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].

Otro problema que reclama atención es el hecho de que, por una vía u otra, virtualmente toda investigación que maneje datos empíricos requiere de la determinación previa de un tamaño de muestra para llevar a cabo el estudio. Es un paso que se debe considerar en una fase temprana de cualquier proyecto de investigación[15]. Para guiar al investigador en esta tarea, se cuenta con innumerables textos y programas informáticos que se dedican a la descripción de procedimientos que se pudieran seguir para realizar el cálculo[16-18].

Recientemente, en el ámbito de las publicaciones científicas en salud, se han venido estableciendo guías que brindan a los autores indicaciones orientadas a acrisolar la elaboración de manuscritos. Las dos más conocidas son las denominadas CONSORT (Consolidated Standards of Reporting Trial) para ensayos clínicos y STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) para los estudios observacionales.

Dichas guías exigen que los manuscritos describan de manera detallada cómo fue determinado el tamaño de muestra empleado. Tal recomendación, sin embargo, no viene avalada con argumentos que la soporten.

Para Bacchetti[19] el método que regularmente se ha empleado para el cálculo del tamaño de muestra no solo es mecánico y ritual, sino que exige un pensamiento especulativo y hondamente subjetivo. El autor nos recuerda que el investigador, al momento de aplicar el método, debe aportar datos que se conocen de manera sumamente vago o que, incluso, se desconocen por ser, justamente los datos que se quieren conocer (por ejemplo, ha de comunicarse de antemano la prevalencia de una determinada enfermedad en una población dada cuando el estudio se emprende precisamente para conocer dicha prevalencia)[20]. En consecuencia, el problema reside en que estos procedimientos se usan con la ingenua convicción de que proveerán, "el" número adecuado de individuos que se deberán interrogar o medir para desarrollar la investigación cuando en realidad es un acto mecánico y casi místico[5, 19, 21, 22]. Por otra parte, es conocida la frecuente manipulación de los datos iniciales para que el cálculo del tamaño de la muestra coincida con lo que se desea, práctica conocida como "retrofitting", "simple size game" o "simple size samba"[5, 21].

El presente artículo se propone sondear el grado en que tres importantes recursos estadísticos (los valores p, los intervalos de confianza y los procedimientos para determinar tamaños muestrales), se emplean de manera ritual en la literatura biomédica actual.

MATERIALES Y MÉTODOS

Este estudio observacional y transversal valora una muestra de artículos originales publicados entre 2014 y 2016 en cuatro revistas de reconocido prestigio (con un índice h entre moderado y elevado) tanto del ámbito hispanohablante como del mundo sajón (Tabla 1).

Por *artículos originales* se entienden aquellos donde se presentan por primera vez resultados de un proceso de experimentación u observación. Se excluyeron los artículos de revisión (meta análisis y revisiones sistemáticas de literatura), teóricos o metodológicos. Para cada uno de los artículos se estableció si se consigna el tamaño de muestra empleado, si dicho tamaño es explicado, si se contemplan valores p e intervalos de confianza y si se usan los extremos de los intervalos de confianza en el contexto de la discusión. Cabe subrayar que, por su naturaleza, para todos los artículos considerados tiene sentido aquilatar el empleo de las técnicas mencionadas.

Inicialmente, se establecieron reglas para el manejo por parte de los observadores a los ítems considerados.

TABLA Nº 1

Revistas seleccionadas según país de origen e índice h según google scholar

REVISTA SELECCIONADA	PAÍS DE ORIGEN	ÍNDICE H SEGÚN GOOGLE SCHOLAR (2017)
British Medical Journal	Reino Unido	145
American Journal of Public Health	Estados Unidos	74
Medicina Clínica	España	21
Revista Cubana de Salud Pública	Cuba	16

Elaborado por: los autores. Fuente: Base de datos.

TABLA Nº 2

. Valores de Kappa correspondiente a la clasificación de varios revisores para 5 variables dicotómicas.

VARIABLE	VALOR KAPPA
a. Se indica tamaño de muestra	1.00
b. Se explica el tamaño de muestra	0.34
c. Se usan "valores p"	0.91
d. Se emplean Intervalos de confianza	0.59
e. Se usan los extremos de los intervalos de confianza para desarrollar la discusión	0.34

Elaborado por: los autores. Fuente: Base de datos.

Posteriormente se realizó una prueba piloto con 16 artículos, cada uno de los cuales fue valorado por los últimos cuatro firmantes del trabajo. Para cada una de las 5 variables, se calculó el coeficiente Kappa de concordancia en su variante para varios clasificadores de variables dicotómicas empleando el programa EPIDAT 4.2. Los resultados se presentan en la Tabla 2.

El pilotaje permitió identificar las discrepancias para la evaluación de cada uno de los ítems. Como pone de manifiesto la Tabla 2, la concordancia entre los examinadores en los ítems b, d y e fue inaceptablemente baja. Para corregir este problema, el grupo investigador analizó pormenorizadamente las causas de dichas discrepancias, lo que condujo a un reajuste de las definiciones operativas. Con las definiciones definitivamente establecidas, se consiguió una adecuada consistencia en la evaluación de todos los ítems. A continuación, se presentan definiciones operativas para cada uno de los ítems considerados:

a. **Indica o define tamaño de muestra:** se comunica explícitamente el tamaño de muestra utilizado.

b. **Se explica tamaño de muestra:** se describen los métodos que se siguieron para determinar el tamaño muestral (sea mediante el uso de fórmulas o no). En los casos en que se utilizan fuentes secundarias de información (encuestas nacionales, reportes de entidades oficiales, etc.) se consideró que se da la explicación si los autores informaron al lector el mecanismo que utilizó la fuente primaria para definir el tamaño de muestra.

c. **Se usan valores p:** se emplean "valores p" con el fin de valorar la existencia o no de diferencias estadísticamente significativas entre los grupos que se investigan.

d. **Se usan intervalos de confianza (IC):** se reportan intervalos de confianza, para la estimación de al menos un parámetro o diferencia de parámetros.

e. **Se usan los extremos de los intervalos de confianza para desarrollar la discusión:** dentro de la discusión del artículo, los autores utilizan los extremos de los IC para enjuiciar sus resultados, o para compararlos con los obtenidos en estudios similares.

Para completar el estudio se decidió tomar números completos de manera que, para cada revista, se acumularan aproximadamente los últimos 25 trabajos publicados en cada cual.

RESULTADOS

Se evaluaron en total 97 artículos de los cuales el 1.0% fue publicado en 2014. El grueso de la muestra estuvo compuesto por artículos publicados en 2016 (56.7%), seguido de aquellos publicados en el año 2015 (42.3%) (Tabla 3).

La declaración del tamaño de la muestra es una práctica virtualmente unánime: se constató en el 97.0% (IC 95%: 95.1- 100) de los artículos evaluados. Sin embargo, sólo en el 62.9% (IC 95%: 53.3 – 72.5) de los trabajos se encuentra una explicación del método que se empleó para tal fin.

La frecuencia con que los autores utilizan los valores p (68.0%; IC 95%: 58.8 – 75.5) es mayor que aquella en que apelan a los intervalos de confianza (63.9%; IC 95%: 54.4-73.5). De otro lado, a pesar de que en más de la mitad de las publicaciones se recurre a los intervalos de confianza para el análisis y presentación de la información, sólo en el 15.5% (IC 95%: 8,3-22.7) dichos intervalos se utilizan en la discusión de los resultados (Tabla 4).

La Tabla 5 presenta los resultados que ofrecen mayor interés: aquellos que conciernen al empleo de valores e intervalos de confianza debido a que existen pautas

TABLA Nº 3

Distribución de artículos originales evaluados según revista y año de publicación.

AÑO DE PUBLICACIÓN	REVISTA CUBANA DE SP n (%)	MEDICINA CLÍNICA n (%)	JOURNAL OF PUBLIC HEALTH n (%)	BRITISH MEDICAL JOURNAL n (%)	TOTAL n (%)
2016	13 (52.0)	5 (20.0)	26 (100)	11 (52.4)	55 (56.7)
2015	11 (44.0)	20 (80.0)	0 (0)	10 (47.6)	41 (42.3)
2014	1 (4.0)	0 (0)	0 (0)	0 (0)	1 (1.0)
TOTAL	25 (100)	25 (100)	26 (100)	21 (100)	97 (100)

Elaborado por: los autores. Fuente: Base de datos.

TABLA Nº 4

Frecuencia de aparición de cada uno de los ítems evaluados según revista

ÍTEM EVALUADO	RCSP % (IC 95%)	MC % (IC 95%)	JPH % (IC 95%)	BMJ % (IC 95%)	TOTAL % (IC 95%)
Indica el tamaño muestra	100 (-)	100 (-)	96.0 (88.3 - 100)	95.2 (87.2 - 100)	97.9 (95.1 - 100)
Explica el tamaño de muestra	64.0 (45.2 – 82.8)	84.0 (69.6 – 98.4)	80.0 (64.3 – 96.7)	38.1 (19.0 – 57.1)	62,9 (53.3 – 72.5)
Se usan valores p	44,0 (24.5 – 63.5)	72.0 (54.4 – 89.6)	64.0 (45.2 – 82.8)	79.2 (59.49 – 92.9)	68.0 (58.8 – 77.3)
Utiliza Intervalos de confianza	24,0 (7.26 – 40.7)	48.0 (28.4 – 67.9)	56.0 (36.5 – 75.5)	90.5 (78.9 - 100)	63.9 (54.4 – 73.5)
Usa los intervalos de confianza en la discusión	0 (-)	28.0 (10.4 – 46.6)	20.0 (4.3 – 35.7)	23.8 (6.7 – 39.8)	15.5 (8.3 – 22.6)

Elaborado por: los autores. Fuente: Base de datos.

TABLA Nº 5

Frecuencia de uso de valores p e intervalos de confianza en los artículos evaluados

ÍTEM EVALUADO	FRECUENCIA	PORCENTAJE	(IC 95%)
Sólo valores p	13	13.4	6.6 – 20.2
Sólo Intervalos de Confianza	9	9.3	3.4 – 14.9
Intervalos de confianza y valores p	53	54.6	44.7 – 64.5
Ninguno de los dos (ni intervalos de confianza ni valores p)	22	22.7	14.3 – 31.0

Elaborado por: los autores. Fuente: Base de datos.

explícitas y bien consensuadas acerca de cómo manejarlos.

Lo más llamativo es que el uso aislado de los valores p se presenta en uno de cada ocho publicaciones (13.4%; IC95%: 6.6-20.2), en tanto que el uso de los intervalos de confianza sin que figuren los valores p es menos frecuente (9.3%; IC95%: 3.4-14.9).

DISCUSIÓN

Casi dos de cada tres artículos publicados siguen la recomendación de las guías CONSORT y STROBE en lo relacionado con explicar el tamaño de muestra. El intervalo de confianza permite apreciar que, muy verosímilmente, lo hacen más de la mitad de los trabajos, una cifra muy alta, si se tiene en cuenta la inutilidad de esta información para un lector[22]. Debe recordarse que hasta el momento no existe un documento que, partiendo de una juiciosa reflexión desde lo epistemológico, ético y práctico, despliegue un argumento sólido que sostenga la necesidad de cumplir con tal demanda[21]. Pero a lo sumo dicha explicación la daría el 75% de los trabajos, lo cual es igualmente expresivo, ya que indica que al menos en uno de cada 4 artículos, tanto autores como editores, parecen comprender tal inutilidad y no exigen que se haga. La situación es entonces ciertamente confusa y, en esa medida, exhorta de más reflexión crítica que la que ha merecido hasta ahora.

Por otra parte, se encontró que uno de cada seis artículos recurren únicamente a los valores p para presentar los resultados. Este dato es alarmante, ya que diversos autores, y desde hace ya bastante tiempo, han comunicado los problemas que entraña la técnica, los inconvenientes que tiene su uso indiscriminado y los errores relacionados con la mala interpretación de sus resultados[4, 8, 10, 23, 24]. De acuerdo con Sarría y Silva[25], las razones por las que aún se siguen utilizando, a pesar de sus limitaciones, pueden ser: a) gran cantidad de investigadores ignoran las objeciones que sobre ellas se han planteado; b) son fáciles de aplicar usando paquetes informáticos; y c) tradicionalmente todo el mundo lo usa; tanto a estudiantes como a investigadores se les instruye para que sigan empleándolas bajo la premisa de que aportan cientificismo y rigor.

La discusión es la sección de un artículo en donde los autores, por una parte, identifican si las hipótesis que sometieron a prueba son ciertas o no, y por otro, el significado de los hallazgos en términos de los cambios que tendrían que hacerse o no, en la manera en que se realizan determinadas prácticas[26]. En este sentido, lo esperado es que los autores utilicen este segmento para construir un argumento que, con base a los resultados encontrados, brindará al lector información acerca de la utilidad de lo encontrado en la investigación.

Otro de los hallazgos de esta investigación es la baja frecuencia con que se utilizan los intervalos de confianza en la discusión de las publicaciones evaluadas. Menos de un tercio de los artículos los contempla en esta sección del artículo, aunque ellos figuren entre los resultados que se comunican. Este, junto con los restantes resultados de este estudio ponen de manifiesto que la utilización de determinadas herramientas estadísticas, en publicaciones del ámbito sanitario a menudo responden más a un ejercicio mecánico determinado por un algoritmo metodológico que a un ejercicio reflexivo sobre la realidad que se quiere conocer y aún más, sobre la verdadera utilidad de los resultados de investigación[27].

Adicionalmente, parece que el interés suele estar más centrado en la utilización de técnicas estadísticas populares como parte de una "moda en el tratamiento de los datos" o con el fin de dar respuesta a lo sugerido por guías de renombre en el ámbito de la ciencia como la CONSORT y la STROBE[27, 28] que en emplearlas como ayuda para discutir los problemas subyacentes.

La selección de las herramientas estadísticas a las que se recurre para dar respuesta a una pregunta de investigación dada es en buena medida ceremonial y en ocasiones simplemente ornamental.

La estadística, como piedra angular en la construcción del conocimiento debe ser usada de manera racional evitando modas o rutinas, que son incongruentes con una reflexión crítica sobre su utilidad y, por tanto, ha de estar lejos de constituirse en un recurso al que se apele mecánicamente dentro del proceso de investigación. Ese es precisamente uno de los reclamos más importantes que hacía Ioannidis y que ha motivado el presente artículo.

REFERENCIAS

1. Ioannidis JP. Why most published research findings are false. PLos Med. 2005; 2(8):e124.
2. Silva LC. Valores p y pruebas de significación estadística: fin de una era. En: La investigación biomédica y sus laberintos: en defensa de la racionalidad para la ciencia del siglo XXI. 1ª ed. España: Ediciones Díaz de Santos; 2008. p. 347-479.
3. Rodríguez AB, Silva LC. Contra la sumisión estadística: un apunte sobre las pruebas de significación. Metas de enferm. 2000; 3(27):35-40.
4. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999; 130(12):995-1004.
5. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. The Lancet. 2005; 365(9467):1348-53.
6. Gerrodette T. Inference without significance: measuring support for hypotheses rather than rejecting them. Marine Ecology. 2011; 32(3):404-18.
7. Läärä E. Statistics: reasoning on uncertainty, and the insignificance of testing null. Ann. Zool. Fennici. 2009; 46(2):138-57.
8. Nicholls N. The insignificance of significance testing. Bulletin of the American Meteorological Society. 2001; 82(5):981.
9. Armstrong JS. Statistical significance tests are unnecessary even when properly done and properly interpreted: Reply to commentaries. International Journal of Forecasting. 2008; 23:335-6.
10. Hubbard R, Lindsay RM. Why P values are not a useful measure of evidence in statistical significance testing. Theory & Psychology. 2008;18(1):69-88.
11. ICMJE. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals: roles and responsibilities of authors, contributors, reviewers, editors, publishers, and owners: defining the role of authors and contributors [Internet]; 2016 Dic [citado 2017 marzo 12]. Disponible en: http://www.icmje.org/icmje-recommendations.pdf
12. Clark ML. Los valores P y los intervalos de confianza: ¿en qué confiar? Rev Panam Salud Pública. 2004; 15(5):293-6.
13. Neyman J. Fiducial argument and the theory of confidence intervals. Biometrika. 1941; 32(2):128-50.
14. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed). 1986;292(6522):746-50.
15. Rondón MA, Rodríguez VA. Algunos conceptos básicos para el cálculo del tamaño de la muestra. Univ Med. 2007; 48(3):334-9.
16. McCrum-Gardner E. Sample size and power calculations made simple. International Journal of Therapy and Rehabilitation. 2010; 17(1):10.
17. Charan J, Biswas T. How to calculate sample size for different study designs in medical research? Indian J Psychol Med. 2013; 35(2):121.
18. Noordzij M, Tripepi G, Dekker FW, Zoccali C, Tanck MW, Jager KJ. Sample size calculations: basic principles and common pitfalls. Nephrol Dial Transplan. 2010; 25(5):1388-93
19. Bacchetti P. Current sample size conventions: flaws, harms, and alternatives. BMC medicine. 2010;8(1):1.
20. Bacchetti P, Deeks SG, McCune JM. Breaking free of sample size dogma to perform innovative translational research. Sci Transl Med. 2011;3(87):87ps24-87ps24.
21. Silva LC, Alonso P. Explicación del tamaño muestral empleado: una exigencia irracional de las revistas biomédicas. Gac Sanit. 2013; 27(1):53-7.
22. Bacchetti P, McCulloch CE, Segal MR. Simple, defensible sample sizes based on cost efficiency. Biometrics. 2008; 64(2):577-85.
23. Anderson DR, Burnham KP, Thompson WL. Null hypothesis testing: problems, prevalence, and an alternative. J. Wild. Manage. 2000:912-23.
24. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods. 2000; 64(4): 912-923.
25. Sarria M, Silva LC. Las pruebas de significación estadística en tres revistas biomédicas: una revisión crítica. Rev Panam Salud Pública. 2004; 15(5): 300-06.
26. Provenzale JM, Stanley RJ. A systematic guide to reviewing a manuscript. J Nucl Med Technol. 2006; 34(2):92-9.
27. Silva LC. Una ceremonia estadística para identificar factores de riesgo. Salud colectiva. 2005; 1(3):309-22.
28. Silva L. Cultura estadística e investigación científica en el campo de la salud. 1ª ed. Madrid: Ediciones Díaz de Santos; 1997