

# Metodología de investigación y lectura crítica de estudios

## Una mirada al valor de $p$ en investigación

**Jacky F. Gil<sup>1</sup>**  
**Javier A. Castañeda<sup>2</sup>**

### Resumen

La búsqueda de datos dentro de la bibliografía científica ofrece tanto al lector espontáneo como al investigador diferentes retos al momento de leer, comprender y analizar los resultados de un artículo publicado, principalmente cuando los resultados contienen información sobre procedimientos estadísticos empleados en el análisis, como es el caso del reporte de valores de  $p$ . Este artículo ofrece al investigador y al usuario de la bibliografía científica elementos y conceptos que le permitan, entre otras: identificar y entender los componentes de una prueba de hipótesis, interpretar adecuadamente los valores de  $p$ , reconocer las ventajas que tiene el uso de intervalos de confianza frente a los valores de  $p$ , diferenciar entre significación estadística y significación clínica en una prueba de hipótesis y reconocer la importancia del cálculo del tamaño de la muestra, desde la planeación de este tipo de estudios.

**Palabras clave:** estadística, tamaño de la muestra, intervalos de confianza.

**Title:** A Look at  $p$ -values in Research.

### Abstract

Researchers and spontaneous readers both have to face several challenges when they are looking for evidence in scientific papers, especially when statistical procedures are developed during the analysis phase, such as the use of hypothesis testing and  $p$ -values. This paper offers concepts and pointers to assist researchers and readers of scientific literature to identify and understand the components within a statistical test of hypothesis, to appropriately interpret  $p$ -values, to recognize the advantages in using confidence intervals rather than  $p$ -values, to establish the difference between statistical significance and clinical relevance in the use of hypothesis testing and to recognize the importance of sample size calculation from the beginning of a study.

**Key words:** Statistic, sample size, confidence intervals.

---

<sup>1</sup> Bioestadístico, profesor instructor, Unidad de Epidemiología Clínica y Bioestadística, Pontificia Universidad Javeriana, Bogotá, Colombia.

<sup>2</sup> Estadístico, profesor instructor, Unidad de Epidemiología Clínica y Bioestadística, Pontificia Universidad Javeriana, Bogotá, Colombia.

Un investigador interesado en conocer la prevalencia de depresión en Colombia realiza una búsqueda en los diversos estudios científicos existentes. A partir de ahí encuentra algunas referencias que considera relevantes y se dispone a extraer los datos que, cree, son de interés. Un ejemplo de estas referencias es, entre otros, el trabajo de Gómez-Res-trepo y colaboradores (1), quienes describen la prevalencia de depresión y los factores asociados con ella en la población colombiana. Los autores anotan en uno de los apartes el siguiente resultado: «La proporción de depresión grave fue mayor en mujeres que en hombres, en el último mes (29,2% frente a 19,6%, respectivamente;  $p < 0,001$ )». Del mismo modo, puede hacerse referencia al artículo de Mantilla y colaboradores (2), quienes describen la prevalencia de sintomatología depresiva en niños y niñas escolares de Bucaramanga, y dentro de los resultados comentan lo siguiente: «En relación con el género la frecuencia de depresión fue mayor en las niñas, con un 13,3%, que en los niños, con un 6,4%. Sin embargo, esta diferencia no fue significativa ( $p = 0,07$ )».

El investigador, luego de leer los fragmentos de los resultados mencionados, se plantea las siguientes preguntas: ¿cómo se debe interpretar la afirmación expresada en el primer estudio? ¿Qué significa en el segundo estudio que la diferen-

cia no es significativa? Con los hallazgos reportados por Gómez-Res-trepo y colaboradores, ¿es posible determinar la magnitud de la diferencia en las prevalencias de depresión por sexo? ¿Cuál es el efecto del tamaño de la muestra en las conclusiones derivadas por los estudios?

El investigador considera pertinente responder a estos interrogantes antes de generar conclusiones y usar esta información en su estudio.

### **Pruebas de hipótesis y valor de $p$**

El reporte de valores de  $p$  dentro de los resultados de una publicación es la prueba del uso de las metodologías de inferencia estadística, es decir, a partir de los resultados de las muestras seleccionadas en un estudio, se derivan conclusiones para las poblaciones de las cuales provienen éstas. De manera particular, los valores de  $p$  indican el uso de pruebas de hipótesis en los parámetros de las poblaciones de interés, motivo por el cual es necesario abordar inicialmente este tema.

Una prueba de hipótesis es una forma de realizar inferencias inductivas (a partir de casos particulares se concluye para casos generales), en la cual inicialmente se plantea una hipótesis de investigación (un

enunciado sobre la naturaleza del problema) y a través de los hallazgos observados en un experimento se decide sobre su factibilidad (3). Estadísticamente, el desarrollo de una prueba de hipótesis requiere plantear dos enunciados excluyentes que se refieran al parámetro que se desea evaluar, denominados *hipótesis nula e hipótesis alterna*.

La hipótesis nula recibe su nombre por afirmar la inexistencia de asociación o de diferencia; mientras que la hipótesis alterna corresponde a la negación de la hipótesis nula y usualmente expresa lo que el investigador desea concluir: que existe diferencia o asociación. El procedimiento para desarrollar la prueba de hipótesis puede semejarse al ejercicio hecho en una corte por el sistema de justicia, donde un acusado (hipótesis nula) es sometido a la evaluación de un conjunto de pruebas (estadísticas muestrales), con el objetivo de valorar su culpabilidad (veracidad de la hipótesis nula). En la corte, el jurado evalúa las pruebas presentadas por la fiscalía y la defensa desde el supuesto de la inocencia del acusado, y se pronunciará declarando al acusado 'culpable' cuando las pruebas sean incompatibles con este supuesto de inocencia; en otro caso, cuando las pruebas no sean lo suficientemente contundentes, declarará al acusado 'no culpable'.

La estadística emplea el mismo método al evaluar una prueba de

hipótesis, pues partiendo del supuesto de que la hipótesis nula es verdadera, evalúa en términos de probabilidad la validez del supuesto, a la luz de la información contenida en la(s) muestra(s). Si la probabilidad de validez del supuesto es muy pequeña, se concluye que la hipótesis nula es falsa, y se acepta el enunciado establecido por la hipótesis alterna; en otro caso, no puede argumentarse la falsedad de la hipótesis nula y, por lo tanto, no se rechaza (4-6).

Si bien puede concluirse un resultado correcto, como rechazar una hipótesis nula cuando es falsa, o no rechazar una hipótesis nula aunque es verdadera; también puede producirse uno de dos posibles resultados erróneos: error tipo I (falso positivo), que consiste en rechazar una hipótesis nula a pesar de que es verdadera, conocido como error alfa ( $\alpha$ ); y un error tipo II (falso negativo), que consiste en no rechazar una hipótesis nula aun cuando es falsa, conocido como error beta ( $\beta$ ). Sin embargo, cuando la muestra es recolectada aleatoriamente por medio de un método de muestreo probabilístico, es posible controlar a priori y cuantificar a posteriori la magnitud en probabilidad de estos errores (Tabla 1). De esta forma, a pesar de que la conclusión derivada de la prueba de hipótesis no sea ciento por ciento confiable, sí tiene una probabilidad controlada de ser acertada (7).

Tabla 1. Decisiones en la prueba de hipótesis

		Verdad en la población	
		Hipótesis nula falsa	Hipótesis nula verdadera
<b>Resultado de la prueba de hipótesis</b>	Rechazar hipótesis nula	Potencia $1-\beta$	Error tipo I $\alpha$
	No rechazar hipótesis nula	Error tipo II $\beta$	$1-\alpha$

El resultado de la prueba de hipótesis se basa principalmente en el cálculo, en términos de probabilidad, de la fuerza de los hallazgos a favor de la hipótesis nula. Esta probabilidad recibe el nombre de valor de  $p$ , y se define como la probabilidad de cometer un falso positivo por efecto del azar, cuando la hipótesis nula es verdadera. Para ejemplificar la noción de valor de  $p$  suponga el siguiente experimento: una moneda legal (no cargada) es lanzada al aire y se obtiene como resultado una 'cara'. Desde el supuesto inicial, la probabilidad de que esto suceda es 0,5. Ahora, en una repetición del experimento, la misma moneda es lanzada, de nuevo, al aire y el resultado vuelve a ser 'cara'. La probabilidad de que esto suceda es  $0,5 \times 0,5 = 0,25$  (aproximadamente en 25 de cada 100 veces que se lance al aire

dos veces una moneda legal se observa como resultado dos 'caras').

Si se repite de manera sistemática el mismo experimento, lanzando la moneda 3, 4, 5, 6 y 7 veces, y cada vez el resultado del lanzamiento es 'cara' y se cuantifica la probabilidad de que esto suceda desde el supuesto inicial, se obtienen los resultados de la Tabla 2. Es probable que ninguna de las personas que observe el experimento tenga comentarios durante las tres primeras repeticiones; sin embargo, al aumentar el número de 'caras' obtenidas de manera consecutiva, es muy posible que algunos observadores duden del supuesto de legalidad de la moneda empleada, puesto que cada vez es menos probable observar un resultado similar cuando se usa una moneda completamente balanceada. El número de

Tabla 2. Cálculo de probabilidades de obtener siempre 'cara' al lanzar una moneda entre 1 y 7 veces al aire

Número de lanzamientos	1	2	3	4	5	6	7
Número de 'caras'	1	2	3	4	5	6	7
Probabilidad	0,5000	0,2500	0,1250	0,0625	0,0313	0,0156	0,0078

lanzamientos necesarios para que cada persona ponga en tela de juicio el supuesto de legalidad de la moneda es personal y subjetivo; no obstante, al observar siete 'caras' consecutivas, seguramente nadie cuestionará que la moneda está cargada, debido a que las posibilidades de observar tal resultado son de 78 en 10.000 (¡realmente baja!).

En investigación, el análisis de los valores de  $p$  necesita reglas objetivas y acordes con la naturaleza del fenómeno que se está evaluado. Por esta razón, cuando se realiza una prueba de hipótesis, es preciso definir un umbral por debajo del cual se rechaza la hipótesis nula. Este umbral recibe el nombre de *nivel de significación*, y representa el máximo nivel en términos de probabilidad que el investigador está dispuesto a cometer un falso positivo ( $\alpha$ ).

En la bibliografía científica los niveles de significación comúnmente utilizados son 0,10, 0,05 y 0,01; valores por arriba de 0,10 no constituyen ninguna prueba y son rechazados dentro de la comunidad científica. Para tomar una conclusión al probar una hipótesis de forma estadística se compara el valor de  $p$  y el nivel de significación alfa que se fijó con antelación. Si el valor de  $p$  es menor que alfa, se rechaza la hipótesis nula (es muy poco probable que la hipótesis nula sea verdadera). En otro caso, no se puede rechazar la hipótesis nula, pero

tampoco se acepta, puesto que puede estarse en presencia de un error tipo II, y se concluye diciendo que las pruebas aportadas por las muestras son insuficientes para rechazar el enunciado descrito por la hipótesis nula (8).

Es importante señalar que el procedimiento descrito sólo es válido cuando el nivel de significación es especificado por el investigador antes de observar el valor de  $p$ . Si el nivel de significación no se establece, el criterio para rechazar la hipótesis nula se restringe a valores de  $p$  menores de 0,01; mientras que para no rechazar la hipótesis nula el valor de  $p$  debe ser mayor o igual a 0,10. Cuando el valor de  $p$  toma un valor en el rango entre 0,01 y 0,10, lamentablemente no se puede tomar una decisión y es necesario recolectar una nueva muestra y repetir la prueba de hipótesis determinando con anterioridad el nivel de significación. De lo anterior, se deriva la importancia de especificar este último incluso desde la escritura del protocolo de investigación.

### **Interpretación de los valores de $p$**

Las dos primeras preguntas planteadas por el investigador del ejemplo se refieren a cómo interpretar y entender los resultados derivados de la prueba de hipótesis en las dos referencias mencionadas; por esta razón, es necesario retomar los

resultados de ambos trabajos para responder estos interrogantes.

El primer paso consiste en establecer un nivel de significación como umbral para comparar el valor de  $p$  obtenido de la prueba de hipótesis, el cual en ambas publicaciones se reporta como  $\alpha = 0,05$ . En segundo lugar, debe establecerse con claridad cuál es la hipótesis nula que se va a probar en cada situación y cuál es la respectiva hipótesis alterna. Finalmente, se debe comparar el  $\alpha$  establecido y el valor de  $p$  encontrado, indicando una conclusión adecuada con la naturaleza del estudio.

Al retomar el primero de los trabajos referenciados, se tiene que la hipótesis nula debe enunciar que la prevalencia (proporción poblacional) de depresión en hombres y mujeres es la misma; mientras que la hipótesis alterna plantea que este trastorno es mayor en la población de mujeres, comparada con la prevalencia en la población de hombres (este tipo de hipótesis alterna se denomina *hipótesis unilateral*, porque establece una única dirección para la evaluación de la diferencia). Dado que el valor de  $p$  reportado, a pesar de no tener una cifra exacta ( $p < 0,001$ ), es menor que el nivel de significación preestablecido (0,05), la conclusión en este caso señala que la magnitud de la diferencia observada en las muestras constituye una fuerte prueba para afirmar que la

prevalencia de depresión en la población de mujeres es mayor, comparada con la prevalencia en la población de hombres. En otras palabras, este resultado señala que la probabilidad de que la diferencia observada entre las prevalencias muestrales de hombres y mujeres se deban solamente al efecto del azar es menor de 1 en 1.000, lo cual es remotamente posible.

Por otra parte, en el trabajo de Mantilla y colaboradores (2), la hipótesis nula plantea que las prevalencias de depresión son iguales para niñas y niños; mientras que la hipótesis alterna enuncia que estas dos prevalencias son diferentes. Este tipo de hipótesis alterna recibe el nombre de *hipótesis bilateral*, puesto que no especifica mayor prevalencia del trastorno depresivo en alguno de los dos grupos. La conclusión en este caso, al ser mayor el valor de  $p$  (0,07) que el valor de  $\alpha$ , es que los datos encontrados en las muestras son insuficientes para afirmar que las prevalencias son distintas en los dos grupos; sin embargo, sería incorrecto concluir que la prevalencia en ambos grupos es igual, pues aunque remota, la probabilidad de cometer un falso negativo existe. De esta forma, la conclusión en términos prácticos es no rechazar la hipótesis nula de igualdad.

En situaciones como la observada en el segundo ejemplo, donde los resultados no logran concluir

contra la hipótesis nula (llamados resultados negativos), los editores de las publicaciones científicas solicitan el reporte de una cifra que cuantifique la capacidad del estudio de rechazar la hipótesis nula cuando realmente el enunciado expuesto por ésta es falso (9), la cual nuevamente es una probabilidad y recibe el nombre de *nivel de potencia estadística de la prueba*, denotado por  $1-\beta$  (complemento de la probabilidad de cometer un falso negativo) (4-6).

Los investigadores deben tener precaución cuando plantean hipótesis bilaterales y cuando se encuentran diferencias estadísticas que permiten rechazar la hipótesis nula de igualdad, porque erróneamente se podría argumentar, por ejemplo, que la característica medida es mayor o menor en alguno de los grupos, basados en las cifras muestrales. Esta conclusión constituye un error, debido a que este enunciado se deriva de una nueva hipótesis planteada a posteriori (luego de ver el valor de  $p$  y las cifras muestrales), la cual requiere diseñar y conducir un nuevo protocolo de investigación para ser probada (10) (excepto cuando el valor de  $p$  a dos colas sea menor de 0,02, valor umbral para hipótesis planteadas a posteriori). Por consiguiente, se recomienda establecer desde etapas tempranas del proceso de investigación la dirección deseada para las pruebas de hipótesis. De otra forma, se asegura la validez de las conclusiones derivadas en el estudio.

En términos prácticos, la decisión de emplear una hipótesis alterna unilateral o bilateral debe estar basada en los conocimientos previos del problema que se va a tratar. Las hipótesis unilaterales solamente deben plantearse en situaciones donde se tenga una expectativa a priori del sentido en el que se puede observar la diferencia de interés o cuando se tiene conocimiento de la imposibilidad biológica, física, etc. de observar una diferencia en una dirección determinada. Por el contrario, las hipótesis bilaterales suelen plantearse cuando se desconoce la dirección en que puede presentarse la diferencia de interés y solamente se quiere establecer la existencia de una diferencia sin concluir específicamente una dirección (8).

Frecuentemente, algunos investigadores suelen comparar los valores de  $p$  obtenidos en estudios diferentes o incluso dentro del mismo estudio, a fin de intentar decidir cuál de los dos es más significativo. Sin embargo, esto representa un error, puesto que los valores de  $p$  corresponden a variables aleatorias, cuyos valores dependen de la muestra estudiada. Por ejemplo, un valor de  $p$  de 0,048 y otro valor de  $p$  de 0,001 tienen la misma interpretación (dicótoma) cuando se ha fijado un nivel de significación del 5%: se rechaza la hipótesis nula, sin poder concluir que uno es más significativo que el otro.



### **Diferencia estadísticamente significativa frente a la diferencia clínicamente significativa**

A pesar de que el resultado de la prueba de hipótesis puede arrojar como resultado el rechazo de la hipótesis nula, al indicar que existe diferencia entre los parámetros poblacionales de interés, esta conclusión no siempre representa el mismo significado en la práctica. En algunas ocasiones se puede encontrar que una diferencia estadísticamente significativa no representa una diferencia de magnitud relevante para la naturaleza del problema que se está estudiando.

Un principio general de la estadística cuando se realiza inferencia señala que a mayor tamaño de la muestra, mayor control del error aleatorio; por lo tanto, mayor potencia de la prueba de hipótesis para detectar diferencias de cualquier magnitud cuando éstas existen. De esta forma, el rechazo de la hipótesis nula depende ampliamente del tamaño de la muestra: cuando es grande casi siempre permitirán rechazar la hipótesis nula, mientras que cuando es pequeño, difícilmente podrán concluir significación estadística (10).

La significación práctica representa la verdadera magnitud de la diferencia, que se considera útil o relevante para la naturaleza del problema. Dentro de la investigación en

salud este concepto es conocido como diferencia clínicamente significativa o tamaño del efecto (11).

La magnitud de la diferencia clínicamente significativa la establece el investigador basándose en múltiples factores, como la gravedad del problema que se va a investigar, la morbimortalidad asociada con el fenómeno, los costos que conlleva la implementación de nuevos tratamientos o la presentación de efectos secundarios, entre otros (12). No obstante, antes de concluir que existe una diferencia clínicamente significativa, es necesario asegurar, en principio, la significación desde el punto de vista estadístico.

### **Valores de $p$ frente a los intervalos de confianza**

A pesar de que el investigador del ejemplo ya está en capacidad de interpretar los resultados publicados en ambos artículos, y concluir que la prevalencia de depresión grave es significativamente mayor en mujeres que en hombres, surgen los siguientes nuevos interrogantes: ¿es posible conocer sobre la magnitud de esta diferencia? ¿Es esta diferencia clínicamente relevante?

Por desgracia, una de las principales debilidades que caracteriza a las pruebas de hipótesis, y que es señalada por muchos autores para desmotivar su uso en la investiga-



ción (14), es la poca información que aporta al conocimiento del tema en estudio, puesto que el valor de  $p$  sólo determina si la diferencia es o no significativa, y de ninguna manera permite cuantificar la magnitud de dicha diferencia.

Puede ocurrir que un lector que esté en búsqueda de un cálculo para la magnitud de la diferencia encontrada pueda verse tentado a utilizar la prevalencia de depresión grave encontrada en la muestra de mujeres (29,2%) y la prevalencia calculada en la muestra de hombres (19,6%), realizar una sustracción directa de estos dos porcentajes y encontrar que la magnitud de la diferencia es un 9,6% (mayor en las mujeres). Sin embargo, este procedimiento es incorrecto, debido a que las prevalencias 29,2% y 19,6% son valores encontrados en la muestra, por lo que es posible que al tomar otras muestras, y producto del error aleatorio, estos valores cambien, lo mismo que la diferencia entre ellos (15).

Para controlar este error de tipo muestral se recomienda construir un intervalo de confianza (IC) para el parámetro de interés, que en este caso corresponde a la diferencia de proporciones (16). El IC del 95% construido para este caso es 7,4% y 11,8%, el cual tiene dos interesantes interpretaciones. La primera es que con un IC del 95%, la prevalencia de depresión grave en mujeres colombianas es mayor que en hom-

bres colombianos entre un 7,4% y un 11,8%. Esta interpretación corresponde a la determinación de la magnitud de la diferencia en las prevalencias en las dos poblaciones.

La segunda interpretación está relacionada con la inclusión del valor 0 dentro del IC. Si está presente, esto significa que no se puede descartar que las dos prevalencias sean iguales. Por el contrario, si el intervalo no incluye el valor 0, que corresponde al caso presente, se puede interpretar que a un IC del 95% existen pruebas para argumentar que las prevalencias en las dos poblaciones son diferentes y que se puede identificar la dirección de dicha diferencia. Para el caso de los datos reportados por Gómez-Restrepo y colaboradores, es de esperar este resultado, porque el IC debe reflejar similares hallazgos que los obtenidos con el valor de  $p$ , con la ganancia de poder cuantificar probabilísticamente la diferencia entre las prevalencias y el error en el cálculo.

De esta forma, se puede decir que el IC es más informativo que un valor de  $p$ ; pero a diferencia del procedimiento de prueba de hipótesis, en la construcción del IC no existe el concepto de falsos negativos. De esta forma, para permitir al lector hacer mejores y más completas interpretaciones, preferiblemente se deben informar los dos resultados en la publicación, intervalos de confianza y valores de  $p$  (17).

## **Importancia del tamaño de la muestra**

Las pruebas de hipótesis, al igual que las demás metodologías estadísticas relacionadas con la inferencia, requieren un cálculo apropiado del tamaño de la muestra, puesto que éste es uno de los más importantes factores que aseguran al investigador obtener resultados con una precisión deseada y a la vez controlar el error aleatorio.

Los factores que influyen en la determinación del número de individuos requeridos para un estudio que pretende probar una hipótesis son:

1. Los índices máximos aceptables de error tipo I y error tipo II, pues entre menores sean, mayor es el tamaño de la muestra.
2. La diferencia mínima que se va a detectar (mínima diferencia que desde el punto de vista clínico resulta ser significativa). A menor diferencia se requieren mayores pruebas para tomar decisiones, lo que conlleva un mayor tamaño de la muestra.
3. La variabilidad de la variable de interés. Cuando la característica de interés en el estudio es poco homogénea, se hacen necesarios mayores tamaños de la muestra.
4. La razón de asignación a los grupos, cuando la asignación de los pacientes en los grupos del estudio es 1 a 1, el tamaño de la muestra resulta ser menor,

comparado con otras razones de asignación.

Como se ve, el cálculo del tamaño de la muestra integra, entre otros, tanto el nivel de significación estadística (nivel  $\alpha$ ) como el nivel de significación clínica (diferencias que clínicamente son relevantes). Dennis y Pérez (11) ilustran las fórmulas empleadas para el cálculo del tamaño de la muestra en estas situaciones y comentan el efecto de cada uno de los factores en el resultado del cálculo.

Para concluir, es importante señalar que el cálculo del tamaño de la muestra, además de ser un ejercicio necesario durante la elaboración de un protocolo de investigación, es también una buena herramienta para evaluar la factibilidad del estudio, en términos del número de pacientes necesarios y la precisión que se desea en los resultados. De esta forma, un estudio conducido con un tamaño pequeño puede ser visto con buenos ojos desde una perspectiva económica (a menor número de pacientes en el estudio, menores serán los costos); pero al tiempo puede carecer de capacidad para detectar diferencias que desde el punto de vista clínico pueden ser relevantes.

Por otra parte, si se emplean tamaños de muestra muy grandes, el estudio puede derivar en la detección de diferencias no relevantes

desde el punto de vista clínico, lo cual representaría un mal uso de los recursos, debido a los gastos innecesarios que esto conlleva. De igual forma, estudios con un número de pacientes mayor o menor al estrictamente necesario no son aceptables desde el punto de vista ético.

### Bibliografía

1. Gomez-Restrepo C, Bohórquez A, Pinto D, Gil JF, Rondón M, Díaz-Granados N. Prevalencia de depresión y factores asociados con ella en la población colombiana. *Revista Panamericana de Salud Pública*. 2004 Dec;16(6):378-86.
2. Mantilla LF, Sabalza L, Díaz LA, Campo-Arias A. Prevalencia de sintomatología depresiva en niños y niñas escolares de Bucaramanga, Colombia. *Revista Colombiana de Psiquiatría* 2004;33(2): 163-71.
3. Goodman SN. Toward evidence-based medical statistics. 1: the p value fallacy. *Ann Intern Med*. 1999;130(12):995-1004.
4. Bernard R. *Fundamentals of biostatistics*. 3rd ed. Belmont (CA): Duxbury Press; 1990.
5. Hogg R, Craig A. *Introduction to mathematical statistics*. New York: MacMillan Publishing; 1978.
6. Gonick L, Smith W. *The cartoon guide to statistics*. New York: Harper Perennial; 1993.
7. Pagano M, Gauvreau K. *Principles of biostatistics*. 2nd ed. Pacific Grove (CA): Duxbury Thomson Learning; 2000.
8. Dawson B, Trapp R. *Bioestadística médica*. 3rd ed. México: Manual Moderno; 2002.
9. Randy SH, Scott MW, Robert SD, Tom AE. Prominent medical journals often provide insufficient information to assess the validity of studies with negative results. *J Negat Results Biomed*. 2002;1(1):1-5.
10. Good P, Hardin J. *Common errors in statistics: and how to avoid them*. New Jersey: John Wiley & Sons; 2003.
11. Dennis R, Pérez A. Cálculo del tamaño de la muestral. En: Ruiz A, Morillo L. *Epidemiología clínica: investigación médica aplicada*. Bogotá: Editorial Médica Panamericana; 2004. p. 141-62.
12. Pita S, Pértega S. Significancia estadística y relevancia clínica serie [Metodología de la investigación]. *Fisterra.com* [Portal de internet]. 2001 [último acceso 2005 junio 30]. Disponible en [http://www.fisterra.com/mbe/investiga/signi\\_estadi/signi\\_estadi.htm](http://www.fisterra.com/mbe/investiga/signi_estadi/signi_estadi.htm).
13. Lozano J, Dennis R. Medidas de frecuencia, de asociación y de impacto. En: Ruiz A, Morillo L. *Epidemiología clínica: investigación médica aplicada*. Bogotá: Editorial Médica Panamericana; 2004. p. 181-97.
14. Clark M. Los valores p y los intervalos de confianza: ¿en qué confiar? [editorial]. *Revista Panamericana de la Salud*;15(5):293-6.
15. Castañeda J, Gil F. Una mirada a los intervalos de confianza en investigación. *Revista Colombiana de Psiquiatría*. 2004;33(2):193-201.
16. Daniel W. *Bioestadística base para el análisis de las ciencias de la salud*. México: Limusa; 2002.
17. Altman D, Machin D, Bryant T, Gardner M. *Statistics with confidence*. 2nd ed. Bristol: BMJ Books; 2000.

*Recibido para publicación:* 3 de marzo de 2005  
*Aceptado para publicación:* 12 de mayo de 2005

Correspondencia

*Jacky F. Gil*

*Hospital Universitario San Ignacio*

*Unidad de Epidemiología Clínica y Bioestadística*

*Carrera 7 No. 40-62, Bogotá*

*fgil@javeriana.edu.co*