

# Bioestadística para clínicos. Más allá de “ $p < 0,05$ ”

A. S. RUBIALES, M. L. DEL VALLE, A. VECINO, L. A. FLORES<sup>1</sup>

*Hospital Clínico Universitario. <sup>1</sup>Centro de Salud Medina del Campo. Valladolid*

## RESUMEN

La metodología de la investigación clínica aparenta ser una cuestión lejana y difícil, limitada y específica de unos pocos. Entre los instrumentos que emplea se encuentra la bioestadística: la ciencia que aplica el análisis estadístico a los problemas y a los objetos de estudio de la biología. La bioestadística no es fin, es un medio: los estudios clínicos no se llevan a cabo para alcanzar significaciones estadísticas, al contrario, la bioestadística ayuda a dar valor y a interpretar sus resultados. Dentro de la bioestadística se distinguen la descriptiva, que informa sobre la distribución de los valores, y la inferencial, que ayuda a conocer poblaciones enteras a partir de los datos de las muestras. También se distinguen los tests paramétricos y no paramétricos para el análisis de datos, según se adapten o no a una distribución normal (campana de Gauss). El valor “p” es la probabilidad de obtener por azar una determinada diferencia o mayor entre varias muestras cuando se acepta que todas proceden de una misma población. Sólo a partir de la definición de la “p” se pueden corregir las interpretaciones erróneas y el uso inadecuado de este valor que se observan con frecuencia en el ámbito de las publicaciones científicas.

*Med Pal 2003; Vol. 10, pp. 208-214*

### PALABRAS CLAVE:

Investigación clínica. Metodología. Bioestadística. Revisión.

## ABSTRACT

Methodology of clinical research seems to be a distant and difficult question, limited and specific for just a few people. One of its means is biostatistics, that is, the science that applies statistical analysis to solve problems and to the objects of study of biology. Biostatistics is not an objective but an instrument: clinical trials are not directed to obtain statistical significances, *au contraire*, biostatistics helps to give value and to interpret the results of clinical trials. Inside biostatistics we can distinguish between descriptive biostatistics, that inform about value distribution, and inferential biostatistics, that assist to know whole population from the data of samples. As well, we can differentiate parametric and non-parametric test to data analysis, designed for data that fit or not, respectively, for a normal distribution (Gauss curve). “p-value” is the probability to obtain by chance a determined difference or even larger between several samples when we accept that all of them come from the same population. Only from this definition of “p-value” we can rectify the erroneous interpretations and the inappropriate use of this value that we can often see around medical literature.

### KEY WORDS:

Clinical research. Methodology. Biostatistics. Review.

## INTRODUCCIÓN

El mundo de la metodología de la investigación clínica se considera como algo digno y respetable. Pero la imagen que suele ofrecer es la de una cuestión lejana y difícil, limitada y específica de unos pocos. Se percibe como algo exclusivo de “iniciados” que se preocupan por el tema y que ¡de manera sorprendente!, lo comprenden o, al menos, aparentan que lo comprenden. Da la impresión de que las cuestiones de metodología, incluida la bioestadística, se ven como algo que merece respeto..., pero dudosamente necesario, y sobre todo, ajeno, muy ajeno a los problemas clínicos del día

a día (1). Aun así, son incontables los textos que intentan acercar al personal sanitario todo el instrumental de la metodología de la investigación clínica. Pero el hecho de que sean tan abundantes y que continúen proliferando, hace sospechar que todavía no han alcanzado este objetivo... ¿Por qué? Probablemente por el escaso interés del lector. Y también porque, a pesar de las buenas intenciones, son pocos los que se atreven a escribir sobre estas cuestiones más con el apoyo de la lógica y del sentido común, que con el respaldo de términos específicos y de fórmulas matemáticas ininteligibles para la mayoría de nosotros.

El objetivo de este trabajo es que no sea sólo “otra vez” en que hay que enfrentarse con un tema áspero y difícil, sino que pueda ser el momento en que realmente se consiguieran extraer ideas, pocas pero claras, que sean capaces de

Recibido: 30-04-03  
Aceptado: 05-05-03

dar luz más adelante. Por este motivo, es lógico que de los lectores se pida un poquito de paciencia y de buena voluntad. La información se intenta dar de manera escalonada. El primer paso es explicar en pocas palabras el qué y el porqué de la bioestadística como un instrumento necesario en la investigación clínica. Y pasar luego a recordar algunos criterios prácticos en el análisis estadístico y en la interpretación de los resultados de un estudio clínico, incluidas las técnicas estadísticas (descriptivas y comparativas) más frecuentes y cómo se deberían emplear en función de las características del estudio.

### ¿POR QUÉ LA BIOESTADÍSTICA?

Por bioestadística se entiende "la ciencia que aplica el análisis estadístico a los problemas y objetos de estudio de la biología". Como la propia palabra indica, combina estadística y biología. Es decir, emplea el método y el material de trabajo de la estadística para intentar resolver problemas del ámbito biológico, en nuestro caso, biomédico. Y su motivo último para existir es que es necesario llevar cuenta de lo que hay. Y que es preciso también que estos datos que nos dicen qué es lo que hay, se puedan presentar de manera resumida y comprensible: para poder comunicarse y para poder compararlos.

Pero la bioestadística es un instrumento, no es fin (2). Esta afirmación, tan sencilla, explica el para qué de la bioestadística. No se hacen estudios clínicos para alcanzar significaciones estadísticas, ese "visto bueno" o esa especie de "bendición" de los datos que otorga un valor "p". Al contrario, el diseño y el análisis estadísticos de los datos son un punto más en el apartado de "Material y métodos" de cada trabajo. Y, por tanto, también ayudan a aproximarse a la respuesta de la pregunta clínica que ha originado el estudio. En cierta medida, la intuición, el olfato clínico y un sano espíritu crítico, son el primer impulso para dar cualquier paso adelante en el ámbito de la Medicina. Pero para acercarse más a la verdad es necesario emplear otro tipo de instrumentos: los que llevan a cabo un análisis cuantitativo de los datos (3). Este análisis estadístico está diseñado para echar una mano a la hora de encontrar respuestas a las cuestiones clínicas diarias y para que estas respuestas se den en los mismos términos clínicos en que se formularon las preguntas. Por este motivo, no se debería explicar la bioestadística desde una terminología fría y matemática, ni tampoco se debería entender como un modo de razonar completamente diferente del clínico.

### ESTADÍSTICA DESCRIPTIVA

Una primera parte de la bioestadística es la meramente descriptiva; otra es la inferencial. La primera, la estadística descriptiva, en más de una ocasión no llega a identificarse con el concepto popular de estadística, ya que se limita a presentar las características que definen un grupo de individuos. Sin embargo, sin la estadística descriptiva no sería posible "hacerse una idea" de cómo es un conjunto de datos sin tener que analizar cada uno de ellos. Aparte de los que son las representaciones gráficas, hay datos concretos que permiten hacer una representación sencilla y aceptable de conjuntos con muchos elementos. Estos datos son los que representan el valor central y lo que serían los límites o los extremos. Para conocerlos es conveniente adelantar algo

que se tratará con más calma algo más adelante, que hay dos tipos principales de distribución de datos: la distribución "normal" o paramétrica, es decir, la que se adapta a la "campana de Gauss" (Fig. 1), y las no paramétricas.

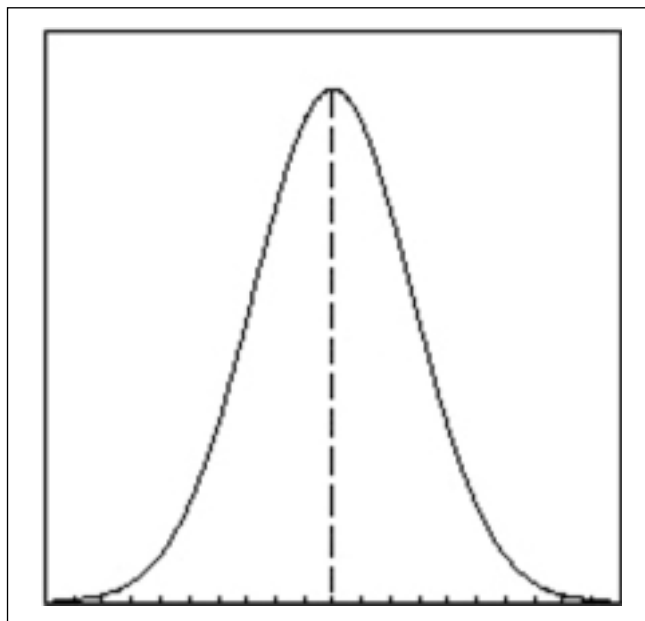


Fig. 1. "Campana de Gauss"; representa lo que se entiende como "distribución normal".

Como se presenta en la tabla I, cuando los datos se adaptan a una distribución normal es necesario conocer la media, es decir, la suma de todos los valores dividido entre el número de individuos y la desviación estándar. Sin embargo, los grupos de datos que no siguen una distribución normal (Fig. 2), se pueden definir aceptablemente con la mediana y los límites y/o los cuartiles. La mediana es el valor que está justo en el medio de todo el conjunto de datos: deja una mitad por delante y la otra mitad por detrás. Los límites son los dos datos más extremos y los cuartiles, los valores que aparecen al dividir los datos en cuatro grupos de igual tamaño, es decir, los que dejan por delante una cuarta parte (el 25% de los datos), dos cuartas partes (la mitad, que corresponde a la mediana) o tres cuartas partes (el 75%).

TABLA I	
VALORES REPRESENTATIVOS EN LAS DIFERENTES DISTRIBUCIONES	
Paramétricas	No paramétricas
Media	Mediana
Desviación estándar	Límites
	Rango intercuartílico

### ¿POR QUÉ LA ESTADÍSTICA COMPARATIVA (INFERENCIAL)?

El sentido común ayuda a reconocer que lo primero y lo más importante no es comparar lo que se consigue, sino

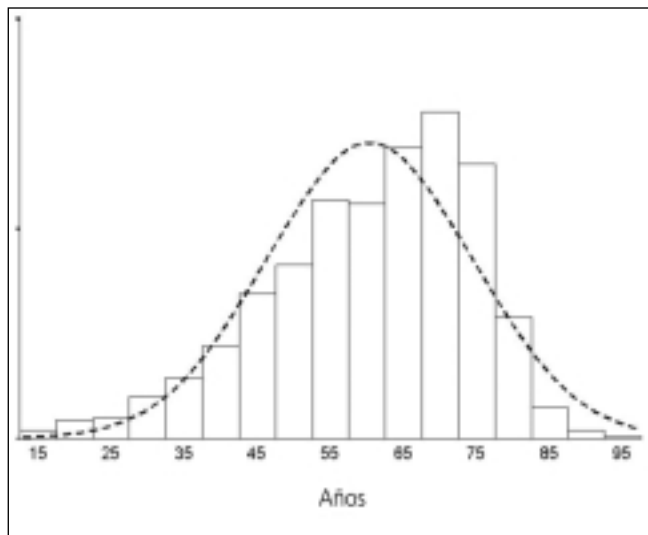


Fig. 2. Distribución de edades de enfermos con cáncer en una consulta de Oncología. La curva no se adapta a lo que sería una distribución normal. En este caso el análisis de los datos debería realizarse con pruebas no paramétricas.

saber qué es lo que se tiene. Sólo entonces se puede sentir la necesidad o la conveniencia de comparar con lo que se tenía antes o con lo que tienen o consiguen los demás... Para ello se emplea la estadística comparativa, inferencial, que se basa en que, como no se puede trabajar con toda la población, hay que analizar los datos de grupos pequeños. Y a partir de esta información, siempre limitada, intentar descubrir, "inferir", qué es lo que sucede realmente en el conjunto de toda la población (4). Y en el caso de comparaciones de dos muestras, lo que se pretende medir es cuál es el riesgo de equivocarse si se acepta que los datos provienen realmente de dos poblaciones diferentes.

En estos estudios en que se comparan dos muestras se emplean tests estadísticos de contraste de hipótesis (5). Estos tests o pruebas de contraste de hipótesis son un modo de comparar los resultados de dos o más grupos. Estos datos que se obtienen son tan sólo una muestra, es decir, sólo están tomados de unos cuantos pacientes (6).

¿Hasta qué punto los datos de un grupo limitado expresan realmente lo que pasa en el conjunto de la población? ¿Las características y las diferencias que encuentro son ciertas o no? Como ya se ha mencionado, lo que se plantea no es ver si dos grupos de datos son diferentes. De hecho, lo que muestra la evidencia es que cuando se toman dos muestras al azar de un mismo grupo, prácticamente nunca son totalmente iguales. Otra cosa es que los resultados se puedan parecer y que de hecho se parezcan, y que se parezcan mucho, incluso que sean "prácticamente" iguales. Lo que interesa ver es hasta qué punto las diferencias (que siempre hay) entre dos muestras pueden hacer pensar que se han obtenido de dos grupos diferentes y no de uno sólo. Pongamos que se analizan dos grupos de pacientes, el primero con 13 individuos y el segundo con 10. Si resulta que un síntoma determinado lo padecen cinco del primer grupo (5+/13) y seis del segundo (6+/10), las proporciones son claramente diferentes (Fig. 3) pero, ¿en qué medida puedo admitir que se deben al azar y que los dos grupos representan una

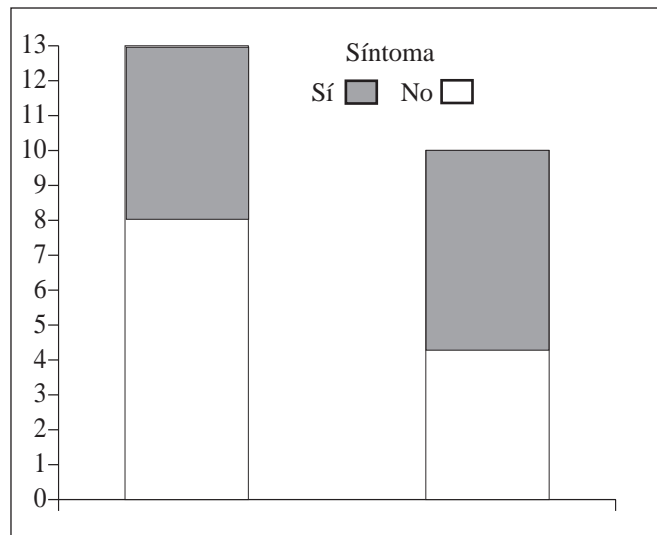


Fig. 3. Un síntoma determinado lo presentan cinco de trece enfermos del primer grupo de pacientes (5+/13) y seis de diez del segundo (6+/10). ¿Pueden proceder las dos proporciones de una misma población?

misma población? Otra posibilidad: si se mueren 45 pacientes de un grupo de 75 que no reciben tratamiento y 35 de los 80 que sí lo reciben, ¿en qué medida puedo asumir que el tratamiento es eficaz?, o si la mediana de la saturación de oxígeno en sangre de 25 personas es de 91% y la de 47 personas es 88%, ¿se debería aceptar que corresponden a dos grupos de enfermos diferentes y no a uno solo?

En resumen, la pregunta que pretende responder la estadística inferencial es: "¿en qué medida puedo extrapolar estos datos a los demás enfermos?"

## PRUEBAS PARAMÉTRICAS Y NO PARAMÉTRICAS

Es conocido que muchas variables tienen una distribución simétrica que tiende a concentrarse en el punto medio: lo que ha venido a llamarse "campana de Gauss" (Fig. 1), que expresa lo que se considera una distribución "normal". Desde el punto de vista matemático es más fácil analizar una variable que se distribuye como esta curva, ya que con los datos de la media y la distribución estándar se puede conocer prácticamente toda la distribución (7). Por este motivo, las pruebas estadísticas paramétricas (para variables que siguen una distribución normal) son, en teoría, más sencillas y facilitan la obtención de resultados con mayor significación estadística. De hecho son las más conocidas en nuestro medio además de ser las más empleadas, también incluso con variables que no se adaptan a una curva gaussiana. Asumir que los datos siguen siempre una distribución normal, es un modo de simplificar los análisis y de obtener "mejores" resultados en los tests estadísticos, pero con una metodología errónea que merma severamente la validez de los resultados.

Hay variables que no siguen una distribución "normal". Unas (como la edad de los pacientes oncológicos) porque tienden a agruparse en torno a un punto, pero con una distribución asimétrica, con más datos y/o más dispersos a un lado que al otro de lo que sería el "pico" de la distribución

(Fig. 2). Otras porque siguen, en ocasiones, una distribución multimodal, es decir, con más de un “pico”. Por último, algunas variables propias del ámbito de la Medicina Paliativa, como las que analizan las cuestiones psicosociales y espirituales, en las que es difícil establecer una graduación y un orden ya que integran un componente cualitativo.

Para cada evaluación estadística se debería emplear una técnica precisa que exige conocer las características de las variables antes de seleccionar el tipo de test (Tabla II). Hay técnicas estadísticas paramétricas y no paramétricas, es decir, para aplicar en la comparación de variables con y sin distribución normal. Por tanto, es preciso conocer qué tipo de distribución tiene la variable antes de emplear cada test estadístico. Hay técnicas estadísticas (como la de Kolmogorov-Smirnov) que valoran si una variable se adapta o no a una distribución normal. Sin embargo, también es aconsejable otra técnica más sencilla, la de observar la distribución gráfica de los resultados. De hecho, si “a ojo” se estima que la gráfica puede ser compatible con una distribución normal, es coherente emplear pruebas de análisis paramétricas (8).

### ¿POR QUÉ LA “p”?

A un porcentaje muy alto de clínicos, los conceptos de bioestadística y de análisis estadístico les lleva a pensar en otro que les resulta casi igual de abstracto: la “p”. Y este valor “p” (cuando es inferior a 0,05), es como el sello de garantía de cualquier resultado: “es cierto, es la verdad”.

Sin embargo, la sorpresa llega cuando se pretende conocer de verdad qué es la “p”. Desde una perspectiva académica, se interpreta este valor “p” como la probabilidad (estimada en “tanto por uno”) de obtener por azar un resultado con estas diferencias o mayores entre los distintos grupos en estudio si se acepta como válida la igualdad entre los grupos, es decir, cuando se acepta que todos proceden de una misma población (en estadística, la hipótesis de igualdad entre los grupos se denomina “hipótesis nula”,  $H_0$ ) (9).

La “p” es, por tanto, una estimación de probabilidad. Pero la valoración que merece cualquier estimación de probabilidad, depende no sólo del valor frío de un porcentaje sino también del riesgo que se asume en el caso de que la estimación “falle”. Si la probabilidad de tener un accidente

severo cada vez que subimos a un avión es del 5%, es decir, uno de cada 20 (0,05 en “tanto por uno”), ¿se viajaría tanto en avión? Aunque se parecen bastante, esta probabilidad de 0,05 (asociada a un riesgo vital) suele tener una valoración más seria de la de una “p” de 0,05 en un estudio clínico. De hecho, el valor “p” marca lo que se considera el riesgo  $\alpha$  de un estudio, es decir, el riesgo de que aparezca un error de tipo I: admitir como cierto, a partir de unos datos limitados, un resultado que es falso en el conjunto de la población. El riesgo  $\beta$  marca el riesgo inverso, o sea, la probabilidad de cometer un error de tipo II: a partir de una muestra, asumir como falso un fenómeno que sí que sucede en el conjunto de la población (Tabla III). El criterio general de la investigación clínica es bastante escéptico y conservador ya que lo que se pretende es evitar “ser timado” por los datos de tantos estudios, o sea, dar la cara por defender datos que a la larga se pueden demostrar erróneos. Así que se suele buscar que este riesgo  $\alpha$  de equivocarse al aceptar algo nuevo sea pequeño (habitualmente por debajo del 5%). Sin embargo, se admite un riesgo de error  $\beta$  más alto (hasta el 20%) porque, por el mismo motivo, en caso de equivocarse las cosas quedan como estaban y se pierde poco: “más vale malo conocido que bueno por conocer”.

Así el valor “p” depende no sólo de las diferencias entre los grupos en comparación, sino también de la cantidad de individuos que entran en el análisis y de la dispersión de los datos. En general, la mayor parte de los tests estadísticos se hacen más robustos, es decir, adquieren una significación mayor, cuanto mayor es el número de elementos que se analizan. Cuanto mayor es el “poder” del estudio, o sea, cuantos más individuos se incluyen, mayor significación estadística alcanzan los resultados (10). El problema de los trabajos con pocos pacientes no es que la estadística concluya que el tratamiento no es bueno; lo que pasa es que con pocos pacientes no se puede concluir nada..., ni a favor de la hipótesis ni en contra. Este es el caso del ejemplo de la figura 3: una diferencia del 21,5% en la incidencia de un síntoma (38,5% en el primer grupo, 60,0% en el segundo) se limita a una “p” de 0,305 (hasta en tres de cada diez de estos muestreos en una misma población se podría obtener por azar una diferencia del 21,5% o mayor). Es un valor “p” muy elevado no porque la diferencia no sea interesante, sino porque el número tan reducido de muestras facilita que estos hallazgos puedan deberse al azar.

**TABLA II**  
**PRUEBAS ESTADÍSTICAS QUE SE EMPLEAN AL COMPARAR MUESTRAS**

		<b>Variables independientes</b>	<b>Variables relacionadas</b>
<i>Dicotómica</i>		$\chi^2$ / Fisher	McNemar
<i>Continua paramétrica</i>	2 muestras	t de Student	t de Student*
	>2 muestras	ANOVA	ANOVA*
	Correlación		Pearson
<i>Continua no paramétrica</i>	2 muestras	Mann-Whitney	Wilcoxon
	>2 muestras	Kruskal-Wallis	Friedman/Cochran
	Correlación		Spearman

Las pruebas paramétricas se aplican en poblaciones que siguen una distribución continua “normal”. Se consideran variables relacionadas o pareadas (no independientes) las que aparecen en grupos de datos que guardan relación entre sí como, por ejemplo, un valor en los mismos pacientes antes y después de recibir un tratamiento. Por correlación se entiende la asociación entre dos variables continuas  
ANOVA: análisis de la variancia; \*: técnica específica para muestras pareadas

TABLA III

**POSIBLES RELACIONES EN LA VALORACIÓN DE LOS TESTS DE HIPÓTESIS PARA DETERMINAR LA EFICACIA DE UN TRATAMIENTO PARTIR DE LOS DATOS DE UNA MUESTRA**

Datos de la muestra	Realidad en la población	
	Es eficaz	No es eficaz
Es eficaz	Verdadero positivo Poder $(1 - \beta)$	Falso positivo Riesgo $\alpha$ (valor "p") Error de tipo I
No es eficaz	Falso negativo Riesgo $\beta$ Error tipo II	Verdadero negativo

Por otra parte, cuanto menor es la dispersión de los datos, algo que va medido por la variancia y otros parámetros similares, también aumenta la significación estadística: los datos muy dispersos quitan fuerza a las conclusiones. Los datos marginales pueden deberse tan sólo al azar o, por el contrario, reflejar lo que es la distribución real de la población (donde, seguro, hay individuos "raros"). Y si se estima que reflejan una parte de la población, habrá que ver si interesa analizarlos dentro del conjunto de datos o no, según que los resultados se quieran extrapolar a toda la población o sólo a los que no se consideran marginales. En los tests de hipótesis, los resultados se pueden dirigir retirando del análisis los valores extremos cuando discrepan de la hipótesis o de los datos esperados o manteniéndolos cuando los apoyan. En todo caso, el remedio es actuar con honradez, de manera que tanto los criterios para aceptar o no los valores de una muestra como las técnicas estadísticas que se vayan a emplear, estén definidas "*a priori*", es decir, antes de tener los datos. Si no, es probable que se juegue con los datos, siempre para alcanzar unos resultados con mayor significación estadística ("nadie tira piedras contra su tejado"), aun con el riesgo de que se alejen de la verdad.

No es preciso comparar todo tipo de datos empleando la "p". Incluso cuando estos datos sean homogéneos y el criterio de comparación, sensato. Por ejemplo, hay que ser prevenido con el uso de la "p" en la "comparación" de las características de los grupos de pacientes que entran en un estudio randomizado. La "p" indica la probabilidad de obtener por azar un resultado. Y cuando se analizan las características de los diferentes subgrupos de pacientes que entran en estos ensayos clínicos aleatorizados, la distribución de los pacientes en los subgrupos es, por definición, aleatoria. Así que no se hace muy necesario conocer con qué probabilidad aparecerían al azar cuando hay certeza de que han aparecido por azar...

Un buen complemento de la "p" es el intervalo de confianza (11). La primera expresa la probabilidad de que una diferencia se deba al azar asumiendo la hipótesis nula (de igualdad). El intervalo de confianza lo que expresa es la precisión de un resultado, es decir, su localización más probable dentro de una población. Este intervalo de confianza depende, como el valor "p", de la dispersión y del tamaño la muestra. Este intervalo ayuda a saber dónde puede encontrarse el valor real de una población a partir de los datos de una muestra. Un intervalo de confianza del 95% define que de cien veces que se obtuviera este dato concreto, en 95 de ellas el valor real de la población se encontraría dentro de los límites del intervalo.

### MÁS ALLÁ DE $p < 0,05$

No se debe valorar un trabajo clínico simplemente porque "da" o porque "no da" (es decir, porque alcanza o no alcanza una " $p < 0,05$ ") (12). Para muchos la "p" es algo así como la calificación de su trabajo: suspenso si la "p" es mayor que 0,05, aprobado si se consigue el aval de una "p" inferior y con tanta mejor nota cuanto más pequeño sea su valor ("*cum laude*" si es menor que 0,001). Así que es necesario que se plantee un análisis no sólo de lo que es la "p", sino también de lo que significa en cada trabajo; un análisis que vaya más allá de "si la p es menor de 0,05, los resultados son buenos". Quedarse en esa interpretación es desconocer lo que quiere decir la "p" y despreciar el trabajo de muchos investigadores.

No es buena metodología de investigación la de tantear resultados y la de cruzar todo tipo de conjunto y subconjunto de datos. En concreto, tomar una base de datos para hacer un análisis preliminar cruzando todos los datos en un programa estadístico, es un modo estupendo de encontrar significaciones estadísticas. El problema es que no toda comparación de resultados tiene valor. No se deben juntar "churras con merinas". De hecho, antes de diseñar un cruce o una comparación es aconsejable plantearse: "y los resultados, ¿qué me van a querer decir?". Y, si no se sabe lo que van a querer decir, probablemente sea mejor no realizar ese cruce o esa comparación.

Tampoco es conveniente tantear diferentes pruebas estadísticas cuando el análisis previsto no es satisfactorio. Si lo que se busca no es tanto la verdad como que el trabajo "salga bien", es decir, que consiga un resultado "estadísticamente satisfactorio" (una "p" lo más baja posible o un intervalo de confianza que excluya el valor nulo; o sea, que los resultados alcancen un "aprobado"), lo natural es buscar alternativas que acerquen a este objetivo. Y si un test estadístico no lo consigue, se va probando con otros hasta elegir (sin ningún criterio metodológico aparente) el que dé mejores resultados o el que, al menos, dé más brillo a los mismos resultados. De hecho, no es raro que para un mismo análisis haya más de una técnica, ya que cada una da un peso relativamente distinto a diferentes factores. Y, aunque suelen aportar significaciones estadísticas similares, a veces el cambio de test regala esas "centésimas" que permiten superar el umbral del valor "p" deseado.

Lo más honrado es que cada trabajo se diseñe para valorar una hipótesis y que uno se atenga a los resultados, "positivos" o "negativos". Sin embargo, cuando el autor se ve apurado porque la falta de significación estadística la ve como un "suspenso", es normal que busque alternativas. Y una es la de multiplicar las comparaciones entre todo tipo de subgrupos (13). Así, si al terminar el estudio los resultados globales no son relevantes, a veces se tiende a "torturar los datos hasta que cantan", o sea, a llevar a cabo análisis múltiples de los resultados con todo tipo de subgrupos (comparables o no) hasta que se obtiene la famosa " $p < 0,05$ ". Y esto es algo que suele ser eficaz, ya que multiplicar los análisis aumenta la probabilidad de encontrar significaciones estadísticas, pero significaciones estadísticas espurias (14). De hecho la "p" marca el riesgo de encontrar unos resultados al azar. Y multiplicar las comparaciones es como comprar más boletos para una rifa: hace que sea más probable encontrar todo tipo de resultados... pero por azar.

La bioestadística, como ya se ha mencionado, no es fin; es sólo un medio. Es un instrumento inerte que hay que saber

emplear y al que hay que aprender a interpretar, o sea, al que hay que llegar a "dar vida" para que sea algo más que series de números. Por esta razón, la relevancia que se dé a cada valor de "p" es esencialmente personal. O sea, que no es adecuado aceptar o rechazar un tratamiento porque la "p" supera o no el límite (arbitrario o consuetudinario) de 0,0515. Es difícil definir la significación "a priori" de manera que todos los que conozcan los resultados asuman que es ese (y no otro) el límite de lo "apto" y "no apto" según el criterio estadístico. Si se presenta como "significativa" cualquier "p < 0,05", ¿sería preciso despreciar siempre un valor de "p" de, por ejemplo, 0,07?, ¿y por tanto no cabrían dudas ante una "p" de 0,04?

Finalmente, no hay que olvidar que el valor "p", por muy pequeño que sea, no define la relevancia clínica de unos resultados (16). Y que esta "p" es un reflejo también del "poder", del número de datos que se analizan. Así, cuantos más individuos se incluyan, más probable será encontrar "significaciones estadísticas" para diferencias mínimas (17), a veces tan pequeñas que llegan a ser clínicamente irrelevantes (18). Por ejemplo, para más de un clínico, los resultados de un fármaco que alcanza un gran alivio de la disnea del enfermo terminal, pueden tener un gran impacto, aunque la significación estadística sea mediocre al haber incluido muy pocos enfermos. En este caso se puede estar consiguiendo un efecto grande sobre un problema difícil. Ciertamente, es posible que los datos se deban al azar, pero abren una puerta al tratamiento de un problema frecuente, severo y de difícil control. Sin embargo, un nuevo analgésico que muestre un alivio discreto del dolor somático, por ejemplo, de 0,3 en una escala de cero a diez, con "p < 0,001", puede no aportar nada o casi nada, si hay otros fármacos que consiguen una analgesia similar o si este nuevo tratamiento es difícil de tolerar o muy caro... En resumen, una cuestión es la significación estadística, y otra diferente la relevancia clínica de unos resultados, es decir, su impacto real en la práctica clínica.

En resumen, la "p" hace el mismo papel que un farol en la noche. A más de uno que llega trastabillado le puede venir muy bien para apoyarse y no caer. Pero su mejor función es la de dar luz para afianzar al caminante y mostrarle el camino (Fig. 4).

## ¿QUÉ ES UN ANÁLISIS MULTIVARIANTE?

Por último, es conveniente repasar en pocas palabras el objetivo de los análisis multivariantes o multivariantes (19). De manera resumida, lo que pretenden es emplear técnicas estadísticas para buscar dentro de un mar de posibles factores, sólo aquellos que sí influyen de manera decisiva en un resultado.

Es obvio que muchos factores que influyen, por ejemplo, en el pronóstico de un enfermo con cáncer avanzado reflejan un mismo problema. Este es el caso de los síntomas, del deterioro del estado general o de la pérdida de peso; aunque cada uno tiene su valor, en el fondo todos reflejan de alguna manera, la progresión del cáncer, el aumento del volumen tumoral. Con las técnicas multivariantes lo que se pretende es el "quitar la paja" y quedarse con el trigo, con los factores que se relacionan de manera independiente.

Este tipo de análisis estadístico puede llevarse a cabo tanto con variables estáticas (como la severidad de la disnea en un momento dado) como con procesos dinámicos, como la supervivencia, que integran el factor "tiempo". En todo caso,



Fig. 4. "La "p" hace el mismo papel que un farol en la noche. A más de uno que llega trastabillado le puede venir muy bien para apoyarse y no caer. Pero su mejor función es la de dar luz para afianzar al caminante y mostrarle el camino.

se requieren técnicas avanzadas de las que se puede disponer en numerosos programas informáticos para análisis estadísticos pero que superan, con creces, los objetivos de este trabajo.

## Y AHORA, ¿QUÉ?

Y después de esta introducción al mundo de la bioestadística... ¿qué?, ¿cuál es el mensaje con el que hay que quedarse? La actitud y la aptitud son algo personal, pero hay consejos prácticos que pueden ser útiles.

Por una parte, hoy en día no es razonable plantearse llevar a cabo un estudio estadístico sin la ayuda de un ordenador. Por este motivo, tanto el que tenga intención de acercarse un poco más al mundo de la bioestadística, como el que se vea abocado a ello porque no le quede más remedio, deberían familiarizarse con alguno de los programas estadísticos dis-

**TABLA IV**  
**CRITERIOS DE MAL USO DE LA ESTADÍSTICA**

1. Asumir que los datos siguen siempre una distribución normal
2. Hacer un análisis preliminar para encontrar significaciones estadísticas cruzando todos los datos en un programa estadístico
3. Tantear diferentes pruebas estadísticas si el análisis previsto no es satisfactorio
4. Presentar como "significativa" cualquier relación con  $p < 0,05$
5. Retirar del análisis los valores extremos cuando alteran los resultados, pero mantenerlos cuando apoyan los resultados
6. Si al terminar el estudio los resultados globales no son relevantes, analizar los datos cruzando todo tipo de subgrupos

ponibles actualmente. Tal vez el que tiene mayor difusión es el SPSS; se trata de un paquete estadístico con muchísimos recursos, tantos que la mayoría de ellos el usuario normal no llega a emplearlos nunca. Por suerte, hay textos accesibles que ayudan a emplear el programa de manera sensata (20,21). En todo caso, como en otras cuestiones informáticas, la selección de un programa depende no sólo de las cualidades técnicas, sino también del hecho de que cubra las necesidades del usuario, de la facilidad y la experiencia en el manejo y de la posibilidad de que la información y que los resultados se puedan compartir con otros colegas.

Pero un buen libro y un buen programa no garantizan un buen diseño ni un buen método. La bioestadística es un

instrumento consistente, pero "no hace milagros": no puede dar respaldo a resultados que se obtienen con una metodología inadecuada. Y el caso es que la metodología inadecuada es un problema a la orden del día: por falta de conocimiento o por afán de obtener el máximo posible a los datos de un estudio. Y eso hace que, poco a poco, vayamos cayendo en errores de diseño y análisis que tienden a perdurar y a transmitirse, a "heredarse" (22), entre diferentes promociones (Tabla IV).

Se han intentado recordar algunos conceptos básicos sobre bioestadística y algunos criterios sobre cómo no se deben emplear estos conceptos. Así que es ahora el momento de intentar aplicarlos de manera que no se limite a un ejercicio matemático, sino que sea un modo de dar luz a los resultados (Fig. 4). En el mundo de la Medicina Paliativa es preciso implantar toda una metodología de investigación clínica, una metodología con matices propios, adaptada a las condiciones de los enfermos en Cuidados Paliativos y a sus necesidades. Y dentro de esta metodología es necesario emplear con criterio y con destreza instrumentos como la bioestadística. De esta manera se robustecerá la validez de los resultados, y la Medicina Paliativa seguirá madurando como disciplina.

**CORRESPONDENCIA:**

Álvaro S. Rubiales  
Servicio de Oncología  
Hospital Clínico Universitario  
Avda. Ramón y Cajal, 3  
47011 Valladolid  
Fax: 983257511  
e-mail: asrubiales@hotmail.com

## Bibliografía

1. Docherty M, Smith R. The case for structuring the discussion of scientific papers. *BMJ* 1999; 318: 1224-5.
2. Beam CA. Statistically engineering the study for success. *AJR* 2002; 179: 47-52.
3. Thomas L. Biostatistics in medicine. *Science* 1977; 198: 675.
4. Altman DG, Bland JM. Generalisation and extrapolation. *BMJ* 1998; 317: 409-10.
5. Plasencia A, Porta Serra M. La calidad de la información clínica: significación estadística. *Med Clin (Barc)* 1988; 90: 122-6.
6. Kazerooni EA. Population and sample. *AJR* 2001; 177: 993-9.
7. Bland M, Peacock J. Statistical questions in evidence-based medicine. Oxford: Oxford University Press, 2001.
8. Glantz SA. Primer of biostatistics. 3rd ed. Nueva York: McGraw-Hill, Inc. 1992.
9. Swinscow TDV. Statistics at square one. 9th ed. Londres: BMJ Books. 1997.
10. Porta Serra M, Moreno V, Sanz F, Carné X, Velilla E. Una cuestión de poder. *Med Clin (Barc)* 1989; 92: 223-8.
11. Gardner MJ, Altman DG. Confidence intervals rather than p values. En: Altman DG, Machin D, Bryant TN, Gardner MJ, eds. *Statistics with confidence*. 2nd ed. Bristol: BMJ Books, 2000. p. 15-27.
12. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Hypothesis testing. *CMAJ* 1995; 152: 27-32.
13. Freemantle N. Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? *BMJ* 2001; 322: 989-91.
14. Tannock IF. False-positive results in clinical trials: multiple significance tests and the problem of unreported comparisons. *J Natl Cancer Inst* 1996; 88: 206-7.
15. Sterne JAC, Smith GD. Sifting the evidence. What's wrong with significance tests? *BMJ* 2001; 322: 226-31.
16. Porta Serra M, Plasencia A, Sanz F. La calidad de la información clínica: ¿estadísticamente significativo o clínicamente importante? *Med Clin (Barc)* 1988; 90: 463-8.
17. Sanz Rubiales A, del Valle Rivero ML, Rey Castro P, Vecino Martínez A. Metaanálisis, significación estadística y beneficio clínico. *Med Clin (Barc)* 2000; 114: 198.
18. Bartelink H. Is neoadjuvant chemotherapy the answer for bladder cancer? *Lancet* 1999; 354: 526-7.
19. Katz MH. Multivariable analysis: a primer for readers of medical research. *Ann Intern Med* 2003; 138: 644-50.
20. Pérez C. Técnicas estadísticas con SPSS. Madrid: Pearson Educación, 2001.
21. Martínez-González MA, Irala Estévez J, Faulin Fajardo FJ. Bioestadística amigable. Madrid: Díaz de Santos, 2001.
22. Greenhalgh T. How to read a paper. 2nd ed. Londres: BMJ Books, 2001.