https://repository.uaeh.edu.mx/revistas/index.php/ICSA/issue/archive

Educación y Salud Boletín Científico Instituto de Ciencias de la Salud

Universidad Autónoma del Estado de Hidalgo

Publicación semestral, Vol. 11, No. 22 (2023) 139-146

ISSN: 2007-4573

Estudios de casos y controles: Propuesta de robustez de análisis para ciencias de la conducta

Case-control studies: Analysis robustness proposal for behavioral sciences

Jesua I. Guzmán-González [a], Franco G. Sánchez-García [b], Luis M. Sánchez-Loyo [c] & Saúl Ramírez-de los Santos [d]

Abstract:

In the field of behavioral sciences, it is essential to carry out adequate interpretations of the results obtained that allow the adequate rejection of the null hypothesis. It is known that the characteristics of the discipline have disadvantages for an adequate statistical approach compared to other applied areas of knowledge, whereby nature there is a greater probability of obtaining data with normal distributions, and therefore, using parametric techniques. Given the above, it is that a compilation of information is carried out on the most important elements suggested to expand and improve the robustness of the interpretation for both parametric and non-parametric techniques.

Keywords:

Statistical analysis, effect size, robust analysis, behavioral sciences

Resumen:

En el campo de las ciencias de la conducta resulta indispensable realizar interpretaciones adecuadas de los resultados obtenidos que permitan el rechazo de la hipótesis nula. Se conoce que las características propias de la disciplina cuentan con desventajas para el abordaje estadístico en comparación con otras áreas aplicadas del conocimiento, donde por naturaleza se tiene mayor probabilidad de obtener datos con distribuciones normales y, por lo tanto, usar técnicas paramétricas. Dado lo anterior es que se realiza una recopilación de información de los elementos de mayor importancia sugeridos para ampliar mejorar la robustez de la interpretación tanto para técnicas paramétricas como para las no paramétricas.

***Palabras Clave**:*

Análisis estadístico, tamaño del efecto, robustez de análisis, ciencias de la conducta

Introducción sobre la relevancia de la robustez de análisis

La productividad en cantidad de estudios de neurociencias y psicología de los últimos años han tomado especial relevancia a nivel mundial dada su aplicación en las ciencias relacionadas a las disciplinas biológicas. A nivel Latinoamérica entre el 2015 y el 2022 se han realizado un total de 4069 productos de investigación en la temática, de los cuales, Brasil, Chile y México han sido los principales productores (Galvez-Contreras et al., 2022). No obstante, desde el decenio pasado, diversos autores señalaron una crisis de

[a] Autor de Correspondencia, Universidad de Guadalajara, https://orcid.org/0000-0002-0043-4365, Email: jesua.guzman2993@alumnos.udg.mx

[b] Universidad de Guadalajara, https://orcid.org/0009-0000-9030-3627, Email: franco.sanchez9043@alumnos.udg.mx

[c] Universidad de Guadalajara https://orcid.org/0000-0001-8800-2622, Email: luis.sloyo@academicos.udg.mx

[b] Universidad de Guadalajara, https://orcid.org/0000-0002-7433-1908, Email: saul.ramirez@cucs.udg.mx

Fecha de recepción: 20/12/2022, Fecha de aceptación: 04/05/2023, Fecha de publicación: 05/06/2023

DOI: https://doi.org/10.29057/icsa.v11i22.10486

confianza cuyo punto central es el análisis estadístico, sugiriendo una controversial confiabilidad en los hallazgos publicados (Begley & Ellis, 2012; Ioannidis, 2005). Se considera que dicha situación es el resultado del tipo de variables y población estudiadas en la disciplina, lo que en la mayoría de los casos obliga a reportar muestras pequeñas con instrumentos limitados (Button et al., 2013), como lo es en el caso de las investigaciones en conducta humana.

A diferencia de otros campos, la investigación anteriormente mencionada cuenta con algunas desventajas predeterminadas naturalmente por los fenómenos de estudio. En otras palabras, es bastante común que se utilicen instrumentos que recojan variables nominales, ordinales o discretas discontinuas. Además, las poblaciones de estudio tienden a ser limitadas y no se obtiene una muestra lo suficientemente grande para poder realizar un análisis probabilístico y paramétrico, obligando en la mayoría de las ocasiones a utilizar métodos no probabilísticos y no paramétricos (Kerlinger, 1966; Goodwin & Goodwin, 2017). Por otro lado, también se ha descrito que pocos estudios en el campo alcanzan una distribución paramétrica. Micceri (1989) demostró que solo el 28.4% de los estudios alcanzan dicha distribución, condición que no ha cambiado radicalmente y se ha criticado duramente (Dudley-Marling, 2010), razón por la cual las propuestas con un diseño de pequeñas muestras (*small-N design*) han recibido bastante apoyo con el fin de aumentar la validez y confiabilidad de los resultados, donde las medidas de robustez y probabilidad de aparición son el centro de la metodología (Smith & Little, 2018).

Test de comprobación de la hipótesis nula, errores, sesgos y poder estadístico

Lo anterior descrito surge en respuesta a que la medidas de mayor uso en la comprobación de hipótesis es el llamado valor p (*p value*) o el test de significancia para la hipótesis nula (TSHN). Esta técnica está basada en un procedimiento que permite observar la probabilidad de que el muestreo entre dos grupos independientes se encuentre fuera de la zona crítica de rechazo, donde lo esperado es que la distribución inter-grupo no sea igual, lo que permite rechazar la hipótesis nula (H_0). No obstante, se han señalado una serie de errores al utilizar el valor p como único para la interpretación de los datos y rechazo de hipótesis (Greenland et al., 2016). El señalado con mayor frecuencia es el de tratar de medir la distancia entre los resultados de dos distribuciones (Colquhoun, 2017), cuando el valor p solo permite saber si estas son diferentes, no la distancia de las diferencias. Este error tiene su base en la estructura de la distribución, con facilidad se puede inferir que si las distribuciones no tocan la zona crítica configurada clásicamente al .05 no tendrían por qué ser iguales, no obstante y como se mencionó anteriormente, en psicología no se obtienen distribuciones tan uniformes tanto por el tipo de datos recogidos como por muestro posible, tampoco un valor más pequeño en el valor p necesariamente es una medida que refleja con exactitud el grado de diferencias si el autor no reporta el poder estadístico (β). Adicionalmente, se sabe que existe la probabilidad de que una distribución de más de 1,000 observaciones puede llegar a comportarse de forma no paramétrica si los datos son de tipo Likert donde la ordinalidad está sujeto a valor cero y máximo (efecto techo), un ejemplo puede ser una escala tipo Likert donde hay valores ordinales del 1 al 5, por el otro lado, puede haber un estudio con menos de 30 observaciones con variables continuas que perfectamente se ajustan al modelo paramétrico, este tipo de condiciones suelen ser los principales sesgos de los que los investigadores tienen que cuidarse (Podsakoff et al., 2003).

Distribución paramétrica y no paramétrica

La distribución de la muestra es el concepto central de la comprobación de hipótesis para comparación de grupos, también conocidos como estudios de casos y controles. El TSHN bajo el enfoque de Neyman-Pearson permite conocer la compatibilidad de la distribución de los datos con la H^0 donde el continuo entre 1 (perfectamente compatible) y 0 (perfectamente incompatible) permite apoyar la noción de diferencias a través del rechazo en la zona crítica (Wasserstein et al., 2019). Kerlinger (1966) sugiere que casi cualquier medida es susceptible a ser analizada para encontrar diferencias, no obstante, también se sabe que no todas las medidas obedecen la distribución normal en estudios de disciplinas biológicas, especialmente en mediciones conductuales (Hill & Dixon, 1982; Kitchen, 2009; Micceri, 1989). Es importante realizar esta distinción porque dependiendo de las pruebas de hipótesis serán los valores reportados. Los estándares de reporte para diferencias de grupos suelen ser al menos una medida de tendencia central y una medida de dispersión, lo cual puede cambiar drásticamente la interpretación de los datos y aumentar el sesgo en la comprobación de hipótesis. Se estila con regularidad que en las distribuciones paramétricas se utilice el promedio (□) y la desviación estándar (σ) ya que el TSHN tiene su base en la conversión de puntuaciones Z, asumiendo claro está que la curtosis y la asimetría se mantienen estables. Teóricamente, una distribución normal tiene estabilidad en sus elementos constituyentes. No obstante, las distribuciones no paramétricas no se comportan según lo previamente descrito, es bastante

común que existan curtosis extremas en comparación con la distribución normal o por el contrario que los extremos de la campana, también conocidas como colas bidireccionales, sea más pesadas al contener mayor frecuencia de datos dentro de su rango. Para tratar este tipo de distribuciones, Wilcoxon (1945) desarrollo una metodología basada en la suma de los rangos, donde en lugar de utilizar los datos brutos de las mediciones estos se acomodan ordinalmente, en este punto es en donde aparece la primera parte del problema, si bien es cierto que aunque se acomoden de forma ordinal el será el mismo, la mediana, otra medida de tendencia central, tomará mayor relevancia al acumular una frecuencia más altas de valores cercana a ella, y por lo tanto, las comparaciones entre grupos serán más visibles utilizándola en contraste con el .

Posteriormente, se desarrolló una metodología para la comprobación de hipótesis diferencias de grupo basada en los rangos. Esta técnica se conoce como la U de Mann-Whitney (1947) que parte de la suma de los rangos, donde la U es un conteo del número de veces en que el puntaje del grupo A y el puntaje del grupo B convergen y por lo tanto se define la probabilidad de aparición de un fenómeno en términos de favorable o desfavorable. Por ejemplo, si se tuviera un grupo control de 10 y uno experimental de 10, esto implicaría que existen un total de 100 pares (10 x 10 = 100); si 70 de esos pares fueran favorables y 30 desfavorables, según la suma de rangos, entonces la U es igual a 30, de tal modo que el dividir la U entre el total de pares (n) permite obtener la proporción de pares favorables representado por (ΣR). La fórmula de la U puede expresarse de la siguiente forma:

$$U_1 = \Sigma R_1 - \frac{n_1(n_1+1)}{2}$$
$$U_2 = \Sigma R_2 - \frac{n_2(n_2+1)}{2}$$

Por ello, pese a que es bastante común encontrar reportes del y la σ en pruebas no paramétricas, estas pueden ser no del todo adecuadas; ya que, revelan poco o nada de la relación de los rangos entre los grupos, y también es importante conocer las formas de distribución y frecuencia en las medidas no paramétricas, para ello el cálculo más adecuado que permite una interpretación certera es de los rangos intercuartílicos (RIQ) en lugar de la σ. Los RIQ cuentan con la peculiaridad de revelar el número de observaciones que hay entre la distancia del primer (Q^1) y el tercer (Q^3) cuartil, donde se agrupan el 50% de la muestra. Adicionalmente, dado que la característica más probable dentro de las distribuciones no paramétricas es que la curtosis sea más alta, y esta no se encuentre necesariamente en el centro de los datos, la mediana tiene mayor movilidad para encontrarse con los datos de mayor frecuencia.

Por otro lado, es bastante común divulgarse entre investigadores que las pruebas no paramétricas son más débiles que sus contrapartes paramétricas. Esto no es del todo cierto, ya que las técnicas no paramétricas resultan más exigentes y poderosas estadísticamente hablando en casos donde las distribuciones tienen colas de distribución más pesadas o tienen una curtosis más extrema (Chernoff & Savage, 1958; Hodges & Lehmann, 1956; Tanizaki, 1997), lo que sugiere que el tratamiento de los datos deberá ser más exigente y permitir menor variabilidad entre las distribuciones. Evidencia sobre la noción de la eficiencia relativa asintótica, un concepto utilizado para mediar la eficacia con una tasa de error Tipo I fija a α, ha demostrado ser muy cercana entre medidas paramétricas y no paramétricas. Dixon (1954) demostró una eficiencia en la prueba t de .955 comparada con el test de suma de rangos de Wilcoxon, que es la base de la prueba de diferencias de grupos no paramétrica más utilizada, la U de Mann-Whitney.

Elementos de robustez

Dado lo anteriormente descrito es que se ha sugerido utilizar estimadores de mayor robustez tanto para distribuciones paramétricas como no paramétricas (Stigler, 1977). Razón por la cual en trabajos previos se sugiere utilizar indicadores independientemente de la distribución (Guzmán-González et al., 2019). Evidencias han demostrado que existe una importante relación entre los valores p, el tamaño del efecto (δ) y el factor bayes (FB) (Wetzels et al., 2011), por lo que, junto con el cálculo de los intervalos de confianza se tendrá mayor seguridad al rechazar la H^0.

Tamaño del efecto y sus derivados

El tamaño del efecto es una medida propuesta para poder determinar cuantitativamente la distancia entre dos distribuciones, el valor δ fue propuesto por Cohen (2013) con el fin de medir la intensidad de la diferencia entre dos medidas independientes a las cuales se les realizó una medición. A nivel Latinoamérica se ha registrado que solo el 10% de las revistas solicitan esta medida de robustez (García et al., 2008). Cohen (2013) mencionó que su cálculo no tiene una interpretación estandarizada, aunque menciona que un valor de .80 es un valor deseable. Se advierte que este no debe entenderse como el valor p porque conllevaría una serie de errores metodológicos y estadísticos (Cohen, 2013; Gurnsey, 2017) y que debe complementarse con otras medidas de robustez (Baguley, 2009). Al respecto, Szucs y

colaboradores (2017) encontraron que los δ empíricos reportados para psicología fueron de δ = .23 para un efecto pequeño, δ = .60 para un efecto moderado y δ = .78 para un efecto amplio, mientras que para neurociencias se encontró un δ = .14 para un efecto pequeño, δ = .44 para un efecto moderado y un δ = .67 para un efecto amplio. Existen algunas consideraciones dado que existen tamaños del efecto para muestras con homogeneidad de varianzas (Cohen, 2013), sin homogeneidad de varianzas o comparaciones pareadas (Glass et al., 1981) y de una sola muestra donde el tamaño de ambos grupos es desproporcional (Hedges, 1981). La base general de la delta es la siguiente:

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 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

Algunos autores han sugerido hasta setenta variantes de medidas de tamaño del efecto (Kirk, 2003). Inclusive se ha propuesto para medidas no paramétricas (Cureton, 1956) denominada correlación de rangos biseriales (r_{rb}). Esta última es de la familia de las relaciones "r" en lugar de las diferencias δ, y se ha propuesto que estas son homologas (Ellis, 2010). De hecho, a partir de un despeje, la δ puede transformarse en r (Friedman, 1968), pero sus propiedades estadísticas hasta ahora no han sido del todo esclarecidas (Chmura-Kraemer, 2014) y se sugiere realizar investigaciones al respecto. La fórmula es la siguiente:

$$r = \frac{\delta}{\sqrt{\delta^2 + 4}}$$

Adicionalmente, se conocen diferentes formas de interpretación con base en el mismo cálculo δ de gran utilidad para pequeñas muestras. Por ejemplo, se encuentra el coeficiente de super posición (Ո) que 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):

$$Ո = 2\phi(\frac{-|\delta|}{2})$$

La interpretación es sencilla, por ejemplo, si se aplicara un cuestionario a dos grupos independientes y se obtiene una δ = 1, el cálculo arroja un Ո = .6170, lo que significa que alrededor de un 62% de la muestra tiene puntuaciones superpuestas o se interceptan entre sí, y que por lo tanto un 38% de las puntuaciones son diferentes entre ambos grupos. Según lo encontrado en Szucs (2017), los Ո para psicología fueron para un efecto pequeño de = .91, con un total de 9% de respuestas diferentes, para un efecto moderado de Ո = .76, y para un efecto amplio de Ո = .69; mientras que para neurociencias se encontró un Ո = .94 para un efecto pequeño, Ո = .82 para un efecto moderado y un Ո = .73 para un efecto amplio, por lo que existen altas probabilidades de un verdadero tamaño del efecto cercano a los valores de superposición.

Factor Bayes

En cuanto a los análisis de probabilidad, no todos requieren de una distribución, existen ciertos enfoques como lo es en el enfoque bayesiano (FB), que se utiliza para calcular los grados de incertidumbre para contrastar hipótesis (De la Fuente et al., 2009). Se ha descrito que la principal ventaja que ofrece el paradigma bayesiano para neurociencias y psicología es el de utilizar a su favor los δ previamente reportados en la literatura para otorgar una mayor precisión a la hipótesis y una posible predicción del δ encontrado. Según lo descrito por Wetzels (2011) se ha observado que un FB en contra de la H^0 es más probable obtener una significancia en el paradigma frecuentista del valor p (Rouder et al., 2009), lo que brinda posibilidades de encontrar verdaderos efectos a estudios de pequeñas muestras. Algunos software de análisis estadísticos ya cuentan con este análisis integrado, Guzmán-González y colaboradores (2019) calcularon los coeficientes necesarios para obtener un adecuado FB en función del grado de incertidumbre en valores δ (Anexo 1).

Dado que para el FB es importante cuantificar el grado de incertidumbre de aparición de un fenómeno, se ha propuesto que, con el fin de disminuir la subjetividad y cuantificar adecuadamente, se utilicen los valores ya previamente descritos (Schönbrodt & Wagenmakers, 2018). Para una revisión más profunda puede consultarse a Guzmán-González y colaboradores (2019).

Propuesta de metodológica y estadística, volviendo a las raíces

La Asociación Psicológica Americana (APA) ha instigado a nuevos investigadores a integrar medidas de robustez con el fin de mejorar la replicabilidad y la confiabilidad de los estudios en materia de ciencias de la conducta (APA, 2008; Appelbaum et al., 2018). Pese a que las medidas propuestas no son del todo novedosas, no suelen utilizarse por diversas razones, por lo que aquí se brinda

una recopilación de los pasos sugeridos para su aplicación en estudios de casos y controles.

El primer paso, que se considera para mejorar la robustez del análisis, es utilizar de nuevo el cálculo del β. Se sugiere utilizar el Software JAMOVI (2020) que ya tiene integrada la paquetería necesaria. Recordando que el cálculo de α y β son importantes para proteger la conclusión de falsos positivos (error I) o falsos negativos (error II) y ambos son igual de graves. Es más probable que el cálculo del β no figure en ninguna parte de la metodología. Algunos autores han insistido en que este cálculo se reporte a priori al estudio para que figure como una fortaleza en el mismo (Hartgerink et al., 2017; Luce & Kahn, 1999; Simmons et al., 2011). Por ejemplo, en la tabla 2 se utiliza un ejemplo del cálculo para muestras independientes típico de un estudio de casos y controles, aunque es posible calcularlo para las tres variantes de diferencias de grupo (independientes, relacionadas, una sola muestra). La relación usada es la siguiente: β = .80, α = .05, bidireccional, N^1 = 20, N^2 = 20, el β es el sugerido por Cohen (2013) para ciencias de la conducta. Los resultados son observados en la tabla 2. Los resultados pueden variar según el tamaño relativo de las muestras. Mientras mayor sea el número de comparación del grupo control (N^1) menor será el δ pero mientras mayor sea el grupo experimental (N^2) mayor será el δ.

Tabla 1
Análisis a priori del poder estadístico

δ	N^1	N^2	β	α
0.909	20	20	0.800	0.050

La relación anterior muestra que para un estudio con esas características se encuentra protegido del error tipo I si se obtiene una δ >= .63, mientras que para el error tipo II si se obtiene una δ >= .90. Es importante resaltar esta relación, dado que el cálculo del poder estadístico tiene el fin de evaluar la sensibilidad del diseño. En la tabla 3 se observan los datos.

Tabla 2
Poder estadístico para β según la detección con δ

Rangos de δ	Probabilidad de detección	Seguridad
0 < δ = 0.636	≤50%	Incertidumbre
0.636 < δ = 0.909	50% – 80%	Poca probabilidad
0.909 < δ = 1.170	80% – 95%	Gran probabilidad
δ = 1.170	≥95%	Certidumbre

En la figura 1 se observan los gráficos de contorno que demuestran la potencia, esta cambia de sensibilidad dependiendo del δ establecido. A medida que se aumentan los tamaños de la muestra, los δ serán más pequeños, y por lo tanto, se vuelven más detectables con gran certidumbre. Por el contrario, si se desea detectar de forma confiable un δ amplio se necesitan tamaños de muestra más pequeños. La curva negra continua en el gráfico muestra el contorno de estas combinaciones con relación al tamaño/muestra con una β de .80, por lo que como se demuestra en la curva de potencia la sensibilidad de la prueba y el diseño es mayor para tamaños más grandes del δ. El diseño del estudio será lo suficientemente sensible para detectar tamaños de | δ | > .909 con gran certidumbre. Por el otro lado, si el δ es menor a .636 la sensibilidad sería mínima y se podría cometer fácilmente el error tipo II.

Cuando el δ observado está lo suficientemente alejado de 0 para ser más extremos en los criterios de rechazo se dice que la H^0 es falsa y se debe de iniciar la discusión para la H^1. Si la H^0 fuera verdadera y δ es cercana a 0, la evidencia al respecto dictaría que debe de rechazarse erróneamente la H^0 como máximo el 5% de las veces. Por el otro lado, si δ > .91, la evidencia excedería el criterio de rechazo, y por lo tanto, correctamente asumir que δ < 0, al menos en el 80% de las veces. La potencia entonces del diseño para detectar | δ | > .91 es de .8 como lo previamente definido.

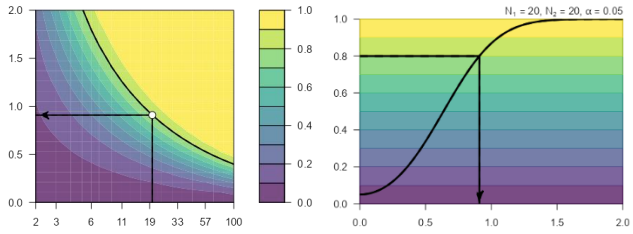

Figura 1
De izquierda a derecha, se puede observar el primer panel con recuadro y una escala de color morado hasta amarillo. El recuadro es la representación gráfica del contorno del poder, en el eje vertical se encuentra el δ hipotético, mientras que en el eje horizontal se encuentra el tamaño relativo de los grupos, para este caso n = 20. A nivel central se encuentra una escala de color que representa el β. En el recuadro de la derecha se encuentra la representación de la curva de poder según el δ, siguiendo la escala central donde el máximo poder estadístico se encuentra en color amarillo, mientras que en el eje horizontal se observa el δ hipotético.

Al respecto de las diferencias entre estudios con muestras paramétricas y no paramétricas, se alienta a que se utilicen las medidas adecuadas para cada una de las distribuciones; sugiriendo que para los reportes de datos los criterios mínimos sean al menos una medida de tendencia central (MTC), una medida de dispersión (MD), el estadístico utilizado (E), los intervalos de confianza al 95% superiores (ICsup) e inferiores (ICinf), la medida del tamaño del efecto (TE), la medida de intersección (Ո) y

el factor bayes ($FB^{10}/^{01}$), reportando significancia con un asterisco según las cifras significativas clásicas (p = .05*, p = .01**, p = .001***), como se muestra en el ejemplo de la tabla 4. Una precisión importante es que si bien es cierto que el cálculo de los intervalos de confianza suele usarse al 95% en las distribuciones paramétricas (z = 1.96, -2 | σ | 2) en distribuciones no paramétricas puede utilizarse el RIQ, donde la distancia entre el Q^1 y el Q^3 es del 50% de la muestra o el equivalente a (-.68 | σ | .68) para construir un intervalo congruente con en RIQ que al mismo tiempo coincide con el percentil 50 y el Q^2.

En el ejemplo de la tabla 4 se observan los resultados de dos puntajes con diferentes distribuciones (n = 20) probadas con la prueba de Shaphiro-Wilks para menos de 50 observaciones, uno discontinuo no paramétrico (p = .026) y otro continuo paramétrico (p = .81) respectivamente. El δ en la muestra no paramétrica se encuentra transformado en la correlación de rangos biseriales (r_{rb}), en ambas muestras el δ es amplio. En cuanto a la conversión, dado que se obtuvo una r_{rb} = .73, al despejarse se obtiene una δ = 2.14, como se observa es discrepante de la obtenida directamente de los datos δ = 1.62 con un ∩ = .42, aunque la significancia con ambos métodos es bastante similar (U = .005, t =.003), lo que sugiere que la manera adecuada es hacerlo con todos los estadísticos no paramétricos para mayor concordancia.

Tabla 3

Ejemplificación de reporte de datos en dos muestras teóricas

Variable	MTC (MD)		E	TE	∩	FB^{10}
	Caso	**Control**				
Puntaje 1	16 (1)	19 (3.5)				
IC			13.50	.73	.71	6.23
IC^{sup}	17.2	20.2				
IC^{Inf}	15.8	17.8				
Puntaje 2	13.9 (1.35)	17.2 (1.33)				
IC						
IC^{sup}	14.7	18	5.47	2.4	.22	155.3
IC^{Inf}	13	16.3				

Nota: p = .05, p = .01**, p = .001****

El coficiente de r_{rb} = .73 es igual a δ 2.14 para el puntaje 1, mientras que para el puntaje 2 el δ = 2.45 es r_{rb} = .77

Respecto a esto último, es importante realizar más investigaciones al respecto dado que en la primera medida se tiene un tope máximo de 1 dado que está bajo la metodología de clásica de las medidas correlativas (r = 0 → 1), mientras que en la δ clásica no se encuentra un límite conocido (δ = 0 → ∞), recordando que el efecto más amplio reportado es menor a 5 empíricamente y fue localizado por Szucs (2017). Esta propiedad se encuentra fuera de los límites de la investigación del presente trabajo, aunque no se haya explorado matemáticamente aún y se desconozca la concordancia entre ambos límites. Por otro lado, el ∩ y los IC pueden ser dos medidas intercambiables dependiendo del fenómeno medido. Es probable que por los tipos de puntajes uno sea más confiable que otro por propias características. Mientras que para el puntaje obtenido en un cuestionario o una prueba es relevante cuantas respuestas en común hay entre dos grupos distintos, para un medición de biomarcador como el cortisol es más relevante saber con qué estabilidad según la distribución se observa por grupo; eso dependerá de la experiencia del investigador y la utilidad que el estadístico pueda brindar a la interpretación de los análisis.

Respecto a la comprobación de la hipótesis para ambas puntuaciones se había calculado de la siguiente manera como puede observarse en la tabla 5. Los cálculos para la prueba de hipótesis sugerían que si δ < .63 no se rechazaba la H^0 se cometería un error tipo I, por el otro lado si δ = > .91 y no se aceptaba la H^0 se cometería el error tipo II. Por esta relación es que algunos investigadores decidieron solo tomar el valor α para protegerse de los sesgos, sin embargo, puede existir el evento en donde se obtenga un α no significativo y obtener un δ lo suficientemente grande para poder decir que se está cayendo en un falso negativo, es por ello que se sugiere tener este cálculo como anteriormente se hacía (Ioannidis, 2005; Nosek et al., 2012).

Tabla 4

Comprobación de límites para falsos positivos y negativos

Pruebas	I	δ	II
Puntaje 1	.63	2.14	.91
Puntaje 2	.63	2.45	.91

Nota: El δ del puntaje 1 se encuentra convertido de r_{rb} a δ, I y II representan el tipo de errores.

Conclusiones

Se articula un cuerpo literario sobre la importancia de realizar buenas prácticas de análisis para la comprobación de hipótesis en estudios de casos y controles, lo anterior con el fin de frenar la crisis de confiabilidad que se encuentra actualmente en las ciencias biomédicas, en particular los subcampos de la psicología y las neurociencias. Esto es relevante en el sentido que Latinoamérica, en especial México, ha tenido una importante producción literaria sobre la temática posicionándolo en tercer lugar de Latinoamérica y en el 33/219 del ranking mundial (SJR, 2022), por lo que con el fin de aumentar la calidad de los productos se articula el presente trabajo. Dentro de las perspectivas más relevantes, primero se debe considerar evaluar la concordancia entre las medidas del efecto entre estadísticas paramétricas y no paramétricas, particularmente de las diferencias entre δ ~ r_{rb} y δ ~ r,

posteriormente, se aconseja a las revistas a solicitar el cálculo del poder estadístico con un β y aumentar el uso del cálculo δ, los IC y el FB como medidas de robustez para las hipótesis presentadas en trabajos metodológica y estadísticamente congruentes.

Las principales limitaciones de los estudios de casos y controles, fuera del campo teórico, está relacionado al tipo muestreo, el cual suele ser no probabilístico y difícilmente se cumplen los criterios de normalidad. Adicionalmente, existen más limitaciones relativas a la población estudiada, ya que éstas suelen ser sin reemplazo, lo cual limita aún más el tamaño del a n, razón por la que se aconseja calcular el tamaño de la muestra en función de un verdadero efecto en lugar de la representatividad, típica del muestreo probabilístico.

Referencias

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

APA. (2008). *Reporting standards for research in psychology: Why do we need them? What might they be?* American Psychologist, 63(9), 839–851. https://doi.org/10.1037/0003-066X.63.9.839

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. American Psychologist, 73(1), 3–25. https://doi.org/10.1037/amp0000191

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

Begley, C. G., & Ellis, L. M. (2012). *Raise standards for preclinical cancer research. Nature*, 483(7391), 531–533. https://doi.org/10.1038/483531a

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

Chernoff, H., & Savage, I. R. (1958). *Asymptotic Normality and Efficiency of Certain Nonparametric Test Statistics*. The Annals of Mathematical Statistics, 29(4), 972–994. http://www.jstor.org/stable/2236941

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. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Routledge. https://doi.org/10.4324/9780203771587

Colquhoun, D. (2017). *The reproducibility of research and the misinterpretation of p -values*. Royal Society Open Science, 4(12), 171085. https://doi.org/10.1098/rsos.171085

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

De la Fuente, E. I., Cañadas, G. R., Guàrdia, J., & Lozano, L. M. (2009). *Hypothesis Probability or Statistical Significance? Methodology*, 5(1), 35–39. https://doi.org/10.1027/1614-2241.5.1.35

Dixon, W. J. (1954). *Power Under Normality of Several Nonparametric Tests*. The Annals of Mathematical Statistics, 25(3), 610–614.

Dudley-Marling, C. (2010). *The myth of the normal curve* (1st ed.). Peter Lang.

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes. In The Essential Guide to Effect Sizes*. Cambridge University Press. https://doi.org/10.1017/cbo9780511761676

Friedman, H. (1968). *Magnitude of experimental effect and a table for its rapid estimation*. Psychological Bulletin, 70(4), 245–251. https://doi.org/10.1037/h0026258

Galvez-Contreras, A. Y., Guzmán-Muñiz, J., Moy-López, N. A., & Gonzalez-Perez, O. (2022). *Contributions of Latin America to scientific research in neuroscience and psychology*. Revista Mexicana de Neurociencia, 23(2). https://doi.org/10.24875/RMN.21000034

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.

Guzmán-González, J. I., Sánchez-García, F., Madera-Carrillo, H., & Medina-Aguayo, F. (2019). *Propuesta de verificación de la robustez del análisis y comprobación de hipótesis en los resultados de estudios en neurociencia cognitiva, psicología y medicina*. Revista Mexicana de Investigación En Psicología, 11(2).

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

Hill, M., & Dixon, W. J. (1982). Robustness in Real Life: A Study of *Clinical Laboratory Data. Biometrics*, 38(2), 377. https://doi.org/10.2307/2530452

Hodges, J. L., & Lehmann, E. L. (1956). *The Efficiency of Some Nonparametric Competitors of the t-Test*. The Annals of Mathematical Statistics, 27(2), 324–335. http://www.jstor.org/stable/2236996

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

Kerlinger, F. N. (1966). *Foundations of behavioral research* (Vol. 1). Holt, Rinehart and Winston. https://psycnet.apa.org/record/1966-35003-000

Kerri A. Goodwin, C., & James Goodwin. (2017). *Research in Psychology: Methods and Design* (8th ed.). John Wiley & Sons.

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

Kitchen, C. M. R. (2009). *Nonparametric vs Parametric Tests of Location in Biomedical Research*. American Journal of Ophthalmology, 147(4), 571–572. https://doi.org/10.1016/j.ajo.2008.06.031

Luce, M. F., & Kahn, B. E. (1999). *Avoidance Or Vigilance? the Psychology of False-Positive Test Results*. Journal of Consumer Research, 26(3), 242–259. https://doi.org/10.1086/209561

Mann, H. B., & Whitney, D. R. (1947). *On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other*. The Annals of Mathematical Statistics, 18(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Micceri, T. (1989). *The unicorn, the normal curve, and other improbable creatures*. Psychological Bulletin, 105(1), 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). *Scientific Utopia*. Perspectives on Psychological Science, 7(6), 615–631. https://doi.org/10.1177/1745691612459058

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). *Common method biases in behavioral research: A critical review of the literature and recommended remedies*. Journal of Applied Psychology, 88(5), 879–903. https://psycnet.apa.org/buy/2003-08045-010

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

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). *Bayes factor design analysis: Planning for compelling evidence*. Psychonomic Bulletin & Review, 25(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

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

SJR. (2022, September 29). *International Science Ranking*. https://www.scimagojr.com/countryrank.php?area=3200

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

Stigler, S. M. (1977). *Do Robust Estimators Work with Real Data?* The Annals of Statistics, 5(6). https://doi.org/10.1214/aos/1176343997

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

Tanizaki, H. (1997). *Power comparison of non-parametric tests: Small-sample properties from Monte Carlo experiments*. Journal of Applied Statistics, 24(5), 603–632. https://doi.org/10.1080/02664769723576

The jamovi project. (2020). Jamovi (1.2). https://www.jamovi.org

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

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

Wilcoxon, F. (1945). *Individual Comparisons by Ranking Methods*. Biometrics Bulletin, 1(6), 80. https://doi.org/10.2307/3001968

Anexo 1

Valores δ y coeficientes para calcular el FB reportados en el trabajo de Guzmán-González y colaboradores (2019)

	Disciplina	δ	20%	30%	40%	50%	60%	70%	80%	90%
Efecto pequeño	Neurociencia cognitiva	.14	.430	.274	.192	.140	.101	.071	.045	.022
	Psicología	.23	.707	.451	.316	.230	.167	.117	.074	.036
	Medicina	.23	.707	.451	.316	.230	.167	.117	.074	.036
	Todos los subcampos	.17	.523	.333	.233	.170	.123	.086	.055	.026
Efecto moderado	Neurociencia cognitiva	.44	1.354	.863	.605	.440	.319	.224	.142	.069
	Psicología	.60	1.846	1.177	.825	.600	.435	.307	.194	.095
	Medicina	.57	1.744	1.118	.784	.570	.414	.290	.185	.090
	Todos los subcampos	.49	1.508	.961	.674	.490	.356	.249	.159	.077
Efecto amplio	Neurociencia cognitiva	.67	2.062	1.314	.922	.670	.486	.341	.217	.106
	Psicología	.78	2.400	1.530	1.073	.780	.566	.397	.253	.123
	Medicina	.77	2.369	1.511	1.059	.770	.559	.392	.250	.121
	Todos los subcampos	.71	2.185	1.393	.977	.710	.515	.361	.230	.112