

# Hablemos de...

## Estadística para pediatras (III). Análisis de datos

ROSARIO MADERO, ELIA PÉREZ Y BELÉN SAN JOSÉ

Sección de Bioestadística. Hospital Universitario La Paz. Madrid. España.  
rmadero.hulp@salud.madrid.org; epefer@gmail.com; bjose.hulp@salud.madrid.org

En las investigaciones científicas, el término «datos» se utiliza para referirse a la información que resulta de observar a los sujetos o unidades experimentales en una muestra extraída de la población sobre la que realizamos el estudio. El análisis estadístico de los datos es el estudio científico de esa información, aplicando métodos estadísticos.

El primer paso en el análisis de los datos es su descripción. Los métodos de la *estadística descriptiva* permiten resumir y resaltar numéricamente aquello que es esencial en los datos. Así, por ejemplo, Francis Galton tenía los instrumentos para medir y describir la talla de las personas antes de establecer la relación entre la talla de los padres y la de sus hijos<sup>1</sup>.

Tras la descripción de los datos observados en la muestra, cabe preguntarse ¿qué conclusiones se puede extraer sobre la población de la que procede la muestra? Los métodos de *inferencia estadística* permiten responder a esa pregunta de una manera objetiva, analizando si los resultados son fiables y si se corresponden con hipótesis planteadas de antemano sobre la población. Cualquier método que permita generalizar a una población todo aquello que se observe en una muestra extraída de ella es inferencia estadística.

### ¿Cómo se describen los datos?

Se llama estadística descriptiva al conjunto de técnicas que facilitan la organización, el resumen y la comunicación de los datos. En la descripción de variables, la principal herramienta es la *distribución de las frecuencias*, que se puede resumir mediante medidas numéricas o expresarse de forma gráfica. La tabla 1 contiene las diferentes formas de sintetizar la información que contienen los datos.

#### Características de resumen o estadísticos

Hay diferentes maneras de resumir numéricamente la distribución de frecuencias según sea la naturaleza de las variables en las que se recogen los datos. Cuando la variable es de tipo cualitativo o nominal, es decir, que no toma valores numéricos y sólo representa las cualidades de los sujetos del estudio, su descripción consiste en contar el número de casos en los que aparecen esas cualidades y se indican sus frecuencias absolutas y relativas en una tabla llamada *tabla de frecuencias*.

Cuando los datos son cuantitativos, las observaciones individuales se corresponden con cantidades numéricas y resulta conveniente complementar la distribución de frecuencias con algunas medidas resumen. Las más importantes son las llama-

#### Puntos clave

- La información recogida de una muestra organizada de manera estructurada para su posterior explotación recibe el nombre de datos.
- La explotación de los datos con el fin de obtener resultados y conclusiones se realiza a través del análisis estadístico.
- El primer paso del análisis es la descripción de los datos a través de los métodos de la estadística descriptiva, que permiten organizar, resumir y comunicar los datos.
- En un segundo paso se pretende extraer conclusiones y resultados que de nuestra muestra se puedan extender a una población; para ello, se aplican las técnicas de la inferencia estadística, cuyos principales métodos son la estimación de parámetros y los test de hipótesis.
- Los test de hipótesis más sencillos pretenden relacionar dos variables ( $\chi^2$ , t de Student, la U de Mann-Whitney), pero estos análisis pueden estar sujetos a factores de confusión o ser insuficientes cuando se pretende modelar o representar un proceso de naturaleza multidimensional. En estos casos se aplican las técnicas de análisis multivariante.

La descripción estadística de los datos consiste en la aplicación de técnicas que facilitan la organización, resumen y comunicación de los datos y se realiza con estadísticos que resumen las características de la distribución de frecuencias o bien gráficamente.

Hay una representación gráfica específica para cada tipo de variable, histogramas para variables continuas y diagramas de sectores o barras para variables discretas y nominales.

Los estadísticos de tendencia central o centralización (medias, mediana y moda) indican el valor medio de los datos y los estadísticos de dispersión (desviación típica y coeficiente de variación) expresan su variabilidad.

Un diagrama de caja se obtiene a partir de 4 valores fundamentales: mediana, media, el percentil 25 (P<sub>25</sub>) y el percentil 75 (P<sub>75</sub>) y además, se representan el máximo y el mínimo de la distribución.

Las técnicas de inferencia estadística permiten extrapolar los resultados obtenidos en una muestra a una población mediante la estimación de los parámetros que definen la distribución en la población y las pruebas de hipótesis estadísticas sobre dichos parámetros.

das de *tendencia central* o *centralización* (medias, mediana y moda), que indican el valor medio de los datos, y las *medidas de dispersión* (desviación típica y coeficiente de variación), que expresan su variabilidad. En un segundo nivel, están las medidas que describen el grado de asimetría (coeficiente de asimetría) y de concentración (coeficiente de kurtosis) de la distribución, así como medidas de posición (percentiles).

Todas estas medidas resumen se llaman *estadísticos*. La más familiar de las medidas de centralización es la media aritmética, que suele identificarse con el promedio y que, junto con la desviación típica, proporciona una información importante sobre la distribución de los datos, independientemente de la forma que tenga ésta. Así por ejemplo, para cualquier tipo de distribución, entre la media y 2 desviaciones típicas se encuentra al menos el 75% de las observaciones y entre la media y 3 desviaciones típicas se encuentra al menos el 89% de las observaciones. Estas cantidades son superiores para la conocida distribución normal.

### Representación gráfica

El objetivo de un gráfico es describir fielmente y de manera sencilla la información contenida en los datos observados. Hay una representación gráfica específica (fig. 1) según sea el tipo de variable estudiada<sup>2</sup>.

El *diagrama de sectores* o pictograma se utiliza para representar variables de tipo nominal en la que los sujetos están distribuidos en un conjunto de categorías excluyentes. El área de cada sector es proporcional a la frecuencia relativa de cada categoría. En la representación gráfica de las variables discretas se utiliza el *diagrama de barras*, que representa en el eje de abscisas los valores que toma la variable, levantando en cada punto una barra con una longitud igual a su frecuencia relativa.

Un *histograma* es un conjunto de rectángulos y cada uno representa un intervalo de agrupación o clase para los datos cuantitativos. Sus bases son iguales a la amplitud del intervalo y las alturas se determinan de manera que su área sea proporcional a la frecuencia de todos los valores incluidos en el intervalo. Los histogramas son la representación más frecuente para los datos agrupados.

Los *diagramas de caja* (*box-plot*) (fig. 2) representan gráficamente algunas de las características importantes de la distribución de los datos, como su centro, su variabilidad, su simetría aproximada o su sesgo, la amplitud total y la amplitud de la porción central de los datos, la conducta aproximada de las colas y las observaciones extremas. Un diagrama de caja se obtiene a partir de 4 valores fundamentales: mediana, media, el percentil 25 (P<sub>25</sub>) y el percentil 75 (P<sub>75</sub>). La caja está limitada por los percentiles 25 y 75. Dentro de la caja se representa la mediana por una línea recta y la media, con un asterisco. Cuanto más se aproximen ambas, más simétrica será la distribución. La altura de la caja representada por el rango intercuartílico (P<sub>75</sub>-P<sub>25</sub>) da una idea de la variabilidad en los datos. Si hay valores infrecuentes, éstos están representados por los *outliers* y extremos.

## ¿Qué es la inferencia estadística?

El conjunto de técnicas cuyo propósito es extender conclusiones a una población a partir de los resultados obtenidos en una muestra constituye la *inferencia estadística*. Los principales métodos de inferencia estadística son la *estimación* de los parámetros que definen la distribución de los datos en una población y las *pruebas de hipótesis estadísticas* sobre dichos parámetros. La *estimación puntual* es el proceso por el cual se calcula de forma aproximada el parámetro a estudiar mediante su correspondiente estadístico observado en la muestra. Por ejemplo, el estimador puntual de la media poblacional  $\mu$  es la media muestral  $\bar{x}$ . La *estimación mediante intervalos de confianza* permite conocer el grado de fiabilidad de los resultados obtenidos y las pruebas de hipótesis son métodos que permiten concluir si una muestra observada concuerda o no con la hipótesis formulada<sup>3</sup>.

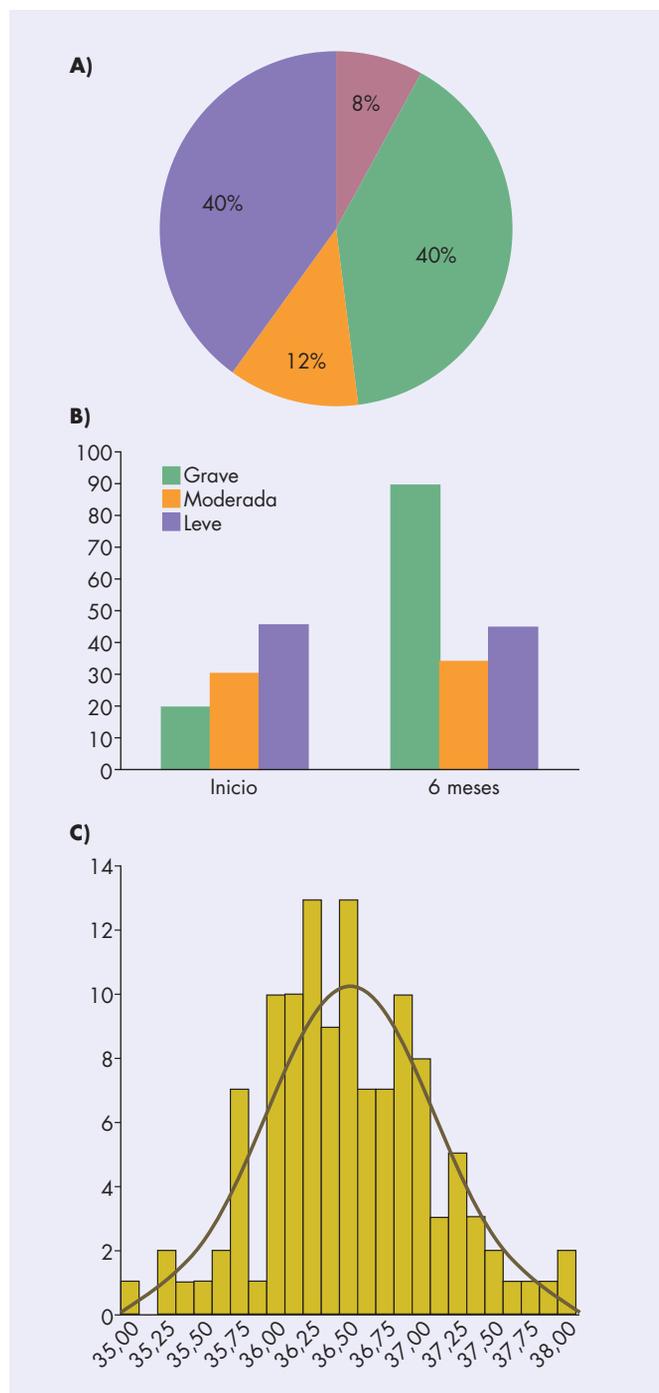
### Intervalo de confianza

El *intervalo de confianza* para un estadístico muestral es aquel que contiene los valores que se consideran como «posibles» para dicho estadístico en la población y que tiene la forma siguiente:

$$\text{Estimador puntual} \pm \text{factor de fiabilidad} * \text{error debido al muestreo del estimador}$$

Tabla 1. Representación gráfica y estadísticos habituales

	Representación gráfica	Medidas de tendencia central	Medidas de dispersión
Nominal (cualitativa)	Diagrama de barras Diagrama de sectores	Moda	
Ordinal	Diagrama de caja ( <i>box-plot</i> )	Mediana	Rango P <sub>25</sub> -P <sub>75</sub>
Continua (cuantitativa)	Histogramas	Medias	Desviación típica Coeficiente de variación



**Figura 1.** Representaciones gráficas según la naturaleza de las variables. A: diagrama de sectores. B: diagrama de barras. C: histograma.

Las pruebas de hipótesis son reglas objetivas que nos permiten rechazar o no una hipótesis planteada de antemano sobre la población, usando la información contenida en la muestra.

Los componentes de una prueba de hipótesis son: la hipótesis nula ( $H_0$ ), la hipótesis alternativa ( $H_1$ ), el error de tipo I o error  $\alpha$ , el error de tipo II o error  $\beta$ , el grado de significación o valor p y la potencia de la prueba.

La precisión de un estadístico se mide por la amplitud de su intervalo de confianza<sup>4</sup>, que depende del error de muestreo y del grado de «confianza» que se quiera asociar al intervalo. Este grado, que establece el investigador fijando el llamado factor de fiabilidad, se expresa en forma de porcentaje, y el valor más habitual es del 95%. El error de muestreo se estima usando alguna medida de la variabilidad y el tamaño muestral ( $n$ ). A los valores extremos del intervalo de confianza se les llama *límites de confianza*. Cuanto más estrecho es un intervalo, más fiable es el estadístico. Por ejemplo, el intervalo de confianza del 95% para una media cuando se desconoce el valor de la varianza poblacional es:

$$\bar{x} \pm t_{(0,05 \ n-1)} \cdot s / \sqrt{n}$$

*Estimador puntual*      *Factor de fiabilidad*      *Error de muestreo*

El estimador puntual es la media muestral, el factor de fiabilidad se calcula mediante el correspondiente valor teórico de la distribución t de Student y el error de muestreo es el cociente entre la desviación típica y la raíz cuadrada del tamaño muestral. El significado práctico de un intervalo de confianza se puede comprender mediante el ejemplo siguiente: que el intervalo de confianza del 95% para la diferencia entre las medias de 2 grupos de pacientes (tratados y no tratados) sea 6-11, significa que si se repitiera el estudio un número suficiente de veces, con muestras de igual tamaño al actual, se confiaría en que en el 95% de los casos el verdadero valor de las diferencias estaría incluido entre 6 y 11, es decir, entre los límites de confianza de ese intervalo<sup>5</sup>.

### Pruebas de hipótesis estadísticas y los valores de p

Las hipótesis en medicina son conjeturas que se hacen sobre el efecto producido por un tratamiento, la genética o el medio ambiente en los pacientes. Son ejemplos de hipótesis: a) los padres altos tienen hijos altos; b) los hombres son más altos que las mujeres; c) fumar produce cáncer de pulmón, o d) la aspirina cura el dolor de cabeza. Si estas hipótesis se expresan mediante los parámetros que definen una distribución, se llaman hipótesis estadísticas. Las *pruebas de hipótesis* son reglas objetivas que nos permiten rechazar o no una hipótesis planteada de antemano sobre la población, usando la información contenida en la muestra. La hipótesis planteada se llama *hipótesis nula*, se la representa por  $H_0$  y no se puede conocer si es cierta o es falsa. Las pruebas de hipótesis nos permiten valorar el grado de consistencia entre lo que se presupone en la hipótesis nula y lo que se observa en la muestra<sup>6</sup>. Se puede plantear otras hipótesis frente a

El estimador puntual de un parámetro es un valor aproximado de éste, obtenido mediante un estadístico muestral y su precisión se mide por la amplitud de su intervalo de confianza.

La amplitud de un intervalo de confianza depende de la variabilidad de los datos, del tamaño muestral y del grado de «confianza» que se quiera asociar al intervalo. Su valor más habitual es del 95%.

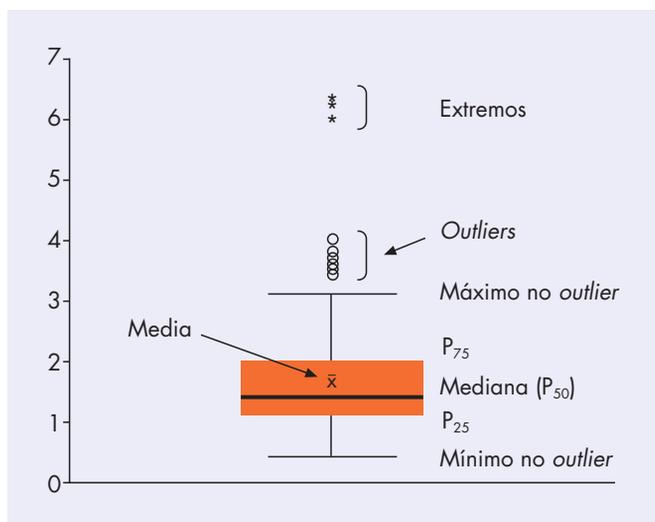


Figura 2. Diagrama de caja (box-plot).

la hipótesis nula que se las conoce como *hipótesis alternativas* y se representan por  $H_1$ . En el ejemplo planteado anteriormente:

$$H_0: \mu_{\text{pacientes tratados}} = \mu_{\text{pacientes no tratados}}$$

$$H_1: \mu_{\text{pacientes tratados}} \neq \mu_{\text{pacientes no tratados}}$$

La prueba o el test estadístico que hay que utilizar, en una situación concreta, dependerá de la hipótesis nula planteada. Siguiendo con el ejemplo anterior, para analizar las discrepancias entre las poblaciones de pacientes tratados y no tratados, se elige un estadístico que refleje las posibles diferencias causadas por el tratamiento en estudio. Para el caso de diferencia de medias se usa un estadístico llamado t de Student y para la diferencia de proporciones<sup>7</sup> se usa el estadístico  $\chi^2$ .

Tras calcular el estadístico correspondiente con los datos muestrales, se estima la probabilidad de que la incertidumbre del muestreo pudiera producir valores menos consistentes con la hipótesis nula que los que hemos encontrado, en el caso de que ésta fuera cierta. A esta probabilidad se le llama grado de significación con respecto a  $H_0$  o *valor p* y es este valor el que nos lleva a rechazar o no  $H_0$ . La figura 3 representa un esquema de una prueba de hipótesis estadística. Tras decidir rechazar o no  $H_0$ , se pueden dar 4 situaciones, en dos de las cuales se tomaría

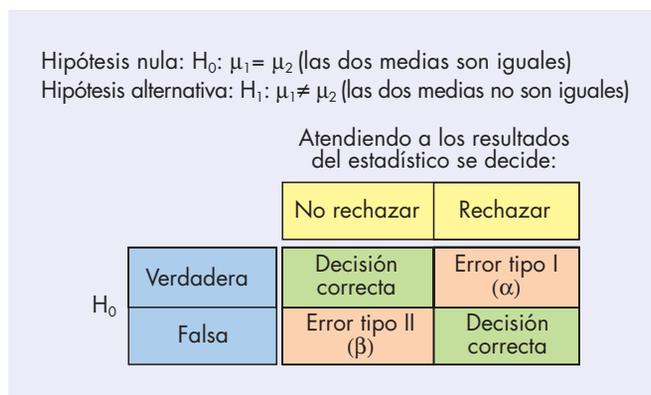


Figura 3. Esquema general de toma de decisiones en una prueba de hipótesis estadística y elementos asociados.

una decisión correcta, al rechazar la  $H_0$  cuando es falsa y al no rechazar la  $H_0$  cuando es correcta. En las otras dos se tomarían decisiones incorrectas, al rechazar  $H_0$  cuando es verdadera (error tipo I o error  $\alpha$ ) y al no rechazar  $H_0$  cuando es falsa (error tipo II o error  $\beta$ ). Se llama *nivel de significación* de la prueba, y se representa por  $\alpha$ , a la probabilidad de cometer un error de tipo I. Se representa por  $\beta$  la probabilidad de cometer un error de tipo II. Las cantidades  $\alpha$ ,  $\beta$  y tamaño muestral están relacionadas, de manera que  $\alpha$  y  $\beta$  sólo pueden ser pequeños incrementando dicho tamaño. Por último, al valor  $1 - \beta$  se le llama *potencia* de la prueba, puesto que representa la capacidad de un test para rechazar una hipótesis nula cuando es falsa. La tabla 2 muestra algunos test estadísticos que se utilizan cuando se trata de relacionar 2 variables. Algunas de esas pruebas se dice que son *paramétricas* si en ellas se compara directamente alguno de los parámetros poblacionales que caracterizan la distribución. Tal es el caso de estadísticos como la t de Student o el análisis de la varianza que se utilizan para contrastar 2 o más medias. En contraposición, se encuentran las llamadas pruebas *no paramétricas* o de distribución libre que permiten comparar distribuciones sin una forma específica y por tanto sin utilizar parámetros. Ejemplos de esta clase son el test de la U de Mann-Whitney o el test de Kruskal-Wallis.

Por otro lado, se distinguen pruebas según se utilicen para datos apareados o no apareados. En las primeras, se trata de estudiar la misma variable en los mismos individuos pero con dife-

Tabla 2. Pruebas estadísticas de relación entre dos variables

	Dicotómica	Nominal	Ordinal	Continua
<b>Dicotómica</b>				
Datos no apareados	$\chi^2$ ( $\pm$ Yates). Test de Fisher	$\chi^2$	U de Mann-Whitney	t de Student (datos independientes)
Datos apareados	Test de simetría de McNemar	$\chi^2$	Test de Wilcoxon	t de Student (datos apareados)
<b>Nominal</b>		$\chi^2$	Kruskal-Wallis (análisis de la varianza)	ANOVA (análisis de la varianza)
<b>Ordinal</b>			Rho Spearman	Rho Spearman
<b>Continua</b>				Correlación, regresión

Se dice que un test estadístico es paramétrico cuando compara directamente parámetros poblacionales y se dice que es no paramétrico si no los utiliza.

Cuando se comparan los datos de una misma variable en los mismos individuos pero en diferente situación experimental, se realizan pruebas con datos apareados y cuando se compara datos de diferentes variables y en diferentes individuos se realizan pruebas de datos independientes.

rente situación experimental. Ejemplos de esta situación se dan al estudiar dos o más observadores independientes la existencia o no de un hallazgo radiológico en una placa simple de tórax o el estudio de la evolución en el tiempo de parámetros antropométricos de sujetos obesos tras cirugía bárica.

### Análisis multivariante

El análisis multivariante es un conjunto de métodos estadísticos y matemáticos que se usa para describir, analizar e interpretar las observaciones multidimensionales, es decir, el material estadístico que proviene de la observación de fenómenos caracterizados por más de una variable. El comportamiento de los pacientes ante un tratamiento debe ser considerado como un fenómeno multifactorial, lo que hace necesario el uso de las técnicas de análisis multivariante. Las posibilidades que ofrece la informática y el espectacular desarrollo de los métodos de cálculo permiten el uso de modelos complejos con un elevado número de variables. La estadística multivariante es más compleja puesto que utiliza métodos de álgebra lineal, cálculo numérico y geometría lineal, pero también es más potente. En general, los métodos multivariantes se diferencian según su área de aplicación (una o varias poblaciones) y según los tipos de variables considerados en el análisis (tabla 3).

El análisis discriminante y la regresión logística tratan de clasificar observaciones en una o más poblaciones utilizando para ello características propias definidas por variables que pueden ser de diferente naturaleza. Cuando el interés de un estudio es conocer la existencia de factores predisponentes o asociados con el tiempo hasta la ocurrencia de un evento, por ejemplo, el tiempo hasta la progresión de una enfermedad o el tiempo de supervivencia, se utilizan técnicas de análisis de supervivencia, entre las más utilizadas en el campo de las ciencias biomédicas está el modelo de regresión de Cox. Las técnicas de análisis de la varianza para varios factores o del análisis de la covarianza se usarán cuando la variable de interés es cuantitativa, mientras que las independientes pueden ser de naturaleza diferente. Esta última es de especial

interés en estudios de pediatría, en que la edad debería utilizarse como covariante en muchos estudios si no se ha controlado su efecto en las fases preliminares del diseño.

## Conclusiones

Muchos clínicos y otros investigadores se sienten desbordados por el complejo tratamiento estadístico que deben aplicar en sus estudios si quieren aprovechar toda la información contenida en sus datos. Es de esperar que los conceptos expuestos de manera somera en estos 3 artículos puedan ayudarles en esa tarea, además de comprender mejor algunos artículos y despejar algunas dudas. No obstante, y siempre que sea factible, es conveniente la consulta con bioestadísticos expertos desde las fases preliminares de un estudio, puesto que las ventajas en términos de tiempo, eficacia y sobre todo de validez científica son incuestionables.

## Bibliografía



● Importante ●● Muy importante

1. ● Sokal RR, Rohlf FJ. Biometry. 2nd ed. San Francisco: WH Freeman and Company; 1981.
2. Peña Sánchez de Rivera D. Estadística 1. Fundamentos. 2.ª ed. Madrid: Alianza Universidad; 1989.
3. ●● Dixon JW, Massey FJ Jr. Introduction to Statistical Analysis. 3th ed. New York: McGraw-Hill; 1983.
4. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ. 1986;292:746-50.
5. Ruiz Maya L, Martín Pliego FJ. Estadística II: Inferencia. Madrid: Ed. AC; 1995.
6. ●● Armitage P, Berry G. Estadística para la Investigación Médica. Barcelona: Ed. Doyma; 1992.
7. Baillar III JC, Mosteller F. Medical Uses of Statistics. 2nd Edition. Boston: NEJM Books; 1992.

Tabla 3. Algunos tipos de análisis multivariante

Variable dependiente	Variabes independientes	Análisis apropiado
Dicotómica	Todas continuas	Regresión logística, análisis discriminante
Dicotómica	Dicotómicas, nominales y continuas	Regresión logística
Nominal	Todas continuas	Análisis discriminante
Continua	Todas continuas	Regresión múltiple
Continua	Dicotómicas y nominales	ANOVA con n factores
Continua	Nominales y continuas	Análisis de la covarianza
Tiempo hasta evento	Dicotómicas, nominales y continuas	Regresión de Cox (análisis de supervivencia)