Inv Ed Med 2012;**1(4)**:225-234

Investigación en Educación Médica

www.elsevier.com.mx

METODOLOGÍA DE INVESTIGACIÓN EN EDUCACIÓN MÉDICA

# La prueba de la hipótesis nula y sus alternativas: revisión de algunas críticas y su relevancia para las ciencias médicas

**Iwin Leenen**

*Departamento de Evaluación Educativa, Facultad de Medicina, Universidad Nacional Autónoma de México. México D.F., México.*

Recepción 7 de mayo 2012; aceptación 9 de junio 2012

**PALABRAS CLAVE**
Prueba de significancia; hipótesis nula; valor *p*; inferencia estadística; inferencia Bayesiana.

**Resumen**
La prueba de significancia de la hipótesis nula constituye la herramienta más generalizada para evaluar hipótesis científicas y tomar decisiones al respecto, no sólo en las ciencias médicas y de la salud, sino también en la biología, la psicología, las ciencias políticas y otras ciencias sociales. Sin embargo, desde su introducción por Sir Ronald Fisher en 1925, la técnica ha sido objeto de un debate acalorado entre, por un lado, críticos que identifican varios problemas conceptuales y de interpretación y, por otro lado, defensores que se aferran a la idea de que (el uso correcto de) la técnica tiene un lugar insustituible en el arsenal de la estadística aplicada. Este ensayo revisa algunas de las críticas de este debate, ilustra su relevancia en el ámbito de la investigación en medicina y discute algunas recomendaciones como remedio o alternativa para la prueba de significancia.

**KEYWORDS**
*Significance test; null hypothesis; p-value; inferential statistics; Bayesian inference.*

**Null hypothesis significance testing and its alternatives: A revision of some criticisms and their relevance for the medical sciences**

***Abstract***
*Arguably, null hypothesis significance testing constitutes the most widely-applied tool for the evaluation of scientific hypotheses and decision making, not only in medical and health sciences but also in biology, psychology, political sciences, and other social sciences. However, since its introduction by Sir Ronald Fisher in 1925, the method has been the center of a heated debate where various criticisms related to conceptual and interpretational problems have been counterattacked by advocates of the technique who argue that the method, if used correctly, does have its place in the statistician's toolbox. In this article, I revise some of the most pertinent criticisms of this debate, illustrate their relevance for research in the field of medical and health sciences and discuss some recommendations that have been proposed as a remedy or an alternative for significance testing.*

**Correspondencia:** Secretaría de Educación Médica, Facultad de Medicina, Universidad Nacional Autónoma de México. Edif. B, 3er piso, Av. Universidad 3000, Circuito escolar CU, C.P. 04510. México D.F., México. Teléfono: 5623 2300, ext. 43034. Correo electrónico: iwin.leenen@gmail.com

Echarle una ojeada a los estudios de investigación publicados en *Investigación en Educación Médica*, así como en otras revistas en el área de medicina nacionales e internacionales, revela que una abrumadora mayoría incluye pruebas de significancia estadística (rechazando o no una hipótesis nula) en las secciones de resultados o discusión. Efectivamente, muchos investigadores -no sólo en las ciencias médicas, sino también en la biología, la psicología, las ciencias políticas y demás ciencias sociales- conceden un papel clave a la prueba de significancia de la hipótesis nula (PSHN), para determinar la validez de sus resultados y decidir cuáles hallazgos merecen ser discutidos. Aunque las ideas básicas de la PSHN datan del siglo XIX,[1] el desarrollo de la teoría se suele situar en los años 20 y 30 del siglo pasado con las publicaciones de Sir Ronald Fisher,[2,3] y Jerzy Neyman e Egon Pearson.[4-6] Existen diferencias filosóficas y conceptuales entre los enfoques de Fisher y de Neyman y Pearson;[7-10] los protagonistas, especialmente Fisher y Neyman, quedaron enredados en una discusión virulenta hasta el fallecimiento de Fisher en 1962. En efecto, la práctica actual de la PSHN resulta un "híbrido anónimo"[9] surgido de los dos corrientes, y es más una mezcla de los dos paradigmas que una teoría coherente sobre la prueba de hipótesis.

No sólo el debate entre Fisher y Neyman tiñó la historia de la PSHN, finalmente, estos dos fundadores defendieron diferentes ángulos de interpretación del mismo procedimiento (en líneas generales coincidieron sobre los cálculos de cómo llegar al resultado). Existe otra controversia, probablemente de más relevancia, que inició poco después[11] de las publicaciones principales de Fisher y que continúa hasta la fecha. Se trata de un debate vigoroso, con numerosos autores,[11-23] que critican severamente la PSHN y otros[24-30] que con igual tenacidad la defienden. Las críticas se pueden dividir en dos grupos. La literatura de las disciplinas aplicadas (es decir, donde se usa la estadística como una herramienta, como en las ciencias médicas o la psicología) ha encaminado la discusión sobre todo hacia las dificultades en la interpretación, mientras que la literatura estadística se enfoca más en los problemas de la construcción formal de las pruebas de significancia.[31] Por lo general, los aliados de la PSHN no niegan que el procedimiento tiene limitaciones y pocos defenderían la idea de que proporcionar el valor *p* obtenido a través de una PSHN, sea un resultado suficiente para un análisis estadístico; en su defensa aluden a que muchas de las críticas no se dirigen a la PSHN misma sino a las personas que la usan o interpretan de forma incorrecta y que, aunque el método tiene defectos, la técnica generalmente lleva a conclusiones correctas. Es decir, los defensores abogan que el *uso correcto* de la técnica puede tener su valor en ciertos estudios y para contestar preguntas específicas de investigación. Para revisiones de la controversia con los argumentos de detractores y defensores de la PSHN, véanse los artículos de Nickerson[32] y de Balluerka y cols.[33]

La estructura del resto del artículo es la siguiente: la siguiente sección inicia con una revisión breve de las ideas fundamentales detrás de la PSHN y la interpretación exacta de un valor *p*, resaltando además las diferencias entre el enfoque de Fisher y el de Neyman y Pearson. En la siguiente sección, se presenta una selección de las críticas sobre la PSHN y se ilustra su posible relevancia en el ámbito de la investigación en medicina. A continuación, se discuten algunas de las enmiendas y enfoques alternativos que se han propuesto para evitar los problemas asociados con la PSHN. Finalmente, la última sección considera algunos elementos de discusión.

## Breve revisión de la teoría de la PSHN

El objetivo de una prueba de significancia es hacer inferencias sobre un parámetro -es decir, una característica numérica de una población- con base en los datos de una muestra extraída de esta población. Específicamente, la PSHN es un instrumento para *excluir* un valor específico (o un rango de valores específicos) como valor(es) plausible(s) para el parámetro.

Para ilustrar los conceptos y argumentos en este artículo, usaré el siguiente ejemplo. Un profesor investigador, que organiza cursos dirigidos a médicos para su desarrollo profesional continuo, desea evaluar la eficacia de dos estrategias para la enseñanza: la educación tradicional, donde el curso de actualización se imparte en una serie de clases presenciales, y la educación a distancia, que ofrece los mismos contenidos a través de una plataforma virtual y donde el contacto profesor-alumno se realiza exclusivamente por vía electrónica. Para este fin, el investigador diseña un estudio en el marco de un curso de actualización sobre los descubrimientos recientes en el tratamiento de pacientes que viven con VIH o SIDA. Asigna aleatoriamente la mitad de los 50 médicos inscritos para el curso a la condición de educación tradicional, mientras que la otra mitad recibirá el curso a distancia. Tanto al inicio (*pre*) como al final (*post*) del curso, aplica a cada médico cuatro pruebas para medir diferentes aspectos de su conocimiento sobre el tema del curso, y calcula una puntuación global sumando los resultados en las cuatro pruebas. Finalmente, obtiene para cada participante la diferencia *pre-post* entre dichas puntuaciones globales.

Una prueba de significancia incluye los siguientes pasos:

1. *Construir un modelo estadístico.* Se trata de un conjunto de supuestos sobre las variables de interés. En nuestro ejemplo se pueden considerar dos variables: (a) la diferencia (*X*) entre pre y post en médicos que reciben el curso tradicional, y (b) la diferencia (*Y*) correspondiente en médicos que participan en el curso impartido a distancia. El investigador supone que *X* y *Y* en sus respectivas poblaciones tienen una distribución normal con media $\mu_X$ y $\mu_Y$, respectivamente. Otros supuestos incluyen que la varianza ($\sigma^2$) de dichas distribuciones es igual y que las observaciones de la muestra se han extraído de forma independiente de sus respectivas poblaciones. Los parámetros del modelo son $\mu_X$, $\mu_Y$ y $\sigma^2$.
2. *Especificar la hipótesis nula.* Aplicado a nuestro ejemplo, se trata de la hipótesis de que no hay diferencia entre las medias de las dos poblaciones: $H_0 : \mu_X - \mu_Y = 0$. Aunque usualmente se especifica la hipótesis nula en estos términos, en general se puede considerar la hipótesis $H_0 : \mu_X - \mu_Y = c$, donde *c* es cualquier número real.

3. *Definir un estadístico de contraste*. Un estadístico se calcula a partir de los datos de la muestra. En nuestro ejemplo, el investigador considerará como estadístico de contraste:

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_{\bar{X}} - \mu_{\bar{Y}})}{\sqrt{\frac{S_X^2 + S_Y^2}{n}}} \quad (1)$$

donde $\overline{X}$ y $\overline{Y}$ son las medias muestrales en ambos grupos, $S_X^2$ y $S_Y^2$ las varianzas muestrales, y $n$ el número de observaciones en cada grupo (25 en este caso).

4. *Identificar la distribución del estadístico de contraste bajo los supuestos del modelo*. Por ejemplo, se ha comprobado que el estadístico $t$ en la Ecuación (1) se distribuye según la distribución $t$ de Student con 48 (en general: $2n - 2$) grados de libertad.[34]
5. *Calcular, bajo el supuesto de la hipótesis nula, el valor del estadístico de contraste en la muestra observada*. Supongamos que el investigador observó que la media en el grupo que asistió a las clases presenciales era de $\bar{x} = 13$ y en el grupo del curso a distancia $\bar{y} = 9$ y que las varianzas observadas son $S_X^2 = 30$ y $S_Y^2 = 45$. En este caso,

$$t_{obs} = \frac{(\bar{x} - \bar{y}) - (\mu_{\bar{x}} - \mu_{\bar{y}})}{\sqrt{\frac{s_x^2 + s_y^2}{n}}}$$

$$= \frac{(13 - 9) - 0}{\sqrt{\frac{30+45}{25}}} = 2.309$$

6. *Calcular la probabilidad de observar $t_{obs}$ o un valor más extremo en la distribución de referencia*. Dicha probabilidad es el valor $p$. En este caso se consideran todos los valores mayores a 2.309 y menores a $-2.309$ más extremos que el valor observado. En la distribución $t$ de Student con 48 grados de libertad, la probabilidad de observar un valor más extremo que $t_{obs}$ es 0.03.
7. *Aceptar o rechazar la hipótesis nula*. Si el valor $p$ es menor que el criterio $\alpha$ de significancia (especificado *a priori*), se rechaza la hipótesis nula; en el caso contrario se acepta. Usualmente se elige $\alpha = 0.05$; en el ejemplo se rechazaría la hipótesis nula.

Para una interpretación correcta del valor $p$ es indispensable tener claro que la PSHN, se desarrolló dentro del marco *frecuentista* (o clásico) de la estadística. Dos ideas son fundamentales en esta perspectiva. Primero, los parámetros del modelo estadístico se consideran constantes, es decir, tienen un valor determinado y fijo; segundo, conceptualmente sería posible repetir el experimento un número infinito de veces (de ahí, "frecuentista"). En las diferentes repeticiones, los parámetros tienen el mismo valor pero las muestras (y, por lo tanto, los datos) fluctúan. La distribución del estadístico de contraste precisamente describe cómo el valor de dicho estadístico, variaría entre las diferentes repeticiones del experimento.

Entonces, el valor $p$ se interpreta como la proporción de veces, en el número infinito de repeticiones conceptuales del experimento, que el estadístico de contraste tiene un valor tan extremo o más extremo que el valor observado en la ejecución del experimento actual. Al respecto es primordial entender que estas repeticiones se realizan (y que la distribución del estadístico de contraste se deriva) bajo los supuestos del modelo estadístico y la hipótesis nula. El valor $p$ es la probabilidad de observar un resultado tan raro o más raro que en la muestra actual, *condicional a que el modelo estadístico y la hipótesis nula sean correctos*. Si el valor $p$ es (muy) bajo (por ejemplo $p < 0.05$), entonces hay dos posibilidades: hemos observado un resultado (muy) raro, o bien, los supuestos del modelo estadístico o la hipótesis nula no son correctos. En este caso, generalmente, se opta por rechazar la hipótesis nula. En nuestro ejemplo, se excluye el valor de cero como plausible valor para la diferencia entre las dos poblaciones (médicos que toman el curso de forma asistencial y aquellos que lo toman en modo virtual), respecto del aprovechamiento del curso (es decir, el cambio pre-post).

El procedimiento que se acaba de describir se acerca más al paradigma desarrollado por Fisher. La perspectiva de Neyman y Pearson sigue las mismas líneas, pero es importante resaltar las siguientes dos diferencias:

1. Neyman y Pearson se ciñeron estrictamente al principio frecuentista. Según ellos, no es posible concluir nada sobre un experimento en particular (siempre es posible que se haya observado un resultado "raro"); únicamente se sabe que, a largo plazo, al realizar muchas PSHN, se equivoca sólo en un porcentaje bajo de los casos (no mayor al nivel de significancia $\alpha$). Estos autores aconsejaron mencionar como resultado de la PSHN únicamente $p < 0.05$, sin interpretar más allá el tamaño de $p$ (es decir, tanto para $p = 0.049$ como para $p = 0.0001$, el informe estadístico simplemente incluye $p < 0.05$). Fisher, por otro lado, opinó que "ningún investigador mantiene un nivel de significancia fijo con el cual rechaza las hipótesis año tras año, y en todas las circunstancias; prefiere interpretar cada caso particular a la luz de la evidencia",[35] y defendió la interpretación del valor $p$ como una cuantificación de la evidencia, aportada por el experimento actual, en contra de (el modelo y) la hipótesis nula (donde valores pequeños para $p$ aportan más evidencia en contra).
2. El paradigma de Neyman y Pearson considera, además de la hipótesis nula ($H_0$), también una hipótesis alternativa ($H_1$). Esto convierte la PSHN en un algoritmo para la toma de decisiones ($H_0$ *vs* $H_1$). En este sentido contrasta con el enfoque de Fisher, donde la PSHN se puede interpretar como un procedimiento para validar un modelo (incluida la hipótesis nula). Puesto que la hipótesis alternativa afecta también las particularidades de la prueba, los resultados de los enfoques de Fisher y Neyman-Pearson pueden diferir. Para una excelente discusión sobre este punto, véase Christensen.[10]

Cabe mencionar que la hipótesis alternativa propició que Neyman y Pearson introdujeran algunos nuevos conceptos que son parte de cualquier curso actual de estadística inferencial, incluyendo el *Error Tipo I* (rechazar una hipótesis nula que es cierta), *Error Tipo II* (aceptar una hipótesis nula que es falsa) y la *potencia* (la probabilidad

de que se rechace la hipótesis nula, cuando la hipótesis alternativa es cierta). La recomendación de presentar valores *p* exactos (en vez de simplemente $p < 0.05$ según Neyman y Pearson), la cual fue adoptada por la mayoría de las revistas científicas, junto con el énfasis en la potencia y el tamaño de la muestra al diseñar un estudio, es testimonio de que la PSHN -como se practica hoy en día- es verdaderamente un híbrido de las ideas de Fisher, por un lado, y de Neyman y Pearson, por otro.

## Tres críticas para la PSHN y su relevancia para la medicina

### La PSHN no proporciona la información que los investigadores realmente quieren

El investigador en nuestro ejemplo busca apoyo para la hipótesis de que los participantes aprovechan mejor los cursos de actualización, si asisten de forma presencial en vez de si lo toman de forma virtual. Específicamente quiere saber cuánta evidencia aporta el experimento para su hipótesis. Interpreta el resultado de la PSHN como que "es poco plausible, a la luz de los datos observados, que la hipótesis nula sea cierta", y concluye que la diferencia entre ambas formas de enseñanza no es nula.

Hay varios puntos que comentar sobre esta forma de llegar a conclusiones derivadas de una PSHN. En primera instancia, nótese que el valor *p*, al excluir el valor de cero como valor plausible para el parámetro, no aporta información sobre los valores que sí son plausibles. Sin embargo, pocas veces interesa meramente falsear la hipótesis de efecto nulo o diferencia nula (estudios sobre percepción extrasensorial sirven tradicionalmente como contraejemplo, pero incluso en esta área no todo el mundo está convencido del mérito de simplemente refutar la hipótesis nula).[36] En el ejemplo, una diferencia promedio real de 0.1 entre las puntuaciones obtenidas por las dos estrategias de enseñanza puede ser enteramente irrelevante; sin embargo, la PSHN no provee ninguna información sobre la plausibilidad de este valor 0.1. En otras palabras, significancia estadística no implica significancia práctica o significancia clínica.[16,32,37-39] Ni siquiera un "valor *p extremadamente* significativo" hace más que excluir el cero como valor plausible para el parámetro.

Además, incluso esta interpretación minimalista ha sido atacada por varios autores.[12,16,40] Interpretaciones del valor *p* en términos de "evidencia en contra de la hipótesis nula" (como lo enunció Fisher mismo) o "la plausibilidad de que la hipótesis nula sea falsa" pueden llegar fácilmente a ser entendidas como la probabilidad de que la hipótesis nula sea falsa en consideración de la evidencia (E) actual, lo cual formalmente corresponde con $P(H_0|E)$. No obstante, interpretar el valor *p* como esta última probabilidad es un error lógico por dos razones. Primera, como se explicó en la sección anterior, el valor *p* se define dentro del marco frecuentista, donde los parámetros son constantes, es decir, no tienen una distribución probabilística (ni siquiera desconocida) y, por consiguiente, no tiene sentido asignar probabilidades a valores distintos del parámetro. Segunda, también explicado en la sección anterior, el valor *p* se calcula *bajo el supuesto de que la hipótesis nula es cierta*; por lo tanto, es imposible interpretarlo como la probabilidad de que la hipótesis nula sea cierta.

La probabilidad a la que se refiere el valor *p* guarda más relación con la probabilidad inversa, $P(E|H_0)$. Esto se conoce como la *verosimilitud*, es decir, la probabilidad de observar los datos que se han observado, condicionales a los supuestos en el modelo estadístico y la hipótesis nula. La probabilidad que realmente interesa -por ejemplo, al investigador de nuestro ejemplo- es la anteriormente mencionada $P(H_0|E)$. Aunque no está definida dentro del marco frecuentista, en el marco Bayesiano sí se define y se conoce como la *probabilidad posterior* (es decir, después de observar la evidencia). Las probabilidades $P(E|H_0)$ y $P(H_0|E)$ no son iguales, pero ¿cuál es la relación entre ambas?

Al respecto, es muy famosa la paradoja de Lindley,[41] quien mostró, dentro de un marco Bayesiano, que existen datos que rechazan una hipótesis nula con un bajo valor *p* y al mismo tiempo llevan a una probabilidad posterior alta. Por ejemplo, es perfectamente posible que, a partir de los mismos datos, se obtenga simultáneamente $P(E|H_0) = 0.05$ y $P(H_0|E) = 0.95$. Obviamente, este resultado pone en duda la validez de la interpretación del valor *p* como "evidencia en contra de la hipótesis nula". Los defensores de la PSHN señalan que la paradoja requiere muestras grandes para manifestarse, y se oponen a los supuestos adicionales que requiere el análisis Bayesiano. Nickerson[32] ha argumentado que bajo condiciones razonables, un bajo valor *p* generalmente conlleva una baja probabilidad posterior (es decir, poca evidencia) para la hipótesis nula. Sin embargo, a pesar de esta correlación entre el valor *p* y la probabilidad posterior, Berger y Sellke[42] han demostrado convincentemente que, incluso bajo los supuestos más favorables para el enfoque frecuentista, los valores *p* sistemáticamente sobreestiman la evidencia en contra de la hipótesis nula. Especialmente cuando $0.01 < p < 0.05$, la conclusión basada en la probabilidad posterior generalmente apunta a evidencia pobre o meramente anecdótica[43] (para un reanálisis de valores *p* en estudios publicados y su apreciación en términos de la evidencia que aportan, véase Wetzels y cols.).[44] Recientemente, Held[45] propuso un método gráfico para relacionar el valor *p* con la probabilidad posterior.

En resumen, el valor *p* resultado de una PSHN, no aporta mucha información de interés para los investigadores. En primera instancia, aceptar la hipótesis nula no permite ninguna conclusión sustancial. Además, si el valor *p* es significativo, únicamente se excluye un solo valor como valor plausible para el parámetro. Aunado a esto, el significado de "plausible" en la última expresión tiene una relación nebulosa con la probabilidad que sí le interesa a los investigadores: la probabilidad posterior de que la hipótesis del estudio sea cierta a la luz de la evidencia recopilada.

### La PSHN transforma un continuo en una decisión dicotómica con base en un criterio arbitrario

Supongamos que la PSHN para el experimento en nuestro ejemplo hubiese resultado en $p = 0.06$. En este caso, el investigador probablemente hubiese puesto bastante más reserva en sus interpretaciones, posiblemente

concluyendo que se requiere más investigación para mostrar la eficacia superior de la enseñanza en modo presencial. Algunos autores[12,19,46,47] han criticado la práctica de distinguir tan radicalmente entre una "diferencia real" y una "vicisitud del azar", dependiendo de que el valor *p* sea justamente menor o justamente mayor que el límite establecido. Más probablemente, dicha práctica tiene su origen en la perspectiva de Neyman y Pearson, que promovieron la PSHN como una metodología de toma de decisiones y que recomendaron rechazar o no la hipótesis nula y actuar correspondientemente. Rozeboom[12] se opone a esta práctica argumentando que el objetivo final de una investigación científica no es una decisión absoluta entre aceptar o rechazar una teoría, sino que se requiere evaluar y cuantificar la credibilidad posexperimental de las hipótesis derivadas de la misma. Algunos críticos de la PSHN incluso opinan que la interpretación absoluta del resultado de la PSHN ha tenido un efecto adverso en el progreso de la ciencia.[48,49]

Aunado a la crítica de dicotomizar el resultado continuo de una PSHN, hay inquietud sobre la arbitrariedad del criterio en el nivel de significancia $\alpha$.[12,50-52] Usualmente, se fija $\alpha = 0.05$, aunque ni siquiera los arquitectos de la metodología defendieron un único nivel de significancia para todos los casos (véase la cita de Fisher en la primera sección y Neyman[53] también manifestó que el nivel de significancia es elección del investigador). Por ejemplo, al realizar un estudio piloto para explorar el posible beneficio de una nueva medicina para la cual existen sospechas de que genere efectos secundarios peligrosos, es lógico exigir que los datos alcancen un nivel de significancia mucho más estricto (digamos, $\alpha = 0.001$), antes de llevar la investigación a una siguiente fase. No hay problema con que se elija $\alpha$ "de forma arbitraria", si esto quiere decir "en función del estudio particular que se está planeando".

Por otro lado, sí es problemático elegir $\alpha$ en función de los resultados, y particularmente del valor *p*, obtenidos.[54,55] A veces, se lee en un mismo reporte $p < 0.05$ y $p < 0.01$, lo cual sugiere que el autor adaptó el nivel de significancia al valor *p* hallado en el estudio (por ejemplo, porque encontró $p = 0.03$ en el primer caso, y $p = 0.006$ en el segundo). Esto es una práctica errada: o bien se incluye el valor *p* exacto en el reporte (lo cual permite al lector evaluar la evidencia con su propio nivel de significancia), o bien se fija el nivel de significancia $\alpha$ *a priori* y se presenta $p < \alpha$ o $p > \alpha$. Al respecto, es muy pertinente la aclaración de Engman[55] cuando señala que, después de haber rechazado la hipótesis nula con $\alpha = 0.05$, ya es irrelevante el valor *p* concreto (por ejemplo, si $p = 0.04$ o $p = 0.0002$). Aunque esta conclusión se adhiere a la norma ortodoxa de Neyman y Pearson, hay que reconocer que es efectivamente incoherente seguir interpretando el valor *p*, el cual se define bajo el supuesto de que la hipótesis nula es cierta, después de haber rechazado la misma.

## El resultado de una PSHN depende de decisiones e intenciones, posiblemente no explícitas o desconocidas, en el plan de estudio

Volvamos a suponer que el experimento de nuestro investigador hubiese resultado en un valor $p = 0.06$. ¿Cómo procedería puesto que el resultado no alcanzó el nivel de significancia tradicional de $\alpha = 0.05$? Una de las opciones que muchos investigadores tendrían en cuenta es la de ampliar la muestra, en consideración de que ampliar la muestra generalmente aumenta la probabilidad de rechazar una hipótesis nula falsa.

Sin embargo, varios autores[31,56,57] han señalado que esta decisión implicaría un cambio en cómo se debe calcular el valor *p* de la PSHN. La razón es que se cambia el modelo estadístico en el que se fundamenta la PSHN y, por lo tanto, el conjunto infinito de repeticiones conceptuales del experimento con que se compara (a través del estadístico de contraste) el experimento actual. Para ilustrar este punto, supongamos que el investigador opta por la siguiente estrategia: si la PSHN inicial, basada en la muestra de 50 personas, resulta en $p \leq 0.05$ o bien $p > 0.20$, el experimento se detiene; si, por otro lado, encuentra $0.05 < p \leq 0.20$, entonces extiende la muestra con 24 personas más (de las cuales 12 reciben el curso de modo presencial y las otras 12 a distancia), y se recalcula el valor *p*, basado en los 74 participantes. Por consiguiente, el conjunto de repeticiones del experimento incluye cuatro tipos de resultados: (i) $p \leq 0.05$ con $n = 50$, (ii) $p \leq 0.05$ con $n = 74$, (iii) $p > 0.05$ con $n = 74$ y (iv) $p > 0.20$ con $n = 50$. Bajo el supuesto de que la hipótesis nula es cierta, se puede calcular que las probabilidades correspondientes de rechazar la hipótesis nula son 0.05, 0.02, 0.13 y 0.80, respectivamente. Es decir, la probabilidad de rechazar la hipótesis nula, siendo cierta, es $0.05 + 0.02 = 0.07$. Aunque este valor parece no diferir mucho del valor nominal de $\alpha = 0.05$, el ejemplo enseña que el cambio de estrategia requiere un ajuste del valor *p* para cada una de las dos pruebas (la primera basada en 50 personas y la segunda con 74 personas), para que el nivel de significancia global corresponda con $\alpha = 0.05$ (además, como señalaron Lindley[41] y Armitage y cols.,[58] si existe la posibilidad de extender ilimitadamente la muestra, paso por paso, la probabilidad de rechazar la hipótesis nula, aunque sea cierta, se acerca a 1).

Por la naturaleza de la investigación en las ciencias médicas, la interrupción/extensión de un experimento no es la excepción. Por ejemplo, si un estudio incluye a pacientes como participantes, éstos típicamente llegan secuencialmente y el experimento se aplica a todos aquellos que lleguen a algún centro de atención en un periodo de, por ejemplo, seis meses. Si los resultados disponibles para una parte de los pacientes indican claramente un efecto adverso del tratamiento bajo estudio, sería poco ético seguir con el experimento. La práctica de realizar evaluaciones intermedias se conoce como la estrategia de *análisis secuenciales*, para la cual se han desarrollado adaptaciones a los procedimientos estadísticos tradicionales para que los valores *p* correspondan con los niveles de significancia nominales.[59-64]

El ejemplo anterior y las adaptaciones a los procedimientos para análisis secuenciales ponen en evidencia que para la PSHN, es necesario involucrar el diseño muestral del estudio al calcular el valor *p*. Es decir, aunque los datos sean idénticos (por ejemplo, 50 participantes del curso estudiados con el plan original, o bien, los mismos 50 participantes estudiados bajo la estrategia adaptada), los valores *p* resultan diferentes para los distintos diseños muestrales.

Sin embargo, no sólo el diseño muestral, también otros aspectos y decisiones del plan de estudio pueden afectar el resultado de una PSHN. Unos ejemplos incluyen:

- ¿Qué haría el investigador si los datos de algunos participantes resultasen poco confiables, por ejemplo, porque sospecha que durante la administración de las pruebas copiaron sus respuestas? ¿Les volvería a aplicar las mismas pruebas en otro momento? ¿Reemplazaría estos participantes con otros? ¿Eliminaría simplemente sus datos y terminaría el estudio con un número menor que los 50 participantes inicialmente planeados?
- ¿Cómo trataría los datos faltantes, por ejemplo, porque, por cualquier razón, algunos participantes no podían participar en una o más de las cuatro pruebas? ¿Usaría métodos de imputación? ¿Qué método usaría?
- ¿Cómo procedería el investigador si la PSHN diese un resultado "marginalmente" significativo (como el $p$ = 0.06 mencionado anteriormente)? Además de ampliar la muestra, podría decidir analizar los datos de otra forma, por ejemplo, evaluando la puntuación de cada prueba por separado, o analizar las puntuaciones de las pruebas mediante un análisis de varianza multivariada (MANOVA).
- ¿Realizará pruebas de significancia bilaterales o unilaterales? En el caso de que se investigasen otros aspectos de los datos, ¿consideraría aplicar ajustes para comparaciones múltiples (tipo Bonferroni, Duncan o Tukey)? ¿Qué método elegiría?

En todos estos casos, las decisiones afectan, en mayor o menor medida, al conjunto de repeticiones hipotéticas del experimento con las cuales se compara el experimento actual y, por consiguiente, el valor $p$. En este sentido, se puede decir que el resultado de la PSHN depende de datos que no han sido observados, o en palabras de Sir Harold Jeffreys, uno de los pioneros de la estadística Bayesiana del siglo XX:

"Lo que el uso del $p$ implica, entonces, es que una hipótesis que puede ser cierta, puede ser rechazada porque no ha predicho resultados observables que no ocurrieron. *Esto parece un procedimiento extraño*".[65]

En general, los investigadores no planean de antemano todas las decisiones que tomarán durante la ejecución de un experimento y el análisis de los datos. Además, en muchas ocasiones es simplemente imposible contemplar todos los imprevistos que pudieran ocurrir. Esto no significa que un investigador no deba explorar sus datos más allá de lo originalmente planeado, o tomar decisiones en función de los resultados que obtenga. Más bien, es importante tener claro que estas intenciones y decisiones no explícitas influyen en el resultado de la PSHN -independientemente de que realmente se apliquen o no en el experimento actual-, lo cual constituye una fuerte crítica para las pruebas de significancia dentro del marco frecuentista.

## Remedios y alternativas para la PSHN

### Intervalos de confianza

Con el fin de resolver algunos de los inconvenientes de la PSHN, varios autores han recomendado acompañar o sustituir el valor $p$ por un intervalo de confianza (IC).[16,66-69] Un IC se define por el conjunto de valores para los cuales no se rechaza la hipótesis nula. En nuestro ejemplo, es como si se realizara una prueba de significancia de la hipótesis nula $H_0 : \mu_X - \mu_Y = c$, para todos los números reales $c$, y se anotara para cuáles valores de $c$ la prueba no rechaza. En este caso, se obtendría que para $0.52 \leq c \leq 7.48$ no se rechaza la hipótesis nula y en consideración que fijamos $\alpha = 0.05$, se dice que [0.52, 7.48] es un IC de 95% para la diferencia entre las medias de las dos poblaciones. Como tal, la práctica de presentar ICs (posiblemente en conjunto con el valor $p$) constituye una respuesta a la crítica de que la PSHN excluye únicamente un valor como valor plausible para el parámetro. Además de proporcionar información sobre significancia (si el IC no incluye el valor de cero, entonces una PSHN declararía el resultado estadísticamente significativo), el IC informa también sobre el posible tamaño del efecto.

En términos generales, se podría decir que el IC representa los valores plausibles para el parámetro. Sin embargo, es importante mantener claro el enfoque frecuentista en el que el IC se inscribe, lo cual implica que el parámetro es una constante.[70] Por lo tanto, la interpretación del IC de nuestro ejemplo como que "el valor del parámetro se encuentra con 95% de probabilidad entre 0.52 y 7.48", a pesar de ser muy común entre los usuarios de la técnica, es incorrecta. El problema con esta interpretación es que no reconoce que son los límites del IC los que varían (entre las repeticiones conceptuales del experimento), no el valor del parámetro, y que la variación deja de existir al momento de fijar dichos límites (en otras palabras, considerando el número infinito de repeticiones conceptuales del experimento, el valor del parámetro se encuentra o bien en ninguno de los casos, o bien en todos, entre 0.52 y 7.48; es decir, la probabilidad es 0 o 1). La interpretación correcta del IC apela al principio frecuentista: "A largo plazo, realizando muchos experimentos, los ICs incluirán el valor del parámetro en por lo menos el 95% de los casos".

Es claro que los intervalos de confianza integran más información que el mero $p < 0.05$ como resultado de una PSHN, información que es valiosa a la hora de evaluar el impacto de un estudio. No obstante, puesto que el IC se define en términos de la PSHN, está abierto a las mismas críticas que dicho procedimiento.[18,31,71] Por ejemplo, la arbitrariedad en la selección del nivel de confianza (comúnmente se fija $\alpha = 0.05$, llevando a ICs de 95%), y la dependencia de las intenciones y decisiones en el plan de estudio son críticas que igualmente aplican a los ICs. Además, cabe mencionar que el IC no es una alternativa para la PSHN en todas las ocasiones, por ejemplo, en el caso de la popular prueba de *ji cuadrada*, que contrasta el supuesto de independencia entre variables o que una muestra se ha extraído de alguna familia de distribuciones (por ejemplo, la normal), no resulta tan claro cómo definir el IC que corresponde en este caso.

### Tamaño del efecto

En algún sentido, los ICs brindan información sobre el tamaño del efecto, por ejemplo, la diferencia entre la educación presencial y la educación a distancia en el estudio de nuestro investigador. Sin embargo, los números en el

IC se expresan en términos de las unidades originales con las que se expresan las medias de ambos grupos; alguien que desconoce los detalles del estudio (y, específicamente, si no está familiarizado con las pruebas de conocimiento aplicadas e ignora cómo se deriva la puntuación global), tendrá dificultades para valorar la significancia práctica de una diferencia entre 0.52 y 7.48. Usando otro ejemplo, al evaluar un tratamiento para reducir la fiebre en bebés que sufren alguna enfermedad (medida en grados centígrados), un efecto de 0.5 de diferencia con otro tratamiento sería espectacular, mientras que al evaluar el efecto de alguna medicina en la presión sanguínea sistólica (medida en milímetros de mercurio mmHg), la misma diferencia numérica de 0.5, sería despreciable. Sin conocimiento del error muestral o la desviación típica de la variable bajo estudio, es imposible interpretar adecuadamente los valores en un IC.

Por estas razones se han desarrollado estadísticas cuyos valores gozan de una interpretación directa, llamadas índices del tamaño del efecto.[72-77] Existen docenas de este tipo de índices que se pueden clasificar en cuatro grupos:[75]

a. *Los índices de diferencias entre grupos* expresan el tamaño de la diferencia entre dos, o más, grupos. Ejemplos incluyen la *d* de Cohen[78] y la Δ de Glass.[79] Se trata de diferencias estandarizadas, es decir, se interpretan en términos de la desviación estándar que existe dentro de los grupos.
b. *Los índices de la fuerza de asociación* usualmente cuantifican la cantidad de varianza que dos o más variables comparten. Como ejemplo más conocido de índices en esta categoría destaca la correlación *r* de Pearson. Otros ejemplos incluyen la correlación múltiple *R*, $\eta^2$,[80] el coeficiente $\varphi$ y el *V* de Cramér.[81]
c. *Los índices ajustados* incluyen una corrección de los índices de la categoría anterior tomando en cuenta el error muestral o la varianza explicada por otras variables, por ejemplo, la correlación múltiple ajustada, $\varepsilon^2$ y $\omega^2$ (que son alternativas para $\eta^2$),[82] y la correlación parcial.
d. *Las estimaciones de riesgo* comparan el riesgo para algún resultado (por ejemplo, fallecimiento después de un infarto) entre dos, o más grupos. Ejemplos incluyen el riesgo relativo (RR) y la razón de momios (OR, por sus siglas en inglés, *odds ratio*).[83]

La elección de uno de los índices del tamaño del efecto depende de varios factores, incluyendo el tipo de variable (categórica o continua), el tipo de análisis (análisis de varianza, análisis de regresión, etc.) y también el área de estudios (por ejemplo, en las ciencias médicas y de salud son muy comunes los distintos índices de estimaciones de riesgo, mientras que en psicología se presentan con mucha frecuencia índices de diferencias entre grupos). Cabe señalar que se han propuesto procedimientos para transformar entre sí los diferentes tipos de índices (para algunos ejemplos, véase Borenstein).[76]

Obviamente, los índices del tamaño del efecto se calculan a partir de los datos recopilados de la muestra, y como tal sirven para estimar el tamaño del efecto en la población. Es decir, se puede distinguir entre el tamaño del efecto como estadístico (calculado a partir de la muestra) y como parámetro (característica de una población). Siguiendo dentro del marco frecuentista, se pueden conceptualizar diferentes repeticiones del experimento y considerar la distribución muestral del (estadístico) tamaño del efecto. Esta lógica permite construir ICs para el tamaño del efecto en la población.[84]

## Adoptar el marco Bayesiano

Varios autores que han criticado la PSHN proponen como solución un cambio del paradigma frecuentista al paradigma Bayesiano.[31,65,85-87] Los estadísticos Bayesianos no hacen referencia a las repeticiones hipotéticas de un experimento para definir el concepto de probabilidad, sino que lo valoran en términos de la incertidumbre o el "grado de creencia" acerca del evento bajo consideración[88] (cabe mencionar que además de esta visión subjetivista, se ha propuesto otra definición de probabilidad Bayesiana, basada en el teorema de Cox, que se ha llamado la visión *objetivista*).[89] La interpretación Bayesiana permite conceptualizar de igual forma la probabilidad de que al tirar una moneda aparezca cruz, la probabilidad de que México gane 10 o más medallas en los próximos juegos olímpicos, y la probabilidad de que supere a algún valor de interés la diferencia entre las medias poblacionales en las pruebas de conocimiento aplicadas a médicos que recibieron un curso de actualización de modo presencial y a aquellos que participan a distancia.

Una discusión profunda de los conceptos básicos de la estadística Bayesiana está fuera del alcance de este artículo; el lector interesado puede consultar el libro de texto de Bolstad[90] o el de Gelman y cols.[88] A continuación, introduzco algunos elementos clave que permitirán comprender el papel del enfoque Bayesiano en el debate sobre la PSHN.

Primero, la distribución (de probabilidad o densidad) *posterior*, $P(\theta|X)$, describe la incertidumbre respecto de los valores de un parámetro $\theta$, después de haber tomado en cuenta los datos observados *X*. La interpretación de la distribución posterior, que permite derivar conclusiones válidas en términos de, por ejemplo, "la probabilidad de que el valor del parámetro supere a 0", es sencilla y corresponde directamente con el interés del investigador. Además de la distribución posterior, el enfoque Bayesiano incluye también la distribución *previa*, $P(\theta)$. Describe la incertidumbre sobre el parámetro $\theta$ que el investigador tiene antes de haber considerado los datos. El teorema de Bayes relaciona la distribución posterior con la distribución previa:

$$P(\theta|X) = \frac{P(X|\theta)\ P(\theta)}{P(X)}$$

En esta ecuación, $P(X|\theta)$ es la *función de verosimilitud* (que da la probabilidad de los datos condicional al valor del parámetro); el término en el denominador $P(X)$ representa la probabilidad incondicional de los datos. Para las aplicaciones Bayesianas es conveniente reconocer que $P(X)$ es una constante, en el sentido de que no depende de $\theta$ (es decir, no influye en la plausibilidad relativa de los valores para $\theta$; simplemente garantiza que el área total debajo de la distribución posterior sea igual a

1, similar al término $1/\sqrt{2\pi}$ en la fórmula de la distribución normal estandarizada). Es la razón por qué a menudo, se encuentra en los textos con el enfoque Bayesiano la siguiente expresión:

$$P(\theta|X) \propto P(X|\theta)\,P(\theta)$$

donde $\propto$ se lee como "es proporcional a".

Los opositores del enfoque Bayesiano rechazan las conclusiones que se sacan de la distribución posterior, ya que las consideran subjetivas por su dependencia de la distribución previa. Efectivamente, diferentes investigadores pueden tener creencias previas distintas sobre los valores plausibles de un parámetro y, por lo tanto, pueden llegar a conclusiones diferentes. Sin embargo, como defensa a esta crítica, los Bayesianos han levantado los siguientes cinco argumentos:

a. Para cualquier análisis estadístico es crucial investigar cómo las conclusiones varían en función de los supuestos del modelo utilizado. En el enfoque Bayesiano típicamente se examina qué tan sensibles son los resultados a la distribución previa (entre otros aspectos del modelo), lo cual implica reajustar el modelo especificando diferentes distribuciones previas y evaluar la robustez de las conclusiones.
b. En general, la sensibilidad de los resultados a la distribución previa disminuye conforme las muestras son más grandes.
c. Para muchos tipos de parámetros se han propuesto distribuciones previas que afectan mínimamente la distribución posterior, denominadas distribuciones previas *no informativas*.[91] Una aproximación relacionada utiliza distribuciones previas *jerárquicas*, lo cual implica que se especifica una familia de distribuciones (en vez de una distribución previa particular), cuyos parámetros se estiman a partir de los datos recopilados.
d. Algunos autores propugnan por la necesidad de considerar información previa al evaluar la evidencia proporcionada por un estudio, enfatizando la importancia del contexto en el cual se realiza la investigación y reconociendo el carácter acumulativo de la ciencia.[18,85] En este sentido, la distribución previa se convierte en una ventaja del enfoque Bayesiano.
e. Dentro del enfoque Bayesiano, se han desarrollado índices que no dependen de la distribución previa. Un ejemplo es el factor de Bayes, que se define en el contexto de contrastar dos modelos (por ejemplo, una "hipótesis nula" y una "hipótesis alternativa") y que indica, como una razón de momios, cuánto apoyo brindan los datos a un modelo en comparación con el otro.[87,92]

A pesar de que el enfoque Bayesiano parece resolver la mayoría de los inconvenientes de la PSHN y los valores *p*, su uso sigue siendo escaso. Las razones por la resistencia hacia la incorporación de métodos Bayesianos en la investigación incluyen (a) la enseñanza académica, que en la mayoría de las facultades se limita a la estadística clásica y que lleva a (b) la poca familiaridad de los investigadores con los conceptos Bayesianos, y (c) la escasa disponibilidad de programas de computo amigables para los análisis Bayesianos. A pesar de que las versiones recientes de los paquetes estadísticos populares como SAS, SPSS y Stata incluyen módulos que permiten emprender análisis Bayesianos, y a pesar de la popularización del software WinBUGS/OpenBUGS (BUGS significa *Bayesian inference Using Gibbs Sampling*), un obstáculo importante sigue siendo la participación que se requiere del usuario, que es inevitable en un análisis Bayesiano (en comparación con un análisis clásico que se puede realizar con un par de clics al ratón, sin conocer a fondo los procedimientos detrás de ellos). Reconociendo este problema, algunos autores[31,93] han desarrollado adaptaciones de índices Bayesianos que son fácilmente calculables a partir de la salida de un análisis clásico y que simultáneamente ofrecen la ventaja de una interpretación Bayesiana.

## Discusión

En este artículo se revisaron algunas de las críticas dirigidas a las pruebas de significancia desarrolladas dentro del marco clásico de la estadística y se dieron algunos ejemplos para ilustrar su relevancia en el contexto médico. Se trata de una selección, ya que las críticas hacia la PSHN son diversas y sacan a la luz tanto problemas de interpretación como de construcción formal. A pesar de la intensidad que tuvo el debate durante las últimas décadas del siglo pasado, al parecer hoy en día éste ha disminuido y actualmente la mayoría de los expertos reconoce las limitaciones del procedimiento. No obstante, las pruebas de significancia siguen siendo la herramienta más popular en la estadística aplicada, tanto en las ciencias médicas y de salud como en las ciencias sociales y del comportamiento. Probablemente esta tenacidad se deba a que no existe "una alternativa mágica a la PSHN, algún otro ritual mecánico objetivo para reemplazarlo"[16] y que la alternativa principal -adoptar el enfoque Bayesiano- no permite un análisis automatizado, que lleve a un resultado sin que el investigador intervenga. En este contexto cabe señalar que varios autores[94-97] rompen una lanza en favor de una educación en estadística que otorga plena importancia al entendimiento correcto de los conceptos básicos (contrario a la enseñanza de recetas -¿en qué botones hay que dar clic?- para llevar a cabo un análisis estadístico); promueven una educación que se fundamenta en el contexto de la investigación científica y, cuando está dirigida a futuros médicos, incluye ejemplos relevantes de la práctica clínica. Invitar a los estudiantes a que reflexionen críticamente sobre los procedimientos estadísticos, es un primer paso para que las diversas herramientas se apliquen con sensatez e inteligencia.

## Agradecimientos

El autor agradece a Pablo Cáceres Serrano de la Pontificia Universidad Católica de Valparaíso, Chile, por sus valiosos comentarios y sugerencias sobre una versión anterior del presente artículo.

## Referencias


1. Cowles M, Davis C. On the origins of the .05 level of statistical significance. Am Psychol 1982;37:553-558.
2. Fisher RA. Statistical methods for research workers. Edimburgo, Reino Unido. Oliver and Boyd. 1925.
3. Fisher RA. The design of experiments. Edimburgo, Reino Unido. Oliver and Boyd. 1935.
4. Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. Biometrika 1928;20A:175-240.
5. Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. Biometrika 1928;20A:263-294.
6. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Phil Trans R Soc Lond A 1933;231:289-337.
7. Lehmann EL. The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? J Am Stat Assoc 1993;88:1242-1249.
8. Berger JO. Could Fisher, Jeffreys and Neyman have agreed on testing? Stat Sci 2003;18:1-32.
9. Hubbard R, Bayarri MJ. Confusion over measures of evidence (*p*'s) versus errors ($\alpha$'s) in classical statistical testing. Am Stat 2003;3:171-182.
10. Christensen R. Testing Fisher, Neyman, Pearson, and Bayes. Am Stat 2005; 59:121-126.
11. Berkson J. Some difficulties of interpretation encountered in the application of the chi-square test. J Am Stat Assoc 1938;33:526-536.
12. Rozeboom WW. The fallacy of the null-hypothesis significance test. Psychol Bull 1960;57:416-428.
13. Meehl PE. Theory-testing in psychology and physics: A methodological paradox. Philos Sci 1967;34:103-115.
14. Bakan D. The test of significance in psychological research. Psychol Bull 1966;66:423-437.
15. Carver RP. The case against statistical significance testing. Harv Educ Rev 1978;48:378-399.
16. Cohen J. The earth is round ($p < .05$). Am Psychol 1994;49:997-1003.
17. Thompson B. In praise of brilliance: Where that praise really belongs. Am Psychol 1998;53:799-800.
18. Goodman SN. Toward evidence-based medical statistics. 1: The *P* value fallacy. Ann Intern Med 1999;130:995-1004.
19. Gigerenzer G. Mindless statistics. J Socio Econ 2004;33:587-606.
20. Hubbard R, Lindsay RM. Why *P* values are not a useful measure of evidence in statistical significance testing. Theory Psychol 2008;18:69-88.
21. McCloskey DN, Ziliak ST. The unreasonable ineffectiveness of Fisherian "tests" in biology, and especially in medicine. Biol Theory 2009;4:44-53.
22. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol 2010;25:225-230.
23. Lambdin C. Significance tests as sorcery: Science is empirical-significance tests are not. Theory Psychol 2012;22:67-90.
24. Cortina JM, Dunlap WP. On the logic and purpose of significance testing. Psychol Methods 1997;2:161-172.
25. Hagen RL. In praise of the null hypothesis statistical test. Am Psychol 1997; 52:15-24.
26. Chow SL. Précis of statistical significance: Rationale, validity, and utility (con discusión). Behav Brain Sci 1998;21:169-239.
27. Gregg AP, Sedikides C. Is social psychology research really so negatively biased? Behav Brain Sci 2004;27:340-341.
28. Mogie M. In support of null hypothesis significance testing. Proc R Soc Lond B Biol Sci 2004;271:S82-S84.
29. Stephens PA, Buskirk SW, Hayward GD, Martínez del Rio C. Information theory and hypothesis testing: A call for pluralism. J Appl Ecol 2005;42:4-12.
30. Stephens PA, Buskirk SW, Martínez del Rio C. Inference in ecology and evolution. Trends Ecol Evol 2007;22:192-197.
31. Wagenmakers EJ. A practical solution to the pervasive problems of *p* values. Psychon Bull Rev 2007;14:779-804.
32. Nickerson RS. Null hypothesis significance testing: A review of an old and continuing controversy. Psychol Methods 2000;5:241-301.
33. Balluerka N, Gómez J, Hidalgo D. The controversy over null hypothesis significance testing revisited. Methodology 2005;1:55-70.
34. Mood AM, Graybill FA, Boes DC. Introduction to the theory of statistics. 3ª ed. Nueva York, EE.UU. McGraw-Hill. 1974.
35. Fisher RA. Statistical methods and scientific inferences. 3ª ed. Nueva York, EE.UU. Hafner. 1973.
36. Alcock JE. Give the null hypothesis a chance: Reasons to remain doubtful about the existence of psi. J Conscious Stud 2003;10(6-7):29-50.
37. LeFort SM. The statistical versus clinical significance debate. J Nurs Scholarsh 1993;25:57-62.
38. Kane RC. The clinical significance of statistical significance. Oncologist 2008;13:1129-1133.
39. Silva-Ayc¸aguer LC, Suárez-Gil P, Fernández-Somoano A. The null hypothesis significance test in health sciences research (1995-2006): Statistical analysis and interpretation. BMC Med Res Methodol 2010;10:44.
40. Falk R, Greenbaum CW. Significance tests die hard: The amazing persistence of a probabilistic misconception. Theory Psychol 1995;5:75-98.
41. Lindley DV. A statistical paradox. Biometrika 1957;44:187-192.
42. Berger JO, Sellke T. Testing a point null hypothesis: The irreconcilability of *P* values and evidence. J Am Stat Assoc 1987;82:112-122.
43. Sellke T, Bayarri MJ, Berger JO. Calibration of *p* values for testing precise null hypotheses. Am Stat 2001;55:62-71.
44. Wetzels R, Matzke D, Lee MD, et al. Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. Perspect Psychol Sci 2011;6:291-298.
45. Held L. A nomogram for *P* values. BMC Med Res Methodol 2010;10:21.
46. Rosnow RL, Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. Am Psychol 1989;44:1276-1284.
47. Hoekstra R, Finch S, Kiers HAL, et al. Probability as certainty: Dichotomous thinking and the misuse of *p* values. Psychon Bull Rev 2006; 13:1033-1037.
48. Savage IR. Nonparametric statistics. J Am Stat Assoc 1957;52:331-344.
49. Schmidt FL. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychol Methods 1996;1:115-129.
50. Glass GV, McGaw B, Smith ML. Meta-analysis in social research. Beverly Hills, CA, EE.UU. Sage. 1981.
51. Anderson DR, Burnham KP, Thompson WL. Null hypothesis testing: Problems, prevalence, and an alternative. J Wildl Manage 2000;64:912-923.
52. Mudge JF, Baker LF, Edge CB, et al. Setting an optimal ¤ that minimizes errors in null hypothesis significance tests. PLoS One 2012;7(2):e32734.
53. Neyman J. Basic ideas and some recent results of the theory of testing statistical hypothesis. J R Stat Soc 1942;105:292-327.
54. Goodman S. A dirty dozen: Twelve *P*-value misconceptions. Semin Hematol 2008;45:135-140.
55. Engman A. Is there life after $P < 0.05$? Statistical significance and quantitative sociology. Quality Quantity 2011, DOI: 10.1007/s11135-011-9516-z.
56. Berger JO, Berry DA. The relevance of stopping rules in statistical inference (con discusión). in: Gupta SS, Berger JO, (editors). Statistical Decision Theory and Related Topics IV. vol. 1. Nueva York, NY, EE.UU. Springer. 1988. 29-72.

57. Lindley DV. The analysis of experimental data: The appreciation of tea and wine. Teach Stat 1993;15:22-25.
58. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. J R Stat Soc Ser A 1969;132:235-244.
59. Wald A. Sequential analysis. Nueva York, NY, EE.UU. Wiley. 1947.
60. Armitage P. Sequential medical trials. 2ª ed. Nueva York, NY, EE.UU. Wiley. 1975.
61. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977;64:191-199.
62. Fairbanks K, Madsen R. *P* values for tests using a repeated significance test design. Biometrika 1982;69:69-74.
63. Whitehead J. The design and analysis of sequential clinical trials. 2ª ed. Statistics in Practice. Chichester, Reino Unido. Wiley. 1997.
64. Goodman SN. Stopping trials for efficacy: An almost unbiased view. Clin Trials 2009;6:133-135.
65. Jeffreys H. Theory of probability. 3ª ed. Oxford, Reino Unido. University Press. 1961.
66. Gardner MJ, Altman DG. Confidence intervals rather than *P* values: Estimation rather than hypothesis testing. Br Med J 1986;292:746-750.
67. Brandstätter E. Confidence intervals as an alternative to significance testing. Methods Psychol Res Online 1999;4(2):33-46.
68. Sterne JAC, Smith GD. Sifting the evidence-what's wrong with significance tests? Br Med J 2001;322:226-231.
69. Ranstam J. Why the *P*-value culture is bad and confidence intervals a better alternative. Osteoarthritis Cartilage 2012;20(8):805-808.
70. Howson C, Urbach P. Bayesian reasoning in science. Nature 1991;350:371-374.
71. Feinstein AR. *P*-values and confidence intervals: Two sides of the same unsatisfactory coin. J Clin Epidemiol 1998;52:355-360.
72. Thompson B. "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? J Couns Dev 2002;80:64-71.
73. Rosnow RL, Rosenthal R. Effect size for experimenting psychologists. Can J Exp Psychol 2003;57:221-237.
74. Vacha-Haase T, Thompson B. How to estimate and interpret various effect sizes. J Couns Psychol 2004;51:473-481.
75. Ferguson CJ. An effect size primer: A guide for clinicians and researchers. Prof Psychol Res Pr 2009;40:532-538.
76. Borenstein M. Effect sizes for continuous data. En: Cooper H, Hedges LV, Valentine JC, editores. The Handbook of Research Synthesis and Meta-Analysis. 2ª ed. Nueva York, NY, EE.UU. Russell Sage Foundation. 2009. 221-235.
77. Fleiss JV, Berlin JA. Effect sizes for dichotomous data. In: Cooper H, Hedges LV, Valentine JC, (editors). The Handbook of Research Synthesis and Meta-Analysis. 2ª ed. Nueva York, NY, EE.UU. Russell Sage Foundation. 2009. 237-253.
78. Cohen J. Statistical power analysis for the behavioral sciences. 2ª ed. Hillsdale, NJ. Lawrence Erlbaum. 1988.
79. Glass GV. Primary, secondary, and meta-analysis of research. Educ Res 1976;5(10):3-8.
80. Peters CC, Van Voorhis WR. Statistical procedures and their mathematical bases. Nueva York, NY, EE.UU. McGraw-Hill. 1940.
81. Cramér H. Mathematical methods of statistics. Princeton, NJ, EE.UU. University Press. 1946.
82. Hays WL. Statistics. 5ª ed. Nueva York, NY, EE.UU. Harcourt Brace. 1994.
83. Sistrom CL, Garvan CW. Proportions, odds, and risk. Radiology 2004; 230:12-19.
84. Thompson B. What future quantitative social science research could look like: Confidence intervals for effect sizes. Educ Res 2002;13(3):25-32.
85. Browner WS, Newman TB. Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. JAMA 1987;257:2459-2463.
86. Trafimow D. Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. Psychol Rev 2003;110:526-535.
87. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med 1999;130:1005-1013.
88. Gelman A, Carlin JB, Stern HS, et al. Bayesian data analysis. 2ª ed. Londres, Reino Unido. Chapman & Hall. 2004.
89. Jaynes ET. Probability theory: The logic of science. Cambridge, Reino Unido. University Press. 2003.
90. Bolstad WM. Introduction to Bayesian statistics. 2ª ed. Hoboken, NJ, EE.UU. John Wiley & Sons. 2007.
91. Kass RE, Wasserman L. The selection of prior distributions by formal rules. J Am Stat Assoc 1996;91:1343-1370.
92. Kass RE, Raftery AE. Bayes factors. J Am Stat Assoc 1995;90:773-795.
93. Johnson VE. Bayes factors based on test statistics. J R Stat Soc Series B Stat Methodol 2005;67:689-701.
94. Gelman A, Nolan D. Teaching statistics: A bag of tricks. Nueva York, NY, EE.UU. Oxford University Press. 2002.
95. Herman A, Notzer N, Libman Z, et al. Statistical education for medical students-Concepts are what remain when the details are forgotten. Stat Med 2007;26:4344-4351.
96. Freeman JV, Collier S, Staniforth D, et al. Innovations in curriculum design: A multi-disciplinary approach to teaching statistics to undergraduate medical students 2008;8:28.
97. Gelman A. Teaching Bayes to graduate students in political science, sociology, public health, education, economics,…. Am Stat 2008;62:202-205.