Las pruebas de significación estadística en tres revistas biomédicas: una revisión crítica

Madelaine Sarria Castro[1] *y Luis Carlos Silva Ayçaguer*[1]

Forma de citar Sarria Castro M, Silva Ayçaguer LC. Las pruebas de significación estadística en tres revistas biomédicas: una revisión crítica. Rev Panam Salud Publica. 2004;15(5):300–6.

RESUMEN

Objetivos. *Caracterizar el empleo de las pruebas convencionales de significación estadística y las tendencias actuales que muestra su uso en tres revistas biomédicas del ámbito hispanohablante.*
Métodos. *Se examinaron todos los artículos originales descriptivos o explicativos que fueron publicados en el quinquenio de 1996–2000 en tres publicaciones:* Revista Cubana de Medicina General Integral, Revista Panamericana de Salud Pública/Pan American Journal of Public Health y Medicina Clínica.
Resultados. *En las tres revistas examinadas se detectaron diversos rasgos criticables en el empleo de las pruebas de hipótesis basadas en los "valores* P*" y la escasa presencia de las nuevas tendencias que se proponen en su lugar: intervalos de confianza (IC) e inferencia bayesiana. Los hallazgos fundamentales fueron los siguientes: mínima presencia de los IC, ya fuese como complemento de las pruebas de significación o como recurso estadístico único; mención del tamaño muestral como posible explicación de los resultados; predominio del empleo de valores rígidos de alfa; falta de uniformidad en la presentación de los resultados, y alusión indebida en las conclusiones de la investigación a los resultados de las pruebas de hipótesis.*
Conclusiones. *Los resultados reflejan la falta de acatamiento de autores y editores en relación con las normas aceptadas en torno al uso de las pruebas de significación estadística y apuntan a que el empleo adocenado de estas pruebas sigue ocupando un espacio importante en la literatura biomédica del ámbito hispanohablante.*

Palabras clave Pruebas de hipótesis, tamaño de la muestra, intervalos de confianza, inferencia.

En la primera mitad del siglo XX, los investigadores del campo de la salud pública raras veces dominaban los métodos que permitieran cuantificar la "evidencia" y complementar con ellos los informes, con frecuencia anecdóticos, de sus investigaciones. En el tiempo transcurrido desde entonces, sin embargo, se desarrollaron y consolidaron lentamente las pruebas de hipótesis basadas en el cálculo de los valores *P*, conocidas también por pruebas de significación estadística (PSE). En efecto, en los años veinte del siglo pasado Ronald Fisher ideó una manera de medir el grado de incompatibilidad de un conjunto de datos con una hipótesis determinada y, con ello, inventó los famosos valores *P*. Unos años más tarde, Jerzy Neyman y Egon Pearson propusieron un procedimiento que consistía en la elección entre dos hipótesis, y poco después empezó a gestarse anónimamente el recurso híbrido surgido de la fusión de ambos aportes que hoy se aplica (1). Fomentada por el creciente acceso a potentes recursos computadorizados y por el desarrollo de numerosos paquetes estadísticos (SPSS, SAS, BMDP, EPINFO, etc.) que cuentan las pruebas de hipótesis entre sus principales atractivos, el uso de las PSE se fue popularizando hasta hacerse casi ubicuo en la investigación biomédica contemporánea.

A lo largo de varias décadas, sin embargo, se fueron acumulando paralela-

[1] Instituto Superior de Ciencias Médicas de La Habana, La Habana, Cuba. La correspondencia debe dirigirse a Luis Carlos Silva Ayçaguer a: Instituto Superior de Ciencias Médicas de La Habana, G y 25, Edif. Paz Borroto, 6to piso, Plaza, Ciudad de La Habana, Cuba. Código postal: 10400. Teléfono: 537-8324991; correo electrónico: lcsilva@infomed.sld.cu

mente cuantiosas y persuasivas objeciones al uso de las PSE. Estas objeciones son de enorme entidad, tanto epistémica como práctica, y abundan no solo en los artículos científicos, sino también en los libros de texto (2–5). Confeccionar un inventario exhaustivo sería ahora improcedente; procede, sin embargo, apreciar una muestra de las críticas realizadas a los valores *P* y a las PSE a lo largo de los últimos sesenta años:

- los valores *P* no cuantifican la probabilidad de que una hipótesis determinada sea cierta a la luz de los datos, que es lo que verdaderamente interesa, sino que la probabilidad de haber obtenido ciertos datos en el supuesto de que sea cierta determinada hipótesis (6–8)
- el rechazo de la hipótesis nula depende vitalmente de un elemento ajeno a la realidad: el tamaño muestral (9–10)
- las PSE se basan en un esquema dicotómico y mecanicista, por lo que no le proporcionan al investigador los recursos inferenciales necesarios para entender a fondo la realidad que examina (7, 11)
- operan formalmente en un vacío de precedentes (divorcio entre las pruebas y los conocimientos previos) (12)
- para valorar la magnitud de *P* se emplean umbrales infundamentados, tales como 0,05 y 0,01 (2, 7, 13)

Como consecuencia, hace ya varios años que diversas revistas punteras de la producción científica internacional dejaron de admitir trabajos en los cuales solo aparecían pruebas de este tipo. Por ejemplo, *British Heart Journal* anunció en un editorial de 1988 que se unía a la exhortación de *British Medical Journal*, que ya desde 1986 pedía a los autores que trabajaran con intervalos de confianza (IC), fuesen acompañados o no de pruebas de significación (14). Tal postura pasó a ser compartida por revistas tan importantes como *American Journal of Public Health*, *The Lancet* y *Annals of Internal Medicine* (revista que, en particular, estimula el empleo de técnicas bayesianas) (15) y, finalmente, fue adoptada en 1988 por el llamado Grupo de Vancouver[2] (16), en cuyo apartado de requisitos técnicos dedicado al empleo de la estadística se consigna textualmente: "Siempre que sea posible, cuantifique los resultados y preséntelos con indicadores apropiados de error o la incertidumbre de la medición (por ej., intervalos de confianza). No dependa exclusivamente de las pruebas estadísticas de comprobación de hipótesis, tales como el uso de los valores *P* que no transmiten información sobre la magnitud del efecto", texto que ha sido ratificado en todas las versiones posteriores, incluida la más reciente (17).

Sin duda, la alternativa más sencilla es la construcción de IC. A pesar de que los IC y las PSE pertenecen a las mismas matemáticas frecuentistas, los IC no desembocan en la interpretación automática de los resultados, propia de los valores *P*. Ellos constituyen un recurso indirecto para resumir tanto la diferencia entre los efectos observados (por ejemplo, entre los porcentajes o las medias correspondientes a diferentes tratamientos), como el grado en que la diferencia observada entre los grupos comparados se aproxima a la verdadera diferencia entre ellos. El argumento central estriba en que los IC proveen más información que las PSE, a la vez que no obligan a dicotomizar las conclusiones (14). Su aplicación en lugar de las PSE o como complemento de ellas también ha sido recomendada como norma por importantes asociaciones científicas, tales como la Asociación Estadounidense de Psicología (18).

Debe señalarse que los IC también se ven determinados por el tamaño muestral y, por ende, son tan criticables en este sentido como las PSE cuando solo se emplean ramplonamente como sus meros sucedáneos (se rechaza la hipótesis de nulidad [H_0] si el IC no contiene el valor que indica que no hay ninguna diferencia entre los grupos y viceversa). Con una muestra suficientemente grande siempre se podrá conseguir un intervalo lo suficientemente estrecho como para que el 0, si por ejemplo se tratara de una diferencia, quede fuera del intervalo y por ende se rechace la hipótesis de que los grupos comparados son iguales. Pero si los IC se emplean de manera adecuada, como recurso para complementar las estimaciones puntuales con una medida del error que pudiera afectarlas, entonces una muestra grande siempre será bienvenida. Al no tener entre sus propósitos valorar si se rechaza o no una hipótesis puntual sino aquilatar la magnitud de un efecto, el investigador podrá conocer mejor dicha magnitud en la medida en que la muestra sea mayor.

Las endebleces intrínsecas del método, sumadas a las dificultades para interpretarlo correctamente, a la insistencia en que cese su uso adocenado y al apoyo que han dado a estos puntos de vista las autoridades y comités citados, conducen a vaticinar una nueva era en que el uso de los IC y la estadística bayesiana desplacen a las PSE.

Aunque cada día son más los autores que critican el empleo de este recurso, las PSE siguen siendo cotidiana moneda de cambio. De ahí nace nuestro interés por examinar los artículos que se publican actualmente con vistas a caracterizar la presencia en ellos de tan controvertido procedimiento; evaluar la manera en que se usan las PSE; detectar interpretaciones incorrectas, y valorar la medida en que se violan las propias normas editoriales de revistas que integran el Grupo de Vancouver. Por una parte, un análisis de esta naturaleza permite aquilatar el grado en que un paradigma que da claras muestras de agotamiento se resiste a cambiar y, por otra, puede tener un valor orientador para autores y editores. Intentos similares se han venido haciendo en el pasado reciente en el mundo sajón (19, 20); nuestro estudio se orienta en esa misma dirección, pero se circunscribe a la literatura sobre temas sanitarios en el ámbito hispanohablante.

[2] "Grupo de Vancouver" es el nombre popular del Comité Internacional de Directores de Revistas Médicas, cuya primera reunión, con carácter informal, se celebró en 1978 en Vancouver, Columbia Británica, Canadá.

MATERIALES Y MÉTODOS

Se realizó un estudio bibliométrico descriptivo en el que se examinaron todos los artículos originales publicados en el período de 1996–2000 en tres publicaciones periódicas: la *Revista Cubana de Medicina General Integral* (RCMGI), la revista española *Medicina Clínica* (MC) y la *Revista Panamericana de Salud Pública/Pan American Journal of Public Health* (RPSP/PAJPH). A nuestro juicio, estas publicaciones son representativas de lo que se produce en el mundo hispanohablante en las dos grandes áreas de la investigación médica: la clínica y la salud pública.

Los trabajos de interés para este estudio se clasificaron en dos categorías:

- Descriptivo: Caracterización resumida de una realidad sanitaria en un lugar y un momento particulares
- Explicativo: Valoración de hipótesis que podrían explicar las causas de un problema

Se identificaron las técnicas estadísticas empleadas para analizar los resultados de cada estudio y se determinó si en el artículo se empleaban recursos estadísticos inferenciales básicos (IC o PSE, estas últimas como complemento de aquellos o bien como recurso exclusivo). Se examinó si en el artículo se daba información innecesaria sobre la elección del tamaño muestral, si se usaban niveles de significación prefijados por inercia, sin que mediase una reflexión acorde a cada situación, y de qué modo se presentaban los resultados de las pruebas de hipótesis (la consignación del valor exacto de *P*, el uso de uno o más valores de alfa (α), el uso de α y la mención de la *P* obtenida, o la omisión de los resultados cuantitativos de la prueba de hipótesis realizada).

En la sección de discusión de cada trabajo valoramos si los autores mencionaban indebidamente el tamaño muestral como posible explicación de los resultados de la PSE. Esto responde al vicio, extendido en ciertos ámbitos, de aducir que no se halló significación pero que quizá se hubiera hallado con una muestra mayor. Lamentablemente, la afirmación siempre es cierta, de modo que no tiene valor. El carácter espurio de tal declaración se hace más claro si se repara en que nadie escribe lo contrario, es decir, que con una muestra más pequeña no se hubiese hallado significación. El modo en que las conclusiones se vinculaban con los resultados estadísticos arrojados por las pruebas de significación se examinó según las siguientes pautas valorativas:

- Se juzgó inadecuado el uso del término "significativo" como sinónimo de relevante, notable o importante, en lugar de para constatar que estadísticamente no regía la nulidad.
- Se consideraron mecánicas aquellas conclusiones elaboradas automáticamente con arreglo a los resultados de la PSE sin un análisis profundo que contemplara todos los elementos que pudieran influir en el proceso o fenómeno estudiado.

Debe tenerse en cuenta que las preguntas de un investigador nunca son estadísticas, pues siempre son sustantivas y guardan relación con algún área del conocimiento. La estadística es un mero intermediario metodológico para hallar respuestas a preguntas de otra índole, de modo que es incorrecto plantear el resultado de una prueba estadística en calidad de conclusión. Por ejemplo, al evaluar un tratamiento novedoso (A), un investigador podría concluir que "el porcentaje de pacientes que mejoraron con el tratamiento A fue significativamente mayor que el de pacientes que recibieron el tratamiento convencional B," en lugar de decir si a la luz de los datos obtenidos, del tratamiento estadístico aplicado, de los efectos secundarios reconocidos, y demás, el tratamiento A es realmente más recomendable que el B, que es, quizá, lo que realmente interesa saber.

RESULTADOS

Tipos de estudio

Las tres revistas elegidas se caracterizan por un claro predominio de trabajos con fines descriptivos y explicativos. A estos dos perfiles responden casi todos (97,3%) los artículos de MC y la inmensa mayoría de los publicados en RCMGI y RPSP/PAJPH durante el quinquenio de 1996–2000 (85,0% y 84,0% respectivamente) (cuadro 1).

Uso de pruebas de hipótesis e intervalos de confianza

El uso de PSE y de IC se examinó solamente en los artículos descriptivos y explicativos. El empleo de recursos inferenciales de este tipo, como era de esperar, fue mucho más acusado en los estudios explicativos que en los descriptivos. Concretamente, el porcentaje de artículos que utilizan recursos estadísticos inferenciales básicos es mayor de 80% en todos los casos y asciende a 98,5% de los artículos publicados en el quinquenio por MC (cuadro 2).

CUADRO 1. Distribución de artículos originales según la naturaleza del estudio en tres revistas biomédicas, 1996–2000

Naturaleza del estudio	RCMGI[a]		RPSP/PAJPH[b]		MC[c]	
	No.	%	No.	%	No.	%
Descriptivo	186	86,5	204	70,6	276	50,5
Explicativo	29	13,5	85	29,4	270	49,5
Total	215	100,0	289	100,0	546	100,0

Nota: En este estudio no se examinaron los trabajos teórico-metodológicos, en los cuales no se emplea inferencia (38, 55 y 15 para RCMGI, RPSP/PAJPH y MC respectivamente).
[a] *Revista Cubana de Medicina General Integral.*
[b] *Revista Panamericana de Salud Pública/Pan American Journal of Public Health.*
[c] *Medicina Clínica.*

CUADRO 2. Frecuencia con que se emplean recursos estadísticos inferenciales básicos (PSE, IC, o ambos) en los trabajos descriptivos y explicativos en tres revistas biomédicas de lengua española, 1996–2000

Naturaleza del estudio	RCMGI[a]		RPSP/PAJPH[b]		MC[c]	
	No./total	%	No./total	%	No./total	%
Descriptivo	62/186	33,3	80/204	39,2	203/276	73.6
Explicativo	24/29	82,8	81/85	95,3	266/270	98,5
Total	86/215	40,0	161/289	55,7	469/546	85,9

[a] *Revista Cubana de Medicina General Integral.*
[b] *Revista Panamericana de Salud Pública/Pan American Journal of Public Health.*
[c] *Medicina Clínica.*

En el período estudiado, RCMGI publicó un total de 215 artículos descriptivos o explicativos y en 86 de ellos (40,0%) se usaron recursos inferenciales. En esa revista prácticamente nunca se emplearon IC, ni solos ni para complementar los resultados de la PSE (cuadro 3). En RPSP/PAJPH se publicaron 289 artículos descriptivos y explicativos y se observó una mayor presencia de trabajos con recursos inferenciales (55,7%). En esta revista se observó un mayor porcentaje de estudios en los que se utilizaron IC (casi 1 de cada 5) que en las otras dos. Finalmente, MC fue la revista donde se encontró un mayor uso de métodos inferenciales (85,9%), pero solamente la décima parte de ellos consistían en IC. El resultado más notable, sin embargo, es que en las tres revistas se observa el uso exclusivo de las PSE en la inmensa mayoría de los trabajos en los que se aplicaron métodos inferenciales. El enfoque bayesiano no se aplicó en ninguno de los artículos publicados por estas tres revistas a lo largo del quinquenio.

Tamaño de la muestra y significación estadística

La declaración de que el pequeño tamaño de la muestra explicaba la falta de significación estadística de los resultados se encontró en aproximadamente 3,8% (27/716) de los artículos en los que se aplicaron métodos inferenciales. Del total de 27 artículos en que se atribuyó la falta de significación al tamaño muestral, dos se publicaron en RPSP/PAJPH y los otros 25 en MC. La cifra no es muy grande, pero cobra importancia cuando se considera que es sumamente raro que se publiquen trabajos en los que no se ha conseguido probar las hipótesis de estudio; consecuentemente, el problema tiene una presencia apreciable.

CUADRO 3. Distribución de artículos que emplean recursos estadísticos inferenciales básicos, según el modo en que estos se usan en tres revistas biomédicas de lengua española, 1996–2000

Recursos estadísticos inferenciales	RCMGI[a]		RPSP/PAJPH[b]		MC[c]	
	No.	%	No.	%	No.	%
PSE[d] sin IC[e]	83	96,5	132	82,0	430	91,7
IC	3	3,5	29	18,0	39	8,3
Total	86	100,0	161	100,0	469	100,0

[a] *Revista Cubana de Medicina General Integral.*
[b] *Revista Panamericana de Salud Pública/Pan American Journal of Public Health.*
[c] *Medicina Clínica.*
[d] Pruebas de significación estadística.
[e] Intervalo de confianza.

Comunicación de los resultados de la prueba de hipótesis

Los resultados de las PSE se expresan de muy diversos modos (cuadro 4). La mayor parte de los artículos publicados que hicieron uso de este recurso estadístico en RCMGI (50,6%) y en MC (33,3%) utilizaron uno o más valores de alfa (0,05 en la inmensa mayoría de los casos) para referirse a los resultados, sin explicitar el valor de *P*. En la revista RCMGI esto ocurrió en aproximadamente la mitad de los artículos. Sin embargo, en RPSP/PAJPH predominó el uso del valor exacto de *P* (51,0%).

"Significación estadística" y conclusiones

En total, en 53 (7,4%) artículos se empleó el término "significativo" en las conclusiones del estudio, y en cada revista se observó el uso incorrecto del término en más de 50% de los artículos publicados. Destacó en este sentido la revista RCMGI, en la cual 12 trabajos de los 16 donde se empleó el adjetivo incurrieron en este error.

DISCUSIÓN

En la actualidad las PSE están inmersas en una inocultable crisis como estándar metodológico. Cada vez son más numerosos los estadísticos e investigadores que entienden la racionalidad de las críticas de que han sido objeto durante años y que simpatizan con la idea de prescindir de ellas (7, 21). La polémica en torno al uso de estas pruebas como recurso inferencial, que hoy en día parece estar llegando a su fin como tal, ha labrado el camino para que se acepte la superioridad de los IC y de las técnicas bayesianas, aunque en este último caso de manera muy embrionaria.

No obstante los claros argumentos en contra del uso de pruebas de hipótesis, resulta imposible lograr un cambio a corto plazo. Muchos consideran estas pruebas como un enfoque inferencial único y coherente, y los resultados de

CUADRO 4. Distribución de artículos que usaron las PSE según la forma de presentación de los resultados, 1996–2000

Presentación de los resultados	RCMGI[a]		RPSP/PAJPH[b]		MC[c]	
	No.	%	No.	%	No.	%
Uso del valor exacto de *P*	32	37,6	76	51,0	152	32,9
Uso de uno o más valores de alfa	43	50,6	42	28,2	154	33,3
Uso de alfa y mención de *P*	3	3,5	24	16,1	117	25,4
No consignan resultados cuantitativos de la PSE[d] realizada	7	8,2	7	4,7	39	8,4
Total	85	100,0	149	100,0	462	100,0

[a] *Revista Cubana de Medicina General Integral.*
[b] *Revista Panamericana de Saludo Pública/Pan American Journal of Public Health.*
[c] *Medicina Clínica.*
[d] Prueba de significación estadística.

nuestro estudio revelan claramente la vigencia de este punto de vista. Numerosos investigadores ignoran o conocen de manera vaga que las PSE han sido objeto de crecientes y fundamentadas dudas desde su aparición y por consiguiente prescinden de las opciones que se proponen en su lugar. Ello explica en buena medida el muy reducido empleo de los IC, aun como recurso complementario, que se registra en las tres revistas examinadas.

La ausencia de los métodos bayesianos en los 1 158 artículos publicados en las tres revistas estudiadas también indica que los investigadores saben poco acerca de ellos, a pesar de que son los métodos inferenciales más antiguos en el campo de la estadística. Basta buscar en Internet para darse cuenta de que estos métodos están gozando de un renovado impulso. Los recursos computadorizados que ahora existen y que no existían hace relativamente pocos años no son especialmente fáciles de usar. El paquete estadístico más conocido y empleado (*Winbugs,* adquirible gratuitamente en http://www.mrc-bsu.cam.ac.uk/bugs/), por ejemplo, exige conocimientos matemáticos avanzados y tiene una interfase complicada. EPIDAT 3.0, un programa concluido en 2003 bajo el auspicio de la Dirección Xeral de Saúde Pública (Galicia, España) y de la Organización Panamericana de la Salud (Washington, D.C., Estados Unidos de América), que se puede obtener también gratuitamente en http://dxsp.sergas.es/, intenta suplir esta carencia. Para obtener información de carácter divulgativo sobre este enfoque se sugiere la lectura de un artículo interesante y ameno de Robert Matthews (22).

Ahora bien, ¿cómo explicar el divorcio entre el criterio teórico prevaleciente, consagrado incluso en las normas de Vancouver, y la realidad? Mucho se han discutido las razones que explican la supervivencia de un procedimiento tan estridentemente cuestionado que podría considerarse incluso una expresión de pseudociencia (23). Tales razones son, entre otras, las siguientes: que es fácil de aplicar usando paquetes informáticos accesibles en todas partes, que todo el mundo lo usa, que tanto a estudiantes como a investigadores se les instruye para que lo empleen, que algunos editores lo exigen, y que aportan una aureola de cientificismo y rigor (24, 25).

Es posible, sin embargo, que el problema tenga connotaciones adicionales; es verosímil pensar que algunos editores prefieran "no buscarse problemas" con los autores y viceversa, acudiendo para ello a los recursos convencionales. Es relativamente natural que los investigadores simples lo hagan si se tiene en cuenta que, como han señalado algunos autores (26), muchos de los objetores teóricos de las PSE las han usado en sus propios estudios después de haberlas criticado. Un solo ejemplo palmario de este fenómeno es sumamente elocuente: Altman y Goodman (27) publicaron un artículo en *JAMA* donde daban cuenta de un estudio bibliométrico sobre la posible influencia de contar con un estadístico dentro de un proyecto a efectos de que fuese aceptado por dos importantes revistas (*British Medical Journal* y *Annals of Internal Medicine*). Sorprendentemente, el trabajo se caracteriza por la máxima ortodoxia metodológica —se ciñe al empleo de las PSE— a pesar de que ambos autores son connotados detractores de ellas: Altman en su calidad emblemática de impulsor de los IC como alternativa (14) y Goodman como crítico severo de las PSE desde la perspectiva bayesiana (1).

Por otra parte, el examen de los artículos permitió apreciar los frecuentes errores en que incurren los autores que utilizan las PSE. Cuando sus resultados no tienen significación estadística, algunos afirman, por ejemplo, que la habrían logrado con una muestra mayor. Esta reacción, que es la típica de un "mal perdedor", es en cierto sentido comprensible en la medida en que constituye un acto de rebeldía ante una dictadura metodológica irracional. No es casualidad que no se haya encontrado ni un solo caso en que los autores adujesen que con una muestra más pequeña quizá no habrían alcanzado la significación. Dada la naturaleza de los valores *P*, el rechazo o la aceptación de una hipótesis puede obedecer de un modo decisivo al tamaño muestral. El hecho de poder, con legitimidad, aprovechar esta notable deficiencia de las PSE para justificar la falta de significación de los resultados habla de la insuficiencia de las PSE como recurso inferencial.

Llama la atención la notable diversidad de formas en que cada revista comunica los resultados de las PSE y, en igual medida, la marcada diferencia en el empleo de PSE en las tres revistas. El hecho de que unos se aferren a valores de α sacralizados, de que otros solo den los valores *P* y de que otros proporcionen ambas cosas revela claramente la falta de coherencia y de consenso en torno al empleo e interpretación de las PSE. De hecho, el empleo de valores rígidos de α (especialmente, del consabido $\alpha = 0{,}05$), ocurre en más de la mitad de los ar-

tículos que emplean PSE en las tres revistas, pese a que esta práctica ha sido rechazada casi universalmente desde el punto de vista teórico, incluso por quienes defienden las PSE. El propio Fisher se había ocupado de subrayar la necesidad de usar flexiblemente los niveles de significación y de no establecer un umbral universal, sino de fijar el valor α a la luz de los conocimientos previos respecto del fenómeno analizado y de las posibilidades prácticas del experimento (22).

A menudo se olvida que el análisis estadístico es solo un elemento más que ha de sumarse al arsenal de conocimientos científicos e información aportada por estudios anteriores para configurar una conclusión. En consecuencia, se cometen muchos errores, tales como convertir en una conclusión algo que no pasa de ser un resultado. En ese contexto, resulta bastante frecuente el uso incorrecto de la palabra "significativo" (o sus derivados) para referirse a un resultado "importante". En muchos casos el adjetivo "significativo" aparece acompañado de un adverbio de cantidad (muy, poco, escasamente, etc.) cuya elección depende de la distancia entre el valor *P* hallado y el nivel de significación con el que explícita o tácitamente se opera. Esto también es un error, pues un valor *P* no indica la magnitud del efecto. Algunos autores que no hallaron una asociación estadísticamente significativa entre las variables que conformaban su hipótesis (resultado no deseado por ellos) concluyeron su estudio con la explicación del resultado de la PSE, en lugar de pronunciarse sobre la pregunta concreta que estaban considerando.

Este error, que se encuentra en trabajos donde la significación hace las veces de conclusiones, constituye un típico ejemplo de mecanicismo y suele encontrarse en discusiones carentes de riqueza teórica y que en la mayoría de los casos no hacen alusión a las hipótesis originales. Este fenómeno, sin embargo, se detectó en pocos artículos; de hecho, el término "significación" se emplea en las conclusiones solo en 53 de los 696 artículos que emplearon pruebas de hipótesis, pero la abrumadora mayoría de los autores lo hace de modo incorrecto. A nuestro juicio, una situación de esta naturaleza es inadmisible, ya que, como bien se ha señalado, lo que hace falta es pensar en términos estadísticos, en lugar de ceñirse a rituales estadísticos (28).

La PSE es un recurso cómodo para los autores que apelan a ella: no hay que pensar ni interesa si los resultados son plausibles o no; basta con que el valor *P* sea pequeño para que con el rechazo (o no) de la hipótesis nula se dé por concluida la investigación (29). Sin embargo, la ciencia no necesita aplicadores de paquetes estadísticos y verbalizadores de los resultados que dichos paquetes arrojan, sino investigadores con gran capacidad analítica que puedan convertir los resultados (estadísticos o no estadísticos) en juicios sustantivos sobre las hipótesis que encaran o las realidades que examinan.

El manejo de las PSE en estas tres revistas revela una situación muy poco satisfactoria, pese a las recomendaciones del Grupo de Vancouver que las tres suscriben. En general, el uso de las PSE se caracterizó por diversos rasgos criticables. Sin embargo, numerosos investigadores que desconocen las peculiaridades y endebleces de esta técnica tienen la certeza de que es óptima, y las alternativas que se han sugerido están lejos de ser incorporadas al quehacer habitual de autores y editores. Esto ocurre pese a la presencia de un consenso abrumador a favor del empleo de técnicas estadísticas de mayor riqueza conceptual y, sobre todo, que propicien mayor reflexión y menos automatismo en la discusión de resultados. El empleo anodino y mecánico de las PSE no es precisamente un ejemplo de reflexión creativa sobre la información empírica obtenida; sin embargo, se encuentra muy arraigado en la práctica cotidiana.

REFERENCIAS

1. Goodman SN. Toward evidence-based medical statistics (1): the p value fallacy. Ann Int Med. 1999;130:995–1004.
2. Jeffreys H. Theory of probability. 3rd ed. Oxford: Oxford University Press; 1961.
3. Lindley DV. Introduction to probability and statistics. Part 2: Inference. Cambridge, UK: Cambridge University Press; 1970.
4. O'Hagan A. Kendall's advanced theory of statistics. Vol 2. B: Bayesian inference. London: Arnold; 1994
5. Lee PM. Bayesian statistics: an introduction. 2nd ed. London: Arnold; 1997.
6. Silva LC, Muñoz A. Debate sobre métodos frecuentistas vs. bayesianos. Gac Sanit. 2000; 14:482–94.
7. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Meth. 2000;5(2):241–301.
8. Berger J, Sellke T. Testing a point null hypothesis: the irreconcilability of P-values and evidence. J Am Stat Assoc. 1987;82**:**112.
9. Berkson J. Some difficulties of interpretation encountered in the application of the chi-square test. J Am Stat Assoc. 1938;33:526–42.
10. Thompson B. In praise of brilliance: where that praise really belongs. Am Psychol. 1998; 53:799–800.
11. Rozeboom WW. The fallacy of the null hypothesis significance test. Psychol Bull. 1960; 56:26–47.
12. Spiegelhalther DJ, Myles JP, Jones DR, Aabrams KR. An introduction to bayesian methods in healthy technology assessment: a review. Br Med J. 2000;319:508–12.
13. Chia KS. "Significant-itis" an obsession with the p values. Scand J Work Environ Health. 1997;23:152–4.
14. Gardner MJ. Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J. 1986;292: 746–50.
15. Davidoff F. Standing statistics right side up. Ann Int Med. 1999;130:1019–21.
16. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. Ann Int Med. 1988;108: 258–65.
17. Comité Internacional de Directores de Revistas Médicas. Requisitos uniformes para la preparación de manuscritos. Rev Panam Salud Publica. 2004;15(1):41–57.
18. Wilkinson L. Task Force on Statistical Inference, APA Board of Scientific Affairs. Statistical methods in psychology journals: guidelines and explanations. Am Psychol. 1999;54: 594–604.
19. Savitz DA, Tolo KA,. Poole C. Statistical significance testing in the *American Journal of Epi-*

demiology, 1970–1990. Am J Epidemiol. 1994; 139(10):1047–52.
20. Vacha-Haase T, Nilson JE A review of statistical significance reporting: Current trends and uses in MECD. Measurement and evaluation in counseling and development, 1998;31: 46–57.
21. Silva LC. Cultura estadística e investigaciones en el campo de la salud: una mirada crítica. Madrid: Díaz de Santos; 1997.
22. Matthews RA. Facts versus factions: the use and abuse of subjectivity in scientific research. European Science and Environment Forum Working Paper. Reimpreso en: Morris J, ed. Rethinking Risk and the Precautionary Principle. Oxford: Butterworth; 2000.
23. Johnson DH. Hypothesis testing: statistics as pseudoscience. Fifth Annual Conference of the Wildlife Society, Buffalo, New York, September 22–26, 1998.
24. Carver RP. The case against statistical significance testing. Harvard Educ Rev. 1978; 48: 378–99.
25. Nester MR An applied statistician's creed. Appl Stat. 1996; 45:401–10.
26. Greenwald AG, Gonzalez R, Harris RJ, Guthrie D. Effect sizes and p-values: what should be reported and what should be replicated? Psychophysiology. 1996; 33:175–83.
27. Altman D, Goodman S. How statistical expertise is used in medical research published. J Am Med Assoc. 2002; 287:2817–20.
28. Gigerenzer G. We need statistical thinking, not statistical rituals. Behav Brain Sci. 1998;21: 199–200.
29. Cohen J. The earth is round ($p < .05$). Am Psychol. 1994;49:997–1003.

Manuscrito recibido el 20 de junio de 2003. Aceptado para publicación, tras revisión, el 30 de enero de 2004.

ABSTRACT

Tests of statistical significance in three biomedical journals: a critical review

Objective. To describe the use of conventional tests of statistical significance and the current trends shown by their use in three biomedical journals read in Spanish-speaking countries.

Methods. All descriptive or explanatory original articles published in the five-year period of 1996 through 2000 were reviewed in three journals: *Revista Cubana de Medicina General Integral* [Cuban Journal of Comprehensive General Medicine], *Revista Panamericana de Salud Pública/Pan American Journal of Public Health*, and *Medicina Clínica* [Clinical Medicine] (which is published in Spain).

Results. In the three journals that were reviewed various shortcomings were found in their use of hypothesis tests based on *P* values and in the limited use of new tools that have been suggested for use in their place: confidence intervals (CIs) and Bayesian inference. The basic findings of our research were: minimal use of CIs, as either a complement to significance tests or as the only statistical tool; mentions of a small sample size as a possible explanation for the lack of statistical significance; a predominant use of rigid alpha values; a lack of uniformity in the presentation of results; and improper reference in the research conclusions to the results of hypothesis tests.

Conclusions. Our results indicate the lack of compliance by authors and editors with accepted standards for the use of tests of statistical significance. The findings also highlight that the stagnant use of these tests continues to be a common practice in the scientific literature.