SOBRE EL USO DE LA PRUEBA DE SIGNIFICACIÓN DE LA HIPÓTESIS NULA EN LA INVESTIGACIÓN PSICOLÓGICA

ON THE USE OF NULL HYPOTHESIS SIGNIFICANCE TESTING IN PSYCHOLOGICAL RESEARCH

*Sergio Alexis Dominguez-Lara**

Facultad de Ciencias de la Comunicación, Turismo y Psicología

Recibido: 29 de setiembre de 2016 Aceptado: 12 de octubre de 2016

RESUMEN

La prueba de significación de la hipótesis nula (PSHN) ha sido cuestionada en los últimos años debido a que invita al investigador a concluir a partir de un criterio dicotómico basado en probabilidades, pero dejando de lado las características del proceso abordado en el análisis (diferencias entre grupos, correlación, etc.). En el presente manuscrito se describe brevemente la lógica subyacente del PSHN, así como un procedimiento que en los últimos años está siendo considerado como una alternativa viable a la PSHN: la magnitud del efecto.

Palabras clave: Estadística inferencial, hipótesis nula, magnitud del efecto.

ABSTRACT

Null hypothesis significance testing (NHST) has been questioned in recent years because it invites the researcher to conclude from a dichotomous criterion based on probabilities, but leaving aside the characteristics of the process addressed in the analysis (between groups, correlation, etc.). The present manuscript briefly describes the underlying logic of the NHST, as well as, a procedure that in recent years is being considered as a viable alternative to NHST: the magnitude of the effect.

Keywords: Inferential statistics, null hypothesis, magnitude of the effect

* sdominguezmpcs@gmail.com
sdominguezl@usmp.pe

ISSN: 1817-0285 (Impresa)
ISSN: 2224-3585 (Digital)

En el proceso de investigación científica bajo un enfoque cuantitativo es frecuente el uso de técnicas estadísticas para avalar las hipótesis de investigación propuestas inicialmente. Cada diseño posee determinadas formas de ser abordado desde técnicas cuantitativas, lo que permite al investigador llegar a conclusiones sustentadas sobre una base empírica. No obstante, muchas veces las conclusiones e interpretaciones derivadas de esos resultados pueden verse sesgadas o distorsionadas si se toma en cuenta únicamente el criterio analítico de rechazo/mantenimiento de la hipótesis nula estadística. Por ello, es necesario rescatar algunas alternativas que han sido planteadas anteriormente y podrían brindar indicadores cuantitativos más esclarecedores.

Pruebas de significación de la hipótesis nula: ¿en qué consiste?

Las pruebas de significación de la hipótesis nula (PSHN) hacen referencia al proceso para rechazar o mantener una hipótesis nula (H_o) considerando valores de probabilidad (comúnmente llamados *p*-valor) asociados a determinados estadísticos (como *t*, *F*, *r*, etc.) y al error tipo I (a) (Frías, Pascual y García, 2002). Generalmente, las H_o se refieren a efectos de valor cero, es decir, diferencias de medias igual a cero (H_o: $m_1 - m_2 = 0$), correlación igual a cero (H_o: $r = 0$), etc.; y las hipótesis alternativas (H_1) se refieren a la presencia del fenómeno, por ejemplo, las diferencias de medias son diferentes de cero (H_1: $m_1 - m_2 \neq 0$), la correlación es diferente de cero (H_o: $r \neq 0$).

En psicología, por lo general, a asume el valor de .05 o .01, dependiendo de la exigencia del investigador. Esto quiere decir que para cualquier H_o planteada, si su estadístico de contraste de hipótesis se halla asociado a una probabilidad de ocurrencia (*p*-valor) menor que a, se rechaza la H_o ya que se considera que el evento no brinda evidencia a favor de la H_o, y la hipótesis alternativa es la opción más plausible.

Por lo tanto, de acuerdo a los planteamientos tradicionales, mientras más pequeña sea la probabilidad asociada al estadístico (p. ej., *p*-valor < a), los datos no apoyan la premisa de efecto cero (p. ej., la H_o) y la presencia del fenómeno es significativa. Por lo tanto, vale la pena rescatarlo e informarlo. En muchas ocasiones los científicos han confiado en la PSHN como la técnica

* sdominguezmpcs@gmail.com
sdominguezl@usmp.pe

ISSN: 1817-0285 (Impresa)
ISSN: 2224-3585 (Digital)

por excelencia del análisis de datos (Monterde-Bort, Pascual y Frías, 2006), incluso en áreas tan complejas como las experimentales.

¿Qué consecuencias traen estos procedimientos?

En el marco de la investigación empírica, concretamente en diseños comparativos (Ato, López y Benavente, 2013), es una práctica habitual utilizar la *t* de Student para explorar si existen diferencias significativas entre las medias de dos grupos, considerando el cumplimiento de sus supuestos (muestra representativa, distribuciones similares, etc.). Luego de ello, es frecuente aceptar como importante una diferencia estadísticamente significativa (*p*-valor < α), sin tener en cuenta la magnitud de las diferencias. Es decir, si el *p*-valor asociado a un valor *t* de Student (o la medida de comparación que use) es menor que α, se rechaza la H_o. Se concluye que hay diferencias estadísticamente significativas, se generaliza a la población (de ser el caso) y se termina el trabajo reportando los hallazgos. No obstante, hay algo que no es considerado rutinariamente.

De este modo, tanto un *p*-valor de .049 como de .001 harían considerar como importante una diferencia. Por ejemplo, en un estudio reciente de Hermida, Tartaglini y Stefani (2015) fueron comparados adultos mayores varones y mujeres respecto a los significados que le asignan a la jubilación. Los resultados que figuran en la Tabla 3 del manuscrito mencionado indican que las diferencias son estadísticamente significativas (*p*-valor < a) en cuatro áreas evaluadas. Es decir, la hipótesis nula (H_o: $m_1 - m_2 = 0$: las diferencias entre varones y mujeres respecto a la dimensión *Descanso* no es estadísticamente significativa) es rechazada en los cuatro análisis.

De este modo, en base a ello las autoras Hermida et al. (2015) indican que:

> Las mujeres, en comparación con los varones, creen más que la jubilación es un *Descanso* (*M2* = 5.12 vs. *M1* = 4.62 [$t_{(298)}$ = -3.126]), *Comienzo* (*M2* = 4.86 vs. *M1* = 4.13 [$t_{(298)}$ = -4.422]) o una *Continuidad* (*M2* = 3.21 vs. *M1* = 2.86 [$t_{(298)}$ = -1.971]); mientras que los varones consideran más a la jubilación como una *Pérdida* (*M1* = 3.95 vs. *M2* = 2.78 [$t_{(298)}$ = 5.928]) (los corchetes son agregados). (p. 63)

* sdominguezmpcs@gmail.com
sdominguezl@usmp.pe

ISSN: 1817-0285 (Impresa)
ISSN: 2224-3585 (Digital)

Luego, concluyen indicando que «los resultados del presente trabajo denotan que las mujeres presentan actitudes más favorables hacia la jubilación y le dan un mayor significado de Descanso, Comienzo o Continuidad» (Hermida et al., 2015, p. 64).

Además, Hermida et al. (2015), señalan lo siguiente:

> La comprobación de la hipótesis sustantiva propuesta sumaría evidencia empírica que apoya la teoría del rol social (Eagly, 1987), que trata sobre el papel que cumplen el género y los roles que devienen de las expectativas sociales acerca de este en el afrontamiento de eventos potencialmente estresantes, tales como la jubilación. (p. 65)

Ante ello, ¿la diferencia observada entre varones y mujeres es tan grande para interpretar teórica o técnicamente los hallazgos?, más adelante se responderá esta pregunta. Este último punto es de especial importancia por dos motivos: el primero, los resultados derivados de la PSHN solo indica que las diferencias no pueden ser atribuidas al azar; y el segundo, si bien existen diferencias, no hay información sobre si las diferencias son interpretables sustantivamente.

Siguiendo con los hallazgos presentados anteriormente (Hermida et al., 2015), ¿qué hubiera pasado si solo se contaba con 20 personas (10 varones y 10 mujeres)? Ya que no se posee la base de datos completa, solo puede hacerse una estimación del *p*-valor para cada comparación haciendo uso del valor *t* y los grados de libertad mediante una calculadora online.[1] En este caso, el nuevo *p*-valor para la comparación respecto a *Descanso*, *Comienzo*, *Continuidad* y *Pérdida* fue de .006, .0003, .064, y < .001, respectivamente. En este nuevo escenario, la tercera comparación pasaría a ser no significativa (es decir, la H_o no hubiera sido rechazada).

Por ello, la utilización de la PSHN como único criterio analítico no es recomendable, debido a su sensibilidad al tamaño muestral (es más probable rechazar H_o cuando más grande sea el tamaño muestral), pero es un método ampliamente aceptado debido a que la simplicidad con la que se rechaza la

[1] http://www.socscistatistics.com/pvalues/tdistribution.aspx

* sdominguezmpcs@gmail.com
sdominguezl@usmp.pe

ISSN: 1817-0285 (Impresa)
ISSN: 2224-3585 (Digital)

H_o garantizaría la objetividad (Borges, San Luis, Sánchez-Bruno y Cañadas, 2001).

De este modo, este asunto es motivo de polémica desde hace décadas (Cohen, 1988), y uno de los puntos importantes a favor de los detractores del uso de la PSHN es que frecuentemente los científicos que usan la PSHN asumen H_o para todos los procedimientos que emplean y no reportan medidas adicionales que ayuden a comprenderla mejor (Dar, Serlin y Omer, 1994). Esto trae consigo que no se lleve a cabo el objetivo de toda ciencia: la acumulación del conocimiento de forma ordenada (Monterde-Bort et al., 2006).

Por tal motivo, el uso de la PSHN sin una guía metodológica apropiada ha conducido a errores en su interpretación (Monterde-Bort et al., 2006; Sánchez-Bruno y Borges, 2005). Inclusive, la PSHN ha sido eliminada de la revista *Basic and Applied Social Psychology* (Factor de Impacto 1.31) (Trafimow y Marks, 2015). Esta acción desató una polémica con voces a favor y en contra (Woolston, 2015), pero brinda evidencias de que un sector de la comunidad psicológica está virando hacia otros métodos, sea complementando o sustituyendo definitivamente la PSHN.

Entre las posibles soluciones se destaca seguir usando la PSHN pero agregando otros resultados, como la magnitud del efecto o *effect size* (ES) (Trafimow y Marks, 2015). En tal sentido, es conveniente implementar otras sugerencias que fueron producto de un consenso: análisis descriptivo de los datos, añadir contenido teórico que ayude a interpretar las ES, incluir intervalos de confianza para todas las ES, así como, la comprobación de los supuestos para el uso de determinadas pruebas estadísticas (Borges et al., 2001; Sánchez-Bruno y Borges, 2005). De este modo, es necesario ir más allá de los hallazgos referidos al procedimiento dicotómico de *retener/rechazar* la H_o, que sería equivalente a reportar la *presencia/ausencia* del fenómeno. Pero, ¿qué es la ES?

Magnitud del efecto (ES) y significancia práctica

Para el caso de la comparación entre grupos, la ES puede ser entendida como una manera de cuantificar el tamaño de la diferencia entre grupos,

* sdominguezmpcs@gmail.com
sdominguezl@usmp.pe

ISSN: 1817-0285 (Impresa)
ISSN: 2224-3585 (Digital)

grandes o pequeñas, que permitan hablar de la importancia de la diferencia encontrada (Coe y Merino, 2002; Ledesma, Macbeth y Cortada de Kohan, 2008). En términos prácticos, si la H_o indicaba la probabilidad de *presencia/ausencia* del fenómeno, la ES indica qué tan *grande/pequeño* es el fenómeno.

La ES también puede (y debe) aplicarse a tratamientos experimentales, para que de ese modo pueda señalarse los beneficios de un tratamiento sobre otro (López et al., 2015), y las decisiones deben estar guiadas por la significancia práctica o importancia del cambio, más que solo en la significancia estadística del procedimiento en cuestión (Frías et al., 2002). El impacto del tratamiento se denomina, en este caso ES; es decir, cuando el cambio de la conducta *inicial* se aproxima a los valores de *normalidad* dentro del ámbito en el cual se está evaluando. De este modo, podría brindarse un indicador que sí haga referencia al efecto del tratamiento, así como de la significación sustantiva o significación práctica (Frías et al., 2002).

Este procedimiento puede complementar a la PSHN tradicional, ya que así podría evitarse una interpretación incorrecta de los resultados (al menos desde el punto de vista de la significación teórica) del *p*-valor asociado al estadístico de contraste. Por ejemplo, en una investigación sobre estilos de aprendizaje en universitarios, se integran de forma coherente los hallazgos de las ES con la interpretación realizada (Freiberg y Fernández, 2015).

Existen diversas fuentes en la literatura científica para la valoración de la magnitud de la ES en sus diferentes aplicaciones (Cohen, 1988; Ferguson, 2009; Fritz, Morris y Richler, 2012). Por ejemplo, en el caso de la *d* de Cohen, estimador paramétrico de la ES para el caso de comparación de dos grupos independientes, normalmente suelen agruparse en tres categorías: pequeña, cuando está alrededor de .30; moderada, cercano a .50, y grande, con valores alrededor de .80 (Cohen, 1988); aunque este último aspecto suele depender del constructo en evaluación.

No obstante, avances posteriores consideran la *d* como un caso de correlación biserial puntual (r_{bis}), y valoran su magnitud según el área de aplicación dentro de la psicología. Por ejemplo, para el caso de tratamientos o intervenciones (Hemphill, 2003) r_{bis} con valores inferiores entre .18 y .30 se consideran de magnitud media. Por otro lado, en el caso de aspectos

* sdominguezmpcs@gmail.com
sdominguezl@usmp.pe

ISSN: 1817-0285 (Impresa)
ISSN: 2224-3585 (Digital)

conductuales y cognitivos, se plantean los rangos para un nivel medio de .10 - .25 y .20 - .40, respectivamente (Bosco, Aguinis, Singh, Field y Pierce, 2015). Por último, en términos de diferencias individuales (Gignac y Szodorai, 2016), los límites para un nivel moderado son .11 y .29. Sin duda, estos puntos de corte resultan menos exigentes que los propuestos por Cohen para una ES moderada (.30 - .50) y se hallan basados en criterios empíricos.

De este modo, la ES, transformación de *t* a r_{bis}: $r_{bis} = \sqrt{t^2/(t^2 + N - 2)}$ (Abrami, Cohen y d'Apollonia, 1988) para las comparaciones entre varones y mujeres revisadas con anterioridad (Hermida et al., 2015) en *Descanso*, *Comienzo*, *Continuidad* y *Pérdida* fueron de .178, .248, .110 y .325, respectivamente. Es decir, solo las diferencias en *Comienzo* y *Pérdida* (diferencia moderada y grande, respectivamente) serían interpretables, la primera con precaución, y la tercera presenta un ES de baja magnitud. Cabe precisar que incluso con el artificio orientado al cambio del tamaño muestral (de 150 a 10 personas por grupo), la magnitud de las ES se mantiene. El lector interesado puede enviar un correo al autor solicitando un módulo de cálculo de la magnitud del efecto.

¿Qué tan grande debe ser el indicador de ES para ser considerado útil?

A pesar de la existencia de las guías mencionadas anteriormente (Bosco et al., 2015; Gignac y Szodorai, 2016; Hemphill, 2003), no existe un estándar para considerar la significación práctica de una ES, ya que tiene un carácter más cualitativo dado que el investigador valora los cambios producto de la intervención y decide si es lo que esperaba o no según planteamientos previos. Por ejemplo, si se evalúa una intervención en habilidades sociales, área comúnmente investigada por los psicólogos, y el investigador hipotetiza un r_{bis} de .30, supuesto valor hipotético en un metaanálisis, entonces una r_{bis} con valor .10 no sería significativo (Frías et al., 2002).

Asimismo, en el ámbito de la información básica, también puede darse el caso de áreas poco exploradas en las cuales aún no hay estudios previos, y si los hay, los resultados no son consistentes. En este caso, aunque la r_{bis} obtenida sea pequeña de acuerdo a determinados estándares, estos se podrían considerar como un hallazgo importante.

* sdominguezmpcs@gmail.com
sdominguezl@usmp.pe

ISSN: 1817-0285 (Impresa)
ISSN: 2224-3585 (Digital)

Reportar o no reportar, he ahí el dilema …

Al inicio de este artículo se indicaba que la H_o representa un efecto cero y que la H_1 se refería a la presencia estadística del fenómeno. No obstante, un mito bastante extendido entre los investigadores (expertos y principiantes) es que si los resultados no son estadísticamente significativos, no sirven ni deben ser reportados, o que algo se hizo mal. Probablemente, hallar un ES bajo o insignificante suscite la misma reacción. Sin embargo, también es posible que sean hallazgos genuinos que obedezcan a las características propias de la población evaluada.

En ciencia puede haber, tanto hallazgos *positivos* (los que apoyan las hipótesis de investigación y que se acostumbra ver en las revistas científicas), *negativos* (que apoyan la H_o), y *no positivos* (que brindan resultados no concluyentes). Estos dos últimos se aluden en el párrafo anterior, y que habitualmente no son publicados, pero cuya omisión puede negar a la comunidad científica ese conocimiento, generando un sesgo en las publicaciones. Estos resultados negativos pueden llegar a abarcar hasta el 50% de la información que se genera (Culebras, 2016), y necesitan ver la luz con el fin de contribuir con el desarrollo de la ciencia (Tárraga-López y Rodríguez-Montes, 2016). ¿Qué sería de la ciencia psicológica si se publica qué intervenciones no son efectivas, o qué variables no se asocian significativamente? Del mismo modo, ¿cuán importante es reportar qué instrumentos de evaluación psicológica no presentan evidencias de validez y confiabilidad para ser usados en el contexto peruano? Son preguntas que aún no tienen respuesta, pero es necesario comenzar a trazar una ruta para responderlas.

Conclusiones

De acuerdo con lo expuesto, puede decirse que con la PSHN solo se conocería la probabilidad de obtener una diferencia de medias (u otro tipo de procedimiento) como la encontrada, pero la ES brindaría información sobre la importancia de tal diferencia (Frías et al., 2002) y la posibilidad de una interpretación sustantiva de esos hallazgos.

* sdominguezmpcs@gmail.com
sdominguezl@usmp.pe

ISSN: 1817-0285 (Impresa)
ISSN: 2224-3585 (Digital)

Sin embargo, no puede dejar de considerarse el aporte de la PSHN al avance de la investigación científica; pero en la actualidad existen procedimientos que pueden ayudar a comprender de forma más precisa los resultados y proveer una interpretación más acorde con estos, como la ES.

Cabe precisar que para fines de presentación del método, solo se ha considerado el caso de diferencia entre grupos independientes, pero la mayoría de los procedimientos empleados en los análisis cuantitativos, tanto paramétrico como no paramétrico, presentan medidas propias de ES (Ferguson, 2009; Fritz et al., 2012), ya que la misma lógica presentada inicialmente puede aplicarse, por ejemplo, al caso de comparaciones entre dos o más grupos (¿Entre qué grupos la diferencias es importante?); todo ello independientemente del nivel de medición de la variable (intervalo ordinal o nominal).

Por otro lado, se alienta a los investigadores, tanto experimentados como en formación, a que utilicen las medidas de la ES, ya que independientemente de que sean métodos de reporte obligatorio en muchas revistas científicas, conocerlos provee mayor solidez a la interpretación de los hallazgos y facilita a su vez la sistematización del conocimiento (p. ej., en metaanálisis).

Finalmente, se sugiere reportar todo tipo de hallazgo: positivo, negativo o no positivo, ya que cada uno de ellos aporta al avance de la ciencia y permite generar un avance ordenado y orientado por la evidencia.

Referencias

Abrami, P. C., Cohen, P. A., & d'Apollonia, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research 58*, 151-179.

Ato, M., López, J., & Benavente, A. (2013). Un sistema de clasificación de los diseños de investigación en psicología. *Anales de Psicología, 29*(3), 1038-1059.

Borges, A., San Luis, C., Sánchez-Bruno, J. S., & Cañadas, I. (2001). El juicio contra la hipótesis nula. Muchos testigos y una sentencia virtuosa. *Psicothema, 13*(1), 173-178.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. Journal of effect size brenchmarks. *Journal of Applied Psychology, 100*(2), 431-449.

Coe, R. & Merino, C. (2002). Magnitud del Efecto: Una guía para investigadores y usuarios. *Revista de Psicología, 21*(1), 147-177.

* sdominguezmpcs@gmail.com
sdominguezl@usmp.pe

ISSN: 1817-0285 (Impresa)
ISSN: 2224-3585 (Digital)

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2.ª ed.). New York: Erlbaum, Hillsdale.

Culebras, J. M. (2016). Resultados negativos, cincuenta por ciento del conocimiento. *Journal of Negative & No Positive Results, 1*(1), 1-2.

Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62*, 75-82.

Ferguson, C. J. (2009). An effect size primer: a guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532-538.

Freiberg, A. & Fernández, M. (2015). Estilos de aprendizaje en estudiantes universitarios ingresantes y avanzados de Buenos Aires. *Liberabit, 21*(1), 71-79.

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

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2-18.

Gignac, G. E. & Szodorai, E. T. (2016). Effect size guideliness for individual differences researchers. *Personality and Individual Differences, 102*, 74-78.

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*(1), 78-80.

Hermida, P. D., Tartaglini, M. F., & Stefani, D. (2015). Actitudes y significados acerca de la jubilación: un estudio comparativo de acuerdo al género en adultos mayores. *Liberabit, 22*(1), 57-66.

Ledesma, R., Macbeth, G., & Cortada de Kohan, N. (2008). Tamaño del Efecto: Revisión teórica y aplicaciones con el sistema estadístico ViSta. *Revista Latinoamericana de Psicología, 40*(3), 425-439.

López, N., Véliz, A., Allegri, R., Soto-Añari, M., Chesta, S., & Coronado, J. (2015). Efectos del ejercicio físico sobre la memoria episódica en ancianas chilenas sanas. *Liberabit, 21*(1), 81-89.

Monterde-Bort, H., Pascual, J., & Frías M. D. (2006). Errores de interpretación de los métodos estadísticos: importancia y recomendaciones. *Psicothema, 18*(4), 848 -856.

Sánchez-Bruno, A. & Borges, Á. (2005). Transformación Z de Fisher para la determinación de intervalos de confianza del coeficiente de correlación de Pearson. *Psicothema, 17*(1), 148-153.

Tárraga-López, P. J. & Rodríguez-Montes, J. A. (2016). ¿Se deben publicar los resultados negativos o no positivos? *Journal of Negative & No Positive Results, 1*(2), 43-44.

Trafimow, D. & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(7487), 1-2.

Woolston, C. (2015). Psychology journal bans P values. *Nature, 519*(7541), 9.

* sdominguezmpcs@gmail.com
sdominguezl@usmp.pe

ISSN: 1817-0285 (Impresa)
ISSN: 2224-3585 (Digital)