tabla de vida se expresan como estimaciones del porcentaje de muerte o de supervivencia, por ejemplo, a los 5 años, la tabla se denomina supervivencia *actuarial* a los 5 años. Las cifras de la parte inferior indican el número de individuos que son seguidos en el grupo de estudio y en el de control hasta un determinado momento tras su ingreso en el estudio.

Para realizar pruebas de significación estadística con los datos de las tablas de vida se emplean con frecuencia la prueba del *log-rank* o la de Mantel-Haenszel. La hipótesis nula en estas pruebas afirma que no existen diferencias entre las curvas del grupo de estudio y las del grupo de control. Estas pruebas comparan los sucesos observados y los esperados si fuese cierta la hipótesis nula de que no hay diferencias entre los grupos. Al realizar estas pruebas de significación estadística, se combinan los datos de cada intervalo de tiempo teniendo en cuenta o ponderando el número de individuos seguidos durante ese intervalo. Así, estas pruebas combinan datos de los distintos intervalos de tiempo para producir una prueba de significación estadística global. La combinación de los datos de múltiples intervalos significa que al realizar la prueba

[9] En otra forma de presentación gráfica de las tablas de vida se puede representar el porcentaje de los que experimentan el efecto adverso, comenzando a partir de 0% al inicio de la parte inferior del eje vertical.

FIGURA 11-1. Tabla de vida típica de un grupo de estudio y uno de control que demuestra el *efecto meseta*, el cual aparece típicamente en el extremo derecho de las representaciones gráficas de las tablas de vida



de significación estadística uno se plantea la siguiente pregunta: si no existen verdaderas diferencias entre los efectos globales de los tratamientos del grupo de estudio y del de control, ¿cuál es la probabilidad de obtener los resultados observados u otros más extremos? En otras palabras, si se ha demostrado una mejoría estadísticamente significativa en el grupo de estudio sobre la base de los resultados de las tablas de vida, es muy probable que un grupo similar de individuos que reciba el tratamiento del grupo de estudio experimente al menos alguna mejora en comparación con el tratamiento del grupo de control.

Interpretación

Como ya se ha mencionado, los datos de las tablas de vida inducen a numerosas interpretaciones incorrectas. Cuando se presenta una tabla de vida es muy importante indicar el número de individuos seguidos en cada intervalo de tiempo en los grupos de estudio y de control. Habitualmente, el número de sujetos seguidos durante todo el tiempo del estudio es bajo. Por ejemplo, en la figura 11-1 solo se siguieron 10 sujetos durante 5 años en el grupo de estudio y 5 en el de control. Esto no es sorprendente, dado que muchas veces se necesita algún tiempo para iniciar un estudio y los individuos seguidos por más tiempo fueron reclutados en el primer año del estudio.

La supervivencia actuarial a los 5 años puede calcularse aunque solo se haya seguido a un paciente durante los 5 años. Por eso, se debe evitar depositar una confianza excesiva en la probabilidad específica de un año, 5 años o en cualquier otra probabilidad final, a no ser que el número de individuos realmente seguidos durante todo el estudio sea elevado.

Al interpretar los resultados de un ensayo clínico controlado es importante examinar el grado de confianza que se puede tener en las estimaciones de la supervivencia. La incapacidad para reconocer esta incertidumbre puede producir el siguiente tipo de interpretación errónea.

Un clínico que examinó las curvas de la tabla de vida de la figura 11-1 llegó a la conclusión de que la supervivencia a los 5 años con el tratamiento en estudio era de 60%, y la del grupo de control, de 35%. Después de aplicar el mismo tratamiento a pacientes similares le sorprendió que de los pacientes a los que se administró el tratamiento estudiado la supervivencia fuera de 55% y la de los pacientes del grupo de control, 50%.

Si el clínico hubiera sabido que las curvas de la tabla de vida no predicen de forma fiable la supervivencia exacta a los 5 años, no le habrían sorprendido los resultados de su experiencia posterior.

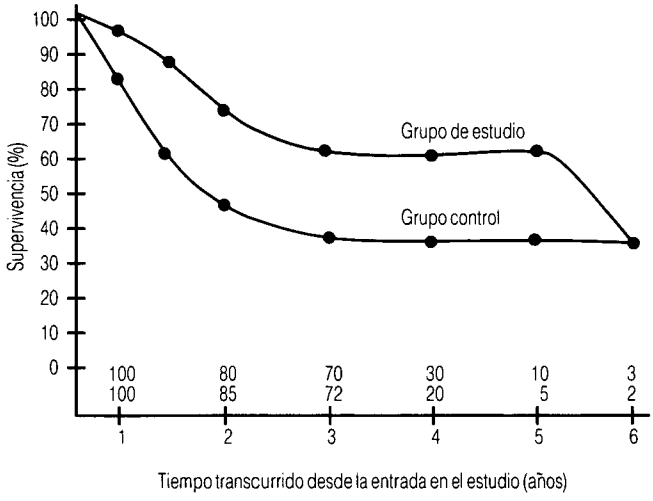
El conocimiento de los procedimientos y de los supuestos subyacentes a las tablas de vida también ayudan a interpretarlas. Muchas curvas de supervivencia tienen una fase plana o de meseta que corresponde a largos intervalos de tiempo en el extremo derecho de la gráfica. Estas se pueden interpretar erróneamente como indicación de una curación cuando un individuo alcanza la fase de meseta de la curva. En realidad, esta fase de meseta se produce habitualmente porque se siguen pocos individuos durante todo el estudio. Entre estos individuos seguidos durante intervalos de tiempo más largos es más probable que las muertes sean pocas y estén muy esparcidas. Dado que la curva de supervivencia solo declina con una muerte, cuando se producen pocos fallecimientos es posible que aparezca una fase de meseta. En consecuencia, para interpretar una tabla de vida es importante comprender el *efecto meseta* (*plateau effect*). No debemos interpretar una meseta como prueba de que se ha producido una curación, a no ser que se haya seguido a un elevado número de pacientes durante largos períodos de tiempo.

Además del peligro de confiar demasiado en la supervivencia actuarial a los 5 años derivada de una tabla de vida y de interpretar erróneamente la meseta, es importante entender completamente el significado de una diferencia estadísticamente significativa entre las curvas de supervivencia. En el estudio ilustrado en la figura 11-1, existía una diferencia estadísticamente significativa entre el desenlace del grupo de estudio y el de control sobre la base del seguimiento actuarial de los pacientes a los 5 años. El estudio se amplió posteriormente por un año adicional y los resultados obtenidos se representaron mediante la curva que aparece en la tabla de vida en la figura 11-2. De acuerdo con esta curva, la supervivencia actuarial a los 6 años fue la misma en ambos grupos. Sobre la base de los datos actuariales a los 6 años, los autores afirmaron que el estudio actuarial a los 5 años fue erróneo, al concluir que el tratamiento estudiado prolongaba la supervivencia.

Recuerde que una diferencia estadísticamente significativa en las curvas de supervivencia implica que los pacientes que reciben un tratamiento evolucionan mejor que los que reciben el otro tratamiento, cuando se tiene en cuenta la experiencia global de cada grupo. Los pacientes de un grupo pueden mejorar solo al principio del tratamiento, a la mitad o solamente al final. Los pacientes que reciben el mejor tratamiento pueden estar realmente peor al principio debido a complicaciones quirúrgicas o, bien, más tarde, por las complicaciones secundarias que surgen en los que sobreviven.

Por lo tanto, cuando se realiza un estudio, es importante conocer lo suficiente de la historia natural de la enfermedad y la esperanza de vida de los individuos para escoger un período de seguimiento que tenga sentido. Es improbable en-

FIGURA 11-2. Las curvas de las tablas de vida pueden unirse después de largos períodos de seguimiento. Incluso en este caso, la diferencia entre ambas curvas puede ser estadísticamente significativa



contrar diferencias en el desenlace si ese período es demasiado corto, por ejemplo, si no alcanza hasta que el tratamiento se termine o llegue a tener un efecto biológico esperado.

Asimismo, los períodos de seguimiento demasiado largos pueden impedir demostrar una diferencia estadísticamente significativa si los riesgos de las enfermedades sobrepasan los beneficios a corto plazo. Por ejemplo, un estudio en el que se valore el desenlace de un tratamiento para la enfermedad coronaria a los 20 años en personas de 65 años de edad podría detectar pocas diferencias a los 20 años, aunque existieran diferencias a los 5 y a los 10 años.

El empleo de una curva de supervivencia y de pruebas de significación estadística proporciona información sobre el éxito del tratamiento en el grupo de estudio y en el de control. Sin embargo, se puede facilitar la interpretación de este efecto considerando si los grupos se diferencian en función de uno o varios factores que influyen en el pronóstico. Estos factores se denominan *variables de confusión* si son distintos en el grupo de estudio y el de control y están relacionados con la probabilidad de un desenlace adverso.

Un método para tratar las diferencias en los factores pronósticos entre grupos consiste en separar o estratificar a los pacientes de acuerdo con su pronóstico al inicio del estudio y luego asignar al azar a los individuos de cada categoría pronóstica o estrato a los grupos de estudio y de control. Este tipo de asignación al azar por bloques o estratificada es una forma de apareamiento por grupos que se usa con frecuencia en los ensayos clínicos controlados. Otra posibilidad consiste en tener en cuenta esas diferencias al final del estudio mediante un *método de ajuste*.

El ajuste según los factores pronósticos exige que la información sobre estos factores, también denominados de riesgo, se recoja al inicio del estudio. Si las diferencias entre los grupos son importantes y esos factores pronósticos del desenlace son potentes, estas diferencias se pueden ajustar. Cuando se usa el método de la tabla de vida, es posible utilizar una prueba de significación estadística ajustada en la

que se suman los desenlaces observados y esperados en cada uno de los diferentes estratos pronósticos, así como en los distintos intervalos de tiempo.[10] Por sí mismas, las curvas de las tablas de vida no suelen estar ajustadas, aunque pueden estarlo. Las pruebas de significación estadística deben tener en cuenta el ajuste según las variables de confusión. Por eso, cuando se interpreta una prueba de significación estadística de los datos de una tabla de vida, es importante saber si sus resultados se ajustaron según variables de confusión importantes.

Hemos subrayado en repetidas ocasiones la distinción entre una asociación estadísticamente significativa y una relación de causa-efecto. El establecimiento de una relación de causa-efecto requiere, en primer lugar, que exista una asociación. En segundo lugar, obliga a demostrar que la causa precede al efecto. En tercer lugar, exige que la modificación de la causa modifique el efecto. Uno de los aspectos intelectualmente interesantes de los ensayos clínicos controlados es que incorporan métodos que ayudan a establecer los tres criterios de causa contribuyente y, de ese modo, la eficacia de un tratamiento.

1. Mediante técnicas de asignación al azar y de ajuste, los investigadores pueden crear grupos de estudio y de control que sean comparables excepto en los efectos del tratamiento administrado. Por eso, cuando existen diferencias notables y estadísticamente significativas en el desenlace, los investigadores pueden concluir que estas diferencias están asociadas al tratamiento.

2. Asignando al azar a los individuos de los grupos de estudio y de control al inicio del estudio, el investigador puede proporcionar pruebas muy evidentes de que el tratamiento precede al efecto y que, por lo tanto, existe una asociación previa que cumple con el criterio No. 2 de causa contribuyente.

3. Administrando un tratamiento que modifica el proceso patológico y comparando los desenlaces del tratamiento en el grupo de estudio y en el de control, los investigadores pueden aportar pruebas de que el tratamiento por sí mismo (la causa) está modificando realmente el desenlace (el efecto), cumpliendo, de esta forma, con el tercer y último criterio de causa contribuyente.

Por consiguiente, los ensayos clínicos controlados pueden ayudar a demostrar que existe una asociación entre un tratamiento y un desenlace, que existe una asociación previa y que la modificación de la causa modifica el desenlace. Estos son los tres criterios necesarios para afirmar que el nuevo tratamiento es la causa de la mejora en el desenlace. Dichos criterios establecen la eficacia del tratamiento. No obstante, siempre es posible que la mejoría observada haya sido producida por unos efectos no identificados ajenos al tratamiento, como sugiere el siguiente estudio.

Se realizó un ensayo clínico controlado sobre un nuevo programa de recuperación posoperatoria de la histerectomía en el que, después de la operación, se asignaron al azar 100 mujeres a una sala habitual y otras 100 a una sala de atención especial con camas experimentales, un equipo para ejercicios posoperatorios y dotada con más enfermeras de plantilla. Las mujeres de la sala especial fueron dadas de alta tras una estancia media de 7 días y las mujeres de la sala habitual, de 12 días. Los resultados fueron estadísticamente significativos. Los investigadores concluyeron que el

[10] Este método puede emplearse con variables de confusión nominales u ordinales. Los métodos de ajuste según las variables de confusión se discuten en el capítulo 29.

nuevo programa de recuperación posoperatoria produjo una reducción considerable en la estancia media.

Antes de concluir que las camas experimentales y el ejercicio posoperatorio causaron la diferencia, no hay que olvidar que también se necesitaron más enfermeras. El interés del investigador en el alta temprana junto con la disponibilidad de un mayor número de enfermeras pudo haber sido la causa del alta más temprana, en lugar de las camas y el ejercicio. En un estudio sin enmascaramiento como este, es posible que el efecto de la observación contribuya por sí mismo a causar el efecto observado. Aunque un ensayo clínico controlado bien realizado puede no establecer definitivamente que el tratamiento causó la mejoría, en la práctica los ensayos clínicos controlados satisfacen la definición de *eficacia*.

Extrapolación

Los pacientes incluidos en muchos ensayos clínicos aleatorios controlados son escogidos para participar porque son el tipo de pacientes que más probablemente responderán al tratamiento. Además, las consideraciones geográficas, de conveniencia para el investigador y de cumplimiento del paciente son habitualmente de capital importancia en la selección de un grupo concreto de pacientes para una investigación. Las pacientes embarazadas, los ancianos, las personas muy jóvenes y aquellas con enfermedades leves no se suelen incluir en los ensayos clínicos controlados a menos que el tratamiento esté diseñado especialmente para estos grupos. Además de estos factores selectivos que están bajo control del investigador, existen otros que pueden limitar la entrada en un ensayo clínico controlado a un grupo de pacientes con características exclusivas. Cada población de un centro sanitario tiene sus propios patrones de remisión de pacientes, de localización y socioeconómicos. Una población de pacientes remitidos a la Clínica Mayo puede ser completamente diferente de la de un hospital comarcal. Los pacientes de atención primaria de las Organizaciones para el Mantenimiento de la Salud (Health Maintenance Organizations, HMO)[11] pueden ser muy diferentes de los que acuden a la consulta externa de un servicio hospitalario. Estas características —que pueden estar fuera del alcance del investigador— pueden influir en los tipos de pacientes incluidos de manera que afecten los resultados del estudio.

El hecho de que el grupo de pacientes incluidos en ensayos clínicos controlados sea diferente del grupo de pacientes a quienes el clínico puede aplicar el nuevo tratamiento muchas veces crea dificultades para extrapolar las conclusiones a los pacientes atendidos en la práctica clínica. Esto no invalida el resultado del ensayo; simplemente significa que el clínico debe usar buen criterio y ser cauteloso al aplicar los resultados en la práctica clínica.

El proceso de extrapolación todavía es especulativo, a pesar de la potencia e importancia que tienen los ensayos clínicos controlados. El uso de muestras de conveniencia o fortuitas (*chunk samples*) en los ensayos clínicos controlados obliga a los clínicos que quieren aplicar sus resultados a examinar la naturaleza de las instituciones y de los pacientes del estudio. Los clínicos deben valorar si el medio y las circunstancias en que trabajan y sus pacientes son comparables con los del estudio. Si no lo

[11] Nota del T. Estas organizaciones son una estructura de prestación de servicios asistenciales caracterizada por asumir la responsabilidad contractual de un tipo predeterminado de asistencia sanitaria a una población definida, que se inscribe de forma voluntaria y que paga unas cuotas fijas, periódicas e independientes del uso de los servicios realizados. Son una alternativa al pago por acto médico dentro del sistema de salud de los Estados Unidos de América.

son, el lector debe preguntarse si las diferencias limitan la capacidad de efectuar extrapolaciones a partir de los resultados del estudio.

Los pacientes y el centro sanitario de estudio que participan en una investigación pueden diferir del contexto clínico habitual de muchas maneras, como se ejemplifica a continuación.

1. Es posible que los pacientes sigan cuidadosamente y se adhieran totalmente al tratamiento. El cumplimiento y el estrecho seguimiento pueden ser fundamentales para el éxito de un tratamiento.

2. Los participantes pueden tener peor pronóstico que los pacientes habituales de la práctica clínica. Por esta razón, puede merecer la pena correr el riesgo de los efectos secundarios del tratamiento en los pacientes estudiados, aunque es posible que esto no sea aplicable a los pacientes atendidos en otro lugar.

3. Los centros de estudio a veces disponen de equipamiento y de personal con habilidades o experiencia que maximizan el éxito del nuevo tratamiento. En otros lugares es posible que esto no sea cierto y que el nuevo tratamiento sea aplicado sin tener experiencia en esas técnicas.

Los clínicos deben tener en cuenta estas diferencias al extrapolar los resultados de un estudio a los pacientes de su práctica clínica, a pesar de que un ensayo clínico controlado haya demostrado la eficacia de un nuevo tratamiento. Estos estudios son capaces de valorar la eficacia o el beneficio de un nuevo tratamiento evaluado en un grupo de pacientes cuidadosamente seleccionado y tratados en las condiciones ideales que se dan en un estudio experimental. Es preciso realizarlos con cuidado cuando se intenta valorar la efectividad del tratamiento tal y como se usa en la práctica clínica. Por esta razón, los médicos motivados y concienzudos que proporcionan asistencia habitual con los equipos usuales a veces no obtienen los mismos resultados que en los ensayos clínicos controlados.

Los estudios de este tipo, en el mejor de los casos, solo son capaces de valorar el beneficio o el tratamiento bajo las condiciones actuales. Sin embargo, no es raro que la introducción de un nuevo tratamiento pueda por sí misma alterar las condiciones actuales y producir efectos secundarios o dinámicos. Los ensayos clínicos tienen una capacidad limitada para valorar los efectos secundarios del tratamiento. Esto es especialmente cierto para aquellos efectos que es más probable que aparezcan cuando el tratamiento se usa ampliamente en la práctica clínica. Imagine el estudio que figura a continuación.

En un ensayo clínico controlado se demostró la eficacia de un nuevo fármaco llamado "Herp-ex" para reducir la frecuencia de ataques en pacientes con herpes genital recurrente grave. Sin embargo, no curaba la infección. Los investigadores se impresionaron mucho con los resultados del estudio y recomendaron su uso en todas las personas con herpes genital.

Si Herp-ex se aprueba para uso clínico, podrían aparecer diversos efectos que no podrían haberse previsto a partir de los resultados del ensayo clínico controlado. Primero, lo más probable es que este medicamento se use ampliando las indicaciones del ensayo original. Es muy posible que se trate también con él a pacientes con ataques moderados o que presentan el primer episodio. La eficacia mostrada en los ataques recurrentes graves de herpes genital no significa que el fármaco sea efectivo para indicaciones distintas de las originales. Segundo, el amplio uso del Herp-ex puede producir cepas de herpes resistentes al fármaco. Finalmente, su uso extendido y su éxito a corto plazo pueden inducir a reducir las precauciones tomadas por los que padecen

herpes genital recurrente. De este modo, con el tiempo, el número de casos de herpes genital puede aumentar realmente a pesar de la eficacia a corto plazo del Herp-ex o debido a ella.

Los ensayos clínicos controlados son nuestra herramienta fundamental para valorar la eficacia de un tratamiento. Cuando se llevan a cabo cuidadosamente, sirven como base para realizar extrapolaciones sobre la efectividad de un tratamiento en la práctica clínica. No obstante, no están diseñados específicamente para valorar su seguridad. Antes de utilizar un tratamiento como parte de un ensayo clínico controlado, se realizan investigaciones en animales y de forma limitada en humanos para excluir sus efectos graves o frecuentes. Sin embargo, los efectos poco frecuentes o a largo plazo no se valoran bien antes de un ensayo clínico controlado o durante su realización.

La seguridad de un tratamiento es más difícil de valorar que su eficacia, especialmente cuando se trata de efectos secundarios poco frecuentes pero graves. La clave del problema radica en el elevado número de personas que necesitarían recibir el tratamiento antes de que se pueda observar este tipo de efectos.

El número de exposiciones necesarias para asegurar una probabilidad de 95% de observar al menos un episodio de un efecto secundario poco frecuente se resume en la *regla de tres*. Según esta regla, para tener una probabilidad de 95% de observar al menos un caso de reacción anafiláctica a la penicilina, que ocurre 1 vez cada 10 000, aproximadamente, se necesitarían 30 000 individuos. Si se desea tener una probabilidad de 95% de observar al menos un caso de anemia aplásica por cloramfenicol —que aparece 1 vez cada 50 000, aproximadamente—, necesitaríamos tratar a 150 000 pacientes. En general, la regla de tres afirma que para tener una confianza de 95% de observar al menos un efecto secundario poco frecuente se necesita tratar aproximadamente tres veces el número de individuos del denominador.[12]

Estas cifras demuestran que no se puede esperar que los ensayos clínicos controlados detecten muchos efectos secundarios poco frecuentes pero importantes. Para hacer frente a este dilema, con frecuencia nos basamos en pruebas realizadas en animales. En dichas pruebas, se administran altas dosis del fármaco a diversas especies animales suponiendo que sus efectos tóxicos, teratogénicos y carcinogénicos se observarán al menos una vez en dichos animales. Si bien este enfoque ha sido de gran ayuda, no ha solucionado definitivamente el problema.

Es aun más difícil detectar las consecuencias a largo plazo de tratamientos preventivos utilizados ampliamente. El dietilestilbestrol (DES) se usó durante muchos años para prevenir los abortos espontáneos. Pasaron décadas antes de que los investigadores notaran el gran aumento de la incidencia de carcinoma de vagina entre las adolescentes cuyas madres habían tomado DES.

Es solo en la práctica clínica que muchos pacientes pueden recibir el tratamiento y en ella es donde resulta más probable observar estos efectos secundarios. La actitud alerta de los clínicos y de los investigadores ha constituido el pilar de nuestra "vigilancia poscomercialización" actual. Hoy día no existe un enfoque sistemático y organizado para detectar efectos secundarios poco frecuentes pero graves después de la comercialización de un medicamento. La FDA debe confiar en los informes recibidos de los médicos en ejercicio. Por ello, los clínicos deben recordar que la aprobación de un fármaco por la FDA no debe considerarse sinónimo de que es totalmente seguro o incluso de que los riesgos están claramente definidos y comprendidos.

[12] Estas cifras suponen que la incidencia previa o espontánea de estos efectos secundarios es cero. Si estas enfermedades también tienen otras causas, el número necesario es aun mayor.

Los ensayos clínicos controlados son fundamentales en nuestro sistema actual de evaluación de la eficacia de los medicamentos y de los procedimientos. Representan un avance del máximo interés. Sin embargo, como clínicos que leemos la literatura médica debemos entender sus ventajas y sus limitaciones. Hemos de estar preparados para extraer conclusiones sobre la aplicación de resultados a nuestros propios pacientes y en nuestros propios contextos. Debemos reconocer que los ensayos clínicos controlados solo pueden proporcionar datos limitados sobre la seguridad y la efectividad del tratamiento investigado.

CAPÍTULO 12

EJERCICIOS PARA DETECTAR ERRORES: ENSAYOS CLÍNICOS CONTROLADOS

Los siguientes ejercicios están diseñados para poner a prueba su habilidad de aplicar los principios del ensayo clínico controlado. Lea cada ejercicio para detectar errores y luego escriba una crítica señalando los tipos de errores que aparecen en cada componente del marco uniforme.

SANGRE SEGURA. UN NUEVO TRATAMIENTO PARA PREVENIR EL SIDA: EJERCICIO NO. 1

Un investigador creyó que había descubierto un método mejor para evitar el peligro de diseminar el síndrome de la inmunodeficiencia adquirida (SIDA) mediante las transfusiones de sangre, matando el virus en las células transfundidas. Su método exigía tratar a todos los receptores de las transfusiones con un nuevo fármaco llamado "Sangre Segura". En el momento de su descubrimiento, la tasa de transmisión del SIDA debida a transfusiones era de 1 por 100 000 transfusiones.

Una vez conseguida la aprobación del estudio del fármaco en seres humanos, diseñó un ensayo clínico controlado para su uso inicial. En el estudio se preguntó a una muestra aleatoria de todos los receptores de transfusiones en una gran área metropolitana si deseaban recibir el fármaco en las dos semanas siguientes a la transfusión.

El grupo de estudio quedó constituido por 1 000 individuos que aceptaron el tratamiento, y el de control, por 1 000 individuos que lo rehusaron. Los integrantes del grupo de control habían recibido una media de tres transfusiones de sangre por cada 1,5 recibida por los tratados con Sangre Segura. Los investigadores lograron realizar pruebas serológicas para seguimiento del virus de la inmunodeficiencia humana (VIH) en 60% de los que recibieron el fármaco y en 60% de los que lo rechazaron, aproximadamente un mes después de recibir la transfusión.

Los que realizaron dichas pruebas de seguimiento no sabían cuáles pacientes habían recibido o no Sangre Segura. El investigador observó que un individuo del grupo de estudio pasó a ser positivo a la prueba de detección de anticuerpos contra el VIH al mes siguiente de iniciar el tratamiento con Sangre Segura. En el grupo de control dos individuos pasaron a ser positivos a la prueba.

El investigador no detectó ningún efecto indeseable atribuible a Sangre Segura durante el período de seguimiento de un mes. Concluyó, por lo tanto, qu el estudio había demostrado que el fármaco era efectivo y seguro, y aconsejó su administración a todos los receptores de transfusiones de sangre.

CRÍTICA: SANGRE SEGURA. UN NUEVO TRATAMIENTO PARA PREVENIR EL SIDA

Diseño del estudio

El investigador intentó realizar un ensayo clínico controlado. Este tipo de estudio es el más indicado para valorar la eficacia de un tratamiento, una vez

que la dosis y el método de administración se han determinado mediante estudios iniciales en seres humanos. Estos estudios no están indicados para realizar investigaciones iniciales en seres humanos.

En el momento del estudio, el riesgo de transmisión del SIDA a través de las transfusiones de sangre era de 1 por 100 000, un riesgo muy bajo. Ya que los ensayos clínicos controlados están destinados a reducir un riesgo ya bajo, es preciso reunir a un número muy elevado de individuos. Se necesitarían millares o incluso millones de individuos para realizar de forma apropiada un ensayo clínico controlado, cuando el riesgo, sin tratamiento, es de 1 por 100 000. Un estudio de esta magnitud no tiene un poder estadístico adecuado. En otras palabras, ese estudio no sería capaz de demostrar una significación estadística para el tratamiento, incluso suponiendo que Sangre Segura fuese capaz de reducir considerablemente la incidencia del SIDA asociado con las transfusiones sanguíneas, por ejemplo, de 1 por 100 000 a 1 por 1 000 000.

Asignación

El investigador identificó una muestra aleatoria de pacientes comparables con los que podrían recibir un tratamiento efectivo. La selección al azar no es un requisito de los ensayos clínicos controlados, pero aumenta la fiabilidad de la extrapolación a los miembros de la población de la que se extrajo la muestra y que no fueron incluidos en el ensayo.

El investigador no asignó al azar a los pacientes a los grupos de estudio y de control. El grupo de control estaba formado por pacientes que rechazaron la administración de Sangre Segura. Este no es un grupo de control ideal, porque los que rehusaron participar podían haber sido diferentes de los que aceptaron, en cuanto a diversos aspectos relacionados con la posibilidad de contraer la infección por el VIH. La asignación al azar, en contraposición a la selección al azar, se considera una característica fundamental de los ensayos clínicos controlados. Por consiguiente, este estudio no fue un verdadero ensayo clínico controlado.

Valoración

Los que valoraron el desenlace de este estudio no sabían cuáles pacientes habían recibido Sangre Segura. Esta valoración a ciegas es objetiva y contribuye a prevenir el sesgo en el proceso de valoración. Sin embargo, la ausencia de enmascaramiento en el proceso de la asignación significa que los pacientes sabían si habían recibido Sangre Segura o no. Esto pudo haber tenido un efecto sobre el desenlace del estudio suponiendo, por ejemplo, que quienes recibieron Sangre Segura creían que estaban protegidos contra el SIDA.

Los investigadores realizaron la prueba de detección de anticuerpos contra el VIH un mes después de que los pacientes recibieran una transfusión. Este período es demasiado corto para valorar adecuadamente si se ha producido o no en un individuo la conversión al estado positivo.

El elevado número de pacientes de los grupos de estudio y de control que se perdieron en el seguimiento constituyó un problema importante al efectuar la valoración, aunque los porcentajes de pérdidas fueron iguales en ambos grupos. Cuando el número de desenlaces adversos es bajo, los sujetos perdidos en el seguimiento son especialmente importantes, porque estos individuos pueden experimentar de forma desproporcionada efectos secundarios o manifestar síntomas a pesar del tratamiento.

Análisis

El investigador no informó sobre pruebas de significación estadística ni intervalos de confianza. Si lo hubiera hecho, no habría sido capaz de demostrar la existencia de una diferencia estadísticamente significativa. Esto no es sorprendente, ya que, con un solo caso más de infección por el VIH, los desenlaces hubieran sido iguales en los grupos de estudio y de control.

En este estudio, el intervalo de confianza hubiera sido muy amplio, indicando que los resultados eran compatibles con la ausencia de diferencias o incluso con una diferencia en la dirección opuesta.

El mayor número de transfusiones de sangre recibidas por los que rehusaron tomar Sangre Segura podría ser una variable de confusión que se debió tener en cuenta mediante un proceso de ajuste en el análisis. El número de transfusiones de sangre es una variable de confusión, dado que es diferente en los dos grupos y está relacionado con el riesgo de desarrollar infecciones por el VIH secundarias a las transfusiones.

Interpretación

Los problemas señalados en el diseño, la asignación, la valoración y el análisis indican que el estudio se debe interpretar con mucha cautela.

El resultado de las pruebas de significación estadística y de los intervalos de confianza implicaría que la diferencia de infecciones por VIH entre los grupos de estudio y de control podrían ser debidas al azar.

El riesgo de desarrollar una infección por el VIH a partir de una transfusión de sangre sin la administración de Sangre Segura es tan pequeño que sería mucho más probable contraer el virus de otra forma. Por lo tanto, ninguna diferencia entre el grupo de estudio y el de control puede atribuirse automáticamente a la administración de Sangre Segura. La diferencia de un grupo a otro en la infección por el VIH puede deberse a diferencias entre otros factores de riesgo del SIDA. No se presentan datos que permitan analizar esos factores, que quizá sean mucho más importantes que las transfusiones de sangre.

Extrapolación

Aunque se demostrara que Sangre Segura es eficaz para prevenir las infecciones por el VIH asociadas con la transfusión, este estudio no permitiría extraer conclusiones acerca de su efectividad o seguridad.

Los ensayos clínicos controlados pueden llevar a conclusiones sobre la eficacia de un tratamiento en las condiciones ideales de una investigación. En cambio, la efectividad implica que el tratamiento ha sido beneficioso en las condiciones habituales de la práctica clínica.

El empleo de Sangre Segura en el medio clínico implicaría su administración a un gran número de individuos. Por eso, los efectos secundarios graves serían importantes por muy raros que fueran. La ausencia de estos efectos entre los que recibieron Sangre Segura no anula la posibilidad de que aparezcan otros efectos poco frecuentes pero graves. De acuerdo con la regla de tres, si se produce un efecto secundario en 1 de cada 1 000 usos, se debe observar a 3 000 individuos para tener una seguridad de 95% de observar al menos un caso de dicho efecto.

Al extrapolar los resultados de un ensayo clínico controlado al uso de un tratamiento en la práctica clínica, es necesario considerar como mínimo lo siguiente:

1. Si el estudio demuestra la eficacia del tratamiento en condiciones ideales.
2. Si los individuos estudiados son similares a los que recibirán el tratamiento.
3. Si los riesgos conocidos o la posibilidad de efectos secundarios raros pero graves no observados en los ensayos clínicos controlados superan los beneficios potenciales.

A veces, también puede ser importante considerar el costo del tratamiento comparado con el de otras opciones.

EJERCICIO NO. 2: VACUNA DE LA GRIPE

Para probar una nueva vacuna contra la gripe, se realizó un ensayo clínico aleatorio. Los participantes del grupo de estudio se eligieron seleccionando al azar a 1 000 familias de una lista de familias de voluntarios para el ensayo. La vacuna se administró a los 4 000 integrantes de las 1 000 familias. Como grupo de control, los investigadores seleccionaron individuos al azar de la guía de teléfonos hasta que 4 000 personas aceptaron la propuesta de recibir una vacuna placebo. Al comparar los grupos, los investigadores observaron que la media de edad del grupo de estudio era de 22 años y la del de control, 35. La posibilidad de que los sujetos poseyeran un termómetro fue dos veces más alta en el grupo de estudio que en el de control, pero, por lo demás, los grupos eran similares cuando se compararon de acuerdo con una larga lista de variables.

Durante la época de gripe del invierno siguiente, se dieron instrucciones a cada individuo de que acudiese a uno de los médicos investigadores siempre que tuviera fiebre, para evaluar clínicamente la posibilidad de que padeciera gripe. Los sujetos del grupo de estudio fueron asignados a un consultorio y los del grupo de control, a otro. Al hacer su evaluación de seguimiento, los investigadores obtuvieron una participación de 95% en el grupo de estudio y de 70% en el de control. Además, observaron que en el grupo de control se produjeron 200 casos de gripe por cada 1 000 personas vacunadas con placebo y seguidas, y 4 casos por 1 000 personas vacunadas en el grupo de estudio. Por último, concluyeron que el nuevo tratamiento reducía el riesgo de gripe a 2% de su tasa previa y recomendaron su administración a la población general para evitar 98% de las muertes causadas por la gripe debida a esta cepa del virus.

CRÍTICA: EJERCICIO NO. 2

Diseño del estudio

Los investigadores no definieron la muestra de estudio que estaban empleando para someter a prueba el nuevo tratamiento. No queda claro si lo estaban probando en familias o en individuos, en sujetos de un grupo de edad o de todas las edades, en voluntarios o en la población general.

Asignación

En los ensayos clínicos aleatorios los investigadores asignan al azar, e idealmente a ciegas, voluntarios al grupo de control y al de estudio. En el caso que nos ocupa, todos los voluntarios fueron asignados al grupo de estudio, se les administró el

tratamiento experimental y se extrajo una muestra separada para constituir el grupo de control. Por lo tanto, la asignación no fue al azar ni a ciegas.

Al vacunar a todos los miembros de una familia se abre la posibilidad de que un tratamiento con éxito parcial parezca casi totalmente efectivo. Si se reduce el riesgo de un miembro concreto de la familia, el riesgo de exposición de los demás miembros de la familia se reduce en gran parte, dado que la exposición familiar es una fuente de transmisión importante. De esta forma, dos factores favorecen la vacunación: la reducción de la exposición y el incremento de la inmunidad.

Los investigadores emplearon una población diferente para seleccionar el grupo de control. Al seleccionar los nombres de la guía telefónica limitaron su muestra a los individuos que aparecían en ella. Con este método no se formó un grupo de control equiparable al de los que recibieron el tratamiento, especialmente en lo que se refiere a los niños. Es probable que estos estén ampliamente representados en una población de familias, pero no en la guía de teléfonos.

Existe la posibilidad de que la distribución de características potencialmente relacionadas con la susceptibilidad a la gripe no sea similar en los grupos de estudio y de control, aunque se obtenga una muestra al azar de la población. Es necesario tener en cuenta las diferencias de edad entre los grupos de estudio y de control (que son de esperar, dado el método de asignación). La diferencia en la frecuencia de posesión de termómetros, aunque sea debida al azar, es importante, porque puede influir en el reconocimiento de la fiebre. Un mayor reconocimiento de la fiebre probablemente resulte en un aumento del número de casos de gripe diagnosticados. Recuerde que la selección al azar de los participantes es un rasgo deseable pero no habitual en un ensayo clínico aleatorio. Sin embargo, la asignación al azar es esencial, porque es la característica distintiva de un ensayo clínico aleatorio.

Valoración

Es probable que el método escogido por los investigadores para diagnosticar la gripe no fuera válido. La gripe es difícil de diagnosticar específicamente y en el estudio no se definieron los criterios estándares para su diagnóstico, como el cultivo de los virus. También, es probable que en los voluntarios el umbral de enfermedad antes de acudir al médico fuera distinto del de los no voluntarios; este hecho, aunado al bajo número de sujetos del grupo de control que poseían termómetros, pudo haber influido en el número de diagnósticos de gripe realizados. El hecho de que los sujetos de los grupos de estudio y de control fueron seguidos en consultorios diferentes sugiere que no se efectuó el enmascaramiento y ello pudo haber influido en la frecuencia de diagnósticos de gripe.

Finalmente, si el seguimiento no incluye a todos los participantes, existe la posibilidad de que los sujetos perdidos en el seguimiento tengan un mejor o peor desenlace que los que pudieron ser seguidos. Una alta proporción de integrantes del grupo de control se perdieron en el seguimiento, lo cual pudo haber influido en la validez de la valoración del desenlace. En general, la valoración del desenlace en este estudio no fue válida.

Análisis

Como la distribución de la edad en el grupo de estudio era diferente de la del grupo de control y la edad es un factor que influye frecuentemente en la susceptibilidad a la infección, es importante ajustar los datos según el efecto de la edad.

Esto se podría haber realizado comparando la tasa de ataque de los individuos de la misma edad para ver si las diferencias entre los vacunados y los no vacunados seguían siendo las mismas. Además, no se aplicó ninguna prueba de significación estadística para determinar cuán probables eran las diferencias observadas, si no existían verdaderas diferencias en las poblaciones de las que se extrajeron las muestras.

Interpretación y extrapolación

Este estudio no consigue cumplir con los estándares mínimos de la asignación, la valoración y el análisis, lo que significa que se debe interpretar con suma cautela. Hay muchas razones, distintas de la de relación causal, que pueden explicar el bajo número de diagnósticos de gripe realizados en el grupo de estudio.

La extrapolación a otras poblaciones requiere una prueba convincente de la existencia de una relación entre los sujetos estudiados. En este estudio faltó la prueba de una relación causal. Aunque se demuestre la existencia de una relación, no se pueden efectuar extrapolaciones de las tasas de ataque de la gripe a las tasas de mortalidad. El estudio no tenía como objetivo estimar la tasa de mortalidad y no proporcionó ninguna prueba de que las tasas de mortalidad de las dos poblaciones fuesen diferentes.

A pesar de que la investigación presenta numerosos problemas de diseño y de que los investigadores extrapolaron mucho más allá de sus datos, es importante darse cuenta de la magnitud del efecto en el grupo de estudio. Una reducción del número de casos de gripe de 98% constituye un efecto muy notable. Este hallazgo exige un análisis riguroso, a pesar de la mala calidad del diseño de la investigación realizada. Si bien es importante, muchas veces es difícil separar la calidad del tratamiento de la calidad de la investigación.

Sección 2

La prueba de una prueba

CAPÍTULO 13

INTRODUCCIÓN A LA PRUEBA DE UNA PRUEBA

El diagnóstico médico puede contemplarse como un intento de tomar las decisiones idóneas manejando información insuficiente. Así, la incertidumbre intrínseca al diagnóstico médico procede de la necesidad de realizar diagnósticos basados en datos inciertos. Los instrumentos diagnósticos empleados en medicina se han considerado tradicionalmente como un medio de reducir la incertidumbre en el diagnóstico. Sin embargo, para utilizar con éxito las pruebas diagnósticas, se debe saber valorar no solo la forma cómo las pruebas reducen la incertidumbre sino también cómo describen y cuantifican la incertidumbre restante.

En épocas pasadas, los instrumentos diagnósticos estaban limitados en gran parte a la historia clínica y al examen físico. Actualmente, estos todavía son potentes instrumentos de diagnóstico. Sin embargo, hoy día se dispone, además de los métodos convencionales, de una tecnología auxiliar con la cual el clínico cuidadoso puede realizar diagnósticos precisos, si la emplea apropiada y selectivamente. La esencia de la práctica de la medicina diagnóstica está constituida por el aprendizaje de cuándo, si es conveniente, debe aplicarse cada elemento de la historia, del examen físico y de la tecnología auxiliar.

El énfasis actual en la calidad con consciencia del costo requiere que los médicos entiendan los principios fundamentales de las pruebas diagnósticas: cuáles son las preguntas que pueden responder y cuáles no, cuáles son las pruebas que aumentan la precisión diagnóstica y cuáles simplemente incrementan el costo.

Para parafrasear a Will Rogers, los médicos tradicionalmente han pensado que nada es cierto, excepto la biopsia y la autopsia. Sin embargo, incluso estos criterios de referencia para el diagnóstico (*gold standards*) pueden errar el objetivo o realizarse demasiado tarde para ser de ayuda alguna. El conocimiento de los principios de las pruebas diagnósticas contribuye a definir el grado de incertidumbre diagnóstica y a aumentar la certeza. Saber cómo vivir con la incertidumbre es una característica central del juicio clínico. El médico habilidoso ha aprendido cúando debe asumir riesgos para aumentar la certeza y cuándo debe, simplemente, tolerar la incertidumbre.

El principio fundamental de las pruebas diagnósticas reside en la creencia de que los individuos que tienen una enfermedad son distintos de los que no la tienen y que las pruebas diagnósticas permiten distinguir a los dos grupos. Las pruebas diagnósticas, para ser perfectas, requerirían que 1) todos los individuos sin la enfermedad en estudio tuvieran un valor uniforme en la prueba, 2) que todos los individuos con la enfermedad tuvieran un valor uniforme pero distinto en la prueba y 3) que todos los resultados de las pruebas fueran consistentes con los resultados del grupo de los enfermos y del de los sanos (figura 13-1).

Si esta fuera la situación en el mundo real, la prueba perfecta podría distinguir la enfermedad de la salud, y el trabajo del médico consistiría únicamente en solicitar la prueba "adecuada". El mundo real, sea para bien o para mal, no es tan simple. Habitualmente, ninguna de estas condiciones está presente. Existen variaciones en cada uno de los siguientes factores básicos: las pruebas, el grupo de enfermos y el de sanos (figura 13-2).

FIGURA 13-1. Condiciones necesarias de una prueba diagnóstica perfecta

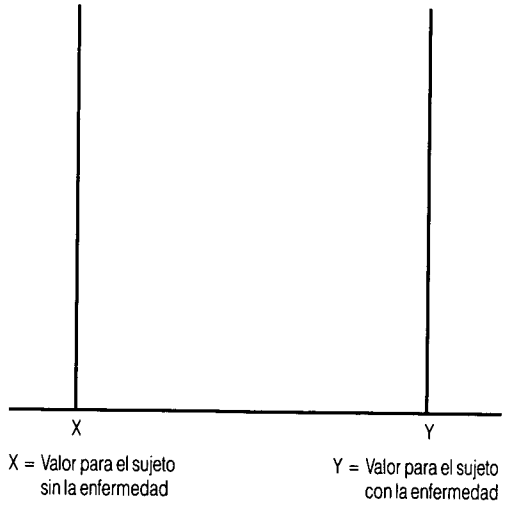
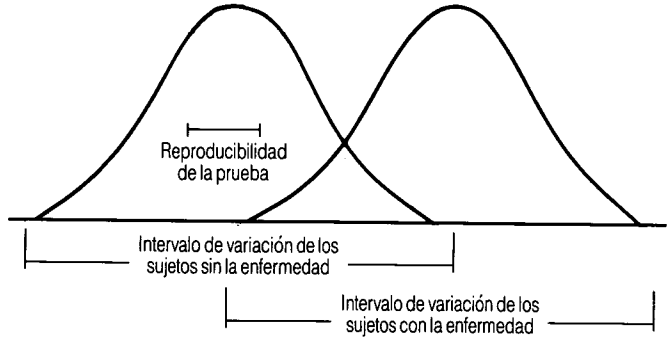


FIGURA 13-2. Los tres tipos de variaciones que afectan a las pruebas diagnósticas

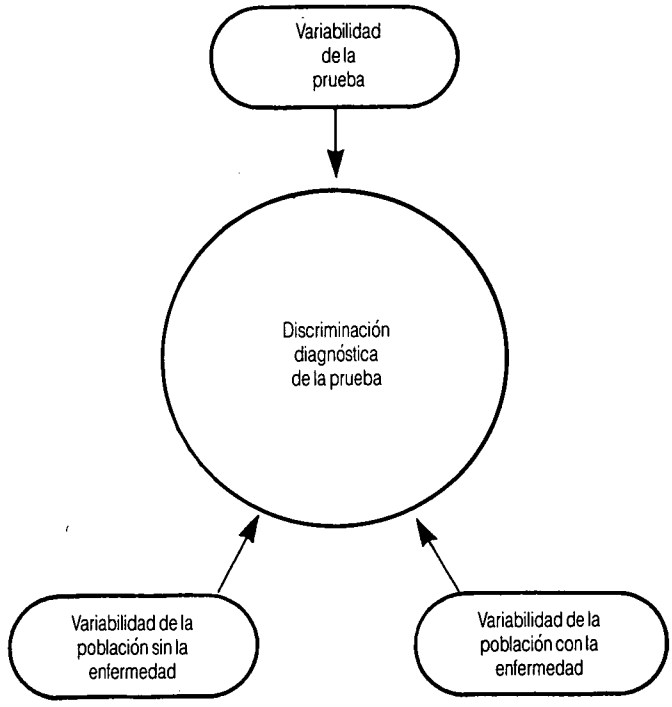


La valoración de las pruebas diagnósticas consiste en gran medida en describir la variabilidad de estos tres factores, y por esa razón se pueden cuantificar las conclusiones, a pesar o a causa de esa variabilidad.

La variabilidad, la reproducibilidad y la exactitud de las pruebas se presentan en el capítulo 14. En el capítulo 15 se revisa y valora la variabilidad de la población de las personas sanas empleando el concepto del intervalo de la normalidad. En los capítulos 16 y 17 se cuantifica la variabilidad de la población de personas enfermas y su relación con la de las sanas por medio de los conceptos de sensibilidad, especificidad y valor predictivo. En esos capítulos se esbozan estos conceptos y se muestran ejemplos de los errores que se cometen al aplicarlos. Seguidamente, se incluyen varios ejercicios para detectar errores, que ofrecen la oportunidad de aplicar esos principios a la evaluación de las pruebas diagnósticas.

Al igual que con el análisis de un estudio, es útil tener una visión panorámica o marco general de la evaluación de una prueba. Este marco se ilustra en la figura 13-3, en la que se representa la variabilidad que existe en las pruebas, en la población sana y en la enferma. También se subraya que estas variaciones deben ser estudiadas e incorporadas en cualquier valoración de la utilidad diagnóstica de una prueba.

FIGURA 13-3. Marco uniforme de la prueba de una prueba



CAPÍTULO 14

VARIABILIDAD DE UNA PRUEBA

Una prueba perfecta produciría los mismos resultados cada vez que se aplicara en las mismas condiciones. Además, sus mediciones reflejarían exactamente el fenómeno que la prueba intenta medir. En otras palabras, una prueba perfecta sería completamente reproducible y exacta. Definamos estos términos y veamos cómo se utilizan.

La *reproducibilidad* es la capacidad de una prueba para producir resultados consistentes cuando se repite en las mismas condiciones y se interpreta sin conocer sus resultados previos. Sin embargo, en la reproducibilidad de una prueba pueden influir diversos factores.

1. Las condiciones del paciente y del laboratorio bajo las que se realiza la prueba pueden no ser las mismas.

2. La prueba puede estar influida por variaciones de interpretación entre observadores. Este efecto se conoce como *variabilidad interobservador*.

3. La prueba puede verse afectada por variaciones en la interpretación que realiza la misma persona en diferentes momentos. Este efecto se conoce como *variabilidad intraobservador*.

Para valorar el rendimiento de la prueba, los investigadores deben estar seguros de que las condiciones técnicas y biológicas de su realización son idénticas cuando se repite. El no seguir esta precaución produce el error que se muestra en el siguiente ejemplo.

Con el fin de evaluar la reproducibilidad de una prueba para medir la concentración sérica de cortisol se extrajeron dos muestras de sangre de los mismos individuos. La primera muestra se extrajo a las 8 de la mañana y la segunda, al mediodía. Los métodos fueron idénticos y la interpretación se realizó sin conocimiento previo de los resultados de la primera prueba. Este método se aplicó a 100 individuos seleccionados al azar. Los autores observaron que los valores de la segunda prueba eran en promedio el doble de la primera prueba aplicada al mismo individuo. Concluyeron que esta gran variación indicaba que la prueba no era reproducible.

Recuerde que la reproducibilidad es la capacidad de la prueba para producir casi los mismos resultados cuando se realiza en las mismas condiciones. En este ejemplo, los investigadores no repitieron la prueba en las mismas condiciones. La concentración de cortisol tiene un ciclo natural durante el día. Al extraer una muestra de sangre a las 8 de la mañana y otra al mediodía, estaban obteniendo muestras de dos momentos distintos del ciclo. Aunque la prueba fuera perfectamente reproducible cuando se realizase en condiciones de laboratorio idénticas, las distintas situaciones de los pacientes a los que se practicó hubieran producido resultados bastante diferentes.

A menos que una prueba se repita sin conocer el resultado de la primera, la segunda lectura puede estar influida por la primera, como ilustra el siguiente ejemplo.

Un investigador, al estudiar la reproducibilidad de un análisis de orina, solicitó a un técnico de laboratorio experimentado que leyera un sedimento urinario y, sin cambiar de lugar el portaobjetos, lo volviera a leer a los cinco minutos. El investigador observó que la lectura, realizada en las mismas condiciones, produjo resultados perfectamente reproducibles.

En este ejemplo, el técnico conocía los resultados de la primera prueba y era probable que estuviese influido por la primera lectura cuando la repitió a los cinco minutos. Una medida de la reproducibilidad requiere que la segunda lectura se efectúe sin conocimiento previo de la primera. Por esta razón, el técnico no debió haber conocido los resultados de la lectura anterior.

Aunque los observadores no sean conscientes de sus propias lecturas previas o de las de otros, existe la posibilidad de que la variación individual induzca a error cuando se comparan los resultados de la prueba. Siempre que hay que formar un juicio en la interpretación de una prueba, existe la posibilidad de variación inter e intraobservador. Es frecuente que dos radiólogos interpreten la misma placa de rayos X de diferente manera, lo cual se conoce como variación interobservador. Un residente puede interpretar el mismo electrocardiograma de forma diferente por la mañana que en medio de una guardia nocturna. Esto se conoce como variación intraobservador. Por sí mismas, estas variaciones no arruinan la utilidad de la prueba. Sin embargo, es necesario que el médico esté siempre alerta a la posibilidad constante de variaciones en la interpretación de los resultados de la prueba. Las inconsistencias de la técnica o de la interpretación contribuyen en cierta medida a la variabilidad de la mayor parte de las pruebas. Por lo tanto, es preciso aplicar criterios que permitan juzgar cuánta variabilidad puede tolerarse.

En general, es importante que la variabilidad de la prueba sea mucho menor que el intervalo de variabilidad debido a factores biológicos. Por esta razón, el grado de variación de la prueba debe ser pequeño en comparación con el intervalo de normalidad de la prueba (véase el capítulo 15).

Es importante distinguir la reproducibilidad de la exactitud. La *exactitud* de una prueba es la capacidad que tiene de producir resultados cercanos a la verdadera medida del fenómeno anatómico, fisiológico o bioquímico. La exactitud de una prueba requiere que esta sea reproducible y que el resultado no muestre una tendencia sistemática a diferir del verdadero valor en una dirección determinada. Cuando disparamos a un blanco, a veces erramos el tiro porque las balas se dispersan alrededor del centro. También puede haber una tendencia a situar todos los tiros a un solo lado ligeramente apartado del centro. Para ser perfectamente exactos y dar cada vez en el centro del blanco, debe existir reproducibilidad, eliminando de esta forma la dispersión. No debe existir un sesgo o una tendencia a disparar siempre hacia un lado. Por lo tanto, una prueba exacta está exenta de los efectos del azar y de errores sistemáticos o sesgos. Una prueba puede ser muy reproducible y a la vez inexacta, si reproduce valores alejados del valor verdadero. El siguiente caso ejemplifica una prueba con alta reproducibilidad pero poca exactitud, o sea, la diferencia entre reproducibilidad y exactitud.

Se realizó un estudio de 100 pacientes que habían sido operados por fracturas naviculares. De cada paciente, se obtuvieron dos radiografías que se interpretaron de forma independiente. Ambas fueron negativas para una fractura navicular en la semana posterior a la lesión. Los autores concluyeron que los radiólogos habían sido negligentes al no diagnosticar esas fracturas.

Por lo general, las fracturas naviculares no se diagnostican por medio de radiografías tomadas en el momento de la lesión. Desde el punto de vista anatómico,

la fractura existe, pero no suele detectarse en la radiografía hasta que no aparecen signos de reparación. Por consiguiente, no se trata de una negligencia de los radiólogos, sino de la inexactitud de la prueba utilizada. La repetición de las radiografías confirmó la reproducibilidad de los resultados negativos obtenidos por los dos radiólogos: no es posible identificar fracturas naviculares recientes en las radiografías. *Exactitud* significa que la prueba producirá resultados semejantes al verdadero valor anatómico, fisiológico o bioquímico. Dado que las radiografías no son siempre un reflejo exacto de la anatomía, el defecto radicaba en la prueba, no en los radiólogos.

Una prueba puede ser bastante exacta cuando se emplea en un estudio científico, pero puede perder su exactitud cuando se aplica en el medio clínico. Es útil pensar en dos tipos de exactitud: 1) exactitud experimental, es decir, la exactitud de la prueba cuando se utiliza en las condiciones especiales de un estudio, y 2) exactitud clínica, es decir, la exactitud de la prueba cuando se emplea en las situaciones clínicas reales. La diferencia entre ambos conceptos se muestra en el siguiente ejemplo.

En un hospital universitario se realizó un estudio de 500 pacientes que siguieron una dieta baja en grasas durante 3 días. Se mostró que el contenido de grasas en las heces recogidas 72 horas después de finalizar la dieta permitía distinguir bien entre los pacientes con malabsorción y los que no la padecían. Con un protocolo de estudio idéntico aplicado a 500 pacientes ambulatorios no se consiguió demostrar con éxito la presencia de malabsorción. Los autores del estudio ambulatorio concluyeron que los resultados obtenidos con los pacientes hospitalizados eran incorrectos.

El rendimiento de una prueba se valora habitualmente en condiciones experimentales ideales, pero las condiciones reales en las que se aplica muchas veces están lejos de ser ideales. En pacientes ambulatorios puede ser bastante difícil recoger las heces a las 72 horas después de una dieta de tres días de duración baja en grasas. El hecho de que los resultados de los datos de pacientes ambulatorios no concuerden con los de los pacientes hospitalizados puede reflejar simplemente la realidad de las condiciones en la práctica. La exactitud es una propiedad necesaria de una buena prueba. No obstante, la exactitud, por sí sola, no garantiza que la prueba sea válida o útil para el diagnóstico. La *validez* implica que la prueba es una medida apropiada del fenómeno estudiado. Una medida muy reproducible y exacta del tamaño de los pulmones puede proporcionar poca información válida o útil para un diagnóstico. Para establecer la utilidad diagnóstica de una prueba, necesitamos valorar la idoneidad con que la prueba distingue entre las personas sanas y las enfermas.

Antes de tratar de determinar la idoneidad de una prueba para distinguir los que no tienen una enfermedad de los que la tienen, centraremos nuestra atención en analizar cómo medimos la ausencia de enfermedad por medio del concepto del intervalo de lo normal.

CAPÍTULO 15

EL INTERVALO DE LO NORMAL

Las poblaciones humanas sanas están sujetas a variaciones biológicas intrínsecas. Uno solo necesita pasear por la calle para apreciar las diferencias entre la gente. La altura, el peso y el color de los individuos cubren un espacio que refleja las variaciones grandes, pero no ilimitadas, que pueden existir entre individuos sanos.

En un mundo con información completa sabríamos cuál es el resultado que debe tener un individuo en una prueba determinada. Esto nos permitiría comparar el resultado obtenido en la prueba con el resultado esperado en esa persona. En la realidad, como raramente sabemos cuál debería ser el resultado en un sujeto, estamos obligados a comparar sus resultados con los de otros individuos considerados sanos. Para llevar a cabo esta comparación utilizamos un *intervalo de lo normal* o de la normalidad. El intervalo de lo normal es un mal necesario basado en la suposición de que un individuo concreto debe ser similar a otros individuos.

El concepto del intervalo de la normalidad representa un esfuerzo para medir y cuantificar el intervalo de valores que existen en individuos considerados sanos. Se puede derivar un intervalo de lo normal de cualquier medición en la que existan múltiples posibles valores numéricos para los sujetos sanos. Estos comprenden exámenes de características físicas tales como la tensión arterial, el tamaño del hígado y el pulso o valores de laboratorio como el hematócrito, la velocidad de sedimentación o la creatinina. Aunque el intervalo de medidas normales suele ser amplio, el concepto no incluye a todas las personas que no están enfermas. Este intervalo excluye a propósito a 5% de los individuos considerados sanos, con el fin de crear un intervalo de lo normal suficientemente amplio para describir la mayor parte de las mediciones de las personas sanas, pero no tan amplio como para incluir todos los posibles valores numéricos. Si el intervalo de la normalidad incluyera las mediciones de todos los individuos sin la enfermedad, sería extremadamente amplio, tan amplio que no sería útil para separar a los enfermos de los sanos. El intervalo de lo normal es descriptivo y no diagnóstico; describe a los individuos sanos, no diagnostica la enfermedad. Los valores que se encuentran fuera de ese intervalo podrían ser el resultado de la variación debida al azar, de cambios fisiológicos no asociados con la enfermedad o de cambios patológicos secundarios a la enfermedad.

CONSTRUCCIÓN DEL INTERVALO DE LO NORMAL

Los valores del intervalo de lo normal se pueden construir de la forma que se detalla a continuación:

1. El investigador localiza a un grupo de individuos que se considera que no están enfermos. Este grupo se conoce como *grupo de referencia*. Estos individuos muchas veces son estudiantes de medicina, trabajadores de hospital u otros voluntarios fácilmente accesibles. En general, simplemente se supone que no están enfermos, aunque en algunas circunstancias pueden llevarse a cabo diversas pruebas y un seguimiento para garantizarlo.

2. El investigador realiza la prueba de interés en todos los individuos del grupo de referencia.

3. Seguidamente, representa gráficamente la distribución de los valores obtenidos mediante la prueba aplicada a los individuos del grupo de referencia.

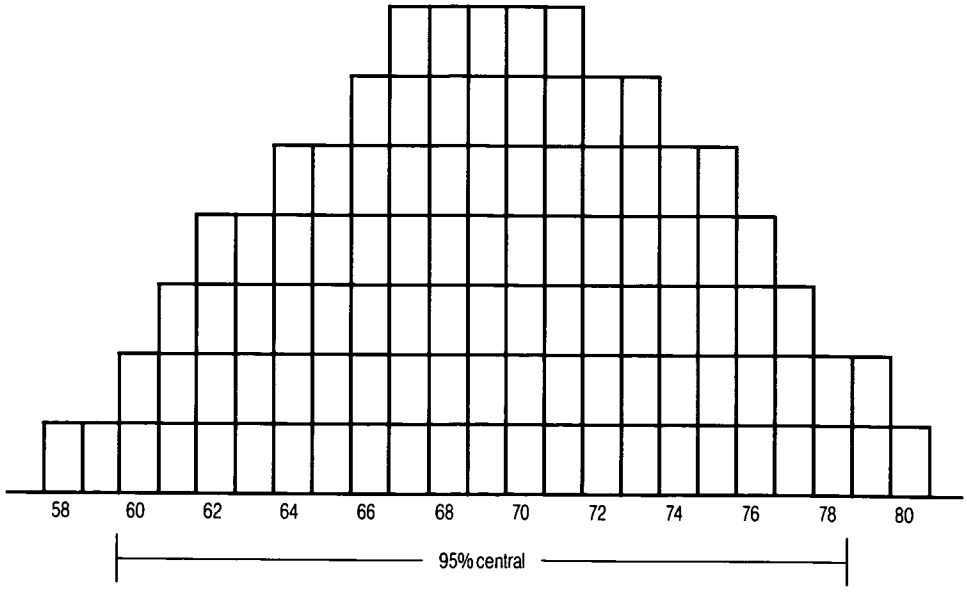
4. Luego calcula el intervalo de lo normal, que comprende 95% de los valores centrales de la población de referencia. En sentido estricto, el intervalo de la normalidad incluye la media más y menos las mediciones incluidas en dos desviaciones estándar de la media. Si no existe alguna razón para hacerlo de otra forma, el investigador generalmente escoge la parte central del intervalo, de forma que 2,5% de los individuos sanos tengan mediciones mayores y 2,5% de los individuos sanos tengan mediciones menores de los valores del intervalo de la normalidad.

Para ilustrar este método, imagine que los investigadores miden la estatura de 100 estudiantes varones de una facultad de medicina y encuentran valores semejantes a los que aparecen en la figura 15-1. Los investigadores escogerían a continuación un intervalo de la normalidad que incluyera 95 de los 100 estudiantes. Si no tuviesen una razón para hacerlo de otro modo, utilizarían la parte central del intervalo, de forma que el intervalo de la normalidad de este grupo de referencia estaría comprendido entre 60 y 78 pulgadas [152 y 198 cm]. Los individuos que quedaran fuera del intervalo no tendrían necesariamente que tener ninguna enfermedad, simplemente podrían ser individuos sanos excluidos del intervalo de la normalidad.

PRINCIPIOS BÁSICOS

En primer lugar, veamos las implicaciones de los principios del intervalo de lo normal y después los errores que pueden resultar si no se comprenden estas implicaciones.

FIGURA 15-1. Estaturas de 100 estudiantes de medicina varones, utilizadas para construir un intervalo de valores normales



1. Por definición, en cualquier prueba determinada aplicada a un grupo, 5% de los resultados se encontrarán fuera del intervalo de lo normal. Por esta razón, "anormal" y "enfermo" no son sinónimos. Cuantas más pruebas se realicen, más individuos se encontrarán fuera del intervalo de la normalidad, por lo menos en una prueba. Llevando esta proposición a su límite, se puede concluir que una persona "normal" es aquella que no ha sido investigada suficientemente. A pesar de lo absurdo de esta proposición, destaca la importancia de comprender que la definición del intervalo de la normalidad sitúa a propósito a 5% de los individuos que no están enfermos fuera de dicho intervalo. Por este motivo, el término *fuera de los límites de lo normal* no debe entenderse como sinónimo de enfermedad.

2. Los valores que se encuentran en el intervalo de la normalidad no garantizan que los individuos no estén enfermos. La capacidad del intervalo de la normalidad de una prueba para discriminar entre los sanos y los enfermos varía de una prueba a otra. A menos que la prueba sea perfecta para descartar la enfermedad —y pocas pruebas lo son—, las mediciones de algunos individuos que tienen la enfermedad se encontrarán dentro de los límites de la normalidad.

3. Los cambios incluidos en los límites normales pueden ser patológicos. Dado que el intervalo de lo normal incluye un amplio intervalo de valores numéricos, las mediciones de un individuo pueden cambiar considerablemente y todavía encontrarse dentro de los límites de la normalidad. Por ejemplo, el intervalo de la normalidad para la enzima hepática AST varía de 8 a 20 U/L, el del potasio sérico puede variar entre 3,5 y 5,4 mEq/L, y el del ácido úrico, desde 2,5 a 8,0 mg por 100 ml. Es importante considerar no solo si los valores de un individuo se hallan dentro de los límites normales, sino también si han cambiado con el tiempo. El concepto del intervalo de lo normal es más útil cuando no se dispone de datos anteriores para comparar pacientes individuales. Sin embargo, cuando se dispone de esos datos, se deben tener en cuenta.

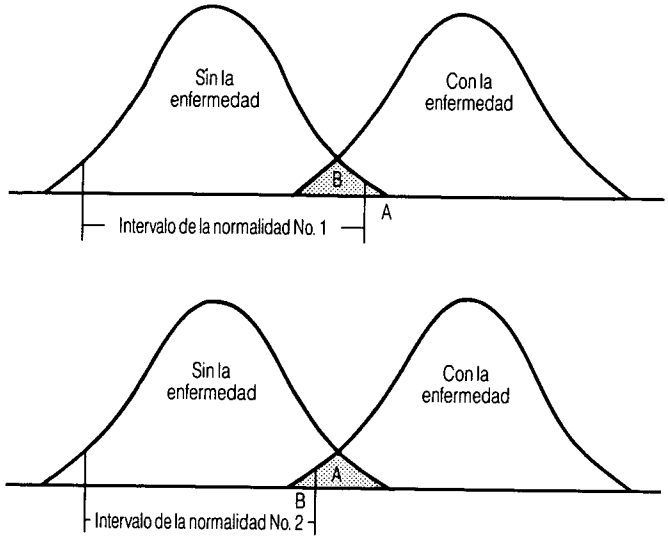
4. El intervalo de lo normal se calcula empleando un grupo concreto de pacientes o una población de referencia considerados sanos. Por consiguiente, cuando se aplica un intervalo concreto a un individuo, se debe averiguar si ese individuo presenta alguna característica que lo diferencia de la población de referencia utilizada para construir el intervalo de la normalidad. Por ejemplo, si para obtener el intervalo de la normalidad del hematócrito se utilizan hombres, este no puede aplicarse necesariamente a las mujeres, que en general tienen hematócritos más bajos.

5. El intervalo de la normalidad no debe confundirse con el intervalo deseable. El intervalo de la normalidad es una medida empírica de cómo son las cosas en un grupo de individuos que se consideran sanos en ese momento. Es posible que amplios sectores de la comunidad tengan resultados de las pruebas más elevados que los ideales y estén predispuestos a desarrollar una enfermedad en el futuro.

6. Los límites superior e inferior del intervalo de lo normal pueden modificarse con fines diagnósticos. El intervalo incluye 95% de los que no presentan una enfermedad o estado concreto. Sin embargo, no es necesario que haya el mismo número de individuos sanos con valores de una prueba por debajo y por encima del intervalo de la normalidad. Existe cierto margen de criterio científico para determinar dónde deben situarse los límites superior e inferior de dicho intervalo. La demarcación de los límites depende del objetivo que persigan el investigador o el clínico al aplicar la prueba.[1] Por

[1] El nivel del intervalo de lo normal fijado definirá la especificidad de la prueba. Este nivel se puede ajustar posteriormente para aumentar (o disminuir) la especificidad. Por esta razón, en la práctica, el intervalo de lo normal siempre se ha ajustado cuando la especificidad no es de 95%.

FIGURA 15-2. Modificaciones del intervalo de lo normal. Positivos falsos, A: individuos que no tienen la enfermedad, con valores por encima del intervalo de la normalidad. Negativos falsos, B: individuos que tienen la enfermedad, con valores dentro del intervalo de la normalidad



ejemplo, suponga que la mayoría de los individuos con una enfermedad tienen niveles cercanos al límite superior del intervalo de lo normal en la prueba. Si el investigador está dispuesto a reducir este límite, se puede prever que un mayor número de individuos que tienen la enfermedad tendrán resultados por encima del intervalo de la normalidad. En este caso, el investigador también paga el precio (o acepta el intercambio) de colocar a una mayor proporción de la población sana fuera del intervalo de la normalidad. A veces merece la pena pagar ese precio, sobre todo cuando es importante detectar el mayor número posible de individuos con la enfermedad o cuando las pruebas de seguimiento para clarificar la situación son baratas y prácticas. La figura 15-2 muestra este intercambio.

Al trasladar el intervalo de la normalidad hacia la izquierda, como en el intervalo de normalidad No. 2, observe que el área B se reduce respecto a la del intervalo No. 1 y que el área A aumenta. En otras palabras, una disminución del número de negativos falsos conduce a un aumento de los positivos falsos y viceversa. En esta situación, la prueba identifica como fuera de lo normal a un mayor número de personas con la enfermedad. Al mismo tiempo, clasifica fuera de los límites normales a más individuos sin la enfermedad; por lo tanto, en el intervalo de la normalidad No. 2 estamos aceptando más lecturas positivas falsas a cambio de menos negativas falsas. El número de positivos y negativos falsos que el médico o el sistema sanitario esté dispuesto a tolerar depende de consideraciones éticas, económicas y políticas, así como del conocimiento médico.

Los siguientes ejemplos ilustran los errores que se pueden cometer al aplicar incorrectamente estos principios.

1. En 1 000 exámenes de salud consecutivos de tipo preventivo se realizaron 12 pruebas de laboratorio (SMA-12) en cada paciente, aunque no se encontraron anormalidades en la historia médica o en la exploración física. El 5% de las SMA-12 estaban

fuera del límite de la normalidad; es decir, se obtuvo un total de 600 pruebas "anormales". Los autores concluyeron que los resultados obtenidos justificaban la realización de las SMA-12 en todos los exámenes de salud rutinarios.

Veamos el significado de estos resultados. Los valores del intervalo de la normalidad incluyen, por definición, solo 95% de todos aquellos que se consideran exentos de la enfermedad. Si aplicamos esta prueba a 1 000 individuos sin la enfermedad, 5% ó 50 individuos tendrán un resultado fuera del intervalo de la normalidad. Si se aplicaran 12 pruebas a 1 000 individuos sin síntomas ni signos de enfermedad, en promedio, los resultados de 5% de las 12 000 pruebas realizadas se encontrarían fuera del intervalo de lo normal. El 5% de 12 000 pruebas es 600 pruebas. Por eso, aunque los 1 000 individuos no tuvieran ninguna enfermedad, se podría prever que 600 pruebas darían resultados fuera del intervalo de la normalidad. Estos reflejarían simplemente el método de calcular el intervalo de la normalidad. Estos resultados no indican necesariamente la presencia de enfermedad y por sí mismos no justifican el realizar múltiples pruebas de laboratorio en todos los exámenes de salud rutinarios.

Al considerar las implicaciones de los resultados de las pruebas, es importante darse cuenta de que no todos los valores fuera del intervalo de lo normal tienen el mismo significado. Es mucho más probable que los valores bastante alejados del límite sean causados por enfermedad que los valores cercanos a dicho límite y asimismo, que los resultados cercanos a los límites del intervalo se deban a la variabilidad de la prueba o a la variabilidad biológica. Por ejemplo, si el límite superior del hematócrito en un hombre es 52, es más probable que el valor 60 esté asociado con una enfermedad que el valor 53.

2. Se midió la concentración de AST [aspartasa aminotransferasa] en 100 alcohólicos, para valorar su función hepática. En la mayoría, los resultados se encontraban dentro del intervalo de la normalidad. Los autores concluyeron que el hígado de estos alcohólicos funcionaba bien.

Este ejemplo ilustra la diferencia entre el intervalo de la normalidad de las pruebas de laboratorio y el no tener la enfermedad. El hecho de que los resultados de las pruebas de laboratorio de esas personas estuvieran dentro de los límites de la normalidad no es suficiente por sí solo para establecer que su hígado funciona perfectamente, dado que en cualquier prueba algunos resultados correspondientes a individuos enfermos se encontrarán dentro de los límites de lo normal. Cuanto menor sea la capacidad de la prueba para diagnosticar la enfermedad, más elevado será el número de individuos enfermos cuyos resultados se encuentran dentro de los límites de lo normal. Puede que ciertas pruebas no permitan distinguir a las personas enfermas de las sanas. Es posible que en ambos grupos la mayor parte de los resultados estén dentro del intervalo de la normalidad. Esto sucedió con los resultados de la AST. La incapacidad de la prueba para discernir entre enfermos y sanos indica que su capacidad de discriminación diagnóstica es baja y que no es útil para el diagnóstico. Subraya, así, la diferencia entre encontrarse dentro del intervalo de la normalidad y no tener la enfermedad.

Las figuras 15-3 a 15-5 muestran las tres posibles relaciones entre la población sana y la población enferma. La figura 15-3 ilustra una prueba que separa completamente a los que tienen la enfermedad de los que no la tienen. La discriminación diagnóstica de esta prueba es perfecta. La figura 15-4 representa la situación usual, de una prueba que separa parcialmente a los que tienen la enfermedad de los que no la tienen. La figura 15-5 muestra el ejemplo de una prueba que no tiene discriminación diagnóstica. En el caso de la AST, la situación se parece mucho a la de la figura 15-5. A pesar de su utilidad en el diagnóstico de muchas enfermedades hepáticas, la medición

FIGURA 15-3. Prueba que separa completamente los resultados de las poblaciones (con discriminación diagnóstica perfecta)

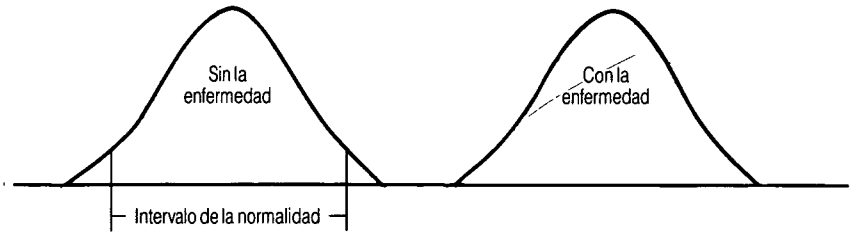


FIGURA 15-4. Prueba que separa parcialmente los resultados de las poblaciones (con discriminación diagnóstica parcial)

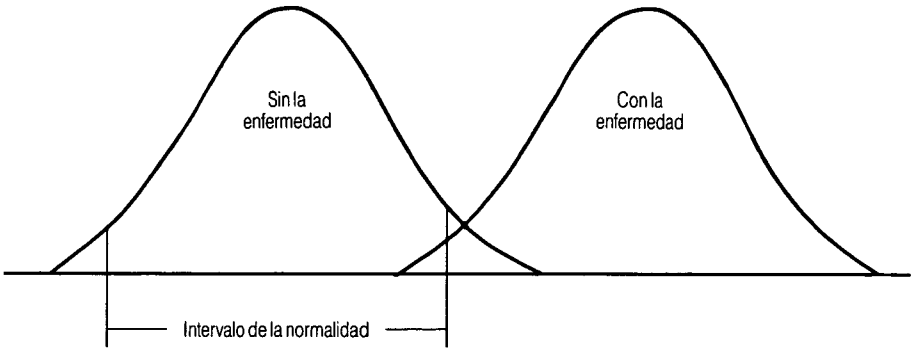
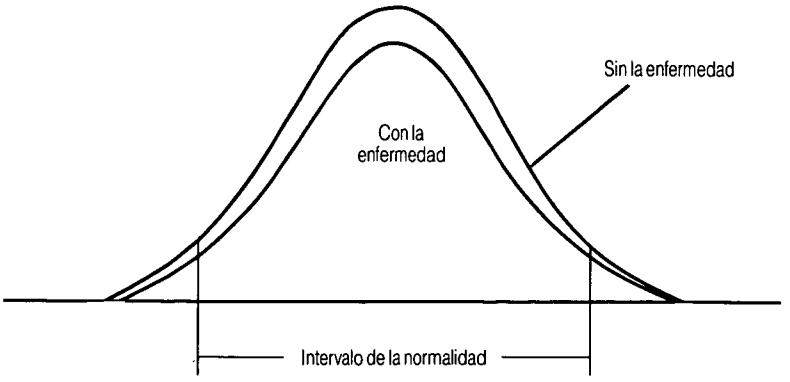


FIGURA 15-5. Prueba que no separa los resultados de las poblaciones (sin discriminación diagnóstica)



de la concentración de AST no es útil para diagnosticar el efecto crónico del alcohol en el hígado. Por eso, a pesar de que se puede calcular el intervalo de la normalidad para cualquier prueba, este intervalo, por sí solo, no indica si la prueba será útil para el diagnóstico. Las mediciones en los individuos con una enfermedad concreta pueden ser idénticas a las de los que no la tienen y viceversa, lo cual indica que la prueba no tiene utilidad diagnóstica para esa enfermedad concreta.

3. Se calculó que el intervalo de la normalidad de la creatinina sérica en 1 000 estadounidenses asintomáticos sin enfermedad renal conocida era de 0,7 a 1,4 mg/dl. Una mujer de 70 años de edad ingresó en el hospital con una concentración de creatinina de 0,8 mg/dl y fue tratada con gentamicina. Cuando fue dada de alta, ese valor era de 1,3 mg/dl. Su médico llegó a la conclusión de que, como la concentración de creatinina se encontraba dentro de los límites de la normalidad, tanto al ingreso como al darla de alta, su paciente no tenía una lesión renal secundaria al tratamiento con gentamicina.

La presencia de un resultado dentro de los límites de la normalidad no garantiza la ausencia de enfermedad. En cada individuo, la medida que indica que no está enfermo puede estar por encima o por debajo de la medida promedio de los demás individuos sin la enfermedad. En este ejemplo, la concentración de creatinina de la paciente aumentó en 60%, si bien todavía se encontraba dentro del intervalo de la normalidad. Dicho cambio sugiere la presencia de un nuevo proceso patológico. Es probable que la gentamicina le haya producido una lesión renal. Cuando se dispone de información previa, es importante considerarla al evaluar el resultado de una prueba. Los cambios, aun dentro del intervalo de la normalidad, *pueden* ser un signo de enfermedad.

4. Se utilizó a un grupo de 100 estudiantes de medicina para calcular el intervalo de valores de la normalidad del recuento de granulocitos. Se escogió un intervalo de forma que incluyera 95 de los 100 recuentos de granulocitos. Los límites del intervalo de la normalidad calculado fueron 2 000 y 5 000. Cuando se preguntó a los autores sobre el recuento de 1 900 en un anciano de raza negra, llegaron a la conclusión de que este se encontraba claramente fuera del intervalo de lo normal y que era preciso realizar más estudios para identificar la causa de ese resultado.

El intervalo de la normalidad depende de la población de referencia sin la enfermedad que se ha seleccionado; esta se define como el intervalo alrededor de un valor promedio que incluye 95% de los individuos de una población de referencia determinada. Sin embargo, la población de referencia sin la enfermedad utilizada para calcular el intervalo de la normalidad puede tener mediciones diferentes de las del grupo de personas en las que queremos usar la prueba.

Es improbable que haya muchos ancianos de raza negra entre el grupo de estudiantes de medicina utilizados para construir el intervalo de la normalidad. De hecho, los hombres de raza negra tienen un intervalo de valores del recuento de granulocitos distinto de los de raza blanca. Por este motivo, el intervalo de lo normal calculado con los estudiantes de medicina puede no reflejar el intervalo de la normalidad aplicable a los ancianos de raza negra. El recuento de granulocitos de este hombre se encontraba probablemente dentro del intervalo de normalidad para su raza, edad y sexo. Como se sabe que los ancianos de raza negra tienen su propio intervalo de la normalidad, este hecho se debe tener en cuenta cuando se interpreten los resultados de la prueba.

5. El intervalo de la normalidad de la concentración de colesterol sérico medida en 100 hombres estadounidenses de raza blanca de 30 a 60 años de edad osciló entre 200 y 300 mg/dl. La concentración de colesterol de un estadounidense de raza blanca de 45

años de edad fue 280 mg/dl. Su médico le dijo que no tenía que preocuparse por el colesterol elevado, dado que su concentración se hallaba dentro de los límites de lo normal.

El intervalo de la normalidad se calcula utilizando los datos recogidos en una población de referencia que en ese momento se considera sana. Es posible que el grupo utilizado esté formado por individuos cuyos resultados en la prueba sean más elevados (o más bajos) que los deseables. Un resultado dentro del límite de la normalidad no garantiza que el individuo se mantendrá sano. Es posible que los hombres estadounidenses, considerados en conjunto, tengan concentraciones de colesterol por encima de los niveles deseables. Si esto es cierto, el paciente con una concentración de colesterol de 280 mg/dl puede muy bien sufrir las consecuencias de la hipercolesterolemia. Cuando existen datos basados en la investigación que sugieren claramente un intervalo deseable de valores para una prueba, es aceptable sustituir el intervalo habitual por el intervalo de la normalidad deseable. Esto se está haciendo cada vez más con el colesterol sérico.

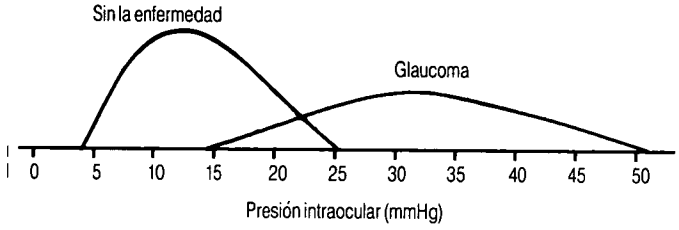
6. En un estudio se demostró que 90% de los que tienen una tensión intraocular mayor de 25 mmHg desarrollarán defectos visuales secundarios al glaucoma en los 10 años siguientes. El 20% de los que tienen tensiones intraoculares de 20 mmHg desarrollarán cambios similares y 1% de los que tienen tensiones intraoculares de 15 mmHg desarrollarán trastornos. Los autores concluyeron que el rendimiento de la prueba se podría mejorar disminuyendo el límite superior de la normalidad de 25 a 15 mmHg, dado que así la prueba podría identificar prácticamente a todos los que se encontraran en riesgo de desarrollar defectos visuales (figura 15-6).

Si el límite superior de la normalidad es de 25 mmHg, casi todo el mundo que no desarrollará glaucoma estará dentro de los límites de la normalidad, aunque un número elevado de los que desarrollarán glaucoma también estarán incluidos en los límites de la normalidad. Por otro lado, si el límite de lo normal se sitúa en 15 mmHg, muy pocos individuos con glaucoma estarán dentro de los límites de la normalidad y un alto número de los que nunca tendrán glaucoma estarán fuera del intervalo de la normalidad.

La capacidad de una prueba para detectar la enfermedad se puede aumentar modificando los límites del intervalo de la normalidad. Si los límites se amplían suficientemente, la prueba incluirá prácticamente a todos los que padezcan la enfermedad. Lamentablemente, esta atractiva solución también sitúa fuera del intervalo de la normalidad a un mayor número de individuos que no tienen ni tendrán la enfermedad. Al aumentar los límites superiores de lo normal, los investigadores incrementan la capacidad de detectar enfermedades futuras, pero solo pagando el precio de seguir a muchos individuos que no tendrán la enfermedad. Al determinar dónde situar los límites superiores, se pueden considerar los factores que se enumeran a continuación.

1. La pérdida de visión por el glaucoma es irreversible en gran parte y puede desarrollarse antes de que el paciente se dé cuenta.
2. El tratamiento suele ser seguro, pero solo parcialmente efectivo para prevenir la pérdida progresiva de la visión.
3. El seguimiento es seguro y acarrea un riesgo bajo, pero el seguimiento de un número elevado de individuos requiere mucho tiempo y es costoso, ya que exige realizar múltiples exámenes repetidos durante largos períodos de tiempo. Además, el seguimiento produce ansiedad en los pacientes.

FIGURA 15-6. Posible distribución de la tensión intraocular de los sujetos con y sin glaucoma



Los factores que hay que sopesar no son todos médicos. Otras consideraciones sociales, psicológicas, económicas o políticas se pueden tomar en cuenta para establecer la línea de demarcación. Es posible que no exista una respuesta correcta. La única salida para esta situación insoluble es que se invente una prueba mejor, en la cual se superponga menos a los que desarrollarán la enfermedad y a los que no lo harán.

El concepto del intervalo de lo normal es un intento para tratar con la variabilidad que existe entre las personas. La comprensión de la utilidad y las limitaciones de este concepto es de capital importancia para comprender las pruebas diagnósticas. El intervalo de la normalidad define los valores numéricos encontrados en 95% de los sujetos considerados sanos de un grupo concreto de referencia. Este intervalo puede no reflejar el nivel deseable y no tiene en cuenta los cambios que se producen respecto de los resultados de pruebas anteriores.

El intervalo de la normalidad *per se* no nos dice nada sobre la utilidad diagnóstica de la prueba. Cada prueba tiene un intervalo de la normalidad que puede o no ayudar a discernir entre los individuos que tienen la enfermedad y los que no la tienen. Para determinar la utilidad de una prueba en el diagnóstico de una enfermedad es necesario examinar los resultados de la prueba en un grupo de individuos con una determinada enfermedad y comparar estos valores con el intervalo de la normalidad de un grupo sin la enfermedad, como haremos en el capítulo 17. Antes de que podamos hacerlo, es preciso que examinemos cómo se puede definir a los individuos que tienen la enfermedad.

CAPÍTULO 16

DEFINICIÓN DE ENFERMEDAD: LA PRUEBA DE ORO

Cuando se aplica cualquier prueba diagnóstica, ya sea a personas que padecen una enfermedad o a las que no la padecen, los resultados representan un recorrido de valores. En los enfermos, la variabilidad de los resultados puede reflejar diferencias en la gravedad de la enfermedad o una respuesta individual a la misma. A pesar de esta variabilidad, es esencial definir un grupo de pacientes que, sin lugar a dudas, padecen la enfermedad.

LA PRUEBA DE ORO

La prueba o criterio utilizado para definir inequívocamente una enfermedad se conoce como *prueba de oro*[1] (*gold standard*). La prueba de oro puede ser una biopsia, un angiograma, una necropsia posterior o cualquier otra prueba reconocida. El uso de un criterio de oro con el fin de identificar definitivamente a los que tienen la enfermedad es un requisito para examinar la utilidad diagnóstica de cualquier prueba nueva o no evaluada. En otras palabras, la utilidad de la nueva prueba se basa en su comparación con la de oro. De este modo, una prueba nueva se compara con una prueba (o pruebas) antigua y más aceptada para determinar si la nueva ofrece el mismo rendimiento que la de referencia. Observe que se parte del supuesto de que, utilizando la mejor de las pruebas antiguas, es posible tener un 100% de posibilidades de realizar un diagnóstico correcto; la suposición de partida es la imposibilidad de "inventar una mejor trampa para ratones", dado que no se puede superar el 100%. Puede existir una trampa para ratones más barata o más práctica pero, por definición, ninguna con la que se atrapen más ratones.

Puede parecer equívoco afirmar que la única forma de evaluar la capacidad diagnóstica de una prueba nueva es suponer que ya es posible realizar diagnósticos perfectos. Lamentablemente, esa es la posición en que nos encontramos al evaluar una prueba nueva. Solo podemos preguntarnos si la prueba está a la altura de la mejor de las pruebas antiguas, esto es, la prueba de oro.

A pesar de la limitación intrínseca de nuestra capacidad para evaluar inicialmente una prueba nueva, el tiempo y las aplicaciones repetidas están del lado de la mejor trampa para ratones. Una vez que se aplica a la práctica clínica, puede hacerse evidente que, en realidad, la prueba nueva predice mejor el curso clínico subsiguiente que la de referencia. Incluso es posible que con el tiempo la prueba nueva sea aceptada como prueba de oro. No obstante, el problema que surge con frecuencia es que, si bien se puede realizar el diagnóstico definitivo, la prueba acarrea un riesgo excesivo o se realiza demasiado tarde para rendir sus máximos beneficios clínicos. Es decir, existe una prueba de oro adecuada que no es práctica en el sentido clínico. En estos casos, es útil comprobar que la prueba nueva está a la altura de la de oro. Debe entenderse, repetimos, que el objetivo de evaluar una prueba se limita a compararla con la mejor prueba disponible. Por esta razón, es preciso estar seguro de que se está utili-

[1] N. del E. Se traducirá como prueba de oro, criterio de oro, y prueba o criterio de referencia, según el contexto.

zando la mejor prueba de oro disponible. Examinemos, a modo de ejemplo, lo que puede suceder cuando la prueba de oro utilizada no es la más adecuada.

Se practicó la autopsia a 100 individuos que fueron ingresados en un hospital con "ondas Q diagnósticas" en su electrocardiograma (ECG) y que fallecieron en la hora siguiente al ingreso, con objeto de determinar si habían sufrido infarto de miocardio (IM). La necropsia, que se utilizó como criterio de oro del IM, reveló pruebas de IM en solo 10 sujetos. Los autores concluyeron que el ECG no era un método útil para realizar el diagnóstico de IM e insistieron en aceptar el diagnóstico anatomopatológico como la prueba de oro.

La utilidad de todas las pruebas diagnósticas se determina comparándolas con pruebas de referencia cuya aptitud para medir las características estudiadas ya se ha establecido con la práctica. Los diagnósticos por necropsia se utilizan frecuentemente como criterio de oro contra el cual se juzgan las otras pruebas. Sin embargo, la necropsia no siempre constituye una forma ideal de medir la enfermedad, como muestra este ejemplo, dado que a veces debe pasar bastante tiempo antes de que se manifiesten los signos patológicos del IM. Es posible que las ondas Q diagnósticas de un ECG reflejen mejor el IM que los cambios patológicos observables en una necropsia. El investigador debe cerciorarse de que el criterio de oro utilizado ha sido realmente establecido como la mejor referencia posible, antes de usarlo como base de comparación.

Por desgracia, incluso las mejores pruebas de referencia disponibles muchas veces no distinguen inequívocamente a los enfermos de los sanos. Puede que los casos de enfermedades leves o en sus fases iniciales no satisfagan los criterios de la prueba de oro. A menudo, los investigadores están tentados de incluir solamente a aquellos individuos que presentan pruebas claras de la enfermedad, tal como se miden con la prueba de referencia. A pesar de la certeza intelectual que parece proporcionar, esto puede redundar en una investigación que se limita a los individuos que tienen una enfermedad grave o en fase muy avanzada. Este peligro se ilustra con el siguiente ejemplo.

Un investigador comparó la capacidad de la citología de la orina para diagnosticar el cáncer de vejiga urinaria con la del diagnóstico inequívoco por biopsia de casos de cáncer invasor de vejiga que cumplían los criterios diagnósticos de la prueba de oro. Mediante el examen citológico se identificó a 95% de las personas que tenían cáncer. Sin embargo, cuando se aplicó en la práctica clínica, la citología de orina solo detectó 10% de los casos.

Al considerar solo los casos avanzados de cáncer invasor de vejiga urinaria, los investigadores habían eliminado los casos dudosos o en etapas iniciales de la enfermedad. Por lo tanto, no debe sorprender que, al aplicar la prueba en la práctica clínica, su rendimiento no fuera tan bueno como el obtenido cuando se comparó con una prueba de oro definitiva.

Por muy tentador que sea estudiar tan solo a los individuos con enfermedades claramente definidas, es engañoso sacar conclusiones sobre la utilidad de una prueba que se ha aplicado exclusivamente a individuos con una enfermedad avanzada o grave. Cuando se valora la discriminación diagnóstica de una prueba, es importante preguntarse si se utilizó el mejor criterio de referencia para definir a las personas con la enfermedad. También es importante preguntarse si con los enfermos estudiados se abarcó todo el espectro de la enfermedad. Debemos reconocer que a veces es imposible lograr ambos objetivos simultáneamente.

Aunque se cumplan estas condiciones, es preciso apreciar que el propósito de probar una prueba se limita a determinar si la prueba estudiada es tan buena como la prueba de referencia establecida. Los métodos empleados no contemplan la posibilidad de que la prueba nueva sea mejor que la de oro.

CAPÍTULO 17

DISCRIMINACIÓN DIAGNÓSTICA DE LAS PRUEBAS

Hoy día es posible medir la capacidad de una prueba para discriminar entre los enfermos y los sanos. Al hacer esa valoración, es importante considerar los tres puntos siguientes:

1. Variabilidad de la prueba: medición de la reproducibilidad del resultado de la prueba. El intervalo de variabilidad debe ser relativamente menor que el intervalo de la normalidad.
2. Variabilidad de la población sana: determinación de los valores del intervalo de la normalidad para la prueba.
3. Definición de la prueba de oro: identificación de los grupos de individuos que definitivamente tienen la enfermedad y de los que no la tienen según la prueba de oro.

SENSIBILIDAD Y ESPECIFICIDAD

Las medidas tradicionales del valor diagnóstico de una prueba son la *sensibilidad* y la *especificidad*. Estas miden la discriminación diagnóstica de la prueba comparada con la del criterio de referencia, que, por definición, tiene una sensibilidad y una especificidad de 100%. La sensibilidad y la especificidad se han seleccionado como medidas, porque son características intrínsecas de una prueba que deben ser idénticas, ya sea que se aplique a un grupo de pacientes en los cuales la enfermedad es rara o a un grupo de pacientes en los que es frecuente.[1] Por esta razón, proporcionan medidas de la discriminación diagnóstica de una prueba, que deben ser las mismas sea cual fuere la probabilidad de enfermedad antes de realizar la prueba. La estabilidad de la sensibilidad y la especificidad permite a los investigadores de Los Ángeles, París o Tokio aplicar la misma prueba diagnóstica y esperar resultados similares a pesar de las diferencias importantes que existen entre las poblaciones. Estas medidas también permiten a los investigadores y a los clínicos comparar directamente el rendimiento de una prueba con el de otras.

La sensibilidad mide la proporción de los individuos con la enfermedad que son identificados correctamente por la prueba. En otras palabras, mide lo sensible que es la prueba para detectar la enfermedad. Puede ser útil recordar la sensibilidad como *positiva en los enfermos* (PEE). La especificidad mide la proporción de los individuos sanos que son correctamente identificados como tales por la prueba. La especificidad se puede recordar como *negativa en los sanos* (NES).

Observe que la sensibilidad y la especificidad solamente indican la proporción o porcentaje de los que han sido correctamente clasificados como sanos o como enfermos. Estas medidas no predicen el número real de individuos que serán clasificados correctamente, cifra que dependerá de la frecuencia de la enfermedad en el grupo al que se aplique la prueba.

[1] Es posible que esto no sea estrictamente cierto, si la proporción de enfermos en estadios iniciales de la enfermedad cambia junto con la frecuencia de la enfermedad. Una prueba puede tener sensibilidad y especificidad distintas para una fase inicial de la enfermedad y para una avanzada.

La sensibilidad y la especificidad son medidas útiles, porque permiten a los lectores y a los investigadores obtener los mismos resultados cuando evalúan una prueba en grupos de pacientes que difieren en la frecuencia de la enfermedad. Sin embargo, los valores numéricos pueden ser diferentes según que se obtengan de un grupo de pacientes en los estadios iniciales de la enfermedad o de otros en estadios avanzados.

Primero mostraremos la forma de calcular la sensibilidad y la especificidad y luego sus implicaciones y limitaciones. Para calcular la sensibilidad y la especificidad de una prueba en comparación con la de oro, se siguen los siguientes pasos:

1. Los investigadores seleccionan una prueba de oro que se usará para identificar los individuos enfermos.

2. Seguidamente, escogen a un grupo de pacientes que según el criterio de referencia padecen la enfermedad y a otro grupo de individuos que según el mismo criterio están sanos. Al aplicar este criterio, es importante saber si los investigadores incluyeron a grupos representativos de individuos con y sin la enfermedad. En otras palabras, ¿representan los individuos seleccionados el espectro completo de los que tienen la enfermedad y de los que no la tienen o representan únicamente los dos extremos del espectro? Una práctica habitual en la selección de estos individuos es la de escoger tantos sujetos sanos como enfermos, definidos según el criterio de referencia. Sin embargo, esta división a medias no es necesaria.[2]

3. Los investigadores deben usar la prueba investigada para clasificar a todos los individuos como positivos o negativos. Para las pruebas cuyos resultados se presentan en valores numéricos, es preciso disponer de un intervalo de la normalidad. Por ejemplo, si la mayoría de los individuos con la enfermedad presentan valores por encima del intervalo de la normalidad, los investigadores usan el límite superior del intervalo de la normalidad como límite de demarcación. A continuación, aplican la nueva prueba a todos los individuos y los clasifican como positivos o negativos.

4. Los investigadores ya han clasificado a cada paciente como sano o enfermo, de acuerdo con la prueba de oro, y como positivo o negativo, según el resultado de la prueba. Ahora, ya pueden calcular el número de individuos en los que la prueba estudiada y la de oro concuerdan y en los que discrepan, y presentar los resultados de la siguiente manera:

PRUEBA EN ESTUDIO	PRUEBA DE ORO ENFERMOS	PRUEBA DE ORO SANOS
Positivos	a = Número de individuos enfermos y positivos	b = Número de individuos sanos y positivos
Negativos	c = Número de individuos enfermos y negativos	d = Número de individuos sanos y negativos
	a + c = Total de individuos enfermos	b + d = Total de individuos sanos

[2] La división a medias proporciona el mayor poder estadístico para un tamaño muestral determinado. Sin embargo, difícilmente veremos aplicar pruebas de significación estadística para valorar pruebas diagnósticas, dado que el tamaño de la muestra generalmente es pequeño y, por esa razón, el poder estadístico suele ser bajo.

5. Finalmente, los investigadores aplican las definiciones de sensibilidad y de especificidad, y calculan directamente sus valores.

$$\text{Sensibilidad} = \frac{a}{a + c} = \begin{array}{l}\text{Proporción de individuos con la enfermedad}\\ \text{según la prueba de oro e identificados}\\ \text{como positivos por la prueba en estudio.}\end{array}$$

$$\text{Especificidad} = \frac{d}{b + d} = \begin{array}{l}\text{Proporción de individuos sanos según la prueba}\\ \text{de oro e identificados como negativos}\\ \text{por la prueba en estudio.}\end{array}$$

Para ilustrar este método numéricamente, imaginemos que se aplica una nueva prueba a 500 individuos que tienen la enfermedad de acuerdo con el criterio de referencia y a 500 individuos que están sanos según el mismo criterio. Podemos construir la tabla de 2 × 2 como sigue:

PRUEBA EN ESTUDIO	PRUEBA DE ORO ENFERMOS	PRUEBA DE ORO SANOS
Positivos	a	b
Negativos	c	d
	500	500

Vamos a suponer que con la nueva prueba 400 de los 500 individuos con la enfermedad son identificados como positivos y que 450 de los 500 individuos sanos son identificados como negativos. Ya estamos en condiciones de rellenar la tabla de 2 × 2:

PRUEBA EN ESTUDIO	PRUEBA DE ORO ENFERMOS	PRUEBA DE ORO SANOS
Positivos	400	50
Negativos	100	450
	500	500

Ahora se pueden calcular la sensibilidad y la especificidad.

$$\text{Sensibilidad} = \frac{a}{a + c} = \frac{400}{500} = 0{,}80 = 80\%$$

$$\text{Especificidad} = \frac{d}{b + d} = \frac{450}{500} = 0{,}90 = 90\%$$

Una sensibilidad de 80% y una especificidad de 90% describen una prueba diagnóstica que, aunque no es ideal, tiene la misma calidad que muchas pruebas que se usan en la medicina clínica para diagnosticar enfermedades.

Observe que la prueba se ha aplicado a un grupo de pacientes, de los cuales 500 tienen la enfermedad y 500 están sanos, según el criterio de referencia. Esta división a medias entre sanos y enfermos es la que se emplea habitualmente al realizar estudios de este tipo. Sin embargo, la sensibilidad y la especificidad habrían sido las mismas independientemente del número de pacientes enfermos y sanos escogidos. Una forma de convencerse de la autenticidad de este importante principio es observar cómo se calculan la sensibilidad y la especificidad, esto es,

$$\text{Sensibilidad} = \frac{a}{a + c} \text{ y especificidad} = \frac{d}{b + d}$$

Observe que a y c —que son necesarios para calcular la sensibilidad— se encuentran en la columna de la izquierda de la tabla. De la misma manera, b y d —que son necesarios para calcular la especificidad— se encuentran en la columna de la derecha de la tabla. De esta forma, el número total de individuos en cada columna no es importante realmente, dado que la sensibilidad y la especificidad se relacionan, respectivamente, solo con la división de los pacientes que se encuentran en una simple columna.

Una vez que se han calculado la sensibilidad y la especificidad, es posible volver atrás y completar la tabla cuando se trabaja con distintos números de individuos enfermos y de sanos definidos según la prueba de oro. Esta vez vamos a suponer que hay 900 individuos sanos y 100 enfermos. En otras palabras, nos encontramos en una situación en la cual 10% de los individuos a los que se aplica la prueba tienen la enfermedad. Por lo tanto, el individuo promedio tiene una probabilidad de 10% de padecer la enfermedad antes de que se realice la prueba.

PRUEBA EN ESTUDIO	PRUEBA DE ORO ENFERMOS	PRUEBA DE ORO SANOS
Positivos	a	b
Negativos	c	d
	100	900

Apliquemos ahora las medidas de la sensibilidad y la especificidad, tal y como hemos hecho previamente.

La sensibilidad es igual a 80%; por lo tanto, 80% de los que tienen la enfermedad serán correctamente identificados como positivos (80% de 100 = 80), y 20% de los que tienen la enfermedad serán incorrectamente identificados como negativos (20% de 100 = 20).

La especificidad es igual a 90%; por consiguiente, 90% de los que no tienen la enfermedad serán correctamente identificados como negativos (90% de 900 = 810), y 10% de los que no tienen la enfermedad serán incorrectamente identificados como positivos (10% de 900 = 90).

Ahora podemos construir la siguiente tabla de 2 × 2.

PRUEBA EN ESTUDIO	PREVALENCIA DE 10% PRUEBA DE ORO ENFERMOS	PRUEBA DE ORO SANOS
Positivos	80	90 positivos falsos
Negativos	20 negativos falsos	810
	100	900

En esta situación, 10% de los pacientes estudiados tienen la enfermedad, según la prueba de oro; por lo tanto, podemos afirmar que, en este grupo de pacientes, la verdadera probabilidad de tener la enfermedad es de 10%.

Comparemos esta tabla con la que construimos al calcular por primera vez la sensibilidad y la especificidad. En realidad, utilizamos un grupo de pacientes cuya probabilidad de tener la enfermedad era de 50%, dado que trabajábamos con 500 individuos enfermos y 500 sanos.

PRUEBA EN ESTUDIO	PREVALENCIA DE 50% PRUEBA DE ORO ENFERMOS	PRUEBA DE ORO SANOS
Positivos	400	50 positivos falsos
Negativos	100 negativos falsos	450
	500	500

Observe que con nuestra división inicial a medias (esto es, con una prevalencia de 50%) se identificaron erróneamente 100 individuos como negativos y 50 como positivos. Sin embargo, en el grupo de pacientes en los que la prevalencia de la enfermedad era de 10%, se identificaron incorrectamente 20 individuos como negativos y 90, también erróneamente, como positivos. El cambio en las cifras se debe únicamente a la diferencia de la frecuencia relativa de la enfermedad o prevalencia en los dos grupos de pacientes estudiados (50% *versus* 10%). Observe que en el ejemplo en que se utilizó una prevalencia de 10% había realmente más positivos que estaban sanos (90) que positivos enfermos (80).

Esto puede sorprender, habida cuenta de que la sensibilidad y la especificidad son relativamente altas. Sin embargo, ilustra un principio que debe conocerse para aplicar los conceptos de sensibilidad y especificidad. A pesar de que la sensibilidad y la especificidad no están influidas directamente por la frecuencia relativa o prevalencia de la enfermedad, el número real de individuos que se identifican erróneamente como positivos o como negativos depende de la frecuencia relativa de la enfermedad.

Ahora analizaremos una situación más extrema, en la cual solo 1% de los integrantes del grupo estudiado tienen la enfermedad. Esta situación es la que

aparece típicamente cuando se realiza el tamizaje de un grupo de individuos que están expuestos a factores de riesgo de una enfermedad común, pero que no tienen signos clínicos. La tabla correspondiente podría parecerse a la siguiente:

PREVALENCIA DE 1%

PRUEBA EN ESTUDIO	PRUEBA DE ORO ENFERMOS	PRUEBA DE ORO SANOS
Positivos	8	99 positivos falsos
Negativos	2 negativos falsos	891
	10	990

En esta situación hemos utilizado de nuevo la misma prueba que tiene una sensibilidad de 80% y una especificidad de 90%. Esta vez encontramos 8 positivos verdaderos y 99 positivos falsos o, dicho de otra forma, 12 positivos falsos por cada positivo verdadero. Por esta razón, la sensibilidad y la especificidad por sí solas no proporcionan indicación suficiente de la utilidad de un resultado para el diagnóstico de una enfermedad en un individuo concreto. Como clínicos y usuarios de una prueba diagnóstica, necesitamos saber algo más que la sensibilidad y la especificidad de la prueba. Hemos de ser capaces de formular preguntas clínicas tales como ¿cuál es la probabilidad de que haya enfermedad si el resultado de la prueba es positivo?; ¿cuál es la probabilidad de que no haya enfermedad si el resultado es negativo? Antes de que podamos responder a estas preguntas, hemos de preguntarnos ¿cuál es la probabilidad de que el paciente tenga la enfermedad antes de realizar la prueba? Esta *probabilidad anterior a la prueba*, junto con la sensibilidad y la especificidad, nos permite calcular la medida denominada *valor predictivo de la prueba*.

VALOR PREDICTIVO DE LAS PRUEBAS POSITIVAS Y NEGATIVAS

Como hemos comentado anteriormente, la principal ventaja que ofrecen la sensibilidad y la especificidad en la valoración de una prueba es que no dependen directamente de la prevalencia o de la probabilidad de la enfermedad anterior a la prueba. Esta ventaja es especialmente útil para los artículos de la literatura médica. Sin embargo, también tienen limitaciones para responder a dos preguntas importantes desde el punto de vista clínico: si la prueba es positiva, ¿cuál es la probabilidad de que el individuo tenga la enfermedad?; si la prueba es negativa, ¿cuál es la probabilidad de que no la padezca? Estas preguntas tienen una importancia práctica para los clínicos.

Las medidas que responden a estos interrogantes se conocen como *valor predictivo*.

Valor predictivo de una prueba positiva = Proporción de los individuos con una prueba positiva que tienen la enfermedad.

Valor predictivo de una prueba negativa = Proporción de los individuos con una prueba negativa que no tienen la enfermedad.

Los términos *prevalencia* y *valor predictivo* aparecen en los artículos de investigación en relación con grupos de individuos. Por fortuna, en la práctica clínica se utilizan términos equivalentes, aunque el médico trata a un solo paciente a la vez. Desde la perspectiva de la actividad clínica, la *prevalencia* de una enfermedad corresponde a la mejor estimación de la probabilidad de enfermedad antes de realizar la prueba. En términos clínicos, la prevalencia se conoce como probabilidad *anterior a la prueba*. El *valor predictivo* significa lo mismo que la probabilidad de que la enfermedad esté presente (o ausente) después de obtener los resultados de la prueba. Por esta razón, los valores predictivos pueden considerarse clínicamente como la *probabilidad posterior a la prueba*. Si los términos *prevalencia* y *valor predictivo* le parecen confusos, puede sustituirlos por los de *probabilidad de la enfermedad* antes y después de realizar la prueba.

Como orientación práctica y sencilla para interpretar los valores de esas medidas, puede ser útil usar las siguientes aproximaciones de las probabilidades anteriores a la prueba:

1% = La probabilidad anterior a la prueba de los que están expuestos a factores de riesgo de una enfermedad común, pero asintomáticos.
10% = La probabilidad anterior a la prueba cuando la enfermedad es improbable, pero clínicamente posible y el clínico desea descartarla.
50% = La probabilidad anterior a la prueba cuando la incertidumbre es considerable, pero la presentación clínica es compatible con la enfermedad.
90% = La probabilidad anterior a la prueba cuando la enfermedad es muy probable clínicamente, pero el clínico desea confirmarla por medio de una prueba diagnóstica.

Mediante las tablas de 2 × 2 mostraremos cómo se calcula el valor predictivo. Recuerde que lo hacemos para una determinada prevalencia o probabilidad anterior a la prueba.

PRUEBA EN ESTUDIO	PRUEBA DE ORO ENFERMOS	PRUEBA DE ORO SANOS
Positivos	a = Número de individuos enfermos y positivos	b = Número de individuos sanos y positivos
Negativos	c = Número de individuos enfermos y negativos	d = Número de individuos sanos y negativos
	a + b = Total de positivos	c + d = Total de negativos

Para calcular el valor predictivo de las pruebas negativas y positivas se emplean las siguientes fórmulas.

$$\text{Valor predictivo de una prueba positiva} = \frac{a}{a+b} = \text{Proporción de individuos con una prueba positiva que realmente tienen la enfermedad (medida con la prueba de oro).}$$

$$\text{Valor predictivo de una prueba negativa} = \frac{d}{c+d} = \text{Proporción de individuos con un resultado negativo que realmente no tienen la enfermedad (medida con la prueba de oro).}$$

Ahora, calculemos estos valores empezando con probabilidades anteriores a la prueba de 90%, 50%, 10% y 1%. Recuerde que el número de positivos y de negativos será diferente para cada prevalencia de la enfermedad.

PROBABILIDAD ANTERIOR DE 90%

PRUEBA EN ESTUDIO	PRUEBA DE ORO ENFERMOS	PRUEBA DE ORO SANOS
Positivos	720	10
Negativos	180	90
	900	100

Probabilidad anterior a la prueba de 90%

$$\text{Valor predictivo de una prueba positiva} = \frac{a}{a + b} = \frac{720}{730} = 98{,}6\%$$

$$\text{Valor predictivo de una prueba negativa} = \frac{d}{c + d} = \frac{90}{270} = 33{,}3\%$$

Empleando el mismo método, los otros valores predictivos son:

Probabilidad anterior a la prueba de 50%

$$\text{Valor predictivo de una prueba positiva} = \frac{a}{a + b} = \frac{400}{450} = 88{,}9\%$$

$$\text{Valor predictivo de una prueba negativa} = \frac{d}{c + d} = \frac{450}{550} = 81{,}8\%$$

Probabilidad anterior a la prueba de 10%

$$\text{Valor predictivo de una prueba positiva} = \frac{a}{a + b} = \frac{80}{170} = 47{,}1\%$$

$$\text{Valor predictivo de una prueba negativa} = \frac{d}{c + d} = \frac{810}{830} = 97{,}6\%$$

Probabilidad anterior a la prueba de 1%

$$\text{Valor predictivo de una prueba positiva} = \frac{a}{a + b} = \frac{8}{107} = 7{,}5\%$$

$$\text{Valor predictivo de una prueba negativa} = \frac{d}{c + d} = \frac{2}{891} = 99{,}8\%$$

Para una prueba con una sensibilidad de 80% y una especificidad de 90%, los datos pueden resumirse de la siguiente forma:

Probabilidad anterior a la prueba	1%	10%	50%	90%
Valor predictivo de una prueba positiva	7,5%	47,1%	88,9%	98,6%
Valor predictivo de una prueba negativa	99,8%	97,6%	81,8%	33,3%

Los cálculos de los valores predictivos tienen importantes implicaciones clínicas. Indican que la probabilidad de que la enfermedad esté presente o ausente después de obtener los resultados de una prueba depende de la mejor estimación posible de la probabilidad de la enfermedad antes de realizar la prueba. Cuando la probabilidad de una enfermedad es moderadamente alta antes de realizar la prueba, por ejemplo de 50%, incluso una prueba negativa, como en el ejemplo utilizado, conduce a una probabilidad de que la enfermedad esté presente de 18,2% (100%–81,8%). Cuando la probabilidad de la enfermedad es relativamente baja antes de realizar la prueba, por ejemplo, 10%, incluso una prueba positiva conduce a una probabilidad de que la enfermedad no esté presente de 52,9% (100%–47,1%).

La situación empeora cuando la prueba se emplea como instrumento de tamizaje. Por ejemplo, podríamos aplicar la prueba a un grupo de individuos expuestos a un factor de riesgo cuya probabilidad de tener la enfermedad activa es de 1%. Nuestro ejemplo de 1% de prevalencia o probabilidad anterior a la prueba nos enseña que cuando se aplica la prueba que tiene una sensibilidad de 80% y una especificidad de 90% a este grupo de individuos, los que dan resultados positivos tienen una probabilidad de enfermedad de 7,5%. Esto es lo que significa un valor predictivo de 7,5% de una prueba positiva. Si no se entiende el efecto de la probabilidad anterior a la prueba sobre el valor predictivo, se puede cometer el error que describimos a continuación.

Se evaluó una prueba nueva y barata para diagnosticar el cáncer de pulmón aplicándola a un grupo de 100 individuos con cáncer de pulmón y a 100 sin la enfermedad. El valor predictivo de la prueba positiva fue de 85%; es decir, que 85% de los que tuvieron pruebas positivas padecían cáncer de pulmón. Los autores concluyeron que la prueba era adecuada para el tamizaje de ese cáncer en la población general, dado que 85% de los que tuvieran resultados positivos padecerían cáncer de pulmón.

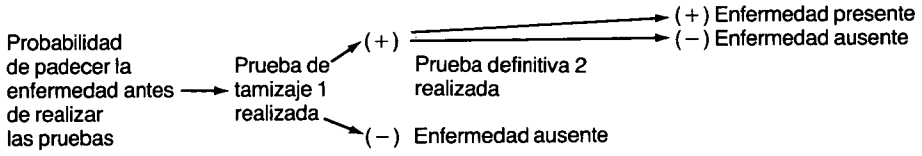
El valor predictivo de una prueba positiva es el porcentaje de personas con un resultado positivo que tienen la enfermedad. Ese valor predictivo depende de la prevalencia de la enfermedad en el grupo de individuos a los que se haya aplicado la prueba. A menudo una prueba se evalúa aplicándola a un grupo de individuos de los que se sabe que la mitad tienen la enfermedad. En este ejemplo, la prueba se aplicó a un grupo en el cual 50% de los individuos padecían la enfermedad (100 con cáncer de pulmón y 100 sin cáncer). Así, la prevalencia o probabilidad anterior de la enfermedad en este grupo era de 50%. Cuanto menor sea la probabilidad anterior de la prueba, menor será el valor predictivo de una prueba positiva.

En la comunidad, la prevalencia de cáncer de pulmón es mucho menor de 50%, incluso entre los fumadores. Por lo tanto, el valor predictivo de una prueba positiva en un individuo promedio de la comunidad —aunque esté expuesto a los factores de riesgo de cáncer de pulmón— será mucho menor de 85%. La capacidad de una prueba positiva para predecir la presencia de enfermedad cambia sustancialmente, según se aplique a grupos de individuos con probabilidades diferentes de presentar la enfermedad. Una prueba positiva puede tener un valor predictivo muy elevado en un grupo de pacientes; no obstante, en otro grupo con una prevalencia o probabilidad anterior distinta de la enfermedad, la misma prueba puede tener un valor predictivo mucho más bajo. La prueba puede ser útil para el diagnóstico en un grupo de pacientes que se sospecha pueden padecer la enfermedad, pero ser inútil para el tamizaje de la población general en la cual la sospecha de enfermedad es baja.

COMBINACIÓN DE PRUEBAS

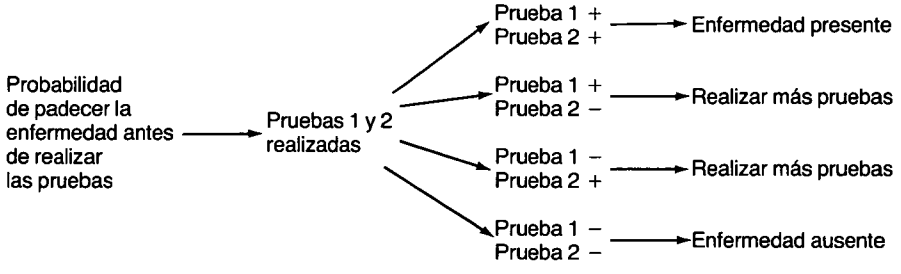
En la práctica clínica y en los artículos de investigación sobre análisis de decisión, que se publican cada vez con mayor frecuencia, los investigadores examinan los efectos de la combinación de pruebas. Hay dos formas básicas de combinar dos pruebas: en serie o en paralelo.

El uso de dos pruebas en serie puede conducir al diagnóstico mediante la siguiente estrategia:[3]



Al usar las dos pruebas en serie, la prueba número 2 solo se realiza en los individuos que son positivos a la prueba 1. Cuando los resultados de ambas pruebas son positivos, la probabilidad de tener la enfermedad se calcula considerando que el valor predictivo de una prueba positiva adicional es igual a la probabilidad de la enfermedad antes de realizar la prueba 2. Las pruebas en serie, aunque suelen tomar más tiempo, permiten a los médicos descartar la enfermedad empleando menos pruebas.[4] El valor predictivo de una prueba positiva o negativa no está influido por la prueba que se realiza en primer lugar.[5] Habitualmente, son razones de seguridad y de costo las que determinan cuál es la prueba que se efectuará primero.

La estrategia de utilizar las pruebas en paralelo exige realizar ambas pruebas al mismo tiempo, y la probabilidad de enfermedad se calcula después de realizadas, tal como se presenta en el siguiente gráfico.[6]



[3] Tomado de Riegelman RK y Povar GJ, eds. *Putting prevention into practice*. Boston: Little, Brown; 1988.
[4] También puede ser de ayuda realizar pruebas en serie cuando ambas pruebas tienen una especificidad baja.
[5] Se presume que las pruebas son independientes. En otras palabras, se parte del supuesto de que la segunda prueba tendrá la misma sensibilidad y especificidad, sea cual fuere el resultado de la primera.
[6] Véase la nota 3.

Esta estrategia en paralelo funciona bien cuando ninguna de las pruebas tiene una sensibilidad especialmente elevada, pero cada una es capaz de detectar un tipo o estadio distinto de la enfermedad. A veces una prueba puede detectar un estadio temprano, mientras que la otra puede detectar uno más avanzado. En otros casos, una prueba puede detectar una enfermedad de evolución rápida o agresiva mientras que la segunda detecta una de progresión gradual o lenta.

Las estrategias en serie y en paralelo suponen que la segunda prueba proporciona una información adicional superior a la que proporciona la primera. En caso contrario, el rendimiento de ambas estrategias es menor que el esperado. Por ejemplo, imagine el siguiente uso de la estrategia en paralelo.

Se estudiaron la exploración mamaria y la termografía como pruebas en paralelo para detectar el cáncer de mama. Se observó que la exploración tenía una sensibilidad de 40% y la termografía, de 50%. Utilizando las dos conjuntamente, se comprobó que con una u otra prueba solo se detectaban 50% de los cánceres. Los investigadores estaban sorprendidos, dado que habían previsto ser capaces de detectar la mayor parte de los cánceres de mama.

La combinación de la exploración mamaria y la termografía añade poco al uso de cada prueba por separado, ya que ninguna de ellas detecta la enfermedad en una fase temprana. Los resultados de las pruebas nos proporcionan básicamente la misma información y, por este motivo, con una tenemos suficiente. Las pruebas de este tipo son poco útiles cuando se aplican conjuntamente, sea en paralelo o en serie.

Por eso, al diseñar una estrategia de diagnóstico para aplicar las pruebas, necesitamos saber algo más que la sensibilidad y la especificidad; necesitamos saber si las pruebas miden fenómenos diferentes o independientes. También es preciso saber los tipos de enfermedad que pasan por alto las pruebas, por ejemplo, si son incapaces de detectar enfermedades en sus fases iniciales o las de evolución lenta.

CAPÍTULO 18

RESUMEN: LA PRUEBA DE UNA PRUEBA

El marco utilizado para probar una prueba nos exige evaluarla mediante los conceptos de reproducibilidad y exactitud, y determinar la variabilidad de los resultados en los que no tienen la enfermedad, mediante el concepto del intervalo de la normalidad, y de los que tienen la enfermedad, mediante su medición con la prueba de oro o de referencia. Seguidamente, esta información se combina para valorar la discriminación diagnóstica de la prueba, que se mide en función de su sensibilidad, especificidad y valor predictivo.

La variabilidad de una prueba se mide mediante su reproducibilidad, o sea, repitiendo la prueba en condiciones idénticas. Las repeticiones de la prueba se interpretan sin conocer los resultados originales. La reproducibilidad no garantiza por sí sola la exactitud de la prueba. Una prueba reproducible puede reproducir resultados inexactos cuando existe un sesgo en una dirección. Aunque se puede prever alguna variabilidad en los resultados, esta debe ser bastante menor que la variabilidad biológica medida con el intervalo de la normalidad.

La variabilidad entre individuos sin la enfermedad se mide mediante el intervalo de la normalidad. Este intervalo muchas veces comprende solo 95% de los valores de los individuos considerados sanos. Además, depende del grupo de referencia seleccionado. Recuerde que ese intervalo es meramente una descripción de cómo son las cosas entre los individuos presuntamente sanos. No es diagnóstico: estar fuera de los límites de la normalidad no equivale a estar enfermo, y estar dentro de los límites de la normalidad no equivale a estar sano; un cambio dentro del intervalo de la normalidad puede ser patológico, y estar dentro de los límites de la normalidad no es necesariamente igual a tener los valores más convenientes. Se pueden ajustar los valores que demarcan los límites de la normalidad, alterando de ese modo la especificidad, pero debe tenerse en cuenta el precio que se paga con el número de positivos y negativos falsos. Finalmente, al emplear el concepto del intervalo de la normalidad para definir a un grupo de individuos sanos, es preciso establecer límites inequívocos, con objeto de determinar cuáles son los pacientes positivos y cuáles los negativos a la prueba.

El grupo de sujetos con la enfermedad se define mediante la prueba de oro, que es el mejor método disponible y generalmente aceptado para diagnosticar la enfermedad. Al definir a los enfermos, es conveniente incluir a individuos que tengan el mismo tipo de estado patológico que encontraremos al aplicar la prueba en el medio clínico.

Después de haber comprobado que la prueba es reproducible, que se ha definido lo que es una prueba positiva y una negativa mediante el intervalo de la normalidad y que, mediante el criterio de referencia, se ha identificado a un grupo de individuos enfermos y a uno de sanos, ya se puede valorar la discriminación diagnóstica de una prueba.

Cuando se aplica la prueba en estudio a los individuos identificados como enfermos o como sanos mediante el criterio de referencia, se calculan su sensibilidad y especificidad comparando los resultados de la prueba estudiada con los de la prueba de referencia. Dado que se supone que la prueba de referencia tiene una dis-

criminación diagnóstica perfecta, o de 100%, la prueba en estudio generalmente no estará a la altura de la de referencia. Esto será cierto aunque la prueba estudiada sea intrínsecamente mejor que la de referencia, ya que cualquier discrepancia se resolverá a favor de la de oro.

La sensibilidad mide la proporción de los individuos que tienen la enfermedad, diagnosticada según el criterio de referencia, que son identificados correctamente como enfermos por la prueba estudiada. La especificidad mide la proporción de los que no tienen la enfermedad, según el criterio de referencia, que son identificados correctamente como sanos por la prueba en estudio. La sensibilidad y la especificidad son importantes porque teóricamente son independientes de la prevalencia o probabilidad anterior a la prueba de que la enfermedad exista en el grupo de individuos estudiados. Esto permite comparar los resultados que obtienen en la prueba distintos grupos de pacientes. Las pruebas se pueden comparar directamente en función de su sensibilidad y especificidad. Sin embargo, es importante reconocer que estas medidas pueden ser distintas en los estadios tempranos de una enfermedad, comparados con los avanzados.

La sensibilidad y la especificidad no responden a la siguiente pregunta clínica: ¿cuál es la probabilidad de tener la enfermedad si la prueba es negativa o si es positiva? El valor predictivo de una prueba positiva nos informa de la corrección con que la prueba confirma la enfermedad en una determinada situación clínica. El valor predictivo de una prueba negativa nos indica la corrección con que la prueba descarta la enfermedad en una determinada situación clínica. Los valores predictivos, a diferencia de la sensibilidad y la especificidad, dependen de la prevalencia o probabilidad de la enfermedad anterior a la prueba.

Desde el punto de vista clínico, esto significa que se debe realizar la mejor estimación posible de la probabilidad de la enfermedad antes de ejecutar la prueba. El valor predictivo informa al clínico sobre la probabilidad de que la enfermedad esté presente después de realizar la prueba. El clínico ha de tener cuidado de no extrapolar el valor predictivo de un contexto clínico a otro. Una prueba muy útil para diagnosticar la enfermedad en presencia de síntomas puede ser prácticamente inútil para el tamizaje de individuos asintomáticos.

Las pruebas pueden combinarse en serie o en paralelo. Cuando se utilizan dos pruebas en serie, la segunda prueba se realiza únicamente si la primera da un resultado positivo. Las pruebas en serie a veces permiten establecer un diagnóstico realizando un menor número de pruebas. Las pruebas en paralelo pueden emplearse cuando ninguna de ellas tiene una sensibilidad suficientemente alta. Esta estrategia funciona bien cuando las dos pruebas detectan diferentes estadios o tipos de la enfermedad, diferenciando, por ejemplo, entre un estadio incipiente y uno avanzado o entre una enfermedad fulminante y una de desarrollo lento. En esta situación, el empleo de las pruebas en paralelo garantiza que se detectará un porcentaje más alto de individuos con la enfermedad. La obtención del máximo beneficio de cualquiera de estas estrategias requiere que los resultados de las dos pruebas midan fenómenos distintos o detecten la enfermedad en etapas diferentes de su evolución.

PREGUNTAS ÚTILES PARA PROBAR UNA PRUEBA

La siguiente lista de preguntas de comprobación ayudará a reforzar los principios necesarios para probar una prueba.

1. Propiedades intrínsecas de una prueba.
   a. Reproducibilidad: ¿producen resultados prácticamente idénticos las

repeticiones múltiples de una prueba realizada en las mismas condiciones?

b. Exactitud: ¿corresponden los resultados de la prueba a los verdaderos valores del fenómeno anatómico, bioquímico o fisiológico?

c. Exactitud clínica: ¿proporciona la prueba mediciones similares a las experimentales cuando se realiza en las condiciones reales de la práctica clínica?

2. Variación biológica: el concepto del intervalo de la normalidad y de la variabilidad en los individuos sanos.
   a. ¿Se ha establecido el intervalo de la normalidad de forma apropiada para que incluya a un porcentaje definido, a menudo 95%, de los individuos considerados sanos?
   b. ¿Se ha distinguido entre estar fuera del intervalo de la normalidad y estar enfermo?
   c. ¿Se ha distinguido entre estar dentro del intervalo de la normalidad y estar sano?
   d. ¿Se puede aplicar de forma generalizada el grupo de referencia utilizado o existen grupos identificables con diferentes intervalos de la normalidad?
   e. ¿Reconocieron quienes aplicaron la prueba que el intervalo de la normalidad es una descripción del grupo presuntamente sano y que los cambios dentro del mismo para un individuo pueden ser patológicos?
   f. ¿Se ha distinguido el intervalo de la normalidad del deseable?
   g. ¿Han justificado los investigadores la modificación del intervalo para cumplir con objetivos diagnósticos específicos?

3. Variabilidad de los individuos enfermos.
   a. ¿Han seleccionado los investigadores la mejor prueba de oro disponible para definir a los pacientes que tienen la enfermedad en estudio?
   b. ¿Han incluido los investigadores a individuos que representen todo el espectro de la enfermedad, para establecer un intervalo realista de los resultados posibles?

4. Discriminación diagnóstica: distinción entre los enfermos y los sanos.
   a. ¿Cuán correctamente identifica la prueba a los enfermos? ¿Cuán alta es su sensibilidad? ¿Con qué frecuencia es positiva en los enfermos?
   b. ¿Cuán correctamente identifica a los que no tienen la enfermedad? ¿Cuán alta es su especificidad? ¿Con qué frecuencia es negativa en los que están sanos?
   c. ¿Se ha reconocido que, si bien en teoría la sensibilidad y la especificidad no son influidas por la probabilidad de la enfermedad posterior a la prueba, estas medidas pueden ser distintas en las fases tempranas y en las avanzadas de la enfermedad?
   d. ¿Se ha distinguido entre la sensibilidad y la especificidad de la prueba y su valor predictivo cuando es positiva y cuando es negativa?

PRUEBAS DE LABORATORIO

Examinemos algunas de las pruebas básicas de laboratorio que se utilizan en la medicina clínica para valorar su exactitud, reproducibilidad, intervalo de la normalidad y discriminación diagnóstica.

Hematócrito

Exactitud y reproducibilidad

El hematócrito es una medida del porcentaje de la sangre total compuesta por glóbulos rojos aglomerados. Los hematócritos de rutina se miden mediante la punción en el dedo, con el fin de valorar la sangre en los capilares, o mediante punción venosa. Ambos métodos permiten medir exactamente la cantidad relativa de eritrocitos en la sangre, pero su reproducibilidad depende de que se preste atención a los detalles técnicos. Se puede esperar que la sangre de los capilares tenga un hematócrito entre 1 y 3% más bajo que la venosa. El hecho de apretar excesivamente el dedo puede extraer más plasma y disminuir erróneamente el hematócrito. En presencia de anemia grave, la punción del dedo ofrece valores menos exactos.

Para valorar la exactitud con que el hematócrito mide el estado fisiológico, cabe recordar que se está midiendo la masa relativa, no absoluta, de glóbulos rojos. Los resultados pueden ser erróneos si el volumen plasmático se ha reducido debido a deshidratación o diuresis. Los individuos con un volumen de plasma reducido pueden tener hematócritos por encima del intervalo de la normalidad. Estas son variaciones normales que pueden confundirse con la policitemia (hematócrito patológicamente elevado).

Intervalo de la normalidad

El intervalo de la normalidad de las concentraciones del hematócrito es diferente en los hombres y en las mujeres. Esto es de conocimiento general en los laboratorios y en sus informes los intervalos normales para hombres y mujeres se presentan por separado. Menos a menudo se reconoce que los intervalos de la normalidad son distintos en las diferentes fases del embarazo, en diversas edades y en las personas que viven a distintas altitudes. El hematócrito suele descender durante el embarazo, empezando en algún momento entre el tercer y el quinto mes. Entre el quinto y el octavo no es inusual que se observe una reducción de 20% respecto a los valores anteriores. Sin embargo, por lo general aumenta ligeramente cerca del término del embarazo y vuelve a sus valores normales seis semanas después del parto.

La edad tiene un efecto marcado sobre el hematócrito, especialmente en los niños. El intervalo de la normalidad del primer día de vida es 54 ± 10 (esto es, oscila entre 44 y 64). Al 14o. día, el intervalo es 42 ± 7, y a los seis meses, 35,5 ± 5. El hematócrito promedio aumenta gradualmente hasta la adolescencia, y alcanza un promedio de 39 entre los 11 y los 15 años de edad. El intervalo de la normalidad de los hombres adultos es 47 ± 5 y el de las mujeres, 42 ± 5.

La presión barométrica baja también tiene un efecto pronunciado sobre el intervalo de la normalidad del hematócrito. Las personas que nacen y viven a grandes alturas tienen en general hematócritos más altos. Por ejemplo, el intervalo de la normalidad a los 1 200 metros es 49,2 ± 4,5 en los hombres adultos y 44,5 ± 4,5 en las mujeres adultas.

El intervalo de valores de los hematócritos normales es bastante amplio. Por esta razón, si un individuo tiene un valor cercano al límite superior de la normalidad, puede llegar a perder hasta una quinta parte del volumen de eritrocitos antes de que se pueda demostrar la existencia de anemia mediante un hematócrito bajo. Las comparaciones con los hematócritos anteriores son importantes para valorar el desarrollo de la anemia. Individualmente, el hematócrito se mantiene dentro de límites

fisiológicos bastante estrechos, por eso, sus cambios constituyen una medida diagnóstica mejor que la de una sola observación.

Discriminación diagnóstica

Al evaluar el hematócrito, también es necesario conocer la forma en que este responde a la pérdida aguda de sangre. Durante una hemorragia aguda se pierde sangre entera y en un primer momento, la restante se aproximará al hematócrito original. Pueden pasar hasta 24 horas o más antes de que la pérdida se compense mediante un aumento del volumen plasmático. Solamente después de esta compensación puede el hematócrito reflejar totalmente la magnitud de la pérdida de sangre. Si no se reconoce este fenómeno, se pueden obtener resultados negativos falsos en la identificación de los sangrados agudos.

El cuerpo suele ser capaz de compensar la desintegración o pérdida de sangre que ocurre lentamente. Por lo tanto, es posible que en los pacientes con sangrado lento o hemólisis no se detecte anemia, aunque el número de reticulocitos esté elevado. Si mediante la prueba del hematócrito se espera diagnosticar las enfermedades que predisponen a las pérdidas de sangre o a las hemólisis, encontraremos un número relativamente alto de negativos falsos. Enfermedades como la beta-talasemia o los déficit de glucosa-6-fosfato deshidrogenasa (G-6FD) se manifiestan muchas veces como anemias compensadas que cursan con un hematócrito normal o en el límite inferior de la normalidad. Los resultados positivos falsos se pueden producir entre los individuos que tienen el volumen plasmático aumentado como una variante de la normalidad. Desde otro punto de vista, estos individuos reflejan el hecho de que para cualquier intervalo de la normalidad habrá individuos sanos con valores fuera de los límites del intervalo de la normalidad.

Nitrógeno ureico en la sangre y creatinina sérica

Exactitud y reproducibilidad

Para medir las concentraciones de nitrógeno ureico en la sangre (NUS) y la creatinina sérica contamos con pruebas automatizadas bastante reproducibles. Dado que esas concentraciones reflejan la acumulación de nitrógeno ureico y creatinina no excretados, constituyen una indicación de la incapacidad del riñón para eliminar esas sustancias. Cuando se utilizan como medidas de la función renal, sirven para valorar la tasa de filtración glomerular, pero reflejan más exactamente esta tasa cuando hay un deterioro importante de la función renal. Como los riñones suelen tener una capacidad de reserva funcional considerable, puede producirse una pérdida funcional importante, por ejemplo, de un riñón, sin que aumente la acumulación de NUS y creatinina. Estas sustancias reflejarán la tasa de filtración glomerular solo después de una pérdida de 50% o superior de la filtración glomerular. Cuanto más desciende la tasa de filtración, más rápidamente aumentan el NUS y la creatinina. Cuando el NUS y la creatinina están claramente elevados, su incremento porcentual refleja mejor el porcentaje de pérdida de la tasa de filtración glomerular que su cambio numérico.

Cuando la concentración de creatinina se encuentra dentro de los límites normales y se precisa disponer de una valoración exacta de la tasa de filtración glomerular, es posible medir la concentración de creatinina en orina de 24 horas y a la vez la sérica, y calcular entonces el aclaramiento de creatinina. Aunque esta prueba presenta problemas relacionados con su reproducibilidad y con el intervalo de la normali-

dad, es capaz de detectar cambios mucho menores de la tasa de filtración glomerular cuando la concentración sérica de creatinina no está elevada.

Intervalo de la normalidad

Los intervalos de la normalidad de uso estándar en laboratorios varían aproximadamente entre 10 y 20 mg/dl para el NUS y entre 0,6 y 1,4 mg/dl para la creatinina sérica. Para utilizar esos intervalos, es necesario comprender los factores que los afectan en ausencia de enfermedad. El NUS refleja el estado proteico y de hidratación de un paciente. Por eso, el NUS puede cambiar marcadamente sin indicar la presencia de una enfermedad específica. La creatinina es un producto muscular y, como tal, varía ampliamente entre individuos sanos, de acuerdo con su masa muscular. La concentración de creatinina suele ser más baja en las mujeres que en los hombres. En los ancianos, considerados como grupo, la masa muscular y la concentración de creatinina son relativamente menores. A pesar de estas importantes diferencias, los laboratorios habitualmente presentan la concentración de creatinina en referencia a un solo intervalo de la normalidad.

La concentración de creatinina, al contrario que la del NUS, varía menos de día a día y de mes a mes en respuesta a factores no renales. Muchas veces, las comparaciones de las concentraciones de creatinina medidas en distintos momentos pueden proporcionar información muy importante sobre el estado renal. Una concentración de creatinina sérica de 1,3 mg/dl en una mujer anciana puede reflejar una pérdida considerable de la tasa de filtración glomerular, especialmente si ha aumentado en comparación con valores anteriores. La misma concentración en un hombre joven y musculoso puede ser estable y no indicar disminución de la tasa de filtración glomerular.

Discriminación diagnóstica

Como hemos comentado anteriormente, a menudo se obtienen resultados negativos falsos para la enfermedad renal utilizando el NUS y la creatinina, dado que la concentración de estos metabolitos no empieza a aumentar hasta que se ha producido una pérdida sustancial de la filtración glomerular. Además, también se producen resultados positivos falsos tanto con la creatinina como con el NUS. La concentración de NUS puede ser baja en enfermedades que producen malnutrición. Por ejemplo, los alcohólicos con frecuencia tienen concentraciones bajas de NUS. El NUS también refleja la degradación hemática que se produce durante una pérdida rápida de sangre en el tubo digestivo. Estas elevaciones del NUS no son resultados positivos falsos en sentido estricto, pues indican la presencia de otras enfermedades distintas de las renales. Sin embargo, son positivos falsos cuando se está tratando de valorar la filtración glomerular. En presencia de enfermedades musculares se puede detectar una elevación falsa de la concentración de creatinina. En este caso, los resultados tampoco son positivos falsos, pero sugieren que la enfermedad no es renal.

Es posible utilizar las pruebas del NUS y la creatinina en paralelo para entender y localizar anormalidades. Normalmente, la razón entre la concentración de NUS y la de creatinina es de 15:1. La elevación desproporcionada del NUS en relación con la creatinina sugiere la presencia de una enfermedad pre o posrenal, más que una enfermedad propiamente renal. La deshidratación puede presentarse como un patrón prerrenal, con elevación desproporcionada del NUS respecto a la creatinina. A veces, alguna enfermedad posrenal, como la obstrucción prostática, también produce una

elevación desproporcionada del NUS respecto a la creatinina. Por este motivo, el uso paralelo o simultáneo de las pruebas del NUS y la creatinina ejemplifica una situación en la cual el uso de dos pruebas ofrecerá más información diagnóstica que el de una sola, dado que miden fenómenos distintos.

Acido úrico

Exactitud y reproducibilidad

La concentración de ácido úrico en la sangre se puede medir de forma reproducible mediante técnicas automatizadas. Para ello existen diversos métodos, cada uno de los cuales proporciona valores ligeramente distintos. Es importante comparar las concentraciones obtenidas con el mismo método. Con las técnicas automatizadas generalmente se obtienen valores algo más altos. La concentración de ácido úrico puede variar en breve plazo; por ejemplo, puede aumentar rápida y significativamente debido a deshidratación.

La concentración de ácido úrico en la sangre mide exactamente la concentración del ácido úrico en el suero. Sin embargo, no valora exactamente todos los parámetros fisiológicos importantes del ácido úrico. No es un indicador exacto del ácido úrico corporal total. Por ejemplo, el ácido úrico cristalizado y depositado no se refleja en la concentración sérica. En relación con el desarrollo de gota, el ácido úrico cristalizado y depositado es frecuentemente el criterio diagnóstico decisivo. Además, la concentración sérica de ácido úrico es solo uno de los factores que influyen en su excreción. Algunos individuos que no tienen una concentración elevada de ácido úrico en el suero pueden excretarlo en grandes cantidades, lo cual los predispone a formar piedras de ácido úrico.

Intervalo de la normalidad y discriminación diagnóstica

El intervalo de valores normales del ácido úrico es bastante amplio y en la mayor parte de los laboratorios varía aproximadamente entre 2 y 8 mg/dl, en función del método empleado. Muchos individuos tienen valores ligeramente por encima de este intervalo y muy pocos por debajo. Son muy pocas las personas con valores ligeramente elevados que desarrollan gota. Esto ha llevado a muchos médicos a afirmar que no se debe tratar a los pacientes con concentraciones ligeramente elevadas de ácido úrico. En realidad, lo que están alegando es que el límite superior de la normalidad se debe aumentar para que la concentración sérica del ácido úrico permita discriminar mejor entre los que están predispuestos a padecer gota y los que no lo están.

Incluso cuando las concentraciones de ácido úrico son elevadas, la discriminación diagnóstica no es buena, porque frecuentemente aparecen resultados positivos falsos. Las personas con insuficiencia renal en muchas ocasiones tienen concentraciones elevadas de ácido úrico, pero raramente padecen gota. Es evidente que existen otros factores, además del ácido úrico, que determinan el desarrollo de esta dolencia. Se ha demostrado que a veces la gota se desarrolla en los individuos predispuestos cuando la concentración de ácido úrico cambia rápidamente. Tanto su reducción como su aumento puede precipitar una crisis de gota. Mucha gente que padece de gota tiene concentraciones séricas de ácido úrico normales durante los episodios de la enfermedad. Estos casos se han documentado mediante la prueba de oro, la demostración de la presencia de cristales birrefringentes en el líquido articular, y no por medio de la concentración de ácido úrico en el suero. Las personas que se sospecha tienen gota a pesar

de que sus concentraciones de ácido úrico están dentro de los límites de la normalidad durante los ataques agudos, deben ser examinadas posteriormente. A menudo, sus concentraciones de ácido úrico habrán aumentado, lo cual demuestra que se obtuvo un resultado negativo falso al inicio del episodio de gota.

A causa de la débil asociación entre la concentración del ácido úrico en el suero y su excreción urinaria, se puede decir que la primera prueba tiene una baja discriminación diagnóstica de las piedras de ácido úrico. El número de resultados positivos y negativos falsos sería muy alto si se empleara la concentración de ácido úrico en el suero como criterio exclusivo de diagnóstico. La demostración de la presencia de piedras es el método de diagnóstico más definitivo. Se ha observado que la excreción elevada de ácido úrico en orina de 24 horas indica predisposición a la formación de piedras de ácido úrico. En este caso la asociación tampoco es exclusiva, porque hay otros factores que influyen en la formación de piedras. El volumen bajo de orina y el pH bajo, especialmente, son factores predisponentes que aumentan la frecuencia de formación de piedras de ácido úrico.

El ácido úrico ilustra diversos principios importantes aplicables a las pruebas diagnósticas.

1. Incluso en presencia de una prueba exacta y reproducible que refleje la anormalidad metabólica relacionada con la enfermedad, el grado de asociación entre la concentración sérica y la presencia de enfermedad puede ser bajo.
2. Es posible que la medición de concentraciones séricas, tanto del ácido úrico como de otras sustancias, no constituya una buena indicación del contenido corporal total y, en particular, de su concentración en las localizaciones patológicas.
3. Puede que la concentración sérica proporcione menos seguridad en el diagnóstico cuando más se necesita, al inicio de los síntomas.

CAPÍTULO 19

EJERCICIOS PARA DETECTAR ERRORES: LA PRUEBA DE UNA PRUEBA

Los siguientes ejercicios se han diseñado con el fin de evaluar la capacidad que usted ha adquirido para aplicar los diversos principios usados en *La prueba de una prueba*. Los ejercicios incluyen varios errores que se han ilustrado con ejemplos hipotéticos. Lea cada ejercicio y luego escriba una crítica señalando los tipos de errores cometidos por los invetigadores. Compare su crítica con la que se proporciona al final de cada ejercicio.

EJERCICIO No. 1: VARIABILIDAD DE LAS POBLACIONES SANA Y ENFERMA

Se realizaron dos investigaciones para evaluar la utilidad de una nueva prueba diagnóstica del cáncer de mama. Previamente, se había observado que los resultados de la prueba variaban en menos de 1% cuando se repetían en condiciones iguales y eran leídos por el mismo intérprete.

En el primer estudio, los investigadores escogieron a 100 mujeres con cáncer de mama metastásico y a 100 mujeres sanas sin signos de enfermedad mamaria. En las mujeres sanas, los resultados de la prueba oscilaron entre 30 y 100 mg/dl, y en las pacientes con cáncer, entre 150 y 200 mg/dl. Dado que la prueba diferenciaba perfectamente a un grupo de otro, los investigadores concluyeron que podía considerarse una prueba ideal para el diagnóstico del cáncer de mama y que debía aplicarse inmediatamente al tamizaje de todas las mujeres.

En un segundo estudio en el que se usó la misma prueba, otro investigador comparó a 100 mujeres recién diagnosticadas de cáncer de mama con 100 que padecían una enfermedad benigna de mama. Los resultados de las pacientes de cáncer variaron entre 70 y 200 mg/dl y los de las pacientes con enfermedad benigna, entre 40 y 180 mg/dl. Los autores de este estudio se percataron de la notable superposición entre los dos grupos y concluyeron que la prueba era inútil.

Un lector de ambos estudios, asombrado de que dos investigadores respetados pudieran obtener resultados tan inconsistentes, concluyó que se debían haber cometido errores al notificarlos.

CRÍTICA: EJERCICIO No. 1

Para revisar estos estudios, es interesante organizar la discusión en torno a los conceptos de variabilidad de la prueba, variabilidad de la población sin la enfermedad y variabilidad de la población con la enfermedad.

Variabilidad de la prueba

La reproducibilidad es la medida de la variabilidad de una prueba. Según se afirma en el estudio, cuando la prueba fue realizada en las mismas condiciones por el mismo intérprete, su variabilidad fue de 1%. La medición de la reproducibi-

lidad exige repetir la prueba para demostrar que los resultados de la segunda lectura no están influidos por los de la primera lectura. Como el mismo intérprete repitió las pruebas, es posible que la lectura de los resultados de la segunda prueba estuviera influida por los resultados iniciales. Si esto fuera cierto, la reproducibilidad podría ser menor que la notificada anteriormente. No obstante, en el resto de la discusión supondremos que los autores tenían razón al creer que la prueba era reproducible.

Variabilidad del grupo de individuos sin la enfermedad

En el primer estudio se incluyeron mujeres sanas y los resultados oscilaron entre 30 y 100 mg/dl. En cambio, en el segundo se estudiaron mujeres con enfermedades mamarias benignas y los resultados estuvieron comprendidos entre 70 y 200 mg/dl. Tal vez estos estudios no sean contradictorios y representen dos segmentos diferentes del grupo de mujeres sin la enfermedad. Es posible que la enfermedad mamaria benigna eleve a valores intermedios la concentración del metabolito medida por la prueba.

La medida adecuada del grupo de personas sin la enfermedad es el intervalo de la normalidad, el cual habitualmente incluye a 95% de los individuos del grupo sin cáncer de mama. Sin embargo, aquí los resultados se presentan en intervalos que agrupan a 100% de los individuos; el intervalo total no nos dice nada sobre la forma en que se agrupan los resultados. Estos podrían estar muy concentrados entre 70 y 100. Sin disponer de todos los datos, o al menos del intervalo de la normalidad, es difícil utilizar estos estudios para comparar pacientes con y sin cáncer de mama.

Variabilidad de los individuos con la enfermedad

En el primer estudio se incluyeron pacientes con cáncer metastásico de mama y se estableció un intervalo de resultados comprendido entre 150 y 200 mg/dl. En el segundo, se estudiaron pacientes recién diagnosticadas y los límites del intervalo de los resultados observados fueron 70 y 200 mg/dl. Es posible que esta discrepancia no refleje un error en la notificación de los datos, sino las diferencias entre los grupos estudiados. Las pacientes recién diagnosticadas de cáncer de mama probablemente representan un amplio espectro de la enfermedad, incluidos los estadios iniciales y los metastásicos. Las pacientes con metástasis de cáncer de mama constituirían solo un extremo del espectro de la enfermedad. Por este motivo, el intervalo más amplio de valores encontrado entre las pacientes recién diagnosticadas podría reflejar el grupo más representativo de pacientes con cáncer de mama incluido en el segundo estudio.

Discriminación diagnóstica

Los datos presentados no nos permiten calcular la sensibilidad o la especificidad de la prueba. Puesto que no se menciona la distribución de los resultados individuales, no se puede establecer un intervalo de la normalidad o una separación entre las pruebas positivas y negativas. Por lo tanto, no es aconsejable llegar a conclusiones sobre la utilidad diagnóstica de la prueba.

En el primer estudio, los investigadores incluyeron tanto a individuos que padecían una enfermedad bastante avanzada como a individuos claramente sanos. No es sorprendente que sus resultados parecieran diferenciar correctamente a los grupos. En el segundo estudio, los investigadores incluyeron a pacientes que representaban un espectro más amplio de la enfermedad y también a las que padecían enfermedades benignas. Por lo tanto, tampoco es sorprendente que se produjera una ma-

yor superposición de los valores numéricos. Tan incorrecto es inferir que el primer estudio es una prueba perfecta como que el segundo es una prueba inútil. La verdad, que probablemente se encuentra entre ambos extremos, exige valorar la sensibilidad y la especificidad utilizando todos los datos de un amplio espectro de la población enferma y de la sana.

EJERCICIO No. 2: EL CONCEPTO DE LA NORMALIDAD

Un investigador intentó establecer los límites de la normalidad para una nueva prueba diagnóstica de la diabetes, como se describe a continuación.

1. Localizó a 1 000 pacientes hospitalizados por enfermedades distintas de la diabetes.
2. Aplicó la prueba a esos 1 000 pacientes.
3. Trazó la distribución de los valores de la nueva prueba, excluyó 2,5% de los valores del extremo superior y 2,5% del inferior e incluyó el 95% restante en el intervalo de la normalidad.

A continuación, aplicó la nueva prueba en la comunidad y realizó pruebas de tamizaje en voluntarios. A los que tuvieron resultados dentro del intervalo de la normalidad les dijo que no tenían diabetes y a los que tuvieron resultados fuera de esos límites, que padecían diabetes. Un año más tarde aplicó la prueba de nuevo a varios individuos cuyos resultados habían correspondido a la zona inferior del límite de la normalidad y al observar que en esta ocasión los resultados se encontraban en la zona superior de dicho límite, les aseguró que no tenían diabetes.

Un paciente obeso con resultados en la zona superior del intervalo de la normalidad y con una historia familiar de diabetes muy marcada le pidió consejo sobre la forma de evitar el desarrollo de la enfermedad. El investigador le respondió que, como sus resultados se encontraban dentro del intervalo de la normalidad, no debía tener motivos de preocupación.

CRÍTICA: EJERCICIO No. 2

Desarrollo del intervalo de la normalidad

Al establecer el intervalo de la normalidad, el investigador debe intentar incluir solamente a los individuos que no tengan la enfermedad estudiada. El investigador del estudio anterior llegó a la conclusión de que los individuos hospitalizados con diagnósticos distintos de la diabetes no padecían esta enfermedad. Sin embargo, la diabetes es muy frecuente y los pacientes diabéticos desarrollan una serie de complicaciones que aumentan el riesgo de ser hospitalizados. Por lo tanto, es probable que una proporción de los individuos internados con diagnósticos principales diferentes también tuvieran diabetes y que el investigador no haya establecido un intervalo de la normalidad de pacientes exentos de la enfermedad.

El investigador utilizó como intervalo de la normalidad 95% de los resultados centrales de un grupo de personas que presuntamente no tenían la enfermedad. Aunque este es el procedimiento habitual, puede que no se preste a la máxima discriminación diagnóstica de la prueba. A veces, la modificación de los límites del intervalo de la normalidad puede mejorar la capacidad de la prueba para discriminar entre los que tienen y los que no tienen la enfermedad. No obstante, debe recordarse que cuando cambiamos los límites del intervalo de la normalidad para obtener menos re-

sultados negativos falsos, pagamos el precio de obtener más resultados positivos falsos o viceversa. Si bien puede merecer la pena pagar ese precio, se necesitan más datos antes de poder saber si este es el caso. De cualquier modo, los datos disponibles no proporcionan los medios adecuados para juzgar si la nueva prueba ayuda a discriminar a los diabéticos de los no diabéticos. Lo único que conocemos es el intervalo de la normalidad definido para la prueba.

Aplicación del intervalo de la normalidad

En el caso descrito, no se ha mantenido la distinción entre el concepto del intervalo de la normalidad y el de enfermedad. El autor ha considerado sinónimos el estar fuera del intervalo de la normalidad y tener diabetes, y el estar dentro del intervalo y no tener diabetes. No se han presentado pruebas de que la nueva prueba sea útil para discriminar a los diabéticos de los no diabéticos y es posible que los primeros se encuentren totalmente dentro del intervalo de la normalidad de esta prueba.

Aunque se hubiera demostrado, desde el punto de vista del diagnóstico, que esta prueba es útil para distinguir a los que tienen diabetes de los que no la tienen, es probable que algunos individuos con diabetes tuvieran valores dentro del intervalo de la normalidad y algunos sin la enfermedad, fuera de dicho intervalo. Por definición, el intervalo de la normalidad excluye a 5% de los individuos que no tienen la enfermedad. Por eso, el investigador no puede limitarse simplemente a aplicar la prueba y a identificar a los individuos como diabéticos o no diabéticos.

Cambios dentro del intervalo de la normalidad

El hecho de que los resultados de una prueba aplicada a un individuo cambien, aunque se mantengan dentro del intervalo de la normalidad, puede ser una manifestación de enfermedad. El concepto del intervalo de la normalidad se ha desarrollado principalmente para individuos sobre los que no disponemos de datos basales anteriores. Cuando este es el caso, es preciso comparar, por medio del intervalo de la normalidad, los resultados individuales con los de aquellos que presuntamente están sanos. Si la misma prueba se ha aplicado al individuo con anterioridad, esta información debe tenerse en cuenta.

Un cambio dentro del intervalo de la normalidad puede representar un gran aumento para un individuo determinado; esto se manifiesta especialmente cuando los resultados anteriores de la prueba se encuentran cerca del límite inferior de la normalidad y los posteriores se desplazan hacia su límite superior. Para esos individuos, los cambios que se producen dentro de los límites normales pueden ser manifestaciones precoces de la enfermedad.

Grupo de referencia

El grupo de referencia empleado en este estudio para fijar el intervalo de la normalidad estaba formado en su totalidad por pacientes hospitalizados. Su intervalo de la normalidad podría ser bastante diferente del de otras poblaciones de pacientes jóvenes, ambulatorios y sanos. Por este motivo, al establecer los límites de la normalidad de un grupo y aplicarlo a otro con características diferentes se pudo haber introducido un error.

Dentro de los límites de la normalidad *versus* lo deseable

Es posible que todos o algunos de los individuos cuyos resultados se encontraban dentro de los límites de la normalidad tuvieran valores más elevados que los deseables. Recuerde que el intervalo de la normalidad refleja cómo son las cosas y no necesariamente cómo deben ser. Posiblemente, una pérdida de peso que en consecuencia disminuya los valores detectados por la prueba prevenga futuros problemas. Esto supone que la prueba discrimina de hecho a los enfermos de los no enfermos, que perder peso influirá en los valores numéricos de los resultados de la prueba y que la reducción de estos últimos mejorará el pronóstico. Sin embargo, lo que interesa en general es que los resultados que caen dentro del intervalo de la normalidad no son necesariamente los deseables.

EJERCICIO No. 3: DISCRIMINACIÓN DIAGNÓSTICA DE LAS PRUEBAS

Se va a evaluar la utilidad de una nueva prueba para el diagnóstico de la tromboflebitis. La prueba de referencia tradicional ha sido la flebografía y con ella se comparará la nueva prueba. Para valorar la reproducibilidad de la nueva prueba, esta se aplica a 100 pacientes consecutivos con flebografías positivas. Los investigadores observan que 98% de los pacientes diagnosticados de tromboflebitis dan resultados positivos a la prueba. Repiten la prueba en el mismo grupo de pacientes y de nuevo observan que es positiva en 98% de los 100 pacientes. A partir de estos datos concluyen que la reproducibilidad de la nueva prueba es de 100%.

Una vez demostrada la reproducibilidad de la prueba, proceden a estudiar su discriminación diagnóstica, para lo cual deben evaluar el éxito de la prueba en comparación con la flebografía, la prueba de oro o de referencia tradicional. Seguidamente, estudian 1 000 pacientes consecutivos con dolor de piernas unilateral, 500 de los cuales tuvieron flebografías positivas y 500, negativas. Los investigadores clasifican a los individuos como positivos o negativos y presentan los datos del siguiente modo:

PRUEBA NUEVA	FLEBOGRAFÍA POSITIVA	FLEBOGRAFÍA NEGATIVA
Positiva	450	100
Negativa	50	400
	500	500

En este ejemplo los investigadores usaron la definición aceptada de sensibilidad, es decir, la proporción de individuos con resultados positivos en la prueba de referencia que tienen resultados positivos en la nueva prueba. De esta manera,

$$\text{Sensibilidad} = \frac{450}{500} = 0{,}90 = 90\%$$

También usaron la definición aceptada de especificidad, es decir, la proporción de individuos negativos a la prueba de referencia que tienen resultados

negativos en la nueva prueba. De este modo,

$$\text{Especificidad} = \frac{400}{500} = 0{,}80 = 80\%$$

Asimismo, calcularon el valor predictivo de una prueba positiva para su grupo de estudio. La definición aceptada del valor predictivo de un resultado positivo es la proporción de individuos con resultados positivos en la nueva prueba que realmente tienen la enfermedad medida con la prueba de oro. De esta manera,

$$\text{Valor predictivo de una prueba positiva} = \frac{450}{550} = 0{,}818 = 81{,}8\%$$

Sobre la base de estos resultados, los investigadores llegaron a las siguientes conclusiones:

1. La nueva prueba es totalmente reproducible.
2. La nueva prueba es menos sensible y menos específica que la flebografía y, por eso, es una prueba intrínsecamente inferior.
3. Cuando se aplica a un nuevo grupo de pacientes, por ejemplo, a un grupo con dolor de piernas bilateral, se puede esperar que el valor predictivo de un resultado positivo con la nueva prueba sea igual a 81,8%.

CRÍTICA: EJERCICIO No. 3

Cuando una prueba se aplica varias veces a los mismos individuos y en las mismas condiciones, el método para valorar su reproducibilidad exige que los resultados de cada individuo sean prácticamente idénticos si la prueba tiene una reproducibilidad de 100%. Los autores declararon que el total de nuevas pruebas positivas fue idéntico cuando se repitieron. Sin embargo, no indicaron si los mismos individuos que fueron positivos cuando se repitió la prueba también fueron positivos la primera vez. Si los mismos individuos no fueron positivos, la prueba no puede considerarse reproducible. Los autores tampoco indicaron si los que realizaron e interpretaron los resultados de las pruebas repetidas conocían los resultados de la primera prueba.

Una prueba de oro es la medida generalmente aceptada de una enfermedad contra la cual se comparan las pruebas nuevas o todavía no probadas, pero, de hecho, no siempre es una medida ideal de la enfermedad para cuyo diagnóstico ha sido diseñada. Una prueba usada como prueba de oro puede considerarse diagnóstica solo por tradición o por aceptación generalizada de su utilidad. No obstante, es posible que una nueva prueba sea una medida más útil de la enfermedad que la aceptada como referencia. Al comparar la sensibilidad y la especificidad de nuevas pruebas con la de oro debemos ser conscientes de que las discrepancias entre las pruebas pueden ser resultado de la imperfección de la prueba de referencia y no de la deficiencia de la nueva prueba.

Cuando los autores concluyeron que la nueva prueba era menos sensible y específica que la flebografía, estaban suponiendo que esta tiene una sensibilidad y especificidad de 100%. Basándose en esta suposición, es imposible que la nueva prueba tenga una sensibilidad y especificidad más altas que la antigua. Si no estamos seguros de que la flebografía siempre es correcta, es prematuro concluir que la nueva prueba es menos útil para diagnosticar la tromboflebitis. Por lo tanto, los autores debie-

ron haber limitado sus conclusiones sobre la sensibilidad y la especificidad de la nueva prueba a una comparación con la flebografía. Si la nueva prueba fuese más segura, barata o práctica que la flebografía, podría llegar a reemplazar a la flebografía en la práctica clínica. A la larga, la experiencia clínica podría demostrar que es lo suficientemente fiable para ser utilizada como prueba de oro. Mientras tanto, lo mejor que se puede esperar de la prueba es que iguale a la prueba de referencia establecida que, por definición, tiene una sensibilidad y especificidad de 100%.

Los autores midieron correctamente la sensibilidad, la especificidad y el valor predictivo de una prueba positiva en su grupo de estudio. Como afirmaron, el valor predictivo de una prueba positiva es la proporción de los positivos a la nueva prueba que realmente tienen la enfermedad medida según el criterio de oro. En este grupo de estudio, la prevalencia de tromboflebitis fue de 50% (500 con tromboflebitis y 500 sin la enfermedad) y, por consiguiente, el valor predictivo de la prueba es 450 dividido por 550, lo que equivale a 81,8%. Sin embargo, el valor predictivo de una prueba es diferente en distintos grupos de pacientes, dependiendo de la prevalencia o probabilidad de la enfermedad anterior a la prueba en el grupo estudiado. No se puede extrapolar directamente un valor predictivo obtenido en un grupo de pacientes a otro grupo en el cual la prevalencia de la enfermedad es distinta. Es de esperar que un grupo de pacientes con dolor de piernas unilateral tenga una prevalencia de tromboflebitis distinta de otro grupo con dolor bilateral.

Los valores predictivos del dolor de piernas bilateral no se pueden estimar basándose únicamente en la sensibilidad y la especificidad que la prueba ha demostrado en los pacientes con dolor unilateral. Sin embargo, si se puede estimar también el porcentaje de individuos con dolor de piernas bilateral que padecen tromboflebitis, entonces es posible estimar los valores predictivos de una prueba en estos pacientes. Supongamos que la prevalencia de tromboflebitis es mucho más baja en los pacientes con dolor de piernas bilateral. Entonces sería de esperar que una nueva prueba positiva tuviera un valor predictivo positivo mucho más bajo que en un grupo de pacientes con dolor unilateral.

Recuerde que, desde el punto de vista clínico, la prevalencia es lo mismo que la probabilidad de padecer la enfermedad antes de realizar la prueba, y que el valor predictivo de una prueba positiva es la probabilidad de padecerla después de obtener un resultado positivo. Dado que la probabilidad de la tromboflebitis en un paciente que presenta dolor bilateral es menos de 50%, la probabilidad de la enfermedad incluso después de un resultado positivo sería mucho menos de 81,8%.

Sección 3

La tasación de una tasa

CAPÍTULO 20

INTRODUCCIÓN A LAS TASAS

Si usted oye el ruido de cascos, lo más probable es que sea un caballo y no una cebra. Esta metáfora de la clínica señala la obvia pero demasiadas veces olvidada verdad de que las enfermedades comunes se producen frecuentemente y las enfermedades raras, raramente. Cuando los clínicos dicen que una enfermedad es frecuente y otra es rara presuponen una diferencia en las *tasas*.

Todos los clínicos instintivamente utilizan el concepto de tasas. Saben que la enfermedad coronaria es mucho más frecuente en un hombre de mediana edad que en una adolescente. Saben que el cáncer de páncreas es mucho más común en las personas de edad avanzada que en los jóvenes. Saben que la anemia de células falciformes es mucho más probable en una persona de raza negra que en una de raza blanca.

En nuestra discusión anterior sobre las pruebas diagnósticas, señalamos que cuanto más baja es la tasa de prevalencia en una población (o sea, cuanto más rara sea la enfermedad), menor será el valor predictivo de una prueba positiva. Cuando se trata de una enfermedad rara, es menos probable que una prueba positiva indique su presencia. Los clínicos emplean este concepto automática y, quizá, inconscientemente. Saben que es improbable que una mujer joven con cambios en la onda T de un electrocardiograma tenga enfermedad coronaria. Saben que es improbable que un hombre joven con dolor abdominal persistente tenga cáncer de páncreas. Saben que es improbable que una persona joven de raza blanca con dolor articular y anemia tenga anemia de células falciformes. El médico puede apreciar el significado de las tasas sobre la base de su experiencia clínica personal; no obstante, es provechoso que haga uso de los artículos de investigación para mejorar su capacidad de valorar científica y objetivamente las tasas de enfermedad. Esta sección tiene como finalidad ayudar al lector a adquirir los conocimientos necesarios para comprender cómo se miden e interpretan científicamente las tasas de enfermedad. Esta comprensión le puede ayudar a escoger el método diagnóstico apropiado y a interpretar los resultados.

Además de facilitar los diagnósticos individuales, la comprensión del significado de las tasas de enfermedad ayuda a valorar los cambios que se producen con el tiempo o como resultado de las intervenciones médicas. Las tasas de enfermedad son un instrumento importante para realizar los tipos de estudios presentados en la *Sección 1, El estudio de un estudio*. La capacidad para reconocer cambios reales y relaciones de causa y efecto verdaderas depende de que se comprendan los principios básicos de la comparación de tasas.

Es posible que no sea evidente la necesidad de estudiar las tasas de las enfermedades. ¿Por qué no comparar simplemente el número de veces que ocurre un suceso? Examinemos el siguiente ejemplo, que muestra los problemas que pueden surgir si se compara únicamente el número de sucesos.

Un panel de revisión hospitalaria evaluó el rendimiento de los médicos del hospital en que usted trabaja. Encontraron que hubo cinco defunciones entre los 1 000 pacientes que usted atendió en el hospital durante el año pasado. El jefe del equipo tuvo solo una defunción entre los 200 pacientes que trató. El panel decidió que

tener cinco veces tantas defunciones como el jefe del equipo indicaba una práctica deficiente de la medicina.

Ahora bien, no es necesario que se prepare para defenderse diciendo: "Puede parecer que está mal, ¡pero realmente lo hago bien!" En lugar de fijarse en el total de defunciones, es más justo considerar cuántas se produjeron en relación con las que podrían haberse producido. Simplemente tiene que señalar que su tasa de mortalidad y la del jefe de su equipo son idénticas: 5 entre 1 000 es lo mismo que 1 entre 200. Las tasas de los sucesos han venido en su ayuda. ¡Vale la pena conocerlas!

Una probabilidad es una proporción en la cual el numerador es el número de veces que ocurre un suceso y el denominador es el número de veces que podría haber ocurrido. Como en todas las proporciones, el numerador está incluido en el denominador. Las tasas realmente son un tipo especial de medida en la que el denominador también incluye una unidad de tiempo.

En medicina, una función importante de las tasas y de las proporciones es la de caracterizar la historia natural de la enfermedad. Habitualmente se usan tres tipos de medidas:

1. Tasa de incidencia: número de casos nuevos que se producen por unidad de tiempo.
2. Prevalencia: probabilidad de tener una enfermedad en un momento dado.
3. Tasa de letalidad: probabilidad de morir de una enfermedad durante un espacio de tiempo a partir de su diagnóstico.

Las tasas de incidencia se definen del siguiente modo:

$$\text{Tasa de incidencia} = \frac{\text{Número de individuos que desarrollan la enfermedad durante un período}}{\text{Total de personas-año}^{1} \text{ en riesgo}}$$

Con frecuencia es difícil saber cuántos individuos y durante cuánto tiempo están en riesgo de padecer una enfermedad. Por eso, las tasas de incidencia suelen estimarse mediante la siguiente fórmula:

$$\text{Tasa de incidencia de la enfermedad} = \frac{\text{Número de individuos que desarrollan la enfermedad durante un período}}{\text{Número de individuos en el grupo de riesgo en el punto medio del período de interés} \times \text{la duración de dicho período}}$$

Si, por ejemplo, se quiere conocer la tasa de incidencia de los casos de úlceras duodenales en Nueva York en 1990, esta tasa se calcularía teóricamente de la siguiente manera:

$$\text{Tasa de incidencia de úlceras duodenales en Nueva York en 1990} = \frac{\text{Número de residentes de Nueva York que desarrollaron una úlcera duodenal en 1990}}{\text{Número de residentes de Nueva York en riesgo de desarrollar una úlcera duodenal durante 1990} \times \text{1 año}}$$

[1] Una persona-año representa un individuo en riesgo de desarrollar la enfermedad durante un 1 año.

Dado que la población de Nueva York fluctúa constantemente, es difícil saber el número real de personas que residieron en la ciudad y por cuánto tiempo vivieron en ella durante 1990. Para calcular la tasa de incidencia aproximada en ese año, se puede usar el censo de Nueva York del 1 de abril de 1990. La tasa aproximada de incidencia de úlceras duodenales en Nueva York en 1990 se calcularía del siguiente modo:

$$\text{Tasa de incidencia de úlceras duodenales en Nueva York en 1990} = \frac{\text{Número de residentes en Nueva York que desarrollaron una úlcera duodenal en 1990}}{\text{Número de residentes de Nueva York en riesgo de desarrollar úlceras duodenales el 1 de abril de 1990 (aproximadamente igual al número de residentes de Nueva York el 1 de abril de 1990)} \times 1 \text{ año}}$$

El tipo de tasa que hemos comentado hasta el momento es una tasa de *incidencia*, que está relacionada con el riesgo de desarrollar una enfermedad durante un espacio de tiempo. El riesgo es el efecto acumulativo de la tasa de incidencia de la enfermedad durante un período específico. Podemos imaginarnos la incidencia como la velocidad a la que uno se desplaza durante un período breve, y el riesgo, como la distancia que uno ha recorrido durante un largo espacio de tiempo, suponiendo que la velocidad es constante.[2] La tasa de incidencia mide los casos nuevos de una enfermedad determinada que se desarrollan por unidad de tiempo, y esto puede ser de ayuda al examinar la causa o etiología de una enfermedad. El riesgo estimado a partir de la probabilidad de desarrollar una enfermedad en un período específico puede contribuir a predecir los sucesos futuros, si se usa con precaución. La enfermedad, una vez desarrollada, puede durar mucho tiempo. Por eso, frecuentemente se usa un segundo tipo de medida que estima la probabilidad de *tener* la enfermedad en un momento determinado. Esta se conoce como *prevalencia* y mide lo frecuente o prevaleciente que es una enfermedad en un momento dado. La prevalencia es muy importante en el diagnóstico, dado que es el punto de partida para estimar la probabilidad anterior a la prueba de que la enfermedad se halle presente. Asimismo, proporciona una estimación de la probabilidad de que la enfermedad esté presente antes de evaluar la historia individual, el examen físico o las pruebas de laboratorio. De esta forma,

$$\text{Prevalencia} = \frac{\text{Número de individuos que tienen la enfermedad en un momento dado}}{\text{Número de individuos que forman parte del grupo en ese momento}}$$

En el ejemplo anterior, la prevalencia de las úlceras duodenales el 1 de abril de 1990 en la Ciudad de Nueva York se calcularía como sigue:

$$\text{Prevalencia} = \frac{\text{Número de residentes de Nueva York con úlceras duodenales el 1 de abril de 1990}}{\text{Número de residentes de Nueva York el 1 de abril de 1990}}$$

[2] Tanto los bioestadísticos como los epidemiólogos establecen una diferencia entre la incidencia acumulativa y las tasas de incidencia. En la incidencia acumulativa, el denominador es el número de individuos en la población al principio de un período determinado. Incidencia acumulativa es sinónimo de riesgo.

Para la mayor parte de las enfermedades, la tasa de incidencia y la de prevalencia se relacionan aproximadamente de la siguiente manera:

$$\text{Prevalencia} = \text{Tasa de incidencia} \times \text{Duración media de la enfermedad}$$

En otras palabras, cuanto más larga sea la duración de la enfermedad, más individuos tendrán la enfermedad en un momento dado y, por lo tanto, más alta será la prevalencia. Las enfermedades crónicas de larga duración, como la diabetes, pueden tener una tasa de incidencia baja, pero una prevalencia elevada en un momento determinado. Las enfermedades agudas de corta duración, como la faringitis estreptocócica, pueden tener una tasa de incidencia elevada, pero una prevalencia baja en un momento dado. Por eso, es importante saber que la prevalencia y la incidencia miden fenómenos distintos. Las tasas de incidencia miden la frecuencia con que se desarrolla un nuevo caso de la enfermedad por unidad de tiempo. La prevalencia mide la probabilidad de *tener* la enfermedad en un momento determinado. Si no se aprecia esta diferencia, se puede cometer el tipo de error que se ilustra con el siguiente ejemplo.

En un estudio sobre la gonorrea asintomática en hombres se tomaron muestras de 1 000 sujetos seleccionados al azar. Se les diagnosticó gonorrea a 10 de ellos. En un segundo estudio, se siguió a un grupo de hombres de la misma población durante un año. Se observó que durante ese lapso de tiempo solo uno de los hombres desarrolló gonorrea asintomática. Al comparar estos estudios, un revisor concluyó que uno de los dos tenía que estar equivocado, ya que las conclusiones eran contradictorias.

Esta aparente incongruencia desaparece, si se distingue entre la tasa de incidencia y la prevalencia. El primer estudio de los casos existentes midió la prevalencia, mientras que el segundo valoró la incidencia. El hecho de que la prevalencia sea mucho más elevada que la incidencia sugiere que la gonorrea asintomática es de larga duración. Esto puede explicarse por el hecho de que, aunque los casos sintomáticos suelen recibir tratamiento, los asintomáticos permanecen en la comunidad sin tratamiento durante un período prolongado.

Además de la tasa de incidencia y de la prevalencia, es necesario definir una tercera medida para caracterizar la historia natural de la enfermedad. Esta medida se conoce como *letalidad*.

$$\text{Letalidad}^3 = \frac{\text{Número de personas fallecidas por una enfermedad durante un período}}{\text{Número de personas diagnosticadas de la enfermedad al inicio del período}}$$

A diferencia de las tasas de incidencia, la letalidad está influida por los éxitos de las intervenciones médicas destinadas a curar las enfermedades. La letalidad es útil para valorar el pronóstico, porque mide la probabilidad de no sobrevivir una vez iniciada la enfermedad. La letalidad durante un período tiene una relación impor-

[3] La letalidad es una proporción que se refiere a la probabilidad de morir de una enfermedad. Cuando esta proporción se multiplica por la tasa de incidencia, se obtiene la tasa de mortalidad.

tante con las tasas de mortalidad de una enfermedad determinada (esto es, el número de defunciones debidas a una enfermedad por persona-año).

$$\text{Tasa de mortalidad} = \text{Tasa de incidencia} \times \text{Letalidad}$$

El no valorar esta relación puede conducir a la confusión que se describe en el siguiente ejemplo.

En un estudio de las tasas de la úlcera duodenal, los autores calcularon correctamente las tasas de mortalidad por úlcera duodenal en los Estados Unidos en 1949 y en 1989. En 1949, la tasa anual de mortalidad fue 5 por 1 000 000 personas-año. Estudios posteriores revelaron que ni la tasa de incidencia ni la prevalencia habían cambiado. Los autores no pudieron interpretar estos datos.

Conociendo la relación que existe entre las tasas de mortalidad y las de incidencia, se entiende que el descenso de las tasas de mortalidad debió ser causado por una reducción de la letalidad. Este descenso de la letalidad puede reflejar el progreso conseguido durante 40 años en el tratamiento de las úlceras duodenales, aunque no se haya progresado en la reducción de la incidencia (casos nuevos) de la enfermedad.

Las tasas de incidencia, la prevalencia y la letalidad miden, respectivamente, la tasa de desarrollo de los casos nuevos de una enfermedad por unidad de tiempo, la probabilidad de tener la enfermedad en un momento dado y la probabilidad de morir por una enfermedad, una vez diagnosticada. Además de estas medidas básicas, en la literatura médica se utiliza con frecuencia una medida conocida como *razón de mortalidad proporcional*, que se define como:

$$\text{Razón de mortalidad proporcional} = \frac{\text{Número de individuos fallecidos por una enfermedad}}{\text{Número de individuos fallecidos por todas las enfermedades}}$$

La razón de mortalidad proporcional mide la probabilidad de que una defunción se deba a una causa determinada. Las razones de mortalidad proporcional son una herramienta útil para determinar cuáles son las causas de muerte más frecuentes. Sin embargo, no nos informan sobre la probabilidad de morir, como muestra el siguiente ejemplo.

Un estudio bien diseñado reveló que los traumatismos fueron la causa de muerte de 4% de las personas mayores de 65 años y que causaron 25% de los fallecimientos entre los menores de 3 años. Los autores llegaron a la conclusión de que los mayores de 65 años tenían una probabilidad mucho menor de morir por traumatismos que los menores de 3 años.

El hecho de que la razón de mortalidad proporcional en los mayores de 65 años sea menor por traumatismos no significa necesariamente que los ancianos tengan una menor probabilidad de morir por esa causa. Dado que entre los mayores de 65 años se producen muchos más fallecimientos, aun 4% de las muertes por traumatismos pueden representar una tasa de mortalidad cercana a la tasa de mortalidad de los menores de 3 años.

Habiendo ya examinado los tipos de tasas y proporciones que se encuentran con mayor frecuencia en la literatura médica, y distinguido esas medidas de las razones, centraremos nuestra atención en los métodos para calcular las tasas de enfermedad.

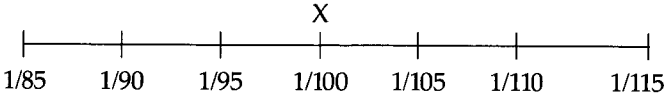
CAPÍTULO 21

MUESTREO DE TASAS

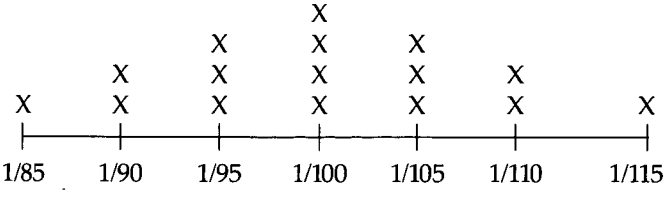
En algunas circunstancias es posible determinar todos los casos de una enfermedad en una población. De ordinario pueden obtenerse tasas de mortalidad para una población, porque los certificados de defunción son documentos legales obligatorios. En consecuencia, ello también permite calcular las tasas de mortalidad de una enfermedad para toda la población. Sin embargo, para la mayor parte de las enfermedades no es factible contar todos los casos en la población y, por esta razón, las técnicas de muestreo son muy útiles. El *muestreo* (*sampling*) es una técnica mediante la cual el investigador selecciona al azar una porción representativa de la población, estudia esa muestra y luego intenta extrapolar los resultados a toda la población escogida para el estudio.

ERROR MUESTRAL

El proceso de extracción de una muestra no es perfecto, aunque se realice correctamente. Para valorar dicho proceso y el error intrínseco introducido por el muestreo es preciso comprender el principio que fundamenta esta técnica. Dicho principio afirma que si se extraen muchas muestras al azar, las tasas calculadas con los datos de esas muestras serán iguales, en promedio, a la tasa de la población original. En otras palabras, cada muestra puede tener una tasa mayor o menor que la de la población original. Por ejemplo, la siguiente figura muestra una tasa de 1 por 100 en la población original,



y si se extraen muestras de esta población, las tasas podrían ser las siguientes:



Observe que, si bien algunas de las tasas obtenidas en determinadas muestras son iguales que las de la población original, otras son mayores o menores. Dado que las muestras solo son exactas en promedio, se dice que una muestra determinada posee un error muestral intrínseco. Si no se aprecia la existencia del error muestral, se puede incurrir en la siguiente interpretación errónea:

Una organización nacional intentó estimar la prevalencia de los portadores de estreptococos realizando cultivos en una muestra aleatoria de 0,1% de todos los niños de la nación. Para verificar los resultados, la misma organización extrajo una segunda muestra aleatoria de 0,1% de los escolares y realizó una segunda encuesta con el mismo protocolo. El primer examen reveló una prevalencia de cultivos positivos de 15 por 1 000, mientras que el segundo examen reveló una prevalencia de 10 por 1 000. Los autores concluyeron que estos resultados inconsistentes eran imposibles, dado que habían usado el mismo método.

Los autores no tuvieron en cuenta el hecho de que el muestreo tiene un error intrínseco. Este error muestral puede explicar las diferencias observadas entre las dos muestras. El ejemplo simplemente señala que dos muestras extraídas de la misma manera pueden producir resultados diferentes solo a causa del azar. Recuerde que un elevado número de muestras proporcionan, en promedio, estimaciones que son idénticas al verdadero valor de la población, pero que puede haber una amplia variación entre dos muestras y entre estas y el verdadero valor de la población.

EL TAMAÑO DE LA MUESTRA

Un segundo principio importante y necesario para comprender el muestreo afirma que cuantos más individuos formen parte de la muestra, más probable será que la tasa estimada con los datos de la muestra se aproxime a la tasa poblacional. Por esta razón, el tamaño de la muestra condiciona la proximidad de la tasa muestral a la poblacional. Esto no es sorprendente dado que, cuando todo el mundo forma parte de la muestra, está garantizado que la tasa muestral es igual a la poblacional.

Examinemos con más detalle este principio. El factor que más influye en la magnitud del error muestral es el tamaño de la muestra. La relación entre el tamaño muestral y la precisión no es de uno a uno, sino que es una función de la raíz cuadrada. Con muestras pequeñas, el aumento del tamaño muestral aumenta notablemente la precisión del estimador muestral de la tasa poblacional. Sin embargo, a medida que aumenta el tamaño de la muestra, la mejora de la precisión disminuye de forma relativa y los aumentos pequeños o moderados del tamaño muestral añaden poco a la precisión del estimador. Por lo tanto, los investigadores intentan equilibrar la necesidad de precisión con los costos económicos que acarrea el aumento del tamaño muestral. La consecuencia del empleo de muestras pequeñas es que estas pueden variar mucho entre sí y en relación con el verdadero valor de la población. El siguiente ejemplo ilustra la necesidad de tener en cuenta los efectos del tamaño muestral en los resultados del muestreo.

Un investigador extrajo una muestra de 0,01% de los certificados de defunción de la nación y encontró que la tasa de mortalidad por cáncer de páncreas era de 50 por 100 000 personas-año. Un segundo investigador, que extrajo una muestra de 1% de dichos certificados, llegó a la conclusión de que la verdadera tasa de mortalidad era de 80 por 100 000 personas-año. Para resolver esta discrepancia, el segundo investigador identificó todas las defunciones causadas por cáncer de páncreas en el país y obtuvo una tasa de 79 por 100 000 personas-año. Finalmente, llegó a la conclusión de que el primer investigador había realizado su estudio de forma fraudulenta.

El primer estudio empleó una muestra que era la centésima parte de la segunda; por lo tanto, es posible que el error muestral del primer estudio fuera mucho mayor. El hecho de que la segunda muestra fuera más exacta se debe probablemente al aumento de precisión que le confiere su mayor tamaño y no a fraude en el primer estudio.

EL MUESTREO ALEATORIO

Aunque se han esbozado dos principios importantes del muestreo, todavía queda un aspecto por considerar. Esos dos principios se basan en el supuesto de que la muestra se ha obtenido al azar, lo que significa que todos los individuos de la población tenían la misma probabilidad (o, al menos, una probabilidad conocida) de ser seleccionados para su inclusión en la muestra. Si no se realiza un muestreo al azar, no se puede estimar de forma precisa la proximidad de los resultados muestrales a los de la población. La necesidad de seleccionar las muestras al azar se ilustra en el siguiente ejemplo:

Un investigador de un hospital comarcal estimó que la tasa de infarto de miocardio de su comunidad era de 150 por 100 000 personas-año. Otro investigador de un hospital privado de la misma comunidad estimó que dicha tasa era de 155 por 100 000 personas-año. Dado que sus resultados eran similares, los investigadores concluyeron que la tasa de infarto de miocardio en su comunidad debía situarse entre 150 y 155 por 100 000 personas-año.

En ninguno de los dos estudios se intentó obtener una muestra al azar de la población. Es posible que los pacientes de infarto de miocardio hubiesen escogido selectivamente esos dos hospitales o los hubiesen evitado. Si se hubieran incluido en la muestra todos los hospitales de la zona, la tasa de infarto de miocardio podría haber sido totalmente distinta. Las tasas en este ejemplo se calcularon a partir de datos disponibles, y esto se conoce como *muestreo fortuito* (*chunk sampling*). Este es el tipo de muestreo más simple, dado que los investigadores calculan las tasas a partir de datos fácilmente disponibles. Sin embargo, la falta de representatividad de las muestras fortuitas implica que los resultados obtenidos a partir de las mismas pueden no ser fiables o fácilmente extrapolables a la población.[4]

Cuando se extrae una muestra al azar, los investigadores suelen procurar que todos los individuos de la población tengan la misma probabilidad de ser seleccionados para inclusión en la muestra. Esto se conoce como *muestreo aleatorio simple* (*simple random sampling*). Muchos investigadores han observado que, si se basan exclusivamente en el muestreo aleatorio, no conseguirán incluir suficientes individuos que posean la característica de interés para el estudio. Por ejemplo, si los investigadores estudian las tasas de hipertensión en los Estados Unidos, es posible que estén interesados especialmente en las tasas de hipertensión de las personas de raza negra o de las orientales. Si simplemente extraen una muestra aleatoria, es posible que no incluyan un número suficiente de orientales o de negros. Por lo tanto, los investigadores podrían extraer muestras por separado de los negros, los orientales y el resto de la población. Este procedimiento de obtención por separado de muestras al azar de los diferentes subgrupos o estratos se conoce como *muestreo aleatorio estratificado* (*stratified random sampling*). Esto es permisible, y muchas veces deseable, siempre que el muestreo dentro de cada grupo sea aleatorio. Existen métodos estadísticos distintos para las muestras estratificadas.

Revisemos los principios y los requisitos del muestreo:

[4] No obstante, en ocasiones nos vemos obligados a realizar extrapolaciones a partir de muestras fortuitas. El ejemplo más común ocurre cuando deseamos extrapolar observaciones de investigaciones a los pacientes que vemos más tarde. El paso del tiempo es un aspecto de una población que no se puede muestrear aleatoriamente.

1. En promedio, las muestras aleatorias de una población tendrán la misma tasa que la población original. Sin embargo, existe un error muestral intrínseco introducido al incluir solamente una parte de la población.

2. La magnitud del error muestral está influido por el tamaño de la muestra obtenida. El aumento del tamaño muestral reduce la magnitud del error muestral, pero el aumento de la precisión desciende a medida que incrementamos el tamaño de la muestra.

3. Los principios del muestreo se basan en el supuesto de que las muestras se obtienen al azar. Mediante el muestreo estratificado es posible garantizar un número suficiente de casos en cada categoría de interés. No obstante, el muestreo debe ser aleatorio en cada categoría o estrato. Si no se realiza un muestreo al azar, no existe ningún método que relacione con precisión la tasa obtenida en la muestra con la verdadera tasa de la población de la que se ha extraído.

CAPÍTULO 22

ESTANDARIZACIÓN DE TASAS

En el capítulo anterior esbozamos los requisitos para calcular tasas con exactitud mediante las técnicas de muestreo aleatorio. Ahora trataremos de comparar las tasas calculadas. Supondremos que las tasas se han calculado correctamente y mostraremos las precauciones que deben tomarse al compararlas, incluso las que se han calculado mediante técnicas de muestreo aleatorio adecuadas.

Compararemos las tasas de muestras extraídas de dos grupos diferentes. Estos grupos pueden ser dos hospitales, dos condados, dos países, dos fábricas, o el mismo hospital, ciudad o fábrica comparado en dos momentos distintos en el tiempo. Compararemos las tasas para determinar la magnitud de las diferencias entre las tasas de las poblaciones o el grado de cambio de las tasas con el tiempo. Estas comparaciones son importantes para las tasas que se calculan a partir de datos obtenidos por muestreo, así como para las calculadas a partir de toda la población.

Cuando se utilizan tasas para comparar probabilidades o riesgos de enfermedad es importante considerar si las poblaciones difieren en algún factor que se sabe que influye en el riesgo de contraer la enfermedad. Esta consideración corresponde al ajuste según las variables de confusión discutido anteriormente.

Es posible que al realizar un estudio, el investigador ya sepa que factores como la edad, el sexo o la raza influyen en el riesgo de desarrollar una enfermedad determinada. En este caso, el investigador ajustaría o "estandarizaría" las tasas según esos factores. La importancia de la estandarización se puede apreciar considerando las tasas del cáncer de pulmón. Dado que la edad es un factor de riesgo conocido para el cáncer de pulmón, se gana poco descubriendo que una comunidad de jubilados tiene una tasa de cáncer de pulmón más elevada que el resto de la comunidad. De forma similar, si una fábrica tiene una fuerza laboral más joven que otra, es erróneo comparar directamente las tasas de cáncer de pulmón de las dos fábricas, sobre todo si uno desea extraer conclusiones sobre la seguridad de las condiciones de trabajo.

Para evitar este problema, las tasas de enfermedad se pueden ajustar para tener en cuenta los factores que ya se sabe que influyen notablemente en el riesgo. Este proceso de ajuste se denomina *estandarización* (*standardization*). La edad es el factor que requiere estandarización con más frecuencia, pero podemos ajustar según cualquier factor del que sepamos que produce un efecto. Por ejemplo, al usar las técnicas de estandarización para comparar tasas de hipertensión de dos muestras, con objeto de estudiar la importancia de las diferencias en el suministro de agua, se puede ajustar según la raza, ya que se sabe que la tasa de hipertensión de las personas de raza negra es más elevada.

El principio utilizado en la estandarización de las tasas es el mismo que se emplea para ajustar según las diferencias entre los grupos de estudio, tal como se comentó en la *Sección 1, El estudio de un estudio*. Los investigadores comparan las tasas entre individuos que son similares en cuanto a la edad o a otros factores según los cuales se ajustarán los datos. Antes de mostrar un ejemplo del método, veamos cuán erróneos pueden ser los resultados si no se realiza una estandarización.

En un estudio se comparó la incidencia de cáncer de páncreas en los Estados Unidos con la de México. La tasa por 100 000 personas-año en los Estados Unidos fue tres veces mayor que la de México. Los autores llegaron a la conclusión de que los estadounidenses tenían un riesgo tres veces más elevado de padecer cáncer de páncreas que los mexicanos, suponiendo que el diagnóstico fuera igualmente exacto en ambos países.

La interpretación de este estudio es superficialmente correcta; si los datos son fiables, el riesgo de cáncer de páncreas es mayor en los Estados Unidos. Sin embargo, se sabe que el cáncer de páncreas se produce con mayor frecuencia en las personas ancianas. Es posible que el hecho de que la población mexicana sea más joven explique la diferencia entre las tasas de cáncer de páncreas. Esta puede ser una cuestión importante si estamos examinando la causa de esta enfermedad. Si la distribución de edad no explica esta diferencia, los autores habrán detectado una diferencia importante e inesperada que exige otra explicación. Por este motivo, los autores deben estandarizar sus datos según la edad y observar si persisten las diferencias.

La estandarización de las tasas frecuentemente se realiza comparando una muestra especial que se está estudiando con la población general. Para realizar este tipo de estandarización empleamos a menudo el denominado *método indirecto* (*indirect method*). Mediante este método se compara el número de sucesos observado, como las defunciones, en la muestra de interés con el número de sucesos que serían de esperar si la muestra estudiada tuviese la misma distribución de edad que la población general. Cuando la muerte es el desenlace de interés, el método indirecto permite calcular una razón conocida como *razón de mortalidad estandarizada* (*standardized mortality ratio*).

$$\text{Razón de mortalidad estandarizada} = \frac{\text{Número observado de muertes}}{\text{Número esperado de muertes}}$$

La razón de mortalidad estandarizada es un instrumento útil para comparar una muestra extraída de una población de interés con la población general. Sin embargo, cuando se interpreta esta razón es importante recordar que a menudo no se espera que una población especial en estudio tenga la misma tasa de mortalidad que la población general.

Por ejemplo, cuando se compara un grupo de empleados con la población general, se debe recordar que, en muchos casos, el estar empleado requiere estar sano o, al menos, no estar incapacitado. La necesidad de tener en cuenta este efecto del empleo se ilustra en el ejemplo que figura a continuación.

En un estudio sobre nuevos empleados de una industria química, la razón estandarizada de mortalidad por todas las causas de muerte fue 1. El investigador concluyó que, como la razón de mortalidad estandarizada era 1, la industria química no presentaba riesgos para la salud de sus trabajadores.

Para interpretar este estudio, es importante recordar que los nuevos empleados por lo común son más sanos que las personas de la población general. Por esta razón, sería de esperar que tuvieran una tasa de mortalidad más baja que la de la población general.[1] Por consiguiente, esta razón de mortalidad estandarizada de 1 ó 100% puede no reflejar los riesgos a que están expuestos estos trabajadores sanos.

[1] N del E. Este hecho se denomina *efecto del trabajador sano* (*healthy worker effect*).

CUADRO 22-1. Comparación de tasas de cáncer de vejiga urinaria

Grupo de edad (años)	Número de sujetos	Número de casos de cáncer de vejiga urinaria	Tasa de cáncer de vejiga urinaria en cada grupo de edad[a]
		FÁBRICA A	
20–30	20 000	0	0 por 100 000
30–40	20 000	10	50 por 100 000
40–50	30 000	20	67 por 100 000
50–60	20 000	80	400 por 100 000
60–70	10 000	90	900 por 100 000
Total	100 000	200	200 por 100 000
		FÁBRICA B	
20–30	10 000	0	0 por 100 000
30–40	10 000	4	40 por 100 000
40–50	20 000	6	30 por 100 000
50–60	50 000	140	280 por 100 000
60–70	10 000	50	500 por 100 000
Total	100 000	200	200 por 100 000

[a] La tasa se obtiene a partir del número de casos y el número de sujetos en el grupo de edad. Las tasas no se pueden sumar a lo largo de la columna.

Cuando se estudian dos grupos de una población o cuando se valoran los *cambios temporales* de una sola población, es preferible utilizar el método directo de estandarización. El método directo funciona de la siguiente manera: Supongamos que los investigadores desean comparar las tasas de cáncer de vejiga urinaria en dos grandes industrias. Los datos del cáncer de vejiga en las dos industrias se presentan en el cuadro 22-1. Observe que las tasas globales de ambas muestras son de 200 por 100 000 trabajadores. Note también que las tasas de cada grupo de edad son tan elevadas o más en la fábrica A que en la B. A pesar de que las tasas de la fábrica B son más bajas, en un primer momento puede parecer sorprendente que las tasas globales sean las mismas. No obstante, examinando el número de indidivuos en cada grupo de edad, es obvio que la fábrica A tiene una fuerza de trabajo más joven que la B. La fábrica B tiene 60 000 trabajadores de 50 a 70 años de edad, mientras que la A solo tiene 30 000 en esos grupos de edad. Como se sabe que el cáncer de vejiga aumenta con la edad, la juventud de la fuerza laboral de la fábrica A reduce la tasa global de esa empresa. Por este motivo, es erróneo examinar solo las tasas globales, ya que la de la industria B se eleva como consecuencia de una estructura de edad más madura. Esto es especialmente aplicable cuando la seguridad del medio ambiente de la fábrica es la cuestión de mayor interés.

Para evitar este problema, los autores deben estandarizar las tasas para ajustarlas según las diferencias de la edad y, de este modo, hacer una comparación más correcta. Para realizar la estandarización, se subdivide cada muestra para indicar el número de individuos, el de casos de la enfermedad y la tasa de incidencia en cada grupo de edad. Cuando se dividen los datos en grupos según una característica como la edad, cada división se denomina un *estrato* (*strata*). El resultado de esta división se muestra en el cuadro 22-1.

CUADRO 22-2. Método de estandarización según la edad

Grupo de edad (años)	Tasa de cáncer de vejiga en la fábrica A	Número de sujetos en la fábrica B	Número de casos que se producirían en la fábrica A si la distribución de edades fuera igual a la de la fábrica B[a]	Número de casos de cáncer de vejiga que realmente se produjeron en la fábrica B
20–30	0/100 000	10 000	0	0
30–40	50/100 000	10 000	5	4
40–50	67/100 000	20 000	13	6
50–60	400/100 000	50 000	200	140
60–70	900/100 000	10 000	90	50
Total		100 000	308	200

[a] Los valores de esta columna se calculan multiplicando los de las columnas precedentes.

A continuación, los autores han de determinar cuántos casos de cáncer de vejiga urinaria hubieran aparecido en la fábrica A si su estructura de edad fuese igual a la de la fábrica B. Seguidamente se detallan las etapas de este proceso.[2]

1. Empezando con el grupo de edad de 20 a 30 años, los autores multiplican la tasa de cáncer de ese grupo en la fábrica A por el número de individuos en el grupo de edad correspondiente de la fábrica B. Este producto es el número de casos que se hubieran producido en la fábrica A si hubiera tenido el mismo número de individuos en ese grupo de edad que la B.

2. A continuación, los autores efectúan este cálculo para cada grupo de edad y suman los totales de casos de los diferentes grupos. Esto produce el total de casos que habrían aparecido si la fábrica A hubiera tenido la misma distribución de edad que la B.

3. Los autores han estandarizado las tasas según la edad y ahora pueden comparar directamente el número de casos que han aparecido en la fábrica B con el de los que se habrían producido en la A si hubiera tenido la misma distribución de edad que la B. Los autores ahora ya han ajustado la fábrica A según la edad a la distribución de edades de la fábrica B.[3]

Vamos a aplicar este procedimiento a los datos del cáncer de vejiga urinaria, tal como se muestra en el cuadro 22-2.

En la fábrica A se habrían producido 308 casos de cáncer de vejiga urinaria si su distribución de edad hubiera sido la misma que en la fábrica B, pero en realidad en la B solo se produjeron 200. Estos resultados constituyen mejores medidas para comparar el riesgo de los trabajadores de desarrollar cáncer de vejiga en cada industria que las frecuencias no ajustadas. Las cifras ajustadas acentúan el hecho de que, a pesar de la igualdad de las tasas globales, la fábrica A tiene una tasa igual o más alta en cada grupo de edad. Por lo tanto, para realizar comparaciones equitativas entre po-

[2] El método que se presenta no es necesariamente el único o el mejor para realizar la estandarización. Por razones estadísticas, es habitual ponderar los estratos según la inversa de la varianza del estimador en cada estrato, como se realiza cuando se utiliza el método de Mantel-Haenszel.

[3] También es posible ajustar según la edad en la dirección opuesta; es decir, ajustar según la edad la fábrica B a la distribución de edad de la fábrica A. Aunque la conclusión general hubiese sido la misma, las estimaciones habrían sido distintas.

CUADRO 22-3. Comparación de las tasas de mortalidad por fibrosis quística, 1969 y 1989

Grupo de edad (años)	Tasa de mortalidad	Población	Número de defunciones
	1969		
0–10	5/100 000	1 000 000	50
10–20	10/100 000	1 000 000	100
20–40	1/100 000	2 000 000	20
		4 000 000	170
	1989		
0–10	3/100 000	1 000 000	30
10–20	6/100 000	1 000 000	60
20–40	4/100 000	2 000 000	80
		4 000 000	170

blaciones que difieren en su estructura de edad y en las que se sabe que la edad influye en el riesgo de padecer la enfermedad, es preciso estandarizar los resultados según esta variable. Si se conocen otros factores que influyen en las tasas, se puede aplicar el mismo proceso para estandarizar o ajustar según esos factores.

La estandarización produce medidas resumidas de grandes cantidades de datos, razón por la cual es tentador estandarizar los datos siempre que se comparan dos grupos.

No obstante, observe que, al estandarizar, los cálculos dan un peso mayor o "ponderan" a los subgrupos más numerosos. Por eso, cuando se produce un cambio importante en un solo subgrupo, especialmente si es pequeño, este efecto puede ser ocultado por el proceso de estandarización. Además, a veces puede que se haya logrado progresar en la disminución de la mortalidad por una causa en los grupos más jóvenes, con aumento en los de más edad. Como se puede observar en el siguiente ejemplo, cuando la mortalidad por una causa determinada se desplaza hacia los grupos de edad más avanzada, el efecto puede quedar encubierto por la estandarización.

Se realizó un estudio sobre las tasas de mortalidad por fibrosis quística, para determinar si los progresos en su tratamiento durante la niñez se reflejaban en las tasas de mortalidad de un estado de gran tamaño con una distribución de población estable. En el cuadro 22-3 se presentan los datos de 1969 y 1989.

Observe que en este ejemplo la mortalidad por fibrosis quística en los grupos de edad de 0 a 10 años y de 10 a 20 años descendió entre 1969 y 1989. Sin embargo, la mortalidad por esta enfermedad aumentó en el grupo de 20 a 40 años durante el mismo período. Este aumento se contrapesa con el descenso entre los más jóvenes, de forma que las tasas de mortalidad global en ambos años fueron 170 por 4 000 000 ó 4,25 por 100 000 personas-año. Resulta tentador intentar estandarizar esos datos y obtener una medida ajustada de la tasa de mortalidad. Sin embargo, si estandarizamos aplicando la distribución de edad de 1989 a las tasas de mortalidad de 1969 (o viceversa), los resultados después del ajuste no serían distintos de los anteriores al ajuste. Esto sucede porque las distribuciones de edad de las poblaciones de 1969 y de 1989 son iguales.

Lamentablemente, la estandarización no nos ayuda a apreciar lo que ocurre en este caso. Tanto las tasas brutas o no ajustadas como las estandarizadas

oscurecen el hecho de que se ha producido un descenso importante de las tasas de mortalidad en los grupos de 0 a 10 y de 10 a 20 años de edad. Para darse cuenta de este cambio, es preciso examinar directamente los datos reales de cada grupo de edad.[4]

Es importante darse cuenta de que tanto las tasas brutas como las ajustadas pueden no revelar diferencias o cambios que solo se producen en uno o en pocos grupos de edad. Sobre todo, es probable que los cambios en un grupo pasen desapercibidos cuando los otros grupos de edad cambian en la dirección opuesta.

Una situación en la que se observan con frecuencia cambios en dirección opuesta es el retraso de la muerte hasta una edad más avanzada sin que se logre la curación. Es importante entender este principio para valorar el error cometido en el siguiente estudio:

Un investigador estudió un nuevo tratamiento para el cáncer de mama, que en promedio prolongaba 5 años la supervivencia en el estadio 2 de la enfermedad. Con toda confianza predijo que, si su tratamiento se aplicaba ampliamente, la tasa global de cáncer de mama descendería sobremanera en los siguientes 20 años.

Este investigador no se dio cuenta de que, cuando se prolonga la vida pero la muerte se retrasa a edades más avanzadas, las tasas de mortalidad globales ajustadas según la edad no mejoran necesariamente. A pesar del éxito de este nuevo tratamiento, los autores no están afirmando que curará la enfermedad. Cuando solo se prolonga la vida, los enfermos pueden morir de la enfermedad a una edad más avanzada. En este caso, las tasas de mortalidad por cáncer de mama pueden descender en los grupos de edad más jóvenes y aumentar en los de edad más avanzada. Por esta razón, las tasas de mortalidad ajustadas según la edad a veces no revelan los progresos realizados.

---

[4] Además, es posible aplicar métodos estadísticos con los que se obtienen medidas conocidas como la *esperanza de vida* (*life expectancy*). La esperanza de vida tiene en cuenta el impacto del aumento de años de vida vividos incluso en ausencia de una curación. Esta medida se obtiene a partir de tablas de vida transversales que valoran la probabilidad hipotética de supervivencia del individuo promedio que alcanza una edad determinada. Suponiendo una experiencia de mortalidad estable, la esperanza de vida se basa en la experiencia de la mortalidad de una población concreta, como la de los Estados Unidos, en un año determinado.

CAPÍTULO 23

ORÍGENES DE LAS DIFERENCIAS ENTRE TASAS

En este capítulo supondremos que se ha demostrado que dos grupos tienen tasas distintas. Estas diferencias entre tasas pueden representar diferencias entre dos grupos o diferencias que se producen con el tiempo en un solo grupo.

CAMBIOS O DIFERENCIAS DEBIDAS A ARTEFACTOS

Las diferencias entre las tasas pueden ser el resultado de cambios reales en la historia natural de la enfermedad o reflejar cambios o diferencias en el método mediante el cual se valora la enfermedad considerada. Las *diferencias por artefactos* implican que, aunque existe una diferencia, esta no refleja cambios en la enfermedad, sino simplemente en la forma en que se mide, busca o define la enfermedad.

Los cambios por artefactos suelen proceder de tres fuentes:

1. Cambios en la *capacidad* de identificar la enfermedad. Representan cambios en la medición de la enfermedad.
2. Cambios en los *esfuerzos* para reconocer la enfermedad. Estos pueden representar esfuerzos para identificar la enfermedad en un estadio más temprano, cambios de los requisitos para su notificación o nuevos incentivos para detectarla.
3. Cambios en la *definición* de la enfermedad. Representan cambios en la terminología utilizada para definir la enfermedad.

El siguiente ejemplo ilustra el primer tipo de cambio por artefactos, es decir, el efecto de un cambio en la capacidad para identificar la enfermedad.

Debido al reciente aumento del interés en el prolapso de la válvula mitral, se realizó un estudio de la prevalencia de esta enfermedad. Mediante un examen completo de todas las historias clínicas de una importante clínica universitaria de cardiología se encontró que en 1969 solo se le había diagnosticado prolapso de la válvula mitral a 1 de cada 1 000 pacientes, mientras que en 1989 se le hizo este diagnóstico a 80 de cada 1 000 pacientes. Los autores concluyeron que la prevalencia de la enfermedad estaba aumentando a una velocidad asombrosa.

Entre 1969 y 1989, el uso de la ecocardiografía aumentó sobremanera la capacidad de documentar el prolapso de la válvula mitral. Además, el reconocimiento creciente de la frecuencia de esta enfermedad ha conducido a identificarla mucho mejor mediante la exploración física. Por eso no es sorprendente que en 1989 se reconociera una proporción mucho más elevada de pacientes de la clínica cardiológica con prolapso de la válvula mitral que en 1969. Es posible que si los conocimientos y la tecnología de 1969 hubieran sido similares a los de 1989, las tasas hubieran sido prácticamente idénticas. Este ejemplo demuestra que los cambios por artefactos pueden explicar grandes diferencias en las tasas de una enfermedad, incluso cuando se hace una revisión completa de todos los casos para calcular su prevalencia.

Los cambios relacionados con los esfuerzos para diagnosticar una enfermedad se pueden producir cuando los médicos intentan diagnosticarla en un estadio temprano. Si se introduce un programa de tamizaje para diagnosticar una enfermedad en un estadio asintomático, es más probable que los individuos sean diagnos-

ticados más tempranamente. Los esfuerzos para diagnosticar tempranamente una enfermedad producen un mayor "adelanto" (*lead time*) en el diagnóstico en comparación con el método anterior. Por otra parte, cuando se dispone de un tratamiento que surte mejor efecto si se utiliza en las etapas tempranas de la enfermedad, se puede mejorar el pronóstico. Sin embargo, si no existe un tratamiento adecuado o el tratamiento temprano no mejora el pronóstico, lo único que se logra es aumentar el intervalo entre el diagnóstico y la muerte. Esto puede producir una diferencia por artefacto si la tasa de muerte se mide en el período inmediatamente posterior al diagnóstico. A continuación se muestra un ejemplo de sesgo debido al diagnóstico adelantado:

Antes de la introducción de un programa de tamizaje para diagnosticar el cáncer de pulmón mediante rayos X, la letalidad a los 6 meses era de 80 por cada 100 casos diagnosticados. Después de la introducción del programa de tamizaje, la letalidad a los 6 meses era de 20 por 100 casos diagnosticados. Los autores llegaron a la conclusión de que el programa de tamizaje había reducido drásticamente la letalidad por cáncer de pulmón y, de esta forma, se había demostrado su importancia.

El programa de tamizaje probablemente diagnosticó casos de cáncer de pulmón en estadios asintomáticos, mientras que los diagnósticos anteriores se habían realizado cuando los síntomas ya estaban presentes. Las pruebas disponibles sugieren que el tamizaje mediante rayos X no modifica el pronóstico a largo plazo del cáncer de pulmón. El tamizaje simplemente modifica el intervalo entre el diagnóstico y la muerte, lo cual produce un sesgo debido al adelanto del diagnóstico. Los cambios observados son probablemente artefactos atribuibles al diagnóstico temprano de la enfermedad.

El siguiente ejemplo ilustra cómo el significado de la terminología puede cambiar con el tiempo y producir un cambio por artefactos en la tasa de los sucesos.

La incidencia del síndrome de la inmunodeficiencia adquirida (SIDA) aumentó anualmente entre 1981 y 1986. En 1987 se produjo un brusco aumento de 50% en la tasa notificada. Un investigador interpretó este incremento como signo de que la epidemia había entrado súbitamente en una nueva fase.

En 1987 se modificó la definición del SIDA establecida por los Centros para el Control de Enfermedades (CDC). Esto significó la inclusión de un mayor número de individuos infectados por el virus de la inmunodeficiencia humana (VIH) entre los casos definidos como SIDA. Cuando se producen alteraciones súbitas en las tasas de incidencia de una enfermedad, se debe sospechar de la influencia de un artefacto tal como el de una nueva definición de la enfermedad. En el caso mencionado, se puede sospechar que un cambio por artefacto se superpuso a un cambio real y complicó la interpretación de los datos.

CAMBIOS REALES

Los cambios de las tasas debidos a artefactos implican que la verdadera tasa de incidencia, la prevalencia o la letalidad no se han modificado aunque aparentemente se haya producido un cambio. Sin embargo, los cambios reales suponen que las tasas de incidencia, la prevalencia o la letalidad han cambiado verdaderamente. En primer lugar, debemos preguntarnos si interviene alguna de las fuentes de diferencias por artefactos. Si no intervienen o no son lo suficientemente importantes para explicar las diferencias, se puede suponer que existen diferencias o cambios reales. Una vez demostrado que se han producido diferencias o cambios reales, tenemos que preguntarnos por qué se han producido. ¿Reflejan un cambio en las tasas de incidencia, en la prevalencia, o en la letalidad o una combinación de estas medidas?

La primera etapa para entender el significado de un cambio real en las tasas consiste en entender en qué momento de la historia natural de la enfermedad se ha producido el cambio. Entonces podremos valorar mejor los efectos de los cambios primarios en las otras tasas de la enfermedad, por ejemplo, en los casos que se comentan a continuación.

1. La letalidad de la enfermedad de Hodgkin ha descendido notablemente en los últimos años. Se considera que los individuos padecen la enfermedad hasta que su curación se verifica mediante un seguimiento a largo plazo. Por este motivo, la prevalencia de la enfermedad ha aumentado. Las tasas de incidencia han permanecido estables; por lo tanto, la tasa de mortalidad, que refleja la tasa de incidencia multiplicada por la letalidad, ha descendido.

2. Las tasas de incidencia del cáncer de pulmón han aumentado de forma sustancial en las últimas décadas. Sin embargo, la letalidad ha seguido siendo muy baja y muchos de los pacientes mueren en los meses que siguen al diagnóstico. Por este motivo, la tasa de mortalidad también ha aumentado drásticamente. A causa de la corta duración de la enfermedad, la prevalencia ha sido siempre baja; no obstante, ha aumentado ligeramente debido al aumento de la tasa de incidencia de la enfermedad.

Estos resultados se pueden representar del siguiente modo:

	TASAS DE MORTALIDAD	LETALIDAD	TASAS DE INCIDENCIA	PREVALENCIA
Enfermedad de Hodgkin	↓	↓ ↓ ↓	→	↑ ↑
Cáncer de pulmón	↑ ↑	→	↑ ↑	↑

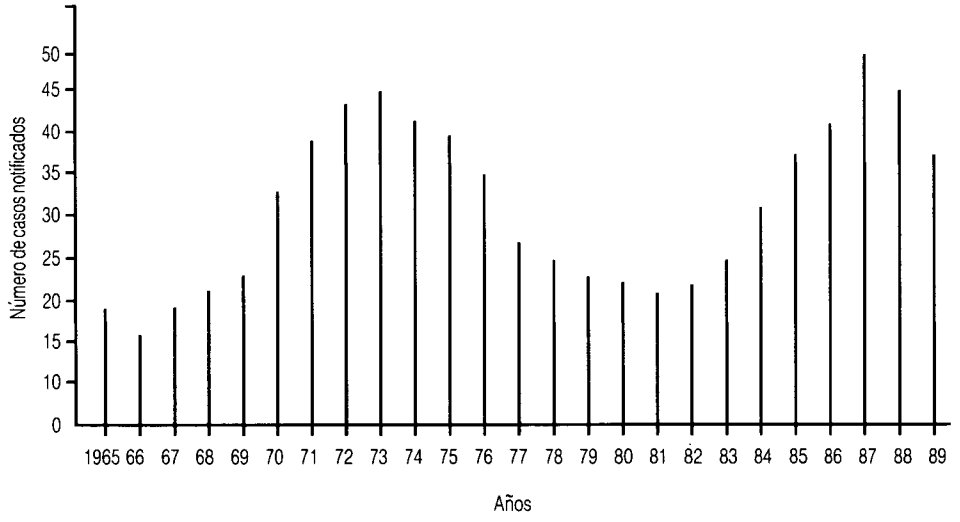
Estos patrones confusos cobran sentido cuando uno se percata de que el cambio principal en la enfermedad de Hodgkin ha sido el descenso de la letalidad, mientras que en el cáncer de pulmón ha sido el aumento de la tasa de incidencia.

Además de entender la fuente de las diferencias o de los cambios en las tasas, se debe resistir la tentación de utilizar los cambios pasados de las tasas para predecir las tasas futuras.

La predicción de las tasas a partir de las actuales o de sus cambios recientes es un trabajo muy difícil. Un cambio reciente de las tasas puede tener alguno de los siguientes significados: 1) prefigurar cambios futuros en la misma dirección; 2) reflejar ciclos o epidemias pronosticables, o 3) ser el resultado de fluctuaciones aleatorias impredecibles que representan una frecuencia inusual de los sucesos. Antes de concluir que es probable que se mantengan los cambios observados en las tasas, es preciso considerar la posibilidad de que se trate de fluctuaciones cíclicas o aleatorias. Si existe un ciclo natural en la frecuencia de la enfermedad, las tasas de año en año pueden ser semejantes a las que aparecen en la figura 23-1.

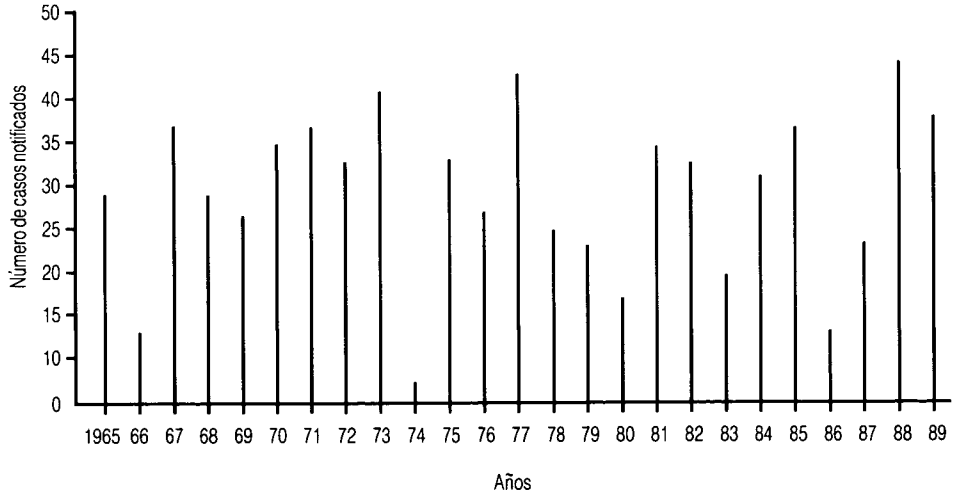
Puede que los investigadores observen el aumento de las tasas de 1981 a 1984, y que al tratar de medir los cambios que ocurrieron entre 1985 y 1987 encuentren un nuevo aumento. Sin embargo, es importante que se den cuenta de que el cambio real que se produjo entre 1981 y 1987 quizá sea parte del ciclo natural de la enfermedad y no implique necesariamente otros incrementos previsibles para 1988, 1990 ó 1995.

FIGURA 23-1. Ciclos o epidemias predecibles en la incidencia anual de una enfermedad



Por el contrario, en lugar de un ciclo predecible de la enfermedad, es posible que se produzca una variación anual, impredecible y aleatoria de la tasa de la enfermedad (figura 23-2). En esta situación, si los investigadores eligen un año en el que la tasa era más alta y lo comparan con el siguiente, en el cual la tasa era más baja solo por azar, pueden creer que están documentando un cambio importante cuando, de hecho, están descubriendo el principio estadístico conocido como *regresión a la media*. La regresión estadística a la media o el retorno al promedio afirma que los valores inu-

FIGURA 23-2. Variaciones impredecibles o debidas al azar de la incidencia anual de una enfermedad



suales son raros por definición y que la probabilidad está en contra de que un suceso raro se repita dos veces consecutivas. De hecho es probable, solo por azar, que la medición siguiente se encuentre más cerca del valor promedio.

Los valores subsiguientes pueden ser más extremos que un valor observado, debido a la fluctuación aleatoria de los sucesos o a las fuerzas que reaccionan frente a la tasa inusual y la desplazan hacia la línea o hacia el promedio o la media. Por ejemplo, si se está estudiando cuánto come un individuo en cada comida, es probable que la cantidad de alimentos ingeridos en la comida que sigue a una comilona sea menor que la habitual. Veamos cómo puede intervenir este principio en un estudio de tasas en el que se observaron diferencias reales que deben ser interpretadas cuidadosamente.

Después de un trágico accidente en una fábrica, en el que fallecieron varios hombres, se inició un programa de prevención de accidentes. Los investigadores encontraron que la tasa de incidencia de accidentes en el momento de la tragedia era inusualmente alta, de 10 por 1 000 días laborables. Cuando se estableció el programa, la tasa descendió hasta 2 por 1 000 días laborables. Los investigadores concluyeron que el programa de prevención de accidentes tuvo un éxito espectacular.

Los investigadores demostraron que se produjo un cambio real. Sin embargo, no demostraron que el programa de prevención de los accidentes fuese la causa del cambio. Es posible que la tasa de 10 por 1 000 fuera inusitadamente alta y que solo por azar retornara a la tasa habitual de 2 por 1 000. Aun más probable es que el trágico accidente haya inducido a los trabajadores a aumentar las medidas de seguridad, produciendo lo que podría denominarse una *regresión psicológica* hacia la media. Los autores partían de una tasa de accidentes desusadamente alta y luego ocurrió una tragedia que podría haber forzado la tasa a regresar hacia la media. Los autores documentaron un cambio real que podía ser explicado mejor por una regresión a la media que por un cambio a largo plazo. Es prematuro concluir que el programa de prevención de accidentes sería de gran ayuda para otros grupos, o incluso para este mismo grupo, si se llevara a cabo en otro momento. Por eso, aunque el principio de regresión a la media puede ser una fuente de cambios reales en las tasas, esta fuente puede tener una importancia limitada y no garantizar la continuidad de los cambios observados.

Es habitual iniciar una investigación cuando se sospecha que las tasas de una enfermedad están aumentando. Por eso es fundamental reconocer el fenómeno de la regresión a la media, ya que puede estar interviniendo siempre que se observan cambios a corto plazo en las tasas.

Otra fuente de diferencias reales que influyen en la predicción de los sucesos futuros se conoce como *efecto de cohorte*. Una cohorte es un grupo de individuos que comparten una experiencia o una exposición en común. Existe la posibilidad de un efecto de cohorte cuando una o diversas cohortes de una población han estado sometidas a una exposición o a una experiencia que las ha hecho especialmente susceptibles a una enfermedad. Las tasas en un grupo de edad determinado que incluye la cohorte susceptible pueden aumentar temporalmente. Esta elevación temporal se conoce como efecto de cohorte. Cuando se halla presente, se puede esperar que las tasas de este grupo de edad concreto desciendan otra vez a medida que transcurra el tiempo y la cohorte susceptible pase al siguiente grupo de edad. La importancia que tiene el saber valorar el efecto de cohorte se ilustra en el siguiente ejemplo.

Un investigador estudió la tasa de incidencia del cáncer de tiroides. Existía preocupación por el hecho de que la irradiación de la cabeza y del cuello, empleada frecuentemente antes de los años cincuenta, hubiese contribuido a aumentar la frecuencia del cáncer de tiroides. Utilizando métodos adecuados, los autores observa-

ron que la incidencia del cáncer de tiroides en 1950 en el grupo de edad de 20 a 30 años fue de 50 por 100 000 personas-año; que en 1960 fue de 100 por 100 000 personas-año, y que en 1970 fue de 150 por 100 000 personas-año. Los autores concluyeron que en 1980 la tasa superaría los 200 casos por 100 000 personas-año y quedaron sorprendidos cuando comprobaron que la incidencia en 1980 fue menor de 150 por 100 000 personas-año y continuó descendiendo en los años ochenta.

Los autores habían demostrado que las tasas de incidencia del cáncer de tiroides habían experimentado cambios reales en el grupo de edad de 20 a 30 años. La fuente de estas diferencias pudo corresponder a un efecto de cohorte. Los individuos de la cohorte que recibieron radiaciones antes de 1950 presentaban un aumento de riesgo del cáncer de tiroides que no afectaba necesariamente a los individuos nacidos después de ese año. En 1980, todos los individuos entre 20 y 30 años de edad habían nacido después de 1950. Por eso, no es sorprendente observar un descenso de la tasa de incidencia de cáncer de tiroides en ese grupo de edad en lugar de un aumento sostenido. El concepto de efecto de cohorte no solo ayuda a predecir las tasas previsibles, sino que también ayuda a reforzar la teoría de que la exposición pasada a la radioterapia aumenta el riesgo de padecer cáncer de tiroides.

La regresión a la media y el efecto de cohorte son dos motivos por los que es peligroso predecir las tasas de una enfermedad a partir de la extrapolación directa de sus cambios recientes. Además de comparar las tasas de muestras del mismo grupo en distintos momentos, también es posible utilizarlas para examinar las diferencias entre muestras de grupos diferentes. La comparación de las tasas de los grupos se realiza frecuentemente para generar hipótesis que luego pueden ser contrastadas mediante los tipos de investigaciones que comentamos antes en *El estudio de un estudio*. Por ejemplo, los investigadores podrían observar diferencias entre las tasas de mortalidad por coronariopatías de los esquimales y de los blancos de Alaska. A partir de lo que sabemos acerca de la dieta de los esquimales de Alaska, se puede formular la hipótesis de que existe una asociación entre el consumo de pescado y la baja mortalidad por enfermedad coronaria.

Si se comparan las tasas de la enfermedad y se ajustan según los factores de riesgo conocidos, es posible demostrar que cuando un factor aumenta en un grupo, la enfermedad también aumenta en el mismo grupo. Esto nos permite establecer *asociaciones de grupo*. Una asociación de grupo implica un aumento de las prevalencias del factor y de la enfermedad en un grupo determinado. Observe que una asociación de grupo no implica necesariamente que aquellos individuos que tienen el factor de riesgo sean los mismos que tienen la enfermedad.

La demostración de que existe una asociación de grupo puede constituir la base de estudios ulteriores que establezcan una asociación a nivel individual y, finalmente, una relación de causa-efecto como en el caso del colesterol y la enfermedad coronaria.

Cuando se utilizan datos de grupos, los investigadores suelen tener poca información sobre los individuos que integran cada grupo. Por esta razón, cuando se comparan tasas con el fin de formular una hipótesis para su estudio posterior, los investigadores deben tener cuidado de no establecer una asociación entre individuos, cuando lo único que se ha establecido es una asociación de grupo. Este tipo de error, conocido como *falacia ecológica*, se ilustra en el siguiente ejemplo.

En un estudio se demostró que la tasa de ahogamientos en Florida es cuatro veces la de Illinois. Los datos del estudio también mostraron que el consumo de helados en Florida es cuatro veces el de Illinois. Los autores llegaron a la conclusión de que comer helados está asociado con ahogarse.

Para establecer una asociación, los autores deben demostrar primero que los que comen más helados son los que están en mayor riesgo de ahogarse. Los datos de grupos no proporcionan ninguna prueba de la existencia de una asociación a nivel individual. Es posible que los que comen helados no sean los que se ahogan. El mayor consumo de helados puede reflejar simplemente la variable de confusión conocida como tiempo cálido, la cual aumenta el consumo de helados y el de ahogados. Estos autores cometieron una falacia ecológica. Para establecer una asociación entre comer helados y ahogarse es necesario demostrar que la asociación se mantiene a nivel individual.

Cuando nos enfrentamos con diferencias entre tasas de grupos o cambios con el tiempo en el mismo grupo, el primer paso consiste en demostrar si la diferencia o los cambios son producidos por artefactos o si son reales. Si no es probable que las diferencias o los cambios sean debidos a la forma en que se mide, se busca o define la enfermedad, entonces los cambios o las diferencias se pueden considerar reales. Seguidamente, se busca la fuente de estas diferencias o de los cambios en la tasa de incidencia, en la prevalencia o en la letalidad. Los cambios o diferencias en las tasas se usan con frecuencia para predecir cambios futuros y para formular hipótesis acerca de la etiología de la enfermedad que se han de utilizar como punto de partida de estudios sobre individuos. Para predecir las tasas futuras, es necesario tener en cuenta el fenómeno de regresión a la media y el efecto de cohorte. Cuando se utilizan tasas de grupos para desarrollar hipótesis, es preciso reconocer que su uso demuestra la existencia de asociaciones entre grupos y no entre individuos. Si no se aprecia esta distinción entre asociaciones individuales y de grupo se puede cometer una falacia ecológica.

CAPÍTULO 24

RESUMEN: LA TASACIÓN DE UNA TASA

Revisemos los pasos necesarios para desarrollar, comparar e interpretar las tasas de enfermedad.

LA SELECCIÓN DE LAS TASAS

Las tasas de incidencia son medidas clínicas importantes, porque miden el ritmo de desarrollo de nuevos casos de una enfermedad. El riesgo es el efecto acumulativo de la tasa de incidencia en un período determinado. La prevalencia ayuda en el diagnóstico, ya que es una aproximación a la probabilidad de tener la enfermedad en un momento determinado. La letalidad mide la probabilidad de morir como consecuencia de la enfermedad una vez que ha sido diagnosticada. La prevalencia es aproximadamente igual a la tasa de incidencia multiplicada por la duración media de la enfermedad. En una población estable, la tasa de mortalidad es igual a la tasa de incidencia multiplicada por la letalidad.

EL MUESTREO DE LAS TASAS

Las tasas pueden calcularse utilizando toda la población. Con más frecuencia, las tasas se estiman a partir de muestras de la población. Las tasas de las muestras, en promedio, son las mismas que las de la población, pero cualquier muestra puede reflejar un error muestral intrínseco. Cuanto mayor sea el tamaño muestral, menor será el error muestral. Sin embargo, la ganancia en la precisión disminuye a medida que aumenta el tamaño de la muestra y el investigador debe contrapesar la exactitud respecto del costo. La capacidad para estimar la tasa de la enfermedad en una población a partir de una muestra exige llevar a cabo un muestreo aleatorio de la población. Es posible estratificar la muestra para garantizar que el tamaño de cada categoría sea suficientemente grande, pero el muestreo debe ser aleatorio dentro de cada categoría.

LA ESTANDARIZACIÓN DE LAS TASAS

Nos interesa comparar las tasas, ya sean obtenidas a partir de toda la población o de una muestra, para estimar las diferencias entre grupos o los cambios con el tiempo. La comparación correcta de las tasas muchas veces requiere ajustarlas según los factores que influyen en ellas, que son distintos en las dos muestras. Frecuentemente se ajusta o se estandariza según factores demográficos como la edad, el sexo y la raza, para buscar otros factores que puedan explicar las diferencias o los cambios en las tasas. Mediante la estandarización indirecta es posible calcular la razón de mortalidad estandarizada, que compara la tasa en un grupo de estudio concreto con la esperada en otro grupo, con frecuencia la población general. El método directo de estandarización permite comparar los grupos directamente después de estandarizar o ajustar las tasas según una característica respecto a la que difieren ambos grupos.

FUENTES DE LAS DIFERENCIAS O DE LOS CAMBIOS

El significado de las diferencias se valora analizando si se deben a artefactos o si son reales. Los cambios en la capacidad para diagnosticar la enfermedad, los esfuerzos para diagnosticar la enfermedad o la definición de la enfermedad pueden producir cambios o diferencias por artefactos que no son debidos a la propia enfermedad.

Cuando las diferencias o los cambios por artefactos no son lo suficientemente grandes como para explicar las diferencias o los cambios observados, entonces se puede suponer que existe una diferencia o un cambio real. Los cambios reales pueden ser el resultado de cambios en la tasa de incidencia, en la prevalencia o en la letalidad. Estos cambios pueden anunciar cambios posteriores en la misma dirección o ser seguidos de un descenso en las tasas, como ocurre cuando finaliza una epidemia. Por otro lado, pueden reflejar fluctuaciones aleatorias de las tasas de la enfermedad. Cuando se usan los cambios de las tasas para predecir tasas futuras, es importante tener en cuenta el fenómeno conocido como la *regresión a la media*. Este efecto se produce con frecuencia cuando se usa una tasa inusual como comparación. Esta tasa inusual puede movilizar fuerzas que desplazan a las futuras tasas hacia la media o incluso por debajo de esta. Los bioestadísticos utilizan el término *regresión a la media* para indicar que después de producirse un valor inusual las mediciones subsiguientes se aproximarán probablemente hacia la media o hacia el valor promedio, solo por azar.

Otro tipo de cambio real en las tasas se conoce como *efecto de cohorte*. El efecto de cohorte, al circunscribir los cambios a un segmento de la población, ayuda a predecir mejor las tasas futuras que la simple extrapolación de los cambios observados. El estudio de las fuentes de los cambios reales de las tasas nos ayuda a comprender mejor sus causas, así como sus consecuencias para los años siguientes.

Cuando se usan las diferencias entre tasas en la formulación de hipótesis para realizar estudios sobre individuos, es preciso reconocer que las asociaciones de grupo se establecen usando tasas. La falacia ecológica se produce cuando las asociaciones observadas en grupos no se reflejan a nivel individual.

PREGUNTAS SOBRE LA TASACIÓN DE UNA TASA

El siguiente grupo de preguntas acerca de la tasación de una tasa se ha diseñado con objeto de repasar los puntos ya tratados y de proporcionar un esquema utilizable en la crítica de los ejercicios para detectar errores o en una investigación real de tasas. El esquema se divide en cuatro partes: selección de las tasas, muestreo de las tasas, estandarización de las tasas y fuentes de las diferencias o cambios en las tasas.

1. Selección de las tasas
   a. ¿Se han identificado las diferencias entre las tasas, las proporciones y las razones?
   b. ¿Han distinguido los autores la tasa de incidencia de la prevalencia y la letalidad?

2. Muestreo de las tasas
   a. ¿Se incluyó en el estudio a toda la población de interés o se utilizaron muestras?
   b. Si se utilizaron muestras, ¿valoraron los autores el error muestral intrínseco?

c. Si se utilizaron muestras, ¿fue aleatoria y representativa la técnica de muestreo o se introdujo algún sesgo en el método de muestreo?
d. Si se utilizaron muestras, ¿fue suficiente el tamaño de las muestras o pudo haberse introducido una gran variación a causa de su pequeño tamaño?

3. Estandarización de las tasas
   a. ¿Fue necesaria la estandarización cuando se compararon las tasas de frecuencia de un suceso en dos muestras diferentes?
   b. Si fue necesaria la estandarización, ¿se estandarizaron las tasas según los factores conocidos que influyen en el desenlace de forma que se pudieran comparar equitativamente?

4. Fuentes de las diferencias o cambios en las tasas
   a. Las diferencias o cambios observados en las tasas, ¿fueron por artefactos debidos a la capacidad de diagnosticar la enfermedad, a los esfuerzos para hacerlo o a cambios de la definición de la enfermedad?
   b. ¿Eran reales las diferencias o cambios debidos a cambios o diferencias en la tasa de incidencia, la prevalencia o la letalidad?
   c. Los cambios o diferencias en las tasas, ¿predicen cambios futuros en la misma dirección o serán las tasas futuras influidas por el fenómeno de la regresión a la media o el efecto de cohorte?
   d. Cuando se emplearon las diferencias entre tasas para formular hipótesis de estudios posteriores, ¿se mantuvo la distinción entre asociación a nivel de grupo y a nivel individual?

EJERCICIOS PARA DETECTAR ERRORES: LA TASACIÓN DE UNA TASA

Los siguientes ejercicios para detectar errores están pensados para que usted adquiera práctica en la aplicación de los principios de tasación de tasas a artículos de investigación simulados. Estos ejercicios incluyen diversos errores que ya han sido ilustrados con ejemplos hipotéticos. Lea cada ejercicio. Luego escriba una crítica señalando los tipos de errores cometidos por los investigadores. Para cada ejercicio se proporciona una crítica de demostración.

EJERCICIO No. 1: CAMBIOS EN EL CÁNCER: ¿CUÁL ES EL PROGRESO?

En un estudio sobre la evaluación del cáncer en los Estados Unidos de América se compararon las tasas de 1969 con las de 1989 para valorar los cambios. Se recogieron datos sobre las tasas de incidencia y mortalidad. Los datos de incidencia se obtuvieron mediante una búsqueda intensa en los registros hospitalarios a partir de una muestra al azar de 1% de los hospitales de la nación. Los datos de mortalidad se obtuvieron revisando todos los certificados de defunción de la nación. La letalidad se calculó por medio de la fórmula para cambios a largo plazo, según la cual

$$\text{Tasas de mortalidad} = \text{tasas de incidencia} \times \text{letalidad}$$

Los datos de este estudio se resumen en el cuadro 25-1.

Además, los investigadores revisaron los ensayos clínicos controlados sobre los tipos de cáncer que causaron 50% de las muertes entre las personas de 20 o más años de edad y observaron que los datos mostraban un aumento de la mediana de supervivencia a los 3 años cuando se aplicaban los nuevos tratamientos desarrollados desde 1969.

Por último, los investigadores calcularon la *razón de mortalidad proporcional* basándose en una revisión de la causa de muerte de todos los certificados de defunción y observaron que la razón de mortalidad proporcional del cáncer había aumentado globalmente de 22 a 24%.

CUADRO 25-1. Cambios en las tasas de cáncer de 1969 a 1989

Edad	Tasa de incidencia	Letalidad	Tasa de mortalidad
0–19	Sin cambios	Disminuye 20%	Disminuye 20%
20–65	Aumenta 1%	Disminuye 1%	Sin cambios
>65	Aumenta 15%	Disminuye 10%	Aumenta 5%

Los investigadores confesaron que estaban totalmente confusos, porque se podía afirmar todo lo siguiente:

1. Basándose en el descenso de las tasas de mortalidad de los individuos menores de 20 años, el descenso de la letalidad en todos los grupos de edad y el aumento de la supervivencia en los ensayos clínicos controlados realizados con personas de 20 o más años de edad, se ha producido un progreso notable.
2. Sobre la base del aumento de las tasas de incidencia entre los que tienen más de 20 años y del aumento de las tasas de incidencia ajustadas según la edad, la situación está empeorando. El aumento de las tasas de mortalidad entre los mayores de 65 años y de las razones de mortalidad proporcional también apoya la interpretación de que la situación está empeorando.
3. Según las tasas de mortalidad global ajustadas por la edad, no han ocurrido cambios.

Los investigadores se dan por vencidos y les preguntan a ustedes, lectores, cómo pueden los datos respaldar resultados tan contradictorios.

CRÍTICA: EJERCICIO No. 1

Todas estas tasas son compatibles y reflejan las diferentes formas de analizarlas. Las tasas de incidencia reflejan el ritmo con que aparecen los nuevos casos de la enfermedad en un período. La letalidad refleja la probabilidad de morir en un período si se desarrolla la enfermedad. Por eso, las tasas de incidencia y la letalidad miden dos fenómenos muy distintos. Las tasas de incidencia reflejan primariamente las causas subyacentes de la enfermedad y pueden cambiar a causa de artefactos como, por ejemplo, las intervenciones médicas que modifican el esfuerzo de detección de la enfermedad, la capacidad de detectarla o de definirla. Los esfuerzos de prevención primaria como las campañas antitabáquicas pueden modificar la incidencia subyacente. No obstante, las tasas de incidencia no reflejan, en general, los esfuerzos terapéuticos constantes de la atención médica. La letalidad, por su parte, es una medida del éxito del tratamiento médico en la curación de la enfermedad.

Las tasas de mortalidad en un período prolongado se relacionan con la incidencia y con la letalidad de la siguiente manera:

$$\text{Tasa de mortalidad} = \text{tasa de incidencia} \times \text{letalidad}$$

Por esta razón, si la tasa de mortalidad y la de incidencia se conocen y son estables, la letalidad se puede estimar con la siguiente fórmula:

$$\text{Letalidad} = \text{tasa de mortalidad/tasa de incidencia}$$

Por lo tanto, en el primer cuadro se empleó correctamente la relación entre la tasa de incidencia y la de mortalidad y la letalidad.

Cuando una intervención médica logra prolongar la vida pero no curar la enfermedad, no tiene un efecto sobre la letalidad a largo plazo ni sóbre la global. Por este motivo, el aumento de 3 años de la mediana de supervivencia entre los que padecen tipos importantes de cáncer es compatible con el descenso más leve de la letalidad observado en la tasa global ajustada según la edad. Se necesitan medidas como la esperanza de vida para tener en cuenta la prolongación de la vida en ausencia de curación.

El aumento de la razón de mortalidad proporcional nos dice muy poco acerca del progreso del cáncer durante esos años. Empero, sugiere que la mortalidad debida a otras enfermedades ha disminuido su frecuencia en comparación con la del cáncer. Las razones de mortalidad proporcional son medidas útiles de la importancia relativa de diversas causas de muerte. El aumento de la razón de mortalidad proporcional sugiere que las muertes por cáncer son cada vez más habituales en relación con las muertes por otras causas.

Este ejercicio demuestra cómo es posible defender conclusiones completamente distintas a partir de los mismos datos. El argumento presentado por el investigador refleja los diferentes conceptos sobre el significado de progreso. ¿Es progreso una tasa de incidencia reducida de una nueva enfermedad? ¿Es progreso una tasa de curación elevada de una enfermedad diagnosticada? O bien, ¿es progreso la prolongación de la vida de los que están enfermos?

EJERCICIO No. 2: TUBERCULOSIS

Un grupo de expertos internacionales en tuberculosis (TB) decidió comparar la tasa de mortalidad de TB en los Estados Unidos con la de la India con objeto de determinar si se podrían aprender algunas cosas. Sabían que la TB era una enfermedad bastante común en los Estados Unidos antes de 1950, pero que había disminuido de forma sustancial. También sabían que la infección se podía controlar, pero no eliminar, en la mayoría de las personas infectadas.

Utilizando un diseño cuidadoso de una técnica de muestreo aleatorio, observaron que la tasa de mortalidad por TB en 1985 era de 200 por 100 000 personas-año en la India y de 20 por 100 000 personas-año en los Estados Unidos. También apreciaron que la tasa de mortalidad de los individuos en el grupo de edad entre 65 y 80 años era de 200 por 100 000 personas-año en la India y en los Estados Unidos.

Dado que los investigadores sabían que la India tenía una estructura de edad mucho más joven que los Estados Unidos y que la edad influye en el riesgo de TB, intentaron estandarizar las tasas. Para ello, aplicaron las tasas de cada grupo de edad de la población estadounidense al número de personas-año de ese mismo grupo en la India. Después de realizar un ajuste directo según la edad observaron que en 1985 se habían producido 200 muertes por 100 000 personas-año en la India por TB. En los Estados Unidos se habrían producido 10 muertes por 100 000 personas-año si este país hubiera tenido la misma estructura de edad que la India. Por último, calcularon la razón de mortalidad proporcional de la TB en los dos países. En la India encontraron que la razón de mortalidad proporcional por TB fue de 1,5% y en los Estados Unidos, de 1%.

Los autores llegaron a las siguientes conclusiones:

1. La gran diferencia entre las tasas de mortalidad de la India y los Estados Unidos puede ser debida al azar.

2. En los Estados Unidos, la tasa de mortalidad del grupo de edad de 65 a 80 años es mucho más elevada que la tasa promedio del país. Por lo tanto, a medida que aumente la edad media de la población de este país, la TB se convertirá en un problema cada vez más importante.

3. Cuando se compararon las tasas no estandarizadas, parecía que la India tenía una tasa de TB 10 veces más alta que la de los Estados Unidos. Una vez estandarizada, se observó que la India tenía una tasa 20 veces más alta. Se debió haber cometido algún error, porque la estandarización de las tasas no debe aumentar las diferencias entre las

tasas que existían antes de llevar a cabo la estandarización. Por eso, el contraste de las tasas globales fue el método de comparación más equitativo.

4. Las razones de mortalidad proporcional son el método más justo para comparar la probabilidad de morir de TB. En consecuencia, la probabilidad de morir por TB era 1,5 veces más alta en la India que en los Estados Unidos.

CRÍTICA: EJERCICIO No. 2

Selección de las tasas

Vamos a evaluar una por una las conclusiones a las que llegaron los "expertos".

1. Es muy improbable que las diferencias entre los Estados Unidos y la India sean exclusivamente debidas al azar, a causa de su amplia separación y al gran tamaño de las muestras estudiadas. Sin embargo, antes de concluir que estas diferencias son reales, hemos de considerar si han existido diferencias por artefactos, tanto debidas al esfuerzo como a la capacidad de diagnosticar la TB entre los dos países y si estos factores pueden explicar las diferencias observadas. Si suponemos que en la India, con sus enormes problemas de salud, es más difícil diagnosticar la TB como causa de defunción que en los Estados Unidos, este hecho aumentaría realmente las ya grandes diferencias observadas. Por consiguiente, no es probable que las diferencias por artefactos en cuanto a esfuerzo o la capacidad de diagnosticar las muertes por TB expliquen o eliminen las diferencias observadas.

2. Los autores observaron que las tasas en el grupo de edad de 65 a 80 años de ambos países eran idénticas. Esto sugiere que la tasa de mortalidad de los ancianos de los Estados Unidos es más elevada que la del resto de la población estadounidense, lo cual puede deberse a alguna susceptibilidad permanente de las personas de ese grupo de edad. Por otro lado, puede reflejar un efecto de cohorte. Recuerde que un efecto de cohorte es una susceptibilidad temporal y única de un grupo o cohorte a una experiencia pasada especial. Sabemos que la TB fue una enfermedad mucho más común en los Estados Unidos antes de 1950. Además, sabemos que los individuos que han estado infectados anteriormente con frecuencia controlan la infección, si bien no se curan totalmente. Estos individuos tienen la posibilidad de desarrollar posteriormente una enfermedad activa, sobre todo cuando envejecen. Por este motivo, es posible que la elevada tasa entre los estadounidenses de edad avanzada se relacione con su experiencia en una era en que la TB activa era bastante frecuente. Si el efecto de cohorte explica la tasa elevada en los ancianos de los Estados Unidos, entonces no es probable que la elevada tasa de los que tienen entre 65 y 80 años continúe aumentando a medida que crezca la proporción de ancianos de dicha población. Las tasas de este grupo de edad realmente pueden descender a medida que las cohortes menos susceptibles, aquellas compuestas por los que no estuvieron expuestos anteriormente a la TB, lleguen a esa edad.

En la India, por otra parte, la tasa de mortalidad del grupo de 65 a 80 años de edad es la misma que la de la población general del país. Por ello, no existe ninguna razón para creer que esté interviniendo un efecto de cohorte. Las tasas de todos los grupos de edad pueden ser uniformemente altas y los ancianos no comparten necesariamente una susceptibilidad única. Esto implica que si la proporción de ancianos en la India aumenta, la tasa de mortalidad por TB en ese grupo de edad no experimentará cambios importantes.

3. Los investigadores concluyeron correctamente que la estandarización de las tasas según la edad era un procedimiento adecuado para compararlas equitativamente. La estandarización es importante cuando las poblaciones tienen una marcada diferencia en su distribución de edad y esta variable se relaciona con el riesgo de morir por TB. La población de la India es generalmente más joven que la estadounidense. Las muertes por TB se concentran en los ancianos y los muy jóvenes, razón por la cual es preciso estandarizar las tasas si se quiere realizar una comparación equitativa. El procedimiento de la estandarización según la edad se realizó correctamente. Al aplicar las tasas de cada grupo de edad de la población estadounidense al número de individuos de ese grupo de edad de la población de la India, los investigadores se preguntaban cuántas muertes se habrían producido en la población de los Estados Unidos si tuviese el mismo número de individuos en cada grupo de edad que la India. Esta cifra se puede comparar directamente con el número de muertes que se produjeron realmente en la India. La estandarización puede hacer que las diferencias parezcan más pequeñas o más grandes. No hay ninguna razón para pensar que la estandarización será el factor responsable de que las grandes diferencias parezcan menores. Como muestra el ejemplo de la fábrica presentado en la discusión de la estandarización (capítulo 22), el ajuste según la edad puede incluso revelar una diferencia importante que no ha sido aparente en las tasas globales.

4. La razón de mortalidad proporcional refleja el número de individuos que mueren de una enfermedad dividido por el número de los que mueren por todas las enfermedades. Como las tasas de mortalidad global son más altas en la India que en los Estados Unidos, el denominador de la razón de mortalidad proporcional es considerablemente más alto en la India. Por eso, cuando se calcula la razón de mortalidad proporcional para la India, dividimos un numerador mayor por un denominador mayor. Por lo tanto, no es sorprendente que obtuviéramos una razón de mortalidad proporcional para la India que solo fue ligeramente más elevada que la de los Estados Unidos. La razón de mortalidad proporcional no se debe considerar como una medida de la probabilidad de morir, sino como una medida de la importancia relativa de una enfermedad en una población determinada.

Sección 4

La selección de una prueba estadística

CAPÍTULO 26

PRINCIPIOS BÁSICOS

La estadística aplicada a la investigación médica persigue tres finalidades: 1) sintetizar numerosas mediciones en un número limitado de datos manejables, 2) realizar estimaciones e inferencias a partir de las muestras extraídas de poblaciones, teniendo en cuenta la influencia del azar, y 3) ajustar los datos según la influencia de las variables de confusión en esas estimaciones e inferencias. Nuestro objetivo en *La selección de una prueba estadística* es arrojar algo de luz sobre la forma en que la estadística puede ayudar a conseguir estos fines. No suponemos que la información brindada en estas pocas páginas pueda reemplazar la participación de un estadístico en las fases de planificación, ejecución e interpretación de la mayor parte de los proyectos de investigación médica; pero sí esperamos proporcionar las herramientas necesarias para que los lectores de la literatura de investigación sepan valorar la sección de los "métodos estadísticos" de tal forma que el análisis, la interpretación y la extrapolación de los resultados de la investigación se puedan comprender totalmente.

Para utilizar la estadística en la investigación médica, en primer lugar es preciso escoger un método estadístico apropiado. En segundo lugar, las mediciones de la investigación deben ser manipuladas de acuerdo con el método seleccionado. Por último, los resultados de estas manipulaciones han de interpretarse correctamente. La primera y la última de estas tareas están íntimamente relacionadas con el tema de la Sección 4, *La selección de una prueba estadística*. Sin embargo, no trataremos de discutir a fondo las manipulaciones de los datos que son necesarias para producir los resultados estadísticos. Sin lugar a dudas, el estudio de estas manipulaciones requiere una comprensión más profunda de los métodos estadísticos, pero, en nuestra opinión, no es preciso tener ese nivel de conocimientos para poder evaluar por qué se selecciona un método determinado y cómo podemos interpretar los resultados de su aplicación.

Empezaremos echando un vistazo a la forma de enfocar las primeras dos finalidades de la estadística. La tercera, ajustar los datos según el efecto de las variables de confusión, se realizará mediante el análisis multivariable, que se presentará en el capítulo 29.

SÍNTESIS DE LAS MEDIDAS

Como se ha afirmado anteriormente, una de las finalidades de los métodos estadísticos consiste en resumir grandes cantidades de datos en un número reducido y manejable de ellos. Para cumplir con esta tarea debemos darnos cuenta, en primer lugar, de que las mediciones realizadas en los sujetos de una investigación son una parte o una *muestra* de un grupo más numeroso de individuos que podrían haber sido incluidos en la misma. Este grupo más numeroso se denomina *población*.[1]

[1] En medicina, habitualmente pensamos en mediciones realizadas en personas, en lugar de animales u objetos. Esto puede crear la falsa impresión de que el término estadístico *población* es el mismo que se utiliza para describir distintos conjuntos de personas en política o en geografía. Aunque una población estadística podría ser uno de esos grupos de personas, no se limita a ellos. Una *población estadística* se define como el conjunto de todas las mediciones posibles (no necesariamente realizadas en personas) de las cuales se selecciona una muestra.

FIGURA 26-1. Una distribución poblacional hipotética de las mediciones de la concentración de bilirrubina sérica

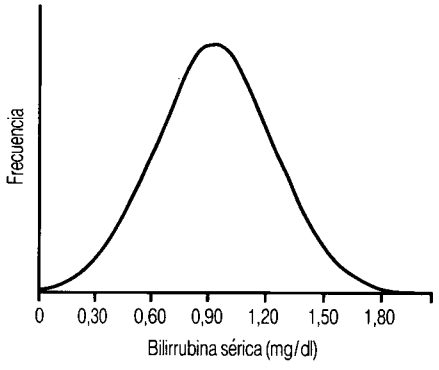
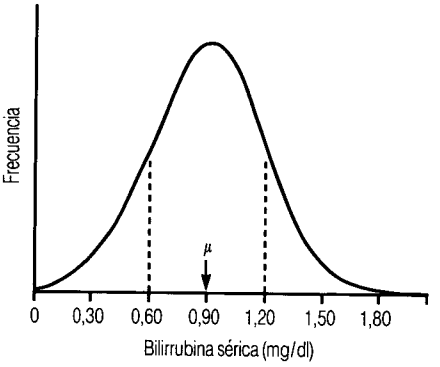


FIGURA 26-2. Una distribución gausiana hipotética de la concentración de bilirrubina sérica con una media de 0,9 mg/dl y una desviación estándar de 0,3 mg/dl. Las líneas discontinuas indican los valores iguales a la media ± la desviación estándar



Si marcamos en una gráfica la frecuencia con que aparecen los distintos valores de una variable en la población, obtendremos una representación gráfica de la *distribución poblacional*. La distribución poblacional describe la frecuencia con que aparecen los valores en la población de la que se extraen las muestras para observación (figura 26-1). No obstante, es difícil asimilar o transmitir la información contenida en los datos representados en esa gráfica.

Los métodos estadísticos ofrecen una medida sintética de la distribución poblacional, en lugar de su descripción gráfica. Cada tipo de distribución poblacional tiene un número limitado de valores sintéticos, denominados *parámetros*, que se utilizan para describir completamente la distribución concreta de las mediciones. Por ejemplo, para describir íntegramente una *distribución gausiana*,[2] se necesitan dos parámetros: la *media*[3] (la posición de la distribución en una escala continua o, más concretamente, su "centro de gravedad") y la *desviación estándar*[4] (la *dispersión* de la distribución, dado que indica cuán alejados de la media se encuentran los valores individuales). La figura 26-2 muestra una distribución gausiana con la media, como medida de posición de la distribución, y la desviación estándar, como medida de dispersión.

---

[2] La distribución gausiana también se conoce como distribución *normal*. Evitaremos la utilización del último término, porque normal tiene otro sentido en medicina. La distribución gausiana es la distribución poblacional que se supone la mayor parte de las veces en estadística.

[3] Con frecuencia el término *promedio* se utiliza como sinónimo de media. En terminología estadística no son lo mismo. La media se calcula sumando todas las mediciones y dividiéndolas por el número de mediciones realizadas. Un promedio, por su lado, se calcula multiplicando cada una de las mediciones por unos valores concretos, denominados *pesos*, antes de sumarlos. Esta suma se divide entonces por la suma de los pesos. La media es un tipo especial de promedio en el cual el peso de cada medición es igual a 1.

[4] La desviación estándar (σ) es la raíz cuadrada de la varianza (σ^2). La varianza es igual a la suma de las desviaciones de los datos (x_i) respecto de la media (μ) al cuadrado. Por lo tanto, la desviación estándar poblacional es

$$\sigma = \sqrt{\frac{\Sigma (x_i - \mu)^2}{N}}.$$

Para demostrar lo que queremos decir con la posición de una distribución, supongamos que la media de la concentración sérica de bilirrubina en la población es de 1,2 ml/dl, en lugar de 0,9 mg/dl. La distribución gausiana de la concentración sérica de la bilirrubina sería entonces como la que aparece en la figura 26-3.

Observe que la forma general de la distribución de la figura 26-3 no se modifica al cambiar la media, pero la posición de su centro de gravedad se mueve 0,3 mg/dl hacia la derecha. No obstante, si hubiésemos cambiado la dispersión de la distribución de la figura 26-2, su forma se habría modificado sin cambiar su posición. Por ejemplo, compare la distribución de la figura 26-2 con la de la figura 26-4, en la cual se ha cambiado la distribución estándar de 0,3 mg/dl a 0,4 mg/dl.

ESTIMACIÓN E INFERENCIA

En muy pocas ocasiones podemos realizar todas las mediciones posibles en una población. No obstante, podemos calcular valores numéricos para *estimar* el valor de los parámetros de la población mediante el empleo de las mediciones observadas en una muestra extraída de esa población. Estas estimaciones muestrales de los parámetros poblacionales son el fin que persiguen los métodos estadísticos. De hecho, ¡esas estimaciones se denominan *estadísticos*! Un estadístico individual utilizado para estimar el valor de un parámetro poblacional determinado se conoce como *estimación puntual*. Estas estimaciones puntuales son los estadísticos que usamos para resumir grandes cantidades de mediciones en unas pocas manejables.

Hasta el momento, solo hemos considerado la primera finalidad de los métodos estadísticos: sintetizar las observaciones. No obstante, es un paso importante para valorar la influencia del azar en esas observaciones. Como hemos afirmado anteriormente, una muestra es un subgrupo de todas las posibles mediciones de una población. En *todos* los métodos estadísticos se supone que la muestra es un subgrupo *aleatorio* de la población de la que se ha extraído. Aunque los subgrupos aleato-

FIGURA 26-3. Una distribución gausiana hipotética de la concentración de bilirrubina sérica con una media de 1,2 mg/dl y una desviación estándar de 0,3 mg/dl. La comparación de esta distribución con la de la figura 26-2 ilustra lo que se pretende decir con posiciones diferentes de las distribuciones poblacionales

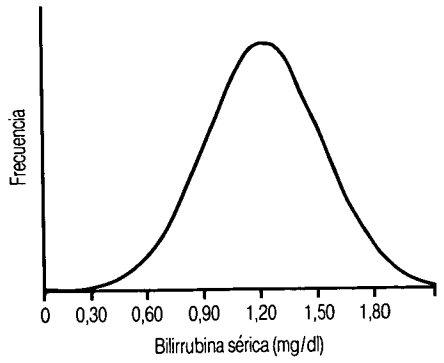
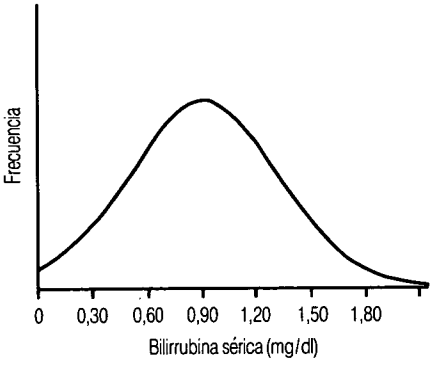


FIGURA 26-4. Una distribución gausiana hipotética de la concentración de bilirrubina sérica con una media de 0,9 mg/dl. La comparación de esta distribución con la de la figura 26-2 ejemplifica lo que se pretende decir con dispersiones diferentes de las distribuciones poblacionales



rios se pueden obtener por distintos métodos, en *La selección de una prueba estadística* solo consideraremos el más simple de todos ellos (y el más habitual), denominado *muestra aleatoria simple*. En una muestra aleatoria simple, todas las mediciones de la población tienen la misma probabilidad de ser incluidas en la muestra.[5] Por consiguiente, el azar dicta cuáles de esas mediciones se incluyen realmente en la muestra.

Cuando se estiman los parámetros poblacionales utilizando estadísticos muestrales, la selección aleatoria de las mediciones realmente incluidas en la muestra determina cuánto se aproxima el estadístico muestral al valor real del parámetro poblacional. Lamentablemente, nunca sabemos cuán correctamente un estadístico refleja el valor del parámetro poblacional correspondiente, porque tendríamos que efectuar mediciones en todos los integrantes de la población para conocer los parámetros poblacionales reales. No obstante, lo que podemos saber es cuánto se espera que varíe el estadístico en relación con el valor hipotético del parámetro poblacional sobre la base de la variabilidad del azar entre las muestras aleatorias. Este conocimiento constituye la base de la *inferencia estadística* o de las *pruebas de significación estadística*.

El marco de la inferencia estadística ha sido descrito en la Sección 1. En ese apartado se señaló que las pruebas de significación estadística se realizan suponiendo que la hipótesis nula es cierta. La hipótesis nula nos proporciona el valor hipotético con el que podemos comparar nuestras estimaciones.

Como se ha comentado en la Sección 1, el "objetivo" en las pruebas de significación estadística es el cálculo del valor *P*.[6] El valor *P* se calcula a partir de las observaciones de la investigación convirtiéndolas, en primer lugar, a una *distribución estándar*. Utilizamos una distribución estándar, porque los valores *P* se pueden obtener a partir de las tablas estadísticas en cualquier lugar de estas distribuciones. Buena parte de lo que se considera metodología de la estadística tiene que ver con la conversión de las observaciones a una distribución estándar.[7]

En la Sección 1 también comentamos que una alternativa al uso de las pruebas de significación estadística para investigar la influencia del azar en las estimaciones muestrales es el cálculo del *intervalo de confianza* o la *estimación por intervalo*.[8] Dentro de un intervalo de confianza, tenemos un nivel de confianza determinado (con frecuencia de 95%) de que está incluido el parámetro poblacional.[9] Generalmente, los intervalos de confianza se calculan modificando mediante el álgebra los cálculos realizados en las pruebas de significación estadística.

Cuando realizamos una prueba de significación estadística o calculamos un intervalo de confianza, podemos usar técnicas *unilaterales* o *bilaterales*. Una prueba de significación estadística *bilateral* o estimación por intervalo se emplea cuando

---

[5] En un sentido general, una muestra aleatoria implica que cualquier individuo en la población tiene una probabilidad conocida de ser incluido en la muestra. Aquí limitamos esas probabilidades conocidas a la condición de que sean iguales.

[6] Recuerde que el valor *P* es la probabilidad de obtener una muestra que sea como mínimo tan distinta de la indicada por la hipótesis nula como la muestra realmente obtenida si la hipótesis nula realmente describe la población. *No* es, como se supone frecuentemente, la probabilidad que el azar haya influido sobre las observaciones muestrales. Esa probabilidad es igual a 1 (es decir, estamos seguros de que el azar ha influido en nuestras observaciones).

[7] Ejemplos de distribuciones estándares son la normal, la de la *t* de Student, la de ji al cuadrado y la de la *F*. Estas distribuciones se presentarán en capítulos posteriores.

[8] Algunas veces, este intervalo se denomina "límites de confianza". En la terminología estadística, los límites de confianza son los valores numéricos que marcan los límites de un intervalo de confianza.

[9] En la estadística clásica, una *estimación por intervalo* significa que, si examinamos un número infinito de muestras de un mismo tamaño, un porcentaje determinado (esto es, el 95%) de las estimaciones por intervalo incluirán el parámetro poblacional. Una visión más moderna entre los estadísticos es que esto equivale a suponer que existe una determinada posibilidad (de 95%) de que el valor del parámetro poblacional esté incluido en el intervalo. Esta última interpretación es la que habitualmente tiene interés para el investigador en medicina.

el investigador no está seguro en qué lado del valor del parámetro implicado en la hipótesis nula se encuentra realmente el parámetro poblacional. Esta es la situación habitual, pero en algunas circunstancias se pueden encontrar en la literatura médica pruebas de significación estadística o estimaciones por intervalo *unilaterales*. Una prueba o intervalo de confianza unilateral se aplica cuando el investigador está dispuesto a suponer que conoce la dirección del efecto estudiado y el análisis solo se centra en el examen de la magnitud o de la fuerza de tal efecto.

Para ilustrar la distinción entre las técnicas unilaterales o bilaterales, imaginaremos un ensayo clínico en el que se mide la tensión arterial diastólica en un grupo de individuos antes y después del tratamiento con un nuevo fármaco antihipertensivo. Antes de examinar los datos resultantes de este estudio, podríamos suponer en nuestra hipótesis de estudio que la tensión arterial diastólica disminuye cuando los pacientes toman el medicamento. En otras palabras, supondríamos que es *imposible* que el medicamento aumente la tensión arterial diastólica. Con este supuesto, la prueba de significación estadística o la estimación por intervalo puede ser unilateral y la potencia estadística de nuestro análisis aumentará. Por otro lado, si nuestra hipótesis de estudio es que la tensión arterial diastólica cambiará cuando los pacientes tomen el medicamento, las pruebas de significación o la estimación por intervalo deben ser bilaterales. Esto se debe a que consideramos *posible*, aunque improbable, que el nuevo medicamento antihipertensivo aumente la presión arterial diastólica.

LA SELECCIÓN DE LOS MÉTODOS ESTADÍSTICOS

Centremos ahora nuestra atención en la selección de los métodos estadísticos para analizar los datos de la investigación médica. Antes de seleccionar un método, debemos tomar dos decisiones: 1) cuál es la variable dependiente y cuál la independiente, y 2) qué tipo de datos constituyen cada una de esas variables. En primer lugar, veamos qué queremos decir con variables dependientes e independientes.

Una *variable* es una característica que se mide en un estudio. Por ejemplo, si medimos la edad, podemos hablar de la edad como una de las variables de nuestro estudio. La mayor parte de los métodos estadísticos distinguen entre variables *dependientes* e *independientes*. Así se indican las funciones o el propósito de una variable en un análisis determinado. Por lo general, una serie de variables diseñadas para investigar una hipótesis de estudio solo incluirá una variable dependiente. Esta variable dependiente puede identificarse como la de interés principal o el desenlace principal del estudio. Queremos contrastar hipótesis o hacer estimaciones, o efectuar ambos procedimientos, acerca de la variable dependiente. Por otro lado, en la serie de variables puede que no haya ninguna variable independiente o que se incluya una o más. Las variables independientes determinan las características que es necesario tener en cuenta o las condiciones en que se contrastan las hipótesis o se realizan las estimaciones.

Para ilustrar la distinción entre variables dependientes e independientes, considere un estudio de cohortes en el que se investiga la relación entre el consumo de tabaco y la enfermedad coronaria. Suponga que solo se miden dos variables en cada individuo: consumo de tabaco (frente a no consumo) y enfermedad coronaria (frente a no enfermedad). Para analizar estos datos, primero decidimos que estamos interesados principalmente en estimar o contrastar una hipótesis sobre el riesgo anual de enfermedad coronaria. Por consiguiente, la enfermedad coronaria es la variable dependiente. Además, deseamos comparar el riesgo de enfermedad coronaria entre los fumadores y los no fumadores. Por este motivo, el consumo de tabaco es la variable independiente.

El número de variables independientes determina el tipo de método estadístico que es apropiado para analizar los datos. Por ejemplo, si nos interesara estimar el riesgo anual de enfermedad coronaria en una comunidad sin tener en cuenta el consumo de tabaco o cualquier otra característica de los individuos, aplicaríamos los métodos estadísticos conocidos como *análisis univariantes*. Estas técnicas se aplican a una serie de observaciones que contienen una variable dependiente y ninguna independiente. Para examinar el riesgo de enfermedad coronaria en relación con el hecho de ser fumador, como en el ejemplo anterior, usaríamos los métodos de *análisis bivariante*. Estos métodos se aplican a grupos de observaciones con una variable dependiente y una independiente. Por último, si nos interesara el riesgo de enfermedad coronaria en los individuos de diversas edades, sexo y hábito de fumar, aplicaríamos los métodos de *análisis multivariante* (*multivariable* en inglés).[10] Estos métodos se utilizan para grupos de observaciones que consisten en una variable dependiente y más de una independiente, como la edad, el sexo y el hábito tabáquico. Los métodos multivariantes se aplican con frecuencia para cumplir la tercera finalidad de los métodos estadísticos: ajustar según la influencia de las variables de confusión.

Las investigaciones médicas suelen incluir diversas series o grupos de variables. Por ejemplo, suponga que hemos realizado un ensayo clínico controlado en el cual los sujetos han recibido el fármaco X o un placebo para facilitar su recuperación de una enfermedad determinada. Dado que nos interesa conocer la influencia de la edad y el sexo en la recuperación (porque la edad y el sexo pueden ser variables de confusión), las incluimos en los registros de datos de la investigación. Por lo tanto, nuestro estudio contiene cuatro variables: tratamiento (fármaco X o placebo), recuperación (sí o no), edad y sexo. En el grupo de datos que incluye las cuatro variables, la recuperación sería la variable de interés, es decir, la variable dependiente. El tratamiento, la edad y el sexo serían las variables independientes, que reflejan nuestro interés en analizar la recuperación en relación con el tratamiento específico que ha recibido el sujeto, su edad y sexo. Sin embargo, incluso antes de contrastar hipótesis o de realizar estimaciones sobre la recuperación, probablemente nos interesaría saber si mediante la asignación al azar de los participantes se obtuvieron distribuciones de edad desiguales en los dos grupos de tratamiento. El grupo de variables que nos permitiría comparar las distribuciones de edad incluye la edad como variable dependiente y el tratamiento como variable independiente, ya que la edad es la variable de interés y el grupo de tratamiento, la condición en la que estamos valorando la edad. Por este motivo, la decisión sobre cuál es la variable dependiente y cuál la independiente *depende de la pregunta que se intenta responder*.

TIPOS DE DATOS

Además de caracterizar la función de las variables en el análisis, para seleccionar la prueba estadística debemos determinar el tipo de datos que constituyen las mediciones de cada variable. Con el fin de categorizar los tipos de datos, realizaremos una primera distinción entre datos *continuos* y *discretos*.

[10] Un error habitual en el uso de la terminología estadística es referirse a las técnicas diseñadas para una variable dependiente y varias independientes como análisis *multivariado* (*multivariate* en inglés). Sin embargo, este término se refiere en rigor a las técnicas diseñadas para tratar con *más de una* variable dependiente. El uso de técnicas multivariadas es raro en la investigación médica. No hemos incluido estas técnicas en nuestro diagrama y mencionamos el término en su aplicación más habitual (variables dependientes nominales multivariantes).

Los datos continuos se definen como los que ofrecen la posibilidad de observar alguno de ellos entre un número infinito de valores regularmente espaciados entre dos puntos cualesquiera de su intervalo de medidas. Son ejemplos de datos continuos la tensión arterial, la concentración de colesterol sérico, la edad y el peso. Para cada una de estas variables podemos escoger dos valores cualesquiera e imaginar mediciones intermedias que sería posible observar, al menos, teóricamente, entre esos valores. Podemos considerar, por ejemplo, las edades de 35 y 36 años. Podríamos imaginar que las edades entre los 35 y 36 años se distinguen por el número de días transcurridos desde el 35o. cumpleaños de la persona. Además, podríamos imaginar el número de horas y de minutos que han transcurrido desde el cumpleaños. Teóricamente, no existe un límite de la precisión con que podríamos medir el tiempo. No obstante, observe que no es necesario que los datos continuos tengan un intervalo infinito de posibles valores, sino un número infinito de posibles valores dentro de su intervalo. Este intervalo puede tener, y de hecho lo tiene frecuentemente, un límite superior y uno inferior. La edad es un buen ejemplo. El límite inferior es cero y es difícil imaginar individuos que tengan edades por encima de los 120 años.

Los datos discretos, por otro lado, solo pueden tener un número finito de valores en su intervalo de medidas. Son ejemplos de datos discretos el número de hijos, el estadio de las enfermedades y el sexo. Para cada una de estas variables podemos seleccionar dos valores entre los cuales no es posible imaginar otros valores. Por ejemplo, no podemos imaginar que el número de hijos de una familia se encuentre entre 2 y 3.

En la práctica, a veces no se puede distinguir claramente entre datos continuos y discretos. Esto ocurre porque no existe ninguna variable en la que podamos medir realmente un número infinito de valores.[11] Este problema se soluciona al reconocer que, si se puede efectuar un elevado número de mediciones en el intervalo de medidas posibles y si los intervalos entre las mediciones son uniformes, esas mediciones son virtualmente continuas. Sin embargo, esto crea otra fuente de confusión, pues permite que se redefinan como continuos datos que son, incluso teóricamente, discretos. Por ejemplo, el número de cabellos en la cabeza es con certeza un dato discreto: no podemos imaginar un valor entre 99 999 y 100 000 cabellos. Con todo, el número de posibles valores dentro del intervalo del número de cabellos es muy elevado. ¿Podemos considerar esta variable como realmente continua? Sí; para casi todos los fines sería totalmente correcto.

Los datos pueden definirse además por su *escala* de medida. Los datos continuos se miden en escalas, denominadas *escala de razón* y *escala de intervalo*,[12] que se definen por estar constituidas por un intervalo constante o uniforme entre mediciones consecutivas. Algunas mediciones discretas se realizan en una *escala ordinal*. Los datos en una escala ordinal tienen una ordenación o posición específica, como en el caso de los datos continuos, pero no es preciso que el intervalo entre mediciones consecutivas sea constante. Un tipo de variable que se mide habitualmente con una escala ordinal es la clasificación conocida como el *estadio de la enfermedad*. Sabemos que el estadio 2 es más avanzado que el 1, pero no podemos afirmar que la diferencia entre los dos estadios sea la misma que la diferencia entre el estadio 2 y el 3.

---

[11] Por ejemplo, podríamos imaginar, aunque no medir, la tensión arterial en, digamos, picómetros de mercurio. Así que, en realidad, ¡todos los datos son discretos!

[12] La distinción entre la escala de razón y la de intervalo consiste en que la primera incluye el valor cero verdadero mientras que la segunda no. Cierto tipo de datos discretos, como los recuentos, tienen intervalos uniformes entre las mediciones y, por lo tanto, también se miden mediante escalas de razón o de intervalo. Otros tipos de datos discretos se miden en escalas ordinales o en escalas nominales.

Cuando no se puede aplicar algún tipo de ordenamiento a los datos discretos, decimos que se midieron en una *escala nominal*. Son ejemplos de características medidas con datos discretos en escalas nominales el tratamiento, el sexo, la raza y el color de los ojos. Los datos que tratamos como nominales incluyen mediciones con dos categorías, aunque se pueda considerar que tienen un orden intrínseco, porque uno es claramente mejor que el otro (por ejemplo, vivo y muerto).

Es importante darse cuenta de que el término variable nominal puede causar confusión. En su uso común, una variable nominal es una característica como el sexo o la raza que tiene dos o más categorías potenciales. Sin embargo, desde un punto de vista estadístico, una variable nominal se limita solamente a dos categorías. De este modo, debemos referirnos a la raza o al color de los ojos como datos nominales que requieren más de una variable nominal. El número de variables nominales es igual al número de categorías potenciales menos uno.

Con el fin de seleccionar una técnica estadística o de interpretar el resultado de una técnica, es importante distinguir entre tres categorías de variables:

1. *Variables continuas* (comprenden datos continuos como la edad y datos discretos que contienen un número elevado de posibles valores como el número de cabellos).
2. *Variables ordinales* (comprenden los datos ordinales con un mínimo de tres valores posibles aunque con un número total limitado, como los estadios de los tipos de cáncer).
3. *Variables nominales* (comprenden los datos nominales que no tienen un orden como la raza, y los datos que solo pueden tomar dos valores posibles, como vivo o muerto).

El orden en el que se han enumerado estos tres tipos de variables indica la cantidad relativa de información que cada una contiene. Esto es, las variables continuas contienen más información que las ordinales y estas, más que las nominales. Por esta razón, las variables continuas se sitúan a un nivel más elevado que las ordinales y estas, a un nivel más elevado que las nominales.

Las mediciones de un nivel de información concreto pueden ser *reescaladas* a un nivel inferior. Por ejemplo, la edad (medida en años) es una variable continua. Podríamos reescalarla de forma legítima y transformarla en una variable ordinal al definir a las personas como niños (0–18 años), jóvenes (19–30 años), adultos (31–45 años), adultos maduros (45–65 años) y ancianos (>65 años). Podríamos reescalarla otra vez para convertirla en una variable nominal. Por ejemplo, podríamos dividir las personas en dos categorías: jóvenes y viejas. Sin embargo, no podemos reescalar las variables a un nivel superior al que se midieron realmente.

Cuando reescalamos una medida a un nivel inferior perdemos información. Es decir, tenemos menos detalles sobre una característica si la medimos en una escala nominal que si la medimos en escala ordinal o continua. Por ejemplo, sabemos menos acerca de una persona cuando la identificamos como de edad madura que si decimos que tiene 54 años. Si una persona tuviera 54 años de edad y midiéramos su edad en una escala continua, podríamos distinguir su edad de la de otra persona que tuviera 64 años. Sin embargo, si la edad se registrara en la escala ordinal antes indicada, no podríamos diferenciar la edad de estos individuos.

La pérdida de información que se produce al utilizar mediciones reescaladas en las técnicas estadísticas tiene el efecto de aumentar el error de tipo II, si todo lo demás se mantiene igual. Es decir, reescalar a un nivel inferior reduce la potencia estadística, lo que hace más difícil establecer el nivel de significación estadística y,

en consecuencia, rechazar una hipótesis nula falsa. Por otra parte, reescalando a un nivel inferior se evita la necesidad de aceptar ciertos supuestos, como la uniformidad de los intervalos, que puede ser un requisito para realizar determinadas pruebas estadísticas. En los siguientes capítulos se describirán con mayor detalle varios ejemplos concretos de determinadas pruebas que requieren estos supuestos y de las que permiten evitarlos.

Hasta aquí, hemos revisado los pasos iniciales que deben darse para seleccionar una prueba estadística. Estos pasos son:

1. Identificar una variable dependiente y todas las variables independientes a partir de la pregunta que se intenta responder con el estudio.
2. Determinar si cada variable es continua, ordinal o nominal.

Una vez completados estos pasos, estamos preparados para iniciar el proceso de selección de una prueba estadística.

EL ESQUEMA

Los capítulos restantes de este libro están organizados como ramas de un esquema diseñado para facilitar la selección e interpretación de los métodos estadísticos. Se han incluido la mayor parte de ellos, aunque no todos los que pueden encontrarse en la literatura médica.

Para usar este esquema (figura 26-5), primero se debe determinar cuál es la variable dependiente entre el grupo de variables. Si hay más de una variable dependiente que usted quiere considerar simultáneamente en el mismo análisis, quizá le interese un análisis multivariante para el cual debe consultar a un estadístico. Si su grupo de variables parece contener más de una variable dependiente, es muy probable que los datos planteen más de una hipótesis de estudio. En ese caso, se deben considerar las variables relevantes para una hipótesis de estudio específica.

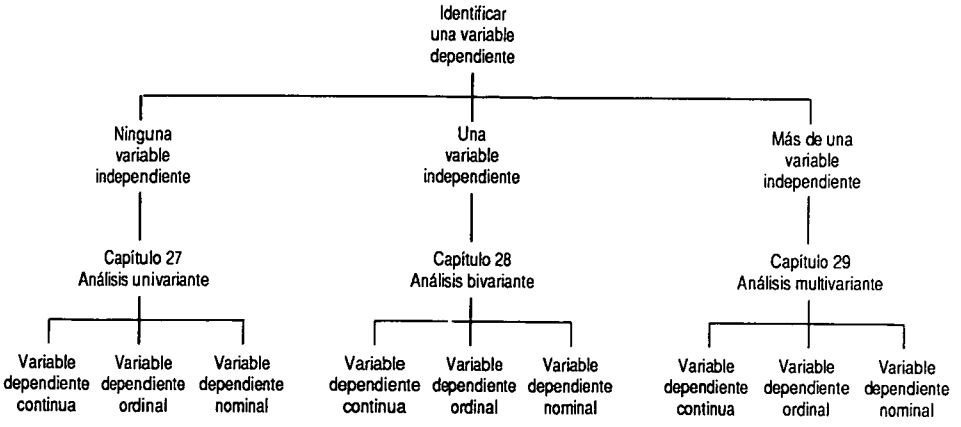
Una vez identificada una sola variable dependiente, el número de variables independientes en la investigación le orientará hacia el capítulo que trata de ese número de variables independientes. Cada capítulo contiene tres grandes divisiones. La primera hace referencia a los grupos de variables en los que la variable dependiente es continua. La segunda se centra en las variables dependientes ordinales y la tercera, en las variables dependientes nominales. Dentro de cada una de estas divisiones se describen las técnicas para variables independientes continuas, ordinales y nominales, cuando se dispone de ellas. El capítulo 30 reúne los esquemas presentados en los capítulos 27, 28 y 29.

RESUMEN

En este capítulo hemos aprendido que los métodos estadísticos utilizados para analizar los datos de una investigación médica tienen tres finalidades. La primera es la de resumir los datos. Las distribuciones de los datos en grandes poblaciones se resumen mediante valores numéricos denominados parámetros. Los valores de estos parámetros poblacionales se estiman a partir de muestras aleatorias mediante estimaciones puntuales denominadas estadísticos.

La segunda finalidad de la estadística es la de tener en cuenta la influencia del azar en las estimaciones puntuales calculadas a partir de las observaciones muestrales seleccionadas al azar de la población. Hay dos enfoques generales para considerar el azar. Uno está constituido por las pruebas de significación estadística. Bajo este enfoque, las observaciones muestrales se comparan con lo que sería de esperar si

FIGURA 26-5. Esquema para identificar el capítulo y la sección en los que se tratan las pruebas estadísticas referentes a un grupo de variables en particular



no existiese una asociación entre variables o una diferencia entre los grupos de la población. Si las observaciones son lo suficientemente inesperadas o no existe una verdadera asociación (o diferencia), rechazamos la hipótesis de que no existe una asociación (o diferencia) en la población. Un enfoque alternativo para considerar el azar es el cálculo de los intervalos de confianza de la estimación puntual. En este caso, podemos suponer con un grado de confianza determinado que el parámetro poblacional se halla incluido en el intervalo de confianza. Aunque las pruebas de significación estadística y la estimación por intervalo son procesos que aparentemente se interpretan de forma distinta, consisten sencillamente en expresiones matemáticas diferentes de un mismo principio.

La tercera finalidad de la estadística es la de ajustar los datos según el efecto de las variables de confusión en nuestras observaciones muestrales. Este objetivo se alcanza mediante el análisis multivariante, que será el tema que nos ocupará en el capítulo 29.

Para cumplir con estas finalidades, debemos seleccionar una técnica estadística apropiada para responder a la cuestión en estudio. Para realizar esta selección, procederemos de la siguiente forma:

1. Decidir cuál es la variable dependiente. Esta será la variable de interés principal en la hipótesis del estudio. Las variables restantes son las variables independientes.

2. Determinar cuántas variables independientes contiene el conjunto de observaciones. Si no existe ninguna, debemos realizar un análisis univariante. Con una variable independiente, el análisis bivariante es el apropiado. Si, por otro lado, la serie contiene más de una variable independiente, usaremos un método multivariante. Recuerde que para los datos nominales, el número de variables es igual al número de categorías potenciales menos una.

3. Definir qué tipo de variable dependiente es la de interés. Si la variable dependiente tiene un número ilimitado de valores uniformemente espaciados, se trata de una variable continua. Si la variable dependiente contiene un número de valores limitado que pueden seguir un orden, se trata de una variable ordinal. Una variable dependiente nominal simplemente identifica la presencia o la ausencia de una condición.

CAPÍTULO 27

ANÁLISIS UNIVARIANTES

Para analizar un conjunto de mediciones que contiene una variable dependiente y ninguna independiente, los métodos estadísticos utilizados son un tipo de *análisis univariante*. En la literatura médica, el análisis univariante tiene tres usos comunes. El primero se encuentra en estudios descriptivos (por ejemplo, en las series de casos) en los que solo se ha examinado una muestra. Por ejemplo, un investigador puede presentar una serie de casos de una enfermedad determinada describiendo diversas mediciones demográficas y patofisiológicas de los pacientes. El propósito del análisis en ese estudio sería el de explicar la influencia del azar en las mediciones de cada característica. Dado que no existen grupos diferentes de personas para comparar, ni interés en comparar una característica con otra, cada característica de las personas enfermas se considera una variable dependiente en un análisis univariante individual.

El análisis univariante también se utiliza comúnmente cuando se extrae una muestra para incluirla en un estudio. Por ejemplo, antes de hacer la selección aleatoria en un ensayo clínico controlado, puede que sea conveniente realizar mediciones en toda la muestra objeto de estudio. Es decir, podríamos determinar el porcentaje y la media de edad de las mujeres en el grupo seleccionado para muestreo al azar antes de asignarlas al grupo de control o al de estudio. Como en el estudio descriptivo comentado antes, cada característica examinada en la muestra es una variable dependiente en un análisis univariante individual.

Por lo general, en los estudios descriptivos o cuando se examina una sola muestra, el interés se centra en la estimación puntual y por intervalo, en lugar de las pruebas de significación estadística. En el esquema univariante se pueden realizar pruebas de hipótesis, pero en la hipótesis nula debe especificarse un valor para el parámetro poblacional. Muchas veces esto no se puede hacer en el análisis univariante. Por ejemplo, es difícil imaginar qué valor se tomará como hipótesis de la prevalencia de hipertensión entre los individuos de una comunidad determinada.[1] Sin embargo, en la tercera aplicación del análisis univariante es fácil imaginar ese valor hipotético. Este es el caso en el que una medición, como la tensión arterial diastólica, se realiza dos veces en el mismo individuo o en uno muy semejante y el interés se centra en la diferencia entre las dos mediciones. En esta aplicación, es lógico imaginar una hipótesis nula que afirme que la diferencia entre las dos mediciones es igual a cero. De este modo, la diferencia entre las mediciones de la tensión arterial diastólica es la variable dependiente. Aunque la diferencia, por su misma naturaleza, es una comparación entre grupos, las diferencias en sí mismas *no* son comparadas con ningún grupo. Por lo tanto, no existe ninguna variable independiente. Cuando se comparan dos mediciones en un mismo individuo o en individuos muy semejantes, estamos tratando con un problema univariante. Por eso, en una investigación que emplea datos apareados y en la que cada par

[1] A primera vista puede parecer que la hipótesis nula sería que la prevalencia en una comunidad determinada es igual a la prevalencia en otra comunidad o a la prevalencia estimada en otro estudio. Sin embargo, es importante recordar que el valor propuesto como parámetro poblacional tiene que ser *conocido sin error*. Esto no sería cierto a no ser que todos los miembros de la comunidad que se compara se incluyeran en el cálculo de la prevalencia.

constituye una observación, los datos se analizan usando métodos univariantes. Los pares pueden consistir en datos de un individuo o de dos individuos que se aparean antes de analizar los datos.

VARIABLE DEPENDIENTE CONTINUA

Comenzaremos a analizar la figura 27-1 preguntando ¿cuál es el aspecto de la distribución poblacional que nos interesa, su posición o su dispersión?[2] A continuación es preciso considerar la estimación puntual que puede emplearse para representar ese aspecto de la distribución poblacional.

En el análisis univariante de una variable dependiente continua se acostumbra suponer que los datos provienen de una población con una distribución gausiana. Por lo tanto, la media se usa habitualmente como medida de posición. La dispersión de las distribuciones gausianas se mide mediante la desviación estándar o, alternativamente, por el cuadrado de la desviación estándar, denominado *varianza*. Para fines de análisis, tanto la varianza como el coeficiente de variación —descrito más adelante— se usan para medir la dispersión de los datos de la distribución poblacional. Por último, cada diagrama clasificará la categoría general de las pruebas estadísticas que se emplean más frecuentemente para calcular los intervalos de confianza o para contrastar las hipótesis estadísticas.

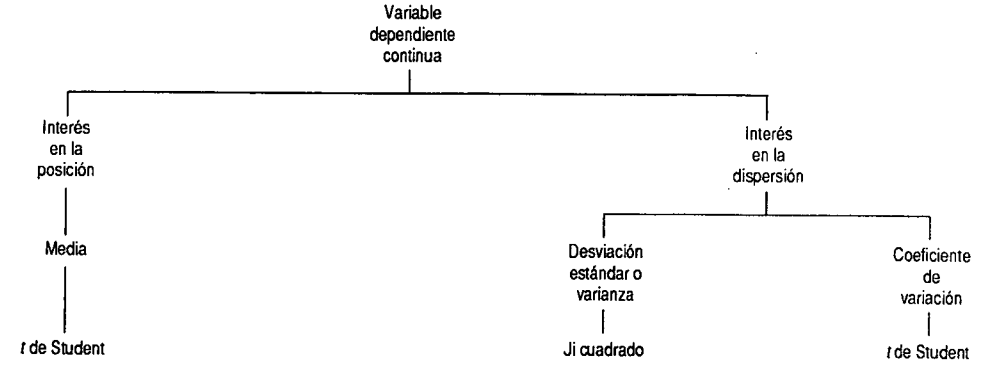
En el capítulo 26 aprendimos que los primeros pasos para escoger una técnica estadística son:

1. Decidir cuál es la variable dependiente.
2. Determinar cuántas variables independientes, si las hubiera, contiene el grupo de observaciones.
3. Definir el tipo de variable dependiente como continua, ordinal o nominal.

A estos pasos, ahora añadimos el siguiente:

4. Seleccionar el parámetro de la distribución poblacional sobre el que desearíamos contrastar hipótesis o efectuar estimaciones. En otras palabras, ¿nos interesa la posición o la dispersión?

FIGURA 27-1. Esquema para seleccionar un método estadístico univariante para variables dependientes continuas (continuación de la figura 26-5)



[2] En los siguientes capítulos centraremos nuestro interés en la posición.

Si seguimos estos pasos en la figura 27-1, observamos que nos conducen al nombre de un tipo general de pruebas estadísticas. Estas pruebas suelen ser apropiadas tanto para determinar la significación estadística como para calcular los intervalos de confianza.

Interés en la posición

Como se ha afirmado anteriormente, la media muestral es una estimación de la posición de la media poblacional. A menudo, la media poblacional es el parámetro que intentamos estimar. Para calcular el intervalo de confianza de la media de una muestra, la *distribución de la t de Student* es la más frecuentemente empleada. La distribución de la *t* de Student es una distribución estándar en la cual se transforman las medias de variables dependientes continuas para facilitar el análisis. Esta distribución es parecida a la gausiana, pero requiere de un parámetro adicional conocido como *grados de libertad*. El propósito de los grados de libertad en la distribución de la *t* de Student es reflejar el papel del azar en la estimación de la desviación estándar.[3]

La distribución de la *t* de Student nos permite construir los intervalos de confianza a partir de la media observada y de su error estándar. En la Sección 3 se señaló que el error estándar de una media disminuye a medida que aumenta el tamaño de la muestra. De forma más precisa, el error estándar es igual a la desviación estándar dividida por la raíz cuadrada del tamaño de la muestra.

El error estándar se emplea en la distribución de la *t* de Student para calcular las estimaciones por intervalo de las medias de las variables continuas. El intervalo de confianza de una media es igual a la estimación muestral de la media + el valor de la *t* de Student para el nivel de confianza deseado y multiplicado por el error estándar. Para una estimación bilateral con un nivel de confianza de 95%, el valor de la *t* de Student es aproximadamente igual a 2 si las muestras contienen 20 casos o más. Sumando y restando a la estimación puntual de la media un valor igual al doble del error estándar, se puede obtener un intervalo de confianza *aproximado*. Es decir, la media poblacional se encuentra en el intervalo comprendido entre la media muestral ± dos errores estándares, con un nivel de confianza de 95%.[4] Por ejemplo, si leemos en un informe de investigación que la media ± el error estándar de la concentración de colesterol sérico en una muestra es igual a 150 ± 30 mg/dl, podemos tener un nivel de confianza de 95% de que la media poblacional se encuentra dentro del intervalo aproximado comprendido entre 120 y 180 mg/dl.

Como se mencionó anteriormente, en el análisis univariante existe una situación especial en la que se pueden aplicar las pruebas de significación estadística. El caso más frecuente es el de un estudio en el que una variable dependiente continua se mide dos veces en el mismo individuo. Por ejemplo, podríamos medir la tensión arterial antes y después de que un paciente reciba un medicamento antihipertensivo. Si lo que realmente nos interesa no son las mediciones antes y después del tratamiento,

---

[3] Al utilizar la distribución de la *t* de Student para realizar estimaciones por intervalo de las medias, se reconoce el hecho de que la desviación estándar se estima a partir de la muestra. Es decir, no se conoce con precisión la desviación estándar.

[4] De forma similar, se pueden estimar otros intervalos de confianza considerando múltiplos del error estándar. Aproximadamente dos tercios de las medias muestrales posibles se encuentran dentro de un error estándar de la media poblacional. Más de 99% de las posibles medias muestrales se encuentran dentro del intervalo de la media poblacional ± tres errores estándares. Sin embargo, es importante recordar que, cuando aplicamos estas interpretaciones a los intervalos de confianza o a sus aproximaciones, estamos suponiendo que la población de todas las medias posibles tiene una distribución gausiana.

sino la diferencia entre las mediciones, nos encontramos frente a un *un diseño apareado*.[5] Este es un problema univariante, dado que la variable dependiente es la diferencia entre las mediciones y no existe una variable independiente. Mediante un diseño apareado, hemos tratado de eliminar la influencia de la variación entre los sujetos en la medición inicial o *de base*.

De la misma manera que se emplea en otros análisis univariantes, la distribución *t* de Student se emplea para contrastar hipótesis o para realizar estimaciones por intervalo para los datos continuos a partir de un diseño apareado. Aunque las pruebas estadísticas utilizadas para analizar los datos de un diseño apareado no son distintas de otras pruebas univariantes, en los textos introductorios de estadística frecuentemente se tratan por separado. En estos casos, la prueba utilizada para examinar la diferencia entre las medias de los datos de un diseño apareado se denomina *prueba de la t de Student para datos apareados*.

Más que la media de la muestra ± el error estándar, con frecuencia vemos los datos univariantes presentados como la media de la muestra ± la desviación estándar. La media muestral ± el error estándar informa del nivel de confianza que podemos tener en nuestra estimación de la media poblacional. El error estándar es un indicador de la *dispersión de las medias muestrales* que podrían obtenerse extrayendo una muestra de la población. Sin embargo, la media de la muestra ± la desviación estándar plantea una cuestión distinta. La desviación estándar de los datos de la muestra estima la *dispersión de las mediciones* en la población. Aproximadamente, el 95% de los valores de una población se encuentran dentro del intervalo de la media poblacional ± dos desviaciones estándares.[6] Por lo tanto, cuando aplicamos una prueba estadística univariante a una variable dependiente continua, podemos estar interesados tanto en la estimación de la posición de la media poblacional y, por ese motivo, en su error estándar, como en la descripción de la dispersión de los valores y, por consiguiente, en la desviación estándar.

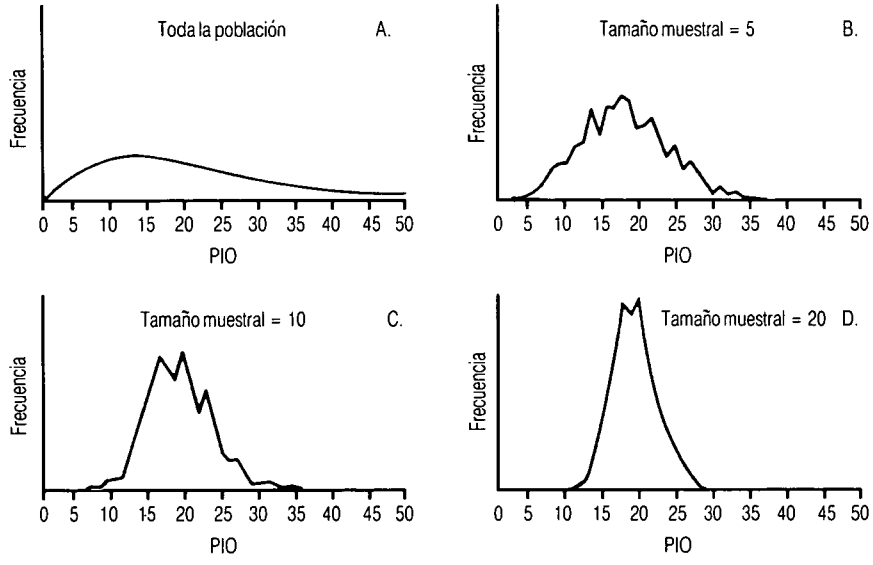
Para ilustrar cómo se escoge entre la presentación de la media ± la desviación estándar y la media ± el error estándar, imaginemos un estudio en el que se describe una serie de casos de una enfermedad determinada. Supongamos que una de las variables medidas en esos pacientes es la concentración del colesterol sérico. Si el objetivo del estudio es estimar los valores de la concentración del colesterol sérico que se podrían observar en los pacientes *individuales* con esa enfermedad, se debe presentar la desviación estándar, dado que estamos interesados en la dispersión de los datos poblacionales. Si, por otro lado, el propósito del estudio es estimar la media de la concentración del colesterol sérico de un *grupo* de pacientes con la enfermedad, se debe presentar el error estándar (o la estimación por intervalo), pues estamos interesados en la dispersión de las medias muestrales obtenidas al azar de la población.

Es importante entender la diferencia entre los supuestos que realizamos cuando interpretamos la media ± el error estándar y la media ± la desviación estándar. Cuando utilizamos el error estándar, suponemos que las medias de las muestras obtenidas al azar de la población siguen una distribución gausiana. En el caso de la media ± la desviación estándar, suponemos que los datos poblacionales por sí mismos

[5] Otro diseño apareado sería el correspondiente a una variable dependiente continua medida en dos individuos apareados que sean similares en las características compartidas que se considera posible que influyan en la magnitud de la variable dependiente.

[6] Asimismo, aproximadamente dos tercios de los datos poblacionales se encuentran dentro del intervalo formado por la media ± una desviación estándar y más de 99%, dentro del intervalo de la media ± tres desviaciones estándares. Para aplicar estas interpretaciones debemos suponer que los datos poblacionales siguen una distribución gausiana.

FIGURA 27-2. Demostración del teorema central del límite. Cuando medimos la tensión intraocular en muchos individuos (A) observamos que la distribución de las mediciones individuales no es gausiana. A pesar de ello, la distribución de la media de la presión intraocular tiende a seguir una distribución gausiana (B-D). Esta tendencia aumenta a la par que el tamaño muestral



siguen una distribución gausiana. A menudo este supuesto será cierto para la media ± el error estándar, como veremos, si escogemos muestras suficientemente grandes. Sin embargo, el supuesto muchas veces no será cierto para la media ± la desviación estándar.

Si los datos poblacionales siguen una distribución gausiana, las medias de las muestras de esa población también seguirán una distribución gausiana. Incluso cuando los datos poblacionales no siguen una distribución gausiana, las medias de un elevado número de muestras obtenidas mediante muestreos aleatorios repetidos de la misma población a la larga seguirán una distribución gausiana (figura 27-2). La probabilidad de que las medias sigan una distribución gausiana aumenta a la par que el número de observaciones en cada muestra. Este importante fenómeno se conoce como el *teorema central del límite* y explica el interés de los estadísticos tanto en las medias como en la distribución gausiana. También les permite a los investigadores médicos emplear los métodos estadísticos que suponen una distribución gausiana para analizar los valores de las medias obtenidas de poblaciones en las que los datos no siguen una distribución gausiana. Esto supone una gran ventaja, ya que muchas de las variables de interés en medicina provienen de poblaciones en las cuales las distribuciones de los datos no son gausianas.

Interés en la dispersión

Con mucho, la media es el parámetro poblacional que se estima con mayor frecuencia en el análisis univariante de las variables continuas. Sin embargo, este no es el único parámetro que podemos estimar con ese tipo de datos y no es siempre el que mejor refleja nuestro interés por una serie de observaciones. Quizá nos in-

terese la dispersión de las mediciones en la población. En este caso, nuestro interés se centra en la varianza o, de forma equivalente, en su raíz cuadrada: la desviación estándar de la población.

Cuando queremos obtener una medida de posición de la población de la cual hemos extraído una serie de observaciones univariantes, generalmente estimamos esa posición con la media de la muestra. El error estándar refleja la dispersión de las medias de la muestra. Empleamos la distribución de la *t* de Student para contrastar hipótesis estadísticas o para realizar estimaciones por intervalo de la media poblacional. Por otro lado, cuando nos interesa la dispersión de los datos de la población por sí mismos, estimamos la desviación estándar o la varianza de la población a partir de nuestras observaciones muestrales. Si deseamos contrastar hipótesis estadísticas o construir intervalos de confianza de la varianza poblacional, empleamos la *distribución de ji al cuadrado*. Sin embargo, el uso de la varianza o de la desviación estándar puede inducir a error si deseamos comparar la dispersión entre grupos distintos. Examinaremos esta situación y una solución habitual.

Una de las propiedades teóricas de los datos que siguen una distribución gausiana es que la desviación estándar y la media son independientes. Es decir, para una media determinada, cualquier desviación estándar es igualmente probable. En la práctica, esto no ocurre con frecuencia. Por ejemplo, considere los pesos corporales desde el nacimiento hasta los 5 años de edad (cuadro 27-1). Queda claro que la variación del peso aumenta con la edad, así como el propio peso. Sin embargo, la asociación entre la media y la desviación estándar hace difícil comparar medidas de dispersión correspondientes a diferentes pesos medios. Por ejemplo, las variaciones de un kilogramo entre lactantes representan una variabilidad mucho mayor para su tamaño que una variación de un kilogramo en niños de 5 años de edad.

Una solución sencilla para este problema consiste en dividir la desviación estándar por la media con el fin de "ajustar" los datos según las diferencias entre las medias. Si multiplicamos esta razón por 100, obtenemos lo que se conoce como el *coeficiente de variación*. En el cuadro 27-2 se presentan los coeficientes de variación de los pesos corporales de niños varones.

CUADRO 27-1. Medias y desviaciones estándares del peso corporal (niños)

	Peso (kg)	
Edad (años)	Media	Desviación estándar
Nacimiento	3,50	0,53
1	10,20	1,01
5	18,50	2,17

(**Fuente:** Smith DS. *Growth and its disorders*. Philadelphia: Saunders; 1977.)

CUADRO 27-2. Medias y coeficientes de variación del peso corporal (niños)

	Peso (kg)	
Edad (años)	Media	Coeficiente de variación
Nacimiento	3,50	15,1%
1	10,20	9,9%
5	18,50	11,7%

(**Fuente:** Smith DS. *Growth and its disorders*. Philadelphia: Saunders: 1977).

El examen de las variaciones absolutas de los pesos, estimadas mediante la desviación estándar, sugiere que la menor variación se observa entre los recién nacidos (cuadro 27-1). Sin embargo, esta variación se da entre niños que, como promedio, pesan menos. La variación del peso *en relación con la media* del peso en cada grupo, tal como muestran los coeficientes de variación, sugiere precisamente lo contrario (cuadro 27-2). La variación del peso al nacer en relación con el peso total al nacer es mayor que en cualquier otra edad considerada.

Por este motivo, el coeficiente de variación es una medida útil para examinar la dispersión relativa de las variables dependientes continuas cuando se cree que la media y la desviación estándar no son independientes y queremos comparar estimaciones univariantes de dispersión. En los intervalos de confianza y las pruebas de hipótesis estadísticas del coeficiente de variación se utiliza la distribución de la *t* de Student.

VARIABLE DEPENDIENTE ORDINAL

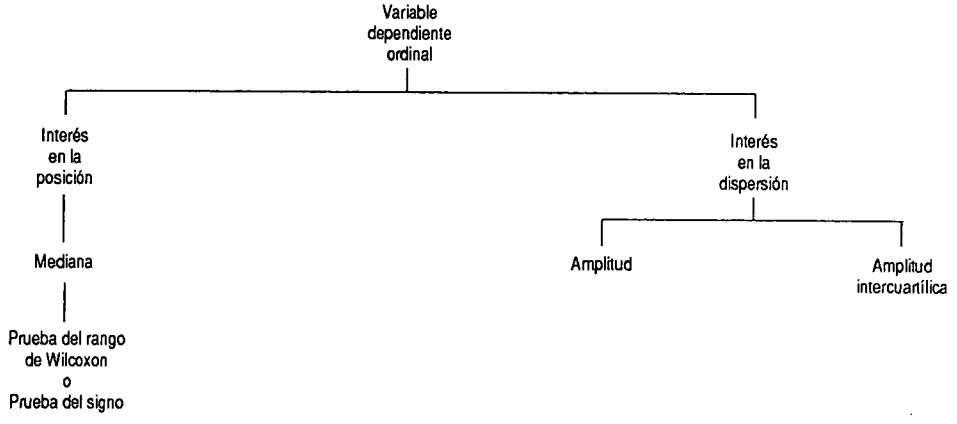
Los métodos estadísticos univariantes para las variables dependientes ordinales se presentan en la figura 27-3.

A diferencia de las variables continuas, con las variables ordinales no suponemos una distribución concreta de los datos poblacionales, tal como la distribución gausiana. Los métodos utilizados para las variables ordinales se denominan por este motivo de *distribución libre* o *no paramétricos*. Es importante darse cuenta de que estos métodos no están libres de supuestos. Por ejemplo, seguimos suponiendo que nuestra muestra es representativa de alguna población de interés.

Interés en la posición

Dado que no suponemos una distribución determinada de los datos medidos en una escala ordinal, no podemos estimar parámetros poblacionales que sinteticen la distribución. No obstante, es posible que nos interese describir la posición de los datos ordinales en una escala continua. Eso lo podemos hacer mediante la *mediana*. La mediana es el punto medio de una serie de datos, seleccionada de forma tal que la mitad de los valores sean más altos y la otra mitad más bajos que la mediana.

FIGURA 27-3. Esquema para seleccionar un método estadístico univariante para una variable dependiente ordinal (continuación de la figura 26-5)

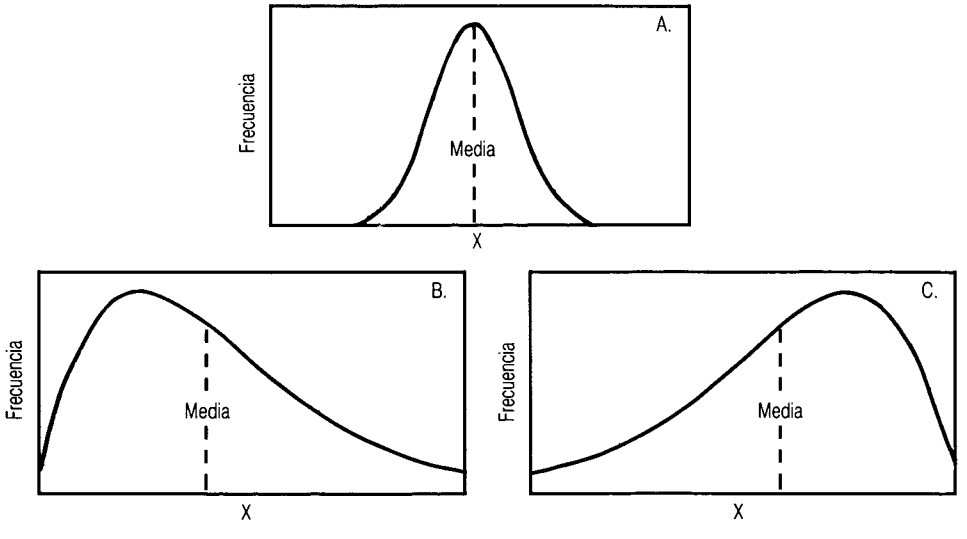


La mediana no tiene una distribución poblacional teórica como medida de su posición, pero puede utilizarse como una estimación *robusta*[7] de la media de una distribución gausiana. La mediana soslaya un supuesto que realizamos cuando calculamos la media: que los intervalos entre las mediciones de una distribución son uniformes y conocidos. Como la mediana se calcula empleando solamente el rango relativo u orden de las mediciones, la estimación de la mediana sería la misma independientemente de que los intervalos sean conocidos y uniformes o no. Por lo tanto, podemos usar la mediana para estimar la media de una población de datos continuos. Esto se lleva a cabo organizando las observaciones muestrales en orden relativo. De este modo, los datos continuos se convierten a una escala ordinal mediante la sustitución de los rangos por las observaciones reales.

En sentido estricto, la mediana puede emplearse como una estimación de la media poblacional solo cuando la distribución de la población es simétrica. Si esto es cierto, la media y la mediana poblacionales tienen el mismo valor (figura 27-4). No obstante, aunque la distribución poblacional sea simétrica, es posible que las observaciones obtenidas en una muestra de esa población sean, sin lugar a dudas, asimétricas. Un motivo habitual de esa asimetría es la posibilidad de incluir *valores extremos o aislados* (*outliers*) en la muestra. Estos valores extremos se producen en la población con muy poca frecuencia. En ocasiones, una muestra incluirá uno o más de estos valores extremos. Cuando esto sucede, las observaciones muestrales sugieren que esos valores extremos han aparecido con una frecuencia mayor de la que realmente tienen en la población.

Debido a que la media es el "centro de gravedad" de una distribución, su valor es influido más por los valores extremos que por los cercanos al centro

FIGURA 27-4. Posición de la media en una distribución simétrica (A) y en distribuciones asimétricas (B,C). *X* indica la posición de la mediana



[7] En términos estadísticos, una estimación robusta es aquella que no se ve sustancialmente influida por desviaciones menores de los supuestos de la prueba.

de la distribución. Por consiguiente, en las muestras que incluyen valores extremos, la media muestral puede ser bastante distinta de la poblacional. La mediana muestral, por su lado, es *resistente* a aquellos valores extremos. Es decir, los valores extremos tienen el mismo impacto sobre la mediana que los valores que se encuentran cerca del centro de la distribución muestral. Por lo tanto, paradójicamente, cuando una muestra de una distribución poblacional simétrica incluye valores extremos, la mediana muestral es un estimador de la media poblacional mejor que la media muestral.

El uso de la mediana para estimar la media poblacional constituye, sin embargo, un inconveniente. Dado que la mediana se basa solamente en la clasificación relativa de las observaciones, contiene menos información que la media. Siempre que utilizamos menos información al aplicar métodos estadísticos corremos un riesgo más elevado de cometer un error de tipo II. En otras palabras, la probabilidad de no poder rechazar una hipótesis nula incorrecta es más alta. Solo vale la pena correr ese riesgo cuando hay razones para sospechar que la información excluida crearía otros errores más graves si se incluyera en el análisis de los datos.

Aunque la mediana se emplea como una estimación robusta y resistente de la media poblacional, es importante recordar que también es por derecho propio una medida legítima de la posición de una distribución. Por ejemplo, si una distribución poblacional es asimétrica, podría interesar menos su centro de gravedad o media que su punto medio o mediana.

Si nos interesa contrastar la hipótesis nula de que la mediana es igual a cero en un análisis univariante, podemos emplear tanto la *prueba del rango con signo de Wilcoxon* como la *prueba del signo*. Habida cuenta de que la mediana no es un parámetro de ninguna distribución determinada, en general no podemos construir un intervalo para ese parámetro. Sin embargo, cuando se emplea la mediana como estimación robusta y resistente de la media poblacional, es correcto realizar una estimación por intervalo de esa media. Para esta estimación se dispone de métodos basados en la prueba del rango con signo de Wilcoxon y en la prueba del signo.[8]

Interés en la dispersión

Como ocurre con la media muestral, el cálculo de la desviación estándar supone que los intervalos entre los valores son conocidos y uniformes. El cálculo de la desviación estándar está influido en gran medida por los valores extremos. Como alternativa, en los artículos de investigación frecuentemente se presenta como medida de dispersión el *recorrido* (*range*) (diferencia entre el valor más alto y el más bajo). Aunque el recorrido es útil para describir la dispersión de un conjunto de observaciones muestrales, no es una buena estimación de la dispersión de los datos poblacionales. Esto se debe al hecho de que los valores de los extremos de la mayor parte de las distribuciones poblacionales raramente se observan en las poblaciones y, por este motivo, tampoco en las muestras. El recorrido se calcula a partir de esos extremos, así que el recorrido calculado en una muestra subestima el recorrido poblacional casi con toda seguridad. Por eso, según se reduce el tamaño muestral, la probabilidad de observar valores extremos también decrece. El resultado es que las estimaciones muestrales del recorrido varían directamente con el tamaño de la muestra.

[8] Del mismo modo, se podría calcular una estimación robusta y resistente de la desviación estándar (descrita más adelante) y emplear la distribución de la t de Student para construir un intervalo de confianza de la media poblacional.

Como alternativa, se puede utilizar el *recorrido intercuartílico* (*interquartile range*) para describir la dispersión de una muestra de observaciones, así como para estimar la dispersión en la población. Los cuartiles dividen una distribución en cuatro partes que contienen el mismo número de observaciones, de la misma forma que la mediana divide una distribución en dos partes iguales. El intervalo entre el valor de los datos que se encuentran un cuartil por debajo de la mediana y un cuartil por encima de la mediana se conoce como recorrido intercuartílico. Dentro de ese intervalo o recorrido se encuentran la mitad de los datos muestrales. Dado que el recorrido intercuartílico no depende de los valores extremos de una distribución, es mucho menos dependiente del tamaño de la muestra que el recorrido.

En una distribución gausiana, dos tercios de los valores poblacionales se encuentran en el intervalo comprendido por la media ± una desviación estándar. Por lo tanto, en una distribución gausiana, la media poblacional ± ⅔ del recorrido intercuartílico se puede considerar una estimación robusta y resistente de la media ± una desviación estándar. Si nos preocupa el supuesto de los intervalos conocidos y uniformes o si la muestra contiene valores extremos de validez cuestionable, podemos estimar la desviación estándar poblacional calculando los dos tercios del recorrido intercuartílico en lugar de usar la desviación estándar calculada a partir de los datos muestrales.

No se realizan pruebas de significación estadística ni cálculo de los intervalos de confianza del recorrido o del recorrido intercuartílico. Por otro lado, si el recorrido intercuartílico se emplea para estimar la desviación estándar poblacional, podemos contrastar una hipótesis estadística o calcular un intervalo de confianza. En ese caso, el método sugerido para la medida de la dispersión podría utilizarse para las variables dependientes continuas.

VARIABLE DEPENDIENTE NOMINAL

Como indica el término, una *variable dependiente nominal* consiste solamente en el nombre de una condición determinada. Además, recuerde que hemos limitado los datos nominales a indicadores de que la condición existe o, por defecto, no existe. Ejemplos de las variables dependientes nominales incluyen vivo o muerto, curado o no curado y enfermo o sano. La cantidad de información contenida en una variable dependiente aislada es bastante limitada, en comparación con la que contienen las variables dependientes continuas, como la edad, o las ordinales, como el estadio de la enfermedad.

Cuando utilizamos variables dependientes nominales solo es necesario referirnos a medidas de posición. Esto puede sorprender, dado que, cuando considerábamos las variables dependientes continuas u ordinales, discutimos la importancia de las estimaciones de la dispersión y de la posición. En las variables dependientes continuas, la dispersión constituye una cuestión importante, porque frecuentemente se supone que siguen una distribución poblacional gausiana caracterizada, en parte, por la independencia entre la posición y la dispersión. Esto equivale a decir que, para una distribución gausiana, el conocimiento de la media no nos dice nada acerca de cuál puede ser la varianza de la distribución. Para una media determinada, son posibles infinitas varianzas. Esto *no* es verdad para las distribuciones aplicables a las variables nominales. Antes bien, esas distribuciones tienen medidas de dispersión que dependen totalmente de las medidas de posición (lo cual significa que pueden calcularse a partir de las medidas de posición o son iguales a un valor constante). Por eso, una vez que conocemos la medida de posición, sabemos o podemos calcular la medida de dispersión.

El método estadístico univariante específico que utilizamos para analizar una variable dependiente nominal (figura 27-5) varía según se trate de una proporción como la prevalencia o de una tasa como la incidencia. Veamos en primer lugar, los métodos aplicables a las proporciones.

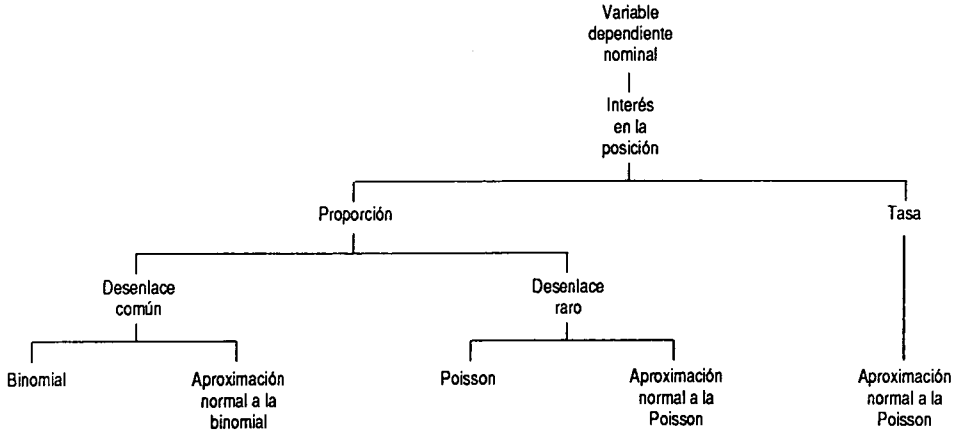
Interés en las proporciones

Para cada medición u observación de una variable compuesta por datos nominales, solo determinaremos la presencia o la ausencia de la condición en estudio. Por ejemplo, podemos determinar si un individuo de una muestra tiene o no una enfermedad concreta. En una muestra constituida por más de una observación podemos estimar la *frecuencia* o el número de veces que la condición ocurre en la población. Por ejemplo, podemos estimar el número de personas que tienen una enfermedad en la población. Más a menudo esa frecuencia nos interesa en relación con el número de observaciones en la muestra. Si dividimos el número de veces que se observa una condición en una muestra por el número de observaciones, estamos calculando la *proporción* de observaciones en la muestra que tienen esa condición. Una proporción calculada a partir de las observaciones muestrales es una estimación puntual de la proporción de la población con la condición. Una forma equivalente de interpretar la proporción de la muestra es estimar la *probabilidad* de la presencia de la condición en la población. Dos proporciones o probabilidades que se calculan habitualmente en la investigación médica son la prevalencia y el riesgo. Estas medidas se comentan en la Sección 1 y en la Sección 3.

Las probabilidades no siguen una distribución gausiana. Se supone que siguen una distribución *binomial* o una de *Poisson*. Se puede aplicar una distribución binomial a toda probabilidad calculada a partir de datos nominales que cumplan los siguientes criterios: 1) la probabilidad de que cualquier observación obtenida mediante un muestreo aleatorio pertenezca a una categoría determinada, denominada *condición nominal*, es la misma para cada observación y 2) las observaciones son independientes entre sí. *Independiente* quiere decir que el resultado de una observación no influye en el resultado de otra.

Una distribución de Poisson es un caso especial de la distribución nominal en la cual el suceso nominal observado, como la muerte o la enfermedad, es

FIGURA 27-5. Esquema para seleccionar un método estadístico univariante para una variable dependiente nominal (continuación de la figura 26-5)



muy infrecuente y el número de observaciones es elevado. El cálculo de la distribución de Poisson es más sencillo que el de la binomial. En general, constituye una buena aproximación a la distribución binomial cuando el número de individuos observado con la condición es 5 o menos y, además, el número de individuos en la muestra es 100 o más.

Las pruebas de significación estadística y el cálculo de los intervalos de confianza de las distribuciones binomial y de Poisson resultan difíciles si deseamos utilizar técnicas *exactas* que realmente usen las distribuciones de Poisson o binomial. Afortunadamente, muchas veces no nos vemos en la necesidad de usar esas técnicas.

Es mucho más sencillo calcular los intervalos de confianza o realizar las pruebas de significación estadística para variables dependientes nominales cuando, en ciertas condiciones, se puede realizar una aproximación a las distribuciones binomial o de Poisson mediante la distribución gausiana. Podemos utilizar una aproximación gausiana, casi siempre denominada *aproximación normal*, a las distribuciones binomial o de Poisson cuando el número de individuos con la condición es mayor de 5 y el número de observaciones es mayor de 10.[9]

Interés en las tasas

En la terminología estadística se reserva el término *tasa* para hacer referencia a una razón que incluya una medida del tiempo en el denominador, en contraposición con el término *proporción*, que solo incluye el número total de observaciones en el denominador. La medida de interés más habitual en la investigación médica que cumple la definición de tasa es la incidencia.

Para ilustrar esta distinción, imagine que hemos observado 100 personas que, al inicio de nuestro período de observación, no tenían cierta enfermedad. A los tres años, 30 de las 100 habían enfermado. Si estuviéramos interesados en conocer la probabilidad de que una persona seleccionada al azar de la población de la que se ha extraído la muestra desarrolle esa enfermedad en un período de tres años, calcularíamos la proporción trianual o el riesgo de padecer la enfermedad dividiendo 30 por 100 = 0,30. Sin embargo, si estuviéramos interesados en la *tasa* con la que aparecen nuevos casos de la enfermedad en la muestra de población, calcularíamos la incidencia de la enfermedad como $30/(100 \times 3) = 0{,}10$ por año. Observe que las probabilidades no tienen unidades y que las tasas se expresan en unidades de 1/tiempo o de sucesos por unidad de tiempo.

Dado que las enfermedades por lo común se producen de forma infrecuente por unidad de tiempo, en el análisis univariante muchas veces se supone que las tasas siguen una distribución de Poisson. Al igual que sucede con las proporciones, es posible aplicar técnicas exactas a las tasas, pero habitualmente las pruebas de significación estadística y la construcción del intervalo de confianza se basan en la aproximación normal. De este modo, se emplean las mismas técnicas para las tasas y las probabilidades, excepto cuando se realizan pruebas de significación estadística y estimaciones por intervalo, para las cuales se emplea la distribución de Poisson o su aproximación normal.

---

[9] En la aproximación normal a la distribución de Poisson o a la binomial, solo necesitamos estimar la probabilidad de observar un suceso, dado que el error estándar se calcula a partir de esa probabilidad. Esto difiere de la distribución gausiana para variables dependientes continuas, en la cual debemos realizar estimaciones separadas para la posición y para la dispersión. Como resultado, no es necesario o, de hecho, apropiado utilizar la distribución de la *t* de Student para tener en cuenta, mediante los grados de libertad, la precisión con que se haya estimado la dispersión. En su lugar, se emplea la distribución normal estándar.

RESUMEN

En este capítulo hemos presentado solamente las técnicas univariantes. Estos métodos se emplean cuando un grupo de observaciones contiene una variable dependiente y ninguna independiente. En su mayor parte, el análisis univariante se centra en el cálculo de los intervalos de confianza más que en las pruebas de hipótesis estadísticas. Una excepción a esta regla es la medición de los valores de una variable dependiente continua dos veces en los mismos individuos o en sujetos muy semejantes. En este caso, la variable dependiente es la diferencia entre dos mediciones. Para contrastar la hipótesis nula de que la diferencia es igual a cero, suele emplearse una prueba de significación estadística para datos apareados.

Durante la presentación del análisis univariante de las variables dependientes continuas, hemos examinado diversos principios importantes del análisis de datos continuos. Uno de ellos, el teorema central del límite, nos ayudó a entender por qué las pruebas estadísticas para las medias se basan tan frecuentemente en la distribución gausiana. Este teorema afirma que las medias tienden a seguir una distribución gausiana, aunque no la sigan en la población de la que proceden.

Otro principio importante es la distinción entre dos medidas de dispersión: la desviación estándar y el error estándar. La desviación estándar es una medida de la dispersión de los datos en la población. Utilizamos la media más y menos la desviación estándar cuando nos interesa comunicar la variabilidad estimada de las observaciones individuales. Por su lado, el error estándar es una medida de la dispersión de las medias de las muestras extraídas de una población. Utilizamos el error estándar cuando nos interesa mostrar la diferencia esperada entre las medias muestrales. Para contrastar hipótesis estadísticas y para construir los intervalos de confianza de las medias empleamos el error estándar.

Al realizar las pruebas de significación estadística y al construir intervalos de confianza para el análisis univariante de las variables dependientes continuas se supone que la población de la que se extrae la muestra sigue una distribución gausiana. Cuando dudamos que sea así, podemos transformar la variable dependiente continua a una escala ordinal. Con una variable dependiente ordinal o con una variable dependiente continua transformada en una variable ordinal podemos realizar cálculos estadísticos paralelos a los comentados cuando tratamos el tema de las variables dependientes continuas, aunque no requieren suponer que la población siga una distribución determinada de los datos. Estos métodos estadísticos se denominan no paramétricos. De forma alternativa, podemos efectuar estimaciones de los parámetros de la distribución gausiana transformando los datos continuos a una escala ordinal y empleando la mediana como una estimación robusta de la media y los dos tercios del intervalo intercuartílico como estimación robusta de la desviación estándar. Esta aproximación es útil cuando la muestra contiene valores extremos o aislados.

El análisis univariante de las variables dependientes nominales se distingue de otros porque en él no se realizan estimaciones independientes de la posición y la dispersión. Las estimaciones de la posición de las variables dependientes nominales pueden ser tasas o proporciones. Los tipos de distribuciones supuestas con más frecuencia para las variables dependientes nominales son la distribución de Poisson y la binomial. La distribución de Poisson se usa siempre que la condición estudiada sea muy poco frecuente. En el análisis se pueden utilizar estas distribuciones directamente o, para simplificar los cálculos, emplear la aproximación normal a las mismas.

CAPÍTULO 28

ANÁLISIS BIVARIANTES

En el análisis bivariante, nos interesa estudiar una variable dependiente y una independiente. Además de determinar el tipo de variable dependiente, para escoger la técnica estadística adecuada es necesario identificar el tipo de variable independiente. Los criterios para clasificar las variables independientes son los mismos que los mencionados anteriormente respecto a las variables dependientes.

En el capítulo 27 pusimos énfasis en la estimación más que en las pruebas de significación estadística. La razón consiste en que es difícil imaginar hipótesis nulas apropiadas para el análisis univariante, excepto el de datos apareados. Esta limitación no es aplicable a los análisis bivariantes o multivariantes.

En general, la hipótesis nula de no asociación entre la variable dependiente y la independiente es importante en el análisis bivariante. Sin embargo, una escuela de pensamiento otorga más importancia al cálculo de los intervalos de confianza que a las pruebas de significación estadística en todos los tipos de análisis estadísticos. El argumento que esgrimen es que los investigadores médicos deben interesarse por estimar la fuerza de las asociaciones y dejar la contrastación de hipótesis a los que deciden la política sanitaria. Sea cual fuere su opinión personal sobre la estimación frente a las pruebas de significación estadística, la literatura médica contiene una mezcla de intervalos de confianza y de pruebas de hipótesis. Por lo tanto, los investigadores médicos y los lectores de la literatura médica deben estar preparados para interpretar apropiadamente ambos enfoques.

Como hemos indicado anteriormente, las pruebas de significación estadística y la estimación están íntimamente relacionadas. Dado que, en la mayor parte de los casos, los intervalos de confianza son simplemente una reordenación algebraica de la ecuación utilizada para las pruebas de significación estadística, la información de un intervalo de confianza se puede utilizar para contrastar la hipótesis nula y, a la inversa, la información de las pruebas de significación estadística puede servir para construir un intervalo de confianza.

Cuando trabajamos con el análisis univariante, podemos basarnos en la siguiente relación entre el intervalo de confianza y la prueba de significación estadística. Una estimación univariante por intervalo de una muestra que no contiene el valor sugerido por la hipótesis nula, denominado *valor nulo*, indica que la prueba para contrastar la hipótesis nula sería estadísticamente significativa. Si la estimación por intervalo contiene el valor nulo, entonces la prueba de significación estadística no sería estadísticamente significativa.

Por ejemplo, suponga que el cambio medio de la tensión arterial diastólica antes y después de una intervención en un ensayo clínico con observaciones apareadas es de 4 ± 1 mmHg, donde 4 mmHg es la media de la diferencia y 1 mmHg es el error estándar de la media de la diferencia. A partir de esta información, podemos calcular un intervalo de confianza bilateral de 95% aproximado igual a:

$$4 \pm 2\,(1) = 2 \text{ y } 6 \text{ mmHg}$$

Una forma de interpretar este intervalo de confianza consiste en afirmar que tenemos un nivel de confianza de 95% de que la media de la diferencia en la población se encuentra en algún lugar entre 2 y 6 mm Hg. Si, en lugar de la estimación del intervalo de confianza, nos interesa contrastar la hipótesis nula de que la diferencia de la media poblacional es igual a cero, observaremos que el valor nulo, cero, se encuentra *fuera* del intervalo de confianza de 95%. El hecho de que el intervalo de confianza de 95% no contiene el valor cero nos dice que sobre la base de una prueba de significación estadística (con una proporción de error de tipo I de 100% − 95% = 5%) rechazaríamos la hipótesis nula.

Lamentablemente, esta relación no se mantiene en las pruebas de significación estadística bivariantes. Por ejemplo, suponga que extraemos muestras de 200 personas de dos comunidades y determinamos la proporción que padece una enfermedad determinada en cada muestra. En este ejemplo, la prevalencia de la enfermedad es la variable dependiente y la comunidad es la variable independiente. Ahora, suponga que encontramos 19 personas con la enfermedad en la primera muestra y 33 en la segunda. Nuestra estimación puntual de la prevalencia de la enfermedad en las dos comunidades es de 19/200 = 0,095 y de 33/200 = 0,165. Mediante la aproximación normal a la distribución binomial encontramos que la estimación por intervalo univariante y bilateral de la prevalencia de la enfermedad en la primera comunidad está comprendida entre 0,0543 y 0,1356. En la segunda comunidad, el intervalo estimado está comprendido entre 0,1136 y 0,2164. Estos resultados se muestran en el cuadro 28-1.

Aunque estos intervalos de confianza univariantes se solapan, sería incorrecto suponer que en una prueba de significación estadística *bivariante* no rechazaríamos la hipótesis nula de que la prevalencia de la enfermedad es igual en las dos comunidades. De hecho, si empleamos una prueba bivariante apropiada para analizar los datos presentados en el cuadro 28-1 con una probabilidad de 5% de cometer un error de tipo I, rechazaríamos la hipótesis nula de que las prevalencias poblacionales son idénticas ($P = 0,04$).[1]

En lugar de calcular dos intervalos de confianza univariantes de las observaciones tales como la prevalencia de la enfermedad en las dos comunidades, podemos calcular un solo intervalo de confianza bivariante para la diferencia o para la razón entre las dos prevalencias. En nuestro ejemplo anterior de dos estimaciones de la prevalencia, el intervalo de confianza bilateral de 95% para la diferencia entre las prevalencias de la comunidad 1 y la 2 está comprendido entre 0,0361 y 0,2999. Al observar

CUADRO 28-1. Estimaciones puntuales y por intervalo de una enfermedad hipotética calculadas en muestras de dos comunidades

Comunidad	Estimación puntual	Intervalo de confianza de 95%
1	0,095	0,0543 − 0,1356
2	0,165	0,1136 − 0,2164

[1] No obstante, podemos hacer algunas afirmaciones sobre la relación entre las estimaciones por intervalo univariantes y las pruebas de inferencia bivariantes. Primero, si los *intervalos de confianza univariantes no se superponen*, podemos suponer que una prueba estadística bivariante de la hipótesis nula de que los parámetros son iguales en las muestras nos conduciría a *rechazarla*. Segundo, si los *intervalos de confianza univariantes se superponen con la estimación puntual* de la otra muestra, podemos suponer que la prueba bivariante de la hipótesis nula de que los parámetros son iguales en las muestras poblacionales nos conduciría a *no rechazar* esa hipótesis nula. Lamentablemente, las situaciones en las que los intervalos de confianza se superponen entre sí pero no lo hacen con las estimaciones puntuales son frecuentes y no proporcionan información fiable sobre los resultados de las pruebas de las hipótesis bivariantes.

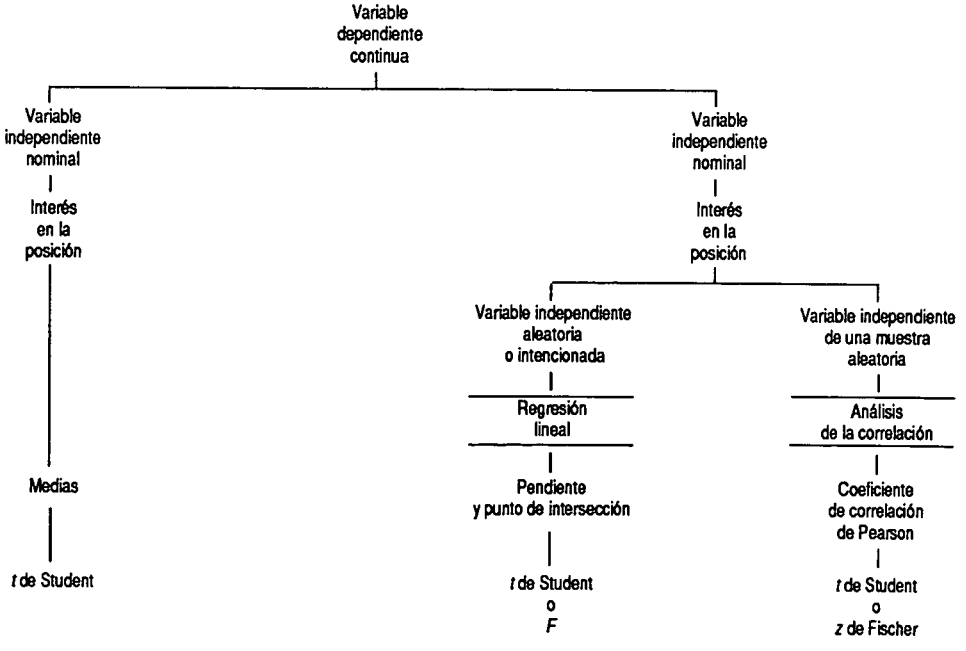
que el intervalo de confianza bivariante no se extiende más allá del cero, podríamos concluir correctamente que la prueba de significación estadística correspondiente conduciría a rechazar la hipótesis nula de que la prevalencia de la enfermedad es igual en las dos comunidades. En otras palabras, podemos rechazar la hipótesis nula de que la diferencia entre las prevalencias es igual a cero.

Aunque hemos utilizado un ejemplo con una variable dependiente nominal para ilustrar la distinción entre intervalos de confianza bivariantes y univariantes y su relación con las pruebas de significación estadística bivariantes, el mismo principio es aplicable a las variables dependientes continuas y ordinales. Por lo tanto, es necesario tener cuidado y no comparar los intervalos de confianza de las variables dependientes en cada grupo como forma de obtener una prueba de hipótesis estadística bivariante sin tener en cuenta el tipo de variable dependiente en consideración. Ahora examinemos más de cerca ciertas cuestiones de interés y los métodos que empleamos para abordarlas en el análisis bivariante.

VARIABLE DEPENDIENTE CONTINUA

Al examinar la figura 28-1 se pueden observar dos cosas. La primera es que no consideramos la asociación entre una variable dependiente continua y una variable independiente ordinal. La razón de esta omisión es que no existen técnicas estadísticas para comparar una variable dependiente continua asociada con una variable independiente ordinal sin transformar la variable continua a una escala ordinal. En segundo lugar, se puede observar que solo hemos considerado el interés en la posición. Esto no significa que no existan técnicas estadísticas para comparar las medidas de dispersión, sino que refleja un interés prácticamente exclusivo en la posición en los análi-

FIGURA 28-1. Esquema para seleccionar un método estadístico bivariante para una variable dependiente continua (continuación de la figura 26-5)



sis bivariantes y multivariantes de los datos de la investigación médica. Los métodos para comparar medidas de dispersión se utilizan para examinar supuestos con objeto de ver si una prueba estadística determinada es apropiada para aplicarla a los datos. No obstante, estas pruebas rara vez aparecen en la literatura médica.

Variable independiente nominal

Una variable independiente nominal divide las observaciones en dos grupos. Por ejemplo, suponga que medimos el tiempo de sangría de mujeres que toman píldoras anticonceptivas (PAC) en relación con el de mujeres que no las toman. La variable dependiente, tiempo de sangría, es continua y la independiente, tomar píldoras/no tomar píldoras, nominal. La variable independiente nominal divide el tiempo de sangría en un grupo de mediciones para la usuarias de PAC y otro grupo de mediciones para las no usuarias. Hemos extraído una muestra de mediciones del tiempo de sangría de una población que contiene un grupo de usuarias de PAC y uno de no usuarias de PAC.

Un supuesto universal en estadística es que nuestras observaciones son el resultado de un muestreo aleatorio. Este supuesto se aplica en el caso de la variable dependiente, pero no en las pruebas estadísticas del muestreo de variables independientes.

En general, hay dos métodos de muestreo de variables independientes que nos interesan en particular.[2] El primer método es el denominado *muestreo aleatorio* (*naturalistic sampling*). En el ejemplo del tiempo de sangría, el muestreo aleatorio significa que seleccionaríamos al azar, por ejemplo, 200 mujeres de una población y luego determinaríamos cuáles son usuarias de PAC y cuáles no lo son. Entonces, si nuestro método de muestreo no estuviese sesgado, las frecuencias relativas de usuarias de PAC comparadas con las de las no usuarias en nuestra muestra serían representativas de la frecuencia del uso de PAC en la población.

El segundo método se denomina *muestreo intencionado* (*purposive sampling*). Si empleamos un muestreo intencionado para estudiar el tiempo de sangría, podríamos seleccionar al azar a 100 mujeres que sean usuarias de PAC y 100 mujeres que no lo sean. Dado que el investigador determina el número de observaciones para cada valor de la variable independiente, la frecuencia relativa de los individuos en la muestra con la variable nominal no es representativa del tamaño relativo de los grupos en la población, aunque nuestro método sea aleatorio y no sesgado. El hecho de que nuestra muestra contenga 100 usuarias de PAC y 100 no usuarias *no* sugiere que la mitad de las mujeres de la población tomen píldoras anticonceptivas.

De este modo, la distinción entre el muestreo aleatorio y el intencionado consiste en si la variable independiente en la muestra es o no representativa de la distribución de esa variable en la población. El muestreo aleatorio es mucho más frecuente en los estudios de cohortes concurrentes. El muestreo intencionado es común en los estudios de casos y controles y en los estudios de cohortes no concurrentes. Como veremos más adelante, el método utilizado para obtener muestras de valores representativos de las variables independientes influirá en nuestra elección de las técnicas estadísticas apropiadas o en la potencia estadística de la técnica seleccionada.

[2] Existe un tercer método de muestreo de variables independientes, que es similar al muestreo intencionado, pero, en lugar de seleccionar las observaciones que tengan valores específicos de las variables independientes, el investigador asigna aleatoriamente un valor, como la dosis, a cada sujeto. Este tercer método de muestreo se emplea en estudios experimentales.

En el análisis bivariante, como en el caso de la asociación entre el tiempo de sangría y la toma de píldoras anticonceptivas, nos interesa la forma de poder comparar el tiempo de sangría entre las usuarias de PAC y las no usuarias. En la comparación de medias, nuestro interés reside en su diferencia.[3] Por ejemplo, nos interesa la diferencia entre los tiempos medios de sangría de las usuarias de PAC y de las no usuarias. El error estándar de la diferencia entre las medias se calcula a partir de las estimaciones de las varianzas de los dos grupos comparados.[4] Para calcular el error estándar de la diferencia en la media de los tiempos de sangría, combinaríamos nuestras estimaciones de la varianza del tiempo de sangría de las usuarias de PAC y la varianza de las no usuarias. Las estimaciones por intervalo y las pruebas de significación estadística aplicadas a inferencias entre medias siguen la distribución de la *t* de Student.

El uso correcto de la distribución de la *t* de Student en las pruebas de significación estadística y el cálculo de los intervalos de confianza no es influido por el método de muestreo de la variable independiente. Sin embargo, en estas técnicas se obtiene la máxima potencia estadística cuando hay un número igual de observaciones para cada una de las categorías potenciales de la variable independiente. Esto equivale a decir que tendríamos la posibilidad más alta de demostrar la significación estadística de una verdadera diferencia en el tiempo medio de sangría en 200 mujeres si utilizáramos un muestreo intencionado, seleccionando 100 usuarias de PAC y 100 no usuarias.

Variable independiente continua

Muchas veces nos interesa utilizar la medida de una variable independiente continua para estimar la medida de una variable dependiente. Por ejemplo, imaginemos que queremos analizar la relación entre la dosis de un fármaco hipotético para el tratamiento del glaucoma y la tensión intraocular. En concreto, deseamos estimar las tensiones intraoculares que esperamos que estén asociadas en la población con diversas dosis del fármaco.

Algunos tipos de cuestiones que pueden plantearse acerca de la estimación de la variable dependiente están relacionadas con la forma de extraer la muestra de valores de la variable independiente continua. Sin tener en cuenta si el muestreo fue aleatorio o intencionado, podemos establecer una ecuación lineal para estimar el valor medio de la variable dependiente (Y_i) para cada valor de la variable independiente (X_i). En nuestro ejemplo, la variable dependiente es la tensión intraocular media y la variable independiente, la dosis del medicamento. La ecuación de una relación lineal en una población se describe mediante dos parámetros: una *pendiente* (β) y un *punto de intersección* (α).

$$Y_i = \alpha + \beta X_i$$

El punto de intersección estima la media de la variable dependiente cuando la variable independiente es igual a cero. Por lo tanto, el punto de intersección de la ecuación lineal de la tensión intraocular y la dosis estimaría la media de la tensión intraocular en los individuos que no han tomado el medicamento. La pendiente

[3] La razón de este interés es que las diferencias entre las medias tienden a seguir una distribución gausiana, mientras que otras combinaciones aritméticas, como la razón de las medias, no lo hacen.

[4] Este error estándar es igual a la raíz cuadrada de la suma de las varianzas de las distribuciones de la media de cada grupo divididas por la suma de los tamaños de las muestras. Conociendo esto, podemos entender mucho mejor por qué no se pueden usar los intervalos de confianza univariantes como sustituto fiable de las pruebas de inferencia bivariantes. La comparación de los intervalos de confianza univariantes equivale a sumar los errores estándares de dos muestras. Esto no es algebraicamente equivalente al error estándar de las diferencias entre medias.

de una ecuación lineal indica cuánto cambia la magnitud de la media de la variable dependiente por cada cambio de unidad en el valor numérico de la variable independiente. Por ejemplo, la pendiente de la ecuación que describe la tensión intraocular en función de la dosis estima cuánto desciende la tensión intraocular por cada unidad que aumenta la dosis.

Si nos interesa este tipo de estimación, necesitamos calcular dos estimaciones puntuales en nuestra muesta de observaciones: la pendiente muestral y el punto de intersección muestral. Para obtener estas estimaciones, utilizamos casi siempre el método denominado *regresión por el método de los mínimos cuadráticos* (*least squares regression*). Este método selecciona los valores de la pendiente y del punto de intersección que minimizan las distancias, o más concretamente, la suma de las diferencias al cuadrado, entre los datos observados en la muestra y los estimados por la ecuación de la recta.[5]

Una forma de presentar las observaciones de los estudios, como las de la dosis del fármaco y la tensión intraocular, consiste en examinar la relación entre la tensión intraocular y la dosis en un *diagrama de puntos* (*scatterplot*) (figura 28-2). Por convención, la variable independiente se sitúa en la *abscisa* o eje horizontal y la variable dependiente, en la *ordenada* o eje vertical. En este ejemplo, nuestro interés se centra principalmente en la tensión intraocular; por lo tanto, la tensión intraocular es la variable dependiente y la dosis del fármaco, la variable independiente.

Con la regresión lineal por el método de los mínimos cuadráticos, podemos estimar el punto de intersección y la pendiente de la relación entre la dosis (X) y la tensión intraocular (Y). Además es posible representar las estimaciones de estos parámetros mediante una ecuación de regresión:

$$Y_i = 37{,}7 + 2{,}3\,X_i$$

Además, podríamos representar la recta de regresión estimada mediante una gráfica (figura 28-3).

FIGURA 28-2. Diagrama de puntos de la tensión intraocular (TIO) después del tratamiento con un medicamento determinado administrado a distintas dosis

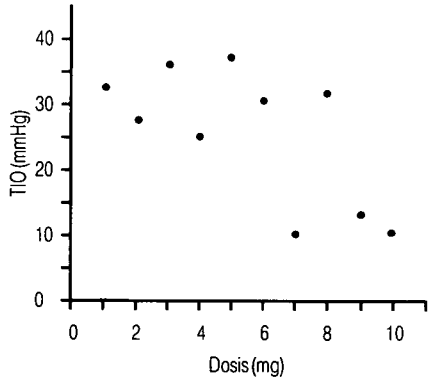
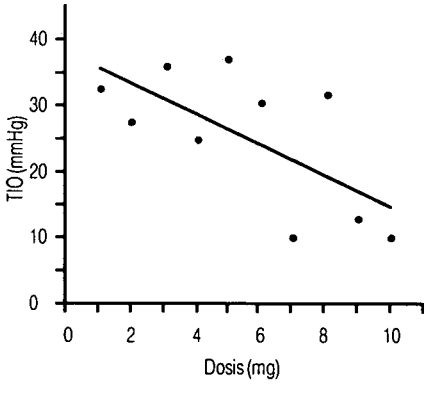


FIGURA 28-3. Regresión de la tensión intraocular (TIO) después del tratamiento con un medicamento determinado en función de la dosis



[5] Las diferencias entre los valores observados de las variables dependientes y los estimados por la ecuación de regresión se conocen como *residuales*. Los residuales indican la precisión con que la ecuación lineal estima la variable dependiente.

En el análisis de regresión se pueden aplicar numerosas pruebas de significación estadística y estimaciones por intervalo. Por ejemplo, podemos considerar la pendiente o el punto de intersección por medio de hipótesis nulas por separado o calcular intervalos de confianza para cada uno de esos parámetros. En este caso se emplea casi siempre la distribución de la *t* de Student.[6]

Podemos considerar la ecuación lineal como un todo, en lugar de considerar por separado la pendiente y el punto de intersección. Para considerar la ecuación como un todo, examinaremos el grado de variación de la variable dependiente que somos *capaces de explicar* mediante la ecuación lineal dividido por el grado de variación que somos *incapaces de explicar* con la ecuación lineal. En el ejemplo del medicamento para tratar la hipertensión intraocular, dividiríamos la variación de la tensión intraocular que es explicada por el conocimiento de la dosis, por la variación de la tensión intraocular que queda inexplicada. A continuación, podemos contrastar la hipótesis nula según la cual la ecuación de regresión no nos permite explicar el valor de la variable dependiente, la tensión intraocular, dado un valor de la variable independiente, la dosis de medicación. Para contrastar esta hipótesis nula se emplea la distribución de *F*.[7]

La estimación por intervalo de la ecuación lineal en su totalidad se lleva a cabo habitualmente mediante la construcción de los intervalos de confianza de las medias esperadas de la variable dependiente, como la tensión intraocular, para distintos valores de la variable independiente, por ejemplo, la dosis del medicamento. Muchas veces construimos estos intervalos de confianza para todos los valores de la variable independiente dentro del recorrido de los valores de la muestra. Estos intervalos de confianza se presentan como una *banda de confianza* que rodea la recta de regresión (figura 28-4).

En la extrapolación de los resultados de estudios analizados con métodos de regresión, algunas veces se especula sobre valores de la variable dependiente que corresponden a valores de la variable independiente que exceden el recorrido de los valores de la muestra. Por ejemplo, podríamos vernos tentados de predecir la tensión intraocular de los pacientes que reciben dosis del medicamento más altas o más bajas que las empleadas en nuestro estudio. No obstante, es peligroso intentar predecir la media de la variable dependiente más allá del recorrido de los valores de la muestra de la variable independiente.

Una de las razones para ser precavidos —en cuanto a predicciones que exceden del recorrido de los valores muestrales de la variable independiente— se manifiesta en las bandas de confianza. La media de la variable dependiente se estima con mayor precisión por la media de la variable independiente. Esto se muestra en la figura 28-4 para la tensión intraocular. En esa figura, podemos observar que la precisión de la predicción de la tensión intraocular desciende a medida que nos alejamos del

---

[6] Los errores estándares de la pendiente y del punto de intersección están en función de la media de los residuales al cuadrado y de la dispersión de los valores de la variable independiente. Cuanto menor sea el grado de ajuste de la ecuación lineal respecto de los valores observados de la variable dependiente, menor será la precisión con que podemos estimar esos parámetros. Por otro lado, cuanto mayor sea la dispersión de los valores muestrales de la variable independiente, mayor será la precisión de estas estimaciones. Esta última relación refleja el hecho que una recta se puede construir, como mínimo, con dos puntos. Cuanto mayor sea la separación entre esos dos puntos, con mayor precisión podremos definir la recta.

[7] En el análisis de regresión con una sola variable independiente, como la regresión bivariante, la raíz cuadrada del estadístico *F* usado para contrastar la regresión global es exactamente igual al estadístico *t* de Student que se obtiene cuando contrastamos la hipótesis nula de que la pendiente es igual a cero.

valor de la media de la dosis del medicamento. Esto se evidencia en el incremento de la banda de confianza de la figura 28-4. Si consideramos valores de la variable independiente que rebasan el intervalo de la muestra, la precisión con que se pueda predecir la media de la variable dependiente es muy baja.

El otro motivo para evitar este tipo de extrapolación es que no podemos estar seguros de que la ecuación lineal sea aplicable a valores de las variables independientes para los cuales no hayamos observado valores correspondientes de la variable dependiente. Es posible que las dosis bajas o altas del medicamento no sigan una relación lineal o, incluso, que vayan en dirección contraria y eleven la tensión intraocular a dosis más altas.

Cuando efectuamos una regresión por el método de los mínimos cuadráticos nos basamos en cuatro supuestos. El primero, común a todas las técnicas estadísticas, es que el muestreo de la variable dependiente se ha realizado al azar. En el análisis de regresión suponemos que las muestras aleatorias de los valores de la variable dependiente se han extraído en relación con cada valor muestral de la variable independiente. En otras palabras, suponemos que hemos extraído muestras al azar de la población de tensiones intraoculares que corresponderían a cada dosis del medicamento estudiado.

Para determinar las estimaciones puntuales de la pendiente y del punto de intersección no estamos obligados a suponer que las muestras aleatorias proceden de una población que sigue una determinada distribución. Sin embargo, cuando realizamos estimaciones por intervalo o aplicamos pruebas de significación estadística, suponemos que la población de la que se extrajo la muestra aleatoria de la variable dependiente sigue una distribución gausiana para cada valor de la variable independiente. En nuestro ejemplo, para calcular la banda de confianza de la figura 28-4, suponemos que, para cada dosis estudiada, la tensión intraocular sigue una distribución gausiana en la población de la que se ha extraído la muestra aleatoria.

El segundo supuesto del análisis de regresión por mínimos cuadráticos consiste en que la dispersión de la variable dependiente en la población es la misma, sea cual fuere el valor de la variable independiente. Es decir, suponemos que la dispersión de la tensión intraocular es la misma independientemente de la dosis del medicamento administrada. Esta igualdad de la dispersión se denomina *homogeneidad de las varianzas* (*homogeneity of variances*) u *homocedasticidad* (*homocedasticity*).

El tercer supuesto es el más obvio y, quizá, el más importante. Para ajustar una ecuación lineal a las observaciones, debemos suponer que la relación entre la variable dependiente y la independiente es de hecho lineal. Por ejemplo, hemos supuesto que una línea recta describe la relación entre la tensión intraocular y la dosis del medicamento en la muestra de la población. La violación de este supuesto reduce la utilidad de la regresión lineal, aunque se cumplan los otros supuestos.[8]

El cuarto supuesto es que la variable independiente se mide con una precisión perfecta. En nuestro ejemplo, suponemos que la dosis del medicamento se conoce exactamente. De hecho, este supuesto se viola con frecuencia. Como efecto de esta violación, la estimación de la pendiente a partir de las observaciones muestrales

[8] Habitualmente se utilizan técnicas gráficas para demostrar los supuestos de distribución gausiana, homocedasticidad y relación lineal. Si uno o más de estos supuestos no se cumplen, se pueden investigar posibles *transformaciones* de la variable dependiente. Esto debe realizarse con cuidado, para garantizar que la variable dependiente transformada no viole otros supuestos del análisis de regresión. Además, se pueden emplear técnicas de regresión ponderada (*weighted*).

será más próxima a cero que la verdadera pendiente poblacional.[9] La violación del supuesto de una medición precisa de la variable independiente dificulta el rechazo de la hipótesis nula de que la ecuación de regresión no explica la variable dependiente. Por lo tanto, si con un análisis de regresión no se logra demostrar una relación estadísticamente significativa entre la variable dependiente y la independiente, uno debe preguntarse si la medición de la variable independiente pudo haber sido lo suficientemente imprecisa para ocultar una verdadera relación.

En investigaciones como la mencionada, en la que se examina la tensión intraocular media y la dosis de un medicamento para tratar el glaucoma, se suelen asignar dosis que no son representativas de todas las que podrían administrarse. En otras palabras, casi nunca se emplea el muestreo aleatorio para investigar una relación dosis-respuesta. Es apropiado usar métodos de regresión lineal sin tener en cuenta si el método de muestreo para obtener los valores de la variable independiente ha sido aleatorio o intencionado. Cuando se utiliza un método de muestreo representativo, como el aleatorio, para obtener la muestra de una variable independiente, se puede emplear otra categoría de técnicas estadísticas conocida como el *análisis de la correlación*.

El análisis de la correlación puede emplearse, por ejemplo, si extrajéramos una muestra aleatoria de los individuos de una población y midiéramos su ingesta de sal y tensión arterial diastólica. En este caso, tanto la variable independiente, la ingesta de sal, como la dependiente, la tensión arterial diastólica, han sido extraídas al azar de la población. La distribución de la ingesta de sal en nuestra muestra aleatoria es representativa de la distribución poblacional de la ingesta de sal.

La distinción entre la variable dependiente y la independiente es menos importante en el análisis de la correlación que en los otros tipos de análisis. En el análisis de la correlación se obtienen los mismos resultados si estas funciones se invierten. En nuestro ejemplo no importa, desde el punto de vista estadístico, si consideramos la tensión arterial diastólica o la ingesta de sal como la variable dependiente cuando realizamos el análisis de la correlación. Sin embargo, los mismos cuatro supuestos se aplican a *ambos* tipos de análisis.

En el análisis de la correlación, medimos cómo cambian conjuntamente la variable dependiente y la independiente. En nuestro ejemplo, mediríamos cuán consistente es la asociación entre el aumento de la ingesta de sal y el aumento de la tensión arterial diastólica. El estadístico calculado que refleja el grado de cambio conjunto de las dos variables se denomina *covarianza (covariance)*. La razón entre la covarianza y el producto de las varianzas de las variables se conoce como *coeficiente de correlación (correlation coefficient)* y se representa con la letra r. El coeficiente de correlación que se emplea más frecuentemente para dos variables continuas es el *coeficiente de correlación de Pearson (Pearson's correlation coefficient)*.

El coeficiente de correlación es una estimación puntual de la *fuerza de la asociación (strength of the association)* entre dos variables continuas. Esta es una dis-

[9] La razón por la cual la medición errónea de la variable independiente siempre hará que la pendiente se aproxime a cero no es evidente inmediatamente. Para apreciar la certeza de la afirmación, imaginemos el caso extremo de que la medición de la variable independiente es tan errónea que equivale prácticamente a un número aleatorio. Por ejemplo, en el caso de la dosis del medicamento empleado para predecir la tensión intraocular, suponga que las etiquetas de los recipientes se han equivocado de forma que no supiéramos cuál es la dosis realmente administrada a un individuo. Si no conocemos la dosis, no podemos explicar la tensión intraocular a partir de la dosis. Es decir, por término medio, no se observarían consistentemente cambios de la tensión intraocular por cada unidad de aumento de la dosis. En una ecuación de regresión, esta situación se representa como una pendiente igual a cero. Errores menos graves en la asignación de dosis nos llevarían a estimar una pendiente poblacional que se situaría entre el valor real de la población y el valor extremo de cero.

FIGURA 28-4. Límites bilaterales de los intervalos de confianza de 95% para la predicción de la media de la tensión intraocular (TIO) después del tratamiento con un medicamento determinado a partir de la dosis administrada

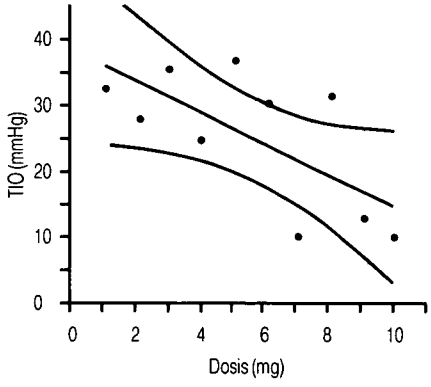
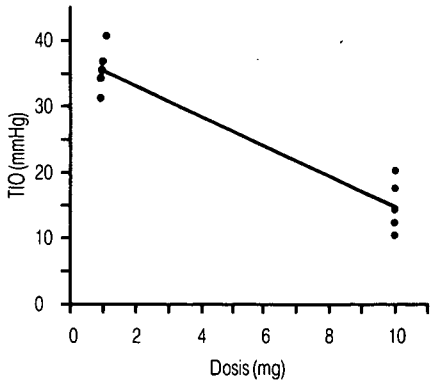


FIGURA 28-5. Regresión de la tensión intraocular (TIO) después del tratamiento con un medicamento determinado en función de la dosis cuando se administra a los pacientes una dosis de 1 mg o de 10 mg



tinción importante entre el análisis de la correlación y el de regresión. El análisis de regresión se puede usar para estimar los valores de la variable dependiente a partir de la variable independiente, pero no estima la fuerza de la asociación entre estas variables *en la población*. El análisis de la correlación estima la fuerza de la asociación entre ambas variables en la población, pero no puede utilizarse para estimar los valores reales de la variable dependiente a partir de la variable independiente.

El coeficiente de correlación tiene un recorrido de valores posibles entre -1 y $+1$. Un coeficiente de correlación igual a cero indica que no existe relación (*lineal*) entre la variable dependiente y la independiente. Un coeficiente de correlación *positivo* indica que el valor de la variable independiente *aumenta* cuando el valor de la variable dependiente *aumenta*. Un coeficiente de correlación *negativo* indica que el valor de la variable independiente *aumenta* cuando el valor de la variable dependiente *desciende*.

La interpretación de la fuerza de la asociación entre la variable dependiente y la independiente es más fácil de entender si elevamos al cuadrado el coeficiente de correlación para obtener el *coeficiente de determinación (coefficient of determination)* (R^2). Si multiplicamos el coeficiente de determinación por 100% obtenemos el porcentaje de la variación de la variable dependiente que es explicado por el valor de la variable independiente. El coeficiente de determinación de las variables continuas se puede considerar como una medida paralela al porcentaje del riesgo atribuible, dado que se refiere a la variabilidad de la variable dependiente que puede atribuirse a la variable independiente. No obstante, recuerde que es apropiado usar el coeficiente de determinación solamente cuando la muestra de la variable independiente, así como de la variable dependiente, se extrae empleando métodos representativos o aleatorios.

Uno de los errores más habituales en la interpretación del análisis estadístico es usar el coeficiente de determinación o el de correlación para realizar estimaciones puntuales sobre una población concreta aunque la muestra de la variable independiente no haya sido extraída mediante un método que garantiza la representatividad de su distribución en esa población. Podemos crear un coeficiente de correlación elevado de forma artificial obteniendo una muestra solamente de los valores extremos de la variable independiente.

Como ejemplo del problema que puede ocurrir cuando se interpretan los coeficientes de correlación, reconsideraremos el ejemplo anterior en el que calculamos una ecuación de regresión para estimar la tensión intraocular a partir de la dosis de un fármaco hipotético para el tratamiento del glaucoma. El extraer una muestra de la variable independiente, la dosis, de forma que tuviésemos una representación uniforme de las dosis dentro del intervalo comprendido entre 1 y 10 mg, como se mostró anteriormente en la figura 28-2, nos conduciría a creer que existe solo una moderada correlación negativa entre la dosis y la tensión intraocular ($r = -0,66$). Por otro lado, podemos tomar la decisión de limitar nuestro estudio a dos dosis del medicamento y asignar aleatoriamente cinco pacientes a 1 mg y cinco pacientes a 10 mg como se muestra en la figura 28-5. En este caso, estimaríamos un coeficiente de correlación negativo mucho más elevado en la población ($r = -0,95$). Sin embargo, las estimaciones de la *ecuación de regresión* en ambos métodos de muestreo son exactamente las mismas.

Para decidir cuál es el método representativo y, de ese modo, legitimar el uso del análisis de la correlación, necesitamos anticipar las dosis que se utilizarán en la práctica clínica. Por ejemplo, ¿recibirán los pacientes todas las dosis entre 1 y 10 mg con frecuencias aproximadamente iguales? Si esto es así, el coeficiente de correlación de $-0,66$ refleja correctamente la asociación entre la tensión intraocular y las dosis que podemos prever que se experimentarán en la práctica. Por otro lado, si los pacientes reciben dosis de 1 mg o de 10 mg con la misma frecuencia, el coeficiente de correlación de $-0,95$ estima la relación dosis-respuesta que puede anticiparse. Si se emplea cualquier otro patrón de administración del fármaco, ninguno de los coeficientes de correlación estima correctamente la relación previsible entre la dosis y la tensión intraocular. Para muchos tipos de datos, es difícil escoger la distribución apropiada de la variable independiente, especialmente en las relaciones dosis-respuesta. Cuando resulta difícil hacerlo, podemos emplear el análisis de regresión, pero debemos evitar el de la correlación.

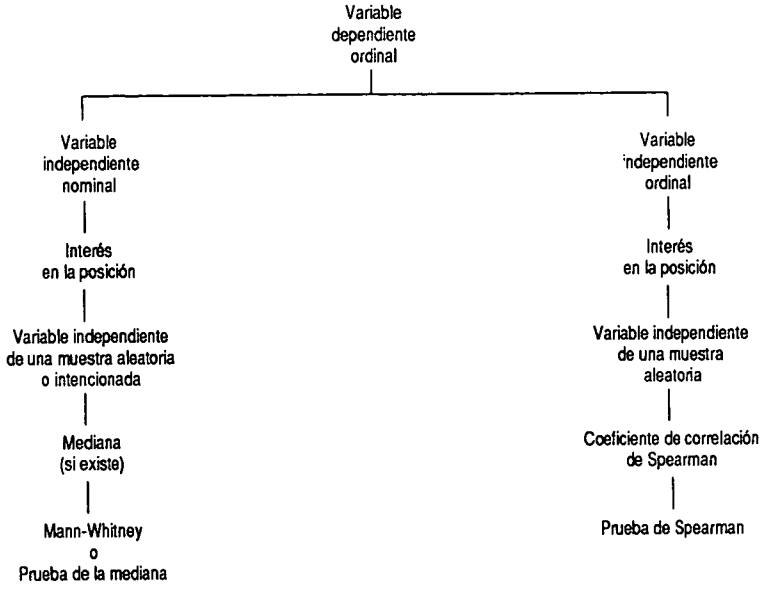
VARIABLE DEPENDIENTE ORDINAL

Al examinar la figura 28-6 observará que no se considera la posibilidad de una variable dependiente ordinal asociada con una variable independiente continua, porque esta última se debe transformar a una escala ordinal. La situación es similar a la que discutíamos en el caso de la variable dependiente continua incluida en un análisis con una variable independiente ordinal. No existen técnicas estadísticas que se utilicen habitualmente para comparar una variable dependiente ordinal con una variable independiente continua sin realizar esa transformación.

Variable independiente nominal

La *prueba de Mann-Whitney* es una prueba de significación estadística aplicable a una variable independiente nominal y a una variable dependiente ordinal. También es aplicable a una variable dependiente continua transformada a una escala ordinal, con objeto de eludir el supuesto de la prueba de la t de Student. La hipótesis nula considerada en la prueba de Mann-Whitney es que las dos muestras de la población no difieren en la posición. Dado que es una prueba no paramétrica, en la hipótesis nula no se especifica ningún parámetro de posición. Muchas veces, oímos hablar de la hipótesis nula de la prueba de Mann-Whitney en términos de la igualdad de las medianas. Esto se aleja de la verdad, pero las medianas de los dos grupos de muestras se pue-

FIGURA 28-6. Esquema para seleccionar un método estadístico bivariante para una variable dependiente ordinal (continuación de la figura 26-5)



den comparar más directamente aplicando una *prueba de las medianas*.[10] La prueba de las medianas generalmente tiene menos potencia estadística que la de Mann-Whitney.

Variable independiente ordinal

Si la variable dependiente es ordinal o continua y transformada a una escala ordinal, podemos estimar la fuerza de la asociación entre la variable dependiente y la independiente mediante un método paralelo al análisis de la correlación. En el caso de las variables ordinales, el coeficiente de correlación más utilizado es el *coeficiente de correlación de Spearman (Spearman's correlation coefficient)*. Este coeficiente se puede calcular sin realizar muchos de los supuestos necesarios para calcular el coeficiente descrito para las variables continuas. Es importante recordar que *todo* coeficiente de correlación puede calcularse a partir de muestras en las cuales *tanto* la variable dependiente *como* la independiente son representativas de la población. En otras palabras, tenemos que emplear el muestreo aleatorio. No existe ningún método no paramétrico que nos exima de este supuesto.

Al igual que ocurre con el coeficiente de correlación calculado para las variables continuas, podemos realizar pruebas de significación estadística y construir intervalos de confianza del coeficiente de correlación de Spearman. También podemos elevar al cuadrado este coeficiente para obtener una estimación no paramétrica del coeficiente de determinación o porcentaje de la variación de la variable dependiente que es explicado por la variable independiente.

[10] Aunque la prueba de las medianas se refiere a medidas de posición específicas, es una prueba no paramétrica, porque en ella no se supone que las medianas de los dos grupos sean parámetros de una distribución poblacional determinada.

VARIABLE DEPENDIENTE NOMINAL

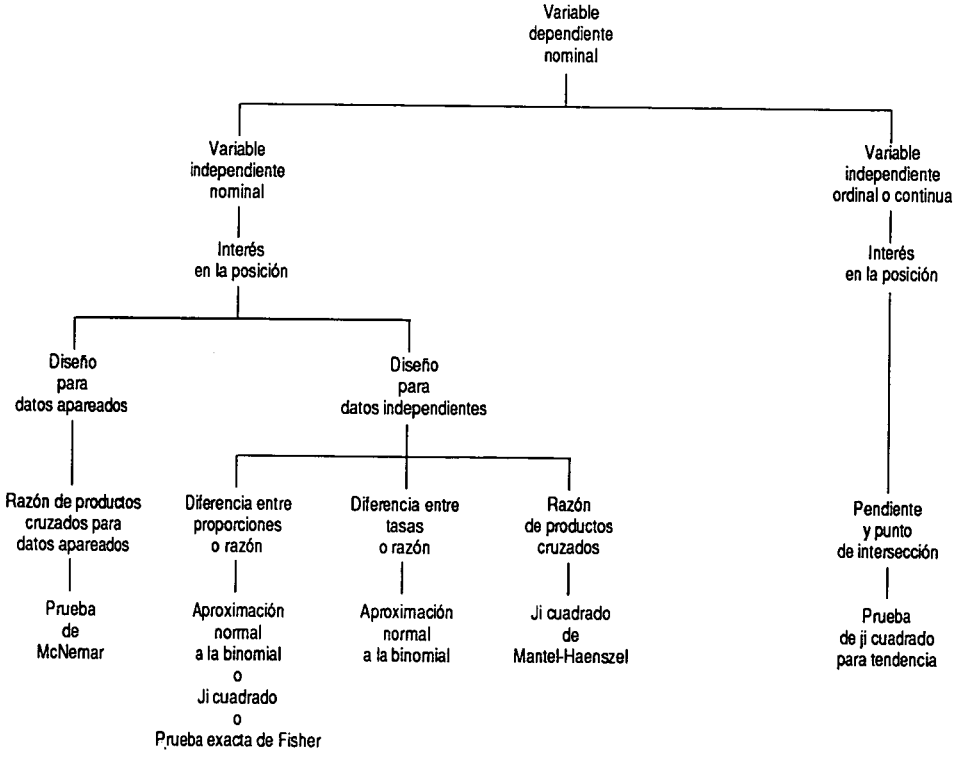
Los métodos estadísticos bivariantes para las variables dependientes nominales se presentan en la figura 28-7.

Variable independiente nominal: diseños apareados

Si nos interesa obtener información sobre una variable dependiente nominal y una independiente nominal, tenemos la posibilidad de escoger entre un diseño para datos apareados y uno para datos no apareados o independientes. Construido de forma apropiada, la potencia estadística de un diseño para datos apareados es más alta que la de un diseño para datos independientes. Recuerde que el apareamiento por parejas es un tipo especial de apareamiento en el cual la variable dependiente y la independiente se miden en cada individuo a partir de un par de individuos similares, y las observaciones de cada par se analizan conjuntamente. Cuando analizamos una variable dependiente nominal mediante un diseño apareado, utilizamos una técnica bivariante en vez de una técnica univariante como hicimos con la variable dependiente continua en un diseño para datos apareados.

En nuestro ejemplo anterior sobre la tensión arterial medida antes y después del tratamiento con un fármaco antihipertensivo, utilizamos un método univariante para examinar la diferencia entre las mediciones de la tensión arterial. Con una variable dependiente continua que se mide por datos apareados es apropiado utilizar

FIGURA 28-7. Esquema para seleccionar un método estadístico bivariante para una variable dependiente nominal (continuación de la figura 26-5)



una técnica univariante, dado que podemos resumir las observaciones de cada par empleando la diferencia entre esas medidas como variable dependiente. Con una variable dependiente nominal medida en grupos apareados, todavía estamos interesados en comparar las mediciones entre pares, pero no podemos resumir los datos nominales de tal forma que nos sea posible utilizar el análisis univariante.

Las variables dependientes nominales permiten obtener cuatro resultados posibles entre los pares. En dos de estos resultados, ambos miembros del par tienen los mismos valores de la variable dependiente nominal. Por ejemplo, si en un ensayo clínico en el cual los individuos se aparean según el sexo y la edad antes de un tratamiento asignado al azar y la variable dependiente fuese la supervivencia, ambos miembros del par podrían sobrevivir o morir. Los pares de este tipo se denominan *pares concordantes (concordant pairs)*.[11] Los dos resultados restantes de las variables dependientes e independientes nominales son aquellos en los cuales los miembros de los pares tienen resultados opuestos. En nuestro ejemplo, estos resultados se producirían cuando un miembro del par muere y el otro sobrevive. Estos se conocen como *pares discordantes (discordant pairs)*.

Consideremos con más detalle el ejemplo de un ensayo clínico que compara la mortalidad entre las personas que fueron tratadas con un determinado fármaco frente a las que fueron tratadas con placebo. Supongamos que nos interesa la influencia de la edad y el sexo en la supervivencia, así que identificamos 50 pares de pacientes de la misma edad y sexo, y asignamos al azar a un miembro del par al grupo que recibe el medicamento y al otro al grupo que recibe placebo. Además, imaginemos que los resultados obtenidos de este ensayo son como los representados en la figura 28-8. En ese caso, habríamos observado 9 + 11 = 20 pares concordantes y 6 + 24 = 30 discordantes.

En este ejemplo, si el tratamiento fuera eficaz, esperaríamos observar diversos pares en los que el miembro tratado con el medicamento sobrevive y el tratado con placebo muere. Asimismo, esperaríamos observar menos pares en los que el miembro tratado muere y el tratado con placebo sobrevive. En otras palabras, esperaríamos observar una diferencia entre las frecuencias de los dos tipos de pares discor-

FIGURA 28-8. Tabla 2 × 2 para datos apareados correspondiente a un ensayo clínico en el cual la mortalidad es la variable dependiente. Los pacientes fueron asignados al azar por parejas de la misma edad y sexo. Las columnas indican el desenlace en el miembro de la pareja no tratado que recibió placebo, y la filas, en el miembro tratado que recibió un medicamento determinado

		PACIENTES NO TRATADOS		
		Vivos	Muertos	
Paciente tratado	Vivo	9	24	33
	Muerto	6	11	17
		15	35	50

[11] Los pares concordantes son análogos a una diferencia entre pares igual a cero para una variable dependiente continua en una prueba apareada de la *t* de Student. Del mismo modo que el cero no influye en la magnitud de la media de las diferencias para una variable dependiente continua, los pares concordantes no contribuyen a la evaluación de la interpretación de una variable dependiente nominal apareada.

dantes, si fueran distintas las probabilidades de supervivencia de los pacientes tratados y los no tratados. Además, cuanto mayor fuera la diferencia entre esas frecuencias, más alta sería la eficacia estimada del tratamiento.

En lugar de examinar la *diferencia* entre las frecuencias de los pares discordantes, lo que habitualmente nos interesa es la *razón* de estas frecuencias. Dicha razón es una estimación de la razón de productos cruzados poblacional (*odds ratio*). En este ejemplo, la razón de productos cruzados para los datos apareados es igual al número de pares en los cuales el miembro tratado sobrevive y el miembro no tratado muere, dividido por el número de pares en los cuales el miembro tratado muere y el no tratado sobrevive, o sea, 24/6 = 4.

Es importante recordar que la razón de productos cruzados para los datos apareados tiene que calcularse a partir de los datos de los pares discordantes. Si hacemos caso omiso del hecho de que los datos son apareados y procedemos como si los datos correspondieran a individuos no apareados, nuestra estimación de la razón de productos cruzados poblacional sería inexacta. Para ilustrar este punto, en la figura 28-9 se presentan los datos de la figura 28-8 como si estos se hubieran analizado sobre la base de 100 individuos separados en lugar de 50 pares. La razón de productos cruzados calculada a partir de los datos presentados de esta forma estaría sobrestimada:

$$\text{Razón de productos cruzados} = \frac{33 \times 35}{15 \times 17} = 4{,}53$$

Para realizar pruebas de significación estadística de pares discordantes se emplea la *prueba de McNemar*. Se pueden aplicar métodos relacionados para calcular los intervalos de confianza de la razón de productos cruzados de las observaciones apareadas.

Variable independiente nominal: datos independientes

En el análisis bivariante de una variable dependiente nominal no apareada, al igual que en el análisis univariante de las variables dependientes nominales, podemos escoger entre medir una proporción como la prevalencia, el riesgo o la ventaja, o medir una tasa como la incidencia. También tenemos la opción de seleccionar el método para comparar dos proporciones o dos tasas. En concreto, podemos decidir comparar estimaciones de grupos utilizando una diferencia o una razón entre las estimaciones.

FIGURA 28-9. Una tabla 2 × 2 para datos independientes correspondiente a los datos apareados de la FIGURA 28-8. Observe cómo difiere esta tabla de la tabla para datos apareados. En esta figura, las columnas indican los resultados en los individuos, y las filas, los grupos de tratamiento a los que fueron asignados los individuos

		SUPERVIVENCIA		
		Vivo	Muerto	
Grupo de tratamiento	Tratados	3	17	50
	No tratados	15	35	50
		48	52	50

Por ejemplo, considere un estudio en el que estimamos la prevalencia de cataratas en las personas expuestas a radiaciones ionizantes cincuenta años después de la exposición. Suponga que la prevalencia de cataratas en 50 personas no expuestas menores de 40 años de edad en el momento de la exposición fue de 2%. En 100 personas de la misma edad expuestas a cierto nivel de radiación ionizante la prevalencia de cataratas fue de 12%, aproximadamente. Como estimación puntual que resume estos datos podemos usar la razón de prevalencias, esto es, la prevalencia de cataratas en los expuestos dividida por la prevalencia en los no expuestos, que es igual a 12%/2% = 6. Por otra parte, también podemos calcular la diferencia de prevalencias o la prevalencia entre los expuestos *menos* la prevalencia en los no expuestos, que es igual a 12% − 2% = 10%.

Desde un punto de vista estadístico, la elección de una razón o de una diferencia entre proporciones o tasas generalmente no tiene importancia. De hecho, en el análisis bivariante se emplean los mismos métodos para construir los intervalos de confianza y las mismas pruebas de significación estadística sin tener en cuenta si la estimación puntual es una razón o una diferencia. Esto se desprende del hecho de que la hipótesis nula de una diferencia igual a cero equivale a la hipótesis nula de que una razón es igual a 1. Cuando una razón es igual a 1, el numerador tiene que ser igual al denominador y, por lo tanto, la diferencia entre el numerador y el denominador tiene que ser igual a cero. Sin embargo, en el análisis multivariante, la distinción entre las diferencias y las razones puede ser muy importante, y se tratará en el capítulo 29.

Es muy probable que en un análisis bivariante de las variables nominales independientes y dependientes de un diseño para datos no apareados nos enfrentemos con varios métodos estadísticos. Como en el análisis univariante de una variable dependiente nominal, estos métodos son de dos tipos: métodos exactos y aproximaciones a la distribución normal. El método exacto para las proporciones bivariantes es la prueba *exacta de Fisher (exact Fisher's test)*.[12] Dos métodos de aproximación habitualmente empleados para las proporciones son la aproximación normal y las pruebas de ji cuadrado.[13] Las tasas casi siempre se analizan utilizando la aproximación normal. Las pruebas de significación estadística y el cálculo de los intervalos de confianza para la razón de productos cruzados se basan habitualmente en la prueba *ji de Mantel-Haenszel*, también una aproximación normal.[14]

Variable independiente continua

Cuando tenemos una variable independiente continua u ordinal y una variable dependiente nominal, podemos considerar la posibilidad de que varios valores de la variable independiente sigan una *tendencia (trend)*. Por ejemplo, quizá nos interese examinar la hipótesis de estudio según la cual la proporción de individuos que desarrollan un accidente vascular cerebral aumenta de forma lineal a medida que se eleva la tensión arterial diastólica, *frente* a la hipótesis nula de que no existe una relación lineal entre esas variables. Este es el mismo tipo de hipótesis que se considera en la regresión

[12] La prueba exacta de Fisher se emplea cuando alguna de las frecuencias previstas según la hipótesis nula en una tabla 2 × 2 es menor que 5.

[13] En realidad, en el análisis bivariante la aproximación normal y la prueba de ji cuadrado son equivalentes. La raíz cuadrada del estadístico ji cuadrado es igual al estadístico de la aproximación normal.

[14] Frecuentemente, una prueba de significación estadística bivariante para variables normales exigirá realizar una "corrección de continuidad" (*correction for continuity*). Esta corrección es un ajuste de las observaciones nominales cuando se transforman en distribuciones *continuas*, como la distribución gausiana, para fines de análisis. El ejemplo más familiar de corrección de continuidad es la corrección de Yates empleada en la prueba de ji cuadrado. Actualmente, los estadísticos no están de acuerdo sobre la utilidad de esta corrección. Por suerte, el uso o no de una corrección de continuidad raramente tiene un impacto importante sobre los resultados del análisis.

lineal simple con la excepción de que en este caso tenemos una variable dependiente nominal en lugar de una variable dependiente continua. En lugar de una regresión lineal simple, realizaremos una *prueba de ji cuadrado para tendencias (chi-square test for trend)*.

Si bien se da un nombre especial a la prueba empleada para investigar la posibilidad de que una variable dependiente nominal siga una tendencia lineal, debemos darnos cuenta de que la prueba de ji cuadrado para tendencias es muy similar a una regresión lineal. Por cierto, las estimaciones puntuales de los métodos que se emplean con más frecuencia para investigar una tendencia son la pendiente y el punto de intersección de una ecuación lineal, que son idénticos a las estimaciones que hemos comentado para la regresión lineal.[15]

Imagine que deseamos investigar la tasa de mortalidad entre las personas con cáncer en los estadios 1, 2, 3 y 4. Como hipótesis razonable de estudio, se podría plantear que la tasa de mortalidad aumenta a medida que avanzan los estadios de la enfermedad. Por lo tanto, deseamos investigar la posibilidad de que la variable dependiente nominal, la tasa de mortalidad, siga una tendencia correspondiente al estadio de la enfermedad. En estas circunstancias, en que tenemos una variable dependiente nominal y una independiente ordinal, es especialmente importante recordar que la prueba de ji cuadrado para tendencias es muy parecida al análisis de regresión lineal. Cuando examinamos la tendencia de una variable independiente ordinal, deben asignarse valores numéricos a las categorías ordinales.[16] La manera como se definan estos valores numéricos determinará el resultado de la prueba de ji cuadrado para tendencias. Es una convención asignar números enteros consecutivos a estas categorías ordinales, a no ser que las categorías sugieran una escala ordinal alternativa. De este modo, la variable ordinal se trata como si realmente tuviera categorías uniformemente espaciadas, como sucedería con los datos continuos. Por fortuna, esta es una prueba muy robusta y, en consecuencia, es improbable que la violación de este supuesto tenga un gran impacto.[17]

RESUMEN

Los métodos bivariantes se utilizan para analizar un conjunto de observaciones que contienen una variable dependiente y una independiente. Las variables independientes pueden ser continuas, ordinales o nominales. Las variables independientes nominales dividen el conjunto de observaciones en dos grupos. Esto permite comparar las estimaciones de la variable dependiente de los dos grupos. En este capítulo hemos aprendido que la comparación de las estimaciones de los grupos en el análisis bivariante no es lo mismo que comparar los intervalos de confianza univariantes de estas variables.

Un supuesto universal de las técnicas estadísticas es que los valores representativos de la variable dependiente se han obtenido mediante un muestreo aleatorio. Por lo tanto, debemos suponer que la distribución de la variable dependiente

---

[15] La estimación puntual de los coeficientes en una prueba de ji cuadrado para tendencias es idéntica a la estimación en la regresión lineal simple. Para la inferencia y la estimación por intervalo, se realiza un supuesto algo distinto que produce intervalos de confianza ligeramente más amplios y valores *P* un poco más altos en la prueba de ji cuadrado que en la regresión lineal. Esta diferencia se reduce a medida que aumenta el tamaño de la muestra.

[16] También se deben asignar valores numéricos a la variable dependiente nominal, pero su elección no influye en el resultado de la inferencia o de la estimación por intervalo debido a la naturaleza dicotómica de la variable.

[17] Si bien se han descrito otros métodos para examinar la tendencia de una variable dependiente nominal respecto de los valores de una variable independiente ordinal que no exigen asignar valores numéricos específicos a las categorías ordinales, no parecen tener el amplio uso del que hemos explicado aquí. Quizá, una de las razones del uso infrecuente de esos métodos alternativos sea que no estiman una ecuación que pueda emplearse para examinar la relación entre la variable dependiente y la independiente.

en la muestra es representativa de su distribución en la población de la que se extrajo la muestra. También es posible obtener la muestra de valores de la variable independiente de forma que sea representativa de la población. El muestreo representativo de la variable independiente se denomina muestreo aleatorio. Por otro lado, podemos escoger la distribución de los valores de la variable independiente en nuestra muestra de tal forma que maximice la potencia estadística o garantice la inclusión de categorías de la variable independiente que raramente ocurren en la población. Este tipo de muestreo se denomina muestreo intencionado y con él se obtienen muestras con valores de la variable independiente que no son representativos de la población de la cual se han extraído.

La distinción entre muestreo aleatorio y muestreo intencionado es especialmente importante en el análisis bivariante de una variable continua dependiente o independiente. En nuestro caso, lo que más interesa es estimar los valores de la variable dependiente para varios valores de la variable independiente. La estimación real de los valores de la variable dependiente se consigue mediante el análisis de regresión. La fuerza de la asociación entre una variable dependiente continua y una independiente continua se estima por medio del análisis de la correlación. El análisis de regresión es apropiado sea cual fuere el tipo de muestreo de los valores de la variable independiente. No obstante, el análisis de la correlación es útil solamente cuando la muestra de la variable independiente se ha obtenido mediante muestreo aleatorio.

Como ocurre en el análisis univariante, las variables continuas en los grupos de datos bivariantes se pueden transformar a una escala ordinal, si sospechamos que la población de la que se han extraído no cumple los requisitos de los análisis de las variables continuas. Los métodos para analizar las variables dependientes ordinales son, en su mayor parte, paralelos a los análisis aplicables a las variables dependientes continuas. Una excepción a esta regla es que no existe un método de uso general para realizar un análisis de regresión con variables dependientes ordinales.

Algunos de los principios generales del análisis bivariante de las variables dependientes nominales son similares a los de las variables dependientes continuas y ordinales. En las tres, las variables independientes nominales dividen a un conjunto de observaciones en grupos para ser comparados. Además, nos interesa estimar la variable dependiente para varios valores de la variable independiente sin tener en cuenta el tipo de variable dependiente. Con las variables dependientes nominales, esto se conoce como análisis de tendencia en lugar de análisis de regresión. Sin embargo, la diferencia de terminología no implica que los métodos sean muy distintos. De hecho, el análisis de regresión realizado con una variable dependiente continua es bastante similar al método más frecuentemente usado para examinar una tendencia con una variable dependiente nominal.

Otros principios generales del análisis bivariante difieren en los tres tipos de variables dependientes. Uno de ellos es el análisis de los datos de un diseño para datos apareados. Con una variable dependiente continua, los datos se analizan usando métodos univariantes. Sin embargo, los datos nominales apareados se deben analizar con métodos bivariantes. Otra diferencia es la forma en que se comparan las estimaciones puntuales cuando la variable independiente es nominal. Para una variable dependiente continua, las medias de los grupos definidos mediante la variable independiente se comparan calculando la diferencia entre esas medias. No obstante, con las variables dependientes nominales es posible comparar proporciones o tasas como diferencias o como razones, en el análisis bivariante. Las pruebas de significación estadística y la construcción de los intervalos de confianza se llevan a cabo utilizando los mismos métodos, tanto si se usan las razones como las diferencias. No obstante, las ventajas (*odds*) siempre se comparan mediante una razón.

CAPÍTULO 29

ANÁLISIS MULTIVARIANTE

En el análisis multivariante tenemos una variable dependiente y dos o más independientes. Estas variables independientes se pueden medir en la misma o en diferentes escalas. Por ejemplo, todas las variables pueden ser continuas o, por otro lado, algunas pueden ser continuas y otras nominales. En los esquemas que figuran en este capítulo solo hemos incluido las variables independientes nominales y las continuas. Aunque en el análisis multivariante se pueden incluir variables independientes ordinales, estas deben transformarse antes a una escala nominal.[1]

El uso de los métodos multivariantes para analizar los datos de la investigación médica presenta tres ventajas generales. En primer lugar, permite investigar la relación entre una variable dependiente y una independiente mientras se "controla" o se "ajusta" según el efecto de otras variables independientes. Este es el método utilizado para eliminar la influencia de las variables de confusión en el análisis de los datos de la investigación médica. Por ese motivo, los métodos multivariantes se utilizan para cumplir con la tercera finalidad de la estadística en el análisis de los resultados de la investigación médica: ajustar según la influencia de las variables de confusión.

Por ejemplo, si nos interesa estudiar la tensión arterial diastólica de las personas que reciben diversas dosis de un fármaco antihipertensivo, podríamos desear controlar el efecto potencial de confusión de la edad y del sexo. Para hacer esto en la fase de análisis de un proyecto de investigación, utilizaríamos un análisis multivariante con la tensión arterial diastólica como variable dependiente y la dosis, la edad y el sexo como variables independientes.

La segunda ventaja que ofrecen los métodos multivariantes es que permiten realizar pruebas de significación estadística de diversas variables manteniendo al mismo tiempo la probabilidad (alfa) escogida de cometer un error de tipo I.[2] En otras palabras, a veces empleamos los métodos multivariantes para evitar el problema de las comparaciones múltiples presentado en la Sección 1.

Como recordatorio del problema de las comparaciones múltiples, imaginemos que tenemos diversas variables independientes que comparamos con una variable dependiente mediante un método bivariante como la prueba de la *t* de Student. Aunque en cada una de estas pruebas bivariantes aceptemos solo un riesgo de 5% de cometer un error de tipo I, la probabilidad de cometer al menos un error de tipo I entre todas estas comparaciones será algo mayor que 5%. La probabilidad de cometer un error de tipo I en alguna comparación determinada se denomina tasa de error *de la prueba* (*testwise*). La probabilidad de cometer un error de tipo I por lo menos en una comparación se denomina tasa de error *del experimento* (*experimentwise*). Los análisis bivariantes

[1] La conversión de una escala ordinal a una nominal produce una pérdida de información que no es necesario justificar. No obstante, la transformación de los datos a una escala continua sugiere que los datos contienen más información de la que realmente poseen, lo cual es a menudo difícil de justificar.

[2] Dado que la probabilidad de cometer un error de tipo I habitualmente se sitúa en el 5%, este será el valor que utilizaremos en el resto de este capítulo.

controlan la tasa de error de la prueba. Por otra parte, muchos métodos multivariantes están diseñados para mantener una tasa consistente de error de tipo I del experimento.

La mayor parte de los métodos multivariantes se aplican para analizar dos tipos de hipótesis nula. La primera se conoce como hipótesis nula *general* (*omnibus*). Esta hipótesis nula plantea la relación entre la variable dependiente y el conjunto de variables independientes considerado como una unidad. La hipótesis nula general es una de las estrategias de los métodos multivariantes para mantener la tasa de error de tipo I del experimento en $\alpha = 0{,}05$. No obstante, un inconveniente de la hipótesis nula general es que no permite investigar las relaciones entre cada una de las variables independientes y la dependiente de forma individualizada. Esto se realiza mediante el segundo tipo de hipótesis nula planteada en las pruebas *parciales* (*partial*) o *por pares* (*pairwise*). Estas pruebas no siempre mantienen una tasa de error de tipo I del experimento igual a $\alpha = 0{,}05$.

La tercera ventaja que ofrece el análisis multivariante es que se puede utilizar para comparar por separado la capacidad de dos o más variables independientes para estimar los valores de la variable dependiente. Por ejemplo, supongamos que hemos llevado a cabo un gran estudio de cohorte para examinar los factores de riesgo de la enfermedad coronaria. Entre las variables independientes medidas se encuentran la tensión arterial diastólica y la concentración de colesterol sérico. Deseamos determinar si ambas variables aumentan el riesgo de padecer una enfermedad coronaria. Sin embargo, el examen de su capacidad para explicar quién desarrollará la enfermedad coronaria mediante un análisis bivariante puede ser engañoso si los individuos con tensión arterial diastólica elevada tienden a ser los mismos que tienen una concentración de colesterol sérico elevada. Por otro lado, si empleamos métodos multivariantes para comparar estos factores de riesgo, podremos separar su capacidad como estimadores del riesgo de enfermedad coronaria de su *aparente* asociación con la enfermedad debida a la asociación entre ellos mismos.

Dadas las ventajas expuestas, los métodos multivariantes se emplean con frecuencia para analizar los datos de las investigaciones médicas. Examinemos ahora más detenidamente esos métodos así como las formas de interpretarlos para aprovechar sus ventajas.

VARIABLE DEPENDIENTE CONTINUA

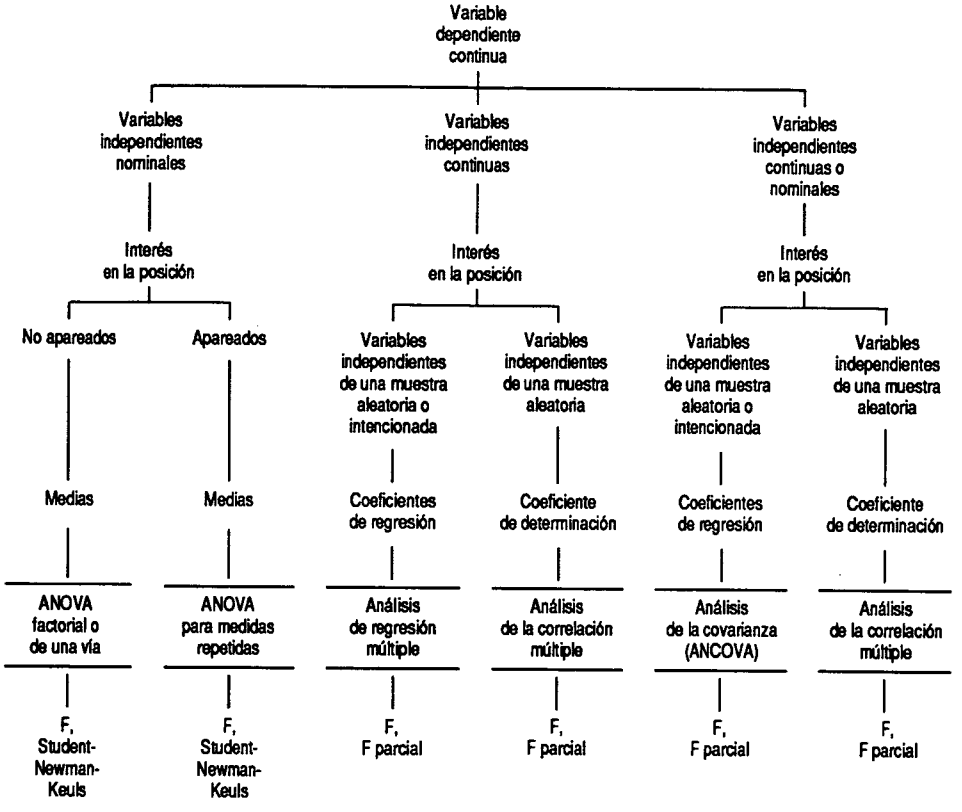
Variables independientes nominales

En el análisis bivariante de una variable dependiente continua y de una variable independiente nominal, esta última tiene el efecto de dividir la variable dependiente en dos subgrupos. En el análisis multivariante, tenemos más de una variable independiente nominal y por eso es posible definir más de dos subgrupos. Los métodos usados con más frecuencia para comparar las medias de la variable dependiente entre tres o más subgrupos son tipos de un análisis estadístico general denominado *análisis de la varianza* (*analysis of variance*) o, a menudo, *ANOVA*[3] (figura 29-1).

El tipo de ANOVA más simple es aquel en el cual k variables independientes nominales separan la variable dependiente en k + 1 subgrupos o cate-

[3] Parece incongruente que un método para comparar medias se denomine análisis de la varianza. La razón de este nombre es que el ANOVA examina la variación entre subgrupos, suponiendo una variación igual dentro de cada subgrupo. Si la varianza entre los subgrupos excede la variación dentro de estos, los subgrupos deben diferir en la posición medida por las medias.

FIGURA 29-1. Esquema para seleccionar un método estadístico multivariante para una variable dependiente continua (continuación de la figura 26-5)



gorías. Por ejemplo, supongamos que nos interesa estudiar la relación entre la glucemia basal y la raza. Además, supongamos que definimos dos variables nominales (k = 2) para indicar la raza: blanca y negra. Estas dos variables nos permiten considerar tres (k + 1 = 3) subgrupos raciales en los cuales determinamos la glucemia basal: blancos, negros y otros. Este tipo de ANOVA se conoce como *ANOVA de una vía* (*one-way ANOVA*).[4] La hipótesis nula general en un análisis de la varianza de una vía es que las medias de los k + 1 subgrupos son iguales entre sí. En nuestro ejemplo, la hipótesis nula general sería que la media de la glucemia basal de los blancos es igual a la de los negros y a la de las personas de otras razas.

Las categorías creadas por las k variables independientes nominales, que definen k + 1 subgrupos, deben ser *mutuamente excluyentes*. Esto significa que un individuo no puede pertenecer a más de una categoría. Por ejemplo, en la investigación médica, se suelen contemplar las razas como categorías mutuamente excluyentes. Para cada individuo se registra una sola categoría de raza. En este contexto es imposible que un individuo sea considerado blanco y negro a la vez.

[4] Cuando k = 1, en el análisis solo se considera una variable nominal. En este caso, estamos comparando solo dos subgrupos y el análisis de la varianza de una vía es exactamente lo mismo que una prueba de la *t* de Student en el análisis bivariante.

Cuando analizamos un grupo de variables como la raza y el sexo, las variables individuales muchas veces no son mutuamente excluyentes. Por ejemplo, un individuo puede ser hombre o mujer sea cual fuere su raza. Por lo tanto, es necesario disponer de otra vía que permita que las variables independientes nominales definan los subgrupos. Habitualmente, la solución de este problema es separar estas variables en *factores* (*factors*). Un factor es un conjunto de variables independientes nominales que define categorías mutuamente excluyentes pero relacionadas. Por ejemplo, suponga que tenemos dos variables independientes que definen la raza y una que define el sexo de las personas de nuestra muestra en las que hemos medido la glucemia basal. Las tres variables independientes de este ejemplo representan realmente dos factores separados: raza y sexo. En lugar de $k + 1 = 4$ subgrupos, definimos $(k_{raza} + 1) \times (k_{sexo} + 1) = 6$ subgrupos entre los cuales deseamos comparar la media de la glucemia basal: hombres blancos, mujeres blancas, hombres negros, mujeres negras, hombres de otras razas y mujeres de otras razas. El tipo de ANOVA que considera varios factores, así como las diferentes categorías dentro de cada factor, se conoce como *ANOVA factorial* (*factorial ANOVA*).

En el ANOVA factorial podemos contrastar el mismo tipo de hipótesis nula general que en el ANOVA de una vía. En nuestro ejemplo, la hipótesis nula sería que la media de la glucemia basal de las mujeres blancas es igual a la de los hombres blancos, los hombres negros, las mujeres negras, los hombres de otras razas y las mujeres de otras razas. Además, podemos contrastar las hipótesis de la igualdad de las medias de la glucemia basal entre los subgrupos de un determinado factor. Esto equivale a decir que podemos examinar el efecto por separado de la raza sobre la media de la glucemia basal o el efecto del sexo sobre la variable dependiente. Las pruebas estadísticas que se emplean para examinar los factores por separado se denominan pruebas de los *efectos principales* (*main effects*). Todas estas hipótesis nulas de los ANOVA se contrastan utilizando la *distribución de F* (*F distribution*).

Los resultados del análisis de un efecto principal tienen en cuenta las posibles relaciones de confusión de las otras variables independientes. En nuestro ejemplo, si contrastamos la hipótesis nula según la cual las medias de la glucemia basal son iguales en los tres subgrupos raciales mediante una prueba de ANOVA del efecto principal de la raza, esta prueba controlaría los resultados según cualquier diferencia en la distribución del sexo de esos grupos raciales. De este modo, el ANOVA factorial nos permite beneficiarnos de la capacidad del análisis multivariante para controlar el efecto de las variables de confusión.

Para interpretar las pruebas de los efectos principales, es necesario suponer que el factor tiene la misma relación con la variable dependiente sea cual fuere el nivel de los otros factores. Es decir, suponemos que la diferencia entre las medias de la glucemia basal de los negros, los blancos y las personas de otras razas es la misma independientemente de que el individuo sea hombre o mujer. Esto no es siempre así. Por ejemplo, las mujeres blancas pueden tener una glucemia basal más elevada que los hombres blancos, pero la glucemia puede ser similar en las mujeres y los hombres negros o, de forma más extrema, los hombres negros pueden tener una glucemia más elevada que las mujeres de esa misma raza. Cuando entre los factores existe este tipo de relación, decimos que existe una *interacción* (*interaction*) entre el sexo y la raza. Usando la terminología médica, podríamos decir que existe un sinergismo entre la raza y el sexo en la determinación de los valores de la glucemia basal. Además de la prueba de los efectos principales, el ANOVA factorial puede usarse para contrastar hipótesis sobre las interacciones.

Como hemos visto, el ANOVA factorial nos permite utilizar la segunda ventaja de los métodos multivariantes para controlar las variables de confusión. En nuestro ejemplo, hemos supuesto que el interés principal se centraba en la relación entre la raza y la glucemia basal, y que deseábamos controlar el posible efecto de confusión del sexo. Otra forma de tratar los datos presentados en este ejemplo sería la de considerar la raza y el sexo como factores que se pueden utilizar para estimar la glucemia basal. En este caso, en lugar de analizar el efecto principal de la raza mientras se controla según el sexo, utilizaríamos el ANOVA factorial, para comparar la relación de la raza y la del sexo con la glucemia basal. De ese modo, el ANOVA factorial nos permitiría examinar por separado la capacidad de la raza y el sexo para estimar la glucemia basal. Este es un ejemplo de la tercera ventaja de los métodos multivariantes.

El ANOVA de una vía y el factorial son métodos útiles para analizar grupos de observaciones que incluyen más de una variable independiente nominal y una variable dependiente que se haya medido una sola vez en cada individuo. La figura 29-1 se refiere a este método como diseño *no apareado* (*unmatched*). Sin embargo, sabemos que a veces se desea medir la variable dependiente repetidamente en el mismo individuo. En el capítulo 27 analizamos el ejemplo sencillo de un estudio en el que la tensión arterial se medía antes y después de un tratamiento antihipertensivo. En aquel ejemplo, la prueba de significación estadística apropiada y también adecuada para construir los intervalos de confianza era la *t* de Student para datos apareados.

A menudo, los estudios realizados en medicina se diseñan de tal forma que incluyen diversas mediciones repetidas de la variable dependiente y, a veces, exigen controlar los datos según varias variables de confusión. Por ejemplo, supongamos que todavía nos interesa estudiar la respuesta de la tensión arterial a la medicación antihipertensiva. Sin embargo, imaginemos ahora que no sabemos cuánto tiempo debe durar el tratamiento para que la tensión arterial se estabilice. En este caso, podríamos diseñar un ensayo clínico para medir la tensión arterial antes del tratamiento y mensualmente durante el primer año de tratamiento. Dado que disponemos de más de dos mediciones de la variable dependiente en cada individuo, denominamos a este diseño *apareado* (*matched*) en lugar de diseño apareado por dúos (o por pares, en el que se aparean dos individuos) (*paired*). Además, supongamos que estamos interesados en los efectos potenciales de confusión de la edad y el sexo. Para analizar las observaciones de este estudio, necesitaríamos un método estadístico distinto de la prueba de la *t* de Student para datos apareados. Un diseño especial del ANOVA nos permite considerar diversas mediciones de la variable dependiente para cada individuo y controlar según los efectos de confusión de otras variables. Este diseño se conoce como *ANOVA para medidas repetidas* (*repeated measures ANOVA*).[5]

En los análisis de la varianza para datos apareados e independientes, la hipótesis nula general mantiene una tasa de error de tipo I del experimento igual a alfa. No obstante, rara vez es suficiente saber que existen diferencias entre las medias dentro de un factor sin conocer específicamente cuál es la categoría en la que difieren esas medias. Es decir, no es suficiente saber que la media de la glucemia basal difiere según la raza sin conocer las razas que contribuyen a esa diferencia. Para examinar las medias de los subgrupos con mayor detalle, empleamos pruebas por dúos.[6] De estas,

[5] En el ANOVA para medidas repetidas, uno de los factores identifica los sujetos individuales, y la variable dependiente se mide para todas las categorías de, como mínimo, otro factor denominado factor "repetido". En ámbitos distintos de la estadística médica este diseño se denomina ANOVA de bloques aleatorios (*randomized block ANOVA*).

[6] En el ANOVA, estas pruebas por dúos o pares se denominan con frecuencia pruebas *a posteriori*. La razón de esta terminología es que algunas pruebas por pares, especialmente las antiguas, exigen haber realizado una prueba de significación estadística de la hipótesis nula general antes de utilizarlas.

la prueba utilizada más ampliamente en grupos de observaciones que incluyen una variable dependiente continua y más de una variable independiente nominal es la *prueba de Student-Newman-Keuls*. Esta prueba permite examinar todos los pares de medias de los subgrupos mientras se mantiene una tasa de error de tipo I del experimento $\alpha = 0{,}05$.[7] Una reorganización algebraica de la prueba de Student-Newman-Keuls permite calcular los intervalos de confianza de la variable dependiente para cada valor de las variables independientes.

Variables independientes continuas

Cuando las variables independientes de un estudio son continuas, podemos escoger entre dos enfoques que corresponden a los tratados en el capítulo 28, cuando considerábamos el análisis de regresión y el de la correlación. Casi siempre nos interesa estimar los valores de la variable dependiente para todos los valores posibles de las variables independientes. En el análisis bivariante, utilizamos la regresión para estimar el valor de la variable dependiente dado un valor de la variable independiente. Cuando tenemos más de una variable independiente continua, el interés en la estimación se puede mantener utilizando el *análisis de regresión múltiple* (*multiple regression analysis*).

En la regresión múltiple se estima la media de la variable dependiente continua mediante una ecuación lineal que es similar a la de la regresión lineal simple, excepto que incluye dos o más variables independientes continuas.

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Por ejemplo, suponga que nos interesa estimar la concentración de cortisol plasmático a partir del recuento de glóbulos blancos (RGB), la temperatura corporal y la producción de orina en respuesta a una sobrecarga de líquidos. Para investigar esta relación, medimos el cortisol (μg/100 ml), los glóbulos blancos (10^3), la temperatura (°C) y la producción de orina (ml) en 20 pacientes. Mediante una regresión múltiple podemos estimar la siguiente ecuación lineal:

$$\text{Concentración de cortisol} = -36{,}8 + 0{,}8 \times \text{GB} + 1{,}2 \times \text{temperatura} + 4{,}7 \times \text{orina}$$

Del mismo modo que en el ANOVA, en la regresión múltiple podemos contrastar una hipótesis general que tiene una tasa de error de tipo I igual a α. En la regresión múltiple, según esta hipótesis, *no* se puede utilizar el conjunto de variables independientes para estimar los valores de la variable dependiente. Para evaluar la significación estadística de la hipótesis nula general se emplea una prueba *F*. Supongamos que, en nuestro ejemplo, obtenemos una *F* estadísticamente significativa. Esto quiere decir que, si conocemos el recuento de glóbulos blancos, la temperatura y la producción de orina de un paciente, podemos estimar o tener una idea aproximada de su concentración de cortisol plasmático.

Además del interés en la hipótesis nula general, en la regresión múltiple casi siempre es deseable examinar individualmente las relaciones entre la variable dependiente y las variables independientes.[8] Los coeficientes de regresión asociados con las variables independientes constituyen una de las formas en las que se re-

[7] Se dispone de otras pruebas por pares para realizar comparaciones como estas o para efectuar comparaciones distintas entre las medias de los subgrupos. Un ejemplo de un tipo de comparación distinta es aquel en el cual deseamos comparar un grupo de control con una serie de grupos experimentales.

[8] El análisis de la relación entre las variables individuales independientes y la dependiente es análogo al examen de los factores en el ANOVA factorial.

CUADRO 29-1. Pruebas *F* parciales de los coeficientes de regresión estimados para variables independientes utilizadas para predecir la concentración plasmática de cortisol

Variable	Coeficiente	*F*	Valor *P*
Recuento de granulocitos	0,8	1,44	0,248
Temperatura	1,2	4,51	0,050
Orina	4,7	9,51	0,007

flejan estas relaciones. Los coeficientes de regresión son estimaciones de las β de la ecuación de regresión. Los resultados del análisis de regresión múltiple permiten efectuar una estimación puntual y calcular los intervalos de confianza de estos coeficientes. En las pruebas de significación estadística de los coeficientes individuales se utiliza una *prueba F parcial* para contrastar la hipótesis nula de que el coeficiente es igual a cero. El cuadro 29-1 muestra las pruebas *F* parciales de las variables independientes utilizadas para estimar la concentración de cortisol plasmático. Aunque en este ejemplo se rechazó la hipótesis general, observamos que solo los coeficientes de la producción de orina y la temperatura son estadísticamente significativos.

En la regresión bivariante, los coeficientes de regresión estiman la pendiente de los valores explicativos lineales de la variable dependiente en función de la variable independiente en la población de la que se extrajo la muestra. En la regresión multivariante, la relación entre la variable dependiente y cualquier variable independiente no es tan directa. El coeficiente de regresión realmente refleja la relación que existe entre los cambios que quedan en los valores numéricos de la variable independiente asociados con cambios de la variable dependiente *después de haber tenido en cuenta los cambios de la variable dependiente asociados con los cambios de los valores de todas las demás variables independientes*. Es decir, la contribución de cualquier variable independiente particular en la regresión múltiple solo es la contribución *que se superpone a las contribuciones de todas las otras variables independientes*. Esto constituye una buena noticia y a la vez una mala noticia. La buena noticia es que los coeficientes de regresión múltiple se pueden considerar como el reflejo de la relación entre la variable dependiente y las variables independientes "que controlan" según los efectos de las otras variables independientes. Por ello, la regresión múltiple se puede utilizar para eliminar el efecto de una variable de confusión continua.

La mala noticia es que "controlar" según el efecto de otras variables independientes es sinónimo de eliminar la variación de la variable dependiente que está asociada con esas otras variables independientes. Si cada una de dos variables independientes puede explicar por sí sola los mismos cambios numéricos de la variable dependiente, en una regresión múltiple las dos *juntas* no tendrán importancia para explicar los cambios de la variable dependiente.[9] No obstante, si se tiene en cuenta este resultado, se puede utilizar la regresión múltiple para examinar por separado la capacidad de las variables independientes para explicar la variable dependiente.

[9] El hecho de que las variables independientes compartan información predictiva se conoce como *multicolinealidad* (*multicollinearity*). Si bien es posible percatarse de que las variables independientes comparten información examinando los coeficientes de correlación bivariantes entre estas variables, el mejor método para evaluar la existencia de multicolinealidad es inspeccionar los modelos de regresión que incluyen y excluyen a cada variable independiente. Existe multicolinealidad si los coeficientes de regresión cambian sustancialmente cuando se consideran modelos diferentes.

Por ejemplo, suponga que nos interesa conocer el gasto cardíaco durante el ejercicio. Como variables independientes se estudian el gasto energético, la frecuencia cardíaca y la tensión arterial sistólica. Sabemos que cada una de estas variables está fuertemente asociada con el gasto cardíaco. Sin embargo, en un análisis de regresión múltiple sería improbable que la asociación entre cualquiera de ellas y la variable dependiente fuera estadísticamente significativa. Este resultado se puede prever, dada la gran cantidad de información sobre el gasto cardíaco que comparten estas variables independientes.

En la regresión múltiple, la construcción de los intervalos de confianza y el cálculo de las pruebas de significación estadística para los coeficientes asociados individualmente con las variables independientes son paralelos a los análisis por pares del ANOVA. En el ANOVA, los análisis por pares se diseñan para mantener una tasa de error de tipo I del experimento igual a α. En la regresión múltiple, la tasa de error de tipo I de la prueba es igual a α, pero la tasa de error del experimento depende del número de variables independientes incluidas. Cuantas más variables independientes examinemos en la regresión múltiple, mayor será la probabilidad de que al menos un coeficiente de regresión parezca significativo aunque no exista una relación entre esas variables en la población de la que se ha extraído la muestra. Por lo tanto, asociaciones estadísticamente significativas entre la variable dependiente y las independientes, que no se esperaba tuvieran importancia antes de analizar los datos, deben interpretarse con cierto escepticismo.[10]

Si todas las variables independientes continuas de un grupo de observaciones son el resultado de un muestreo aleatorio de alguna población de interés, podríamos estimar la fuerza de la asociación entre la variable dependiente y todas las variables independientes. Esto es paralelo a nuestro interés en el análisis de la correlación bivariante. En el análisis multivariante, el método utilizado para medir el grado de asociación se denomina análisis de la correlación múltiple. El resultado del análisis de la correlación múltiple se puede expresar tanto como un coeficiente múltiple de determinación o como su raíz cuadrada, el *coeficiente de correlación múltiple* (*multiple correlation coefficient*). Es importante recordar que estos estadísticos reflejan el grado de asociación entre la variable dependiente y todas las variables independientes. Por ejemplo, suponga que en nuestro ejemplo obtenemos un coeficiente de determinación de 0,82, lo que quiere decir que 82% de la variación de la concentración del cortisol plasmático de los pacientes puede explicarse conociendo el recuento de glóbulos blancos, la temperatura y la producción de orina. La prueba *F* estadísticamente significativa correspondiente a la prueba de la hipótesis nula de la regresión múltiple también contrasta la hipótesis nula según la cual el coeficiente de determinación poblacional es igual a cero. A partir de estos mismos cálculos se pueden derivar los intervalos de confianza de los coeficientes de determinación.

Variables independientes nominales y continuas

Muchas veces nos encontramos con una serie de observaciones en las que algunas de las variables independientes son continuas y algunas nominales. Por

[10] Esta perspectiva de la inferencia estadística y de la estimación por intervalo es un ejemplo de la aproximación bayesiana. En la inferencia bayesiana, consideramos el valor *P* y la probabilidad anterior, independiente de los datos, de la hipótesis nula como verdadera para determinar la probabilidad de la hipótesis nula a la luz de los datos.

ejemplo, suponga que diseñamos un estudio para explicar el gasto cardíaco a partir del gasto energético durante el ejercicio. Además, esperamos que la relación entre el gasto cardíaco y el energético sea diferente entre ambos sexos. En este ejemplo, nuestras observaciones comprenderían una variable dependiente continua, el gasto cardíaco; una variable independiente continua, el gasto energético; y una variable independiente nominal, el sexo.

Para examinar estos datos, que contienen una variable dependiente continua y una mezcla de variables independientes continuas y nominales, utilizamos una prueba denominada *análisis de la covarianza* (*analysis of covariance*) o ANCOVA. Las variables independientes continuas en el ANCOVA se relacionan con la variable dependiente de la misma forma que en la regresión múltiple. Las variables independientes nominales se relacionan con la variable dependiente de la misma forma que las variables independientes nominales se relacionan con la variable dependiente continua en el ANOVA. Por lo tanto, el ANCOVA es un método híbrido que contiene aspectos de la regresión múltiple y del ANOVA.

Un uso común del ANCOVA que es similar al del ANOVA es el estudio de la estimación de una variable dependiente continua a partir de una variable independiente nominal mientras se controla el efecto de una segunda variable. En el ANCOVA, la variable que se controla es continua. Un ejemplo de esto lo constituye la capacidad de controlar los efectos de confusión de la edad cuando se estudia la asociación entre una variable independiente nominal, como el tratamiento frente al no tratamiento, y una variable dependiente continua, como la tensión arterial diastólica.

El ANCOVA también se puede considerar como un método de análisis de regresión múltiple en el cual algunas de las variables independientes son nominales en lugar de continuas. Para incluir una variable independiente nominal en una regresión múltiple, tenemos que transformarla a una escala numérica. Una variable nominal expresada numéricamente se denomina variable *ficticia o indicadora* (*indicator* o "*dummy*" variable).[11]

Con frecuencia, los valores numéricos asociados con una variable nominal son el cero y el 1. En este caso, el valor 1 se asigna arbitrariamente a las observaciones en las cuales está representada una de las dos categorías potenciales de la variable nominal; y el cero, a la categoría no representada. Por ejemplo, si introdujéramos el sexo femenino en una regresión múltiple, podríamos asignar el valor 1 a las mujeres y el cero a los hombres.

Para ver cómo se pueden interpretar las variables indicadoras en la regresión múltiple, reconsideremos el ejemplo anterior: tenemos una variable independiente nominal para describir el sexo y una variable independiente continua, el gasto energético, para describir la variable dependiente continua del gasto cardíaco. El modelo de regresión múltiple en este ejemplo se expresa del siguiente modo:

[11] Aunque podemos considerar el ANCOVA como una extensión del ANOVA o de la regresión múltiple, esto no significa que la interpretación del ANCOVA sea distinta según el método aplicado. En el ejemplo del gasto cardíaco descrito como función del sexo y del gasto energético, podríamos realizar un ANCOVA como un ANOVA con un factor, el sexo, que controle el efecto del gasto energético como si este constituyera una variable de confusión. Al hacerlo, obtendríamos resultados idénticos a los de una regresión. En realidad, el ANOVA, el ANCOVA y la regresión múltiple son ejemplos del mismo método estadístico conocido como *modelo lineal general* (*general linear model*). El ANCOVA se puede representar como una regresión múltiple en la que las variables independientes son representaciones numéricas de variables nominales. Los "efectos principales" se miden mediante coeficientes asociados con las variables indicadoras; y las "interacciones", mediante el producto de estas variables indicadoras. En la regresión, estas también se denominan interacciones.

$$\hat{Y} = \alpha + \beta_1 X + \beta_2 I$$

donde

$\hat{Y}$ = gasto cardíaco
X = gasto energético
I = indicador del sexo masculino
(1 para las mujeres, 0 para los hombres)

Dado que los hombres están representados por I = 0 y cero multiplicado por β_2 es cero, la ecuación de regresión múltiple para los hombres es igual a la siguiente ecuación bivariante de regresión:

$$\hat{Y} = \alpha + \beta_1 X$$

También podemos representar la ecuación para las mujeres como una regresión bivariante. En este caso, la variable indicadora o ficticia es igual a 1 y $1 \times \beta_2 = \beta_2$. Dado que β_2 y α son constantes para las mujeres, podemos describir las relaciones entre el gasto cardíaco y el energético entre las mujeres como:

$$\hat{Y} = (\alpha + \beta_2) + \beta_1 X$$

Si comparamos la ecuación de regresión para los hombres con la de las mujeres, podemos observar que el coeficiente de regresión asociado con la variable independiente nominal (β_2) es igual a la diferencia entre los puntos de intersección (el gasto cardíaco, cuando el gasto energético es igual a cero) para los hombres y para las mujeres.

Uno de los problemas que surgen cuando usamos la variable indicadora para comparar la relación entre el gasto cardíaco y el energético de los hombres con esta relación en las mujeres es que debemos suponer que los hombres y las mujeres se diferencian solamente en los puntos de intersección de sus ecuaciones de regresión individuales. Es decir, suponemos que un aumento de una unidad en el gasto energético se asocia con el mismo aumento en el gasto cardíaco en los hombres y en las mujeres. Esto implica que la pendiente de la relación entre el gasto cardíaco y el energético para los hombres es la misma que para las mujeres. Muchas veces no estamos dispuestos a aceptar este supuesto de la igualdad de las pendientes. Cuando esto sucede, podemos crear otro tipo de variable en el enfoque de la regresión múltiple del ANCOVA multiplicando una variable independiente continua por la nominal transformada a una escala numérica. Esta nueva variable se denomina *término de interacción* (*interaction term*).[12] En nuestro ejemplo, la ecuación del ANCOVA que incluye un término de interacción entre el gasto energético (X) y el sexo (I) sería:

$$\hat{Y} = \alpha + \beta_1 X + \beta_2 I + \beta_3 XI$$

Para los hombres, esta ecuación es de nuevo una ecuación de regresión bivariante, dado que I = 0 y, por lo tanto, $0 \times \beta_3 = 0$:

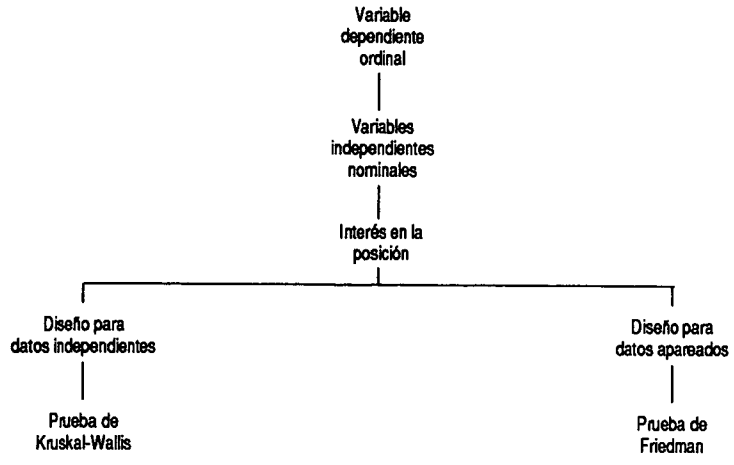
$$\hat{Y} = \alpha + \beta_1 X$$

Para las mujeres, dado que I = 1, la ecuación es

$$\hat{Y} = (\alpha + \beta_2) + (\beta_1 + \beta_3) X$$

[12] Los términos de interacción no se limitan al producto de una variable continua y una nominal. Muchas veces podemos observar interacciones que son el producto de dos variables nominales. También es posible considerar una interacción entre dos variables continuas, pero la interpretación de este producto es mucho más complicada.

FIGURA 29-2. Esquema para seleccionar un método estadístico multivariante para una variable dependiente ordinal (continuación de la figura 26-5)



El coeficiente para la variable indicadora (β_2) indica la diferencia entre los puntos de intersección para los hombres y para las mujeres. El coeficiente del término de interacción (β_3) nos informa de la diferencia entre las pendientes de ambos sexos. Por consiguiente, tenemos tres variables independientes: una variable continua, una variable nominal expresada como variable indicadora y un término de interacción. En esta situación, un ANCOVA es semejante a tener una regresión bivariante por separado para cada una de las dos categorías identificadas por la variable independiente nominal. En este ejemplo, podemos estimar mediante regresiones separadas la relación para los hombres y para las mujeres. Además, el ANCOVA nos permite comparar estas dos ecuaciones de regresión por medio del contraste de las hipótesis de los coeficientes de regresión de las variables indicadoras y de los términos de interacción.

VARIABLE DEPENDIENTE ORDINAL

En los análisis univariante y bivariante, disponíamos de métodos estadísticos para analizar las variables dependientes ordinales y para posibilitar la transformación de las variables dependientes continuas a una escala ordinal, cuando no se podían cumplir los supuestos necesarios para utilizar los métodos estadísticos diseñados para las variables dependientes continuas. Esto también es cierto para los métodos multivariantes con variables dependientes ordinales.

Idealmente, desearíamos disponer de métodos para las variables dependientes ordinales que fueran paralelos a los métodos multivariantes para las variables dependientes continuas: ANOVA, ANCOVA y regresión múltiple. Lamentablemente, esto no es así. Las únicas técnicas multivariantes aceptadas para las variables dependientes ordinales son aquellas que pueden usarse como equivalentes no paramétricos de ciertos diseños del ANOVA.[13] Por eso, la figura 29-2 se limita a los métodos que pueden emplearse *exclusivamente* con variables independientes nominales y una va-

[13] Aunque no es de uso amplio, el *análisis de regresión logística ordinal* (*ordinal logistic regression*) es un método prometedor que podría finalmente ganar aceptación como forma de incluir variables independientes continuas en el análisis multivariante de variables dependientes ordinales.

riable dependiente ordinal. Para poder aplicar esos métodos, las variables independientes continuas u ordinales deben transformarse a escalas nominales.

Por un momento, reconsideremos el ejemplo anterior de la glucemia basal medida en personas de tres categorías raciales (negra, blanca y otras) y de ambos sexos. En este ejemplo, nuestro interés se centraba en determinar los efectos independientes de la raza y el sexo en la glucemia. Para analizar estos datos, utilizamos un ANOVA factorial. Si estuviéramos preocupados por el cumplimiento de los supuestos del ANOVA[14] en relación con la glucemia basal, podríamos transformar estos datos a una escala ordinal mediante la asignación de rangos relativos a las mediciones de la glucemia basal. Entonces podríamos aplicar la *prueba de Kruskal-Wallis* a los datos transformados. Esta prueba es apropiada para realizar las pruebas de significación estadística de una variable dependiente ordinal y dos o más variables independientes nominales en un diseño de una vía o uno factorial. También existen técnicas no paramétricas para realizar comparaciones por pares entre los subgrupos de la variable dependiente.

Como hemos comentado anteriormente, los métodos estadísticos para las variables dependientes ordinales se conocen como no paramétricos, porque no exigen realizar supuestos acerca de los parámetros poblacionales. Los métodos no paramétricos permiten contrastar hipótesis relacionadas principalmente con la distribución general de la población. La distinción entre hipótesis paramétricas y no paramétricas, por lo tanto, reside en que en las segundas se hacen afirmaciones sobre la distribución de los valores para la población *general*, mientras que en las hipótesis paramétricas se realizan afirmaciones sobre medidas *específicas* resumidas o parámetros como la media poblacional.

Al analizar los datos de un estudio en el que se mide una variable dependiente continua tres o más veces en los mismos individuos o en individuos apareados, probablemente escogeríamos un ANOVA para medidas repetidas. Por otro lado, si la variable dependiente fuese ordinal o continua y deseáramos convertirla en ordinal para obviar los supuestos del ANOVA, todavía podríamos beneficiarnos del diseño apareado. Una prueba no paramétrica paralela al ANOVA para medidas repetidas es la *prueba de Friedman*.

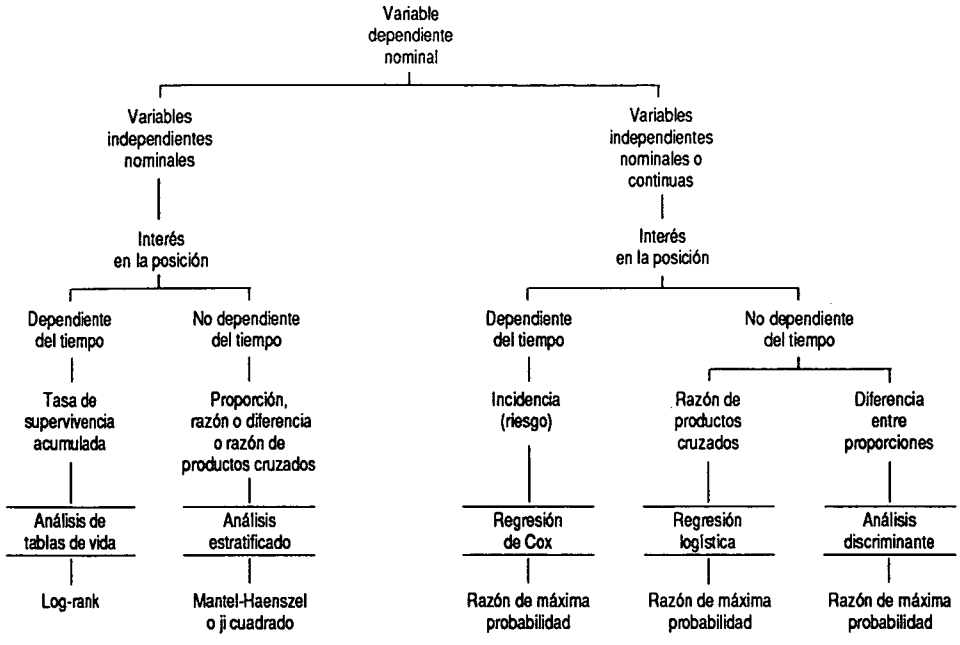
Cuando empleamos métodos multivariantes diseñados para variables dependientes ordinales con objeto de analizar grupos de observaciones que contienen una variable dependiente continua transformada a una escala ordinal, debemos tener en cuenta una desventaja potencial: que la técnica no paramétrica tiene menor potencia estadística que la paramétrica correspondiente si la variable dependiente continua no viola los supuestos de la prueba paramétrica. Esto se aplica a todas las técnicas estadísticas realizadas con variables continuas transformadas a una escala ordinal. Por eso, si se cumplen los supuestos de una prueba paramétrica, es aconsejable utilizarla para analizar una variable dependiente continua antes que la técnica no paramétrica paralela.

VARIABLE DEPENDIENTE NOMINAL

En la investigación médica, a menudo nos interesan los desenlaces de vida o muerte, o curación o no curación, medidos como datos nominales. Además, a causa de la complejidad de los fenómenos médicos, casi siempre es deseable me-

[14] Los supuestos del ANOVA y del ANCOVA son los mismos que los descritos anteriormente para el análisis de regresión.

FIGURA 29-3. Esquema para seleccionar un método estadístico multivariante para una variable dependiente nominal (continuación de la figura 26-5)



dir diversas variables independientes para considerar hipótesis separadas, para controlar según variables de confusión y para investigar la posibilidad de sinergismo o de interacción entre las variables. En consecuencia, los análisis multivariantes con variables dependientes nominales se emplean con frecuencia o se deben emplear en el análisis de los datos de la investigación médica.

Hemos separado las técnicas estadísticas multivariantes para variables dependientes nominales en dos grupos: las que son aplicables cuando las variables independientes son todas nominales y las que lo son para una combinación de variables independientes nominales y continuas (figura 29-3). Los análisis del primer grupo se limitan a las variables independientes nominales o a las transformadas a una escala nominal. Por otro lado, se pueden usar variables independientes nominales y continuas en el análisis del segundo grupo. No existe ningún método establecido para considerar las variables independientes ordinales, si no se transforman a una escala nominal.

Variables independientes nominales

Cuando analizamos una variable dependiente nominal y dos o más variables independientes nominales, nos interesan las medidas de posición, al igual que en el análisis bivariante de una variable dependiente nominal y una independiente nominal. Por ejemplo, podemos estar interesados en proporciones, tasas o ventajas (*odds*). Sin embargo, en el análisis multivariante de las variables nominales dependientes e independientes nos interesan aquellas mediciones de la frecuencia de la enfermedad al mismo tiempo que ajustamos según las otras variables independientes.

Por ejemplo, suponga que nos interesa comparar la prevalencia del cáncer de pulmón entre los bebedores de café en relación con la de los no bebedores.

En este caso, la prevalencia del cáncer de pulmón es la variable de interés y, por lo tanto, la variable dependiente nominal. Beber café (sí o no) es la variable independiente nominal. Al mismo tiempo, podríamos desear ajustar según el efecto de confusión potencial del consumo de cigarrillos. Para ello, podemos incluir otra variable independiente nominal que identifique a los fumadores respecto de los no fumadores.

Cuando tenemos dos o más variables independientes en un conjunto de datos y todas son nominales o han sido transformadas a una escala nominal, el enfoque general para ajustar según las variables independientes muchas veces es un *análisis estratificado* (*stratified analysis*). Como se ha descrito en la Sección 1, los métodos de análisis estratificado exigen separar las observaciones en subgrupos definidos por los valores de las variables independientes nominales que se consideran variables de confusión. En nuestro ejemplo sobre la prevalencia del cáncer de pulmón y del consumo de café, comenzaríamos el análisis estratificado dividiendo nuestras observaciones en dos grupos: uno compuesto por fumadores y otro, por no fumadores.

Dentro de cada subgrupo, como el de los bebedores y el de los no bebedores de café, estimaríamos la prevalencia de cáncer de pulmón en los fumadores y en los no fumadores por separado. Estas estimaciones separadas se conocen como estimaciones puntuales *específicas del estrato* (*stratum-specific*). Las estimaciones puntuales específicas del estrato se combinan empleando un sistema de *ponderación* (*weighting*) de los resultados de cada estrato. Es decir, combinaríamos la información de cada estrato utilizando uno de los muchos métodos disponibles para determinar cuánto impacto debe tener cada estimación específica del estrato en la estimación combinada.[15] La estimación combinada resultante se considera una estimación puntual ajustada o estandarizada para todos los estratos en conjunto con los efectos de la variable de confusión eliminados.

En el esquema hemos indicado dos tipos de variables dependientes: las tasas, que son *dependientes del tiempo*, y las proporciones, que no son dependientes del tiempo. Por dependiente del tiempo queremos decir que la frecuencia con la que se observa un desenlace nominal depende del tiempo de seguimiento de las personas. Por ejemplo, considere la muerte como una variable dependiente del tiempo. Si no estamos estudiando personas con una tasa de mortalidad inusualmente elevada, esperaríamos observar una proporción baja de personas fallecidas si siguiéramos al grupo durante, por ejemplo, un año. Por otro lado, si siguiéramos a este grupo durante 20 años, esperaríamos observar una proporción de muertes mucho más alta. Hasta ahora solo hemos presentado métodos multivariantes para variables dependientes nominales que no son dependientes del tiempo. Por ejemplo, hemos analizado la prevalencia de diversas enfermedades. La prevalencia no depende del tiempo, puesto que se refiere a la frecuencia de una enfermedad en un momento dado.

Las variables dependientes del tiempo pueden causar problemas de interpretación si los grupos que se comparan difieren en los períodos de seguimiento, lo cual sucede casi siempre. Estos problemas se pueden solventar si consideramos la incidencia como la variable dependiente, ya que la tasa de incidencia tiene una

[15] El sistema de ponderación de las estimaciones específicas del estrato es una de las formas en que se diferencian los distintos métodos de análisis estratificado. En la estandarización directa, el sistema de ponderación se basa en la frecuencia relativa de cada estrato en una población de referencia. Desde un punto de vista estadístico, los sistemas de ponderación más útiles son los que reflejan la precisión de las estimaciones específicas de los estratos.

unidad de tiempo en el denominador y, de ese modo, toma en cuenta el tiempo de seguimiento. Lamentablemente, la incidencia es una medida que puede interpretarse de forma errónea. Para la mayoría de las personas es difícil comprender intuitivamente el significado de *casos por año-persona* (*cases per person-year*). Por el contrario, es mucho más fácil comprender el *riesgo*. Recuerde que el riesgo es la proporción de personas que desarrollan un desenlace durante un período de tiempo determinado. No obstante, observe que el riesgo es una variable dependiente del tiempo, pues se calcula para un período de tiempo determinado. Del mismo modo, no es posible interpretar el riesgo calculado a partir de los datos que representan diversos períodos de tiempo, como lo es para la incidencia, porque el riesgo no contiene ninguna dimensión temporal en el denominador.

Si nos interesa el riesgo y los datos contienen observaciones realizadas en personas seguidas durante períodos de tiempo distintos, debemos emplear técnicas estadísticas especiales para ajustar según las diferencias en los períodos de seguimiento. Cuando todas las variables independientes son nominales, los métodos que utilizamos son tipos de *análisis de las tablas de vida* (*life-table analysis*). En estos métodos, los períodos de seguimiento, por ejemplo intervalos de 1 año, se consideran como un grupo de variables independientes nominales. Cada intervalo de 1 año se utiliza para estratificar las observaciones del mismo modo que se estratifican los datos según las categorías de una variable de confusión como el grupo de edad. La supervivencia acumulada (*cumulative survival*),[16] que es igual a 1 menos el riesgo, se determina combinando estas probabilidades ajustadas de sobrevivir cada período.

Generalmente, se emplean dos métodos para analizar la tabla de vida: el método de *Kaplan-Meier* o del *producto límite* (*product limit*) y el de *Cutler-Ederer* o *actuarial* (*actuarial*). Estos métodos se diferencian en la forma de manejar los datos de las personas cuyo seguimiento termina en un período.[17] En el método de Kaplan-Meier, se supone que el seguimiento termine al final de cierto intervalo de tiempo. Por su lado, en el método de Cutler-Ederer se supone que los tiempos de finalización del seguimiento se distribuyen uniformemente durante el período. Como consecuencia de estos supuestos diferentes, las estimaciones de riesgo del método de Cutler-Ederer tienden a ser ligeramente más altas que en el de Kaplan-Meier. Existen métodos estadísticos para calcular las estimaciones por intervalo y para realizar pruebas de significación estadística para ambos métodos.

Variables independientes continuas o nominales

El análisis estratificado que hemos presentado para las variables dependientes nominales, dependientes e independientes del tiempo, y para las variables independientes nominales tiene para muchos investigadores el atractivo de que parece más simple y controlable que otros tipos de análisis. No obstante, el análisis estratificado presenta algunas limitaciones. Este tipo de análisis se ha diseñado para examinar la relación entre una variable dependiente nominal y una independiente nominal mien-

[16] Las tablas de vida se diseñaron inicialmente para considerar el riesgo de muerte, pero pueden utilizarse para calcular el riesgo de cualquier desenlace irreversible.

[17] En el análisis de la tabla de vida, el seguimiento durante un período puede finalizar por diversos motivos. El más común es la terminación del estudio. A menudo, los estudios se diseñan para reclutar a los sujetos durante gran parte del período de estudio y suspender el seguimiento en una fecha concreta. Los sujetos reclutados al inicio del período contribuirán a los datos de cada período de análisis de la tabla de vida. Los sujetos reclutados hacia el final del estudio se siguen durante períodos más cortos y su seguimiento termina al finalizar el estudio. Otros sujetos pueden "perderse" durante un período de seguimiento, porque abandonan el estudio, porque fallecen debido a causas no relacionadas con el estudio, etc.

tras se controla según el efecto de una variable de confusión nominal. Este análisis no permite examinar directamente variables explicativas alternativas, investigar las interacciones o el sinergismo, considerar las variables continuas de confusión sin transformarlas a una escala nominal ni estimar la importancia de las variables de confusión. Muchas veces, estas son características de gran interés para los investigadores médicos.

Los métodos de análisis que permiten investigar simultáneamente las variables independientes nominales y continuas y sus interacciones son paralelas en su enfoque general a la regresión múltiple tratada anteriormente. Sin embargo, los métodos que empleamos aquí difieren de la regresión múltiple en tres aspectos. La primera diferencia, como se indica en el esquema, es que la regresión múltiple es un método de análisis de variables dependientes continuas, mientras que ahora estamos interesados en variables dependientes nominales. La segunda diferencia es que en la mayor parte de los métodos aplicables a las variables dependientes nominales, no se utiliza el método de los mínimos cuadráticos empleado en la regresión múltiple para encontrar el mejor ajuste de los datos. Casi siempre, los coeficientes de regresión de las variables dependientes nominales se estiman utilizando el método de la *máxima verosimilitud* (*maximum likelihood*).[18]

La tercera diferencia es quizá la más importante para los investigadores médicos que interpretan los resultados del análisis de regresión con variables dependientes nominales. Aunque este tipo de análisis proporciona estimaciones de los coeficientes de regresión y de sus errores estándares, el resto de la información que resulta del análisis es distinto del de la regresión múltiple. La razón consiste en que estos coeficientes de regresión no proporcionan estimaciones paralelas a los coeficientes de correlación. Por eso, sin un coeficiente de determinación, no es posible determinar el porcentaje de la variación de la variable dependiente que es explicado por el grupo de variables independientes.[19]

Para los desenlaces dependientes del tiempo, el método de regresión habitualmente empleado es el *modelo de Cox* (*Cox model*).[20] En este modelo, el grupo de variables independientes y, si se desea, sus interacciones, se emplean para estimar la incidencia[21] de la variable dependiente nominal,[22] como la incidencia de la muerte. Se puede utilizar una simple combinación algebraica de los coeficientes de cierto modelo de Cox para estimar la curva de supervivencia en una serie de valores de variables independientes. Cuando todas las variables independientes son nominales, el modelo de Cox estima las curvas de supervivencia que son muy semejantes a las que resultan del análisis de la tabla de vida de Kaplan-Meier. Por eso, cada vez se observa con más frecuencia el uso de este modelo en la investigación médica, tanto para la construcción de curvas de las tablas de vida como para ajustar los datos según las variables de confusión.

Las variables dependientes nominales que no dependen del tiempo se analizan frecuentemente mediante uno o dos métodos multivariantes: el *análisis discriminante* (*discriminant analysis*) y la *regresión logística* (*logistic regression*).

---

[18] El método de la máxima verosimilitud selecciona las estimaciones de los coeficientes de regresión para maximizar la probabilidad de que los datos observados hubieran resultado del muestreo de una población con estos coeficientes.

[19] Se ha propuesto un sustituto para el coeficiente de determinación, pero los estadísticos no están convencidos de su utilidad.

[20] Este método también se conoce como la *regresión de Cox* (*Cox regression*) o *modelo de riesgos proporcionales* (*proportional hazards regression*).

[21] En el modelo de Cox, casi siempre se utiliza el término *riesgo* (*hazard*) como sinónimo de incidencia.

[22] En realidad, el modelo de Cox predice el logaritmo neperiano de la razón de la incidencia ajustada según las variables independientes dividida por la incidencia no ajustada según estas variables.

Como se deduce de su nombre, el análisis discriminante está diseñado para discriminar entre subgrupos definidos por una variable dependiente nominal. Aquí, nos hemos limitado al análisis que abarca una variable dependiente y, por lo tanto, solo estamos interesados en discriminar entre dos subgrupos. No obstante, una de las ventajas del análisis discriminante es la facilidad con que puede extenderse al análisis de más de dos subgrupos. De este modo, puede utilizarse para datos nominales con más de dos categorías potenciales, como un método estadístico multivariante.

El análisis discriminante es muy similar a la regresión múltiple por el método de los mínimos cuadráticos,[23] y permite estimar un coeficiente de determinación y estadísticos relacionados. Los coeficientes de regresión estimados en el análisis discriminante se pueden utilizar para predecir la probabilidad de pertenencia a un subgrupo de individuos con un determinado grupo de valores en las variables independientes.

Algunos estadísticos consideran que dos características del análisis discriminante imponen limitaciones. Ambas están relacionadas con el hecho de que el análisis discriminante es prácticamente una regresión múltiple con una variable dependiente nominal. La primera es que el análisis discriminante está basado en los mismos supuestos que el análisis de regresión múltiple. El problema estriba concretamente en el supuesto de que la variable dependiente sigue una distribución gausiana. Esto no sucede con una variable nominal. Por suerte, el análisis de regresión múltiple es un método robusto que permite una violación considerable de sus supuestos antes de que esta violación influya en los resultados.

La segunda limitación del análisis discriminante es que supone que la probabilidad de pertenencia a un subgrupo sigue una línea recta o una función lineal. Si esto es así, el análisis discriminante es el método apropiado. No obstante, una característica de una función lineal es que, teóricamente, está comprendida entre $-\infty$ y $+\infty$. Dado que las probabilidades pueden tomar valores entre 0 y 1, es posible predecir valores absurdos de la variable dependiente para ciertos valores de las variables independientes. Algunos estadísticos consideran que esta capacidad para hacer predicciones imposibles es un inconveniente del análisis discriminante.

Como alternativa, a menudo las variables dependientes nominales que no dependen del tiempo se analizan mediante la regresión logística. Existen tres diferencias importantes entre la regresión logística y el análisis discriminante. La primera es que la regresión logística no está tan estrechamente relacionada con la regresión múltiple como para compartir el supuesto de que una variable dependiente sigue una distribución gausiana. La segunda es que la variable dependiente no se expresa directamente como la probabilidad de pertenencia a un grupo. La tercera es que las técnicas de regresión logística no se pueden ampliar fácilmente para considerar más de una variable nominal.

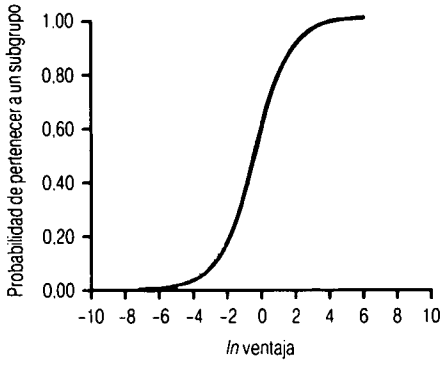
En la regresión logística, la variable dependiente es el logaritmo neperiano de la ventaja (*odds*) de pertenencia a un grupo.[24] Con esta presentación de la variable dependiente, la transformación resultante para estimar las probabilidades de pertenencia a un subgrupo se reduce al intervalo comprendido entre 0 y 1.[25] Específi-

[23] De hecho, el análisis discriminante solamente se diferencia del método de los mínimos cuadráticos de regresión de una variable dependiente nominal en un multiplicador constante.

[24] Esto se conoce como *transformación logit* (*logit transformation*).

[25] Otro modelo de regresión que tiene la propiedad de estimar las probabilidades del intervalo comprendido entre 0 y 1 es el *análisis probit* (*probit analysis*). Este tipo de análisis no se ve con frecuencia en la literatura médica, excepto en los ensayos clínicos de medicamentos con animales de laboratorio.

FIGURA 29-4. Ejemplo de una curva sigmoidea correspondiente a la probabilidad de pertenencia a un subgrupo determinada a partir del *ln* de la ventaja (*log odds*)



camente, estas transformaciones siguen una curva *sigmoidea* dentro del intervalo comprendido entre 0 y 1 (figura 29-4). Por consiguiente, la regresión logística satisface a los estadísticos que se preocupan porque el análisis discriminante permite valores imposibles.[26]

Los coeficientes de regresión que se calculan con el análisis de la regresión logística se usan con frecuencia para estimar la razón de productos cruzados o de ventajas (*odds ratio*). Veamos, mediante un ejemplo, cómo se interpretan estas razones de productos cruzados calculadas con la regresión logística. Supongamos que hemos llevado a cabo un estudio transversal en un grupo de personas con arco senil y que las hemos comparado con otro grupo de personas en quienes el mismo oftalmólogo ha practicado un examen de la refracción. Hemos registrado la edad, el sexo y la concentración de colesterol sérico de cada sujeto. Supongamos que hemos obtenido los coeficientes de regresión logística que aparecen en el cuadro 29-2, al analizar estos datos mediante una regresión logística con la aparición o no del arco senil como variable dependiente.

Algo que podemos decir a partir de los datos del cuadro 29-2 es que la edad, el sexo y la concentración de colesterol sérico son estimadores estadísticamente significativos de la aparición de un arco senil. Sin embargo, no es fácil interpretar los coeficientes de regresión para determinar la fuerza de la asociación de la ventaja (*odds*) de tener arco senil con, por ejemplo, el sexo. Esto se facilita si convertimos estos coeficientes a una razón de productos cruzados. Para el sexo, el coeficiente de regresión logística de 1,50 equivale a una razón de productos cruzados de 4,5. Esto significa que, controlando según los efectos de la edad y la concentración de colesterol sérico, las mujeres tienen 4,5 veces más ventajas de tener un arco senil que los hombres.

Normalmente no pensamos en las razones de productos cruzados en relación con variables continuas. No obstante, la capacidad de incluir variables continuas independientes es una de las ventajas de la regresión logística sobre el análisis estratificado. También pueden interpretarse los coeficientes de regresión logística de las

[26] Sin embargo, no existe ninguna garantía de que el modelo logístico sea *biológicamente* apropiado para analizar cualquier grupo determinado de observaciones. La calidad de las pruebas determinará el grado con que el análisis discriminante y el logístico se ajustarán a un grupo de observaciones.

CUADRO 29-2. Coeficientes de regresión de una regresión logística en la cual la presencia de arco senil es la variable dependiente

Variable	Coeficiente	Valor *P*
Edad	0,10	0,002
Sexo (mujer)	1,50	0,030
Colesterol	0,30	0,010

variables independientes continuas con las razones de productos cruzados. Para ello, debemos seleccionar un incremento de la variable continua para el que se pueda calcular la razón de productos cruzados. Por ejemplo, podemos escoger el cálculo de la ventaja del arco senil para un incremento de 10 años como el de las personas con 60 años respecto de las de 50 años. En este ejemplo, la razón de productos cruzados es de 2,7. Además, el diseño concreto de la regresión logística implica que podríamos obtener la misma razón de productos cruzados para *cualquier* diferencia de 10 años de edad.

RESUMEN

El análisis multivariante nos permite analizar grupos de observaciones que incluyen más de una variable independiente. Al proporcionar un método para tomar en cuenta varias variables independientes a la vez, el análisis multivariante ofrece tres ventajas: 1) poder controlar el efecto de las variables de confusión, 2) evitar frecuentemente el problema de las comparaciones múltiples, y 3) poder comparar la capacidad de las variables independientes para estimar los valores de la variable dependiente.

Los métodos multivariantes aplicables a variables dependientes continuas son, en su mayor parte, extensiones de los análisis bivariantes que permiten considerar más de una variable independiente. Para las variables independientes nominales, la extensión de la técnica bivariante de la *t* de Student es el análisis de la varianza (ANOVA). En el ANOVA podemos examinar las variables independientes nominales que indican diversas categorías de una característica concreta o analizar grupos de variables independientes nominales conocidas como factores. En el ANOVA se pueden contrastar dos tipos de hipótesis nulas. La hipótesis nula general afirma que todas las medias son iguales. Las hipótesis nulas por pares afirman que las medias de una pareja concreta son iguales. Ambos tipos de hipótesis se contrastan con una tasa de error de tipo I del experimento igual a alfa = 0,05 independientemente del número de medias comparadas.

Un tipo especial de ANOVA muy útil en la investigación médica es el ANOVA para medidas repetidas. Esta técnica es una extensión de la prueba univariante de la *t* de Student aplicada a datos apareados. Mediante el ANOVA para medidas repetidas se pueden analizar grupos de observaciones en las cuales la variable dependiente se mida más de dos veces en el mismo individuo o podemos emplearlo para controlar según el efecto de las variables de confusión potenciales, o para ambos propósitos a la vez.

La asociación entre una variable dependiente continua y dos o más variables independientes continuas se investiga mediante el análisis de regresión múltiple, una extensión de la regresión lineal bivariante. La capacidad de considerar más de una variable independiente en el análisis de la regresión múltiple permite controlar el efecto de las variables de confusión y comparar la capacidad de varias variables in-

dependientes para estimar los valores de la variable dependiente. Las relaciones entre la variable dependiente y las independientes deben interpretarse reconociendo que los coeficientes de regresión múltiple están influidos por la capacidad de las otras variables independientes para explicar la relación. La fuerza de una asociación entre una variable dependiente continua y un conjunto de variables independientes continuas se estima mediante el coeficiente de correlación múltiple.

Muchas veces tenemos una variable dependiente continua, una o más variables independientes nominales y una o más variables independientes continuas. Este grupo de observaciones se analiza mediante el análisis de la covarianza (ANCOVA). El ANCOVA comparte características de la regresión múltiple y del análisis de la varianza.

De la misma forma que en el análisis bivariante, los métodos multivariantes para las variables dependientes ordinales se pueden considerar como paralelos no paramétricos de las pruebas para variables dependientes continuas. Sin embargo, en el análisis multivariante los únicos métodos usados habitualmente son paralelos a los del ANOVA.

Con las variables dependientes nominales, las pruebas que se emplean son tipos especiales del análisis de la regresión o métodos que exigen estratificar los datos. La estratificación exige que todas las variables independientes sean nominales o que hayan sido transformadas a una escala nominal. Las técnicas de regresión pueden incluir variables dependientes nominales o continuas.

Para ambos métodos, existe una distinción adicional en el análisis de las variables dependientes nominales que consiste en determinar si las medidas de posición son dependientes del tiempo o no. El análisis de la tabla de vida es una técnica de estratificación para las variables nominales que son dependientes del tiempo. Una técnica de regresión paralela es la regresión de Cox. La regresión logística es el método más empleado para analizar las variables dependientes que no dependen del tiempo. Los coeficientes de la regresión logística se pueden convertir en razones de productos cruzados. Otra técnica es el análisis discriminante. Una ventaja del análisis discriminante es que puede extenderse a más de una variable dependiente nominal.

CAPÍTULO 30

RESUMEN ESQUEMÁTICO

En este capítulo presentamos en su totalidad el esquema necesario para seleccionar una prueba estadística. El esquema resumido puede utilizarse de dos maneras. La primera consiste en empezar en la página 317 y seguir el esquema hasta descubrir cuáles son los tipos de técnicas estadísticas apropiados para una investigación determinada. Para usar el esquema de esta manera, primero debe identificar una variable dependiente y luego 0, 1 o más variables independientes. Seguidamente, ha de decidir el tipo de la variable dependiente (esto es, nominal, ordinal o continua). Una vez que haya tomado estas decisiones, usted encontrará un número que le conducirá a la siguiente parte del esquema aplicable a sus datos.

Todas las partes subsiguientes del esquema se han construido de la misma forma. Si sus datos contienen variables independientes, deberá identificar el tipo.[1] A continuación, en algunos diagramas tendrá que decidir cuál es el parámetro poblacional que le interesa, la posición o la dispersión.[2] Si existen limitaciones o supuestos especiales aplicables a las técnicas estadísticas apropiadas para analizar sus datos, será necesario determinar si se cumplen. En el caso de que no se cumplan, puede transformar su variable o variables a una escala inferior y consultar el esquema para buscar la parte que corresponda a la variable transformada.

Siguiendo el esquema, llegará a una medida de síntesis o a una estimación puntual útil para sus datos, que muchas veces va seguida de una clasificación general de las pruebas estadísticas. Al final de los esquemas encontrará el nombre de las técnicas que se emplean más frecuentemente para las pruebas de significación estadística y para la construcción de los intervalos de confianza de datos como los suyos.

Cuando utilice el esquema observe que:

1. Las medidas subrayadas con una sola línea son estimaciones muestrales puntuales.
2. Las técnicas subrayadas con una línea doble se usan para realizar pruebas de significación estadística o para construir intervalos de confianza.
3. El tipo de las pruebas se indica con líneas horizontales por encima y por debajo.
4. La palabra "o" indica que cualquiera de las pruebas mencionadas es aceptable para responder a la misma pregunta; sin embargo, la prueba situada en primer lugar tiene más potencia estadística o se usa más frecuentemente, o ambas cosas a la vez.
5. Otras condiciones que es necesario cumplir para utilizar una determinada técnica estadística aparecen sin líneas por encima o por debajo.

[1] Recuerde que, para fines estadísticos, una variable nominal solo se refiere a dos categorías de una característica. Si la característica tiene k categorías, se necesitarán k-1 variables nominales.

[2] El término interés en la posición se usa para el análisis bivariante y multivariante, así como para el univariante, en el cual dicho término tiene un significado más intuitivo. En el análisis bivariante y multivariante, nos interesa disponer de una medida que sitúe la fuerza de una relación o la magnitud de una diferencia en una serie de valores posibles.

6. Cuando aparece una coma entre dos pruebas de significación estadística, la primera prueba se usa para evaluar una hipótesis nula general y la segunda, para evaluar las comparaciones por pares.

La primera forma de utilizar el esquema es aplicable a las personas que están interesadas en seleccionar una prueba estadística para analizar un conjunto de datos. Por otra parte, como lectores de la literatura médica, lo que nos interesa más a menudo es comprobar si la prueba seleccionada por otros es apropiada. En este caso, el esquema puede utilizarse como una ayuda para encontrar el nombre de la prueba seleccionada y, siguiendo el esquema hacia atrás, determinar si la prueba es una elección lógica para los datos analizados.

FIGURA 30-1. Esquema principal para determinar cuál de los esquemas que siguen son aplicables a un conjunto de datos determinados. Los números de la parte inferior indican los esquemas que se deben utilizar.

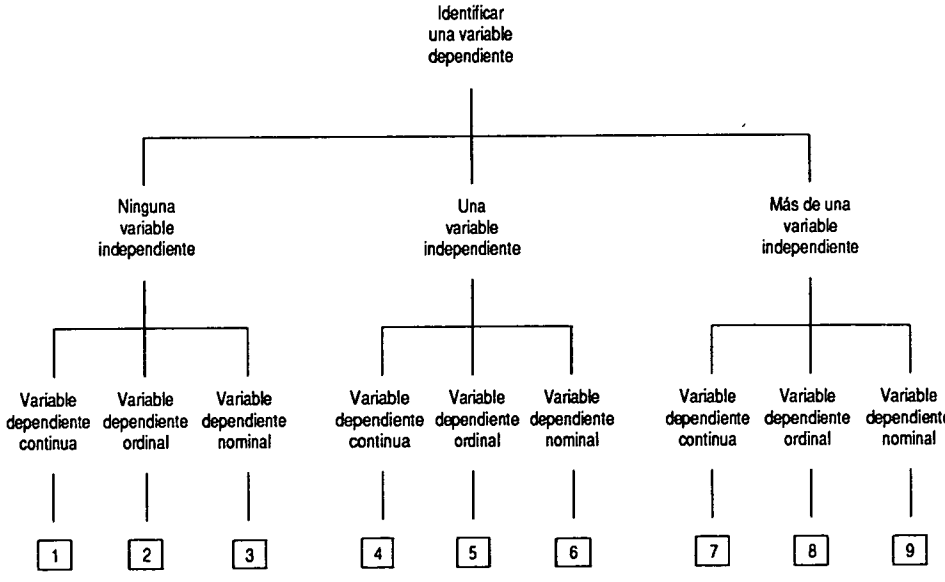


FIGURA 30-2. Esquema para seleccionar una técnica estadística univariante para una variable dependiente continua.

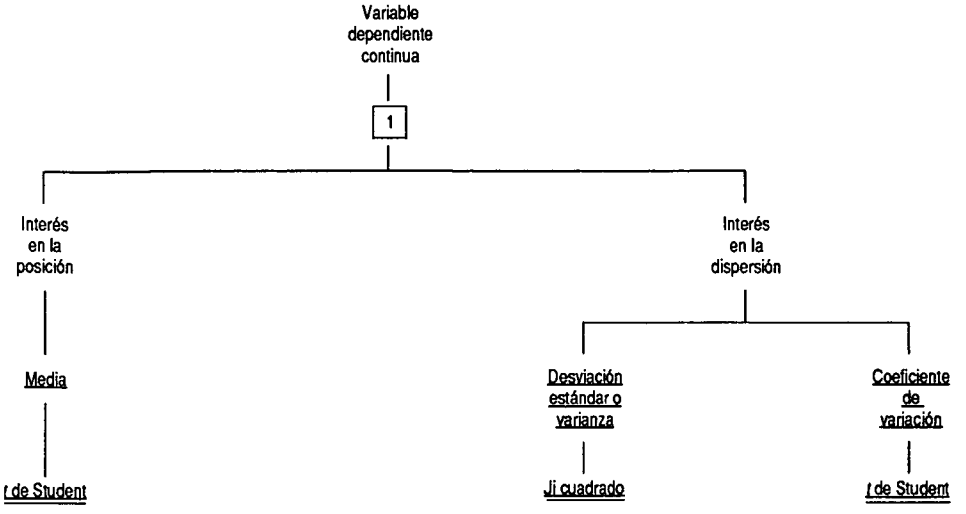
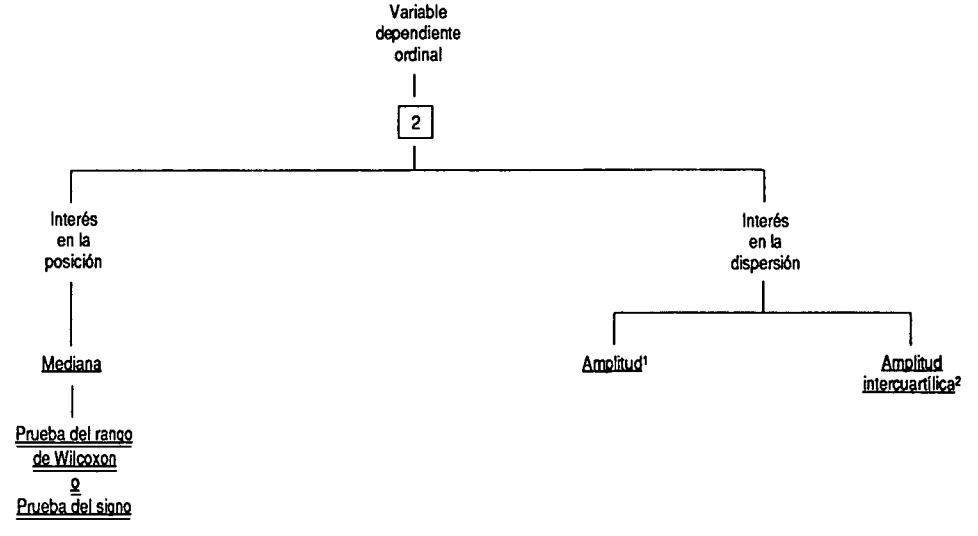


FIGURA 30-3. Esquema para seleccionar una técnica estadística univariante para una variable dependiente ordinal.



[1] La amplitud se incluye aquí solo por su extendido uso. Sin embargo, es difícil interpretarla, como se comentó en el capítulo 27.

[2] Las pruebas de significación estadística y los intervalos de confianza no se aplican a la amplitud intercuartílica, excepto cuando esta se emplea como aproximación a la desviación estándar.

FIGURA 30-4. Esquema para seleccionar una técnica estadística univariante para una variable dependiente nominal.

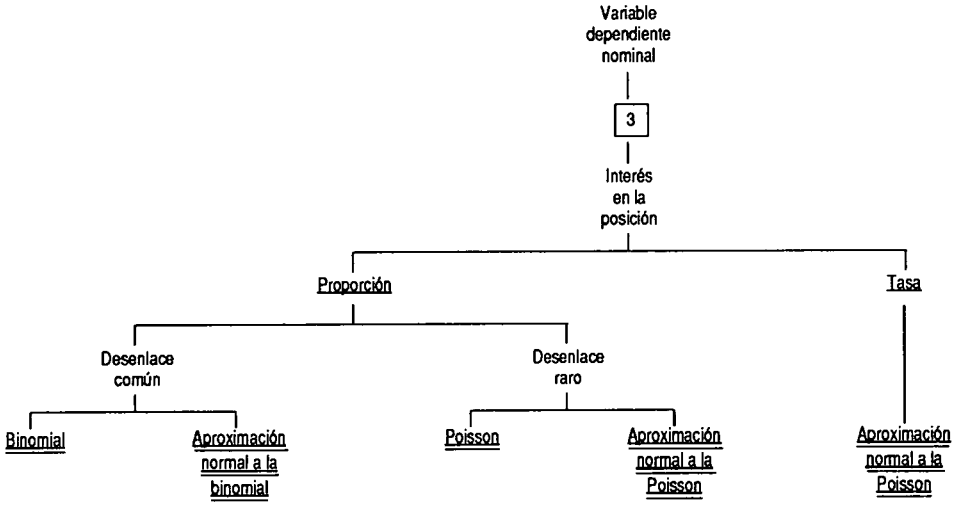


FIGURA 30-5. Esquema para seleccionar una prueba estadística bivariante para una variable dependiente continua.

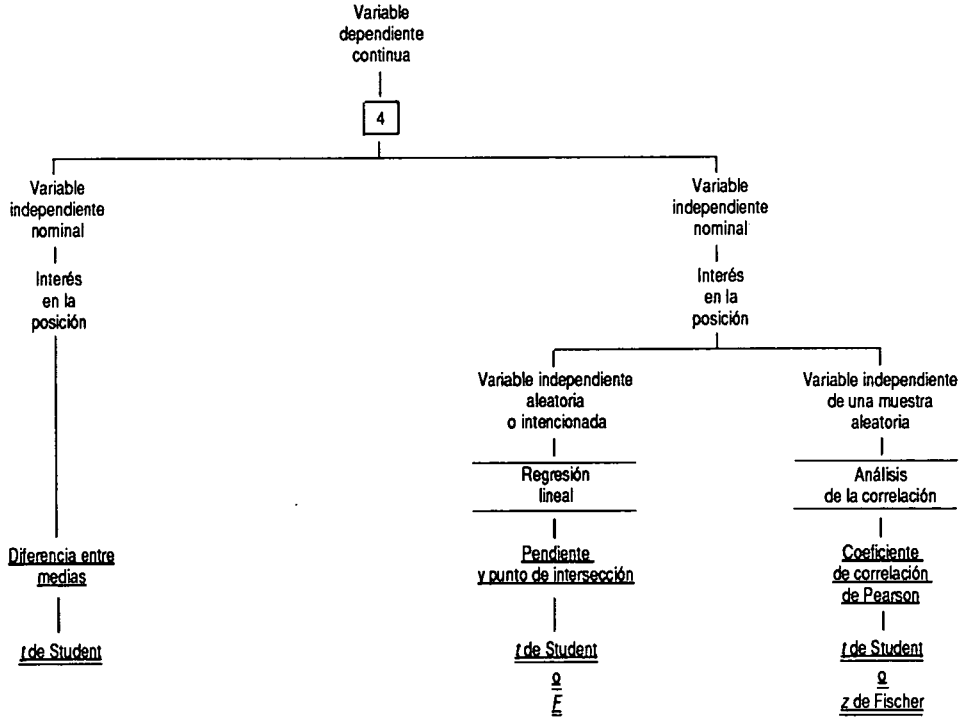


FIGURA 30-6. Esquema para seleccionar una técnica estadística bivariante para una variable dependiente ordinal.

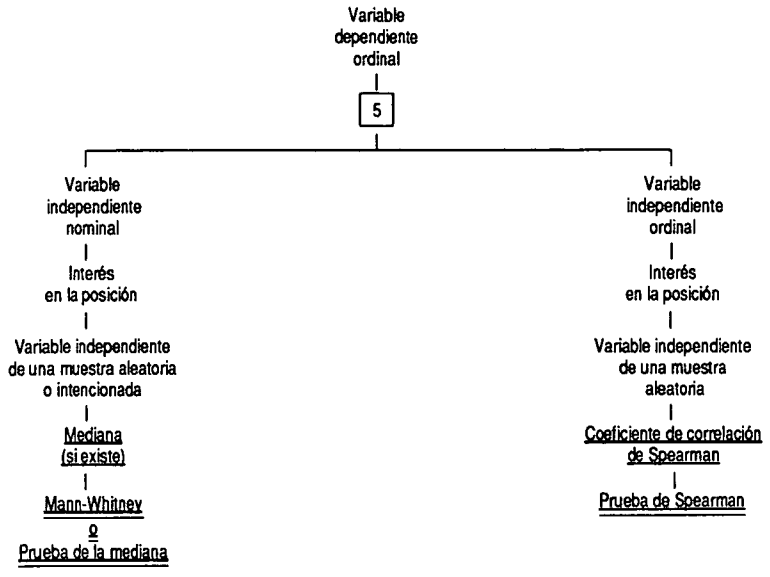


FIGURA 30-7. Esquema para seleccionar una técnica estadística bivariante para una variable dependiente nominal.

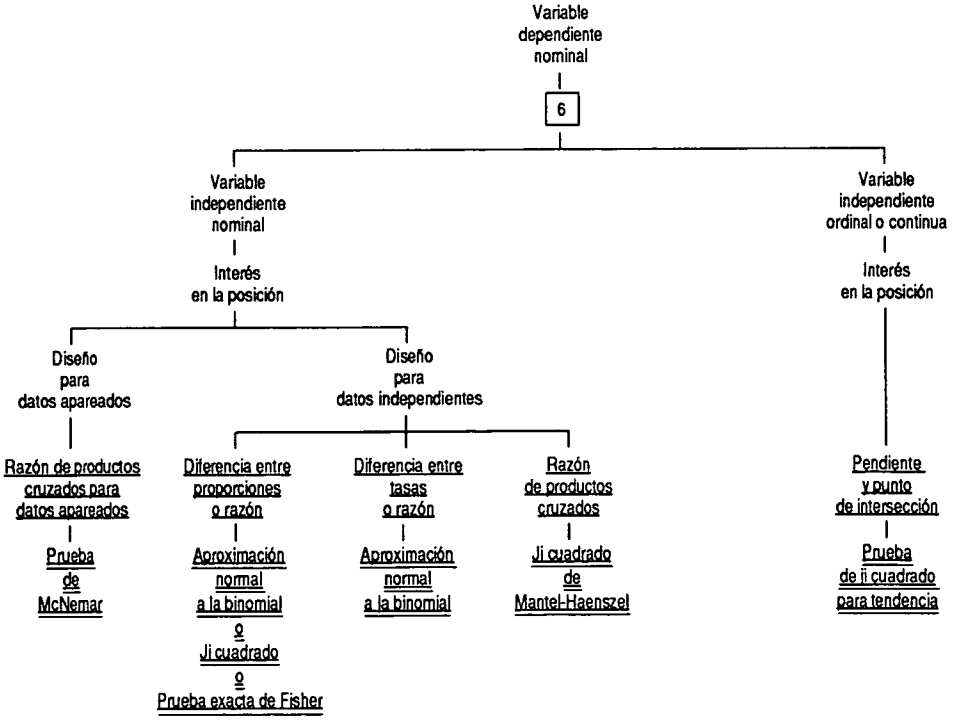


FIGURA 30-8. Esquema para seleccionar una técnica estadística multivariante para una variable dependiente continua.

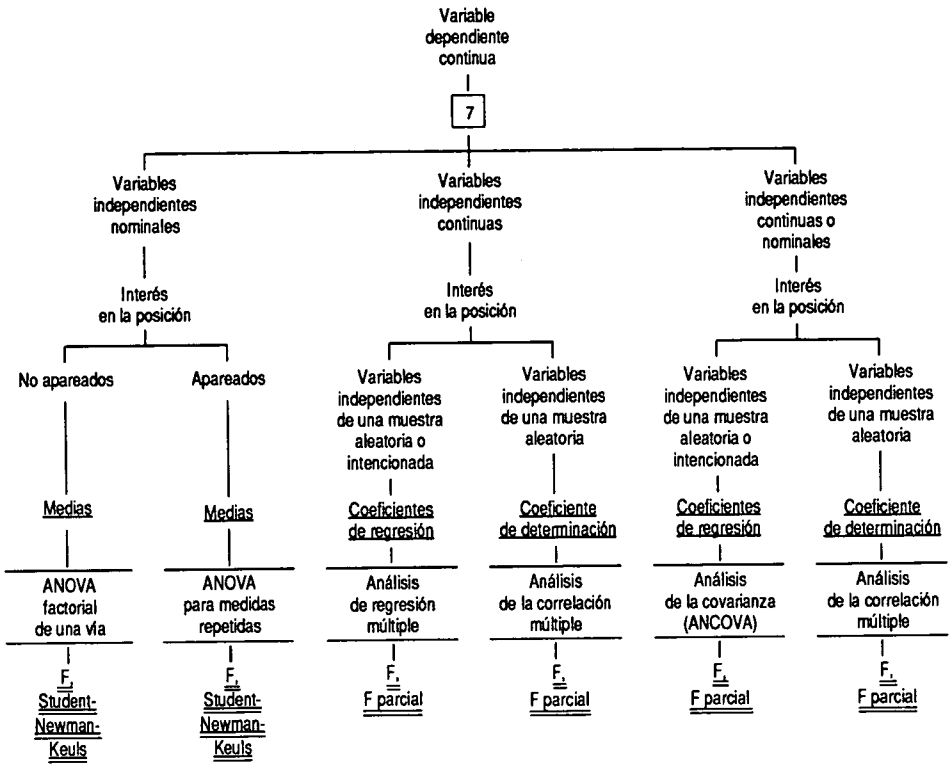


FIGURA 30-9. Esquema para seleccionar una técnica estadística multivariante para una variable dependiente ordinal.

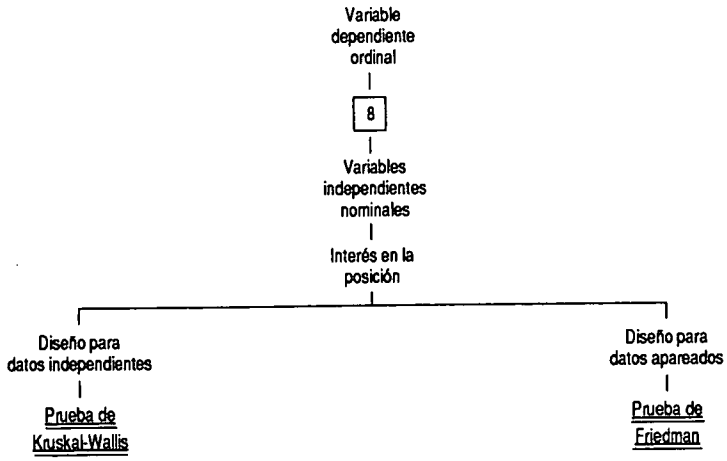
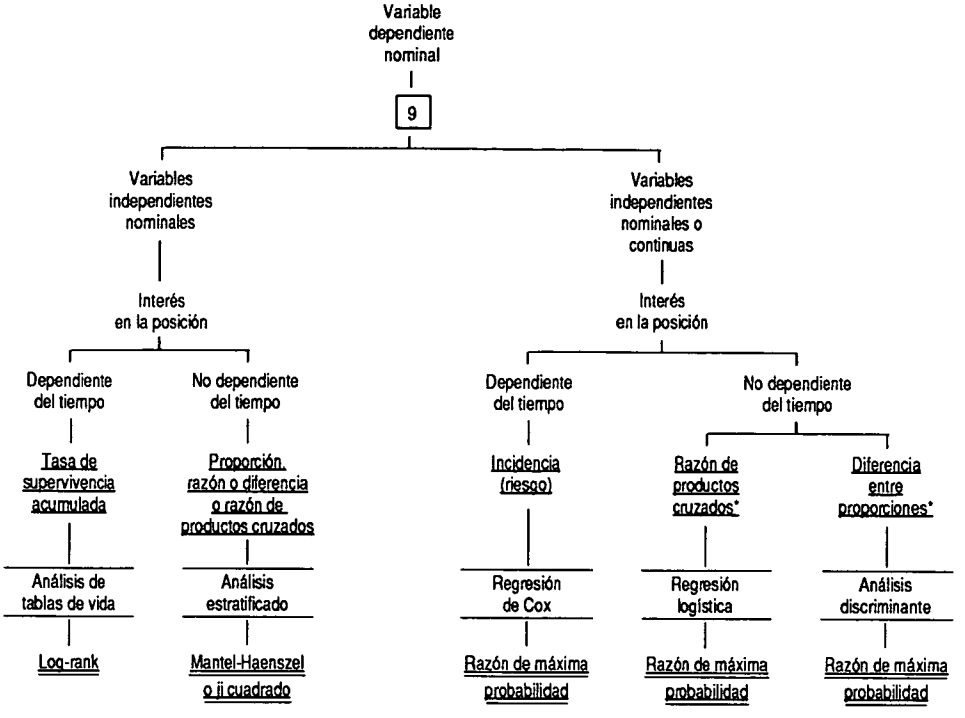


FIGURA 30-10. Esquema para seleccionar una técnica estadística multivariante para una variable dependiente nominal (* = véase la discusión de los métodos para elegir estimaciones puntuales en el capítulo 29).



GLOSARIO

AJUSTE (Adjustment) Conjunto de técnicas que se emplean después de la recogida de datos para controlar o tener en cuenta el efecto de las variables de confusión, sean conocidas o potenciales (sinónimos de ajustar: controlar, tener en cuenta, estandarizar).

ANÁLISIS (Analysis) Comparación del desenlace del grupo de estudio con el del grupo de control o testigo.

APAREAMIENTO (Matching) Análisis conjunto de dos o más observaciones realizadas en el mismo individuo o en individuos similares. El apareamiento por dúos (o pares) es un tipo especial de apareamiento en el que dos observaciones se analizan conjuntamente.

APAREAMIENTO POR DÚOS (O PARES) (Pairing) Forma especial de agrupación en la cual cada individuo del grupo de estudio se empareja con uno del grupo de control y se comparan sus desenlaces. Cuando se emplea este tipo de agrupación, se deben utilizar pruebas estadísticas especiales. Estas técnicas pueden aumentar la potencia estadística del estudio.

APAREAMIENTO POR GRUPOS (Group matching) Método de apareamiento empleado durante el proceso de asignación en una investigación en la cual los individuos del grupo de estudio y del grupo de control se seleccionan de tal forma que la distribución de cierta variable o variables sea prácticamente idéntica en ambos (sinónimo: apareamiento por frecuencias).

ASIGNACIÓN (Assignment) Selección de individuos para los grupos de estudio y de control.

ASIGNACIÓN A CIEGAS (Blind assignment) Proceso mediante el cual los individuos se asignan al grupo de estudio o al de control sin que ni ellos ni el investigador sepan a cuál grupo se asignan. Cuando los sujetos y el investigador están "cegados", el estudio a veces se denomina *estudio doble ciego*.

ASIGNACIÓN AL AZAR (Randomization) Método de asignación en el cual los individuos tienen una probabilidad conocida, aunque no necesariamente igual, de ser asignados a un grupo determinado, sea el de estudio o el de control. Se diferencia de la selección al azar en que los individuos que se asignan pueden ser o no representativos de la población (sinónimo: asignación aleatoria).

ASOCIACIÓN (Association) Relación entre dos o más características u otras medidas, que es más intensa de lo que se esperaría solamente por azar. Cuando se usa para establecer el primer criterio de causa contribuyente, la asociación implica que las dos características aparecen en el mismo individuo con más frecuencia de la esperada exclusivamente por azar.

ASOCIACIÓN DE GRUPO (Group association) Situación en la que una característica y una enfermedad se presentan más frecuentemente en un grupo de individuos que en otro. La asociación de grupo no implica necesariamente que los individuos que presentan dicha característica sean los mismos que tienen la enfermedad (sinónimo: asociación ecológica, correlación ecológica).

CAMBIOS O DIFERENCIAS POR ARTEFACTOS (Artifactual differences or changes) Cambios o diferencias entre las medidas de la frecuencia de un fenómeno que son consecuencia de la forma en que se mide, busca o define la enfermedad.

CASOS Y CONTROLES (Case-control) Estudio que se inicia con la identificación de los individuos que tienen la enfermedad (casos) y los individuos que no la tienen (con-

troles o testigos). Los casos y los controles se identifican desconociendo si estuvieron o no expuestos individualmente a los factores que se desea investigar. Estos factores se determinan a partir de la información existente (sinónimo: retrospectivo).

CAUSA CONTRIBUYENTE (Contributory cause) Se afirma que una causa es contribuyente cuando se cumplen las siguientes condiciones: 1) existe una asociación entre la causa y el efecto, 2) la causa precede al efecto en el tiempo, y 3) al alterar la causa, se modifica la probabilidad de que aparezca el efecto.

CAUSA DIRECTA (Direct cause) La causa contribuyente directa más conocida de la enfermedad (por ejemplo, el virus de la hepatitis B es una causa directa de la hepatitis B, mientras que las jeringas contaminadas son una causa indirecta). La causa directa depende de los conocimientos actuales y puede cambiar cuando se descubren mecanismos más inmediatos.

CAUSA INDIRECTA (Indirect cause) Causa contribuyente que actúa a través de un mecanismo biológico que está más estrechamente relacionado con la enfermedad que con la causa directa (por ejemplo, las agujas contaminadas son una causa contribuyente indirecta de la hepatitis B, mientras que el virus de la hepatitis B es una causa contribuyente directa) (véase causa directa).

CAUSA NECESARIA (Necessary cause) Una característica cuya presencia se requiere para producir o causar la enfermedad.

CAUSA SUFICIENTE (Sufficient cause) Una característica es una causa suficiente si su presencia produce o causa la enfermedad.

COEFICIENTE DE CORRELACIÓN (Correlation coefficient) Estadístico utilizado para estudiar la fuerza de una asociación entre dos variables, cada una de las cuales se ha extraído por muestreo de la población de interés mediante un método representativo o aleatorio.

COEFICIENTE DE DETERMINACIÓN (Coefficient of determination) (R^2) Cuadrado del coeficiente de correlación. Este estadístico indica la proporción de la variabilidad de una variable (la variable dependiente), que es explicada conociendo un valor de una o más variables (las variables independientes).

COHORTE (Cohort) Grupo de individuos que comparten una exposición, una experiencia o una característica (véase estudio de cohorte, efecto de cohorte).

CRITERIO DE REFERENCIA (Gold standard) Criterio empleado para definir de forma inequívoca la presencia de la condición o enfermedad en estudio (sinónimo: prueba de referencia, prueba de oro).

CRITERIOS DE APOYO (Supportive criteria) Cuando no es posible establecer una causa contribuyente, se pueden utilizar otros criterios para emitir un juicio sobre la existencia de una causa contribuyente. Estos criterios son la fuerza de la asociación, la relación dosis-respuesta, la consistencia de la asociación y la plausibilidad biológica (sinónimo: criterios secundarios o accesorios).

DATOS CONTINUOS (Continuous data) Tipo de datos con un número ilimitado de valores espaciados uniformemente (por ejemplo, la tensión arterial diastólica, la colesterolemia).

DATOS NOMINALES (Nominal data) Aquellos datos que se dividen en categorías. Si los datos nominales contienen más de dos categorías, estas no se pueden ordenar (por ejemplo, la raza o el color de los ojos). Los datos nominales necesitan más de una variable nominal si existen más de dos posibles categorías.

DATOS ORDINALES (Ordinal data) Datos sobre un número limitado de categorías que tienen un orden inherente de menor a mayor. Sin embargo, los datos ordinales no predeterminan el espacio que existe entre las categorías (por ejemplo, estadios 1, 2, 3 y 4 de un cáncer).

DESENLACE (Outcome) Resultado de una investigación sobre la medición empleada

en el proceso de valoración. En los estudios de casos y controles, el desenlace es una característica previa, mientras que en los estudios de cohortes y en los ensayos clínicos controlados es un suceso futuro (sinónimo: resultado final).

DESVIACIÓN ESTÁNDAR (Standard deviation) Medida de la dispersión de los datos empleada habitualmente. El cuadrado de la desviación estándar se denomina varianza (sinónimo: desviación típica).

DISTRIBUCIÓN (Distribution) Frecuencias absolutas o relativas de todos los posibles valores de una característica. Las poblaciones y las distribuciones muestrales se pueden describir matemática o gráficamente. Uno de los objetivos de la estadística es estimar parámetros de las distribuciones poblacionales.

DISTRIBUCIÓN GAUSIANA (Gaussian distribution) Una distribución de los datos que se supone en numerosas pruebas estadísticas. La distribución gausiana se representa en una curva simétrica, continua y acampanada, en la cual el valor de la media corresponde al punto más alto (sinónimo: distribución normal).

DISTRIBUCIÓN NORMAL (Normal distribution) Véase distribución gausiana.

EFECTIVIDAD (Effectiveness) Grado en que un tratamiento produce un efecto beneficioso cuando se administra bajo las condiciones habituales de la atención clínica a un grupo concreto de pacientes.

EFECTO (Effect) Un desenlace o resultado que es producido, al menos en parte, por un factor etiológico conocido como causa.

EFECTO DE COHORTE (Cohort effect) Aquel cambio en las tasas que puede ser explicado por la experiencia o la característica que comparte un grupo o cohorte de individuos. La existencia de un efecto de cohorte implica que las tasas actuales no se pueden extrapolar directamente al futuro.

EFECTO DE LA OBSERVACIÓN (Effect of observation) Tipo de sesgo que se origina cuando el mero proceso de observación modifica el desenlace del estudio.

EFICACIA (Efficacy) Grado en que un tratamiento produce un efecto beneficioso cuando se valora bajo las condiciones ideales de una investigación. La eficacia es al tratamiento lo que la causa contribuyente es a la etiología de la enfermedad.

ENSAYO CLÍNICO ALEATORIO (Randomized clinical trial) Véase ensayo clínico controlado.

ENSAYO CLÍNICO CONTROLADO (Controlled clinical trial) Investigación en la que el investigador asigna los individuos al grupo de estudio o al de control empleando un proceso conocido como asignación al azar (sinónimo: ensayo clínico aleatorio, estudio experimental).

ERROR DE MUESTREO (Sampling error) Error introducido por las diferencias debidas al azar entre la estimación obtenida en la muestra y el valor verdadero en la población de la que se ha extraído dicha muestra. El error de muestreo es inherente al uso de métodos de muestreo, y el error estándar cuantifica su magnitud.

ERROR DE TIPO I (Type I error) Error que se comete cuando los datos indican un resultado estadísticamente significativo a pesar de que no existe una verdadera asociación o diferencia en la población. El nivel alfa es el tamaño del error de tipo I tolerado, habitualmente, de 5%.

ERROR DE TIPO II (Type II error) Error que se comete cuando con las observaciones muestrales no se consigue demostrar la existencia de una significación estadística, a pesar de que existe una asociación o diferencia verdadera en la población.

ERROR DEL INSTRUMENTO (Instrument error) Un sesgo en la valoración que se produce cuando el instrumento de medida no es apropiado para las condiciones del estudio o no es suficientemente exacto para medir el desenlace o resultado final del estudio.

ERROR ESTÁNDAR (Standard error) Grado de dispersión de las estimaciones puntuales obtenidas en muestras de un tamaño determinado.

ESPECIFICIDAD (Specificity) Proporción de sujetos sin la enfermedad, según la prueba de referencia, que obtienen resultados negativos en la prueba que se estudia (sinónimo: negativo para la enfermedad).

ESTADÍSTICO (Statistic) Valor calculado a partir de los datos de una muestra y utilizado para estimar un valor o parámetro de la población de la que se ha extraído dicha muestra (sinónimos: valor muestral, estadígrafo).

ESTANDARIZACIÓN (Standardization) DE UNA TASA (Of a rate) Proceso que permite tomar en cuenta o ajustar los datos según los efectos de un factor como la edad o el sexo sobre las tasas calculadas (véase ajuste).

ESTIMACIÓN (Estimate) Un valor o intervalo de valores calculados a partir de una muestra de observaciones que se emplea como aproximación al valor correspondiente en la población, es decir, al parámetro (véase también estimación por intervalo, estimación puntual).

ESTIMACIÓN POR INTERVALO (Interval estimate) Véase intervalo de confianza.

ESTIMACIÓN PUNTUAL (Point estimate) Valor único calculado a partir de las observaciones muestrales que se utiliza como estimación del valor poblacional o parámetro.

ESTRATIFICACIÓN (Stratification) En general, por estratificación se entiende la división en grupos. El mismo término se puede utilizar para hacer referencia al proceso de control según las diferencias entre los factores de confusión, mediante la obtención de estimaciones separadas para los grupos de individuos que tienen los mismos valores de la variable de confusión. La estratificación también puede referirse a un método de muestreo intencionado y diseñado para representar en exceso categorías poco frecuentes de una variable independiente.

ESTUDIO DE COHORTE (Cohort study) Estudio que se inicia con la identificación de individuos con y sin el factor que se va a investigar. Estos factores se determinan sin saber cuáles individuos padecen o padecerán la enfermedad. Los estudios de cohortes pueden ser concurrentes o no concurrentes (sinónimo: prospectivo).

ESTUDIO DE COHORTE CONCURRENTE (Concurrent cohort study) Estudio de cohorte en el que la asignación de un sujeto al grupo de estudio o al de control se determina al iniciar la investigación y en el que se sigue la evolución de los individuos de ambos grupos para determinar si desarrollan la enfermedad (sinónimo: estudio de cohorte prospectivo).

ESTUDIO DE COHORTE NO CONCURRENTE (Noncurrent cohort study) Estudio de cohorte en el cual la asignación de un individuo a un grupo se determina a partir de la información existente en el momento en que se inicia el estudio. El estudio de cohorte no concurrente extremo es aquel en el cual el desenlace se determina a partir de los registros existentes (sinónimo: estudio de cohorte retrospectivo).

ESTUDIO PROSPECTIVO (Prospective study) Véase estudio de cohorte.

ESTUDIO RETROSPECTIVO (Retrospective study) Véase casos y controles.

ESTUDIO TRANSVERSAL (Cross-sectional study) Estudio que identifica en el mismo momento a los individuos con y sin la condición o la enfermedad en estudio y la característica o exposición de interés.

EXACTITUD (Accuracy) Capacidad de una prueba para producir resultados que se aproximen al verdadero valor del fenómeno; falta de error sistemático o aleatorio; precisión sin sesgo.

EXPERIMENTO NATURAL (Natural experiment) Investigación en la cual la modificación de un factor de riesgo se produce en un grupo de individuos, pero no en un

grupo de control. A diferencia del ensayo clínico controlado, en el experimento natural la modificación no es producida por la intervención del investigador.

EXTRAPOLACIÓN (Extrapolation) Conclusiones sobre el significado del estudio para una población objetivo compuesta por individuos o datos no representados en la muestra estudiada.

FACTOR DE RIESGO (Risk factor) Característica o factor que se ha observado que está asociado con un aumento de la probabilidad de que aparezca una enfermedad. Un factor de riesgo no implica necesariamente la existencia de una relación de causa-efecto. En este libro, el factor de riesgo implica que al menos se ha establecido una asociación a nivel individual.

FALACIA ECOLÓGICA (Ecological fallacy) Tipo de error que puede cometerse cuando a partir de una asociación a nivel de grupo se deduce una relación inexistente a nivel individual.

GRADOS DE LIBERTAD (Degrees of freedom) Parámetro de muchas distribuciones estadísticas estándares. Los grados de libertad permiten tomar en cuenta el número de parámetros poblacionales que se deben estimar en una muestra para poder aplicar ciertas pruebas estadísticas.

GRUPO DE CONTROL (Control group) Grupo de personas que se selecciona para comparación con el grupo de estudio. Idealmente, el grupo de control es idéntico al de estudio excepto en que no posee la característica estudiada o no ha sido expuesto al tratamiento que se investiga (sinónimo: grupo de referencia o grupo testigo).

GRUPO DE ESTUDIO (Study group) En un estudio de cohortes o en un ensayo clínico controlado, este es el grupo de individuos que posee las características o está expuesto a los factores estudiados. En los estudios de casos y controles o en los transversales, corresponde al grupo de individuos que padecen la enfermedad investigada.

GRUPO DE REFERENCIA (Reference group) Grupo de individuos, presuntamente sanos, del que se extrae una muestra de sujetos en los que se realizarán mediciones para establecer un intervalo de normalidad (sinónimo: población de referencia).

HIPÓTESIS DE ESTUDIO (Study hypothesis) Afirmación de la existencia de una asociación entre dos o más variables en la población de la que procede la muestra. Una hipótesis de estudio es unilateral cuando solo considera las asociaciones en una dirección; es bilateral cuando no se especifica la dirección de la asociación.

HIPÓTESIS NULA (Null hypothesis) Afirmación de que no existe una asociación o diferencia verdadera entre las variables en la población de la que se extrajo la muestra estudiada.

INFERENCIA (Inference) En términos estadísticos, la inferencia es el proceso lógico que tiene lugar durante las pruebas de significación estadística (véase prueba de significación estadística).

INTERPRETACIÓN (Interpretation) Extracción de conclusiones sobre el significado de cualquier diferencia observada entre el grupo de estudio y el de control incluidos en la investigación.

INTERVALO DE CONFIANZA DE 95% (Confidence interval) (95%) En términos estadísticos, es el intervalo de valores numéricos en el que se encuentra el valor poblacional que se está estimando con un nivel de confianza de 95% (sinónimo: estimación por intervalo).

INTERVALO DE NORMALIDAD (Range of normal) Medida del intervalo de valores obtenidos en una prueba correspondientes a los sujetos que no padecen la enfermedad. Con frecuencia, hace referencia al 95% de los valores centrales o a la media de los valores de los individuos sin la enfermedad, más o menos dos desviaciones es-

tándares (sinónimo: valores normales).

LETALIDAD (Case fatality) Número de muertes causadas por una determinada enfermedad dividido por el número de personas diagnosticadas de esta enfermedad al inicio del período de estudio. La letalidad es una estimación de la probabilidad de morir a consecuencia de la enfermedad. La tasa de letalidad incluye el número de años-persona como unidad de tiempo en el denominador.

MEDIA (Mean) Suma de todas las mediciones dividida por el número total de valores sumados. "Centro de gravedad" de la distribución de las observaciones. Forma especial de promedio.

MEDIANA (Median) Punto medio de la distribución. La mediana es el valor que deja la mitad de los valores por arriba y la otra mitad por debajo.

MÉTODO DE LA TABLA DE VIDA (Life-table method) Método para organizar los datos que permite examinar la experiencia de uno o más grupos de individuos durante un intervalo de tiempo cuando la evolución de algunos individuos se sigue durante períodos más prolongados que la de otros (sinónimo: Kaplan-Meier, tablas de vida de Cutler-Ederer, tablas de vida de cohortes o clínicas). Estas tablas son distintas de las transversales y actuales, las cuales permiten calcular la esperanza de vida).

MUESTRA (Sample) Subgrupo de una población obtenido por un investigador para extraer conclusiones o para realizar estimaciones sobre la población.

MUESTRA ALEATORIA (Naturalistic sample) Un grupo de observaciones obtenidas de una población de forma tal que la distribución muestral de los valores de la variable independiente es representativa de su distribución en la población (sinónimo: muestra al azar).

MUESTRA FORTUITA (Chunk sample) Muestra que se extrae de una población por lo fácil que resulta obtener datos de ella, sin tener en cuenta el grado en que es aleatoria o representativa de dicha población.

MUESTRA INTENCIONADA (Purposive sample) Grupo de observaciones obtenidas a partir de una población de forma tal que el investigador determina la distribución muestral de los valores de la variable independiente sin que sea necesariamente representativa de su distribución en la población.

NEGATIVO FALSO (False-negative) Individuo cuyo resultado en una prueba es negativo, pero que tiene la enfermedad según la prueba de referencia.

NEGATIVO VERDADERO (True–negative) Individuo que no padece la enfermedad según la prueba de referencia y obtiene resultados negativos en la prueba estudiada.

NO SESGADO (Unbiased) Sin error sistemático asociado.

NÚMERO DE PACIENTES QUE SE DEBEN TRATAR (Number needed to treat) Valor recíproco de la diferencia de riesgos. Es el número de pacientes similares a los pacientes estudiados que sería necesario tratar para conseguir un desenlace positivo más o un desenlace negativo menos.

PARÁMETRO (Parameter) Valor que sintetiza la distribución de una población. Uno de los objetivos del análisis estadístico consiste en estimar los parámetros poblacionales a partir de las observaciones muestrales (sinónimo: valor poblacional).

POBLACIÓN (Population) Grupo numeroso compuesto con frecuencia, pero no necesariamente, por individuos. En estadística, el objetivo es extraer conclusiones acerca de una o más poblaciones mediante la obtención de subgrupos o muestras compuestos por individuos pertenecientes a la población.

POBLACIÓN OBJETIVO (Target population) Grupo de individuos a los que se desea extrapolar o aplicar los resultados de una investigación. La población objetivo puede ser, y de hecho lo es frecuentemente, distinta de la población de la que se extrae la muestra en una investigación.

POSITIVO FALSO (False-positive) Individuo cuyo resultado en una prueba es positivo, pero que no tiene la enfermedad según la prueba de referencia.

POSITIVO VERDADERO (True-positive) Individuo que padece la enfermedad según la prueba de referencia y obtiene resultados positivos en la prueba estudiada.

POTENCIA (Power) Capacidad de un estudio para demostrar significación estadística, cuando existe una diferencia o una asociación verdadera de una fuerza determinada en la población de la que se ha extraído la muestra (sinónimo: poder estadístico, poder de resolución).

PRECISO (Precise) Sin error aleatorio asociado (una medición imprecisa puede desviarse del valor numérico verdadero en cualquier dirección).

PREVALENCIA (Prevalence) Proporción de individuos con una enfermedad determinada en un momento dado. La prevalencia también puede interpretarse como la probabilidad de que un individuo elegido al azar de una población tenga la enfermedad (sinónimo: probabilidad anterior a la prueba).

PROBABILIDAD (Probability) Proporción en la cual el numerador es el número de veces que ocurre un suceso y el denominador, ese mismo número sumado al número de veces que no ocurre ese suceso.

PROPORCIÓN (Proportion) Fracción cuyo numerador está formado por un subgrupo de individuos incluido en el denominador.

PRUEBA BILATERAL (Two-tailed test) Prueba de significación estadística en la que se toman en cuenta las desviaciones de la hipótesis nula en cualquier dirección. El uso de una prueba bilateral implica que el investigador deseaba considerar las desviaciones en cualquier dirección antes de recoger los datos (sinónimo: prueba de dos colas).

PRUEBA DE SIGNIFICACIÓN ESTADÍSTICA (Statistical significance test) Técnica estadística para calcular la probabilidad de que la asociación observada en una muestra hubiera podido ocurrir por azar si no existiera esa asociación en la población origen (sinónimo: inferencia, contraste de hipótesis).

PRUEBA UNILATERAL (One tailed test) Prueba de significación estadística en la cual solo se toman en cuenta las desviaciones respecto de la hipótesis nula en una sola dirección. El empleo de una prueba unilateral implica que el investigador no considera posible una desviación verdadera en dirección opuesta (sinónimo: prueba de una cola).

PRUEBAS ESTADÍSTICAS NO PARAMÉTRICAS (Nonparametric statistics) Pruebas estadísticas en las que no existen supuestos sobre la distribución de los parámetros en la población de la que se extrajo la muestra. En estas pruebas se aceptan otros supuestos, como el relativo a la aleatoriedad del muestreo. Se aplican con mayor frecuencia a los datos ordinales, si bien pueden emplearse también para analizar datos continuos transformados a una escala ordinal (sinónimo: pruebas de distribución libre).

RAZÓN (Ratio) Fracción en la cual el numerador no es necesariamente un subconjunto del denominador, como ocurre en la proporción.

RAZÓN DE MORTALIDAD ESTANDARIZADA (Standardized mortality ratio) Fracción cuyo numerador es el número de muertes observadas y cuyo denominador corresponde al número de muertes esperables sobre la base de una población de referencia. La razón de mortalidad estandarizada implica que para ajustar los datos según los factores de confusión se ha empleado la estandarización indirecta. Los términos *razón de mortalidad estandarizada* y *razón de mortalidad proporcional* no son sinónimos.

RAZÓN DE MORTALIDAD PROPORCIONAL (Proportionate mortality ratio) Fracción cuyo numerador está formado por el número de personas que mueren de una enferme-

dad concreta durante un período determinado y cuyo denominador es el número de individuos fallecidos por todas las causas en el mismo período.

RAZÓN DE PRODUCTOS CRUZADOS (Odds ratio) Medida de la fuerza o del grado de una asociación aplicable a todos los tipos de estudios que utilizan datos nominales, pero que habitualmente se aplica a los estudios de casos y controles y a los estudios transversales. Se calcula como el cociente del número de sujetos expuestos al factor de riesgo respecto al de los no expuestos entre los que presentan la enfermedad, dividido por el cociente del número de sujetos expuestos al factor de riesgo respecto al de los no expuestos cuando no está presente la enfermedad (sinónimos: razón de ventajas, desigualdad relativa, razón de momios).

RECORRIDO (Range) Diferencia entre los valores máximo y mínimo de una población o de una muestra (sinónimo: amplitud).

REGRESIÓN A LA MEDIA (Regression to the mean) Principio estadístico que indica que es improbable que los sucesos infrecuentes vuelvan a suceder. Es más probable, solo por azar, que las mediciones siguientes a un resultado infrecuente sean más cercanas a la media. Además, es posible que existan factores psicológicos o sociales que contribuyan a forzar a los sucesos posteriores a "regresar" hacia valores más cercanos a la media.

RELACIÓN DOSIS-RESPUESTA (Dose-response relationship) Una relación dosis-respuesta está presente cuando los cambios en los niveles de una exposición están asociados de forma consistente en una dirección con los cambios en la frecuencia del desenlace. La existencia de una relación dosis-respuesta es un criterio que apoya el que una causa sea contribuyente.

REPRODUCIBILIDAD (Reproducibility) Capacidad de una prueba para producir resultados consistentes cuando se repite en condiciones similares y se interpreta sin conocimiento de los resultados previos (sinónimo: fiabilidad, repetibilidad).

RIESGO (Risk) Probabilidad de que ocurra un suceso durante un período determinado. El numerador del riesgo es el número de individuos en los que aparece la enfermedad durante dicho período, mientras que el denominador es el número de sujetos sin la enfermedad al inicio del período.

RIESGO ABSOLUTO (Absolute risk) La probabilidad de que ocurra un suceso durante un período determinado. Si no está presente el factor de riesgo, el riesgo absoluto es igual al riesgo relativo multiplicado por la probabilidad media del suceso durante el mismo período.

RIESGO ATRIBUIBLE POBLACIONAL PORCENTUAL (Population attributable risk percentage) Porcentaje del riesgo en la comunidad, *incluidos los individuos expuestos al factor de riesgo y los no expuestos*, asociado con la exposición al factor de riesgo. El riesgo atribuible poblacional no implica necesariamente una relación de causa-efecto (sinónimos: fracción atribuible (poblacional), proporción atribuible (población), fracción etiológica (población)).

RIESGO ATRIBUIBLE PORCENTUAL (Attributable risk percentage) Porcentaje del riesgo *entre aquellos individuos expuestos al factor de riesgo* que está asociado con dicho factor. Si existe una relación de causa–efecto, el riesgo atribuible es el porcentaje de la frecuencia de la enfermedad que se esperaría que disminuyera entre los expuestos al factor de riesgo si ese factor se pudiera suprimir completamente (sinónimos: riesgo atribuible, riesgo atribuible (en los expuestos), fracción etiológica (en los expuestos), porcentaje de reducción del riesgo, tasa de eficacia protectora).

RIESGO RELATIVO (Relative risk) Razón entre la probabilidad de que suceda un desenlace en un período determinado en los expuestos al factor de riesgo y la probabilidad de que suceda entre los no expuestos al factor de riesgo en el mismo período.

El riesgo relativo es una medida de la fuerza o del grado de asociación aplicable a los estudios de cohorte y a los ensayos clínicos aleatorios. En los de casos y controles, la razón de productos cruzados se puede utilizar frecuentemente como una aproximación al riesgo relativo.

ROBUSTO (Robust) Se dice que una prueba estadística es robusta si se pueden violar sus supuestos sin que ello repercuta sustancialmente en las conclusiones.

SELECCIÓN AL AZAR (Random selection) Método para obtener una muestra que asegura que cada individuo de la población tiene una probabilidad conocida, aunque no necesariamente igual, de ser seleccionado para formar parte de la muestra.

SENSIBILIDAD (Sensitivity) Proporción de sujetos que padecen la enfermedad, según la prueba de referencia, y obtienen resultados positivos en la prueba que se estudia (sinónimo: positivo para la enfermedad).

SESGO (Bias) Un factor que produce la desviación sistemática de un resultado en una dirección, en relación con los valores reales. El uso de este término está limitado a las desviaciones originadas por defectos en el diseño del estudio.

SESGO DE ADELANTO DIAGNÓSTICO (Lead-time bias) Diferencia entre tasas debida a artefactos que se produce cuando el tamizaje de la enfermedad conduce a un diagnóstico temprano que no mejora el pronóstico.

SESGO DE NOTIFICACIÓN (Reporting bias) Sesgo de información que se produce cuando es más probable que los individuos de un grupo declaren sucesos pasados que los de otros grupos de estudio o de control. Es muy posible que se produzca sesgo de notificación cuando un grupo está sometido a una presión desproporcionada para dar información confidencial.

SESGO DE RECUERDO (Recall bias) Sesgo de información que se produce cuando es más probable que los individuos de un grupo recuerden los sucesos pasados que los de otros grupos de estudio o de control. El sesgo de recuerdo es especialmente probable en los estudios de casos y controles que tengan que ver con enfermedades graves y en los que las características estudiadas sean sucesos frecuentes y recordados de forma subjetiva.

SESGO DE SELECCIÓN (Selection bias) Sesgo que se produce en el proceso de asignación cuando la forma como se escogen los grupos de estudio y de control determina que estos grupos difieran en uno o más de los factores que afectan al desenlace del estudio. Tipo especial de variable de confusión que surge más como consecuencia del diseño del estudio que por azar (véase variable de confusión).

TASA (Rate) Habitualmente se emplea para indicar cualquier medida de la frecuencia de una enfermedad o desenlace. Desde un punto de vista estadístico, las tasas son aquellas medidas de la frecuencia de la enfermedad que incluyen una medida de tiempo en el denominador (por ejemplo, la incidencia).

TASA DE INCIDENCIA (Incidence rate) Tasa en la cual los nuevos casos de la enfermedad se contabilizan por unidad de tiempo. La tasa de incidencia se calcula teóricamente como el número de individuos que desarrollan la enfermedad en un período determinado dividido por el número de años-persona en riesgo.

TASA DE MORTALIDAD (Mortality rate) Es una medida de la incidencia de muerte. Esta tasa se calcula dividiendo el número de muertes que han ocurrido durante un período por el producto del número de individuos y el número de unidades de tiempo del período de seguimiento.

TÉCNICAS DE REGRESIÓN (Regression techniques) Métodos estadísticos útiles para describir la asociación entre una variable dependiente y una o más variables independientes. Las técnicas de regresión se utilizan con frecuencia para ajustar el efecto según las variables de confusión.

VÁLIDO (Valid) Una medición es válida si es apropiada para la cuestión que se está investigando o si mide lo que intenta medir.

VALOR P *(P value)* Probabilidad de realizar una observación al menos tan alejada de la condición descrita en la hipótesis nula como la observada en nuestro conjunto de datos si la hipótesis nula fuera cierta. El cálculo del valor *P* constituye el "objetivo" de las pruebas de significación estadística.

VALOR PREDICTIVO DE UNA PRUEBA NEGATIVA (Predictive value of a negative test) Proporción de aquellos sujetos con resultados negativos en una prueba, que no padecen la enfermedad según la prueba de referencia. Esta medida incorpora la prevalencia de la enfermedad. Desde el punto de vista clínico, el valor predictivo de una prueba negativa es la probabilidad de que un individuo no padezca la enfermedad cuando el resultado de la prueba es negativo (sinónimo: probabilidad posterior a la prueba).

VALOR PREDICTIVO DE UNA PRUEBA POSITIVA (Predictive value of a positive test) Proporción de los sujetos con resultados positivos en una prueba, que padecen la enfermedad según la prueba de referencia. Esta medida incorpora la prevalencia de la enfermedad. Desde el punto de vista clínico, el valor predictivo de una prueba positiva es la probabilidad de que un individuo padezca la enfermedad cuando el resultado de la prueba es positivo (sinónimo: probabilidad posterior a la prueba).

VALORACIÓN (Assessment) Determinación del desenlace o resultado final en los grupos de estudio y de control.

VALORACIÓN A CIEGAS (Blind assessment) Evaluación de un desenlace o resultado final en los individuos incluidos en el estudio sin que la persona que la realiza sepa si pertenecen al grupo de estudio o al de control.

VARIABILIDAD INTEROBSERVADOR (Interobserver variability) Variabilidad en las medidas realizadas por diversos observadores.

VARIABILIDAD INTRAOBSERVADOR (Intraobserver variability) Variabilidad en las medidas realizadas por el mismo observador en distintas ocasiones.

VARIABLE (Variable) En su acepción general, variable se refiere a una característica que se mide en el estudio. En términos estadísticos rigurosos, una variable es la representante de esas mediciones en el análisis. Los datos medidos en una escala continua u ordinal se expresan por medio de una variable, como ocurre con las variables nominales que solo tienen dos categorías. Sin embargo, los datos nominales con más de dos categorías deben expresarse con más de una variable.

VARIABLE DE CONFUSIÓN (Confounding variable) Característica o variable que se distribuye de forma diferente en el grupo de estudio y en el de control y que afecta al desenlace estudiado. Una variable de confusión puede deberse al azar o a un sesgo. Cuando se debe a un sesgo en el proceso de asignación, el error resultante se denomina *sesgo de selección* (sinónimo: factor de confusión).

VARIABLE DEPENDIENTE (Dependent variable) En general, la variable del desenlace de interés en cualquier tipo de estudio. El desenlace o resultado que uno pretende explicar o estimar.

VARIABLE INDEPENDIENTE (Independent variable) Variable que se mide para determinar el valor correspondiente de la variable dependiente en cualquier tipo de estudio. Las variables independientes definen las condiciones bajo las cuales se examinará a la variable dependiente.

VARIANZA (Variance) Véase desviación estándar.

INDICE ALFABÉTICO

Cómo estudiar un estudio y probar una prueba: lectura crítica de la literatura médica ofrece al lector un enfoque activo y progresivo para revisar la literatura médica. La obra está destinada a enseñar a los estudiantes, residentes, médicos y a los trabajadores de la salud en general a leer artículos de revistas biomédicas—una fuente vital de información actual—de forma crítica y eficiente. Su principal valor estriba en que no supone que el lector tenga conocimientos previos en epidemiología, bioestadística o matemáticas.

Su estructura consta de cuatro partes independientes en las que se explica cómo evaluar los estudios, las pruebas diagnósticas, las razones y las tasas, así como las pruebas estadísticas que aparecen en los artículos de las revistas biomédicas. Diversos elementos de ayuda al aprendizaje—como resúmenes comprimidos de artículos hipotéticos, listas de preguntas para comprobación, ejercicios para detectar errores, y un esquema no matemático para evaluar las pruebas estadísticas—ayudan al lector a poner a prueba su capacidad analítica, a la vez que pueden utilizarse como guía en la lectura y discusión de los artículos científicos. Además, el libro contiene exámenes de autoevaluación y un glosario.

Cómo estudiar un estudio y probar una prueba: lectura crítica de la literatura médica constituye una herramienta esencial para los estudiantes y profesionales de la salud que usan la literatura médica para mantenerse al corriente de los rápidos cambios que caracterizan el progreso de las ciencias de la salud.

ORGANIZACION PANAMERICANA DE LA SALUD
Oficina Sanitaria Panamericana, Oficina Regional de la
ORGANIZACION MUNDIAL DE LA SALUD