Convivencia del nivel de significación y tamaño del efecto y otros retos de la práctica basada en la evidencia

Francisco Rivera de los Santos.
Profesor de Psicología. Universidad de Huelva.

Septiembre 2017. Boletín Psicoevidencias nº 48. ISSN 2254-4046.

"La práctica clínica de un profesional de Psicología debe basarse en evidencias empíricas obtenidas por el método científico" (*American Psychological Association*, APA). Este mantra se repite de forma recurrente en los pasillos y las aulas de la universidad, en las consultas y reuniones de profesionales sanitarios o en congresos y jornadas científicas, entre otros. Sin embargo, la repetición de esta afirmación no garantiza su generalización en todos los puestos de decisión clínica.

Bajo esa premisa científica se construye precisamente la Práctica Basada en la Evidencia en Psicología (conocida como *Evidence-Based Practice in Psychology*, EBPP, en la APA). Este paradigma no es más que la integración o combinación de la mejor investigación disponible con la experiencia clínica del/a profesional en relación con las características, la cultura y las preferencias del paciente[1].

No obstante, todo esto requiere que los/as profesionales de la Psicología y de la salud mental sean capaces de generar evidencias científicas basadas en los **criterios de objetividad, sistematicidad y replicabilidad** en cuestiones psicológicas donde sea necesario profundizar o aclarar conocimiento. Y, complementariamente, se hace necesario que los/as profesionales sean capaces de "saber leer" dichas evidencias en una situación de toma de decisiones, además de disponer del tiempo necesario para ello.

El disponer de tiempo para buscar y leer evidencias empíricas y ofrecer la formación necesaria para capacitar en materia de investigación son, principalmente, variables que se encuentran dentro de las competencias de las administraciones implicadas, por ejemplo, el sistema sanitario. Es decir, se necesita que estas administraciones apuesten claramente por dichos preceptos y fomenten la cultura científica mediante la dotación de recursos.

De hecho, la capacidad de generar evidencia científica no es un camino fácil, ya que requiere formación y experiencia en materia de investigación (tanto en aspectos metodológicos como estadísticos). Ya que es solo a través de este conocimiento como los/as profesionales de salud mental podrán abordar progresivamente objetivos de investigación cada vez más complejos y, por tanto, más cercanos a la realidad de su práctica profesional.

En relación con aspectos metodológicos, la variedad de enfoques con amplia tradición en Psicología[2], junto con la incorporación de diseños con mayor presencia en el ámbito biomédico, como los ensayos clínicos[3], ofrecen garantías de adaptabilidad para los distintos enfoques y objetivos dentro de la investigación.

Pero incluso aplicando correctamente los diferentes diseños metodológicos y las técnicas estadísticas, aún son muchos los retos a los que se enfrenta la investigación actual en salud mental en lo referido al manejo y desarrollo de evidencias.

Retos

En la vorágine actual de publicaciones, en la que solo en 2016 se han publicado más de 135.000 artículos únicamente en las revistas incluidas en el JCR adscritas a materias referidas a Psicología (lo que supone algo más de 4.500 días seguidos de lectura solo para lo publicado en 2016), los/as profesionales de este ámbito de estudio se enfrentan a cuatro asignaturas pendientes:

1. la generación de metanálisis y revisiones sistemáticas, que faciliten la integración de resultados de numerosos artículos[4],
2. la replicabilidad de los resultados obtenidos para ofrecer evidencias empíricas que afiancen los resultados obtenidos en investigaciones previas[5], donde se incluiría igualmente el acceso de los datos para su re-análisis posterior por otros investigadores,
3. la publicación de resultados negativos, ya que aún queda implícito un cierto sesgo de publicación hacia resultados positivos, reminiscencia de una tradición en la que era incluso un criterio explícito (por ejemplo, en la *Journal of Experimental Psychology* se indicaban en 1962 que los manuscritos que no rechazaran la hipótesis nula nunca serían publicados[6]), y
4. la necesidad de complementar las pruebas de significación estadística con las pruebas del tamaño del efecto, lo que afianzaría las conclusiones de los diferentes artículos y facilitaría la comparabilidad entre publicaciones.

En relación con este último punto sobre aspectos estadísticos, los avances informáticos han hecho asequibles un gran número de análisis estadísticos, cada vez más versátiles y ambiciosos, que han mejorado la investigación. Pero de esa variedad parte también la **complejidad en la selección del análisis estadístico más adecuado** para el objetivo propuesto y el conocimiento de los supuestos necesarios para su aplicación.

Por tanto, una metodología adecuada y un análisis de datos correcto constituyen las bases para la generación de evidencias empíricas que contribuyan al avance de Psicología. No obstante, la simple aplicación de la metodología más adecuada y un desarrollo estadístico acorde a los objetivos suponen requisitos indispensables, pero no suficientes para garantizar la calidad de la evidencia científica.

Nivel de significación (*p-value*)

En este sentido, destaca el nivel de significación (*p-value*) como uno de los elementos clave en estadística, ya que supone tanto un eje central del proceso como una fuente de conflictos y debates. El *p-value* vinculado a una prueba estadística podría definirse como la probabilidad, suponiendo como cierta la hipótesis nula, de obtener un resultado al menos tan extremo como el que realmente se ha obtenido[7]. No obstante, un proceso tan aparentemente impoluto no está carente de limitaciones, procedentes de una incorrecta aplicación, mal uso o limitada interpretación[8].

Por ejemplo, su dependencia del tamaño muestral, del nivel de confianza establecido y del tamaño del efecto en el contraste, hace vulnerable la prueba de significación a la presencia de errores fundamentales que pueden alterar el resultado de una investigación. Incluso la propia regla de aceptación o rechazo (decisión dicotómica tradicionalmente asociada al valor 0.05 –correspondiente a un nivel de confianza del 95%-), no ha disuadido a multitud de investigadores/as a concluir que un contraste de variables con una *p* igual a 0.049 sí presente relación, mientras que otro contraste con *p* igual a 0.051, no.

Sin embargo, estas limitaciones evidentes que muestra la prueba de significación basada en el *p-value* no han conllevado una disminución de su uso ni un aumento claro en el empleo de otras medidas de asociación que, complementando las pruebas de significación, minimicen las consecuencias de la decisión dicotómica que implica. Sí es cierto que las posturas enfrentadas entre partidarios y detractares de las pruebas de significación han ayudado a plantearse la necesidad del *desidealizar* a las pruebas de significación para ser complementadas, por ejemplo, con las pruebas del tamaño del efecto[9].

Tamaño del efecto (*effect size*)

Las pruebas del tamaño del efecto son un conjunto de técnicas que fueron sistematizadas y difundidas hace ya varias décadas[10,11], y cuya aplicación ya era recomendada por los manuales de la APA en su cuarta edición[12,13]. En concreto, el tamaño del efecto es un índice, usualmente estandarizado, que indica la magnitud de una relación o efecto[11]. Es decir, se trata de un índice que permite analizar y cuantificar la intensidad de la relación entre dos variables o la diferencia entre dos grupos, todo ello de forma independiente al tamaño de la muestra.

Existen numerosas herramientas online que facilitan el acceso al cálculo de las pruebas de tamaño del efecto (destaca por su diversidad: **www.psychometrica.de/ effect_size.html**) y los parámetros necesarios para el cálculo son relativamente asequibles de obtener (medias, desviaciones tipos, porcentajes, recuentos…). A su vez, la interpretación de este índice es sencilla e intuitiva y facilita la comparabilidad entre investigaciones, por lo que su uso se está generalizando de forma progresiva.

Aunque en sus comienzos el propio Jacon Cohen en 1988[11] señaló la no conveniencia de poner niveles de interpretación al tamaño del efecto (es decir, bajo, medio, alto…), en la actualidad el uso de puntos de corte se ha generalizado para facilitar su comprensión. La cautela de Cohen se debía a que esa concepción podía ser heredada de la necesidad de dicotomización, procedente de la interpretación del *p-value,* junto con la idea de que en cada ámbito y disciplina de estudio un mismo valor podría categorizarse con distinto nivel (relacionado con el nivel de causalidad propio de cada área).

Entre las principales pruebas del tamaño del efecto destacan:

- ***d de Cohen***, aplicado en la comparación de medias entre grupos, que establece los puntos de corte en 0.00-0.19 despreciable, 0.20-0.49 bajo, 0.50-0.79 medio y a partir de 0.80 alto. Este índice es de los más usados, existiendo actualmente generalizaciones para cuando se cumplan las condiciones de aplicación de las pruebas paramétricas o cuando sea necesario recurrir a pruebas no paramétricas. Igualmente se han añadido aplicaciones para diseños experimentales con grupo control y experimental y medida pre y post intervención. Igualmente existen diversas formulaciones que permiten la traducción de cualquier otro tamaño del efecto a los valores propios de la *d* de Cohen, lo que convierte a este tamaño del efecto en la prueba más utilizada en estudios de metanálisis.

- ***Odds Ratio* (*OR*) y *Risk Ratio* (*RR*),** tradicionalmente aplicado para estimar la probabilidad de que se desarrolle un determinado evento (enfermedad, habitualmente) en los sujetos expuestos a un factor de riesgo con respecto al grupo de los sujetos no expuestos (OR

aplicados en diseños de caso y control y RR en estudio de cohortes). Aunque el cálculo del OR y RR difieren (modificándose el número y denominador en cada caso), se establece que cuando el intervalo de confianza del *RR* u *OR* incluye el 1 se concluye que existe la misma probabilidad en ambos grupos. Si no fuera así, se marcan en el caso del OR como un nivel bajo cuando el valor puntual del OR abarca de 1.5 a 1.99, medio de 2.00 a 2.99 y grande cuando supera el 3.00. En RR se adopta como bajo los niveles comprendidos entre 2 y 2.99, medio de 3.00 a 3.99 y grande cuando supera el 4.00.

- ***Phi* y *V* de Cramer**, se aplica en contrastes de proporciones cuando se trata de un contraste entre dos variables con dos categorías cada una de ellas (*Phi*) o cuando alguna presente más de dos categorías (*V* de Cramer). Establece como puntos de corte los valores 0.00-0.09 como despreciable, 0.10-0.29 como bajo, 0.30-0.49 como medio y a partir de 0.50 como alto.

- ***Eta-squared* (η^2) y Coeficiente de determinación (R^2)** para anova y regresión lineal múltiple. Estos índices establecen los siguientes puntos de corte: despreciable de 0.000-0.019, bajo de 0.020-0.129, medio de 0.130-0.259 y alto a partir de 0.260.

Estas son solo algunas de herramientas estadísticas utilizadas para el cálculo del tamaño del efecto, cuyo uso es sencillo, asequible y necesario. No solo complementa al *p-value* (el *p-value* puede informar acerca de si existe o no un efecto o relación, pero no revelará el tamaño del efecto de dicha relación), sino que reduce la posibilidad del falso positivo (error tipo I o *alpha*) y del falso negativo (error tipo II o *beta*) de la prueba de significación y permite la comparabilidad entre investigaciones[14].

Así pues, conocemos el camino (metodología científica), disponemos de herramientas adecuadas (técnicas estadísticas, entre otras, en las que cobra especial relevancia las pruebas de tamaño del efecto) y existen multitud de objetivos de investigación que necesitan ser abordados.

Aún hay dificultades y retos por superar, pero en la mayoría de profesionales ha calado ya el paradigma de la práctica basada en la evidencia, así que para una disciplina tan relativamente joven como es la Psicología, solo queda seguir avanzando en el desarrollo de evidencias empíricas para contribuir a una toma de decisiones consciente y fundamentada en la evidencia científica.

Referencias

[1] Anderson, N. B. (2006). Evidence-based practice in psychology. American Psychologist, 61(4), 271-285.

[2] Montero, I. & León, O. G. (2007). A guide for naming research studies in Psychology. International Journal of Clinical and Health Psychology, 7(3), 847-862.

[3] Argimón Pallás, J. M. & Jiménez Villa, J. (2013). Métodos de investigación clínica y epidemiológica. 4ª Edición. Elsevier. Barcelona.

[4] Meca, J. S. & Ausina, J. B. (2010). Revisiones sistemáticas y meta-análisis: Herramientas para la práctica profesional. Papeles del psicólogo, 31(1), 7-17.

[5] Asendorpf, J. B., Conner, M., De Fruyt, F. et al. (2013). Recommendations for increasing replicability in psychology. European Journal of Personality, 27(2), 108-119.

[6] Melton, A. W. Editorial Journal of Experimental Psychology, 64(6), Dec 1962, 553-557.

[7] Fisher, R. A. (1925). Statistical methods for research workers. Genesis Publishing Pvt Ltd.

[8] Ronald L. Wasserstein & Nicole A. Lazar (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108.

[9] Frías Navarro, M. D., Pascual Llobell, J. & García Pérez, J. F. (2000). Tamaño del efecto del tratamiento y significación estadística. Psicothema, 12(2), 236-240.

[10] Cohen, J. (1969). Statistical power analysis for the behavioral sciences. New York: Academic Press.

[11] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd edition). Hillsdale, NJ: Lawrence Earlbaum Associates.

[12] Wilkinson, Leland. APA Task Force on Statistical Inference (1999). «Statistical methods in psychology journals: Guidelines and explanations». American Psychologist 54 (8): 594-604. doi:10.1037/0003-066X.54.8.594.

[13] American Psychological Association (A.P.A.) (1994). Publications manual of the American Psychological Association (4th edition). Washington, DC.

[14] Sullivan, G. M. & Feinn, R. (2012). Using effect size—or why the p value is not enough. Journal of graduate medical education, 4(3), 279-282.