La controversia sobre el contraste de hipótesis: Situación actual en psicología y recomendaciones didácticas

Controversy around statistical tests: Current situation in psychology and didactical implications

Carmen Díaz-Batanero, Oscar M. Lozano-Rojas y Fermín Fernández-Calderón

Universidad de Huelva, España

Resumen

La controversia sobre el uso e interpretación de los contrastes de hipótesis en psicología ha sido muy amplia en los últimos 30 años y sociedades como la American Psychological Association ha publicado recomendaciones para mejorar la práctica estadística en este campo. En este trabajo analizamos las investigaciones recientes que estudian el uso de métodos estadísticos en psicología, mostrando que aunque ha habido un avance, queda mucho por hacer para solucionar el problema. Finalizamos con algunas recomendaciones para la mejora de la enseñanza de la estadística en psicología

Palabras clave: Controversia, contraste de hipótesis, situación actual

Abstract

The significance tests controversy in psychology was very wide in the past 30 years and associations such as the American Psychological Association published recommendations to improve the statistics practice in this field. In this paper we analyse recent research that study the statistical methods in psychology and show some advance that is still not enough to solve the problem. We finish with some recommendations to improve the teaching of inference in psychology.

Keywords: Controversy, statistical tests, current situation

1. Introducción

La psicología ha recibido numerosas contribuciones de la estadística para su desarrollo como ciencia, pero del mismo modo, la estadística se ha beneficiado de métodos creados para la solución de problemas psicológicos. Al igual que otras ciencias, como la biología o la agronomía, los problemas surgidos de un contexto específico han proporcionados soluciones, que luego se han utilizado en la solución de otros en áreas científicas diferentes, al tiempo que la estadística, como conjunto de métodos para aprender de la experiencia se ha enriquecido progresivamente (Cowles, 2005).

Las primeras aplicaciones de la estadística en psicología tienen relación con la medición de características psicológicas. El uso de métodos cuantitativos en el estudio de los procesos mentales empieza con Gustav Fechner (1801-1887), quien se propuso establecer una relación entre un estímulo y la sensación perceptiva. El trabajo de este autor fue el inicio de la psicología experimental y de la medición en psicología. Según Stigler (1999), estas primeras aplicaciones tienen en común el uso de métodos experimentales que permitían determinar una línea base de conducta y las fluctuaciones con respecto a dicha línea base. Los psicofísicos del siglo XIX usaron la estadística para describir las reacciones de los sujetos experimentales ante estímulos físicos como magnitud del estímulo más error sensorial.

Díaz-Batanero, C., Lozano-Rojas, O. M. y Fernández-Calderón, F. (2019). La controversia sobre el contraste de hipótesis: Situación actual en psicología y recomendaciones didácticas. En J. M. Contreras, M. M. Gea, M. M. López-Martín y E. Molina-Portillo (Eds.), *Actas del Tercer Congreso Internacional Virtual de Educación Estadística.* Disponible en www.ugr.es/local/fqm126/civeest.html

Es interesante reflexionar sobre cómo el desarrollo histórico de la estadística y de la psicología está muy entrelazado desde el inicio de la psicología científica. La preocupación central del estudio de la psicología es la explicación de la variabilidad del comportamiento humano: en qué medida una persona se comporta de forma diferente a otra y cuáles son las razones para dicha variabilidad. La curiosidad sobre la diversidad y variabilidad ha llevado al ser humano a establecer sistemas de clasificación, medida y predicción donde la estadística ha desempeñado un papel crucial y lo más importante, donde ambas disciplinas han tenido una influencia recíproca (Stigler, 1999).

Ronald Fisher (1980-1962) junto con Galton y Pearson son los fundadores de los métodos modernos de estadística inferencial, siendo Fisher el autor que adoptó el análisis de la varianza como una herramienta de la investigación experimental. La progresión de los trabajos de Fisher también llevó a una división entre psicología experimental y correlacional (Cowles, 2005). Sin embargo las contribuciones de Fisher no se limitan a éstas, si no que es sin duda su formulación de las pruebas de significación y estimación de parámetros las aportaciones más significativas y con mayor trascendencia para el análisis de datos. La colaboración entre Jerzy Neyman y Egon Pearson dio lugar a la teoría de las pruebas de hipótesis, que, según ellos, mejoraba el planteamiento de las pruebas de significación de Fisher reconociendo explícitamente el papel de las hipótesis rivales (Gigerenzer et al., 1997).

Sin embargo, en la actualidad no existe una única metodología de la inferencia, lo que origina una serie de debates sobre la correcta aplicación de la estadística en diferentes ciencias, entre ellas la psicología. El objetivo de este trabajo es analizar la situación actual de lo que se ha venido a denominar la controversia en torno al contraste de hipótesis dentro de la psicología y proporcionar algunas recomendaciones en relación con la enseñanza del tema. El trabajo utiliza y actualiza el material previo inédito elaborado por la autora para un concurso (Díaz-Batanero, 2018).

2. Controversia sobre el uso de la inferencia en psicología

El debate filosófico relacionado con la inferencia y sus diferentes aproximaciones (Batanero, 2000) ha tenido reflejo en la práctica de la estadística en la investigación en psicología, generando en las últimas décadas numerosas críticas al uso y abuso de la inferencia. Las principales críticas enumeradas se pueden aglutinar por un lado en cuanto al procedimiento matemático en sí y por otro en cuanto al uso que se hace del procedimiento.

En cuanto al procedimiento en sí, cabe destacar que la práctica habitual del contraste de hipotesis constituye una mezcla de las propuestas de Fisher y de Neyman-Pearson (Gigerenzer, 1993). En esta lógica híbrida se establece una hipótesis nula de diferencias cero o correlación cero, sin que haya necesidad de establecer las predicciones del investigador o sus hipótesis sustantivas y se usa un nivel de significación de .05 como convención. Si los resultados son significativos, se acepta la hipótesis de investigación.

Tal como señala Gigerenzer (2004), este procedimiento hubiera sido rechazado tanto por Fisher como por Neyman y Pearson. Por un lado, las investigaciones experimentales siempre se hacen esperando un determinado efecto, pero la hipótesis nula según este procedimiento siempre se plantea estableciendo una diferencia cero o correlación cero. Es decir, la hipotesis nula se plantea a priori como falsa aunque su falsedad nunca se pone realmente en duda; es más, los cálculos se realizan asumiendo que ésta es cierta (Falk y Greenbaum, 1995).

Por otro lado, se ha señalado que el significado del resultado no significativo no es claro, dejando patente el trato asimétrico de la hipótesis nula y alternativa (Chow, 1996). Para Fisher las pruebas de hipótesis nula eran la forma más primitiva de análisis estadístico y deberían ser usadas sólo para situaciones en las que el conocimiento previo es muy limitado. Neyman y Pearson hubieran también rechazado este procedimiento porque eran partidarios de someter a prueba dos o más hipótesis estadísticas simultáneamente.

El segundo conjunto de críticas al contraste de hipótesis se centran en el mal uso e interpretación de los mismos. Numerosas revistas tomaron como criterio la obtención de resultados estadísticamente significativos para aceptar artículos en la década de los 70, aunque algunos de estos artículos presentaban resultados de poco interés práctico (Frías, Pascual y García, 2002). No era habitual calcular intervalos de confianza para los parámetros estimados, ni tampoco se solía dar información sobre el tamaño de los efectos (Falk y Greenbaum, 1995) o calcular la potencia del contraste (Valera, Sánchez y Marín, 2000).

Estas prácticas, junto con los errores de interpretación de los resultados del contraste de hipótesis produjeron como reacción numerosas críticas al uso de la inferencia estadística (Borges, San Luis, Sánchez, y Cañadas, 2001; De la Fuente y Díaz-Batanero, 2004; Morrison y Henkel, 2006; Nickerson, 2000; Harlow, Mulaik y Steiger, 2016; Valera, Sánchez y Marín, 2000; Verdam, Oort y Sprangers, 2014).

La amplitud que llegó a alcanzar esta controveersia en varios campos científicos se pone de manifiesto, por ejemplo, en el hecho de que la revista *Psychological Science* publicó un monográfico sobre el tema en 1997, donde Hunter (1997, p. 3) indicaba:

> El uso actual del test de significación es un desastre. Mientras que la mayoría de investigadores creen falsamente que el test de significación tiene una tasa de error del 5%, los estudios empíricos muestran que la tasa de error medio en psicología es del 60%- 12 veces mayor que lo que cree el investigador. La tasa de error para la inferencia usando el test de significación es mayor que la tasa de error si se usara el lanzamiento de una moneda para sustituir el estudio empírico.

El Board of Scientific Affairs de la APA creó la Task Force on Statistical Inference para analizar este debate y la práctica estadística en la investigación en psicología. Dicho comité trató de analizar la controversia sobre la aplicación de la estadística en psicología y de sugerir posibles mejoras en la metodología de investigación (Wilkinson y TSFI, 1999). Como consecuencia del debate, publicaron una sección dedicada a la presentación de resultados estadísticos del Manual de Publicaciones de la APA e impulsaron la discusión sobre posibles cambios en las prácticas actuales de análisis de datos en nuestra disciplina. Este comité y otros autores han hecho varias recomendaciones sobre muestreo, diseño, medición, instrumentos, procedimientos e interpretación de resultados (Borges et al., 2001; Fidler, 2002).

Una de las principales recomendaciones sugeridas en esta reforma estadística iniciada en 1997 por la Task Force on Statistical Inference ha estado ampliamente dirigida a persuadir o requerir alguna medida de la variabilidad en el muestreo, tal como el error estándar o intervalos de confianza (Fidler, 2006; Valera, Sánchez y Marín, 2000). La sexta edición del American Psychological Association (APA) Publication Manual (APA, 2010) incluye expresamente la recomendación de la inclusión de los intervalos de confianza y una guía de uso. Esta recomendación se basa en el hecho de que los intervalos de confianza son más sencillos de interpretar y los estudiantes de psicología los comprenden mejor que el test de hipótesis (Hunter, 1997).

3. Situación actual

Sin embargo, cuando examinamos la situación actual no percibimos mucho avance en la mejora del uso de la estadística en la investigación. Los resultados en cuanto al aporte de esta información por parte de los investigadores no son tan prometedores como hubiera sido deseable (Byrd, 2007; Sesé y Palmer, 2012). Cumming et al. (2007) en un análisis de 10 revistas de mayor impacto internacional en psicología durante los años posteriores a las recomendaciones del Task Force informan de un incremento en la incorporación de intervalos de confianza, aunque insuficiente.

Este incremento pasa desde el 3.7% de artículos con esta información en el año 1998 al 10.6% en el año 2006 y lo que es más llamativo, tan sólo un 24% de los artículos que informaban de los intervalos de confianza, aportaban una interpretación del mismo. Fidler, Thomason, Cumming, Finch, y Leeman (2004) señalan resultados similares, donde a pesar del aumento del uso de intervalos de confianza, principalmente motivada por la exigencia por parte de las revistas, esta información no se incorpora ni se comenta en la discusión de los resultados. Coulson, Healeay, Fidler y Cumming (2010) sugieren que un enfásis en el pensamiento meta-analítico junto con el uso de intervalos de confianza puede mejorar la comprensión y por tanto el mejor uso de la inferencia estadística.

La segunda de las medidas más recomendada es incluir el análisis de la potencia estadística y proporcionar información sobre el tamaño de los efectos (Frías, Pascual y García, 2002; Thompson, 2007; Vacha-Haase, 2001). Kline (2013) sugiere que la utilización del test de hipótesis como una regla de determinación dicotómica (cuando existe o no un efecto) sólo debe de aplicarse en fases iniciales de la investigación. Una vez detectada la existencia de un efecto, los siguientes pasos deben ir encaminados a la deteción de la magnitud de dicho efecto y la evaluación de la significación substantiva.

Cuando se analizan los trabajos pulicados que incluyen el tamaño del efecto, García, Ortega y De la Fuente Sánchez (2011) informan que el porcentaje de artículos publicados en revistas españolas de Psicología indexadas en JCR con esta información están en un rango de 2.9% – 31.8%, siendo además poco habitual la discusión acerca del significado sustantivo de las mismas. Tambien el estudio de Caperos y Pardo (2013) mostró que entre los articulos evaluados tan sólo un 24.3% incluian un estadístico de tamaño del efecto. Sesé y Palmer (2012) señalan que además las medidas del tamaño de efecto que se incluyen en la mayor parte de los estudios son estadísticos no ajustados, que no son inmunes a la existencia de valores atípicos o violaciones de los supuestos de aplicación.

Otras medidas propuestas son emplear procedimientos especiales (por ejemplo el test de Bonferroni) para tratar las situaciones en que se necesita realizar múltiples contrastes en la misma muestra, prestar atención a la violación de supuestos de aplicación y aplicar métodos robustos en caso de incumplimiento (Moses, 1992; Wilkinson y TSFI, 1999).

De nuevo, diferentes estudios muestran que la implementación de estas propuestas es aún deficiente (Monterde-Bort, Pascual, Frías, 2005). Un trabajo reciente de Badenes-Ribera, Frías-Navarro, Pascual-Soler y Monterde-i-Bort (2016) realizado entre profesores e investigadores en Psicología señala aún lagunas de conocimientos: métodos estadísticos robustos que no son conocidos, creencias sobre la relación entre la significación estadística la importancia de un hallazgo, el uso de expresiones incorrectas

en relación al p-valor o el desconocimiento del objetivo del estudio a priori de la potencia estadística.

En resumen, como indicó Nickerson (2000, p. 241), "La conclusión es que el contraste de hipótesis se puede interpretar o usar mal con facilidad, pero cuando se aplica con buen juicio puede ser una ayuda efectiva a la interpretacion de los datos experimentales". Compartiendo esta opinión, se debe continuarr la enseñanza y aplicación del contraste de hipótesis. No obstante, también se considera que la enseñanza de los métodos estadísticos requiere un cambio, para hacer a los estudiantes más conscientes de la necesidad de la replicación de resultados y la determinación de cúales resultados son substantivamente significativos (Kline, 2013).

4. Recomendaciones para la enseñanza

La revisión a la amplia literatura existente en torno a la controversia filosófica y al mal uso de los test de hipótesis en la investigación suscita la reflexión sobre el enfoque que es deseable dar a la enseñanza de estos contenidos. A pesar de las numerosas críticas descritas respecto al excesivo uso de las pruebas de hipótesis, que provocan un pensamiento dicótomico, diversas razones hacen que tradicionalmente se opte por este enfoque de enseñanza: persistencia de una tradición frecuentista en la mayor parte de publicaciones científicas, mayor acceso al software para el cálculo, mayor cantidad de libros de texto y manuales con esta orientación.

Nuestra recomendación es tener en cuenta, sin embargo, la problemática filosófica asociada al contraste de hipótesis, y al tipo de prueba que proporcionan y su diferencia con la demostración matemática deductiva; todo ello se debe dar a conocer a los estudiantes. Respecto al contraste de hipótesis sugerimos además en lo posible, considerar un enfoque informal (Batanero y Díaz-Batanero, 2015), donde se comenzará con la metodología de Fisher que es más intuitiva, pasando posteriormente al metodo de Neymann y Pearson y haciendo diferenciar a los estudiantes las situaciones en que conviene aplicar uno y otro enfoque.

Se realizarán algunas actividades prácticas que permitan dar a conocer a los estudiantes en forma intuitiva la metodología bayesiana, partiendo de material previamente elaborado (por ejemplo, Díaz-Batanero, 2005). Igualmente, se podría realizar una actividad práctica de aplicación de la metodología de remuestreo, basada simplemente en la simulación. Se considera que estas medidas favorecerán una formación más integral en los alumnos, así como desarrollar un punto de vista más crítico a la hora de realizar e interpretar cálculos estadísticos en investigaciones.

Adicionalmente, siguiendo las recomendaciones de la APA sobre la práctica de la estadística en Psicología se debe disminuir el énfasis en el contraste de hipótesis, complementando siempre los resultados con intervalos de confianza y estimación de los efectos que se incluirán en cada uno de los temas.

Cabe destacar que el análisis de datos es una parte del proceso investigador, al igual que la delimitación de los problemas y las hipótesis, el diseño de la recogida de datos, la medida de las variables y la elaboración de informes. Por ello, es conveniente ofrecer a los estudiantes una visión integrada del proceso investigador (Pedhazur y Schmelkin, 1991). Es importante, por tanto, contextualizar adecuadamente el aprendizaje de las herramientas analíticas, indicando sus conexiones con diseños de investigación.

Agradecimiento: Proyecto EDU2016-74848-P.

Referencias

Badenes-Ribera, L., Frías-Navarro, D., Monterde-i-Bort, H. y Pascual-Soler, M. (2015). Interpretation of the p-value: A national survey study in academic psychologist from Spain. *Psicothema, 27*, 290-295. doi: 10.7334/psicothema2014.283.

Batanero, C. (2000). Controversies around significance tests. *Mathematical Thinking and Learning*, 2(1-2), 75-98.

Batanero, C. y Díaz-Batanero C. (2015). Aproximación informal al contraste de hipótesis. En J. M. Contreras, (Ed.), *Actas de las II Jornadas de Didáctica de la Estadística, Probabilidad y Combinatoria* (pp. 135-144). Granada: Sociedad Española de Investigación en Educación Matemática.

Batanero, C., Díaz-Batanero, C. y López-Martín, M. M. (2017). Significados del contraste de hipótesis, configuraciones epistémicas asociadas y algunos conflictos semióticos. En J. M. Contreras, P. Arteaga, G. R. Cañadas, M.M. Gea, B. Giacomone y M. M. López-Martín (Eds.), *Actas del Segundo Congreso International Virtual sobre el Enfoque Ontosemiótico del Conocimiento y la Instrucción Matemáticos.* Disponibles en: enfoqueontosemiotico.ugr.es/civeos.html

Borges, A., San Luis, C., Sánchez, J. A. y Cañadas, I. (2001). El juicio contra la hipótesis nula: muchos testigos y una sentencia virtuosa. *Psicothema, 13*(1), 174-178.

Byrd J. K. (2007). A call for statistical reform in EAQ. *Education Administration Quaterly*, *43*(3), 381-391. doi: 381–39110.1177/0013161X06297137

Caperos, J. M. y Pardo, A. (2013). Consistency errors in p-values reported in Spanish psychology journals. *Psicothema*, *25*, 408-414. doi: 10.7334/psicothema2012.207

Chow, L. S. (1996). *Statistical significance: Rationale, validity and utility.* Londres: Sage.

Coulson, M., Healey, M., Fidler, F. y Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Frontiers in Psychology*, *1* (26). doi: 10.3389/fpsyg.2010.00026.

Cowles, M. (2005). *Statistics in psychology: An historical perspective*. Londres: Psychology Press.

De la Fuente, E. I. y Díaz-Batanero, C. (2004). Controversias en el uso de la inferencia en la investigación experimental. *Metodología de las Ciencias del Comportamiento*, Volumen especial 2004, 161-167.

Díaz-Batanero, C. (2005). *Apuntes sobre inferencia bayesiana*. Granada: La autora.

Díaz-Batanero, C. (2018). *Proyecto docente para el concurso a profesora titular*. Huelva: La autora.

Falk, R. y Greenbaum, C. W. (1995) Significance tests die hard: The amazing persistence of a probabilistic misconception, *Theory and Psychology, 5*(1), 75-98.

Fidler, F. (2002). The fifth edition of the APA publication manual: Why statistics recommendations are so controversial. *Educational and Psychological Measurement*, *62 (5)*, 749-770.

Fidler, F. (2006). Should psychology abandon p-values and teach CIs instead? Evidence-based reforms in statistics education. En C. Reading (Ed.), *Proceedings of the Sseventh International Conference on Teaching Statistics*. International Association for Statistical Education.

Fidler, F., Thomason, N., Cumming, G., Finch, S. y Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science, 15*(2), 119-126.

Frías, M. D., Pascual, J. y García, J. F. (2002). La hipótesis nula y la significación práctica. *Metodología de las Ciencias del Comportamiento, 4 (especial),* 181-185.

García, J. G., Ortega, E., y De la Fuente Sánchez, L. (2011). The use of the effect size in JCR Spanish Journals of Psychology: From theory to fact. *The Spanish Journal of Psychology, 14*(2), 1050-1055. doi: 10.5209/rev_SJOP.v14.n2.49.

Gigerenzer, G. (1993). The superego, the ego and the id in statistical reasoning. En G. Keren y C. Lewis (Eds.), *A handbook for data analysis in the behavioural sciences: Methodological issues* (pp. 311 - 339). Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33,* 587-606.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. y Krüger, L. (1997). *The empire of chance. How probability changed science and everyday life*. Nueva York: Cambridge University Press.

Harlow, L. L., Mulaik, S. A. y Steiger, J. H. (2016). *What if there were no significance tests?* Mahwah, NJ: Erlbaum. 2ª edición.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8(1),* 3 – 7.

Kline, R.B. (2013). *Beyond significance testing: Statistic reform in the behavioral sciences.* Washington, DC: American Psychological Association

Monterde-Bort, H., Pascual, J. y Frías, M. D. (2005). Incomprensión de los conceptos metodológicos y estadísticos: La encuesta "USABE". Presentado en el *IX Congreso de Metodología de las Ciencias Sociales y de la Salud.* Granada.

Morrison, D. E., y Henkel, R. E. (Eds.) (2006). *The significance tests controversy. A reader*, 2ª ed.. Chicago: Aldine.

Moses, L. E. (1992). The reasoning of statistical inference. En D. C. Hoaglin y D. S. Moore (Eds.), *Perspectives on contemporary statistics* (pp. 107 – 122). Washington, DC: Mathematical Association of America.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods, 5*(2), 241-301.

Pedhazur, E. J., y Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Lawrence Erlbaum.

Sesé, A. y Palmer, A. (2012). The current use of statistics in clinical and health psychology under review. *Clínica y Salud, 23*(1)

Stigler, S.M. (1999). *Statistics on the table. The history of statistical concepts and methods.* Londres: Harvard University Press.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*(5), 423-432.

Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement*, 61, 219-224.

Valera, S., Sánchez, J. y Marín, F. (2000). Contraste de hipótesis e investigación psicológica española: Análisis y propuestas. *Psicothema, 12(2),* 549-582.

Verdam, M. G., Oort, F. J. y Sprangers, M. A. (2014). Significance, truth and proof of p values: reminders about common misconceptions regarding null hypothesis significance testing. *Quality of Life Research, 23*(1), 5-7.

Wilkinson, L. y Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594 – 604.