RICYDE. REVISTA INTERNACIONAL DE CIENCIAS DEL DEPORTE
http://www.ricyde.org
doi:10.5232/ricyde

RICYDE. Revista Internacional de Ciencias del Deporte
Volumen XI - Año XI
Páginas:101-103 ISSN:1885-3137
Nº 40 - Abril - 2015

EDITORIAL
http://dx.doi.org/10.5232/ricyde2015.040ed

La significación estadística no es suficiente. [Statistical significance is not enough].

María I. Barriopedro
Universidad Politécnica de Madrid

Introducción

En el contexto de la actividad investigadora en Ciencias del Deporte, la forma actual de las Pruebas de Significación de la Hipótesis Nula (PSHN) es el método inferencial inductivo utilizado por excelencia en los informes de investigación de este ámbito. Las críticas al uso de estas pruebas son tan numerosas (Balluerka, Gomez e Hidalgo, 2005; Kline, 2004; Nickerson, 2000) que sería difícil abordarlas de manera exhaustiva en un trabajo como el presente. Estas críticas varían desde aquellas que se centran en la mala utilización de las mismas en los informes de investigación hasta aquellas que cuestionan su utilidad científica y proponen su abandono (Anderson, Burnham y Thompson, 2000; Harlow, Mulaik y Steiger, 2013).

La polémica sobre las PSHN ha sido tan intensa en las últimas décadas, que algunas asociaciones científicas y profesionales, como la *American Psychological Association (APA) o la American Education Research Association (AERA)*, recomiendan cambios en las políticas editoriales de las revistas científicas con respecto a la utilización de las mismas (Thompson, 1996; Wilkinson y Task Force on Statistical Inference, 1999). Los cambios propuestos no suponen alternativas al modelo de inferencia estadística clásica sino una forma de compensar alguna de las limitaciones de las PSHN. Estas recomendaciones hacen referencia fundamentalmente a tres aspectos: a) la necesidad de incluir estimaciones del tamaño del efecto, b) la necesidad de tener en cuenta la potencia en los estudios y c) la conveniencia de utilizar intervalos confidenciales.

Tamaño del efecto

El tamaño del efecto es el grado en que un fenómeno está presente en la población o el grado en que la hipótesis nula es falsa (Cohen, 1988). Ya en la cuarta edición del Manual de publicación de la APA (1984) se señala que los valores de significación estadística (valores *p*) no son índices aceptables del tamaño del efecto, puesto que dependen del tamaño de la muestra. Así, por ejemplo, es posible encontrar valores de p muy pequeños asociados a diferencias de medias despreciables obtenidas en muestras muy numerosas.

La significación estadística nada nos dice de la magnitud, relevancia o significación práctica de un efecto (Fritz, Scherndl y Kühberger, 2012; Kirk, 1996). El valor p sólo nos informa de la probabilidad de encontrar unos datos al menos tan discrepantes como los obtenidos si la hipótesis nula fuera verdadera. En su quinta edición, el Manual de publicación de la APA (2001), recogiendo las sugerencias del grupo de trabajo sobre inferencia estadística (Wilkinson y col., 1999), recomienda informar del tamaño del efecto junto con el valor de *p*. A pesar de esta recomendación, todavía son muchas las investigaciones publicadas que no incluyen índices del tamaño del efecto. La última edición señala que para que el lector pueda apreciar la magnitud o la importancia de los hallazgos de un estudio, casi siempre es necesario incluir alguna medida de tamaño del efecto en la sección de resultados (APA, 2010, p. 34). Wilkinson y col. (1999) subrayan además la importancia de informar del tamaño del efecto de cara a los estudios de potencia de futuras investigaciones y futuros meta-análisis (p.599).

Potencia

De las numerosas críticas a las PSHN, la más destacable es que los investigadores al utilizarlas no suelen tener en cuenta la potencia estadística en la planificación de sus investigaciones (Bakker y Wicherts, 2011; Clark-Carter, 1997; Cohen, 1992; Frías, García y Pascual, 1994; Sedlmeier y Gigerenzer, 1989). Los investigadores son conscientes de la necesidad de controlar la probabilidad de cometer Error de Tipo I (α), habitualmente fijada en 0,05, pero parecen menos conscientes de la necesidad de controlar la probabilidad de cometer Error de Tipo II (β), es decir, aceptar la hipótesis nula cuando es falsa. La potencia (1- β) de una prueba estadística es la probabilidad complementaria a la probabilidad de cometer Error de Tipo II. La potencia de una prueba depende del nivel de significación o probabilidad de cometer Error de tipo I (α), del tamaño del efecto y del tamaño de la

muestra. Aunque es posible realizar distintos análisis sobre la potencia, Wilkinson y col. (1999) recomiendan realizar estudios a priori, con objeto de determinar el tamaño de muestra necesario para, fijado un nivel de significación (α), obtener una potencia determinada para detectar un tamaño de efecto hipotetizado. Desde los trabajos de Cohen (1988, 1992) se suele requerir por convención una potencia mínima de 0,80, dado que habitualmente es más grave señalar que existe un efecto cuando no lo hay (Error de Tipo I) que señalar que no existe efecto cuando si lo hay (Error de Tipo II). La sexta edición del Manual de publicación de la APA señala la necesidad de tomar en cuenta seriamente la potencia estadística suministrando información que evidencie que el estudio tiene la suficiente potencia para detectar efectos de interés sustantivo (APA, 2010, p.30). Por tanto, los análisis de potencia a priori, son fundamentales. El software libre G*Power (Faul, Erdfelder, Lang y Buchner, 2007) es de inestimable ayuda para este tipo de análisis (Cumming, 2014).

Intervalos Confidenciales

La aplicación de las PSHN resulta en una decisión dicotómica con respecto a la hipótesis nula (rechazarla/no rechazarla) que no nos informa acerca del verdadero valor del parámetro. Para paliar este problema, algunos investigadores han propuesto sustituir las PSHN por la confección de intervalos de confianza (Cumming, 2014; Cumming y Finch, 2005; Reichard y Gollob, 1997). Aunque los intervalos de confianza no aportan nada relevante en el contexto de corroboración de teorías, donde se trata de establecer los factores a los que es sensible una variable dependiente, en los contextos de investigación aplicada aportan una información indispensable para valorar la importancia práctica del resultado (Botella y Barriopedro, 1995). Por otro lado, algunos estudios parecen indicar que la interpretación de los intervalos confidenciales genera menos errores que la significación obtenida en las PSHN (Coulson, Healey, Fidler y Cumming, 2012; Fidler and Loftus, 2009).

Wilkinson y col. (1999) remarcan la utilidad de los intervalos confidenciales y la importancia de compararlos con los obtenidos en estudios previos relacionados. El manual de la APA en su quinta edición (2001) señala que los intervalos confidenciales representan en general la mejor estrategia de presentación de resultados y por tanto recomienda encarecidamente su uso. En su sexta edición, no solo señala la necesidad de incluir medidas del tamaño del efecto sino también sus intervalos confidenciales (APA, 2010, p.34). El cálculo de intervalos confidenciales para el tamaño del efecto presenta ciertas dificultades al basarse en distribuciones no centrales y requerir procedimientos iterativos (Kline, 2013a). Afortunadamente en la actualidad pueden encontrase numerosos recursos que facilitan estos cálculos (Cuming, 2013; Kelley y Preacher, 2012; Kline, 2013b).

Conclusiones

La controversia sobre las PSHN ha puesto de manifiesto la insuficiencia de las mismas para realizar una valoración rigurosa de los datos obtenidos en una investigación. La inclusión en el apartado de resultados de los informes de investigación de índices del tamaño del efecto junto con sus intervalos confidenciales contribuirá a incrementar la calidad de nuestros análisis de datos y de la interpretación de los mismos. Por otro lado, informar del tamaño del efecto facilitará tanto la planificación de estudios posteriores mediante el análisis de la potencia como la realización de futuros meta-análisis. Para terminar, consideramos importante señalar la necesidad de ofrecer evidencia en los informes de investigación acerca del cumplimiento de los supuestos en los que las clásicas PSHN se basan (Wilkinson y col., 1999) y utilizar métodos robustos cuando no se tengan garantías del cumplimiento de los mismos (Wilcox, 2010).

Referencias

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC.

American Psychological Association. (2001). Publication manual of the American Psychological Association (5th ed.). Washington, DC.

American Psychological Association. (2010). Publication manual of the American Psychological Association (6th ed.). Washington, DC.

Anderson, D. R.; Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management*, 64, 912-923.
http://dx.doi.org/10.2307/3803199

Bakker, M. & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, 43, 666–678.
http://dx.doi.org/10.3758/s13428-011-0089-5

Balluerka, N.; Gomez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. Methodology: *European Journal of Research Methods for the Behavioral and Social Sciences*, 1, 55–70.
http://dx.doi.org/10.1027/1614-1881.1.2.55

Botella, J. y Barriopedro, M. I. (1995). Análisis de datos. En Fernández-Ballesteros, R. (dir.): *Evaluación de Programas*. Madrid: Síntesis.

Clark-Carter, D. (1997). The account taken of statistical power in research published in the British Journal of Psychology. *British Journal of Psychology*, 88(1), 71-83. http://dx.doi.org/10.1111/j.2044-8295.1997.tb02621.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. San Diego, CA: Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155. http://dx.doi.org/10.1037/0033-2909.112.1.155

Coulson, M.; Healey, M.; Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. *Frontiers in Quantitative Psychology and Measurement*, 1, 26. http://dx.doi.org/10.3389/fpsyg.2010.00026

Cumming, G. (2013). The new statistics: Estimation for better research. Consultado 22/11/2014. Recuperado de: http://www.thenewstatistics.com.

Cumming, G. (2014). The new statistics why and how. *Psychological science*, 25(1), 7-29. http://dx.doi.org/10.1177/0956797613504966

Cumming, G. & Finch, S. (2005). Inference by Eye: Confidence Intervals and How to Read Pictures of Data. *American Psychologist*, 60(2), 170. http://dx.doi.org/10.1037/0003-066X.60.2.170

Faul, F.; Erdfelder, E.; Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191. http://dx.doi.org/10.3758/BF03193146

Fidler, F. & Loftus, G. R. (2009). Why figures with error bars should replace p values. Zeitschrift für Psychologie. *Journal of Psychology*, 217(1), 27-37. http://dx.doi.org/10.1027/0044-3409.217.1.27

Frías, D.; García, J. F., & Pascual, J. (1994). Estudio de la Potencia de los Trabajos Publicados en «Psicológica». Estimación del Número de Sujetos Fijando Alfa y Beta. C. Arce y J. Seoane (Coords.), III Simposium de Metodología de las Ciencias Sociales y del Comportamiento, 1-057.

Fritz, A.; Scherndl, T., & Kühberger, A. (2012). A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough?. *Theory & Psychology*, 23(1), 98–122. http://dx.doi.org/10.1177/0959354312436870

Harlow, L. L.; Mulaik, S. A., & Steiger, J. H. (Eds.). (2013). What if there were no significance tests?. Psychology Press.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56, 746–759. http://dx.doi.org/10.1177/0013164496056005002

Kelley, K. & Preacher, K. J. (2012). On Effect Size. *Psychological Methods*, 17, 137–152. http://dx.doi.org/10.1037/a0028086

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10693-000

Kline, R. B. (2013a). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: APA Books. http://dx.doi.org/10.1037/14136-000

Kline, R. B. (2013b). *Beyond significance testing (2nd edition): Reader and instructor resource guide*. Recuperado de: http://pubs.apa.org/books/supp/kline/

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241–301. http://dx.doi.org/10.1037/1082-989X.5.2.241

Reichardt, C. S. & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical tests and viceversa. En L.L. Harlow, S.A. Mulaik y J.H. Steige (Eds). What if there were no significance tests?. Hillsdale, NJ: Erlbaum.

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105(2), 309. http://dx.doi.org/10.1037/0033-2909.105.2.309

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, 25(2), 26–30.

Wilkinson, L. & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604. http://dx.doi.org/10.1037/0003-066X.54.8.594

Wilcox, R. R. (2010). *Fundamentals of modern statistical methods: Substantially Improving Power and Accuracy* (2nd ed.). New York: Springer