Cómo estudiar un estudio y probar una prueba: lectura crítica de la literatura médica[1]

Segunda edición

Richard K. Riegelman y Robert P. Hirsch

PARTE XI:

Capítulo 29. Análisis multivariante

[1] Título 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.

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 Parte 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

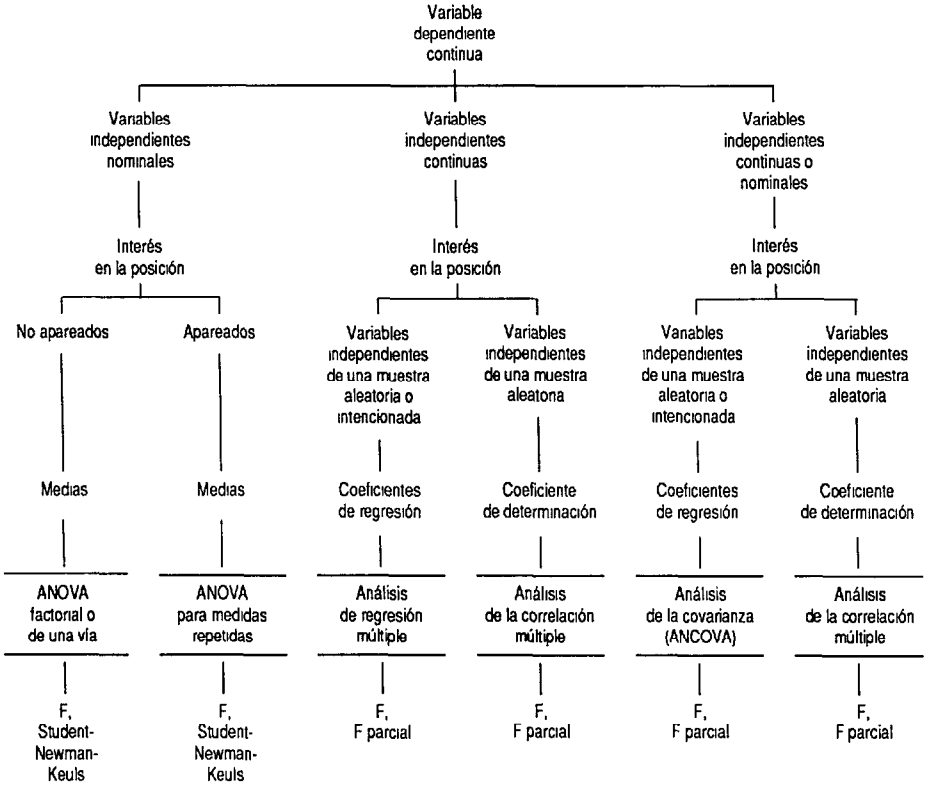
Variables independientes nominales

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, $0 \times \beta_3 = 0$:

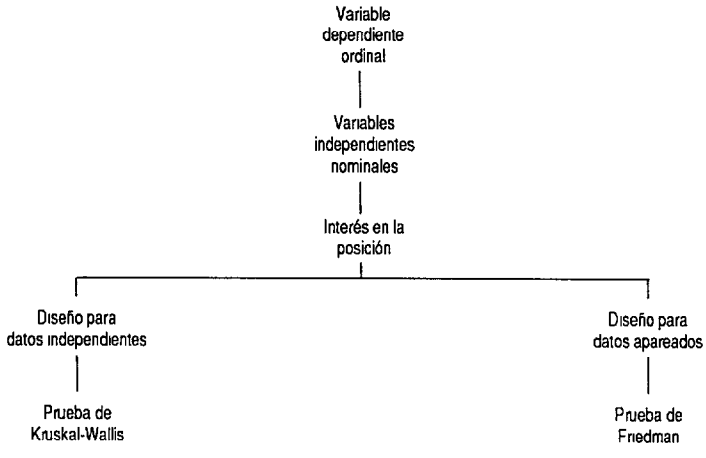
$$\hat{Y} = \alpha + \beta_1 X$$

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*.

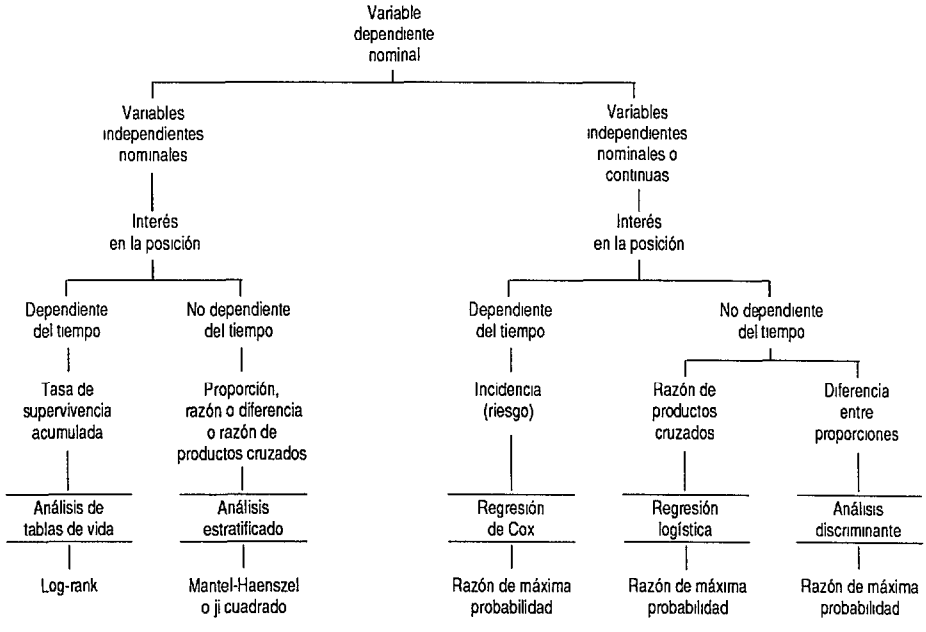
Cuando empleamos métodos multivariantes diseñados para variables dependientes ordinales con objeto de analizar grupos de observaciones que contienen una variable dependiente continua transformada a una escala ordinal, debemos tener en cuenta una desventaja potencial: que la técnica no paramétrica tiene 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. 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 Parte 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-

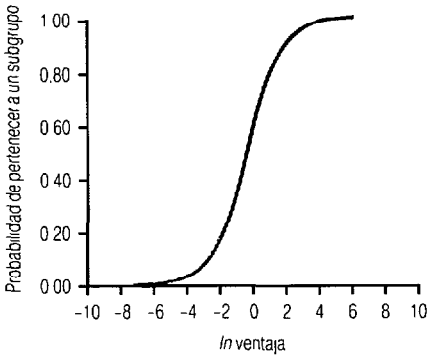
---

[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.