

Cómo estudiar un estudio y probar una prueba: lectura crítica de la literatura médica¹

Segunda edición

Richard K. Riegelman y Robert P. Hirsch

PARTE X:

Capítulo 28. Análisis bivariantes

¹Título original: *Studying a Study and Testing a Test. How to Read the Medical Literature*. Second edition. © Richard K. Riegelman, Robert P. Hirsch. Publicado por Little, Brown and Company, Boston, Massachusetts 02108, Estados Unidos de América. Los pedidos del libro en inglés deben dirigirse a esta dirección.

Versión en español autorizada por Little, Brown and Company; se publica simultáneamente en forma de libro (Publicación Científica 531) y como serie en el *Boletín de la Oficina Sanitaria Panamericana*. Traducción de José María Borrás, revisada por el Servicio Editorial de la Organización Panamericana de la Salud.

© Little, Brown and Company, 1989. Todos los derechos reservados. Ninguna parte de esta publicación puede ser reproducida ni transmitida en ninguna forma ni por ningún medio de carácter mecánico o electrónico, incluidos fotocopia y grabación, ni tampoco mediante sistemas de almacenamiento y recuperación de información, a menos que se cuente con la autorización por escrito de Little, Brown and Company.

ANÁLISIS BIVARIANTES

En el análisis bivalente, 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 bivalente. 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.

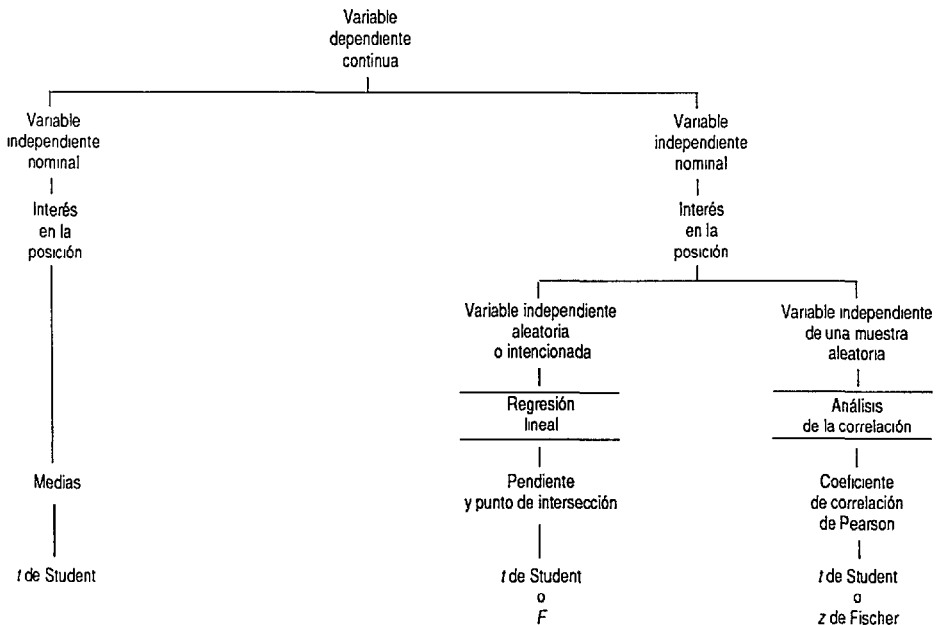
que el intervalo de confianza bivalente no se extiende más allá del cero, podríamos concluir correctamente que la prueba de significación estadística correspondiente conduciría a rechazar la hipótesis nula de que la prevalencia de la enfermedad es igual en las dos comunidades. En otras palabras, podemos 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 bivalente 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 bivalente.

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 análisis

FIGURA 28-1. Esquema para seleccionar un método estadístico bivalente 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 gaussiana, 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 muestra 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 cuadrados* (*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 cuadrados, 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

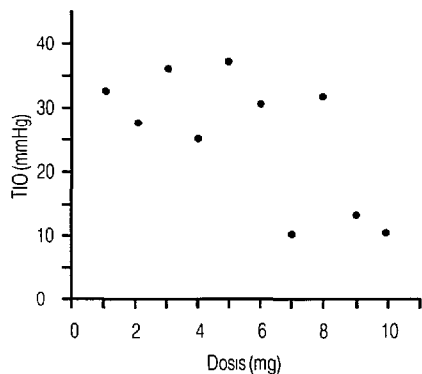
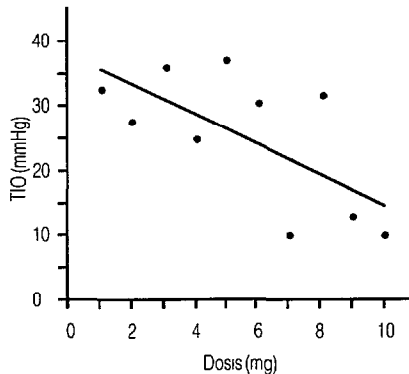


FIGURA 28-3. Regresión de la tensión intraocular (TIO) después del tratamiento con un medicamento determinado en función de la dosis



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 bivalente, 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 cuadrados 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 gaussiana 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 gaussiana en la población de la que se ha extraído la muestra aleatoria.

El segundo supuesto del análisis de regresión por mínimos cuadrados 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 gaussiana, 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 extrayéramos 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

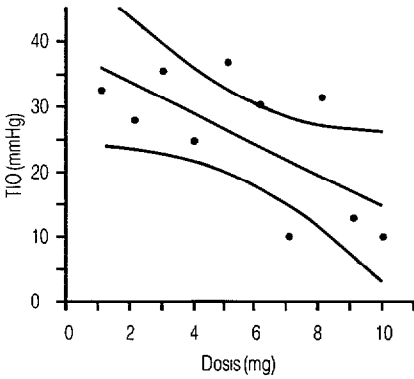
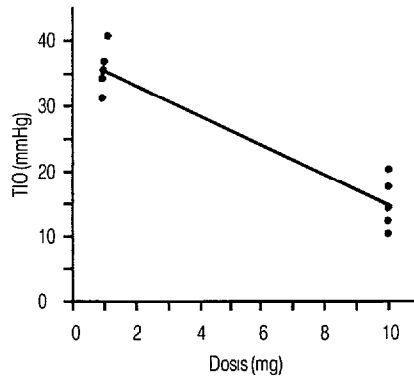


FIGURA 28-5. Regresión de la tensión intraocular (TIO) después del tratamiento con un medicamento determinado en función de la dosis cuando se administra a los pacientes una dosis de 1 mg o de 10 mg



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.

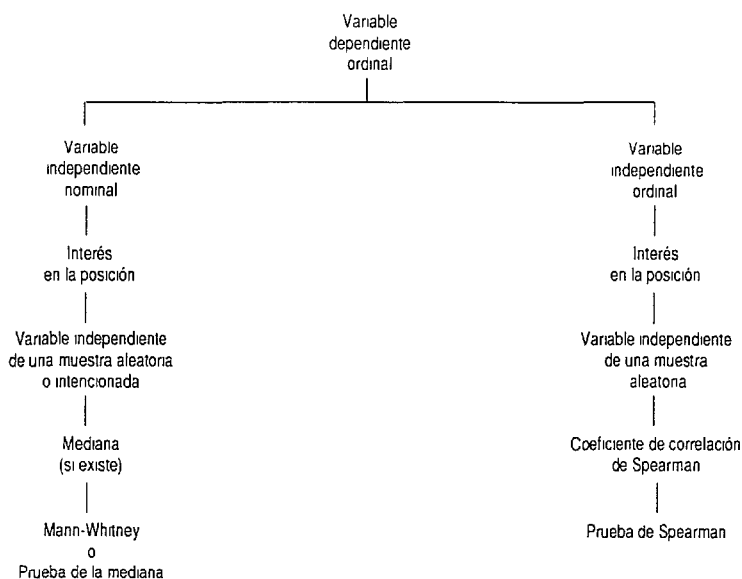
VARIABLE DEPENDIENTE ORDINAL

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 bivalente 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

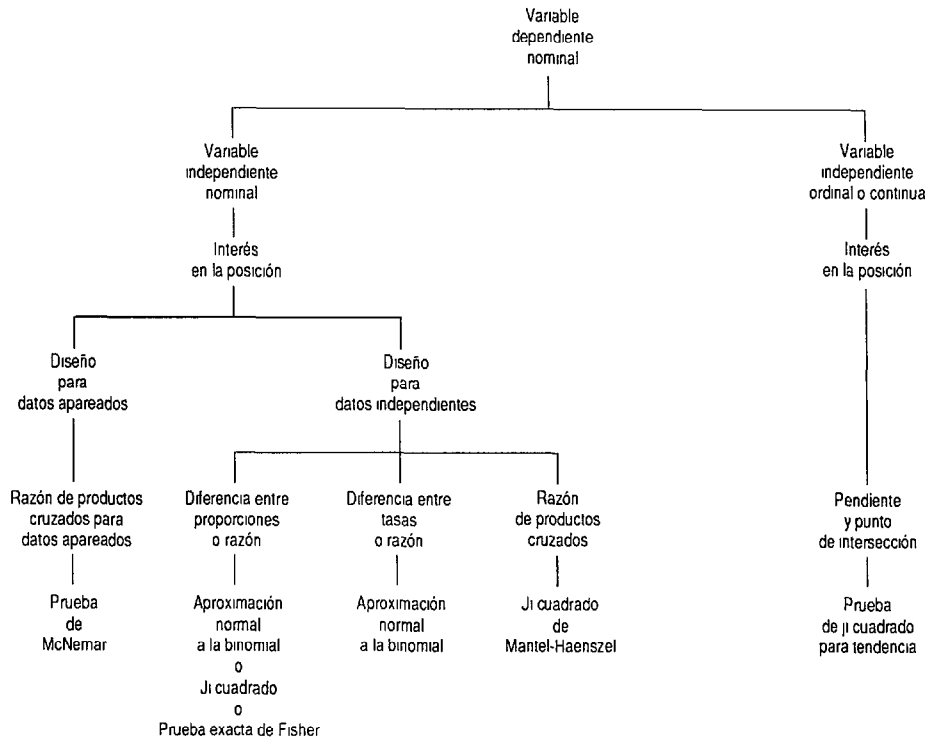
Los métodos estadísticos bivariantes para las variables dependientes nominales se presentan en la figura 28-7.

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 bivalente 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 bivalente 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 gaussiana, 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.