Propuesta de verificación de la robustez del análisis y comprobación de hipótesis en los resultados de estudios en neurociencia cognitiva, psicología y medicina

Jesua Guzmán-González[1], Franco Sánchez-García[2], Humberto Madera-Carrillo[3], Felipe Medina-Aguayo[4]

[1] *Centro de Evaluación e Investigación en Psicología, Departamento de Psicología Básica, Centro Universitario de Ciencias de la Salud, Universidad de Guadalajara. Guadalajara, Jalisco, México*
[2] *División de Innovación y Desarrollo Tecnológico en Psicología, Colegio de Profesionales de la Psicología del Estado de Jalisco. Guadalajara, Jalisco, México*
[3] *Departamento de Disciplinas Filosóficas, Metodológicas e Instrumentales, Centro Universitario de Ciencias de la Salud, Universidad de Guadalajara. Guadalajara, Jalisco, México*
[4] *Departamento de Probabilidad y Estadística, Centro de Investigación en Matemáticas, AC, Guanajuato, Guanajuato, México*

Cómo citar este artículo (estilo APA) / Citing this article (APA style):
Guzmán-González, J., 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*(1), pp 11-22

Resumen
El p-value es el análisis estadístico más utilizado para comprobar hipótesis en áreas de la salud; sin embargo, su uso único presenta desventajas que afectan los estudios por depender del tamaño de la muestra y necesitar una distribución normal. Esto justifica buscar formas más flexibles y robustas para comprobar hipótesis. El objetivo de este trabajo es recopilar procedimientos estadísticos complementarios sensibles para muestras típicamente clínicas. Durante el proceso, se eligió el tamaño del efecto y la inferencia bayesiana. Ambos presentan una correlación entre ellos y con el p-value, lo que permite disminuir la incertidumbre al comprobar supuestos. Se propone que el grado de certidumbre de la inferencia bayesiana se base en literatura previa y valores δ reportados, y se añada el apartado de intervalos de credibilidad en la tabla de mediciones generales, además de incluir las hipótesis sometidas a prueba y mostrar las evidencias a favor de H^0 y H^1. La implementación de las propuestas presentadas como complemento de la estadística frecuentista apoyará la interpretación de hipótesis con mayor fuerza estadística y mejor toma de decisiones. Se sugiere integrarlas a la revisión académica o de cursos, y la realización de convenciones para esparcir su uso adecuado.
Palabras clave: intervención clínica, interpretación, análisis bayesiano, tamaño del efecto

Results of studies in cognitive neuroscience, psychology, and medicine. A proposal to verify the robustness of the analysis, and hypotheses

Abstract
The history and growth of humanity was determined, not only by the advances of the results of scientific discoveries and investigations, but also by regrettable situations such as wars and great epidemics that have killed millions of human beings. Each one of these situations has marked milestones that have generated opportunities for change and evolution at multiple levels. The article recounts and analyzes the history, the symptomatic descriptions, the etiology and the context of the moment, in which diseases, large viral outbreaks and infections developed. Epidemics and pandemics that are the result of context variables such as poverty, lack of hygiene, individualism, the hyperkinetic rhythm of life. A

Dirigir toda correspondencia al autor a la siguiente dirección:
Jesua Guzmán-González
Sierra Mojada 950, Puerta 7, Edificio A, Col. Independencia, CP 44340, Guadalajara, Jalisco, Departamento de Psicología Básica, Edificio I segunda planta.
RMIP 2019, Vol. 11, Núm. 1, pp. 11-22
www.revistamexicanadeinestigacionenpsicologia.com

detailed historical bibliographic study was carried out where some epidemics were selected. The research developed from ancient Greece and was divided for its organization into Pandemic Pests, Historical Pandemics, Pandemic Flu, Pandemics and more recent epidemics. The purpose is to analyze the variables vulnerability, resilience, and crisis, and find common patterns in the different epidemics such as xenophobic attitudes, poverty and health, disease as punishment, social crisis, hygiene, emotions such as uncertainty, fear, anguish and anxiety.

Keywords: clinical intervention, interpretation, Bayesian analysis, effect size

INTRODUCCIÓN

Dada la reciente crisis de confianza en la investigación clínica e interventora (Begley & Ellis, 2012; Button *et al.*, 2013; Ioannidis, 2005; John *et al.*, 2012; Nosek *et al.*, 2012; Nosek & Bar-Anan, 2012; Pashler & Wagenmakers, 2012; Simmons *et al.*, 2011), los investigadores se vieron forzados a explorar formas más robustas, válidas y fiables para apoyar los resultados en investigaciones. El procedimiento más común utilizado es la interpretación del valor p (p-value), que calculan la mayoría de los software estadísticos; sin embargo, en forma reciente, se ha sugerido utilizar medidas complementarias para mejorar la potencia estadística que permita disminuir el grado de incertidumbre, dado que este es insuficiente para llegar a una adecuada conclusión (Solla *et al.*, 2018).

El valor p ("*p*"), también conocido como *test* de significancia para la hipótesis nula (TSHN), es un procedimiento que permite observar la probabilidad de que el muestreo de los datos se encuentre dentro o fuera de lo esperado en términos de distribución, y se asume que la hipótesis nula (H^0) es cierta. A partir de este supuesto, hay dos marcos de interpretación: en el marco fisheriano es interpretado como una medida continua de compatibilidad entre los datos observados y la H^0 (Greenland *et al.*, 2016), mientras que, en el marco de Neyman-Pearson, el objetivo de la prueba estadística es la toma de decisiones, de tal manera que, basados en los resultados de la prueba estadística, y sin saber si la hipótesis es verdadera o falsa, se asume un tipo de hipótesis dependiendo de lo buscado en el estudio, por ejemplo, comprobar diferencias de grupo o correlaciones en función de lo esperado.

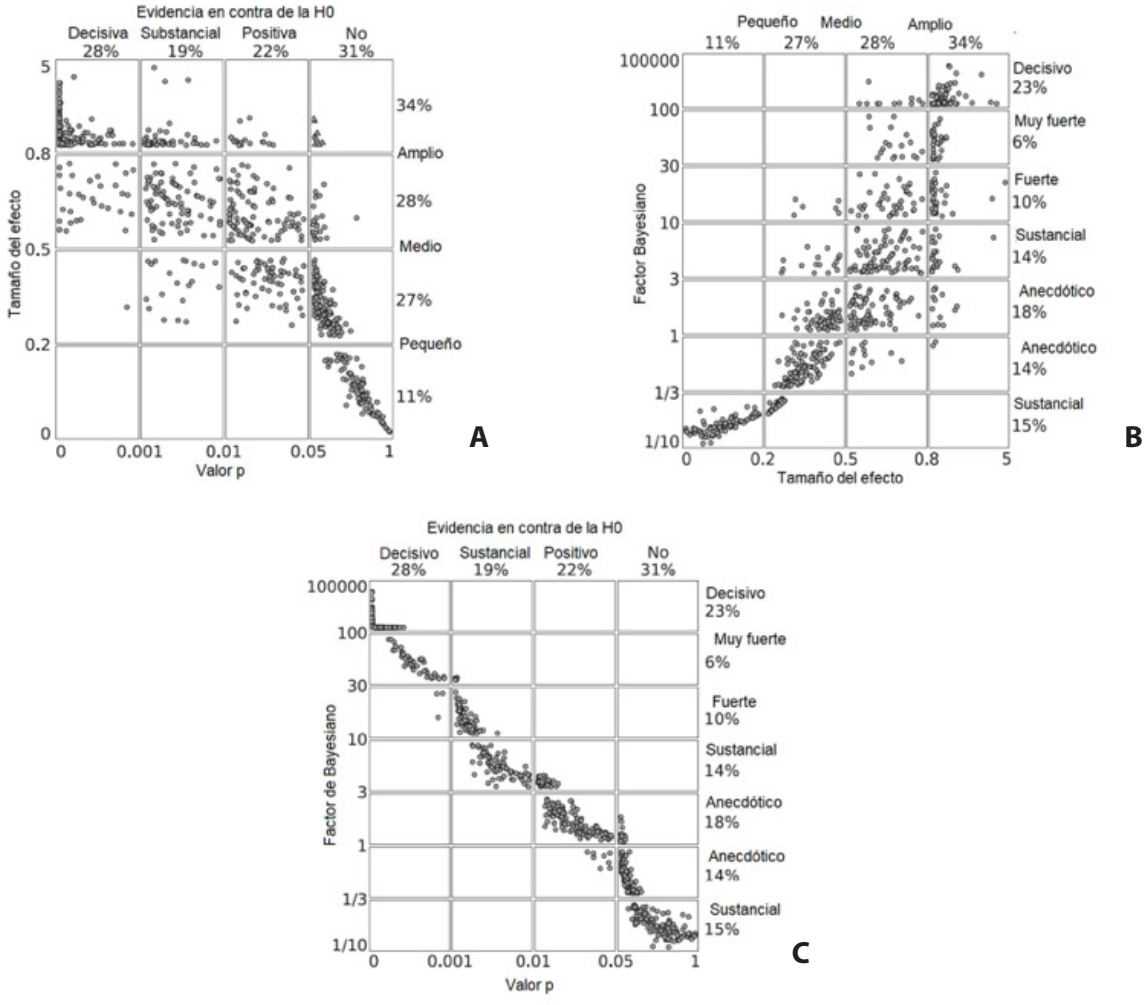
En este sentido, la compatibilidad de los datos con la H^0 puede establecerse con el continuo entre 1 (perfectamente compatible) y 0 (perfectamente incompatible), por lo que su interpretación depende de la lógica y las consideraciones estadísticas; lo más usual es que se asuma que un valor $p < .05$ considera incompatibilidad con una distribución que apoye la H^0 (Wasserstein *et al.*, 2019); de hecho, es bastante usual que se admita como una forma de observar la ausencia de relación o de efecto entre dos o más fenómenos medidos dada la paridad o la incompatibilidad de lo obtenido (Colquhoun, 2017), aunque dicho estadístico no haya sido creado para esa función.

De acuerdo con lo anterior, es necesario resaltar que el uso único de la *p* tiene desventajas importantes; incluso, Greenland (2016) enumera al menos 25. No obstante, existen algunas que afectan de modo directo los estudios de intervención; una de las más importantes es que este valor es altamente dependiente del tamaño de la muestra; por ello, un muestreo pequeño dificulta alcanzar una significancia, o bien, el hecho de que este no siempre alcanza una distribución normal.

En cuanto a este último punto, Micceri (1989) encontró en una muestra de cerca de 440 estudios clínicos y de intervención llevados a cabo en la época que tan solo el 28.4% alcanzaban una distribución paramétrica en términos de simetría; por ello, eso obliga a los académicos a buscar formas más flexibles, pero igual de válidas para la comprobación de hipótesis en este tipo de estudios. Por lo tanto, el objetivo de nuestro estudio es recopilar procedimientos estadísticos complementarios que permitan robustecer la evidencia obtenida para la confirmación de hipótesis, sobre todo para estudios de intervención. Para ello, existen algunas alternativas que potencian el poder estadístico donde, por cualquier razón, la muestra es relativamente pequeña, o bien, no se alcanza la normalidad distribucional.

Estos cálculos son procedimientos complementarios requeridos para mejorar las pruebas de hipótesis (Hand, 2012) y, a pesar de que el procedimiento idóneo debería ser calcular a priori el tamaño de la muestra y la potencia estadística, por lo regular se calculan por separado, pese a que la tendencia más reciente en términos de evidencia científica aconseja calcularlos en el mismo estudio, dado que se ha descrito una correlación importante entre el valor δ, el FB y el valor p. De esa manera, una congruencia entre los tres permite disminuir la incertidumbre en la comprobación de los supuestos. Esto puede observarse en un estudio de Wetzels y colaboradores (2011) que encontró, en un total de 855 pruebas t, que los valores p, los valores δ y el FB tenían una gran probabilidad de coincidir en el sustento de la hipótesis (ver Figura 1); esta potencia estadística contribuye a que los investigadores no sobreestimen la evidencia a favor de un efecto solo con el valor p.

Figura 1. Esquema de relación entre el valor p, el tamaño del efecto y el factor bayes



Nota: Relación entre el tamaño del efecto y los valores p (A), relación entre el tamaño del efecto y el factor bayesiano (B) y por último relación entre el valor p y el factor bayesiano (C). Diagrama de dispersión donde los puntos indican una coherencia relativa entre los tres cálculos realizados a partir de la muestra dada (855), mientras que los indicados por triángulos indican grandes inconsistencias, adaptado de Wetzels (2011).

δ = D DE COHEN, VALOR DELTA O TAMAÑO DEL EFECTO

El cálculo de δ es una medida cuantitativa que ayuda a entender el grado en el que un fenómeno está presente en un grupo de estudio; para datos paramétricos, se desarrolló la d de Cohen (2013), la cual puede entenderse de dos formas: como la forma 'r' en términos de asociación o la 'd' en términos de diferencias de medias entre grupos (Ellis, 2010). Este cálculo ha sido poco utilizado en revistas psicológicas según revisiones sistemáticas: solo alrededor de un 10% (García *et al.*, 2008). La operación matemática para el cálculo de la forma d se expresa como:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sigma}$$

Mientras que, para la forma r, medida de la cual se puede inferir relación, es:

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

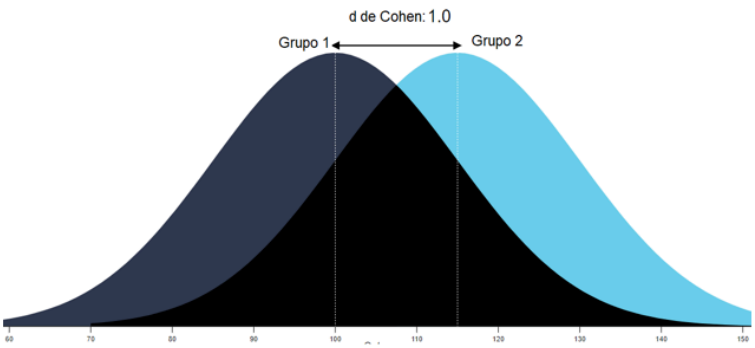
El propio Cohen (2013) menciona que este efecto no tiene una interpretación estandarizada y que, a pesar de que se menciona un ".80" como un valor adecuado, también advierte que no debe considerarse como una ley sobre los datos y evitar entenderse como un valor p a toda costa (Gurnsey, 2017), ya que asumir puntos de corte puede ser engañoso (Baguley, 2009). De hecho, los autores sugerimos observar cuáles son los valores δ específicos reportados para cada fenómeno estudiado, porque es probable que una ciencia como la química, que es relativamente más estable en sus variables, presente un valor δ mayor en comparación con la sociología, que tiene variables menos estables.

Al respecto, un trabajo de Szucs y Ioannidis (2017) recopila los valores reportados comúnmente en la literatura y concluye que, para la literatura de la neurociencia cognitiva, hay un δ pequeño de .14, moderado de .44 y amplio de .67; para la medicina, un δ pequeño de .23, moderado de .57 y amplio de .77, mientras que para la psicología, un δ pequeño de .23, moderado de .60 y amplio de .78. Además, existen formas de mejorar la interpretación de este valor, como el coeficiente de superposición (OVL o ∩) (Al-Saleh & Samawi, 2007; Reiser & Faraggi, 1999), que es el resultado de multiplicar dos veces la función de la distribución acumulativa para la distribución normal por el negativo de la mitad de la δ (para consultar su cálculo de manera sencilla en el programa R, revisar a Ventura-León [2018] junto con otros coeficientes).

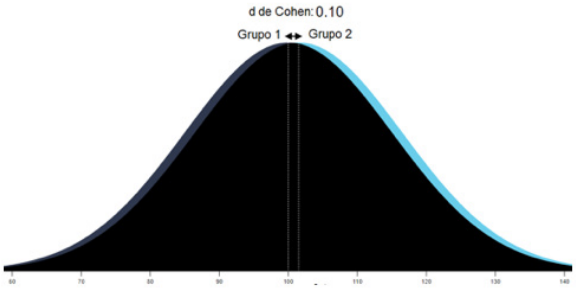
Con dicho coeficiente podemos observar qué tanto porcentaje de los datos obtenidos se superpone entre sí, es decir, si hubiera un d = 1.0, entonces obtenemos un ∩ = .6170 y, por lo tanto, podemos asumir que cerca del 62% de las puntuaciones de los grupos estudiados se superponen/interceptan entre sí, como en la Figura 2A; por el otro lado, en la Figura 2B se tiene un d = 0.1, ∩ = .96 y, en consecuencia, una intersección del 96% de las puntuaciones. En este caso es muy poco probable obtener una significancia estadística.

Figura 2. Representación gráfica del tamaño del efecto cuando se compraran grupos

A



B



Nota: El término de superposición en este caso se entiende como la zona más oscura donde se interponen ambas distribuciones (ſl), mientras que ambos extremos más claros son los grupos estudiados. El presente gráfico se hizo con ayuda del visualizador creado por Magnusson (Magnusson, 2020)

Para medidas no paramétricas, existe un cálculo muy semejante derivado de la correlación de rangos biseriales de Glass (1966), en el que se utiliza la prueba de suma de rangos U desarrollada por Mann y Whitney (1947), a partir de la prueba no paramétrica de Wilcoxon (1945). En su artículo original, Mann y Whitney definen la U como un conteo del número de veces en las que el puntaje de y, el puntaje del grupo 1, precede el rango del puntaje de x, que proviene del grupo 2.

En ese sentido, la U es fácil de interpretar dado que, dependiendo del número de pares, se puede saber qué tan favorable es para una hipótesis. De esa manera, interpretar el tamaño del efecto entre ambos grupos requiere primero entender su lenguaje bajo el supuesto de la correlación similar a la 'r' que clásicamente se utiliza; 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 (f), que puede ser usado como una medida del tamaño del efecto con la U de Mann-Whitney (Kerby, 2014).

McGraw y Wong (1992) discutieron en su trabajo la mejor forma de poder utilizar el valor δ; en ese sentido, cuando se tiene que interpretar con base en el marco de las correlaciones biseriales, en el cual el resultante de la ecuación de Wendt (1972) proviene de considerar la U y el tamaño de la muestra permite entender la dirección de los datos, la fórmula es la siguiente:

$$r = \frac{2(\bar{R}_1 - \bar{R}_2)}{n_1 + n_2}$$

En términos prácticos, a mayor nivel de correlación, menor efecto entre variables. Aunque también existe la posibilidad de realizar la ecuación inversa para determinar un δ a partir de r con la fórmula de Friedman (1968):

$$d = \frac{2r}{\sqrt{1-r^2}}$$

LA ESTADÍSTICA BAYESIANA

El enfoque bayesiano debe diferenciarse de la estadística inferencial clásica, que utiliza comúnmente los supuestos del TSHN. A esta última también se le conoce como estadística frecuentista, por considerar que la probabilidad de que un evento ocurra puede ser interpretada como la frecuencia relativa, o proporción de veces, de que el evento en cuestión suceda en una serie de experimentos. En ese sentido, se requiere entender que utilizar de manera adecuada el TSHN significa realizar una primera línea de pruebas estadísticas que deben complementarse, ya que, en cierta medida, estas permiten separar los verdaderos efectos de una probabilidad aleatoria de obtener lo que el investigador desea (Benjami-

ni & Cohen, 2017). En contraste, la estadística bayesiana tiene algunas bondades que pueden aprovecharse, en especial en muestras pequeñas.

Una de las principales ventajas del paradigma bayesiano es que favorece el utilizar valores δ conocidos de la propia disciplina, así que otorga al investigador herramientas de mayor precisión para discutir los nuevos hallazgos, o bien, a la toma de decisiones en el estudio; de nuevo, sugerimos utilizar el trabajo de Szucs (2017) a fin de consultar en su análisis los valores delta encontrados en las disciplinas descritas en esa propuesta.

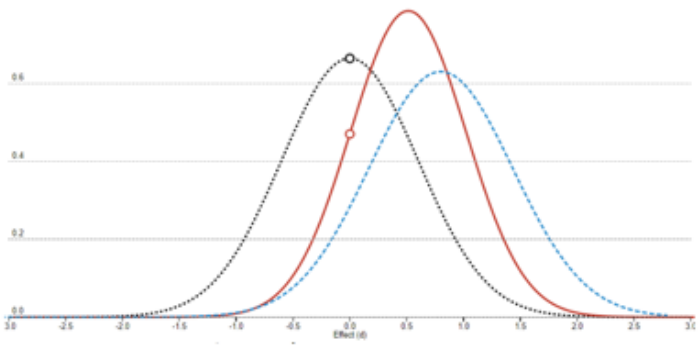
Dicho esto, vale la pena resaltar que los datos tampoco necesitan comprobación del supuesto paramétrico, ya que las estimaciones por intervalos de credibilidad son más precisas y permiten calcular un grado de certidumbre para contrastar hipótesis (De la Fuente *et al.*, 2009). Por otro lado, también ofrece métodos que pueden incorporar información crucial de modo directo en el modelo estadístico, como aspectos cualitativos. Lamentablemente, el enfoque bayesiano no es tan popular como los valores p en ciertas ramas de la ciencia, en parte, debido a que no se ha llegado a acuerdos entre científicos para su uso adecuado (Greenland *et al.*, 2016), sobre todo en el establecimiento de los valores δ a priori o valores esperados.

El beneficio que nuestro trabajo sugiere aprovechar de la estadística bayesiana tiene que ver con el soporte de la hipótesis nula; es bien sabido que, cuando esta no se rechaza, el valor p oscilará entre .05 y 1 usando el enfoque frecuentista. Sin embargo, para el FB se calcula un δ esperado a priori, uno observado y uno agregado (llamado también posterior) para apoyar diferencias de grupo (≠) o una probabilidad de que un grupo tenga un medición mayor o menor que otro (A < B o A > B)).

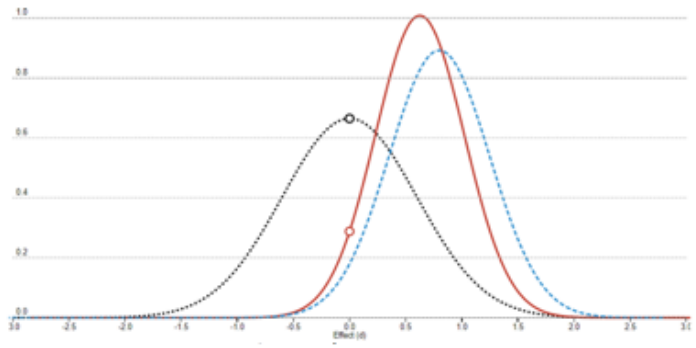
La interpretación se realiza con base en la evidencia; puede ser que la evidencia a favor de una hipótesis sea muy débil y, a medida que se recopilen más datos, el FB favorezca la hipótesis alternativa; de hecho, es probable que, a medida que el tamaño de la muestra aumente, la tendencia para obtener una significancia aparecerá (Rouder *et al.,* 2009), como se observa en la Figura 3, en la que se utilizan el mismo rango δ, mismo valor δ entre grupos y se puede observar cómo, conforme la n aumenta, se alcanza mayor significancia.

No obstante, es necesario resaltar que, a pesar de que la muestra es pequeña, el análisis del FB permite generar grados de certidumbre que evidencian una tendencia. Si esta es favorable cuando la *n* es pequeña, es probable que se mantenga mientras mayor sea el tamaño de la muestra, posiblemente en futuros estudios o intervenciones. Para ello, los valores utilizados deben ser lo más precisos posibles, dado que la probabilidad de que el δ del estudio se presente entre el rango δ empleado depende de la interacción de estos dos más el grado de confianza que se elija para el estudio, y eso obedece, en gran medida, a los objetivos, el tipo de estudio y la disciplina, como discutiremos más delante.

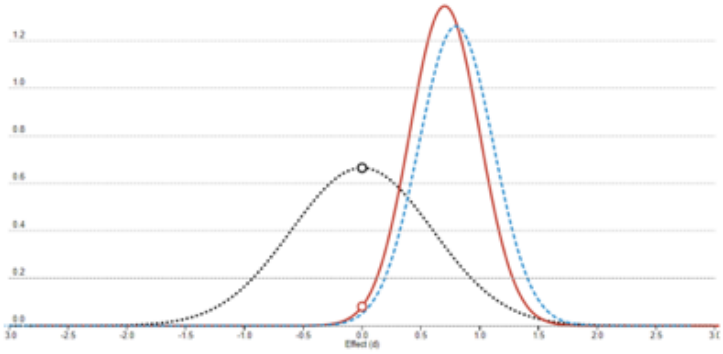
Figura 3. Ejemplos de apoyo de hipótesis en función de la muestra



A)
σδ = 0.6
n = 5
δ = 0.8
FB01 = .707
FB10 = 1.41
Valor p = .206



B)
$^{\sigma}\delta = 0.6$
$n = 10$
$\delta = 0.8$
FB01 = .434
FB10 = 2.31
Valor p = .073



C
$^{\sigma}\delta = 0.6$
$n = 20$
$\delta = 0.8$
FB01 = .122
FB10 = 8.20
Valor p = .011*

Nota: Visualización de inferencia bayesiana configurada para ver la distribución de densidad en un estudio hipotético (δ = 0.8) donde progresivamente aumenta la muestra, los parámetros son los encontrados en la literatura previa (0.6). Se puede observar que las gráficas tienen tres líneas de colores diferentes. La línea negra punteada es la distribución de la densidad de los valores δ previos, también la literatura lo llama evidencia previa necesaria para el FB. La línea roja lisa es la distribución de densidad posterior del valor δ, es decir, la evidencia encontrada en el estudio actual o lo que la literatura llama "posterior". El círculo negro y rojo en las curvas antes descritas representan la probabilidad de 0 bajo la información previa y posterior, respectivamente. Por último, la línea punteada azul indica la relación de probabilidad en función de la densidad de Savage-Dickey que se utilizó para calcular el FB, es la que permite obtener el soporte de la hipótesis. El presente gráfico se hizo con ayuda del visualizador creado por Magnusson (2020)

En el enfoque bayesiano clásico, el grado de certidumbre sería otorgado por el investigador experto, y en el propuesto en este trabajo ese grado de creencia se tomaría de la literatura previa, específicamente de los valores δ reportados (Schönbrodt & Wagenmakers, 2018; Szucs & Ioannidis, 2017; Wetzels *et al.*, 2011). Esta distribución previa será más o menos dispersa en función de la certidumbre de la evidencia; en ese sentido, más que un valor, resulta que es un rango de valores delta (δ) en el cual la evidencia indica si hay mayor probabilidad de obtener un resultado fuera o dentro de este rango, de tal manera que una mayor cercanía (o lejanía) corrobora evidencias ante una hipótesis. Para ello, se necesitan fijar tres valores: la probabilidad de aparición del rango, en la que debemos elegir la posibilidad de cuantificar el grado de confianza que tenemos de que una medición, en este caso la del estudio, se presente en este rango.

Los autores proponemos utilizar la Tabla 1, en la que se hicieron los cálculos del rango δ reportados en la literatura (Szucs & Ioannidis, 2017). Además, es necesario aclarar que la distribución de densidad que usamos es la de Cauchy, dado que posee colas pesadas, en comparación con la

normal, y que es no-informativa, lo cual favorece que los datos influyan en la posterior y, por tanto, en la toma de decisiones (Rouder *et al.*, 2009), y facilita cuantificar la probabilidad de que un valor pueda entrar en un cierto rango de ella (Arnold & Beaver, 2000; Kent & Tyler, 1988).

En la Tabla 1 se encuentran los parámetros de dispersión que deben usarse en la distribución Cauchy a fin de obtener la probabilidad deseada, es decir, que delta se encuentre entre - δ y δ, donde δ es el valor reportado; por ejemplo, en la primera línea donde δ = 0.14, a mayor porcentaje de probabilidad se requiere una menor dispersión, esto es, la densidad estará más concentrada en el origen. Los cálculos se realizaron en el programa R (2020) para determinar opciones, la probabilidad y los rangos δ con los valores reportados con los siguientes códigos, por si algún lector se encuentra interesado en valores específicos "d<-0.14//probas<-c(.2,.3,.4,.5,.6,.7,.8,.9)//vec<-d/qcauchy(1/2+probas/2,0,1)//print(vec)" y "*pcauchy(.23,0,0.07) - pcauchy(-.23,0,0.07)* = 0.20".

Tabla 1. Valores δ reportados en el trabajo de Szucs y Ioannidis (2017) separados por disciplina y porcentaje de probabilidad de aparición dentro de la distribución

	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
	Neurociencia cognitiva	.67	2.062	1.314	.922	.670	.486	.341	.217	.106
Efecto amplio	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

Cuando se obtiene evidencia a favor de la H^0, significa que el rango δ y el valor δ obtenido en el estudio no difieren entre sí, es decir, H^0: δ = 0, por lo que medir el soporte de evidencia con base en los datos para H^0 frente a H^1 requiere valores relativamente distantes entre la priori y lo observado, ya que un probable δ cercano a 0 con valores positivos es tan probable como su correspondiente valor negativo (Rouder *et al.*, 2009). La lógica detrás de la interpretación tiene que ver con el hecho de que la evidencia a favor de una hipótesis tiene "X" veces más probabilidad

de aparecer bajo H^0 frente a H^1, o viceversa.

Otro punto importante que resalta las ventajas del FB es que este permite conocer la evidencia para soportar la H^0 y la hipótesis alterna (H^1) al mismo tiempo, expresada de la siguiente manera: FB_{01}, o bien, si se busca encontrar evidencia relativa de la H^1 sobre la H^0 sería FB_{10}. Expresiones como $p(\mu H^0 > \mu H^1)$ o $p(\mu H^0 < \mu H^1)$ ayudan a los lectores a saber qué tipo de hipótesis se contrasta, a diferencia de la organización del valor p, que provee evidencia en contra de H0; en esta se planea mostrar los grados de certidumbre o probabilidad de contraste. La palabra 'p' implica probabilidad y μ (mu) es la medición del grupo.

PROPUESTA DE REPORTE DE RESULTADOS

En nuestra propuesta de presentación de resultados, los autores sugerimos que dentro de la tabla de mediciones generales y características de la población, identificadas como sociodemográficas en algunas investigaciones, se añada el apartado de intervalos de confianza al 95%, o bien, si no se tiene acceso a ellos, las estimaciones de los intervalos de credibilidad fijados en 2.5% (menos dos desviaciones estándar) y 97.5% (más dos desviaciones estándar), la media de la muestra (por grupo) y su desviación estándar, o si esta no cumple con características paramétricas, entonces reportar las medianas y los rangos intercuartílicos.

En la Tabla 2, que contendría los resultados del estudio, sugerimos que deben incluirse todas las hipótesis sometidas a prueba que tengan relación con la totalidad de las comparaciones de dos en dos; además, siempre deben mostrarse las evidencias a favor de la H^0 y H^1 para que el lector las compruebe y solo reporte las evidencias superiores al nivel de anecdótico. Las variables independientes pueden ser numéricas (discretas o continuas) o cualitativas (nominales u ordinales); eso dependerá del nivel de análisis y la prueba a contrastar, de tal manera que, al incluirse en las tablas, pueda facilitarse el proceso analítico del lector. En el apartado de la probabilidad bayesiana, observamos que se encuentran tres apartados con la siguiente leyenda: "A ≠ B; A < B; A > B)"; este elemento significa la evidencia a favor de una hipótesis según el acomodo de ella.

Tabla 2. Pruebas de diferencias de grupo

Medidas	t	p	δ	ſl	FB^{10}		FB^{01}	
Variable explicada [AB]	.080	.937	.019	.99	μ A ≠ B	.792	μ A ≠ B	1.262*
					μ A > B	.813	μ A > B	1.229*
					μ A < B	.771	μ A < B	1.297*

Nota: A = Medición previa; B = Medición posterior
> la H^1 especifica que la medida A es mayor que la medida B; < la H^1 especifica que la medida A es menor que la medida B
* = Soporte de evidencia significativa

CONCLUSIÓN

Complementar los análisis del valor p con el análisis del tamaño del efecto y del factor Bayes resulta de utilidad, dado que permite tener una mayor certidumbre en la interpretación de resultados y, por lo tanto, el investigador puede discutir con precisión los hallazgos encontrados en su propia investigación. Estos métodos complementarios brindan mejores resultados si se interpretan en conjunto, ya que resulta útil explorar e interpretar los datos a partir de los datos aquí brindados en software que consideran el FB, como el programa R (2020), JASP (2020) o JAMOVI (2020).

Motivamos, entonces, el uso de estas alternativas en aras de reducir las desventajas que podría presentar el uso exclusivo de un método estadístico, así como la integración de su revisión de manera académica o la creación de cursos estadísticos que incluyan estos temas. Asimismo, recomendamos concertar acuerdos científicos de expertos en el tema para poder llegar a convenciones para el uso adecuado de este paradigma. Finalmente, revisar la bibliografía aquí presentada puede contribuir al desarrollo de métodos alternos que permitan adecuarse a los datos encontrados comúnmente en estas disciplinas científicas, especialmente en muestras pequeñas comúnmente encontradas en las disciplinas previamente mencionadas.

REFERENCIAS BIBLIOGRÁFICAS

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

Arnold, B. C. & Beaver, R. J. (2000). The skew-Cauchy distribution. *Statistics & Probability Letters, 49*(3), 285-290. https://doi.org/10.1016/S0167-7152(00)00059-6

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

Benjamini, Y. & Cohen, R. (2017). Weighted false discovery rate controlling procedures for clinical trials. *Biostatistics, 18*(1), 91-104. https://doi.org/10.1093/biostatistics/kxw030

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

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

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

Ellis, P. D. (2010). *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

García, J., Ortega, E. y 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. (1966). Note on rank biserial correlation. *Educational and Psychological Measurement, 26*(3), 623-631. https://doi.org/10.1177/001316446602600307

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 Publications inc.

Hand, D. J. (2012). Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. En G. Cumming (ed.). *International Statistical Review, 80* (2). https://doi.org/10.1111/j.1751-5823.2012.00187_26.x

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

JASP Team (2020). JASP (0.13.1). University of Amsterdam.

John, L. K., Loewenstein, G. & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. https://doi.org/10.1177/0956797611430953

Kent, J. T. & Tyler, D. E. (1988). Maximum likelihood estimation for the wrapped Cauchy distribution. *Journal of Applied Statistics, 15*(2), 247-254. https://doi.org/10.1080/02664768800000029

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

Magnusson, K. (2020). *Interpreting Cohen's d effect size: An interactive visualization.*

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

McGraw, K. O. & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*(2), 361-365. https://doi.org/10.1037/0033-2909.111.2.361

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. & Bar-Anan, Y. (2012). Scientific Utopia: I. Opening Scientific Communication. *Psychological Inquiry, 23*(3), 217-243. https://doi.org/10.1080/1047840X.2012.692215

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

Pashler, H. & Wagenmakers, E. (2012). Editors' Introduction to the special section on replicability in Psychological Science. *Perspectives on Psychological Science, 7*(6), 528-530. https://doi.org/10.1177/1745691612465253

R Core Team (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.*

Reiser, B. & Faraggi, D. (1999). Confidence intervals for the overlapping coefficient: The normal equal variance case. Journal of the Royal Statistical Society: Series D (The Statistician), 48(3), 413-418. https://doi.org/10.1111/1467-9884.00199

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

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

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." *The American Statistician, 73*(sup1), 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

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80. https://doi.org/10.2307/3001968