ARTÍCULOS ORIGINALES

Aplicación de las pruebas de hipótesis en la investigación en salud: ¿estamos en lo correcto?

PEDRO MONTERREY[1]
CARLOS GÓMEZ-RESTREPO[2]

Resumen

Introducción. Las pruebas de hipótesis son comúnmente utilizadas en medicina. No obstante, por lo general, no se conoce la historia de su desarrollo ni cómo se han generado a partir de posiciones filosóficas antagónicas que, en ocasiones, combinamos inconscientemente.

Objetivo. Describir la historia de las pruebas de hipótesis y profundizar en el manejo del valor de P y los intervalos de confianza.

Método. Se hace la revisión histórica y la aplicación estadística en algunos ejemplos clínicos.

Resultados. Se describe la historia de las pruebas de hipótesis con los planteamientos de Fisher contrarios a los de Neyman y Pearson. Se esbozan algunos ejemplos en que se observa la importancia de tener en cuenta cómo se analizan los datos y la información complementaria que provee el valor de P y los intervalos de confianza.

Conclusión. La presentación explícita de los valores de P, su uso combinado con los intervalos de confianza y la valoración de los resultados a la luz de su plausibilidad biológica, son los componentes centrales en el uso adecuado de las pruebas de hipótesis.

Palabras clave

Pruebas de hipótesis, valor de P, intervalos de confianza, plausibilidad biológica.

1 Ph.D. en Matemáticas, especialización en Epidemiología Nutricional; profesor asociado, Departamento de Epidemiología Clínica y Bioestadística, Pontificia Universidad Javeriana, Bogotá, D. C., Colombia

2 Médico psiquiatra, psicoanalista, M.Sc. en Epidemiología Clínica; profesor asociado, Departamento de Epidemiología Clínica y Bioestadística y del Departamento de Psiquiatría y Salud Mental, Pontificia Universidad Javeriana, Bogotá, D. C., Colombia

Recibido: abril 16/2007. Revisado: julio 14/2007. Aceptado: julio 31/2007.

Abstract

Introduction: Hypothesis testing tools are commonly used in medicine. Despite this, in general, there is no knowledge about how these tools were developed and how their generation has surged out of opposing philosophical positions that we commonly combine in an unconscious fashion.

Objective: To describe the history of hypothesis testing tools and to gain a stronger insight into the handling of the P value and confidence intervals.

Methods: Historical review and statistical application to some of the clinical examples mentioned.

Results: The history of hypothesis tests are described along with the opposing postulations of Fisher in regards to those of Neyman and Pearson. Some examples that show the importance of having in mind how data is analyzed and the additional information that p values and confidence intervals bring, are outlined.

Conclusion: The explicit presentation of P values, its combined use along with confidence intervals, and the analysis of the results having into account its biological plausibility are key components in the proper use of hypothesis testing.

Key words

Hypothesis testing, p value, confidence intervals, biological plausibility.

Introducción

Desde el siglo XIX y hasta principios del siglo XX las publicaciones científicas estaban plagadas de presentaciones de casos y análisis que se traducían en juicios subjetivos los que, en muchos casos, no eran relevantes y producían confusiones que entorpecían el avance de las diferentes disciplinas científicas. Esta situación ocurría porque no existía una formalización en el manejo de los datos, ni procedimientos de análisis que introdujeran un criterio de objetividad en las decisiones que se tomaban. Ante estas necesidades científicas y editoriales surgieron las pruebas de hipótesis estadísticas.

Las pruebas de hipótesis fueron creadas en el período entre 1915 y 1933 como resultado de la labor de dos grupos o tendencias: por un lado, Ronald Fisher (1890-1962) y, el por otro, Jerzy Neyman (1894-1981) en conjunto con Egon Pearson (1895-1980). Ambas tendencias tuvieron como antecedente la famosa prueba de ji al cuadrado de Karl Pearson (1857-1936). Los procedimientos de Fisher y Neyman-Pearson se desarrollaron a partir de posiciones filosóficas antagónicas, por lo que la historia de las pruebas de hipótesis no ha estado exenta de controversias, desacuerdos científicos y agrias disputas personales que lamentablemente se reflejan en la actualidad y han conducido a dificultades en su aplicación y aceptación[1-3]. Al aplicar las pruebas de hipótesis muchos investigadores combinan de manera ecléctica, elementos de los dos enfoques antagónicos: "…las pruebas de hipótesis estadísticas son presentadas en los artículos de las revistas siguiendo normalmente a Neyman-Pearson pero según la guía práctica de Fisher…"[3]. Este trata-

miento ecléctico es consecuencia de lo que aparece explicado en muchos libros de texto y que se refleja en la enseñanza de la estadística, tanto en el posgrado como en el pregrado.

Independientemente de la visión teórica y las opiniones o la posición científica de quien las analice, las pruebas de hipótesis estadísticas se han convertido, para muchos, en un instrumento fundamental del análisis de los datos y, para otros, en la única técnica para realizarlo, algo así como la piedra filosofal del conocimiento científico, cuyo uso invalida o glorifica un resultado. El papel preponderante de las pruebas de hipótesis, como criterio de validez de un análisis y de la calidad de un reporte científico, ha conducido al uso y abuso de la técnica. Su uso desmedido y en ocasiones indiscriminado ha producido errores, ha conducido a la obtención de conclusiones erróneas en algunas investigaciones, lo que ha determinado que para algunos esa técnica sea indeseable, superficial, prescindible e, incluso, peligrosa por la posibilidad de generar errores y conclusiones equivocadas, por no decir absurdas o tontas. A esta confusa situación han contribuido, infortunadamente, los sistemas computacionales para el análisis estadístico, los libros de texto y los propios criterios de enseñanza de la estadística[2-4].

Los sistemas de computación, en lugar de cumplir la loable función de poner la técnica al alcance de todos, han simplificado y vulgarizado su uso al permitir la generalización de la filosofía de "la caja negra" para el análisis, en la que el investigador maneja técnicas y conceptos que no comprende bien pero cuya ejecución deja al "sistema de computación" y después reduce la decisión a un simple, por no decir simplista o ingenuo, procedimiento dicotómico que reduce su problema de análisis y decisión a un simple sí/no, según un "místico" valor de P.

Esta situación ha sido favorecida, por no decir propiciada, por los programas docentes de estadística y los libros de texto, los que han ayudado a que esto ocurra al promover un mecanicismo ingenuo en los análisis que conducen a que el investigador, paradójicamente, realice el análisis de sus datos sin mirar los datos: "sólo decide sí/no, según el valor de P". Esta filosofía ha conducido a algunos a no considerar la plausibilidad biológica y la lógica de los resultados que van obteniendo, lo que se ha traducido en la publicación de conclusiones erróneas (4,5). Por otra parte, el valor P se convirtió en una especie de "dios" cuando muchos editores de revistas y pares favorecían la publicación de aquellos estudios con un P significativo. Esto llegó a

ser tan significativo que se introdujo el famoso sesgo de publicación que, hoy en día, tiende a ser de menor magnitud.

En los últimos tiempos, las revistas científicas dedicadas a la epidemiología y la bioestadística han reflejado un profundo conflicto entre defensores y partidarios del uso de las pruebas de hipótesis. Monterrey *et al*.[5] presentan un análisis de esa situación, sus causas y consecuencias. Para que se tenga una idea de la magnitud e importancia del asunto, basta citar el siguiente párrafo que aparece en las normas de Vancouver que, como se sabe, rigen la elaboración de publicaciones en el área de la salud y son seguidas por las revistas más prestigiosas del área:

> "… Describir los métodos estadísticos con suficiente detalle, para permitir a un lector conocedor, con acceso a los datos originales, la verificación de los resultados que se presentan. Cuando sea posible, se deben cuantificar los hallazgos y presentarlos con los correspondientes indicadores de los errores de medición e incertidumbre (tales como intervalos de confianza). Evite depender exclusivamente de las pruebas de hipótesis estadísticas, y del uso de valores de P, los que fallan cuando se pretende transmitir información importante acerca de la medida de un efecto"[6].

Como se puede apreciar, las normas de Vancouver no recomiendan el uso único de las pruebas de hipótesis; sin embargo, en la literatura biomédica se han presentado situaciones más extremas que han conducido a no considerar estos métodos, como puede verse en los trabajo de Walter[7] y de Gardner y Altman[8]; éste último, por ejemplo, se utiliza como paradigma para la presentación de datos en la revista *British Medical Journal*.

Una breve lectura a cualquier texto básico de estadística permite ver que en su mayoría está dedicado a las pruebas de hipótesis y, por tanto, los cursos de estadística mayoritariamente se dedican a enseñar pruebas de hipótesis, Además, es un hecho que la mayoría de los artículos científicos que se publican basan sus análisis en la aplicación de las pruebas de hipótesis; sin embargo, existen prohibiciones y barreras para su uso. Existen opiniones encontradas, pero muchos consideran aún válido el uso de las pruebas de hipótesis, claro que haciendo modificaciones en la forma de aplicarlas actualmente. La inmensa mayoría de los artículos sobre el tema son difíciles de leer para aquéllos que no conozcan en profundidad la técnica o no tengan una formación básica en la teoría de probabilidades. Dadas estas dificultades, el objetivo del pre-

sente artículo es mostrar cómo utilizar las pruebas de hipótesis en un contexto teórico coherente, analizando la plausibilidad biológica de los resultados, y presentar los alcances, posibilidades y limitaciones de esa técnica.

Fundamentos para la aplicación de las pruebas de hipótesis

Para comprender lo que son las pruebas de hipótesis estadísticas hay que partir del concepto de hipótesis de investigación y establecer sus diferencias con las hipótesis estadísticas.

La hipótesis de investigación es un enunciado que representa la posible respuesta a la pregunta de investigación[9]; como tal, tiene dos "valores de verdad": verdadero o falso, lo que determina las correspondientes respuestas a las preguntas de investigación. La investigación se realiza para determinar el valor de verdad que corresponde a la hipótesis de investigación, es decir, se realiza un estudio como resultado del cual se obtienen datos que contienen la información necesaria para dar respuesta a la pregunta de investigación al decidir si la hipótesis de investigación es verdadera o falsa. El reto de la estadística como disciplina es desarrollar métodos para obtener la información contenida en los datos y analizar esas evidencias para tomar tal decisión. Las pruebas de hipótesis son uno de tales instrumentos, pero no el único, aun cuando sea el más conocido y utilizado.

Las hipótesis estadísticas son enunciados que, en términos de conceptos estadísticos, representan o caracterizan la información contenida en los datos. Para decidir acerca de la validez de una hipótesis de investigación, pudiera ser necesario trabajar con varias hipótesis estadísticas; por ejemplo, en un estudio para determinar si la práctica de deportes tiene influencia sobre el perfil lipídico de los adultos varones, la hipótesis de investigación pudiera ser: la práctica diaria de ejercicios físicos modula los valores del perfil lipídico de los adultos varones. Para decidir si esta hipótesis es verdadera o falsa, se realiza un estudio en el que se toman dos grupos de individuos, uno que practica diariamente deportes (grupo A) y otro sedentario (grupo B), y se comparan las variables del perfil lipídico; en este caso, la decisión se tomaría considerando varias hipótesis estadísticas referidas a los valores medios (promedio) de cada una de las variables del perfil lipídico.

El enfoque teórico de Fisher para abordar un problema de prueba de hipótesis se fundamenta en la realización de una inferencia inductiva; consiste en plantear *una* hipótesis de interés, que en el ejemplo pudiera ser H_0: $\mu_A=\mu_B$, es decir, el valor medio del colesterol total es igual entre los gru-

pos A y B y, una vez obtenidos los datos, calcular un valor, que se identifica con la letra p, que es una medida de la evidencia que arrojan los datos contra la hipótesis cuya validez se desea comprobar. En este caso lo importante es saber si la hipótesis es cierta o se rechaza. Este es el valor de p que calculan todos los sistemas computacionales para el análisis estadístico; "…uno de los errores más comunes al interpretar el valor P es considerarlo como la probabilidad de que la hipótesis nula sea cierta…" (4).

Ante la ausencia de una hipótesis alternativa en el proceso de análisis de Fisher, Neyman y Pearson plantearon un proceso de decisión de tipo deductivo, diseñado *a priori* sobre la base de los datos, que consideraba una hipótesis o decisión alternativa a H_0 (hipótesis nula) y las dos tasas de error que se desea cometer en el proceso de decisión: los errores de tipo I (α) y II (β). En el ejemplo, las alternativas pudieran ser, por ejemplo, H_A: $\mu_A \neq \mu_B$ o H_A: $\mu_A < \mu_B$.

En el criterio de Fisher, el valor de P se establece *a posteriori*, es decir, sobre la base de los datos; en el de Neyman y Pearson, los datos se obtienen con una confiabilidad dada *a priori* por los errores (α y β).

Lamentablemente y de forma incierta, pero muy extendida, ambos enfoques se han fusionado en el procedimiento que aplican los investigadores en la actualidad y que se enseña en muchos cursos de estadística. Hubbard y Bayarri[2] señalan al respecto:

> "…Los libros de texto actuales sobre el análisis estadístico en economía, ciencias sociales o bioestadística, ya sea en el nivel posgraduado o de pregrado, presentan usualmente el tema de las pruebas de hipótesis como si fueran el evangelio: una teoría única, unificada y sin controversias. Es muy infrecuente que esos textos mencionen, menos aún que discutan teóricamente, que esa teoría que presentan es un híbrido anónimo entre las ideas de Fisher, por un lado, y de Neyman y Pearson, por el otro...".

El enfoque de Fisher parece ser más razonable para la investigación en salud que el de Neyman-Pearson. Parafraseando a Fisher: "…el procedimiento de Neyman-Pearson es más apropiado para problemas de control estadístico de la calidad…"[4] .

Siguiendo el enfoque de Fisher y sus consideraciones sobre el uso de los valores de P, es aconsejable proceder según los siguientes tres pasos al aplicar las pruebas de hipótesis:

Paso 1. Identificar el tipo de prueba que corresponde al problema y la hipótesis nula.

Paso 2. Obtención del valor de P.

Este valor de P es parte de la información que brindan los sistemas de cómputo. Todas las pruebas de hipótesis descansan en la identificación de un estadígrafo, que es conocido como el estadígrafo de la prueba. El valor de P refleja la posición del valor observado del estadígrafo en las colas de su distribución si H_0 fuera cierta y responde a la pregunta de cuán extremo es el valor observado en la distribución del estadígrafo empleado; por ello, el valor de P refleja o cuantifica la evidencia que contienen los datos contra la hipótesis nula.

Paso 3. Publicación explicita del valor de P junto con los valores descriptivos del proceso, es decir, junto con la media, la desviación estándar, la moda, etc.

La interpretación del valor de P descansa en el hecho de que los valores pequeños son evidencias contra la validez de la hipótesis nula[5]. Usualmente el investigador prefija el error de tipo I (α), casi siempre lo identifica como 0,05 y, cuando P<0,05, decide rechazar H_0. Esta forma de proceder es inadecuada pues mezcla las ideas de Fisher con las de Neyman y Pearson al comparar α, una tasa de error (*a priori*), con P, una medida de evidencia (*a posteriori*)[2-4]; de esta manera, se pierde el sentido del valor de P como evidencia *a posteriori* contra H_0 y predomina la lógica de Neyman-Pearson pero de forma incompleta pues, en general, no se controla el error de tipo II.

Divulgar explícitamente el valor de P permite a los diferentes lectores hacerse una idea clara y personalizada de la evidencia encontrada; para su interpretación es aconsejable utilizar los criterios de Sterne y Smith[4] que se resumen en el cuadro 1. Es por esto

Cuadro 1
Criterios para la interpretación de los valores de P

Valor P	Criterio de análisis del valor P
0,1<P<1,0	Débil evidencia contra la hipótesis nula
0,01<P<0,1	A medida que P disminuye, aumenta la evidencia contra la hipótesis nula
0,001<P<0,01	
0,0001<P<0,001	A medida que P disminuye, hay fuerte evidencia contra la hipótesis nula
P<0,0001	

que no es recomendable publicar el resultado de los análisis estadísticos, utilizando pruebas de hipótesis, con símbolos como **, NS, P<0,05, etc.; lo importante es escribir el valor de P obtenido para que el lector pueda enjuiciar la fuerza de la evidencia contra H_0 que se obtuvo en el estudio. La publicación explícita del valor de P es importante, fundamental para los ejercicios de integración de resultados conocidos como metaanálisis; de hecho, la evidencia a favor o en contra de H_0 surge de la combinación de diferentes estudios.

El valor de P cambia entre estudios y este cambio obedece a las leyes de la probabilidad; por ejemplo, si H_0 es cierta, el valor P queda determinado por realizaciones de una variable aleatoria con distribución uniforme[3], por lo que la probabilidad de encontrar estudios con p<0,05, cuando H_0 es cierta, es de 0,05; es decir, el 5% de las investigaciones deben rechazar H_0 siendo cierta, si se sigue la regla de rechazarla cuando p<0,05. Por eso, es importante publicar tanto los hallazgos positivos como los negativos; lamentablemente "…la literatura médica muestra una fuerte tendencia a acentuar lo positivo, los resultados positivos son reportados más frecuentemente que los nulos…"[4]. Por otra parte, el publicar el valor de P obtenido en el estudio permite al lector pensar en un sinnúmero de opciones y tomar su propia posición respeto al hallazgo que se reporta. No es lo mismo, en un experimento clínico que compara dos medicamentos, obtener un resultado con un valor P de 0,00001, que uno de 0,045, de 0,06, o una P de 0,10 o 0,5.

Como la evidencia contra H_0 queda determinada por valores pequeños de P, Fisher propuso utilizar 1:20 (0,05) como una buena medida, pero rápidamente alertó que esa no es una regla inflexible y que el valor debía adecuarse a las características del problema. El umbral de aceptación puede y debe cambiar; sería una buena ayuda tomar como punto de partida los valores que pudieran ser relevantes para el error de tipo I de Neyman-Pearson. En ningún caso es válido ni correcto considerar universalmente el umbral 0,05, como se hace usualmente; este valor fue popularizado por Fisher al ser empleado, junto a 0,01, en su famoso libro *Statistical methods for research workers*, pero estos valores fueron escogidos ante la imposibilidad de publicar tablas con valores más generales que se encontraban protegidas por los derechos de autor de la revista *Biometrika*. Tanto Fisher, como Neyman y Pearson, recomendaron flexibilidad en la determinación de los umbrales o las tasas de error, según el caso. En general, la determinación de un valor de umbral depende del problema y sus características; por ejemplo, no sería necesario tener un valor de P del 5% si estamos investi-

gando una droga curativa para la esquizofrenia que, como sabemos, no existe; en este caso, posiblemente con un 10% podríamos sentirnos satisfechos. Sin embargo, para un medicamento que combata los síntomas de la esquizofrenia, dado que existen otros que lo hacen, podríamos estar satisfechos con un 5% o menos.

Papel de los intervalos de confianza en los problemas de decisión: consideraciones para su uso

Existe una tendencia a desechar las pruebas de hipótesis y sustituirlas por intervalos de confianza[8,10]. La propuesta es bastante conocida pero, tal vez, no del todo comprendida por algunos en su real significado.

Un intervalo de confianza es un intervalo que se construye a partir de los datos, es decir, sus extremos son aleatorios y cambian de muestra en muestra. Al depender de la muestra, tienen un componente aleatorio determinado por el muestreo; por ello, se estudian en el marco de sus propiedades probabilísticas. Concretamente, los intervalos de confianza se construyen prefijando la probabilidad de que contengan el verdadero valor del parámetro que se va a estimar. Esta afirmación es difícil de entender y ha generado malas interpretaciones. Cuando se trata de estimar la estatura promedio de los adultos varones se dice, por ejemplo, que debe estar entre 170 y 175 cm, con un nivel de confianza de 95%. Es un error entender esta afirmación en el sentido de que, con probabilidad del 95%, la media está entre 170 y 175; significa en realidad que este intervalo fue construido a partir de un "procedimiento" que en el 95% de las veces arroja un intervalo que contiene al parámetro. Esa es la confianza que se tiene en la estimación que se presenta y, por ello, es una medida de su calidad. La mejor manera de leerlo o interpretarlo sería que si yo repitiera el estudio de la estatura en adultos varones en 100 ocasiones, 95 de los 100 intervalos de confianza obtenidos contendrían el valor real del parámetro (media).

Los intervalos de confianza son utilizados como alternativas en el proceso de decisión de Neyman y Pearson. Para ello se toma el valor del parámetro que establece H_0 y se decide acerca de la validez o no de esa hipótesis según si el valor contenido en ella se encuentra o no en el intervalo. De esa forma, si en la situación anterior H_0 establece que $\mu=177$, como 177 no pertenece al intervalo se concluiría que H_0 es falsa pues el valor de ì debe ser uno de los del intervalo, entre 170 y 175, lo que excluye la posibilidad de que pueda ser 177. El error de muchos está en pensar que en esto consiste la sustitución de las pruebas de hipótesis por el análisis con

intervalos de confianza. Un poco de historia basta para comprender la falacia. Los intervalos de confianza fueron introducidos por Neyman en 1937 como parte integrante de su teoría de las pruebas de hipótesis: "… los valores p y los intervalos de confianza son esencialmente recíprocos…"[11]. Y más allá, posiblemente provean información complementaría el uno al otro. Otro ejemplo, en un estudio sobre eficacia de un medicamento nuevo para la depresión, se puede decir que el medicamento obtuvo un riesgo relativo de 2 y un valor de P de 0,04; esto es significativo para lo que se proponía, pero se reafirma más la certeza en la evidencia y sus niveles de incertidumbre al decir que se obtuvo lo anterior descrito más un intervalo del 95% de (1,2 - 2,1). Este último dato aporta mayor información que presentar solamente el primero.

La mejor postura debe ser unir en los análisis el valor de P con los intervalos de confianza pues, analizados correctamente, ambos dan informaciones complementarias. El valor de P tiene la virtud de dar un valor objetivo, independiente del investigador, que permite caracterizar si la asociación es real siguiendo criterios establecidos *a priori* o si no lo es[12]. Por otra parte, la amplitud o diámetro del intervalo de confianza refleja la variabilidad del estudio, los niveles de incertidumbre que el diseño de muestreo o los propios errores no lograron corregir, "…el ancho de los intervalos de confianza da una clara indicación de cuán poco informativo es el estudio…"[13]. Es por ello que el intervalo de confianza da un ángulo del problema complementario a las pruebas de hipótesis, además de que, al brindar información sobre el posible valor del parámetro, permite tomar una posición respecto al significado biológico de los resultados obtenidos y proponer otro tipo de modelos causales.

Algunos ejemplos sobre cómo utilizar las pruebas de hipótesis

A continuación se identificará un problema de investigación y, tomándolo como referencia, se presentarán varias situaciones alternativas de cómo podría presentarse el análisis de los datos utilizando las pruebas de hipótesis.

La depresión mayor constituye uno de los problemas más importantes en salud pública en nuestro país y en el mundo. Se calcula que la depresión será la primera o la segunda causa en aportar a los años de vida ajustados por discapacidad (DALY o AVAD) en el mundo, para el año 2010. En este contexto los estudios clínicos para el análisis de la eficacia de diferentes tratamientos resultan especialmente importantes. Se toma como problema de referencia un experimento clínico aleatorio para analizar la efectividad de un nuevo tratamien-

to no farmacológico (tratamiento A), que se compara con un procedimiento tradicional de probada eficiencia (tratamiento B); como criterios de diagnóstico se consideran la entrevista estructurada de CIDI para depresión y la evaluación del puntaje y cambio de la depresión mediante el puntaje obtenido en la escala de Hamilton. El interés en el análisis de los datos del ensayo clínico sería determinar si en promedio los puntajes en la escala de Hamilton para depresión de los individuos sometidos al tratamiento A (μ_A) son diferentes, más precisamente menores, que los puntajes de los sometidos al tratamiento B (μ_B), en cuyo caso se podría afirmar que los datos del estudio arrojan evidencia a favor de la mayor eficacia del tratamiento A. En la terminología de las pruebas de hipótesis, en la formulación de Neyman y Pearson, el problema consistiría en decidir acerca de la validez de H_0:$\mu_A=\mu_B$ o de H_A:$\mu_A\neq\mu_B$, tomando como base los resultados experimentales.

Situación 1. En un estudio de integración de los resultados de la literatura se analizan tres publicaciones cuyos valores de P, resultantes de aplicar la prueba estadística para la comparación de medias de dos poblaciones independientes, se presentan en el cuadro 2.

Si se toma un umbral de 1:20 = 0,05 para identificar valores de P pequeños y mantenerse en un error de tipo 1 de esa cuantía y utilizando como elemento central para la interpretación de la evidencia obtenida en cada una de las publicaciones el criterio de Sterne y Smith[4], que aparece resumido en el cuadro 1, se tiene lo siguiente.

(1) El estudio A, con P=0,000000, presenta un valor P pequeño, lo que es una fuerte evidencia contra H_0.

(2) El estudio B, con P=0,04500 presenta un valor P relativamente pequeño, pero no muy pequeño, por lo que la evidencia que arroja contra H_0 no es muy fuerte.

(3) El estudio C, con P=0,882301, presenta un valor P elevado (re-

Cuadro 2
Resultados de tres publicaciones hipotéticas

Estudio	Valor de P
A	0,000000
B	0,045000
C	0,882301

cuerde que los valores P oscilan entre 0 y 1), por lo que la evidencia contra la hipótesis nula es muy débil.

Los datos para el análisis de las situaciones siguientes se presentan en el cuadro 3.

Situación 2. El valor P=0,0000 indica una fuerte evidencia contra H_0, lo que es confirmado por el intervalo de confianza que estima que la diferencia entre las medias de los dos tratamientos es un número entre -11,2 y -8,2. El radio del intervalo

$$\frac{11.2 - 8.2}{2} = \frac{3}{2} = 1.5$$

El valor obtenido, 1,5 unidades de la prueba, muestra que el nivel de error es bajo y, como una diferencia de puntajes de mínimo 8 unidades es relevante como índice de depresión, queda confirmada la relevancia de la disminución de la depresión que se logra al aplicar el tratamiento A.

Situación 3. El valor P=0,0149 es un valor P relativamente pequeño, pero no muy pequeño, por lo que la evidencia que arroja contra H_0 no es muy fuerte. El intervalo de confianza para estimar la diferencia de las medias entre los tratamientos (A y B) oscila entre -17,9 y -2,0, y tiene una amplitud muy grande, por lo que el nivel de imprecisión en la estimación de la diferencia de medias es elevado y, por consiguiente, el análisis que se haga acá es impreciso. Por ello, sus conclusiones deben tomarse cautelosamente y, como muestran los valores numéricos, no son evidencias fuertes a favor de que existan diferencias.

En una situación como está debe valorarse si el tamaño de muestra utilizado es adecuado y analizar todas las fases de diseño y conducción del estudio en busca de posibles sesgos.

Cuadro 3
Resultados hipotéticos de tres estudios (situaciones)

	Tratamiento A		Tratamiento B		Comparación (A – B)	
Situación	n	Media±DE	n	Media±DE	P	Intervalo de confianza (95%)
2	25	20 ± 2,1	25	30 ± 2,2	0,0000	-11,2 ~ -8,7
3	25	22 ± 14,0	25	30 ± 14,1	0,0149	-17,9 ~ -2,0
4	25	22 ± 0,5	25	23 ± 0,7	0,0000	-1,3 ~ -0,7

Situación 4. El valor P=0,0000 indica una fuerte evidencia contra H_0, es decir, se puede aceptar que hay diferencias en el puntaje de los individuos según el tratamiento; los del tratamiento A son los que obtienen una media[22] menor que la de los que reciben el tratamiento B[23]. Eso es un hecho, pero una observación del intervalo de confianza muestra una información adicional: que la diferencia entre las medias es un número que se estima en máximo 1,3, es decir, que la media del tratamiento A es a lo sumo 1,3 unidades mayor que la del tratamiento B. Diferencias de esta cuantía no son relevantes clínicamente desde el punto de vista de la presencia de depresión, así que, aunque existen diferencias entre los tratamientos que se comparan, estas diferencias no son clínicamente relevantes. Esta decisión es muy importante pues la sustitución del tratamiento usual (B) sólo se justifica si el nuevo tratamiento es más eficiente clínicamente o si tiene una mejor relación costo-beneficio.

Conclusiones

En este escrito se presentan datos sobre la importancia de conocer los orígenes de la prueba de hipótesis y de las posiciones antagónicas que tiene sus representantes máximos, Fisher, por un lado, y Newman con Pearson, por el otro. Se esboza cómo el planteamiento de Fisher podría ser más útil para las ciencias de la salud. Sin embargo, es prioritario poder pensar en la necesidad de tener un valor p especificado y divulgarlo explícitamente en las publicaciones; no presentarlo de forma aproximada en las publicaciones resaltando la decisión que toma con él el autor del estudio; esta forma de proceder aporta más información al lector. Es fundamental complementar las pruebas de hipótesis con los intervalos de confianza y valorar las conclusiones en el marco del significado de las mediciones y la plausibilidad biológica de las conclusiones a que se arriba.

Bibliografía

1. Lehman E. *The Fisher, Neyman-Pearson theories of testing hypothesis: one theory or two?* J Am Stat Assoc 1993; 88: 122-49.
2. Hubbard R, Bayarri M. *Confusion over measures of evidence (p´s) versus errors (*a´*s) in classical statistical testing.* Am Stat 2003; 57): 171-82.
3. Moran JL. *A farewell to P values?* Criti Care Resusc 2004; 6: 130-7.
4. Sterne JA, Smith GD. *Sifting the evidence-what´s wrong with significance tests?* Br Med J. 2001; 322: 226-31.
5. Monterrey PA, Cortés LY, Días ME. *Utilidad y limitaciones de las pruebas de hipótesis en epidemiología nutricional.* ¿Cómo proceder frente a un problema? Perspectivas en Nutrición Humana. 2003; 9: 72-87.

6. http: //www.rbccv.org.br/english/normsOfVancouver.asp
7. Walker AM. *Cómo presentar los resultados en los estudios epidemiológicos*. Bol Of Sanit Panam. 1993; 115): 148-54.
8. Gardner MJ, Altman DG. *Confidence intervals rather than P values: estimation rather than hypothesis testing*. Br Med J 1986; 292: 746-50.
9. Polit DF, Hungler BP. *Investigación científica en ciencias de la salud*. Principios y métodos. McGraw-Hill Interamericana, HealthCare Group; 2000.
10. Goodman SN. Toward evidence-based medical statistics. 1. The P value fallacy. Ann Intern Med 1999; 130: 995-1004.
11. Feinstein AR. *P-values and confidence intervals: two sides of the same unsatisfactory coin.* J Clin Epidemiol 1998; 51: 355-60.
12. Fleiss JL. *Confidence intervals vs. significance tests: quantitative interpretation.* Am J Public Health 1986; 76: 587-8.
13. Thompson WD. *Statistical criteria in the interpretation of epidemiologic data*. Am J Public Health. 1987; 77: 191-4.