REVISTA DE LA FACULTAD DE CIENCIAS MÉDICAS
UNIVERSIDAD DE CUENCA

Uso inercial de técnicas estadístico-inferenciales: su posible impacto en el hallazgo de resultados falsos en salud.

Luis Carlos Silva Ayçaguer[1], Gino Montenegro Martínez[2], Elizabeth García Restrepo[2], Natalia Eugenia Gómez Rúa[2], Víctor Hernán Arcila Quiceno[3].

VOLUMEN 35 / Nº2 / DICIEMBRE 2017

FECHA DE RECEPCIÓN: 01/11/ 2017
FECHA DE APROBACIÓN: 11/12/2017
FECHA DE PUBLICACIÓN:14/12/2017

ENSAYO/ESSAY

CONFLICTO DE INTERESES:
LOS AUTORES DECLARAN QUE NO EXISTE CONFLICTO DE INTERESES.

1. Docente doctorado Salud Pública Universidad CES.
2. Estudiante doctorado Salud Pública Universidad CES.
3. Estudiante doctorado Salud Pública Universidad CES. Docente Universidad Cooperativa de Colombia. Grupo de Investigación de Ciencias Animales GRICA.

Correspondencia:
lcsilvaa@yahoo.com.

RESUMEN

Objetivo: Sondear el grado en que importantes recursos estadísticos, en particular los valores p, los intervalos de confianza y los procedimientos para determinar tamaños muestrales, se emplean en la literatura biomédica de manera ritual. **Metodología:** se seleccionaron 25 artículos originales publicados en cada una de 4 revistas indexadas del campo biomédico. Para cada uno de ellos se evalúo si cumplían con las indicaciones de las guías STROBE y CONSORT en lo concerniente al tamaño de la muestra, así como el uso de los valores p, de los intervalos de confianza y la utilización de estos en la discusión del artículo. **Resultados:** el 97.0% de los artículos reporta el tamaño de la muestra, pero sólo el 62.9% explica cómo fue determinado. El valor p se usa con mayor frecuencia (68.0%) que los intervalos de confianza (63.9%). Solo el 15.5% usa los intervalos de confianza en la discusión. **Conclusión:** las herramientas estadísticas más convencionales se emplean en buena medida de manera más ceremonial que funcional.

Palabras clave: Estadística, Tamaño de la Muestra, Intervalos de Confianza.

ABSTRACT

Objective: To evaluate the degree in which important statistical resources, particularly p-values, confidence intervals and procedures for determining sample sizes, are used in the biomedical literature through a ritual way.

Methodology: A total of 25 original articles published in each of 4 journals indexed in the biomedical field were selected. For each of them, it was assessed whether they follow the instructions of the STROBE and CONSORT guidelines regarding the sample size, as well as the use of p-values, confidence intervals, and the use of these in the discussion of the article.

Results: The 97.0% of the articles reported the sample size, but only 62.9% explained how it was determined. The p-value is used more frequently (68.0%) than the confidence intervals (63.9%). Only 15.5% uses confidence intervals in the discussion section.

Conclusion: the most conventional statistical tools are used more in ceremonial way rather than in a functional one.

Key words: statistics, Predictive Value of Tests sample size, Confidence Intervals.

INTRODUCCIÓN

A comienzos del siglo XX, el epidemiólogo y estadístico de la Universidad de Stanford, John Ioannidis, publicó un artículo que daba por sentado que los hallazgos de la investigación contemporánea en salud arriban en su mayoría a conclusiones falsas[1]. ¿Cómo puede explicarse una afirmación de ese calibre? Varias son las causas, pero ahora nos concentraremos en una de ellas: los resultados de muchos estudios se examinan usando métodos estadísticos inadecuados u obsoletos.

Parte de dichos métodos se ocupan de describir o caracterizar poblaciones de interés, en tanto otros fueron concebidos para la realización de inferencias. Algunos de estos últimos son los que más se cuestionan, como ocurre con las pruebas de significación estadística (PSE). Otros, como los intervalos de confianza (IC), han sido y son actualmente mucho más aceptados. No obstante, raramente se enfatiza la necesidad de examinar sus extremos para elaborar conclusiones: si bien se comunican tales límites, luego se actúa como si no se hubieran calculado.

Los valores p y las PSE vieron la luz entre 1925 y 1929 gracias a Ronald Fisher. Por su conducto se procura medir el grado de discrepancia entre los datos y la llamada "hipótesis nula", que afirma que determinada condición o intervención no tiene efecto alguno o, para decirlo en términos más estadísticos, que los parámetros que se comparan son idénticos. Si la p resultante fuera pequeña, según la propuesta fisheriana, tal hipótesis sería puesta en cuestión y demandaría un análisis más profundo[2, 5]. En consecuencia, la utilización de los valores p se concebía como una herramienta que, combinada con otra información disponible, sería útil para construir conclusiones a partir de lo observado[4].

Años más tarde, Jerzy Neyman y Egon Pearson propusieron un método alternativo al que denominaron "prueba de hipótesis". Esta técnica estaba orientada a la toma de decisiones: rechazar una hipótesis y aceptar la opuesta únicamente a partir de unos datos[3]. Según este planteamiento, el problema consistía en decantarse por una de dos hipótesis complementarias, y el procedimiento proporcionaba una regla formal para adoptar tal decisión.

El método que actualmente se usa en las PSE es una combinación de las ideas propuestas por Fisher y por el binomio Pearson-Neyman[6]. En efecto, a comienzos de la década de 1950 ambos métodos se unen de manera anónima, como un intento de conciliar éstas dos perspectivas originalmente contrapuestas. El método combinado toma de Fisher el valor p para ser utilizado como un índice que mide la fuerza de la evidencia. De Neyman y Pearson, se toma la idea de que la finalidad del uso de estas técnicas es tomar una "decisión" sobre la hipótesis nula (rechazarla o no), aunque su aplicación actual no establece una hipótesis alternativa concreta que habría de aceptarse en caso de rechazo[2, 3]. El modus operandi contemporáneo produce un veredicto dicotómico: se ha encontrado significación estadística si p es menor que 0,05 y no se encontrado en caso contrario.

En lo que va de siglo, se han venido multiplicando los cuestionamientos sobre la utilidad de las PSE[7, 10], pero desde muchos años antes tales críticas venían poniéndose de manifiesto en documentos relevantes, tales como las recomendaciones del Comité Internacional de Editores de Revistas Médicas (ICMJE, por sus siglas en inglés). Dicho Comité ha venido recomendando desde hace muchos años y aún recomienda que, al momento de presentar los resultados de investigación, éstos no dependan exclusivamente de los valores p e insta a la utilización de una medida que presente la magnitud del efecto y el grado de incertidumbre asociado. Es así como los intervalos de confianza (IC), hoy en día, representan una alternativa aceptable para suplir, o por lo menos complementar, el uso de los valores p[11].

El IC se define como un recorrido de valores entre los que se "espera" que se encuentre el "verdadero valor" de un parámetro poblacional con cierto grado de confianza[12].

La teoría de los IC fue propuesta por Jerzy Neyman en la década de 1930. El término es la traducción de "przedzial ufnósci", de origen polaco. Anterior a una publicación que explicara la teoría que entraña la construcción de los IC, una guía que contenía los IC para las medias y los coeficientes de regresión fue publicada por Pytkowski en 1932[13]. A finales de la década de 1980 diversos autores, entre los que se destacan Martin Gardner y Douglas Altman, se pronuncian subrayando la inequívoca preeminencia que deberían de tener los IC por encima de los valores p[2, 14].