Estudio primario
Medwave 2015 Ago;15(7):e6238 doi: 10.5867/medwave.2015.07.6238

Herramientas estadísticas en los artículos publicados en una revista de salud pública durante el periodo 2013-2014: estudio bibliométrico transversal

Statistical tools in published articles of a public health journal in 2013 and 2014: bibliometric cross-sectional study

Autores: Víctor Arcila Quiceno**[1]**, Elizabeth García Restrepo**[1]**, Natalia Gómez Rúa**[1]**, Gino Montenegro Martínez**[1]**, Luis Carlos Silva Ayçaguer**[1,2]**

Filiación:
[1] Facultad de Medicina, Universidad CES, Medellin, Colombia
[2] Escuela Nacional de Salud Pública, La Habana, Cuba

E-mail: gnmontenegro@gmail.com

Citación: Arcila Quiceno A, García Restrepo E, Gómez Rúa N, Montenegro Martínez G, Silva Ayçaguer LC. Statistical tools in published articles of a public health journal in 2013 and 2014: bibliometric cross-sectional study. *Medwave* 2015 Ago;15(7):e6238 doi: 10.5867/medwave.2015.07.6238
Fecha de envío: 2/7/2015
Fecha de aceptación: 24/8/2015
Fecha de publicación: 31/8/2015
Origen: no solicitado
Tipo de revisión: con revisión por dos pares revisores externos, a doble ciego

Resumen

INTRODUCCIÓN
El desarrollo de proyectos de investigación suele demandar del uso de herramientas matemáticas para expresar en términos numéricos o gráficos diversas magnitudes, tales como frecuencias, diferencias, o asociaciones.

OBJETIVOS
El presente estudio se propone describir la utilización que se hace de las herramientas estadísticas básicas, haciendo especial énfasis en el uso de las pruebas de significación estadística y los intervalos de confianza, para la presentación de resultados de investigación en los artículos publicados en una reconocida revista de la salud pública en Colombia.

MÉTODOS
Se realizó una revisión de los 84 artículos originales publicados en dicha revista entre 2013 y 2014.

RESULTADOS
El recurso más empleado es la utilización de análisis de frecuencias (89,3%), seguido de los valores p (65,5%) y los intervalos de confianza (53,6%); el 48,9% de los artículos utiliza a la vez intervalos de confianza y valores p para la presentación de resultados y el 29,8% ninguno de los dos. El 16,7% de los artículos sólo utiliza valores p y el 4,8% sólo intervalos de confianza.

CONCLUSIONES
La estadística descriptiva es una herramienta que se utiliza con asiduidad en la presentación de resultados y que las críticas y advertencias sugiriendo que se evite el uso exclusivo de las pruebas de significación estadística en la presentación de resultados no son cabalmente tenidas en cuenta en el análisis y presentación de los resultados de investigación.

Abstract

INTRODUCTION
Research projects use statistical resources to express in numerical or graphic terms different magnitudes like frequencies, differences or associations.

OBJECTIVES
The purpose of this paper is to describe the statistics tools utilization, with special emphasis in the use of conventional statistical tests and confidence intervals, to communicate results in a renowned public health peer reviewed journal in Colombia.

METHODS
We included the 84 articles published in the journal between 2013 and 2014.

RESULTS
The most used resource is frequency analysis (89,3%), followed by p values (65,5%) and confidence intervals (53,6%); 48,9% of the papers used confidence intervals together with p values; 29.8% use neither of them; 16.7% of the articles only used p values and 4.8% only confidence intervals.

CONCLUSIONS
Descriptive statistics is a tool widely used in research results presentation; the critics and caveats suggesting to avoid the exclusive use of the statistical signification test in the results presentation are not followed in the analysis and presentation of the research results.

Introducción

El desarrollo de proyectos de investigación suele requerir del uso de herramientas matemáticas para expresar en términos numéricos o gráficos diversas magnitudes, tales como frecuencias, diferencias, o asociaciones. En diferentes campos de la ciencia han emergido como recurso que ayuda a visualizar y argumentar las respuestas de que son objeto las preguntas de investigación. Tales herramientas se podrían clasificar en las descriptivas o medidas de resumen (como la media, la desviación estándar o las frecuencias), y las de naturaleza inferencial, que incluyen las llamadas pruebas de significación estadística y la construcción de intervalos de confianza, por solo mencionar las más elementales y utilizadas [1].

La utilidad de las medidas descriptivas está fuera de toda duda y, en principio, muchos resultados interesantes e incisivos pueden conseguirse gracias a ellas. Sin embargo, el empleo de los recursos inferenciales, es más polémico.

Las pruebas de significación estadística se introdujeron en la década de 1920 como un recurso que revolucionó las técnicas inferenciales. En su inicio, se trató de un procedimiento propuesto por Ronald Fisher quien introdujo el concepto de "hipótesis nula" que, por lo general, establece que determinados rasgos no se distinguen entre un grupo y otro. El aporte de Fisher giraba en torno a su más famosa invención: los famosos valores p [2].

En 1928 Jerzy Neyman y Egon Pearson publican un artículo [3] donde proponen un proceder íntimamente enlazado con la propuesta de Fisher pero operativa y epistemológicamente distinto, concebido para elegir una de dos posibilidades: la llamada hipótesis nula y una hipótesis alterna. Desde mediados de ese siglo y hasta actualidad se aplica un híbrido de estos dos métodos [1],[4]: dados unos datos obtenidos a partir de una observación o un experimento, se calcula el valor de "p" utilizando recursos asociados a la pruebas de significación estadística de que se trate. Cuando ese valor de p es menor que determinado umbral prefijado (casi siempre igual a 0,05) se rechaza implícitamente la hipótesis nula afirmando que "se ha hallado significación" y consignando el valor de p obtenido.

A pesar de que su uso se ha popularizado en las más diversas disciplinas, las pruebas de significación estadística han sido objeto de numerosas críticas a cargo de renombrados especialistas del área de la estadística, centradas tanto en sus debilidades lógicas, como en sus limitaciones epistemológicas y prácticas[5],[6],[7],[8],[9]. Silva et al., [4] hacen una compilación de algunas de ellas, entre las que se encuentran las siguientes:

- Se puede tener un valor p tan pequeño como se quiera con solo aumentar el tamaño de la muestra. Siendo así, las conclusiones dependen de los recursos disponibles (aquellos que permiten tener un gran tamaño de muestra) más que de la realidad misma que se estudia.
- En la práctica, las pruebas de significación estadística se limitan a dar una respuesta binaria (aceptación o rechazo; significación o no), sin exigir una interpretación razonada de los resultados. Esta ortodoxia circunscrita a lo dicotómico ofrece una visión muy simplista que pasa por alto la valoración sobre los efectos que demanda la ciencia.

Se debe tener en cuenta que estamos hablando de un concepto matemático, por lo que una asociación estadísticamente significativa puede no ser clínicamente

relevante y puede no ser causal; asimismo, una asociación estadísticamente no significativa no necesariamente es irrelevante. Puesto que es bien conocida la "cacería de significaciones" que tanto seduce a autores y editores, cabe reparar especialmente en que es frecuente encontrar y exaltar asociaciones "estadísticamente significativas pero conceptualmente espurias", como se enfatizó recientemente en la encumbrada revista Nature [10]. Pese a todo lo anterior, la reproducción acrítica del método sigue siendo parte de una especie de ritual que se retroalimenta en los distintos textos y cursos de estadística.

A la luz de estas deficiencias, a lo largo de varias décadas se han propuesto alternativas a la utilización de las pruebas de significación estadística. Dentro de ellas, la de mayor prominencia es la que reclama la comunicación de la magnitud del efecto acompañado de una medición del error asociado a su estimación. Los intervalos de confianza no solo cumplen tal exigencia sino que proveen más información que las pruebas de significación estadística a la vez que operan con una lógica que no "obliga" a dicotomizar las conclusiones.

Aunque podrían citarse decenas de trabajos que se adhieren a esta sugerencia, basta señalar que, desde 1988, el Comité Internacional de Editores de Revistas Médicas (ICMJE por sus siglas en inglés), también conocido como el "grupo de Vancouver", incorpora dentro de las recomendaciones para los autores que se incluya el cálculo de intervalos de confianza y sugiere explícitamente que se eviten los análisis que dependen exclusivamente de las pruebas de significación estadística y de los valores p [11].

Por su utilidad potencial para aquilatar las tendencias en el empleo de la estadística y pensando en que puede servir de referente para que los autores reconsideren los paradigmas inferenciales al uso, nos hemos propuesto describir la utilización reciente de las herramientas estadísticas más simples y también la de los valores p e intervalos de confianza para la presentación de resultados de investigación en los artículos publicados en una reconocida revista de la salud pública en Colombia.

Métodos

Se examinaron todos los artículos originales de la Revista Nacional de Salud Pública (RNSP) de la Universidad Nacional de Colombia en los años de 2013 a 2014. Se trata de un conjunto singular de artículos, circunscrito a un bienio dado y a una revista específica. Sin embargo, podemos entenderla como una muestra de un*superuniverso* en el sentido definido por Hagood hace ya varios decenios [12]: el de los trabajos que actualmente se realizan en esta disciplina y que se publican tras un peer review riguroso en el ámbito latinoamericano.

Se incluyeron en el estudio todos los artículos originales; es decir, aquellos donde se presentan por primera vez los resultados de un proceso de experimentación u observación, y se excluyeron los artículos de revisión, teóricos o metodológicos.

Para cada uno de los artículos, el grupo investigador identificó la presencia o no de los siguientes 12 recursos estadísticos: Tasas, frecuencias (absolutas o porcentuales), medias, medianas, desviaciones estándar, gráficos (de barras, circulares, de líneas, de tendencia u otros), valores p e intervalos de confianza.

Inicialmente, se establecieron reglas para el manejo que daría el grupo investigador a los ítems considerados. Posteriormente se realizó una prueba piloto con 20 artículos, cada uno de los cuales fue valorado por los primeros cuatro firmantes del trabajo. Para cada una de las 12 variables se calculó el coeficiente Kappa en su variante para varios clasificadores de variables dicotómicas empleando el programa EPIDAT en su versión 4.1[13]. Los resultados se presentan en la Tabla 1.

Recurso estadístico	Kappa
Valor p	0,65
Tasas	0,23
Frecuencias	0,33
Medias	0,48
Mediana	0,53
Desv. Estan	0,54
Inter. Conf	0,33
Histograma	0,19
Barras	0,71
Circulares	1,00
Líneas	0,27
Otros	0,47

Tabla 1. Valores de Kappa correspondiente a varios revisores de variables dicotómicas para cada recurso estadístico considerado en el estudio (estudio piloto con 20 artículos).

Tal resultado, ciertamente pobre, pone de manifiesto la dificultad para conseguir consistencia entre observadores, incluso en casos aparentemente sencillos. Se realizó un examen cuidadoso de las discrepancias lo cual dio lugar a ajustes adicionales. Puesto que las discrepancias fueron en su casi totalidad debidas a que las reglas operativas de los conceptos involucrados habían sido interpretadas de manera diferente por los observadores, tales ajustes consistieron básicamente en refinar o precisar dichas definiciones operativas con el fin de asegurar una adecuada consistencia en la información que habría de recogerse.

En la valoración final se incluyeron los 84 artículos que cumplían las condiciones de inclusión en el bienio. Cada investigador realizó una lectura total de todos los artículos de la muestra.

En todos los casos, se consideró como un resultado positivo el caso en que el recurso de que se tratara apareciera al menos una vez en cualquier punto de los resultados (no así, si dicho recurso solo fuera mencionado en la introducción u otra zona del artículo en alusión a un trabajo previo).

Resultados

Al momento de la revisión de los documentos, para el año 2013, la revista tenía disponible en su página Web (http://www.revistas.unal.edu.co/index.php/revsaludpublica) seis ediciones y para el 2014 contaba con tres ediciones (Tabla 2).

Año de publicación	Número de ediciones del año revisadas	Número de artículos originales	%
2014	3	23	27,4%
2013	6	61	72,6%
Total	9	84	100,0%

Tabla 2. Artículos originales según año de publicación en la Revista de Salud Pública.

El recurso estadístico que se utiliza con mayor asiduidad en la presentación de los resultados de investigación (89,3%) resultaron ser las frecuencias, sean estas absolutas o relativas, usualmente expresadas en porcentajes, seguido de los valores p (presentes en el 65,4% de los trabajos) y de los intervalos de confianza (53,6%); por otra parte, la tasa fue el recurso que se empleó con menor frecuencia (14,3%).

En relación con los gráficos, los de barras y de líneas o tendencia, fueron los que más se emplearon (15,4%), aunque de manera general los gráficos no se utilizan con frecuencia apreciable a la hora de presentar los resultados de investigación (Tabla 3).

Recurso estadístico	Frecuencia	%
Tasa	12	14,3
Frecuencia (directas o porcentuales)	75	89,3
Media	45	53,6
Mediana	14	16,7
Desviación estándar	32	38,1
Gráfico de Barras	13	15,5
Gráficos Circulares	3	3,6
Gráfico de líneas o tendencia	13	15,5
Otros gráficos	12	14,4
Valor p	55	65,5
Intervalo de confianza	45	53,6

Tabla 3. Frecuencias para el empleo los recursos descriptivos. Revista de Salud Pública Años 2013-2014 (n=84).

De manera general, en casi la mitad de los artículos (48,9%; IC95%=37,7-59,9) se utilizan simultáneamente los intervalos de confianza y valores p. La utilización aislada del valor p se produce en uno de cada seis trabajos (16,7%; IC95%=9,9-26,4) y el uso de intervalos de confianza sin que figuren valores p es sumamente reducido (4,8%; IC95%=1,3-11,7) (Tabla 4).

	Frecuencia	%	IC 95%
Sólo valor *P*	14	16,7	9,9 - 26,4
Sólo Intervalo de Confianza	4	4,8	1,3 – 11,7
Intervalo de Confianza y valor *P*	41	48,9	37,7 – 59,9
Ninguno de los dos (ni valor *P ni Intervalo de Confianza)*	25	29,8	20,3 – 40,7
Total	**84**	**100,0**	

Tabla 4. Frecuencia de utilización de valores p e Intervalos de Confianza.

Discusión

Uno de los hallazgos interesantes aportados por esta investigación concierne a la utilización de la estadística descriptiva como una única herramienta para presentar los resultados de investigación. Partiendo Partiendo del supuesto de que el aparato editorial de la revista elegida ha sido celoso en la corroboración de que las preguntas que se plantearon los autores han sido efectivamente respondidas, el hecho de que el 29,8% de los artículos la utilicen sin recurrir ni a las pruebas de significación estadística ni a los intervalos de confianza revela la capacidad expresiva de estos recursos. La utilidad de las medidas meramente descriptivas está fuera de toda duda y, en principio, muchos resultados interesantes y penetrantes pueden conseguirse gracias a ellas; nuestros resultados de algún modo convalidan esta realidad.

Una de las funciones de la estadística es proporcionar alternativas cuantitativas objetivas de manera que se eviten en lo posible la subjetividad y los sesgos en el proceso de obtención de nuevos conocimientos. Las pruebas de significación suelen ser consideradas como la expresión poco menos que ideal para satisfacer ese afán de objetividad, pues se cree que pueden generar conclusiones independientemente de las personas que las emplean [14].

Aunque la objetividad es un deseo natural y legítimo, *sensu stricto*, es al mismo tiempo una meta inalcanzable. La estadística no puede resolver totalmente este conflicto, pues todo proceso inferencial tendrá siempre un componente subjetivo. Si bien las técnicas estadísticas pueden ser muy útiles, tienden a promover tal convicción, que suele comprometer la obligación de examinar la realidad a través de un pensamiento integral [15].

En el área de la salud y propiamente en temas relacionados con la salud pública, no se han realizado reflexiones sistemáticas en nuestro ámbito acerca de la utilización de las pruebas de significación estadística como una herramienta supuestamente capaz de medir, cuantificar o valorar la existencia o no de diferencias entre los objetos que se investigan. El valor de p, sin embargo, no proporciona por sí mismo información acerca de la importancia global o el significado cualitativo de los resultados para la práctica clínica, ni aporta información sobre lo que podría suceder en el futuro, o en la población general [16].

En este mismo sentido, pensar en que el cálculo de los valores p pueden dar información suficiente que permita responder las preguntas de investigación en un mundo complejo puede configurar una falacia, ya que desde su formulación, el cálculo de los valores p solo remite a conclusiones dicotómicas. Consecuentemente, su uso puede, principalmente en dos formas, amenazar la construcción de conocimiento útil para la toma de decisiones en el mundo real [17].

Por una parte, las preguntas de investigación que se encaran bajo la lógica de los valores p serán solo aquellas que se preguntan si existen o no diferencias entre los grupos que se comparan. Por otra parte, apelar a los valores p como herramienta única para la presentación de resultados de investigación no sería suficiente, ya que la información que podría aportar solamente ayudaría a construir un discurso parcializado, fragmentado y a menudo inconexo con la vida misma.

Cabe recordar la idea debida a Manterola y Pineda [17], quienes afirman que la significación estadística no es nada más que eso, "la significación estadística", que en ocasiones puede ser positiva y clínicamente irrelevante, o negativa, sin que eso signifique necesariamente que no hay diferencias reales entre las variables en estudio.

A pesar de que ya han pasado 27 años desde que el Comité Internacional de Editores de Revistas Médicas (CIERM) recomendó evitar que se dependa exclusivamente de las pruebas de hipótesis para la presentación de los resultados de investigación, teniendo en cuenta que no logran transmitir información sobre la magnitud del efecto [11] y después de décadas de objeciones a cargo de diversos autores del área de la estadística [5],[8],[17],[18],[19],[20],[21]; las pruebas de significación estadística sin que siquiera vengan acompañadas de intervalos de confianza se siguen utilizando de manera regular en la presentación de los resultados de investigación. Cohen [22] , en un trabajo intensamente citado en las últimas dos décadas, consideraba que el uso del valor p y el planteamiento de las pruebas de hipótesis no pasan de ser una ritualización

presente durante más de 40 años; la presente revisión extiende 20 años más dicha realidad. Nuestros resultados revelan que tales advertencias no son cabalmente tenidas en cuenta en la investigación salubrista en nuestro ámbito geográfico.

En un estudio anterior, que tuvo un objetivo similar al de la presente investigación [4], se encontró que el 21% (IC95%=16-26) (n=306) de los artículos revisados del periodo 2005 a 2006, utilizan sólo valores p para la presentación de resultados de investigación. Este resultado se asemeja mucho al hallado por nosotros, (16,7%, IC95%=9,9 - 26,4); el extremo inferior del intervalo de confianza conduce a confiar en que por lo menos el 10% de los trabajos incurren en esta práctica explícitamente condenada por las autoridades metodológicas menos cuestionadas.

Frente al uso de los intervalos de confianza sin apelar a los valores p, en el estudio mencionado [4] se encontró que el 14% (IC95%= 9-17) de los artículos del período 2005-2007 estaban en ese caso. En nuestro caso el resultado fue algo "peor", (4,8%; IC95%=1,3 – 11,7).

El presente estudio tiene como insuficiencia que los datos se reducen a un período breve y conciernen a una sola revista. No obstante, la idea de ofrecer una semblanza de las prácticas actuales se cumple a pesar de tal limitación.

Conclusiones

La estadística descriptiva es una herramienta que se utiliza asidua y adecuadamente en la presentación de resultados. En materia inferencial, sin embargo, las críticas y advertencias que sugiere el Comité Internacional de Editores de Revistas Médicas, según las cuales ha de evitarse el uso exclusivo de las pruebas de significación estadística para el análisis y presentación de los resultados de investigación distan de ser tenidas en cuenta por una parte importante de los autores. Se pone de manifiesto asimismo que los editores de la revista estudiada no siempre exigen su cumplimiento.

Notas

Conflictos de intereses

Los autores han completado el formulario de declaración de conflictos de intereses del ICMJE, y declaran no haber recibido financiamiento para la realización del artículo y no tener otros conflictos de intereses con la materia del artículo. Los formularios pueden solicitarse al autor o la *Revista*.

Referencias

1. Läärä E. Statistics: reasoning on uncertainty, and the insignificance of testing null. Annales Zoologici Fennici. BioOne. 2009:138-57. | CrossRef |
2. Fisher RA. Statistical methods for research workers. Edinburgh: Oliver and Boyd;1950.
3. Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. Biometrika;1928:175-240.
4. Silva-Ayçaguer 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 May 19;10:44. | CrossRef | PubMed |
5. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999 Jun 15;130(12):995-1004. | PubMed |
6. Nicholls N. Commentary and analysis: the insignificance of significance testing. Bull Am Meteorol Soc. 2001;82(5):981-6. | Link |
7. Armstrong JS. Statistical significance tests are unnecessary even when properly done and properly interpreted: Reply to commentaries.J of Forecasting 2007;(3):335-36. | Link |
8. Hubbard R, Lindsay RM. Why P values are not a useful measure of evidence in statistical significance testing. Theory Psychol. 2008;18(1):69-88. | CrossRef |
9. Ayçaguer LCS. Cultura estadística e investigación científica en el campo de la salud: una mirada crítica. Madrid: Díaz de Santos; 1997
10. Nuzzo R. Scientific method: statistical errors. Nature. 2014 Feb 13;506(7487):150-2. |CrossRef | PubMed |
11. Uniform requirements for manuscripts submitted to biomedical journals. International Committee of Medical Journal Editors. Br Med J (Clin Res Ed). 1988 Feb 6;296(6619):401-5.| PubMed |
12. Hagood MJ. Statistics for sociologists. New York: Reynal and Hitchcock; 1941.
13. EPIDAT: Análisis Epidemiológico de Datos sergas.es [online]. | Link |
14. Silva LC, Benavídes A. Apuntes sobre subjetividad y estadística en la investigación. Rev Cuba Salud Pública. 2003;29(2):170-3. | Link |
15. Fethney J. Statistical and clinical significance, and how to use confidence intervals to help interpret both. Aust Crit Care. 2010 May;23(2):93-7. | CrossRef | PubMed |
16. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005 Aug;2(8):e124. | PubMed |
17. Manterola C, Pineda V. El valor de« p» y la« significación estadística»: Aspectos generales y su valor en la práctica clínica. Rev Chil Cir. 2008;60(1):86-9. | CrossRef |
18. Rozeboom WW. The fallacy of the null-hypothesis significance test. Psychol Bull. 1960 Sep;57:416-28. | PubMed |
19. Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of P values and evidence. J Am Stat Assoc. 1987;82(397):112-22. | CrossRef |
20. Anderson DR, Burnham KP, Thompson WL. Null hypothesis testing: problems, prevalence, and an alternative. J Wildl Manag. 2000;64(4):912-23. | CrossRef |
21. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods. 2000 Jun;5(2):241-301. | PubMed | Link |
22. Cohen J. The earth is round (p<. 05): Rejoinder. 1995;50(12). | CrossRef |

Correspondencia a:
[1] Tr 5 # 41-15
Bogotá
Colombia