ARTÍCULO ORIGINAL

# Uso de pruebas estadísticas inferenciales en la Revista Cubana de Medicina Militar

## Use of Inferential Statistical Tests in the Cuban Journal of Military Medicine

**Carlos Rafael Quevedo Fonseca, Rubén Arturo Cárdenas Díaz**

Universidad de Ciencias Médicas de las FAR. La Habana, Cuba

**RESUMEN**

**Introducción:** desde hace años, existe un debate sobre el uso de las pruebas estadísticas inferenciales en los reportes de resultados de investigación, se destaca la crítica al empleo de las pruebas de significación estadística y sus limitaciones.
**Objetivos:** determinar la frecuencia de empleo de las pruebas de significación estadística (PSE) e intervalos de confianza (IC) por tipos de estudio publicado, cómo se reflejan los resultados de estas, la influencia del tamaño de la muestra, así comosu vinculación con las conclusiones.
**Resultados:** en el periodo 2010 - 2015 de 150 artículos originales, 98 % fueron descriptivos o explicativos y de ellos, el 95 % emplea las PSE, solas o con IC. Predomina el uso de las PSE solas (69 % de los trabajos). En el 25 % se explica la selección del nivel de significación utilizado y el 53 % de los estudios reflejan las cifras exactas de las pruebas realizadas. Solo el 15 % menciona la influencia del tamaño de la muestra en relación con los resultados de las pruebas estadísticas. En las conclusiones, el 86 % de los artículos se refieren adecuadamente a los objetivos del estudio.
**Conclusiones:** predomina el uso de las PSE e IC, fundamentalmente de las PSE, más de la mitad de los trabajos mencionan los resultados precisos de las pruebas, la mayoría no argumenta la relación de estos resultados con el tamaño de la muestra y los autores elaboran las conclusiones de acuerdo con los objetivos planteados en el estudio.

**Palabras clave:** pruebas de significación estadística; intervalos de confianza; pruebas de hipótesis; inferencia; tamaño de la muestra.

**ABSTRACT**

**Introduction:** For years there has been a debate about the use of inferential statistical tests in the reports of research results, highlighting the criticism to the use of tests of statistical significance and its limitations.
**Objectives:** To determine the frequency of use of statistical significance tests (SST) and confidence intervals (CI) by published study types, how the results are reported, and the influence of sample size, as well as their relationship with the conclusions.
**Results:** In the period 2010-2015 of 150 original articles, 98% were descriptive or explanatory and of them, 95% used SST alone or with CI. The use of SST alone (69% of the articles) predominates. In 25% the significance level selection is explained and 53% of the studies reflect the exact figures of the tests performed. Only 15% mentions the influence of sample size on the results of statistical tests. In the conclusions, 86% of the articles refer adequately to the objectives of the study.
**Conclusions:** SST and CI use predominate, mainly SST, more than half of the studies mention the precise results of the tests, most do not argue the relation of these results to the sample size and the authors elaborate the conclusions in accordance with the objectives set out in the study.

**Keywords:** statistical significance test; confidence intervals; hypothesis test; inference; sample size.

## INTRODUCCIÓN

Desde hace varios años han sido publicados diversos artículos en revistas médicas y de otros campos de la ciencia donde se analizan los argumentos a favor y en contra del informe de resultados de investigación, empleando las pruebas de significación estadística (PSE) con los valores P, los intervalos de confianza (IC) o ambos.[1]

Al margen de la necesidad o no del empleo de análisis estadísticos en determinados estudios publicados, resulta de enorme importancia para analizar los resultados, tanto para los autores como para los revisores, conocer el uso (y el mal uso) de estos estadígrafos, fundamentalmente, a partir de las propias publicaciones donde son ampliamente empleados.

El uso del valor P no aporta mucho de lo que puede necesitar realmente el investigador, además, no aporta datos de la importancia, relevancia o significación clínica cuando se investiga en este campo. El IC permite conocer la magnitud y precisión del efecto observado. El tamaño de la muestra garantiza mayor o menor precisión en el intervalo (junto con la variabilidad de los datos). Aunque tanto las PSE como los IC pertenecen a la misma base matemática "frecuentista", estos últimos proporcionan más información que las PSE. [2]

En el uso de las PSE, aparentemente y a menudo, se obvia el juicio, es decir, el proceso cognoscitivo de analizar el resultado, que no se puede dejar solo al valor de la P, pues interpretar de la misma forma el rechazo de una hipótesis nula por una alternativa particular, independientemente de si la muestra es pequeña o grande, no es correcto. De la misma forma, no siempre el rechazo de la hipótesis nula proporciona el mismo grado de evidencia ante una alternativa específica, independientemente del poder estadístico del test usado.[3]

Por otra parte, el IC describe la incertidumbre del resultado de un estadígrafo y dentro de su rango de valores, podemos decir que razonablemente, yace el verdadero valor del efecto. Cuando el intervalo es estrecho, se conoce el tamaño del efecto con mayor precisión; si el intervalo es más amplio, la incertidumbre es mayor. Cuando es demasiado amplio, tenemos poco conocimiento del efecto y se necesita más información.[4-5]

A diferencia del valor P, el IC señala la dirección del efecto bajo estudio y este último puede ser utilizado como medida de significación estadística, ya que al no incluir el valor "cero efecto", puede considerarse el resultado estadísticamente significativo.[4-5]

A partir del debate y la crítica al uso de las estadísticas frecuentistas, del interés despertado por el artículo de *Sarría y Silva*,[2] los autores nos interesamos en examinar el empleo de los procedimientos estadísticos bajo escrutinio en la Revista Cubana de Medicina Militar. Los objetivos planteados fueron: determinar la frecuencia con que se utilizan las PSE e IC de acuerdo con el tipo de estudio publicado, cómo se reflejan los resultados de las pruebas estadísticas, la influencia en estos del tamaño de la muestra y la vinculación adecuada o no con las conclusiones.

## MÉTODOS

Se realizó un estudio descriptivo de tipo bibliométrico, en el que fueron examinados todos los artículos publicados en la sección de "trabajos originales" de la Revista Cubana de Medicina Militar durante el periodo 2010 hasta 2015.

De todos los artículos originales publicados, fueron analizados los procedimientos estadísticos utilizados y se identificaron los trabajos que emplearon técnicas estadísticas inferenciales comunes (PSE, IC o ambos).

Estos trabajos se clasificaron en descriptivos (cuando caracterizaron o exploraban un fenómeno) y explicativos (cuando se trataba de definir relaciones causales) y en ellos se examinó la frecuencia con que fueron empleadas las PSE e IC, si los niveles de significación fueron prefijados de forma argumentada teniendo en cuenta el fenómeno específico en estudio y la presentación de los resultados de las pruebas de hipótesis con los valores numéricos exactos, o su ausencia.

Además se analizó si se mencionan adecuadamente, al analizar los resultados de las pruebas de significación estadística, el tamaño de la muestra como explicación de estos resultados y por último, si las conclusiones se refieren directamente a los resultados estadísticos en lugar de los objetivos del artículo.

Los resultados de este análisis fueron tabulados mediante frecuencias absolutas para su presentación a los lectores.

## RESULTADOS

Desde el año 2010 hasta el 2015 - ambos incluidos - fueron publicados 150 artículos en la sección "trabajos originales" de la Revista Cubana de Medicina Militar, de ellos el 98 % es de tipo descriptivo o explicativo; hubo 3 (2 %) que no estuvieron enmarcados en estas categorías. (Tabla 1).

**Tabla 1.** Distribución de artículos originales en la Revista Cubana de Medicina Militar según el tipo de estudio

| Tipos de estudio | No. | % |
|---|---|---|
| Descriptivos | 125 | 83 |
| Explicativos | 22 | 15 |
| Otros | 3 | 2 |
| **Total** | **150** | **100** |

En el 95 % de los trabajos descriptivos o explicativos fueron empleadas las PSE, solas o con IC. Es predominante el uso las PSE solas, en el 69 % (68 % de los descriptivos y 73 % de los explicativos). Se señalan como dudosos 4 trabajos donde, a pesar de estar recogidos en los Métodos, el uso de pruebas estadísticas no son mencionadas en los acápites Resultados ni Discusión. Los IC solos no se utilizaron en ningún artículo. (Tabla 2).

**Tabla 2.** Frecuencia de empleo de pruebas estadísticas inferenciales (PSE, IC o ambos), de acuerdo con el tipo de estudio, en los artículos originales de la Revista Cubana de Medicina Militar

| Pruebas estadísticas | Descriptivos | | Explicativos | | Total | |
|---|---|---|---|---|---|---|
| | No. | % | No. | % | No. | % |
| PSE | 44 | 68 | 16 | 73 | 60 | 69 |
| IC y PSE | 19 | 29 | 4 | 18 | 23 | 26 |
| Dudoso | 2 | 3 | 2 | 9 | 4 | 5 |
| **Total** | **65** | **100** | **22** | **100** | **87** | **100** |

Al analizar la presencia, en los Métodos, de la explicación sobre el nivel de significación empleado, solo en 10 trabajos (11 %) se pudo constatar (7 descriptivos - 11 % de estos - y 3 explicativos - 14 %). En los demás fue fijada sin explicación añadida, como "un 5 % de error" o "95 % de confianza" o "p <= 0,05"; o no se menciona qué resultados fueron considerados significativos.

En cuanto al reflejo de las cifras precisas de las PSE y/o IC en los Resultados; en el 53 % de los trabajos se refleja, más en los explicativos (64 %) que en los descriptivos (49 %). (Tabla 3).

Al discutir los resultados obtenidos, solo en el 15 % de los artículos (17 % descriptivos y 18 % explicativos), se menciona la influencia del tamaño de la muestra en los resultados de las pruebas estadísticas. (Tabla 4).

En cuanto a las Conclusiones, elaboradas a partir de dar salida a los objetivos del trabajo; se observa que en el 86 % de los artículos (88 % descriptivos y 82 % explicativos) estas fueron apropiadamente redactadas. (Tabla 5).

**Tabla 3.** Reflejo, en Resultados, del uso de las pruebas estadísticas en los artículos que usan pruebas de inferencia estadística en los artículos originales de la Revista Cubana de Medicina Militar

| Reflejo del uso de las pruebas en los Resultados | Descriptivos | | Explicativos | | Total | |
|---|---|---|---|---|---|---|
| | No. | % | No. | % | No. | % |
| Recogen el uso de las pruebas con resultados precisos | 32 | 49 | 14 | 64 | 46 | 53 |
| No recogen resultados precisos o no mencionan las pruebas | 33 | 51 | 8 | 36 | 41 | 47 |
| **Total** | **65** | **100** | **22** | **100** | **87** | **100** |

**Tabla 4.** Argumentación, en la Discusión, sobre la influencia del tamaño de la muestra en los resultados de los artículos que usan pruebas de inferencia estadística en los artículos originales de la Revista Cubana de Medicina Militar

| Argumentación en la discusión | Descriptivos | | Explicativos | | Total | |
|---|---|---|---|---|---|---|
| | No. | % | No. | % | No. | % |
| Argumentan la relación con el tamaño de la muestra | 11 | 17 | 4 | 18 | 15 | 17 |
| No argumentan relación con tamaño de muestra | 54 | 83 | 18 | 82 | 72 | 83 |
| **Total** | **65** | **100** | **22** | **100** | **87** | **100** |

**Tabla 5.** Vinculación apropiada o no de las conclusiones con los resultados de las pruebas estadísticas en los artículos originales de la Revista Cubana de Medicina Militar

| Reflejo en las conclusiones | Descriptivos | | Explicativos | | Total | |
|---|---|---|---|---|---|---|
| | No. | % | No. | % | No. | % |
| Se refieren directamente a las pruebas estadísticas | 8 | 12 | 4 | 18 | 12 | 14 |
| Se refieren a los objetivos del trabajo | 57 | 88 | 18 | 82 | 75 | 86 |
| **Total** | **65** | **100** | **22** | **100** | **87** | **100** |

## DISCUSIÓN

Se observó que fueron empleadas las pruebas de inferencia en más de la mitad de los artículos originales, con amplio predominio de las PSE. En muchos de los casos tanto las PSE como los IC fueron utilizados de forma inercial, incluyendo el nivel de significación, sin contextualizar el problema en estudio, lo cual conduce a más errores que aciertos.[6] Los que aún apoyan el uso de las PSE realizan recomendaciones específicas para su empleo, por sus reconocidas limitaciones y llegan a atribuir estas últimas a errores, incomprensiones e interpretaciones incorrectas.[7-8]

*Greenland y cols.* plantean que la interpretación de las pruebas estadísticas y los conceptos subyacentes imponen una demanda cognitiva intensa, además, no resultan simples, intuitivas y a prueba de errores. Como consecuencia hay una epidemia de definiciones cortas e interpretaciones que, aunque erradas, están presentes en la literatura científica[3] y dan lugar a superficialidad o evidencias de desconocimiento en el empleo de las pruebas estadísticas.

Es importante recordar que el valor P no cuantifica la magnitud de la diferencia encontrada entre dos grupos, sino la probabilidad de haber observado esa diferencia si en realidad no existe ninguna[4] y que significación estadística no es sinónimo de relevancia clínica. Sin embargo, el amplio uso de las PSE, por encima de los IC, señala desconocimiento de los autores (y revisores y comité editorial...) lo cual lleva a ignorar las alternativas que se ofrecen.[2]

En un reciente editorial de la revista Medicentro Electrónica, los autores llaman la atención en las ventajas del uso de los IC, así como el empleo en las conclusiones de los resultados estadísticos en lugar de los objetivos del trabajo.[9] Incluso cuando se considera necesario, puede ser calculada la P a partir del IC, sin olvidar que estos son más útiles que el valor P.[10-11]

Una P con un valor "significativo" no rechaza de forma absoluta una hipótesis; no olvidar que en el cálculo de las probabilidades influyen varios factores, incluido el azar. Cuando una muestra es grande, hay menos influencia del azar; pero también cualquier resultado puede ser estadísticamente significativo al igual que puede no serlo con una muestra pequeña. Pocos artículos mencionan la influencia del tamaño de la muestra en el resultado del análisis estadístico, pero mayormente de forma mecánica, señalando la falta de significación por lo pequeña de la muestra; lo cual es obvio, por las propias características y limitaciones de las PSE. Igualmente, cuando se utiliza el IC no se realizan comentarios de estos resultados y su relación con el tamaño de la muestra, aún en presencia de IC extremadamente amplios y consecuentemente menos precisos.

La pertinencia de las pruebas empleadas debe ser tenida en cuenta también, pues se ha recomendado evitar interpretaciones estrictas de los valores de P y los IC en los estudios observacionales, donde carecen de bases teóricas, así como dejar de interpretarlos como si midieran la probabilidad de la hipótesis.[12]

Por otra parte, la selección del valor de α se realiza convencionalmente, sin examinar específicamente el fenómeno, práctica rechazada incluso por los defensores del uso de las PSE. *Sarría*[2] menciona que el propio Fisher subraya la necesidad del uso flexible de los niveles de significación, de acuerdo con lo que se conoce del fenómeno estudiado y las posibilidades prácticas del experimento; pero no establecer un umbral universal, generalmente α = 0,05.

La mención de las cifras de los resultados de las PSE o IC mostró una situación más favorable que otros aspectos examinados, pues la mayor parte de los autores ponen de manera explícita el valor de la P en el contexto de las pruebas realizadas y los IC con sus límites. Asimismo, la gran mayoría de los artículos mostraron conclusiones que respondían a los objetivos del estudio, en lugar de referirse directamente a las pruebas estadísticas.

Incrementar la calidad de las publicaciones no solo garantiza mejor calidad de la ciencia que comunican y mayor impacto práctico, sobre todo de la investigación clínica,[13] sino que influye en la autoridad y reputación del medio. Se estima que el 85 % de la investigación científica publicada no tiene valor, y uno de los factores para disminuir esto es la adopción de métodos estadísticos más apropiados, así

como la preparación de los investigadores en metodología de la investigación y la alfabetización en estadísticas.[14] Para mejorar la calidad de la investigación y reducir la ciencia sin valor, se necesita de la llamada cultura de pospublicación: revisar, criticar lo negativo o cuando falla la replicación; también en la incorporación de los resultados a la práctica y a las políticas, y como marco de discusión.[15]

Varios de los autores citados insisten en las ventajas del uso de los IC sobre las PSE,[1,2,4,11,16] en parte por las limitaciones de esta última y también por la superioridad de los IC.

Actualmente la crítica a los problemas de las pruebas estadísticas inferenciales se ha avivado, con el editorial de la revista *Basic and Applied Social Psychology*, en el que se prohibió el uso de PSE e IC en los trabajos que le presenten a publicación, y entre otros argumentos esperan un incremento en la calidad en los manuscritos que le lleguen, al liberar a los autores del uso de estas pruebas como obstáculo al pensamiento creativo.[17]

Los autores coindicen en que es necesario mejor preparación en los investigadores a la hora de realizar su trabajo, de redactarlo; así como de los revisores para evaluarlos. Y naturalmente, el comité de redacción o comité editorial, que tiene que dar el veredicto final, partiendo desde una clara política de evaluación, hasta el impacto profesional y social de los artículos que son aprobados para publicar.

Se concluye que predomina el uso de las PSE e IC en los artículos originales publicados en la Revista Cubana de Medicina Militar, fundamentalmente de las PSE, más de la mitad de los trabajos mencionan los resultados precisos de las pruebas inferenciales realizadas, la gran mayoría no argumenta la relación de estos resultados con el tamaño de la muestra y los autores elaboran las conclusiones de acuerdo con los objetivos planteados en el estudio.

## REFERENCIAS BIBLIOGRÁFICAS

1. Spanos A. Recurring controversies about P values and confidence intervals revisited. Ecology. 2014 [cited 2016 Nov 22];95(3):645-51. Available from: http://onlinelibrary.wiley.com/doi/10.1890/13%AD1291.1/full .

2. Sarria Castro M, Silva Ayçaguer LC. Las pruebas de significación estadística en tres revistas biomédicas: una revisión crítica. RevPanam Salud Pública. 2004 [cited 2016 Nov 22];15(5):300-6. Disponible en: http://www.sld.cu/galerias/pdf/sitios/anestesiologia/significacion_estadistica.pdf

3. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical Tests, P-values, Confidence Intervals, and Power: A Guide to Misinterpretations. The American Statistician, Online Supplement 1 2016 [cited 2016 Nov 11]. Available from: http://link.springer.com/content/pdf/10.1007%2Fs10654-016-0149-3.pdf

4. Clark ML. Los valores P y los intervalos de confianza: ¿en quién confiar? Revista Panamericana de Salud Pública. 2004 [citado 27 May 2016];15(5):293-6. Disponible en: http://www.scielosp.org/pdf/rpsp/v15n5/21999.pdf

5. Doll H, Carney S. Statistical approaches to uncertainty: p values and confidence interval sunpacked. EBM Notebook. 2005 [cited 2016 May 27];10:102-3. Available from: http://www.cebm.net/wp-content/uploads/2014/12/Statistical-approaches-to-uncertainty.pdf .

6. Bland JM, Altman DG. Best (but oft forgotten) practices: testing for treatment effects in randomized trials by separate analyses of changes from base line in each group is a misleading approach. Am J Clin Nutr. 2015 [cited 2016 Nov 11];102(5):991-4. Available from: http://ajcn.nutrition.org/content/102/5/991.full

7. Grieve A. How To Test Hypothesis. PSI Journal Club. 2015 [cited 2016 Nov 19]. Disponible en: http://www.psiweb.org/docs/default-source/resources/psi-subgroups/publications/journal-club/201507---Sept-2015-Journal-Club/andy-grieve---2015-(04)-sep---how-to-test-hypotheses-if-you-must-psi-virtual-journal-club.pdf .

8. Murtaugh PA. In defense of P values. Ecology. 2014 [cited 2016 Nov 11];95(3):611-7. Available from: https://ir.library.oregonstate.edu/xmlui/bitstream/handle/1957/49298/MurtaughPaulStatisticsDefensePValues.pdf?sequence=1

9. Batista Hernández NE, Hernández Moreno VJ, Guirado Blanco O. Pensamiento estadístico: el valor "p" y los intervalos de confianza. Medicent electron. 2016 [citado 19 Nov 2016];20(2):93-94. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1029-30432016000200001

10. Altman DG, BlandJM. How to obtain a P value from a confidence interval. BMJ. 2011 [cited 2016 Nov 19];343:d2304. Available from: http://www.bmj.com/content/343/bmj.d2304

11. Du Prel JB, Hommel G, Röhrig B, Blettner M. Confidenceintervalor P-value? Dtsch ArzteblInt. 2009 [cited 2016 Nov 01];106(19):335-9. Available from: http://dx.doi.org/10.3238%2Farztebl.2009.0335

12. Poole C. Low P-Values or Narrow Confidence Intervals: Which Are More Durable?Epidemiology. 2001 [cited 2016Nov 11];12(3):291-4. Available from: http://www.tc.umn.edu/~alonso/Poole_Epidemiology_2001.pdf

13. Loannidis JPA. Why Most Clinical Research Is Not Useful. PLoSMed. 2016 [cited 2016 Nov 11];13(6):1-10. Available from: http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002049

14. Loannidis JPA. How to Make More Public Research True. PloS Med. 2014 [cited 2016 Nov 11];11(10):e1001747. Available from: http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001747

15. Bastian H. A Stronger Post-Publication Culture Is Needed For Better Science. PLoS Med. 2014 [cited 2016 Nov 11];11(12):1-3. Available from: http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001772

16. Attia A. Why should researchers report de confidence interval in modern research? Middle East Fertility Society Journal. 2005 [cited 2016 May 27];1(10):78-81. Available from: http://www.bioline.org.br/pdf?mf05015

17. Trafimow D, Marks M. Editorial. Basic and Applied Social Psychology. 2015 [cited 2016 Nov 11];37(1):1-2. Available from: http://www.tandfonline.com/doi/pdf/10.1080/01973533.2015.1012991

Recibido: 28 de diciembre de 2016.
Aprobado: 28 de enero de 2017.

*Dr. C. Carlos Rafael Quevedo Fonseca.* Universidad de Ciencias Médicas de las FAR.
Dirección electrónica: quevedo@infomed.sld.cu