Análisis de poder como justificación de la muestra: estrategia analítica para la búsqueda del verdadero efecto

Jesua Guzmán González[1], Franco Sánchez García[1], Luis Miguel Sánchez Loyo[2] y Saúl Ramírez de los Santos[3]

[1]*Doctorado en Biociencias, Centro Universitario de los Altos, Universidad de Guadalajara,* [2]*Departamento de Estudios en Lenguas Indígenas, Centro Universitario de Ciencias de la Sociales y Humanidades, Universidad de Guadalajara, y* [3]*Departamento de Psicología Básica, Centro Universitario de Ciencias de la Salud, Universidad de Guadalajara*

Cómo citar este artículo (estilo APA) / Citing this article (APA style):
Guzmán-González, J., Sánchez García, F., Sánchez Loyo, L., & Ramírez de los Santos, S. (2023). Análisis de poder como justificación de la muestra: estrategia analítica para la búsqueda del verdadero efecto. *Revista Mexicana De Investigación En Psicología, 15*(1), 13-22.

Resumen

El *Manual de publicación de la Asociación Americana de Psicología*, en su séptima edición (2020), destaca la importancia de privilegiar la calidad en la publicación de artículos científicos. Uno de sus capítulos se relaciona con los estándares de reporte de resultados para artículos de revista, que se centra en la precisión de los resultados, con especial atención en los aspectos cuantitativos, para aumentar la capacidad de identificar efectos reales en la investigación mediante el análisis de poder (AP), que, en este trabajo, se presenta como una herramienta esencial en el diseño y análisis de estudios científicos; también, se sugiere como una estrategia analítica para calcular el tamaño de la muestra. Se explica que el AP se deriva de la propuesta estadística de Neyman-Pearson para la comprobación de hipótesis. Se proporcionan definiciones clave, como el poder estadístico (β), el valor alfa crítico (α), el tamaño del efecto (δ) y el tamaño de la muestra (n). Mediante el enfoque del AP, se puede coadyuvar a mejorar la calidad y transparencia en la investigación científica, así como la capacidad de reconocer efectos reales en la investigación. Se subraya la relevancia de la estadística en la construcción del conocimiento y la necesidad de estrategias para incrementar la confiabilidad de la producción científica.

Palabras clave: tamaño de la muestra, análisis de poder, tamaño del efecto, estadística de robustez, estrategia analítica.

Power analysis as justification of the sample: Analytical strategy for searching for the true effect

Abstract

The *Publication manual of the American Psychological Association* in its seventh edition, published in 2020, highlights the importance of seeking quality in the publication of scientific articles. One of the highlighted chapters is the Journal Article Results Reporting Standards, which focuses on the precision of the results, with emphasis on quantitative aspects to increase the ability to detect real effects in research through power analysis (PA). The present work presents as an important tool in the design and analysis of scientific studies, suggesting it as an analytical strategy to calculate the sample size. It is explained that the PA is derived from the Neyman-Pearson statistical proposal for hypothesis testing. Key definitions are provided, such as statistical power (β), critical alpha value (α), effect size (δ), and sample size (n). It can help improve the quality and transparency in scientific research, through the approach of power analysis as an essential tool to improve the ability to detect real effects in research. The importance of statistics in the construction of knowledge is highlighted and the need for strategies to improve the reliability of scientific production is recognized.

Dirigir toda correspondencia al autor a la siguiente dirección:
Jesua Guzmán-González
Psic.jesua-guzman@outlook.com
RMIP 2023, Vol. 15, Núm. 1, pp. 13-20.
www.revistamexicanadeinvestigacionenpsicologia.udg.mx

Keywords: sample size, power analysis, effect size, robust statistics, analytical strategy

INTRODUCCIÓN

El *Manual de publicación de la Asociación Psicológica Americana,* en su séptima versión (2020) presenta mayor atención a aspectos de la publicación como la diversidad, inclusividad y elementos que consideran importantes para educar, crear, persuadir o expandir un conocimiento específico (Bradley et al., 2020). Dentro de este último elemento, se encuentra el capítulo denominado estándares de reporte de resultados para artículos de revista o, por sus siglas en inglés, JARS (Journal Article Reporting Standards). El capítulo se centra en una serie de recomendaciones que permiten mejorar la calidad del artículo en términos de precisión, claridad y transparencia de los resultados. Este capítulo, respecto de los estándares cuantitativos, dedica un apartado al tamaño de la muestra, el poder estadístico y la precisión de los resultados reportados a fin de aumentar el poder estadístico del artículo científico conocido como análisis de poder (AP).

El AP es definido como una medida que se utiliza en el diseño y análisis de estudios científicos para evaluar la capacidad, alcance y fuerza de una estrategia analítica de detectar un verdadero efecto o una relación real entre variables (Aberson, 2019), lo cual resulta con mayor poder en comparación con el uso único del abordaje de Neyman-Pearson del test de la significancia de la hipótesis nula (Greenland et al., 2016). En otras palabras, es la probabilidad de que una estrategia analítica, por lo regular mediante pruebas estadísticas, pueda identificar la existencia de una diferencia o efecto significativo en el fenómeno a estudiar.

Las recomendaciones del citado manual se han señalado por décadas con poco éxito (Button et al., 2013; Cohen, 1988) debido a que, en vez de profundizar en métodos compatibles y de robustez con muestras pequeñas que mantengan en niveles adecuados el resultado del AP (Smith y Little, 2018), con frecuencia se decide aumentar el número de la muestra (*n*) sin considerar que eso no necesariamente es la solución que mejor se ajusta al fenómeno. De hecho, una muestra mayor no siempre permite encontrar el efecto deseado por algunas simples razones. La primera es que es probable que un mayor tamaño del efecto (TE) se asocie a muestras más pequeñas y, a la inversa, se podrá encontrar un TE pequeño en muestras más grandes. Eso no depende *per se* del tamaño de la muestra, sino del TE deseado que haya sido reportado antes en la literatura especializada. En general, el AP ha demostrado ser la forma más efectiva durante el proceso de interpretación, recolección y búsqueda de hallazgos relevantes para la comprobación de hipótesis, incluso más efectiva que aumentar la *n* de la muestra en el estudio (Bartlett y Charles, 2022).

ANÁLISIS DE PODER EN LAS HIPÓTESIS: LOS PASOS HACIA LA VERDAD

Los seres humanos tienen la capacidad de desarrollar inferencias y generar nociones del mundo gracias al desarrollo cognitivo (Colombo y Seriès, 2012). Estas nociones de la realidad son parte integral de las hipótesis dada su naturaleza especulativa y tendencia a la generalización, y varían en grado de ajuste y alcance del fenómeno observado. Conocemos alrededor de 45 tipos de hipótesis en el método científico actual (Ortiz-Uribe, 2003); sin embargo, esta capacidad no es libre de errores porque es probable que se cometan diversas falacias (Bordes-Solanas, 2011). Es importante aclarar que los errores, en cuanto a las falacias se trata, no se refieren necesariamente a la equivocación, sino a la imprecisión, en otras palabras, a un sesgo. Gutiérrez-Cabria (1994) argumenta que la estadística es una herramienta que puede prevenir la aparición de estos sesgos, y fungir como una línea de defensa ante las falacias. Además, es una herramienta indispensable para la construcción del conocimiento, razón por la cual conocer de manera adecuada la estrategia analítica permite saber, primero, que las nociones creadas nunca son, en esencia, erróneas o ciertas, sino que deben entenderse bajo un grado de ajuste a la realidad percibida, y segundo, que la estadística es solo una herramienta útil para describir y contrastar esa descripción del mundo que el ser humano percibe como real.

Mediante una integración de lo anterior, podemos partir de que el ser humano se encuentra formulando hipótesis de su entorno en forma constante más o menos precisas, ante lo cual la estadística presenta un rol importante cuidando que las inferencias sean lo menos imprecisas posible. Lo anterior es generalizable a todas las ciencias, incluidas las consideradas "blandas", como las ciencias humanas (Ortiz-Uribe, 2003). Véase el ejemplo de las ciencias nomotécnicas (como la sociología) en comparación con las ciencias naturales (como la física), en las que, a pesar de que la primera es de contenido metafísico y la segunda de contenido físico, ambas son susceptibles a los sesgos en la percepción. Es necesario aclarar en este punto que no existe un método correcto o incorrecto para entender un fenómeno, sino que, al igual que las hipótesis, habrá métodos que tengan un grado de ajuste mayor o menor para el fenómeno observado. Aunado a esto, es una realidad que, en la actualidad, las ciencias consideradas como "blandas" se encuentran en una crisis de confianza dada la baja posibilidad de replicación de

hallazgos (Ioannidis, 2005), por lo que se han desarrollado múltiples estrategias para mejorar la construcción del conocimiento y la producción científica.

Se sabe que no resulta conveniente llevar a cabo una evaluación de una muestra determinada sin haber hecho un AP. Debido a lo antes descrito y a que las funciones de la potencia estadística no son lineales y son específicas de la estrategia estadística del estudio (Cohen, 1988), el AP se deriva del poder estadístico propuesto por Neyman-Pearson para comprobación de la hipótesis (Neyman y Pearson, 1993). Entonces, antes de hablar de este, definiremos algunos de sus elementos constituyentes, es decir, cuáles son los tipos de error, entender la distribución muestral (central y no central) y conocer las variables definiendo el significado del poder estadístico (β), el valor alfa crítico (α), el TE (δ) y la *n* (véase tabla 1), ya que, si especificamos tres de estas, podemos encontrar la cuarta (Aberson, 2019). La tabla 1 contiene el resumen de estas variables y las definiciones clásicas de estos elementos.

Tabla 1
Resumen de conceptos en estadística constituyentes del AP y su uso común

Símbolo	Definición	Uso
Alfa (α)	Área bajo la curva que representa la zona crítica necesaria de superar, que se interpreta como la probabilidad de aceptar un falso positivo	Tradicionalmente, este se configura al 0.05, lo que se traduce a tener la probabilidad de un 5% de aceptar el error de observar un fenómeno que no está ahí. Existen otras disciplinas más exigentes, como los ensayos clínicos en los que se exige un 0.001
Beta (β)	Área bajo la curva que representa la zona crítica necesaria a superar, que se interpreta como la probabilidad de aceptar un falso negativo	Tradicionalmente, este se configura al 0.80, lo que se traduce a tener la probabilidad de un 20% de aceptar el error de no observar un fenómeno que está ahí. No existe un consenso sólido que indique cuál sería el β deseado, pero se considera que este valor es el mínimo aceptable
Delta (δ)	Representa la magnitud de distancia entre dos medidas obtenidas de grupos independientes	Este cálculo depende casi exclusivamente del fenómeno estudiado; sin embargo, aunque se conocen algunos valores empíricos que ayudan a un punto de partida para las ciencias de la conducta (δ = 0.23 como efecto pequeño, δ = 0.60 como efecto moderado y δ = 0.78 como efecto amplio), se recomienda consultar los metaanálisis del fenómeno específico para averiguar su efecto particular
Muestra (n)	Bajo la perspectiva del AP, este es el número necesario de observaciones para apreciar un verdadero efecto	El número de observaciones mínimas necesarias para evidenciar un verdadero efecto entre dos grupos

Como mencionamos, la estadística es una herramienta esencial para evitar caer en errores interpretativos y funciona como una línea de defensa para prevenir imprecisiones en la comunidad científica con hallazgos y, también, para no engañarnos a nosotros mismos con sesgos. La teoría estadística predominante, llamada frecuentista, es un marco de trabajo matemático basado en la distribución normal, que supone que la probabilidad de aparición de un fenómeno se acumula con el tiempo. Dada esta aproximación, se considera que pueden existir fundamentalmente dos tipos de sesgos: afirmar que existe un fenómeno cuando no lo hay (error tipo I), también denominado falso positivo, y negar que existe un fenómeno cuando en realidad lo hay (error tipo II), conocido como falso negativo.

La estrategia más común es mediante el valor α, también conocido como valor p (*p value*). Este es un procedimiento estadístico que permite evitar el error tipo I en fenómenos susceptibles de ser medidos (Pepper, 1972). Clásicamente, es la primera línea de defensa contra ses-

gos. La mayoría de los software estadísticos lo calculan por programación estándar; sin embargo, dicho estadístico se ha visto limitado en potencia para brindar una adecuada conclusión (Solla et al., 2018) ante la ausencia de conocimiento estadístico previo. El α, bajo el supuesto de interpretación del marco fisheriano, es interpretado como una medida continua de compatibilidad entre los datos observados que sirven para rechazar la H^0, de tal forma la compatibilidad de los datos con la H^0 puede establecerse con el continuo entre 1 (perfectamente compatible) y 0 (perfectamente incompatible) (Wasserstein et al., 2019).

Además, el α parte del teorema del límite central, el cual asume que todo fenómeno observado tenderá a centralizarse de tal manera que la "normalidad" se observará en la frecuencia de aparición de un fenómeno en el universo observable (μ); por lo tanto, las muestras provenientes de dicho universo (M) pueden acercarse o alejarse del μ en dependencia de variables que convergen en el mismo fenómeno. Estas muestras pueden utilizarse para compararse con la μ, comparar dos muestras M^1 y M^2, o bien, asociar algún fenómeno de naturaleza medible, y luego se someterá a prueba el grado de ajuste en el que se diferencia o se acerca a lo esperado. Estas pruebas de hipótesis, en general provenientes de técnicas estadísticas, se centran en la fuente de variabilidad dentro de la clásica distribución normal.

Por otro lado, el poder estadístico es definido como 1- β (Cohen, 1988, 1992), donde la β es la probabilidad de error permisible para un falso negativo (tipo II), es decir, la probabilidad de no detectar un efecto como significativo si la hipótesis neutra es falsa y se cumplen los supuestos de la prueba de significancia. Esto significa que, si existe un poder más alto en el estudio, si un verdadero efecto está presente, se detecta con mayor frecuencia. Cohen (1988) menciona que, para estudios en psicología, el nivel recomendado de la β es de 0.80, y surge de la afirmación de Neyman y Pearson (1933) que los falsos positivos son cuatro veces peores para la ciencia que los falsos negativos. Por lo tanto, si los "falsos positivos" (α) se mantienen al (0.05) 5%, los falsos negativos (β) no deberían ser mayores del 20%. Cabe mencionar que esta afirmación es arbitraria y sigue en estudio una mejor decisión al respecto.

Ambas medidas se encuentran presentes dentro de la distribución muestral y, clásicamente, se describe a la H^0 como la probabilidad de distribución centrada en cero. Se puede rechazar la H^0 si los resultados observados son mayores que el valor crítico observado en el α. Esta crea una zona de rechazo en una de las colas de distribución; si los resultados observados se encuentran dentro de la zona de rechazo, se concluye que los datos serían improbables si se supone que la hipótesis alterna fuese verdadera, y si se rechaza la hipótesis nula. Con esta posibilidad, entonces existen dos maneras de declarar la zona de rechazo: hipótesis no direccional y direccional. La primera corresponde a la suposición de que existirá un efecto independientemente de la dirección del efecto, es decir, en cualquiera de las dos colas extremas de la distribución muestral, por ejemplo, diferencias de medias entre grupos ($M^1 \neq M^2$), lo que quiere decir que la H^0 se interpretaría como que no existirán diferencias entre grupos. La segunda corresponde a un fenómeno en el que existe una dirección o un sesgo de la diferencia, por ejemplo, que el primer grupo es mayor que el segundo ($M^1 > M^2$) o viceversa ($M^1 < M^2$). En este último caso, conocido como prueba de una sola cola, el área de rechazo se posicionará en dependencia de la dirección de interés, por ejemplo, si existe un efecto techo en la distribución de los datos en el que no se puede obtener más allá del μ y se busca objetivar que el grupo es mayor o menor del presentado. Se puede encontrar que el primer grupo es más grande que el segundo, pero no importa cuán grande sea la diferencia, no se puede rechazar la hipótesis nula porque es contraria a la predicción direccional.

Respecto a los TE, estos son una familia de cálculos que han cobrado relevancia conforme la investigación estadística aumenta. Cohen (1992) menciona el siguiente supuesto:

"El principal producto de la indagación en la investigación es una o más medidas del tamaño del efecto, no los valores p" (Cohen, 1992, p. 1310).

Estimar el TE permite cuantificar la magnitud de un fenómeno observado en cualquiera de sus variantes: diferencia de medias (grupos), varianza compartida (correlaciones), área bajo la curva (Salgado, 2018) y la proporción de aparición de un fenómeno en contraste con otro (Rouder et al., 2009). La versión más conocida de cálculo del TE es la δ, también conocida como d de Cohen, la cual es una medida cuantitativa que permite entender el grado en el que un fenómeno (categoría, agrupación, característica, condición, etcétera) distancia la distribución muestral entre los grupos. En la propuesta original, Cohen (2013) sugirió que la δ no tiene una interpretación estandarizada, contrario a entenderse como el valor p, lo cual implicaría una serie de errores metodológicos y estadísticos (Cohen, 1988; Gurnsey, 2017), por lo que, actualmente, se considera que debe complementarse con la finalidad de mejorar la precisión de la interpretación (Baguley, 2009).

Existen diferentes fórmulas para estimar los TE: para muestras con homogeneidad de varianzas (Cohen, 2013), sin homogeneidad de varianzas o comparaciones

pareadas (Glass et al., 1981) y de una sola muestra en la que el tamaño de ambos grupos es desproporcional (Hedges, 1981).

Cohen (2013)

$$\delta = \frac{M^1 - M^2}{\sigma}$$

Glass (1981)

$$\Delta = \frac{M^1 - M^2}{\sigma^{Control}}$$

Hedges (1981)

$$g = \frac{M^1 - M^2}{\sigma^*}$$

Nota: En la δ de Cohen debe obtenerse la σ estándar en común, o bien, sumar la σ de ambos grupos y dividirla entre dos. En la Δ de Glass se considera que, si las σ difieren y se viola la homogeneidad de las varianzas, no es apropiado utilizar el método de la δ, por lo que se puede tabular la σ del grupo control; esta configuración obtiene mayor proporcional al tamaño del grupo control. Por último, en la g de Hedges lo recomendado es ponderar* la σ de cada grupo por su tamaño de muestra.
M = media

Cabe aclarar que la δ de Cohen (1988) fue diseñada para datos con distribución paramétrica, aunque existen alternativas viables para muestras no paramétricas debido a que se conocen 70 variantes de medidas del TE (Kirk, 2003). Al respecto, la correlación de rangos biseriales (r_{rb}) (Cureton, 1956), modificada por Wendt (1972) y basada en la fórmula de la U de Mann Whitney, es la que se conoce con mejores propiedades para las distribuciones no paramétricas (Guzmán-González et al., 2023). Esta resulta de la modificación de la aproximación biserial cuyo centro del cálculo es el coeficiente de correlación entre rangos (Y y Z). Esta correlación es adecuada para la distribución no paramétrica porque utiliza una relación monotónica entre Y y Z similar a la de la correlación de la Tau-b de Kendall o la ρ de Spearman, que siguen la aproximación matemática de la *U* de Mann-Whitney (Kerby, 2014) basada en rangos. La fórmula es la siguiente:

$$r_{rb} = 1 - (2U)/ (n_1 * n_2)$$

Aunque se sabe poco acerca de sus propiedades estadísticas, se han hecho cálculos de validez (Chmura-Kraemer, 2014) y concordancia entre la r_{rb} y la δ de Cohen clásica que demuestran que son válidas para su uso estadístico (Guzmán-González et al., 2023). La fórmula de conversión es la siguiente:

$$r= \delta/\sqrt{(\delta\text{^}2+4)}$$

Respecto a la pregunta sobre cuál magnitud del TE es adecuada, (Szucs & Ioannidis, 2017) encontraron δ empíricos reportados para estudios en el campo de la psicología donde la δ fue de 0.23 para un TE pequeño, δ = 0.60 para un TE moderado y δ = 0.78 para un TE amplio, mientras que para las neurociencias se encontró un δ = 0.14 para un TE pequeño, δ = 0.44 para un TE moderado y un δ = 0.67 para un TE amplio. Lo anterior cobra relevancia dado que, a pesar de que estos cálculos son procedimientos que se consideraban complementarios, tras las sugerencias de la Asociación Americana de Psicología (Wilkinson, 1999), se decidieron integrar al reporte de resultados a fin de mejorar las pruebas de hipótesis (Hand, 2012). Lamentablemente, solo el 10% de las revistas en Latinoamérica han integrado este cálculo (García et al., 2008), lo que pudiera contribuir en su menor impacto y visibilidad científica.

La evidencia científica aconseja calcularlos en el mismo estudio para mejorar la robustez de análisis, dado que se ha reportado una correlación importante entre el valor δ, el factor Bayes y el valor p, lo que permite disminuir la incertidumbre en la comprobación de los supuestos. Respecto a lo anterior, un estudio realizado por Wetzels et al. (2011) encontró que, en un total de 855 pruebas t, los valores p, los valores δ y el factor Bayes tenían una gran probabilidad de coincidir en el sustento de la hipótesis (véase figura 1). Esta potencia estadística permite que los investigadores no sobreestimen la evidencia a favor de un efecto solo con el valor p (Wetzels et al., 2011).

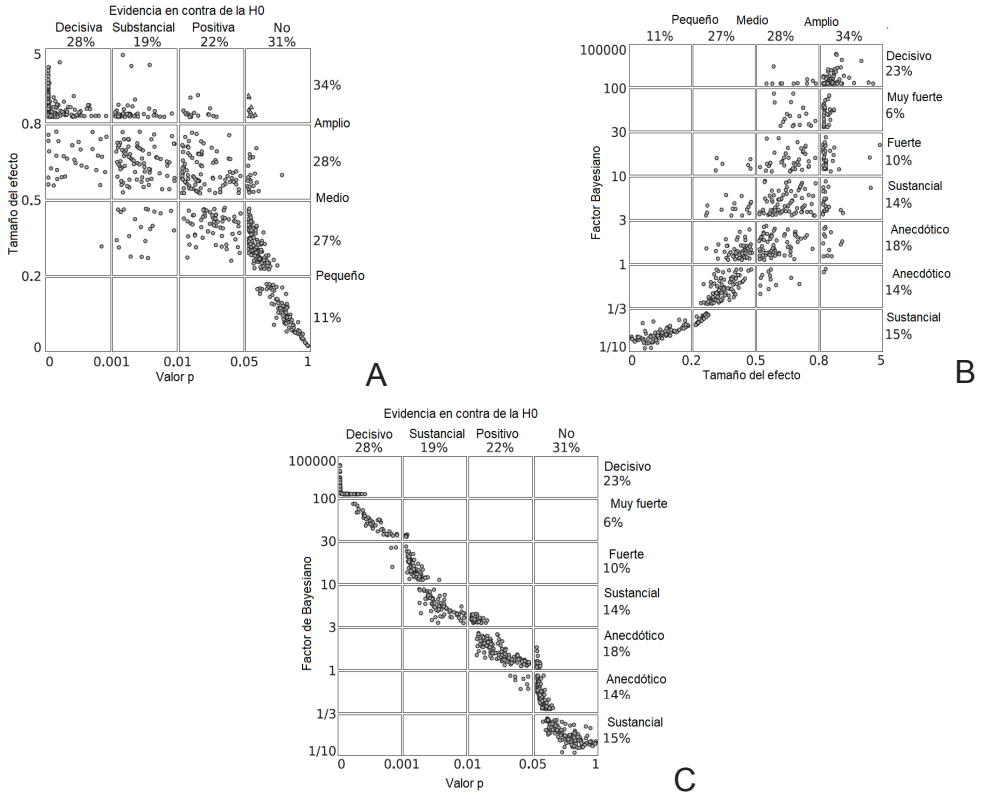


Figura 1. Relación entre el TE y los valores p (A), relación entre el TE y el factor bayesiano (B) y, por último, relación entre el valor p y el factor bayesiano (C). Diagrama de dispersión en el que los puntos indican una coherencia relativa entre los tres cálculos realizados a partir de la muestra dada (855), mientras que los indicados por triángulos señalan grandes inconsistencias, adaptado de Wetzels (Wetzels et al., 2011).

Lo anterior es tomado para nuestro trabajo porque se han identificado cuáles medidas son las que pudieran predecir un verdadero efecto. Aunque pudiera parecer contraintuitivo no necesariamente un valor p por debajo de su significancia, implica un verdadero fenómeno, ya que, por efectos de una distribución atípica, se puede obtener una significancia y sugerir que existe un fenómeno donde no lo hay. El verdadero efecto dependerá más de aquello que se desea comprobar que de una medida estandarizada. Una explicación de esto es la siguiente: se trata de un fenómeno tan amplio que es visible incluso sin pruebas estadísticas, por ejemplo, un grupo A con un promedio (M) de 65.42 y una desviación estándar (σ) de 9.8, y un grupo B con una M = 53.14 y σ = 9.8. Al hacer el cálculo sencillo del TE, la δ = 1.22 supone un efecto amplio, entonces la probabilidad de una significancia es mayor; es posible que en este estudio exista una significancia positiva (p => 0.05), y eso será independiente de la muestra; por lo tanto, aunque se evalúen, por ejemplo, 15 sujetos o 150, si el valor de δ es alto, se obtendrá una significancia dado que las M y σ tendrán una variabilidad similar, es decir, las M serán similares si el fenómeno es estable.

Otro ejemplo es que, si la μ del grupo A es bien conocida al igual que la μ del grupo B, conforme la muestra aumente solo se irá reduciendo la σ si el M es general. De manera inversa, si se ha encontrado un valor de δ pequeño, ocurre un fenómeno ligeramente diferente, ya que, para encontrar efectos de tales características que tengan significancia, se requiere una muestra mayor, que es lo que realizan la mayoría de los investigadores. Un mayor efecto necesita menos muestra, mientras que un menor efecto demanda mayor muestra. Referente a lo anterior, es conveniente primero conocer cuál es el TE buscado en la literatura y tratar de replicarlo. Una forma sencilla de hacerlo es buscar en metaanálisis, que son la forma de evidencia científica de mayor jerarquía.

Además, a partir del TE se pueden llevar a cabo diferentes abordajes; entre ellos, el que pudiera ser de utilidad para estudios en conducta es el coeficiente de superposición (Ω), que se calcula multiplicando dos veces la función de la distribución acumulativa para la distribución

normal por el negativo de la mitad de la δ (Al-Saleh & Samawi, 2007):

$$\Omega = 2\phi(\frac{-|\delta|}{2})$$

La interpretación que pudiera enriquecer la explicación del fenómeno es la siguiente: por ejemplo, si se aplicara un cuestionario a dos grupos independientes (casos y controles) y se obtuviera una δ = 1, como resultado se obtendría un Ω = .6170; al multiplicar por 100 el resultado, se puede interpretar que, alrededor de un 62% de los grupos, tienen puntuaciones que coinciden (están superpuestas) o se interceptan entre sí; por lo tanto, eso significa también que cerca de un 38% de las puntuaciones son diferentes entre ambos grupos. La riqueza proviene de explicar este grado de compatibilidad entre grupos, en especial para desempeños en tareas o cuestionarios. Si se busca profundizar en estos cálculos, se puede consultar a Ventura-León (2018)los investigadores reportan que la diferencia entre dos distribuciones es pequeña (d > .20, quien reporta los códigos en R para realizar el cálculo.

JUSTIFICACIÓN DE LA MUESTRA

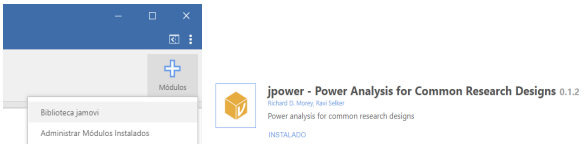
Desde el reconocimiento de la crisis de confiabilidad (Hartgerink et al., 2017; Ioannidis, 2005), se han hecho esfuerzos para mejorar los métodos usuales que determinan el tamaño de la muestra (Simmons et al., 2011) logramos dos cosas. Primero, mostramos que a pesar del respaldo nominal de los psicólogos empíricos a una baja tasa de hallazgos falsos positivos (≤ .05. Existen al menos seis maneras de poder justificar el tamaño de la muestra (Lakens, 2022); sin embargo, la de mayor uso es la de representatividad. Desde esta perspectiva, se considera que la forma clásica para muestreos aleatorios simples, que cabe aclarar que existen diferentes aproximaciones en dependencia del tipo de muestreo, es mediante la siguiente ecuación:

$$n = (\frac{Z_{\alpha/2}}{e})^2$$

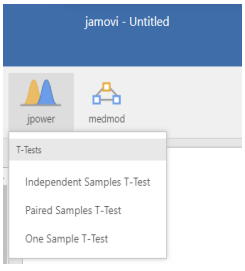
La anterior forma tiene un uso central en la epidemiología (Milton, 2001; Polit, 2002) debido a que busca una sensibilidad suficiente para poder efectuar una prueba *Z* diseñada para muestras mayores de 130 sujetos. No obstante, esta aproximación es poco sensible en estudios observacionales o diseños experimentales en los que la población a estudiar suele ser pequeña o se tiene acceso a pocos sujetos en los que se objetiva el fenómeno. Por ejemplo, con esta aproximación, mientras menor sea la población, mayor es el número de la muestra; inversamente, si se tiene mayor población, menor será la muestra para representarla, es decir, si se tienen 100 sujetos de población, la muestra debe ser de 80, pero si se tiene una población de 1,000, la muestra tendrá que ser de 278, o si son 10,000, será de 370. Conforme la población vaya aumentando, la muestra se estabilizará, de tal modo que si ahora se tiene una población de 100,000, se necesitarán tan solo 383 en la muestra, o si son 1'000,000, será de 385 fijo por más que se incremente la población. Por ende, es poco realista que algún investigador que estudie fenómenos poco típicos, enfermedades raras o que cuente con recursos limitados pueda utilizar esta fórmula para justificar su *n*.

Por ello, sugerimos utilizar el AP, que no es una propuesta novedosa, pero ha sido poco utilizada. Existen tres formas comunes de usarlo: el primero es denominado *a priori*; el segundo, calculando la sensibilidad; y el tercero es hacer un *post-hoc*. Con base en el objetivo de nuestro trabajo, solo profundizaremos en el primero, ya que es el que permite calcular la *n* en función de un δ de interés en particular para un fenómeno específico.

A continuación, mostramos una explicación gráfica a manera de guía utilizando el programa Jamovi 2.3.2 (2022) de libre acceso. Primero, tras abrir el programa, se selecciona el botón de módulos ubicado en la esquina superior derecha y se da clic en "Biblioteca Jamovi". Luego, se instala el módulo "jpower".

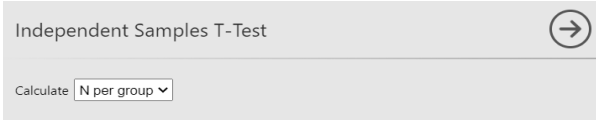


Segundo, ya instalado el módulo "japower", se seleccionará de la barra de módulos y se dará clic al tipo de muestra que se tiene (independientes, pareadas o una muestra).

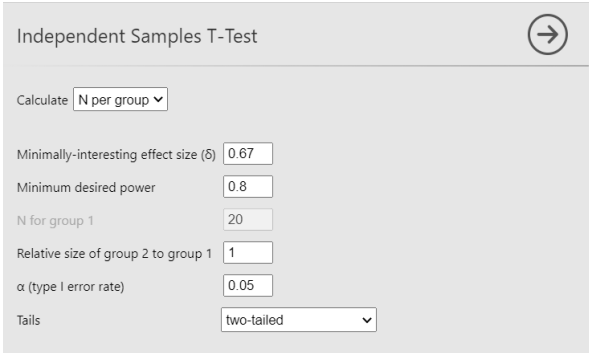


Tercero (se realizará el ejemplo para muestras independientes), del lado izquierdo de la pantalla, en la opción

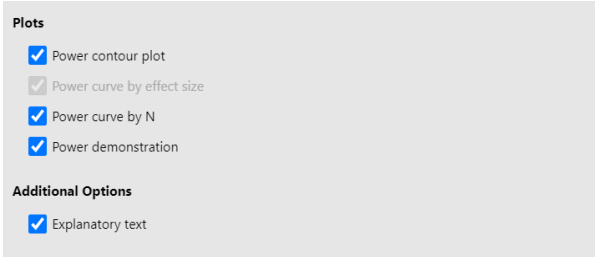
calcular se seleccionará "N por grupo". Para muestras pareadas o una muestra, se lleva a cabo el mismo procedimiento, pero se omite el dato de tamaño relativo del grupo 2 al grupo 1, lo que significa que, si el tamaño relativo es de 1, habrá 36 sujetos por grupo; si este valor se modifica por ejemplo a 1.5, ahora habrá 36 sujetos en un grupo y 45 en el otro. Lo anterior es buscando un δ de 0.67



Cuarto, se ingresarán los valores del TE mínimo de interés (δ), el poder mínimo deseado que puede consultarse en la literatura previa o metaanálisis, el tamaño relativo del grupo 2 al grupo 1, el α (rango de error tipo I) y las colas (a dos colas o a una cola). En el ejemplo se supondrán los valores deseados para el estudio: δ = 0.67 para un efecto amplio como es sugerido en estudios de neurociencias, un poder de 0.80, un tamaño relativo de 1, un α = 0.05 y a dos colas.



Quinto, se seleccionarán las gráficas en caso de ser deseadas, así como un texto de explicación si se necesitara.



Sexto, del lado derecho se podrá observar el cálculo de N realizado, así como las descripciones del poder por TE. En el ejemplo realizado se calcula que se necesitarán al menos 36 sujetos o muestras por grupo como previamente se describió.

A Priori Power Analysis

		User Defined		
N_1	N_2	Effect Size	Power	α
36	36	0.670	0.800	0.0500

Power by Effect Size

True effect size	Power to detect	Description
0 < d = 0.468	≤50%	Likely miss
0.468 < d = 0.670	50% – 80%	Good chance of missing
0.670 < d = 0.862	80% – 95%	Probably detect
d = 0.862	≥95%	Almost surely detect

El resultado de las tablas anteriores indica el poder para detectar el error tipo II, por lo que el TE seguro para identificar el verdadero efecto debe ser 0.67 o superior, cualquier resultado menor corre el riesgo debido a que tiene menor poder estadístico.

DISCUSIÓN Y CONCLUSIONES

Este trabajo busca principalmente servir como una guía para estudiantes, profesores e investigadores que requieren realizar el cálculo de su muestra para sus estudios de manera válida y con consenso científico. Como hemos comentado en anteriores párrafos, lo propuesto aquí no es algo novedoso y ha habido esfuerzos tanto por las comunidades científicas, como la Asociación Americana de Psicología (2020) y las revistas especializadas, para que se utilice con la finalidad de generar un abordaje adecuado de los fenómenos.

Más allá de proponer un método estándar, el trabajo tiene el objetivo contrario: posicionar y difundir el conocimiento de que existen múltiples formas válidas para efectuar el abordaje estadístico en la ciencia, contrario a la noción de que siempre deben utilizarse los métodos clásicos y obtener siempre los mismos resultados porque son "más robustos", ya que el número no hace a la ciencia, sino su precisión. La literatura ha insistido en que existen múltiples herramientas matemáticas igual de válidas y robustas, pero de baja difusión entre la comunidad científica. Este tipo de información "divergente" y también con sustento científico debe difundirse con mayor frecuencia.

En particular, usar el AP permite realizar el cálculo de la muestra para buscar un TE preciso para el fenómeno en particular, casi de manera personalizada para cada línea de investigación. Por lo anterior, sugerimos a los líderes de líneas de generación del conocimiento abordar sus propios estudios y hallazgos empleando el TE a fin de ser cada vez más precisos en sus resultados y que se re-

porte la transparencia necesaria para la comunidad científica. Lo anterior se encuentra respaldado con la noción de que la productividad en equipos de ciencia es, en general, mayor en comparación con la de los investigadores individuales (Wuchty et al., 2007), por lo que instruir a estudiantes, posdoctorandos y técnicos académicos a entender los fenómenos de estudio desde el AP probablemente beneficiará la producción del conocimiento.

REFERENCIAS BIBLIOGRÁFICAS

Aberson, C. L. (2019). *Applied Power Analysis for the Behavioral Sciences*. Routledge. https://doi.org/10.4324/9781315171500

Al-Saleh, M. F., & Samawi, H. M. (2007). Interference on Overlapping Coefficients in Two Exponential Populations. *Journal of Modern Applied Statistical Methods*, *6*(2), 503–516. https://doi.org/10.22237/jmasm/1193890440

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617. https://doi.org/10.1348/000712608X377117

Bartlett, J., & Charles, S. (2022). Power to the People: A Beginner's Tutorial to Power Analysis using jamovi. *Meta-Psychology*, *6*. https://doi.org/10.15626/MP.2021.3078

Bordes-Solanas. (2011). *Las trampas de Circe: falacias lógicas y argumentación informal* (1st ed.). Cátedra teorema.

Bradley, L., Noble, N., & Hendricks, B. (2020). The APA Publication Manual: Changes in the Seventh Edition. *The Family Journal*, *28*(2), 126–130. https://doi.org/10.1177/1066480720911625

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Chmura-Kraemer, H. (2014). *Wiley StatsRef: Statistics Reference Online* (N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, & J. L. Teugels (Eds.)). Wiley. https://doi.org/10.1002/9781118445112

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (Erlbaum (Ed.); 2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cohen, J. (1992). Things I have learned (so far). In *Methodological issues & strategies in clinical research.* (pp. 315–333). American Psychological Association. https://doi.org/10.1037/10109-028

Colombo, M., & Seriès, P. (2012). Bayes in the Brain—On Bayesian Modelling in Neuroscience. *The British Journal for the Philosophy of Science*, *63*(3), 697–723. https://doi.org/10.1093/bjps/axr043

Cureton, E. E. (1956). Rank-biserial correlation. *Psychometrika*, *21*(3), 287–290. https://doi.org/10.1007/BF02289138

García, J., Ortega, E., & De la Fuente, L. (2008). Tamaño del efecto en las revistas de Psicología indizadas en Redalyc. *Informes Psicológicos*, *10*(11), 173–188.

Glass, G. V., McGaw, B., & G.V., S. (1981). *Meta-Analysis in Social Research*. SAGE publiclations inc.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Gurnsey, R. (2017). *Statistics for research in Psychology: A modern approach using estimation.* SAGE publiclations inc.

Gutiérrez-Cabria, S. (1994). *Filosofía de la estadística*. Servei de Publicacions Universitat de Valéncia.

Guzmán-González, J. I., Sánchez-García, F., Ramirez-Vega, H., Sánchez-Loyo, L. M., & Ramírez-de los Santos, S. (2023). Tamaño del efecto para distribuciones No-paramétricas: concordancia entre medidas parala robustez de análisis en ciencias de la conducta. *Metodología, Instrumentación, Lógica, Estadística, Evidencias Y Epistemología En Salud*, *1*(14), 43–54. https://mileees.cucs.udg.mx/ojs/index.php/MILEEES/article/view/89

Hand, D. J. (2012). Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. In G. Cumming (Ed.), *International Statistical Review* (Vol. 80, Issue 2). https://doi.org/10.1111/j.1751-5823.2012.00187_26.x

Hartgerink, C. H. J., Wicherts, J. M., & van Assen, M. A. L. M. (2017). Too Good to be False: Nonsignificant Results Revisited. *Collabra: Psychology*, *3*(1). https://doi.org/10.1525/collabra.71

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kerby, D. S. (2014). The Simple Difference Formula: An Approach to Teaching Nonparametric Correlation. *Comprehensive Psychology*, *3*, 11.IT.3.1. https://doi.org/10.2466/11.IT.3.1

Kirk, R. E. (2003). The importance of effect magnitude. In S. Davis (Ed.), *Handbook of Research Methods in Experimental Psychology* (pp. 83–105). Blackwell.

Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.33267

Milton, S. (2001). *Estadística para la biología y ciencias de la salud* (3rd ed.). Edit. McGraw.

Neyman, J., & Pearson, E. S. (1993). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society*, *29*(4), 492–510.

Ortíz-Uribe, F. G. (2003). *Diccionario de metodología de la investi-*

gación científica (1st ed.). Noriega Editores.

Pepper, S. C. (1972). Systems Philosophy as a World Hypothesis. *Philosophy and Phenomenological Research*, *32*(4), 548. https://doi.org/10.2307/2106292

Polit, H. (2002). *Investigación científica en ciencias de la salud* (5th ed.). McGraw Hill.

Publication manual of the American Psychological Association (7th ed.). (2020). American Psychological Association. https://doi.org/10.1037/0000165-000

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Salgado, J. F. (2018). Transforming the Area under the Normal Curve (AUC) into Cohen's d, Pearson's r pb , Odds-Ratio, and Natural Log Odds-Ratio: Two Conversion Tables. *The European Journal of Psychology Applied to Legal Context*, *10*(1), 35–47. https://doi.org/10.5093/ejpalc2018a5

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, *25*(6), 2083–2101. https://doi.org/10.3758/s13423-018-1451-8

Solla, F., Tran, A., Bertoncelli, D., Musoff, C., & Bertoncelli, C. M. (2018). Why a P-Value is Not Enough. *Clinical Spine Surgery*, *31*(9), 385–388. https://doi.org/10.1097/BSD.0000000000000695

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

The jamovi project. (2022). *The jamovi project* (2.3).

Ventura-León, J. (2018). Otras formas de entender la d de Cohen. *Revista Evaluar*, *18*(3). https://doi.org/10.35670/1667-4545.v18.n3.22305

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "p < 0.05." *American Statistician*, *73*(1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wendt, H. W. (1972). Dealing with a common problem in Social science: A simplified rank-biserial coefficient of correlation based on the U statistic. *European Journal of Social Psychology*, *2*(4), 463–465. https://doi.org/10.1002/ejsp.2420020412

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical Evidence in Experimental Psychology. *Perspectives on Psychological Science*, *6*(3), 291–298. https://doi.org/10.1177/1745691611406923

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Wuchty, S., Jones, B., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. *Science*, *326*, 1036–1039. https://doi.org/10.1126/science.1136099

Recibido: septiembre 25, 2023
Última revisión: octubre 9, 2023
Aceptado: octubre 13, 2023